Neurocomputing 151 (2015) 481–500

Contents lists available at ScienceDirect

Neurocomputing

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

Minimum class spread constrained support vector machine

JianWen Tao [a,*], Shitong Wang [b], Wenjun Hu [c]

[a] School of Information Science and Engineering, Ningbo Institute of Technology, Zhejiang University, Ningbo, China
[b] School of Digital Media, Jiangnan University, Wuxi, China
[c] School of Information and Engineering, Huzhou Teachers College, Huzhou 313000, China

ARTICLE INFO

Article history:
Received 23 January 2014
Received in revised form
8 June 2014
Accepted 13 September 2014
Communicated by Deng Cai
Available online 28 September 2014

Keywords:
Pattern classification
Novelty detection
Within-class spread
Support vector machines (SVMs)
Kernel trick

ABSTRACT

While classical classification methods such as support vector machine and its extensions (SVMs) obtain strong generalization capability by maximizing the separation margin between binary classes, they are usually sensitive to affine (or scaling) transformation of data. The optimal solutions of SVMs may be misled by the spread of data and preferentially separate classes along large margin directions. To this end, in this paper, we propose a novel minimum class spread constrained support vector machine (MCSSVM) for pattern classification problems by simultaneously considering the maximization of inter-class separation margin and the minimization of within-class spread of data, which corresponds to a distribution-dependent regularization on the classification function by constraining the within-class spread of input data. The basic idea of MCSSVM is to find a class distribution constrained hyperplane such that one class (or normal class) can be enclosed in a minimum q-spread tube, while the other class (or abnormal class) is farthest from this tube. MCSSVM can simultaneously achieve both maximum inter-class margin and minimum within-class spread so as to enhance the generalization capability of the proposed classifier. Moreover, the proposed method only requires simple extensions to existing maximum margin formulations such as SVMs and still preserves their computational efficiency. Generalization bound for MCSSVM is theoretically derived and the validity of MCSSVM is examined by classification of toy and real-world classification problems, which demonstrate the superiority of our method in comparison to other related state-of-the-art algorithms.

1. Introduction

Pattern classification aims at learning a classifier (or classification decision function) on a finite training dataset, which has a sound generalization capacity for unseen testing data. Classification can be achieved by a linear or nonlinear separating surface in the sample space. In order to find out the optimal separating surface, in terms of statistical learning theory [3], its VC dimensional complexity satisfies [8,11]: $VC \le \min\left\{\lceil D_{\max}{}^2/\Delta_{\min}{}^2\rceil, d\right\}+1$, where $D_{\max}$ denotes the maximal diameter of the smallest hypersphere which can cover all the samples, Δ_{min} denotes the minimal margin between two classes, and d denotes the dimensional number of samples. Therefore, in order to make VC minimal, we should try to make$\Delta_{\min}$ become the largest (or maximizing inter-class margin), and meanwhile try to make $D_{\max}$ minimal, namely minimizing within-class volume (or within-class scatter). Thus, the inter-class margin and within-class compactness are two important factors impacting on the performance of classifiers. The former leads to the classical support vector machine (SVM) [3,7], which is so-called linear unconstrained classifier, while the latter leads to its dual version, i.e., the hypersphere SVM [8], which is also called gap-tolerant classifier [11].

Traditionally, how to control the distribution spread (or scatter in this context) of data has been an important issue in pattern classification domain. There have existed many effective methods used to address this issue up to date [23]. For instance, classical linear discriminant analysis (LDA) [25] finds the projections of the data such that the inter-class separation margin is large as much as possible while the within-class scatter is small. However, the second-order statistics in LDA are inappropriate for many real-world data sets, and thus the classification performance of LDA is typically poorer than that of SVMs [23]. The estimation of spread should be tied to the margin criterion which is superior to the second-order assumptions about the data distribution [3,23]. Along this study line, ellipsoidal kernel machines [11] were proposed to normalize data in feature space by estimating the boundary hyper-ellipsoids while avoiding second-order assumptions. Besides, very recently, relative margin machine (RMM) was initially proposed by Shivaswamy et al. [17], in which margin was measured in a relative sense rather than in the absolute sense. Similarly, Dredze et al. [26] and Crammer et al. [27] consider a distribution on the perceptron hyperplane. These distribution assumptions permit update rules that resemble the whitening

* Corresponding author.
E-mail address: jianwen_tao@aliyun.com (J. Tao).

http://dx.doi.org/10.1016/j.neucom.2014.09.017
0925-2312/

of the data [28], thus alleviating adversarial affine transformations and producing changes to the basic maximum margin formulation, which are similar in spirit to those which RMM provides.

Though the maximum margin classifiers like SVMs work well in practice, their solutions are very sensitive to the affine or scaling transformation of the input data space [23] due to the reason that they exclusively measure margin by the data near the separation boundary regardless of how spread the remaining data is away from the separating hyperplane. For example, by transforming both training and testing data by an invertible linear transformation, the solution of SVM and its resultant classification performance can be significantly changed, which can naturally occur in many real world problems in pattern classification, especially in high dimensions [23]. Since simultaneously augmenting the inter-class margin and diminishing within-class scatter may be beneficial to the improvements of classification performance, one may consider distributions over large-margin classifier solutions, which provide a different estimate than the maximum margin setting in SVMs. This has shown empirical improvements over SVMs [1,12,11,17,23]. Note that, while RMM has achieved better classification performance than SVMs in some real-world applications by controlling the data spread of input space, it may still suffer from several practical problems such as class data imbalance and distribution inconsistency of two class data, which will be discussed in Section 4.2.

