A Multiclass Kernel Perceptron Algorithm

Jianhua Xu
Department of Computer Science
Nanjing Normal University
Nanjing 210097, China
E-mail: xujianhua@njnu.edu.cn

Xuegong Zhang
Department of Automation
Tsinghua University
Beijing 100084, China
E-mail: zhangxg@tsinghua.edu.cn

Abstract—**Original kernel machines (e.g., support vector machine, least squares support vector machine, kernel Fisher discriminant analysis, kernel perceptron algorithm, and etc.) were mainly designed for binary classification. How to effectively extend them for multiclass classification is still an ongoing research issue. Rosenblatt's linear perceptron algorithm for binary classification and its corresponding multiclass linear version are the simplest learning machines according to their algorithmic routines. Kernel perceptron algorithm for binary classification was constructed by extending linear perceptron algorithm with Mercer kernel. In this paper, a multiclass kernel perceptron algorithm is proposed by combining multiclass linear perceptron algorithm with binary kernel perceptron algorithm, which can deal with multiclass classification problem directly and nonlinearly in a simple iterative procedure. Two artificial examples and four benchmark datasets are used to evaluate the performance of our multiclass method. The experimental results show that our algorithm could achieve the good classification performance.**

I. INTRODUCTION

Kernel machines (e.g., support vector machine (SVM), kernel Fisher discriminant analysis (KFD), least squares support vector machine (LS-SVM), and etc.) were originally designed for binary classification [1][2][3][4][5]. Currently there are two strategies to deal with multiclass classification. One strategy is to reduce multiclass classification into several binary classification problems. The "one-against-other" and "one-against-one" techniques were used to construct multiclass SVM methods [6][7]. Such two techniques can also be utilized to design multiclass classifiers according to other binary kernel machines. The other strategy is to consider all classes in one optimization problem by extending binary classification learning machines, which can design so-called "all-together" methods. Through extending binary SVM, several "all-together" multiclass classifiers were presented [3][8][9]. It is noted that the number of variables in their quadratic problems is about the product of the number of total samples and the number of classes. It is very difficult to solve these optimization problems directly. Hsu and Lin [10] used decomposition technique to solve two "all-together" methods [3][8], and compared two types of multiclass SVM methods using large benchmark datasets elaborately. It was shown that their classification accuracy is very similar, the "one-against-one" method is more suitable for practical applications, and "all-together" methods need fewer support vectors. However, in [9] the experimental results showed that the classification accuracy of "all-together" method is better than that of "one-against-other" method for small benchmark datasets generally. For multiclass SVM classifiers only, there exist some contradictive experimental results up to now. Thus, how to effectively extend binary kernel machines for multiclass classification and elaborately analyze their performances is still an ongoing research issue.

For binary classification, Rosenblatt's linear perceptron algorithm with fixed increment learning rule (or simply linear perceptron algorithm) is considered as the simplest learning machine in machine learning and pattern recognition [11]. In [12], it was extended to construct an "all-together" linear classifier considering all classes simultaneously. Based on the kernel trick in SVM, KFD and LS-SVM, a nonlinear kernel perceptron algorithm for binary classification was presented by extending linear perceptron algorithm [13][14], which is one of the simplest kernel machines according to its algorithmic structure. In this paper, we combine binary kernel perceptron algorithm with multiclass linear perceptron algorithm to construct a multiclass kernel perceptron algorithm. It is a simple iterative procedure that can deal with all training data simultaneously in one optimization problem. Two artificial examples and four benchmark datasets are used to evaluate its performance in detail experimentally. The experimental results demonstrate that our multiclass method could behave well.

This paper is organized as follows. In next section, binary and multiclass linear perceptron algorithms are simply reviewed. In section III, a kernel perceptron algorithm is introduced briefly and a multiclass kernel perceptron algorithm is proposed in details. Then the experimental results from two artificial examples and four benchmark datasets are provided and analyzed. Finally, the conclusions are presented.

