A Parallel Classification Method for Genomic and Proteomic Problems

Mario R. Guarracino
ICAR-CNR, Italy

Claudio Cifarelli
University of Rome La Sapienza, Italy

Onur Seref, Panos M. Pardalos
Center for Applied Optimization - University of Florida, USA

Outline

- Introduction
- Applications of classification methods
- Motivation
- Existing solutions
- P-Regec
- Experimental results
- Conclusion and future work

Introduction

- *Supervised learning* refers to the capability of a system to learn by examples (*training set*).
  - S*upervised* means the examples are provided by an external teacher.
- The trained system is able to provide an answer (*output*) for each new question (*input*).
- *Binary classification* is among the most successful methods for supervised learning.

F. Cucker and S. Smale *On the mathematical foundation of learning*. Bulletin of the American Mathematical Society, 39(1), 1-49, 2001.

Applications

- Many applications in biology and medicine:
  - Tissues that are prone to cancer can be detected with high accuracy,
  - New DNA sequences or proteins can be tracked down to their origins,
  - Protein folding provides important information on protein expression level,
  - Identification of new genes or isoforms of gene expressions in large datasets,
  - Analysis and reduction of data spatiality and principal characteristics for protein determination.

V. Boginski, P. Pardalos, A. Vazacopoulos *Network-based Models and Algorithms in DM & KD. Handbook of Combinatorial Optimization, Kluwer (5), 217-258, 2004*

Parallel algorithms

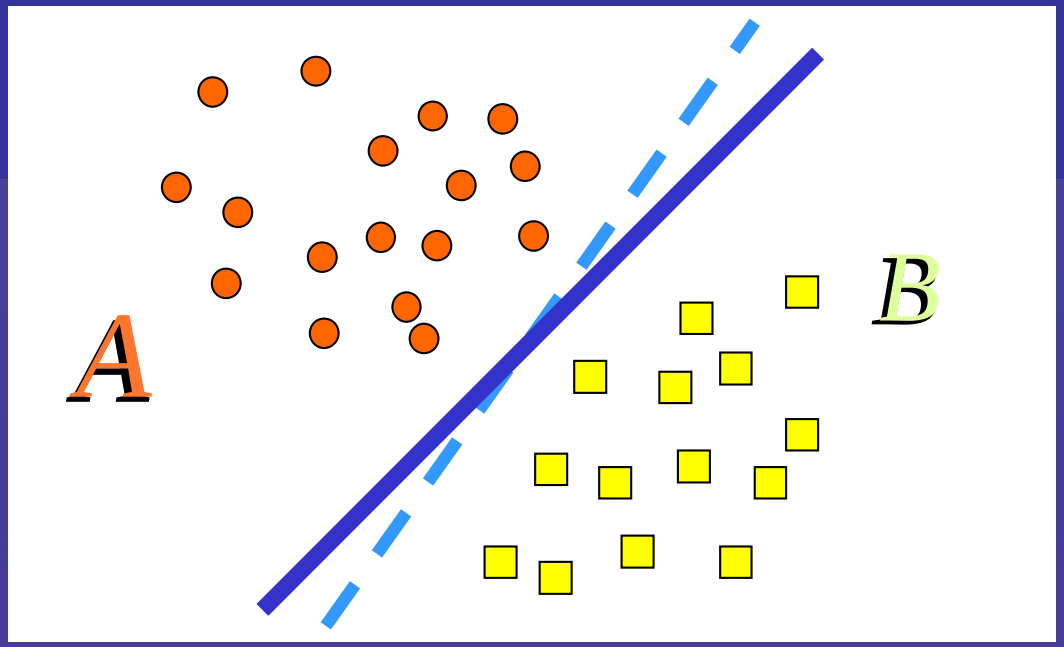
- Genomic and proteomic classification problems have peculiar characteristic:
  - Test sets are two orders of magnitude larger than train
  - Thousands of features
  - Frequent updates
  - Dimension of datasets is rapidly increasing
- Need for tools specifically designed to target genomic problems.

Existing methods

- Many methods available for classification.
  - Support Vector Machines, Fisher linear discriminant analysis, Nearest neighbor classification trees, Genetic algorithms, …
- Existing parallel methods either suffer for accuracy degradation or for poor performances.
- Need of fully parallelized methods.

