Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

# On Classification Methods for Mathematical Models of Learning

Mario R. Guarracino

**RT-ICAR-NA-2005-13** **Novembre 2005**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR) – Sede di Napoli, Via P. Castellino 111, I-80131 Napoli, Tel: +39-0816139508, Fax: +39-0816139531, e-mail: napoli@icar.cnr.it, URL: *www.na.icar.cnr.it*

Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

# On Classification Methods for Mathematical Models of Learning[1]

Mario R. Guarracino[2]

***Rapporto Tecnico N.:***
**RT-ICAR-NA-2005-13**

***Data:***
**Novembre 2005**

[1] anche Rapporto Tecnico 3.246.1572 del Gruppo di Ottimizzazione e Ricerca Operativa, Dipartimento di Matematica, Università degli Studi di Pisa, Febbraio 2005

[2] *High Performance Computing and Networking Institute Italian National Research Council*
*Via P. Castellino, 111, 80131, Napoli (Italy)*

# On Classification Methods for Mathematical Models of Learning

**Mario Rosario Guarracino**

*High Performance Computing and Networking Institute*

*Italian National Research Council*

*Via Pietro Castellino, 111 80131 Naples (It)*

*e-mail:* mario.guarracino@icar.cnr.it

*February 2005*

## Abstract

The present paper intends to describe some classification methods based on computational kernels developed in the field of generalized eigenvalue problems. It will be illustrated how some numerical difficulties can be overcome and how to obtain a simple iterative algorithm for binary and n-ary classification. Finally, some hints will be given on how eigenvalues techniques can be used in mathematical models of learning.

## 1 Introduction

Mathematical modeling of learning has a key role in many scientific and technological problems and can be considered as one of the most interesting problems of this new century.
*Supervised learning* is referred to the capability of a system to learn from a set of examples, that is a set of input/output couples; starting from that set, the system is able to give an answer (output), as soon as a new question (input) is provided. The term supervised originates from the fact that the desired output on a set of specific input points is provided by an external teacher.

Systems for supervised learning can find application in many fields. Let's suppose a bank needs to classify customer loan requests in "good" and "bad", depending on their ability to pay back, or an inland revenue tries to discover more tax evaders starting from the characteristics of known ones. Furthermore, a car built-in system could detect if a walking pedestrian is going to cross the street. More examples related to numerical interpolation, handwriting recognition and Montecarlo methods for numerical integration can be found in [5].

A word about the notation used in the paper. All vectors are column vectors, unless transposed to row vectors by a prime $'$. Scalar product of two vectors

$x$ and $y$ in $\mathbb{R}^n$ will be denoted by $x'y$, 2-norm of $x$ will be denoted by $\|x\|$. Finally, the unit vector will be denoted by $e$.
The remainder of the work is organized as follows. Section 2 describes a model for learning and how that represents a generalization of binary classification introduced by Support Vector Machine (*SVM*) methods. In Section 3, it is shown how a binary classification algorithm can be brought back to a generalized eigenvalue problem; furthermore, first results on the use of eigenvalue problem techniques for binary classification are detailed and it is shown how they can be generalized to n-ary classification. In Section 4, work related to algorithms for the evaluation of some eigenvalues of large symmetric matrices is reported. Open problems are described in Section 5 and finally, in Section 6, conclusions are drawn and future work is proposed.

## 2 Related work

To provide an answer to problems outlined in the previous section, different mathematical models of learning have been proposed in literature. In [13], Poggio and Smale have proposed a simple algorithm to determine, starting from a data set, a function interpolating data in a *predictive* way, in analogy with physical studies, in which models are conceived to forecast physical phenomena.
Given the set of examples $S_m = (x_i, y_i)_{i=1}^m$, with $x_i \in \mathbb{R}^n,\ y_i \in \mathbb{R}$, define a positive definite kernel $K(t, s)$:

$$K(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}. \tag{1}$$

Introduce the following function:

$$f(x) = \sum_{i=1}^{m} c_i K(x_i, x). \tag{2}$$

with $c \in \mathbb{R}^m$ solution of:

$$(m\gamma I + K)c = y, \tag{3}$$

where $I$ is the identity matrix, $K$ is a definite positive matrix with elements $K_{i,j} = K(x_i, x_j)$, and $\gamma$ is a constant obtained by Tikhonov regularization method[1]. The function is such that $f(x_i) = y_i$ and, for each $\overline{x} \neq x_i$, returns a value $\overline{y}$ that depends on the example set $(x_i, y_i)_{i=1}^m$.
In the remaining of the article, it is shown how the method can be used to deal

