Linear Penalization Support Vector Machines for Feature Selection

Jaime Miranda, Ricardo Montoya, and Richard Weber

Department of Industrial Engineering,
Faculty of Physical and Mathematical Sciences,
University of Chile
{jmiranda, rmontoya, rweber}@dii.uchile.cl

Abstract. Support Vector Machines have proved to be powerful tools for classification tasks combining the minimization of classification errors and maximizing their generalization capabilities. Feature selection, however, is not considered explicitly in the basic model formulation. We propose a linearly penalized Support Vector Machines (LP-SVM) model where feature selection is performed *simultaneously* with model construction. Its application to a problem of customer retention and a comparison with other feature selection techniques demonstrates its effectiveness.

1 Introduction

One of the tasks of Statistics and Data Mining consists of extracting patterns contained in large data bases. In the case of classification, discriminating rules are constructed based on the information contained in the feature values of each object. By applying the discriminating rule to new objects, it is possible to conduct a classification that predicts the class to which these new objects belong.

Building a classifier, it is desirable to use the smallest number of features possible in order to obtain a result considered acceptable by the user. This problem is known as feature selection [5] and is combinatorial in the number of original features [8].

Recently, Support Vector Machines (SVM) have received growing attention in the area of classification due to certain significant characteristics such as an adequate generalization to new objects, the absence of local minima and representation that depends on few parameters [3, 11, 12]. Nevertheless, among SVM systems there are few approaches established in order to identify the most important features to construct the discriminating rule.

In the present paper we develop a methodology for feature selection based on SVM. The focus chosen is the penalization for each feature used, which is carried out *simultaneously* with the construction of the classification model.

The paper is structured as follows: Section 2 presents the proposed methodology. Section 3 describes its application for customer retention in a Chilean bank. Section 4 shows the areas of future development and summarizes the main conclusions.

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 188 – 192, 2005.

2 Penalized Support Vector Machines (LP-SVM) for Feature Selection

To determine the most relevant features, and to take advantage of SVM capabilities to solve classification problems, we propose a modification of the mathematical model. A penalization for each feature used is incorporated into the objective function. This way, the following three objectives are proposed at the moment of constructing the mathematical formulation of the problem to solve:

1. The capacity for generalization: Minimizing the norm of the normal to the separating hyper plane.
2. Classification errors: Minimizing the sum of the slack variables added to the problem, penalizing each slack variable used.
3. Feature selection: Minimizing the number of features when building the discrimination model, penalizing each feature used.

The current mathematical formulation of the Support Vector Machines only includes the first two of the three objectives. On the other hand, our approach differs from that proposed by Bradley and Mangasarian where only the objectives 2. and 3. are used [1]. The formulation presented in [2] uses second order cone programming for feature selection where a bound assures low misclassification errors and selecting the most appropriate features is performed by minimizing the respective L_1 norm. However, only by explicitly optimizing all three objectives, we can assure that the advantages provided by SVM are combined with the attempt to select the most relevant features.

We need the following notation in order to develop our model.

$$\text{Let } \vec{x} \in \Re^m: \qquad \left|\vec{x}\right|_j = \begin{cases} x_j & \text{if } x_j > 0 \\ 0 & \text{if } x_j = 0, \qquad j = 1,\ldots,m \\ -x_j & \text{if } x_j < 0 \end{cases}$$

Step Function:

$$\left(\vec{x}_*\right)j = \begin{cases} 1 & \text{if } x_j > 0 \\ 0 & \text{if } x_j = 0, \qquad j = 1,\ldots,m \\ -1 & \text{if } x_j < 0 \end{cases}$$

$$\text{We note that if } \vec{x} \in \Re_+^m \qquad \left(\vec{x}_*\right)j \in \{0,1\} \;\; j = 1,\ldots,m$$

Number of Components:

$$x_j \in \{0,1\} \;\; \forall_j \Rightarrow \vec{\mathrm{e}}^T \cdot \vec{x} = \sum_{j=1}^{m} x_j = \text{number of strictly positive components of } \vec{x}$$

Objective function of the proposed model:

$$\frac{1}{2}\left\|\vec{w}\right\|^2 + C_1 \sum_{i=1}^{n} \xi_i + C_2 \vec{e}^{\,T} \cdot \left|\vec{w}\right|_* \tag{1}$$

where $\frac{1}{2}\left\|\vec{w}\right\|^2 + C_1\sum_{i=1}^{n}\xi_i$ corresponds to the traditional formulation of SVM objective function and $\vec{e}^{\,T} \cdot \left|\vec{w}\right|_*$ is the sum of non-negative components of w and refers to the number of selected features.

However, the formulation (1) has the inconvenience of not being a continuous function, because it

- incorporates the modulus $\left(\left|\vec{w}\right|\right)$ and
- a discontinuous step function $\left(\left|\vec{w}\right|_*\right)$.

Replacing the modulus by auxiliary variables and the step function by a concave exponential approximation we obtain the following model:

$$\underset{\vec{w},\vec{v},\varepsilon_i,b}{\text{Minimize}} \quad \frac{1}{2}\left\|\vec{w}\right\|^2 + C_1\sum_{i=1}^{n}\xi i + C_2 \vec{e}^{\,t} \cdot \left(\vec{e} - \varepsilon^{-T\cdot\vec{v}}\right)$$

$$\begin{aligned} \text{Subject to:} \quad & y_i\left(\vec{x}_i \cdot \vec{w} + b\right) - 1 + \xi_i \geq 0 \qquad & i = 1,\ldots,n \\ & \xi_i \geq 0 \qquad & i = 1,\ldots,n \\ & -\vec{v} \leq \vec{w} \leq \vec{v} & \end{aligned} \tag{2}$$

Where $\vec{w}, \vec{v} \in \Re^m$ and $\xi_{i,} b \in \Re$. This model will be called LP-SVM. We shall use Cross Validation in order to obtain the best model parameters in a particular application [7].

