FACULTAS ARTIS INFORMATICÆ • UNIVERSITAS MASARYKIANA •

Faculty of Informatics
Masaryk University
Czech Republic

Knowledge Management and New Paradigm of Advanced Machine Learning

Habilitation Thesis

Parag Kulkarni

Brno, October 2011

Abstract

Machine Learning and Knowledge Management are the two major research areas and have received great attention in recent years. The intelligent system building cannot be possible without efficient knowledge management and machine learning. Since traditional machine learning approaches do not consider systemic interactions, they fail in complex decision scenarios. The voyage of building smart societies, smart organizations and smart businesses need strong KM and machine learning platform. Increasing competition and the availability of multitudes of information and diminishing traditional entry barriers have left next generation businesses with the requirement to compete on knowledge grounds. This fuelled my research in KM and Machine Learning. The holistic learning based on context and systemic knowledge management are two major research outcomes of my work, This work introduces a new paradigm for machine learning and propose new framework for knowledge management. This work introduces a paradigm of systemic machine learning that can help in building real smart systems and societies. Learning is dynamic scenario, exploration and smart decision making are a few other facets of this work. In this work there is emphasis on building smart businesses and knowledge efficient organizations, multidisciplinary solutions and combining management, engineering and services for creating better value. With the information explosion, numerous information systems and complex decision scenarios there is need of truly smart and learning systems and effective knowledge management. This work definitely builds the platform for next generation smart systems those can create measurable value for societies based on KM and machine learning.

Abstrakt

Strojové učení a management znalostí jsou dvě významné oblasti výzkumu a v posledních letech je jim věnována velká pozornost. Budování inteligentních systémů není možné bez efektivního managementu znalostí a strojového učení. Poněvadž tradiční přístupy ke strojovému učení neuvažují se systémovými interakcemi (v kontextu celého systému), nedaří se jim řešit scénáře komplexního rozhodování. Cesta k budování “smart” společností, “smart” organizací a “smart” businessů potřebuje silnou platformu KM (Knowledge Management) a strojového učení. Zvyšující se konkurence a dostupnost spousty informací spolu se snižováním tradičních bariér vstupu zanechává novou generaci businessů s požadavkem soutěžit na poli znalostí. Právě toto mne pohánělo k výzkumu v oblasti KM a strojového učení. Dva hlavní výstupy mé výzkumné práce jsou holistické učení založené na kontextu a systémový management znalostí. Tato práce představuje nové paradigma strojového učení a navrhuje nový rámec pro management znalostí. Práce uvádí paradigma systémového strojového učení, které může pomoci při budování reálných “smart” systémů a společností.To, že učení spočívá v dynamických scénářích, prozkoumávání a provádění “smart” rozhodnutí, je další stránkou této práce. Práce klade důraz na budování “smart” businessů a znalostně efektivních organizací, multidisciplinárních řešení a kombinování managementu, engineeringu a služeb pro vytvoření lepší hodnoty. Informační explozí, řadou rozličných informačních systémů a komplexních rozhodovacích scénářů je nastolena jasná potřeba skutečně chytrého a učícího se systému a efektivního managementu znalostí. Tato práce rozhodně vytváří platformu pro příští generaci chytrých (“smart”) systémů, které mohou vytvářet znatelnou hodnotu pro společnosti založené na KM a strojovém učení.

Knowledge Management and New Paradigm of Advanced Machine Learning

Parag Kulkarni

Habilitation Thesis

Submitted to

Masaryk University,

Faculty of Informatics

Part I. Preface

Part II. Selected Publications

1. Introduction

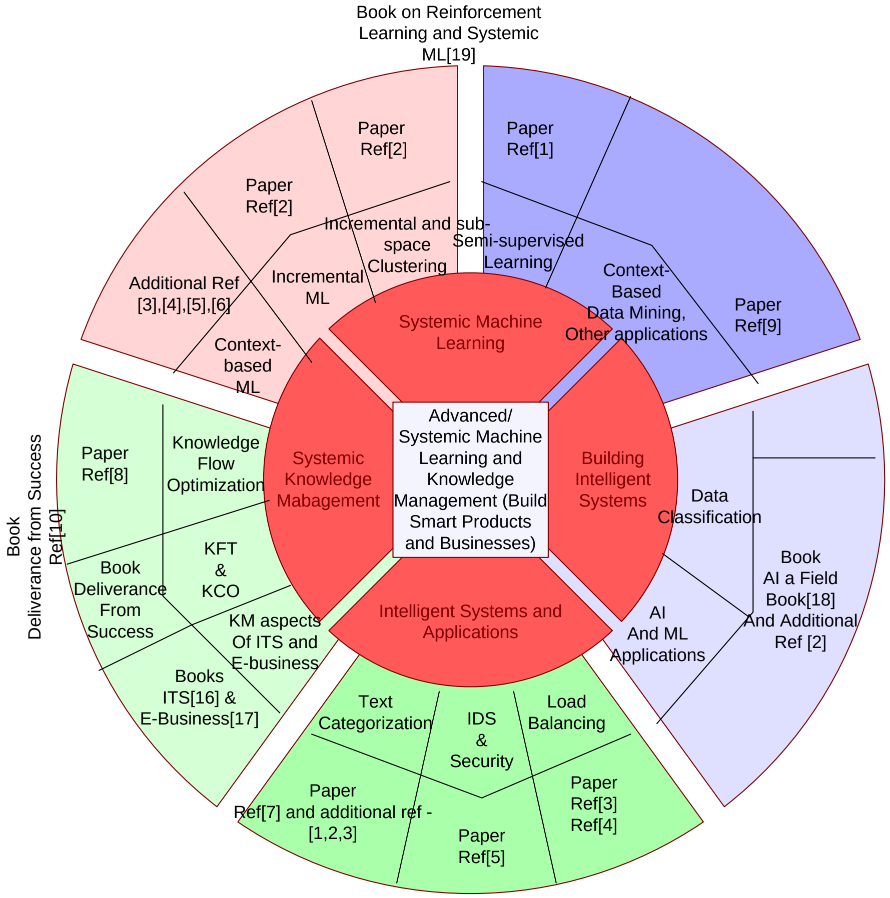
This preface of my thesis presents selected major publications, with both this preface and the referenced publications constituting the thesis as a whole. The preface introduces in thirteen pages the topics investigated as well as research projects undertaken in the areas of Machine Learning, Knowledge Management and Intelligent Systems, with direct references to publications attached in Part II of this thesis. The overall areas of these references are set in the following style:

> **Attached Paper (or Document Not Attached)**
> **x.** Reference to a given paper that is attached as a part of this thesis or (for documents that are not attached) to a standard in which I have had considerable involvement as described in the related text of this preface.

My research in the area of machine learning and knowledge management reveals nine important components:

1. Semi-supervised machine learning
2. Clustering and subspace clustering
3. Intelligent systems and machine learning applications.
4. Intelligent intrusion detection systems
5. Context based machine learning and decision making
6. Text classification and categorization
7. Collaborative and holistic machine learning
8. Classification and intelligent decision-making
9. Knowledge management and deliverance from success

The areas and research papers referred are presented pictorially below to provide overview and inter-relationship among the research topics, publications and projects.


Book on Reinforcement Learning and Systemic ML[19]
Paper Ref[2]
Paper Ref[2]
Additional Ref [3],[4],[5],[6]
Incremental and sub-space Clustering
Incremental ML
Context-based ML
Paper Ref[1]
Semi-supervised Learning
Context-Based Data Mining, Other applications
Paper Ref[9]
Systemic Machine Learning
Advanced/ Systemic Machine Learning and Knowledge Management (Build Smart Products and Businesses)
Systemic Knowledge Mabagement
Building Intelligent Systems
Intelligent Systems and Applications
Book Deliverance from Success Ref[10]
Paper Ref[8]
Knowledge Flow Optimization
KFT & KCO
Book Deliverance From Success
KM aspects Of ITS and E-business
Books ITS[16] & E-Business[17]
Data Classification
Book AI a Field Book[18] And Additional Ref [2]
AI And ML Applications
Text Categorization
IDS & Security
Load Balancing
Paper Ref[7] and additional ref - [1,2,3]
Paper Ref[5]
Paper Ref[3] Ref[4]

There are also four publications (Ref [16, 17, 18, 19]) referred in diagram are provided in the list as other publications but not attached. The discussion in thesis is restricted to the attached publications and project and work very closely related to these publications. The broad area of research comes under knowledge management, management and engineering with reference to machine learning and intelligent systems. In my career I have focused on research in areas of machine learning and knowledge management. There is emphasis on building smart businesses and knowledge efficient organizations. This needs multidisciplinary solutions and combining management, engineering and services for creating better value. This work is also closely associated with concept of SSME. Increasing competition and the availability of multitudes of information and diminishing traditional entry barriers have left next generation businesses with the requirement to compete on knowledge grounds. This fuelled my research in KM and machine learning. Systemic machine learning, systemic knowledge management, intelligent systems and their applications are the four major research areas where research impact can be observed. The representative publications in this area and major areas of contribution are depicted in the diagram.

The information explosion, many information systems and complex decision scenarios created numerous research opportunities. The increasing expectations from intelligence and distributed knowledge and information sources pose new challenges for knowledge management and machine learning. My research focuses on knowledge management and its machine learning aspect with systemic perspective to cope up with this challenge.

2. Semi-supervised Machine Learning

Over the years a lot of evolution has taken place in the field of Machine Learning. Right from simple memorization based learning it has grown to more complex intelligent systems based on inference and complex distributed interactive learning systems. Machine learning paradigms though remained mostly data centric but have evolved over the years. The purpose of learning is to build knowledge, manage knowledge and take right decisions or rather the best possible decisions to create value. *Machine learning* includes various supervised and unsupervised and exploration based learning techniques. Both techniques have their advantages and disadvantages. The objective of machine learning remains to make machine intelligent and handle complex and not so complex scenarios.

In practical dynamic scenario where new scenarios and new information reveals over the time, supervised or unsupervised learning in isolation are incomplete and hence a pressing need of semi-supervised learning is felt. *Semi-supervised learning* is defined as learning from both labelled and unlabeled data. Can system learn from labelled and unlabeled data and that too simultaneously, was the major research question. Semi-supervised learning itself is a new paradigm of learning which is not explored to fullest extent. There was need to explore this paradigm with reference to existing methods used for semi-supervised learning. This was even important from the broader research objective of knowledge management and knowledge augmentation.

In 2008 we worked on exploring different methods for semi-supervised learning and further building a new model for semi-supervised learning. The broader objective was to build adaptive learning system. Ensemble learning has its own pros and cons. In fact ensemble learning is a special case of adaptive learning.

> **NN Pise and P. Kulkarni, "A Survey of Semi-Supervised Learning Methods", Proc. International Conference on Intelligence and Security, 2008. CIS'08, page 30-34, ieeexplore.ieee.org, 2008**

This work further applied to a few specific domains like text classification, context based document mining and semi-supervised learning for intrusion detection.

While working on this approach we worked on evolution of clustering algorithm. Where a lot of work on incremental clustering is carried out. We proposed a new closeness factor based clustering and incremental learning method This method is enhanced and applied to various applications like software lifecycle forecasting, traffic pattern analysis and network security through various research projects. The work undertaken in this regard is building a semi-supervised learning system which can use labelled and unlabeled data. The decision about the qualification of data point for learning was crucial. In this work a new entropy and boosting based algorithm implemented to help in deciding the qualifying data points from learning. Clustering is an act of collecting similar data sets. Most of the clustering techniques require entire data set for clustering. New data can be evolved over a period of time and handling this with re-clustering by the existing methods would be expensive. Further we even proved that some important knowledge built in past is lost in this process. In dynamic scenario, in order to accommodate the new data and to determine the cluster update, incremental semi-supervised learning is required. An incremental clustering approach that is fast, effective and accurate is proposed. The proposed incremental clustering is based on a novel method of closeness factor that identifies the similarity among the data series. It is necessary to consider the behavioral pattern of the data in the learning. The proposed method effectively modifies the knowledge and at the same time maintains the patterns for reuse thus saving the clustering time and helping in decision making. While the learning takes place, the cluster quality is also maintained. The learning outcome is tested on wine, iris and wine quality data sets showed results worth to notice compared with existing methods like k-means, hierarchical as well as COBWEB clustering. From application perspective, the method was applied for decision making and learning in education sector. The aim was to determine the student's capability for a particular course. It was noticed that over a period of time there was a need of shifting or re-assigning a student from a course on the basis of a quarterly evaluation; hence there is a need to incremental learning in a dynamic environment. The proposed method was applied in this problem domain and the results obtained proved to be very useful and accurate, improving the overall grades of the students. This method can prove very useful even in case of other applications where dynamic learning scenario is manifested.

3. Intelligent Subspace Clustering.

Many real-world data sets consist of a very high dimensional feature space. Clustering groups a set of physical or abstract objects into classes containing similar objects. A cluster of data objects can be treated collectively as one group in many applications. As a data mining function, cluster analysis can be used as a standalone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, or to focus on a particular set of clusters for further analysis. Most clustering techniques use the distance or similarity between objects as a measure to build clusters. In high dimensional spaces, subspace clustering algorithms automatically identify lower dimensional subspaces of the higher dimensional feature space in which clusters exist. Further, distances between points become relatively uniform in high dimensional databases. In such cases, density based approaches give better results.

In this research, we propose a new clustering algorithm, Intelligent Subspace Clustering (ISC), which is based on density notion of subspace clustering. The concept of hierarchical clustering is used at dimension level in ISC. That is, we find low dimensional Subspace Clusters first and then try to combine these low dimensional subspace clusters to form a higher dimensional meaningful subspace cluster. At each dimension level, objects will be assigned to subspace clusters using the density notion of clustering.

However, traditional density based clustering approaches use the same input parameters such as Density Threshold (μ) and Epsilon Distance (ε) at different levels of dimensionalities while finding subspace clusters. As a result, the clusters are biased with respect to the dimensionality. To find clusters those are hidden in various subspaces, these parameters have to be set depending upon number of dimensions considered for clustering. ISC determines ϵ i.e. distance dynamically and adaptively at various dimensionality levels, which helps in finding meaningful clusters.

ISC tries to overcome three major limitations of the existing state-of-art techniques. It determines the input parameter such as ϵ i.e. distance at various levels of subspace clustering which helps in finding meaningful clusters. The uniform parameters approach is not suitable for different kind of databases. ISC implements dynamic and adaptive determination of meaningful clustering parameters based on hierarchical filtering approach. Third and most important feature of ISC is the ability of incremental learning and dynamic inclusion and exclusions of subspaces which lead to better cluster formation.

The experimental evaluation proved it to be a successful clustering approach. Thus, it will benefit large application domains such as DNA database and correspondingly in DNA analysis. It will help to identify customer groups, identify frauds or unusual transactions in financial databases and lots of other application areas like information integration system, text-mining, CAD database etc. It will be a very effective, specialized data mining approach/innovation.

> **S Jahirabadkar and P Kulkarni, "ISC–Intelligent Subspace Clustering, A Density based Clustering approach for High Dimensional Dataset", Journal of World Academy of Science, Engineering issue 31 , pp 69-73, 2009http://www.waset.org/journals/waset/v31.php**

The data size and number of parameters in complex data set in real life are astounding. Further the data sets are distributed and are at different locations. Hence there is need of distributed subspace clustering. The further research work is in progress where this work is combined to achieve distributed, incremental and semi-supervised learning for complex and dynamic scenarios.

4. Intelligent Systems and Machine Learning Applications

Over the years AI has become a multi-disciplinary topic and played a key role in decision-making for different applications. These intelligent systems try to make best use of available data. Right from simple rule based systems, pattern based decision systems to very complex inference mechanisms can be used. Initially we worked on a differential technique for load balancing. There are techniques like Sender Initiated Diffusion (SID), Receiver Initiated Diffusion (RID). The new concept of differential measurement based load balancing fetched very good results while balancing the load on network of workstations.

P Kulkarni and I Sengupta, "A new approach for load balancing using differential load measurement", The International Conference on Information Technology: Coding and Computing, ITCC, computer.org, 2000, pp 355-359

This research further led to developing intelligent system for load balancing that can work for different types of networks. In case of variable loads, dynamic scenarios and different network densities the algorithm needs to capture information and decide the optimal node to transfer the load. To handle this problem we proposed algorithm Load Graph Based Transfer Method (LoGTra). This method was working based on load graph and tokens to detect receiver node. This method is further enhanced to Dual Token Based Load Balancing (DTLB) and m-LoGTra. Even some of the complex scenarios can be handled with help of multiple tokens using M-LoGTra. This research helped to over come the typical issues in balancing load in case of low density networks and dynamic load scenarios.

P Kulkarni and I Sengupta, Dual and multiple token based approaches for load balancing, Journal of Systems Architecture, Elsevier, 51/1 pp 95-112, Feb 2005

5. Intelligent Intrusion Detection System

Machine learning and intelligent systems are used in network security to great extent. The Internet as well as local networks are expanding at a speed never seen before. This on one hand, improves the life quality and convenience, on the other hand, provides a platform for network criminals and security breach, i.e. any network in the world is under constant attacks by malicious users internally or externally. Intrusion Detection System (IDS) tries to analyze the behaviour of network and network traffic based on different parameters to detect attacks and intrusions.

One of the important applications of advanced machine learning is in building intelligent security systems. Recently pattern based network security methods have gained importance in addressing network security issues, including network intrusion detection. Intrusion detection is one of the most challenging tasks in network security. Pattern based network security models for intrusion detection is often ineffective in dealing with dynamic changes in intrusion patterns and characteristics. Consequently, unsupervised learning methods have been given a closer look for network intrusion detection. We investigate multiple centroid-based unsupervised clustering algorithms for intrusion detection, and propose a simple yet effective self-labeling heuristic for detecting attack and normal clusters of network traffic audit data. Further we worked on systemic and semi-supervised intrusion detection. It is very tricky, as the overprotective system may result in a number of false alarms while failing to detect intrusion may result in loss of information or data.

VK Pachghare and VA Patole and P Kulkarni, "Self Organizing Maps to Build Intrusion Detection System, International Journal of

Computer Applications, Foundation of Computer Science, USA, Vol.1, No.8, pp 1-4, 2010

This work is motivated by the experience of setting up the open-source Bro NIDS in a medium-scale university environment. When we set it up for the first time, problems arose immediately. The system usually exhausted the host's memory after running for a couple of hours. Hence it missed a significant fraction of the network data as its processing was not able to keep up with the packet stream in real-time. This happened when we experimented with the most popular open-source system. Later we developed advanced boosting algorithm for intrusion detection. This research proposed and implemented a new methodology for intrusion detection. This method makes effective use of unlabelled data. Further this method reduced number of false positive and can help in handling the dynamic scenario due to ability to use unlabelled data effectively. Another important aspect of this work is ability to learn incrementally. This research can help in building better intrusion detection systems overcoming some of the existing problems.

6. Context Based Learning and Decision-Making

Context is the key while building knowledge and making decisions. This context can be proved to be very useful in various learning application. It can be either text categorization or gesture based learning context plays a key role. Recently we have been working on context based machine learning and its different applications.

Ayesha Butalia, AK Ramani,. Parag Kulkarn, "Emotional Recognition and towards Context based Decision", Number 3 - Article 8, International Journal of Computer Applications, Foundation of Computer Science, USA, 2010, pp 42-54

With reference to overall context the emotions in sequence of images can be identified. This overall context can help in building the complete story based on sequence of actions. The most important part in this process is identification of context. This research work includes context predictor and context building. This paper proposes a basic framework for context based gesture recognition. This work is further extended to complex problem of extracting story based on sequence of images. This is very relevant in case of Bharat Natyam a one of the gesture based Indian dance type. Theoretically all the aspects including cues, facial and gesture are important for non verbal communication. Adding facial features to gestures don't cause much difference to the results, rather can include complexity. Adding gesture again to it causes ambiguity, and probably the results go down. Hence facial features are the most important aspect for emotion recognition. Further, ANN, fuzzy are extensively used for this purpose. Adding rough set approach gives better results, and rough set is easy to implement too. Some context based research in these areas have been done using support vector machine and Bayesian networks. Furthermore, context based approach can be further extended with the help of rough-fuzzy approach towards refining artificial intelligence. This research work has many application and we are working on research projects those help in context based decision-making for farmers, context based decision making for software development and context based business decision-making.

7. Text Classification and Categorization

Knowledge needs to be built and augmented for future use. The information comes in different forms. One of the most common forms is unstructured text. Majority of text information is in unstructured form. When the prime source of text is images then the text is to be captured in the form of OCR form. The OCR output may contain many errors. Mining the data and document based on context may required to cope up with this problem. We worked on novel location diagram (LD) based text categorization. This method is further combined with term based inference and a novel classifier is built. We built Web-based intelligent paperless document management where users can collect, store, and share all document from various locations. Also provided are systems and methods requiring minimal data reentry because of data extraction capabilities. The business method to make office paperless was equipped with document classifier which can classify documents based on similarity with Location Diagram (LD) based method. This method builds a unique location map out of each document. The location map is rich in information and describes the relationships among different key words, key phrases and other place holder to make the representation of document in different way. The location map builds the context for the document with reference to metadata.

> **Malaney, P. Kulkarni, etal, Patent "Business method using the automated processing of paper and unstructured electronic documents." No. 7,747,495 USPTO, June 19 2010**

The ability to represent unstructured documents and classify them accurately and efficiently helps in building paperless office. This is very useful in processing of forms, documents, loan papers and insurance and healthcare information.

On front of context based text categorization we focused on text disambiguation, sub-categorization and context building. This work focuses on building context map based on properties of documents. There are 3 research papers those listed but not attached (Additional research publications [1, 2, 3]) depict research effort to make context based learning possible. A new approach for smart text categorization is proposed in these papers. The novelty and originality of the smart text categorization lies in the fact that completely novel algorithms are developed to find bag-of-words. The algorithms categorize 90 % of the test documents of Reuters 21578 into their relevant categories. Also, the 80-90% Non-Reuters documents were categorized into their relevant categories. Similarly, the novel method for unsupervised word sense disambiguation is proposed. A rich variety of techniques have been researched, from dictionary-based methods that use the knowledge encoded in lexical resources, to supervised machine learning methods. Among these, supervised learning approaches have been the most successful algorithms to date. The novel context based, unsupervised/semi-supervised method using knowledge based resource called WordNet is proposed. This resulted in improving accuracy. This work has built a platform for context based learning and that can further allow text based systemic learning and will allow systemic decision-making possible. This work has built foundation for our research projects and work for multi-perspective and context based decision-making. Further this work led to new invention of Context Vector Machine (CVM). The research work on CVM can solve problems in document mining and text as well as image search faced using traditional search paradigm.

8. Collaborative and Holistic Machine Learning:

Learning in isolation has its own limitations. With distributed information sources and a number of intelligent agent there is need of collective and collaborative machine learning. These concepts of holistic learning are very relevant for knowledge management. The work carried out to develop a new knowledge management and learning framework. Mind maps tries to represent the conceptual picture of the system and dependencies as mind portrays it. The holistic machine learning is expected to take into account the complete systems, various parts of the system and their interdependencies. But in real life scenario the complete information is not available. Further information becomes available in bits and pieces and over the time. These dependencies make the overall system view complicated.

To attack this problem in our research we proposed a new concept of semi-constrained influence diagram. This representation tries to accommodate dependencies and partial information. There could be a number of such semi-constrained influence diagrams for a particular decision scenario. A number of such semi-constrained influence diagrams are combined to form a representative influence diagram. These are combined with reference to a particular decision scenario. This not only allows to build holistic scenario but also helps to build a multi-perspective view and helps in multi-perspective decision making.

> **Swamy BK, Kulkarni P, Intelligent Decision Making Based on Pattern Matching and Mind-Maps, WSEAS International Conference on Computers, Athens, pp 493-498, 2006, ISSN 1109-9526 http://direct.bl.uk/bld/PlaceOrder.do?UIN=192046024&ETOC=RN&from=searchengine**

There are two important aspects of this work, one knowledge management while other is machine learning. These two concepts are further explored to build a holistic machine learning system. This concept includes multi-perspective learning, incremental learning and taking into account dependencies among different parts of the system.

9. Classification and Intelligent Decision Making

Classification remains a key part of decision making. I have worked on many research projects on classification. I worked on classifier for numeric data, images and text. This work helped to come up with new concept of classification.

Shankar Lal, Parag Kulkarni, Amarjeet Singh, "Classification Based on Parametric Partitioning of Solution Space" International Journal on Intelligent Systems, 2010, pp 165-191, Vol 19, No 2, 2010

Classification is an important aspect of data mining activity and involves supervised learning along with limited use of other types of learning like unsupervised learning, semi-supervised learning and reinforcement learning. The field of classification is crippled by high time complexity and subdued accuracy. In many algorithms complexity of learning task and/or classification task becomes prohibitive for real life problems in various fields. This results in making compromises in terms of reducing the problem size by removing less relevant parameters. In many cases this introduces loss of information. This work was undertaken by carrying out detailed analysis of network usage behaviour. The work presents an Incremental, Simple, Efficient and Accurate (I-SEA) algorithm to consider contribution of all available relevant parameters while keeping the computational complexity and accuracy within acceptable limits. I-SEA partitions the training sample space during the pre-processing stage as part of learning. Partitioning is done based on parameter values, taking one at a time, resulting into equivalent classes with respect to the parameter.

10. Knowledge Management and Deliverance from Success

I have worked on research on knowledge management aspects of business, IT strategy and its implementation. This work carried out based on extensive study of different start up organizations. The objective of this work was to come up with novel KM based methodology for business. This research work resulted in a new framework for knowledge management. The organizational aspects, knowledge management practices and engineering implementation aspects are considered in this work. The knowledge management research included the business aspects of KM and machine learning and decision making aspects of KM. The work focused on knowledge augmentation and representation. This work introduces a new concept of "Buddhiyogi Organization".

Knowledge management has become the necessary and inevitable part of management techniques in the organization. The pressing need for framework to incorporate knowledge management and even need for taking measures to minimize the knowledge losses in the organization motivated us to research and study knowledge management aspects in the organizations from different perspectives. This work focuses on Knowledge Building and Knowledge Flow Management within and across the organizations. The book further describes new models and methodologies to overcome the issues about inefficiencies in knowledge flow resulting in knowledge losses. The work emphasizes on building a "Buddhiyogi Organization" that possesses pure knowledge and has minimal knowledge loss. This work introduces our two researched concepts in knowledge management Viz KCO – Knowledge Circuit Optimization and KFT – Knowledge Fidelity Treatment. Here KFT is more about purification of knowledge. For this purpose various context-based methods and pattern analysis methods along with traditional knowledge management is used. The process helps in removing the knowledge blocks and allowing it to flow. The knowledge flow analysis is also important aspect of this whole exercise. In this process the objective is to minimize knowledge loss. In next stage there is continuous need for Knowledge Circuit Optimization (KCO). The knowledge circuits are formed within the organization and across the organization. These are formed even in system, subsystems. The knowledge circuits are systemic in nature. Looking at them in isolation and treating them like that may lead to lot of inefficiencies. Further with every change and every new revelation the circuits need to be optimized. In absence of KCO there are chances of developing knowledge blocks. Further various machine learning based methods to analyze these dependencies and interdependencies for KCO are proposed. We would like to claim that KCO is one of the new innovations to handle Knowledge Management in complex scenarios.

We have contributed to the field of knowledge management, Machine Learning, Systemic machine learning with the objective of building intelligent, efficient and smart systems

10. Conclusions and Remarks

Knowledge management is about providing right knowledge, to right person at right moment. This is not possible without research in systemic knowledge management and decision-making. The decision-making for the organizations many times driven just with the aspect of what customer wants. Though it is true that this is one of the most important points in decision-making but there are many factors those need to be considered. Hence I think that systemic decision-making and systemic knowledge management can bring long term and sustainable success to organizations. The research undertaken has again and again revealed the need of holistic intelligent and understanding of systemic dependencies. The paradigm of systemic machine learning has incremental machine learning, multi-perspective machine learning and context based machine learning. Hence it becomes one of the most important milestones of the voyage of making system intelligent. Systemic machine learning can help in building knowledge and making effective decisions. I have a book in press with IEEE that represents the various aspects in systemic machine learning and this paradigm with reference holistic machine learning.

Many times the information is even collected and processed from a particular perspective. There is need to understand impact of decisions and actions on system and that to over the period of time. In absence of knowing this impact decisions and learning cannot handle complex problems. Reinforcement learning does take into account environment and based on response from the environment reinforces to maximize the rewards. But it does not take into account systemic dependencies.

The all research work depicted in thesis like incremental machine learning, systemic machine learning, multi-perspective machine learning and context based learning are the research milestones and steps to build a systemic intelligent framework.

In future we need to look at systemic organization and further research on building buddhiyogi organization. These concepts and research created a foundation for building

"Buddhiyogi Organization". The research focuses on smart organization, smart systems and smart decisions. This research has gone through various stages and has allowed us to come up with a new paradigm for machine learning. This new paradigm of holistic machine learning will definitely help to solve complex and dynamic problems.

The submitted thesis includes a collection of ten papers, where for the co-authored papers it was either my initiation or introduction of the topic to co-authors that lead to the research and publication. In all instances I was either the major contributor or one of the major contributors This thesis documents a large body of work undertaken to enhance existing approaches in knowledge management and machine learning. A few of my ideas are also contributed to commercially successful products those have 100+ installations. These products include IDeaS' e-yield product, Siemens' SECURA+ product, Capsilon's Loan Katalyst and ReasonEdge's RE Modeler. My research ideas contributed to two research project those are in progress.

Following is a list of additional scientific publications on the topics: Knowledge Management and Advanced Machine Learning.

1. *Yashodhara V. Haribhakta, Parag Kulkarni, Pradnya D. Balvir: "Finite State Transducer For Verb Sub-categorization". Proc. IKE 2009: PP 252-257, Las Vegas, USA*
2. *Yashodhara V. Haribhakta, Parag Kulkarni, Balaji A. Bandewar, Dnyaneshwar A. Dewadkar: "A Hybrid Approach For Part Of Speech (POS) Tagging. IKE 2009: 379-383, Las Vegas, USA, July 2009*
3. *Yashodhara V. Haribhakta, Parag Kulkarni: Smart Text Categorization. IKE 2008: 567-570*
4. *BK Swamy, P.Kulkarni, Paper title "Intelligent knowledge flow detection in Organization based on pattern matching", Proc. Third International conference on AI in Engineering and Technology, Kota Kinabalu, Sabah, Malaysia, Nov. 2006*
5. *Prachi Joshi,. Parag Kulkarn, "A Novel Approach for Clustering Based on Pattern Analysis", Number 4, Article 8, International Journal of Computer Applications, 2011*
6. *Parag Kulkarni, Mrudula Kulkarni, "Deliverance from Success", CTC Publication, Mar 2007 (No of pages 200) ISBN 81-89194-08-9*
7. *P. Mulay, P.Kulkarni, "Knowledge Management in Software Development Process" Accepted for Special Issue IJBIS, Inderscience to appear in 2012*
8. V. K. Pachghare, Parag Kulkarni & Deven M. Nikam *"Neural Network Algorithms for Building Pattern Based Network Security", International Conference on VLSI and Communication, 2009, Kerala*
9. *V.K. Pachghare, Parag Kulkarni & Deven M. Nikam "Neural Network Algorithms for Building Pattern"Pages:* 33-37 *International Journal of Scientific Computing, Serial Publications, ISSN 0973-578X*
10. *V. K. Pachghare, Parag Kulkarni & Deven M. Nikam Intrusion Detection System using SELF" International Conferences on Intelligent Agent & Multi-Agent Systems (IAMA09), 22-24 July, Chennai, India*
11. *U. Adsule, V. K. Pachghare and P. Kulkarni , "Efficient Intrusion Detection System Models", 2009 International Conference on Computer and Network Technology, July 24 to 27, 2009 GRT Grand, Chennai, India International Association of Computer Science & Information Technology (IACSIT)*

12. *Pachghare, P. Kulkarni, “An Overview of Intrusion Detection System”, International Journal of Computer Sciences and Engineering Systems, Vol. 3, No. 3 (2009)*
13. *Pachghare, P.Kulkarni, “Self Organizing Maps to Build Intrusion Detection System” International Journal of Computer Applications, Vol. 1, No. 8 (Feb 2010)*
14. *Vivek A. Patole, V. K. Pachghare, Parag Kulkarni, Paper Titled “Performance Analysis of Intrusion Detection System at Training Time Using : AdaBoost Algorithm”, SAM'109th International Conference on Security and Management (July 12-15, 2010, USA) – Communicated and Accepted*
15. *P.Mulay, P. Kulkarni, “An Automated Forecasting Tool (AFT) achieved Clustering Entity Relationship Model”, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.12, December 2008*
16. *P.Mulay, P. Kulkarni, “Support Vector Machine based, project simulation with focus on Security in software development”, IJCSNS International Journal of Computer Science and Network Security, Vol. 8 No. 11 pp. 393-400*
17. *Parag Kulkarni, PK Chande, “IT Strategies-for business”, Oxford University Press, 2008 (No of pages 424) ISBN-10: 0195694473*

Books in Press:

18. *Parag Kulkarni, Sunita Jahirabadkar, Pradip Chande, “E-Business Models”, in press, Oxford University Press, Expected to be published in Dec 2011 (Pages 610)*
19. *Parag Kulkarni, Prachi Joshi, “AI – A Field Book”, In Press, Pearson Education - Expected to publish in winter of 2012, (No of pages 1100 approximately)*
20. *Book "Reinforcement and Systemic Machine Learning for Decision Making" Wiley - IEEE, ISBN: 978-0-470-91999-6 to be published in Dec 2011 Pages 352*

*A detailed list of publications is attached separately

1

NN Pise and P. Kulkarni, "A Survey of Semi-Supervised Learning Methods", Proc. International Conference on Intelligence and Security, 2008. CIS'08, page 30-34, ieeexplore.ieee.org, 2008

A Survey of Semi-Supervised Learning Methods

Mr. Nitin N. Pise
Maharashtra Institute of Technology
Sr. No. 124, Kothrud, Pune, India
nnpise@yahoo.com

Dr. Parag Kulkarni
Capsilon Research Laboratory
Kalyaninagar, Pune, India
parag.kulkarni@yahoo.com

Abstract

In traditional machine learning approaches to classification, one uses only a labelled set to train the classifier. Labelled instances however are often difficult, expensive, or time consuming to obtain, as they require the efforts of experienced human annotators. Meanwhile unlabeled data may be relatively easy to collect, but there has been few ways to use them. Semi-supervised learning addresses this problem by using large amount of unlabeled data, together with the labelled data, to build better classifiers. Because semi-supervised learning requires less human effort and gives higher accuracy, it is of great interest both in theory and in practice. The paper discusses various important approaches to semi-supervised learning such as self-training, co-training(CO), expectation maximization (EM) ,CO-EM, Then how graph-based methods are useful is explained. All semi-supervised learning methods are classified into generative and discriminative methods. But experimental results show that the hybrid algorithm gives better classification accuracy.

Keywords: Semi-Supervised, learning, labelled, unlabelled, data, training, test, expectation maximization , classifier, methods

1. Introduction

Recently, there has been a lot of interest in the continuum between completely supervised and unsupervised learning (Nigam, 2001; Ghani, Jones, &Rosenberg, 2003). Classification [1] [23] is a supervised task, where supervision is provided in the form of a set of labelled training data, each data point having a class label selected from a fixed set of classes (Mitchell, 1997). The goal in classification is to learn a function from the training data that gives the best prediction of the class label of unseen (test) data points. Generative models for classification learn the joint distribution of the data and class variables by assuming a particular parametric form of the underlying distribution that generated the data points in each class.

In many practical learning domains (e.g. text processing, bioinformatics), there is a large supply of unlabeled data but limited labelled data, and in most cases it can be expensive to generate that labelled data. Consequently, semi-supervised learning, learning from a combination of both labelled and unlabeled data [21], has become a topic of significant recent interest. The framework of semi-supervised learning is applicable to both classification and clustering [1]. Supervised classification has a fixed known set of categories, and category-labelled training data is used to induce a classification function. In this setting, the training can also exploit additional unlabeled data, frequently resulting in a more accurate classification function. Several semi-supervised classification algorithms that use unlabeled data to improve classification accuracy have become popular in the past few years, which include co-training (Blum & Mitchell, 1998), transductive support vector machines (Joachims, 1999), and using Expectation Maximization to incorporate unlabeled data into training (Ghahramani & Jordan, 1994; Nigam, McCallum, Thrun, & Mitchell, 2000). Unlabeled data have also been used to learn good distance measures in the classification setting (Hastie & Tibshirani,1996). A good review of semi-supervised classification methods is given in [2, 18].

A variety of semi-supervised techniques have been developed for both generative and discriminative models. A straightforward, generative semi-supervised method is the expectation maximization (EM) algorithm. The EM approach for naive Bayes text classification models is discussed by Nigam [8].He proposed two augmentations to the basic EM to improve classifier accuracy. The first extension introduces a weighting factor which dynamically adjusts the strength of the unlabeled data

Generative [20] semi-supervised methods rely on a model for the distribution of the input data, and can fail either when this model is wrong, or when the structure of the input data is not correlated with the classification task Discriminative semi-supervised methods [22], including probabilistic and non-probabilistic approaches, such as transductive or semi-

978-0-7695-3508-1/08 $25.00 © 2008 IEEE
DOI 10.1109/CIS.2008.204

supervised support vector machines (TSVMs, S3VMs) and a variety of other graph based methods [5, 18] assume high density within class and low density between classes, and can fail when the classes are strongly overlapping.

2. Self-Training

Self-training is a technique commonly used for semi-supervised learning. In self training a classifier is first trained with the small amount of labelled data. The classifier is then used to classify the unlabeled data. Typically the most confident unlabelled points, together with their predicted labels, are added to the training set. The classifier is re-trained and the procedure repeated. Note the classifier uses its own predictions to teach itself. The procedure is also called self-teaching or bootstrapping (not to be confused with the statistical procedure with the same name). The generative model and EM approach of section 2 can be viewed as a special case of 'soft' self-training. One can imagine that a classification mistake can reinforce itself. Some algorithms try to avoid this by 'unlearn' unlabeled points if the prediction confidence drops below a threshold.
Self-training has been applied to several natural language processing tasks.

3.Co-Training and Expectation Maximization (EM)

Co-training (Blum & Mitchell, 1998) (Mitchell, 1999) assumes that (i) features can be split into two sets; (ii) each sub-feature set is sufficient to train a good classifier; (iii) the two sets are conditionally independent given the class. Initially two separate classifiers are trained with the labelled data, on the two sub-feature sets respectively. Each classifier then classifies the unlabeled data, and 'teaches' the other classifier with the few unlabeled examples (and the predicted labels) they feel most confident. Each classifier is retrained with the additional training examples given by the other classifier, and the process repeats.

In co-training, unlabeled data helps by reducing the version space size. In other words, the two classifiers (or hypotheses) must agree on the much larger unlabeled data as well as the labelled data.

Nigam and Ghani (2000) perform extensive empirical experiments to compare co-training with generative mixture models and EM. The use of EM for semi-supervised learning has been proposed in (Miller & Uyar, 1997). More recently, Nigam et al (2000) [8] have studied its application to text classification problems, in which each text class is modelled with a multinomial distribution, corresponding to a naive Bayes classifier. They also consider an extension in which each class is modelled with a mixture of multinomials. Their result shows co-training performs well if the conditional independence assumption indeed holds. In addition, it is better to probabilistically label the entire U, instead of a few most confident data points. They name this paradigm co-EM. Finally, if there is no natural feature split, the authors create artificial split by randomly break the feature set into two subsets. They show co-training with artificial feature split still helps, though not as much as before. Collins and Singer (1999); Jones (2005) used co-training, Co-EM and other related methods for information extraction from text. Balcan and Blum (2006) show that co-training can be quite effective, that in the extreme case only one labeled point is needed to learn the classifier. Zhou et al. (2007) give a co-training algorithm using Canonical Correlation Analysis which also need only one labelled point. Disgust et al. (Dasgupta et al., 2001) provide a PAC-style theoretical analysis.

4. Transductive SVMs

Transductive support vector machines (TSVMs) builds the connection between p(x) and the discriminative decision boundary by not putting the boundary in high density regions. TSVMs an extension of standard support vector machines with unlabeled data. In a standard SVM only the labelled data is used, and the goal is to find a maximum margin linear boundary in the Reproducing Kernel Hilbert Space. In a TSVM the unlabeled data is also used. The goal is to find a labelling of the unlabeled data, so that a linear boundary has the maximum margin on both the original labelled data and the (now labelled) unlabeled data. The decision boundary has the smallest generalization error bound on unlabeled data (Vapnik, 1998). Intuitively, unlabeled data guides the linear boundary away from dense regions.

However finding the exact transductive SVM solution is NP-hard. Major effort has focused on efficient approximation algorithms. Early algorithms (Bennett & Demiriz, 1999) (Demirez & Bennett, 2000) (Fung & Mangasarian, 1999) either cannot handle more than a few hundred unlabeled examples, or did not do so in experiments. The SVM-light TSVM implementation (Joachims, 1999) is the first widely used software.

5. Graph-Based Methods

Graph-based semi-supervised methods define a graph where the nodes are labelled and unlabeled examples in the dataset, and edges (may be weighted) reflect the similarity of examples. Graph methods are nonparametric, discriminative, and transductive in nature. Some of the graph-based methods like mincut, harmonic, local and global consistency, manifold regularization are discussed in [13].

MISSL (Multiple-Instance Semi-Supervised Learning) [6] that transforms any MI problem into an input for a graph-based single-instance semi-supervised learning method that encodes the MI aspects of the problem simultaneously working at both the bag and point levels. MISSL makes use of the unlabeled data. MISSL does not need to repeatedly run the supervised algorithm with different training data each time..

5.1 Regularization by Graph

Many graph-based methods can be viewed as estimating a function f on the graph..One wants f to satisfy two things at the same time: 1) it should be close to the given labels yL on the labeled nodes, and 2) it should be smooth on the whole graph. This can be expressed in a regularization framework where the first term is a loss function, and the second term is a regularizer.

5.1.1 Mincut

Blum and Chawla (2001) pose semi-supervised learning as a graph mincut (also known as st-cut) problem. In the binary case, positive labels act as sources and negative labels act as sinks. The objective is to find a minimum set of edges whose removal blocks all flow from the sources to the sinks. The nodes connecting to the sources are then labeled positive, and those to the sinks are labeled negative. Equivalently mincut is the *mode* of aMarkov random field with binary labels (Boltzmann machine).

Many of these graph based methods [6] do not scale well since the cost is cubic in the size of the graph. More recently, Zhu and Lafferty (2005) convert the original graph into a much smaller backbone graph by applying a mixture model to L [U. Since the computational cost is very dependent on the size of the graph, learning on this smaller graph is much more efficient.

6. Multi-view Learning

Co-training [8] and related multi-view learning methods [4] assume that multiple classifiers are trained over multiple feature views (splits) of the same labelled examples. As capacity control, these classifiers are encouraged to make the same prediction on any unlabeled example. However, multiple feature views often do not naturally exist in practice, and these methods resort to artificially creating random feature splits.

7. Generative Discriminative Hybrids

In many approaches, parameters are separated into two subsets, one of which is trained discriminatively and the other generatively. Raina, present a model for document classification in which documents are split into multiple regions. For a newsgroup message, regions might include the header and the body. In this model, each region has its own set parameters that are trained generatively, while the parameters that weight the importance of each region in the final classification are trained discriminatively. Experimental results show that this hybrid algorithm gives better classification accuracy than either naive Bayes or logistic regression alone.

8. Semi-supervised subspace learning

Semi -supervised subspace learning algorithm [7] incrementally learns an adaptive subspace by preserving the semantic structure of the image space, based on user interactions in a relevance feedback driven query-by-example system. The algorithm is capable of accumulating knowledge from users, which could result in new feature representations for images in the database so that the system's future retrieval performance can be enhanced. Experiments on a large collection of images have shown the effectiveness and efficiency of the algorithm.

9. Semi- Supervised Classification with Markov Random Fields

[9] describes the use of Boltzmann machines in semi-supervised classification. It treats the labelled/ unlabeled dataset as a Markov random field, and derive a Boltmann machine learning algorithm for it to learn the feature weights, label noise and labels for unlabeled data at all once. Some Markov Chain Monte Carlo Methods needed for learning are presented in [9].

10. With Label Propagation

[10] gives a simple iterative algorithm, label propagation, to propagate labels through the dataset along with high density areas defined by unlabeled data. The problem is formulated as a particular form of label propagation, where a node's labels

propagate to neighbouring nodes according to their proximity.

11. Combining Active Learning with Gaussian Fields and Harmonic Functions

Active learning [16] and semi-supervised learning are combined under a Gaussian random field model. Labelled and unlabeled data are represented as vertices in a weighted graph, with edge weights encoding the similarity between instances. The semi-supervised learning problem is then formulated in terms of a Gaussian random field on this graph, the mean of which is characterized in terms of harmonic functions. Active learning is performed on top of the semi supervised learning scheme by greedily selecting queries from the unlabeled data to minimize the estimated expected classification error (risk); in the case of Gaussian fields the risk is efficiently computed using matrix methods. [11] presents experimental results on synthetic data, hand written digit recognition, and text classification tasks. [17] explains the use of simple Gaussian field and harmonic function algorithm on the FreeFoodCam [17] dataset.

12. Other Methods

[12] presents an algorithm based on convex optimization for constructing kernels for semi-supervised learning. The kernel matrices are derived from the spectral decomposition of graph Laplacians, and combine labelled and unlabeled data in a systematic fashion. A nonparametric kernel approach is presented that incorporates order constraints during optimization.

13. Conclusion and Future Work

To fully utilize the potential of unlabeled data, the abilities and limitations of existing methods [15] must be understood. Semi-supervised learning is still a very young discipline, more work is needed to develop the field and to gain deeper understanding how it relates to and compares with established classification methods such as K-NN and SVM.

The first author is doing research in learning methods and proposes a new adaptive learning methodology where learning can be either supervised, unsupervised or semi-supervised and it can switch from supervised to unsupervised or semi-supervised as per the scenario by considering all perspectives.

14. Acknowledgements

We would like to thank Prof. Anil D. Sahasrabudhe, Director and Prof. A.A. Sawant, the Head of Computer Engineering and Infornation Technology who motivated and provided facilties for the research. We are very grateful to the anonymous referees for their valuable suggestions to improve this paper. We are also thankful to the Board of College and University Development, University of Pune, India for promoting and funding research activities in University of Pune. We also thank for all the encouragement, motivation, support provided by Dr. V. D. Karad, Director, MIT, Pune, Dean Prof. S.S. Darade Patil, Prof. P. B. Joshi, Principal Dr. L. K. Kshirsagar. We are thankful to Infosys Technologies for partial sponsorship.

15. References

[1]Sugato Basu, Mikhail Bilenko, Raymond J. Mooney, “A Probabilistic Framework for Semi-supervised Clustering”, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004),

[2] Xiaojin Zhu, “Semi-Supervised Learning Literature Survey”, Computer Sciences TR 1530, University of Wisconsin – Madison, Dec 2007

[3] Pedro J. Moreno, Shivani Agarwal, “An Experimental Study of EM-Based Algorithms for Semi-Supervised Learning in Audio Classification”, Proceedings of the ICML-2003 Workshop on The Continuum from Labelled to Unlabeled Data, Washington DC, 2003.

