A classification method based on linear algebra computational kernels

Mario R. Guarracino
High Performance Computing and Networking Institute
Italian National Research Council

mario.guarracino@icar.cnr.it

Introduction

- *Supervised learning* refers to the capability of a system to learn from examples (*training set*).
- The trained system is able to provide an answer (*output*) for each new question (*input*).
- S*upervised* means the desired output for the training set is provided by an external teacher.
- *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

- A bank classifies customer loan requests in *good* and *bad*, depending on their ability to pay back.
- Inland revenue tries to discover more tax evaders starting from the characteristics of known ones.
- A car built-in system detects if a walking pedestrian is going to cross the street.
- A decision support system automatically discards medical analysis of patients not showing a specific pattern.

T. Poggio and S. Smale The Mathematics of Learning: Dealing with Data
AMS Notice 537-544, 2003.

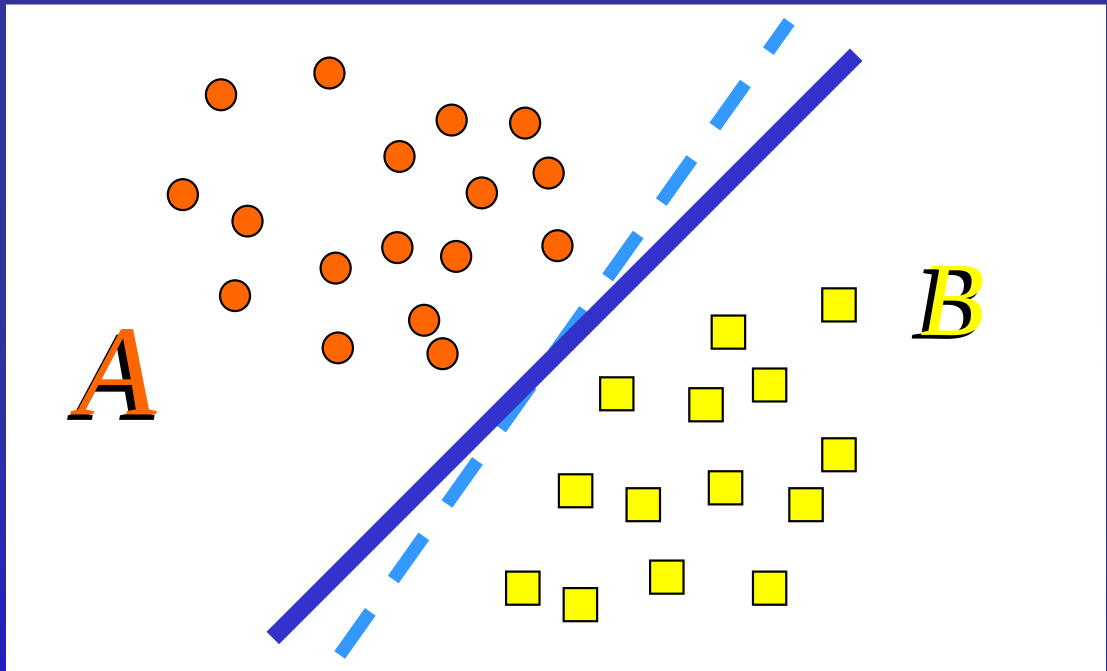
Applications

- 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*

Linear discriminant planes

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

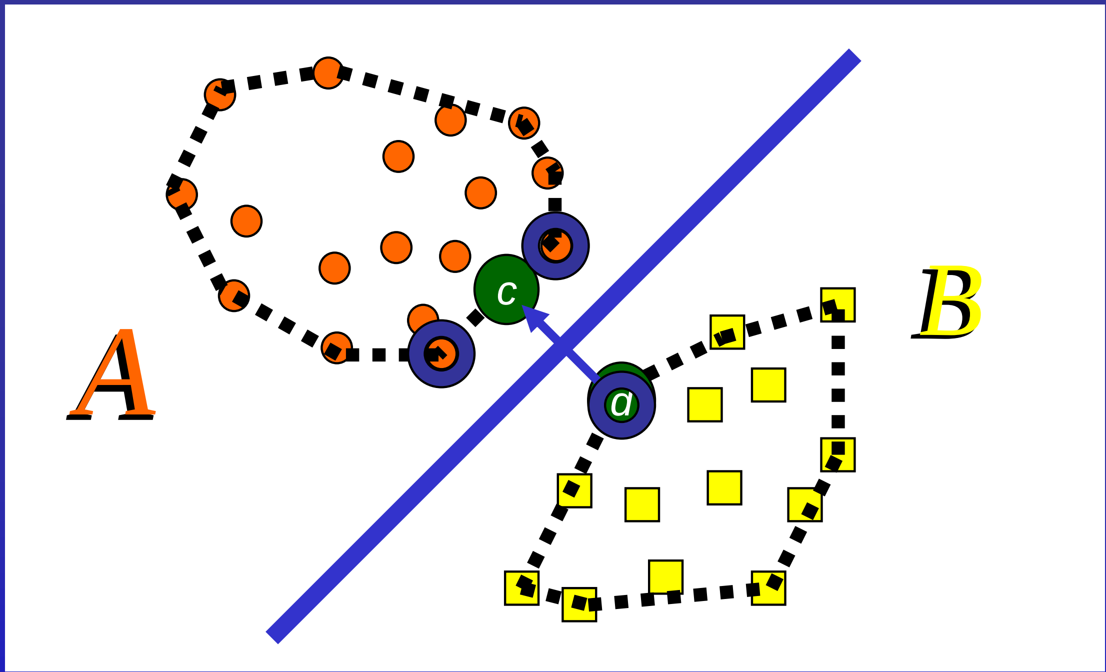


- There are infinitely many planes that correctly classify the training data.

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

Best plane

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



$$\min_{a} \frac{1}{2}\|c-d\|^2$$

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

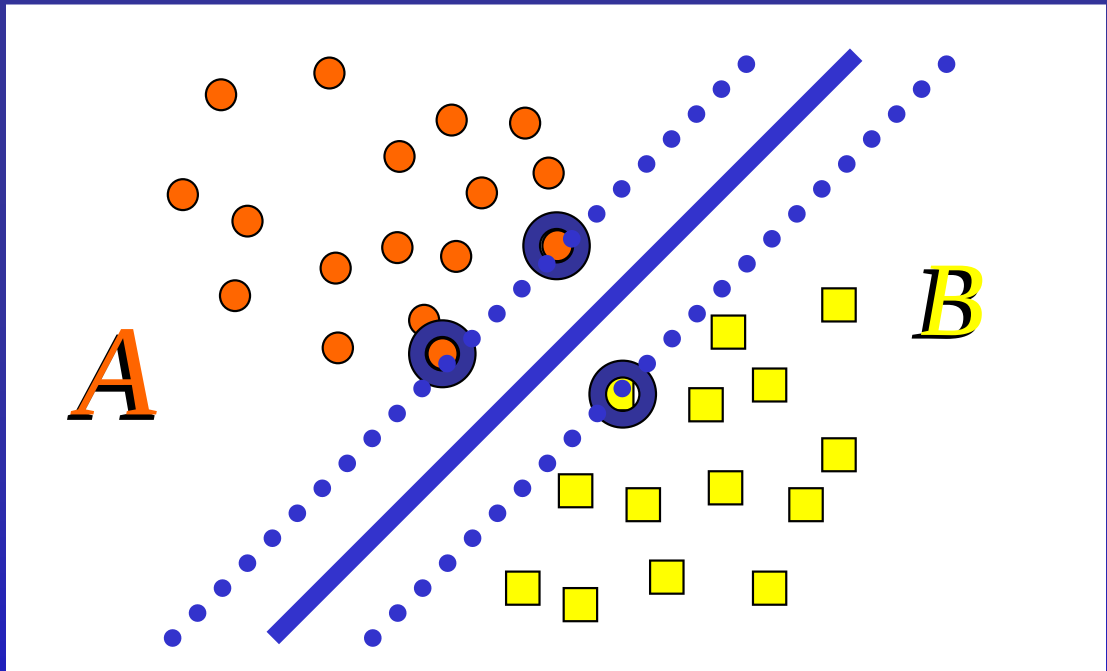
$$s.t. \sum_{x_i \in A} \alpha_i = 1 \quad \sum_{x_i \in B} \alpha_i = 1$$

$$\alpha_i \geq 0$$

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

SVM classification

- A different approach, yielding the same solution, is to maximize the margin between *support planes*
  - Support planes leave all points of a class on one side



$$\min_{a} \frac{1}{2}\|w\|^2$$

$$s.t.$$

$$Aw + b \geq e$$

$$Bw + b \leq -e$$

- Support planes are pushed apart until they “bump” into a small set of data points (*support vectors*).