Linear discriminant planes

- Consider a binary classification task of linearly separable sets.
  - There exists a plane that classifies all points in the two sets

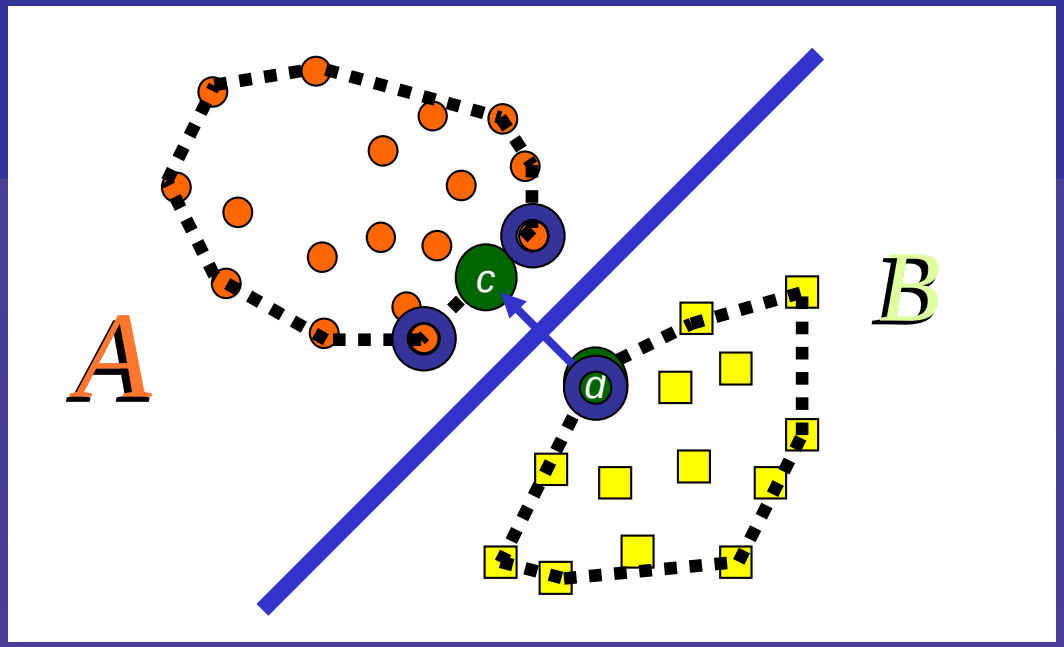


- There are infinitely many planes that correctly classify the training data.

K. Bennet and C. Campbell *Support Vector Machines: Hype or Hallelujah?*, SIGKDD Expl., 2, 2, 1-13, 2000.

Best plane

- To construct the *further* plane from both classes, we examine the *convex hull* of each set.