[4] Gregory Druck, Chris Pal, Xiaojin Zhu, Andrew McCallum, “Semi-Supervised Classification with Hybrid Generative/Discriminative Methods”, KDD’07, August 12–15, 2007, San Jose, California, USA.

[5] Xiaojin Zhu,” Semi-Supervised Learning with Graphs”, Doctoral Thesis, Carnegie Mellon University, May 2005.

[6] Rouhollah Rahmani, Sally A. Goldman, “MISSL: Multiple-Instance Semi-Supervised Learning”, 23 rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

[7] Xiaofei He, “Incremental Semi-Supervised Subspace Learning for Image Retrieval”, ACM MM’04.

[8] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. “Text classification from labeled and unlabeled documents using EM”, Machine Learning, 1–34, Kluwer Academic Publishers, Boston, 2000.

[9] Xiaojin Zhu, Zoubin Ghahramani, “Towards Semi-Supervised Classification with Markov Random Fields”, School of Computer Science, Carnegie Mellon University, USA, June 2002.

[10] Xiaojin Zhu, Zoubin Ghahramani, "Learning from Labeled ans Unlabeled Data with Label Propogation",m Field", School of Computer Science, Carnegie Mellon University, USA, June 2002.

[11] Xiaojin Zhu, John Lafferty, Zoubin Ghahramani, "Combining Active Learning and Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions," Proceedings of the ICML-2003 Workshop on The Continuum from Labelled to Unlabeled Data, Washington DC, 2003.

[12] Xiaojin Zhu, Jaz Kandola, John Lafferty, Zoubin Ghahramani, "Nonparametric Transforms of Graph Kernels for Semi-Supervised Learning", School of Computer Science, Carnegie Mellon University, USA, Gatsby Computational Neuroscience Unit, UK.

[13] Xiaojin Zhu ,"Semi-Supervised Learning Tutorial", International Conference on Machine Learning, 2007.

[14] X. Zhu, Z. Ghahramani and J. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, 20th International Conference on Machine Learning, 2003.

[15] Ira Cohen, Fabio G. Cozman, Nicu Sebe, Marcelo C. Cirelo,Thomas S. Huang, "Semi Supervised Learning of Classifiers: Theory, Algorithms, and Their Application to Human-Computer Interaction", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 26, No. 12, December 2004, pp 1553-1567.

[16] Marko Grobelnik, Dunja Mladenic, "Text Mining, Link Analysis and Semantic Web", Jozef Stefan Institute Ljubljana, Slovenia, KDD 2007.

[17] Maria-Florina Balcan, Avrim Blum, Patrick Pakyan Choi, John Lafferty, Mugizi Robert Rwebangira, Xiaojin Zhu ,"Person Identification in Webcam Images: An Application of Semi-Supervised Learning", Proc. of the 22 nd ICML Workshop on Learning with Partially Classified Training Data, Bonn, Germany, 2005.

[18] Pavan Kumar Mallapragada, Rong Jin, Anil K. Jain, and Yi Liu, "SemiBoost: Boosting for Semi-supervised Learning", Department of Computer Science and Engineering, Michigan State University.

[19] Rie Kubota, Ando, Tong Zang, "A High-Performance Semi-Supervised Learning Method for Text Chunking", IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A.

[20] Christopher M. Bishop, Julia Lasserre, "Generative or Discriminative? Getting the Best of Both Worlds", BAYESIAN STATISTICS 8, pp. 3-24.,Oxford University Press, 2007

[21] Massih-Reza Amini, Patrick Gallinari, "The Use of Unlabeled Data to Improve Supervised Learning for Text Summarization", *SIGIR'02*, August 11-15, 2002, Tampere, Finland.

[22] Rie Kubota, Ando, Tong Zang, "Two-view Feature Generation Model for Semi-supervised", Proceedings of the 24 th International Conference on Machine Learning, Corvallis, OR, 2007

[23]Sugata Basu, "Semi-supervised Clustering: Probabilistic Models,Algorithms and Experiments", Doctoral Thesis, University of Texas, Austin , August 2005

2

S Jahirabadkar and P Kulkarni, ."ISC–Intelligent Subspace Clustering, A Density based Clustering approach for High Dimensional Dataset", Journal of World Academy of Science, Engineering issue 31 , pp 69-73,
2009http://www.waset.org/journals/waset/v31.php

ISC – Intelligent Subspace Clustering, A Density based Clustering approach for High Dimensional Dataset

Sunita Jahirabadkar, and Parag Kulkarni

***Abstract*—**Many real-world data sets consist of a very high dimensional feature space. Most clustering techniques use the distance or similarity between objects as a measure to build clusters. But in high dimensional spaces, distances between points become relatively uniform. In such cases, density based approaches may give better results. Subspace Clustering algorithms automatically identify lower dimensional subspaces of the higher dimensional feature space in which clusters exist. In this paper, we propose a new clustering algorithm, ISC – Intelligent Subspace Clustering, which tries to overcome three major limitations of the existing state-of-art techniques. ISC determines the input parameter such as ϵ – distance at various levels of Subspace Clustering which helps in finding meaningful clusters. The uniform parameters approach is not suitable for different kind of databases. ISC implements dynamic and adaptive determination of Meaningful clustering parameters based on hierarchical filtering approach. Third and most important feature of ISC is the ability of incremental learning and dynamic inclusion and exclusions of subspaces which lead to better cluster formation.

***Keywords*—**Density based Clustering, High Dimensional Data, Subspace Clustering, Dynamic Parameter Setting.

I. Introduction

THE process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. The dissimilarities between objects are accessed based on the attribute values describing the objects. A cluster of data objects can be treated collectively as one group in many applications [4]. As a data mining function, cluster analysis can be used as a stand alone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis.

Typical clustering methods compute similarities between objects based on an entire set of selected attributes. Many of the real world datasets consist of objects modeled by high dimensional data. Each object is described by hundreds of attributes. For instance, In many Computer Vision applications, such as motion segmentation, face clustering with varying illumination, Pattern Classification, Temporal Video Segmentation etc., image data is huge-dimensional. Other examples for high-dimensional feature vectors representing complex objects can be found in the area of Molecular Biology [13], CAD databases etc. However, when the number of measured attributes is large, it may be the case that two given groups differ at only a subset of the measured attributes, and so only a subset of the attributes are "relevant" to the clustering. In such cases, traditional clustering methods may fail because the differences between any two groups, averaged over all the attributes, are small [1]. Subspace Clustering algorithms are clustering algorithms that look for and build clusters not necessarily in the whole space, but also in subspaces of the attributes. Formally, a subspace cluster can be defined as a pair (Subspace of the feature space, Subset of data points).

Generally, the subspace clusters may be hierarchically nested, i.e. several subspace clusters of low dimensionality may together form a subspace cluster of higher dimensionality. Detecting such relationships of subspace clusters is obviously a hierarchical problem. Fig. 1 illustrates a simple example of a hierarchy of subspace clusters in a 3-dimensional feature space: the 2-dimensional clusters C1 and C2 are embedded within cluster C3 which is a 3-dimensional cluster.

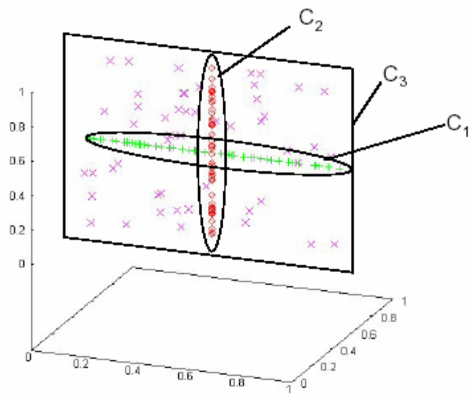


Fig. 1 Hierarchy of Subspace Clusters

The resulting hierarchy is different from the result of a conventional hierarchical clustering algorithm, e.g., a Dendrogram. In a Dendrogram, each object is placed in a singleton cluster at the leaf level, whereas the root node represents the cluster consisting of the entire data set.

This concept of hierarchy will be used at dimension level in ISC i.e. we find low dimensional Subspace Clusters first and

S. Jahirabadkar is with the Computer Engineering Department of Cummins College of Engineering, Pune University, Pune (India) as Asst. professor (e-mail: sunita.jahirabadkar@cumminscollege.in).

P. Kulkarni is with Capsilon Research Labs, Pune (India) as Chief Scientist and Director – Research. He is Alumnus of IIT and IIM (e-mail: parag.kulkarni@capsilon.com).

then try to combine these low dimensional Subspace Clusters to form a higher dimensional meaningful subspace cluster. At each dimension level, objects will be assigned to subspace clusters using the density notion of clustering. Thus research area of our paper, we call as, "Density based Hierarchical Subspace Clustering". ISC determines meaningful clustering parameters dynamically and adaptively based on hierarchical filtering approach. Thus the most important feature of ISC is the ability of incremental learning and dynamic inclusion and exclusions of subspaces which lead to better cluster formation.

The remainder of this paper is organized as follows. We start by reviewing the related work in Density based Subspace Clustering approaches for clustering High Dimensional Dataset in section II. In Section III, we detail our new clustering algorithm, ISC – Intelligent Subspace Clustering. Section IV discusses the results along with future research direction and the conclusion is in Section V.

II. Related Work

Subspace Clustering is a very important technique to seek clusters hidden in various subspaces (Dimensions) in a very high dimensional database. There are very few approaches to Subspace Clustering. These approaches can be classified by the type of results they produce. The first class of algorithms allows overlapping clusters, i.e., one data point or object may belong to different clusters in different projections e.g. CLIQUE [5], ENCLUS [6], MAFIA [7], SUBCLU [8] and FIRES [9]. The second class of subspace clustering algorithms generate non-overlapping clusters and assign each object to a unique cluster or noise e.g. DOC [10] and PreDeCon [11].

The first well-known Subspace Clustering algorithm is CLIQUE, CLUstering in QUEst [5]. CLIQUE is a grid-based algorithm, using an apriori-like method which recursively navigates through the set of possible subspaces. A slight modification of CLIQUE is the algorithm ENCLUS, Entropy based CLUStering [6]. A more significant modification to CLIQUE is MAFIA, Merging of Adaptive Finite IntervAls [7], which is also a grid-based but uses adaptive, variable sized grids in each dimension. The major disadvantage of all these techniques is caused by the use of grids. Grid-based approaches are based on positioning of grids. Thus clusters are always of fixed size and depend on orientation of grid. Density based Subspace Clustering is one more approach. The first of this kind, DOC proposes a mathematical formulation regarding the density of points in subspaces. But again, the density of subspaces is measured using a hypercube of fixed width w, so it has the similar problems [10].

Another approach SUBCLU (density connected SUBspace CLUstering) is able to effectively detect arbitrarily shaped and positioned clusters in subspaces [8]. Compared to the grid-based approaches SUBCLU achieves a better clustering quality but requires a higher runtime. SURFING is one more effective and efficient algorithm for feature selection in high dimensional data [12]. It finds all subspaces interesting for clustering and sorts them by relevance. But it just gives relevant subspaces for further clustering. The only approach which can find subspace cluster hierarchies is HiSC [14]. However it uses the global parameters such as Density Threshold (μ) and Epsilon Distance (ε) at different levels of dimensionalities while finding subspace clusters. Thus its results are biased with respect to the dimensionality.

To find clusters those are hidden in various subspaces, parameters like ϵ – distance (epsilon distance) has to be set depending upon number of dimensions considered for clustering. ISC determines ϵ – distance dynamically and adaptively at various dimensionality levels, which helps in finding meaningful clusters. The most important feature of ISC is the ability of incremental learning and dynamic inclusion and exclusion of subspaces which lead to better cluster formation.

III. Intelligent Subspace Clustering Algorithm - ISC

Algorithm ISC is based on the density notion of hierarchical subspace clustering. Thus the aim of ISC is to detect clusters of lower dimensional subspaces contained in clusters of higher dimensional subspaces. Our general idea is to evaluate whether two points are contained in a common subspace cluster, using the density based clustering approach. For example, two points that are in a 1-d subspace cluster may be contained in a 2-d cluster that consists of the two 1-d projections.

Let D be a data set of n-normalized feature vectors of dimensionality d. Let $A = \{A_1,\ldots,A_d\}$ be the set of all attributes A_i of D. Any subset $S \subseteq A$ is called a Subspace. The projection of an object $p \in D$ into a subspace $S \subseteq A$ is denoted by $\pi_S(p)$.

For any $\varepsilon \in R^+$ the ε -neighborhood of an object $p \in DB$ is denoted by $N_\varepsilon(p)$. It can be defined as all those objects, the distance between some object p and other object is less than ε. The parameter μ specifies the density threshold, initially as an input parameter. Based on these two parameters dense regions can be specified with the help of core objects. An object $p \in DB$ is called core object in D if its ε - neighborhood in D contains at least μ objects.

Usually clusters contain several core objects located inside a cluster.

The aim of ISC is to detect clusters of lower dimensional subspaces contained in clusters of higher dimensional subspaces. The general idea is to evaluate whether two points are contained in a common subspace cluster. For example, two points that are in a 1-d subspace cluster may be contained in a 2-d cluster that consists of the two 1-d projections.

For example, in Fig. 2 each of the two lines forms a 1-dimensional subspace cluster. The plane is a 2-dimensional subspace cluster and it includes the two 1-dimensional subspace clusters. In order to detect the lines, a search for 1-dimensional subspace clusters has to be applied, whereas in order to detect the plane, a search for 2-dimensional subspace clusters has to be performed. Moreover, searching subspace clusters of different dimensionality is basically a hierarchical problem, because the information that a point belongs to some i-dimensional subspace cluster that is itself embedded into an j-dimensional subspace cluster where i < j, can only be uncovered by using a hierarchical approach.

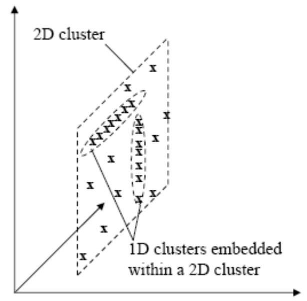


Fig. 2 1-D clusters embedded within 2-D cluster

The algorithm ISC starts at 1-dimension and will iterate till d-dimensions. At each dimension level, it applies a density based clustering. The dimensions are going to form a tree-like structure with single dimensional clusters at leaf nodes and d-dimensional clusters at root node. It uses the concept that several subspace clusters of low dimensionality may together form a larger subspace cluster of higher dimensionality [14]. The parameter such as ε - Epsilon distance will be decided differentially and adaptively depending upon the subspace dimensionality and the minimum and maximum distances among the data points at that level. If this distance has to be measured between 2 subspace clusters where Subspace Dimensionality $\neq 1$, we use single-link linkage method which tries to find the minimum distance between two clusters C_i and C_j as the distance between the two closest objects within those clusters [2]. The strategy will be to merge those points into common clusters which will have Subspace Distances smaller than ε - distance, adaptively calculated at that level. A hierarchy of subspace clusters will be build accordingly.

None of the previously proposed algorithms uses density based approach for subspace clustering to detect such hierarchies of nested Subspace Clusters using differential input parameters.

ISC can be divided into 4 major tasks.

1. Application of algorithm RANK to find dimensions in the descending order of relevance / interestingness.
2. Application of density based clustering on all 1-d data (starting with highest ranked attribute). {Algorithm DBSCAN}
3. Continue combining smaller dimensional clusters to form higher dimensional clusters by selecting next ranked dimensions to add with.
   - Density approach of clustering to detect clusters even at any Subspace Dimensionality > 1
   - The ε – distance will be the distance between clusters. It will be called as Subspace Distance.
   - Single-Link linkage method to find the distance between two clusters (set of points).
   - Combine those clusters whose Subspace Distance is less than ε – distance.
4. At each dimension level –
   - Depending upon minimum and maximum distance between Subspace Distances, the parameters such as ε – distance and threshold (μ) will be changed to accommodate correct points in the clusters. This leads to adaptive parameter setting.
   - Those Subspaces which do not contain any core points are directly removed which makes it more efficient.

A. *Algorithm RANK*

Algorithm RANK measures the "Interestingness" of a dimension with respect to no. of data points taking part in building subspace clusters along that dimension. For each object, we will first compute a Subspace Preference Vector (p_vector). This vector will contain "1" for all those dimensions for which the point can become part of the cluster along that dimension and "0" otherwise. For this we search ε - neighborhood of the point in that dimension. If it contains no. of objects higher than μ, the point can participate in cluster along that dimension. So p_vector contains "1" in that dimension position. All these sets are then compared and the attributes with highest number of "1"s are ranked as most interesting. Subsequently in descending order we tag remaining dimesnions with corresponding rankings. The pseudo code of algorithm is given in Fig. 3.

```
Algorithm RANK (Database D, real μ)
// Finds dimensions in descending order of interestingness
          For each p ϵ D DO
     Initialize $p\_vector_{p(i)}(v^1,\ldots,v^d)$ ; // preference vector
 For dim = 1,d DO              for  each data point
                $N^{\{ai\}}_{\varepsilon}(p) = \{x \mid DIST_{\{ai\}}(p, x) \leq \varepsilon\};$  // find no. of objects
                S = sum(x);                        in ϵ-neighborhood
                if s > μ then
                    $p\_vector_{p(i)} = p\_vector_{p(i)} \cup 1;$
                Else
                    $p\_vector_{p(i)} = p\_vector_{p(i)} \cup 0;$
                End if
            End For       // all dimensions
          End For         // all data points

          For dim = 1,d DO
            For each p ϵ D DO
                 Sum $p\_vector_{p(i)}[d];$
            Update vector p_vector;  // vector with sorted dims
          End For
End
```

Fig. 3 Algorithm to find interestingness of dimensions

Once the vector p_vector is generated, we check for any dimension which does not contain any core objects. These dimensions indicate least importance in forming Subspace Clusters along those dimensions. We remove such dimensions from the vector p_vector to reduce the computations and thus to improve the efficiency of our algorithm.

B. *Algorithm ISC*

First we apply algorithm RANK which gives a list of dimensions in the descending order of Interestingness.

Then we apply DBSCAN [3], the robust density based clustering algorithm with input parameters μ (Density Threshold) to one dimensional dataset. DBSCAN also needs ε – distance as another parameter which we calculate by using the dissimilarity matrix formed with that dimension and the minimum and maximum distance found there in. For this we will start with that dimension which is having highest rank given by algorithm RANK. DBSCAN is able to detect arbitrarily shaped clusters by one single pass over the data. DBSCAN checks the ε-neighborhood of each point p in the database. If $N_\varepsilon(p)$ of an object p consists of at least µ objects, i.e., if p is a core object, a new cluster C containing all objects of $N_\varepsilon(p)$ is created. Then, the ε - neighborhood of all points q ∈ C which has not yet been processed is checked. If object q is also a core object, the neighbors of q which are not already assigned to cluster C are added to C and their ε -neighborhood is checked in the next step. This procedure is repeated until no new point can be added to the current cluster C. Then the algorithm continues with a point which has not yet been processed, trying to expand a new cluster.

The algorithm ISC (see Fig. 4) will start at 1-dimension and will iterate till d-dimensions according to the ranks stored in vector p_vector. At each dimension level, we first calculate ε - distance to be used at that dimensionality level. Then DBSCAN will be applied considering these parameters. Again those Subspace Clusters with null core objects will be removed to reduce the computations.

We will need to define Subspace Distance which will be the distance between two subspace clusters to be combined in higher dimensions. For this Single-Link method [2] will be used to find Subspace Distance between two subspace clusters. It is the distance between any two clusters C_i and C_j as the minimum distance between two closest objects.

```
Algorithm ISC (Database D, real µ )
    RANK(D, µ );
    Remove dimensions with null core objects;
    Apply DBSCAN on RANK(1); // on the highest ranked dim
    For dim = 2,d DO // in vector p_vector
       Calculate ϵ – distance;
       Apply DBSCAN on RANK(1)+dim(i);
                  Remove Subspaces with null core objects;
    End For;
End;
```

Fig. 4 Algorithm ISC

C. Input Parameters

ISC applies DBSCAN which is a robust density based clustering at each level of dimensionality to find Subspace Clusters. DBSCAN needs two input parameters ε - distance and µ – Density Threshold used to define:

1) An ε - neighborhood of a point p

2) A core object (a point with a neighborhood consisting of more than µ objects)

3) A concept of a point q, density-reachable from a core object p (a finite sequence of core objects between p and q, such that each next belongs to an ε - neighborhood of its predecessor)

4) A density-connectivity of two points p, q (they should be density-reachable from a common core object) from different subspace clusters.

But ISC takes only µ as input parameter. The parameter ε specifies the locality of the neighborhood from which the local Subspace Distance of each point in Subspace Clusters is determined. Obviously, this parameter is rather critical because if it is chosen too large, the subspace image may be blurred by noise points, whereas if it is chosen too small, there may not be a clear subspace preference observable, although existing. Further, as the dimensions goes on increasing, the distance between data points already become larger. So we may need to increase it a little bit at higher level to accommodate high level subspace clusters. ISC successfully identifies this parameter at various levels of dimensions when tested with scientific data with nearly 30-40 dimensions. This gives the ability of incremental learning and dynamic inclusion and exclusions of subspaces which lead to better cluster formation.

IV. Experimental Evaluation

In this section, we present a broad evaluation of ISC. We implemented ISC as well as the two basic methods DBSCAN and RANK in JAVA. All experiments were run on Microsoft Windows XP platform with a 2.0 GHz CPU and min 2.0 GB RAM. We evaluated ISC using several synthetic datasets. We tried to vary the dimensions of the data sets from 8 to 40, the number of clusters from 2 to 5, the subspace dimensionality of the clusters from 2 to 8. The density of the clusters was chosen randomly. All attribute values were normalized between 0 and 10.

In all experiments, ISC could generate Subspace Clusters hidden in the data.

V. Conclusion

In this paper, we first motivated the need for Hierarchical Subspace Clustering for very high dimensional dataset. Later we proposed a new approach ISC (Intelligent Subspace Clustering) which uses the density based clustering to find Subspace Clusters embedded in higher dimensional clusters. By determining the ϵ – distance parameter dynamically and adaptively at each dimension level, it allows for incremental learning by allowing modifying parameters adaptively. This leads to better cluster formation at higher dimensionality. The experimental evaluation proved it to be a successful clustering approach. Thus, it will benefit large application domains such as DNA database and correspondingly in DNA analysis, It will help to identify customer groups, identify frauds or unusual transactions in financial database and lots of other application areas like information integration system, text-mining, CAD database etc. It will be a specialized, very effective Data Mining tool.

References

[1] Michael Steinbach, Levent Ertöz and Vipin Kumar, “The Challenges of Clustering High Dimensional Data”, [online]. Available : http://www-users.cs.umn.edu/~kumar/papers/high_dim_clustering_19.pdf

[2] R. Sibson. SLINK, “An optimally efficient algorithm for the single-link cluster method”, The Computer Journal, 16(1):30{34,1973.

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with Noise”, In Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR, 1996.

[4] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufman, 2001.

[5] R. Agrawal, J. Gehrke, D. Gunopulos, and. Raghavan, “Automatic subspace clustering of high dimensional data for data mining applications”, In Proceedings of the SIGMOD Conference, Seattle, WA, 1998.

[6] C. H. Cheng, A. W.-C. Fu, and Y. Zhang, “Entropy-based subspace clustering for mining numerical data”, In Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Diego, CA, pages 84{93, 1999.

[7] S. Goil, H. Nagesh, and A. Choudhary, “MAFIA: Efficient and scalable subspace clustering for very large data sets”, Technical Report CPDC-TR-9906-010, Northwestern University, 1999.

[8] K. Kailing, H.P. Kriegel, and P. Kroger, “Density-connected subspace clustering for high-dimensional data”, In Proceedings of the 4th SIAM International Conference on Data Mining (SDM), Orlando, FL, 2004.

[9] H.P. Kriegel, P. Kroger, M. Renz, and S. Wurst, “A generic framework for efficient subspace clustering of high-dimensional data. In Proceedings of the 5th International Conference on Data Mining (ICDM), Houston, TX, 2005.

[10] C. M. Procopiuc, M. Jones, P. K. Agarwal, and T. M. Murali, “A Monte Carlo algorithm for fast projective clustering. In Proceedings of the SIGMOD Conference, Madison, WI, 2002.

[11] C. Bohm, K. Kailing, H.P. Kriegel, and P. Kroger, “Density connected clustering with local subspace preferences”, In Proceedings of the 4th International Conference on Data Mining (ICDM), Brighton, U.K., 2004.

[12] C. Baumgartner, Plant C, Railing K, Kriegel H. -P, Kroger P, “Subspace Selection for Clustering High-Dimensional Data”, In proceedings of 4th IEEE Int. Conference on Data Mining (ICDM 04), PP 11-18, Brighton, UK, 2004.

[13] Daxin Jiang, Chun Tang , Aidong Zhang: “Cluster Analysis for Gene Expression Data: A Survey”, IEEE Transactions on Knowledge and Data Engineering, Issue Date : November 2004, pp. 1370-1386.

[14] Elke Achtert, Christian Bohm, Hans-Peter Kriegel, Peer Kroger, Ina Muller-Gorman, Arthur Zimek, “Finding Hierarchies of Subspace Clusters”, In Proceedings of 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Berlin, Germany, 2006.

3

P Kulkarni and I. Sengupta, "A new approach for load balancing using differential load measurement", The International Conference on Information Technology: Coding and Computing, ITCC, computer.org, 2000, pp 355-359

A New Approach For Load Balancing
Using Differential Load Measurement

Parag Kulkarni
Ways India Limited
International Infotech Park
NaviMumbai India
paragkulkarni@ways.com

Indranil Sengupta
Dept. of Computer Science and Engg.
I.I.T., Kharagpur, 721 302
India
isg@cse.iitkgp.ernet.in

Abstract

In distributed systems uneven arrival of the tasks may overload few hosts keeping rest of the hosts lightly loaded or even idle. This results in overall poor system performance in spite of the system capabilities. To make utilization of capabilities of the distributed systems load balancing can be used. Load balancing always comes with lot of overheads. Here we propose an algorithm based on Differential Load Measurement (DLM) for the load balancing. This algorithm helps us in selecting proper hosts for load balancing which can lead us to performance improvement of the system.

1 Introduction

One of the main advantages of a distributed system is that it leads to improved system performance and better sharing of resources. However, the overall performance of such systems may be poor due to imbalance in the loads as distributed to the various nodes [6]. *Load balancing* is one of the options to overcome this problem, which of course comes with a lot of overheads. So whether to go for load redistribution for load balancing and if yes which of the available alternatives should we choose — these are some of the critical decisions to be taken care of.

A number of algorithms have been proposed for load balancing and load distribution [9, 10, 5, 13]. The network partitioning issues along with inter-cluster and intra-cluster transfers for decision-making of load balancing [11] is also one of the major parameters which can be considered for the transfer [1].

Proximity of the nodes (in terms of hops) can also be considered while deciding the pair of the nodes to be selected for the task transfer [7]. *Receiver Initiated Diffusion (RID)* gives better performance in most of the cases as compared to *Sender Initiated Method (SIM)* and *Sender Initiated Diffusion (SID)* [2].

The gradient model uses the concept of gradient map [3]. *Dimension Exchange Method (DEM)* requires synchronization and followed by iterative steps of load balancing [3]. Hybrid compiler time and decision process / run time modeling also shows very good results [4]. A method called *Load Graph Based Transfer Method* is proposed for load balancing in distributed systems [8]. The Contraction within Neighborhood method and its adaptive version uses load contraction within neighborhood but does not keep it limited only to neighbors [12].

In this paper we propose a load balancing scheme using *Differential Load Measurement (DLM)*. This is a centralized approach. This can be made semi-distributed if it is applied on clustered basis. In DLM we use the differential load for the selection of the node pairs for load balancing. As per the differential loads we decide the momentums of the load for the transfer. Obviously higher load difference and less distance gets more priority. The transfers which are not cost-effective are avoided. This analogy is further developed on the cluster basis to make it applicable to large networks. Even for interconnected LANs, we can consider a single LAN as a group and apply the same method. In the case of LAN the node ordering (virtual) can be performed on the basis of load, and then the DLM is used to select the node.

2 Outline of the model

The policies of load determination may vary from system to system. Our system has its own policy to determine the load of the nodes. The load of the host is represented as a vector. This vector is called as a *load vector* and can be represented as a *load matrix*. The effective load of the host is calculated accordingly.

Let us consider a large system spread over a large area. Let the number of nodes in the system be n. We linearly order these nodes from 1 to n based on their proximity (number of hops separating them). Then we plot ithe load of every node on a graph in order and connect consecutive points in this graph with straight lines (Figure 1).

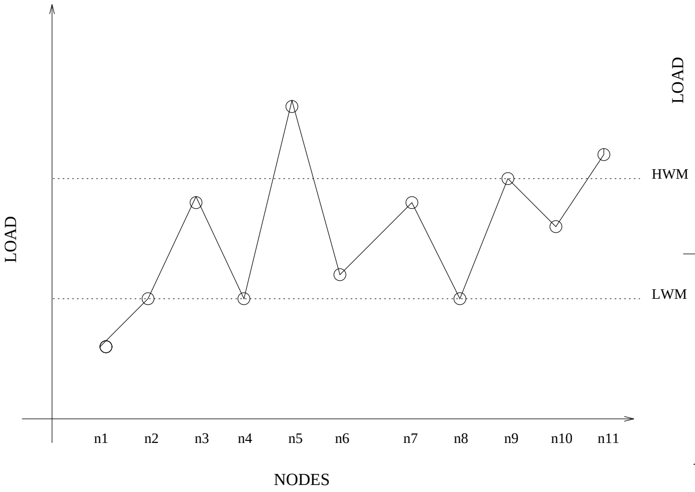


Figure 1. Plot of Loads against Nodes

Figure 1 also shows the '*Highest Water Mark (HWM)* and *Lowest Water Mark (LWM)* levels for the system with dotted lines. *HWM* is the load threshold above which we say that the node is heavily loaded, and *LWM* is the load threshold below which we say that the system is lightly loaded. HWM and LWM are decided on parameters susc as queue length, memory utilization, resource utilization, etc. The HWM and LWM values can be either absolute or relative. These values should be selected very carefully. Initially the values can be decided on the basis of load but as the algorithm progresses the values are tuned as per the history and the performance of the algorithm so as to get better performance.

In Figure 1, node $n5$ is the most heavily loaded. Since the load of $n5$ is greater than *HWM*, it is required to be balanced to improve the performance of the system. We have to select an appropriate node for shifting some of this load. In order to determine the node where the load of $n5$ is to be shifted we propose a simple procedure. We release two balls of some standard mass *m* from the highest point *A* corresponding to node $n5$. The balls are released, one each on both sides of the hill (Figure 2).

Both balls will slide on the slopes with acceleration determined by the requirement of the load transfer. Here the momentum of ball is determined by gravitation, friction and the mass of the ball. The friction on the balls, acting in opposite directions, is indicative of task shifting overheads. The friction will vary from node to node depending on load difference, overheads, etc.

The ball will acquire some acceleration and the of mentioned parameters decide whether it will able to clear the next hill or not. In this way the final position of the node where the ball will settle ultimately (valley) is determined. So, for the above case, the ball released on the left side slope of the hill settles at B as it could not clear the next hill. On the other hand the ball released on the right side of the hill settles at C clearing one intermediate hill. The node having a lower load among B and C is finally selected for the transfer of the load.

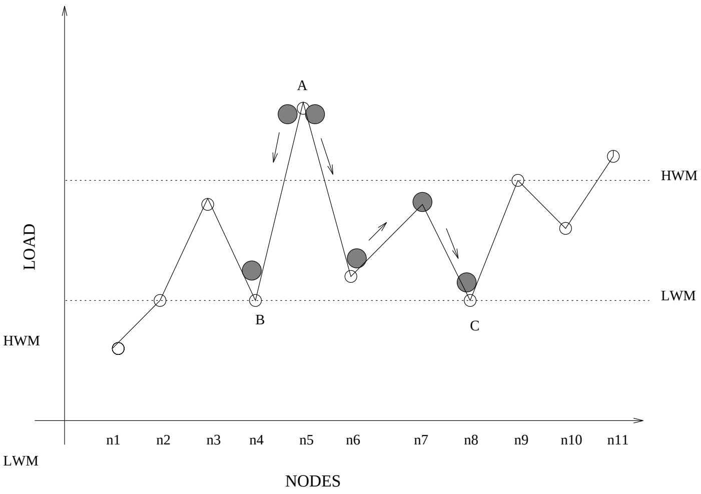


Figure 2. Movement of the Ball

This is fine for small systems but cannot be that effective for very large systems. In systems spread over large areas, to get even better results we first select the node with highest load. Then we proceed to perform some sort of grouping on the both sides of the node on the basis of proximity of these nodes with respect to the heavily loaded node. Then we order these groups so as to get a single valley in each group (Figure 3). Then we shall apply the same procedure to select the node for the task transfer. This approach reduces chances of improper selection of the node.

For large systems or interconnected LANs the same approach is extended. Here we will select the heavily loaded nodes in the cluster and arrange the other nodes on both the sides. This will be done for each cluster and then the same approach will be followed for the selection of the node for the load transfer. Figure 3 shows the arrangement of the nodes so that a heavily loaded node lies at the center. In the figure node A is heavily loaded and has load greater than HWM. It is then taken at the top. The rest of the hosts are arranged on the both side of the nodes.

Then two balls will be released on the two sides of the heavily loaded node. The valley with the lower load between the two nodes where the ball will settle is selected for the load transfer. This approach will avoid intra-cluster transfers if those are not cost-effective. But it will permit

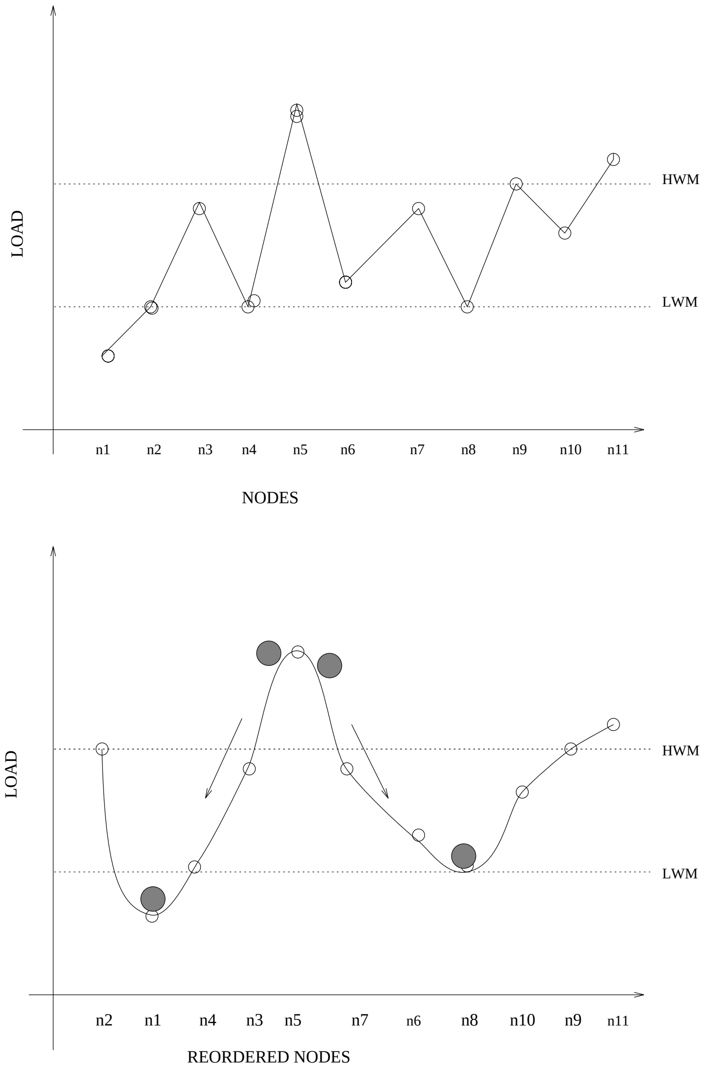


Figure 3. Arranging the nodes and clustered based approach for large systems

intra-cluster transfers if those are cost-effective. This algorithm is executed at some of the hosts. For large systems more than one host can be responsible for processing the algorithm for the particular group. The node where this processing is carried out is selected dynamically based on previous load history.

For small systems or machines connected in LAN, this ordering cannot be expected to be done on the basis of proximity of the nodes. In such situations node ordering can be done on existing loads so as to get maximum mileage out of the proposed algorithm.

3 Mathematical formulation of the model

Let us present a simple analysis of ball movement. First of all let us define mathematical symbols and notations.

g	:- acceleration
ρ	:- friction of the surface of the hill
θ	:- angle with vertical plane
G_c	:- cost Gradient
h	:- height of the hill $(h_1 - h_2)$
w	:- space between two nodes on the graph
O_t	:- task overheads
v	:- velocity of the ball
P_s	:- system parameters
k_1, k_2, k_3	:- constants
k	:- constant
m	:- Mass of the ball

Potential energy of the ball at point A figure 2 is given by

$$PE_1 = mgh_1$$

Potential energy of the ball at point B is given by

$$PE_2 = mgh_2$$

So difference between potential energies

$$\begin{aligned} PE_1 - PE_2 &= mgh_1 - mgh_2 \\ &= mg(h_2 - h_1) \\ \text{So, } PE &= mgh \end{aligned}$$

Force acting on the ball due to gravitation is given by

$$F = mg\cos(\theta)$$

where

$$g = \text{gravity} = kP_s$$

where the system parameter P_s is assumed to be uniform throughout the system.

$$\rho = \rho 1 + k(1 - e^{-(k_1 G_c + k_2 O_t)})$$

We can see the nature of change in ρ with respect to $k_1 G_c + k_2 O_t$ from the equation.

Hence, as the cost and overhead increases friction will also increase ultimately restricting the transfer.
Force acting in opposite direction due to friction is

$$\text{friction} = \rho m g \sin\theta$$

$$\text{Effective force} = mg\cos(\theta) - \rho mg \sin(\theta)$$

Hence by repeated application of the same analysis the position where the ball will settle down ultimately can be determined.

4 Algorithm

The steps of the algorithm are as follows.

Algorithm: *Differential Load Measurement*

Step 1: Determine the load values for all the nodes.

Step 2: Determine proximity of the nodes.

Step 3: Order all the nodes with respect to their proximity.

Step 4: Plot loads of the nodes as per their ordering.

Step 5: Connect all points on the graph with a straight line.

Step 6: Choose the node with highest load, provided its load is greater than *HWM*. If there is no node with load greater than *HWM*, no load redistribution is done.

Step 7: Grouping is done on the both sides of the node on basis of proximity and the network size.

Step 8: The ordering in each group is done to get single valley in each group.

Step 9: ρ is calculated for both the sides of the hills.

Step 10: On the basis of above parameters the final settling position of the ball on both the sides, if released on the slopes, is determined.

Step 11: Out of these two nodes the node with less load is chosen for the load transfer from the current node provided the load of that node is less than *LWM*.

Step 12: The same procedure is repeated with all the nodes having load greater than *HWM*.

Here the position where the ball settles is the node which is suitable for the task transfer. The friction and the mass of the ball are decided so as to get more cost-effective transfers and to improve the performance of the system.

5 Results

The above algorithm have been implemented on PVM using C++ tasks and for variable number of systems.

Figure 4 shows the comparison of the algorithm with other algorithm which uses just LWM and HWM for the load balancing for different task arrival rates. Figure 5 shows a comparison of this algorithm with the algorithm without load balancing and an algorithm which uses just *LWM* and *HWM* for the load balancing. For higher network sizes and for substantial task overheads for task transfer this algorithm gives even better results. It is also compared with Receiver Initiated Diffusion (RID).

Table 1. Performance vs. Number of Hosts

Number of Hosts	Optimal tine reqd.	Without Load Balancing	RID	Load balancing DLM
7	31.5	62.0	42.8	40.1
10	21.9	41.0	30.7	28.0
20	35.0	71.0	52.2	50.7
30	15.3	28.0	16.0	15.1
40	27.1	42.0	33.0	31.5

Table 2. Performance vs. Task Arrival Rate

Task Arrival Rate	Optimal tine reqd.	Without Load Balancing	RID	Load balancing DLM
0.19	58	84	72	71
0.27	30	46	39	37
0.45	41	63	46	43
0.63	11	17	12.5	12
0.75	19	31	25	23.5

For even larger systems we expect even better performance. However performance would vary across systems and applications and with the change of parameters, change in proximity.

6 Conclusions

Load balancing can be used to improve the performance of the distributed systems. To keep the over heads of the load balancing minimum we are required to perform it carefully. Load balancing is to be done only if it is cost-effective. DLM keeps track on selection of node pairs for load balancing and permits only cost-effective transfers. This ultimately enhance the system performance.

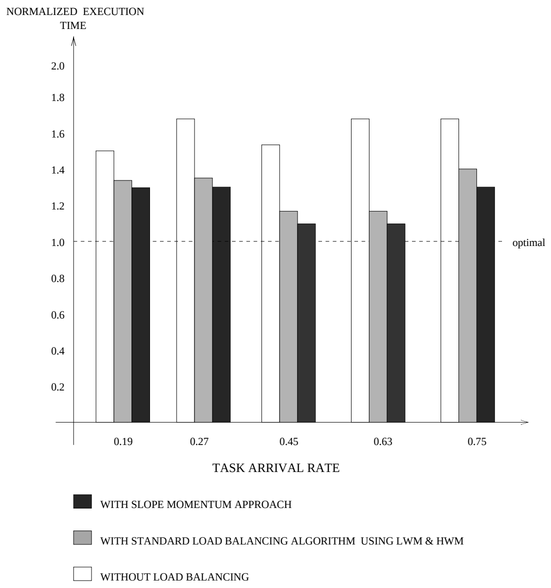


Figure 4. Performance Vs Task Arrival Rate

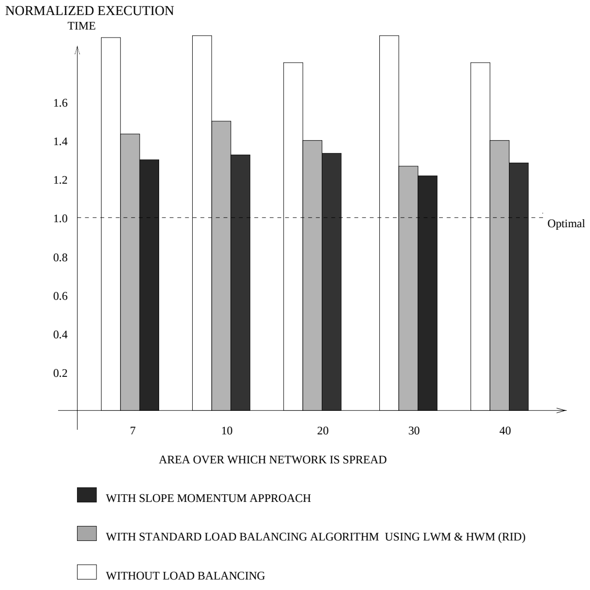


Figure 5. Performance Vs Number of Nodes

References

[1] A.Hac and T.J.Janson. A study of dynamic load balancing in distributed systems. *Proc. ACM SIGCOMM Symp. Communication Architecture and Protocols*, pages 348–356, August 1986.

[2] D.Eager, E.Lazowaska, and J.Jahorjan. A comparison of receiver-initiated and sender-initiated dynamic load sharing. *Dept. of Computer Science, University of Washington, Technical Report.*, April 1985.

[3] D.J.Evans and Wunbutt. Load balancing with network partitioning using host groups. *Parallel computing*, 20:325–345, March 1994.

[4] M. Z. et.al. Customized dynamic load balancing for a network workstations. *Journal of Parallel and Distributed computing*, 43:156–162, March 1997.

[5] J.Xu and K. Hawng. Heuristic methods for dynamic load balancing in message passing multicomputers. *Journal of Parallel and Distributed Systems.*, 18:1–13, 1993.

[6] M.Livy and M.Melman. Load balancing in homogeneous broadcasta distributed systems. *Proc. Conf. Performance, ACM*, 99(7):47–55, January 1982.

[7] P.Kulkarni and I.Sengupta. Fault tolerant system for distributed learning environment and dynamic load balancing on network of workstations. *International Conference on Distributed Learning, IGNOU, New Delhi, India*, February 1998.

[8] P.Kulkarni and I.Sengupta. Load graph based transfer method for dynamic load balancing on network of workstations. *National Conference on Communication IIT, Kharagpur, India*, February 1999.

[9] S.Zhou and D.Farrari. An experimental study of load balancing performance. *Proc. 7th International Conf. on Distributed Computing*, pages 490–497, September 1987.

[10] T.Y.Suen and J. S.K.Wong. Efficient task migration algorithm for distributed systems. *Journal of Parallel and Distributed Systems.*, 3(4):488–499, 1992.

[11] M. Willebeck and L. M. et.al. Strategies for dynamic load balancing on highly parallel computers. *IEEE Trans. Parallel and Distributed Systems*, 4(9):979–993, September 1993.

[12] W.W.Shu. Chare kernel based implementation on multicomputers. *Ph. D. Thesis, University of Illinois, USA*, 1990.

[13] Y.Chow and Kohler. Models of dynamic load balancing in a heterogeneous multiple processor system. *IEEE Trans. on Computers*, c-28(5):354–361, May 1979.

4

P Kulkarni and I Sengupta, Dual and multiple token based approaches for load balancing, Journal of Systems Architecture, Elsevier, 51/1 Page 95-112, Feb 2005

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Systems Architecture 51 (2005) 95–110

JOURNAL OF SYSTEMS ARCHITECTURE

www.elsevier.com/locate/sysarc

Dual and multiple token based approaches for load balancing

Parag Kulkarni [a,*], Indranil SenGupta [b]

[a] *Capsilon Research Labs, Capsilon India, Pune, India*
[b] *IIT, Kharagpur, India*

Received 26 June 2000; received in revised form 1 November 2002; accepted 18 October 2004

Abstract

In distributed systems uneven arrivals of the tasks may overload a few hosts while some of the hosts may be lightly loaded. This load imbalance prevents distributed systems from delivering its performance to its capacity. Load balancing has been advocated as a means of improving performance and reliability of distributed systems. We propose a distributed load balancing algorithm *LoGTra* to deal with this problem. *LoGTra* uses load graph and token based policy. The extensions to *LoGTra* based on dual tokens *DTLB* and multiple tokens *m-LoGTra* are proposed in this paper. *m-LoGTra* allows host to generate multiple tokens. This allows system to search host for load balancing in multiple directions and can avoid starvation of remote hosts. Local maxima and local minima are responsible for initiating transfer and that limits the number of hosts generating tokens at a time. As overheads are kept under control by limiting number of tokens, the algorithm promises for improved performance.

