Simultaneous feature selection and classification via Minimax Probability Machine

Liming Yang*
College of Science, China Agricultural University, Beijing, 100083, China

Laisheng Wang
College of Science, China Agricultural University, Beijing, 100083, China
E-mail: wanglaish@sina.com

Yuhua Sun
Department of Mathematics and Mechanics,USTB, Beijing, 100083, China
E-mail: syuhua68@sina.com

Ruiyan Zhang
Science China Press,Beijing,100717,China
E-mail: zhangry@scichina.org

Received: 27-12-2009
Accepted: 13-08-2010

Abstract

This paper presents a novel method for simultaneous feature selection and classification by incorporating a robust L_1-norm into the objective function of Minimax Probability Machine (MPM). A fractional programming framework is derived by using a bound on the misclassification error involving the mean and covariance of the data. Furthermore, the problems are solved by the Quadratic Interpolation method. Experiments show that our methods can select fewer features to improve the generalization compared to MPM, which illustrates the effectiveness of the proposed algorithms.

Keywords: Minimax probability machine, Feature selection, Probability of misclassification, Machine learning.

1. Introduction

Feature selection for classifiers is an important research tool with many applications[1,2] in machine learning field. Feature selection has two main objectives: 1) to select a small feature subset and 2) to maintain high classification accuracy. This paper addresses the issue of constructing linear classifiers using a small number of features when data is summarized by its moments.

Given the data set , $D=\{(x_i,y_i)|x_i \in R^n, y_i=\pm 1, i=1,2,\cdots m\}$, finding useful features for a linear classifier $f(x)=\mathrm{sgn}(w^T x-b)$ is equivalent to searching for a sparse *w*, such that the most elements of *w* are zero. This can be understood as when the *ith* component of *w* is zero, the *ith* component of the observation vector *x* is irrelevant in deciding the class of *x*. Using the L_0-norm of *w*, $\|w\|_0 = number\ of \{i|w_i \neq 0\}$, the problem of feature selection can be designed to minimize the L_0-norm, but this problem is generally NP-hard[3]. A tractable convex approximation to the problem can be obtained by replacing the L_0-norm with the L_1-norm $\|w\|_1$. Therefore, the problem of feature selection for classifiers can be posed as:

$$\begin{aligned} &\min_{w,\,b} \|w\|_1 \\ &s.t.\quad y_i(w_i^T x-b) \geq 0, i=1,2,\cdots m. \end{aligned} \tag{1}$$

A solution to Eq. (1) yields the desired sparse weight vector *w*. The above formulation can be categorized as an embedded approach[4], where the feature selection

[*] Address: College of Science, China Agricultural University. Correspondence should be addressed to cauylm@126.com.

Published by Atlantis Press

process is embedded into the classification framework. Other feature selection methods for classifiers, like the filter approach and the wrapper approach[5,6], are not discussed in this paper. The interested reader is referred to cited references for additional information on these methods.

Although Minimax Probability Machine (MPM)[7] has been recently shown to have advantages over other methods in the machine learning processes, the featureselection for MPM is still a novel and challenging subject. This paper develops a novel and critical extension algorithm for MPM by incorporating a robust L1-norm into the objective function of MPM to suppress the dimension of the input space and reduce the sensitivity to outliers. As a result, the problem can be solved by the Quadratic Interpolation (QI) algorithm[8].

2. Minimax Probability Machine

MPM provides a worst-case bound on the misclassification error of future data when data is summarized by its moments. Compared with traditional probability models, MPM avoids making assumptions with respect to the data distribution.

Following is a simplified explanation of MPM. A more detailed description can be found in Ref. 7. Let X_1 and X_2 denote n dimension random vectors representing two classes of data, with means $X_1 \sim (\mu_1, \Sigma_1)$ and $X_2 \sim (\mu_2, \Sigma_2)$ covariance matrices and respectively, where $\mu_1, \mu_2 \in R^n$, and $\Sigma_1, \Sigma_2 \in R^{n\times n}$. The objective of MPM is to formulate the hyper plane: $H(w,b) = \{x | w^T x = b\}$ such that class X_1 (or class X_2) is placed in the half space $H_1(w,b) = \{x | w^T x > b\}$ (or $H_2(w,b) = \{x | w^T x < b\}$) with maximal probability with respect to all distributions. This can be formulated as:

$$\begin{aligned} &\max_{\alpha,w,b} \ \gamma \\ s.t.\ &\inf \Pr\{X_1 \in H_1\} \geq \gamma, \\ &\inf \ \Pr\{X_2 \in H_2\} \geq \gamma. \end{aligned} \tag{2}$$

where γ represent the lower bounds of the accuracy for future data. Furthermore, Eq. (2) can be expressed as a second order cone program (SOCP)[9].