II. MULTICLASS LINEAR PERCEPTRON ALGORITHM

In this section, we simply review linear perceptron algorithms for binary and multiclass classification. Let the training set be

$$\{(\mathbf{x}_1, y_1), \mathrm{L}\ , (\mathbf{x}_l, y_l)\}\ , \qquad (1)$$

0-7803-9422-4/05/$20.00 ©2005 IEEE

where $\mathbf{x}_i \in R^d$ denotes the ith input vector and l the size of training set. For binary classification (ω_1, ω_2), $y_i \in \{+1, -1\}$. That is, if $\mathbf{x}_i \in \omega_1$ then $y_i = +1$, otherwise $y_i = -1$. For multiclass classification $(\omega_1, ..., \omega_c)$, $y_i \in \{1, 2, ..., c\}$, where c is the number of classes.

The linear discriminant function for binary classification is

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b, \qquad (2)$$

where $\mathbf{w} \in R^d$ and $b \in R$ are the weight vector and threshold. For some new vector $\mathbf{x}$, the decision rule is that if $f(\mathbf{x}) > 0$ then $\mathbf{x} \in \omega_1$, otherwise $\mathbf{x} \in \omega_2$.

Finding out the weight vector and threshold in (2) can be done by linear perceptron algorithm [11][12]. For some training sample $(\mathbf{x}_q, y_q)$, if it is misclassified by current weight vector and threshold, these parameters can be updated using the following rule,

$$\text{if } y_q f(\mathbf{x}_q) \le 0, \text{ then } \mathbf{w} \Leftarrow \mathbf{w} + y_q \mathbf{x}_q,\ b \Leftarrow b + y_q. \qquad (3)$$

If there exist a pair of $(\mathbf{w}, b)$, all training samples can be classified correctly (i.e., $f(\mathbf{x}_q) y_q > 0, q = 1, ..., l$), the sample set is said to be linearly separable. It has been proven that this procedure converges to a solution within limited steps starting from any arbitrary initial values when the training samples are linearly separable (i.e., perceptron convergence theorem) [12].

In [12], this binary perceptron algorithm was extended to construct a multiclass linear perceptron algorithm considering all classes at once. In this case, c linear discriminant functions have to be defined, i.e.,

$$f_i(\mathbf{x}) = \mathbf{w}_i^{\mathrm{T}}\mathbf{x} + b_i \quad i = 1, ..., c, \qquad (4)$$

where $\mathbf{w}_i$ and b_i denote the weight vector and threshold of the ith discriminant function. Now, for some input vector $\mathbf{x}$, if $f_i(\mathbf{x}) > f_j(\mathbf{x})$ for all $j \ne i$, assign this vector to the ith class, i.e., $\mathbf{x} \in \omega_i$.

For some training sample $\mathbf{x}_q \in \omega_i$, if there is at least one $j \ne i$ for which $f_i(\mathbf{x}_q) \le f_j(\mathbf{x}_q)$, this vector is referred to as a misclassified sample. In multiclass linear perceptron algorithm, such a misclassified sample is used to modify some weight vectors and thresholds in (4) according to the learning rule,

$$\begin{aligned} \mathbf{w}_i &\Leftarrow \mathbf{w}_i + \mathbf{x}_q, b_i \Leftarrow b_i + 1 \\ \mathbf{w}_j &\Leftarrow \mathbf{w}_j - \mathbf{x}_q, b_j \Leftarrow b_j - 1 \end{aligned} . \qquad (5)$$

That is, the weight vector and threshold for the desired class is increased by the misclassified sample, the vector and threshold for the incorrectly chosen class is decreased, and all others are left unchanged.

For multiclass classification problem, if there also exist the weight vectors and thresholds that can classify all training samples correctly, this training set is still said to be linearly separable. Based on Kesler's construction [12], this multiclass problem can be reduced into the binary one. Thus its convergence for linearly separable case can be proven using perceptron convergence theorem for binary classification.

It is attractive that linear perceptron algorithms for binary and multiclass classification are the simplest learning machines according to their algorithmic routines in machine learning and pattern recognition.