Keywords: Distributed system; Load balancing; Task transfer; Token

1. Introduction

In last two decades distributed applications have grown increasingly complex and there is continuous demand for more computation power, efficiency and performance. Distributed systems are capable of giving performance comparable to highly parallel computers with better resource utilization and fault tolerance. In a distributed system, however, the CPU power is fragmented over a number of different hosts which generally means a process can access the computation resources of the host over which it is executing. This means that if more processes are executing on one host than another each process is not getting fair share of

[*] Corresponding author. Address: L-2, Suyog Nagar, Behind Shivaji Housing Society, Off. Senapati Bapat Road, Pune, 411016, India. Tel.: +91 20 2567 3941; fax: +91 20 412 0609.

E-mail addresses: parag.kulkarni@capsilon.com, paragakulkarni@yahoo.com (P. Kulkarni), isg@cse.iitkgp.ernet.in (I. SenGupta).

1383-7621/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.sysarc.2004.10.001

the system resources leading to load imbalance. Load imbalance deprive the system from using all resources to their fullest capacity.

In a network of computer systems it is very likely that while some hosts are fully laden with excess jobs queuing for execution other machines are idle [1]. Overall performance of the system may be increased by transferring some of the excess jobs on to the more lightly loaded hosts. In short load balancing is distributing tasks (load) among available hosts as evenly as possible. Unless otherwise specified we shall assume that all systems are architecturally homogeneous whereby any processor can service requests issued by program running at any of the other sites.

Irrespective of how efficiently it is implemented, remote execution incurs overhead due to process migration. Naive use of remote execution facilities may, therefore, cause degradation rather than improvement in performance.

Load balancing is the finest form of load distribution. It attempts to ensure that load is distributed equally among the hosts as far as possible. Load balancing is nothing but equalizing load at each host. It generally attempts to minimize execution time by maximizing productive utilization of resources. This can be done by giving processors more work when they become idle or by attempting to predict how much work each of them can handle and distributing work to them. Load leveling occupies the ground between the two extremes of load sharing and load balancing.

A centralized algorithm based on differential load measurement and proximity gradient promises good performance [2]. Algorithms to give minimal interconnection penalties when a slow interconnecting network is involved is discussed in [3]. In most of the algorithms referred the load balancing is limited to neighbors. Load balancing is done by diffusion to result in balancing across complete network. The concept of diffusion is combined with token based approach in *LoGTra*. A distributed algorithm based on load graph [4] and its extensions are discussed in [5].

In this paper we propose extensions to *LoGTra* [4]. *LoGTra*, is based on load graph and is a token based approach. The extensions to this algorithm, Dual Token Based Load Balancing (*DTLB*) and Multiple Token Based LoGTra (*m-LoGTra*) are proposed in this paper. In *LoGTra*, only tokens those are at local maxima can generate tokens. Here in this extension we made some changes in this assumption. This section is followed by Section 2 that gives an overview of related work. Section 3 discusses about multiple token based extensions of *LoGTra* followed by Section 4, that gives outline and algorithm for this approach. Section 5 gives an analytical model for this algorithm and analyzes complexity of the same. Section 6 gives results that is followed by discussion about *m-LoGTra* and results for the same in Section 7. Section 8 gives concluding remarks.

2. Related work

Load balancing has been an area of active research since the emergence of distributed systems in the early 1970s. The general problem has been examined for a broad range of computing environments, at many different levels and using variety of strategies. General overviews may be found in [6].

The division of system load can take place statically or dynamically [7]. Load scheduling is an essential component of scheduling in distributed systems [8,9]. In static load distribution approach jobs are assigned to hosts probabilistically or deterministically and is effective when scheduler is pervasive.

Load distribution can be done by load sharing, load balancing or load leveling [7]. Load sharing in distributed system was studied by Wang [10]. Process transfer can be used to achieve balanced load and is performed either non-preemptively through process placement or preemptively through process migration. Migration is more costly than placement. In addition migration may not result in performance improvement due to overheads for process migration [11,12].

A study carried out by Devarkonda and Iyer [13] showed that CPU, memory and IO requirements of a process can be predicted prior to execution using statistical pattern recognition method. Goswami [14] claimed that prediction based heuristics can be more effective and studied a few prediction based as well as non-prediction based heuristics.

Zhou et al. [15] discussed design, implementation and performance of a load sharing system for large, heterogeneous distributed system.

Kruger Liveny [16] have shown that load balancing can potentially reduce the mean and standard deviation of task response times relative to load sharing algorithms.

In sender initiated method heavily loaded node initiates load transfer [17]. I.e. an overloaded source searching for and sending tasks to an underloaded target. In a receiver initiated method lightly loaded node initiates load balancing [1,18]. i.e. an underloaded target searches for an overloaded source. A similar strategy called neighborhood averaging is proposed in [19]. Receiver Initiated Diffusion (RID) can be thought of as the converse of the Sender Initiated Diffusion strategy in that it is initiated by receiver. The load diffusion is initiated by any processor whose load drops below a prespecified threshold. Under symmetrically initiated algorithm [20] both sender and receiver initiates load distributing activities for task transfers. These algorithms have advantages of both sender initiated and receiver initiated algorithms.

The gradient model is demand driven approach [21]. The basic concept of this approach is that underloaded processors inform other processors in the system about their state and over-loaded processor responds by sending a portion of their load to the nearest lightly loaded processor in the system.

In Dimension Exchange Method (*DEM*) [22,23] small domains are balanced first and then combined to form large domains until ultimately the entire system is balanced. Marc H. Willbeek et al. made comparative study of Sender Initiated Diffusion, Receiver Initiated Diffusion, HBM, Dimension Exchange Method and Gradient Model approaches of load balancing [24].

Pankaj Mehara et al. [25] studied the problem of generating synthetic workload for load balancing. Mitzenmacher [26] proposed an approach studying limiting, deterministic models representing the behavior of the systems as the number of servers goes to infinity.

Shu [27] proposed an approach called Adaptive Contraction Within Neighborhood (ACWN). *ACWN* is an extension to *DEM* and load balancing is not just kept limited to neighbors [28,29]. *ACWN* is based on *RID*.

Zaki et al. [30] analyzed local, global, centralized and distributed interrupt based receiver initiated dynamic load balancing strategies on network of workstations with transient external load per processor. James D. Teresco [3] has discussed penalties involved when slow interconnection network is involved. [31] compared computation and local distribution costs. A framework for dynamic load balancing in which the job traffic is modeled by adversary is given by Muthukrishnan and Rajaraman [32]. A load balancing algorithm *DASUD* (Diffusion Algorithm Searching Unbalanced Domain) is proposed to deal with some problems incurred by *SID* when tasks are considered to be indivisible [33]. This algorithm also checks whether domain of the host is balanced or not. Diffusive load balancing policies suitable for dynamic load balancing are studied in [34]. Comparison of *SID* and *DASUD* is given in [35].

A few load balancing implementations have been reported in literature. The early systems [36–38] were used remotely place long-running processes on lightly loaded hosts. A slightly more advanced system is Condor [39,40] which supports process check-pointing and migration. The majority of systems have used UNIX processes as unit of migration.

3. Multiple token based extensions of LoGTra

Load balancing has been advocated as a means of improving performance and reliability of distributed systems. As load balancing allows migration of processes it can help in meeting deadlines. Also execution of tasks on remote hosts in case of emergency improves reliability of the system.

A new load balancing approach *LoGTra* has been proposed [4,5] to deal with this problem. *LoGTra* is a token based approach and a token is used to select node to transfer load from heavily loaded node. In *LoGTra* there is possibility that a node, which is lightly loaded and at local minima may not get justified load share. For example token generated at local maxima may land up with the node which is lightly loaded but not local minima. This may keep local minima lightly loaded even after load balancing. To end this possibility in dual and multiple token based approaches both local maxima as well as local minima can generate tokens. When there is need of immediate balancing of load at any node then that node can generate multiple tokens to intensify hunt for lightly loaded node.

Dual Token based approach (DTLB) and *multiple token approach (m-LoGTra)* can help us to further intensify our hunt of node pair for load balancing. At first, all the nodes will check loads of their neighbors for the task transfer. If there is a neighbor, with suitable load difference for the task transfer, the transfer is initiated. In next stage nodes with all incoming edges and load less than *LWM* (*Low Water Mark*) along with the nodes having all outgoing edges and nodes with load above *HWM* (*High Water Mark*) can generate tokens. *HWM* is level above which node is treated heavily loaded and *LWM* is the level below which a node is treated lightly loaded. In this paper these levels are determined on the basis of historical data and tuned as algorithm progresses. This node will then take decision about load transfer on the basis of data provided with the token. The implementation of the algorithm can be started from any of the nodes, which is heavily loaded and is a local maxima or lightly loaded and local minima.

In this load balancing algorithm we can divide the process in to three steps.

(1) Load estimation. Where each node determines own load.
(2) Finding out node pairs for load balancing.
(3) Actual task migration.

In *Dual Token based Load Balancing* (*DTLB*) both nodes those at local maxima as well as those at local minima are allowed to generate tokens. As the nodes which are local maxima and load value greater than *HWM* have excess load these token are named as *supply tokens*. In contrast, the tokens generated by the nodes at local minima are called as *demand tokens*. In *multiple token based LoGTra*, which hereafter referred as *m-LoGTra*, both the nodes at the local maxima as well as at local minima can generate multiple tokens and send those in different directions in hunt of a node suitable for load transfer.

4. Outline of the DTLB algorithm

Each node will determine its load status with respect to neighboring nodes based on load graph i.e. whether it is a local maxima. Then it can generate supply token, which is a general request for load. In case if it has all incoming edges i.e. it is a local minima then it can generate demand token. Token is a permission to share load of the node or share load with the node. A token will contain token type, address of the host that has generated token and path followed by token. E.g. if host with load n has generated a demand token and so far visited node m it will contain information:

(n, d, m)

Demand token holder can share own load with the token initiator, while the supply token holder can share load of the token initiator. In the first phase of the algorithm, each node will inform its neighboring node about its load. Therefore, at the end of initial stage each node will have the information of the load levels of its neighboring nodes. After this stage, all the nodes will try to balance load among neighbors. Then as discussed above, local maxima and minima can Generate token based on their load values. The token carries an array with it which contains information of the load positions of the last three nodes it has come across and the pair of nodes between which the task transfer has taken place recently. It also contains the information about the initiator of the token and the counter indicating the hops it can make. Any node with the token can initiate task transfer if the load difference is adequate.

The required difference between loads for task transfer is decided based on *HWM*, *LWM* and the distance between the nodes in terms of hops. Along with this, any node can initiate transfer if it finds its neighboring node has sufficiently high load. Token continues to travel until it reaches to a node that wants to go for the transfer based on data held by the token. A situation may arise when a node which has generated supply token has received a demand token then it can respond it. But if a node, which has generated demand token, when receives supply token from some other node is not allowed to respond it without getting response for its own demand token. This allows the responders of demand tokens to start task transfers immediately after receiving demand token.

After initiating the token the following rules of token transfer are followed to find out if there is any node, which is suitable for the cost-effective task transfer. For demand and supply tokens exactly opposite transfer rules are followed. The important points about token transfer are listed below.

Rules of token transfer:

(1) All the nodes with all outgoing edges and are heavily loaded generate supply token. Similarly nodes with all incoming edges and are lightly loaded generate demand token.
(2) Token is accompanied with
    (a) Information about the initiator.
    (b) A counter initialized to value N where N is the number indicating maximum transfers of the token permitted.
    (c) An array of three indicating three nodes along with their loads which the token came across most recently.
(3) Supply token can travel only in direction of the edges of LG and demand token can travel only in opposite direction of these edges.

Refer to Fig. 1, here nodes N1,...,N10 are different hosts connected in network. All the numbers associated with each of the nodes gives the load of corresponding host. The links among different host show connectivity and all links are directed from node with higher load to that with lower one. Node $n1$ has all outgoing edges and no incoming edge. None of its neighboring nodes has such load that task transfer can be initiated. Now as this node has all outgoing edges, it can initiate a supply token. Therefore, as shown in Fig. 1, it has initiated a token with down counter set to value 5. The value of down counter can be set based on diameter of the network. Similarly node $n10$ which has load value below LWM and all incoming edges, can initiate demand token. The demand token generated by $n10$ and the supply token generated by $n1$ proceeds as per the token transfer rules.

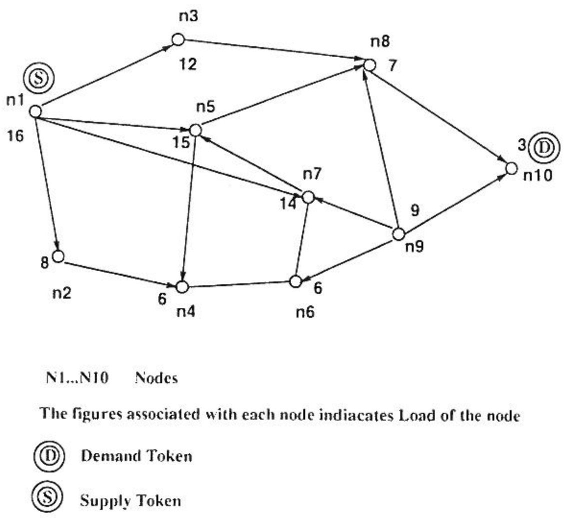
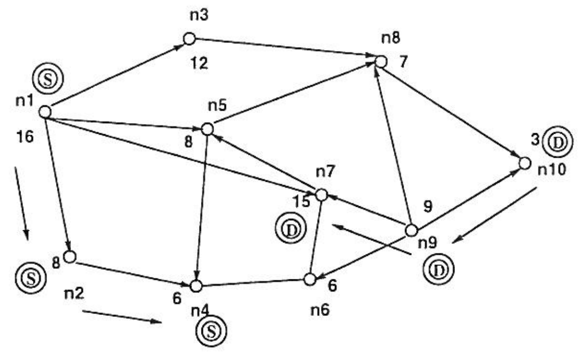


Fig. 1. Demand and supply token transfers in load graph.

The supply token travels from $n1$ to $n2$, which is the node with lowest load among its neighboring nodes. As this node is not suitable for the task transfer the token travels to node $n4$, which is suitable to initiate task transfer. Then the transfer can be initiated. Similarly demand token travels from $n10$ to node $n9$ and then as $n9$ is not suitable for the load balancing token will proceed to node $n7$, which is suitable for the task transfer and hence load is transferred as shown in Fig. 2. After the transfers each node will again inform its neighbors about its new load level. The new local maxima and minima are determined and supply and demand tokens are generated.

Along with the advantages of $LoGTra$ of more flexibility of node selection and restricting transfers those are not cost effective by strictly allowing only local maxima and minima to generate tokens, $DTLB$ helps in avoiding starvation of local minimas by allowing them to generate tokens. DTLB intensifies the hunt for nodes for load balancing among hosts.

In this algorithm, apart from the task transfer among the neighboring nodes, with the token



N1...N10 Nodes

The figures associated with each node indiacates Load of the node

Ⓓ Demand Token

Ⓢ Supply Token

Fig. 2. Token transfer.

transfer method care is taken to avoid the starvation of those nodes for which neighbors are not suitable for the task transfer. *DTLB* further avoids this starvation by taking care of both local maxima and minima.

The algorithm for *DTLB* is as follows

Algorithm for Dual Token Based Load Balancing

(1) *M*: Minimum difference between load of token holder and load of token initiator necessary for task transfer
(2) LWM: Lowest water mark
(3) HWM: Highest water mark
(4) *T*: Token

```
{
  For node 1 to N
  {
    determine load of the node;
    determine value of M;
    if (node = neighbor(node))
    {
        inform load of the node;
    }
  }
  For node 1 to N
  {
    Determine HWM and LWM on the basis
          of Load of neighbors;
    Determine value of M (allowed load
          difference);
    If ($Loadofnode_i$ < LWM)
    {
     if (load of Neighborofnode
           -Loadofnode > M)
     {
          initiate transfer;
     }
    }
}

For a Node
{
  If (Load > HWM) and ((loadofall
        neighbors) < (loadnode))&&
        ((loadofnode-loadofneighbor) < M)
  {
   Initiate a sender token T
   with a token count;
   while ((tokencounti > 0)|||| (load
   of node – load of token holder) < M)
   {
         forward T to neighbor
              with least load;
         token count–;
   }
   if ((loadofnode – loadoftoken holder) > M)
   {
       Initiate the transfer;
   }
  }
}

For a Node
{
  (Load < HWM) and (loadofallneighbor) >
  (loadnode)
  &&(loadofneighbor – loadofnode) < M
  {
   Initiate a demand token T with
   a token count;
  }
  while ((tokencount > 0) or (loadofnode
  – loadoftokenholder < M))
  {
   forward T to neighbor with highest load;
   token count–;
```

```
        }
        if (loadoftokenholder – loadofnode > M)
        {
                Initiate the transfer;
        }
    }
}
```

Fig. 3 shows the position of the graph after the task transfer has taken place. In new graph *n*8 is can generate demand token and *n*5 can generate supply token. The nodes will inform their neighbors about their loads periodically.

Network intimation frequency (*NIF*) and token count (T_c) can be tuned on the basis of empirical result to get maximum mileage out of this algorithm. These values are set dynamically. This setting improves performance.

In this algorithm we have given priority to the transfer among neighbors. However, the algorithm is not restricting the transfer from the remote node to any other node if it is cost-effective.

5. Analytic model for DTLB

To analyze the algorithm Let

$p_i(k)$ probability that k tasks at node i at time t_i

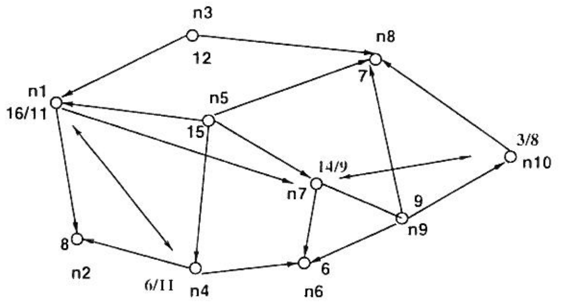
λ_i task arrival rate at node i due to external demand

e_i effective arrival rate at node i from all sources

mu_i service rate at node i

r_{ij} probability that a task from node i will be transferred to j

N number of nodes in the network



N1...N10 Nodes

The figures associated with each node indiacates Load of the node

Bold numers indicate loads of the nodes after load balancing

Fig. 3. Graph after load transfer.

Total flow rate into node i

$$e_i = \lambda_i + \sum_{k=1}^{N} r_{ki} * e_k - \sum_{i=1}^{N} r_{ik} * e_i \tag{1}$$

$$p_i(k) = p_i(0) * (e_i/\mu_i)^k/k! \quad \text{where } k = 0, 1, \ldots, N \tag{2}$$

Assuming that arrival is Poisson we can state that [41,42]

$$p_i(0) = (e_i/\mu_i^k)/k\,! + (e_i/\mu_i)/(1 - e_i/\mu_i) \tag{3}$$

One should expect that this system is subjected to Poisson input at the rate of e_i. The length of queue at node i can be given with the following expression

Length of queue node i

$$= p(0)(e_i/\mu_i)(e_i/\mu_i)/(1 - e_i/\mu_i)^2 \tag{4}$$

Now suppose we want to maintain L_{q_i} less than some particular value M. From this equation one can derive value of e_i. Now suppose we want to maintain L_{q_i} less than some particular value M. From this equation one can derive value of e_i.

$$e_i^2 * (\mu * p(0) - \mu_i) - 2 * \mu_i * e_i^2 - M^2 * \mu_i^3 < 0 \tag{5}$$

Solving this equation value of e_i can be obtained.

We have already stated that

$$e_i = \lambda_i + \left(\sum_{i=1}^{N} r_{ki} * e_k - \sum_{k=1}^{N} r_{ik} * e_i\right) \tag{6}$$

$$\rho_i = \left(\sum_{i=1}^{N} r_{ki} * e_k - \sum_{k=1}^{N} r_{ik} e_i\right) \tag{7}$$

$$e_i = \lambda_i + \rho_i \tag{8}$$

Now let us assume that for a particular node p1

$$\sum_{i=1}^{N} r_{ki} * e_k - \sum_{k=1}^{N} r_{ik} e_i > 0$$

i.e. the incoming tasks due to task transfer are more than those leaving the node. In this case, there exists a particular node p2 where

$$\lambda_{p2} > \lambda_{p1}$$

Because the task will not transferred to a node unless it has lesser task arrival rate and for that time period say t_1 to t_2

$$\lambda_{p2} > \lambda_{p_t} + \rho_t \quad \text{for } t = t_1 \text{ to } t = t_2 \tag{9}$$

As we are not allowing the transfer of the task unless load difference is greater than M. As we are not allowing the transfer of the task unless load difference is greater than M the total task arrival rate in spite of the positive value of transferred task will always be less than the service rate

$$\lambda_i + \rho_i * e_i < \mu_i \tag{10}$$

As in the network in spite of task transfers the arrival rate at each host is less than the service rate at each node. Which clearly indicates that system is stable and converges.

Let for complete network the task arrival rate is less than that of the task service rate. In this case the system should be stable if load is balanced. In this particular case if there is a node for which task arrival rate is more than that of service rate the system may not be stable. If we can able to treat this node, system can be made stable.

Now let us take an example where at a particular node i $\lambda_i > \mu_i$ ultimately an unstable queue is developed and for the whole network $\lambda < \mu$. In this particular system without load balancing the system will not converge and our scheme can lead the system towards a stable solution. Due to unstable queue development this node will be the local maxima and will have less load as compared to node i. Therefore, the load graph will show all out going edges. Hence, the tasks will be transferred out of that node that means ρ_i will be negative.

$$e_i = \lambda_i + \rho_i \tag{11}$$

and ρ_i is negative

$$\lambda_i - \mu_i < \rho_i \tag{12}$$

This will ultimately make the system stable. The complexity of *DTLB* is of the order of n^2.

5.1. Complexity analysis

Let N be the number of nodes in the network. In the initial phase each of N nodes will inform their neighbors about their load status. The number of neighbors for any node cannot be more than $(N - 1)$. Let us consider a completely connected graph where maximum transfer in initial phase can be given by

$$\textit{message transfers per node} = N - 1 \tag{13}$$

$$\textit{maximum message transfers} = N * (N - 1) \tag{14}$$

In worst case all the nodes will generate tokens. However, this is possible if the network is connected as a bus. The product of nodes generating tokens and the connectivity is of the order of N.

As the connectivity of the graph increases the performance of the system is bound to improve. However, as the connectivity still increases the performance becomes steady for fully connected networks. The overheads of the message transfers to run the algorithm increases with the connectivity.

In this algorithm, in the initial phase every node will inform its neighbor about the load. In worst case, when the complete graph is connected the message transfers will be $N * (N - 1)$. Now, for fully connected network there can be at the most one local maxima and one local minima so the token can be generated by at the most two node. So the complexity is of the order of N^2. In worst case in a network there can be $N - 1$ local maxima (star connected network). However, in this case as the graph density is very low the message transfers in initial phase are of the order of N. And the algorithm complexity remains of the order of N^2. We can observe very easily that as the graph connectivity increases the number possible maximum local maxima and minimas reduces. Even though both local maxima and minima can generate the tokens complexity of the algorithm remains of the order of N^2.

6. Results using DTLB

Simulation study for the proposed algorithm is carried out with randomly generated tasks. Artificial tasks are generated with variable execution times. The arrival of the tasks is assumed as Poisson. Also the tasks are randomly assigned to the hosts. We have taken results with different task arrival rates, different graph densities and different network sizes. Here graph density stands for connectivity of graph in percentage where fully connected graph density is 100%. All the load in tables are normalized and are in terms of units. Where as time of execution is also normalized. The lower values indicate better performance. The time of execution for the same set of artificial tasks is found out without load balancing, with receiver initiation and with Dimension Exchange Method. Tables 1–3 shows the simulation results of for randomly generated tasks, with different number of hosts and different task arrival rates. In receiver initiation diffusion only receiver initiates transfer. At a time RID addresses its neighbors and in process diffusion all across the network takes place. Table 4 shows simulation results for randomly generated tasks, with different graph densities for *DTLB* and *DASUD*.

Diffusion policies have a few limitations. If load is not divisible *LoGTra* can handle it. As token based approach prevents multiple transfers and transfers are limited to pair of nodes. In this paper we compared performance of algorithm with two basic diffusion algorithms *SID* and *RID*.

The performance is bound to vary from the system to system and load variation. For lower graph connectivity and higher sizes of networks there is remarkable performance improvement using this algorithm. The performance for higher connectivities drops down a bit due to high communication overheads. When connectivity of the network or

Table 1
Performance vs graph density

Graph density (%)	Optimal time required	Without load balancing	Load balancing DTLB	RID method	DEM method
20	11	30	21	23	21.5
40	25	48	33	36	35
60	10	25	16	17	16.5
80	30	58	32	34	33
100	40	61	43	43	44

Table 2
Performance vs number of hosts

Number of hosts	Optimal time required	Without load balancing	Load balancing DTLB	RID method	DEM
7	8	14	11	11	11.2
10	11	17	13	13	13
20	13	24	14	14.7	14.5
30	7.5	16	8	9	8.5
50	17	32	18	19.5	19
100	20	37	22	26.5	25

Table 3
Performance vs task arrival rate

Task arrival rate	Optimal time required	Without load balancing	Load balancing DTLB	RID method	DEM
0.19	58	84	68	72	70
0.27	30	49	37	39	39
0.45	41	63	48	46	45.5
0.63	11	17	11	12.5	12.5
0.75	19	31	23	25	25
0.92	28	46	35.5	37	35

Table 4
Performance vs graph density

Graph density (%)	Optimal time required	Without load balancing	Load balancing DTLB	DASUD method
20	17.5	37	21	24.5
40	27	51	29	31
60	19	35	26	27
80	28	54	37	38
100	30	51	41	41

graph density is very high number of remote hosts are less. When network density is close to 100%, all the hosts are connected. But when graph density is low in diffusion method chances of starvation of remote host are very high. *DTLB* can handle this situation with its hunt for local maxima and minima. For different network sizes, task arrival rates and graph densities substantial improvement can be obtained using *DTLB*. Only for very high graph densities, improvement over *RID*, is not that significant.

Figs. 4 and 7 shows the outputs for the different graph densities. As the graph density increases the performance also improves and Figs. 5 and 8 shows the results obtained for the different size of the networks. It can be clearly seen that as the size of the network increases the improved performance can be obtained. Fig. 6 gives the performance of the algorithm for various task arrival rates. Fig. 7 shows comparison of this algorithm with the algorithm without load balancing and an algorithm which uses just sender initiated transfer and using *LoGTra*. As graph density increases the performance improvement by *DTLB* over *RID* goes on reducing and for fully connected network performance improvement due to *RID* is slightly more than that by *DTLB*. In fully connected network all the hosts are neighbor of one another and there is no remote host.

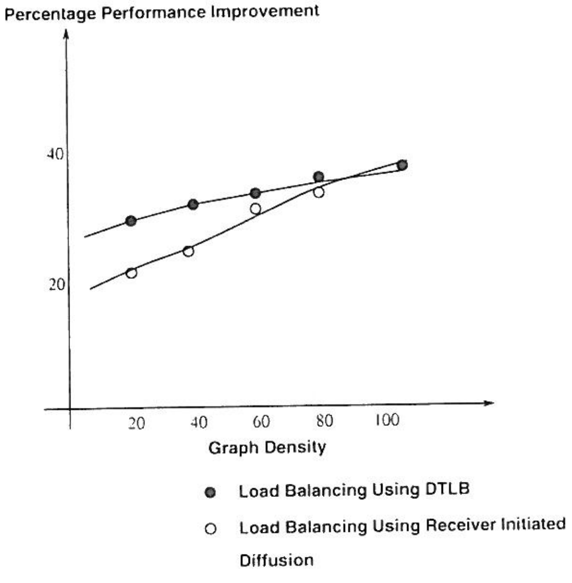


Fig. 4. Execution time vs graph density.

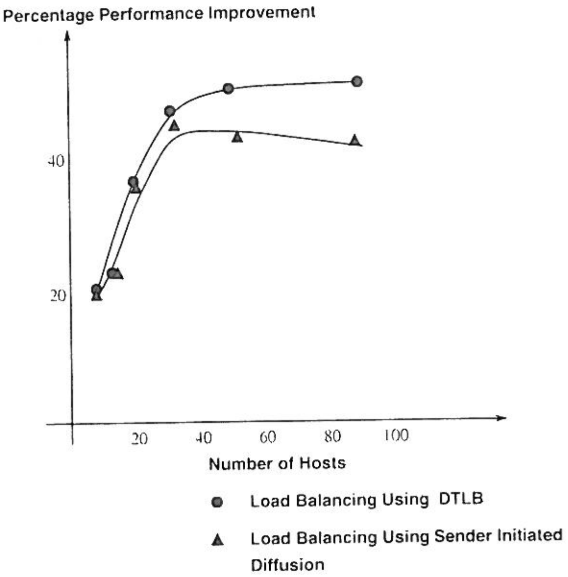


Fig. 5. Execution time vs number of hosts (network size).

Fig. 9 gives the performance with respect to various values of *NIF*. The performance improvement is very less for the very low value of *NIF*. Initially as the *NIF* increases the performance also improves. However, after certain value it drops down

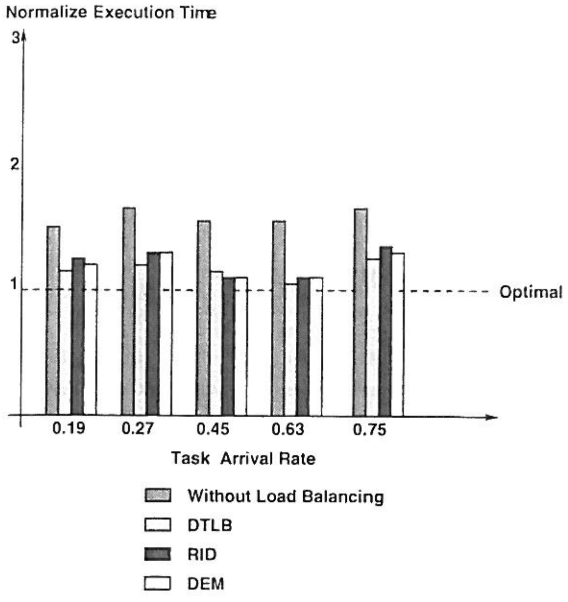


Fig. 6. Execution time vs task arrival rate.

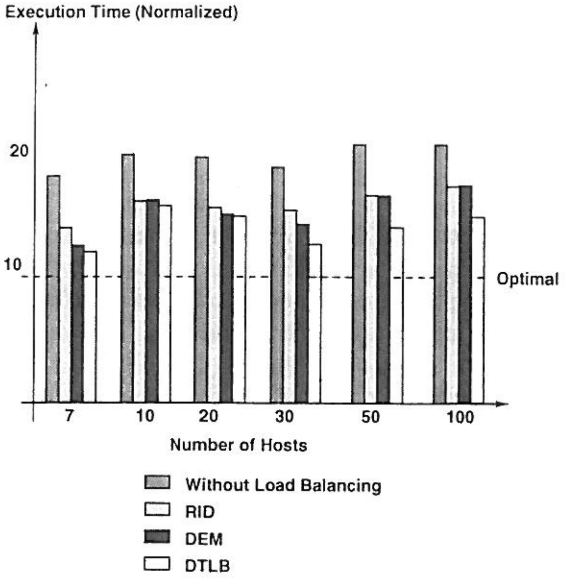


Fig. 8. Execution time vs number of hosts.

with increase in *NIF*. The nature of this graph is very similar to that of performance of *LoGTra* with respect to *NIF*.

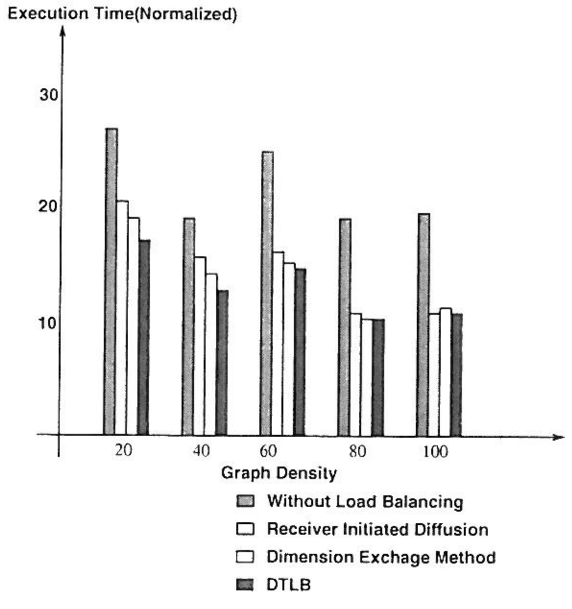


Fig. 7. Execution time vs graph density.

7. LoGTra with multiple tokens

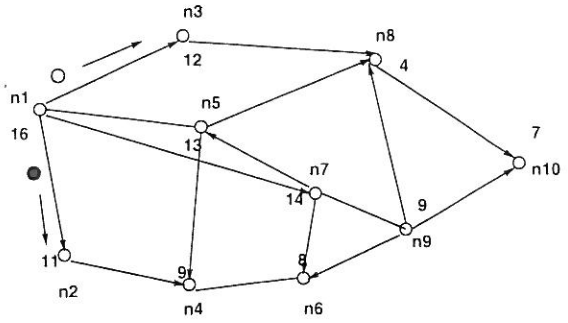
An extension to this approach is *LoGTra with multiple tokens* (*m-LoGTra*), where multiple tokens can be generated by both local maximal as well as local minimal. The nodes can generate more than one tokens and this number can be decided based on necessity of load balancing. There will be multiple tokens generated by the node if the node has very high load or very low load and there is immediate need to balance the load.

Consider a simple graph where a node *n*1 has load more then *HWM* and is at local maxima and no neighboring node is suitable for the task transfer. As load of the node is very high, it is required to be balanced. In this case this node will generate two tokens and forward them in different directions. As the number of tokens increases the overhead also increases so the multiple tokens should be generated only in the case of high imbalance and immediate necessity of balancing load. However, if as per the situation one or more tokens are generated the performance can be improved. More than one token can help to get a better node as well as can help in searching node for load balancing. In addition, a node can get more than one suitable nodes for task transfer and as per requirement it can distribute its excess load among them equally. The comparison of the *multiple token base LoGTra* (*m-LoGTra*) and *LoGTra* shows the following facts.

(1) For more number of hosts multiple token based LoGTra gives better performance as compared to *LoGTra*.
(2) The performance improvement is more evident for higher task arrival rates.

Consider Fig. 10 here node n_1 is at local maxima and is heavily loaded. In addition, there is need of immediately balancing the load as per the task deadlines. With the *m-LoGTra* two tokens are generated. The node can decide the number of tokens generated on the basis of urgency. The maximum token generated by a node is kept limited to maximum number of edges. For two tokens substantial improvement in performance is observed as compare to one token. However, same is not true for increasing the number of tokens further. With reference to experimental results, we do not recommend this number to exceed number of outgoing edges.

In Fig. 10, one token is forwarded in the direction of the neighbor with the lowest load. Another token is forwarded in the direction of the node with the second lowest load value. From n_1 one token is forwarded to node n_2 (refer Fig. 11), then node n_4 and then node $n6$. Similarly, second token is forwarded in the direction of node n_3 and then to node n_8. As between node n_8 and n_6, n_6 has lower load, n_6 can be selected for the task transfer.



N1...N10 Nods

The figures associated with each node indiacates Load of the node

● ○ Tokens

Fig. 10. Multiple token generation.

In some cases load can be distributed to both nodes to get more balanced results. This is decided on the basis of load difference of the source with both the nodes. Here it distributes load to both hosts. In this way *Multiple Tokens* can help in

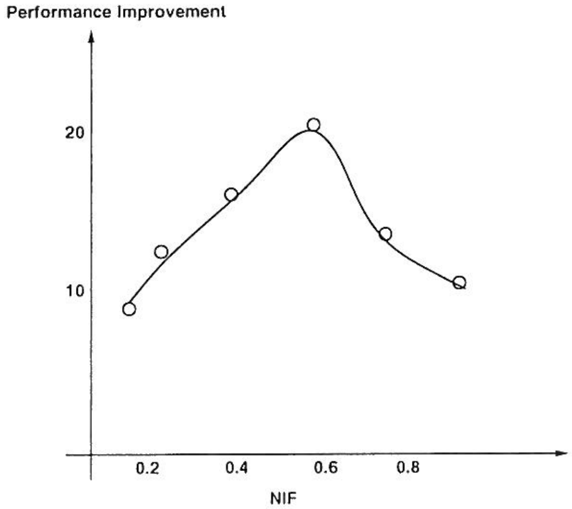
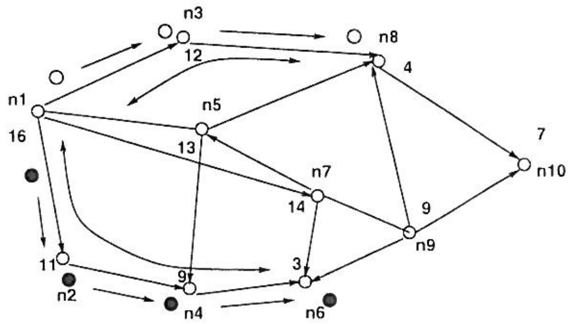


Fig. 9. NIF vs performance improvement.



N1...N10 Nodes

The figures associated with each node indiacates Load of the node

● ○ Tokens

Fig. 11. Multiple token generation (token transfer).

selection of the node pairs. Fig. 12 shows load values after the task transfers. The new *LG* is shown and with respect to new *LG* node n_7 can generate token.

The complexity of *m-LoGTra* is of the order of N^2 as we have already discussed with *DTLB*. As multiple token generation is used on selective basis it will not affect the complexity and the complexity more or less remains same as that of *LoGTra*. In case if all the nodes will generate multiple tokens and the number of tokens generated by each node is a function of number of nodes. Then the complexity of algorithm will be of the order of N^3.

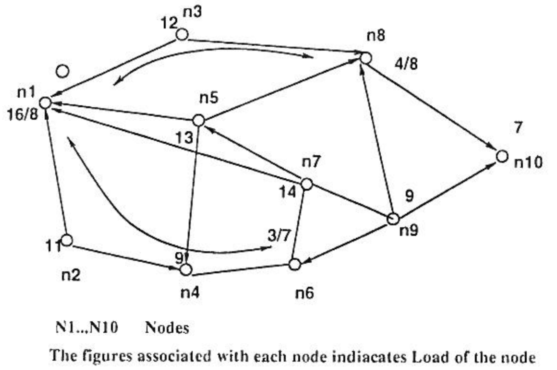

12
n3
n8
4/8
n1
16/8
n5
13
7
n10
n7
14
9
n9
3/7
11
9
n2
n4
n6
N1...N10 Nodes
The figures associated with each node indiacates Load of the node

Fig. 12. Multiple token generation (load values after the task transfer).

Table 5
Performance vs network size

Network size	Optimal time required	Without load balancing	Load balancing (LoGTra)	DTLB method	Multiple token approach	RID	DEM	SID
5	85	152	105	103	104	105	105	107
10	70	128	92	92	93	95	94	97
25	91	163	120	117	116	114	115	120
70	51	87	60	59	58	65	63	70
120	98	175	130	125	123	140	133	146

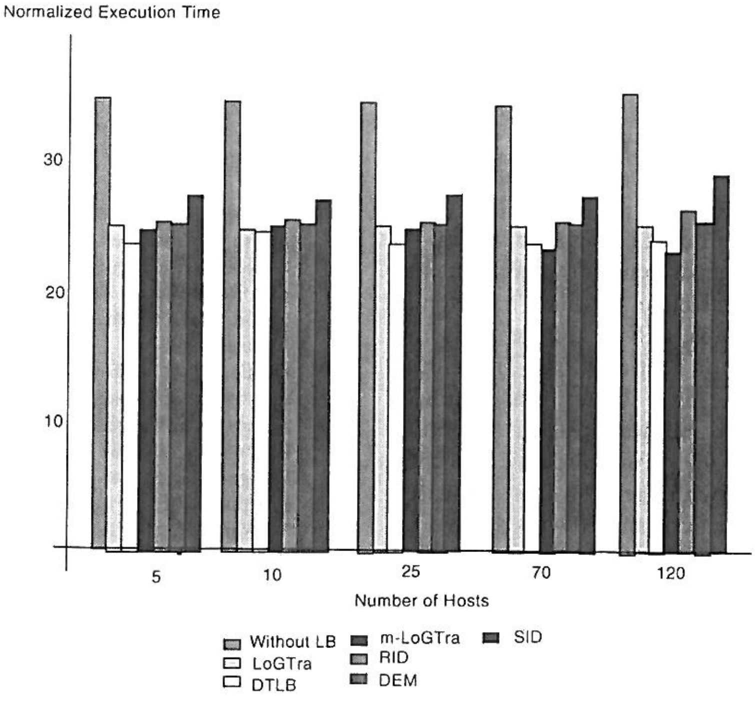


Fig. 13. Multiple token generation (performance with reference to number of hosts).

7.1. Results

This section gives simulation results for *m-LoGTra*. The comparison of this algorithm with *DTLB* and *LoGTra* is also given.

The comparison is made with increasing network size and it is observed that the performance of DTLB is better in most of the cases as compared to *LoGTra* and *Multiple token policy* for the small network sizes. Refer Table 5 which gives the performance of algorithms with respect to network size. For large networks *multiple token policy* gives better performance. In small network sizes overheads added by *m-LoGTra* compensate the advantages but that is not the case with large networks. For small networks DTLB can be preferred over *m-LoGTra*. As the number of tokens to be generated by a node is decided dynamically *m-LoGTra* can be preferred over *DTLB* and *LoGTra* for high task arrival rate. Similarly for real time systems *m-LoGTra* has an edge over other algorithms.

Fig. 13 clearly shows that the performance of *LoGTra* can be improved using *DTLB* and *multiple token approach*.

8. Conclusion

As in this algorithm local minima along with local maxima can generate tokens. better load balancing can be achieved. There is possibility in *LoGTra* that it may not able to treat local minimas effectively, especially in case of large networks.

As tokens can be generated by the nodes at the local minima and local maxima which have loads below *LWM* and above *HWM* this restricts the transfers those are not cost-effective. This limits overheads and also gives better performance. Algorithm *m-LoGTra* allows hosts to generate multiple tokens. This helps in searching a proper node to balance load with the host. In many cases with this option load can be shared with more than one host, which helps in balancing load equally among all the hosts. This gives better results for larger networks. Concluding remarks:

1. *DTLB* can be preferred when graph density is less and network size is small.
2. *m-LoGTra* gives better performance for large networks.

As this method gives load balancing beyond neighbors and takes care of local maxima as well as minimas it can be preferred over other methods for larger network size. The algorithm can be further improved to include automatic selection of number of tokens generated by any host. Also use of the same algorithm for real time applications can be explored.

References

[1] M. Livy, M. Melman, Load balancing in homogeneous broadcast distributed systems, in: Proc. Conf. Performance, ACM, vol. 99, January 1982, pp. 47–55.

[2] P. Kulkarni, I. SenGupta, A new approach for load balancing using differential load measurements (dlm), in: Proc. IEEE International Conference on Information Technology, ITCC'2000, LasVegas, USA, March 2000, pp. 355–359.

[3] J. Tereso, PhD dissertation, Rensselaer Polytechnic Institute, 2001.

[4] P. Kulkarni, I. SenGupta, Load graph based transfer method (LoGTra) for dynamic load balancing on network of of workstations, in: Proc. International Conference on Parallel and Distributed Systems PDCS'98, Las Vegas, USA, October 1998, pp. 189–191.

[5] P. Kulkarni, I. SenGupta, Dual token based load balancing (DTLB), in: Proc. International Conference on Applied Informatics, AI'2000, Insbruch, Austria, February 2000.

[6] C. Grantz, R. Silverman, S. Stuart, A distributed batching system for parallel processing, Software Practice and Experience 7 (January) (1989).

[7] K.P. Bubendorfer, Resource based policies for load distribution, PhD thesis, Victoria University of Wellington, August 1996.

[8] D. Eager, E. Lazowaska, J. Jahorjan, Adaptive load sharing in homogeneous distributed systems, IEEE Transactions on Software Engineering 5 (May) (1986) 662–675.

[9] F. Douglas, Ousterhout, Transparent process migration: Design alternatives and implementation ons, Software Practice and Experience 8 (August) (1991) 757–785.

[10] Y. Wang, R. Morris, Load sharing in distributed systems, IEEE Transactions on Computers C-34 (March) (1985) 204–217.

[11] M. Powell, B. Miller, Process migration in demos/mp, in: Proceedings 9th ACM Symposium on Operating System Principles, vol. 17:5, October 1987, pp. 110–118.

[12] M. Theimer, K. Lantz, D. Cheriten, Preemptible remote execution facilities for v system, in: Proceedings ACM

Symposium on Operating System Principles, December 1985, pp. 2–12.

[13] M. Deverkonda, R. Iyer, Predictability of process resource usage: A measurement based study of unix, IEEE Transactions on Software Engineering 15 (December) (1989).

[14] K. Goswami, M. Deverkonda, R. Iyer, Prediction based dynamic load sharing heuristics, IEEE Transactions on Parallel and Distributed Systems 4 (June) (1993) 638–648.

[15] S. Zhou, X. Zheng, J. Wang, P. Delisle, Utopia: A load sharing facility for large, heterogeneous distributed computer systems, Software: Practice and Experience 23 (December) (1993) 1305–1336.

[16] N. Shivratri, P. Krueger, M. Singhal, Load distribution for locally distributed systems, IEEE Transactions on Computers (December) (1992) 33–44.

[17] P. Krueger, R. Finkel, An adaptive load balancing algorithm for multicomputers, Technical Report at University of Wisconsin, vol. 39, April 1984, p. 539.

[18] L. Ni, C. Xu, T. Gendreau, A distributed algorithm for load balancing, IEEE Transactions on Software Engineering SE-10 (October) (1985) 1153–1161.

[19] V. Salatore, A distributed and adaptive dynamic load balancing scheme for parallel processing of medium grain tasks, in: Proceedings of Fifth Distributed Memory Computers Conf., April 1990, pp. 995–999.

[20] P. Krueger, M. Liveny, The diverse objectives of distributed scheduling policies, in: Proc. Seventh International Conference on Distributed Computing Systems, IEEE C.S. Press, Los Alamitos, March 1987, pp. 242–249.

[21] F. Lin, R. Keller, The gradient model load balancing method, IEEE Transactions on Software Engineering SE-13 (January) (1987) 32–38.

[22] G. Cybenko, Dynamic load balancing for distributed memory multiprocessors, Journal on Parallel and Distributed Computing 7 (October) (1989) 279–301.

[23] K. Dragon, J. Gustafson, A low cost hypercube load balancing algorithm, in: 4th International Conference Hypercube Concurrent Computers and Applications, April 1989, pp. 583–590.

[24] M. Willebeek, L. Mair, A. Reeves, Strategies for dynamic load balancing on highly parallel computers, IEEE Transactions on Parallel and Distributed System 4 (September) (1993) 979–993.

[25] P. Mehara, Benjamin Wah, Synthetic workload generation for load balancing experiments, IEEE Parallel and Distributed Technology (1995) 4–19.

