Optimization Methods and Software
Vol. 00, No. 00, Month 200x, 1–12

A Classification Method based on Generalized Eigenvalue Problems

Mario R. Guarracino†, Claudio Cifarelli‡, Onur Seref§, Panos M. Pardalos§

† High Performance Computing and Networking Institute,
National Research Council, Italy
‡ Department of Statistic, Probability and Applied Statistics,
University of Rome "La Sapienza", Italy
§ Center for Applied Optimization,
University of Florida, Gainesville, FL, 32611-6595 USA
(Received 00 Month 200x; In final form 00 Month 200x)

Binary classification refers to supervised techniques that split a set of points in two classes, with respect to a training set of points whose membership is known for each class. Binary classification plays a central role in the solution of many scientific, financial, engineering, medical and biological problems. Many methods with good classification accuracy are currently available. This work shows how a binary classification problem can be expressed in terms of a generalized eigenvalue problem. A new regularization technique is proposed, which gives results that are comparable to other techniques in use, in terms of classification accuracy. The advantage of this method relies in its lower computational complexity with respect to the existing techniques based on generalized eigenvalue problems. Finally, the method is compared with other methods using benchmark data sets.

1 Introduction

Supervised learning refers to the capability of a system to learn from a set of examples, which is a set of input/output couples. This set is called the *training set*. The trained system is able to provide an answer (output) for a new question (input). The term *supervised* originates from the fact that the desired output for the training set of points is provided by an external teacher.

Supervised learning systems can find applications in many fields. A bank prefers to classify customer loan requests as "good" or "bad" depending on their ability to pay back. The Internal Revenue Service tries to discover tax evaders starting from the characteristics of known ones. As another example, a built-in system in a car could detect if a walking pedestrian is going to cross the street. There are many applications in biology and medicine. The tissues

that are prone to cancer can be detected with high accuracy, or the new DNA sequences or proteins can be tracked down to their origins. Given its amino acids sequence, finding how a protein folds provides important information on its expression level. More examples related to numerical interpolation, handwriting recognition and Montecarlo methods for numerical integration can be found, for example, in [4, 6].

Support Vector Machine (SVMs) algorithms [24] are the state-of-the-art for the existing classification methods. These methods classify the points from two linearly separable sets in two classes by solving a quadratic optimization problem in order to find the optimal separating hyperplane between these two classes. This hyperplane maximizes the distance from the convex hulls of each class. These techniques can be extended to the nonlinear cases by embedding the data in a nonlinear space using *kernel functions* [21].

SVMs have been one of the most successful methods in supervised learning with applications in a wide spectrum of research areas, ranging from pattern recognition [10] and text categorization [8] to biomedicine [11, 14], brain-computer interface [7, 23], and financial applications [22, 26]. The robustness of SVMs originates from the strong fundamentals of statistical learning theory [24]. The training part relies on optimization of a quadratic convex cost function. Quadratic programming (QP) is an extensively studied field of mathematics and there are many general purpose methods to solve QP problems such as quasi-newton, primal-dual, and interior-point methods. The general purpose methods are suitable for small size problems, whereas for large problems chunking [15] and decomposition [17] methods use subsets of points to optimize SVMs. SVM-Lite [9] and LIBSVM [5] are among the most preferred implementations that use chunking and decomposition methods efficiently. There are also efficient algorithms that exploit the special structure of the optimization problem such as Generalized Proximal SVMs (GEPSVM) [12].

The binary classification problem can be formulated as a generalized eigenvalue problem [12]. This formulation differs from SVMs since, instead of finding one hyperplane that separates the two classes, it finds two hyperplanes that approximate the two classes. The prior study requires the solution of two different eigenvalue problems. The aim of this work is to present *Regularized General Eigenvalue Classifier* (ReGEC), a classification method that uses a new regularization technique for the solution of the underlying generalized eigenvalue problem. Our work differs from original method by the fact that the regularization technique we use permits us to solve only one eigenvalue problem instead of two. Thus, our method halves the execution time compared to the original method and provides comparable accuracy results.

The notation used in the paper is as follows. 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 $\|x\|$ and

the unit vector will be denoted by e.
The remainder of the the paper is organized as follows. Section 2 describes how the generalized eigenvalue classifier differs from the generic SVM methods. In Section 3 regularization technique is presented. In Section 4, numerical experiments are reported, and finally, in Section 5, conclusions are drawn and future work is proposed.

