Proceedings of the Federated Conference on Computer Science and Information Systems pp. 41–46
DOI: 10.15439/2016F453
ACSIS, Vol. 8. ISSN 2300-5963

A new way for the exploration of a dataset based on a social choice inspired approach

Michel HERBIN, Amine AÏT YOUNES, Frédéric BLANCHARD
Université de Reims Champagne-Ardenne
CReSTIC, France
Email: {michel.herbin, amine.ait-younes, frederic.blanchard }@univ-reims.fr

***Abstract*—The exploration of a data set consists in grouping similar data. The classical statistical methods often fail when there is is no minimal assumption on the clusters. Our approach is based on the links between data, but the pairwise comparison between data and the importance of the links depend heavily on context where data lies. We propose to analyze a dataset through methods of the social choice theory where data plays both the role of a candidate and the role of a voter. The candidates are ranked by the voters and each voter gives a score to each candidate according to his ranking. We propose one specific election for each voter based on his preferences. The voters of these elections have weights computed according to their respective behaviors. In this approach, the conventional similarity indices between data are used to define the electoral behavior of each data.**

***Index Terms*—exploratory data analysis, social choice theory, representatives, vote, data reduction**

I. Introduction

ONE OF the first steps in the exploration of a data set consists in grouping similar data. For this purpose, lot of clustering methods are proposed in literature to detect clusters within a dataset. The methods often fail when there is neither a minimal assumption on the clusters nor a minimal model of the clusters. For instance the classical k-means method [8] assumes both that data could be grouped around mean values or mean vectors and that the number of clusters is known. Unfortunately the first assumption leads to important constraints on the shape of the clusters in the data space and this condition is seldom corroborated. Other approaches of clustering are based on links between data. The hierarchical agglomerative clustering methods are probably the most known methods for exploring the datasets using such links. The links are usually drawn from pairwise comparisons between data and they are based on distances or pseudo-distances [4]. But the pairwise comparison between data and the importance of the links depend heavily on context where data lies. Indeed the ranges of values of a comparison index could change when data are not in the same clusters. In other words, the links could be well suited to connect two data in one cluster and they are not adapted for the other clusters. Thus this paper proposes a new way to define the links between data through the ranks to overcome this constraint of cluster context.

We propose to analyze a dataset through methods of the social choice theory where data plays both the role of a candidate and the role of a voter [5]. The social choice inspired approach brings a metaphorical meaning that help to understand the concepts (as in bioinspired or human-inspired algorithms [10]).

The candidates are ranked by the voters and each voter gives a score to each candidate according to his ranking. Then the scores of the voters are aggregated using generally the sum of scores obtained by the candidates. In the classical procedure of election, each voter has the same weight in the aggregation. Thus this procedure is the same for all clusters. In this paper the election procedures differ from one cluster to another. We propose one specific election for each voter based on his preferences (i.e. one election per voter). The voters of these elections have weights computed by comparison of their respective behaviors. The weights differ from one election to another. The links between data are defined using these elections where each voter selects one candidate for representing itself within the dataset. The chainings between the voters and their representatives define data communities. Thus the partitions of the dataset with these communities give us a new way to explore the dataset.

The following section describes the procedure of election that we propose in this paper. It leads to a graph that permit us to structure the dataset. Then we study and we assess this method for structuring a dataset. Finally we discuss and we conclude this work.

II. Dataset and Voters

A. Collective preference

Let Ω be a dataset with n elements:

$$\Omega = \{X_1, X_2, ...X_n\}$$

In the framework of the social choice theory [9] [3], Ω is both a set of n voters and a set of n candidates. Thus each data is a voter of Ω and it also becomes an alternative that the other voters could prefer as a representative in Ω (i.e. an elected candidate of Ω).