$$\begin{aligned} &\min_w \sqrt{w^T \Sigma_1 w} + \sqrt{w^T \Sigma_2 w} \\ &s.t. \quad w^T(\mu_1 - \mu_2) = 1. \end{aligned} \tag{3}$$

3. Feature Selection via MPM (S-MPM)

In this section, we present a novel method for feature selection and classification simultaneously based on MPM. More specifically, we designed a feature selection framework such that the maximum misclassification Bayes error rate is minimized. This indicates that using as few relevant features as possible minimizes the probability of misclassification error.

3.1. *Problem Definition*

Let $\gamma = 1 - \alpha$ and then MPM (2) can be equivalently expressed as:

$$\begin{aligned} &\min_{\alpha,w,b} \ \alpha \\ s.t.\ &\sup \ \Pr\{X_1 \in H_2\} \leq \alpha, \\ &\sup \ \Pr\{X_2 \in H_1\} \leq \alpha. \end{aligned} \tag{4}$$

where α is the upper bound on the misclassification probability in a worst-case setting. This optimization exactly leads to minimizing the expected upper bound of the misclassification Bayes error for two class data. According to the above analysis, the feature selection for classifiers can be designed to minimize the L1-norm of w. Thus, we incorporate a robust L1-norm of w into the objective function of MPM (4) by weighting L1-norm by $1-\lambda$ with a suitably chosen parameter $\lambda \in (0,1)$, which leads to a feature selection framework based on MPM, or S-MPM for short.

$$\begin{aligned} &\min_{\alpha,w,\,b} \ (1-\lambda)\|w\|_1 + \lambda\alpha \\ s.t.\ &\sup \ \Pr\{X_1 \in H_2\} \leq \alpha, \\ &\sup \ \Pr\{X_2 \in H_1\} \leq \alpha, \\ &X_1 \sim (\mu_1, \Sigma_1), X_2 \sim (\mu_2, \Sigma_2). \end{aligned} \tag{5}$$

On the training dataset, the error rate of the classifier, with as few of the useful features as possible, is upper bounded by α . Here the positive parameter λ is a scalar regularization parameter that controls the balance between the prediction accuracy and the number of selected features for the classifier. Thus, the S-MPM classifier is a combination of the MPM and the L1-norm, where the MPM minimizes the upper bound of the misclassification error of predicting future data, and the L1-norm encourages sparseness for the classifier.

3.2. *Model Interpretation*

MPM is mainly focused on maximizing the probability of predicting future data, which is not explicitly connected with the issue of the feature selection for classifiers and the generalization of the model as described here. We will show that S-MPM makes it possible to reduce the selected features and improve the generalization of the model. The advantage of doing so is twofold:

(i) As a generalized model of MPM, S-MPM includes and expands the MPM; when λ= 1, Eq. (5) is equivalent to the MPM. Moreover, this model includes another special model when λ= 0, which formulates feature selection using moments as proposed in Ref.10.

(ii) S-MPM can effectively control the course of dimensionality and also reduce the sensitivity for the classifier. Thus it improves the robustness to outliers due to including L_1-norm in the objective function of S-MPM model.

3.3. *Solving S-MPM Optimization*

For simplicity, we will assume that both Σ_1, Σ_2 are positive definite. Our results can be extended to general positive semi-definite cases.

The following multivariate generalization of the Chebychev Cantelli inequality[11] will be used in the sequel to derive an upper bound on the misclassification probability of a random vector taking values in a given half space.

Lemma. Given $w \in R^n, w \neq 0, b \in R\pounds X \in -R^n$, the mean and covariance of X be $\mu \in R^n, \Sigma \in R^{n\times n}$. Let $H(w,b) = \{z | w^T z < b, z \in R^n\}$ be a given half space. Then the following inequality holds:

$$\Pr\{X \in H\} \geq \frac{s^2}{s^2 + w^T \Sigma w}. \qquad (6)$$

where $s = (b - w^T \mu)_+$, $(x)_+ = \max(x, 0)$.

