SUPPORT VECTOR MACHINE FOR MULTIPLE FEATURE CLASSIFCATION

Bing-Yu Sun *Moon-Chuen Lee*
Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Shatin, Hong Kong.

ABSTRACT

In this paper an effective method of using SVM classifier for multiple feature classification is proposed. Compared with traditional combination methods where all needed base classifiers should be trained before the decision combination, the proposed approach is to train individual classifiers and combine the decisions of these base classifiers at the same time. Thus the complexity of the training can be reduced because our proposed method involves solving only one optimization problem while several optimization problems should be solved for traditional methods. Furthermore, during the combination, our proposed approach takes into account both a base classifier's performance on the training data and its generalization ability while traditional combination approaches consider only a base classifier's performance on the training data. The experiments proved the efficiency of our proposed approach.

1. INTRODUCTION

The Support Vector Machine (SVM) was introduced by Vapnik [1] as a method for classification and function approximation and currently it has been successfully applied in many areas such as face detection, hand-written digit recognition, and so on [2] [3]. In this paper, we focus on the classification problem only.

The aim of multiple feature classification is to improve the performance of the classification by using several features. So far many multiple feature classification approaches have been developed [4][5]. This paper focuses only on the problem of using SVM classifiers for multiple feature classification problems.

A simple method of multiple feature classification is to concatenate all features to a single feature with a very high dimension [6]. This approach may get certain successes, however, on the one hand, it suffers from the curse of dimensionality, and on the other hand, sometimes the available features may be of different forms and it is hard to lump them together. So a more commonly used strategy is to train different SVM classifiers using different features and then combine the outputs of these classifiers. The key of this decision combination strategy is how to combine the output of these individual classifiers. Currently many decision combination approaches have been developed, including majority voting [7], methods based on Dempster-Shafer theory [8], the combination based on Bayesian theory [8], linear combination [9], etc. However, in these approaches all needed classifiers should be trained before decision combination, which would increase the training time. To resolve this problem, an efficient multiple feature classification approach for SVM is proposed. In this strategy the training of the individual classifiers and the decision combination are performed at the same time. So this approach can simplify the complexity of the training during the training.

2. SUPPORT VECTOR MACHINES

This section briefly reviews the basics of SVM in pattern recognition. More detailed treatment on principles and applications of SVM (such as in regression estimation and operator inversion) can be found in [1] [2].

An SVM is a binary classifier trained on a set of labeled patterns called training samples. Let $(\mathbf{x}_i, y_i) \in R^l \times \{\pm 1\}, i = 1, \ldots, N$ be such a set of training samples with inputs $\mathbf{x}_i \in R^l$, and outputs $y_i \in \{\pm 1\}$. The objective in training an SVM is to find a hyperplane which divides these samples such that all the points with the same label will be on the same side of the hyperplane, i.e., to find $\mathbf{w}$ and b. After the training, we obtain the classifier decision function, given by:

$$f_{\mathbf{w},b}(\mathbf{x}) = \mathrm{sgn}(\mathbf{w} \cdot \mathbf{x} + b) \quad (1)$$

where $\mathbf{w}$ is a coefficient vector and b is the bias of the hyperplane; sgn stands for a bipolar sign function. The hyperplane of the classifier should satisfy the following:

$$y_i[\mathbf{w} \cdot \mathbf{x}_i + b] \geq 1, i = 1, 2, \ldots, N \quad (2)$$

Among all the separating hyperplanes satisfying (2), the one with the maximal distance to the closest point is called the optimal separating hyperplane (OSH), which will result in an optimal generalization. On the other hand, in many practical situations, we may not have such an ideal hyperplane. To allow for possibilities of violating (2), some slack variables $e_i \geq 0$, can be introduced into (2), and we obtain :

$$y_i[\mathbf{w} \cdot \mathbf{x}_i + b] \geq 1 - e_i, i = 1, 2, \ldots, N \quad (3)$$

According to the structural risk minimization inductive principle, the training of an SVM is to minimize the guaranteed risk bound as follows:

$$\min J(\mathbf{w},e,b)=\frac{1}{2}\mathbf{w}\cdot\mathbf{w}+\frac{1}{2}C\sum_{i=1}^{N}e_i^2 \tag{4}$$

subject to (3).

