Available online at www.sciencedirect.com
SCIENCE DIRECT
Pattern Recognition 39 (2006) 2258–2264
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

A novel and quick SVM-based multi-class classifier

Yiguang Liu[a, b, *], Zhisheng You[a], Liping Cao[c]

[a] Institute of Image & Graphics, School of Computer Science and Engineering, Sichuan University, Chengdu 610064, PR China
[b] Center for Nonlinear and Complex Systems, School of Electronic Engineering, University of Electronic Science and Technology, Chengdu 610054, PR China
[c] Sichuan University Library, Sichuan University, Chengdu 610064, PR China

Received 1 December 2005; received in revised form 25 May 2006; accepted 31 May 2006

Abstract

Use different real positive numbers p_i to represent all kinds of pattern categories, after mapping the inputted patterns into a special feature space by a non-linear mapping, a linear relation between the mapped patterns and numbers p_i is assumed, whose bias and coefficients are undetermined, and the hyper-plane corresponding to zero output of the linear relation is looked as the base hyper-plane. To determine the pending parameters, an objective function is founded aiming to minimize the difference between the outputs of the patterns belonging to a same type and the corresponding p_i, and to maximize the distance between any two different hyper-planes corresponding to different pattern types. The objective function is same to that of support vector regression in form, so the coefficients and bias of the linear relation are calculated by some known methods such as SVM$^{\text{light}}$ approach. Simultaneously, three methods are also given to determine p_i, the best one is to determine them in training process, which has relatively high accuracy. Experiment results of the IRIS data set show that, the accuracy of this method is better than those of many SVM-based multi-class classifiers, and close to that of DAGSVM (decision-directed acyclic graph SVM), emphatically, the recognition speed is the highest.

Keywords: SVM; Multi-class classifier; SVM$^{\text{light}}$ approach; Objective function

1. Introduction

Distinguishing binary different types is easy for SVM, but how to use SVM to recognize multi-class patterns is also not perfectly solved [1–17]. Traditionally, in using SVM to recognize many types, the pattern space is partitioned into many subspaces, each of which includes only two pattern types, where an ordinary SVM is adaptable. Originating from this idea, many techniques, such that one-versus-one method implemented by max-wins voting (Max–Wins), one-versus-all method using winner-takes-all (1-v-r) strategy and directed acyclic graph SVM (DAGSVM) are proposed [2,3,5,17]. By an empirical study, Duan et al. [3] call PWC_PSVM (one SVM-based multi-class classifier using Platt's posterior probabilities together with the pairwise coupling idea of Hastie and Tibshirani.) has superior generalization performance over 1-v-r and Max–Wins. Angulo et al. [1] introduce a "Support Vector Classification-Regression" machine for K-class classification purposes (K-SVCR), as the approach adopts 1-versus-1-versus-rest structure during the decomposing phase, its computation load is very heavy. Lee et al. [4] design a loss function deliberately tailored to target the coded class with the maximum conditional probability for multi-category classification problems. Representing each pattern category in binary format, and to each bit of that representation is assigned a conventional SVM, totally $\lceil \text{Log}_2 \text{K} \rceil$ SVMs are required to classify K classifications, this is the method named as Mx-ary SVM [6]. The k-SVM approach needs to construct k two-class discriminants using k quadratic programmings, the M-SVM approach is analogical, and requires the solution of a single quadratic

* Corresponding author. Institute of Image & Graphics, School of Computer Science and Engineering, Sichuan University, Chengdu 610064, PR China. Tel.: +86 28 85412565.
E-mail address: lygpapers@yahoo.com.cn (Y. Liu).

0031-3203/$30.00 © 2006 Published by Elsevier Ltd on behalf of Pattern Recognition Society.
doi:10.1016/j.patcog.2006.05.034

programming [7]. Ref. [8] introduces an approach involving conventional SVMs, least-square SVMs and Bayes' formula, etc., apparently, the way requires much computation. It must be noted that: at present there exists no theory which shows that good generalization performance is guaranteed for SVMs while very high VC dimension would normally bode ill for generalization performance [9,10].

As for the realization, the SVM$^{\text{light}}$ approach makes large-scale SVM training more practical [11]. Also, combining SVM and nearest neighbor classifiers provides a powerful alternative to SVMs [12], especially in place where computation time and accuracy are primary important. Based on a measurement of similarity among samples, a heuristic method for accelerating SVM training is fascinating [13], the most attractive point of this idea is to make SVM training fast especially for large-size training data. A hybrid method of DAGSVM and Max–Wins algorithm is also powerful [14], whose cumulative recognition rate is as good as the Max–Wins algorithm, and the execution time is almost as fast as DAGSVM. The mean field approach can be used in SVM-based classification problem [15], actually, all category approaches based on SVM admit the same dual problem formulation [16].

2. Objective function of categorization

In a special space, if the distances between the samples belonging to different types to a certain base are distinguishingly different, then the larger the distance difference between two pattern types is, the easier the two pattern categories can be distinguished, the better the general performance of the classifier for these two pattern categories is. The classification base needs to satisfy the following traits: (1) the base is effective everywhere, i.e., each sample in this space possesses a distance from the base; (2) the distance can be definitely calculated, or those of two different patterns are comparable when they cannot be definitely calculated. (3) in order to easily calculate the distance, it is best that the base is hyper-point, line or plane.