[1] In the paper it is stated that the conditioning of the problem (sensibility of solutions to input variation, measured in spectral norm with $|\lambda_{max}|/|\lambda_{min}|$) is good for large *my*. Since $(m\gamma I + K)$ has the same eigenvalues of $K$ shifted by $m\gamma$, if $m\gamma \to \infty$, the ratio tends to 1, which is a desirable behaviour, but the numerical solution of the shifted problem has to deal with a matrix whose elements (extra-diagonal vs. diagonal) can have a difference of many orders of magnitude.

with different problems, obtaining results comparable with other classification methods.

The outlined method can be derived from Tikhonov regularization method, starting from the identification of the function $f$ minimizing:

$$\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2 \tag{4}$$

and adding a term that assures the problem is *well-posed* in Hadamard's sense[2]:

$$\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2+\gamma\|f\|^2. \tag{5}$$

In the case in which $y$ is binary and the *loss function* is $V(f(x,y)) = (f(x)-y)^2$, we obtain the *proximal SVM*, and with $V(f(x,y)) = (1-yf(x))_+$ *SVM* classification.

On the other side, in a recent technical report by Mangasarian [11], it is showed how to treat a binary classification problem:

$$\min_{w,\gamma\neq 0}\frac{\|Aw-\gamma\|}{\|Bw-\gamma\|}, \tag{6}$$

in which the solution is a pair of parallel hyperplanes separating $A$ and $B$ sets, as a generalized eigenvalue problem. Indeed, set:

$$G=[A \quad -e]'[A \quad -e], \quad H=[B \quad -e]'[B \quad -e], \quad z=[w' \quad \gamma]', \tag{7}$$

equation (6), becomes:

$$\min_{z\in \mathbb{R}^m}\frac{z'Gz}{z'Hz}, \tag{8}$$

the Raleigh quotient of generalized eigenvalue problem $Gx=\lambda Hx$. If we call $z_{min}=[w_1 \quad \gamma_1]$ and $z_{max}=[w_m \quad \gamma_m]$ the eigenvectors related to the eigenvalues of smallest and largest modulo, respectively, it follows that the distance of each point of $A$ from $x'w_1-\gamma_1=0$ is less that the one from $x'w_m-\gamma_m=0$ and, *mutatis mutandis*, the distance of each point of $B$ from $x'w_m-\gamma_m=0$ is less than $x'w_1-\gamma_1=0$.

Those hyperplanes are incident and therefore different from the ones obtained with SVM. Studies devoted to non separable problems [1, 2] still hold and it is simple to deduce a formulation in case of eigenvalues.

[2] Following Hadamard, the problem of finding solutions to the equations $f(x)=y$ is said to be *well-posed* provided solutions exist, are unique, and depend continuously on $n$.

**Example 1**
Let:

$$A = \begin{bmatrix} 2 & 0 \\ 2 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \tag{9}$$

Building $G$ and $H$ as in (7), we obtain:

$$G = \begin{bmatrix} 8 & 2 & -4 \\ 2 & 1 & -1 \\ -4 & -1 & 2 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 2 \end{bmatrix} \tag{10}$$

Smallest and largest eigenvalues of the problem $Gx = \lambda Hx$ are $\lambda_1 = 0$ and $\lambda_3 = \infty$ and the respective eigenvectors:

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

The resulting lines are $x = 2$ and $x - y = 0$, which precisely describe the starting sets $A$ and $B$.

**Example 1a**
Consider the same problem of Example 1 embedded in $\mathbb{R}^3$ and suppose points lay on the hyperplane $z = 0$, then the same result is obtained in terms of eigenvalues and eigenvectors, with the only difference that those ones will be:

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

Note $A$ and $B$ can be rank-deficient matrices, as in the previous examples. Matrices $G$ and $H$ are always rank-deficient, since the product of matrices of dimension $n + 1 \times n$ is of rank at least $n$, which introduces a difficulty from a numerical point of view, since we need to deal with an infinite eigenvalue.