SVM classification

- Support Vector Machines are the state of the art for the existing classification methods.

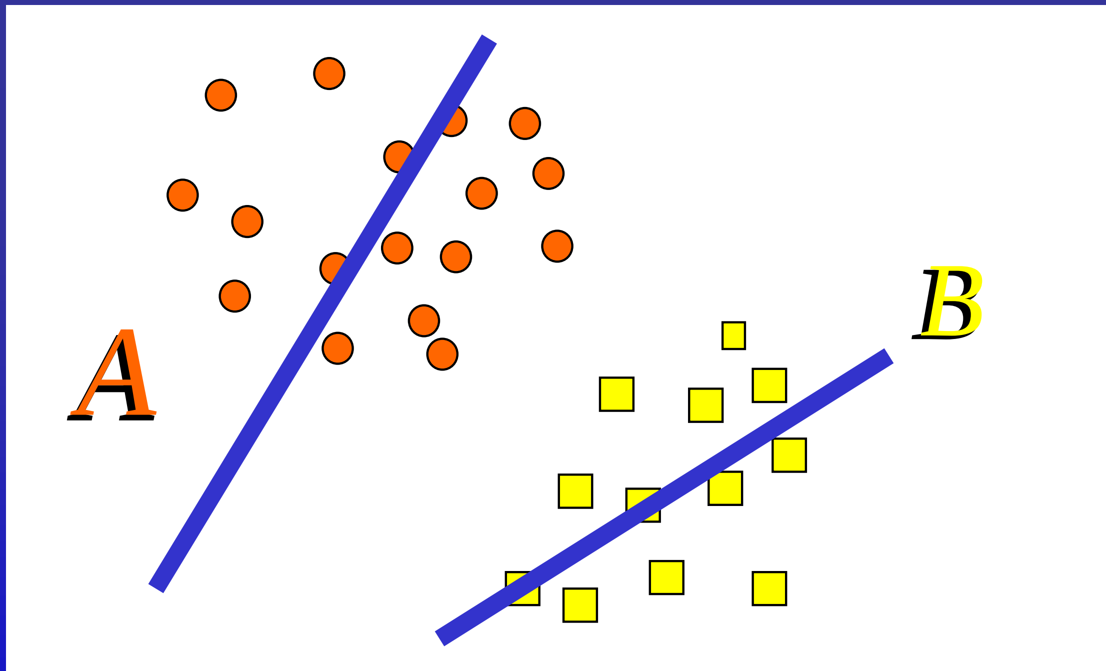
- Their robustness is due to the strong fundamentals of statistical learning theory.

- The training relies on optimization of a quadratic convex cost function, for which many methods are available.
  - Available software includes SVM-Lite and LIBSVM.

- These techniques can be extended to the nonlinear discrimination, embedding the data in a nonlinear space using *kernel functions*.

A different approach

- The binary classification problem can be formulated as a generalized eigenvalue problem (GEP).
- The problem can be restated as: find two hyper planes that *describe* the two classes.



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

GEP formulation

Find a plane $x'w_1=\gamma_1$ the closer to A and the farther from B:

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

GEP technique

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

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

$$Gx = \lambda Hx.$$

GEP technique

Conversely, to find the plane closer to B and further from A we need to solve:

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

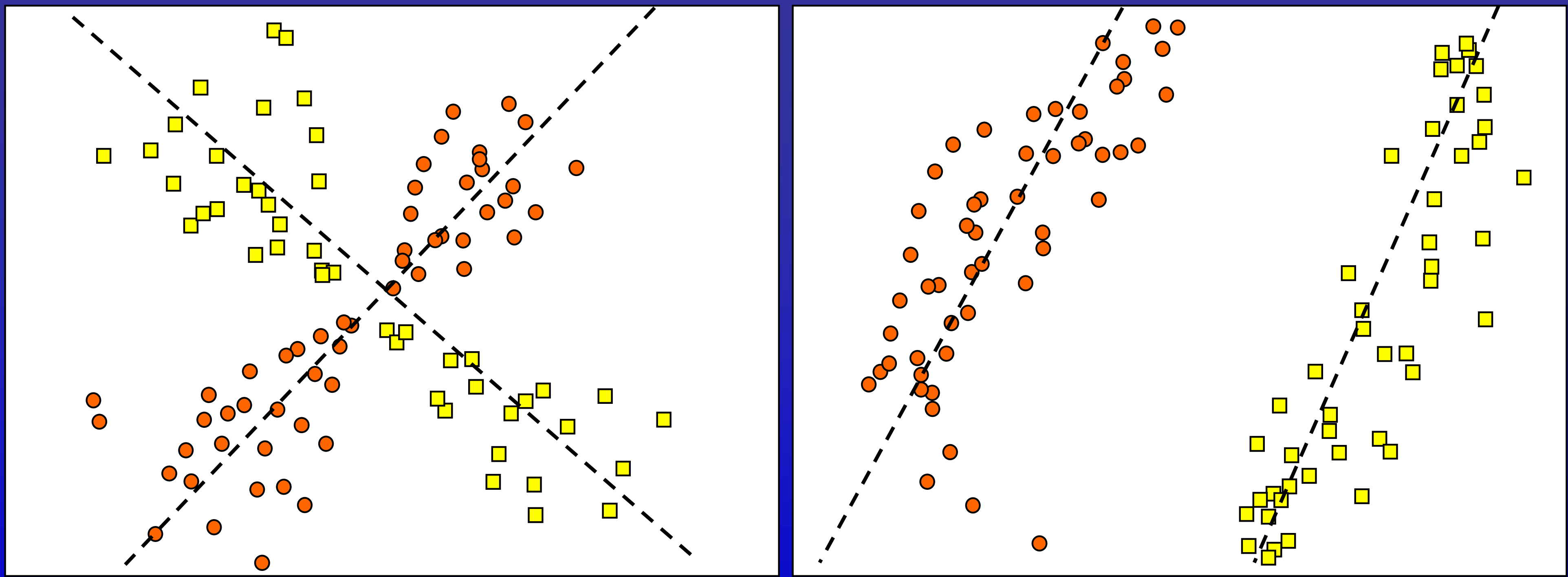
which has the same eigenvectors of the previous problem and reciprocal eigenvalues.

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

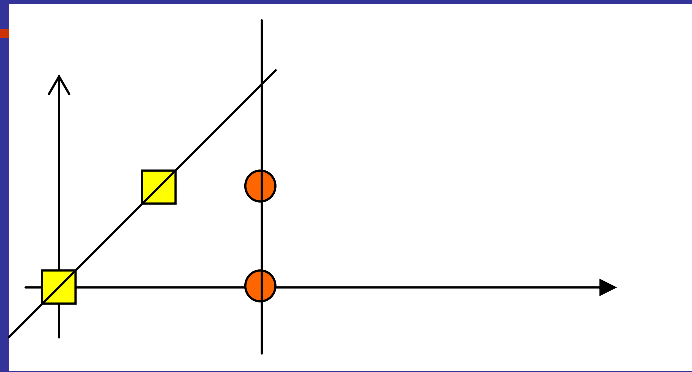
GEP technique

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 $x'w_1 - \gamma_1 = 0$.



Example



Let:

$$A = \begin{bmatrix} 2 & 0 \\ 2 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix},$$

Set G=[A -e]' [A -e] and H=[B -e]' [B -e], we obtain:

$$G = \begin{bmatrix} 8 & 2 & -4 \\ 2 & 1 & -1 \\ -4 & -1 & 2 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 2 \end{bmatrix}.$$

Min and max eigenvalues of $Gx=\lambda Hx$ are $\lambda_1=0$ and $\lambda_3=\infty$ and the respective eigenvectors:

$$x_1=[1\ 0\ 2], \quad x_2=[1\ -1\ 0].$$

The resulting planes are $x = 2$ and $x - y = 0$, which are closer to one set and further from the other.

Regularization

- A and B can be rank-deficient.
- G and H are always rank-deficient,
  - the product of matrices of dimension $(n+1 \times n)$ is of rank at least $n \Rightarrow 0/\infty$ eigenvalue.
- Do we need to regularize the problem to obtain a well posed problem?

