Knowledge-Based Systems 59 (2014) 1–8

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A novel feature selection method for twin support vector machine

Lan Bai [a], Zhen Wang [a], Yuan-Hai Shao [b,*], Nai-Yang Deng [c]

[a] Mathematics School of Jilin University, Changchun 130012, PR China
[b] Zhijiang College, Zhejiang University of Technology, Hangzhou 310024, PR China
[c] College of Science, China Agricultural University, Beijing 100083, PR China

ARTICLE INFO

Article history:
Received 28 June 2013
Received in revised form 25 January 2014
Accepted 31 January 2014
Available online 10 February 2014

Keywords:
Pattern recognition
Feature selection
Twin support vector machine
Feature ranking
L_1 norm
Multi-objective mixed-integer programming

ABSTRACT

Both support vector machine (SVM) and twin support vector machine (TWSVM) are powerful classification tools. However, in contrast to many SVM-based feature selection methods, TWSVM has not any corresponding one due to its different mechanism up to now. In this paper, we propose a feature selection method based on TWSVM, called FTSVM. It is interesting because of the advantages of TWSVM in many cases. Our FTSVM is quite different from the SVM-based feature selection methods. In fact, linear SVM constructs a single separating hyperplane which corresponds a single weight for each feature, whereas linear TWSVM constructs two fitting hyperplanes which corresponds to two weights for each feature. In our linear FTSVM, in order to link these two fitting hyperplanes, a feature selection matrix is introduced. Thus, the feature selection becomes to find an optimal matrix, leading to solve a multi-objective mixed-integer programming problem by a greedy algorithm. In addition, the linear FTSVM has been extended to the nonlinear case. Furthermore, a feature ranking strategy based on FTSVM is also suggested. The experimental results on several public available benchmark datasets indicate that our FTSVM not only gives nice feature selection on both linear and nonlinear cases but also improves the performance of TWSVM efficiently.

1. Introduction

Feature selection [18,15,6,11,19] is an important problem since it removes the irrelevant features and maintains the relevant features that are as close to the class as possible. The benefit of feature selection is twofold. On one hand, it is meaningful because it can identify the features that contribute most to classification. On the other hand, feature selection is helpful for solving the classification problem since it can not only reduce the dimension of input space and speed up the computation procedure but also improve the classification accuracy. There are mainly two types of feature selection methods: one is the general methods independent of any classifiers, e.g., Fisher score [38] and Laplacian score [12]; the other one is the wrapper-type method dependent on the classifier, e.g., the methods based on Bayesian network [14], based on neural networks [32] and based on support vector machine (SVM) [2,23]. The wrapper-type methods attract more attentions and often improve the performance of the original classifier as well [14,37,23,2,32].

It is interesting to investigate the wrapper-type feature selection methods based on the twin support vector machine (TWSVM) [16,35]. TWSVM generates two nonparallel fitting hyperplanes and has superior performance than SVM [3,21,7,43] on both the classification accuracy and learning speed in many practical applications [16,35,33,30]. Particularly, it is suitable for some special datasets, e.g., "cross-planes" [16]. However, in contrast to SVM which owns many wrapper-type feature selection methods such as SVM-RFE [23] and RFSVM [2], TWSVM has no any corresponding one. The reason is that the current SVM-based feature selection methods cannot be used to TWSVM directly. In fact, for each feature, SVM provides a single weight corresponding to the single separating hyperplane whereas TWSVM provides two weights corresponding to two fitting hyperplanes, leading to some difficulties for feature selection.

In this paper, we propose a novel feature selection method for TWSVM, called FTSVM for short. Our FTSVM includes two forms: linear FTSVM and nonlinear one. The former is formulated by the following steps: first of all, a basic L_1-TWSVM is proposed by introducing the L_1-norm regularization terms, due to the success in L_1-SVM to obtain a sparse feature weight [44,20,8,13]; then, a feature selection matrix, a diagonal matrix with element either 1 or 0, is introduced in the proposed L_1-TWSVM, resulting to two mixed integer programming problems (MIPPs); finally, the two MIPPs are solved simultaneously as a multi-objective mixed integer programming problem (MOMIPP) [39] by an greedy algorithm. The nonlinear FTSVM is constructed by kernel trick [16,35], which

* Corresponding author. Tel.: +86 057187313551.
E-mail address: shaoyuanhai21@163.com (Y.-H. Shao).

http://dx.doi.org/10.1016/j.knosys.2014.01.025
0950-7051/

needs also to solve a MOMIPP by the proposed greedy algorithm. In addition, based on FTSVM, a feature ranking strategy is suggested, which ranks the features according to their contributions to the objective in the MOMIPP. The experimental results show that our FTSVM not only gives nice feature selection but also improves the performance of TWSVM.