2 Related work

SVM algorithm for classification consists of finding a hyperplane that separates the elements belonging to two different classes. The separating hyperplane is usually chosen to maximize the margin between the two classes. The margin can be defined as the minimum distance between the separating hyperplane and the points of either class. The optimum hyperplane is the one that maximizes the margin. The points that are closest to the hyperplane are called *support vectors*, and are the only points needed to train the classifier. Consider two matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{k\times m}$, that represent the two classes, each row being a point in the feature space. The quadratic linearly constrained problem to obtain the optimal hyperplane (w, b) is:

$$\begin{aligned} & \min f(w) = \frac{w'w}{2} \\ s.t.\ & (Aw + b) \geq e \\ & (Bw + b) \leq -e. \end{aligned} \tag{1}$$

Mangasarian et al. [12] proposes to classify these two sets of points A and B using two hyperplanes, each closest to one set of points, and furthest from the other. Let $x'w - \gamma = 0$ be a hyperplane in $\mathbb{R}^m$. In order to satisfy the previous condition for the points in A, the hyperplanes can be obtained by solving the following optimization problem:

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

The hyperplane for the B can be obtained by minimizing the inverse of the objective function in (3). Now, let

$$G = [A \quad -e]'[A \quad -e], \quad H = [B \quad -e]'[B \quad -e], \quad z = [w' \quad \gamma]', \tag{3}$$

Figure 1. Separation obtained with generalized eigenvectors.

then equation (2), becomes:

$$\min_{z \in \mathbb{R}^m} \frac{z'Gz}{z'Hz}. \tag{4}$$

The expression in (4) is the Raleigh quotient of the generalized eigenvalue problem $Gx = \lambda Hx$. The stationary points are obtained at and only at the eigenvectors of (4), where the value of the objective function is given by the eigenvalues. When H is positive definite, the Raleigh quotient is bounded and it ranges over the interval determined by minimum and maximum eigenvalues [16]. H is positive definite under the assumption that the columns of $[B \quad -e]$ are linearly independent. The inverse of the objective function in (4) has the same eigenvectors and reciprocal eigenvalues. Let $z_{min} = [w_1 \quad \gamma_1]$ and $z_{max} = [w_2 \quad \gamma_2]$ be the eigenvectors related to the eigenvalues of smallest and largest modulo, respectively. Then $x'w_1 - \gamma_1 = 0$ is the closest hyperplane to the set of points in A and the furthest from those in B and $x'w_2 - \gamma_2 = 0$ is the closest hyperplane to the set of points in B and the furthest from those in A. This is depicted in the examples shown in figure 2.

A standard technique in SVMs to obtain a greater separability between sets is to embed the points into a nonlinear space, via kernel functions. In this work we use the *Gaussian kernel*,

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{\sigma}}. \tag{5}$$

In (5), x_i and x_j denote two points in the feature space. This technique usually allows one to obtain better results, as shown in several applications. Results regarding nonlinearly separable problems [1, 2] still hold and a formu-

lation for the eigenvalues problem can easily be derived. This formulation is given in the next section.

3 Regularized method

Recall that A and B are the matrices containing the two classes of training points, with each row representing a point in the future space. Let G and H be as defined in (3). Note that even if A and B are full rank, matrices G and H are always rank-deficient. The reason is that G and H are matrices of order $m+1$, and their rank can be at most m. The added complexity due to singularity of the matrices means that special care has to be given to the solution of the generalized eigenvalue problem. Indeed, if the null spaces of G and H have a nontrivial intersection, i.e. $Ker(A) \bigcap Ker(B) \neq 0$, then the problem is singular and a regularization technique is needed to solve the eigenvalue problem.

Mangasarian et al. proposes to use Tikhonov regularization applied to a two-fold problem:

$$\min_{w,\gamma \neq 0} \frac{\|Aw - e\gamma\|^2 + \delta\|z\|^2}{\|Bw - e\gamma\|^2}, \tag{6}$$

and

$$\min_{w,\gamma \neq 0} \frac{\|Bw - e\gamma\|^2 + \delta\|z\|^2}{\|Aw - e\gamma\|^2}, \tag{7}$$

where δ is the regularization parameter and the new problems are still convex. The minimum eigenvalues-eigenvectors of these problems are approximations of the minimum and the maximum eigenvalues-eigenvectors of equation (4). The solutions $(w_i, \gamma_i), i = 1, 2$ to (6) and (7) represent the two hyperplanes approximating the two classes of training points.

