Prior Knowledge in the Classification of Biomedical Data

Danilo Abbate, Roberta De Asmundis, and Mario Rosario Guarracino

Abstract. Standard data analysis techniques for biomedical problems cannot take into account existing prior knowledge, and available literature results cannot be incorporated in further studies. In this work we review some techniques that incorporate prior knowledge in supervised classification algorithms as constraints to the underlying optimization and linear algebra problems. We analyze a case study, to show the advantage of such techniques in terms of prediction accuracy.

Keywords: Supervised classification, Neural Networks, Support Vector Machines, Generalized Eigenvalue Classifier.

1 Introduction

The widespread availability of biomedical data is posing new and challenging problems to standard analysis algorithms. These problems are related to the quality of data, that are often affected by errors and uncertainty. This is the case of high throughput genomic and proteomic technologies, where the signal to noise ratio is very low. Other questions raise when data produced by comparable experimental protocols are available, because there is no clear strategy to systematically take advantage of previous results and knowledge. In the case of supervised classification, where models are built from data for which the class membership is known, available labeled data is added to the training sets. This has two major drawbacks. First, enlarging the training set increases the computational time needed to elaborate the model. Then, if data are affected by errors or uncertainties, these are introduced in the new classification model, reducing its generalization capabilities.

Danilo Abbate, Roberta De Asmundis, and Mario Rosario Guarracino
High Performance Computing and Networking Institute,
National Research Council (ICAR-CNR), 80131 Naples, Italy
e-mail: mario.guarracino@cnr.it

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 1–8.
springerlink.com

In this paper we show how to introduce prior knowledge in Support Vector Machines (SVM) [12], Generalized Eigenvalue Proximal SVM (GPSVM) [8], and Radial Basis Functions (RBF) Neural Networks [1]. The idea is if knowledge can be expressed in terms of regions of the data space, in which all points belong to a given class, then the geometrical expression of such regions can be used to constrain the underlying mathematical programming problem. The advantage of such strategy is that, although no points are added to the training set, the model is constrained to take into account available knowledge. We provide a case study that highlights the advantages of such strategy, in terms of classification accuracy.

2 Classification Algorithms

Support Vector Machines

SVM are the state of the art supervised classification methods, widely accepted in many application areas. SVM find a plane $\mathbf{w}^T\mathbf{x} + b = 0$ with the objective to separate the elements belonging to two different classes. To this extend, we determine two parallel planes $\mathbf{w}^T\mathbf{x} + b = \pm 1$, of maximum distance, leaving all points of the two classes on different sides. Elements with the minimum distance from both classes are called *support vectors* and are the only elements needed to train the classifier.

Let us consider a data set composed of n pairs $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i \in \mathbf{R}^m$ is the feature vector of a point, and $y_i \in \{-1, 1\}$ is the class label. The optimal separating plane is the solution to a quadratic linearly constrained problem.

The advantage of this method is that a very small number of support vectors are sufficient to define the optimal separating plane. In some cases, the relationship between points and class labels can be nonlinear and it is impossible to find a separating plane. In such a case, data can be nonlinearly embedded to a higher dimensional space in which the linear separation can be found. This nonlinear mapping can be implicitly done by kernel functions, which represent the inner product of the elements in the nonlinear space.

The nonlinear classification model cannot describe the discriminating function in terms of inequalities involving linear relations among features. This can be perceived as a problem in case of medical diagnosis, in which doctors prefer to find simple correlations between the results of a clinic exams and the diagnosis or prognosis of an illness. On the other hand, it is generally accepted that results achieved by nonlinear models provide higher classification accuracy. Furthermore, the number of exams to consider for a diagnosis can be very high and cannot be correlated only with the experience. Finally, methods that provide explicit classification rules are not guaranteed to find a set of rules small enough to be easy readable.

Generalized Eigenvalue Classifier

GEPSVM is an efficient algorithm in which the binary classification problem can be formulated as a generalized eigenvalue problem.

