Algorithms Implemented for Cancer Gene Searching and Classifications

Matej Troják

- Rôznorodé typy rakovinových buniek
- Najlepšie hodnotné algoritmy
  - implementované pre Bio image
  - implementované pre DNA array

Rakovina

- Diagnostika – klinický výskyt
- Vplyv prostredia
- 1 gén z tisícok
- Zdravie jednotlivcov
- Vytváranie liekov

Požiadavky na algoritmus

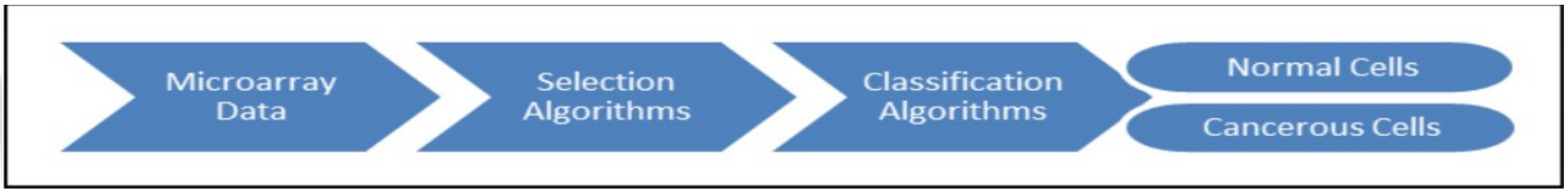
- Rýchlosť
- Presnosť
- Ľahká implementácia
- Testovateľnosť

Zaradenie algoritmov

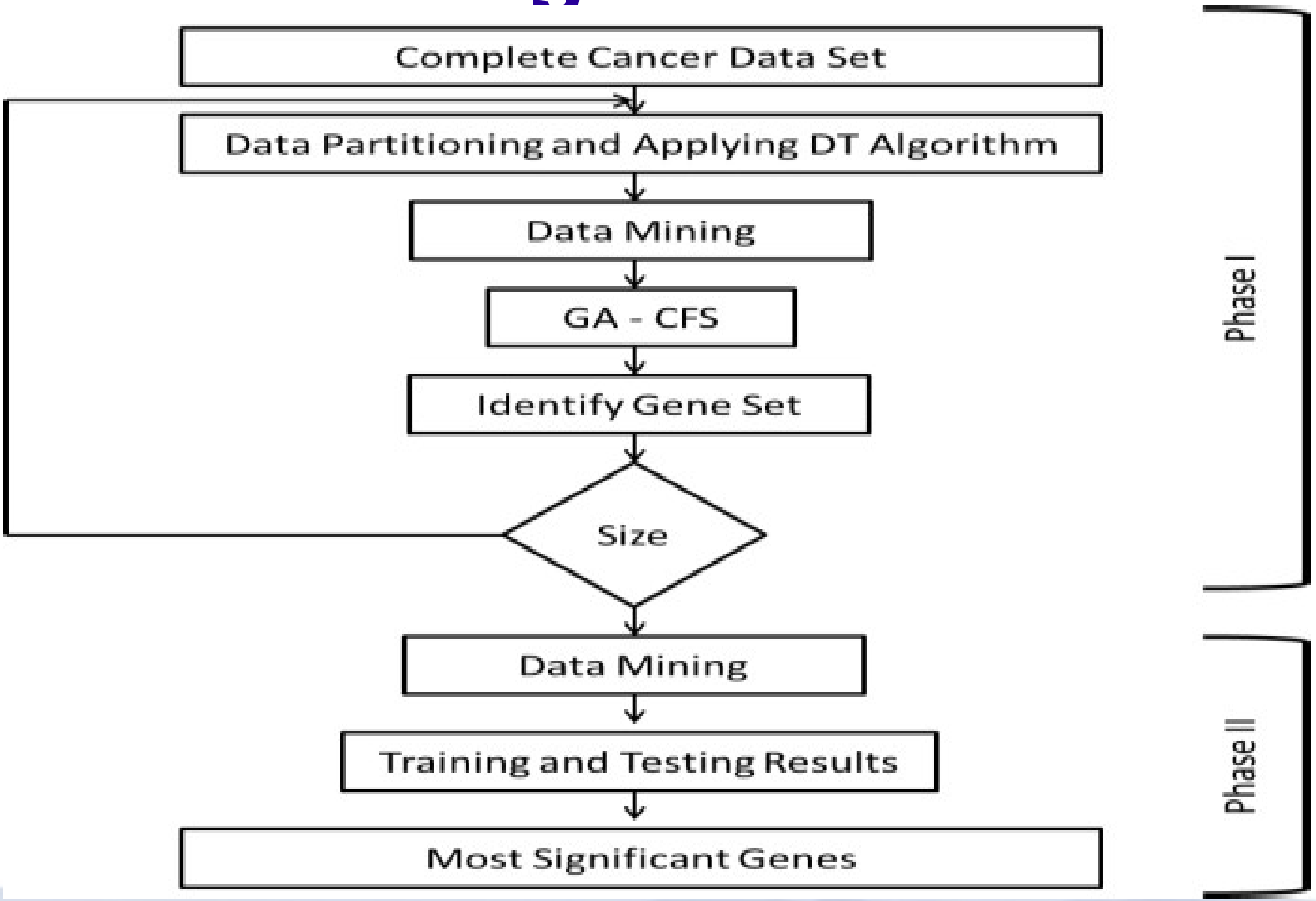
- Genetic Algorithms (GA)
  - Optimálna metóda
  - Správanie ako evolúcia
- Tabu Search (TS)
  - Heuristická metóda
  - Flexibilné využitie pamäte