In practice, if $\beta G - \alpha H$ is nonsingular for every α and β, it is possible to transform the problem into another problem that is nonsingular and that has the same eigenvectors of the initial one. We start with the following theorem whose proof can be found in [20], p. 288.

THEOREM 3.1 *Consider the generalized eigenvalue problem* $Gx = \lambda Hx$ *and*

the transformed $G^*x = \lambda H^*x$ *defined by:*

$$G^* = \tau_1 G - \delta_1 H, \quad H^* = \tau_2 H - \delta_2 G, \tag{8}$$

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

$$\Omega = \begin{pmatrix} \tau_2 & \delta_1 \\ \delta_2 & \tau_1 \end{pmatrix} \tag{9}$$

is nonsingular. Then the problem $G^*x = \lambda H^*x$ *has the same eigenvectors of the problem* $Gx = \lambda Hx$. *An associated eigenvalue* λ^* *of the transformed problem is related to an eigenvalue* λ *of the original problem by*

$$\lambda = \frac{\tau_2 \lambda^* + \delta_1}{\tau_1 + \delta_2 \lambda^*}.$$

In the linear case Theorem 3.1 can be applied. By setting $\tau_1 = \tau_2 = 1$ and $\hat{\delta_1} = -\delta_1$, $\hat{\delta_2} = -\delta_2$, the regularized problem becomes

$$\min_{w,\gamma \neq 0} \frac{\|Aw - e\gamma\|^2 + \hat{\delta_1}\|Bw - e\gamma\|^2}{\|Bw - e\gamma\|^2 + \hat{\delta_2}\|Aw - e\gamma\|^2}. \tag{10}$$

If $\hat{\delta_1}$and $\hat{\delta_2}$ are non negative, Ω is non-degenerate. The spectrum is now shifted and inverted so that the minimum eigenvalue of the original problem becomes the maximum of the regularized one, and the maximum becomes the minimum eigenvalue. Choosing the eigenvectors related to the new minimum and maximum eigenvalue, we still obtain the same ones of the original problem.

This regularization works for the linear case if we suppose that in each class of the training set there is a number of linearly independent rows that is at least equal to the number of the features. This is often the case and, since the number of points in the training set is much greater than the number of features, $Ker(G)$ and $Ker(H)$ have both dimension 1. In this case, the probability of a nontrivial intersection is zero.

In the nonlinear case the situation is different. Using the kernel function (5), each element of the kernel matrix is

$$K(A, B)_{i,j} = e^{-\frac{\|A_i - B_j\|^2}{\sigma}}. \tag{11}$$

Let

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

then, problem (2) becomes:

$$\min_{u,\gamma \neq 0} \frac{\|K(A,C)u - e\gamma\|^2}{\|K(B,C)u - e\gamma\|^2}. \tag{12}$$

Now the associated eigenvalue problem has matrices of order $n + k + 1$ and rank at most m. This means a regularization technique is needed, since the problem can be singular.
We propose to generate the following two proximal surfaces:

$$K(x,C)u_1 - \gamma_1 = 0, \quad K(x,C)u_2 - \gamma_2 = 0 \tag{13}$$

by solving the following problem

$$\min_{u,\gamma \neq 0} \frac{\|K(A,C)u - e\gamma\|^2 + \delta\|\tilde{K}_B u - e\gamma\|^2}{\|K(B,C)u - e\gamma\|^2 + \delta\|\tilde{K}_A u - e\gamma\|^2} \tag{14}$$

where $\tilde{K_A}$ and $\tilde{K_B}$ are diagonal matrices with the diagonal entries from the matrices $K(A,C)$ and $K(B,C)$. The perturbation theory of eigenvalue problems [25] provides an estimation of the distance between the original and the regularized eigenvectors. If we call z an eigenvector of the initial problem and $z(\delta)$ the corresponding one in the regularized problem, then $|z - z(\delta)| = \mathcal{O}(\delta)$, which means their closeness is in the order of δ.

As mentioned in the previous section, the minimum and the maximum eigenvalues obtained from the solution of (14) provide the proximal planes P_i, $i = 1, 2$ to classify the new points. A point x is classified using the distance

$$dist(x, P_i) = \frac{\|K(x,C)u - \gamma\|^2}{\|u\|^2}. \tag{15}$$

and the class of a point x is determined as

$$class(x) = argmin_{i=1,2}\{dist(x, P_i)\}. \tag{16}$$

In the next section we present comparisons of accuracy and speed of the

proposed method to the original generalized proximal classifier as well as the widely used SVMs implementations.