[26] M. Mitzenmacher, On the analysis of randomized load balancing schemes, Theory of Computing System 32 (1999) 361–386.

[27] W. Shu, L. Kale, A dynamic load balancing strategy for the chare kernel systems, in: Proc. ACM Super-computing Conference, 1989, pp. 995–999.

[28] L. Kale, Comparing the performance of two dynamic load distribution methods, in: Proc. Conf on Parallel Processing, 1988, pp. 8–12.

[29] W. Shu, Chare kernel and its implementation, PhD thesis, University of Illinois, 1990.

[30] M.J. Zaki, W. Li, S. Parthsarathy, Customized dynamic load balancing on network of workstations, Journal of Parallel and Distributed Computing 43 (1997) 156–162.

[31] Z. Lun, V. Taylor, Dynamic load balancing samr in distributed computation, Available from: <http://www.sc200/org/papers>, 2000.

[32] S. Muthukrishnan, R. Rajaraman, An adversarial model for distributed dynamic load balancing, The Journal of Interconnection Networks 3 (2) (2002) 35–47.

[33] A. Cortes, E.A. Ripoll, M. Senar, On convergence of sid and dasud load-balancing algorithms, Available from: <http://pirdi.uab.es/document/pirdi8.ps>, 1998.

[34] A. Corradi, L. Leonardi, F. Zambonelli, On effectiveness of different diffusive load balancing policies in dynamic applications, Euromicro Workshop on Parallel and Distributed Processing (PDP-99), no. 1, 1999.

[35] A. Cortes, E.A. Ripoll, M. Senar, Performance comparison of dynamic load-balancing strategies for distributed computing, Hawaii International Conference on System Sciences, 1999.

[36] M. Cierniak, M.J. Zaki, W. Li, Compile time scheduling on heterogeneous network of workstations, Presentation (1997).

[37] M. Richmond, M. Hitchens, A new process migration algorithm, Technical Report, Basser Dept. of Computer Science, University of Sidney, Australia, vol. 509, September 1996, pp. 31–32.

[38] M. Cierniak, W. Li, M.J. Zaki, Loop scheduling for heterogeneity, in: Proceedings of Fourth International Symposium on High Performance Distributed Computing, August 1995, p. 78.

[39] T. Chou, J. Abraham, Load redistribution under failure in distributed systems, IEEE Transactions on Computers C-32 (September) (1988) 799–808.

[40] J. Kurose, R. Chipalkatti, Load sharing in soft real time distributed computers systems, IEEE Transactions on Computers C-36 (August) (1987) 993–1000.

[41] W. Giffin, Transform Techniques for Probability Modeling, Academic Press, 1975.

[42] D. Gross, D. Harris, Fundamentals of Queuing Theory, Wiley, New York, 1974.

Dr. Parag Kulkarni is Ph.D. form IIT, Kharagpur (www.iitkgp.ernet.in). He is working in IT industry for more than 13years. He is on research panel and Ph.D. guide for University of Pune, BITS and Symbiosis deemed University. He has conducted five tutorials at various international conferences and was a keynote speaker for three international conferences. He has also worked as a referee for International Journal for Parallel and Distributed Computing, IASTED conferences. He is member of IASTED technical committee. Presently he is Chief scientist and Research Head at Capsilon Research Labs, Capsilon India, Pune. His areas of interest include image processing, security systems, decision systems, expert systems, classification techniques, load balancing and distributed computing.

Dr. Indranil SenGupta is faculty member of Computer Department at IIT Kharagpur for more than 15 years. Presently he is working as a Professor in Computer Department at IIT, Kharagpur. He has more than 50 international publications. His areas of interest include VLSI, security systems, Computer Networks, Fault tolerance, load balancing and distributed computing.

5

VK Pachghare and VA Patole and P Kulkarni, "Self Organizing Maps to Build Intrusion Detection System, International Journal of Computer Applications, Foundation of Computer Science, USA, Vol.1, No.8, pp 1-4, 2010

Self Organizing Maps to Build Intrusion Detection System

Mr. Vivek A. Patole
Dept. of Computer Engineering &
Information Technology,
College of Engineering Pune,
Pune, India

Mr. V. K. Pachghare
Assistant Professor,
Dept. of Computer Engineering &
Information Technology,
College of Engineering Pune,
Pune, India

Dr. Parag Kulkarni
Chief Scientist and Research Head
Capsilon Research Labs,
Capsilon India,
Pune, India

ABSTRACT

With the rapid expansion of computer usage and computer network the security of the computer system has became very important. Every day new kind of attacks are being faced by industries. Many methods have been proposed for the development of intrusion detection system using artificial intelligence technique. In this paper we will have a look at an algorithm based on neural networks that are suitable for Intrusion Detection Systems (IDS) [1] [2]. The name of this algorithm is “Self Organizing Maps” (SOM). Neural networks method is a promising technique which has been used in many classification problems. The neural network component will implement the neural approach, which is based on the assumption that each user is unique and leaves a unique footprint on a computer system when using it. If a user’s footprint does not match his/her reference footprint based on normal system activities, the system administrator or security officer can be alerted to a possible security breach. At the end of the paper we will figure out the advantages and disadvantages of Self Organizing Maps and explain how it is useful for building an Intrusion Detection System.

Keywords: Neural Networks, Intrusion Detection System, Self Organizing Maps.

1. INTRODUCTION

Over the last few decades information is the most precious part of any organization. Most of the things what an organization does revolve around this important asset. Organizations are taking measures to safeguard this information from intruders. The rapid development and expansion of World Wide Web and local networks and their usage in any industry has changed the computing world by leaps and bounds [1][2].

INTRUSION DETECTION SYSTEM is a system that identifies, in real time, attacks on a network and takes corrective action to prevent them. They are the set of techniques that are used to detect suspicious activity both at network and host level. There are two main approaches to design an IDS.

1) MISUSE BASED IDS (SIGNATURE BASED)
2) ANOMALY BASED IDS.

In a misuse based intrusion detection system, intrusions are detected by looking for activities that correspond to know signatures of intrusions or vulnerabilities [3]. While an anomaly based intrusion detection system detect intrusions by searching for abnormal network traffic. The abnormal traffic pattern can be defined either as the violation of accepted thresholds for frequency of events in a connection or as a user’s violation of the legitimate profile developed for normal behavior.

One of the most commonly used approaches in expert system based intrusion detection systems is rule-based analysis using Denning’s profile model [3]. Rule-based analysis depends on sets of predefined rules that are provided by an administrator. Expert systems require frequent updates to remain current. This design approach usually results in an inflexible detection system that is unable to detect an attack if the sequence of events is slightly different from the predefined profile [4]. Considered that the intruder is an intelligent and flexible agent while the rule based IDSs obey fixed rules. This problem can be tackled by the application of soft computing techniques in IDSs. Soft computing is a general term for describing a set of optimization and processing techniques. The principal constituents of soft computing techniques are Fuzzy Logic (FL), Artificial Neural Networks (ANNs), Probabilistic Reasoning (PR), and Genetic Algorithms (GAs) [4].

SYSTEM ARCHITECTURE

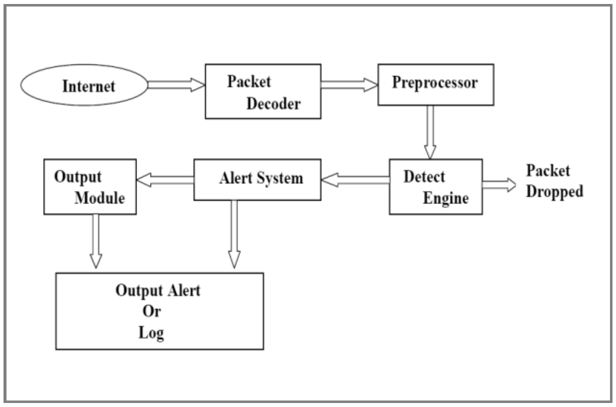


Figure 1. System Architecture.

2. TYPES OF NETWORKING ATTACKS

There are four major categories of networking attacks. Every attack on a network can be placed into one of these groupings [4].

- **Denial of Service (DoS):** A DoS attacks is a type of attack in which the hacker makes a memory resources too busy to serve legitimate networking requests and hence denying users access to a machine e.g. apache, smurf, Neptune, ping of death, back, mail bomb, UDP storm, etc.
- **Remote to User attacks (R2L):** A remote to user attack is an attack in which a user sends packets to a machine over the internet, and the user does not have access to in order to expose the machines vulnerabilities and exploit privileges which a local user would have on the computer, e.g. xlock, guest, xnsnoop, phf, sendmail dictionary etc.
- **User to Root Attacks (U2R):** These attacks are exploitations in which the hacker starts off on the system with a normal user account and attempts to abuse vulnerabilities in the system in order to gain super user privileges, e.g. perl, xterm.
- **Probing:** Probing is an attack in which the hacker scans a machine or a networking device in order to determine weaknesses or vulnerabilities that may later be exploited so as to compromise the system. This technique is commonly used in data mining, e.g. satan, saint, portsweep, mscan, nmap etc.

Two different attack types were included for this study: *SYN Flood (Neptune)* and *Satan*. These two attack types were selected from two different attack categories (denial of service and probing) to check for the ability of the intrusion detection system to identify attacks from different categories.

SYN Flood (Neptune) is a denial of service attack to which every TCP/IP implementation is vulnerable (to some degree). For distinguishing a Neptune attack, network traffic is monitored for a number of simultaneous SYN packets destined for a particular machine. *Satan* is a probing intrusion, which automatically scans a network of computers to gather information or find known vulnerabilities. The purpose of classifiers in IDSs is to identify attacks from all four groups as accurately as possible.

3. SELF ORGANIZING MAPS (SOM)

The Self-Organizing Map is a neural network model for analyzing and visualizing high dimensional data. It belongs to the category of competitive learning network. The SOM Fig. 1 defines a mapping from high dimensional input data space onto a regular two-dimensional array of neurons. It is a competitive network where the goal is to transform an input data set of arbitrary dimension to a one- or two-dimensional topological map. The model was first described by the Finnish professor Teuvo Kohonen and is thus sometimes referred to as a Kohonen Map. The SOM aims to discover underlying structure, e.g. feature map, of the input data set by building a topology preserving map which describes neighborhood relations of the points in the data set [5].

The SOM is often used in the fields of data compression and pattern recognition. There are also some commercial intrusion detection products that use SOM to discover anomaly traffic in networks by classifying traffic into categories. The structure of the SOM is a single feed forward network, where each source node of the input layer is connected to all output neurons. The number of the input dimensions is usually higher than the output dimension.

The neurons of the Kohonen layer in the SOM are organized into a grid, see figure 2 and are in a space separate from the input space. The algorithm tries to find clusters such that two neighboring clusters in the grid have codebook vectors close to each other in the input space. Another way to look at this is that related data in the input data set are grouped in clusters in the grid [5]. The training utilizes competitive learning, meaning that neuron with weight vector that is most similar to the input vector is adjusted towards the input vector. The neuron is said to be the ’winning neuron’ or the Best Matching Unit (BMU). The weights of the neurons close to the winning neuron are also adjusted but the magnitude of the change depends on the physical distance from the winning neuron and it is also decreased with the time.

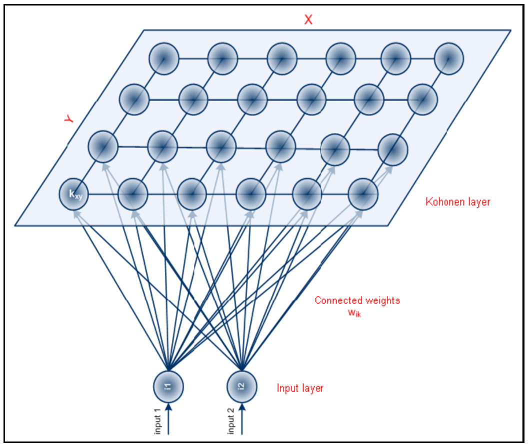


Figure 2. The self-organizing (Kohonen) map

The learning process of the SOM goes as follows:

1. One sample vector x is randomly drawn from the input data set and its similarity (distance) to the codebook vectors is computed by using Euclidean distance measure [6]:

$$\|x - m_c\| = \min_i \{\|x - m_i\|\} \text{ -------- (1)}$$

2. After the BMU has been found, the codebook vectors are updated. The BMU itself as well as its topological neighbors are moved closer to the input vector in the input space i.e. the input vector attracts them. The magnitude of the attraction is governed by the learning rate. As the learning proceeds and new input vectors are given to the map, the learning rate gradually decreases to zero according to the specified learning rate function type. Along with the learning rate, the neighborhood radius decreases as well. The update rule for the reference vector of unit i is the following:

$$m_i(t+1) = \begin{cases} m_i(t) + \alpha(t)[x(t) - m_i(t)], i \in N_c(t) \\ m_i(t), i! \in N_c(t) \end{cases}$$

(2)

3. The steps 1 and 2 together constitute a single training step and they are repeated until the training ends. The number of training steps must be fixed prior to training the SOM because the rate of convergence in the neighborhood function and the learning rate are calculated accordingly.

After the training is over, the map should be topologically ordered. This means that *n* topologically close input data vectors map to *n* adjacent map neurons or even to the same single neuron.

4.1 Mapping Precision

The mapping precision measure describes how accurately the neurons respond to the given data set. If the reference vector of the BMU calculated for a given testing vector xi is exactly the same xi, the error in precision is then 0. Normally, the number of data vectors exceeds the number of neurons and the precision error is thus always different from 0. A common measure that calculates the precision of the mapping is the average quantization error over the entire data set [6]:

$$E_q = \frac{1}{N}\sum_{i=1}^{n} ||x_i + m_c||$$

(6)

4.2 Topology Preservation

The topology preservation measure describes how well the SOM preserves the topology of the studied data set. Unlike the mapping precision measure, it considers the structure of the map. For a strangely twisted map, the topographic error is big even if the mapping precision error is small. A simple method for calculating the topographic error [6]:

$$E_q = \frac{1}{N}\sum_{i=1}^{n} u(x_x)$$

(7)

Where $u(x_x$) is 1 if the first and second BMUs of x_k are not next to each other. Otherwise $u(x_x)$ is 0.

4. EXPERIMENTS

5.1 Data Collection

If the network traffic has been examined carefully for different types of events such as downloading, port scanning, surfing etc., it is possible to identify the formal distinctions between them. The idea behind this work is to collect distinct and various kinds of network packets. To collect data we can use any packet sniffer which is available readily. Here in this case we have developed our own packet sniffer. Apart from capturing live packets we also a standard DARPA dataset, which we will be using for training purpose [7]. The dataset contain both packets with intrusion and without intrusion.

5.2 Vector Extraction

After the process of data collection, the features should be extracted from the data in order to obtain better classification results [8]. We slide time one by one, obtain new vectors and these obtained vectors will be used to provide input to self organizing map. Table 1. Shows examples of obtained vectors for time window length is equal to 5. We accepted window length as 20 for our application. Since the data are collected in every 20 seconds an input vector corresponds to time interval of 400 seconds

V1	V2	V3	V4	V5
16.1	14.3	24	125	128
14.6	25	125	133	11.2
23	122	134	10.5	16
122	129	10.3	18	19
132	10.6	16	19	15.5

Table 1. Table showing Extracted Vectors.

5.3 Training Self Organizing Map

For training purpose we constructed a 30x30 Self Organizing Map in order to perform clustering. The data that was used for it was DARPA dataset [7]. We used batch training algorithm with training length 100 and starting radius 15. Self organizing map was largely successful in classifying the IP packets.

5. RESULTS

After the data collection, vector extraction and training of the Self Organizing Maps we pass the packets through the SOM. The result is shown in the fig. The results Fig. 3 show input vectors classification, which represents behavior and its mapping to particular neurons, which form single possible user behavior states. Form states as intrusion – Intrusion, possible intrusion – Intrusion? Normal – Norm. From the test result SOM network represents suitable core for IDS systems [9] [10].

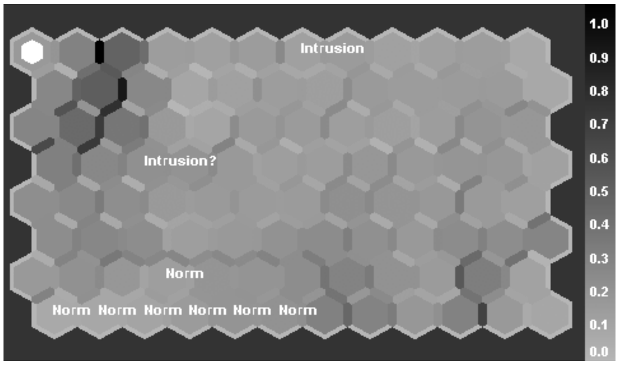


Figure 3. Results of the Experiment

6.1 Advantages and Disadvantages of SOM

6.1.1 Advantages:

1. Simple and easy-to-understand algorithm that works.
2. Topological clustering
3. Unsupervised algorithm that works with nonlinear data set.
4. The excellent capability to visualize high- dimensional data onto 1 or 2 dimensional space makes it unique especially for dimensionality reduction.

6.1.2 Disadvantages:

Time consuming algorithm, this is because as the no. of neurons affects the performance of the algorithm. And as the number increases the computation increases which results in increasing computational time [6] [7].

6. CONCLUSIONS

The Self Organizing Map is an extremely powerful mechanism for automatic mathematical characterization of acceptable system activity. In the above paper we have described how we can use Self Organizing Maps for building an Intrusion Detection System. We have explained the system architecture and the flow diagram for the SOM. We have also presented the pros and cons of the algorithm.

Our actual experiments show that even a simple map, when trained on normal data, will detect the anomalous features of both buffer overflow intrusions to which we exposed it. This approach is particularly powerful because the self organizing map never needs to be told what intrusive behavior looks like [11]. By learning to characterize normal behavior, it implicitly prepares itself to detect any aberrant network activity.

7. REFERENCES

[1] Damiano Bolzoni, Sandro Etalle, Pieter H. Hartel, and Emmanuele Zambon. Poseidon: a 2-tier anomaly-based network intrusion detection system. In Proceedings of the 4th IEEE International Workshop on Information Assurance, 13-14 April 2006, Egham, Surrey, UK, pages 144–156, 2006.

[2] D. A. Frincke, D. Tobin, J. C. McConnell, J. Marconi, and D. Polla. A framework for cooperative intrusion detection. In Proc. 21st NIST-NCSC National Information Systems Security Conference, pages 361–373, 1998.

[3] Denning D, “An Intrusion-Detection Model”, IEEE Transactions on Software Engineering, Vol. SE-13, No 2, Feb 1987.

[4] Simon Haykin, “Neural Networks: A Comprehensive Foundation”, Prentice Hall, 2nd edition, 1999.

[5] Kohonen, T, “Self-Organizing Maps”, Springer Series in Information Sciences. Berlin, Heidelberg: Springer. 1997.

[6] P. Lichodzijewski, A. Zincir-Heywood, and M. Heywood. “Dynamic intrusion detection using self organizing maps”, 2002.

[7] McHugh, J.: Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by lincoln laboratory. ACM Trans. on Information and System Security 3 (2000) 262–294.

[8] Wenke Lee and Salvatore J. Stolfo, “A framework for constructing features and models for intrusion detection systems”, ACM Trans. Inf. Syst. Secur., 3(4):227–261, 2000.

[9] Rhodes, B., Mahaffey, J., Cannady, J., “Multiple Self-Organizing Maps for Intrusion Systems”

[10] Bishop, C. M, “Neural Networks for Pattern Recognition”, Oxford: Clarendon-Press, 1996.

[11] Lane, T., and Brodley, C. E. 1999. Temporal sequence learning and data reduction for anomaly detection. ACM Transactions on Information and System Security 2(3):295—331.

6

Butalia, A.K. Ramani,. Parag Kulkarn, "Emotional Recognition and towards Context based Decision", Number 3 - Article 8, International Journal of Computer Applications, Foundation of Computer Science, USA, 2010, pp 42-54

Emotional Recognition and towards Context based Decision

Ayesha Butalia
MIT College of Engg
Pune, India

Dr. A.K. Ramani
Devi Ahilya University,
Indore, India

Dr. Parag Kulkarni
Capsilon,
Pune, India

ABSTRACT

Non-verbal communication may be used to enhance verbal communication or even provide developers with an alternative for communicating information.

Emotion or Gesture recognition is been highlighted in the area of Artificial Intelligence and advanced machine learning. Emotion or gesture is an important feature for an intelligent Human Computer Interaction. This paper basically is a literature survey paper which reveals with the research work already dealt with in this area. Facial expression has been concluded as the most important part involved in it. Even Facial features are also distinguished out of which eyes and mouth is probably more prominent. Neural networks are the widely used. Approaches towards Rough Fuzzy definition can be probably resolve the complexity. Context based recognition can be added so as to resolve the ambiguity involved in different scenarios.

Keywords

Facial expression, emotion recognition, rough sets, fuzzy sets, neural networks, context.

1. INTRODUCTION

Emotion recognition is very important for human-computer intelligent interaction. Facial expressions, gestures, and body postures can portray emotions in a non-verbal way. These methods are frequently employed by actors in theatrical plays or in movies, or even by virtual characters such as those found in computer games, animated storybooks, and website e-assistants. Signals for emotion expressions (“cues”), such as a raised fist and narrowing of the eyes, substantially influence the viewers’ assumptions on the emotional state of the person portraying it. To enable humans to recognize emotions of virtual characters, the characters’ cues must be portrayed according to the human counterparts. Previous research has shown that emotions can be effectively portrayed through non-verbal means [Atkinson et al. 2004; Coulson 2004; Ekman 2003].

Emotion Recognition is generally performed on facial or audio information by artificial neural network, fuzzy set, support vector machine, hidden Markov model, and so forth. Although some progress has already been made in emotion recognition, several unsolved issues still exist. For example, it is still an open problem which features are the most important for emotion recognition. It is a subject that was seldom studied in computer science. However, related research works have been conducted in cognitive psychology. In recent years, there has been a growing interest in improving all aspects of the interactions between humans and computers. It is argued that to truly achieve effective human-computer intelligent interaction (HCII), there is a requirement for computers to be able to interact naturally with users, similarly to the way human-human interaction. HCII is becoming more and more important in such applications as smart home, smart office, and virtual reality, and it will be popular in all aspects of daily life in the future. To achieve the purpose of HCII, it is essential for computers to recognize human emotion and to give a suitable feedback. Consequently, emotion recognition attracts significant attention in both industry and academia. There are several research works in this field in recent years and some successful products such as AIBO, the popular robot dog produced by Sony.

In the first section of the study, we first look at existing systems for synthesizing cues for facial expressions, gestures and body postures. This is followed by examining the emotion recognition problems that arise from utilizing the various systems. Lastly, the systems are integrated together and the implications that arise from the integration are analyzed.

The second section deals with the comparison of Roles of Postures, Facial and Gestures in Emotion Recognition The study showed that hand gestures aided in the emotion recognition rate for postures which others [Coulson 2004] had previously assumed as unimportant. Additionally, it was discovered that the emotion recognition rate using gestures can be greatly improved when emblematic actions are combined with functional actions. This study also confirmed our assumption that an integrated system covering facial expression, gesture and body postures is capable of enhancing the emotion recognition rate beyond the rates of the single systems.

Usually, emotion recognition is studied by the methods of artificial neural network (ANN), fuzzy set, support vector machine (SVM), hidden Markov model (HMM), and based on the facial or audio features, and the recognition rate often arrives at 64% to 98% [1–3]. In the third section, the comparison of Rough Set theory, Adaptive Neuro - Fuzzy Inference Systems (ANFIS) and Rough – ANFIS approach is analyzed.

Lastly, the section describes the context approach towards facial recognition.

2. FUNDAMENTALS OF EMOTION EXPRESSION

This section looks at how cues are able to produce emotions. In addition, it provides an overview of how certain factors may affect emotion recognition in a 3D virtual agent.

2.1 Cues

Cues are non-verbal signals involving either the movement/positioning of individualized parts of the body or the movement/positioning of a group of body parts in concert with each other [Ekman 1978]. Cues such as a raised fist and narrowing of the eyes can indicate that the individual is angry. Movement/positioning classes like facial expressions, body postures, and gestures can involve one or more of these cues, which people (subconsciously) use for interpreting the emotional state.

2.2 Facial Expressions:

Facial expressions involve facial cues that are displayed using body parts from the head region (e.g., eyebrows, mouth, lips). Common facial expressions such as the raising of the lips (facial cue) as part of a smile (facial expression) is interpreted by others to be a display of emotion of the actor; happiness in this example [Ekman 1978].

2.3 Body Posture:

Body cues involved in body postures are displayed using body parts such as the torso, arms and legs. They are another component of non-verbal emotion expression. For example, the clenching of a fist and raising it to appear like the actor is trying to attack someone is usually interpreted by others as a display of anger [Ekman 2003].

2.4 Gestures and Actions:

Gestures are actions/movements of body parts and they are another component of non-verbal communication of emotion. For example, a high frequency gesture such as jumping up and down very quickly can be interpreted by others to be a sign of happiness [Raouzaiou et al. 2004].

2.5 Factors Affecting Emotion Recognition:

Although the exact factors which can influence the interpretation of emotion have not yet been thoroughly researched upon, four factors have recently surfaced based on current experiments and research. They are gender, job status, culture, and age.

Men and women express emotions differently [Brody and Hall 1992] in terms of the frequencies of occurrence (men often experience anger more often than women). It was also proven that recognition of ambiguous facial expressions is influenced by the gender of the person performing it [Condry 1976; Devine et al. 2000] whereby “masculine” emotions (e.g., anger) are assigned to men while “feminine” emotions (e.g., happiness) are assigned to women. As such, there is a need to be mindful of these gender stereotypes
when trying to synthesize emotions.

Stereotypes of job status are known to exist too [Algoe et al. 2000]. For example, managers are often associated with “masculine” emotions and character traits while nurses are associated with “feminine” emotions and character traits. If the virtual agent is assigned human jobs (usually identified by the type of uniform they are wearing), ambiguous emotion expressions may lead others to wrongly assign “masculine” and “feminine” emotions to it. Culture can also affect the interpretation of emotions [Bianchi-Berthouze el al. 2006]. It was discovered that the Japanese are less animated in their body expressions for emotion than the Sri Lankans leading to the same emotion being read differently.

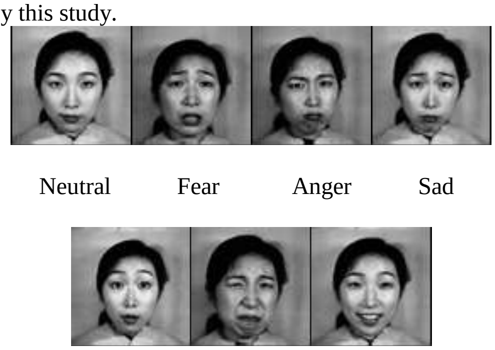
Lastly, there is neurological evidence to suggest that age can affect the interpretation of emotions. It was shown that people in the 60-80 years old age group tend to suffer from emotion processing impairments and therefore require stronger or more cues to be displayed before being able to associate an emotion.

2.6 Emotion Blending and Transition:

Human beings are capable of feeling multiple emotions simultaneously [Ekman 2003]. These emotions may transition/morph in time from one state of emotion to another (e.g., a loud noise may suddenly cause a passerby to feel surprise momentarily, which might later transition into a feeling of fear if the passerby feels that his/her life is in eminent danger), or they may also be displayed at the same point in time (e.g., the loss of a loved one in a car accident may cause a person to feel both angry and upset at the same time). Emotion blending is the mechanism by which multiple emotional expressions are altered or combined simultaneously to convey more subtle information about the performer. Unfortunately, such a process is a complicated one and has not yet been well understood and researched by behavioral psychologists and animators.

3. CONCEPTS FOR REPRESENTING AND MODELING FACIAL EXPRESSIONS, POSTURES, AND GESTURES

This section looks at the approaches taken by others to represent and model emotion expression. It also describes the approach taken by this study.

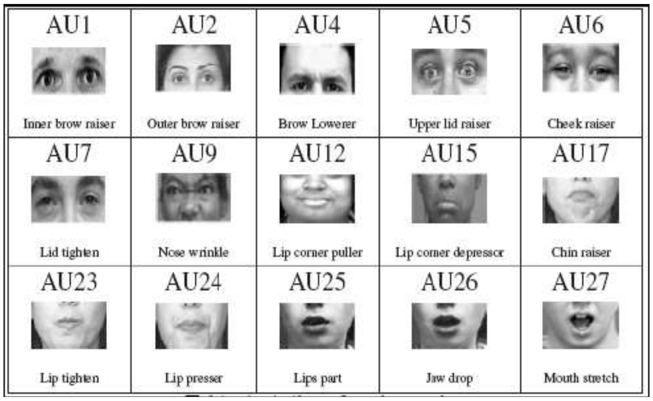


Neutral Fear Anger Sad

Surprise Disgust Happy

Figure 1. 6 Basic Emotions and Neutral Expression

Table 1a. Some AU and their associated facial change obtained from Ekman’s study [Ekman 1978].



AU1	AU2	AU4	AU5	AU6
Inner brow raiser	Outer brow raiser	Brow Lowerer	Upper lid raiser	Cheek raiser
AU7	AU9	AU12	AU15	AU17
Lid tighten	Nose wrinkle	Lip corner puller	Lip corner depressor	Chin raiser
AU23	AU24	AU25	AU26	AU27
Lip tighten	Lip presser	Lips part	Jaw drop	Mouth stretch

Table 1b. Table of the six basic emotions and the AUs involved.

Basic Expressions	Involved Action Units
Surprise	AU1, 2, 5, 15, 16, 20, 26
Fear	AU1, 2, 4, 5, 15, 20, 26
Disgust	AU2, 4, 9, 15, 17
Anger	AU2, 4, 7, 9, 10, 20, 26
Happiness	AU1, 6, 12, 14
Sadness	AU1, 4, 15, 23

3.1 Representing and Modeling Postures

There exist a variety of sources which offer more or less detailed descriptions of emotional postures [Birdwhistell 1975; Boone and Cunningham 2001; Darwin 1872]. For instance, in the descriptions put forward by these authors, anger is described as involving a jutting chin, angular body shape, forward weight transfer, chest out and angled forwards, and a bowed head. Unfortunately, prior to Coulson's study [Coulson 2004], no formal research has been done to quantify the anatomical features which produce the emotional posture (i.e., posture was mostly descriptively documented).

Coulson found that the anatomical descriptions obtained from the studies mentioned earlier could be translated into joint rotations, which he then attempted to quantify via testing on human volunteers. The approach taken by this study to modeling postures relies on Coulson's findings for the angle representation of each emotion as it is the only study which quantifies the respective joint angles.

Gestures in this study are modeled by animating the virtual agent since gestures are essentially non-static displays of emotion. As there is a relative paucity of studies on dynamic emotion gestures [Atkinson et al. 2004], the approach taken by this study relies on actors' knowledge of gestures.

Raouzaiou et al. (2004) and Atkinson et al. (2004) formulated a short table of emotions that depict general hand and head gestures for each emotion. This study relies upon those data to provide a basic framework for generating gestures as it is also compatible with the joint-angle system used for modeling of posture.

Table 2. Table of gestures extracted from Raouzaiou's study [Raouzaiou et al. 2004].

Emotion	Gesture Class
Joy	Hand clapping-high frequency
Sadness	Hands over the head-posture
Anger	Lift of the hand-high speed
	Italianate gestures
Fear	Hands over the head-gesture
	Italianate gestures
Disgust	Lift of the hand-low speed
	Hand clapping-low-frequency
Surprise	Hands over the head gesture

Table 3. Recognition rate from integrating posture and facial expression.

Emotion(Postures)	No. of correct recognition	Emotion Recognition rate
Happy	16	100.00%
Anger	16	100.00%
Sad	16	100.00%
Surprise	9	56.25%
Fear	14	87.50%
Disgust	1	6.25%

Table 4. Recognition rate from facial expressions alone.

Emotion(Postures)	No. of correct recognition	Emotion Recognition rate
Happy	16	100.00%
Anger	16	100.00%
Sad	13	81.25%
Surprise	10	62.50%
Fear	11	68.75%
Disgust	2	12.50%

Table 5: Recognition rate from integrating gesture, posture, and facial expression.

Emotion(Postures)	No. of correct recognition	Emotion Recognition rate
Happy	16	100.00%
Anger	16	100.00%
Sad	16	100.00%
Surprise	4	25.00%
Fear	12	75.00%
Disgust	2	12.50%

Conclusion: Infact Gesture lowered the recognition rate when added to Posture and Facial features. There was not much difference in the recognition through facial rather than from facial and posture together. Infact posture recognition would add much complexity in real life recognition. A single error to it could make lot of difference and consuming time also.

4. CLASSIFICATION BASED ON RS, ANFIS, RS-ANFIS FOR FACIAL EXPRESSION:

Models and automated systems have been created to recognize the emotional states from facial expressions. The leading method Facial Action Coding System [23], for measuring facial movements in behavioral science was developed by Ekman and Friesen in 1977. Other methods such as electromyography, which directly measures the electrical signals generated by the facial muscles and deducing the facial behavior from it, are both obtrusive non-comprehensive. According to the survey [7], FACS is the leading method for measuring facial expression in behavioral science. It uses 46 defined Action Units to correspond into each independent motion of the face. However this method takes over 100 hours of training to achieve minimal competency for a human expert [8]. Faster automation approaches, such as measurement of facial motion through optic flow [9, 10] and analysis of surface textures based on principal component analysis (PCA) [11]. Newer techniques include using Gabor wavelets [12], linear discriminant analysis [13], local feature analysis [14], and independent component analysis [15]. The techniques are

benchmarked [8] and best classification accuracy of about 95% for the recognition of the twelve facial actions, was obtained using Gabor filter representation. Human experts and
naïve human tester were benchmarked as well; scored about 94% and 78% respectively, and experiments were supported by Zhang et. al. [16].

Most of these systems use a set of feature vectors to represent facial images, without describing the relationship between the feature vectors. A method for emotion recognition by transforming the feature vector data into tree structure representation, which encodes the feature relationship information among the face features was then proposed **[Automated knowledge engg].** Sixty Localized Gabor Features (LGF) and one Global Gabor Feature are obtained as a feature vector and transforming them into a Facial Emotion Tree Structure (FETS) representation. Tsoi [17] proposed using tree structures to preserve and make use of these relationships and processing them by specific machine learning models [1, 18-20]. Cho and Wong proposed using Gabor features in tree structure representation for face recognition with achieving high accuracy rate [1]. Gabor Feature extraction makes use of Gabor wavelets, which capture the properties of spatial localization, orientation selectivity, spatial frequency selectivity, and quadrature phase relationship, seem to be a good approximation to filter response profiles encountered experimentally in cortical neurons. A probabilistic based recursive neural network is proposed for classification of the FETS in this paper. This method is benchmarked against Support Vector Machines (SVM) [21], K nearest neighbors (KNN) [22], Naïve Bayes algorithm [23] where the flat vector representations were used in the recognition experiments. QuadTree tree structure processed using our probabilistic recursive neural network is also benchmarked. We made use of the Japanese Female Facial Expression (JAFFE) [24] database to illustrate the performance of the recognition system. Our proposed emotion recognition system is illustrated in Figure 2. This system constitutes the low-level feature extraction and the high-level tree structure representation for emotion recognition. The details of the major components in the proposed system will be described in the following sections.

Although some progress has been made in emotion recognition, several unsolved issues still exist. For example, it is still an open problem which features are the most important for emotion recognition. It is a subject that was seldom studied in computer science. However, related research works have been conducted in cognitive psychology [4–6]. Affective computing is becoming an important research area in intelligent computing technology. Furthermore, emotion recognition is one of the hot topics in affective computing. It is usually studied based on facial and audio information with technologies such as ANN, fuzzy set, SVM, HMM, etc. Many different facial and acoustic features are considered in emotion recognition by researchers.

4.1 Rough Sets (RS):

Rough set (RS) is a valid mathematical theory for dealing with imprecise, uncertain, and
vague information; it was developed by Professor Pawlak in 1980s [28, 29]. RS has been
successfully used in many domains such as machine learning, pattern recognition, intelligent data analyzing, and control algorithm acquiring [30–32]. The most advantage of RS is its great ability of attribute reduction (knowledge reduction, feature selection).

4.2 The problem here is:

a) Rough Sets are not efficient in dealing with data sets made up of continuous attribute values. Various quantization or Discretization techniques exist to address this issue, but no global technique applicable to a variety of data sets exists.

b) Rough Sets are purely rule based and due to the limited number of rules that such a system has, it fails to efficiently evaluate new cases or unseen situations. A Rough system can thus be said to be very fragile at its boundaries.

c) Adaptive Neuro – Fuzzy Inference Systems inherently are disadvantaged in modeling systems which have a large number of inputs or outputs. Though time efficient, the amount of pre – processing required before an ANFIS model can be generated, can be enormous depending upon the given problem or situation.

4.3 Adaptive Neuro - Fuzzy Inference Systems (ANFIS):

The ANFIS proposed by Jang [24] can be described as adaptive networks which are similar to Fuzzy Inference Systems functionally. They also have outlined a methodology to help decompose the parameter set (of the adaptive network nodes), so as to help implement the Hybrid Learning algorithm within the said systems. A successful attempt to represent the Sugeno and Tsukamoto Fuzzy models through ANFIS has been undertaken and further, it is also observed that the Radial Basis Function Network when subjected to certain constraints is equivalent to the ANFIS functionality. The idea of such an approach towards modeling systems is to interpret fuzzy rules in terms of a neural network. The fuzzy sets can be representative of the weights and the input – output functions along with the rules are representative of the neurons of a network. The (hybrid) learning as well is implemented as in a connectionist system i.e. the system learns by continuously modifying and adapting the neural structure and the parameters. The advantage of such a system is that the learning can be interpreted from perspectives of both neural and fuzzy systems. And more importantly such a system enables viewing the problem solution in a linguistic fashion Architecture: In the simplest terms, the structure of an ANFIS consists of a first layer where the inputs are mapped to their respective and relevant input membership functions. These membership functions taken it on to the rules layer further and then onto the output membership functions and finally to the output characteristic function which computationally produce the final single1 valued output. The following synopsis about the ANFIS architecture has been adopted from [25]. Consider a set of standard Sugeno style if then rules i.e.
Rule1: if x is A and y is B, then F1 = px + qy + r and;
Rule2: if x is C and y is D, then F2 = sx + ty + u.
Modeling the above using an ANFIS would result in a 5 layer network (excluding the input and the output layers) with the following sequential operations being accomplished,
Layer 1: The objective function of the nodes (all of the nodes in the same layer are characterized by the same objective function) is basically to assign a relevant fuzzy membership function to the said inputs i.e. ( )

$$O_{i,\,i(1-2)} = \mu_{A,\,i(x)}$$

$$O_{i,i(3-4)} = \mu_{B,i-2(y)} \tag{1}$$

Moreover, initially a user chosen parameterized general membership function is also adopted and the parameters of μ_{Ai}, μ_{Bi} are assigned some random values to start off with.

Layer 2: The outputs of the individual preceding layers are multiplied to retrieve the first set of real parameters which are called the premise parameters. i.e.

$$O_{2,i} = w_i = \mu_{Ai(x)}\mu_{Bi(y)} \tag{2}$$

Layer 3: The outputs of this layer are normalized or weighted weights i.e.

$$O_{3,i} = \overline{w_i} = \frac{w_{i(1-2)}}{w_1 + w_2} \tag{3}$$

Layer 4: Unlike the nodes from layers 2 and 3, layer 4 nodes are adaptive with the following output, (the final output individual vector components are computed). The parameters of p, q, r, s, t, u, are all called Consequent parameters.

$$O_{4,i} = \overline{w_i}F_i \tag{4}$$

Layer 5: Usually composed of a single static node, the operation of summing up the incoming individual vector components of the final single valued output.

$$O_{5,i} = \sum_i \overline{wi}Fi \tag{5}$$

As described earlier, the ANFIS network shall comprise of a minimum of 5 layers with at least two of them being adaptive in nature. This structure generates two sets of varied parameters to solve for, namely the premise and the consequent parameters. These parameters and their values are generated appropriately during the learning process (as described in the following section).
Hybrid Learning: Conventionally the Gradient optimization methods or back propagation techniques have been used to identify and optimize the various nodal parameters associated with adaptive networks. But Jang [24] proposed a hybrid learning rule approach to identify the parameters. This method involves combining the back propagation steepest descent and least squares method for a faster identification. The process of continuous parameter update can be achieved by two methods i.e. off – line or batch learning which involves updating the parameters iteratively at the end of all training data pair runs and on- line learning wherein the parameters are updated posthumously after every layered data run.

Batch Learning: Consider the following output function foran adaptive network

$$O = f(i,S) \tag{6}$$

where '*O*' is the output function, 'i' inputs set and '*S*' is the parameter set.
Now if there exists a function *'h'* such that applying it to Eq (1) would generate a new function linear in certain parameters of the set '*S*' i.e.

$$S = S1 \oplus S2 \tag{7}$$

therefore applying *'h'* to the objective function,

$$h(O) = h \circ f(i,S) \tag{8}$$

where *'h'* is linear in *S2* parameters.

Assuming a set of values for the parameter set S_2. We can generate *P* number of linear equations in *S2* parameters.
Representing the same in a matrix we get,

$$A\theta = y \tag{9}$$

and by the Least Squares Estimator the solution to the above equation can be given as,

$$\theta = (A^T A)^{-1} A^T y \tag{10}$$

The above described procedure is implemented within forward pass of the network. Once all of the S_2 parameters are identified, then the backward pass is initiated. The errors are calculated and the gradient vector is determined. At the end of all training pairs, in the backward pass, the parameters in S_1 are updated by steepest descent method.

$$f = (\overline{w_1}x)p + (\overline{w_1}y)q + (\overline{w_1})r + (\overline{w2}x)s + (\overline{w_2}y)t + (\overline{w_2})u \tag{11}$$

$\overline{w_1}$ and $\overline{w_2}$ are the premise parameters and p, q, r, s, t, and u are the consequents.
Data sets used: Following is a brief outline of the data sets which have been used, and a little bit more about them individually.
a) Breast Cancer:
Source: Orange Data Mining [26]
Attributes (9 in all): recurrence, age, menopause, Tumor
Size, inv nodes, node caps, Deg
Malig, breast and breast quad.
Missing Values: No
Data Objects: 190 (one per patient)
Decision Classes: 2
b) Lung Cancer:
Source: Orange Data Mining [26]
Attributes (56 in all): Attributes 1 to 56.
Missing Values: Yes
Data Objects: 32 (one per patient)
Decision Classes: 2
Data Reliability: Though emphasis was laid on the conceptual implementation in this work, an effort has been made to use reliable data. The data is a common testing data available for cost free downloads from various severs. The UCI data repository and the Orange Data Mining Repository are just a couple examples. This data is widely used by AI researchers in order to establish a performance scale for their respective techniques.

Table 6a. Breast Cancer

Attribute	**Rough**	**ANFIS**	**Rough-ANFIS**
Number of Rules	152	512	64
Error Rate	0	0.175	0.08
Attributes in use	6 out of 9	9 out of 9	6 out of 9
Reduct Cardinality	1(100%)	1(100%)	-NA-

Table 6b. Lung Cancer

Attribute	**Rough**	**Rough-ANFIS**
Number of Rules	32	27
Error Rate	0	0.1 (per unity)
Attributes in use	3 out of 56	3 out of 56
Reduct Cardinality	1(100%)	-NA-

Neuro Fuzzy systems: Neural networks exhibit lack of interpretability and Fuzzy systems lack the capability of effective learning. Thus Neuro Fuzzy systems present the learning capability of the neural networks and the fuzzy interpretation skills both in one tool. Moreover with respect to dynamic systems, neural networks provide the requisite skills for knowledge acquisition through learning and the fuzzy systems top it up with their ability to automatically approximate the knowledge bases for non – deterministic events. Due to the presence of these characteristics belonging to both the techniques, Neuro Fuzzy systems are widely used in machine learning applications. ANFIS is one of the kinds.

Rough Sets framework was chosen predominantly for its complimentary behavior when amalgamated with other approaches. Rough Sets Theory (RST) is well equipped to deal effectively with *imprecise, noisy and missing information* [27]. Unlike in other approaches, RST effectively sets the accuracy and precision value as per the requirement of the user for various classificatory processes. Further the concept of indiscernibility relations coupled with the concept of Reducts provides *discernibility – preserving elimination of irrelevant information* [27]. Also the issues arising due to multi and partial memberships of the objects in various sets have been reasonably addressed using the RST. And similar to the ES answering module, Rough Set models can be made capable of providing a description of analysis which led to the final decision. Fuzzy set theory and Rough Set theory are complimentary and not competitive. Amalgamation of the said techniques or theories would result in constructive determination since each of them refers to different aspects of imprecision i.e. Fuzzy set theory represents imprecision in the form of a partial membership whereas Rough sets accommodate the imprecision in the form of indiscernibility relations and the set upper and lower approximations.

4.4 Facial Feature Recognition: Parametric Feature Representation:

The contours of the facial features, generated by the facial feature detection method (Fig. 1), are utilized for further analysis of shown facial gestures. First, we carry out feature points' extraction under two assumptions: (1) the face images are non-occluded and in frontal view, and (2) the first frame is in a neutral expression. We extract 22 fiducial points: 19 are extracted as vertices or apices of the contours of the facial features (Fig. 2), 2 represent the centers of the eyes (points X and Y), and 1 represents the the middle point between the nostrils (point C). We assign a certainty factor to each of the extracted points, based on an "intra-solution consistency check". For example, the fiducial points of the right eye are assigned a certainty factor CF Ꞓ[0, 1] based upon the calculated deviation of the actually detected inner corner $B_{current}$ from the pertinent point $B_{neutral}$ localized in the first frame of the input sequence. The functional form of this mapping is:

CF = *sigm*(*d*($B_{current}$, $B_{neutral}$); 1, 4, 10) where *d*(p_1, p_2) is the block distance between points p_1 and p_2 (i.e., maximal difference in x and y direction) while *sigm*(x; α, β, γ) is a Sigmoid function. The major impulse for the usage of the inner corners of the eyes as the referential points for calculating CFs of the fiducial points of the eyes comes from the stability of these points with respect to non-rigid facial movements: facial muscles' contractions do not cause physical displacements of these points. For the same reason, the referential features used for calculating CFs of the fiducial points of the eyebrows, nose/chin and mouth are the size of the relevant eyebrow area, the inner corners of the nostrils and the medial point of the mouth respectively. Eventually, in order to select the best of sometimes redundantly available solutions (e.g., for the fiducial points belonging to the mouth), an intersolution

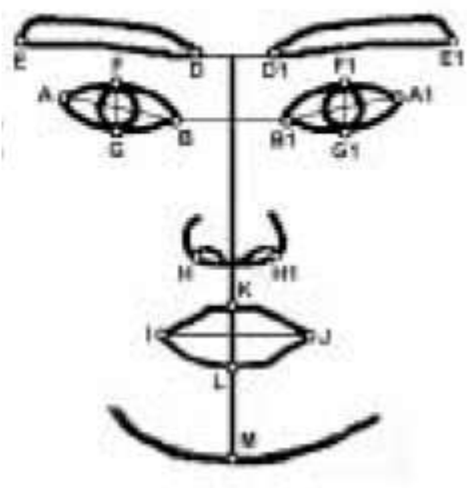


Figure 2: Feature points (fiducials of the features' contours)

E, E1: outer corner of the eyebrow
D, D1: inner corner of the eyebrow
A, A1: outer corner of the eye
B, B1: inner corner of the eye
F, F1: top of the eye
G, G1: bottom of the eye
H, H1: Outer corner of the nostril
K: top of the upper lip
L: bottom of the lower lip
I, J: mouth corner
N: tip of the chin

consistency check is performed by comparing the CFs of the points extracted by different detectors of the same facial feature.

