Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Pattern Recognition Letters 26 (2005) 2187–2194
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Training TSVM with the proper number of positive samples

Ye Wang *, Shang-Teng Huang

Computer Science and Engineering Department, Shanghai JiaoTong University, Shanghai 200030, PR China

Received 20 September 2004; received in revised form 14 March 2005
Available online 13 June 2005

Communicated by R.P.W. Duin

Abstract

The transductive support vector machine (TSVM) is the transductive inference of the support vector machine. The TSVM utilizes the information carried by the unlabeled samples for classification and acquires better classification performance than the regular support vector machine (SVM). As effective as the TSVM is, it still has obvious deficiency: The number of positive samples must be appointed before training and it is not changed during the training phase. This deficiency is caused by the pair-wise exchanging criterion used in the TSVM. In this paper, we propose a new transductive training algorithm by substituting the pair-wise exchanging criterion with the individually judging and changing criterion. Experimental results show that the new method releases the restriction of the appointment of the number of positive samples beforehand and improves the adaptability of the TSVM.

Keywords: Transductive support vector machine; Training algorithm

1. Introduction

Different from the traditional inductive learning method, the transductive learning method is a way to utilize the information of the labeled samples together with that of the unlabeled samples. Nowadays, under the condition that obtaining labels is more expensive than getting samples, the transductive method is significantly important for the machine learning. Based on the theory of the statistical learning (Vapnik, 1995), the support vector machine (SVM) was proposed for the pattern recognition by Vapnik (1998). The SVM shows better performance than other learning method in the classification field of the small training sets in a nonlinear and high-dimensional feature space.

The transductive support vector machine (TSVM) (Joachims, 1999) is a SVM combining with the transductive learning procedure. The iteration algorithm of the TSVM utilizes the

* Corresponding author. Fax: +86 21 6293 4107.
E-mail addresses: wangye34@sjtu.edu.cn, wangye@cs.sjtu.edu.cn (Y. Wang).

0167-8655/$ - see front matter
doi:10.1016/j.patrec.2005.03.034

information of the unlabeled samples for classification and predicts the optimal labels for them. The TSVM is suitable in the case when the distribution of the training samples differs from that of the test samples.

After reviewing the training algorithm of the TSVM, we propose in this paper that the *individually judging and changing* criterion instead of the *pair-wise exchanging* criterion is utilized to minimize the objective function. The training algorithm is also changed accordingly. Experimental results show that the new method releases the restriction of the appointment of the number of positive samples beforehand. It could be used in much wider ranges.

The remaining of this paper is organized as follows: The training algorithm of the TSVM and its deficiency are described in Section 2. An improved label exchanging criterion, a new training method and the theoretical proof of convergence are developed in Section 3. Experimental results of comparing SVM, TSVM and the new method are shown in Section 4. Finally, conclusions are given in the last section.

2. Review of the TSVM

The transductive learning is completely different from the inductive learning. Not caring the decision hyper-plane or decision rules, the transductive learning focuses on the construction of a transductive mechanism that can label test samples with the information carried by either the labeled or the unlabeled samples in the training phase. The idea of the transductive learning originates from the dilemma of the machine learning, where the labeled samples are sparse and expensive while the unlabeled samples are plentiful and cheap. Through the learning process of the transductive learning, the helpful information concealed in the unlabeled samples is transferred into the final classifier. This mechanism makes classification results more accurate.

2.1. Principle

The TSVM is an application of the transductive learning theory to the SVM. The principle of the regular SVM is described as follows. An SVM classifier tries to find a decision hyper-plane that classifies training samples correctly (or basically correctly). The margin between this hyper-plane and the nearest sample should be as large as possible. In the training phase, the SVM resolves the following optimization problem to find the decision hyper-plane,

$$\min \frac{1}{2} w^{\mathrm{T}} w + C \sum_i \xi_i \tag{2.1}$$

such that

$$y_i(w^{\mathrm{T}} x_i + b) \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0$$

where b is a threshold; C is an influencing parameter for trade-off and ξ is a slack variable. In the decision phase, the SVM labels the test samples according to the side of the hyper-plane that they lie on. The decision formula is given by

$$D(x) = w^{\mathrm{T}} x + b \tag{2.2}$$

According to the principle of the transductive learning, the optimization problem of the TSVM (Joachims, 1999) is given by

$$\min \frac{1}{2} w^{\mathrm{T}} w + C \sum_i \xi_i + C^* \sum_j \xi_j^* \tag{2.3}$$

such that

$$y_i(w^{\mathrm{T}} x_i + b) \geqslant 1 - \xi_i, \ \ \xi_i \geqslant 0;$$

$$y_j(w^{\mathrm{T}} x_j + b) \geqslant 1 - \xi_j^*, \ \ \xi_j^* \geqslant 0$$

where ξ and ξ^* are slack variables for the training and test samples, respectively; C and C^* are the influencing parameters for the training and test samples, respectively.

2.2. Training algorithm

To acquire optimal solutions of Eq. (2.3), Joachims presented a training algorithm. The main steps of this algorithm are listed as follows:

(1) Specify C and C^* and train a SVM with the training samples. Specify N_{p} (the number of positive samples in the test set) according to the proportion of positive samples in the training set.