

# Twin Minimax Probability Machine for Handwritten Digit Recognition

<sup>1</sup>Zhijie Xu, <sup>2</sup>Jianqin Zhang and <sup>3</sup>Hengyou Wang

<sup>1,3</sup>*School of Science, Beijing University of Civil Engineering and Architecture, Beijing 102616, China*

<sup>2</sup>*Urban Geomatics of National Administration of Surveying, Mapping and Geoinformation, Beijing University of Civil Engineering and Architecture, Beijing 100044, China*

*xuzhijie@bucea.edu.cn<sup>1</sup>, zhangjianqin@bucea.edu.cn<sup>2</sup>, wanghengyou@bucea.edu.cn<sup>3</sup>*

## Abstract

*Handwritten digit recognition is a task of great importance in many applications. There are different challenges faced while attempting to solve this problem. It has drawn much attention from the field of machine learning and pattern recognition. Minimax probability machine (MPM) is a novel method in machine learning and data mining. In this paper, we present an extension algorithm for MPM, which is named twin minimax probability machine (TWMPM). TWMPM generates two hyperplanes to improve the classification accuracy. Experiment results on several data sets from the UCI repository demonstrate that TWMPM can improve performance of MPM in most cases. The proposed method is used for recognizing the handwritten digits provided in the MNIST data set of images of handwritten digits (0-9). The testing accuracies are improved comparing with MPM.*

**Keywords:** *Handwritten Digit Recognition, Minimax Probability Machine, Two Hyperplanes*

## 1. Introduction

Machine recognition of handwritten digits is widely used and plays an important role in, for example, office automation, bank check recognition, mail resorting, and data entry applications. It has always been a challenging task due to the writers does not always write the same character in exactly the same way, so that the handwritten digits are not always of the same size, thickness, orientation and position relative to the margins.

Many systems and classification methods have been proposed for handwritten digit recognition in the past years. In early years, template matching based recognition techniques were used [1]. This kind of methods design templates or prototypes artificially. Because of the shape variability of handwritten digits, these methods are not able to yield high recognition accuracies. Recently, classification methods based on machine learning have drawn more and more attention from the field of handwriting recognition [2]. Learning from large sample data, the recognition accuracies are improved significantly. Besides of Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) is another popular method based on learning, which has been applied to solve this problem and studied actively and applied widely in many other pattern recognition applications.

A common paradigm for handwritten digit recognition based on machine learning is concentrated around two primary stages, feature extraction and classification. Feature extraction aims at reducing the dimensionality and capturing the essence of the pattern, so

that which is advantageous with respect to recognition tasks. In the stage of classification, a mapping from inputs  $x$  to outputs  $y$ , where  $y \in \{0, 1, \dots, 9\}$ , is constructed, and predictions on novel inputs can be made using the learned decision rules.

In this paper, we consider the handwritten digit recognition method using the combination of structural feature and the statistical feature and an extension algorithm for MPM.

Minimax probability machine (MPM) [3] is a novel method in the field of machine learning and data mining. It has shown many advantages and has drawn much more attention in machine learning. MPM makes use of recent work on moment problems and semidefinite optimization to obtain distribution-free results for linear discriminants. And nonlinear classification can be realized by exploiting Mercer kernels. Based on MPM, Huang *et al.* [4] propose a model named Biased Minimax Probability Machine to handle the tasks of learning from imbalanced data, and the performance is shown to be the best comparing with other competitive methods. The minimax probability machine regression (MPMR) algorithm has been proposed in [5]. The applications involve medical diagnosis, face recognition, image retrieval, intrusion detection and so on [6-9].

TWSVM has been proposed recently and has shown significant advantage in many applications. TWSVM seeks two nonparallel proximal hyperplanes such that each hyperplane is closer to one of two classes and as far as possible from the other class. It is implemented by solving two smaller Quadratic Programming Problems (QPPs) rather than a single large QPP in the classical SVM. Experimental results in [10] and [11] have shown that TWSVM outperforms both standard SVM and GEPSVM in the most cases. In addition, TWSVM is excellent at dealing with some certain probability model data (such as "Cross Planes" data). Thus, the methods of constructing the nonparallel hyperplanes and the extensions of TWSVM have been studied extensively [11-16].