The dataset is provided with a pairwise comparison index between data. We call D this index. In this paper, we use Euclidean distance as pairwise comparison index. But we need only two properties of D. When X_i, X_j, and X_k are three data in Ω, we should have:

- $D(X_i, X_i) \leq D(X_i, X_j)$,
- $D(X_i, X_j) \leq D(X_i, X_k)$ if X_j is more similar to X_i than X_k is (in other terms : X_j is prefered to X_k by X_i)

978-83-60810-90-3/$25.00

In the following, any pairwise comparison index should respect these two properties.

With using such a pairwise comparator, each data X_i is considered as a voter which can rank the other data. The ranks of X_i are defined between 1 and n. The ranking function is called R_{X_i} and we have:

- $R_{X_i}(X_i) = 1$,
- $R_{X_i}(X_j) \leq R_{X_i}(X_k)$ if $D(X_i, X_j) \leq D(X_i, X_k)$.

The data X_i is a voter that selects the candidates using R_{X_i} as preference indicator. The vote of X_i is realized with a score of Borda which is a classical method of social choice theory [6] [1]. In this paper, the score of Borda given by the voter X_i to the candidate X_j is defined as:

$$S_{X_i}(X_j) = \frac{n - R_{X_i}(X_j)}{(n-1)}$$

where X_j is a candidate and X_i is a voter.

The classical election procedure attributes the sum of the scores of the voters for each candidate. Thus the candidate X_j obtains the global score $S(X_j)$ defined by:

$$S(X_j) = \sum_{i=1}^{n} S_{X_i}(X_j)$$

This procedure leads to nominate the best candidate as the one with the highest score. But each voter has the same weight in this overall vote. This overall election does not take into account that two voters could belong to two different clusters.

In the following, we will change the paradigm. We consider that each voter has his own election procedure that is adapted to itself. The following describes the specific procedure for each voter.

B. Individual preference

Each voter will choose its candidate with its own election procedure. Let X_i be a voter that chooses the candidates. Each data X_j is also a voter of the election that X_i proposes. All the voters of Ω have weights that are specific of the election procedure of X_i. The weight of X_i itself is equal to one. The more similar to X_i a voter X_j is, the higher the weight of X_j is in this election. The weights are based on the similarity between the voters and the similarities with X_i are used for the election that X_i proposes.

Let us describe the similarity of the behaviors of two voters. We consider that two voters X_i and X_j are similar when their respective ranking function R_{X_i} and R_{X_j} are similar. The correlation of Spearman [11] is classically used to evaluate the correlation between ranks. The higher the correlation is close to 1, the more ranks are correlated. In this paper the correlation gives us an index of the similarity of the behavior of two voters. Spearman correlation between X_i and X_j is defined by:

$$Cor(X_i, X_j) = 1 - \frac{6 * \sum_{k=1}^{n} (R_{X_j}(X_k) - R_{X_i}(X_k))^2}{n^3 - n}$$

$Cor(X_i, X_j)$ lies between -1 and 1. We consider that X_i and X_j have similar behavior when the Spearman correlation is greater then a positive threshold which is a significance level. If we call t this level, then X_i and X_j become similar when $Cor(X_i, X_j) \geq t$.

Let $w_{X_i}(X_j)$ be the weight given to the voter X_j for the election based on the preferences of X_i.

We define this weight by:

$$w_{X_i}(X_j) = max(0, \frac{Cor(X_i, X_j) - t}{1 - t})$$

The weight lies between 0 and 1. It is equal to zero when X_i and X_j are not similar.

In the election based on the preferences of X_i, each candidate X_j obtains a score $Score_{X_i}(X_j)$ defined by:

$$Score_{X_i}(X_j) = \sum_{k=1}^{n} w_{X_i}(X_k) \times S_{X_k}(X_j)$$

Thus this election is based on a sum of scores weighted by the similarity of the voters with X_i. Other voters similar to X_i participate in the election of the representative of X_i.