Let us consider two matrices $A \in \mathbf{R}^{n\times m}$ and $B \in \mathbf{R}^{k\times m}$, with $m \ll n+k$, representing the two classes, each row being a point in the feature space. Mangasarian et al. [8] propose to classify these sets of points A and B using two planes in the feature space, each closest to one set of points, and furthest from the other.

Suppose that points in classes A and B are not linearly separable, then a nonlinear embedding of each point $\mathbf{x}$ can be obtained using a Radial Basis Function kernel. Each component of the transformed point is given by $K(\mathbf{x},C_i) = exp(\|\mathbf{x}-C_i\|^2/\sigma)$, where C_i is the i-th row of $C = \left[A^T, B^T\right]^T \in \mathbf{R}^{(n+k)\times m}$, and σ is a parameter.

The two planes $K(\mathbf{x},C)\mathbf{u}_1 - \gamma_1 = 0$ and $K(\mathbf{x},C)\mathbf{u}_2 - \gamma_2 = 0$ in the feature space, can be obtained solving the generalized eigenvalue problem [6]:

$$\min_{\mathbf{u},\gamma\neq 0} \frac{\|K(A,C)\mathbf{u} - \mathbf{e}\gamma\|^2 + \delta\|\tilde{K}_B\mathbf{u} - \mathbf{e}\gamma\|^2}{\|K(B,C)\mathbf{u} - \mathbf{e}\gamma\|^2 + \delta\|\tilde{K}_A\mathbf{u} - \mathbf{e}\gamma\|^2}. \tag{1}$$

Here $\tilde{K}_A$ and $\tilde{K}_B$ are diagonal matrices with the diagonal entries from the matrices $K(A,C)$ and $K(B,C)$; $\mathbf{e}$ is a vector of 1s of proper dimension, $\mathbf{u}$ is the coefficient vector of the plane, γ is the plane intercept and δ is the regularization parameter. The eigenvectors related to the minimum and the maximum eigenvalues of (1), provide the coefficients of the proximal planes P_i, $i = 1,2$. The class of a new point $\mathbf{x}$ is determined as

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

where $dist(\mathbf{x},P_i)$ is the distance of a point $\mathbf{x}$ from plane P_i.

RBF Neural Networks

A RBF neural network is divided into two operative blocks: an inner hidden layer, and the output layer. The hidden layer creates a response localized on the input vector $\mathbf{x}$; the binary output will then be calculated as a weighted sum of these localized responses. Training a RBF network is a procedure divided into two phases: in the first one the parameters of the radial bases function are calculated using an unsupervised learning algorithm. In this phase the data set is divided in $\bar{n}+\bar{k}$ clusters. We define as $\bar{\mathbf{x}}$ the $\bar{n}+\bar{k}$ points closest to each centroid. In the second part of the training, we search for values of the weights w_i which determine the binary output:

$$h(\mathbf{x}) = \sum_{i=1}^{\bar{n}+\bar{k}} w_i K(\mathbf{x},\bar{\mathbf{x}}_i), \quad \bar{n} \ll n, \quad \bar{k} \ll k. \tag{3}$$

Such weights are calculated by minimizing the following error function, with respect to w_i:

$$E = \frac{1}{2}\sum_{i=1}^{n+k}(h(\mathbf{x}_i) - y_i)^2 \tag{4}$$

where y_i is the label of the point $\mathbf{x}_i$.

3 Prior Knowledge

SVM

We are now showing how it is possible to obtain, with a linear program [9], a nonlinear separating surface using a kernel function $K(\mathbf{x},C) : \mathbf{R}^m \times \mathbf{R}^{(n+k)\times m} \rightarrow \mathbf{R}^{n+k}$, to embed the points in a higher dimensional space. We recall that the resulting plane, projected in the feature space [11], has equation:

$$K(\mathbf{x},C)\mathbf{u} - \gamma = 0. \tag{5}$$

In standard SVM, parameters $\mathbf{u} \in \mathbf{R}^{n+k}$ and $\gamma \in \mathbf{R}$ are determined solving the following quadratic optimization problem [7], for some $\nu > 0$:

$$\begin{array}{ll} \min\limits_{\mathbf{u},\gamma,\mathbf{y}\in\mathbf{R}^{(n+k)+1+(n+k)}} & \nu\mathbf{e}^T\mathbf{y} + \frac{1}{2}\mathbf{u}^T\mathbf{u} \\ s.t. & D(K(C,C)\mathbf{u} - \mathbf{e}\gamma) + \mathbf{y} \geq \mathbf{e}, \quad \mathbf{y} \geq 0. \end{array} \tag{6}$$

where D is a diagonal matrix, with the diagonal elements equal to the labels of the corresponding element of the training set C, $\mathbf{y}$ is a vector of slack variables. Such condition places the points belonging to the two classes $+1$ and -1 on two different sides of the nonlinear separation surface (5). Problem (6) corresponds to the following linear programming problem [9]:

$$\begin{array}{lr} \underset{\mathbf{u},\gamma,\mathbf{y},\mathbf{s}}{min} & \nu\mathbf{e}^T\mathbf{y} + \mathbf{e}^T\mathbf{s} \\ s.t. & (K(C,C)\mathbf{u} - \mathbf{e}\gamma) + \mathbf{y} \geq \mathbf{e}, \\ & -\mathbf{s} \leq \mathbf{u} \leq \mathbf{s}, \\ & \mathbf{y} \geq 0, \end{array} \tag{7}$$

where $\mathbf{s} \in \mathbf{R}^{n+k}$ is a vector of non negative slack variables.

In order to improve the results obtained by a classifier solely from the training set, it is possible to impose the knowledge of an expert into the learning phase of the function (5) [10]. Such expertise is represented by the following implication, which represents a knowledge region $\Delta \subset \mathbf{R}^m$ in the input space in which all points $\mathbf{x}$ are known to belong to class $+1$:

$$g(\mathbf{x}) \leq 0 \Rightarrow K(\mathbf{x},C)\mathbf{u} - \gamma \geq \alpha, \forall \mathbf{x} \in \Delta, \alpha \in \mathbf{R}^+, \tag{8}$$

where $g(\mathbf{x}) : \Delta \subset \mathbf{R}^m \rightarrow \mathbf{R}$.

To add positive nonlinear knowledge (8) to problem (7) we solve:

$$\begin{aligned} \underset{\mathbf{u},\gamma,\mathbf{y},\mathbf{s}}{min} \quad & \nu\mathbf{e}^T\mathbf{y}+\mathbf{e}^T\mathbf{s} \\ s.t. \quad & D(K(C,C)\mathbf{u}-\mathbf{e}\gamma)+\mathbf{y}\geq\mathbf{e}, \\ & -\mathbf{s}\leq\mathbf{u}\leq\mathbf{s},\ \ \mathbf{y}\geq 0, \\ & K(\mathbf{x}_i,C)\mathbf{u}-\gamma-\alpha+\nu g(\mathbf{x}_i)+z_i\geq 0, \\ & \nu\geq 0, z_i\geq 0,\quad i=1,\ldots,l. \end{aligned} \tag{9}$$

Here $z_1,\ldots,z_l$ are non negative slack variables used to allow small deviation in prior knowledge and ν is a parameter.

To add negative nonlinear knowledge just consider the following implication:

$$f(\mathbf{x})\leq 0\Rightarrow K(\mathbf{x},C)\mathbf{u}-\gamma\leq-\alpha,\forall\mathbf{x}\in\Lambda,\alpha\in\mathbf{R}^+, \tag{10}$$

where $f(\mathbf{x}):\Lambda\subset\mathbf{R}^m\rightarrow\mathbf{R}$ represents the region in the input space where implication (10) forces the classification function to be less than or equal to $-\alpha$, in order to classify the points $\mathbf{x}\in\{\mathbf{x}|h(\mathbf{x})\leq 0\}$ as -1.