A first solution

- Mangasarian et al. proposes GEPSVM:

$$\min_{w,\gamma\neq 0} \frac{\|Aw - e\gamma\|^2 + \delta\|z\|^2}{\|Bw - e\gamma\|^2},$$

$$\min_{w,\gamma\neq 0} \frac{\|Bw - e\gamma\|^2 + \delta\|z\|^2}{\|Aw - e\gamma\|^2},$$

- Only minimum eigenvalue/eigenvector for each problem.
- No gain using computational kernels for single eigenvalue and eigenvector computation.

A useful theorem

Consider GEP $Gx=\lambda Hx$ and the transformed $G_1x=\lambda H_1x$ defined by:

$$G^* = \tau_1 G - \delta_1 H, \quad H^* = \tau_2 H - \delta_2 G,$$

for each choice of scalars τ_1, τ_2, δ_1 and δ_2, such that the 2×2 matrix

$$\Omega = \begin{pmatrix} \tau_2 & \delta_1 \\ \delta_2 & \tau_1 \end{pmatrix}$$

is nonsingular.

Then $G^*x=\lambda H^*x$ has the same eigenvectors of $Gx=\lambda Hx$.

Y. Saad, Numerical Methods for Large Eigenvalue Problems, Halsted Press, New York, NY, 1992.

Linear case

- In the linear case, the theorem can be applied. For $\tau_1=\tau_2=1$ and $\delta_1=\delta_2=\delta$, the transformed problem is:

$$\min_{w,\gamma\neq 0} \frac{\|Aw - e\gamma\|^2 + \delta\|Bw - e\gamma\|^2}{\|Bw - e\gamma\|^2 + \delta\|Aw - e\gamma\|^2}.$$

- As long as $\delta \neq 1$, matrix Ω is non-degenerate.
- This transformation works if, in each class of the training set, there is a number of linearly independent points equal to the number of features.
  - *prob* $(Ker(G) \cap Ker(H) \neq 0) = 0$

M.R. Guarracino On Classification Methods for Mathematical Models of Learning. Rapp. Tec. 3.246.1572 del Gruppo di Ottimizzazione e Ricerca Operativa, Dip. di Matematica, Univ. di Pisa, Febbraio 2005.

Nonlinear case

- A standard technique to obtain greater separability between sets 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 of the order of the training set and rank at most the number of features.

Nonlinear case

- Matrices are deeply 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*, submitted to OMS, 2005.

Numerical experiments

- Performance on benchmark data sets publicly available.
  - Data from UCI, Odewahn, and IDA repository.

- Accuracy results for linear and nonlinear kernel SVMs and GEPSVMare taken from literature.

- Kernel parameters have been taken from literature.

Classification accuracy: linear kernel

dataset	*n+k*	*dim*	*ReGEC*	*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

Elapsed time: linear kernel

dataset	*ReGEC*	*GEPSVM*	*LIBSVM*	*SVMlight*
NDC	0.1e-03	0.2e-03	0.8991	22.002
ClevelandHeart	1.9e-04	3.6e-04	9.9e-03	0.3801
PimaIndians	1.2e-04	2.4e-04	15.873	48.809
GalaxyBright	0.3e-3	0.5e-3	1.2027	21.128

Results computed on Xeon 3.2GHz, 6GB RAM, RH Linux, Matlab 6.5.

Matlab function *eig* used for GEPSVM and ReGEC.
Latest releases of libsmv and SVMlight used.

Classification accuracy: gaussian kernel

dataset	*n+k*	*test*	*m*	δ	σ	*ReGEC*	*GEPSVM*	*SVM*
Breast-cancer	200	77	9	1.e-03	50	73.40	71.73	73.49
Diabetis	468	300	8	1.e-03	500	74.56	74.75	76.21
German	700	300	20	1.e-03	500	70.26	69.36	75.66
Thyroid	140	75	5	1.e-03	0.8	92.76	92.71	95.20
Heart	170	100	13	1.e-03	120	82.06	81.43	83.05
Waveform	400	4600	21	1.e-03	150	88.56	87.70	90.21
Flare-solar	666	400	9	1.e-03	3	58.23	59.63	65.80
Titanic	150	2051	3	1.e-03	150	75.29	75.77	77.36
Banana	400	4900	2	1.e-05	0.2	84.44	85.53	89.15

