Laboratoire d'Informatique
Fondamentale de Lille

UNIVERSIDAD
DE MÁLAGA

Gene Selection in Cancer Classification using PSO/SVM and GA/SVM Hybrid Algorithms

Enrique Alba, José García-Nieto,
Laetitia Jourdan and El-Ghazali Talbi

INRIA
FUTURS

CEC 2007 IEEE

Outline

- Motivations & Objectives
- Gene Selection and Classification. Methodology
- Algorithms Descriptions. Operators and Details
- Datasets
- Experimental Results and Comparisons
- Conclusions and Further Work

Motivations

- Microarray experiments produce gene expression patterns that provide information about cell function
- Allowing to analyze thousands of genes (Breast cancer 24481, Lung 12533, …)
- However, expression data are highly redundant and noisy (most of genes are believed to be uninformative)
- Large number of genes and small number of samples
- Extracting and analyzing information from large datasets is highly complex
- Reduction techniques improving the learning accuracy are critically important (Data mining + Metaheuristics)

Objectives

- Distinguish (Classify) tumor samples from normal ones (2 classes)
- Discover reduced subsets with informative genes, achieving high accuracies
- Geometric PSO (GPSO) for feature selection
- Classification with Support Vector Machines
- Algorithms comparisons. GPSO vs. GA
- Experimentation using 6 public cancer datasets

Feature Selection (FS) I

- FS can reduce the dimensionality of the datasets
- Two models of FS: *Wrapper* and *Filter*

Depending on whether the selection is coupled with a learning scheme or not

- Support Vector Machines (SVM), a wrapper method was used in this work.

Advantageous since the features are selected by optimizing the discriminate power of the induction algorithm used

- FS problem definition

Given a set of features $F = \{f_1, ..., f_i, ..., f_n\}$, find a subset $F' \subseteq F$, that maximizes a scoring function $\Theta : \Gamma \rightarrow G$ such that

$$F' = argmax_{G \subset \Gamma}\{\Theta(G)\},$$

where Γ is the is the space of all possible feature subsets of *F* and *G* a subset of Γ

Feature Selection (FS) II

- Evaluation of solutions by means of SVM to assess the quality of the gene subset represented
- After this, 10-Fold Cross Validation is applied to calculate the rate of correct classification
- Fitness Aggregative Function (minimization):

$$fitness(x) = \alpha \cdot (100/accuracy) + \beta \cdot \#features$$

- Adapted initialization method:
  - The population (swarm) was divided into four subsets of individuals (particles), such that:
    - 10% of individuals $\longrightarrow$ *N* genes (*N* 1's in the solution)
    - 20% of individuals $\longrightarrow$ 2*N genes*
    - 30% of individuals $\longrightarrow$ 3*N* genes
    - 40% of individuals $\longrightarrow$ randomly

N = 4 in experiments

FS Methodology

Cancer Dataset

Genes / Expression Levels

G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	Classes
-214	-153	-58	88	-295	-558	199	-176	252	206	-41	-831	Normal
-139	-73	-1	283	-264	-400	-330	-168	101	74	19	-743	Normal
-76	-49	-307	309	-376	-650	33	-367	206	-215	19	-135	Tumor
-135	-114	-256	12	-419	-585	158	-253	49	31	363	-934	Normal
-106	-125	-76	168	-230	-284	4	-122	70	252	155	-471	Tumor
-138	-85	215	71	-272	-558	67	-186	87	193	325	-631	Tumor
-72	-144	238	55	-399	-551	131	-179	126	-20	-115	-103	Normal
-413	-260	7	-2	-541	-790	-275	-463	70	-169	-20	-143	Tumor

Accuracy 75%

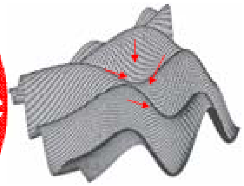
Number of features

75(4)