AUs of the FACS system are anatomically related to contractions of facial muscles [4]. Contractions of facial muscles produce motion on the skin surface and changes in the shape and location of the prominent facial features. Some of these changes are observable from changes in the position of the fiducial points. To classify detected changes in the position of the fiducial points in terms of AUs, these changes should be represented first as a set of suitable feature parameters. Motivated by the FACS system, we represent these changes as a set of mid-level feature parameters describing the state and motion of the fiducial points. We defined a single mid-level feature parameter, which describes the state of the fiducials. This parameter, which is calculated for each frame for various fiducial points by comparing the currently extracted fiducial points with the relevant fiducial points extracted from the neutral frame, is defined as:

inc/dec(AB) = $AB_{neutral} - AB_{current}$, where AB

$$= \sqrt{\{(xA - xB)^2 + (yA - yB)^2\}}$$

If inc/dec(AB) < 0, distances increases.

4.4.1 Action Unit Recognition:

The last step in automatic facial gesture analysis is to translate the extracted facial information (i.e., the calculated feature parameters) into a description of shown facial changes, e.g., into the AU codes.

Table 7. The description of 22 AUs to be recognized and the related rules for AU recognition

AU	AU description & the related rule
1	Raised inner portion of the eyebrow(s) IF ***inc/dec***(BD) < 0 OR ***inc/dec***(B1D1) < 0 THEN AU1
2	Raised outer portion of the eyebrow(s) IF ***inc/dec***(AE) < 0 OR ***inc/dec***(A1E1) < 0 THEN AU2
3	Eyebrows pulled closer together (frown) IF ***inc/dec***(DD1) > 0 THEN AU4
4	Raised upper eyelid(s) IF ***inc/dec***(FG) < 0 OR ***inc/dec***(F1G1) < 0 THEN AU5
5	Raised cheeks (smile); IF AU12 OR AU13 THEN AU6
6	Raised lower eyelid(s) IF ***not***(AU12) AND ((FG > 0 AND ***inc/dec***(GX) > 0) OR (F1G1 > 0 AND ***inc/dec***(G1Y) > 0)) THEN AU7
7	Lips pulled towards each other IF ***not***(AU12 OR AU13 OR AU15 OR AU18 OR AU20 OR AU23 OR AU24 OR AU35) AND KL > 0 AND ***inc/dec***(CK) < 0 THEN AU8
8	Mouth corner(s) pulled up IF (***inc/dec***(IB) > 0 AND ***inc/dec***(CI) < 0) OR (***inc/dec***(JB1) > 0 AND ***inc/dec***(CJ) < 0) THEN AU12
9	Mouth corner(s) pulled sharply up IF (***inc/dec***(IB) > 0 AND ***inc/dec***(CI) > 0) OR (***inc/dec***(JB1) > 0 AND ***inc/dec***(CJ) > 0) THEN AU13
10	Mouth corner(s) pulled down IF ***inc/dec***(IB) < 0 OR ***inc/dec***(JB1) < 0 THEN AU15
11	Mouth pushed medially forward (as when saying "fool") IF ***not***(AU28) AND IJ ≥***t1*** AND ***inc/dec***(IJ) > 0 AND ***inc/dec***(KL) ≤0 THEN AU18
12	Mouth stretched horizontally IF ***inc/dec***(IJ) < 0 AND ***inc/dec***(IB) = 0 AND ***inc/dec***(JB1) = 0 THEN AU20
13	Tightened lips IF KL > 0 AND ***inc/dec***(KL) > 0 AND ***inc/dec***(IJ) ≤0 ND ***inc/dec***(JB1) ≥0 AND ***inc/dec***(IB) ≥0 THEN AU23
14	Lips pressed together IF ***not***(AU12 OR AU13 OR AU15) AND KL > 0 AND ***inc/dec***(KL) > 0 AND IJ > ***t1*** AND ***inc/dec***(IJ) > 0 THEN AU24
15	Parted lips IF ***inc/dec***(KL) < 0 AND ***inc/dec***(CM) ≥0 THEN AU25
16	Parted jaws IF ***inc/dec***(CM) < 0 AND CM ≤***t2*** THEN AU26
17	Mouth stretched vertically; IF CM > ***t2*** THEN AU27
18	Lips sucked into the mouth; IF KL = 0 THEN AU28
19	Cheeks sucked into the mouth; IF IJ < ***t1*** THEN AU35
20	Widened nostrils IF ***not***(AU8 OR AU12 OR AU13 OR AU18 OR AU24) AND ***inc/dec***(HH1) < 0 THEN AU38
21	Compressed nostrils IF ***not***(AU8 OR AU15 OR AU18 OR AU24 OR AU28) AND ***inc/dec***(HH1) > 0 THEN AU39
22	Dropped upper eyelid(s) IF ***not***(AU7) AND ((FG > 0 AND ***inc/dec***(FG) > 0 AND ***inc/dec***(FX) > 0) OR (F1G1 > 0 AND ***inc/dec***(F1G1) > 0 AND ***inc/dec***(F1Y) > 0)) THEN AU41

4.4.2 Feature transformation and Retention:

Definition 1.1.: A decision information system is a continuous value information system, and it is defined as a quadruple *s* = (*U, C U D, V, f),* where *U* is a finite set of objects, *C* is the condition attribute set, and *D* = {*d*} is the decision attribute set. For all *c Є C, c* is continuous value attribute.

A facial expression information system is a continuous value information system according to the above Definition.
If a condition attribute value is a continuous value, indiscernibility relation cannot be used directly since it requires that the condition attribute values of two different samples are equal, which is difficult to satisfy. Consequently, a process of discretization must be taken, in which information may be lost or changed. The result of attribute reduction would be affected. Since all measured facial attributes are continuous value and imprecise to some extent, the process of discretization may affect the result of emotion recognition. We argue that it is suitable for the continuous value information systems that the attribute values are taken as equal if they are similar in some range. Based on this idea, a method based on tolerance relation that avoids the process of discretization is proposed.

Definition 1.2.: A binary relation *R(x, y)* defined on an attribute set *B* is called a tolerance relation if it satisfies

(1) symmetrical: $\forall_{x, y \in U} (R(x, y) = R(y, x));$ (12)

(2) reflexive: $\forall_{x \varepsilon U} (R(x, x) = 1).$ (13)

From the standpoint of a continuous value information system, a relation could be set up for a continuous value information system as follows.

Definition 1.3.: Let an information system *S* = (*U, C U D, V, f*) be a continuous value information system; a relation *R(x, y)* is defined as

$$R(x, y) = \{(x, y) \mid x \in U \wedge y \in U \wedge \forall_{a \in c} (| a_x - a_y | \le \varepsilon, 0 \le \varepsilon \le 1)\}. \qquad (14)$$

Apparently, *R(x, y)* is a tolerance relation according to Definition 2.4 since *R(x, y)* is symmetrical and reflexive. In classical rough set theory, an equivalence relation constitutes a partition of *U*, but a tolerance relation constitutes a cover of *U*, and equivalence relation is a particular type of tolerance relation.

Definition 1.4.: Let *R(x, y)* be a tolerance relation based on (12) and (13),

$$n_R(x_i) = \{x_j \mid x_j \in U \wedge \forall_{a \in c} (| a_x - a_y | \le \varepsilon)\}$$

is called a tolerance class of *x_i*, and $| n_R(x_i) | = | \{x_j \mid x_j \in n_R(x_i), 1 \le j \le U\} |$ is the cardinal number of the tolerance class of $x_{i.}$

According to above Definition, for all *x Є U*, the bigger the tolerance class of *x* is, the more uncertainty it will be and the less knowledge it will contain. On the contrary, the smaller the tolerance class of *x* is, the less uncertainty it will be and the more knowledge it will contain. Accordingly, the concept of knowledge entropy and conditional entropy could be defined as follows.

Definition 1.5.: Let *U* = {x_1, x_2, . . . , $x_{|U|}$}, *R(x, y)* be a tolerance relation; the knowledge entropy *E_R_* of relation *R* is defined as

$$E(R) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{n_R(x_i)}{|U|}). \qquad (15)$$

Definition 1.6.: Let *R* and *Q* be tolerance relations defined on *U*, a relation satisfying *R* and *Q* simultaneous can be taken as *RUQ*, and it is a tolerance relation too. For all *xi ЄU*, $| n_{R \cup Q}(x_i) = n_R(x_i) \cap n_R(x_i);$ therefore, the knowledge entropy of *R U Q* can be defined as

$$E(R \cup Q) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{n_{R \cup Q}(x_i)}{|U|}. \qquad (16)$$

Definition 1.7.: Let *R* and *Q* be tolerance relations defined on *U*; the conditional entropy of *R* with respect to *Q* is defined as

$$E(Q \mid R) = E(R \cup Q) - E(R).$$

Let *S* = (*U, C U D, V, f)* be a continuous value information system, let relation *K* be a tolerance relation defined on its condition attribute set *C*, and let relation *L* be an equivalence relation _a special tolerance relation_ defined on its decision attribute set *D*. According to Definitions 2.7, 2.8, and 2.9, we can get

$$\begin{aligned} E(D \mid C) &= E(L \mid K) \\ &= E(K \cup L) - E(K) \\ &= -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{n_{K \cup L}(x_i)}{|U|} - (-\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{n_K(x_i)}{|U|}) \qquad (17) \\ &= -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{n_{R \cup Q}(x_i)}{|U|}) \end{aligned}$$

where the conditional entropy *E(D | C)* has a clear meaning; that is, it is a ratio between the knowledge of all attributes (condition attribute set plus decision attribute set) and the knowledge of the condition attribute set.
A novel attribute reduction algorithm is proposed based on rough set theory and domain-oriented data-driven data mining (3DM).

4.4.3 Handling Facial Dynamics:

Yacoob and Davis [35] proposed an approach for analyzing and representing the dynamics of facial expressions from image sequences. This approach is divided into three stages: locating and tracking prominent facial features (i.e., mouth, nose, eyes, and brows), using optical flow at these features to construct a mid-level representation that describes spatio-temporal actions, and applying rules for classification of mid-level representation of actions into one of the six universal facial expressions. Matsuno *et al.* [37] proposed an approach for recognizing facial expressions from static images based on precomputed parameterization of facial expressions. Their approach lays a grid over the face and warps it based on the gradient magnitude using a physical model. The amount of warping is represented in a multivariate vector that is compared to learned vectors of four facial expressions (happiness, sadness, anger, and surprise). Mase [36] used optical flow computation for recognizing and analyzing facial expressions in both a top-down and bottom- up approach. In both cases, the focus was on computing the motion of facial *muscles* rather than of facial *features.* Four facial expressions were studied: surprise, anger, happiness, and disgust. The top-down approach assumed that the face's image can be divided into muscle units that correspond to the action units (AU's) suggested by Ekman and Friesen [38]. Optical flow is computed within rectangles that include these muscle units, which in turn can be related to facial expressions. However, Mase did not report any results on mapping the optical motion results into facial expressions. This approach relies heavily on locating rectangles containing the appropriate muscles-a difficult image analysis problem since the muscle units correspond to smooth, featureless surfaces of the face. Furthermore, it builds upon a model that is suitable for synthesizing facial expressions but remains untested in analysis of facial expressions (for more details see *[39]).* The bottom-up approach covered the area of the face with evenly divided rectangular regions over which feature vectors derived from an optical flow computation are computed. The feature vectors are defined over a 15-dimensional space that is computed based on the means and variances of the optical flow. The recognition of expressions is based on a k-nearest-neighbor voting rule. The optical flow calculation was averaged within each window to smooth the results over edges. Furthermore, the optical flow is treated on a per-frame basis without considering the time-sequence of frames. The experiments considered the expressions of just one face andthe results were compared with the performance of human subjects that were asked to classify the displayed emotions. Terzopoulos and Waters [40] proposed an approach for synthesis and analysis of facial expressions based on physical modeling of the muscles of the face. They devised a six level representation of the face that consists of expression level (which includes the six primary expressions); control level (implements a subset of the FACS-facial action coding system for controlling muscles), muscle level (models the muscles' contraction and expansion as springs), physics level (models the facial tissue's deformations), geometry level (provides a geometric representation of the face as a mesh of polyhedral elements that depend on the curvature of surface), and the image level (visualizes the data). The analysis part assumes that 11 principle contours are initially located manually on the face. These contours are tracked throughout the sequence by applying an image force field that is computed from the gradient of the intensity image. The estimation of the muscle contractions takes place after the contours' reference points have been determined. In addition to assuming a frontal view, it was assumed that the projection is orthographic. Once the muscle contractions have been estimated, they were resynthesized onto the 3-D range data model of the subject to recreate the muscle contractions. It remains to be determined whether the computation of muscle mapped onto a 3-D physical model of the human face to activate muscles that create the same facial expressions on this model.

Cottrell and Metcalfe [41] proposed a back propagation neural network that recognizes facial expression (along with gender and identity) from static images. Their network compresses the input images in a "holistic" manner into 40 hidden units that were fed into a three-layer classification network. The expression network was able to distinguish some of the positive emotions but was less able to handle negative expressions. The ability of the network to generalize to new faces was poor. Essa [42] recently proposed a physically based approach for modeling and analyzing facial expressions. This approach extends the FACS model to the temporal dimension (thus calling it FACS+) to support combined spatial and temporal modeling of facial expressions. It is assumed that a mesh is originally overlaid on the face; then its vertices are tracked based on the optical flow field throughout the sequence. The emphasis is on accuracy of capturing facial changes, which is important for synthesis. Recognition results were reported in [42] on six subjects displaying four expressions and eyebrow raising.

5. TOWARDS CONTEXT BASED APPROACH TOWARDS FACIAL RECOGNTION

5.1 Contextual features

There are several sources of potential errors with a gesture based recognition system. For example when people search for their cursor on the screen, they perform fast short movements similar to head nods or head shakes, and when people switch attention between the screen and keyboard to place their fingers on the right keys, the resulting motion can appear like a head nod. These types of false positives can cause trouble, especially for users who are not aware of the tracking system.

In research on interactions with ECAs, it has been shown that contextual information about dialog state is a productive way to reduce false positives [33]. A context-based recognition framework can exploit several cues to determine whether a particular gesture is more or less likely in a given situation. To apply the idea of context-based recognition to non-embodied interfaces, i.e. windows-based interfaces, here we define a new set of contextual features based on window manager state.

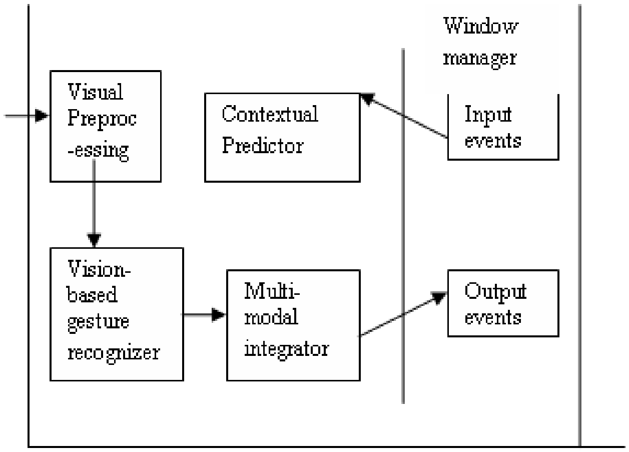


Figure 3: Framework for context-based gesture recognition. The contextual predictor translates contextual features into a likelihood measure, similar to the visual recognizer output. The multi-modal integrator fuses these visual and contextual likelihood measures.

We want to find contextual features that will reduce false positives that happen during interaction with conventional input devices, and contextual features that can be easily computed using pre-existing information. For our initial prototype we selected two contextual features: *fd* and *fm*, defined as the time since a dialog box appeared and time since the last mouse event and respectively. These features can be easily computed by listening to the input and display events sent inside the message dispatching loop of the application or operating system. We compute the dialog box feature *fd* as

$f_d(t) = C_d$ if no dialog box was shown
$t – t_d$ otherwise

where *td* is the time-stamp of the last dialog box appearance and *Cd* is default value if no dialog box was previously shown. The same way, we compute the mouse feature *fm* as

$f_m(t) = C_m$ if no mouse event happened
$t – t_m$ otherwise

where *tm* is the time-stamp of the last mouse event and *Cm* is default value if no mouse event happened recently. In our experiments, *Cd* and *Cm* were set to 20. The contextual features are evaluated at the same rate as the vision-based gesture recognizer (about 18Hz).

We wish to learn a measure of likelihood for a gesture given only the contextual features described in the previous section. This measure will later be integrated with the measure from our vision-based head gesture recognizer to produce the final decision of our context-based gesture recognizer (see Figure 3). The measure of likelihood is taken to be the distance to a separating surface of a multi-class Support Vector Machine (SVM) classifier that predicts the gesture based on contextual features only. The SVM classifier learns a separating function whose distance *m*(*x*) to training labels is maximized. The margin *m*(*x*) of the feature vector *x*, created from the concatenation of the contextual features, can easily be computed given the learned set of support vectors x_i, the associated set of labels y_i and weights w_i, and the bias *b*:

$$m(x) = \sum_{i=1}^{l} y_i w_i K(x_i, x) + b \qquad (18)$$

where *l* is the number of support vectors and *K*(*xi*, *x*) is the kernel function. In our experiments, we used a radial basis function (RBF) kernel:

$$K(xi, x) = e^{-\gamma \|x_i - x\|^2} \qquad (19)$$

where *γ* is the kernel smoothing parameter learned automatically using cross-validation on our training set. After training the multi-class SVM, we can easily compute a margin for each class and use this scalar value as a prediction for each visual gesture. The contextual predictor was trained using a subset of twelve participants. Positive and negative samples were selected from this data set based on manual transcription of head nods and head shakes.

5.2 Context-Based Scene Recognition Using Bayesian Networks

Scene understanding is an important problem in intelligent robotics. Since visual information is uncertain due to several reasons, we need a novel method that has robustness to the uncertainty. Bayesian probabilistic approach is robust to manage the uncertainty, and powerful to model high-level contexts like the relationship between places and objects. At first, image pre-processing extracts features from vision information and objects existence information is extracted by SIFT that is rotation and scale invariant. This information is provided to Bayesian networks for robust inference in scene understanding.

5.2.1 Visual Context-Based Low-Level Feature Extraction:

It would be better to use features that are related to functional constraints, which suggests
to examine the textural properties of the image and their spatial layout [43]. To compute texture feature, a steerable pyramid is used with 6 orientations and 4 scales applied to the gray-scale image. The local representation of an image at time *t* is as follows:

$$v_x^L(x) = \{v_t, k(x)\}_{k=1,N}, \quad \text{where N=24} \qquad (20)$$

It is desirable to capture global image properties, while keeping some spatial information. Therefore, we take the mean value of the magnitude of the local features
averaged over large spatial regions:

$$m_t(x) = \sum_{x'} | v_t^L x' | \, w(x' - x), \quad \text{where w(x) is averaging window} \qquad (21)$$

The resulting representation is down-sampled to have a spatial resolution of 4x4 pixels, leading to the size of m_t as 384(4 x 4 x 24), whose dimension is reduced by PCA (80 PCs). Then, we have to compute the most likely location of the visual features acquired at time *t* . Let the place be denoted as $Q_t \in \{1,\ldots,N_p\}$ where = $N_p = 5$. Hidden Markov model (HMM) is used to get place probability as follows:

$$P(Q_t = q \mid v_{1:t}^G) \alpha p(v_{1:t}^G \mid Q_t = q) P(Q_t = q \mid v_{1:t-1}^G$$

$$p(v_{1:t}^G \mid Q_t = q) \sum_{q'} A(q', q) P(Q_{t-1} = q' \mid v_{1:t-1}^G \qquad (22)$$

where *A*(*q*',*q*) is the topological transition matrix. The transition matrix is simply learned from labeled sequence data by counting the number of transitions from location *i* to location *j* . We use a simple layered approach with HMM and Bayesian networks. This presents several advantages that are relevant to modeling high dimensional visual information: learning each level independently with less computation, and although environment changes, only first layer requires new learning with the remaining unchanged [45]. The HMM is for extracting place recognition and BNs are for high-level inference.

5.2.2 High-Level Context Extraction with SIFT

Scale-Invariant Feature Transform (SIFT) is used to compute high-level object existence information. Since visual information is uncertain, we need a method that has robustness to scale or camera angle change. It was shown that under a variety of reasonable assumptions the only possible scale-space kernel was the Gaussian function [44]. Therefore, the scale space of an image is defined as a function, $L(x, y, \sigma)$ that is produced by the convolution of a variable-scale Gaussian, $G(x, y, \sigma)$, with an input image, $I(x, y)$:

$$L(x, y, \sigma) = Gx, y, \sigma) * I(x, y), \qquad (23)$$

where * is the convolution operation in x and y , and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x2+y2)/2\sigma^2} \qquad (24)$$

To efficiently detect stable key-point locations in scale space, scale-space extrema in the difference-of-Gaussian function are convolved with the image, $D(x, y, \sigma)$, which can be computed from the difference of two nearby scales separated by a constant multiplicative factor k:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, k\sigma)) * I(x, y)$$

$$= L(x, y, k\sigma) - L(x, y, \sigma) \qquad (25)$$

Extracted key-points are examined in each scene image, and the algorithm decides that the object exists if match score is larger than a threshold.

5.2.3 Context-Based Bayesian Network Inference

A Bayesian network is a graphical structure that allows us to represent and reason in an uncertain domain. The nodes in a Bayesian network represent a set of random variables from the domain. A set of directed arcs connect pairs of nodes, representing the direct dependencies between variables. Assuming discrete variables, the strength of the relationship between variables is quantified by conditional probability distributions associated with each node [46].

Consider a BN containing n nodes, Y_1 to $n\ Y_n$, taken in that order. The joint probability for any desired assignment of values < y_1 ,…, y_n > to the tuple of network variables < Y_1 ,…, Y_n > can be computed by the following equation:

$$p(y_1, y_2, \ldots\ldots, y_n) = \prod_i P(y_i \mid Parents(Y_i)) \qquad (26)$$

where $Parents(Y_i)$ denotes the set of immediate predecessors of Y_i in the network.

BN consists of 4 types of nodes: (1) 'PCA Node' for inserting global feature information of current place, (2) 'Object Node' representing object existence and correlation between object and place, and (3) 'Current Place Node' representing the probability of each place. Let the place be denoted $Q_{t} \in \{1,\ldots,N_p\}$ where = $N_p = 5$ and object existence is denoted by $O_{t,i} \in \{1,\ldots,N_{object}\}$ where = $N_{object} = 14$ Place recognition can be computed by the following equation:

$$CurrentPlace = \arg\max P(Q_t = q \mid v^G_{1:t}, O_{t,i}, \ldots., O_{t,N_{object}} \qquad (27)$$

The BNs are manually constructed by expert, and nodes that have low dependency are not connected to reduce computational complexity. Fig. 2 shows a BN that is actually used in experiments.

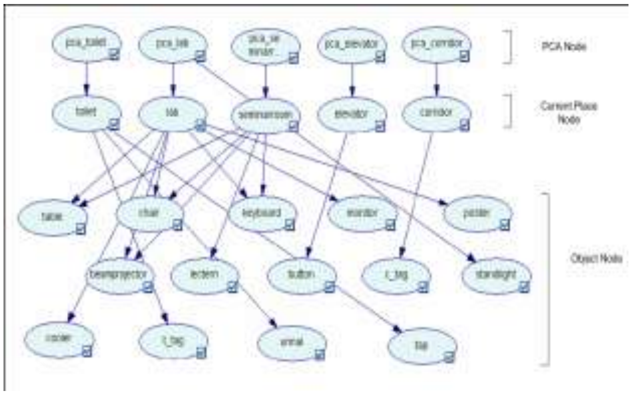


Figure 4. A BN manually constructed for place and object recognition

The work has been verified[34] that the context-based Bayesian inference for scene recognition shows good performance in the complex real domains. Even though the global feature information extracted is the same, the proposed method could produce correct result using contextual information: relationship between object and place. But SIFT algorithm showed low performance when objects had insufficient textual features, and this lack of the information caused to the low performance of scene understanding. To overcome it, we need a method that disjoints objects with ontology concept, and extracts SIFT key-points in each component. Besides, we could easily adopt more robust object recognition algorithm to our method.

6. CONCLUSION AND FUTURE WORKS

Theoretically all the aspects including cues, facial and gesture are important for non verbal communication. Adding facial features to gestures don't cause much difference to the results, rather can include complexity. Adding gesture again to it causes ambiguity, and probably the results go down. Hence facial features are the most important aspect for emotion recognition. Further, ANN, Fuzzy are extensively used for this purpose. Adding Rough set approach gives better results, and rough set is easy to implement too. Some Coxtext based research in these areas have been done using support vector machine and Bayesian networks Furthermore, context based approach can be further extended with the help of Rough-fuzzy approach towards refining artificial intelligence.

7. ACKNOWLEDGE

This is the research work associated to the Ph.D work going on Under Devi Ahilya University, Indore, and College of Engg, Pune. My sincere thanks the centre and their faculty for their extensive support.

8. REFERENCES

[1] R.W. Picard, *Affective Computing*, MIT Press, Cambridge, UK, 1997.

[2] R. W. Picard, "Affective computing: challenges," *International Journal of Human Computer Studies*, vol. 59, no. 1-2, pp. 55–64, 2003.

[3] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[4] X. Sui and Y. T. Ren, "Online processing of facial expression recognition," *Acta Psychologica Sinica*, vol. 39, no. 1, pp. 64–70, 2007 _Chinese_.

[5] Y. M. Wang and X. L. Fu, "Recognizing facial expression and facial identity: parallel processing or interactive processing," *Advances in Psychological Science*, vol. 13, no. 4, pp. 497–500, 2005 _Chinese_.

[6] Paul Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Palo Alto, CA: Consulting Psychologist Press, 1978.

[7] Paul Ekman and E.L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*, New York: Oxford University Press, 1997.

[8] Gianluca Donato, Marian Steward Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski, *Classifying Facial Actions.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21(10), pp. 974-989, 1999.

[9] K. Mase, *Recognition of facial expression from optical flow.* IEICE Transactions E, vol. 74(10), pp. 3474-3483, 1991.

[10] M. Rosenblum, Y. Yacoob, and L. Davis, *Human expression recognition from motion using a radial basis function network architecture.* IEEE Trans. Neural Networks, vol. 7(5), pp. 1121-1138, 1996.

[11] A. Lanitis, C. Taylor, and T. Cootes, *Automatic interpretation and coding of face images using flexible models.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19(7), pp. 743-756, 1997.

[12] J.G. Daugman, *Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression.* IEEE Trans. Pattern Anal. Machine Intell., vol. 36, pp. 1169-1179, 1988.

[13] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, *EigenFaces vs. FisherFaces: Recognition Using Class Specific Linear Projection.* IEEE Trans. Pattern Anal. Machine Intell., vol. 19(7), pp. 711-720, 1996.

[14] P.S. Penev and J.J. Atick, *Local feature analysis: a general statistical theory for object representation* Network: Computation in Neural Systems, vol. 7(3), pp. 477-500, 1996.

[15] M.S. Bartlett and T. Sejnowski, *Viewpoint invariant face recognition using independent component analysis and attractor networks*, in *Advances in Neural Information Processing Systems*, M. Mozer, M. Jordan, and T. Petsche, Editors, MIT Press: Cambridge, MA, 1997.

[16] J. Zhang, Y. Yan, and M. Lades. *Face recognition: Eigenface, elastic matching, and neural nets*, in proceedings of *IEEE*, vol. 85(9), pp. 1423-1435, 1997.

[17] A. C. Tsoi, *Adaptive Processing of Data Structure : AnExpository Overview and Comments*, Faculty Informatics, Univ. Wollongong, Wollongong, Australia, 1998.

[18] A. Sperduti and A. Starita, *Supervised neural networks for classification of structures.* IEEE Trans. Neural Networks, vol. 8, pp. 714-735, 1997.

[19] P. Frasconi, M. Gori, and A. Sperduti, *A General Framework for Adaptive Processing of Data Structures.* IEEE Trans. Neural Networks, vol. 9, pp. 768-785, 1998.

[20] Siu-Yeung Cho, Zheru Chi, Wan-Chi Siu, and Ah Chung Tsoi, *An Improved Algorithm for learning longterm dependency problems in adaptive processing of data structures.* IEEE Transactions on Neural Networks, vol. 14(4), pp. 781-793, 2003.

[21] J. Platt, *Fast Training of Support Vector Machines using Sequential Minimal Optimization*, in *Advances in Kernel Methods - Suppoort Vector Learning*, B. Scholkopf, C. Burges, and A. Smola, Editors, MIT Press, pp. 185-208, 1998.

[22] D. Aha and D. Kibler, *Instance based learning algorithms. Machine Learning.* Machine Learning, vol. 6, pp. 37-66, 1991.

[23] M. White, "Effect of photographic negation on matching the expressions and identities of faces," *Perception*, vol. 30, no. 8, pp. 969–981, 2001.

[24] Jang. J. S. R; "ANFIS: Adaptive Network Based Fuzzy Inference Systems"; IEEE Transactions on Systems, Man and Cybernetics; May 1993.

[25] Jang. J. S. R, Sun. C. T, Mizutani. E; Neuro – Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence; Prentice Hall; 1997.

[26] Demsar J, Zupan B; "Orange: From Experimental Machine Learning to Interactive Data Mining"; White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana; 2004.

[27] Kudo. Y, Murai. T; "Missing Value Semantics and Absent Value Semantics for Incomplete Information in Object-Oriented Rough Set Models"; In Bello et. al. [13]; 2008.

[28] Z. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.

[29] Z. Pawlak, "Rough classification," *International Journal of Man-Machine Studies*, vol. 20, no. 5, pp. 469– 483, 1984.

[30] Z. Pawlak, "Rough set theory and its applications to data analysis," *Cybernetics and Systems*, vol. 29, no. 7, pp. 661–688, 1998.

[31] R. W. Swiniarski and A. Skowron, "Rough set methods in feature selection and recognition," *Pattern Recognition Letters*, vol. 24, no. 6, pp. 833–849, 2003.

[32] N. Zhong and A. Skowron, "A rough set-based knowledge discovery process," *International Journal of Applied*

Mathematics and Computer Science, vol. 11, no. 3, pp. 603–619, 2001.

[33] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In Proceedings of the International Conference on Multi-modal Interfaces, October 2005.

[34] Seung-Bin Im and Sung-Bae Cho, Context-Based Scene Recognition Using Bayesian Networks with Scale-Invariant Feature Transform, ACIVS 2006, LNCS 4179, pp. 1080 – 1087, 2006. © Springer-Verlag Berlin Heidelberg 2006.

[35] "Computing spatio-temporal representations of human faces," in *IEEE Con$ Comput. Vision and Pattern Recognition*, 1994, pp. 70-75.

[36] K. Mase, "Recognition of facial expression from optical flow," *IEICE Trans.*, vol. E74, no. 10, pp. 3474-3483, Oct. 1991.

[37] K. Matsuno, C. Lee, and S. Tsuji, "Recognition of human facial expressions without feature extraction," *ECCV*, pp. *5* 13-520, 1994.

[38] *The Facial Action Coding System.* San Francisco, CA: Consulting Psychologists Press, 1978.

[39] V. Bruce, *Recognizing Face*, London: Lawrence Erlbaum, 1988.

[40] D. Terzopoulos and K. Waters, "Analysis and synthesis of facial image sequences using physical and anatomical models," *IEEE Truns. Pattern Anal. Machine Intell.*, vol. 15, pp. 569-579,'June 1993.

[41] G. W. Cottrell and J. Metcalfe, "EMPATH: Face, gender, and emotion recognition using holons," in *Advances in Neural Information Processing Systems*, R. P. Lippman, J. Moody, and D. S. Touretzky, Eds. San Mateo, CA: 1991, vol. 3, pp. 564-571.

[42] I. A. Essa and A. Pentland, "A vision system for observing and extracting facial action parameters," in *Proc. Computer Vision and Pattern Recognition, CVPR-94*, Seattle, WA, June 1994, pp. 76-83.

[43] A. Torralba, K.P. Mutphy, W. T. Freeman and M. A. Rubin, "Context-based vision system for place and object recognition," *IEEE Int. Conf. Computer Vision*, vol. 1, no. 1, pp. 273-280, 2003.

[44] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. J. Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[45] N. Oliver, A. Garg and E. Horvitz, "Layered representations for learning and inferring office activity from multiple sensory channels," *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 163-180, 2004.

[46] R.E. Neapolitan, *Learning Bayesian Network*, Prentice hall series in Artificial Intelligence, 2003.

7

Patent - Malaney, P. Kulkarni, etal, “Business method using the automated processing of paper and unstructured electronic documents.” No. 7,747,495 USPTO, June 19 2010

USPTO PATENT FULL-TEXT AND IMAGE DATABASE

Home Quick Advanced Pat Num Help

Bottom

View Cart Add to Cart

Images

(1 of 1)

United States Patent	**7,747,495**
***Malaney* , et al.**	**June 29, 2010**

Business method using the automated processing of paper and unstructured electronic documents

Abstract

The invention features a business method that takes advantage of the ability to convert unorganized information in the form of paper documents, document images, and electronic documents and converts the information to an organized electronic form referred to as Knowledge Objects. The invention further encompasses forming electronic Business Objects, such as documents and data sets, useful for business decision making and information exchange. The methods of the invention may utilize computerized storage and computerized decision-making systems to enable making more rapid critical business decisions.

Inventors:	***Malaney*; Sanjeev** (La Jolla, CA), **Kulkarni; *Parag*** (Pune, **IN**), **Viswanathan; Krishnaswami** (Pune, **IN**), ***Malaney*; Vikram** (Mumbai, **IN**)
Assignee:	**Capsilon Corporation** (San Francisco, CA)
Appl. No.:	**11/552,495**
Filed:	**October 24, 2006**

Related U.S. Patent Documents

Application Number	Filing Date	Patent Number	Issue Date
60730237	Oct., 2005		

Current U.S. Class:	**705/35** ; 358/1.15; 358/400; 707/E17.008; 707/E17.09; 715/209
Current International Class:	G06Q 99/00 (20060101)
Field of Search:	707/102,E17.09,E17.008 358/1.15,400 715/209

References Cited [Referenced By]

U.S. Patent Documents

4453268	June 1984	Britt
4468809	August 1984	Grabowski et al.
4899299	February 1990	MacPhail
5020019	May 1991	Ogawa
5159667	October 1992	Borrey et al.
5168565	December 1992	Morita
5297042	March 1994	Morita
5323311	June 1994	Fukao et al.
5414781	May 1995	Spitz et al.
5418946	May 1995	Mori
5463773	October 1995	Sakakibara et al.
5479574	December 1995	Glier et al.
5574802	November 1996	Ozaki
5579519	November 1996	Pelletier
5588149	December 1996	Hirose
5768580	June 1998	Wical
5812995	September 1998	Sasaki et al.
5819295	October 1998	Nakagawa et al.
5832470	November 1998	Morita et al.
5867799	February 1999	Lang et al.
5903904	May 1999	Peairs
5930788	July 1999	Wical
5940821	August 1999	Wical
5983246	November 1999	Takano
5991709	November 1999	Schoen
5999893	December 1999	Lynch, Jr. et al.
6055540	April 2000	Snow et al.
6061675	May 2000	Wical
6094653	July 2000	Li et al.
6101515	August 2000	Wical et al.
6125362	September 2000	Elworthy
6185576	February 2001	McIntosh
6243723	June 2001	Ikeda et al.
6266656	July 2001	Ohno
6442555	August 2002	Shmueli et al.
6460034	October 2002	Wical
6477528	November 2002	Takayama
6505195	January 2003	Ikeda et al.
6542635	April 2003	Hu et al.

6553358	April 2003	Horvitz
6553365	April 2003	Summerlin et al.
6556982	April 2003	McGaffey et al.
6556987	April 2003	Brown et al.
6625312	September 2003	Nagarajan et al.
6647534	November 2003	Graham
6701305	March 2004	Holt et al.
6718333	April 2004	Matsuda
2002/0052835	May 2002	Toscano
2002/0143704	October 2002	Nassiri
2003/0182304	September 2003	Summerlin et al.
2004/0199460	October 2004	Barash
2005/0080721	April 2005	Kearney et al.
2005/0108001	May 2005	Aarskog
2005/0134935	June 2005	Schmidtler et al.
2005/0209955	September 2005	Underwood et al.
2007/0024899	February 2007	Henry et al.
2007/0033078	February 2007	Mandalia et al.

Foreign Patent Documents

0460995	Dec., 1991	EP
0891593	Jan., 1999	EP
0891593	Mar., 2000	EP
1158424	Nov., 2001	EP
1199647	Apr., 2002	EP
1256886	Nov., 2002	EP
1256886	Nov., 2002	EP
1365329	Nov., 2003	EP
1365329	Nov., 2003	EP
1376420	Jan., 2004	EP
1378375	Jan., 2004	EP
1385329	Jan., 2004	EP
1385329	Jan., 2004	EP
1199647	Dec., 2005	EP

Other References

Amar Gupta, Sanjay Hazarika, Maher Kallel, Pankaj Srivastava; Optical Image Scanners and Character Recognition Devices: A Survey and New Taxonomy, 2003; MIT, Working Paper #3081-89. cited by examiner .
PCT/US07/86673 Search Report dated Apr. 10, 2008. cited by other .
PCT/US06/41542 Search Report dated Aug. 20, 2008. cited by other.

Primary Examiner: Badii; Behrang
Attorney, Agent or Firm: Sonsini; Wilson Goodrich & Rosati

Parent Case Text

INCORPORATION BY REFERENCE

This application claims priority to the U.S. Provisional Application No. 60/730,237 filed Oct. 24, 2005 entitled "A Business Method Using the Automated Processing of Paper Documents" which is hereby incorporated by reference herein in its entirety. All publications and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication or patent application was specifically and individually indicated to be incorporated by reference.

Claims

What is claimed is:

1. A method of doing business by processing a group of documents comprising the steps of: (1) performing optical character recognition from said discrete documents using said device to generate one or more sets of text-based information; (2) classifying at least some of said discrete documents using said sets of text-based information, wherein multiple classification engines are employed and said classifying is based on a consensus of said classification engines; (3) classifying at least some of the discrete documents using Image Based Classification; (4) verifying any of said remaining discrete documents that are not classified in said steps of classifying by employing a Location Diagram with said remaining discrete documents or a portion thereof; (5) collating said at least two of said discrete documents; (6) versioning and sequencing at least two of said discrete documents; (7) locating said fields containing data in said at least two of said discrete documents; (8) extracting data from said fields of said at least two discrete documents to generate extracted data; (9) scrubbing values from said extracted data to generate values therefrom; (10) forming Knowledge Objects; (11) storing said values in a data storage device; (12) forming Business Objects; (13) displaying at least some of said values to a user.

2. The method of claim 1, wherein said step of performing optical character recognition is performed by, or with the assistance of, a computer.

3. The method of claim 1, wherein said step of classifying is performed by, or with the assistance of, a computer.

4. The method of claim 3, wherein said step of classifying is performed by Image Based Classification.

5. The method of claim 1, wherein said step of verifying is performed by, or with the assistance of, a computer.

6. The method of claim 1, wherein said step of collating is performed by, or with the assistance of, a computer.

7. The method of claim 1, wherein said step of versioning and sequencing is performed by, or with the assistance of, a computer.

8. The method of claim 1, wherein said step of extracting data is performed by, or with the assistance of, a computer.

9. The method of claim 1, wherein said step of scrubbing is performed by, or with the assistance of, a computer.

10. The method of claim 1, wherein said step of formation of Knowledge Object is performed by, or with the assistance of a computer.

11. The method of claim 1, wherein said step of outputting is performed by, or with the assistance of, a computer.

12. The method of claim 1, wherein said step of formation of Business Object is performed by, or with the assistance of, a computer.

13. The method of claim 1, wherein said step of displaying is performed by, or with the assistance of, a computer.

14. The method of claim 1, wherein ambiguities at any step are escalated to a human operator.

15. The method of any of claim 1, wherein an ambiguity in a classification step, a field location step, or a data extraction step are escalated to a human collaborator.

16. The method of any claims of claims 1-14, wherein said group of documents is a group of mortgage loan documents.

Description

FIELD OF THE INVENTION

The field of invention is the process or method of doing business by processing paper documents, image files, and/or electronic documents; using a computer to analyze, collate and capture information from the documents; optionally using a computer for making decisions based on this information; and exchanging the organized information between organizations electronically. The field of the invention also includes the method of performing such analysis, collation, and information capture as well as an apparatus for conducting such analysis, collation, and capture.

BACKGROUND OF THE INVENTION

The mortgage banking industry is faced with the daunting task of organizing, inputting and accessing a vast number and array of divergent types of documents and manually entering several hundred fields of information from a subset of these documents in order to make a loan to a borrower. Although many attempts have been made to streamline the process, most recently by the Mortgage Bankers Association (MBA) which established standards for representing information in a mortgage transaction, the problem of identifying and capturing information from paper documents, image files, native PDF files, and other electronic files in the loan origination process has yet to be solved in order to take advantage of these standards. In the United States alone, mortgage bankers are faced with the idiosyncratic documents from a minimum of fifty states where some mortgage documents differ from state to state and may have further individual variations within each state. In addition, once the loan is made to the borrower, there is a huge secondary market for mortgages, where existing mortgage loans are bundled and sold to large investment firms. These investment entities, in order to pursue a rational risk management policy presentable to their owners and/or shareholders, must organize and analyze these mortgage documents for asset risk and compliance with local, state and federal laws. Values necessary to compare and analyze these loans must be extracted from paper documents or images of the document, then tabulated, analyzed and the resultant data and documents made readily available in order for informed decision-making to occur. In January 2000, the MBA formed the Mortgage Industry Standards Maintenance Organization (MISMO). This group has driven the development of industry specifications that allow seamless data exchange using standard electronic mortgage documents called SMART DOcS.TM.. The SMART Doc XML specification is the foundation of the eMortgage efforts of lenders, vendors, and investors, as it provides for the electronic versions of key mortgage documents. This specification enables electronic mortgage loan package creation by providing a standard for creating and processing uniform electronic transactions for use in electronic mortgage commerce.

Nor is this dilemma restricted to the mortgage industry. In other industries, including the finance industry, the hospitality industry, the health care field and the insurance industry, there is a constant need to collate documents into logically related groups, and capture key information to enable information exchange. These documents must be further collated in order to identify and store multiple revisions of the same type of document, along with extracting data and inferred information from the documents, together with making the resultant transaction data and underlying documents available in an electronically accessible manner.

Unfortunately, the manual organization, collation of paper documents, and extraction of information is very time consuming and slows the process of making business decisions. Additionally, there is an increased possibility of error due to manual processing. Validation of these decisions is very difficult since the paper documents are stored separately from the electronic databases maintained by the processing organizations. Thus, there is a clear need for process automation and well organized and easily searchable electronic storage of the documents as well as extraction of relevant information contained within the documents.

In other methods or processes known in the art, automated document identification or classification methods fall into one of three categories: (1) they are either completely dependant on image based techniques for classification; (2) they use simple keyword search techniques, Bayesian and/or Support Vector Machine ("SVM") algorithms for text classification; or (3) they rely on document boundary detection methods using image and text based classification techniques. These methods are inadequate to deal with the wide variation in documents typically seen in the business environment and are not capable of separating multiple revisions of the same document type to enable information to be captured from the most current version of the document, hence limiting the utility of such systems.

Although it is known in the art to view paper documents by conversion into simpler electronic forms such as PDF files, these files, in general, do not allow extracting information beyond Optical Character Recognition (OCR). The OCR quality is highly dependant on image quality and the extraction is frequently of very poor quality. Finally, these methods or apparatuses do not offer a complete solution to the dilemma of analyzing and manipulating large paper document sets. Thus, the automated systems currently available generally have at least the following problems:

(1) such systems are limited to document boundary detection, document classification and text extraction and do not offer advanced document collation with separation of very similar documents, and domain-sensitive scrubbing of extracted information into usable data;

(2) techniques based on the current methods of out-of-context extraction and keyword-based classification cannot offer the consistent extraction of information from documents for automated decision making, or formation of Business Objects such as SMART DocS.TM. for information exchange between two organizations using industry standard taxonomy;

(3) similarity among documents may lead to misclassification when using pattern-based classification, especially in cases where the optical character recognition quality of the document is poor;

(4) extraction processes that handle structured data using a template-based matching generally fail even with a slight shifting of images, and those with rules-based templates can return false results if there are significant variations of the document;

(5) such systems cannot handle both structured and unstructured documents equally efficiently and reliably to serve an entire business process;

(6) such systems frequently are wed to the strengths and weaknesses of a particular algorithm and are thus not able to handle wide variations in analyzed documents with acceptable accuracy without manual rule creation;

(7) such systems cannot locate the information across the documents and variations;

(8) neither do such systems provide a complete solution to a business problem; and

(9) such systems do not have intelligent scrubbing of extracted information to enable the creation of electronic transaction sets such as MISMO SMART Doc.TM. XML files.

To analyze complicated documents, workers in several industries, for example, mortgage banking, currently analyze documents using a manual collation process; a manual stacking process; a wide variety of manual classification methods; and manual extraction methods, in particular a manual search and transcription. These methods suffer from the disadvantages of requiring substantial investment of human capital and not being automated sufficiently to handle bulk processing of documents and the information contained in those documents.

The number and kind of documents accompanying a mortgage loan are very specific to the mortgage loan industry, and as mentioned above, vary from state to state, and may vary in the jurisdictions within a particular state. However, the documents related to a given loan for the purchase of a property or properties in any jurisdiction may be assembled into electronic images by scanning (or direct entry, if already in an electronic form) before, during and after funding of the loan to form a partially, or preferably, complete document set, referred to herein as the "Dox Package." These documents originate from a number of sources, including banks and/or credit unions. Moreover, the order of these documents are assembled and filed depends very much on the individuals involved, their timeliness and their preferences, organization, or disorganization in sorting the various forms and other documents containing the required information. Further, even though some standardization of documents has occurred, such as Form 1003 published by FNMA, certain data essential for further analysis may still be found at disparate locations in idiosyncratic documents. For example, each bank and credit union formats an individual's bank statement in a different manner, yet the data from each format must be extracted for income verification. Additionally, depending on the stage of loan processing, not all of the documents may be present in a Dox Package at a given point in time.