This paper is organized as follows. A briefly review of L_2-TWSVM is in Section 2, and the standard L_1-TWSVM is proposed in Section 3. Our feature selection and ranking methods are formulated in Section 4, and experiments are arranged in Section 5. Finally, Section 6 gives the conclusions. For convenience, in Table 1, we present some notations used in the paper.

2. Review of L_2-TWSVM

Consider the binary classification problem with m_1 training samples belong to positive class represented by A_1 and m_2 training samples belong to negative class represented by A_2 in the n-dimensional real space R^n, a classifier attempts to predict the new samples belong to either positive or negative class.

TWSVM [16] seeks two nonparallel hyperplanes

$$f_1(x) = xw_1 + b_1 = 0 \quad \text{and} \quad f_2(x) = xw_2 + b_2 = 0 \tag{1}$$

such that each one is the fitting hyperplane of one of the positive and negative classes, a new sample will be predicted to one class if it is closer to the corresponding fitting hyperplane. TWSVM keeps the fitting hyperplane as close as possible to its corresponding training samples and far away from the others. The purpose of L_2-TWSVM [35] is very the same as TWSVM, where the difference is L_2-TWSVM introduces the L_2 regularization terms of the weight vectors into TWSVM to minimize the structural risk. L_2-TWSVM solves following quadratic programming problems (QPPs)

$$\begin{aligned} \min_{w_1,b_1,\xi_1} \quad & \frac{1}{2}\left(\|w_1\|_2^2 + b_1^2 + c_{11}\|A_1w_1 + b_1e\|_2^2\right) + c_{12}e^\top\xi_1 \\ \text{s.t.} \quad & -(A_2w_1 + b_1e) + \xi_1 \geqslant e, \quad \xi_1 \geqslant 0, \end{aligned} \tag{2}$$

and

$$\begin{aligned} \min_{w_2,b_2,\xi_2} \quad & \frac{1}{2}\left(\|w_2\|_2^2 + b_2^2 + c_{21}\|A_2w_2 + b_2e\|_2^2\right) + c_{22}e^\top\xi_2 \\ \text{s.t.} \quad & A_1w_2 + b_2e + \xi_2 \geqslant e, \quad \xi_2 \geqslant 0, \end{aligned} \tag{3}$$

where $c_{11}, c_{12}, c_{21}, c_{22} > 0$ are parameters, $\xi_1 \in R^{m_2}$ and $\xi_2 \in R^{m_1}$ are slack vectors.

Table 1
The symbols used in this paper.

Symbol	Domain	Description
$m_1,\ m_2$	Z^+	Number of positive and negative training samples
m	Z^+	Number of training samples, $m = m_1 + m_2$
n	Z^+	Dimension of training samples
x_i	R^n	The ith training sample
y_i	{+1,-1}	Class of the ith training sample
$A_1,\ A_2$	$R^{m_1\times n},\ R^{m_2\times n}$	Positive and negative training matrices
A	$R^{m\times n}$	Training matrix, $A = [A_1; A_2]$
e		Vector of ones of appropriate dimension
$w_{1,2}, b_{1,2}$	R^n, R	Normal vectors and biases
$u_{1,2}, \gamma_{1,2}$	R^m, R	Normal vectors and biases
$\xi_{1,2}$	$R^{m_{2,1}}$	Slack vectors
$c_{11,12,21,22}$	R^+	Parameters
σ	R^+	Parameter
λ	(0, 1)	Parameter
E	diag (1 or 0)	Diagonal matrix
$K(\cdot,\cdot)$		Kernel function
$\|\|\cdot\|\|_{1,2}$		$L_{1,2}$ norm
$(\cdot)_+$		Replaces the negative elements with 0

The geometric meanings of (2) and (3) are clear. For example, for (2), its objective function makes the fitting hyperplane $f_1(x) = 0$ of the positive class fit the positive class samples A_1, while the constraints keep the negative class far from this hyperplane to some extent (the bias of each negative training sample to $f_1(x) = 0$ is no more than -1).