III. MULTICLASS KERNEL PERCEPTRON ALGORITHM

On the basis of reviewing kernel perceptron algorithm for binary classification briefly, we generalize it to construct a multiclass kernel perceptron algorithm in this section.

In order to enhance the classification ability, based on kernel functions satisfying Mercer condition, a nonlinear form of binary linear perceptron algorithm was proposed [13][14], which is referred to as kernel perceptron algorithm. Now, the nonlinear discriminant function with kernel is defined as,

$$f^k(\mathbf{x}) = \sum_{m=1}^{l} \alpha_m k(\mathbf{x}_m, \mathbf{x}) + \beta, \qquad (6)$$

where $\alpha_m, m = 1, ..., l$ and β are its parameters to be solved, $k(\cdot,\cdot)$ indicates the kernel functions satisfying Mercer condition [2][3], which represents a inner product of two vectors in some feature space. The widely used kernel functions are polynomial and RBF kernels, i.e.,

$$k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^{\mathrm{T}}\mathbf{y})^p \text{ and} \qquad (7)$$

$$k(\mathbf{x}, \mathbf{y}) = \exp(-\tfrac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}), \qquad (8)$$

where p indicates the degree of polynomial kernel and σ the width of RBF kernel. In this case, for some new vector $\mathbf{x}$, if $f^k(\mathbf{x}) > 0$, then $\mathbf{x} \in \omega_1$; otherwise $\mathbf{x} \in \omega_2$.

Here, for some input training vector $(\mathbf{x}_q, y_q)$, if it is misclassified by current parameters, it is utilized to modify those parameters in (6) through the following learning rule,

$$\begin{aligned} &\text{if } y_q f(\mathbf{x}_q) \le 0, \\ &\text{then } \alpha_i \Leftarrow \alpha_i + y_q k(\mathbf{x}_i, \mathbf{x}_q),\ \beta \Leftarrow \beta + y_q \end{aligned} . \qquad (9)$$

A training set is considered as linearly separable case in the feature space if there exist the parameters that can classify all training samples correctly. But it is possible that this training set is nonlinearly separable in the original input space. The convergence of kernel perceptron algorithm was proven by percpetron convergence theorem through a proper transform [14].

Based on generalizing technique in section II from binary classifcation to multiclass classification for linear perceptron algorithm, now we extend this binary kernel perceptron algorithm to construct a multiclass kernel perceptron algorithm. It must include c nonlinear discriminant functions with kernel, i.e.,

$$f_i^k(\mathbf{x}) = \sum_{m=1}^{l} \alpha_m^i k(\mathbf{x}_m, \mathbf{x}) + \beta_i, i = 1, \ldots, c \quad , \tag{10}$$

where α_m^i and β_i imply the parameters of the *i*th nonlinear discriminant function. For a new input vector $\mathbf{x}$, it is assigned to the *i*th class (i.e., $\mathbf{x} \in \omega_i$) if $f_i^k(\mathbf{x}) > f_j^k(\mathbf{x})$ for all $j \neq i$.

Similarly, for some training sample $\mathbf{x}_q \in \omega_i$, if there is at least one $j \neq i$ for which $f_i^k(\mathbf{x}_q) \leq f_j^k(\mathbf{x}_q)$, this vector is defined as a misclassified sample. In this case, we can use such a misclassified sample to update some parameters in (10) using the learning rule as follows,

$$\begin{aligned} \alpha_m^i &\Leftarrow \alpha_m^i + k(\mathbf{x}_m, \mathbf{x}_q), \beta_i \Leftarrow \beta_i + 1 \\ \alpha_m^j &\Leftarrow \alpha_m^j - k(\mathbf{x}_m, \mathbf{x}_q), \beta_j \Leftarrow \beta_j - 1 \end{aligned} . \tag{11}$$

It implies that, the parameters for the desired class are increased by the misclassified sample, the parameters for the incorrectly chosen class are decreased, and all others remain unchanged.