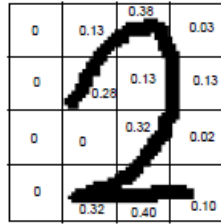
In this paper, we combine the idea of TWSVM with MPM to propose an extension algorithm for MPM, which is named twin minimax probability machine (TWMPM). TWMPM takes advantage of TWSVM and MPM. On one hand, TWMPM generates two hyperplanes to improve the performance of MPM. On the other hand, TWMPM avoids making distributional assumptions about the class-conditional densities. Experiment results on several data sets from the UCI repository demonstrate that TWMPM can improve performance of MPM in most cases. Combining with the extraction of four different feature groups, the proposed method is used for recognizing the handwritten digits provided in the MNIST data set of images of hand written digits (0-9). The testing accuracies are improved comparing with MPM.

The remaining parts of the paper are organized as follows. In Section 2, the features extracted for the classification method are presented. In Section 3, MPM and TWSVM are introduced briefly, and the details of TWMPM are described. In section 4, experimental results are presented and discussed. Finally, conclusions are presented in section 5.

## 2. Feature Extraction

Four different groups of features including the structural features and the statistical features are extracted from each digit image: coarse grid features, stroke density features, contour features and Kirsch edge features.

The first 16 features are the simple coarse grid features. The image is divided into  $4 \times 4$  patches of equal size. The proportions of black pixels in each patch are computed as the features (see Figure 1).



**Figure 1. Coarse Grid Features**

The second feature family contains 28 stroke density features. The image is scanned from the horizontal, vertical and diagonal directions, respectively. The scan is conducted with an interval of 4 rows or columns, and the times the scanning lines of different directions intersects with the strokes are provided as features (7 features every direction) (see figure 2).



**Figure 2. Stroke Density Features**

The third group of features is composed of 4 contour features. Scan the image from left to right, top to bottom, right to left and bottom to top, respectively. Compute the proportion of the white pixels before the first black pixel the scanning line intersects in all pixels (see figure 3).



**Figure 3. Contour Features**

The last group of features is the Kirsch edge features. Due to the strokes of the handwritten digits are simple, the local directional information of the edges would be an appropriate feature for recognizing the characters. The outer edges are determined by scanning the image from left to right, top to bottom, right to left and bottom to top, respectively. For each edge pixel, 4 kirsch features are computed using the Kirsch templates of the horizontal, vertical and both diagonal directions. The feature vectors are again linearly rescaled to 20 features coming from the left and right edges each, 16 features coming from the top and bottom edges each.

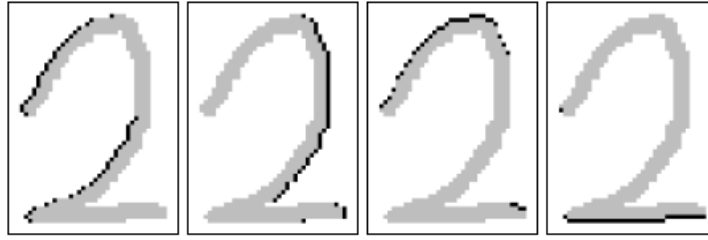


Figure 4. Kirsch Edge Features

### 3. TWMPM

In this section, we give a brief outline of MPM and TWSVM and present the twin minimax probability machine.

