On the regularization of generalized eigenvalues classifiers

Mario R. Guarracino' , Mara Sangiovanni' , Gerardo Severino' , Gerardo Toraldo' , and Marco Viola'

Citation: **1776**, 040005 (2016); doi: 10.1063/1.4965317
View online: http://dx.doi.org/10.1063/1.4965317
View Table of Contents: http://aip.scitation.org/toc/apc/1776/1
Published by the American Institute of Physics

On the Regularization of Generalized Eigenvalues Classifiers

Mario R. Guarracino[1,b)], Mara Sangiovanni[1,a)], Gerardo Severino[2,c)], Gerardo Toraldo[3,d)] and Marco Viola[4,e)]

[1]*High Performance Computing and Networking Institute, National Research Council of Italy - Naples, Italy*
[2]*University of Naples Federico II Division of Agricultural, Forest and Biosystems Engineering - Naples, Italy*
[3]*University of Naples Federico II, Department of Mathematics and Applications - Naples, Italy*
[4]*Sapienza University of Rome, Department of Computer, Control and Management Engineering - Rome, Italy*

[a)]Corresponding author: mara.sangiovanni@icar.cnr.it
[b)]mario.guarracino@cnr.it
[c)]gerardo.severino@unina.it
[d)]toraldo@unina.it
[e)]marco.viola@uniroma1.it

Abstract. Generalized Eigenvalues Classifiers (GEC), which originated from the GEPSVM algorithm by Mangasarian, proved to be an efficient alternative to the Support Vector Machines (SVMs) in the solution of supervised classification tasks. However real-life datasets are often characterized by a large number of redundant features and by a great number of points whose labels are difficult (or too expensive) to assign. In this work we start from the Regularized Generalized Eigenvalue Classifier (ReGEC) and show how regularization terms can be used to enable the classifier to solve two different problems, strictly connected to that of supervised classification: feature selection and semi-supervised classification. Numerical results, obtained on some standard benchmark data sets, show the efficiency of the proposed solutions.

INTRODUCTION

Supervised classification represents one of the most used techniques in machine learning. The goal is to train a model from available data, which are labeled as belonging to an *a-priori* known number of classes; then such model is used to predict the class labels for new data. The term *supervised* refers to the fact that the labels used to build the model in the training step are assigned by a supervisor, and their correctness is not questioned. The problem of binary supervised classification can be efficiently solved by means of the SVM algorithm, introduced by Vapnik in the '90s [1], which aims to find a surface separating the two classes. Recently, thanks to the work of Mangasarian [2], some effective alternatives to the SVM have been developed, based on the idea of building two surfaces approximating the two classes instead of a single, separating one. The methodologies following this approach are generally known as Generalized Eigenvalues Classifiers (GEC), since their mathematical formulation leads to the solution of a generalized eigenvalues problem.

In many fields of knowledge, the technological gave rise to huge collection of data, often characterized by a high (up to many thousands) number of features. Classifying those data, and extracting meaningful information from them, represents a real challenge for the machine learning community. Here we deal with two issues related to the processing of high-dimensional data:

- *the feature selection problem*: among all the features which describe the data, we are interested in the ones that mostly affect the classification; more precisely, we would like to identify the most relevant ones, that is the smallest set of features that can correctly describe (i.e. classify) the phenomenon under study, and thus reducing the computational effort of analyzing new data;
- *the semi-supervised classification problem*: the labeling process, which is essential in supervised classification tasks, can be both expensive and time-consuming, especially on high-dimensional datasets; we would like to

Numerical Computations: Theory and Algorithms (NUMTA–2016)
AIP Conf. Proc. 1776, 040005-1–040005-4; doi: 10.1063/1.4965317
Published by AIP Publishing. 978-0-7354-1438-9/$30.00

investigate semi-supervised approaches, in which good classification models are built exploiting also the information coming from the unlabeled data.

Here we present two extensions of the ReGEC algorithm (Guarracino et al. [3]), namely ReGEC_L1 and LapReGEC, in which the original algorithm is extended with a regularization term in order to address the problems above stated.

Notation: a real column vector is indicated by a bold letter ($\mathbf{v}$) and matrices by capital letters (M). The transposed of a vector $\mathbf{v}$ and a matrix M are $\mathbf{v}'$ and M', respectively. $[M \ \ \mathbf{v}]$ represents the matrix M with an added column equal to $\mathbf{v}$. Norms are L2, unless otherwise stated.