The above linear L_2-TWSVM has been extended to nonlinear classifier by kernel trick [35]. Define the inner product by the kernel function $K(\cdot,\cdot)$, nonlinear L_2-TWSVM seeks two kernel generated surfaces

$$f_1(x) = K(x,A)u_1 + \gamma_1 = 0 \quad \text{and} \quad f_2(x) = K(x,A)u_2 + \gamma_2 = 0, \tag{4}$$

where A is the whole training set, including A_1 and A_2.

The corresponding QPPs solved in nonlinear L_2-TWSVM are

$$\begin{aligned} \min_{u_1,\gamma_1,\xi_1} \quad & \frac{1}{2}\left(\|u_1\|_2^2 + \gamma_1^2 + c_{11}\|K(A_1,A)u_1 + \gamma_1e\|_2^2\right) + c_{12}e^\top\xi_1 \\ \text{s.t.} \quad & -(K(A_2,A)u_1 + \gamma_1e) + \xi_1 \geqslant e, \quad \xi_1 \geqslant 0, \end{aligned} \tag{5}$$

and

$$\begin{aligned} \min_{u_2,\gamma_2,\xi_2} \quad & \frac{1}{2}\left(\|u_2\|_2^2 + \gamma_2^2 + c_{21}\|K(A_2,A)u_2 + \gamma_2e\|_2^2\right) + c_{22}e^\top\xi_2 \\ \text{s.t.} \quad & K(A_1,A)u_2 + \gamma_2e + \xi_2 \geqslant e, \quad \xi_2 \geqslant 0. \end{aligned} \tag{6}$$

3. L_1-TWSVM

As stated above, L_1-SVM which is sparser than L_2-SVM [8,13] is proposed by replacing the L_2 norm regularization term in L_2-SVM with L_1 norm. Therefore, we propose the L_1-TWSVM that minimizes L_1-norm of the feature weight vectors so that the features are sparser than L_2-TWSVM. Similar to L_1-SVM, we introduce the L_1-norm of $w_i, i = 1,2$, and straightly transform the other parts in L_2-TWSVM from L_2 norm to L_1-norm as follows

$$\begin{aligned} \min_{w_1,b_1,\xi_1} \quad & \|w_1\|_1 + c_{11}\|A_1w_1 + b_1e\|_1 + c_{12}\|\xi_1\|_1 \\ \text{s.t.} \quad & -(A_2w_1 + b_1e) + \xi_1 \geqslant e, \quad \xi_1 \geqslant 0, \end{aligned} \tag{7}$$

and

$$\begin{aligned} \min_{w_2,b_2,\xi_2} \quad & \|w_2\|_1 + c_{21}\|A_2w_2 + b_2e\|_1 + c_{22}\|\xi_2\|_1 \\ \text{s.t.} \quad & A_1w_2 + b_2e + \xi_2 \geqslant e, \quad \xi_2 \geqslant 0. \end{aligned} \tag{8}$$

Different from solving two dual QPPs in L_2-TWSVM [35], the above problems can be solved as two differentiable linear programming problems (DLPPs) [9,22]. Define $w_i = p_i - q_i$, $A_iw_i + b_ie = s_i - t_i,\ i = 1,2$, then (7) and (8) are equivalent to

$$\begin{aligned} \min_{p_1,q_1,s_1,t_1,b_1,\xi_1} \quad & e^\top(p_1 + q_1) + c_{11}e^\top(s_1 + t_1) + c_{12}e^\top\xi_1 \\ \text{s.t.} \quad & A_1(p_1 - q_1) + b_1e = s_1 - t_1, \\ & A_2(p_1 - q_1) + b_1e \leqslant -e + \xi_1, \\ & p_1, q_1, s_1, t_1, \xi_1 \geqslant 0, \end{aligned} \tag{9}$$

and

$$\begin{aligned} \min_{p_2,q_2,s_2,t_2,b_2,\xi_2} \quad & e^\top(p_2 + q_2) + c_{21}e^\top(s_2 + t_2) + c_{22}e^\top\xi_2 \\ \text{s.t.} \quad & A_2(p_2 - q_2) + b_2e = s_2 - t_2, \\ & A_1(p_2 - q_2) + b_2e \geqslant e - \xi_2, \\ & p_2, q_2, s_2, t_2, \xi_2 \geqslant 0. \end{aligned} \tag{10}$$

Note $s_i = A_i(p_i - q_i) + b_ie + t_i, i = 1,2$, the final problems we solved are