Contents lists available at ScienceDirect

Applied Soft Computing

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

Enhanced algorithm for high-dimensional data classification

Xiaoming Wang [a,*], Shitong Wang [b]

[a] *School of Computer and Software Engineering, Xihua University, Chengdu 610039, China*
[b] *School of Digital Media, Jiangnan University, Wuxi 214122, China*

ARTICLE INFO

ABSTRACT

Minimum class variance support vector machine (MCVSVM) and large margin linear projection (LMLP) classifier, in contrast with traditional support vector machine (SVM), take the distribution information of the data into consideration and can obtain better performance. However, in the case of the singularity of the within-class scatter matrix, both MCVSVM and LMLP only exploit the discriminant information in a single subspace of the within-class scatter matrix and discard the discriminant information in the other subspace. In this paper, a so-called twin-space support vector machine (TSSVM) algorithm is proposed to deal with the high-dimensional data classification task where the within-class scatter matrix is singular. TSSVM is rooted in both the non-null space and the null space of the within-class scatter matrix, takes full advantage of the discriminant information in the two subspaces, and so can achieve better classification accuracy. In the paper, we first discuss the linear case of TSSVM, and then develop the nonlinear TSSVM. Experimental results on real datasets validate the effectiveness of TSSVM and indicate its superior performance over MCVSVM and LMLP.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

In the past decade, kernel methods [1] have been widely studied and applied [2–4]. Support vector machine (SVM), as the most well-known representative of kernel methods, is a powerful machine learning method based on Vapnik's Statistical Learning Theory [5] and draws lots of researchers' attentions [6–10]. Different from other methods which usually attempt to minimize the misclassification errors on the training set (empirical risk minimization), SVM minimizes the structural risk, which is the probability of misclassifying a previously unseen sample [5,11]. The essential point of SVM is to find a separating hyperplane which achieves the maximal margin among two different classes of data. The basic idea of SVM has also been used to deal with regression problems and the corresponding method is called support vector regression (SVR) [12], which is a very effective regression method. The regression problem of SVR can actually be viewed as a classification problem of SVM in the dual space. Therefore, SVR can be seen as the most common application form of SVM and they share many same properties, for example the structural risk minimization principle. The one-class SVM algorithm [13] also stems from SVM and is designed as a data description method. One-class SVM shows appealing performance in the field of the outlier detection. It employs the strategy of constructing a hyperplane in the feature space in order to distinguish the normal data points from the given dataset which possibly contains the abnormal ones. Actually, the hyperplane separates the normal data points from the origin with maximum margin. Obviously, one-class SVM embodies the basic idea of SVM.

However, SVM does not take the class distribution into consideration and may result in a non-robust solution [14], i.e., the solution of SVM is easily impacted by outliers or noises. This means that its generalization performance or ability to correctly classify unknown samples may be impaired. In order to overcome the drawback of

SVM, a modified class of SVM called minimum class variance support vector machine (MCVSVM) was presented in [14] which is inspired from the optimization of Fisher's discriminant analysis (FDA) [15,16]. Unlike SVM, in which only the samples in the boundaries are taken into consideration, the solution of MCVSVM takes full advantage of both the samples in the boundaries and the distribution of the classes and gives a robust solution. In [17], MCVSVM were further extended to directly solve multiclass classification problems and a novel class of multiclass classifiers called minimum within-class variance multiclass classifiers (MWCVMC) were introduced. Similarly to MCVSVM, MWCVMC is also inspired by multiclass FDA and SVM. Moreover, the authors investigated and solved MWCVMC based on indefinite kernels and dissimilarity measures via pseudo-Euclidean embedding in [17]. In [18], the authors introduced the intrinsic manifold structure of the data space into MCVSVM and a novel method called minimum class locality preserving variance support vector machine MCLPVSVM is proposed. Following the way that MCVSVM incorporates the distribution of the classes by using the within-class covariance matrix, MCLPVSVM incorporates the intrinsic geometry of the data and local structure by using the locality preserving within-class covariance matrix when we define the distance for the sample point to the decision hyperplane. MCLPVSVM are very closely related to MCVSVM and share some properties with the SVM and the MCVSVM.

Similarly to FDA, MCVSVM encounters the singularity of the within-class scatter matrix in the high-dimensional data classification task, e.g., in the small sample size (SSS) problem [19–21]. Here the singularity problem is that the within-class scatter matrix is not invertible but its inverse is necessary during solving the optimization problem of MCVSVM. In order to tackle this problem, a good method is to first perform eigenanalysis to the within-class scatter matrix and then implement MCVSVM only in the non-null space [22,23] which is spanned by the eigenvectors corresponding to non-zero eigenvalues [14]. This method essentially removes the null space which is spanned by the eigenvectors corresponding to zero eigenvalues and dodges the singularity problem. However, as was pointed out in [24], the null space generally contains the most discriminative information. Obviously, it can be observed that MCVSVM is just carried out in a single subspace, which leads to a loss of some significant discriminant information in the high-dimensional data space.