Fitness = 0.75 x 0.75 + 0.25 x 4

Alpha & Beta parameters

Provided by Metaheuristic GPSO/GA



Solution

0	0	1	0	1	1	0	0	1	0	0	0

Feature Selection

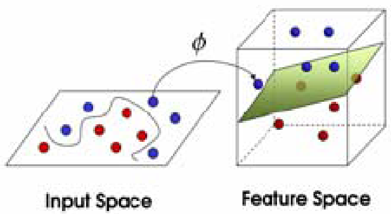
Cross Validation

4 Selected features

G3	G5	G6	G9	Classes
-58	-295	-558	252	Normal
-1	-264	-400	101	Normal
-307	-376	-650	206	Tumor
-256	-419	-585	49	Normal
-76	-230	-284	70	Tumor
215	-272	-558	87	Tumor
238	-399	-551	126	Normal
7	-541	-790	70	Tumor

Support Vector Machines

Training classifier



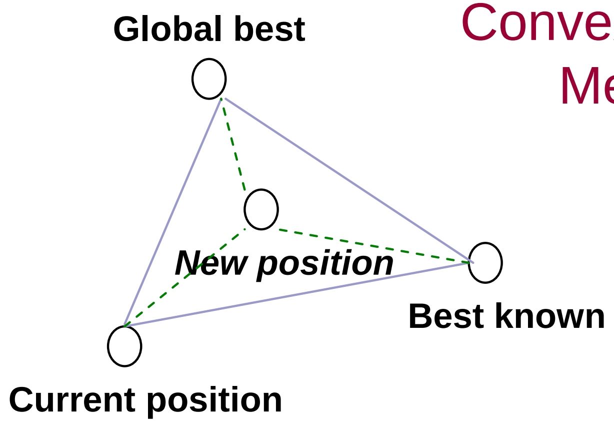
G3	G5	G6	G9
-58	-295	-558	252
-1	-264	-400	101
-307	-376	-650	206
-256	-419	-585	49
-76	-230	-284	70
215	-272	-558	87
238	-399	-551	126
7	-541	-790	70

prediction

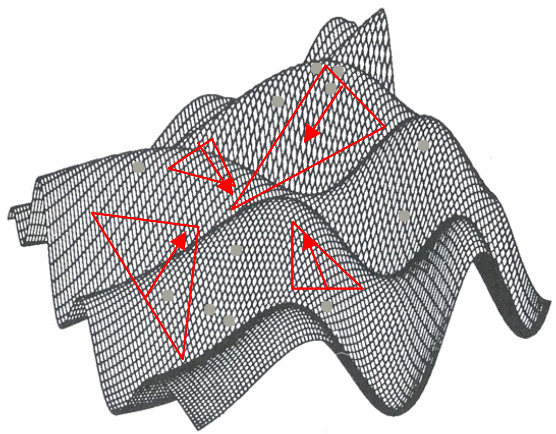
Predicted Classes
Normal
Tumor
Tumor
Normal
Tumor
Tumor
Normal
Normal

Geometric PSO

- Based on Poli & Moraglio 2006, a new binary representation PSO algorithm
- Provide support for more representations: continuous, permutations,…
- Using Metric Space frameworks: Hamming, Euclidean, Manhattan
- Operators
  - Movement by Three Parent Geometric Crossover
  - Without velocity factor
  - Application of Mutation (BitFlip)
  - Adaptation of Three Parent Saving Pattern for FS



Convex Combination
Metric Space



Geometric PSO

- Pseudocode

$S \leftarrow SwarmInitialization()$
while not stop condition **do**
  for each particle x_i of the swarm S **do**
    evaluate(x_i)
    if $fitness(x_i)$ is better than $fitness(h_i)$ **then**
      $h_i \leftarrow x_i$
    end if
    if $fitness(h_i)$ is better than $fitness(g_i)$ **then**
      $g_i \leftarrow h_i$
    end if
  end for
  for each particle x_i of the swarm S **do**
    $x_i \leftarrow 3PMBCX((x_i, w_1), (g_i, w_2), (h_i, w_3))$
    mutate(x_i)
  end for
end while
Output: best solution found