4 Numerical results

The aforementioned methods have been tested on benchmark data sets publicly available. Results regard their performance in terms of classification accuracy and execution time. We used data from different repositories: UCI repository [3], Odewahn et al. [19], and IDA repository [18]. These repositories are widely used to compare the performance of new algorithms to the existing methods. The results regarding the linear kernel have been obtained using the first two repositories. The third one has been used in the non-linear kernel implementation. For each data set, the latter repository offers 100 predefined random splits into training and test sets. For several algorithms, results obtained from each trial, including SVMs, are recorded. The accuracy results for the linear kernel SVMs and GEPSVM are taken from [12] and for the non linear kernel from [18]. Execution times and the other accuracy results have been calculated using an Intel Xeon CPU 3.20GHz, 6GB RAM running Red Hat Enterprise Linux WS release 3 with Matlab 6.5, during normal daylight operations. Matlab function *eig* for the solution of the generalized eigenvalue problem has been used for GEPSVM and ReGEC. The latest releases for LIBSVM [5] and SVMlight [9] have been used to compare these methods with SVMs.

In tables 1 and 2, classification accuracy using linear and gaussian kernels have been evaluated. Tables columns represent: data set name, the number of elements in the training set (n+k), the number of elements in the test set and the accuracy results for ReGEC, GEPSVM and SVMs. In table 1, the accuracy results have been evaluated using ten fold cross validation. In table 2, the random splits of IDA repository have been used. In the linear case comparable accuracy results have been obtained by the three methods. Using the gaussian kernel, ReGEC and GEPSVM show similar behavior yielding always results slightly lower than SVMs.

In tables 3 and 4, elapsed time is reported. In the linear case ReGEC and GEPSVM outperform SVMs implementations (LIBSVM and SVM light) in all cases. Furthermore ReGEC is at least twice faster then GEPSVM. When the gaussian kernel is used, SVMs implementations achieve better performances with respect to the eigenvalues based methods. In all cases, ReGEC is faster than GEPSVM.

Table 1. Classification accuracy using linear kernel.

dataset	n+k	test	m	ReGEC	GEPSVM	SVM
NDC	300	30	7	87.60	86.70	89.00
Cleveland Heart	297	30	13	86.05	81.80	83.60
Pima Indians	768	77	8	74.91	73.60	75.70
Galaxy Bright	2462	240	14	98.24	98.60	98.30

Table 2. Classification accuracy using 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

Table 3. Elapsed time in seconds using linear kernel.

dataset	ReGEC	GEPSVM	LIBSVM	SVM ligh
NDC	0.1e-03	0.2e-03	0.8991	22.002
Cleveland Heart	1.92e-04	3.58e-04	9.90e-03	0.3801
Pima Indians	1.21e-04	2.36e-04	15.8737	48.8092
Galaxy Bright	0.3e-3	0.5e-3	1.2027	21.128

Finally, a graphical representation of the classification surfaces obtained by ReGEC, GEPSVM and SVMs is given in figure 4 relatively to Banana dataset. The three methods show similar class regions. SVMs obtain smoother borders and more regular regions. These differences depend upon the fact that in SVMs the surfaces are characterized by the support vectors and the penalties terms, while in the eigenvalues methods all the points contribute to the solution surfaces. This behavior depends on the fact that eigenvalues methods always

Table 4. Elapsed time in seconds using gaussian kernel.

dataset	**ReGEC**	**GEPSVM**	**LIBSVM**	**SVM light**
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

Figure 2. Separation surfaces obtained with ReGEC, GEPSVM and LIBSVM on Banana dataset.

maximize the classification accuracy on the training set with respect to kernel and regularization parameters.

5 Conclusions and future work

Research activities related to supervised learning have an important role in many scientific and engineering applications. In the present work a novel regularization technique and its application has been proposed and tested against other methods on a number of datasets. Results show that the proposed method *i*) has a classification accuracy comparable to other methods, *ii*) has a computational performance comparable to most of the other methods, and *iii*) is much faster then the others in the linear case.