On the other hand, the large margin linear projection (LMLP) classifier [25] is rooted only in the null space. LMLP takes full advantage of the singularity of the within-class scatter matrix, and classifies projected points in one-dimensional space by itself. Experimental results indicated the effectiveness of LMLP for the high-dimensional classification problems such as face recognition and image classification. However, LMLP obviously ignore the discriminant information in the non-null space.

So, in the case of the high-dimensional data classification where the singularity of the within-class scatter matrix occurs, both MCVSVM and LMLP only exploit the discriminant information in a single subspace. Thus, when employing MCVSVM or LMLP, it is obvious that certain discriminative information resides in the other subspace which is discarded. So, MCVSVM and LMLP are opposite extremes to each other, and lose the discriminant information residing in the complementary subspace of the within-class scatter matrix since the non-null space and the null space are orthogonal and complementary with respect to discriminative power.

Aiming at the drawbacks of MCVSVM and LMLP, in the paper, we propose a novel classification algorithm called twin-space support vector machine (TSSVM) to deal with the high-dimensional data classification task. Formally, the optimization problem of TSSVM is a combination of the optimizations of MCVSVM and LMLP. In essence, however, TSSVM is different from MCVSVM and LMLP. The key difference between TSSVM and the other two methods is that the former explicitly exploits the information of both the null space and the non-null space of the within-class scatter matrix, whereas the latter two are only rooted in a single subspace. In the paper, we discuss the linear case of TSSVM. After that, we first present a novel alternative version of the nonlinear MCVSVM and develop a kernelization algorithm of LMLP since in [25] the authors did not present the nonlinear LMLP, and then propose the nonlinear TSSVM. Experimental results on real datasets validate the effectiveness of TSSVM and indicate its superior performance over MCVSVM and LMLP.

The rest of this paper is organized as follows. Section 2 briefly reviews the related work. In Section 3, the linear case of TSSVM is discussed. In Section 4, an alternative version of the nonlinear MCVSVM and the nonlinear LMLP are first proposed, and then the nonlinear TSSVM is defined and solved. The experimental results are reported in Section 5. Finally, conclusions are drawn in Section 6.

2. Related work

In this paper, we suppose a training dataset which contains two classes of N samples, represented by $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, with training samples $\mathbf{x}_i \in R^M$ and corresponding labels $y_i \in \{1, -1\}$, where $i = 1, \ldots, N$ and M is the dimension of the sample space. Note, in the paper we focus on the classification problem in the high-dimensional sample space where the singularity of the within-class scatter matrix occurs.

2.1. Fundamentals

For the above given training dataset, the within-class scatter matrix $\mathbf{S}_W$ and the between-class scatter matrix $\mathbf{S}_B$ are respectively defined as [24]

$$\mathbf{S}_W = \sum_{k=1}^{2} \sum_{\mathbf{x} \in \mathbf{X}_k} (\mathbf{x} - \mathbf{u}_k)(\mathbf{x} - \mathbf{u}_k)^T \tag{1}$$

$$\mathbf{S}_B = (\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{u}_1 - \mathbf{u}_2)^T \tag{2}$$

where $\mathbf{X}_1 = \{\mathbf{x}_i | y_i = 1, i = 1, \ldots, N\}$, $\mathbf{X}_2 = \{\mathbf{x}_i | y_i = -1, i = 1, \ldots, N\}$, $\mathbf{u}_k = \frac{1}{N_k} \sum_{\mathbf{x} \in \mathbf{X}_k} \mathbf{x}$ and T denotes vector transpose. Here, N_k is the cardinality of $\mathbf{X}_k$.

Definition 1 ([26]). Suppose $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M\}$ are M orthonormal eigenvectors of $\mathbf{S}_W$ and the first r ($r = rank \, \mathbf{S}_W$) ones $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r\}$ are corresponding to non-zero eigenvalues, then we define the non-null space $\boldsymbol{\Omega}$ and the null space $\boldsymbol{\Psi}$ of $\mathbf{S}_W$ as $\boldsymbol{\Omega} = span\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r\}$ and $\boldsymbol{\Psi} = span\{\boldsymbol{\beta}_{r+1}, \ldots, \boldsymbol{\beta}_M\}$, respectively.