Algoritmy pre analýzu

- Microarray
  - mnoho rozmerné dáta → nízka výkonnosť klasifikácie
  - Zložité priradenie odpovedajúcich génov
- Identifikácia “tichých” génov
- Zaradenie informačných génov



Integrated Gene-Search Algorithm



Hybrid Algorithm

- Genetic Algorithm + Particle Swarm Optimization + Support Vector Machine + Analysis of Variance
- Význam pri rakovine vaječníkov

Table 1. The Proposed Algorithm Accuracy of classification for various approaches

	The hybrid process of SVM and GA (%)	The hybrid process of SVM and PSO (%)	The proposed algorithm (%)
Colon	95.65%	97.13%	99.13
Breast	96.23%	97.95%	98.55

Bio Image

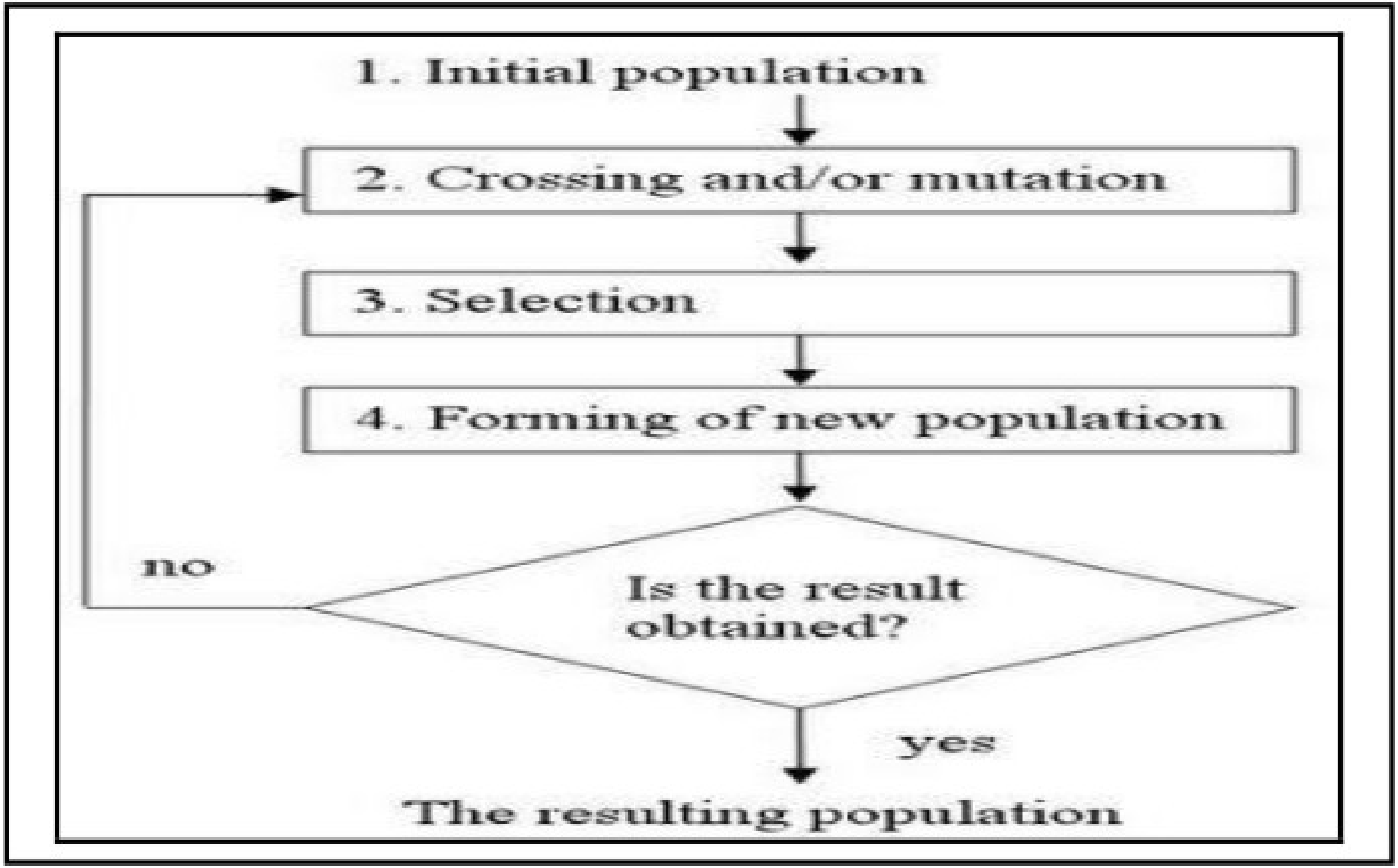
- CAIMAN system (Cancer Image Analyzis)
  - Cellular migration
  - Vasculature analysis
  - Shading correction
- www.caiman.org.uk