#### 3.1. Minimax Probability Machine (MPM)

MPM provides a worst-case bound on the probability of misclassification of future data points when data is summarized by its moments. Compared with traditional probability models, it avoids making assumptions with respect to the data distribution. We sketch the ideas of MPM simply and a more detailed description can be found in [3]. Let  $X_1$  and  $X_2$  denote  $n$  dimension random vectors representing two classes of data, with means and covariance matrices as  $X_1 \sim (\mu_1, \Sigma_1)$  and  $X_2 \sim (\mu_2, \Sigma_2)$  respectively, where  $\mu_1, \mu_2 \in \mathbb{R}^n$ , and  $\Sigma_1, \Sigma_2 \in \mathbb{R}^{n \times n}$ . The objective of MPM is to formulate the hyperplane:  $H(w, b) = \{x | w^T x = b\}$  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:

$$\begin{aligned} & \max_{\alpha, w, b} \gamma \\ & s.t. \quad \inf \Pr\{X_1 \in H_1\} \geq \gamma \\ & \quad \quad \inf \Pr\{X_2 \in H_2\} \geq \gamma. \end{aligned} \quad (1)$$

Where  $\gamma$  represent the lower bounds of the accuracy for future data classification. Moreover, (1) can be expressed as a second order cone program (SOCP).

$$\begin{aligned} & \min_w \sqrt{w^T \Sigma_1 w} + \sqrt{w^T \Sigma_2 w} \\ & s.t. \quad w^T (\mu_1 - \mu_2) = 1 \end{aligned} \quad (2)$$

#### 3.2. Twin Support Vector Machine (TWSVM)

Given the following training set for the binary classification:

$$T = \{(x_1, y_1), \dots, (x_l, y_l)\}, \quad (3)$$

where  $(x_i, y_i)$  is the  $i$ -th data point, the input  $x_i \in \mathbb{R}^n$  is a pattern, the output  $y_i \in \{-1, 1\}$  is a class label,  $i = 1, \dots, l$ , and  $l$  is the number of data points. In addition, let  $l_1$  and  $l_2$  be the number of data points in positive class and negative class respectively ( $l = l_1 + l_2$ ).

Furthermore the matrix  $A \in \mathbb{R}^{l_1 \times n}$ ,  $B \in \mathbb{R}^{l_2 \times n}$  consists of the  $l_1$  inputs of positive class and the  $l_2$  inputs of negative class respectively. For the linear case, TWSVM seeks to find two nonparallel hyperplanes in  $n$ -dimensional input space

$$(w_+ \cdot x) + b_+ = 0 \text{ and } (w_- \cdot x) + b_- = 0, \quad (4)$$

where  $w_+, w_- \in \mathbb{R}^n$ ,  $b_+, b_- \in \mathbb{R}$ . Here, each hyperplane is close to the examples of one class and far away from the examples of the other class. TWSVM is in spirit of GEPSVM [17]. But both of GEPSVM and TWSVM are different from the standard SVM. For TWSVM, each hyperplane is generated by solving a QP problem looking like the primal problem of the standard SVM. The primal problems of TWSVM can be presented as follows:

$$\begin{aligned} \min_{w_+, b_+, \xi} \quad & \frac{1}{2} \|Aw_+ + e_+ b_+\|_2^2 + C_1 e_-^T \xi \\ \text{s.t.} \quad & -(Bw_+ + e_- b_+) + \xi \geq e_-, \\ & \xi \geq 0 \end{aligned} \quad (5)$$

and

$$\begin{aligned} \min_{w_-, b_-, \eta} \quad & \frac{1}{2} \|Bw_- + e_- b_-\|_2^2 + C_2 e_+^T \eta \\ \text{s.t.} \quad & (Aw_- + e_+ b_-) + \eta \geq e_+, \\ & \eta \geq 0 \end{aligned} \quad (6)$$

Where  $C_1$  and  $C_2$  are nonnegative parameters, and  $e_1, e_2$  are vectors of ones of appropriate dimensions. In the QP problem (5), the objective function tends to keep the positive hyperplane close to the examples of positive class and the constraints require the hyperplane to be at a distance of at least 1 from the examples of negative class. The QP problem (6) has the similar property.