C. Communities of voters

The representative of X_i becomes the one which have the highest score within Ω for the election based on the preferences of X_i. So each voter X_i has one representative in Ω elected by the specific election of X_i : $Rep_{Score}(X_i)$.

$$Score(Rep_{Score}(X_i)) = \max_{k=1}^{n}(Score_{X_i}(X_k))$$

We define a graph in Ω where the vertices are the voters and the edges are the links between the voters and their representatives. Each connected components of this graph defines a community of voters. The more we claim a high correlation between voters, the more the size of communities is reduced. In other words, the higher the threshold t is close to 1, the more the communities are small and the number of communities increases within Ω. These communities give a data structuration to study a dataset when we have neither assumption nor model for the clusters.

If the threshold t is close to 1, the representative of each voter X_i is based only on the preference of X_i. If this threshold decrease, other voters similar to X_i participate in the election of the representative of X_i.

Each data Xi has two representatives: the favorite candidate of Xi and the elected candidate of the local election of Xi. For each data X_i we define an individual loss indicator, wich represents the correlation loss between X_i and theses two representatives. The collective $loss$ indicator for the data is the sum of all the individual loss.

$$loss = \sum_{k=1}^{n} loss_{Ind}(X_k)$$

$$loss_{Ind}(X_i) = Cor(X_i, Rep_S(X_i)) - Cor(X_i, Rep_{Score}(X_i))$$

with

$$S(Rep_S(X_i)) = \max_{k \neq i}(S_{X_i}(X_k))$$

$$Score(Rep_{Score}(X_i)) = \max_{k \neq i}(Score_{X_i}(X_k))$$

III. Experimental Study

This section is devoted to the study of our method for structuring a dataset with a pairwise comparator. First let us present an example of the different steps of the dataset structuration with our method. In a second section, we assess the quality of this structuration using simulated data. Third the quality is assessed when using one real dataset.

TABLE I: Number of communities of voters, number of unique representatives and loss, using the simple example of Fig.1 when the threshold t of correlation varies between 0 and 1.

	t	nbcom	nbrep	loss
1	0.00	2	4	0.85
2	0.05	2	4	0.85
3	0.10	2	4	0.85
4	0.15	2	4	0.85
5	0.20	2	4	0.85
6	0.25	2	4	0.85
7	0.30	2	4	0.85
8	0.35	2	4	0.77
9	0.40	2	4	0.77
10	0.45	2	4	0.77
11	0.50	2	4	0.78
12	0.55	2	4	0.78
13	0.60	2	5	0.59
14	0.65	2	5	0.59
15	0.70	2	5	0.52
16	0.75	3	8	0.32
17	0.80	3	9	0.23
18	0.85	3	11	0.17
19	0.90	5	14	0.00
20	0.95	5	14	0.00
21	1.00	20	20	0.00

A. Workflow for structuring a dataset

We propose to explore a dataset with 20 simulated data in dimension 2 (see Fig.1-A). The pairwise comparisons are based on Euclidean distance. We conduct the overall election with a classical Borda's procedure. This overall election permits us to propose the best candidate which could be considered as the representative of the whole dataset (see Fig.1-B). Then we proceed to the elections based on the individual preferences for obtaining linking each data with another one. These links allow to define communities of voters. The procedure of the election with the individual preferences is based on a threshold of correlation. Fig.1-C shows the number of communities when the correlation threshold increases.

The higher the threshold, the higher the number of communities is. The highest threshold leads to the highest number of unique representatives. The higher the threshold, the lesser the losses are (both individual and collective). When the threshold is equal to 0.5, 0.95 and 0.99 (Fig.1-D, Fig.1-E, Fig.1-F) :

- the number of communities is 2, 5 and 6 (resp.)
- the number of unique representatives is 4, 14 and 15 (resp.)
- the collective loss is 0.78, 0 and 0 (resp.)

This number of communities is less than the number of data. That gives a new way for the exploration of a dataset.