$$\min_{\alpha,\beta} \frac{1}{2}\|c-d\|^2$$

$$c = \sum_{x_i \in A} \alpha_i x_i \quad d = \sum_{x_i \in B} \beta_i x_i$$

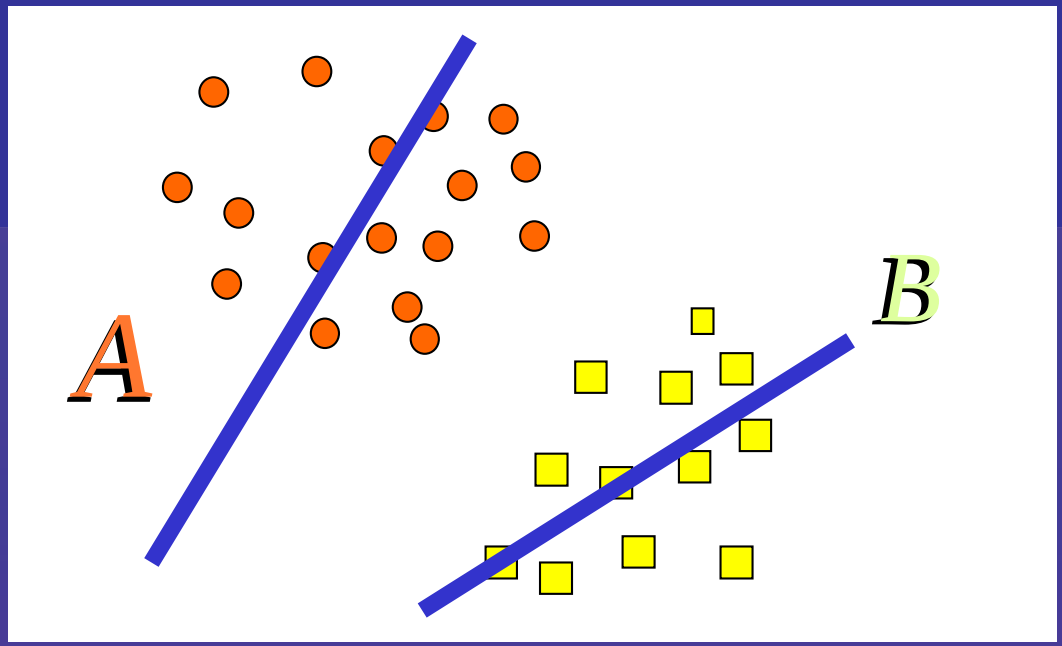
$$s.t. \sum_{x_i \in A} \alpha_i = 1 \quad \sum_{x_i \in B} \beta_i = 1$$

$$\alpha_i,\ \beta_i \geq 0$$

- The best plane bisects closest points in the convex hulls.

A different approach

- The problem can be restated as: find two hyperplanes that *describe* the two classes.



$$\min_{w,\gamma \neq 0} \frac{\|Aw - e\gamma\|}{\|Bw - e\gamma\|}$$

- The binary classification problem can be formulated as a generalized eigenvalue problem (GEP).

O. L. Mangasarian and E. W. Wild Multisurface Proximal Support Vector Classification via Generalized Eigenvalues. Data Mining Institute Tech. Rep. 04-03, June 2004.

GEC method

$$\min_{w,\gamma \neq 0} \frac{\|Aw - e\gamma\|}{\|Bw - e\gamma\|}$$

Let:

$$G = [A \quad -e]'[A \quad -e],\ H = [B \quad -e]'[B \quad -e],\ z = [w' \quad \gamma]'$$

Previous equation becomes:

$$\min_{z \in R^m} \frac{z'Gz}{z'Hz}$$

Raleigh quotient of Generalized Eigenvalue Problem

$$Gx = \lambda Hx.$$

GEC method

Conversely, the plane closer to B and further from A:

$$\min_{w,\gamma \neq 0} \frac{\|Bw - e\gamma\|}{\|Aw - e\gamma\|}$$

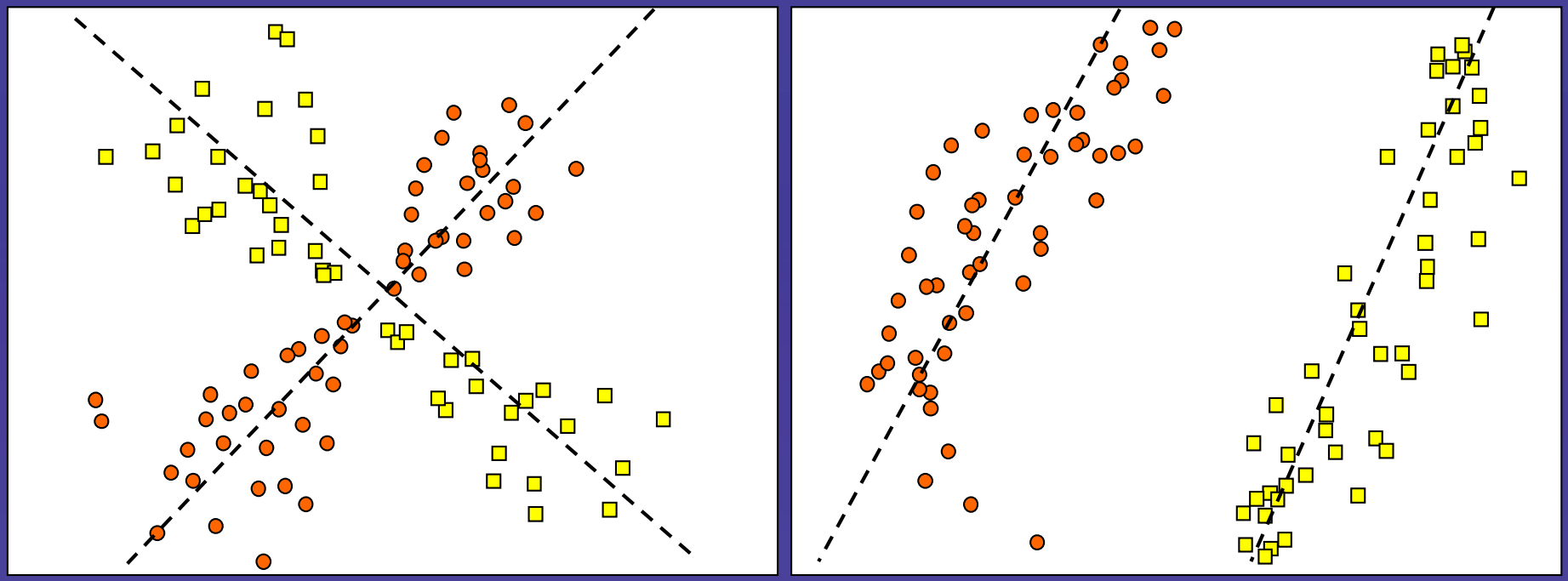
Same eigenvectors of the previous problem and reciprocal eigenvalues.