A solution to this problem is proposed in [11][3], determining parameter $\delta$ in case of Tikhonov regularization of the problem:

$$\min_{w,\gamma \neq 0} \frac{\|Aw - \gamma\| + \delta\|z\|}{\|Bw - \gamma\| + \delta\|z\|}. \tag{11}$$

In fact, the regularized problem has the same eigenvectors of the starting problem *iff* $\delta = 1$[4]. In order to have the same eigenvectors, a shifting technique is needed, which transforms problem (11) in:

$$\min_{w,\gamma \neq 0} \frac{\|Aw - \gamma\| + \delta\|Bz\|}{\|Bw - \gamma\| + \delta\|Az\|}. \tag{12}$$

This is a particular instance of the problem to determine a not degenerate matrix pair $G_1$ and $H_1$ such that $G_1 y = \mu H_1 y$ has the same eigenvectors of the starting problem. It is possible to prove the following theorem:

[3] In a more recent version of the same report a slightly different approach is proposed

[4] Problem $(A + \delta)x = \lambda(B + \delta)x$ has the same eigencouples $(\overline{\lambda}, \overline{x})$ of the starting problem *iff* $A\overline{x} + \delta\overline{x} = \overline{\lambda}B\overline{x} + \overline{\lambda}\delta\overline{x}$, i.e. *iff* $\delta\overline{x} = \overline{\lambda}\delta\overline{x}$, which is equivalent to $\delta = 1$.

**Theorem 2.1** *Consider the generalized eigenvalue problem $Gx = \lambda Hx$ and the transformed $G_1x = \lambda H_1x$ defined by:*

$$G_1 = \tau_1 G - \sigma_1 H, \quad H_1 = \tau_2 H - \sigma_2 G, \tag{13}$$

*for each choice of scalars $\tau_1$, $\tau_2$, $\sigma_1$ and $\sigma_2$, such that the $2 \times 2$ matrix*

$$\Omega = \begin{pmatrix} \sigma_2 & \tau_1 \\ \tau_2 & \sigma_1 \end{pmatrix} \tag{14}$$

*is nonsingular. Then the problem $G_1x = \lambda H_1x$ has the same eigenvectors of the problem $Gx = \lambda Hx$.*

**Proof 2.1** *See [14], p. 288.*

Equation 12 is obtained applying previous theorem with $\sigma_1 = \sigma_2 = -\delta$ e $\tau_1 = \tau_2 = 1$.

**Example 2**
Let:

$$G_1 = G + 8 * H, \quad H_1 = H + 2 * G,$$

we have:

$$G_1 = \begin{bmatrix} 16 & 10 & -12 \\ 10 & 9 & -9 \\ -12 & -9 & 18 \end{bmatrix}, \quad H_1 = \begin{bmatrix} 17 & 5 & -9 \\ 5 & 3 & -3 \\ -9 & -3 & 6 \end{bmatrix} \tag{15}$$

$G_1$ and $H_1$ are not degenerate and the smallest eigenvalue is transformed in $\mu_1 = 0.5$, while the largest in $\mu_3 = 8$.

# 3 Preliminary results

As it has been shown in the previous section, the problem of determining costants in (14), so that $det(\Omega) \neq 0$, still holds. A choice can be:

$$\sigma_1 = max(diag(G)), \quad \sigma_2 = max(diag(H)). \tag{16}$$

and $\tau_1 = \tau_2 = 1$.

In this way, if $\sigma_1\sigma_2 \neq 1$, matrix $\Omega$ is not degenerate.