Then, the expected upper bound of the misclassification error rate can be expressed as:

$$\Pr\{X \notin H\} \leq \frac{w^T \Sigma w}{s^2 + w^T \Sigma w}. \qquad (7)$$

Using Eq. (7), constraint for class X_1 in Eq. (5) can be handled by setting

$$\Pr\{X_1 \in H_2\} \leq \frac{w^T \Sigma_1 w}{(w^T \mu_1 - b)_+^2 + w^T \Sigma_1 w} \leq \alpha. \qquad (8)$$

which results in two constraints:

$$w^T \mu_1 - b \geq \sqrt{\frac{1-\alpha}{\alpha}} \cdot \sqrt{w^T \Sigma_1 w}\,, \quad w^T \mu_1 - b \geq 0. \qquad (9)$$

Let $\sqrt{\alpha/(1-\alpha)} = \eta$. Similarly, applying (7) to the other constraint, Eq. (5) can be formulated as:

$$\begin{aligned} \min_{\alpha, w, b}\ & (1-\lambda)\|w\|_1 + \lambda\alpha \\ s.t.\ & \sqrt{w^T \Sigma_1 w} \leq \eta(w^T \mu_1 - b), \\ & \sqrt{w^T \Sigma_2 w} \leq \eta(b - w^T \mu_2), \\ & w^T \mu_1 - b \geq 0,\ b - w^T \mu_2 \geq 0. \end{aligned} \qquad (10)$$

Let $\Sigma_1 = C_1 C_1^T, \Sigma_2 = C_2 C_2^T, C_1, C_2 \in R^{n\times n}$. Without loss of generality, we can restrict w such that: $w^T \mu_1 - b \geq 1$ and $b - w^T \mu_2 \geq 1$. Furthermore, by introducing two vectors $u \geq 0, v \geq 0, u, v \in R^n$ such that $w = u - v$, then $\|w\|_1 = e^T(u+v)$, finally, the problem (10) can be formulated as:

$$\begin{aligned} \min_{\eta, u, v, b}\ & (1-\lambda)e^T(u+v) + \lambda\frac{\eta^2}{1+\eta^2} \\ s.t.\ & \|C_1^T(u-v)\| \leq \eta((u-v)^T \mu_1 - b), \\ & \|C_2^T(u-v)\| \leq \eta(b - (u-v)^T \mu_2), \\ & (u-v)^T \mu_1 - b \geq 1, \\ & b - (u-v)^T \mu_2 \geq 1, u, v \geq 0. \end{aligned} \qquad (11)$$

This is a fractional programming that minimizes the sum of convex-convex ratios. However, finding its global optima in general has been shown to be difficult[12]. In recent years, although some progress in the special structure of the objective function has been made, most of the corresponding algorithms apply only to the sum of linear ratios. To the best of our knowledge, as of today there have been no reports of an effective method for globally solving the sum of nonlinear ratios problems.

In this paper, we have solved the problem using the Quadratic parabolic Interpolation algorithm[8]. More precisely, in Eq. (11), if we fix η to a specific value within (0,1), the optimization (11) is equivalent to minimizing L_1-norm and becomes a SOCP. If we denote the value of the optimization as a function, the above procedure corresponds to finding an optimal η to minimize. This means finding the minimum point by updating a three-point pattern (η_1, η_2, η_3) repeatedly. The new η denoted by $\overline{\eta}_k$ is given by the quadratic interpolation from the three-point pattern. Then a new three-point pattern is constructed by $\overline{\eta}_k$ and two of η_1, η_2, η_3. This method has been shown to converge super-linearly to a local optimum point[8]. The algorithm is described below:

(i) Given $\varepsilon > 0$ and taking $\eta_1 < \eta_2 < \eta_3$, $\eta_1, \eta_2, \eta_3 \in (0,1)$. Let k=0 and

$$f(\eta) = (1-\lambda)e^T(u+v) + \lambda\frac{\eta^2}{1+\eta^2}. \qquad (12)$$

(ii) Let

$$\overline{\eta}_k = \frac{1}{2}\cdot\frac{(\eta_2^2-\eta_3^2)f(\eta_1)+(\eta_3^2-\eta_1^2)f(\eta_2)+(\eta_1^2-\eta_2^2)f(\eta_3)}{(\eta_2-\eta_3)f(\eta_1)+(\eta_3-\eta_1)f(\eta_2)+(\eta_1-\eta_2)f(\eta_3)}. \qquad (13)$$

(iii) Solve the Eq.(11) and calculate $f(\overline{\eta}_k)$.