To solve nonlinear recognition problems, we can map the data to another dot product space (called the feature space) F via a nonlinear map $\varphi: R^N \to F$, and then perform the above analysis in F. Two commonly used kernel functions for SVMs are polynomial kernels and Gaussian RBF kernels [1].

3. PROPOSED APPROACH FOR MULTIPLE FEATURE CLASSIFICATION USING SVM

Suppose there are *K* kinds of features available, and accordingly we should have *K* individual SVM classifiers:

$$y_j(\mathbf{x})=f_j(\mathbf{x})=\mathbf{w}_j\cdot\varphi_j(\mathbf{x})+b_j, j=1,2...,K \tag{5}$$

where f_j is the decision function of the *j*-th SVM classifier and $\mathbf{w}_j,\varphi_j,b_j$ are the coefficient vector , the mapping function, and the bias of this function respectively.

Then for a given sample $\mathbf{x}$, the output of the multiple SVM classifier system will be:

$$\begin{aligned} f(\mathbf{x})&=\sum_{j=1}^{K} f_j(\mathbf{x})=\sum_{j=1}^{K}\mathbf{w}_j\cdot\varphi_j(\mathbf{x})+\sum_{j=1}^{K}b_j \\ &=\sum_{j=1}^{K}\mathbf{w}_j\cdot\varphi_j(\mathbf{x})+b \end{aligned} \tag{6}$$

In traditional combination approaches, before combination, all base SVM classifiers should be trained, i.e., the values of $\mathbf{w}_j,b_j$ have been obtained. So in this situation, the label of a testing sample can be determined directly by using (6). Different from those approaches, our approach is to train and combine classifiers at the same time. To get the decision function of multiple SVM classifier system, i.e., to get the values of the value of $\mathbf{w}_j,b$, similar to (4), we can construct following optimization problem:

$$\min f(\mathbf{w},b,\xi) \quad =\frac{1}{2}\sum_{j=1}^{K}\left\|\mathbf{w}_j\right\|^2+\gamma\sum_{i=1}^{N}\xi_i^2 \tag{7}$$

$$\text{st. } y_i(\sum_{j=1}^{K}\mathbf{w}_j\cdot\varphi_j(\mathbf{x}_i)+b)\geq 1-\zeta_i, i=1,2,...,N$$

where the constant γ is used to control the trade-off between the generalization and the classification errors represented by slack variable ξ. Compared with (4), the different is that the first term of the objective function of (7) is to evaluate the sum of the generalization ability of all the available SVM classifiers while in (4) it is to evaluate the generalization ability of a single SVM classifier. And accordingly, we can get a new LaGrange equation:

$$\begin{aligned} L(\mathbf{w},b,e,\alpha)=f(\mathbf{w},b,\zeta)-\\ \sum_{i=1}^{N}\alpha_i\{y_i(\sum_{j=1}^{K}\mathbf{w}_j\cdot\varphi_j(\mathbf{x}_i)+b)-1+\zeta_i\} \end{aligned} \tag{8}$$

with LaGrange multipliers $\alpha_k>0$. To derive the LaGrange multipliers from (8), we can do:

$$\frac{\partial L}{\partial \mathbf{w}_j}=0\to\mathbf{w}_j=\sum_{i=1}^{N}\alpha_i\varphi_j(\mathbf{x}_i) \tag{9}$$

$$\frac{\partial L}{\partial b}=0\to\sum_{i=1}^{N}\alpha_i=0 \tag{10}$$

Then the corresponding optimization problem can be turned into:

$$\max W(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j\left\{\sum_{k=1}^{K}\varphi_k(\mathbf{x}_i)\cdot\varphi_k(\mathbf{x}_j)\right\} \tag{11}$$

$$s.t.\ \sum_{i=1}^{N}\alpha_i y_i=0$$

Obviously, above optimization problem is a convex QP one with linear constraints; we can easily get its global optimal solution. The solutions of (11) are the values of $\mathbf{w}_j,\varphi_j,b$, thus we can get the label of a testing sample through (6).