Once the solutions  $(w_+, b_+)$  and  $(w_-, b_-)$  of the problems (5) and (6) are obtained, a new point  $x \in \mathbb{R}^n$  is assigned to class  $i$  ( $i = +1, -1$ ), depending on which of the two hyperplanes in (4) is closer to, *i. e.*

$$\text{class } i = \arg \min_{k=+,-} \frac{|w_k^T x + b_k|}{\|w_k\|_2}. \quad (7)$$

Where  $|\cdot|$  is the absolute value. For the nonlinear case, we can refer to [11]. As for  $K$ -class classification problem, multi-TWSVM [18] constructs  $K$  hyperplanes, each hyperplane close to one of the class, and at the same time, far from the other classes. A new points closer to which hyperplane will be classified to which class.

### 3.3. TWMPM

We now present the twin minimax probability machine for linear classification in this section. Let  $X_1$  and  $X_2$  denote  $n$  dimension random vectors representing two classes of data, with means and covariance matrices as  $X_1 \sim (\mu_1, \Sigma_1)$  and  $X_2 \sim (\mu_2, \Sigma_2)$  respectively, where  $X_1 \in \mathbb{R}^{l_1 \times n}$ ,  $X_2 \in \mathbb{R}^{l_2 \times n}$ ,  $\mu_1, \mu_2 \in \mathbb{R}^n$ , and  $\Sigma_1, \Sigma_2 \in \mathbb{R}^{n \times n}$ . The goal of TWMPM is to find two nonparallel hyperplanes:

$$(w_+ \cdot x) + b_+ = 0 \text{ and } (w_- \cdot x) + b_- = 0, \quad (8)$$

so that the label of a new point  $x \in \mathbb{R}^n$  can be inferred according to which hyperplane it is closer to, where  $w_+, w_-, x \in \mathbb{R}^n$  and  $b_+, b_- \in \mathbb{R}$ . We want to determine the two hyperplanes such that each of them is closer to one of two classes and is at least one distance from the other with maximal probability with respect to all distributions having these means and covariance matrices. This is formulated as the following optimization problems:

$$\begin{aligned} \min_{w_+, b_+, \alpha} & \|Aw_+ + e_+ b_+\|_2 + \lambda_1 \alpha \\ \text{s.t.} & \sup \Pr\{w_+^T X_2 + b_+ \geq -1\} \leq \alpha \end{aligned} \quad (9)$$

and

$$\begin{aligned} \min_{w_-, b_-, \beta} & \|Bw_- + e_- b_-\|_2 + \lambda_2 \beta \\ \text{s.t.} & \sup \Pr\{w_-^T X_1 + b_- \leq 1\} \leq \beta. \end{aligned} \quad (10)$$

Consider the constraint in (9). In order to bound the probability  $\Pr\{w_+^T X_2 + b_+ \geq -1\}$ , without making any distributional assumptions, the multivariate generalization of the Chebychev Cantelli inequality [19] will be used, which is presented in Theorem 1.

**Theorem 1.** Given  $w \in \mathbb{R}^n$ ,  $w \neq 0$ ,  $b \in \mathbb{R}$ ,  $X \in \mathbb{R}^n$ , the mean and covariance of  $X$  is  $\mu \in \mathbb{R}^n$ ,  $\Sigma \in \mathbb{R}^{n \times n}$ . Then the following inequality holds:

$$\Pr\{w^T X \geq b\} \leq \frac{w^T \Sigma w}{s^2 + w^T \Sigma w}, \quad (11)$$

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

Then the constraint in (9) can be expressed as

$$\frac{w_+^T \Sigma_2 w_+}{(-1 - b_+ - w_+^T \mu_2)_+^2 + w_+^T \Sigma_2 w_+} \leq \alpha. \quad (12)$$

That is equivalent to the following inequalities:

$$\begin{aligned} -1 - b_+ - w_+^T \mu_2 & \geq \sqrt{\frac{1-\alpha}{\alpha}} \sqrt{w_+^T \Sigma_2 w_+} \\ 1 + b_+ + w_+^T \mu_2 & \leq 0 \end{aligned} \quad (13)$$