B. Assessment of the links structuring a dataset

TABLE II: Number of communities of voters, number of unique representatives and loss, using the three classes of Fig.2 and criterion of assessment when the threshold of correlation varies between 0 and 1.

	t	nbcom	nbrep	loss
1	0.00	3	20	6.99
2	0.05	3	19	6.65
3	0.10	3	19	6.37
4	0.15	3	18	6.49
5	0.20	3	16	6.23
6	0.25	3	17	6.09
7	0.30	3	17	6.00
8	0.35	3	16	6.01
9	0.40	3	16	6.03
10	0.45	3	16	5.95
11	0.50	3	16	5.82
12	0.55	3	17	5.70
13	0.60	3	19	5.47
14	0.65	3	21	5.00
15	0.70	3	24	4.58
16	0.75	3	31	4.04
17	0.80	3	35	3.03
18	0.85	4	47	2.32
19	0.90	7	58	1.37
20	0.95	10	74	0.52
21	1.00	150	150	0.00

In this paper, we place ourselves resolutely in the context of the exploratory analysis of data without any a priori assumption on eventual classes, we only use an index of pairwise comparison. But the use of classes gives the most classical way to evaluate a structuration of a dataset. So this paper uses classes to assess only the links that we propose between data. The detection of classes (i.e. the clustering) is out of the scope of this paper.

The assessment of our method for structuring a dataset is performed using a dataset with known classes. Each data belongs to one class and it has the label of its class. Using our structuration each data is also linked to a representative in the dataset. A data is well represented when its own label is equal to the label of its representative. In such case the link between a voter and its representative remains inside a class of the dataset. We propose a structuration of the dataset with graphs. The vertices of the graph are labeled and the edges are labeled when their extremities have the same label. We compute the number of the labeled edges.

The percentage of such edges could assess the quality of the structuration through a graph. Unfortunately the classes are unknown in the first step of data exploration. Thus we propose to use the loss indicator instead of this percentage of labeled links.

The higher is this quality criterion and the lower the number of communities is, then the better the structuration is.

Table II gives the values of this criterion when the threshold of correlation lies between 0 and 1. The dataset is simulated in dimension 2 (see Fig.2) and the number of detected