Accuracy results have been obtained using ten random splits provided by IDA repository

Elapsed time: gaussian kernel

dataset	ReGEC	GEPSVM	LIBSVM	SVMlight
Breast-cancer	0.0698	0.3545	0.0229	0.1188
Diabetis	1.1474	5.8743	0.1323	0.2022
German	3.8177	25.2349	0.2855	0.4005
Thyroid	0.0243	0.1208	0.0053	0.0781
Heart	0.0316	0.2139	0.0172	0.1372
Waveform	0.5962	4.409	0.0916	0.2228
Flare-solar	1.8737	16.2658	0.1429	4.4524
Titanic	0.0269	0.1134	0.0032	7.1953
Banana	0.4989	3.1102	0.0344	1.3505

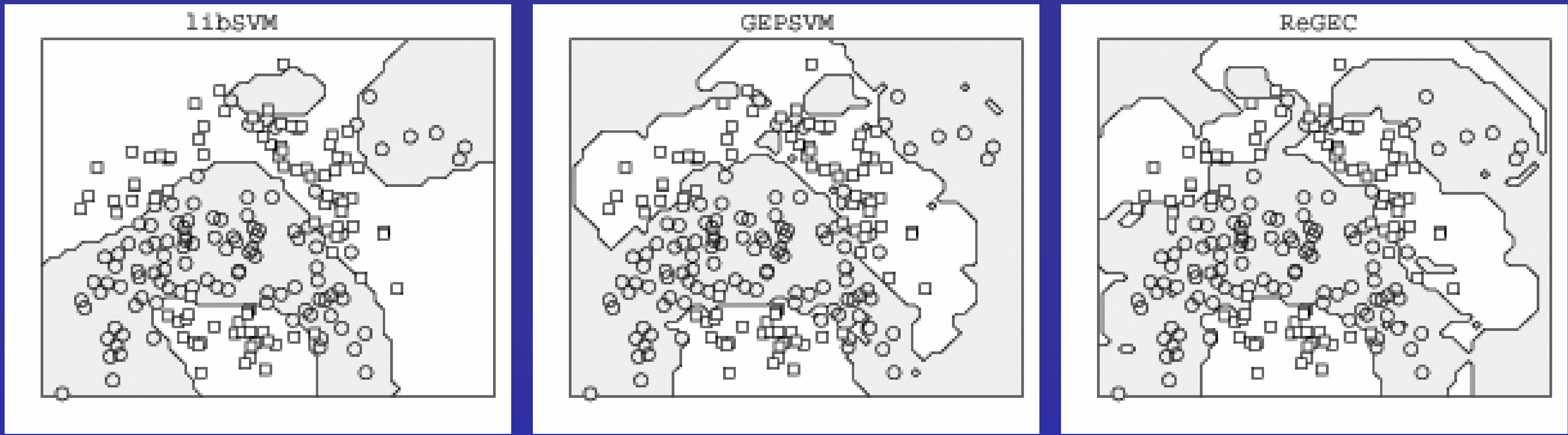
Work in progress

- A parallel eigensolver for large sparse matrices has been implemented and tested.
- Its generalization to GEP is in sight.
- A parallel version of ReGEC has been implemented. Tests of its performance are in progress.

M.R. Guarracino - *HPEC: High Performance Eigenvalue Computation, a software for the evaluation of large sparse eigenvalue problems* - Parallel Matrix Algorithms and Applications, Marsiglia, Ottobre 2004.
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., in print, 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, submitted, 2005.
M.R. Guarracino, F. Perla, P. Zanetti - *A Sparse Nonsymmetric Eigensolver for Distributed Memory Architectures,* Int. J. of Parallel, Emergent and Distributed Systems, submitted, 2005.

Future work

- Develop a *chunking technique* for ReGEC



- (Semi) Automatic determination of parameters.
- Test iterative projection methods with respect to quality assessment of computed solutions.

Conclusions

- Supervised learning will continue to be an active research field.
- Many problem are still open and in need of answers.
- Binary and n-ary classification will play a central role in biomedical applications.