If $\eta_1 < \overline{\eta}_k < \eta_2$, *and* $f(\overline{\eta}_k) < f(\eta_1)$, $f(\overline{\eta}_k) < f(\eta_2)$, then $\eta_3 := \eta_2, \eta_2 := \overline{\eta}_k$, namely, use $(\eta_1, \overline{\eta}_k, \eta_2)$ as new (η_1, η_2, η_3) in the new iteration.

If $\eta_2 < \overline{\eta}_k < \eta_3$, *and* $f(\overline{\eta}_k) < f(\eta_2)$, $f(\overline{\eta}_k) < f(\eta_3)$, then $\eta_1 := \eta_2, \eta_2 := \overline{\eta}_k$, namely, use $(\eta_2, \overline{\eta}_k, \eta_3)$ as new (η_1, η_2, η_3) in the new iteration. k:=k+1.

(iv) If $|f(\overline{\eta}_k) - f(\overline{\eta}_{k-1})| < \varepsilon$, then obtain $\overline{\eta}_k$, keep *w,b* in memory, then stop, else (ii).

4. Feature Selection for Fisher Discriminants via MPM (FS-MPM)

In this section, we describe the design of the feature selection for the Fisher discriminant classifier[13] based on MPM. Using the above notation, let $X = X_1 - X_2$ define the difference between the class conditional random vectors, and then X lies in the halfspace. We can derive the Fisher discriminant classifier based on S-MPM (called FS-MPM) by considering the following formulation.

$$\begin{aligned} &\min_{\alpha,w,b} \ \alpha \\ &s.t. \ \sup \Pr\{X \notin H\} \leq \alpha \pounds X \sim\!\dashv(\mu,\Sigma). \end{aligned} \qquad (14)$$

Similarly, we incorporate a robust L$_1$-norm into of *w* the objective function of the Eq. (14), and then the problem can be formulated as:

$$\begin{aligned} &\min_{\alpha,w,b} \ (1-\lambda)\|w\|_1 + \lambda\alpha \\ &s.t. \ \sup \Pr\{X \notin H\} \leq \alpha, \ X \sim (\mu,\Sigma). \end{aligned} \qquad (15)$$

As X_1 and X_2 are independent, the mean of X is $\mu = \mu_1 - \mu_2$ and covariance is $\Sigma = \Sigma_1 + \Sigma_2$. Using the Chebychev bound (6), the constraint of Eq. (15) can be lower bounded by

$$\Pr\{X \notin H\} \leq \frac{w^T\Sigma w}{(w^T\mu)_+^2 + w^T\Sigma w} \leq \alpha, w^T\mu \geq 0. \qquad (16)$$

which results in two constraints:

$$w^T\mu \geq \sqrt{\frac{1-\alpha}{\alpha}}\cdot\sqrt{w^T\Sigma w}, \ \ w^T\mu \geq 0. \qquad (17)$$

Finally, FS-MPM (15) can be reformulated as:

$$\begin{aligned} &\min_{\alpha,w,b} \ (1-\lambda)\|w\|_1 + \lambda\alpha \\ &s.t. \ \sqrt{w^T\Sigma w} \leq \sqrt{\frac{\alpha}{1-\alpha}}\, w^T\mu_1, \\ &\qquad w^T\mu \geq 0. \end{aligned} \qquad (18)$$

A similar analysis is carried out for Eq. (18), and then Eq. (18) involves solving the following problem

$$\begin{aligned} &\min_{\eta,u,v,b} \ (1-\lambda)e^T(u+v) + \lambda\frac{\eta^2}{1+\eta^2} \\ &s.t. \ \ \|C^T(u-v)\| \leq \eta((u-v)^T\mu), \\ &\qquad (u-v)^T\mu \geq 1, \ u,v \geq 0. \end{aligned} \qquad (19)$$

Here $w = u - v, u \geq 0, v \geq 0$, $\Sigma = CC^T$, $C \in R^{n\times n}$ and $\sqrt{\alpha/(1-\alpha)} = \eta$. This is also a nonlinear fractional programming. Here, the QI method is also used to find the local solution of the problem (19).

5. Experimental Design and Results

In order to evaluate the proposed algorithms, we compared our algorithms with the original MPM in 7 real data sets (Wine, Ionosphere, Hepatitis, Sonar, Spam, German credit, Australian credit) from the UCI machine learning repository[14]. These data sets have 13, 34, 19, 60, 57, 20 and 14 features, respectively.