Table 2. Proposed Algorithm Performance Estimation

Algorithm	Dimension (pixels)	Size (kb)	Time ± (s)
Migration	285 x 203	1001	62.6 ± 9.6
	127 x 900	1700	81.4 ± 16.7
Tracing	220 x 164	108	66.2 ± 20.3
	768 x 576	1300	207.4 ± 14.6
Shading	285 x 203	100	59.5 ± 14.1
	1270 x 900	1700	65.0 ± 15.6

Genetic Algorithm

- Inicializace
- Začátek cyklu
- Nové jedince:
  - křížení, mutace, reprodukce
- Zdatnost jedinců
- Konec cyklu → 2
- Konec algoritmu


1. Initial population
2. Crossing and/or mutation
3. Selection
4. Forming of new population
Is the result obtained?
no
yes
The resulting population

Correlation-based feature selection

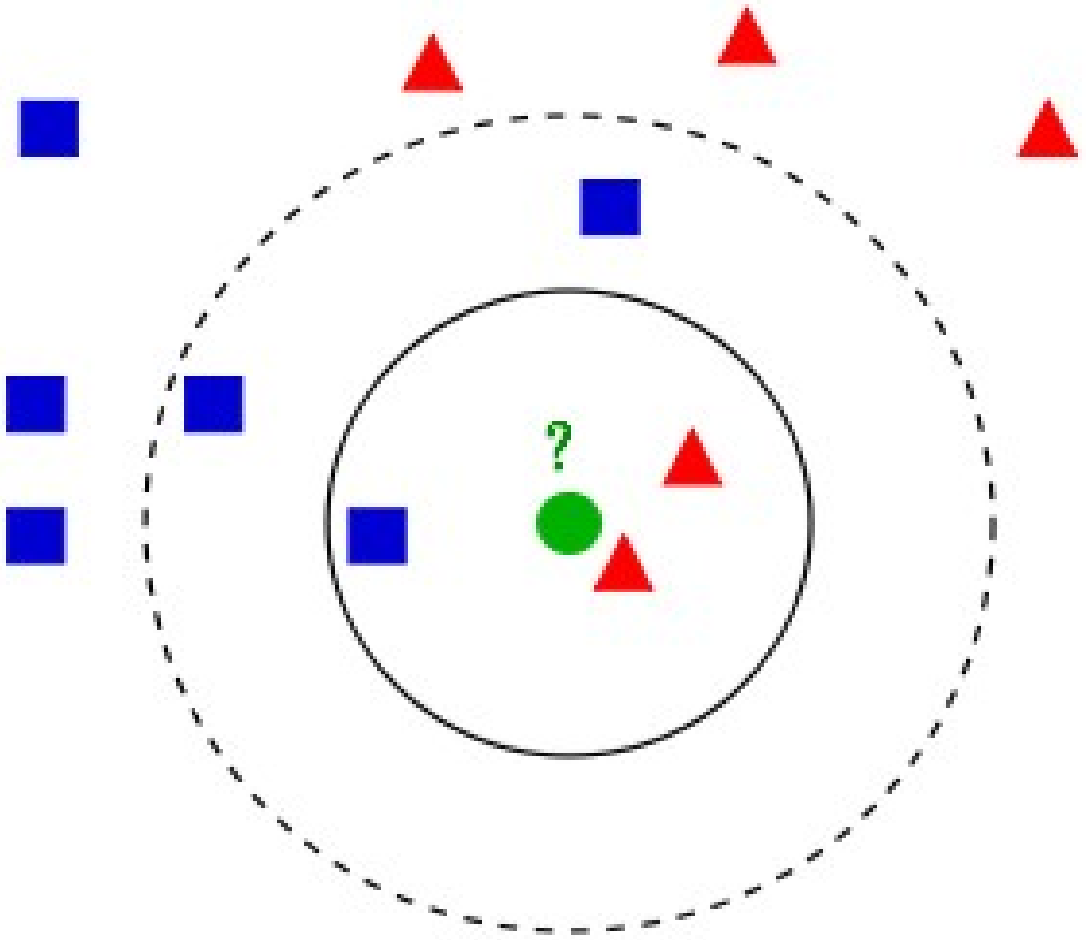
- Podmnožina redukovaná na základe určitých kritérií
- V relácii s pôvodnou množinou, nie v relácii medzi sebou

Particle Swarm Optimization

- Simulácia sociálneho správania
- Podobný s GA

Methods/ Technology involved	Importance	Area/s	Advantages	Disadvantages	Problems
Filter Selection Techniques	Compute the importance of each feature (gene) and then select the top ranked	Gene Selection	Simple Fast Easy scales to very high dimensional data	Univarate that means each feature is considered and treated separately, ignoring any correlation between features	Low classification performance