GENERALIZED EIGENVALUES CLASSIFIERS

The ReGEC algorithm

Let $\mathbf{x}_i \in \mathbb{R}^n, \forall i = 1, \ldots, l$ be the input samples, and $y_i \in \{-1, +1\} (i = 1, \ldots, l)$ their labels, and let consider two matrices $A \in \mathbb{R}^{l_+ \times n}$ and $B \in \mathbb{R}^{l_- \times n}$, representing the points of the two classes: a row of the matrix A is a point $\mathbf{x}_i$ in the feature space such that $y_i = +1$ (l_+ points), while a row of the matrix B is a point $\mathbf{x}_j$ such that $y_j = -1$ (l_- points). The GEPSVM algorithm by Mangasarian and Wild [2] aims to find two hyperplanes, each closest to one set of points and furthest from the other. Identifying a hyperplane with the couple $(\mathbf{w}, \gamma) \in \mathbb{R}^n \times \mathbb{R}$ of its coefficients, the two proximal hyperplanes can be determined by:

$$\begin{aligned} (\mathbf{w}_+, \gamma_+) &= \arg \min_{(\mathbf{w},\gamma) \neq \mathbf{0}} \frac{\|A\,\mathbf{w} - \mathbf{e}\,\gamma\|^2}{\|B\,\mathbf{w} - \mathbf{e}\,\gamma\|^2} \\ (\mathbf{w}_-, \gamma_-) &= \arg \min_{(\mathbf{w},\gamma) \neq \mathbf{0}} \frac{\|B\,\mathbf{w} - \mathbf{e}\,\gamma\|^2}{\|A\,\mathbf{w} - \mathbf{e}\,\gamma\|^2} \end{aligned} \tag{1}$$

where $\mathbf{e}$ is a column vector of 1s of suitable size. By defining

$$G = [A \ \ -\mathbf{e}]' [A \ \ -\mathbf{e}], \quad H = [B \ \ -\mathbf{e}]' [B \ \ -\mathbf{e}], \quad \mathbf{z} = [\mathbf{w}' \ \ \gamma]', \tag{2}$$

the first in (1) can be rewritten as

$$\mathbf{z}_+ = \arg \min_{\mathbf{z} \neq \mathbf{0}} \frac{\mathbf{z}'\,G\,\mathbf{z}}{\mathbf{z}'\,H\,\mathbf{z}} \tag{3}$$

i.e. the minimization of a generalized Rayleigh quotient. Since G and H are symmetric positive semi-definite matrices, from the Courant-Fischer min-max principles, if H is positive definite, solving problem (3) is equivalent to finding the eigenvector related to the minimum eigenvalue of the generalized eigenvalues problem

$$G\mathbf{z} = \lambda H \mathbf{z}. \tag{4}$$

A similar result can be proved in the case in which there exists a real value c such that $G - cH$ is positive semi-definite [4] [5]. Mangasarian proposed to use a Tikhonov's regularization term for the numerator, i.e. to shift the spectrum of the matrix G by a positive scalar δ. This means that the original problem is perturbed and eigenvectors of the new problem are computed. The approach we will follow is the one proposed by Guarracino et al. [3], who in their ReGEC algorithm solve the problem

$$\min_{\mathbf{z} \neq \mathbf{0}} \frac{\mathbf{z}'\,(G + \delta H)\,\mathbf{z}}{\mathbf{z}'\,(H + \delta G)\,\mathbf{z}}. \tag{5}$$

A peculiar feature of (5) is that it preserves the eigenvectors of (4), with the eigenvalues λ which undergo the shift

$$\lambda^r = \frac{\lambda + \delta}{1 + \delta\,\lambda}. \tag{6}$$

Feature selection via L1-regularization

In this section we will focus on the first of the two problem posed, the problem of feature selection. The solution proposed in [6] consists in a L1-norm regularization term added in (5), which, by formulating the problem as a minimization on the sphere $\|\mathbf{z}\|_2 = 1$, becomes

$$\min_{\|\mathbf{z}\|_2 = 1} \frac{\mathbf{z}'\,(G + \delta H)\,\mathbf{z}}{\mathbf{z}'\,(H + \delta G)\,\mathbf{z}} + \nu \|\mathbf{w}\|_1, \tag{7}$$

where $\mathbf{z} = [\mathbf{w}' \ \ \gamma]'$. Each component of the vector $\mathbf{w}$ is related to a feature of the classification problem. It is well known [7] that L1-regularization has the property of forcing components to be exactly zero, therefore producing a sparse solution and fostering a natural feature selection process. From the mathematical point of view the main drawback of equation (7) is its non-differentiability, which is due to the presence of the L1-norm term. In [6] the issue is addressed by replacing the L1-norm with a two times differentiable approximation, i.e. by solving the problem