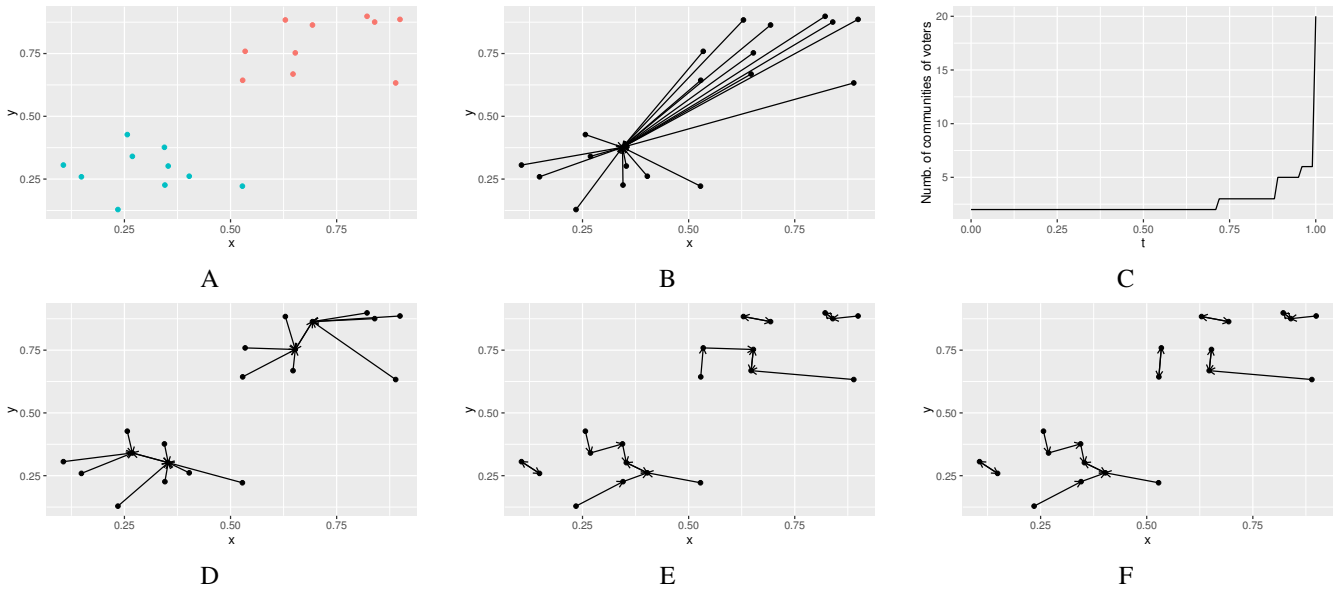

Fig. 1: Twenty simulated data in dimension 2 (A), the overal election selecting one representative (B) and elections based on individual preferences leading to several communities whose number depends on the correlation threshold (C). 2, 5, and 6 communities respectively obtained with a correlation threshold equal to 0.5, 0.95, 0.99 (D, E, F).

communities of voters is displayed when the threshold of correlation between voters increases from 0 to 1. Fig.2 gives also three examples of the communities when the threshold is respectively equal to 0, 0.5 and 0.9.

TABLE III: Number of communities of voters, number of unique representatives and loss, using the three classes of Fig.3 and criterion of assessment when the threshold of correlation varies between 0 and 1.

	t	nbcom	nbrep	loss
1	0.00	1	74	35.49
2	0.05	1	70	32.88
3	0.10	1	69	30.32
4	0.15	1	72	28.65
5	0.20	1	69	26.90
6	0.25	1	70	24.89
7	0.30	1	74	23.56
8	0.35	4	79	21.55
9	0.40	4	75	19.58
10	0.45	3	76	17.25
11	0.50	3	82	15.09
12	0.55	5	92	11.73
13	0.60	7	101	9.50
14	0.65	7	107	7.94
15	0.70	11	118	6.54
16	0.75	13	128	5.27
17	0.80	14	134	3.99
18	0.85	19	149	3.10
19	0.90	25	166	1.96
20	0.95	36	186	0.88
21	1.00	380	380	0.00

In the following we simulated a dataset with three classes that are hardly distinguishable because of their shapes and their overlapping. The dataset ($n = 380$) is simulated in dimension 2 with three uniform distributions in two rectangular crowns with 200 and 80 data and one rectangle with 100 data (see Fig.3). The number of voter communities is displayed when the threshold of correlation between voters increases from 0 to 1. Fig.3 gives also three examples of the communities when the threshold is respectively equal to 0.5, 0.8 and 0.99,

- the number of communities is 3, 14 and 71 (resp.)
- the number of unique representatives is 82, 134 and 212 (resp.)
- the collective loss is 15.09, 3.99 and 0.11 (resp.)

the number of communities are respectively equal to 8, 16 and 106. In such a case the classical clustering methods fail to detect meaning clusters. Indeed classical clustering methods are often based on statistics such as means or medoids. They use these statistics to determine the clusters and they make the assumption that data could be well represented with such statistics. Unfortunately these statistical approaches are unadapted in this case. Table III gives the values of our assessment criterion when the threshold of correlation lies between 0 and 1.

C. Assessment with real data

We use the databases from Machine Learning Repository of UCI [2] to assess our method with real data. Iris is the classical database that has 150 iris plants with 4 attributes and three clusters. Table IV gives the results we obtain with this dataset. Fig.4 displays the number of voter communities and the percentage of labeled links when the correlation threshold increases from 0 to 0.99, and the loss indicator.