Canonical PSO

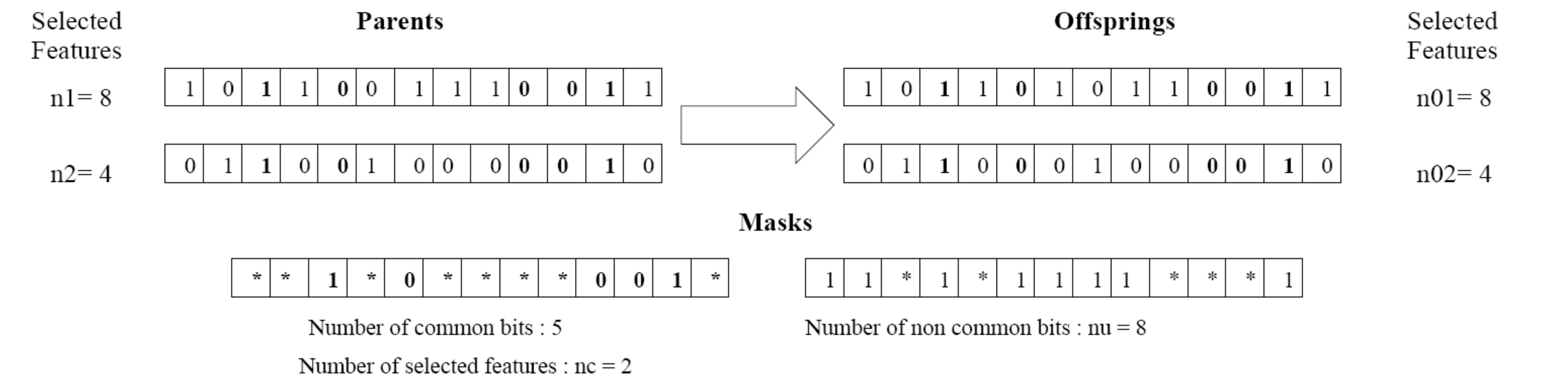
Movement

Three Parent Geometric Crossover

Genetic Algorithm

- Generational evolution
- Elitist
- Operators:
  - Deterministic tournament Selection
  - Subset Size-Oriented Common Feature Crossover Operator (SSOCF)
  - Uniform Mutation (bitflip)

SSOCF



Data Sets

Kent Ridge Bio-medical Data Set Repository

http://sdmc.lit.org.sg/GEDatasets/Datasets.html

- ALL-AML Leukemia. 7129 gene expression levels and 72 samples
- Breast Cancer. 24481 gene expression levels and 97 samples
- Colon Tumor. 2000 gene expression levels and 62 samples
- Lung Cancer. 12533 gene expression levels and 181 samples
- Ovarian Cancer. 15154 gene expression levels and 162 samples
- Prostate Cancer. 12600 gene expression levels and 136 samples

Experiments

- Configurations
  - SVM: Linear Kernel using the libsvm library
  - Metaheuristics parameters

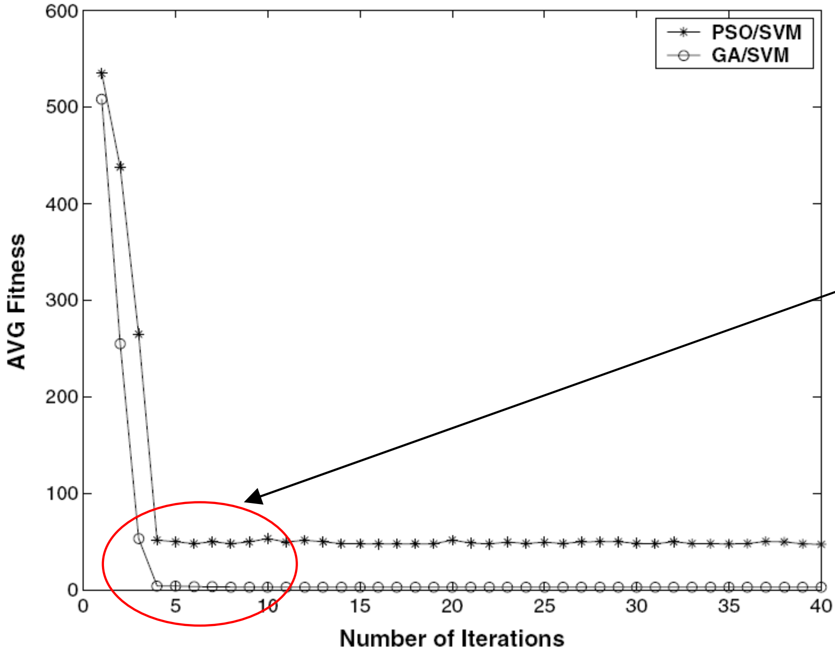
PSO		GA	
Parameter	**Value**	**Parameter**	**Value**
Swarm size	40	Population size	40
Number of generations	100	Number of generations	100
Neighborhood size	20	Probability of crossover	0.9
Probability of mutation	0.1	Probability of mutation	0.1
(w1, w2, w3)	(0.33, 0.33, 0.34)	-	-