$$\min_{\|\mathbf{z}\|_2=1} \frac{\mathbf{z}'\,(G+\delta H)\,\mathbf{z}}{\mathbf{z}'\,(H+\delta G)\,\mathbf{z}} + \nu \sum_{i=1}^{n} \sqrt{z_i^2+\varepsilon}, \tag{8}$$

which has been called ReGEC_L1 method. It's worth to note that problem (8) is a non convex problem and it is no more solvable by means of a generalized eigenvalues problem. Therefore an algorithm for constrained differentiable optimization is needed, and, as described in [6], we use a Riemannian Trust Region (RTR) algorithm [8, 9]. A possible alternative approach is to use global optimization algorithms either heuristic or deterministic which have been successfully used in different contexts (see, for instance [10, 11]).

Semi-supervised Classification via graph-based regularization

The second problem we address is the semi-supervised classification. We have the labeled data $(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \mathbb{R},\ i = 1, \ldots, l$, and also a set of unlabeled data $\mathbf{x}_i \in \mathbb{R}^n,\ i = l+1, \ldots, l+u$. As before we consider the two matrices $A \in \mathbb{R}^{l_+ \times n}$ and $B \in \mathbb{R}^{l_- \times n}$, representing the labeled points of the two classes.

Following the example of the LapSVM by Belkin et al. [12], information on the distribution of labeled and unlabeled data can be embedded in the training step by means of a graph-based regularization [13], i.e. by treating the points as nodes of an undirected graph. This strategy is based on the so called *manifold assumption* which states that the data belong to a low dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^n$ and their labels change smoothly along the tangent direction.

We modify (5) by adding at the numerator a "smoothness" regularization term, i.e. $(J\mathbf{z})^T L J\mathbf{z}$, where the matrix J is defined as $J = [X, \mathbf{e}]$, with $X \in \mathbb{R}^{(l+u)\times n}$ matrix containing the whole training set (labeled + unlabeled points) and $\mathbf{e}$, as before, column vector of 1s. The matrix $L \in \mathbb{R}^{(l+u)\times(l+u)}$ is the *k-nearest neighbors graph (k-NNG) Laplacian matrix* built for the points of X, which is defined as $L = \mathcal{D} - \mathcal{A}$, where $\mathcal{A} = (a_{ij})$ is the adjacency matrix of the graph and the diagonal matrix $\mathcal{D}$ is called the degree matrix, i.e. $d_{ii} = \sum_{j=1}^{l+u} a_{ij}\ (\forall i = 1, \ldots, l+u)$. Therefore the minimization problem to be solved is

$$\mathbf{z}_+ = \arg\min_{\mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}^T\,(G+\delta H+\mu\, J^T L J)\,\mathbf{z}}{\mathbf{z}^T\,(H+\delta G)\,\mathbf{z}}. \tag{9}$$

Differently from problem (8), the problem in (9) can be still formulated as the resolution of a generalized eigenvalues problem.

NUMERICAL RESULTS

The proposed algorithms were tested on five benchmark datasets taken from the UCI repository [14] and already used in [6] to test the Regec_L1 algorithm against other feature selection methodologies. For each data set 20 splits were considered, each partitioning the data into training and testing sets containing respectively 70% and 30% of the samples. We used the values already determined in the Regec_L1 paper to set the λ, ε and τ parameters. For LapReGEC we exploited a grid search in the parameters space. Moreover for each split and for each percentage of labeled training data in the set $\{30\%, 20\%, 10\%\}$, 4 holdouts were considered, dividing the original training set in labeled and unlabeled samples. Table 1 shows the mean test accuracy obtained on each dataset with the ReGEC, ReGEC_L1 and LapReGEC algorithms; as regards ReGEC_L1, the mean number of selected features is showed; as for LapReGEC the accuracy results are shown for the three different percent ages of unlabeled training data.

When comparing the accuracy results of ReGEC and ReGEC_L1, and considering the "feature" column for ReGEC_L1, we note that, despite the strong feature selection, just a very little deterioration of the accuracy performances can be observed. LapRegec was tested using only a subset of the labeled data, ranging from 30% to 10%. On three datasets (namely *Heart, Ionosphere* and *WDBC*) the accuracies are as good as the ones obtained when the whole labeled dataset is used. On the *Parkinsons* and *Sonar* datasets the results are not as stunning, but still the LapRegec algorithm is able to recover the essential structure of the data with as few labeled samples as the 10%.