We only need to evaluate the eigenvectors related to min and max eigenvalues of $Gx = \lambda Hx$.

GEC method

Let $[w_1\ \gamma_1]$ and $[w_m\ \gamma_m]$ be eigenvectors associated to min and max eigenvalues of $Gx=\lambda Hx$:

- $a \in A$ closer to $x'w_1 - \gamma_1 = 0$ than to $x'w_m - \gamma_m = 0$,
- $b \in B$ closer to $x'w_m - \gamma_m = 0$ than to $x'w_1 - \gamma_1 = 0$.



Classification accuracy: linear kernel

dataset	*train*	*s*	*GEC*	*GEPSVM*	*SVMs*
NDC	300	7	87.60	86.70	89.00
ClevelandHeart	297	13	86.05	81.80	83.60
PimaIndians	768	8	74.91	73.60	75.70
GalaxyBright	2462	14	98.24	98.60	98.30

Accuracy results have been obtained using ten fold cross validation

Nonlinear case

- A standard technique to obtain greater separability between classes is to embed the points into a nonlinear space, via kernel functions, like the *gaussian kernel* :

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{\sigma}}$$

- Each element of kernel matrix is:

$$K(A, C)_{i,j} = e^{-\frac{\|A_i - C_j\|^2}{\sigma}}$$

where

$$C = \begin{bmatrix} A \\ B \end{bmatrix}$$

K. Bennett and O. Mangasarian, *Robust Linear Programming Discrimination of Two Linearly Inseparable Sets*, Optimization Methods and Software, 1, 23-34, 1992.

Nonlinear case

- Using a gaussian kernel the problem becomes:

$$\min_{w,\gamma \neq 0} \frac{\|K(A,C)u - e\gamma\|^2}{\|K(B,C)u - e\gamma\|^2}$$

to produce the proximal surfaces:

$$K(x,C)u_1 - \gamma_1 = 0, \quad K(x,C)u_2 - \gamma_2 = 0$$

- The associated GEP involves matrices in the order of the training set and rank at most the number of features.

ReGEC method

- Matrices are rank deficient and the problem is ill posed.
- We propose to generate the two proximal surfaces:

$$K(x,C)u_1 - \gamma_1 = 0, \quad K(x,C)u_2 - \gamma_2 = 0$$

solving the problem:

$$\min_{w,\gamma \neq 0} \frac{\|K(A,C)u - e\gamma\|^2 + \delta\|\tilde{K}_B u - e\gamma\|^2}{\|K(B,C)u - e\gamma\|^2 + \delta\|\tilde{K}_A u - e\gamma\|^2}$$

where $\tilde{K}_A$ and $\tilde{K}_B$ are main diagonals of $K(A,C)$ and $K(B,C)$.