If we can find out the parameters that can classify all training samples correctly, such a training set is considered as linearly separable set in the feature space. We can prove the convergence of our method for linearly separable case using multiclass linear perceptron convergence [12] and a proper transform in [14].

In real world applications, we do not know whether the training set is linearly separable or not previously in the original input space or feature space, the maximal iterations can be set to terminate iterative procedure. When the iterative procedure is stopped, the current parameters are accepted as the final solution.

In our multiclass kernel perceptron algorithm, there only exist two parameters (i.e., kernel parameter and iterations) that affect the generalization ability of our method. That is, we have to adjust or choose them elaborately to achieve the good classification accuracy.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

We present two types of experiments to demonstrate the performance of our multiclass kernel perceptron algorithm: experiments with artificial data that can be visualized in two dimensional plane and experiments with benchmark datasets. Some experimental results from multiclass linear perceptron algorithm are provided in order to compare linear method with kernel one.

A. Experiments in two dimensional plan

To illustrate our multiclass kernel perceptron, we first provide two artificial examples: linearly separable problem with five classes and four spirals problem.

In the first example, five classes of sample vectors are represented in Fig.1 and Fig.2 as five different symbols: circles, crosses, points, squares and stars. The decision hyperplanes from multiclass linear perceptron algorithm are shown in Fig.1, which are piecewise linear. When linear kernel and polynomial kernel of 2 degree are used, the corresponding hyperplanes are illustrated in Fig.2. The decision boundaries in Fig.2 (a) are piecewise linear, while ones in Fig.2 (b) nonlinear. In this example, all training samples are classified correctly.

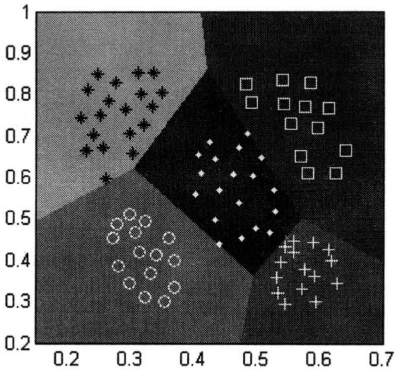
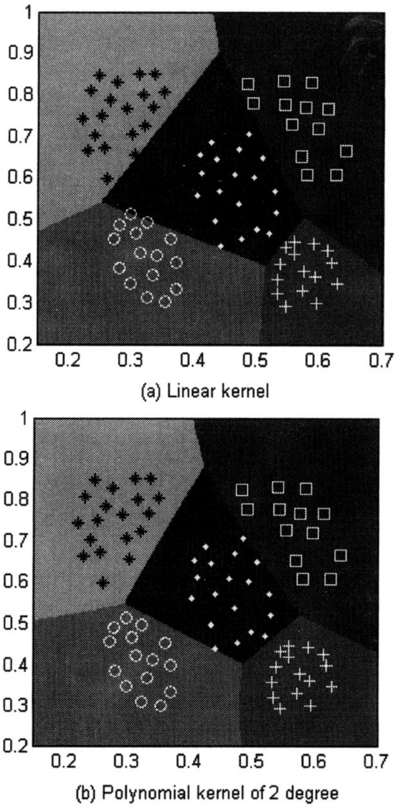


Fig. 1. Hyperplanes from Multiclass Linear Perceptron Algorithm for Linearly Separable Problem with Five Classes



(a) Linear kernel

(b) Polynomial kernel of 2 degree

Fig. 2. Hyperplanes from Multiclass Kernel Perceptron Algorithm with Linear Kernel and Polynomial Kernel of 2 Degree

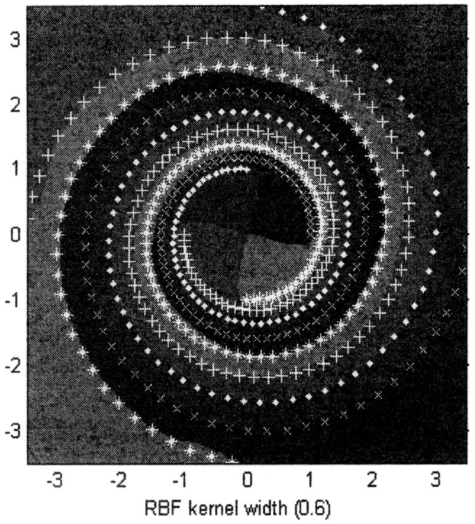


Fig. 3. Decision Hyperplanes for Four Spirals Problem Obtained by Multiclass Kernel Perceptron Algorithm with RBF Kernel Width 0.6