As mentioned above, following the funding of the loan, loans are frequently bundled with many other similar loans and sold on the secondary market. At this stage, entire lots of mortgage-secured loans are bundled and sold with minimal quality control. In current usage in the secondary mortgage market, a randomly selected ten percent sample of mortgage documents (Dox Packages) are analyzed in detail (largely by manual means) and taken as representative for the lot. Obviously, if more loans, or substantially all the loans in a bundle, could be evaluated, better decisions could be made regarding the marketing of mortgage-backed loans on the secondary market. Hence, pricing of these loans in the market would be more efficient. Thus, there is a clear need for the automated analysis, collation of documents, and extraction of information in the mortgage loan industry, as well as other industries with no automated or standardized data input in place.

The following patents and applications may also be relevant in describing the background of the instant invention: U.S. Patent Application No. 2005 0134935 and U.S. Pat. Nos.: 6,754,389, 6,751,614, 6,742,003, 6,735,347, 6,732,090, 6,728,690, 6,728,689, 6,718,333, 6,704,449, 6,701,305, 6,691,108, 6,675,159, 6,668,256, 6,658,151, 6,647,534, 6,640,224, 6,625,312, 6,622,134, 6,618,717, 6,611,825, 6,606,623, 6,606,620, 6,604,875, 6,604,099, 6,592,627, 6,585,163, 6,556,987, 6,556,982, 6,553,365, 6,553,358, 6,542,635, 6,519,362, 6,512,850, 6,505,195, 6,502,081, 6,499,665, 6,487,545, 6,477,528, 6,473,757, 6,473,730, 6,470,362, 6,470,307, 6,470,095, 6,460,034, 6,457,026, 6,442,555, 6,411,974, 6,397,215, 6,362,837, 6,353,827, 6,298,351, 6,289,337, 6,266,656, 6,259,812, 6,243,723, 6,233,575, 6,216,134, 6,212,517, 6,199,034, 6,185,576, 6,185,550, 6,178,417, 6,175,844, 6,157,738, 6,128,613, 6,125,362, 6,101,515, 6,094,653, 6,088,692, 6,061,675, 6,055,540, 6,044,375, 6,038,560, 5,999,893, 5,999,647, 5,995,659, 5,991,709, 5,983,246, 5,966,652, 5,960,383, 5,943,669, 5,940,821, 5,937,084, 5,930,788, 5,918,236, 5,909,510, 5,907,821, 5,905,991, 5,873,056, 5,867,799, 5,854,855, 5,848,186, 5,835,638, 5,832,470, 5,819,295, 5,812,995, 5,794,236, 5,768,580, 5,717,913, 5,706,497, 5,696,841, 5,694,523, 5,689,342, 5,598,557, 5,588,149, 5,579,519, 5,574,802, 5,568,640, 5,535,382, 5,519,608, 5,500,796, 5,479,574, 5,463,773, 5,428,778, 5,426,700, 5,423,032, 5,418,946, 5,414,781, 5,323,311, 5,297,042, 5,287,278, 5,204,812, 5,181,259, 5,168,565, 5,159,667, 5,107,419, 5,091,964, 5,051,891, 5,048,099, 5,021,989, 5,020,019, 4,899,299, 4,860,203, and 4,856,074.

SUMMARY OF THE INVENTION

In one aspect the instant invention features a method of doing business by processing a Dox

Packages wherein each Dox Package has at least two pages wherein minimal human intervention is involved in the extraction of information and/or data. In preferred embodiments, the Dox Package has documents related to a mortgage.

In another aspect the instant invention features a method of doing business by processing a group of Dox Packages wherein each Dox Package has at least two pages wherein the information is extracted from the Dox Packages and organized ten times as fast as a human operator. In preferred embodiments, the Dox Package has documents related to a mortgage.

In one aspect, the instant invention features a method of doing business by processing a group of documents, i.e., a Dox Package, where the process comprises some or all of the following steps:

(1) providing at least two of the discrete documents pages containing one or more fields from the group of documents to a device that can provide optical character recognition (OCR), and performing optical character recognition from the discrete documents using the device to generate one or more sets of text-based information;

(2) classifying at least some of the discrete document pages using the sets of text-based information, wherein multiple classification engines are employed and classification is based on a consensus of the classification engines, i.e. their vote;

(3) classifying at least some of the discrete document pages using Image Based Classification (as defined herein);

(4) verifying any of the remaining discrete document pages that are not classified in the step of classifying by employing a Location Diagram wherein the Location Diagram may be constructed using Feature Vectors with the remaining discrete document pages or a portion thereof;

(5) collating at least two of the discrete document pages that form discrete documents;

(6) determining the version number of each document and verifying the page sequence to form a unique document with a specific revision/version identity;

(7) extracting data from the fields of a discrete document to generate extracted data;

(8) scrubbing values from the extracted data to generate values therefrom;

(9) outputting the values to a data warehouse such as a data storage device or a hard drive;

(10) displaying at least some of the values to a user;

(11) forming required relationships between extracted information to form Knowledge Objects; and

(12) collating Knowledge Objects to form Business Objects such as MISMO SMART Docs.

In one aspect, the instant invention features a method of doing business by processing a group of documents using a computer where the process comprises some or all of the following steps:

(1) providing at least two of the discrete documents pages containing one or more fields from the group of documents to a device that can provide optical character recognition (OCR), and performing optical character recognition from the discrete documents using the device to generate one or more sets of text-based information;

(2) classifying at least some of the discrete document pages using the sets of text-based information, wherein multiple classification engines are employed and classification is based on a consensus of the classification engines, i.e. their vote;

(3) classifying at least some of the discrete document pages using Image Based Classification;

(4) verifying any of the remaining discrete document pages that are not classified in the step of classifying by employing a Location Diagram wherein the Location Diagram may be constructed using Feature Vectors with the remaining discrete document pages or a portion thereof;

(5) collating at least two of the discrete document pages that form discrete documents;

(6) determining the version number of each document and verifying the page sequence to form a unique document with a specific revision/version identity;

(7) extracting data from the fields of a discrete document to generate extracted data;

(8) scrubbing values from the extracted data to generate values therefrom;

(9) outputting the values to a data warehouse such as a data storage device or a hard drive;

(10) displaying at least some of the values to a user;

(11) forming required relationships between extracted information to form Knowledge Objects; and

(12) collating Knowledge Objects to form Business Objects such as MISMO SMART Docs.

In one aspect, the instant invention features an apparatus for analyzing a group of documents using the methods described herein wherein said apparatus comprises a computer. In this aspect, the instant invention features an apparatus for processing a group of documents where the apparatus performs all or some of the following steps:

(1) providing at least two discrete documents pages containing one or more fields from the group of documents to a device that can provide optical character recognition (OCR), and performing optical character recognition from the discrete documents using the device to generate one or more sets of text-based information;

(2) classifying at least some of the discrete document pages using the sets of text-based information, wherein multiple classification engines are employed and classification is based on a consensus of the classification engines, i.e. their vote;

(3) classifying at least some of the discrete document pages using Image Based Classification;

(4) verifying any of the remaining discrete document pages that are not classified in the step of classifying by employing a Location Diagram wherein the Location Diagram may be constructed using Feature Vectors with the remaining discrete document pages or a portion thereof;

(5) collating at least two of said discrete document pages that form discrete documents;

(6) determining the version number of each document and verifying the page sequence to form a unique document with a specific revision/version identity;

(7) extracting data from the fields of a discrete document to generate extracted data;

(8) scrubbing values from the extracted data to generate values therefrom;

(9) outputting the values to a data warehouse such as a data storage device or a hard drive;

(10) displaying at least some of the values to a user;

(11) forming required relationships between extracted information to form Knowledge Objects; and

(12) collating Knowledge Objects to form Business Objects such as MISMO SMART Docs.

In a still other aspect, the instant invention features a method of analyzing a bundle of loans assembled for sale on the secondary market wherein over 30%, over 40%, over 50%, over 60%, or over 70% of the mortgage documents are analyzed and the data/information is extracted.

In certain embodiments in any of the aspects of the instant invention, ambiguities in the processing of the documents are escalated to a human collaborator, in particular this may occur during or following the classification step, the field location step, and/or the data extraction step. In one embodiment of the instant invention, the step of performing optical character recognition is performed by, or with the assistance of, a computer. In another embodiment of the instant invention, the step of classifying is performed by, or with the assistance of, a computer. In still another embodiment of the instant invention, the step of verifying is performed by, or with the assistance of, a computer. In a further embodiment of the instant invention, the step of collating is performed by, or with the assistance of, a computer. In a still further embodiment of the instant invention, the step of extracting data is performed by, or with the assistance of, a computer. In another embodiment of the instant invention, the step of scrubbing is performed by, or with the assistance of, a computer. In still another embodiment of the instant invention, the outputting is performed by, or with the assistance of, a computer. In still a further embodiment of the instant invention, the step of displaying is performed by, or with the assistance of, a computer. In one embodiment of the instant invention, ambiguities at any step are escalated to a human operator. In another embodiment of the instant invention, the group of documents being analyzed is a group of mortgage loan documents. In other embodiments of the instant invention, the groups of documents being analyzed may be home appraisals, credit reports, and a single loan file where it is frequently used for underwriting purposes.

In any of the above aspects, the invention also features a method of operating a business where a purpose of the business is to offer the method/apparatus of preferred embodiments of the instant invention as a service. In another aspect, the instant invention features advertising the method/apparatus of the instant invention and/or advertising the availability of a service featuring the method/apparatus of the instant invention.

BRIEF DESCRIPTION OF THE DRAWINGS

The novel features of the invention are set forth with particularity in the specification, drawings and figures and in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

FIG. 1 depicts an overview of the business methods of the instant invention.

FIG. 2 depicts a detailed diagrammatic view of the business methods of the instant invention, i.e., the system flow of a preferred embodiment of the invention.

FIG. 3 depicts an embodiment of the Document Learner process, i.e., the flow of the classification learner.

FIG. 4 depicts an embodiment of the Business Object formation elements.

FIG. 4A depicts an embodiment of the relationship of Knowledge Objects within a Business Object.

FIG. 4B depicts the process of Dox Package creation in one embodiment of the invention.

FIG. 4C depicts the process of document creation in one embodiment of the invention.

FIG. 4D depicts MISMO transaction data-set creation in one embodiment of the invention.

FIGS. 5A and 5B depict screen shots of output obtained through the use of the instant invention. That is, using a Dox Package analyzed by the method/apparatus as described herein, the exemplary data in the figure was available for analysis.

DETAILED DESCRIPTION OF THE INVENTION

While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

Workers in a variety of organizations and/or industries, such as the mortgage industry, especially the secondary market for the re-sale of mortgage loans, face the enormous problem of tracking a vast array of information presented to them in the form of paper documents arriving in a bewildering array of formats, and require that information transferred to an electronic form for rapid analysis and decision-making. Extracting exact data and/or information from idiosyncratic document sets with accuracy is essential for the data to be useful for decision-making.

As noted above, the MBA formed the Mortgage Industry Standards Maintenance Organization (MISMO) to address this problem. This group has driven the development of industry specifications that allow seamless data exchange using standard electronic mortgage documents called SMART DocS.TM.. However, in order for the mortgage industry to fully utilize this standardization, every piece of software in the industry would have to be re-created to generate data to adhere to this standard. Hence, the industry requires a practical solution to enjoy the increased velocity and standardization that SMART Doc XML standards bring to the loan origination process using the current forms of data available such as paper images, and native PDF files. In a preferred embodiment, it is one of the objects of our invention to provide such a solution.

It is always difficult and time-consuming to determine the exact nature and identity of documents present in such a document set. For example, with reference to the mortgage industry, mortgage documents in some states, e.g., California, contain reports concerning the seismic environment of the subject property. In other states, such documentation might be rarely, if ever, be found in the package of documents associated with the sale of property, or the refinancing thereof (such a document package is referred to herein as the "Dox Package"). Further, without knowing the type of document or specific revision of the document being reviewed, up until now, it has been difficult or impossible to extract the required information from it by automated means. The exact documents provided in a Dox Package may prove insufficient because at a particular point in time, not all required pages of the documents may be available. Additionally, there may be a confusing variety or subvarieties for any given type of document, and further, essential information may be scattered across many or all the pages in the Dox Package. And for added complication, individual pages may arrive in a scrambled order in any given packet, and portions of the packet may arrive for analysis at different times. Obtaining accurate information in an organized form is the challenge solved by the instant invention. If a human were to enter the information into a computer, the process would be labor intensive and would be expected to take much longer. In preferred embodiments, a Dox Package may consist of at least two pages, at least three pages, at least five pages, at least ten pages, at least twenty pages, at least fifty pages, at least one hundred pages, or more, Further, as used herein, a Dox Package includes sets of documents in which all the information/data contained therein is not readily available in electronic or digital forms. Thus, a Dox Package may consist of a variety of documents some of which are electronic documents but some of which are paper copies only, or images, such as PDFs or TIFFs, of such paper documents.

The instant invention, in some embodiments, can extract the information from the heterogeneous set of documents that forms Dox Package and enter that information into a computer database much faster than, and in some embodiments, with minimal or no intervention from, a human operator; in some cases ten times as fast, twenty times as fast, thirty times as fast, forty times as fast, fifty times as fast, or more. Additionally, in one embodiment, the instant invention can extract and enter information from a Dox Package with human review of, at most, one page in ten, one page in twenty, one page in thirty, one page in forty, one page in fifty, one page in sixty, one page in seventy, one page in eighty, one page in ninety, one page in one hundred, or one page in over a hundred.

As used herein, a "Knowledge Object" is a matrix of the information and its association with reference to a particular business process. When a Knowledge Object is not specific to a process and/or a complete domain, it can be cluster of information. Knowledge Objects are intended to be useful and available for decision-making.

The term "Knowledge Object," as used herein, refers to a set of facts preferably along with their relationship and association with other Knowledge Objects in a given Dox Package. Knowledge Object is a matrix of relevant information entities such as facts, image field coordinates, value type, intended to address and assist decision making in businesses.

As used herein, a "Business Object" is a collected and organized set of information extracted from a Dox Package intended for a business purpose and ready to use to illustrate relationships and/or the utility of Knowledge Objects. It gives a business-centered view of the extracted and organized knowledge for the decision-making process. An example of a Business Object is a MISMO standard SMART Doc

As used herein, the term "Dox Package" refers to the pile, stack, or file of documents that is delivered, handed, and/or made available to the operator of the instant invention. In certain preferred embodiments, the Dox Package comprises mortgage documents and documents in support of a mortgage, or secondary financing thereof.

As used herein, "Taxonomy" refers list of document types (or document classes) expected in any Dox Package.

Documents within the Dox Package or taxonomy may consist of multiple pages, but all pages are preferably logically related to the reference page (as defined below).

The term "escalation" as used herein refers to a subroutine within the method/apparatus in embodiments of the instant invention that when the method/apparatus finds a document and/or page it cannot assign or identify, it escalates the document and/or page out of the program, or automated document analysis, and displays the document to a human collaborator. In preferred embodiments, the page is displayed on a split screen with the "heading region" of the document page amplified at the top of the screen and the entirety of the document shown in the bottom of the split screen. The instant inventors have determined that the identity of most documents can be determined by clues obtainable in the header region.

As used herein, the term "buckets" is a location to store related pages during the processing involved in preferred embodiments of the instant invention. Buckets may later be correlated and classified to the operative taxonomy so that a given bucket becomes a document within the taxonomy system.

The term "forensic page analysis" as used herein refers to a detailed extraction and mapping of the image that forms a sheet or an image of a sheet wherein this mapping is used to identify the page and/or sheet. Forensic page analysis generates a Location Diagram and Feature Vectors.

As used herein, the term "reference page" refers to the most readily identifiable document in a set of documents or pages within a Dox Package. Frequently, it is the first page of a document, but that is not required by the definition as the first page of a document may be a cover page, such as a fax cover page. An example of a reference page is the front page of a Form 1003. The "reference page" herein is the page of a document that represents the maximum logical properties or identifying properties of the document with all subsequent document members able to be classified as having affinity towards this "reference page." This "reference page" could be, but is not necessarily, the first page of the document within a bucket or with the classified documents.

As used herein, the term "field" refers to the region of a document where specific items of information might be found. Thus, on a Form 1003 there is a field for a name where an individual's name is found; the individual's name is a "fact" and may also referred to herein as a "text snippet" when the fact is extracted from a field.

Thus, fields are converted into facts by extracting the information and "scrubbing" the text output to create a value that can be utilized and/or consumed by a computer in the operation of embodiments of the instant invention.

As used herein, the term "information fields" refers to the content of the blanks on the forms, e.g., in the context of the mortgage field, the price of the property, the amount financed, the address, etc. or specific content from an unstructured document such as stated interest rate in a promissory note.

The term "Feature Vector" as used herein refers to a manner of mapping documents wherein the relationship of keywords to fields or keywords to other keywords is mapped both as to physical distance and direction.

The meaning of the term "Location Diagram" as used herein is best explained by an example. Each file is present in three formats: (1) the original .tiff image format, (2) the text format from simple OCR output, and (3) a grid format, i.e., a text pictorial representation of the document. All three formats are used in classification and extraction.

Assuming that A, B, C, D and E are five phrases, the overall representation that may come in a single feature-vector may be represented as follows: (1) A and B form a meaning X; (2) A is primary key; (3) B is p columns and q rows away from A; (4) with similar information about other key phrases being recorded.

These overall positions form a Location Diagram.

Here, the Location Diagram is a relative position map of key phrases represented in unique way by their vectors of relative distances. The structured files are represented in flexible structure maps called grid files.

Collation is done to segregate documents in groups to represent: (1) the Class-version, (2) the document identity (doc id), (3) page, and (4) versions and/or occurrences.

As used herein the term "collate" refers to the process of taking a bucket comprising a document, or a pages of a document, or sheets classified to the same taxonomy identified niche; analyzing the sheets located therein, preferably as well as all the sheets in a Dox Package, and sorting them into the correct buckets whereby all sheets belong to a document will be correctly sorted, and preferably different versions or dates of documents collected together. Thus, the term's definition comprises the dictionary meaning of "collates" whereby a collation occurs through a process that assembles pages in their proper numerical or logical sequence, and/or through a process examines gathered sheets in order to arrange them in the proper sequence. Collation also refers to the process of organizing Knowledge Objects into Business Objects.

OCR is generally referred to as the process of recognizing characters on an image file and converting them to ASCII text characters format.

As used herein, the acronym "NLP" refers to natural language processing, as is known to one of skill in the art.

As used herein, the term "Image Based Classification" refers to methods to classify documents using features and/or references other than text such as the visual page layout, the white-space distribution, and graphic patterns.

The purposes of instant invention include conducting a business and making business decisions using an automated acquisition and analysis of information from a Dox Package. This invention thus, in part, provides:

(1) a comprehensive method/apparatus that extracts relevant information from electronic images of paper documents to electronic data and assembles the extracted information with a very high level of accuracy and very little human intervention;

(2) a comprehensive method/apparatus that facilitates decisions at all levels by those with an interest in the documents or data therein by providing data with a quantifiable level of accuracy;

(3) a comprehensive method/apparatus for classification, collation, and identifying the version of documents together with relevant information extraction where the overall method/apparatus being enabled by an automatic document learner; and/or

(4) a decision-engineering framework specific to a given business application to overview and analyze the extracted information. In preferred embodiments, the documents and/or information may be converted in an XML file format such as those defined for the mortgage industry by MISMO.

Although there are a few superficially similar classifiers and extractors in the present-day art, the instant invention has several advantages over the art by fulfilling some or all of the purposes noted above, and in its unique combination of document processing features which include some or all of the following features:

(a) it is enabled with automated document learner providing learning and classification at the level of a page, the level of a zone within a page, or the level of a field within a page;

(b) it is easily adaptable to any given business due to its learning ability;

(c) it provides incremental learning to allow the system/process to rapidly accommodate new variations of the same documents as well as new types of documents;

(d) it features incremental learning that enables the system to accommodate variations and adapt to the changes in patterns of documents;

(e) it provides validation and verification of located and extracted information specific to the business domain while minimizing extraction mistakes and providing a high confidence level in the accuracy of the results;

(f) it provides a Location Diagram-based extraction that allows for accurate extracting of information even with significant changes in the document formatting;

(g) it provides, via Location Diagram-based information extraction, the accurate extraction of information even when page boundary information is lost during the OCR process, including data slipping to other pages, and/or the format or organization of the document changes;

(h) it provides, via the Location Diagram-based classification and identification, the ability to provide the sequential number and order of pages based on intelligence built during learning the document set in the form of Location Diagrams;

(i) it provides the ability to separate multiple revisions of the same document type into unique documents by identifying the reference page of each document type and the Feature Vector affinity of associated pages of that document by using distance measurement algorithms; and

(j) it provides the ability to further collate the information with the help of the grid of information created; and

(k) it provides the ability to flexibly distribute collated documents or extracted information to a user, or sets of different documents or information to different users or decision systems using standards such as MISMO SMARTDocs or custom XML tags.

One of the advantages of embodiments of instant invention is the number of discrete pages it can analyze. Although other document analysis methods and apparatuses exist, the instant invention may handle more pages and more diverse pages than what was present in the art prior to the instant invention. Thus, in embodiments of the instant invention, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13, 15, 18, 20, 25, 30, 35, 40, or more pages may be analyzed in on Dox Package. Also in embodiments of the instant invention, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13, 15, 18, 20, 25, 30, 35, 40, or more document types may be analyzed in on Dox Package

Thus, the instant invention provides a method/apparatus that analyzes, and collates documents, even individual versions of similar documents, preferably based on both their logical and their numerical sequence to systematically order groups of pages to enhance usability and to analyze them based on these grouped sets. These grouped sets are meaningful and comprehensive entities and are placed in their unique context for the specific business being supported. This collation takes place in spite of potential extreme variation in documents and in forms and the sequence of the documents or forms being input into the process. In preferred embodiments, the method/apparatus of the instant invention is directed to a specific business, the mortgage loan business, for example.

It is an object of the instant invention to provide comprehensive processes and systems that can convert relevant information from electronic images of papers and/or documents already in an electronic form to an electronic database with minimal human intervention. Further, "Knowledge Objects" are formed based on the extracted information. These Knowledge Objects may be further utilized to form "Business Objects." The Business Objects are collations of Knowledge Objects centered on specific business requirements and can be used for subsequent decision making. An additional object of the invention is to provide a managing tool that can help in learning and configuring the overall process.

It is also an object of the instant invention to classify documents and uniquely identify documents and revisions of the same document type, and extract information with the aid of automatic learners.

The method/apparatus of the instant invention may collate images of sets of pages for any given type of document package (referred to herein as the "Dox Package") presented to the operator or the apparatus of the instant invention. It is expected that documents in such a Dox Package may include images of paper documents, such as those in electronic .pdf files, native pdf files, or documents received by fax servers, for example in .tiff format. The instant invention, however, is not limited to the handling of such paper documents or images thereof. Thus, as used and defined herein, documents, sets of documents, pages, sets of pages, paper documents, form documents, physical pages, paper form, paper images, sheets, and the like includes documents and the like that exist in digital form, including documents, papers and forms, such as Microsoft.RTM. .doc documents and in other proprietary document formats, and the use of such are included within the scope of the present invention. Such documents may also contain embedded images, such as digital signatures or imported graphics or other documents, and likewise are included within the scope of use of the present invention.

In many preferred embodiments of the instant invention, documents are presented or utilized following the OCR conversion of original, signed or executed, documents or a text dump of the native pdf document. Along with mapping to standard MISMO taxonomy, the method/apparatus is also capable of generating its own taxonomy of buckets based on document features observed or recognized by the method/apparatus during analysis of the Dox Package. In this collation process, each page analyzed is assumed to hold a unique position within an individual document, and this page's position is determined and assigned. The method/apparatus initially assigns each page from the Dox Package the most logical bucket and the most appropriate position within the bucket; a page can belong to one and only one logical group. The position or a particular page and the sequence of pages is determined based on the page's purpose, location, readability and usability by the method/apparatus of the instant invention. After being assigned, the location of the sheet or page is preferably repeatedly re-evaluated and thus the accuracy of its position assignment, and the ultimate quality of the data, is increased.

In the case of ambiguity it cannot resolve in the assignment of a document to a bucket or to a page location within a bucket, the method/apparatus of the instant invention, in preferred embodiments, provides for escalation to a human collaborator or assistor to supplement the basic machine and expert-system-based collation. The level of ambiguity that triggers escalation may be preset, modified, or created during operation. In escalation, the human collaborator can determine the identity and classification of the ambiguous document and where it should be assigned to provide clues to the method/apparatus of the instant invention.

The present invention, in preferred embodiments, utilizes Location Diagram concepts and integrates multiple components including image processing, intelligent collation, feedback learning, a document classifier, a verifier, a versioning engine, an information locator, a data extractor, a data scrubber, and manual collaboration. Taking advantage of structured and unstructured properties of documents, the instant invention can convert representations of form documents into grid format, i.e., a text-pictorial representation. Using grid format, the instant invention can extract more and important features from the documents that then can help in formation of a Knowledge Object with very high level of accuracy and minimum human intervention. By using the method/apparatus of some embodiments of the instant invention, human review of pages within the Dox Package may only be required for one page in ten, one page in twenty, one page in thirty, one page in forty, one page in fifty, one page in sixty, one page in seventy, one page in eighty, one page in ninety, one page in one hundred, or less.

(a) Objects of Invention and Their Description

Numerous paper transactions occur in various business fields such as the mortgage industry, the health care field, the various insurance industries, including the health care insurance industry, financial banking, etc. The papers, documents and other information involved in these transactions generally are not random but rather all have interrelationships within a specific business context. Dox Packages obtained during the course of business, or images thereof, typically are not very well organized especially due to the fact that they may be created or obtained by different entities and/or at different points in time. There is need for segregation and subsequent coherent organization of these documents, as well as extracting information from these documents, and organizing and collating the extracted information, e.g. into MISMO standard SMART Doc.TM., custom XML tag based, other commonly used data file formats, or those to be developed. The need for segregation, organization and collation of documents in the Dox Package arises from a number of reasons: (1) checking for completeness of the Dox Package, i.e., whether all documents required, necessary or desirable to those entities having an interest in the information contained in the Dox Package, are present in the Dox Package; (2) legal aspects of the information contained within the Dox Package; (3) business aspects of the information within the Dox Package, (4) extracting data from a large number of "hard-copy"-only documents or images thereof, which may only be possible from a `representative number` dur to time or money constraints, (5) requiring rapid and inexpensive access to the data contained in the documents for analysis; and (6) having available or distributing documents or sets of documents in a segregated manner based on type of document or other criteria; and (7) making decisions based on the extracted data, including compilations, aggregations, and analyzed or processed sets of such data, optionally with an automated rules engine.

To address these needs and other needs associated with the collation documents and extraction of information, the inventors have devised a method and apparatus to accomplish these tasks to collate and analyze documents and sets of documents, and extract information from specific versions of these documents. The instant invention, in preferred embodiments, provides a comprehensive process and system which can convert information on papers or images to an analyzed and organized electronic form where it can be used for business decision-making.

The present invention, in some embodiments, solves the problems of sorting into versions, sequencing and collating documents and extracting information for specific industries. Thus one object of the instant invention is to provide users with separated, collated and sequenced documents. Users of the instant invention provide the method/apparatus a document set obtained in their course of business, a `Dox Package,` which is then collated and analyzed to meet their business requirements. In preferred embodiments, all documents are provided at once in one location, although such documents may be provided at different times and from different locations. A feature of the invention is that paper documents that do not have all the data contained therein in a segregated digital form are readily used with the instant invention.

This invention, in some preferred embodiments, comprises a comprehensively automated process that can convert data from documents in paper form to electronic form without with little, if any, human intervention. The instant invention may collate and classify documents based on Location Diagrams, which are based on Feature Vectors and connectivity/relationships among them. Further, the engine used in the instant invention can locate and extracts information from documents based on these Location Diagrams with additional scrubbing. The product is equipped with learners, which work on Location Diagram distance maximization within and across the document classes to optimize results, a "studio" (user-friendly interface) and a warehouse for storage and making data available as required by the operator of the instant invention or others designated by the operator. This invention may use methods of solving Location Diagrams based on simultaneous equation- and weight-based confidence measurements. The invention may provide significant benefit to all industries that handle sets of documents, and in particular, large, disparate sets of documents, by accelerating and improving accuracy to current decision-making process when compared to existing and traditional methods/technologies.

The instant invention, in some embodiments, provides a method/apparatus that collates and analyzes a set of documents. The apparatus automatically employs various algorithms to identify groups or logical units of documents. These algorithms work to complement one another to yield higher quality results. Further, the method/apparatus of the instant invention utilizes and takes into account discontinuities, for example, a page break in the middle of a sentence, to assemble pages of a document. Each of these logical units is a complete document identified as to its business identification and mapping to location within the taxonomy.

Further, the invention's method/apparatus preferably measures relatedness among various pages; to accomplish this the method/apparatus works on the principle of a reference page. As used herein, a "reference page" is a page that represents the maximum or near-maximum logical properties of a particular document, and thus all the subsequent document members have affinity towards this reference page. A reference page frequently is, but is not required to be, the first page of a given document. Using the principles of the instant invention, the logical sequence of a Dox Package is related to its purpose, location, readability and usability. Grouping and collating using the principles of the instant invention is concerned with completeness, usability, integrity, and unique occurrence.

The classification and collation unit as used in the instant invention in a preferred embodiment has an Image Based Classifier, a set of text based classifiers, versioning engine, an intelligent collation engine and a verifier. The text based classifier preferably has a set of classification engines and each classification engine confidence is prioritized based on its strengths in handling particular types of documents as will be determined by the particular application and recognized by one of skill in the art operating the instant invention.

The reference page identification method in one embodiment uses a hybrid approach where an affinity determination method is used in connection with an input dictionary, but can also provide feedback to enhance and/or enhance the input dictionary. This dictionary preferably not only provides a list of words but also gives quantitative relevance of words and phrases with reference to each class of document. Keywords and keyphrases have a high affinity towards a given document. For example, word `W1` is defined as having a very high chance of occurrence in document `D1` (e.g., the word `interest` (`W1`) in a mortgage note (`D1`)) then, according to the uses and principles of the instant invention, the word `W1` has high affinity towards document `D1.` This affinity may be determined using Bayesian analysis and is represented as a probability or a conditional probability. Other Feature Vectors such as font size and type may also be considered in determining the affinity of a page to the reference page of document being examined. There is no limit to the number of Feature Vectors that might be considered for affinity analysis.

The method/apparatus employs a multi-level approach to identify documents. Typically, the first pass, or Level-1 approach identifies some of the reference pages efficiently and quickly. Level-1 analysis may identify some reference pages along with their respective classes. Using the instant invention, attempts are made to identify classes for the remaining pages. In preferred embodiments, Level-1 uses various statistical algorithms, e.g., algorithms based on SVM and Bayesian. In preferred embodiments, the Level-1 reference page identifier is integrated with multi-algorithm classifier which selects the best of set of algorithms based on input data.

These reference pages are mapped to a taxonomy class by measuring the association of Feature Vectors and the relevance of the reference page using supervised learning. The closeness of other pages with reference to reference page is measured. This closeness is used to establish association of these pages with respect to the reference page. The pages in the document are arranged in logical/numerical sequence using this relevance.

The classifier takes advantage of various methods like word phrase frequency, Bayesian analysis, and SVM, but is not limited to these methods and has the capability to give priority and higher weight to the most suitable method to be used for the given document for maximum accuracy and usability.

In some of these preferred embodiments, Location Diagrams and Feature Vectors are neither required nor generated. Documents identified by Level-1 algorithms as ambiguous or as having affinity for more than one taxonomy class proceed to Level-2 analysis. Thus, all the documents that could not be handled in the Level-1 process effectively or routinely by the classifier are sent for verification in a Level-2 analysis. The verifier used in the Level-2 analysis is preferably capable of resolving the ambiguous document classes leftover from the Level-1 analysis. The instant invention also can resolve and relate documents belonging to multi-class families and documents that are within families or a group of classes that are similar. The verifier produces the final identification in these multi-class scenarios, using a combination of voting and critical-feature-based class verification.

In preferred embodiments, with the Level-2 analysis, all the documents that are unable to be classified in the Level-1 analysis are processed using critical-feature-based verification approach. In other embodiments, all documents are processed using the Level-2 analysis. The Location Diagram Map approach used in the Level-2 analysis in some preferred embodiments of the instant invention provides the required discrimination and accuracy to handle ambiguous documents and correctly classify or collate them. In preferred embodiments, this Level-2 reference page identifier uses critical-feature-based verification and voting, along with the verification algorithm and is referred to as the "verifier."

In preferred embodiments, the collation process provides for documents to be given a logical and/or numerical sequence. Thus in accordance with the instant invention, a Dox Package is collated with reference to a prescribed or developed taxonomy where the taxonomy classes are characteristic of the industry, (e.g., industry standards like MISMO) or required or desirable by the industry, yet may be adjusted by the operator/user of the instant invention.

Each document and/or page within the Dox Package is mapped to a class according to the taxonomy. The method/apparatus of the instant invention classifies these documents and collates sets of pages for industry standard taxonomy like MISMO, or any given taxonomy. A further feature in some embodiments of the instant invention is that the method/apparatus of the invention is also capable of generating its own taxonomy based on document features it observes. The overall method assigns most logical document structure based on the taxonomy and most appropriate position within each document for each page.

The separation, collation, and sequencing of the documents is taxonomy-based set by users' business requirements or defined by the field of use, such as MISMO standards, for the documents processed in the instant invention. The initial grouping into buckets and then refined into documents is utilized to further extract information specific to each document and Dox Package and is an important feature of preferred embodiments of the instant invention from the business perspective.

The instant invention may thus assign meanings to the documents and put them in their proper business context by the use of the separation, sequencing and collating methods described above. Each document, group, or subset formed within the Dox Package is based on the document's, group's or subset's use in the relevant business.

In preferred embodiments, the system has human collaboration along with its basic machine learning and expert system based collation.

In preferred embodiments of the instant invention, the method/apparatus of the instant invention is also equipped with a fact extractor for use with the pages, documents or sets of documents in the Dox Package. This fact extraction capability provides for locating and extracting the information/fields required for various business/compliance requirements and transforms the information contained therein to facts or data that can be subject to further use or manipulation. Preferably, the fact extractor is also equipped with weight-based confidence measurement. The fact extractor enables, in part, facts of all types and coming in various forms in the original documents to be accessed, extracted and/or manipulated. As with one feature of the instant invention that ultimately provides for pages, documents or sets of documents to be separated, classified or collated, the fact extraction feature of the instant invention allows for human collaboration for exceptional or problematic documents, although such human intervention is not required. The instant invention can handle all types of fields, e.g., OMR, tables, descriptions, numbers and the like, that will be known to those of skill in the art, depending on the particular business application. The decision system that is optionally used as part of the instant invention provides logical decisions based on this information obtained or extracted and the relevant business context. The preferable need-based human collaboration built into the system makes it possible to extract information and/or data from fields with a very high level of accuracy and coverage.

In preferred embodiments, the instant invention also provides a decision-engineering-framework specific to the business application to organize and utilize the extracted information. Thus, the information extracted from a Dox Package is preferably presented in a usable format, such as a spreadsheet or XML tag file format. Further, automated decisions may be made on the information obtained by an automated rules engine such as Microsoft BizTalk, ILOG jrules, etc.

In an preferred embodiment, an appraisal report regarding a piece of property (most of which are created as PDF files, if they are available electronically at all) and extract the information (including unstructured information), to create an XML output. This output can be used for a variety of purposes such as it may be furnished to a company that evaluates and scores the accuracy risk of the appraiser's information, to generate a report similar to an AVM to a mortgage banker for a business decision. In preferred embodiments, the instant invention may convert the information from an appraisal into electronic data over 100 times as fast as a human operator and with better accuracy.

In another preferred embodiment, the instant invention can extract information automatically from a credit report. This information may be furnished, for example, to a mortgage lender for their risk assessment process.

Thus, in preferred embodiments, the instant invention provides for collation of all the pages, documents, or sets of documents within a Dox Package into a taxonomy classification to meet the business needs of the operator and/or a particular industry. Virtually any Dox Package from any industry may be analyzed by preferred embodiments of the instant invention. Thus, the collated documents are mapped as to a taxonomy such as MISMO or any other industry-specific or user-specific taxonomy. As part or in addition to this, information is extracted from this Dox Package. This information is scrubbed and transformed into discrete data and/or facts. The facts and its related information is used to form an information matrix called a Knowledge Object. The Knowledge Objects are transformed in a particular or required business context to create Business Objects. The Business Objects are then used for business decision-making. In preferred embodiments, the instant invention therefore facilitates extraction of critical information for businesses from the documents, and provides for manipulation, compilation, analyzing and/or access to the facts or data or creation of transaction sets that comply with the MISMO SMART Docs standard and/or other custom XML tag file formats.

Advantages of the Instant Invention

Methods currently available do not meet all the objects of the instant invention, but rather have contributed to the shortcomings, problems and challenges present in the art. Preferred embodiments of the instant invention provide advantages over the current state of the art and these embodiments improve upon them because of all or some of the following reasons:

a) the instant invention, in preferred embodiments, offers a comprehensive process which takes unorganized documents or document images and yields extracted information suitable for business decision-making;

(b) the instant invention, in preferred embodiments, provides for an automated method with exception-based human collaboration ("escalation") to collate with increased speed and accuracy;

(c) the instant invention, in preferred embodiments, provides superior accuracy and quantifiable measures for accuracy;

(d) the instant invention, in preferred embodiments, is the only comprehensive collation solution which can collate pages, documents, or sets of documents identified by revision numbers, for business decision making purposes;

(e) the instant invention, in preferred embodiments, is not limited to document separation by boundary detection algorithms;

(f) the instant invention, in preferred embodiments, provides for the mapping of documents and document images to a MISMO taxonomy, as well as other industry standard and custom taxonomies;

(g) the instant invention, in preferred embodiments, locates and extracts information from documents and document images sorted into buckets with a high degree of accuracy;

(h) the instant invention, in preferred embodiments, provides, among others, the features of intelligent scrubbing and fact conversion and/or other data manipulation features; the fact conversion converts extracted information into data or facts that offer value to businesses and provide direct input into an automated rules engine using custom or industry standard XML formats such as those specified by MISMO;

(i) the instant invention, in preferred embodiments, provides an automated learner which can accommodate and incorporate new document types, and the intelligence to deal with variations in the number and type of documents and field locations;

(j) the instant invention, in preferred embodiments, can incrementally learn to adapt to changes in the patterns between and/or within documents;

(k) the instant invention, in preferred embodiments, validates and verifies collated documents, and Knowledge Objects to improve accuracy;

(l) the instant invention, in preferred embodiments, provides a Location Diagram-based extraction for accurate extractions in case of slippage, variations and changes in format; and

(m) the instant invention, in preferred embodiments, features in some embodiments a collation confidence matrix to be able to assess the confidence level of the method or algorithm, plus the instant invention, in preferred embodiments, may effectively use all clues gathered during all phases of document processing and analysis to validate the accuracy of the result. Thus, by use of the instant invention, business decisions, such as whether to invest in a bundle of loans on the secondary market, may be based on extracted information from a large number of the associated Dox Packages, or a majority of the associated Dox Packages, or almost all of the associated Dox Packages.

One of the advantages of embodiments of instant invention is the number of discrete pages it can analyze. Although other document analysis methods and apparatuses exist, the instant invention may handle more pages and more diverse pages than what was present in the art. Thus, in embodiments of the instant invention, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13, 15, 18, 20, 25, 30, 35, 40, or more pages may be analyzed in on Dox Package. Also in embodiments of the instant invention, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13, 15, 18, 20, 25, 30, 35, 40, or more document types may be analyzed in on Dox Package

Illustration of the Instant Invention

As described above, the instant invention is, in preferred embodiments, a process and system for separating, organizing and retrieving information from various documents, for example from a Dox Package. The system preferably employs a collator, a classifier, an extractor, a scrubber, a verifier, a version engine, a voting engine, a transformer for creation of Knowledge Objects and Business Objects, a decision engine and a learner for classification and extraction.

A high-level exemplary overview of one embodiment of the instant invention is provided by FIG. 1 depicting the method/apparatus of the instant invention. Here, unorganized information is captured by the apparatus from various office devices such a computer, a FAX, an e-mail system, a scanner, or uploaded to a FTP or a Web site 101. Further the captured documents or information, unorganized and unidentified when acquired, are organized into an information matrix known as Knowledge Objects by referencing a Knowledge Warehouse 102, and stored in an information data warehouse 103. Knowledge Objects are then transformed into Business Objects, such as electronic documents and transaction sets such as MISMO standard XML files 104. The Business Objects are stored in business data warehouse or delivered to users of the system and external organizations 105. Finally, a Work-flow and Decision engine uses the Business Objects to facilitate both manual and automated business decisions, and collaboration 106.

A detailed exemplary overview of the instant invention is provided by FIG. 2 depicting one preferred embodiment of the method/apparatus of the instant invention. It will be recognized by those of skill in the art that FIG. 2 is only one example or embodiment of the instant invention; other embodiments of the instant invention may be recognized by reference to FIG. 2 and/or the description herein. For example, each of the steps described in FIG. 2 may be modified; further, many of the steps are optional so that one or more steps may be eliminated. Also, other steps may be added. Similarly, the order of the steps may be changed or rearranged in numerous ways. Each of these embodiments is within the spirit and scope of the invention as defined by the appended claims.

Capture Documents:

Pages, documents, sets of documents, a Dox Package, or Dox Packages are sent electronically to the system for classification or/and extraction of data 200. Such documents may be input in any sequence and by or through any manner known to those of skill in the art such as from a fax machine, scanner, e-mail system or any other electronic communication device. The document or documents may be in text, electronic, paper, or image form, or a mixture of formats. If needed, in preferred embodiments, the document(s) are captured by techniques or in a manner known to those of skill in the art. The Dox Package is separated into Image type documents or Text type documents as they are captured.

Image Pre-processing:

The document image quality, in particular, from documents obtained by low resolution scans or facsimile transmission may not be good enough for direct OCR. Therefore, primary image processing may optionally be done to bring the image to the requisite quality for OCR, and Image based classification 202. In preferred embodiments, noise is removed from the image by technologies such as de-skew techniques and de-speckle techniques, a change or changes in DPI, and/or image registration correction or by a combination of the above and/or similar techniques.

Image Based Classification (IBC)

In the preferred embodiment the IBC 203 attempts to identify one or more discrete pages using Image Based Features like lay out, white space distribution, and other features registered in the collection of document feature descriptors by the Document Learner.

OCR (Optical Character Recognition)

If required, as in the case of image type documents, and in preferred embodiments, the portions of or the entire image of a page or document is converted into text using OCR by means known in the art 204. In some preferred embodiments, the OCR program is available commercially. In preferred embodiments, the OCR engine is supported with a general as well as a business-domain-specific dictionary to increase the accuracy. The OCR output may optionally be in text and xml formats, or may be in other formats.

In preferred embodiments, once a image type document is OCRed, the output file is converted to a grid-based matrix format to form a text-pictorial representation of the document (Document Grid File). Text type documents, such Microsoft word documents are also converted into a Document Grid File.

In preferred embodiment the output from image preprocessing and OCR is used for the Image Quality Detection IQD 205.

Identification:

All pages or documents are placed in buckets using a preliminary analysis of features discovered in the Document Grid File, without detailed validation. The order of the presented pages, sheets, and/or documents presented to the method/apparatus is recorded by the system, for example using a computer database.

In preferred embodiments, the method/apparatus then attempts to identify one or more reference pages and then the documents are grouped logically based on the reference page and/or affinity 206. Numerous classification engines known in the art can be used, separately or together, including a Word Map 207, a SVM classifier 208, a Location Diagram 209, a Bayesian classifier 210, and a critical-feature-based identifier 211, but any manner known to those of skill in the art can be used. In preferred embodiments, the engines are used in a particular order. In some preferred embodiments, if all the classification engines agree as to the classification of a page or document, the result is accepted as the identification, and taxonomy classification, of the document; in other or the same preferred embodiments, if most of the classification engines agree, then the result is accepted as the identification of the document. If the document, page or sheet is not identified at this point, further analysis is performed with the aid of a human collaborator (i. e., via escalation). Further, in preferred embodiments, discontinuities are used to identify pages from a single document, e.g., a sentence or a table separated by a page break.

Preferably, all pages are revisited and checked with regard to their affinity towards the reference page. This method of confirmation in preferred embodiments works by measuring affinity of the pages in the vicinity of a reference page towards that reference page, but also reviews pages far removed (distance measurement) from the reference page to guard against, and correct, pages being shuffled during document assembly of the Dox Package or input into the system. In preferred embodiments, the page "footer" description is measured for closeness, an example of distance measurement, against the reference page using fuzzy logic matching techniques, and other mathematical techniques as known to one of skill in the art.

Taxonomy Classification/Mapping:

In preferred embodiments, each document page is classified into, or sorted into, one of the taxonomy classes, as defined by the MISMO standards committee or pre-programmed by user using the document learner, or a class designated by a human collaborator. If the putative class identified by the system is unknown and it cannot be classified by the system, the document or representation thereof may be stored in a feedback folder for further manipulation. Taxonomy classification is also done in multiple levels to identify class, sub-class, and version of the document. Taxonomy classification is preferably performed using multiple classification engines. All the outputs of the taxonomy classifier may be flagged or designated as one of four types: (1) classified, (2) multi-class, (3) ambiguous, (4) unknown. Document pages flagged as unknown are submitted to an OCR program from a different manufacturer 212, and re-identified 213 using the same Identifier Engines, 207, 208, 209, 210, 211.

The pages or documents thus far classified may be further evaluated automatically. In preferred embodiments, those documents that fall into categories 2 and 3 are forwarded to the verifier.

Verifier:

Documents that are flagged as either multi-class or ambiguous, or both, are routed through the verifier, although any document used in the system may be routed through the verifier. The document verifier performs a very accurate form of location-based checking for verification of class 214.

Further voting and probability algorithms are preferably used to determine the class for the remaining pages. 215.

The Information Sequencing process is used to create a sequence matrix from the information acquired during the previous steps. 216.

The automatic version detection and page sequencing for some or all documents is done using the Versioning and Sequencing engine. 217. This is done using the Feature Vectors specific to versions and sequencing matrix as captured 216.

Classification Exception Handler:

Document pages that are still not mapped to a taxonomy class due to bad image quality, a new variation of a document, or for other reasons that do not result in immediate identification or classification are flagged as Unknown. Document pages that fall below the confidence threshold value that may be preset or varied by the user, even after the verifier, are sent to exception handling client (Classification client) (i.e., via escalation) 218. There, human collaborators can verify the class, assign a class, or note that the document cannot be identified.

If a human collaborator verifies or changes the class, this information is sent to a feedback box for an incremental learning. During escalation, in preferred embodiments, the human collaborator is presented with an image comprising the header and footer region of the page or document in question, and optionally with an image of the entire page if the image quality is poor. Frequently the identification of the page or sheet may be made in reference to the header and footer information, although display parameters, such as position of the various images on the human collaborator's computer screen and zoom capabilities, may be varied by the human collaborator. Escalation may occur before, but preferably occurs after, the verifier step.