Table 1. Accuracy results of ReGEC, ReGEC_L1 and LapReGEC on five UCI benchmark datasets. For ReGEC_L1 the number of feature selected is also shown, together with the total number of features of the data. For LapReGEC, results are shown when 30%, 20%, and 10% of all the training labels are used, respectively.

Dataset	ReGEC	ReGEC_L1	ReGEC_L1	LapREGEC		
			sel. feat. (tot.)	30%	20%	10%
Heart	81.79 ±3.7	80.93 ±3.4	7.5 (13)	83.73 ±2.58	83.01 ±2.13	82.28 ±2.52
Ionosphere	85.75 ±2.9	83.21 ±2.9	7.4 (34)	90.24 ±1.98	89.56 ±2.13	83.88 ±3.55
Parkinson	93.58 ±3.9	93.58 ±3.9	3.2 (22)	79.17 ±2.31	79.52 ±2.34	80.52 ±3.49
Sonar	83.20 ±4.7	82.66 ±4.3	9.0 (60)	71.54±3.34	68.95±4.00	65.70 ±2.90
WDBC	95.38 ±1.2	95.49 ±1.5	4.5 (30)	94.03±1.05	94.00 ±1.21	93.60 ±1.34

CONCLUSIONS AND FUTURE WORK

In this paper, we showed two extensions of ReGEC based on exploiting regularization terms to both improve the accuracy of the classification, and to address the problems arising from dealing with large datasets. ReGEC_L1 algorithm exploits an embedded feature selection method that is able to correctly classify data, selecting a small number of features. This might be a very useful approach when data contains a large number of not relevant, redundant and noisy features. LapReGEC is a semi-supervised method that can obtain good classification accuracies also when very small subsets are used. This approach is essential when only small portion of large datasets are annotated. Since real data is often affected by both those issues at the same time, we are planning to develop a complete methodology comprising both the feature selection and the semi-supervised approach and their extensions to multiclass classification tasks [15], with the aim to address the classification of large data sets in the most flexible and effective way.

ACKNOWLEDGMENTS

This work has been funded by MIUR PON02-00619 project. Mario Guarracino work has been conducted at National Research Institute University Higher School of Economics and has been supported by the RSF grant n. 14-41-00039.

References

[1] C. Cortes and V. Vapnik, Machine Learning **20**, 273–297.
[2] O. L. Mangasarian and E. W. Wild, Pattern Analysis and Machine Intelligence, IEEE Transactions on **28**, 69–74 (2006).
[3] M. R. Guarracino, C. Cifarelli, O. Seref, and P. M. Pardalos, Optimisation Methods and Software **22**, 73–81 (2007).
[4] P. Lancaster and Q. Ye, in *The Gohberg Anniversary Collection* (Springer, 1989), pp. 247–278.
[5] Q. Ye, *Variational principles and numerical algorithms for symmetric matrix pencils* (Mathematics and Statistics, University of Calgary, 1989).
[6] M. Viola, M. Sangiovanni, G. Toraldo, and M. R. Guarracino, Optimization Letters 1–13 (2015).
[7] R. Tibshirani, Journal of the Royal Statistical Society. Series B (Methodological) 267–288 (1996).
[8] P.-A. Absil, C. Baker, and K. Gallivan, Journal of Computational and Applied Mathematics **189**, 274–285 (2006).
[9] P.-A. Absil, C. G. Baker, and K. A. Gallivan, Foundations of Computational Mathematics **7**, 303–330 (2007).
[10] D. di Serafino, S. Gomez, L. Milano, F. Riccio, and G. Toraldo, Journal of Global Optimization **48**, 41–55 (2010).
[11] D. di Serafino, G. Liuzzi, V. Piccialli, F. Riccio, and G. Toraldo, Journal of Optimization Theory and Applications **151**, 175–190 (2011).
[12] M. Belkin, P. Niyogi, and V. Sindhwani, The Journal of Machine Learning Research **7**, 2399–2434 (2006).
[13] K. Sinha, in *Data Classification: Algorithm and Applications*, Data Mining and Knowledge Discovery Series, edited by C. C.Aggarwal (Chapman and Hall/CRC, 2014), pp. 511–536.
[14] K. Bache and M. Lichman, URL http://archive. ics. uci. edu/ml **901** (2013).
[15] M. R. Guarracino, A. Irpino, and R. Verde, "Multiclass generalized eigenvalue proximal support vector machines," in *Complex, Intelligent and Software Intensive Systems (CISIS), 2010 International Conference on* (2010), pp. 25–32.