In four spirals problem, there are four classes ("+", "*", "x", and ".") of two dimension vectors, where each class samples lie in a spiral respectively, as shown in Fig.3. When RBF kernel with width 0.6 is utilized, the decision hyperplanes are illustrated in Fig.3. It is noted that all training vectors are recognized correctly and the boundaries are very smooth. But we also observe that some training samples are very close to boundaries.

B. Experiments with benchmark datasets

To show the performance of our multiclass method further, we choose four ten-fold cross validation datasets without missing attributes and with numerical attributes from [15], which are summarized in Table I. The last column indicates the smallest average error rates from different method in [15]. In our experiments, we scale each attribute to the interval [0,1].

TABLE I

BENCHMARK DATASETS USED IN OUR EXPERIMENTS

Dataset	Dimension	Class	Size	Error rate (%)
Iris	4	3	150	3.33
Sonar	60	2	208	14.53
Vowel	10	11	990	6.26
Wine	13	3	178	2.25

Firstly we examine mulitclass linear perceptron algorithm on these benchmark datasets. Since we do not know whether these sets are linearly separable or not, the maximal iterations are set in advance. Table II shows the average error rates and variance on training and testing sets, and corresponding maximal iterations. It is shown that the wine dataset is linearly separable for its training sets.

TABLE II

EXPERIMENTAL RESULTS FROM MULTICLASS LINEAR PERCEPTRON ALGORITHM

Dataset	Iterations	Training error rate	Testing error rate
Iris	3000	2.82 ± 1.93	5.33 ± 6.53
Sonar	2000	13.03 ± 7.56	27.01 ± 6.13
Vowel	5000	76.93 ± 5.13	77.17 ± 5.72
Wine	1000	0.00 ± 0.00	2.81 ± 4.50

For muliticlass kernel perceptron algorithm, we utilize the polynomial kernel and RBF kernel respectively. Table III shows the smallest error rates and variances on training sets and testing sets from different maximal iterations and different kernel parameters. The better results are denoted by boldface.

TABLE III

EXPERIMENTAL RESULTS FROM MULTICLASS KERNEL PERCEPTRON ALGORITHM

Dataset	Iterations	Kernel	Training error rate	Testing error rate
Iris	5000	$p=2$	3.33 ± 1.88	**4.67 ± 4.23**
	2000	$\sigma=0.3$	3.78 ± 2.70	5.33 ± 4.00
Sonar	2000	$p=9$	1.49 ± 1.06	16.37 ± 8.15
	2000	$\sigma=0.3$	0.00 ± 0.00	**14.00 ± 4.68**
Vowel	500	$p=16$	4.15 ± 0.85	9.60 ± 2.18
	500	$\sigma=0.2$	0.00 ± 0.00	**2.73 ± 1.43**
Wine	5000	$p=4$	0.00 ± 0.00	3.38 ± 3.64
	5000	$\sigma=0.6$	0.00 ± 0.00	**2.26 ± 3.72**

From Table II and III, it is observed that our multiclass kernel perceptron algorithm can effectively improve the classification performance compared with multiclass linear one. According to Table I and III, we can find out that our results on sonar and vowel datasets are better than those from [15], and the result on wine dataset is almost equal to that from [15]. Although our result on iris dataset is worse than the best one from [15], only two methods (RISE and Globoost) in [15] can obtain better performance than our method does. It is noted that in general the error rates based on RBF kernel are prior to those based on polynomial kernel. These results demonstrate that our multiclass method could behave well in these practical applications.