In complex distribution pattern space, seeking a good base is difficult, even it does not exist completely. Mapping the patterns from a low-dimension space to a high-dimension one can make the base take on, or be easily determined.

Let $\aleph$ denote the feature space spanned by all input patterns, $\Re$ denote a high-dimension space, and the mapping $\psi : \aleph \rightarrow \Re$ mapped the patterns in $\aleph$ to $\Re$. Assume all samples are categorized into m types, each type possesses n samples for training, a real positive number p_i ($p_i \neq p_j$ for $i \neq j, i, j = 1, \ldots, m$) is the object value corresponding to the ith type samples $\mathbf{x}_{ij}$ ($i = 1, \ldots, m,\ j = 1, \ldots, n$), the base $\beta \subset \Re$ is assumed to be a hyper-plane

$$\beta: \mathbf{w}^{\mathrm{T}}\mathbf{y} + b = 0, \tag{1}$$

where $\mathbf{y} \in \Re$, $\mathbf{w}$ is a pending vector whose dimension number is dependent upon $\psi(\mathbf{x}_{ij})$, b is a pending number. In order to make $\mathbf{w}^{\mathrm{T}}\psi(\mathbf{x}_{ij}) + b$ close to p_i, we have following constraint set:

$$\begin{cases} \mathbf{w}^{\mathrm{T}}\psi(\mathbf{x}_{ij}) + b \leqslant p_i + \varepsilon + \xi_{ij}^{+},\ \xi_{ij}^{+} \geqslant 0, \\ \mathbf{w}^{\mathrm{T}}\psi(\mathbf{x}_{ij}) + b \geqslant p_i - \varepsilon - \xi_{ij}^{-},\ \xi_{ij}^{-} \geqslant 0, \end{cases} \tag{2}$$

where $\varepsilon > 0$ is the non-sensitive quantity, $\xi_{ij}^{+} \geqslant 0$ and $\xi_{ij}^{-} \geqslant 0$ are parameters waiting for extraction. The smaller the parameters ξ_{ij}^{+} and ξ_{ij}^{-} are, the less the empirical risk minimization is. The distance from the mapped pattern $\psi(\mathbf{x}_{ij})$ to the base β can be given as

$$d_{\mathbf{x}_{ij}} = \frac{|\mathbf{w}^{\mathrm{T}}\psi(\mathbf{x}_{ij}) + b|}{\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}}. \tag{3}$$

Denote

$$\Delta d_{ik} \equiv \left| \sum_{j=1}^{n} d_{\mathbf{x}_{ij}} - \sum_{h=1}^{n} d_{\mathbf{x}_{kh}} \right| = \frac{1}{\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}} \times \left| \sum_{j=1}^{n} |\mathbf{w}^{\mathrm{T}}\psi(\mathbf{x}_{ij}) + b| - \sum_{h=1}^{n} |\mathbf{w}^{\mathrm{T}}\psi(\mathbf{x}_{kh}) + b| \right|. \tag{4}$$

To categorize the mapped patterns easily, it is needed to make the distance difference Δd_{ik} ($i, k = 1, \ldots, n$) maximal so as to minimize the structure risk minimization. From Eqs. (2) and (4), it follows that

$$\Delta d_{ik} \leqslant \frac{1}{\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}} \sum_{j,h=1}^{n} (2\varepsilon + \xi_{ij}^{+} + \xi_{kh}^{-} + |p_i - p_k|), \tag{5}$$

thus Δd_{ik} is proportional to $1/\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}$, i.e.,

$$\Delta d_{ik} \propto 1/\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}. \tag{6}$$

Eq. (2) tells that the smaller the parameters ξ_{ij}^{+} and ξ_{ij}^{-} are, the more $\mathbf{w}^{\mathrm{T}}\psi(\mathbf{x}_{ij}) + b$ is close to p_i. From Eq. (5), we can see, the smaller $\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}$ is, the larger Δd_{ik} is, the more easily the two pattern types are categorized. Taking into account the two sides, we need to solve the following quadratic optimization problem: minimize the functional

$$\phi(\mathbf{w}, \xi^{+}, \xi^{-}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\left(\sum_{i=1}^{m}\sum_{j=1}^{n} \xi_{ij}^{+} + \sum_{i=1}^{m}\sum_{j=1}^{n} \xi_{ij}^{-} \right), \tag{7}$$

subject to constraints

$$\begin{aligned} &\mathbf{w}^{\mathrm{T}}\psi(\mathbf{x}_{ij}) + b \leqslant p_i + \varepsilon + \xi_{ij}^{+},\ \xi_{ij}^{+} \geqslant 0, \\ &i = 1, \ldots, m;\quad j = 1, \ldots, n, \\ &\mathbf{w}^{\mathrm{T}}\psi(\mathbf{x}_{ij}) + b \geqslant p_i - \varepsilon - \xi_{ij}^{-},\ \xi_{ij}^{-} \geqslant 0, \\ &i = 1, \ldots, m;\quad j = 1, \ldots, n, \end{aligned} \tag{8}$$

where ξ^{+} and ξ^{-} denote $(\xi_{ij}^{+})_{m \times n}$ and $(\xi_{ij}^{-})_{m \times n}$ formally, and C is a predefined control quantity about error of sample