If $\beta G + \alpha H$ happens to be singular for every $\alpha$ and $\beta$, the probability matrices $G_1$ and $H_1$ are singular is null. Indeed, if the number $m$ of points (rows) of each matrix is greater then the dimension $n$ of the space of characteristics, and those matrices have full rank, resulting matrices will be rank deficient *iff* $G$ and $H$ have the same null space, i.e. *iff* $\exists \overline{x} : G\overline{x} = 0 \wedge H\overline{x} = 0$, which can happen with probability 0. That can be empirically verified with a simple Matlab macro:

```
for i=10:110
   for j=1:1000
      a=rand(i,i);
      b=rand(i,i);
      G=[a -ones(i,1)]'*[a -ones(i,1)];
      H=[b -ones(i,1)]'*[b -ones(i,1)];
      G1=G+max(diag(G))*H;
      H1=H+max(diag(H))*G;
      if (rank(G1) + rank(H1) < 2*(i +1))
         G, H;
      end
   end
end
```

The macro generates 100.000 random matrices of variable dimension between 10 and 110 and it checks matrices $G_1$ and $H_1$ are not deficient. The 100.000 tests that have been carried out have not produced a degenerate matrix.

A problem arises when $G$ and $H$, which are singular by definition, have a nontrivial intersection of their null spaces, i.e. when $m \ll n$. In this particular case, it is possible to project the operators in the complement of such intersection, which can be done provided we can compute a basis of the common null space, a task that can be computationally unfeasible for large matrices.

In short, whatever is the applied regularization technique, *the problem is solvable with eigenvalue based techniques if and only if input matrices $A$ and $B$ have maximum rank*, i.e. if there is a number of independent input points for learning at least equal to the number of characteristics.

This regularization technique, which derives from Theorem 3.1, has been applied to the data set generated by NDC (http://www.cs.wisc.edu/dmi/svm/ndc/). Using the macros made available by the authors Musicant and Mangasarian, 300 points with 7 characteristics have been produced, divided by the generator in two classes of 156 and 144 points each. For the learning phase, a subset of 30 points have been used to determine the hyperplanes describing such sets. The method has correctly classified 87,6% of cases, which is comparable with 86,7% obtained with Tikhonov regularization.

Now it is worthwhile noting that it is possible to generalize the binary classification problem to an n-ary one in a very simple way. Let's start with a few examples.

**Example 3**
Consider the following data sets:

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix}, \quad C = \begin{bmatrix} 3 & 1 \\ 4 & 2 \end{bmatrix}, \tag{17}$$

Starting from $A$ and $B$, evaluate $G$ and $H$:

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

with $e$ unit vector in $\mathbb{R}^2$:

$$G = \begin{bmatrix} 5 & 5 & -3 \\ 5 & 5 & -3 \\ -3 & -3 & 2 \end{bmatrix}, \quad H = \begin{bmatrix} 13 & 8 & -5 \\ 8 & 5 & -3 \\ -5 & -3 & 2 \end{bmatrix},$$

Computing $G_1 = G + max(diag(G)) * H$ e $H_1 = H + max(diag(H)) * G$, we obtain:

$$G_1 = \begin{bmatrix} 70 & 45 & -28 \\ 45 & 30 & -18 \\ -28 & -18 & 12 \end{bmatrix}, \quad H_1 = \begin{bmatrix} 78 & 73 & -44 \\ 73 & 70 & -42 \\ -44 & -42 & 28 \end{bmatrix},$$

from which $\lambda_{max} = 5$, $\lambda_{min} = 0.0769$ and the respective eigenvectors:

$$x_{max} = [-1, 1, 0], \quad x_{min} = [1, -1, 1],$$

which describe the lines $x - y = 0$ and $x - y = 1$ to which points in $A$ and $B$ belong.
Repeating the same procedure with $C$ and $B$ we obtain:

$$x_{max} = [1, -1, 2] \quad x_{min} = [1, -1, 1],$$

while with $A$ and $C$ we have

$$x_{max} = [-1, 1, 0] \quad x_{min} = [1, -1, 2].$$

What we can infer from the previous example is $x_{max}$ eigenvector depends on $A$ and $x_{min}$ on $B$. From those observations it is possible to deduce a simple algorithm for n-ary classification. Let $A_i,\ i = 0,\ k - 1$ be matrices of separable points of dimension $m \times n \times k$.

```
for i=0:k-2
   G=[A(:,:,i) -ones(m,1)]'*[A(:,:,i) -ones(m,1)];
   H=[A(:,:,i+1) -ones(m,1)]'*[A(:,:,i+1) -ones(m,1)];
   G1=G+max(diag(G))*H;
   H1=H+max(diag(H))*G;
   [Lambda, x] = eig[G1, H1] # Lambda of largest modulo
end
```