V. CONCLUSIONS

Rosenblatt's linear binary perceptron algorithm and its corresponding multiclass form are considered as the simplest learning machines in machine learning and pattern recognition. Kernel perceptron algorithm for binary classification is one of its nonlinear generalizations based on Mercer kernel. In this paper, a multiclass kernel perceptron algorithm considering all training samples in one optimization problem simultaneously is proposed, which combines multiclass linear perceptron algorithm with binary kernel perceptron algorithm. Such a method can deal with multiclass classification directly rather than reduce a multiclass classification problem into several binary classification problems. According to its algorithmic structure, our method is the simplest multiclass kernel algorithm and easier to be implemented in any programming languages (e.g., Matlab, C/C++, Fortran, and etc.). The experimental results on two artificial examples and four benchmark datasets illustrate that our multiclass kernel perceptron algorithm can work well.

Our further work is to reduce the computational consuming, consider margin into iterative procedure and examine more benchmark databases.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China under grant no. 60275007 and Natural Science Foundation of Jiangsu Province under grant no. BK2004142.

REFERENCES

[1] C. Cortes and V. N. Vapnik, "Support Vector Networks," *Machine Learning*, vol. 20, pp. 273-297, Sep. 1995.

[2] V. N. Vapnik, *The Nature of Statistical Learning Theory* (2nd Edition), New York: Springer-Verlag, 1999.

[3] V. N. Vapnik, *Statistical Learning Theory*, New York: Wiley, 1998.

[4] S. Mika, G. Ratsch, J. Weston, B. Scholkopf and K. R. Muller, "Fisher Discriminant Analysis with Kernels," *Proceedings of Neural Networks for Signal Processing IX*, Winsconsin, USA, 1999, pp. 41-48.

[5] J. A. K. Suykens and J. Vandewalle, "Least squares Support Vector Machines," *Neural Processing Letters*, vol. 9, pp. 293-300, Jun. 1999.

[6] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Miller, E. Sackinger, P. Simard and V. Vapnik, "Comparison of Classifier Methods: A Case Study in Handwriting Digit Recognition," *Proceedings of International Conference on Pattern Recognition*, Jerusalem, Israel, 1994, vol. 2, pp. 77-83.

[7] U. Kerbel, "Pairwise Classification and Support Vector Machines," *Advances in Kernel Methods – Support Vector Learning*, B. Scholkopf, C. J. C Burges and A. J. Somla (Eds.), Cambridge MA: MIT Press, 1999, pp. 255-268.

[8] K. Crammer and Y. Singer, "On the Learnability and Design of Output Codes for Multiclass Problems," *Machine Learning*, vol. 47, pp. 201-233, May/Jun. 2002.

[9] V. Franc and V. Hlavac, "Multi-class Support Vector Machine," *Proceedings of International Conference on Pattern Recognition*, Quebec City, Canada, 2002, pp. 236-239.

[10] C. W. Hsu and, C. J. Lin, "A Comparison of Methods for Multiclass Support Vector Machines," *IEEE Transactions on Neural Networks*, vol. 13, pp. 415-425, Mar. 2002.

[11] F. Rosenblatt, "The Perceptron: Probabilistic Model for Information Storage and Organization in the Brain, " *Psychological Review*, vol. 65, pp. 386-408, Nov. 1958.

[12] R. O. Duda, P. E. Hart and G. S. David, *Pattern Classification* (2nd Edition), New York: John Wiley & Sons, 2001

[13] J. Xu, X. Zhang and Y. Li, "Large Margin Kernel Pocket Algorithm," *Proceedings of 2001 IEEE International Conference on Neural Networks*, Washington DC, USA, 2001, pp. 1480-1485

[14] J. Xu, X. Zhang and Y. Li, "A Nonlinear Perceptron Algorithm Based on Kernel Functions," *Chinese Journal of Computers*, Vol. 25, pp. 689-695, Jul. 2002. (In Chinese).

[15] http: // www.grappa.univ-lille3.fr / ~torre / guide.php ? id=datasets, =methods and =results.