5.1. *Experimental Design*

We used the following performance measurements to evaluate our methods:

(i) Test-Set Accuracy (TSA): including Test-Set Accuracy on Class 1 (TSA1), on Class 2 (TSA2) and on both classes (TSA).

(ii) The Number of selected features (NSFs);

(iii) Receiver Operating Characteristic (ROC)[15,16]. The ROC curve plots a series of true positive rates (TPR) versus false positive rates (FPR). Moreover, when the ROC curves are generated with good shapes and evenly distributed along their length, they can be used to evaluate learning algorithms by using the area under the curve. The larger the area under the curve, the higher the sensitivity for a given specificity that results in better performance of the method.

The experimental results are obtained by averaging 10-fold cross-validation for each dataset. The experiments use SeDuMi[17] as a solver, and the results are given in Table 1-7 respectively. The ROC curves are illustrated in Fig. 1-7. The optimal parameter (para.) λ of Eq. (11) and Eq. (19) is tuned by 5-fold cross-validation on the training set to maximize the test accuracy.

5.2. *Experimental Results*

Tables 1-7 summarize the test set accuracies, the number of selected features and the optimal parameter values.

(i) TSA analysis. We compared only the performance of the S-MPM (11) with MPM (2). The results presented in Tables 1-6 show that S-MPM achieves noticeably better performance than MPM in Sonar, Ionosphere, Hepatitis, Wine, Spam and German credit datasets, especially, for Wine dataset. In Table 7, our models are very close to MPM for Australian credit dataset.

(ii) Comparison of the feature selection. Tables 1-6 show that S-MPM can always select fewer features than MPM but improve the test accuracy in all 6 datasets, with the exception of the Australian credit data. In Table 7, we observe that our models and MPM show no significant difference in terms of the feature selection comparison in the Australian credit data.

By analyzing the results of the simulations, we observed that our models were able to always select fewer features, and the TSAs were consistently better than the ones for MPM in 6 of the analyzed datasets: Ionosphere, Wine, Sonar, Spam and German credit datasets. This means that our models can maintain high test accuracy by removing a few irrelevant features compared to the MPM in most of the data set.

(iii) ROC curve analysis. Figs. 1-6 illustrate that the S-MPM performs significantly better than MPM in 6 datasets: Sonar, Ionosphere, Hepatitis, Wine, Spam and German credit datasets, as shown by the fact that the S-MPM curve is noticeably above the one for MPM. In Fig. 7, two curves are very close in the Australian credit dataset. In addition, not all the portions of the ROC curve are of great interest. In general, those with a small FPR and a high TPR are most important. In light of this, we show the critical portions of Figs. 1-6 with more detail when the FPR is in the range of [0, 0.3] and the TPR is in the range of [0.7, 1.0], respectively. This again demonstrates the superiority of the S-MPM.

In summary, the experimental results demonstrate that our algorithms achieve better performances by removing a few irrelevant features than MPM in terms of the TSA comparison, ROC curve analysis and NSFs criterion in the majority of the datasets. Effectively reducing the number of dimensionality will greatly decrease the computational complexity and reduce the memory requirement.

Table 1 TSAs and NSFs for Sonar .

Sonar	TSA	TSA1	TSA2	NSFs	Para.
MPM	0.545	1.000	0.000	60	--
S-MPM	0.787	0.911	0.762	5	0.90
FS-MPM	0.818	----	----	54	0.90

Table 2 TSAs and NSFs for Ionosphere.

Ionosphere	TSA	TSA1	TSA2	NSFs	Para.
MPM	0.615	1.000	0.000	34	--
S-MPM	0.870	0.920	0.880	5	0.90
FS-MPM	0.930	----	----	12	0.60

Table 3 TSAs and NSFs for Hepatitis.

Hepatitis	TSA	TSA1	TSA2	NSFs	Para.
MPM	0.535	0.800	0.000	19	--
S-MPM	0.635	0.600	0.650	9	0.90
FS-MPM	0.500	----	----	5	0.20

Table 4 TSAs and NSFs for Wine.

Wine	TSA	TSA1	TSA2	NSFs	Para.
MPM	0.350	0.447	0.444	13	--
S-MPM	0.900	0.900	0.890	2	0.50
FS-MPM	1.000	----	----	4	0.50

Table 5 TSAs and NSFs for Spam.