Apply Filter for Classification: For documents having peculiar properties, such as a specific variation of a class of documents, a filter may optionally be applied. An example is if two documents are very close in format and data, but they differ in a very specific property and because of that they belong to different class. A weighted filter, that is a Location Diagram with primary key set for the distinguishing property or feature, is applied so that those can be classified accurately and rapidly. This technique is also used for determination of different versions of documents. For example, two notes may have very similar contents but differ in specific feature such as the absence or presence of an interest rate adjustment clause, need to be put in different classes for business decisions involving an Adjustable Rate Note and/or a Fixed Rate Note. Collating (Class Specific):

Within each taxonomy class as determined to this point, document pages are collated using methods of analysis based on Location Diagrams and Feature Vectors as may be understood by one of skill in the art 219. These methods of analysis determine the sequence, page numbers of pages and sheets within documents. This process of collation is capable of determining not only the class of a document page (which in most cases is determined earlier at the Classification step or the Verifier step) but also the exact identification of a document including the version number of each particular document within the Dox Package. The collation methods also correctly identifies the pagination within the document, and also notes and records the presence of duplicate documents. For example, during collation, the method/apparatus of the instant invention may find and note as identical two identical mortgage notes in a single taxonomy class. This collation process of the instant invention is differentiated from classification technologies known in the art by its ability to distinguish closely related documents. An example of this is that the method/apparatus of the instant invention can pick two mortgage notes out of a Dox Package, correctly paginate them, and identify and log them as separate, but otherwise identical, documents. Pages or documents are then segregated into a logical group determination, and the pages are mapped to a predetermined business-specific or user-determined taxonomy.

In preferred embodiments of the instant invention, the collation process is based on incremental learning and various artificial intelligence ("AI")-based techniques, which may include one or more of the following, such as:

(1) the Location Diagram- and Feature Vector-based feature extraction and page mapping;

(2) SVM and NLP;

(3) an intelligent filter technique taking advantage of header and footer based information;

(4) collation by finding common threads within or between pages, documents, or sets of documents;

(5) finding disagreements based on affinities;

(6) inference-based mapping; and

(7) feature based discontinuity detection and collation, as well as human collaboration.

The collate confidence matrix which is the result of the above-described collation process is preferably used for final formation of documents. The collate confidence matrix represents affinity among various pages, positions of the pages within sets and the confidence of mapping to a particular taxonomy.

Extraction:

In preferred embodiments of the invention, extraction of information or data from the documents or Dox Package that has been captured using the method or apparatus of the invention, and preferably extraction is first done automatically from readily identifiable fields 221 and image snippets of other fields location are re-submitted to the OCR step with a field specific dictionary before repeating the extraction process 220. Using a Location Diagram-based method allows the location of fields even in case of variation between pages or documents within or between Dox Packages 221. Values missed by automatic extraction of these methods may be located by an automatic field locator 222. The automatic field locator uses auto field location based on Location Diagrams 223 and Image based field locator 224. In preferred embodiments of the instant invention, if automatic field locator cannot locate values, the region of the page and/or sheet in question is escalated and the field may be identified with the assistance of a human collaborator by escalation 225. In preferred embodiments, the human collaborator may be shown only the relevant region of a page or sheet (Image Snippet) and may identify the region containing the data to be extracted by simply mousing over the region with the values extracted by further processing 226 and, in preferred embodiments, the location of the value within the document then sent to the feedback folder for future reference in regard to learning and optimization of the system. In related preferred embodiments, the human collaborator indicates exactly where the field is located.

Relevant information, as defined by a pre-determined business-specific application or set by a user, is extracted from documents that have been successfully classified. In some preferred embodiments, each time a Location Diagram is resolved to select a field region, an overall weight may be associated with that solution and used to improve future selection of fields in a particular class of document.

Scrubbing and Verification of Extracted Information:

Extracted values are scrubbed to get exact value 227. Scrubbing further transforms the extracted value to a specific data type. The accuracy of the scrubbed value is verified. Thus, the system provides multiple confidence levels for decision-making. The system generates a Knowledge Object from the scrubbed results. The values with very high extraction confidence but very low scrubbing confidence are sent to a human assisted Field Location Process ("Manual FLP") 225. The system generates a field value from scrubbed results that pass the confidence threshold for the overall process.

Extraction Exception Handler:

Extracted data falling below confidence threshold value is sent to exception handling client (Manual Extraction Process ("MEP")) (i.e., via escalation) 225 & 226. Human collaborators can verify and/or change the data and/or extracted information in reference to the Dox Package. In preferred embodiments, each field subjected to MEP is extracted by a minimum of two human collaborators and the system compares the extracted value. In the event of a discrepancy, the value in question can be sent to additional human collaborators.

In all the steps involving human collaboration, the method/apparatus of the instant invention may optionally keep track of which data was viewed by human collaborators, and how long they viewed the data, in order to detect potential fraud or illicit activities. Information related to exceptions may also be used for statistical learning. In preferred embodiments, the human collaborator mouses over the exact value to be extracted. This is referred to herein as a "snippet" or a "text snippet" and the method/apparatus can pull the snippet and subject it to further scrubbing and processing 227. These snippets of required/specific values may also be extracted and used for formation of Knowledge Objects.

Transformation:

The processes preferred embodiments of the instant invention typically extracts the fields (as they appear in the document) required for various business and/or compliance requirements, then transforms them into facts that can be used further for decision making by an automated rules engine or search engine by packaging these facts and other related information such as text and image snippets, x,y coordinate location of these facts from a Location Diagram into an entity referred to herein as a Knowledge Object. A Knowledge Object 228 is an information matrix with the relationship among all the information entities clearly defined 229.

Knowledge Objects can be used to form Business Objects. A Business Object is a collated set of Knowledge Objects created for use in particular business context such as a MISMO SMART Doc XML file, custom transaction set or electronic document. Business Objects give data a business centered view of the information captured by the method/apparatus. 230 Business Objects are stored and used for business decision making by a Decision Engine. These Knowledge Objects and Business Objects are stored in an electronic data repository which can further be used by a decision engine, 231 a rule engine, or a search engine to make various decisions and/or accelerate, support, or validate decisions.

Further Features

Business Object Formation:

FIG. 4 depicts Business Object formation. The relationships among all Knowledge Objects is established by a method called Collation 401. The output of collating Knowledge Objects is done by referring to a knowledge map which has a business-process-specific knowledge representation of the Business Object required for making business decisions 402. For example a organized Dox Package in the form of a MISMO SMART Doc 403, XML representations of industry standard documents 404 405 406 Industry standard transaction sets defined by MISMO 407 408 409.

FIG. 4A depicts the relationships among the Knowledge Objects. The relationship between Knowledge Object P1 410 and Knowledge Object P2 411 is shown in the figure. I1 and I3 is the set of common features belonging to P1 and P2, I2 is the set of data elements, I4 is set of location co-ordinates (snippets and regions) and I5, I16, I17, and I8 are the other attributes of P2. Since I1 and I3 are common to P1 and P2, the knowledge map is referenced to determine if they have affinity to the same category of Business Object such as a Promissory Note.

As an example FIG. 4B depicts the process of Dox Package creation. Here Document-1, Document-2, Document-n 412 413 414 have their individual attributes. (Attributes from left to right 415 416.) Based on these attributes these documents are mapped to the taxonomy. Here the collation process is used to determine affinity to a Dox Package based on common attributes such as Loan number or borrower name.

FIG. 4C depicts the process of Document formation. Here pages page-1 to page-n 417 418 419 420 421 based on closeness among pages, Feature Vectors and affinity 422 are mapped to different documents and their copies, revisions.

FIG. 4D depicts a Business Object MISMO AU.S. Transaction set 428. The Knowledge Objects extracted from various forms like 1003 429, 1004 434, and Note 433 are combined to form a transaction set for underwriting of a loan using a rules engine.

Incremental Learning:

The system of preferred embodiments of the instant invention performs incremental learning and tuning based on feedback and/or unclassified documents. All Feature Vectors are retuned without actual calculation of relative distances. The incremental learning is based on statistical analysis of exception and tuning.

The system keeps watch on statistical data of the collate, classification and extraction to dynamically tune various control parameters and optimize results. Further, in preferred embodiments, the method/apparatus can readily keep track of where human collaborators reviewed data and how long they accessed the data, thus enabling an operator of the instant invention a certain level of protection against fraud.

Learning:

FIG. 3 depicts the flow of learning in one embodiment of the instant invention. The document samples for the document to be learned, and document-specific dictionaries and generic, as well as domain, dictionaries are loaded in to the Learner's Knowledge Base. 301. The Learner reads the document samples, and if document specific dictionaries are not available, then one is generated from the sample documents. 302. For some specific files, human input such as very specific key phrases and location are provided for learning, if required. 303 Text Feature Vectors are created using image processing, machine learning and Location Diagram based techniques and other methods known in a manner known to those of skill in the art. Here the Feature Vector represents various text features including frequencies, relative locations and Location Diagrams. 304. The distances among the Feature Vectors representing different classes, locating different information are maximized. Weights are assigned to Feature Vectors based on their uniqueness and distance from the other Feature Vectors. 305. Using Statistical techniques thresholds are calculated. 306. If the Feature Vectors uniquely identify document 307 document is flagged and Feature Vectors are loaded 308. Otherwise the Feature Vector is re-tuned to prevent misclassification by maximize the distance to from the wrongly classified document class 309. Similarly Feature Vectors are created based on image features. 310. These sets of Feature Vectors are then mapped to a class. 311. The Feature Vectors are tuned to optimize the results. 312. The documents are flagged and corresponding Feature Vectors are loaded in the system. 313. The text and image based learning process complements each other and can be performed in any order. The output of the learner is a collection of reference-sets that are then stored in a Knowledge Base of the Classifier and Extractor methods to reference.

Regarding Classification:

The system can prepare reference-sets of known classes with title of the class, i.e., taxonomies. The system can use either a dictionary specific to the endeavor domain (i.e., real estate) or a dictionary specific to a document classes.

The system can, based on reference-sets, generate a dictionary for each class. This dictionary also contains a weight for each word. The weight for each word plus a weight for combinations of words is determined based on frequency and Bayesian analysis of word features with reference to document identity.

Learning also generates Feature Vectors based on Location Diagrams for each set (reference-set). The Feature Vectors generated represent precisely that set of documents, or at least most of the documents, in that reference-set.

The method/apparatus can maximize distance between Feature Vectors derived from Location Diagrams to eliminate overlap and give high weight to properties those are specific to the document.

The method/apparatus can also load Feature Vectors from an outside source.

To address the needs of assigning a unique position to each page in a set of documents to its business context, as well as other needs associated with the given business, the instant invention features in one embodiment a method/apparatus that identifies and collates individual documents and revisions of the same document type within a set. The method/apparatus automatically identifies discontinuities using various algorithms to identify groups or logical units of documents. The instant invention takes advantage of its computer and human collaboration and to utilize the strengths of both. The output of the method/apparatus is a Business Object like MISMO Smart Docs. The Business Object is a business-centered Knowledge Object representation useful to a business decision maker. Further, the method/apparatus of preferred embodiments of the instant invention has a method for making decisions based on business processes to select and organize the Business Objects and provide automated decisions in some situations. The Business Object contains a complete collated and bucketed set of documents, complete relationships of KOs for specific process, etc. Further collated documents and information is presented with the business identification furnished and mapped to the business-specific or user-provided taxonomy. Further this method/apparatus measures relatedness among various pages and sorts and identifies documents on the principle of the reference page.

The instant invention, in preferred embodiments, collates pages from the input set of documents into a logical/numerical sequence. The fields required for different business processes are extracted from these collated and taxonomy-mapped buckets. In preferred embodiments, the instant invention also provides for fact transformation so that the information extracted from the pages in the document set is converted into usable form and can be used directly according to various business-specific manners. The instant invention, in preferred embodiments, provides the formation of Knowledge Objects and additionally ready-to-use Business Objects.

The processes of the instant invention, in preferred embodiments, typically extract the information fields required for various business and/or compliance requirements, then transforms them into facts that can be used further for decision making. The decision system used for analyzing the document set provides logical decisions based on the information within and the business context. The instant invention offers a collation system and complete organization and fact extraction solution that forms the information matrix, Knowledge Objects. This allows information flow from paper documents from a wide variety of types of images to decision-making based on error free analysis using the techniques of intelligent mapping available to the operator of the instant invention. The invention is highly scalable because of its dynamic learning ability based on Feature Vectors and ability to create Business Object based on requirement and business process.

The applications for this Business-Object-creation based on Knowledge Objects as are created by processes such as intelligent document collation and extraction of information are not limited to the mortgage and insurance industries. In fact, this method is useful where there is any business process that uses information from unorganized set of documents. All the places where unorganized information from the documents need to be used for business decisions this business method is useful. It can be used for Knowledge Object creation based on information extraction from various sources of images, paper documents, and PDF files. Further, this system can be great help for many processes, both inside the legal field and otherwise, that are based on signed documents and files with information available within the set of documents is distributed across a variety of pages.

Thus, some preferred embodiments of the instant invention feature:

(1) output of data from Dox Packages as Business Objects (e.g. MISMO SMART Doc) that is business-type specific (Underwriting, Servicing, Closing process etc.);

(2) a complete process right from information/document capture to creation of Business Objects which can directly used for automated decision making and also to advise manual decisions;

(3) unique flow with new algorithms;

(4) novel, user-adjustable, and very business-specific representation of information; and/or

(5) making data, or rather the Business Object, available to make facilitate e-mortgage processing as envisioned by MISMO.

In some embodiments, one of the major purposes of the process of the instant invention is Knowledge Object and Business Object creation. The final output is a Business Object and not only a set of classified or sorted documents. Further, in some embodiments, the purpose of the instant invention is not classification or extraction but to create Business Objects like MISMO Smart-Docs from Knowledge Objects, thereby accelerating automated and manual business decisions.

The basic method used for classification is different from current methods. Also the manner in which and sequence the instant invention uses various complementary technologies, such as filtering and voting, makes the method of the instant invention more accurate.

Additionally, the flow of preferred embodiments of the instant invention is uniquely valuable in yielding Business Objects. Various algorithms are used in a manner and sequence to obtain optimal accuracy. Also, the process of preferred embodiments of the instant invention emphasizes feature/knowledge extraction out of Dox Packages with classification and document separation an allied output. The instant invention, in preferred embodiments, locates the knowledge portion within a Dox Package irrespective of slippage and page numbers. Thus, the instant invention may provide information for the downstream business process directly from Dox Package capture to Business Object creation and decision-making based on the Business Objects.

The assembly of technology and algorithms unique to the instant invention in some embodiments may include at least some, or all of, the following in preferred embodiments:

(1) The intelligent information locator of the instant invention may help the business process by locating the business critical information. The location algorithm uses a novel method to provide accuracy.

(2) The method of preferred embodiments of the instant invention identifies all available sources and multiple occurrences of the same information across the Dox Package, i. e., to different versions of the same type document; this enables the user to compare this information and make decisions based on the most recent or relevant information.

(3) The image- and text-based information locator of the instant invention, in preferred embodiments, takes advantage of image and text properties of the documents while locating the information.

(4) The instant invention recognizes that the document boundaries in business context are not as significant as the multiple occurrences and sets of Knowledge Objects that suggest the presence of more than one form of the same types.

(5) The information locator may also indicate versions and facilitates relevant decisions.

(6) The Location Diagram-based method may be used for rapid location of data, and, which in turn, returns the data association with the image.

(7) The Location Diagram based method locates may collect information from proper page irrespective of similarities among the pages, as well as new variations among the forms.

(8) The Location Diagram-based locator can locate appropriate information based on the version of the form.

(9) The Image and Location Diagram based locator can locate the information on forms irrespective of poor quality of images/OCR output.

(10) The system of the instant invention either may extract or make available the relevant portion or the Dox Package for knowledge extraction by an operator by increasing extraction efficiency by up to 5X over prior methods.

(11) The instant invention, in preferred embodiments, features less turnaround or learning time.

(12) The instant invention, in preferred embodiments, features incremental learning as to locations.

(13) The instant invention, in preferred embodiments, features automatic and semiautomatic learning for added flexibility.

(14) The instant invention, in preferred embodiments, features the verifier for verifying location.

(15) The instant invention, in preferred embodiments, features a scrubber which can scrub extraction output.

(16) The instant invention, in preferred embodiments, features the ability of establishing knowledge-based relationship among all the relevant knowledge portions resulting in a rich Knowledge Object that can help in Creation of Business Objects.

(17) The instant invention, in preferred embodiments, features collation of Knowledge Objects to create Business Objects.

(18) The instant invention, in preferred embodiments, features efficient decision making based on Business Objects.

EXAMPLES

FIG. 5 depicts screen shots of output obtained through the use of one embodiment of the instant invention. That is, using a Dox Package analyzed by the method/apparatus as described herein, the following exemplary data was available for analysis for making business decisions.

The invention illustratively described herein can suitably be practiced in the absence of any element or elements, limitation or limitations that is not specifically disclosed herein. Thus, for example, the terms "comprising," "including," "containing," etc. shall be read expansively and without limitation. Additionally, the terms and expressions employed herein have been used as terms of description and not of limitation, and there is no intention in the use of such terms and expressions of excluding any equivalent of the invention shown or portion thereof, but it is recognized that various modifications are possible within the scope of the invention claimed. Thus, it should be understood that although the present invention has been specifically disclosed by preferred embodiments and optional features, modifications and variations of the inventions embodied herein disclosed can be readily made by those skilled in the art, and that such modifications and variations are considered to be within the scope of the inventions disclosed herein. The inventions have been described broadly and generically herein. Each of the narrower species and subgeneric groupings falling within the generic disclosure also form the part of these inventions. This includes within the generic description of each of the inventions a proviso or negative limitation that will allow removing any subject matter from the genus, regardless or whether or not the material to be removed was specifically recited. In addition, where features or aspects of an invention are described in terms of the Markush group, those schooled in the art will recognize that the invention is also thereby described in terms of any individual member or subgroup of members of the Markush group. Further, when a reference to an aspect of the invention lists a range of individual members, as for a non-limiting example, `the letters A through F, inclusive,` it is intended to be equivalent to listing every member of the list individually, that is `A, B, C, D, E and/or F,` and additionally it should be understood that every individual member may be excluded or included in the claim individually. Additionally, when a reference to an aspect of the invention lists a range of individual numbers, as for a non-limiting example, `0.25% to 0.35%, inclusive,` it is intended to be equivalent to listing every number in the range individually, and additionally it should be understood that any given number within the range may be included in the claim individually.

The steps depicted and/or used in methods herein may be performed in a different order than as depicted and/or stated. The steps are merely exemplary of the order these steps may occur. The steps may occur in any order that is desired such that it still performs the goals of the claimed invention.

From the description of the invention herein, it is manifest that various equivalents can be used to implement the concepts of the present invention without departing from its scope. Moreover, while the invention has been described with specific reference to certain embodiments, a person of ordinary skill in the art would recognize that changes can be made in form and detail without departing from the spirit and the scope of the invention. The described embodiments are considered in all respects as illustrative and not restrictive. It should also be understood that the invention is not limited to the particular embodiments described herein, but is capable of many equivalents, rearrangements, modifications, and substitutions without departing from the scope of the invention. Thus, additional embodiments are within the scope of the invention and within the following claims.

* * * * *

Images

View Cart Add to Cart

Top

Home Quick Advanced Pat Num Help

8

Swamy BK, Kulkarni, P, Intelligent Decision Making Based on Pattern Matching and Mind-Maps, WSEAS International Conference on Computers, Athens, pp 493-498, 2006, ISSN 1109-9526 http://direct.bl.uk/bld/PlaceOrder.do?UIN=192046024&ETOC=RN&from=searchengine

Intelligent Decision Making Based on Pattern Matching & Mind-Maps

B. K. SWAMY [a], PARAG A. KULKARNI [b]
[a] Department of Computer Engineering
AISSMSP college Pune-411001
INDIA
bk_swamy64@rediffmail.com, http://www.aicoep.org
[b] Capsilon Research Labs
Capsilon India Pune-411014
INDIA
parag.kulkarni@capsilon.com, http://www.capsilon.com

Abstract: - This paper will represent a novel technique of decision making based on Mind-maps. The Mind-maps of the previously learned activities are stored in database. These Mind-maps are clustered based on their closeness factor and number-weighting. The closeness may be activity specific or decision specific. The classification of Mind-maps is done based on SVM (Support Vector Machine). Every Mind-map has its own behavioral features and a feature extraction based on activity centric mechanism is used to generate a feature vectors. These feature vectors are used for clustering and classification. As the approach takes in to account the decision and related data the way human being perceives it, it has great potential to address difficult decision problems.

Keywords: Mind-map, Support vector machine, Closeness factor, Number-weighting, Clustering.

1 Introduction

Mind-maps are the representation of the relationship between various objects, events and actions in real world the way human mind perceives it. Buzan[1] Elaborated how we can effectively represent various problems. A Mind-map is a diagram for linking words and ideas to a central key word or idea. It is used to visualize, classify, stucture, and generate ideas, as well as an aid instudy, problem solving, and decision making. It is similar to a semantic network or cognitive map but there are no formal restrictions on the kinds of links used. Most often the map involves images, words, and lines. The elements are arranged intuitively according to the importance of the concepts and they are organized into groupings, branches, or areas.

Mind-map creation is done using a map title (title branch), which holds the subject of the map. When the title is on paper, expand the map by filling in the thoughts (child branches) may have about the subject. Each thought may again have its own thoughts, describing that thought in more detail. This way, a Mind-map grows bigger and bigger as shown in Fig.1.

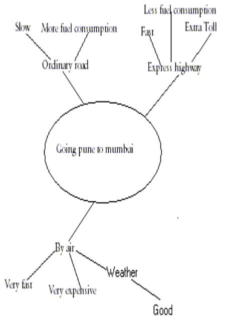


Fig. 1 Example : Mind-map to travel Pune to Mumbai

Once all the relevant information, thoughts and emotions have been collated on to the Mind-map, there is Number-Weighting method for making a dyadic choice [1] as shown in Fig.3.

The main philosophy behind these pattern matching and Mind-map techniques is to identify the best matching past trends to current ones and use the knowledge of how the time series behaved in the past in those situations to make predictions for the future.

A range of closeness factor strategies can be adopted for this matching such as the fuzzy Single Nearest Neighbor (SNN) approach.

This decision-making technique can be used for the applications such as business, finance and medical field. Mind-map techniques can be extended for other applications such as: Brainstorming, Presentations, Note-making, Planning and Summaries.

2. Related work

Some of the research work surveyed in the field of Mind-map is summarized as follows: Tony Buzan developed Mind-maps as an efficient way of using the brain's ability for association. Association plays a dominant role in nearly every mental function, and words themselves are no exception. Every single word and idea has numerous links. The book (Tony Buzan and Bary Buzan, 2003) elaborates how we can effectively represent various problems using Mind-map. But they leave room for storing previously learned activities under the class and using them for decision making based on best matching past trends to current ones and use knowledge of how the time series behaved in the past in those situations to make predictions for the future and hence decision making.

The Mind-map by Vanda North, past president of the International Society for Accelerated Learning and Teaching, and Co-founder of The Brain Trust, gives a clear example of the number-weighting Mind-map. Vanda had to weight a number of personal and professional factors in deciding whether to move her business headquarters or remain where she was (see Fig.2).

Some of the research work surveyed in the field specially applications of support vector machines (SVMs) is summarized as follows: For the pattern recognition case, SVMs have been used for isolated handwritten digit recognition (Cortes and Vapnik, 1995; Scholkopf, Burges and Vapnik, 1996; Burges and Scholkopf, 1997), object recognition (Blenz et al., 1996), speaker identification (Schmidt, 1996), and text categorization (Joachims, 1997). SVM is proved to be a best classification technique for the pattern and text classification. Following this track, we propos to use SVM for classification of Mind-maps.

3. SVM based clustering and classification

Classification is done using SVM technique. The SVM based classification is further used for clustering. The number of classes they have close proximity with decision space are searched with reference to new situation for decision making. We also aim at providing an effective means to collect relevant information to make good decisions.

Mind-map based learning technique establish various links between key activities. These links are divided in three type:

a. Activity link
b. Belongs to link
c. Result link

There are few other types of links also. These links helps in deciding relationships among various nodes. To determine closeness among various maps we use SVM, which will use various features of the map as input. These features are extracted using simple key word based method.

Mind-map1 = f (f1, f2,…fn)

Categorization has become one of the key techniques for handling and organizing mind-map data. Categorization techniques are used to find required information from mind-maps. Since building mind-map classifiers by hand is difficult and time-consuming, it is advantageous to learn classifiers from examples. SVMs are a new learning method introduced by V. Vapnic et. el.[4]. They are well founded in terms of computational learning theory and very open to theoretical understanding and analysis. After reviewing the standard feature vector representation of mind-map, we will identify the particular properties of mind-map in this representation.

The goal of mind-map categorization is the classification of mind-maps into a fixed number of predefined categories. Each mind-map can be in multiple, exactly one, or no category at all. Using machine learning, the objective is to learn classifiers from examples, which perform the category assignments automatically. This is a supervised learning problem. Since categories may overlap, each category is treated as a separate binary classification problem. Each such problem answers the question of whether or not a mind-map should be assigned to a particular category.

The first step in mind-map categorization is to transform mind-maps, which typically are strings of characters, into a representation suitable for the learning algorithm and the classification task. Fig. 2 shows an example feature vector for a list of mind-maps.

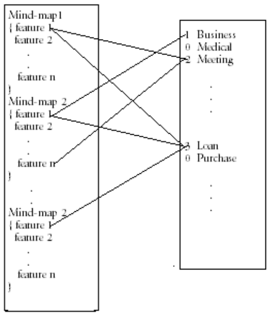
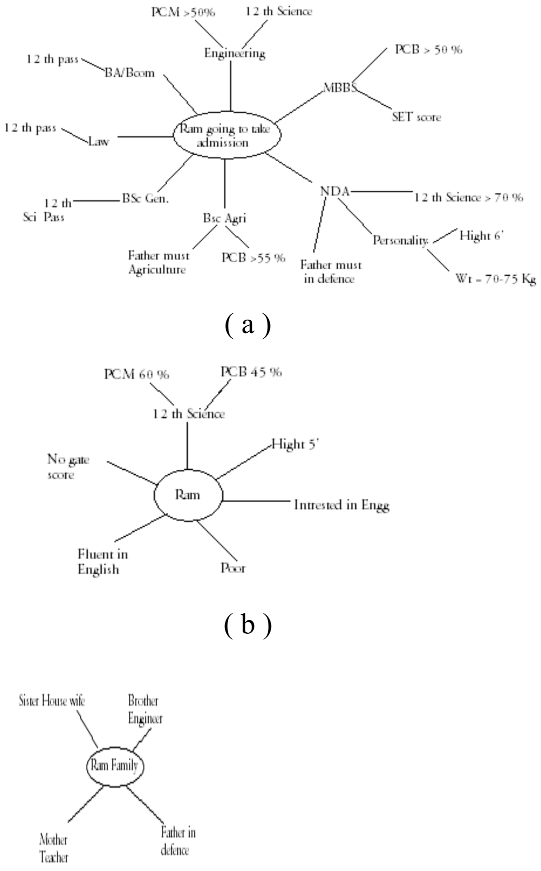


Fig.2 Example: Representing mind-maps as a feature vector.

3.1 Clustering of Mind-maps

Following three Mind-maps (Fig.3 a, b & c) are all ram related Mind maps. For taking decision about course admission all these three related Mind-maps must be clustered. Fig.4 represents the feature vector of the Mind-maps. After clustering we get the four possible alternative:

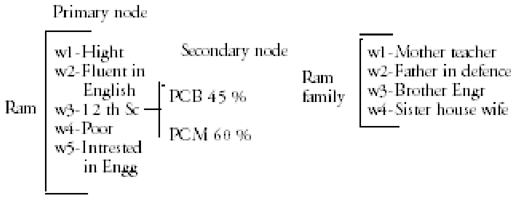


(a)

(b)

(c)

Fig. 3 Mind-map (a) Ram Admission (b) Ram and (c) Ram family



Feature vector representation of Mind-map Ram

Feature vector representation of Mind-map Ram family

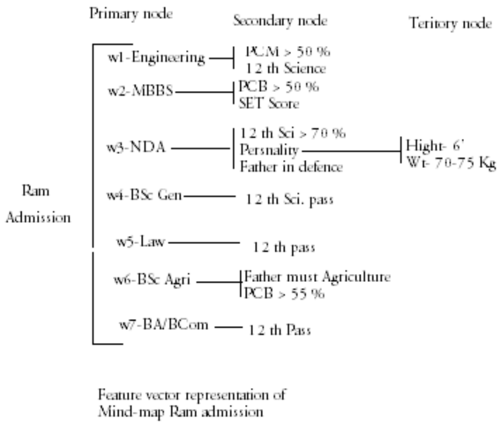


Fig. 4 Feature vector representation of Mind-map (a), (b) & (c)

1. Ram is eligible for Engineering
2. Ram is eligible for BSc Gen.
3. Ram is eligible for Law
4. Ram is eligible for BA/Bcom

But as per the feature vector ram is interested in Engineering, therefore the best choice/decision is to take admission for Engineering.

4. Decision making based on mind-map

4.1 General Decision Making

In general decision making the Mind-map helps to balance competing factor. Let's take the example of deciding whether or not to buy a new car. We require a certain degree of comfort and quality but we don't have a great deal of money. We may therefore have to go for a second-hand car and so will have to weigh up the financial saving against the reduction in reliability and durability.

The Mind-map does not make the choice. However it dramatically increases our ability to make the choice by highlighting the key trade-offs.

4.2 Simple Decision Making

A simple choice of this kind is known as a dyadic decision (derived from the Latin dyas, meaning 'two'). Dyadic decisions are the first stage in creating order. They can be broadly categorized as evaluation decisions, and they involve simple choices such as: yes/no, better/worse, stronger/weaker, more effective/less effective, more efficient/less efficient, more expensive/less expensive.

4.3 Making the Choice

Once all the relevant information, thoughts and emotions have been collated on to the Mind-map, there is Number-Weighting method for making a dyadic choice[1].

4.4 Number-Weighting

If, after completion of the Mind-map, the decision is still not clear, the number-weighting method can be used. In this method, each specific Key Word on either side of the Mind-map is given a number from 1 to 100 according to its importance

4.5 If the Weightings are Equal

If, after completion of the Mind-map, and none of the previous methods has generated a decision, there must be an equal weighting between 'YES' and 'NO'. In a case like this, either choice will be satisfactory, and we may find it useful simply to toss a coin (the ultimate dyadic device) – heads for one option, tails for the other.

During the coin tossing we should monitor our emotions very carefully, in case we find that we really do have a preference. We may think we have decided that the choice is equal but our parabrain may already have made its superlogical decision.

If the coin shows heads, and our first reaction is one of disappointment or relief then our true feelings will finally be revealed and we will be able to make an appropriate choice.

4.6 Dealing with Indecision

In a very few instances all the above decision making methods will fail and we will be left swinging to and fro like a pendulum.

At this point the brain is undergoing a subtle shift from the dyadic (two option) choice to a triadic (three-option) choice. The decision is no longer simply 'yes' or 'no'. It is now:

1. Yes.

2. No.
3. Continue thinking about the choice.

The third option is not only counter-productive but becomes more so the longer it is maintained. Eventually it becomes the choice because that is where our mental energy is being directed.

The simplest solution to this problem is to decide not to make the third decision! In other words, the minute we recognize this spiraling whirlwind on our mental horizon, we should immediately choose 'Yes' or 'No" (the first or second option). The basic principle here is that it is more fruitful to have made some decision and to be implementing it, than to be in a state of paralysis [2].

5. Outline of the IDMBPMM algorithm

Some fundamental details involved in Intelligent Decision Making Based on Pattern matching and Mind-maps framework building is approached, divided in three Subsections: Mind-map creation, classification of Mind-maps and decision making as shown in Fig. 4.

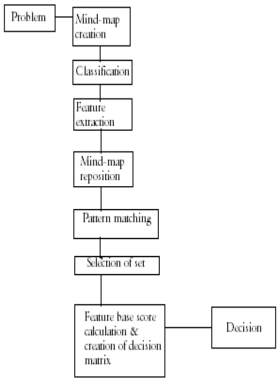


Fig. 4 Framework of decision system based on mind-map and pattern matching.

Steps to construct Algorithm:

1. Problem: Define, understand, and analyze the problem to be solved.

2. Mind-map creation: Mind-map creation is done using a map title (title branch), which holds the subject of the map. When the title is on paper, expand the map by filling in the thoughts (child branches) may have about the subject. Each thought may again have its own thoughts, describing that thought in more detail. This way, a Mind-map grows bigger and bigger as shown in Fig.1.

3.Mind-map classification: The classification of Mind-maps is done based on SVM (Support Vector Machine).

4. Feature extraction: Every Mind-map has its own behavioral features and a feature extraction based on activity centric mechanism is used to generate a feature vectors. These feature vectors are used for clustering and classification.

5. Mind-map reposition*:* The Mind-maps of the learned activities are stored in database.

Pattern Matching: Pattern matching and Mind-map techniques is to identify the best matching past trends to current ones and use the knowledge of how the time series behaved in the past in those situations to make predictions for the future.

A range of closeness factor strategies can be adopted for this matching.

6. Selection of set: Mind-maps are clustered based on their closeness factor and number-weighting.

7. Creation of decision matrix: Creating a chart that allows a team or individual to systematically identify, analyze, and rate the strength of relationships between sets of mind-maps. The matrix is especially useful for looking at large numbers of decision factors and assessing each factor's relative importance.

Steps to use/construct decision matrix

- Identify alternatives: Depending upon the team's needs, these can be product/service features, process steps, projects, or potential solutions. List these across the top of the matrix.
- Identify decision/selection criteria: These key criteria may come from a previously
  prepared affinity diagram or from a brainstorming activity. Make sure that everyone has a clear and common understanding of what the criteria mean.

Also ensure that the criteria are written so that a high score for each criterion represents a favorable result and a low score represents an unfavorable result. List the criteria down the left side of the matrix.

- Assign weights: If some decision criteria are more important than others, review and agree on appropriate weights to assign (e.g., 1, 2, 3).
- Design scoring system: . Before rating the alternatives, the team must agree on a scoring system. Determine the scoring range (e.g., 1 to 5 or 1, 3, 5) and ensure that all team members have a common understanding of what high, medium, and low scores represent.
- Rate the alternatives: For each alternative, assign a consensus rating for each decision criterion. The team may average the scores from individual team members or may develop scores through a consensus-building activity.
- Total the scores: Multiply the score for each decision criterion by its weighting factor. Then total the scores for each alternative being considered and analyze the results.

8. Decision: Depending upon highest total rating the best alternative can be selected.

6. Results

Based on analysis made for health care decision making using pattern based forecasting based on closeness factors [3] and pattern based decision making based on Mind-map with increasing feature vector are given below.

Number of feature vectors	Closeness factor based technique	Mind-map based technique
10	80	83
20	75	81
30	69	79
40	65	78
50	61	75

Dimensions are chosen in accordance with the complexity of the problem. For each type of problems 100 sample outputs are tested to arrive at the conclusion.

7. Conclusion and future work

This paper has proposed a new Mind-map based technique for decision making. When the numbers of dimensions are very high, this technique outperforms closeness factor based statistical technique. As this method works on the way human being perceives the problem. For the problem where inference is very important and complexity is high. The Mind-map based decision technique has potential to give very high success rate.

The combination of this technique with neural network can bring interesting insight into decision-making and can help to improve decision-making accuracy further.

References:

[1] Tony Buzan and Barry Buzan, “The Mind-map Book”, BBC Worldwide Limited, 2003.

[2] Singh S.A., “Long Memory Pattern Modelling and Recognition System for Financial Time-Series Forecasting”, Pattern Analysis and Applications, 1999.

[3]Parag Kulkarni, “Forecasting and Decision Making based on Patterns”, National Conference on AI, IPE Hyderabad, 2003.

[4]Cortes and Vapnik, Burges and Scholkopf, “Isolated handwritten digit recognition”,Support Vector Machines, 2000.

[5]Christopher J. C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition”, Data Mining and Knowledge Discovery 2, 121-167, 1998.

[6]Thorsten Joachims, “Transductive Inference for Text Classification using Support Vector Machines”, In European Conference on Machine Learning, 1998.

9

Shankar Lal, Parag Kulkarni, Amarjeet Singh, "Classification Based on Parametric Partitioning of Solution Space" International Journal on Intelligent Systems, 2010, pp 165-191, Vol 19, No 2, 2010

Parametric Decomposition of Sample Space for Classification

Shankar Lal,[1] Parag Kulkarni[2] and Amarjit Singh[1]

[1]Defence Institute of Advanced Technology, (Deemed University), Girinagar, Sinhgad Road, Pune- 411025, Maharashtra, India; and [2]Capsilon India, 6th Floor Marisoft 1, Marigold Premises, Kalyani Nagar, Pune – 411014, Maharashtra, India

ABSTRACT: In classification task training, sample size and high feature dimensionality enhance the computational complexity of algorithms. Therefore, many algorithms tend to work with reduced feature set and hence ignore contribution of all relevant parameters. This paper presents an **I**ncremental, **S**imple, **E**fficient and **A**ccurate (*I-SEA*) algorithm to consider contribution of all available relevant parameters while keeping the computational complexity and accuracy within acceptable limits. *I-SEA* partitions the training sample space during the preprocessing stage as part of learning. Partitioning is done based on parameter values, taking one at a time, resulting into equivalent classes with respect to the parameter. During classification, these sets are selected based on parameter values of the test case and the result is collated. The algorithm supports incremental learning at a cost of O(m). The average case complexity of classification is $O(m\ (\log_2 n)^2)$.

KEYWORDS: incremental learning, parametric decomposition, classification, data mining, algorithm

1. INTRODUCTION

The task of representation and classification poses a great challenge in the computing world. Consider a classification problem represented by a data set

Correspondence: Prof. Shankar Lal, Defence Institute of Advanced Technology, (Deemed

University), Girinagar, Sinhgad Road, Pune- 411025, Maharashtra, India; shankarlal@diat.ac.in

consisting of sample objects in the form of decision table and a set of test samples to be classified based on the training data. Within the scope of this paper we do not consider the creation of a training dataset, but rather will concentrate on developing and establishing the efficacy of the **I**ncremental, **S**imple, **E**fficient, and **A**ccurate (**I-SEA**) Algorithm for classification. The average computational efficiency has been derived, which proves it highly efficient in comparison with other algorithms for classification. Better Accuracy has been established by comparing the empirical results of the **I-SEA** algorithm with other contemporary algorithms in the field on some popular standard data sets.

Dimensionality of large real-time databases is one of the major challenges faced by the community addressing the classification task. Large-size databases are crippled not only by the number of records but also by the number of attributes (parameters) (Lin & Chen 1997). Various search algorithms peruse the records based on these parameters for matching. The permutation, combinations, correlation, and interplay of the parameters results in increased search time in an exponential manner (Wang R. et al. 2006).

We will use 1-NN and ID3 as a benchmark for comparing the runtime performance of the algorithm. In 1-NN effort is made to find out which of the trained sample point is nearest to the test point (Bian 2006; Cover & Hart 1967). Based on the class label of nearest trained sample point, we classify the test point. To do so, we need some measure of the distance between the test point and various sample points. If the number of objects in the training sample set is |T|=n, and the number of parameters is |P|=m, then finding the distance will be of the order of $O(mn^2)$. Having a smaller |T| will result in poor statistics and hence higher rates of misclassification, therefore reducing the |T| will have undesired effect on accuracy.

Attempts have been made to reduce |P|, resulting in a feature reduction approach (Wang Y et al. 2005; Wang L et al. 2002; Liu & Motada 1998), which helps in controlling the complexity to some extent. There are various approaches for feature reduction, depending largely on the problem domain and availability of primary and secondary features. One approach for feature reduction is using the concept of “Redutcs” and “Core” in the Rough Set Theory (Anazida et al. 2006, 2007; Pawlak 2005, 1991). Reduct and Core are “necessary and sufficient” for preserving the classification structure derived from the training sample (Pawlak 2005; Ye & Chen 2002). However, constructing the reduct has been reported as NP-hard (Jing-Kai et al. 2005; Wong & Ziarko 1985), and some investigators have reported it as an NP-complete problem (Wang R. et al. 2006; Ziarko, 1991; Min et al. 2006).

In addition, several authors (Wang Y. et al. 2005; Wang L. et al. 2002; Wenkee & Salvatore, 1999) have presented studies in various domains and tried to obtain a reduced feature set for optimized classification valid in respective problem domains, with varying degree of accuracy. Despite the success of Feature reduction, this approach has several shortfalls. The assumption of successfully reducing the features is not always true, as has been demonstrated by the Rough Set theory. In several problems, even after reduction we are left with sufficiently large number of attributes (Wenkee & Salvatore 1999; Lee H. et al. 2006; Lee 1999).

The metrics for quantifying relevance of parameter can be devised (Wenkee & Salvatore 1999; Miao & Wang, 1998; Tamilarsan et al. 2006), but ignoring the least relevant parameter can also affect the classification task. Therefore, in the I-SEA algorithm we consider all necessary parameters. I-SEA has the capability of identifying the circumstances, when the given list of parameters is not sufficient to classify and hence, more parameters can be scrutinized or secondary parameters can be devised for classification in *incremental* manner.

The I-SEA Algorithm considers the value of single parameters, one at a time, to decompose the solution space. Each value of one parameter induces partition in the solution space, which results in disjoint equivalence classes consisting of the objects in the training data set. Each equivalence class represents objects in the solution set for the particular value of parameter. We preserve FILES for such solution sets for all the values of all the parameters. The FILES are indexed on the value of partitioning parameter. This approach leads to faster retrieval of objects from a set related to a particular value of a parameter, while classifying the test sample, which amounts to a preprocessing stage for which worst case complexity is $O(mn^2)$. The above discussion is valid for discrete parameters only. In the case of continuous parameters, the number of values of a parameter could be large. Therefore continuous parameters should be **discretisized** by suitable means.

To classify a test case, we select the appropriate partial solution set based on the parameter and its value. This approach results in a highly efficient content-based retrieval of the set, which holds all the objects matching the particular value of the parameter under consideration. Thus, associated with every parameter for the test case, we get a solution set. Taking all the parameters into consideration will result in “m” solution sets. The intersection of such solution sets gives us the desired result.

In case we have to add a new record or object to the leaning structure, this can be done by updating the '**m**' existing partial sets, depending on the value of parameters of the new object to be learned. The values of parameter are assumed to be within the closed set of the original problem. Although the model can be easily extended to the open set, this paper will not discuss this aspect. The details of learning, classification, and the incremental learning algorithm and its analysis are presented in the subsequent sections.

Approaches such as ID3 and its derivatives (Rokach & Maimon 2001; Mitchell 1997) also consider parameters based on information gain theory and develop a decision tree to achieve the decomposition of sample space. Such approaches, however, are highly prone to fall into local minima. Backtracking is not available in ID3 (Maimon & Rokach, 2002, 2001; Maimon & Last 2000). The I-SEA algorithm leads either to a solution (Single class) or to information that a solution could not be reached (null solution) with current learning and hence requires more training, or ends up in a confused state with multiple classes signifying need of more parameters to resolve.

Another subtle difference between ID3 and I-SEA algorithm is in the approach of building the solution tree. In ID3, the complete solution tree is built beforehand. For any change in base configuration the tree has to be modified, which is a costly operation with complexity of $O(mn^2 \log_2 n)$ (Duda et al. 2001). In I-SEA, the parameters are not given any precedence and no hierarchy is built beforehand. For each test case, the solution is built parameter by parameter. In the case of any new training, the sample is to be added to the pre-existing solution sets; it can be done efficiently in O(m) time. This helps the I-SEA algorithm to be flexible for incremental learning from new training sample.

A major emphasis in ID3 is on a criterion for the selection (Mitchell 1997) of parameters for attribute decomposition, in the case of I-SEA algorithm, all parameters are considered equally important for partitioning and hence, any specific sequence is not relevant. Intersection of the partial solution sets in any order will result same final solution set. The I-SEA algorithm is discussed in detail in the sections to follow.

2. RELATED WORK

The Field of Classification is full of challenges. Many researchers have contributed to the growth of this field (Mukkamala et al. 2005; Rao et al. 2003; Dubois

& Prade 1990; Domingos & Pazzani 1997). Yet, the complexity of the problem is such that it has not yet been generalized. Based on the problem domain, the authors emerged with effective algorithms, but no single algorithm can claim success across the problem domains. This deficiency has resulted in large number of approaches to solve the problem, each successful in some manner (Witcha et al. 2006; Peng et al. 2004; Cai & Guan 2003; Ghosh et al. 2000). In this section we discuss some of the approaches that try to improve upon the complexity of the problem.

2.1 Feature Selection Using Rough Set theory

Rough sets have been introduced as a tool to deal with inexact, uncertain, or vague knowledge in many branches of artificial intelligence (Pawlak 2005). The investigation of the rough set methodology for data mining is a challenging research area with promise of high payoffs in many domains (Srinivasa et al. 2007).

An information system defined as $S = (U, A, \{d\}, V, f)$, where U is a non-empty, finite set of objects called the universe, V is a set of values of attributes in A such that f: $U \to V_a$ for any a in A, where V_a is called the domain of a, and $\{d\}$, includes decision attributes. The information system is also called a decision table. Each non-empty subset $B \subseteq A$ determines an indiscernibility relation as follows:

$RB=\{ (x,y)$ in $U\times U$:$a(x)=a(y)\ \forall a$ in $B \}$.

RB partitions U into a family of disjoint subsets

U/RB called a quotient set of U:

$U/RB= \{(x)B$: x in $U\}$,

where $(x)B$ denotes the equivalence class determined by x with respect to B, i.e.,

$(x)B= \{y$ in U: (x,y) in $RB\}$.

Given a set (or concept) $X \subseteq U$, and an equivalence relationship $RB \subseteq A$, one can define X's lower and upper approximations:

Lower Approximation, $\underline{R}\,B(X) = \{x$:($x$ in U) and $((x)B \subseteq X)\}$

Upper Approximation, $\overline{R}\,B(X) = \{x$:(x in U) and $((x)B \cap X \neq \emptyset)\}$

The lower approximation $\underline{R}\ B\ (X)$ is the set of objects that belong to X with certainty, while the upper approximation $\overline{R}\,B\ (X)$ is the set of objects that possibly belong to X. The pair ($\underline{R}\,B\ (X)$, $\overline{R}\,B\ (X)$) is referred to as the Pawlak rough set of X with respect to B (Wikipedia 2008; Pawlak 2005, 1991).

Reduct is a subset of attributes such that $(x)_{RED} = (x)_A$, that is, the equivalence classes induced by reduced attribute set *RED* is the same as the equivalence class structure induced by full attribute set *A*. Attribute set *RED* is minimal in the sense that for any attribute. In other words, no attribute can be removed from the set *RED* without changing the equivalence classes $(x)_A$. A reduct can be thought of as a sufficient set of features to represent the classification structure. Reduct of an information system is not unique (Wikipedia 2008; Pawlak 2005, 1991). There may be many subsets of attributes that preserve the equivalence class structure expressed in the information system.