IV. Discussion and Conclusion

In this paper we describe and we implement a method for exploring a data set. The main originality of this method lies in

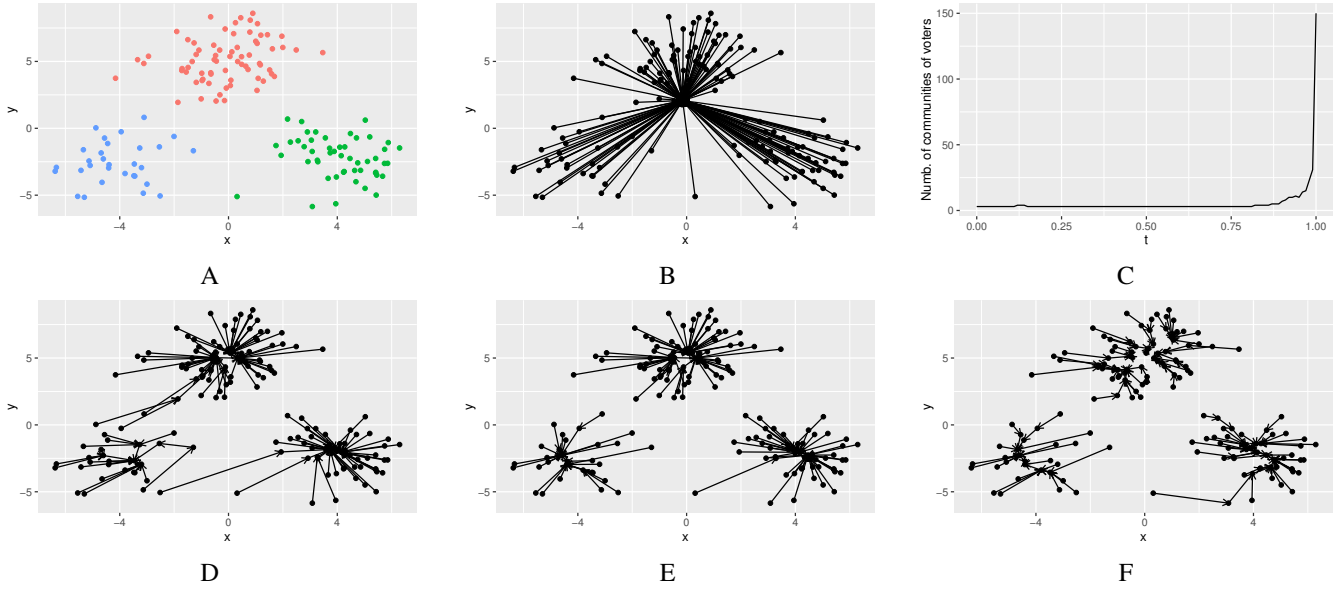

Fig. 2: Simulated data with three classes (three multinomial distributed subsamples)