Now we present an analysis of the computational cost of our proposed method. The proposed method involves solving a QP and the complexity of solving this problem is determined primarily by the size of the Hessian matrix. The size of this Hessian equates to the number of the training samples; so the computational cost of our proposed approach is primarily determined by the number of the training samples. While traditional combination approaches need to solve several QP problems, each trained by one feature, and the computational cost for solving each of these problems is also primarily determined by the number of the training samples, our method only need to solve one optimization problem. So compared to traditional combinational approaches, our method is simpler.

4. EXPERIMENT RESULTS

To evaluate the performance of the proposed method for combining the decisions from individual SVM classifiers, we performed two sets of experiments using texture image dataset and color image dataset. Both of these two sets of experiments are concerned with classification problems, involving multiple features. The Gaussian RBF kernel [1] has been used as the kernel of each SVM classifier; and the kernel parameters for each classifier are determined by a cross-validation method. When several features are used to train an SVM classifier, each dimension would be normalized to the same scale with the range between -1 and 1. In each experiment, we also concatenate all features to a long feature and one SVM classifier is trained by using this feature to do the comparison. We adopt the one-versus-all method to realize the multi-class classification [10].

4.1. Texture Image Classification

Nine texture images from Brodatz's texture album [11], including D1, D3, D6, D11, D16, D17, D20, D21, D24 were used for texture classification experiments. Each image is of size 512×512 pixels with 256 gray values. From each original image, 250 sub-images of size 32×32 are extracted randomly; 50 of them are used for training and the remaining 150 images for testing. Five most commonly used feature domains, including autocorrelation (ACF), edge frequency (EF), wavelet transform (WT), discrete cosine transform (DCT) and Gabor transform (GB) were used to represent the texture images [12]. Furthermore, the formed long feature (concatenating all features) is also used.

Four combination methods, including majority voting, LSE weighted average, simple average, and our proposed approach, have been implemented, and their respective classification error rates recorded. As the preliminary experiment results showed that the individual classifiers could achieve very high testing accuracy, nearly 100%, some white noise was then added to the testing images. Table.1 shows the experiment results based on the noisy images. It can be seen that using multiple features can improve the performance of classification in comparison with the best performance achievable by any single feature. Moreover, the table also shows that our proposed combination approach outperforms the other methods

To further compare the performances of difference combination approaches, we also calculate the statistical significances of the obtained results. The standard variances of the accuracies of different combination approaches are as shown in Table.1. Compared to other three combination approaches, our proposed approach is more stable. The standard variance of the LSE weighted average method is biggest, which shows that the performance of this method is not stable compared with other combination approaches.

4.2 Color image classification

In this experiment the color image dataset is used [13]. This dataset consists of 10 semantic categories and each category has 100 images. Fig.2 shows 10 images of this dataset, each image drown from one class. Fig.3 shows different images from same categories. In this paper these images are represented in terms five different feature domains [14],including First and second color moment in Lab space (CM),Color coherence vector in LUV space (CCV), Gabor texture feature (GB), Color histogram in HSV space (CH) and Wavelet texture feature (WT). From the 100 images for each category, 50 samples are selected randomly for training and the remaining are used for testing.

Fig.1 Ten categories of used color images

Fig.2 Different images from the same category (Left: Category A; Right: Category B)

Table 2 shows the experiment results. It can be seen that the performances of the single-feature classifiers vary greatly and all these features have poor performances. Again, by combining the single-feature classifiers, we can obtain improved performance. The standard variances of different combination approaches are also shown in Table.2, from which it can be seen that our approach is more stable than other combination approaches.

5. CONCLUDING REMARKS

This paper proposes an effective method for solving multiple feature classification problems with SVM. Traditional method of using multiple features is to concatenate all features to a single yet high dimensional one, or to train different individual classifiers by using different features and then combine the decisions of these classifiers. Both of these two approaches have their own weaknesses: the concatenation approach would suffer from the curse of dimension and its performance can not be guaranteed while the decision combination approach needs to train several classifiers and also it is very difficult to get a satisfactory combination approach. To improve the overall performance, we propose an efficient method where the training and the combining of the individual SVM classifiers are performed at the same time. Compared to traditional approaches, our approach can use all kind of features and furthermore need not to train each classifier respectively before the combination. So our proposed approach is more efficient while simpler.