M. R. Guarracino, C. Cifarelli, O. Seref, P. M.Pardalos, *A Classification Method based on Generalized Eigenvalue Problems*, Optimization Methods and Software, OMS, to appear.

Classification accuracy: gaussian kernel

dataset	*train*	*test*	*s*	*ReGEC*	*GEPSVM*	*SVM*
Breast-cancer	200	77	9	73.40	71.73	73.49
Diabetis	468	300	8	74.56	74.75	76.21
German	700	300	20	70.26	69.36	75.66
Thyroid	140	75	5	92.76	92.71	95.20
Heart	170	100	13	82.06	81.43	83.05
Waveform	400	4600	21	88.56	87.70	90.21
Flare-solar	666	400	9	58.23	59.63	65.80
Titanic	150	2051	3	75.29	75.77	77.36
Banana	400	4900	2	84.44	85.53	89.15

ReGEC algorithm

Let $A \in R^{m \times s}$ and $B \in R^{n \times s}$ be the training points in each class.
Choose appropriate δ_1, δ_2 and $\sigma \in R$

% Build G and H matrices
g = [*K*(*A*, *C*, σ), *-ones*(*m*, 1)];
h = [*K*(*B*, *C*, σ), *-ones*(*n*, 1)];
G = *g′* × *g*;
H = *h′* × *h*;

% Regularize the problem
$G^* = G + \delta_1 \times diag(H)$;
$H^* = H + \delta_2 \times diag(G)$;

% Compute the classification hyperplanes V(:,1) V(:,m+n+1)
[*V*,*D*] = *eig*(G^*;H^*);

P-ReGEC implementation details

- Parallel ReGEC (P-ReGEC) only uses matrix-matrix products, kernel evaluation and a generalized eigenvalue problem solution.
- Parallel implementation is based on BLACS and MPI message passing libraries, and ScaLAPACK numerical linear algebra library.
- Matrices involved in the algorithm are distributed among nodes.
  - Memory is used efficiently and no replication of data occurs.
- The current data model of PBLAS assumes matrix operands to be distributed according to the *block scattered composition*.

P-ReGEC implementation details

- PBLAS routine PDGEMM is used to evaluate matrix-matrix multiplications.
- The evaluation of the generalized eigenvalue problem $G^*x=\lambda H^*x$ is performed by PDSYGVX routine.
- We developed the auxiliary routines for parallel kernel computation, and for diagonal matrices operations.
- The operation count of P-ReGEC is exactly the same as the sequential algorithm.
- P-ReGEC has a computational complexity $O(n^3)$, and a communication complexity of $O(n^2)$.

TIS Dataset

- The dataset consists of genomic sequences of Translation Initiation Site (TIS)
  - Extracted from GenBank, 10.000 elements with 927 features, divided in 2 classes.
- The problem consists in finding TIS at which the translation from mRNA to proteins initiates.
  - There is a great potential for improvement of the accuracy and speed.
- It provides a significant case study for the analysis of genomic sequences.

Performance evaluation

- Results in terms of execution time and parallel efficiency.
- Cluster of 16 P4s 1.5GHz, 512MB RAM, fast ethernet.
- RedHat 9 - kernel 2.4.20,
  - gcc 2.96, mpich 1.2.5, BLACS 1.1, ScaLAPACK 1.7, LAPACK 3.0, BLAS with ATLAS optimization.
- Tests have been performed on idle nodes
- Wall clock time in seconds of the slower executing node measured with MPI_WTIME().

Elapsed time (nodes vs dimension)

	1	2	4	8	16
500	2.99	3.59	3.07	3.51	4.00
1000	21.90	17.79	12.29	12.61	12.43
2000	162.12	89.79	55.95	46.59	40.54
3000	532.42	260.39	143.93	109.63	87.30
4000	1487.87	562.70	290.02	205.95	155.39
5000	2887.51	1050.02	265.92	342.22	247.36
6000	-	1921.13	812.64	523.99	365.92
7000	-	3414.97	1298.75	753.63	514.66
8000	-	-	1875.02	1046.08	693.84
9000	-	-	2733.95	1421.28	913.16

For problems of sufficient dimension, the time decreases as the number of processor increases.