Similarly, the constraint in (10) can be expressed as

$$\begin{aligned} -1 + b_- + w_-^T \mu_1 & \geq \sqrt{\frac{1-\beta}{\beta}} \sqrt{w_-^T \Sigma_1 w_-} \\ -1 + b_- + w_-^T \mu_1 & \geq 0 \end{aligned} \quad (14)$$

Let  $\Sigma_1 = C_1^T C_1$ ,  $\Sigma_2 = C_2^T C_2$ ,  $\sqrt{\frac{1-\alpha}{\alpha}} = \frac{1}{\xi}$ ,  $\sqrt{\frac{1-\beta}{\beta}} = \frac{1}{\eta}$ , where  $\xi, \eta \in \mathbb{R}$ ,  $C_1, C_2 \in \mathbb{R}^{n \times n}$ ,

then the problems (9) and (10) turn to be the following problems (15) and (16):

$$\begin{aligned} \min_{w_+, b_+, \xi} & \|Aw_+ + e_+ b_+\|_2 + \lambda_1 \frac{\xi^2}{1 + \xi^2} \\ \text{s.t.} & \|C_2 w_+\|_2 \leq \xi(-w_+^T \mu_2 - b_+ - 1) \\ & w_+^T \mu_2 + b_+ + 1 \leq 0 \end{aligned} \quad (15)$$

$$\begin{aligned} \min_{w_-, b_-, \eta} & \|Bw_- + e_- b_-\|_2 + \lambda_2 \frac{\eta^2}{1 + \eta^2} \\ \text{s.t.} & \|C_1 w_-\|_2 \leq \eta(w_-^T \mu_1 + b_- - 1) \\ & w_-^T \mu_1 + b_- - 1 \geq 0. \end{aligned} \quad (16)$$

For a fixed  $\xi$  and  $\eta$ , the optimization over  $w_+, b_+, w_-, b_-$  in (15) and (16) turn to be convex optimization problem. We can solve them in various ways, for example using interior-point methods. So we can try different  $\xi$  and  $\eta$  to find the optimums that minimize the objective functions. In this paper, we use the Genetic algorithm to find the optimal  $\xi$  and  $\eta$ , which is briefly summarized as follows.

---

**Algorithm 1.** Genetic algorithm for problem (15) and (16).

---

1. Initialise the population.

a) Initialise *popsiz*e, *crossover\_rate*, *mutation\_rate*, *sample\_rate*.

b) Create a random population of candidate solutions (individuals) of *popsiz*e.

c) Solve (15) and (16). Using the objective function values as the fitness function, evaluate all individuals.

d) Store the best evaluated individual as best-ever individual.

2. Selection and generation.

a) Sample the individuals according to their fitness, so that the individuals with higher fitness appear with higher probability, i.e. the  $j$ th individual is selected with the probability of

$$f_j / \sum_k f_k.$$

b) Apply crossover with probability *crossover\_rate*.

c) Apply mutation with probability *mutation\_rate*.

d) Evaluate all individuals using the fitness function.

e) Update the best-ever individual.

3. If termination criteria satisfied, provide the best-ever individual as a solution. Otherwise go to step 2.

---

Once  $w_+, b_+, w_-, b_-$  are obtained from (15) and (16), the separating planes

$$(w_+ \cdot x) + b_+ = 0 \text{ and } (w_- \cdot x) + b_- = 0 \quad (17)$$

are known. A new point  $x \in \mathbb{R}^n$  is then assigned to class  $i$  ( $i = +1, -1$ ), depending on which of the two hyperplanes in (17) it is closer to, i.e.

$$\text{class } i = \arg \min_{k=+,-} \frac{|w_k^T x + b_k|}{\|w_k\|_2}. \quad (7)$$

where  $|\cdot|$  is the absolute value. For nonlinear classification tasks, the kernelization trick can be used. To save space, we omit the kernelization in this paper.

## 4. Experiments

In this section, we show the performance of TWMPM on several data sets.