The proposed method has been used in our experiments on texture image classification and color image classification. The results show that the proposed method outperforms concatenating approach, i.e., the method that concatenates all features and other three combination methods, namely LSE weighted average, simple weighted

Table.1. The accuracy of texture image classification

	Multiple features				Single feature					
	Improved approaches	Traditional approaches			EF	ACF	WT	DCT	GB	All
		LSE	SA	MV						
Accuracy (Standard Variance)	0.989 (0.004)	0.961 (0.009)	0.976 (0.004)	0.976 (0.005)	0.87	0.94	0.83	0.64	0.81	0.95

Table.2. The accuracy of color image classification

	Multiple features					Single feature					
	Improved approaches		Traditional approaches			CM	CCV	GB	CH	WT	All
	SA	WA	LSE	SA	MV						
Accuracy (Standard Variance)	0.698 (0.009)	0.722 (0.01)	0.568 (0.012)	0.60 (0.015)	0.51 (0.011)	0.386	0.554	0.158	0.53	0.196	0.6

average, and majority voting for solving the above classification problems.

Further work is needed to address some important issues related to the use of our proposed combination method to solve practical problems, including how to implement an efficient weighted average approach, how to use this approach to combine results from other types of classifiers and so on.

ACKNOWLEDGEMENT

The work reported in this article has been supported in part by the Hong Kong Research Grants Council under CUHK4377/02E.

REFERENCES

[1] V. Vapnik, *The nature of Statistical Learning Theory*, New York: Wiley, 1998.

[2] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Network,* vol. 12, pp. 181–201, Mar. 2001.

[3] M. Pontil and A. Verri, "Support vector machines for 3-D object recognition," *IEEE Trans. Pattern Anal. Machine Intell*., vol. 20, pp. 637–646, 1998.

[4] C. Y. Suen, C. Nadal, T. A. Mai, R. Legault, and L. Lam, "Recognition of totally unconstrained handwritten numerals based on the concept multiple experts," *Frontiers in Handwriting Recognition,* C. Y. Suen, Ed., in Proc. Int. Workshop on Frontiers in Handwriting Recognition**,** Montreal, Canada, Apr. 2-3, 1990, pp. 131-143.

[5] T. K. Ho, J. J. Hull, and **S.** N. Srihari, "Combination of structural classifiers," *Proc. Workshop Syntactic and Structural Pattern Recognition*, pp. 123-137, June 1990.

[6] R. D. Zilca and Y. Bistritz*, "*Feature concatenation for speaker identification", European Signal Processing Conference, Tampere, Finland, 2000.

[7] L. Louisa Lam and C Y. Suen, "Application of majority voting to pattern recognition: An analysis of its behavior and performance", *IEEE Transactions On Systems, Man, and Cybernetics—Part A: Systems And Humans*, Vol. 27, no. 5, pp.553-567, 1997.

[8] L. Xu, A. Krzyzak, and C.Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. Syst., Man, Cybern*.. vol. 22, no. 3, pp. 418-435, 1992**.**

[9] G.. Gumera, F. Boli, "A theoretical and experimental analysis of linear combiners for multiple classifier system", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.27, pp. 942-956, 2005.

[10] C.-W. Hsu and C.-J. Lin, "A comparison on methods for multi-class support vector machines," *Technical Report, Dept. of Computer Science and Information Eng., Nat'l Taiwan Univ.*, 2001

[11] P. Brodatz, *Textures: A Photographic Album for Artists and Designers*. New York: Dover, 1966.

[12] M. Sonka, V. Hlavac, and R. Boyer, *Image Processing, Analysis, and Machine Vision,* Thomoson Learning and PT Press, 1999

[13] J. Z. Wang, J. Li and G. Wiederhold, "SIMPLIcity: semantics-sensitive integrated matching for picture libraries", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol 23, no.9, pp. 947-963, 2001.

[14] A. Vailaya, M.A.T Figueiredo, A.K Jain and H. Zhang, "Image classification for content-based indexing" , *IEEE Transactions on Image Processing*, Vol.10, no.1, pp.117-130, 2001.