Lemma 2 ([26]). For the non-null space $\boldsymbol{\Omega}$ and the null space $\boldsymbol{\Psi}$ of $\mathbf{S}_W$ as defined above, $\boldsymbol{\Psi}$ is orthogonal complement of $\boldsymbol{\Omega}$. Also, any arbitrary $\mathbf{w} \in \mathbf{R}^M$ can be denoted by $\mathbf{w} = \boldsymbol{\zeta} + \boldsymbol{\eta}$, where $\boldsymbol{\zeta} \in \boldsymbol{\Omega}$ and $\boldsymbol{\eta} \in \boldsymbol{\Psi}$.

2.2. Minimum class variance support vector machine

In order to overcome the drawback of SVM, a modified class of SVM called MCVSVM was proposed in [14], which takes into consideration both the samples in the boundaries and the distribution of the classes and gives a robust solution. In the case where the training samples are not linearly separable the optimization problem of MCVSVM is defined as

$$\min_{\mathbf{w}, \, b} \quad \mathbf{w}^T \mathbf{S}_W \mathbf{w} + C \sum_{i=1}^{N} \xi_i \tag{3}$$

$$s.t. \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad \xi_i \geq 0, \quad i = 1, \ldots, N$$

If the within-class scatter matrix $\mathbf{S}_W$ is nonsingular, similarly to SVM, the optimization problem (3) of MCVSVM could be efficiently solved by switching to its Wolfe dual problem using a Lagrangian formulation of the problem [27]. However, when encountering the singularity of $\mathbf{S}_W$, e.g., in the case of the SSS problem, ones cannot directly use this way to tackle the optimization problem (3). In Section 3.1 we will discuss the case in detail.

2.3. Large margin linear projection (LMLP) classifier

Unlike MCVSVM and various LDA methods which try hard to avoid the singularity of $\mathbf{S}_W$, LMLP takes full advantage of the characteristic. If $\mathbf{S}_W$ is singular, there is at least one non-zero vector $\boldsymbol{\eta}$ such that $\boldsymbol{\eta}^T \mathbf{S}_W \boldsymbol{\eta} = 0$. In this case, LMLP defines its optimization problem as follows [25]

$$\max_{\boldsymbol{\eta}} \quad \boldsymbol{\eta}^T \mathbf{S}_B \boldsymbol{\eta} \tag{4}$$

$$s.t. \quad \boldsymbol{\eta}^T \mathbf{S}_W \boldsymbol{\eta} = 0, \quad \boldsymbol{\eta}^T \boldsymbol{\eta} = 1$$

Note, $\boldsymbol{\eta}^T \mathbf{S}_W \boldsymbol{\eta} = 0$ implies $\boldsymbol{\eta} \in \boldsymbol{\Psi}$. Thus, the above optimization problem (4) can be rewritten as

$$\max_{\boldsymbol{\eta}} \quad \boldsymbol{\eta}^T \mathbf{S}_B \boldsymbol{\eta} \tag{5}$$

$$s.t. \quad \boldsymbol{\eta} \in \boldsymbol{\Psi}, \quad \boldsymbol{\eta}^T \boldsymbol{\eta} = 1$$

In the linear case, similarly to SVM, LMLP wants to find a linear decision function or hyperplane to classify the samples. In LMLP, however, the threshold is set to 0. Thus, suppose $\boldsymbol{\eta}^*$ is the optimal solution of the optimization problem (5) of LMLP, then the decision function of LMLP is given as

$$f(\mathbf{x}) = sgn((\boldsymbol{\eta}^*)^T \mathbf{x}) \tag{6}$$

As pointed out in [25], LMLP is a special kind of SVM which assumes that the samples are linearly separable. Thus, LMLP inherits the excellent characteristics of SVM to some extent.

3. Twin-space support vector machine

In this section, we first analyze and illustrate the problems with MCVSVM and LMLP in the case of the singularity of $\mathbf{S}_W$, then define the optimization problem of our method TSSVM. Finally, how to solve the optimization problem is discussed.

3.1. The problems with MCVSVM and LMLP

In the case where $\mathbf{S}_W$ is nonsingular, ones can solve the optimization problem (3) of MCVSVM by its dual optimization problem. When $\mathbf{S}_W$ is singular, however, this way cannot be directly employed to solve (3) because during switching (3) into its dual problem the inverse of $\mathbf{S}_W$ is necessary but it cannot be directly obtained. A method to deal with this case is to employ principal component analysis (PCA) [28] to transform the samples into a low-enough-dimensional subspace where the new within-class scatter matrix is nonsingular, and then the MCVSVM algorithm is carried out in the new subspace. This method needs to predefine the dimensionality of the low-dimensional subspace.