Spam	TSA	TSA1	TSA2	NSFs	Para.
MPM	0.500	1.000	0.000	57	--
S-MPM	0.846	0.802	0.800	12	0.90
FS-MPM	1.000	----	----	4	0.50

Table 6 TSAs and NSFs for German credit.

German	TSA	TSA1	TSA2	NSFs	Para.
MPM	0.478	0.871	0.087	20	--
S-MPM	0.730	0.809	0.639	7	0.50
FS-MPM	0.750	----	----	4	0.50

Table 7 TSAs and NSFs for German credit.

German	TSA	TSA1	TSA2	NSFs	Para.
MPM	0.939	0.875	0.994	14	--
S-MPM	0.928	0.849	0.996	14	0.50
FS-MPM	1.000	----	----	9	0.50

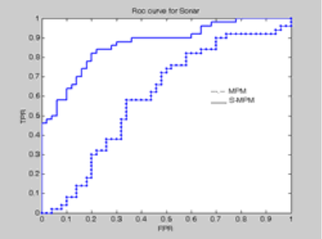

Fig. 1. ROC curves of Sonar.

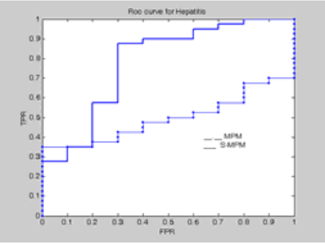

Fig.2. ROC curves of Hepatitis.

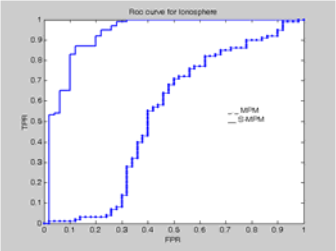

Fig. 3. ROC curves of Ionosphere.

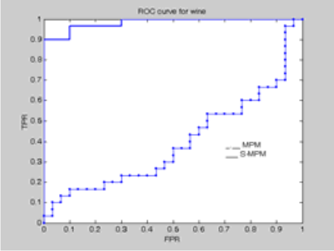

Fig. 4. ROC curves of Wine.

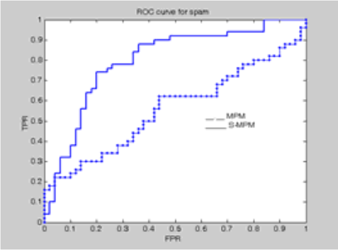

Fig. 5. ROC curves of Spam.

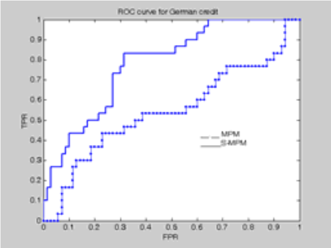

Fig. 6. ROC curves of German credit.

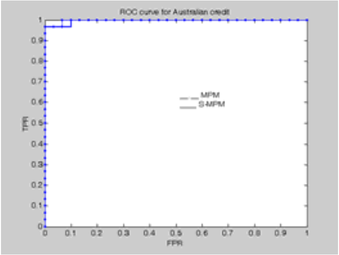

Fig. 7. ROC curves of Australian credit.

6. Conclusion and Remarks

This paper proposes two feature selections for MPM by incorporating a robust L$_1$-norm into the objective function of MPM to accomplish feature selection and classifier training simultaneously, and demonstrates their performances on public datasets. Through detailed comparisons, our models always select the least number of features and maintain high test accuracy. This indicates that the proposed models are superior to the MPM in most datasets, and simulation results show also the effectiveness of the proposed algorithms.

The approach in this paper can also be extended to formulate nonlinear version using very few support vectors. Assume that the discriminating hyper plane be, $\{x \mid \beta^T k(x) = b\}$, which divides the feature space into two subsets $\{x \mid \beta^T k(x) > b\}$ and $\{x \mid \beta^T k(x) < b\}$, where the kernel k is a function obeying the Mercer conditions[18]. We would like to find a decision hyper plane utilizing very small number of these vectors or, in other words, the goal is to find sparse vector β, which can be approximated by the L$_1$-norm of β .