In the last years there has been a wide effort devoted to the implementation of algorithms for the efficient computation of eigenvectors corresponding to extremal eigenvalues of large, sparse and symmetric matrices on distributed

memory multiprocessors (see for example [13]). Therefore eigenvalue based techniques are attractive for the classification of very large sparse data sets. Future work will regard the verification and comparison of the proposed classification method on large data sets with respect to other methods using high performance computers.

References

[1] K. Bennet and C. Campbell, 2000, Support vector machines: Hype or hallelujah? *SIGKDD Explorations*, 2(2):1–13.

[2] K. Bennett and O. Mangasarian, 1992, Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–34.

[3] C.L. Blake and C.J. Merz, 1998, Uci repository of machine learning databases. www.ics.uci.edu/∼mlearn/MLRepository.html.

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

[5] C.J. Lin C.W. Hsu, C.C. Chang, 2004, A practical guide to support vector classification. http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf.

[6] F. Giannessi, 1982, Complementarity problems and their applications to structural engineering. In Pitagora, editor, *Methods and algorithms for optimization*, 507–514, Bologna.

[7] T. Ebrahimi G.N. Garcia and J.M. Vesin, 2003, Joint time-frequency-space classification of eeg in a brain-computer interface application. *Journal on Applied Signal Processing*, pages 713–729.

[8] T. Joachims, 1998,Text categorization with support vector machines: Learning with many relevant features. In Claire Ndellec and Cline Rouveirol, editors, *Proceedings of the European Conference on Machine Learning*, 137–142, Berlin.

[9] T. Joachims, 1999, Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning. MIT-Press.*

[10] S. Lee and A. Verri, 2002, Pattern recognition with support vector machines, In *SVM 2002*, Niagara Falls, Canada, Springer.

[11] D. Lin N. Cristianini C. Sugne T. Furey M. Ares M. Brown, W. Grundy and D. Haussler, 2000, Knowledge-base analysis of microarray gene expressiondata by using support vector machines. *PNAS*, 97(1):262–267.

[12] O. L. Mangasarian and E. W, 2004, Wild. Multisurface proximal support vector classification via generalized eigenvalues. Technical Report 04-03, Data Mining Institute, September.

[13] F. Perla M.R. Guarracino. 1995, A parallel block lanczos' algorithm for distributed memory architectures. *Parallel Algorithms and Applications*, 4(1-2).

[14] W. S. Noble, 2004, *Kernel Methods in Computational Biology*, chapter Support vector machine applications in computational biology, pages 71–92. MIT Press.

[15] R. F. E. Osuna and F. Girosi, 1997, An improved training algorithm for support vector machines. In *IEEE Workshop on Neural Networks for Signal Processing*, 276–285.

[16] B. N. Parlett. 1998, The Symmetric Eigenvalue Problem, 357. *SIAM*, Philadelphia,PA.

[17] J. Platt, 1999, Fast training of SVMs using sequential minimal optimization, *Advances in Kernel Methods: Support Vector Learning*, 185–208. MIT press, Cambridge, MA.

[18] J.Weston B. Schlkopf S. Mika, G. Rtsch and K. R. Mller. 1999, Fisher discriminant analysis with kernels. *IEEE Neural Networks for Signal Processing*, IX:41–48.

[19] R. Pennington R. Humphreys S. Odewahn, E. Stockwell and W. Zumach, 1992, Automated star/galaxy discrimiation with neural networks. *Astronomical Journal*, 103(1):318–331.

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

[21] J. Shawe-Taylor and N. Cristianini. 2004, *Kernel methods for pattern analysis.* Cambridge University Press, Cambridge, UK.

[22] H. Ince T. B. Trafalis, 2002, Support vector machine for regression and applications to financial forecasting. In *International Joint Conference on Neural Networks (IJCNN'02)*, Como, Italy. IEEE-INNS-ENNS.

[23] T. Hinterberger J. Weston M. Bogdan N. Birbaumer T. N. Lal, M. Schroeder and B. Schlkopf, 2004, Support vector channel selection in bci. *IEEE Transactions on Biomedical Engineering*, 51(6):1003–1010.

[24] V. Vapnik, 1995, *The Nature of Statistical Learning Theory*. Springer-Verlag.
[25] J. Wilkinson, 1965, *The Algebraic Eigenvalue Problem*. Clarendon Press.
[26] C. J. Hsu W. H. Chenb S. Wuc Z. Huang, H. Chen, 2004, Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37:543–558.