The set of attributes that is common to all Reduct sets is called the Core. The core is the set of attributes that is possessed by every reduct and therefore cannot be removed from the information system without causing collapse of the equivalence class structure. The core may be thought of as the set of necessary attributes for the classification structure to be preserved (Wikipedia 2008; Pawlak 2005, 1991).

Finding optimal reducts is NP-Complete (Lee 1999; Mukkamala et al. 2005; Rao et al. 2003) and NP-Hard (Jing-Kai et al. 2005; Wong & Ziarko 1985), hence there is no fast and reliable way to find them in deterministic way. Also for incremental learning, we must find out the reduct and core for each new training point introduced as the set of reduct and core can vary with the addition of a new training sample. Therefore it is not suitable for incremental learning.

2.2 Feature Selection Based on Parameters Relevance

Based on the relevance of parameters, a suitable set of parameters is identified for the classification task, thereby reducing the computational complexity for various classification algorithms to large extent. The technique has been successfully applied to arrive at the 6 most relevant parameters of 41 parameters for the multiclass classification task in the field of network intrusion detection (Wenkee & Salvatore 1999; Lee W. et al. 1999; Sung & Mukkamala 2003; Tamilarsan et al. 2006).

As part of the feature selection experiments, Chi-Square (χ^2) analysis, logistic regression, normal distribution, and beta distribution experiments are performed to arrive at the relevance of feature (Wenkee & Salvatore 1999; Lee W et al. 1999; Sung & Mukkamala 2003; Tamilarsan et al. 2006). A set of sufficient features, which reduces the error probability, is used. Using features selected by relevance, authors have reported accuracy ranging from 48% to 99.6%. Noteworthy is that for the same problem, different methods lead to different sets of the most relevant parameters, each resulting in different accuracy (Anazida et al. 2006, 2007). Because the problem involves a large training set and 41 dimensions, reduction to just 6 parameters makes it faster but uncertain in terms of error surface. One cannot deny that a higher number of relevant parameters will result in better accuracy.

2.3 ID3 Algorithms

Iterative Dichotomiser 3 (ID3) is an algorithm used to generate a decision tree based on attribute decomposition (Mitchell 1997). Several variants of this algorithm have been developed, all demonstrating similar powers in learning and classification efficiency (Mansour & David, 2000). The original algorithm lacks in backtracking, and subsequent algorithms have been refined to incorporate this feature. As part of learning, a decision tree is created by: (i) taking all unused attributes and counting their entropy concerning training samples, (ii) choosing an attribute for which entropy is minimum, and (iii) making a node in the tree containing the attribute with minimum entropy.

The entropy is given by, Entropy(S) = $\Sigma_{i=1,c}$ (– p_i $\log_2$ p_i), where p_i is the proportion of S belonging to class i. c is the possible values of class that a target may attain. Based on the entire attribute, a decision tree can be grown that can be subsequently used for classifying a test sample (Mitchell 1997).

The complexity of creating the ID3 decision tree for a 2-class problem is $O(mn^2$ log2 n) (Duda et al. 2001), where m is the dimension and n is the training sample size. The classification complexity is of the order of $O(\log_2 n)$, which makes it quite efficient during classification, however the algorithm is highly prone to converge to local solutions (Maimon & Rokach, 2002; Mitchell, 1997). The algorithm is highly sensitive to the training points. The alteration of even a single training point may lead to radically different classification. The algorithm is not suitable to incremental learning. For any change in the dimensionality of the training sample or modification of a single training sample, the decision tree must be rebuilt, which is a costly operation.

2.4 In *k*-Nearest Neighbours Algorithm

An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbours (Bian, 2006; Cover & Hart, 1967). *k* is a positive integer, typically small. If *k* = 1, then the object is simply assigned to the class of its nearest neighbor. The neighbors are taken from a set of objects from training set that has the correct classification. In this approach, no explicit training step is required. To identify neighbors, the objects are represented by position vectors in a multidimensional feature space. In its simplest form it uses “Euclidean distance”, although other distance measures, such as the “Manhattan distance”, could be used instead (Mitchell 1997).

The *k*-nearest neighbor algorithm is sensitive to the local structure of the data. A major drawback to using this technique to classify a new vector to a class is that the classes with the more frequent examples tend to dominate the prediction of the new vector, as they tend to come up in the *k* nearest neighbors when the neighbors are computed due to their large number. The accuracy of the *k*-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. The complexity of classification is $O(mn^2)$, which amounts to a computationally intensive task.

3. “I-SEA” CLASSIFIER MODEL

The classifier developed in this paper has been done keeping the simplicity of approach, computational efficiency of algorithm, accuracy of classification, and flexibility of learning into consideration. The model is divided mainly into two phases. Phase-I is the learning phase, which takes training samples and their respective classes as input and partitions objects into equivalence classes based on parameter values. Phase-II is the classification phase, in which the test case is taken as input and its class is decided as the output. An optional phase of Incremental Learning has been introduced for learning new training samples. The model is developed considering a set of data is available. The Formal model is developed as follows.

3.1 Definitions

This sub section contains definitions related to the classifier modeling.

Definition 1: Let the information set I be represented as, I = {U,A} where U is Universal set of objects and A is finite set of attributes.

Definition 2: Let universal set, U = {O_1, O_2, O_3, O_4, O_5, ……. O_n}, where n represents number of training samples points.

Definition 3: Let attribute set, A = {P_1, P_2, P_3, P_4, P_5,….. P_m, P_d}, where **m** represents number of attributes.

Definition 4: $P_d \in A$ is Decision attribute. 'd' can be m+1 or 0 or any other possible position in the set A, as may be dictated by the problem domain at hand.

Definition 5: VP_i: Set of possible values of attribute P_i. Say, VP_1={0,1,2,5}; VP_2={0,1,3,4} ; VP_3={1,2,3};……; VP_m={…..}, Assuming all values are discrete and ∀i VP_i is closed.

Definition 6: Let C be the set of classes, C = {c_1, c_2, c_3, c_4, c_5,….. c_r}, where I is *r-class* system. ∀i $O_i \in U$, $\exists c_q \in C$, $1 \le q \le r$| $P_d(i)= c_q$. i.e. each sample in U has a decision value from C stored under column P_d in respective row.

3.2 Phase – I: Learning Phase

Lemma 1: ∀i∀x: $P_{i\neq d}$ = x, x∈ VP_i, induces k_i =| VP_i | partitions of the U, namely SOL_i, where SOL_i ={ SP_{ix1}, SP_{ix2}, SP_{ix3}, …… SP_{ixki}}, $1 \le i \le m$, $i \neq d$, SP_{ix} is an equivalent class within SOL_i.

e.g. SOL_1 ={ SP_{1x1}, SP_{1x2}, SP_{1x3}, …… SP_{1xk1}}, k_1= |VP_1|; for P_1

SOL_2 ={ SP_{2x1}, SP_{2x2}, SP_{2x3}, …… SP_{2xk2}}, k_2= |VP_2|; for P_2

……

SOL_m ={ SP_{mx1}, SP_{mx2}, ……….. SP_{mxkm}}, k_m= |VP_2|; for P_m

Total Number of such solution sets obtained will be $K=\Sigma_{i=1,m}\, k_i$

3.3 Phase-II: Classification Phase

Consider a test case T with parameter values as {t1, t2, t3,…, tm,td=?}, where td is unknown. Assume no parameter is missing. Aim of classification is to predict td from the SOLi generated during learning phase –I .

Lemma 2: ∀i: From SOL_i for (P_i= t_i), we get S_i ={SP_{ixj} | x_j= t_i, $1 \le i \le m$ and $1 \le j \le k_i$}. Thus we have *'m'* Partial Solution set, PS={S_1, S_2, S_3, S_4,…. S_m}.

Lemma 3: $\forall i$, $1 \le i \le m$ $(P_i = t_i) \Rightarrow (\wedge_{1\le j \le n} O_j, \forall O_j \in S_i) \wedge \neg (\vee_{1\le j \le n} O_j, \forall O_j \notin S_i)$, where $S_i \in PS$, as defined in Lemma 2.

Theorem 1: $\wedge_{1\le i \le m} (P_i = t_i) \Rightarrow \cap_{i=1,m} S_i$

Proof: $\forall i$, $1 \le i \le m$ $(P_i = t_i)$ selects objects from set S_i as per above definitions.

$(P_1 = t_1)$ will select objects from S_1, fulfilling the property $(P_1 = t_1)$ and it will mark all other objects not in S_1 as not possible candidates for solution.

$\therefore$ $(P_1 = t_1) \Rightarrow (\wedge_{1\le j \le n} O_j, \forall O_j \in S_1) \wedge \neg (\vee_{1\le j \le n} O_j, \forall O_j \notin S_1)$

Similarly,

$(P_2 = t_2)$ will select objects from S_2,

$\therefore$ $(P_2 = t_2) \Rightarrow (\wedge_{1\le j \le n} O_j, \forall O_j \in S_2) \wedge \neg (\vee_{1\le j \le n} O_j, \forall O_j \notin S_2)$

$(P_1 = t_1) \wedge (P_2 = t_2)$ will select objects and place them in $S_1 \cap S_2$, fulfilling the property $\{(P_1 = t_1) \wedge (P_2 = t_2)\}$

Similarly, Intersection of all partial solution sets, $S_1 \cap S_2 \cap S_s \cap S_4 \ldots\ldots \cap S_{m-1} \cap S_m$ for all the "m" parameters will contain only those objects for which,

$(P_1 = t_1) \wedge (P_2 = t_2) \wedge (P_3 = t_3) \wedge (P_4 = t_4) \wedge \ldots\ldots (P_{m-1} = t_{m-1}) \wedge (P_m = t_m)$

Hence, in general, we get $\wedge_{1\le i \le m} (P_i = t_i) \Rightarrow \cap_{i=1,m} S_i$

e.g. Let $U = \{O_1, O_2, O_3, O_4, O_5, \ldots\ldots O_{10}\}$,

Let $S_1 = \{O_1 \wedge O_5 \wedge O_7\}$, $S_2 = \{O_2 \wedge O_5 \wedge O_7 \wedge O_{10}\}$ and $S_3 = \{O_5 \wedge O_6 \wedge O_9 \wedge O_{10}\}$.

We get. $S_1 \cap S_2 \cap S_3 = O_5$ (1)

Now consider,

$(P_1 = t_1) \Rightarrow (O_1 \wedge O_5 \wedge O_7 \quad) \wedge \neg (O_2 \vee O_3 \vee O_4 \vee O_6 \vee O_8 \vee O_9 \vee O_{10})$

$(P_2 = t_2) \Rightarrow (O_2 \wedge O_5 \wedge O_7 \wedge O_{10}) \wedge \neg (O_1 \vee O_3 \vee O_4 \vee O_6 \vee O_8 \vee O_9)$

$(P_3 = t_3) \Rightarrow (O_5 \wedge O_6 \wedge O_9 \wedge O_{10}) \wedge \neg (O_1 \vee O_2 \vee O_3 \vee O_4 \vee O_7 \vee O_8)$

$(P_1 = t_1) \wedge (P_2 = t_2) \wedge (P_3 = t_3) \Rightarrow (O_1 \wedge O_5 \wedge O_7) \wedge \{\neg O_2 \wedge \neg O_3 \wedge \neg O_4 \wedge \neg O_6 \wedge \neg O_8 \wedge \neg O_9 \wedge \neg O_{10}\}$

$\wedge (O_2 \wedge O_5 \wedge O_7 \wedge O_{10}) \wedge \{\neg O_1 \wedge \neg O_3 \wedge \neg O_4 \wedge \neg O_6 \wedge \neg O_8 \wedge \neg O_9\}$

$\wedge (O_5 \wedge O_6 \wedge O_9 \wedge O_{10}) \wedge \{\neg O_1 \wedge \neg O_2 \wedge \neg O_3 \wedge \neg O_4 \wedge \neg O_7 \wedge \neg O_8\}$

$(P_1 = t_1) \wedge (P_2 = t_2) \wedge (P_3 = t_3) \Rightarrow (O_1 \wedge \neg O_1) \wedge (O_7 \wedge \neg O_7) \wedge (O_2 \wedge \neg O_2) \wedge (O_{10} \wedge \neg O_{10}) \wedge (O_5)$ (2)

$(P_1 = t_1) \wedge (P_2 = t_2) \wedge (P_3 = t_3) \Rightarrow (O_5) = \{O_1, O_5, O_7\} \cap \{O_2, O_5, O_7, O_{10}\} \cap \{O_5, O_6, O_9, O_{10}\}$ (3)

From Eq. (1) and Eq. (3) it follows

$$(P_1 = t_1) \wedge (P_2 = t_2) \wedge (P_3 = t_3) \Rightarrow S_1 \cap S_2 \cap S_3 \quad (4)$$

Using principles of strong mathematical induction, the general case can be proved. Therefore $\wedge 1 \le i \le 3$ (Pi= ti) $\Rightarrow \cap$ i=1,3 Si.

***3.3.1** Decision making.*

Definition 7: Final Solution Set, FSS: Taking intersection of all $S_i \in$ PS, We get, Final Solution Set, FSS = $\cap_{i=1,m} S_i$, FSS $\subset$ U.

Definition 8: Decision Vector: A vector table representing class of the objects in the training sample.

Definition 9: Class Frequency Vector: A tuple, CFV = $\{(c_q,f_q)\forall q \mid 1 \le q \le r\}$, where f_q represent the number of time $P_d = c_q$ has occurred in the FSS, is termed as Class Frequency Vector.

Definition 10: Final Solution Cardinality: FSC is defined as number of object in the final solution set and is given as FSC = $|\cap_{i=1,m} S_i|$

Definition 11: Hamming Distance: Hamming distance H_{dis}, between the test case and any object in the training sample with same number of attributes is the number of attributes for which the respective parameters are different. Put another way, it measures the minimum number of mismatch in the parameter values.

Decision conditions

Condition 1: if FSC >0, then there are some samples in the training set for which H_{dis}=0. For all objects in FSS, find the classes from the decision vector.

case I: If class is unique then we get a solution

case II: If multiple classes are reported then the classifier is termed as "CONFUSED".

Condition 2: if FSC = 0, then $|\cap_{i=1,m} S_i| = \phi$, $H_{dis}>0$, implies that training set is insufficient. Objects with minimum H_{dis}, can be selected and based on the class frequency vector (CFV) a probabilistic decision can be made.

***Decision 1*:**FSC = 0 $\Rightarrow$ T $\notin$ U, Training dataset is not trained enough to predict t_d.

Action to be taken: (i) Report as Failed to get a solution.

(ii) Request for more training.

(iii) Resort to incremental learning if class given externally

***Decision 2*:** FSC $\ge$1 $\Rightarrow$ (SS::Single Solution) $\vee$ (MS::Multiple Solutions) found.

$\forall O_i \in$FSS, create a class frequency vector tuple CFV = $\{(c_q,f_q)\forall q \mid 1 \le q \le r\}$,

where f_q represents the number of time $P_d = c_q$ has occurred in the FSS.

Case SS: if $\exists$i such that $(c_i,f_i>0)$ and $\forall j \ne i$, $(c_j,f_j=0)$, then $t_d=c_i$, $\therefore$T is classified as c_i.

Action to be taken: (i) Report as success

(ii) Deliver the Class c_i as result.

Case MS: if ∃i such that (c_i,max(f_i>0)) and ∃j≠i, (c_j,f_j>0), then Classifier is in CONFUSED state.

Action to be taken:

(i) Report as CONFUSED

(ii) (Classify T as t_d=c_i) OR (Suggest to add more attributes to resolve)

4. "I-SEA" ALGORITHM APPROACH

```
Algorithm (Pseudo Code)
Phase I : Partition Phase(One time Learning from n
records in TDF)
Input: Training Data Set File (TDF)
Output:
     (i) Decision Vector File (DVF)
     (ii)Sets based on parameter values.
     (iii) Training Sample Size (tsize)
Step 1.1:
      Store n in tsize; // n is training sample size
      open TDF; //Creating DVF
      for (all records in TDF){
        read $P_d$; //$P_d$ is decision value of the object
        store $P_d$ in DVF;
      } close TDF,DVF; //Worst Case:O(n)
Step 1.2: Create $SP_{ixj}$
    open TDF;
      for (i = 1 to m){
       populate $VP_i$[];//[Def. 5]
      }
      for (i= 1 to m){//m parameters excluding $P_d$
        k = | $VP_i$ |; //k is cardinality,[Def.5,Lemma 1]
        for(j=1 to k){//for parameter $P_i$
          str = $VP_i$[j];//str holds $j^{th}$ value of $P_i$
          fname= SP+i+"\"+str;// name indexed on $P_i$ and
                              its value
        open fname;cnt=0;
        for(obj=1 to n){ // n is number of objects in
                  the training data
           if(value $O_{obj}$ for $P_i$ in TDF matches str){
             append obj in fname;
        cnt++;
```

```
          }
        }
        insert cnt as first element in fname;
        close file fname;
        rewind TDF;
      }
    }
    close TDF; //Worst case $O(mn^2)$;
Phase II : Classification Phase
Input: Test Case T={$P_1$, $P_2$, $P_3$, $P_4$, $P_5$,… ,$P_m$}
Uses: Equivalent classes created in the phase I.
Output: Classification results
    (a)CLASS of T or (b)CONFUSED or(c)NO SOLUTION
FOUND
Complexity:Worst:$O(mn\log_2 n)$;Average:$O(m(\log_2 n)^2)$
Step 2.1: LOAD Decision Vector Table from DVF
   DVT[1:n]; // Decision Vector Table [Def. 8]
     CFV[1:n]; //Class frequency Vector [Def. 9]
     S[1:n]..S[m:n]//m Ordered Partial Solution sets
   FSS[1:n];// Final Solution Set
     for(i=1 to n){
      decision= read DVF;
      DVT[i]=decision; // Populate DVT[]
     }
Step 2.2: Create Partial Solution Sets
     for (i=1 to m){
      str = $P_i$ ;//i^th parameter of test case
      fname= SP+i+"\"+str;
      open fname;
      cnt = read fname;//no. of records in fname
      for(j = 1 to cnt){
    obj = read fname;
        S[i,j] = obj;//list of all objects in fname;
      }
     }
Step 2.3: Create Final Solution Set, //Form [Def. 7]
     FSS[1:n]=Intersect(S[1:n],S[2:n]..S[m:n]);
Step 2.4: Decision Making
     fsc=|FSS|; // Final Solution Cardinality
     if(fsc==0){
      return "NO SOLUTION FOUND"; //|FSS| = $\phi$
     }
     if(fsc>=1){
      maxdec=0;
```

```
        for(i= 1 to fsc){
          dec=DVT[FSS[i]];
          CFV[dec]++;
         If( dec>= maxdec) maxdec=dec;
        }
        classcnt =0;
        for(j=1 to maxdec){
          if((CFV[j]>0) classcnt++;
        }
        if(classcnt ==1){
          CLASS = CLASS corresponding to CFV[dec]>0;
          return "CLASS FOUND";
        }
        else return "CONFUSED STATE";
```

Incremental Learning Phase:(Learning from new records on need basis)

```
Input: (i) New record with a class decision value
       (ii) Decision Vector File (DVF)
       (iii) Equivalent classes created in the phase I.
       (iv) tsize

Output:
    (i) Revised Decision Vector File (DVF) and tsize
    (ii)Selectively Revised Sets based on parameter
    values (only one set per parameter will have to be
    revised).
Step 3.1: open DVF; //Revising DVF
       read $P_d$; //$P_d$ is decision value of the object
       append $P_d$ to DVF;
      } close DVF; //Worst Case:O(1)
Step 3.2: Create $SP_{ixj}$
   open tsize; original training sample size
   read n from tsize;
   n++;
   update tsize by new value of n;
   close tsize;
   obj=n; //changed value of training samples size
     for (i= 1 to m){//m parameters excluding $P_d$
        str = value of $P_i$;
        fname= SP+i+"\"+str;
       open fname; //Create if does not exist
         read cnt from fname;
    cnt++;
       append obj to fname;
```

```
        insert cnt as first element in fname;
        close file fname;
    }
     }//Worst case O(m);
```

5. ANALYSIS

Let us consider the P_1, P_2, P_3, P_4, P_5,…, P_m , *'m'* parameter problem. Each parameter causes partitioning of k_1, k_2, k_3, k_4, k_5,….., k_m, corresponding to the respective parameters, i.e. P_1, creates cardinality of equivalent set as k_1, P_2 as k_2…P_m as k_m.

For a given test case, let us select the parameters in ascending order based on k_i. This will result in parameter throwing the smallest partial solution set as first set. Since our partial solution sets are ordered according to their position in the sample space, we are thrown with a problem of finding an intersection of two ordered sets (Yates 2004). Thus, the first intersection can be achieved in $k_1 \log k_2$ comparisons. This will throw the maximum size of the resulting partial solution as k_1. The second intersection can be achieved in worst case of $k_1 \log k_3$. This will again throw the maximum size of the partial solution as k_1. Thus for all the m parameters, the number of comparisons during the intersection process is capped by C.

Hence Comparisons, C

$$C = k_1 \log k_2 + k_1 \log k_3 + k_1 \log k_4 + k_1 \log k_5 + \mathbf{\ldots..} + k_1 \log k_m$$

$$= k_1 \{\log k_2 + \log k_3 + \log k_4 + \log k_5 + \mathbf{\ldots..} + \log k_m\}$$

$$C = k_1 \sum_{i=2}^{m} \log_2 k_1$$

Under the assumption of equal distribution, k_i will be some fraction of n, say

$k_1 = \frac{n}{a_j}$, we get,

$$C = \frac{n}{a_1}\left\{ \log_2 + \frac{n}{a_2} + \log_2 + \frac{n}{a_3} + \log_2 + \frac{n}{a_4} + \log_2 + \ldots + \frac{n}{a_m} \right\}$$

$$C = \frac{n}{a_1}\left\{ \log_2\left(n^{(m-1)}\right) - \log_2\left(a_2.a_3.a_4\ldots .a_m\right)\right\}$$

In Worst case we have, $a_l = a_2 = a_3 = a_4$ …..= a_m = 2, we get,

$$C = \frac{n}{a_1}\left\{ (m-1)\ ((\log_2 n)\ -1)\right\}$$

Ignoring small constants, we get,

$$C = \frac{n}{a_1}\ [m \log_2 n] \qquad (5)$$

Case 1: Worst case a_l = 2, we get O(mnlog$_2$ n)

Case 2: For average case we have a_l = x, ranging from 2 to n, $a_2 = a_3$ …..= a_m =2,

From Eq. (5) we have,

$$C = \frac{n}{a_1}\ [m \log_2 n]$$

$$C_{avg} = \frac{n[m \log n]}{n} \sum_{x=2}^{n} \frac{1}{x}$$

$$C_{avg} = m\ (\log_2 n)\ (\log_2 n) \qquad (6)$$

Therefore average Classification complexity is **O(m (log$_2$ n)^2)**

6. AN ILLUSTRATIVE EXAMPLE

A new algorithm can be best understood with help of an illustrative example covering various facets of the algorithm. In this example, we present an information system and various test cases. In some test cases, some parameters are missing. The example shows resilient nature of the algorithm. Let us consider the sample information system I ={U,A} as given in Table 1.

TABLE 1

Sample Information System for Illustrative Example

Object	**P_1**	**P_2**	**P_3**	**P_4**	**P_5**	**Decision**
O_1	1	2	0	1	1	D_1
O_2	1	2	0	1	1	D_1
O_3	2	0	0	1	0	D_2
O_4	0	0	1	2	1	D_3
O_5	2	1	0	2	1	D_4
O_6	0	0	1	2	2	D_5
O_7	2	0	0	1	0	D_2
O_8	0	1	2	2	1	D_6
O_9	2	1	0	2	2	D_7
O_{10}	2	0	0	1	0	D_2

Where,

$U = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8, O_9, O_{10}\}$;

$A = \{\{P_1, P_2, P_3, P_4, P_5, \}$, Decision column$\}$

$VP_1=\{0,1,2\}$; $VP_2=\{0,1,2\}$; $VP_3=\{0,1,2\}$; $VP_4=\{1,2\}$; $VP_5=\{0,1,2\}$;

Considering all the parameters we get, $A_p = \{P_1, P_2, P_3, P_4, P_5\}$

The resulting equivalence classes are as follows:

$D_1 = \{O_1, O_2\}$;

$D_2 = \{O_3, O_7, O_{10}\}$;

$D_3 = \{O_4\}$;

$D_4 = \{O_5\}$;

$D_5 = \{O_6\}$; $D_6 = \{O_8\}$; $D_7 = \{O_9\}$

Phase I: Partitioning/Learning

Construct Decision files for all values of P_1 in VP_1

[$P_1 = 0 \rightarrow \{O_4, O_6, O_8\}$: store in SP10), ($P_1 = 1 \rightarrow \{O_1, O_2\}$: SP11],

[$P_1 = 2 \rightarrow \{O_3, O_5, O_7, O_9, O_{10}\}$: SP12]

Construct Decision files for all values of P_2 in VP_2

[$P_2 = 0 \rightarrow \{O_3, O_4, O_6, O_7, O_{10}\}$: SP20), ($P_2 = 1 \rightarrow \{O_5, O_8, O_9\}$: SP21],

[$P_2 = 2 \rightarrow \{O_1, O_2\}$: SP22]

Construct Decision file for all values of P_3 in VP_3

[P_3= 0 → {O_1, O_2, O_3, O_5, O_7, O_9,O_{10}}: SP30), (P_3= 1 → {O_4, O_6}: SP31],
[P_3= 2 → {O_8}: SP32]

Construct Decision file for all values of P_4 in VP_4

[P_4= 1 → {O_1, O_2, O_3, O_7, O_{10}}: SP41),(P_4= 2 → {O_4, O_5, O_6, O_8, O_9} SP42]

Construct Decision file for all values of P_5 in VP_5

[P_5= 0 → {O_3, O_7, O_{10}}: SP50), (P_5= 1 → {O_1, O_2 O_4, O_5 O_8}: SP51],
[P_5= 2 → {O_6, O_9 }: SP52]

Phase II: Classification Test Cases

Case 1: consider a test sample O_t= {0, 1, 2, 1 ,0}

Get the solution S_1 from SP10 = {O_4, O_6, O_8}; S_2 from SP21 = {O_5, O_8, O_9};
S_3 from SP32 = {O_8}; S_4 from SP41 = {O_1, O_2, O_3, O_7, O_{10}} and
S_5 from SP50 = {O_3, O_7, O_{10}}

Occurrence Frequency Vector Case 1:

O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
1	1	1	1	1	1	2	3	1	2

$\cap_{i=1,5} S_i = \phi$, O_t= {0,1,2,1,0} is closest to O_8 → D_6; P_4 and P_5 have not participated in the decision making.

Case 2: consider a test sample O_t= {2, 0, 0, 1, 0}

Get the solutions S_1 = {O_3, O_5, O_7, O_9, O_{10}}; S_2 = {O_3, O_4, O_6, O_7, O_{10}}
S_3= {O_1, O_2, O_3, O_5, O_7, O_9,O_{10}}; S_4= {O_1, O_2, O_3, O_7, O_{10}} and
S_5= {O_3, O_7, O_{10}}.

Occurrence Frequency Vector Case 2:

O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
2	2	5	1	2	1	5	1	3	5

$\cap_{i=1,5} S_i$ = {O_3, O_7, O_{10}}, O_t belongs to class {O_3, O_7, O_{10}} → D_2

Case 3: consider a test sample O_t= {1, 2, 0, 1, 1}

Get the solution S_1 = {O_1, O_2}; S_2= {O_1, O_2}; S_3= {O_1, O_2, O_3, O_5, O_7, O_9, O_{10}};
S_4= {O_1, O_2, O_3, O_7, O_{10}} and S_5= {O_1, O_2 O_4, O_5 O_8}.

Occurrence Frequency Vector Case 3:

O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
5	5	2	1	2	0	2	1	1	2

$\cap_{i=1,5} S_i$ = {O_1, O_2}, O_t belongs to class {O_1, O_2,} → D_1

Case 4: One Parameter missing from Reduct can be tolerated

Consider a test sample O_t= {1, 2, 0, _, 1}

Get the solution S_1 = {O_1, O_2}; S_2= {O_1, O_2}; S_3 ={O_1, O_2, O_3, O_5, O_7, O_9, O_{10}};

Get the solution S_4 from SP4? = Data unknown/missing

Get the solution S_5 from SP51 = $\{O_1, O_2\ O_4, O_5\ O_8\}$

Occurrence Frequency Vector Case 5:

O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
4	4	1	1	2	0	1	1	1	1

$\cap_{i=1\text{-}3,5}\ S_i = \{O_1, O_2\}$, O_t belongs to class $\{O_1, O_{2,}\}$ → D_1

Case 5: One Parameter missing from Core leads to confusion

Consider a test sample $O_t = \{2, 1, 0, 2, \_\}$

Get the solution $S_1 = \{O_3, O_5, O_{7,}\ O_9, O_{10}\}$; $S_2=\{O_5, O_8, O_9\}$;

$S_3 = \{O_1, O_2, O_{3,}\ O_5, O_7, O_9, O_{10}\}$; $S_4=\{O_4, O_5, O_6, O_8, O_9\}$

Get the solution S_5 from SP5? = Data unknown/missing

Occurrence Frequency Vector:

O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
1	1	2	1	4	1	2	2	4	2

$\cap_{i=1,4}\ S_i = \{O_5, O_9\}$, O_t belongs to class $\{O_5, O_{9,}\}$ → D_4 , D_7 reflects confusion and needs more training data to resolve the classes.

Case 6: Parameter not in V_a

Consider a test sample $O_t = \{2, 1, 0, 0, 1\}$

Get the solution S_1 from SP12= $\{O_3, O_5, O_{7,}\ O_9, O_{10}\}$

Get the solution S_2 from SP21 = $\{O_5, O_8, O_9\}$

Get the solution S_3 from SP30 = $\{O_1, O_2, O_{3,}\ O_5, O_7, O_9, O_{10}\}$

Get the solution S_4 from SP40 = Decision File does not exist implies a new class for which following action to be taken by updating the solution sets.

(a) File FP40 to be created = $\{O_{11}\}$;

(b) Other files to be updated to include O_{11}

Update new solution sets will be as follows:

Updated SP12= $\{O_3, O_5, O_{7,}\ O_9, O_{10}\ O_{11}\}$; SP21 = $\{O_5, O_8, O_9\ O_{11}\}$;

SP30 = $\{O_1, O_2, O_{3,}\ O_5, O_7, O_9, O_{10}\ O_{11}\}$; SP40 = $\{O_{11}\}$ and

SP51 = $\{O_1, O_2\ O_4, O_5\ O_8\ O_{11}\}$

Occurrence Frequency Vector:

O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}
2	2	2	1	4	0	2	2	3	2	5

$\cap_{i=1,5}\ S_i = \{O_{11}\}$, O_t belongs to class $\{O_{11}\}$ → D_8 reflects addition of new class and information system also gets updated.

7. EXPERIMENTAL SETUPS

In order to evaluate performance of the I-SEA algorithm, we used three data sets from the “UCI” repository; namely LETTER, WINE, and MONKS. The experiment helped in establishing the comparative accuracy as compared with existing algorithms.

7.1 Data Set “LETTER”

The “LETTER Image Recognition” data set as available on the UCI site was used (Slate 1991). Data set uses 17 parameters, out of which the first parameter represents the class. It is a 26 class problem. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15. The objective is to identify the alphabets based on the 16 parameters given in the database. Authors of the data set have reported best accuracy of 80%.

7.1.1 *Learning Phase.* The data file was split into two parts. The first part of 16000 records was used for learning purposes. The remaining 4000 records were used for testing the classifier. Based on the 16000 records, the equivalent classes were created, which resulted in 256 files (16 parameter × 16 values of parameters). These files contain the record number of the record fulfilling the criterion of the parameter value. The 256 files thus created were indexed on the parameter and its value. A class vector file was also created consisting of 16000 entries, one for the class of each record.

7.1.2 *Classification Phase. Definitions related to this section.*

Definition 12: Parameter Value, pv is value of the parameter $p_i= t_i$

Definition 13: Confusion Zone Size: Number of classes in the final solution set has been treated as cardinality of confusion zone.

Definition 14: Single Solution: Percentage of cases having only one class suggested by the objects in the final solution set.

Definition 15: Multiple Resolved: Percentage of cases where multiple classes have been suggested, however the class could be resolved on the basis of class occurring maximum number of times.

Definition 16: Misclassification: Percentage of cases where the Test Case has been misclassified.

Definition 17: Confused: Percentage, where multiple classes could not be resolved.

Definition 18: Success: percentage sum of single solution and multiple resolved represented as success.

Definition 19: Failed: Percentage of cases where solution could not be achieved.

During the classification phase, 4000 records in the second part of the data set were used as test records. Classification was tried by varying various parameters to study the performance of the algorithm. *Delta0* was used for accepting only those objects in the partial solution set, which exactly matched the parameter. *Delta1* was used to include the objects lying within a window of (pv-1, pv, pv+1). Similarly, *Delta2* was used to include the objects lying within a window of (pv-2, pv, pv+2). Boundary conditions were handled by taking the last possible window size for a particular case.

TABLE 2

Results of accuracy for different window sizes

Window Size	Single Solution %	Multiple Resolved %	Failed %	Mis-Classified %	Success %	Avg. Confusion Zone
Delta0	33.4	49.8	6.27	0.27	83.2	3.01
Delta1	65.9	22.4	7.8	0.05	88.3	1.34
Delta2	17.45	59.06	0.05	0.4	76.51	6.4

TABLE 3

Effect of change in number of parameters on Accuracy %

Number of parameters	10	12	13	14	15	16
Percent accuracy	54.17	68.77	77.37	85.42	87.06	88.3

Delta0 transformed in the Occam Razor approach and resulted in a success rate of 83% with a high confusion zone (Table 2). *Delta1* was able to achieve an 88.3% success rate with a reduced confusion zone. *Delta2* was able to achieve merely a

76.51% success rate with a confusion zone higher than even *Delta0*. As can be seen in Figure 1 and Figure 2, the Single Solution and the Success rate are maximum for *Delta 1*. Figure 3 and Table 3 show reduced accuracy with a reduction in the number of parameters. A noteworthy feature is the extremely low percentage of *Misclassification* (Table 2).

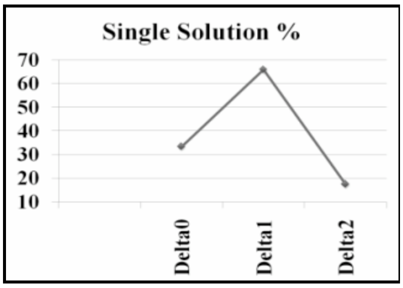


Fig. 1: Window size vs % of single solution

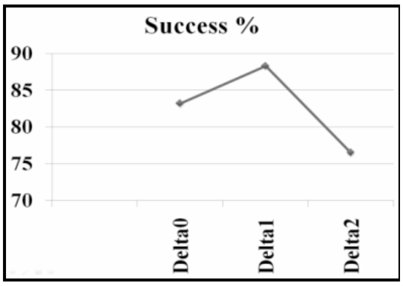
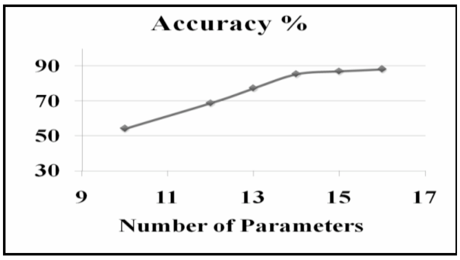


Fig. 2: Window size vs success



7.2 Data Set "WINE"

The WINE recognition data available on the UCI site was used (Aeberhard 1991). The WINE data set categorizes wines on the basis of 13 constituents found in three types of wines. All 13 attributes are continuous. In my experimental setup, I considered a reduced set of training data of 150 records. During testing, all 178 records were used for testing. Only Delta0 was used and all 13 parameters considered. Overall success rate achieved by I-SEA is 98.3%.

7.3 Date Set "MONKS"

MONK's problems are a collection of three binary classification problems (MONK1, MONK2 and MONK3), over six-attribute discrete domain (Thrun 1992). When the I-SEA algorithm was applied to MONK1 with 6 given parameters, the success rate was 81.94%. We experimented with the introduction of an additional parameter, derived by taking the ratio of attribute_1 and attribute_2. The nature of the problem is such that the resultant parameter turned out to be perfect discriminator, resulting in an accuracy of 100%. In the case of MONK2 and MONK3, the success rate achieved was 65.7% and 93.51%, respectively, which is comparable to the other algorithms.

7.4 Result Comparisons

Rokach and Maimon (2001) compared their algorithm D-IFN with respect to several data sets, for various algorithms. For comparison's sake, we selected the data presented by the authors for the data sets being used in this paper. Table 4 shows that the I-SEA Algorithm performed far better than the optimum classifier Naïve Bayes in all sets mentioned here and has produced better accuracy as compared with all other algorithms mentioned in the table.

8. CONCLUSIONS AND FUTURE WORK

Any classifier can be judged by its computational complexity and accuracy. In this paper we have developed the I-SEA Algorithm. During the Learning phase, the worst case complexity is O(mn2). Its average complexity during classification is O(m(log2n)2).

TABLE 4

Percent Accuracy Comparison of SEA with other Algorithms

Data Set	Naïve Bayes	C4.5	IFN	D-IFN	SEA
LETTER	73.29	74.96	69.56	79.07	Delta1: 88.3
WINE	96.63	85.96	91.45	96.63	98.3
MONK1	73.39	75.81	75.00	92.74	81.94 (100%)[1]
MONK2	56.21	52.07	63.87	62.13	65.7
MONK3	93.44	93.44	92.38	92.62	93.51

[1]Sec. parameter used

The cost of Incremental learning is O(m). Therefore, Incremental learning is better than many existing algorithms on performance issues. Experimentally the classification accuracy is comparable to most algorithms. The misclassification percentage is extremely low.

The algorithm is still in its infancy and therefore throws a broad scope for future work. The algorithm is yet to be tried on large data sets. The algorithm is flexible and can be extended to incorporate incremental learning, which will give immense power to the algorithm. We wish to apply the algorithm in field of “Behavioral Monitoring of Systems”, in general and in “Computer Networks” in particular.

REFERENCES

Anazida Z., Mohd. Aizaina Maarof and Siti Mariyam Shamsuddin. 2006. Feature selection Using Rough Set in Intrusion Detection, *Proceedings of IEEE TENCON*, Hongkong, 1-4.

Anazida Z., Mohd. Aizaina Maarof and Siti Mariyam Shamsuddin. 2007. Feature Selection Using rough-DPSO in Anomaly Intrusion Detection. *LNCS*, **4705**, Springer Berlin / Heidelberg, 512-524.

Bian H.Y. 2002. Fuzzy-Rough Nearest Neighbor Classification: an Integrated Framework, *Proceedings of IASTED, International Symposium on Artificial intelligence and Applications,* 160-164.

Cai Z. M., Guan X. H. et al. 2003. A new approach to intrusion detection based on rough set theory, *Chinese Journal of Computers*, **26(3)**, 361-366.

Cover T.M., Hart P.E. 1967. Nearest neighbor pattern classification, *IEEE Transactions on Information Theory, IT*, **13**, 21-27.

Domingos P., Pazzani M. J. 1997. On the optimality of the simple Bayesian classifier under zero one loss, *Machine Learning*, **29(23)**, 103-130.

Dubois A. D., Prade H. 1990: Rough Fuzzy Sets and Fuzzy Rough Sets, *International Journal of General Systems*, **17**, 191-209.

Jing-Kai Liang, Yang Zhang, Yan-Bin Qu. 2005. A Heuristic Algorithm Of Attribute Reduction In Rough Set, *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, 3140-3142.

Lee H., Chung Y., and Park D. 2006. An Adaptive Intrusion Detection Algorithm Based on Clustering and Kernel- Method, *Proceedings of Advances in Knowledge Discovery and Data Mining*, 10th Pacific-Asia Conference, PAKDD, Singapore, 603-610.

Lee W. and Stolfo S.J.. 1999. A data mining framework for building intrusion detection model, *Proceedings of the IEEE Symposium on Security and Privacy*, Oakland, CA, IEEE Computer Society Press, 120-132.

Lin T.Y., Chen R. 1997. Finding Reducts in Very Large Databases. *Proceedings of Joint Conference, Information Science Research*, 350-352.

Maimon O., Rokach Lior. 2002. Improving Supervised Learning by Feature Decomposition, *Proceedings of the Second International Symposium on Foundations of Information and Knowledge Systems, LNCS*, Springer, 178-196.

Mansour Yishay, David McAllester A. 2000. Generalization Bounds for Decision Trees, *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc.**,** 69-74.

Miao D.Q., Wang J. 1998. Analysis on Attribute Reduction Strategies of Rough Set, *Chinese Journal of Computer Science and Technology*, **13(2)**, 189-192.

Mukkamala S., Sung A. H., and Abraham A. 2005. Intrusion detection using an ensemble of intelligent paradigms, *Journal of Network and Computer Applications*, **28**, 167-182.

Pawlak Z. 1982. Rough Sets, *Int'l J. Computer and Information Sciences*, **11**, 341-356.

Peng Hong, Dongna Zhang, Tiefeng Wu. 2004. An Intrusion Detection Method Based on Rough Set and SVM Algorithm, *Communications, Circuits and Systems, ICCCAS,* **2**, IEEE,1127-1130.

Rao X., Dong C.X. and Yang S.Q. 2003. An Intrusion detection system based on Support Vector Machine, *Journal of Software*, **14(4)**, 798-803.

Rokach Lior, Maimon Oded. 2001. Theory and Application of Attribute Decomposition, *Proceedings of the First IEEE International Conference on Data Mining*, IEEE Computer Society Press, 473-480.

Srinivasa K. G., Jagadish M., Venugopal K. R., Patnaik L. M. 2007. Data Mining based Query Processing using Rough Sets and Genetic Algorithms, *IEEE Symposium on Computational Intelligence and Data Mining*, 275-282.

Sung Andrew H., Mukkamala Srinivas. 2003. Identifying Important Features for Intrusion Detection Using Support Vector Machines and Neural Networks, *Proceedings of the Symposium on Applications and the Internet (SAINT'03), Elec. Ed.,* 209-217.

Tamilarsan A., S. Mukkamala, Sung A. H., Yendrapalli K. 2006. Feature Ranking and Selection for Intrusion Detection Using Artificial Neural Networks and Statistical Methods. *International Joint Conference on Neural Networks,* Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, 4754-4761.

Wang L., Yu G., Wang G., and Wang D. 2002. Method of Evolutionary Neural Network-based Intrusion Detection, *Journal North Eastern University Natural Science*, **23**, 107-110.

Wang R., Duoqian Miao, Guirong Hu. 2006. Discernibility Matrix Based Algorithm for Reduction of Attributes, *Proceedings of the International Conference on Web Intelligence and Intelligent Agent (WI-IATW'06)*, IEEE/WIC/ACM*,* 477-480.

Wang Y., Yang H. H., Wang X.Y. 2005. An Intrusion Detection Experimental System Using Evolutionary Neural Network, *Journal of East China University of Science and Technology (Natural Science Edition)*, **31(3)**, 362-366.

Wenkee Lee, Salvatore Stolfo J. 1999. A Framework for Constructing Features and Models for Intrusion Detection Systems, *Proceedings of the IEEE Symposium on Security and Privacy* , Oakland, CA, IEEE Computer Society Press, 227-261. (Wenkee and Salvatore, 1999)

Witcha Chimphlee, Abdul Hanan Abdullah, Mohd Noor Md Sap, Surat Srinoy, and Siriporn Chimphlee. 2006. Anomaly-Based Intrusion Detection using Fuzzy Rough Clustering, *Proceedings of International Conference on Hybrid Information Technology*, **01**, IEEE Computer Society, USA, 329-334.

Wong S.K.M., Ziarko W. 1985. On Optional Decision Rules in Decision Tables. *Bulletin of Polish Academy of Science,* **33**, 693-696.

Yates Ricardo Baeza. 2004. A Fast Set Intersection Algorithm for Sorted Sequences, *LNCS*, **3109**, Springer Berlin/ Heidelberg, 400-408.

Ye D.Y., Chen, Z.J. 2002."A New Discernibility Matrix and the Computation of a Core", *Chinese Journal of Electronics*, **30(7)**, 1086-1088.

Books:

Duda Richard O., Hart Peter E., Stock David G. 2001. *Pattern Classification,* 2nd ed, Wiley Student Editions, 394-412.

Lee W. 1999. A data mining framework for constructing features and models for intrusion detection systems, *PhD Dissertation*, Columbia University.

Liu, Motoda. 1998. *Feature Selection for Knowledge Discovery and Data Mining*, Norwell, MA, USA, Kluwer Academic Publishers, 17-70.

Maimon O., Last M. 2000. Knowledge Discovery and Data Mining: The Info Fuzzy network (IFN) methodology, Kluwer Academic Publishers.

Mitchell Tom M. 1997. *Machine Learning*, Carnegie Mellon University, McGraw-Hill Intl. Eds., 52-76.

Pawlak, Z. 1991. *Rough Sets - Theoretical Aspects of Reasoning about Data.* Dordrecht, Kluwer Academic Publishers.

Edited collections:

Aeberhard Stefan. 1991. Wine Identification Problem. Dept. of Computer Science, North Queensland, *http://archive.ics.uci.edu/ml/datasets/Wine/* , UCI dataset.

Ghosh A.K., Michael C. and Schatz M.. 2000. A real-time intrusion system based on learning program behavior, *Recent Advances in Intrusion Detection, (RAID 2000),* 1907/2000, edited by Debar H, Ludovic M., Toulouse, Spinger-Verlag, 93-109.

Maimon O. and Rokach L. 2001. Data Mining by Attribute Decomposition with semiconductors manufacturing case study, *Data Mining for Design and Manufacturing: Methods and Applications*, edited by D. Bracha, Kluwer Academic Publishers, 311-336.

Min Fan, Qihe, Hao Tan, Leiting Chen, Wang G. et al. (Eds.). 2006. The M-Relative Reduct Problem, *Rough Sets and Knowledge Technology, LNAI*, 4062, Springer-Verlag Berlin Heidelberg, 170-175.

Slate David J. 1991. Letter Image Recognition Data: Odesta Corporation; Evanston, IL 60201, *http://archive.ics.uci.edu/ml/datasets/Letter+Recognition/* , UCI dataset.

Thrun Sebastian. 1992. Monks Problem. School of Computer Science, CMU, Pittsburgh, *http://archive.ics.uci.edu/ml/datasets/MONK%27s+Problems/* , UCI dataset.

Wikipedia 2008. Rough Set, *http://en.wikipedia.org/wiki/Rough_set*.

Ziarko W.P. 1991. The Discovery, Analysis, and Representation of Data Dependencies in Databases, *Knowledge Discovery in Databases*, edited by Piatesky-Shapiro, G. and Frawley, W. J., AAAI Press/MIT Press, 177-195.