Assume that $k_1 = k(X_1)$ be a random vector corresponding to class 1 while $k_2 = k(X_2)$ be another random vector belong to class 2. Let the means of k_1 and k_2 be $\overline{k}_1$ and $\overline{k}_2$ respectively and the covariance be $\overline{\Sigma}_1$ and $\overline{\Sigma}_2$ respectively. Using the Chebychev bound (6), the feature selection for nonlinear MPM can be formulated as:

$$\begin{aligned} &\min_{\alpha,\beta,\,b}\ (1-\lambda)\|\beta\|_1+\lambda\alpha \\ s.t.\quad &\sqrt{\beta^T\overline{\Sigma}_1\beta}\le\sqrt{\frac{\alpha}{1-\alpha}}(\beta^T\overline{k}_1-b), \\ &\sqrt{\beta^T\overline{\Sigma}_2\beta}\le\sqrt{\frac{\alpha}{1-\alpha}}(b-\beta^T\overline{k}_2), \\ &\beta^T\overline{k}_1-b\ge 0,\ b-\beta^T\overline{k}_2\ge 0. \end{aligned} \tag{20}$$

A similar analysis is carried out for nonlinear S-MPM. It can also be reformulated as a fractional programming which can be solved by the QI algorithm. We believe that this can have significant advantages for data-mining problems.

In this paper, we have only considered the binary cases because multi-class problems can be easily approached via standard techniques, such as the one vs. others and the one vs. one technique.

Acknowledgements

This work is supported by National Nature Science Foundation of China (10771213) and Chinese Universities Scientific Fund (2009, 2, 05).

References

1. M. Prasad. Online Feature Selection for Classifying Emphysema in HRCT Images. *International Journal of Computational Intelligence Systems.* **1**(2), 127-133(2008).
2. Z. Wei and D.Miao, N-grams based feature selection and text representation for Chinese Text Classification. *International Journal of Computational Intelligence Systems*.**2**(4), 365 – 374(2009).
3. D. Donoho and X. Huo, Uncertainty principles and ideal atomic decomposition, *IEEE Trans. on Information Theory*, **47** (7), 2845–2862(2001).
4. X. Peng, I. King and M.R.Lyu. Feature Selection based on Minimum Error Minimax Probability Machine. *IJPRAI*, **21**(8),1279-1292(2007).
5. C. Bhattacharyya, L.R.Grate and A.Rizki, Simultaneous classification and relevant feature identification in high-dimensional spaces: application to molecular profiling data. *Signal Processing* **83**, 729-743(2003).
6. Shlens J., A Tutorial on Pricipal Component Analysis, *www.snl.salk.edu/~shlens/pub/notes/pca.pdf (2009).*
7. G. R. G. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan. Minimax probability machine. *In Advances in Neural Information Processing Systems*, **14**(2002).
8. M.S.Bazarra, H.D. Sherali, and C.M. Shetty. *Nonlinear Programming: Theory and Algorithms*. (3rd Edition, IIIE Transactions,2008).
9. M. Lobo, L. Vandenberghe, S. Boyd and H. Lebret. Applications of second order cone programming. *Linear Algebra Appl*. **284**, 193–228 (1998).
10. C. Bhattacharyya. Second order cone programming formulations for feature selection. *Journal of Machine Learning Research*,**5**,1417–1433(2004).
11. W. Marshall and I. Olkin. Multivariate Chebychev inequalities. *Annals of Mathematical Statistics*, **31**(4),1001-1014(1960).
12. R.W.Freund and F. Jarre,: Solving the sum-of-ratios problem by an interior-point method. *Journal of Global Optimization*, **19**, 83-102 (2001).
13. S.-J. Kim, A. Magnani, and S. Boyd. Robust Fisher discriminant analysis. *In Advances in Neural Information Processing Systems*.(2006).
14. C.L. Blake and C.J. Merz: UCI repository of machine learning databases (University of California. 1992) *www.icsuci.edu/~mlearn/MLRepository.html.*
15 T.Fawcett. An introduction to ROC analysis, *Pattern Recognition Letters* .**27**, 861–874(2006).
16. K. Huang, H.Yang, I. King and M.L. Lyu. Maximizing sensitivity in medical diagnosis using biased minimax probability machine. *IEEE Transactions on Biomedical Engineering*, **53**(5) ,821–831(2006).
17. J.F. Sturm. Using SeDuMi 1.03, a MATLAB toolbox for optimization over symmetric cones. (1999). *http://www2.Unimaas.nl/sturm/software/sedumi.html.*
18. C.J.C.Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*. **2**(2), 121-167(1998).