Now we can finally formulate the linear program (7) with nonlinear knowledge included in the cost function:

$$\begin{aligned} \underset{\mathbf{u},\gamma,\mathbf{y},\mathbf{s},\nu,\mathbf{p},z_i,q_i}{min} \quad & \nu\mathbf{e}^T\mathbf{y}+\mathbf{e}^T\mathbf{s}+\mu(\sum_{i=1}^{l}z_i+\sum_{j=1}^{t}q_j) \\ s.t. \quad & D(K(C,C)\mathbf{u}-\mathbf{e}\gamma)+\mathbf{y}\geq\mathbf{e}, \\ & -\mathbf{s}\leq\mathbf{u}\leq\mathbf{s},\ \ \mathbf{y}\geq 0, \\ & K(\mathbf{x}_i,C)\mathbf{u}-\gamma-\alpha+\nu g(\mathbf{x}_i)+z_i\geq 0, \\ & \nu\geq 0,\ z_i\geq 0,\ i=1,\ldots,l \\ & -K(\mathbf{x}_j,C)\mathbf{u}+\gamma-\alpha+pf(\mathbf{x}_j)+q_j\geq 0, \\ & p\geq 0,\ q_j\geq 0,\ j=1,\ldots,t \end{aligned} \tag{11}$$

where μ is a positive weight, and p is a parameter.

The LP problem (11) minimizes the margin between the two classes constraining the classification model to leave the two prior knowledge regions Δ and Λ in the corresponding half spaces.

GEPSVM

It is possible to add nonlinear prior knowledge to GEPSVM formulating the model in terms of a constrained generalized eigenvalue problem. The latter has been extensively studied and a procedure for its solution has been proposed by Golub in [4].

If G, H and $\mathbf{z}$ are defined as:

$$\begin{aligned} G &= [K(A,C), -\mathbf{e}]^T [K(A,C), -\mathbf{e}] + \delta \left[\tilde{K}_B, -\mathbf{e}\right]^T \left[\tilde{K}_B, -\mathbf{e}\right], \\ H &= [K(B,C), -\mathbf{e}]^T [K(B,C), -\mathbf{e}] + \delta \left[\tilde{K}_A, -\mathbf{e}\right]^T \left[\tilde{K}_A, -\mathbf{e}\right], \\ \mathbf{z} &= \left[\mathbf{u}^T, \gamma\right]^T \in \mathbf{R}^{n+k+1}, \end{aligned} \tag{12}$$

constraints can be expressed by the equation:

$$V^T \mathbf{z} = 0, \tag{13}$$

where $V \in \mathbf{R}^{(n+k+1)\times p}$ is a matrix of rank r, with $r < p < n+k+1$. The constrained formulation of the eigenvalue problem (1) with positive knowledge becomes:

$$\begin{aligned} \min_{\mathbf{z}\in\mathbf{R}^{n+k+1}} \quad & \frac{\mathbf{z}^T G \mathbf{z}}{\mathbf{z}^T H \mathbf{z}} \\ s.t. \quad & V^T \mathbf{z} = 0. \end{aligned} \tag{14}$$

Let Δ be the set of class $+1$ points describing nonlinear positive knowledge, then the constraint matrix V represents knowledge imposed on class $+1$ points, hence it will be:

$$V = \left[K(\Delta, C),\ -\mathbf{e}\right]^T \tag{15}$$

Matrix V needs to be rank deficient in order to have a non-trivial solution. The set of constraints (13) requires all points in Δ to have null distance from the plane, and thus to belong to class $+1$. Similarly, we can add a negative knowledge.

RBF Neural Networks

As for GEPSVM, [5], a classification model calculated by the RBF network must pass through the prior knowledge points.

Prior knowledge is then added as a set of constraints to problem (4) to obtain the following minimization problem:

$$\begin{aligned} \min_{w_i} \ & \frac{1}{2} \sum_{i=1}^{n+k} (h(\mathbf{x}_i) - y_i)^2 \\ s.t. \ & V^T \mathbf{x} \geq 0. \end{aligned} \tag{16}$$

The constraints of this problem force the solution of equation (4) to pass through the points represented by the matrix V. Algebraically, this means the solution of the least squares problem has to be searched in the subspace generated by prior knowledge points. As pointed out by Golub [3], the original problem is reduced with a *QR* decomposition, or with a *singular value decomposition* as shown by Bjorck [2].