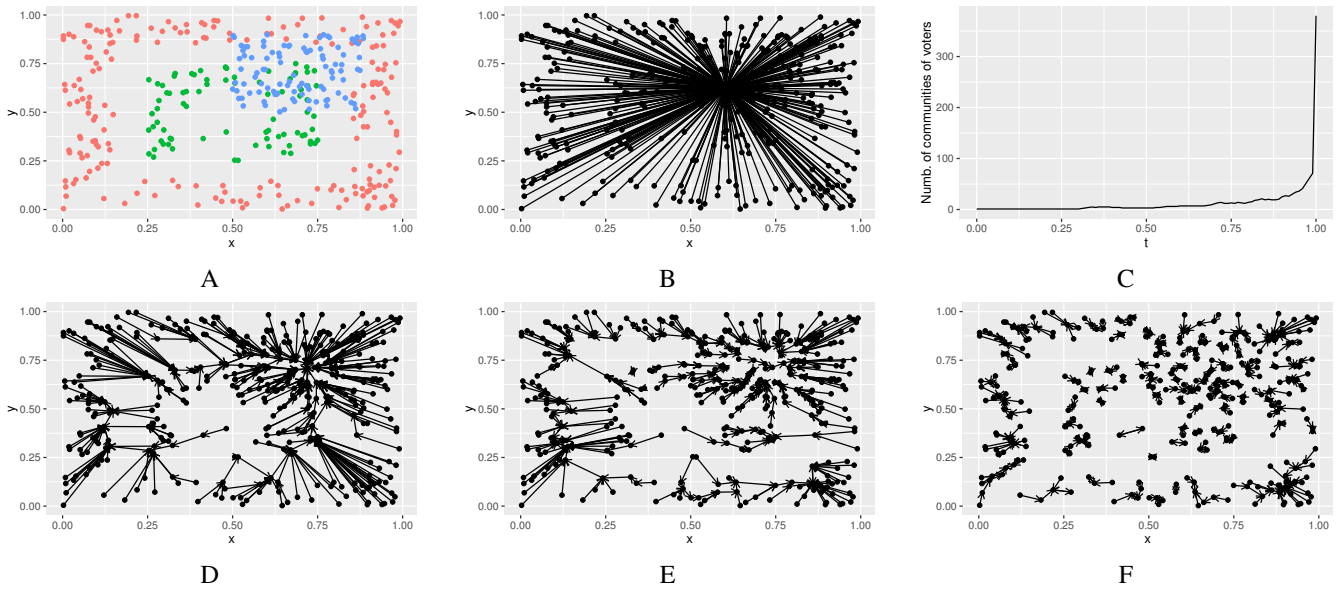

Fig. 3: Simulated data with three uniform distributions in two rectangular crowns of respectively 200 and 80 data and one rectangle of 100 data. The three classes are hardly distinguishable with classical clustering methods.

the definition of links between data. These links are based on a local election mechanism with individual preferences that connects each data to another data designated by the local election process. In this approach, the conventional similarity indices between data are used to define the electoral behavior of each data. As the preferences of the users in a recommender system, the voters then have weights corresponding to the similarity of electoral behaviors. However this approach by recommender systems is not used in this paper and the robustness of our method when data is incomplete or imperfect could be studied in future work.

Another important contribution of this work is to reduce the size of a data set from the exploration of a set of n data to a set of p communities where p is much smaller than n. This approach of dimensionality reduction has the advantage that it makes no assumption about the shape or the exact number of communities. It thus constitutes a preliminary step to a more meaningful clustering and it leads to select a more suitable method for the exploring dataset. This extension of our work could also be involved in further work.

