Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition 40 (2007) 41–51

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Domain described support vector classifier for multi-classification problems

Daewon Lee, Jaewook Lee*

Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, Kyungbuk 790-784, Republic of Korea

Received 8 November 2005; received in revised form 1 May 2006; accepted 6 June 2006

Abstract

In this paper, a novel classifier for multi-classification problems is proposed. The proposed classifier, based on the Bayesian optimal decision theory, tries to model the decision boundaries via the posterior probability distributions constructed from support vector domain description rather than to model them via the optimal hyperplanes constructed from two-class support vector machines. Experimental results show that the proposed method is more accurate and efficient for multi-classification problems.

Keywords: Multi-class classification; Kernel methods; Bayes decision theory; Density estimation; Support vector domain description

1. Introduction

Support vector machines (SVMs), originally formulated for two-class classification problems, have been successfully applied to diverse pattern recognition problems and have become in a very short period of time the standard state-of-the-art tool. The SVMs, based on the *structured risk minimization* (SRM), are primarily devised in order to minimize the upper bound of the expected error by optimizing the trade-off between the empirical risk and the model complexity [1–3]. To achieve this, they construct an optimal hyperplane to separate *binary class* data so that the margin is maximal.

Since many real-world applications are multi-class classification problems, several approaches to extend two-class SVMs to a multi-class SVM for multi-category classifications have been proposed. Most of the previous approaches try to decompose a multi-class problem to a set of multiple binary classification problems where two-class SVMs can be trained and applied. For example, one-against-all algorithm transforms a c-class problem into c two-class problems where one class is separated from the remaining ones; one-against-one (pair-wise) algorithm converts the c-class problem into $c(c-1)/2$ two-class problems where pairwise optimal hyperplanes for each pair of classes are constructed and max-voting strategy is used to predict their classes, and so on (cf. [4,5]). These approaches, however, have some drawbacks inherent in the architecture of multiple binary classifications: some unclassifiable regions may exist if a data point belongs to more than one class or to none, resulting in low accuracy in correct classification. Also, to train two-class SVMs multiple times for the same data set repeatedly often results in a highly intensive time complexity for large scale problems.

To overcome such drawbacks, in this paper, we propose a novel support vector classifier for multi-classification problems. The proposed classifier, based on the Bayesian optimal decision theory, tries to model the posterior probability distributions via support vector domain description (SVDD) [6,7] rather than to model the decision boundaries by constructing optimal hyperplanes. The performance of the proposed method is confirmed through simulation.

The organization of this paper is as follows. In Section 2, we review the Bayesian optimal decision theory and briefly outline a SVDD algorithm. A novel method for multi-classification problems is proposed in Section 3

* Corresponding author. Tel.: +82 54 279 2209.
E-mail addresses: woosuhan@postech.ac.kr (D. Lee), jaewookl@postech.ac.kr (J. Lee).

0031-3203/$30.00 © 2006 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2006.06.008

with an illustrative example and Section 4 provides the theoretical basis of the proposed method. In Section 5, simulation results are given to illustrate the effectiveness and the efficiency of the proposed method.

2. Previous works

In this section, we first review the Bayesian optimal decision theory and describe the existing density estimation algorithms. Then we briefly outline the SVDD algorithm employed in our proposed method.

2.1. Bayesian optimal decision theory

According to the Bayesian decision theory, an optimal classifier can be designed if we know the prior probabilities $p(w_i)$ and the class-conditional densities $p(\mathrm{x}|w_i)$, that is, with Bayes formula, the posterior probabilities are given by

$$p(w_i|\mathrm{x}) = \frac{p(\mathrm{x}|w_i)p(w_i)}{p(\mathrm{x})} = \frac{p(\mathrm{x}|w_i)p(w_i)}{\sum_{i=1}^{c} p(\mathrm{x}|w_i)p(w_i)}, \tag{1}$$

where c is the number of output class labels. The optimal decision rule to minimize the average probability of error can then be shown to be the Bayesian decision rule [8,9] that selects the w_i maximizing the posterior probability $p(w_i|\mathrm{x})$ as follows:

$$\text{Decide } w_i \text{ if } p(w_i|\mathrm{x}) > p(w_j|\mathrm{x}) \quad \text{for all } j \neq i. \tag{2}$$

In typical classification problems, estimation of the prior probabilities presents no serious difficulties (normally all are assumed to be equal or N_i/N). However, estimation of the class-conditional densities is quite another matter. During the last decades, lots of density estimation algorithms have been proposed and the existing density estimation algorithms may generally be categorized into three approaches: parametric, semi-parametric, and nonparametric methods.

Parametric methods assume a specific functional form of $p(\mathrm{x}|w_i)$ to contain a number of adjustable parameters. The simplest and the most widely used form is a normal distribution given by

$$p(\mathrm{x}|w_i, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{(d/2)}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathrm{x}-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(\mathrm{x}-\mu_i)\right). \tag{3}$$

The drawback of such an approach is that a particular form of parametric function might be incapable of describing the true data distribution.

Second, semi-parametric methods have a form of finite mixtures of Gaussians as follows:

$$p(\mathrm{x}|w_i, \theta_1, \ldots, \theta_M) = \sum_{k=1}^{M} p(\mathrm{x}|w_i, \theta_k, k)p(k), \tag{4}$$

where $p(\mathrm{x}|w_i, \theta_k, k)$ is a kth component in the form of Gaussian function and $p(k)$ are mixing parameters. In semi-parametric methods, training data do not provide any *component labels* to say which component was responsible for generating each data point. To select the number of components and to estimate its parameters, however, we need to incorporate with an iterative scheme such as an EM algorithm, which often proved to be highly computationally extensive.

The third approach is nonparametric methods which estimate the class-conditional density function as a weighted sum of a set of kernel functions, $K(\cdot,\cdot)$, to be determined entirely by the data

$$p(\mathrm{x}|w_i) = \sum_{j=1}^{N} \beta_j K(\mathrm{x}_j, \mathrm{x}). \tag{5}$$

Though such methods have the most descriptive capability, they typically suffer from the problem that the number of parameters grows with the size of the data set, so that the models can quickly become unwieldy.

2.2. Support vector domain description

The existing methods for density estimation have a tradeoff between a descriptive ability and a computational burden. To solve this problem, our proposed method utilizes a so-called trained kernel support function that characterizes the support of a high dimensional distribution of a given data set, inspired by the SVMs. We first review a support vector domain description (SVDD) procedure (also called a one-class support vector machine). Then we build a trained kernel support function, to be used as a pseudo-density function, via SVDD.

The basic idea of SVDD is to map data points by means of a nonlinear transformation to a high dimensional feature space and to find the smallest sphere that contains most of the mapped data points in the feature space [6,7]. This sphere, when mapped back to the data space, can separate into several components, each enclosing a separate cluster of points. More specifically, let $\{\mathrm{x}_i\} \subset \mathscr{X}$ be a given training data set of $\mathscr{X} \subset \Re^n$, the data space. Using a nonlinear transformation Φ from $\mathscr{X}$ to some high dimensional feature space, we look for the smallest enclosing sphere of radius R described by the following model:

$$\begin{aligned} \min \quad & R^2 + C\sum_j \xi_j \\ \text{s.t.} \quad & \|\Phi(\mathrm{x}_j) - \mathbf{a}\|^2 \leqslant R^2 + \xi_j, \\ & \xi_j \geqslant 0 \quad \text{for } j = 1, \ldots, N, \end{aligned} \tag{6}$$