To address the shortcomings existing in SVMs and RMM, we propose a novel minimum class spread constrained support vector machine (MCSSVM) model by controlling the within-class spread of data whilst maximizing the separating margin between classes. One distinctive feature of our method is to recover a large margin solution normalized by explicitly controlling the within-class distribution spread of the input data. We may use a 2D classification problem illustrated in Fig. 1 to explain the difference among SVMs, RMM and our method. In Fig. 1(a), both classes are separated respectively by SVM with the largest margin, and MCSSVM and RMM with both minimal spread of data and maximal separation margin. Intuitively, the separation hyperplanes of RMM and MCSSVM are more reasonable than that of SVM. But, the classification hyperplane of RMM was obviously biased for large class data due to the class data imbalance. Hence, our method MCSSVM is superior to both SVM and RMM. Next, we will further explore the scaling effect, which is a particular affine transformation. To explore the scaling effect in a controlled manner, first, the projection w of the separation line recovered by the optimal RMM classifier is obtained. Therefore an orthogonal vector v (such that $w^T v = 0$, where a^T denotes the transposition of the vector or matrix a) can be obtained. The examples (training, test and validation) are then projected onto the axes defined by w and v. Each projection along w is preserved while the projection along v is scaled by a factor $s > 1$. This merely stretches the data further along directions orthogonal to w (i.e., along the optimal classification boundary). More concisely, given an example x, the following scaling transformation is applied:

$$\begin{bmatrix} w & v \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & s \end{bmatrix} \begin{bmatrix} w & v \end{bmatrix}^T x.$$

Fig. 1(b) shows the test accuracies of SVM, RMM and MCSSVM across a range of scaling values s. As s grows, SVM further deviates from the optimal RMM classifier and attempts to separate the data along directions of large margin. While both RMM and MCSSVM remain resilient to scaling, our proposed method MCSSVM maintains a relatively higher accuracy than those of SVM and RMM. The situation depicted in Fig. 1 occurs whenever the data can be enclosed by a (non-spherical) ellipsoid [32]. Fig. 1(b) shows that the optimal hyperplane of SVM is not a scale invariant and predictions of class labels may change if the data is rescaled before learning. Meanwhile, RMM may not obtain the best classification performance in the case of class data imbalance. But these drawbacks existing in SVM and RMM can be overcome by MCSSVM via simultaneously considering the within-class spread and the inter-class separation margin.

Compared with the state-of-the-art methods, several main contributions of this paper can be summarized as follows:

(1) We propose a robust support vector classification model coined as MCSSVM via constraining the within-class spread and maximizing the inter-class margin of input data, the optimal solution of which is insensitive to the affine or scaling transformation of input space. It is justified that MCSSVM can simultaneously augment both the inter-class margin and within-class compactness.
(2) The proposed model is presented by constructing a minimal within-class q-spread tube enclosing all input data from one class while keeping another class far away from this tube as large as possible. For this end, two additional variables q and ρ as well as two tunable parameters μ, ν are introduced into the model, so as to simultaneously control the within-class spread and the inter-class margin of input data. Moreover, it is theoretically proved that the parameters μ, ν can control both

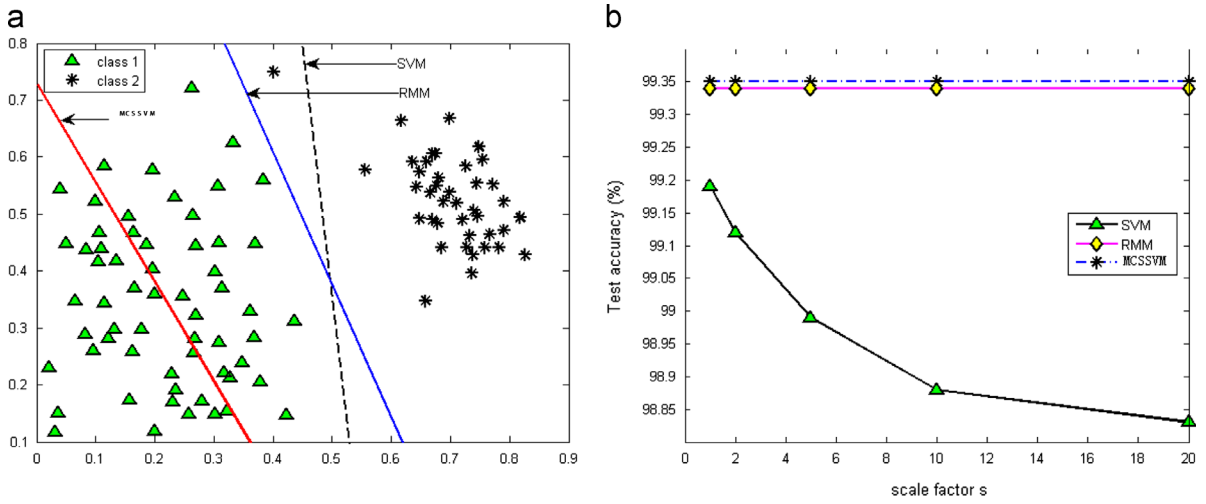

Fig. 1. (a) Data points from two classes (triangles and stars) are separated by the hyperplanes of SVM, RMM and MCSSVM, respectively; and (b) classification accuracy vs. scaling factor s