### 4.1. UCI Datasets

In this section, we perform the methods on the UCI Machine Learning Repository datasets, which is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms [20]. For each dataset, 90% of the points are randomly selected for training, and the rest points (10%) are used as testing data. For each dataset, the experiment is repeated 10 times. The average accuracies are summarized in Table 1. For comparison, the results of MPM with linear kernel are also listed. The results demonstrate that TWMPM outperforms MPM in most cases.

**Table 1. The Testing Accuracy (%) on UCI Datasets**

Data set (Ins×Fea×Class)	TWMPM Accuracy(%)	MPM Accuracy(%)
Breast cancer 683×10×2	97.6	97.2
Pima diabetes 768×8×2	75.3	73.8
Ionosphere 351×34×2	86.5	85.4
Sonar 208×60×2	74.4	75.1

### 4.2. MNIST Data Set

In this section, we use TWMPM and MPM to classify the MNIST digits. The MNIST database of handwritten digits contains a training set of 60,000 examples, and a test set of 10,000 examples. Each example is a handwritten digit of 0-9, and the size is 28×28 pixel. Table 2 shows the composition of the database. For each number, we randomly choose 200 samples from the training set and 50 samples from the testing set, forming a training set of 2000 samples and a testing set of 500 samples. Figure 5 shows a fragment from the data set. Using MPM and TWMPM to classify these handwritten digits, the average recognition accuracy on the training set and the testing set are showed in Table 3. These results illustrate the feasibility and effectiveness of the proposed approach in handwritten digit recognition.

## 5. Conclusion

In this paper, we propose an extension algorithm for MPM, which is named twin minimax probability machine (TWMPM). Given the reliable estimation of the mean and

**Table 2. The Composition of the Database**

Class	Training set: 60000 examples	Test set: 10000 examples
0	5923 (9.87%)	980 (9.80%)



1	6742 (11.24%)	1135 (11.35%)
2	5958 (9.93%)	1032 (10.32%)
3	6131 (10.22%)	1010 (10.10%)
4	5842 (9.74%)	982 (9.82%)
5	5421 (9.04%)	892 (8.92%)
6	5918 (9.86%)	958 (9.58%)
7	6265 (10.44%)	1028 (10.28%)
8	5851 (9.75%)	974 (9.74%)
9	5949 (9.92%)	1009 (10.09%)

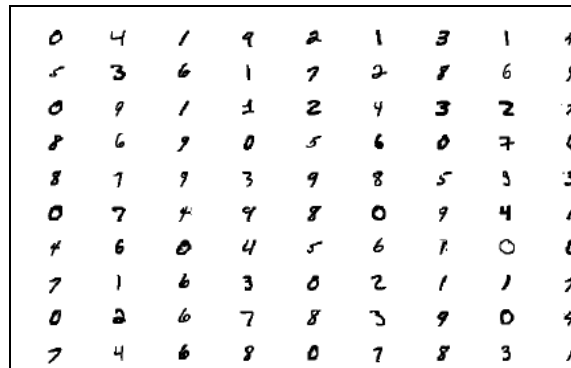


Figure 5. A Fragment from the MNIST Database

Table 3. The Recognition Accuracy of TWMPM and MPM on MNIST Datasets

Recognition Accuracy (%)	TWMPM	MPM
Training Set	76.75	74.8
Testing Set	75.2	71.2

covariance of data, TWMPM generates two hyperplanes by controlling the upper bound of the misclassification probabilities and the distances between the hyperplanes and the points of two classes. Using two hyperplanes, TWMPM improves the performance of MPM. We evaluate our method on several datasets from the UCI Machine Learning Repository datasets and MNIST database of handwritten digits. Experimental results demonstrate that the performance of TWMPM is better than MPM in the case of linear classification. In the future work, more efficient methods to solve the optimization problem and the multi-class classification method are under our consideration.

### Acknowledgements

This work has been partially supported by the Key Laboratory for Urban Geomatics of National Administration of Surveying, Mapping and Geoinformation Foundation (20111216N), Beijing Municipal Commission of Education, Science and Technology Development (KM201210016014) and Funding project for training excellent talent of Beijing city (2011D005017000005).