4 A Case Study

The prior knowledge introduced in the classification methods discussed above, has been tested on the UCI data set Thyroid composed of data coming from 215 patients. For each patient 5 cytological and clinical features are provided, which are useful to divide patients in two classes: *sick* and *not sick*. The first class is composed of 65 patients, while the second of 150 healthy patients. The features are: the percentage of T3-resin, total serum thyroxin measured by the isotopic displacement method, total serum triiodothyronine measured by radioimmuno assay, TSH measured by radioimmuno assay, and the maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value.

The results reported in Table 1 are comparison among GEPSVM, SVM and RBF Neural Network methods with and without prior knowledge. The values of accuracy, sensitivity and specificity have been obtained with the *leave one out* cross validation.To simulate the prior knowledge, points were chosen as the misclassified support vectors, obtained training the SVM on the complete data set during the *leave one out* cross validation.

Table 1 Values of accuracy, sensitivity and specificity obtained using GEPSVM, SVM and RBF methods The second line of each block in the table shows the results obtained introducing prior knowledge.

Method	Accuracy	Sensitivity	Specificity
GEPSVM	93.02%	87.69%	95.33%
GEPSVM with knowledge	99.07%	96.62%	100.00%
SVM	93.95%	92.23%	96.00%
SVM with knowledge	98.90%	96.92%	99.33%
RBF	85.12%	55.38%	98.00%
RBF with knowledge	90.23%	72.31%	98.00%

We note that all methods have a better prediction accuracy and higher values of sensitivity and specificity.

5 Conclusion

In this work, we described some classification methods that can take advantage of prior knowledge. We provided a case study to show the gain in terms of accuracy, sensitivity and specificity. Results confirm that prior knowledge substantially increase the classification accuracy of the considered methods. Further work need to be devoted to the automatic knowledge discovery in databases, when data are affected by noise and uncertainty.

Acknowledgements. Danilo Abbate and Roberta De Asmundis spent a research period at ICAR CNR as graduate students. This work has been partially funded by COST action ICT-0702 and PRIN 20078MHYS4.

References

1. Bishop, C.M.: Neural networks for pattern recognition. Oxford University Press, Oxford (1995)
2. Bjork, A.: Numerical methods for least squares. SIAM, Philadelphia (1996)
3. Golub, G., van Loan, C.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore (1996)
4. Golub, G.H., Underwood, R.: Stationary values of the ratio of quadratic forms subject to linear constraints. Z. Angew. Math. Phys (ZAMP) 21(3), 318–326 (1970)
5. Guarracino, M., Abbate, D., Prevete, R.: Nonlinear knowledge in learning models. In: Proceedings of Workshop on Prior Conceptual Knowledge in Machine Learning and Knowledge Discovery, European Conference on Machine Learning, pp. 29–40 (2007), `http://www.ecmlpkdd2007.org/CD/workshops/PRICKLWM2/P_Gua/GuarracinoPriCKL/Guarracino.pdf`
6. Guarracino, M.R., Cifarelli, C., Seref, O., Pardalos, P.: A classification method based on generalized eigenvalue problems. Optim. Methods Softw. 22, 73–81 (2007)
7. Lee, Y., Mangasarian, O.L.: Ssvm: A smooth support vector machine for classification (1999), `http://citeseer.ist.psu.edu/lee99ssvm.html`
8. Mangasarian, O.L., Wild, E.W.: Multisurface proximal support vector classification via generalized eigenvalues. Tech. Rep. 04-03, Data Mining Institute (2004)
9. Mangasarian, O.L., Wild, E.W.: Nonlinear knowledge-based classification. Tech. rep., Data Mining Institute Technical Report 06-04, Computer Science Department, University of Wisconsin, Madison, Wisconsin (2006)
10. Pardalos, P.M., Abbate, D., Guarracino, M.R., Chinchuluun, A.: Neural network classification with prior knowledge for analysis of biological data. In: Proceedings of the International Symposium on Mathematical and Computational Biology, Biomat 2008, Brazil, pp. 223–234. World Scientific, Singapore (2008)
11. Schölop, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)
12. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)