Efficiency (nodes vs dimension)

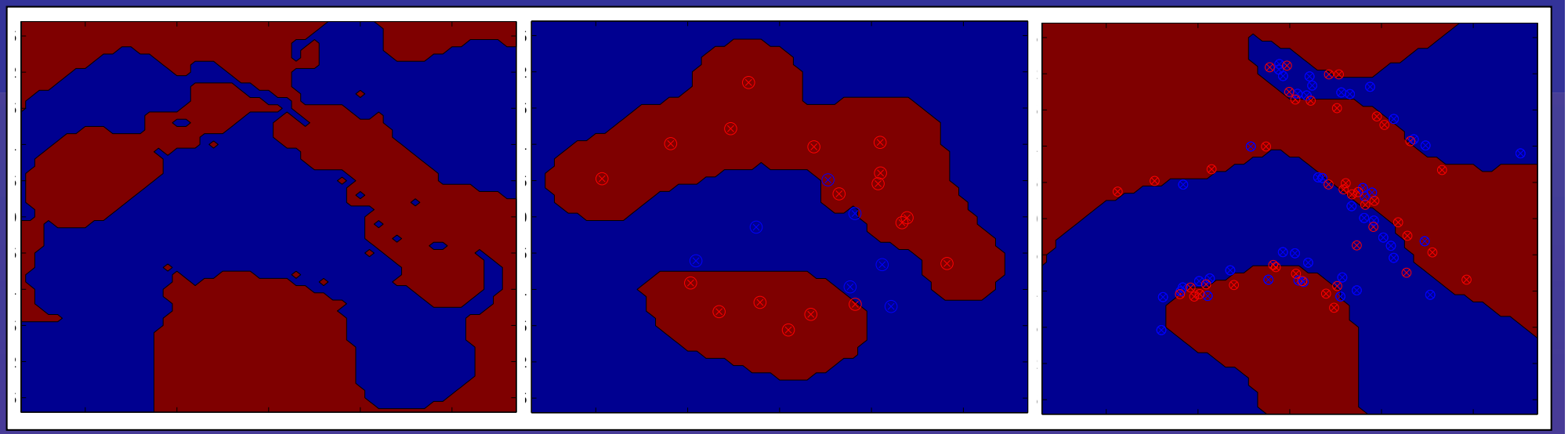
	2	4	8	16
500	0.4175	0.2442	0.1066	0.0468
1000	0.6157	0.4458	0.2172	0.1102
2000	0.9027	0.7244	0.4349	0.2499
3000	1.0223	0.9248	0.6071	0.3812
4000	1.3221	1.2825	0.9031	0.5984
5000	1.375	2.7146	1.0547	0.7296
6000	1	1.182	0.9166	0.6563
7000	1	1.3147	1.1328	0.8294
8000	-	1	0.8962	0.6756
9000	-	1	0.9618	0.7485

$$Eff = \frac{T_1}{p * T_p},$$

Super-linear speed-up due to memory effects

Work in progress

- Implementation and testing of a parallel eigensolver for large sparse matrices.
- Develop an *incremental learning* technique for P-ReGEC.



M.R. Guarracino, F. Perla, P. Zanetti - *HPEC: a software for the evaluation of large sparse eigenvalue problems on multicomputers*, Int. J. of Pure and Appl. Math., to appear, 2005.
M.R. Guarracino, F. Perla, P. Zanetti - *A parallel block Lanczos algorithm and its implementation for the evaluation of some eigenvalues of large sparse symmetric matrices on multicomputers* - Int. J. of Appl. Math. and Comp. Sc, vol. 16 n. 2, 2006.

Conclusion and future work

- P-ReGEC is a fully parallel classification method based on generalized eigenvalue problem.
  - it is as accurate as state-of-the-art methods.
  - it is efficient and scalable on multicomputers architectures.
- It is suitable for genomic and proteomic data analysis.

- In future, development and test of:
  - parallel feature selection techniques to reduce problem complexity,
  - parallel bi-clustering techniques to speed-up computation.

Conclusions

- Supervised learning will continue to be an active research field.
- Many problems in genomics are still open and in need of answers.
- Parallel and distributed methods and tools will play a central role in the next years.