# 4 Algorithms for eigenvalue problems

At the end of 90s there has been a wide effort devoted to the implementation of algorithm for the efficient computation of eigenvectors corresponding to extremal eigenvalues of large, sparse and symmetric matrices. Among those there is the method proposed by Lancozs (see, for example [3]), which uses a projection of the operator $M$ on a *Krylov* subspace, i.e. the space spanned by the vectors computed by the power method:

$$\mathcal{K}_k(M, v) = span\{v, Mv, \dots, M^{k-1}v\}.$$

In such subspace the operator assumes a tridiagonal form:

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{k-1} \\ & & & \beta_{k-1} & \alpha_k \end{bmatrix}.$$

It is possible to prove, for increasing values of $k$, $T$ extremal eigenvalues are increasingly better approximations by defect of extremal eigenvalues of matrix $M$.

Since $T$ is tridiagonal, and of much smaller dimension with respect to $M$, it is possible to iteratively solve huge problems fast and accurately.
What can seem a trivial generalization of a slow algorithm, results in one among the most successful methods for the solution of such problems, especially in the hermitian case. Furthermore, Lanczos method can be easily extended to generalized problems.

In order to obtain a sufficient computational power to solve huge problems in a short period of time, distributed memory multiprocessors, such as clusters of workstations, need to be used.

Lanczos algorithm, in its initial formulation, is not suited for an efficient implementation for that class of computers. For this reason, it has been taken into account its block formulation, which uses a block of $s$ vectors for the projection. It has been proposed a new parallel block algorithm for the target architecture, in which processors are logically configured as a ring. It is based on a *block-column wrap around* data distribution among processors, which consists, in case of a matrix, in cyclically assigning blocks of columns of equal size to each processor. Such data decomposition is well suited for the underlying logical topology.

In accordance with such distribution, an efficient parallelization of basic linear algebra computational kernels, such as dense matrices products and QR factorizations, are used.

Performance results confirm the benefits of the proposed approach; results have been firstly reported at Cornelius Lanczos International Conference [6], and then in [8]. In [9], it is shown how to deal with variable sparsity of the matrices.

In [7] a parallel algorithm for the hermitian case is proposed. It can be implemented on a multicomputer logically configured with a mesh topology. It can benefit from the use of *de facto* standard parallel computational kernels, as the ones can be found in ScaLapack (http://www.netlib.org/scalapack), which is an ongoing project at UTK for the development of algorithms and software for numerical parallel dense linear algebra problems. Moreover, an algorithm is proposed for the nonsymmetric case, which is based on a block oblique projection over a Krylov subspace.

The implementation of the modified block Lanczos algorithm, and its performance evaluation have been presented in [10].

# 5 Open problems

Well-posedness of a problem does not imply well conditioning. On the other hand, the determination of a problem formulation in which the function mapping input data to solutions is Lipshitz, with a small constant, assures less sensitivity of the problem to input data perturbation and better computed solutions. At a first glance, formulations involving eigenvalue problems have an advantage: conditioning of eigenvectors depends on the distance of the relative eigenvalue from the rest of the spectrum. Since we are interested in extremal eigenvalues, techniques already exist to enhance that characteristic.
The previous considerations motivate the identification of a link between eigenvalue problems, classification problems and supervised learning theory. The problem can be formulated as follows:

*1. Is it possible to find a connection, as it has been done with binary classification and generalized eigenvalue problems, between Smale Poggio theory on supervised learning and eigenvalue problems?*

In classification problems there can be some characteristics that influence the process more then others, and others that don not have any effect at all on computed solutions. The problem of identifying main characteristics can be stated as:

*2. Is it possible to determine a solution of a classification problem minimizing the dimension of characteristics space in which the solution is analyzed?*

This problem has been stated and solved in [4, 12], giving a method that discerns, in case of medical diagnosis, the examinations that permit to classify patients in healthy and ill, automatically detecting those not influencing the diagnosis. That is achieved requiring the solution (descriptor vector of separating hyperplanes) to have as many zero components as possible.