3 Applying LP-SVM for Customer Retention

In the case of customer retention, companies spend much of their budget trying to understand customers. It is known that the cost of acquiring a new customer is between 5 and 7 times higher than retaining an old one [10] and that, moreover, to increase customer retention by 5% means increasing financial results by 25% [6]. Data coming from customers is very diverse, so it has to be processed previously, selecting the most important features that represent the customer in order to respond to the market requirements more accurately and efficiently.

For financial institutions such as banks, it is important to understand the typical behavior of customers that are about to leave the institution and want to close their current account. In this case the bank would take actions in order to retain these customers. Although we know how the behavior of each one of our customers evolves, it is not possible to manually follow up on each of them because of the portfolio size and, furthermore, we do not know for certain what behavior is associated with a customer who is about to close his/her account so as to be able to pinpoint him/her exactly.

We apply the proposed methodology LP-SVM to a database of a Chilean bank concerned about retention of its customers. We built a classification model using Support Vector Machines, adding the approach suggested in this publication for feature selection and compared it with two other techniques for feature selection.

We analyzed a data set from the respective database that contains information on the customers who voluntarily closed their current accounts over a 3 months period prior to September 2001 and those who were still active at that day. The information contained in this database includes, among others, the following features for each customer: age, sex, antiquity, marital status, level of education, average salary over the last 3 months, number of products acquired and transaction data.

Each customer in the database belongs to a class depending on whether he/she closed his/her current account or remains active. The variable that indicates the class to which the customer belongs is equal to 1 if the customer closed his/her account and -1 if the customer is still active.

The database we shall study in this application is that for September 2001. It contains 1,937 customers, 995 (51.37%) of which closed their current accounts within the 3 months period prior to September 2001 and the remaining 942 (48.63%) are considered to be active customers. For each customer we have 36 feature values.

The following table presents the results applying three methods to a validation set (holdout sample): LP-SVM, Clamping [9] as a wrapper technique combined with a MLP-type neural network as classifier (C-NN), and a decision tree (DT). The underlined value indicates the best model regarding the respective number of selected features.

As can be seen, LP-SVM performs best among the three methods in 7 of 9 cases.

Table 1. Percentage of correct classification in validation set

Number of selected features:	**10**	**12**	**13**	**14**	**15**	**20**	**25**	**30**	**36**
LP-SVM	50.0	84.0	83.7	82.4	82.0	81.1	82.4	82.9	72.8
C-NN	65.0	67.3	67.1	60.5	63.4	67.0	68.3	68.6	68.7
DT	70.3	72.0	72.8	72.8	73.8	73.8	72.8	73.8	73.8

4 Conclusions and Future Work

We presented LP-SVM, a new approach for feature selection using SVM where the feature selection step is performed *simultaneously* with model construction. A comparison with other techniques for feature selection and classification shows the advantages of LP-SVM.

Future work has to be done in various directions. First, it would be interesting to apply the proposed formulation to the non-linear case where Kernel functions are used for feature space transformation.

It would also be interesting to apply feature selection *simultaneously* to model construction for regression problems. A hybrid methodology where feature selection and model building using Support Vector Regression are performed sequentially has been presented in [4].

Acknowledgements

This project has been funded by the Millennium Science Nucleus on Complex Engineering Systems (www.sistemasdeingenieria.cl) and the project Fondecyt 1040926.

References

1. Bradley, P., Mangasarian, O. (1998): Feature selection vía concave minimization and support vector machines. In *Machine Learning proceedings of the fifteenth International Conference* (ICML'98) 82 –90, San Francisco, California, 1998. Morgan Kaufmann.
2. Bhattacharya, Ch. (2004): Second Order Cone Programming Formulations for Feature Selection. Journal of Machine Learning Research , 1417-1433
3. Cristianini, N., Shawe-Taylor, J. (2000): *An Introduction to Support Vector Machines.* Cambridge University Press, Cambridge, UK.
4. Guajardo, J., Miranda, J., Weber, R. (2005): A Hybrid Forecasting Methodology using Feature Selection and Support Vector Regression. Presented at HIS2005 Hybrid Intelligent Systems, Rio de Janeiro, November 2005
5. Hand, D. (1981): *Discrimination and Classification.* Wiley, Chichester
6. Kotler, P. (2000): *Marketing Management: Analysis, Planning, Implementation, and Control.* Prentice Hall International, New Jersey, 10th edition
7. Montoya, R., Weber, R. (2002): Support Vector Machines Penalizado para Selección de Atributos. XI CLAIO, Concepción, Chile, 27-30 de octubre de 2002 (in Spanish)
8. Nemhauser, G., Wolsey, L. (1988): *Integer and Combinatorial Optimization.* John Wiley and Sons, New York.
9. Partridge, D., Cang, S. (2002): Revealing Feature Interactions in Classification Tasks. In: A. Abraham, J. Ruiz-del-Solar, M. Köppen (eds.): Soft Computing Systems - Design, Management and Applications. IOS Press, Amsterdam, Berlin, 394-403
10. Reichheld, F., Sasser, E. (1990): Zero defections: Quality comes to services. Harvard Business Review September-October:105 –111
11. Shawe-Taylor, J., Cristianini, N. (2004): Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge
12. Vapnik, V. (1998): Statistical Learning Theory John Wiley and Sons, New York.