- Executions
  - Two algorithms: GPSO (MALLBA Library), GA (Paradiseo Framework) and six datasets
  - 10 independent runs each one

Results

- Performance Analysis
  - Both algorithms obtain acceptable results in few iterations

Dataset	GPSO	GA	Huerta et al.	Juliusdottir et al.	Deb et al.	Guyon et al.	Yu et al.	Liu et al.	Shen et al.
Leukemia	97.38(3)	97.27(4)	100(25)	-	100(4)	**100(2)**	87.44(4)	-	-
Breast	86.35(4)	**95.86(4)**	-	-	-	-	79.38(67)	-	-
Colon	**100(2)**	100(3)	99.41(10)	94.12(37)	97(7)	98(4)	93.55(4)	85.48(-)	94(4)
Lung	99.00(4)	**99.49(4)**	-	-	-	-	98.34(6)	-	-
Ovarian	**99.44(4)**	98.83(4)	-	-	-	-	-	99.21(75)	-
Prostate	**98.66(4)**	98.65(4)	-	88.88(20)	-	-	-	-	-



In few iterations the average of fitnes Decrease quickly

GAsvm obtains generally lower average than GPSOsvm, whose solutions have in turn higher diversity

Results

- Algorithm Robustness
  - The total accuracy and the number of selected features in all the cases didn't deviate from each other by more than 5.5

Dataset	PSO_{SVM}			GA_{SVM}		
	Best	Mean	Std Dev.	Best	Mean	Std Dev.
Leukemia	100(3)	97.38(3)	3.80	100(4)	97.27(4)	3.82
Breast	90.72(4)	86.35(4)	4.11	100(4)	95.86(4)	5.33
Colon	100(2)	100(2)	0.0000	100(3)	100(3)	0.0000
Lung	99.44(4)	99.00(4)	0.50	100(4)	99.49(4)	0.41
Ovarian	100(4)	99.44(4)	0.38	100(4)	98.83(4)	3.18
Prostate	100(4)	98.66(4)	1.14	100(4)	98.65(4)	3.24

- Examples of Selected Gene Subsets

Dataset	PSO_{SVM}		GA_{SVM}	
Leukemia	100(3)	*U39226_at, L12052_at, X99101_at*	100(4)	*Z26634_at, HG870-HT870_at X52005_at, L02840_at*
Breast	90.72(4)	*NM_012269, NM_002850 AL162032, AB022847*	100(4)	*NM_005014, AF060168 NM_021176, NM_013242*
Colon	100(2)	*U29092, M55543*	100(3)	*M90684, M94132 X62025*
Lung	99.44(4)	*31820_at, 33389_at 39057_at, 40772_at*	100(4)	*31573_at, 33226_at 36245_at, 37076_at*
Ovarian	100(4)	*MZ49.784115, MZ3546.2884 MZ4362.0866, MZ9159.3641*	100(4)	*MZ420.40671, MZ825.16557 MZ1024.6857, MZ1166.0749*
Prostate	100(4)	*35106_at, 35869_at 36754_at, 37107_at*	100(4)	*41447_at, 34299_at 39556_at, 39813_s_at*

Conclusions

- Two hybrid algorithms for gene selection and classification of high dimensional DNA Microarray were presented
- New algorithm GPSO for feature selection was applied
- GPSOsvm vs. GAsvm were experimentally assessed on six well-known datasets
- Results of 100% accuracy and few genes per subset (3 and 4)
- Use of adapted initialization method
- Use of adapted operators for FS (3PMBCX & SSOCF)
- Biological analysis of selected gene subsets

Further Work

- Develop and test new combinations of other metaheuristic algorithm with classification methods (KNN,…)
- Use of Multiobjective approaches
- Application of Parallel approaches
- Gene selection and classification of new real datasets

Thanks!

&

Questions