If we consider Examples 1 and 1a, it is clear that if the problem is projected in a space of greater dimension, eigenvector components related to the added dimensions will still be zero. Therefore, with respect to starting problems, useless characteristics would not be taken into account in solution vectors. This leads to suppose that a formulation based on eigenvalues has, for its nature, the characteristic to determine the optimal subspace in which the initial problem has to be projected to be solved, without loss of information.

Another question regards the minimum number of examples needed to obtain a certain kind of learning. Often, the number of available examples for training is small and we want to make a decision from a few examples, which is somehow similar to the human cognitive process.

*3. What can be said on the quality of a computed solution, starting from the knowledge at hand? If data are available, for which a solution is not known, how can they be used?*

Krylov subspace methods determine a subspace in which the projection of the operator maintains information about the eigenvectors of the starting operator related to extremal eigenvalues and, as the dimension grows, increasingly better approximation can be computed. That characteristic could be used to find the minimum number of examples needed to obtain a certain classification, using information provided from all input data, and not only from a part that has been chosen in some way. This feature can be introduced in the proposed algorithm, obtaining an iterative procedure.

# 6 Conclusions and future work

Research activities related to mathematical models of learning have an important role in the near future.

They find application both in the filtering of *multimensional media*, with respect to analysis, classification, segmentation of media, and in the development of *global knowledge and ubiquitous services*, in the case of learning techniques in the field of knowledge discovery and distribution. A schedule for a short term activity could be:

- Verification of the possibility of reformulating Smale Poggio theory in terms of eigenvalue/eigenvectors.
  Evaluation of gains in terms of conditioning and computational complexity.

- Testing of generalized eigenvalue problems techniques to discern the importance of every characteristic in problem solution.
- Testing of iterative projection methods with respect to quality assesment of computed solutions.

# References

[1] K. P. Bennett and O. L. Mangasarian Robust Linear Programming Discrimination of Two Linearly Inseparable Sets. Computer Sciences Technical Report 1054a, 1991. Optimization Methods and Software 1, 23-34, 1992.

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

[3] G. H. Golub and C. F. Van Loan, Matrix Computations, 2nd ed., Johns Hopkins University Press, Baltimore, 1989.

[4] P. S. Bradley, O. L. Mangasarian and W. N. Street, Feature Selection via Mathematical Programming. Mathematical Programming Technical Report 95-21, December 1995. INFORMS Journal on Computing 10, 209-217, 1998.

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

[6] M.R. Guarracino, F. Perla - A Parallel Version of a Block Lanczos' Algorithm for Distributed Memory Architectures - Cornelius Lanczos Centenary Conference, Raleigh, 1993.

[7] M.R. Guarracino, Metodi Numerici per il Calcolo degli Autovalori per Matrici Sparse e di Elevate Dimensioni in Ambiente Parallelo - PhD Thesis, 1996.

[8] M.R. Guarracino, F. Perla - A Parallel Block Lanczos' Algorithm for Distributed Memory Architectures, Parallel Algorithms and Applications vol.4 n. 1-2, 1995.

[9] M.R. Guarracino, F. Perla A Parallel Modified Block Lanczos' Algorithm for Distributed Memory Architectures - in Proceedings of 3rd Euromicro Workshop on Parallel and Distributed Processing, IEEE pub, pp. 424-431, 1995.

[10] M.R. Guarracino HPEC: High Performance Eigenvalue Computation, a software for the evaluation of large sparse eigenvalue problems - communication at Parallel Matrix Algorithms and Applications, Marsiglia, 2004.

[11] O. L. Mangasarian and E. W. Wild Multisurface Proximal Support Vector Classification via Generalized Eigenvalues Data Mining Institute Technical Report 04-03, June 2004.

[12] O. L. Mangasarian Machine Learning via Polyhedral Concave Minimization. Mathematical Programming Technical Report 95-20, November 1995. "Applied Mathematics and Parallel Computing – Festschrift for Klaus Ritter", H. Fischer, B. Riedmueller, S. Schaeffler, editors, Physica-Verlag, Germany, 175-188, 1996.

[13] T. Poggio and S. Smale The Mathematics of Learning: Dealing with Data Amer. Math. Soc. Notice 537-544, 2003.

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