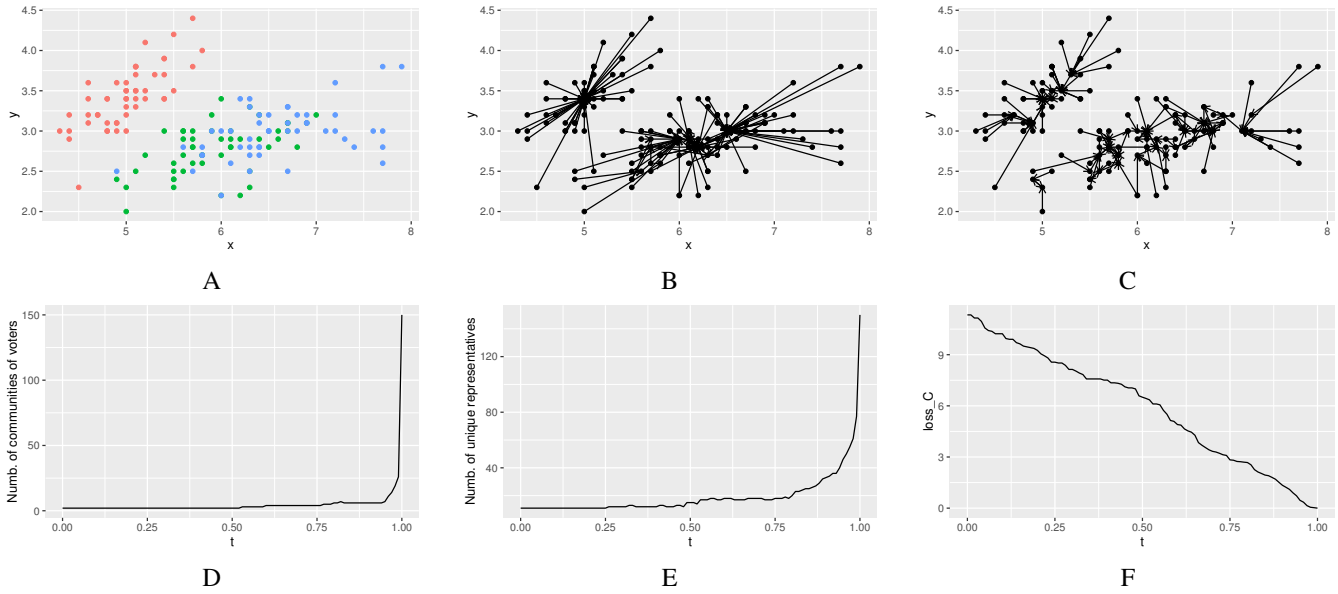

Fig. 4: Iris Data ($n = 150$) with three classes of 50 data in dimension four. Top : data projections in dimension two using sepal width and sepal length and detection of communities when the correlation threshold is equal to 0.5 and 0.95. Bottom : Number of voter communities and the percentage of labeled links when the correlation threshold increases from 0 to 0.99, and the loss indicator

TABLE IV: Number of communities of voters, number of unique representatives and loss, using the three classes of the Iris data (see Fig.4) and criterion of assessment when the threshold of correlation varies between 0 and 1.

	t	nbcom	nbrep	loss
1	0.00	2	11	11.33
2	0.05	2	11	10.57
3	0.10	2	11	10.23
4	0.15	2	11	9.64
5	0.20	2	11	9.25
6	0.25	2	11	8.56
7	0.30	2	12	8.14
8	0.35	2	12	7.58
9	0.40	2	12	7.51
10	0.45	2	12	7.11
11	0.50	2	15	6.51
12	0.55	3	17	6.06
13	0.60	4	18	4.90
14	0.65	4	17	4.27
15	0.70	4	18	3.33
16	0.75	4	18	2.83
17	0.80	6	20	2.67
18	0.85	6	25	1.94
19	0.90	6	33	1.34
20	0.95	7	46	0.46
21	1.00	150	150	0.00

We are currently working on application for sensor network data analysis.

ACKNOWLEDGMENT

This work is partially supported by the EC SCOOP project (INEA/CEF/TRAN/A2014/1042281).

REFERENCES

[1] A. Aït Younes, F. Blanchard and M. Herbin, "New similarity index based on the aggregation of membership functions through OWA operator", *Federated Conference on Computer Science and Information Systems*, FedCSIS 2015, 163–168, Lódz, Poland, 2015.

[2] K. Bache, M. Lichman, "UCI Machine learning repository", *http://archive.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Sciences*, 2013.

[3] J.P. Barthélémy and B.Montjardet, "The median procedure in cluster analysis and social choice theory", *Mathematical Social Sciences*, 1:235–267, 1981.

[4] A. Bellet, A. Habrard, M. Sebban, "A Survey on Metric Learning for Feature Vectors and Structured Data", *Technical report*, arXiv:1306.6709, 2014.

[5] F. Blanchard, C. de Runz, M. Herbin, H. Akdag, "Représentativité et graphe de représentants : une approche inspirée de la théorie du choix social pour la fouille de données relationnelles", *Atelier Fouille de Données Complexes, Conférence Extraction et Gestion des Connaissances*, EGC, 73-83, Brest, France, 2011.

[6] M. de Borda, "Memoire sur les elections au scrutin", *Academie Royale des Sciences*, Paris, 1784.

[7] A.K. Jain, M.N. Murty, P.J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, 31(3), 264–323, 1999.

[8] A.K. Jain, "Data clustering: 50 years beyond K-means", *Pattern Recognition Letters*, 31, 651–666, 2010.

[9] J.N. Mordeson, D.S. Malik, T.D. Clark, "Application of Fuzzy Logic to Social Choice Theory", Chapman and Hall/CRC, 2015.

[10] M. Parsapoor, U. Bilstrup, "An Emotional Learning-inspired Ensemble Classifier (ELiEC)", *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, 137-141, 2013

[11] C. Spearman, "General intelligence objectively determined and measured", *Am J Psychol*, 15, 201-293, 1904.