## References

- [1] S. Mori, C. Y. Suen and K. Yamamoto, "Historical review of OCR research and development", In: Proc. IEEE, vol. 80, no. 7, (1992), pp. 1029–1058.
- [2] S. Arora, D. Bhattacharjee and M. Nasipuri, "Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition", International Journal of Computer Science Issues, vol. 7, no. 3, no. 6, (2010), pp. 18–26.
- [3] G. Lanckriet, L. E. Ghaoui, C. Bhattacharyya and M. I. Jordan, "Minimax probability machine", In: Advances in Neural Information Processing Systems, (2001), pp. 801–807.
- [4] K. Huang, H. Yang, I. King and M. R. Lyu, "Learning classifiers from imbalanced data based on biased minimax probability machine", In: Computer Vision and Pattern Recognition, 2004, CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 2, (2004), pp. II–558. IEEE.
- [5] T. Strohmann and G. Z. Grudic, "A formulation for minimax probability machine regression", In: Advances in Neural Information Processing Systems, (2002), pp. 769–776.
- [6] C.-H. Hoi and M. R. Lyu, "Robust face recognition using minimax probability machine", Multimedia and Expo, Icme 04, IEEE International Conference on. IEEE, (2004), vol. 2, pp. 1175 - 1178.
- [7] K. Huang, H. Yang, I. King, M. R. Lyu and L. Chan, "Biased minimax probability machine for medical diagnosis", In: AMAI (2004).
- [8] X. Peng and I. King, "A biased minimax probability machine-based scheme for relevance feedback in image retrieval", Neurocomputing, vol. 72, no. 7, (2009), pp. 2046–2051.
- [9] Z. g. Chen and D. y. Li, "Intrusion detection based on kernel fisher discriminant analysis and minimax probability machine classifier", Journal of University of Electronic Science and Technology of China, vol. 36, no. 6, (2007), p. 1192.
- [10] R. Khemchandani and S. Chandra, "Twin support vector machines for pattern classification", Pattern Analysis and Machine Intelligence, IEEE Transactions on vol. 29, no. 5, (2007), pp. 905–910.
- [11] M. A. Kumar and M. Gopal, "Application of smoothing technique on twin support vector machines", Pattern Recognition Letters, vol. 29, no. 13, (2008), pp. 1842–1848.
- [12] M. Arun Kumar and M. Gopal, "Least squares twin support vector machines for pattern classification", Expert Systems with Applications, vol. 36, no. 4, (2009), pp. 7535–7543.
- [13] S. Ghorai, A. Mukherjee and P. K. Dutta, "Nonparallel plane proximal classifier", Signal Processing, vol. 89, no. 4, (2009), pp. 510–522.
- [14] R. Khemchandani and S. Chandra, "Optimal kernel selection in twin support vector machines", Optimization Letters, vol. 3, no. 1, (2009), pp. 77–88.
- [15] X. Peng, "Efficient twin parametric insensitive support vector regression model", Neurocomputing, vol. 79, (2012), pp. 26–38.
- [16] Y. H. Shao and N. Y. Deng, "A coordinate descent margin based-twin support vector machine for classification", Neural Networks, vol. 25, (2012), pp. 114–121.
- [17] O. L. Mangasarian and E. W. Wild, "Multisurface proximal support vector machine classification via generalized eigenvalues", Pattern Analysis and Machine Intelligence, IEEE Transactions on vol. 28, no. 1, (2006), pp. 69–74.
- [18] Z. Wang, J. Chen and M. Qin, "Non-parallel planes support vector machine for multi-class classification", In: Logistics Systems and Intelligent Management, International Conference on IEEE, (2010), pp. 581 - 585.
- [19] A. W. Marshall and I. Olkin, "Multivariate chebyshev inequalities", The Annals of Mathematical Statistics vol. 31, no. 4, (1960), pp. 1001–1014.
- [20] A. Asuncion and D. J. Newman, "Uci machine learning repository", (2007).