Wrapper Selection Technique	Selects subset of features that is useful to build a good classifier or predictor	Gene Selection	The ability to take into account the correlation between features and the interaction with the classifier	Prone to high risk of over fitting It require very intensive computation	Unfeasible for feature selection in high-dimensional data More complex

K-Nearest Neighbor



Selection Algorithms	Classification Algorithms
Genetic Algorithm (GA)	Support Vector Machine (SVM)
Correlation-based heuristics (Correlation-based feature selection) (CFS)	Bootstrapped SVM
Particle Swarm Optimization (PSO)	K-Nearest Neighbors (KNN)
Analysis of Variance (ANOVA)	Naïve Bayes
Information Gain (IG)	Neural Networks (NN)
Relief Algorithm (RA)	Decision Tree (DT)
t-statistics (TA)	Bagging and Stacking Algorithms
	Fuzzy Model

Cancer / Algorithm	Ovarian	Prostate	Lung	Colon	Breast	Bladder	Leukemia	Brain	Lymphoma	CNS
Genetic Algorithm	✓	✓	✓	✓	✓	✓	✓		✓	✓
Correlation based heuristics	✓	✓	✓	✓	✓	✓	✓			
Decision tree	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Support Vector Machine	✓	✓	✓	✓	✓	✓	✓	✓		
Particle Swarm Optimization	✓			✓	✓					
Analysis of variance	✓			✓	✓					
Fuzzy Model	✓			✓	✓					
Information Gain	✓	✓	✓	✓	✓		✓	✓		
Relief Algorithm	✓	✓	✓	✓	✓		✓	✓		
t-statistics	✓	✓	✓	✓	✓		✓	✓	✓	✓
K nearest Neighbor	✓	✓	✓	✓	✓		✓	✓		
Naïve Bayes	✓	✓	✓	✓	✓		✓	✓		
Neural Network	✓	✓	✓	✓	✓		✓	✓		