Pattern Recognition 47 (2014) 3168–3178

Contents lists available at ScienceDirect

Pattern Recognition

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

A locality correlation preserving support vector machine

Huaxiang Zhang [a,b,*], Linlin Cao [a,b], Shuang Gao [a,b]

[a] Department of Computer Science, Shandong Normal University, Jinan 250014, Shandong, China
[b] Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, China

ARTICLE INFO

Article history:
Received 17 March 2013
Received in revised form
13 November 2013
Accepted 1 April 2014
Available online 13 April 2014

Keywords:
Support vector machine
Kernel methods
Locality correlation preservation
Fuzzy membership

ABSTRACT

This paper proposes a locality correlation preserving based support vector machine (LCPSVM) by combining the idea of margin maximization between classes and local correlation preservation of class data. It is a Support Vector Machine (SVM) like algorithm, which explicitly considers the locality correlation within each class in the margin and the penalty term of the optimization function. Canonical correlation analysis (CCA) is used to reveal the hidden correlations between two datasets, and a variant of correlation analysis model which implements locality preserving has been proposed by integrating local information into the objective function of CCA. Inspired by the idea used in canonical correlation analysis, we propose a locality correlation preserving within-class scatter matrix to replace the within-class scatter matrix in minimum class variance support machine (MCVSVM). This substitution has the property of keeping the locality correlation of data, and inherits the properties of SVM and other similar modified class of support vector machines. LCPSVM is discussed under linearly separable, small sample size and nonlinearly separable conditions, and experimental results on benchmark datasets demonstrate its effectiveness.

1. Introduction

The support vector machines (SVMs) [1] are based on the idea of structural risk minimization to classify instances, and have been shown to be more powerful tools than traditional machine learning approaches. Different from other classification approaches which simply minimize the training error, SVMs aim to maximize the class margin for linearly separable data. When dealing with nonlinear classification data, SVMs adopt kernel-based mapping approaches to map the data in the original space to a high-dimensional feature space, in which the data become linearly separable or near-linearly separable, and obtain a maximum margin hyperplane determined by some support vectors. As the solution of SVMs is exclusively determined by support vectors, and all other data are irrelevant to the decision hyperplane [2], they have sparseness and unique solution properties. In SVMs, classification problems are reformulated by convex quadratic programming (QP) problems which can guarantee unique and global solutions.

Although the SVM algorithm works effectively for classification problems, it ignores the inner structure of the data, and the generalization capability of its solution is low [3,4]. Several efforts have been made to deal with its disadvantages. Inspired by the optimization of Fisher's discriminant ratio [5], a modified version of SVM has been constructed by merging Fisher's discriminant and SVMs [6]. It takes consideration of the local similarity of the elastic graph nodes according to their discriminant power, and experimental results show that it outperforms typical SVMs. In order to overcome the initiated singular within-class scatter matrix problem when the number of training vectors is larger than the feature dimensionality, the method is extended to a minimum class variance support vector machines (MCVSVM) [3], and the solution of MCVSVM can be found through principal component analysis (PCA) dimensionality reduction [7]. When the data are nonlinearly separable, it has also been proved that, under kernel PCA (KPCA) [8], the classification problem can be reformulated into an equivalent linear MCVSVM problem. Different from typical SVMs, the optimization problem of MCVSVM considers the class distribution while ensuring class separability, so it has the advantages of Fisher's linear discriminant analysis (FLDA) [9] and SVMs. But MCVSVM ignores the embedded manifold structure of data.

In order to overcome the drawbacks of SVM and MCVSVM, an algorithm called minimum class locality preserving variance support machine (MCLPV_SVM) is proposed [4], which explicitly considers the manifold structure of data in the optimization problem. MCLPV_SVM merges the locality preserving projections (LPP) [10–12] with SVMs, and preserves the local manifold structure of the data by a nearest-neighbor graph for each class. A locality preserving within-class scatter matrix is defined and used in the optimization problem of MCLPV_SVM. MCLPV_SVM inherits

* Corresponding author at: Department of Computer Science, Shandong Normal University, Jinan 250014, Shandong, China.
E-mail address: huaxzhang@hotmail.com (H. Zhang).

http://dx.doi.org/10.1016/j.patcog.2014.04.004
0031-3203/

advantages of both SVM and MCVSVM. Another similar work called the Laplacian support vector machine (LSVM) [13] also combines LPP and SVM, but it aims at semi-supervised learning problems.

Slack variables are introduced to allow some of the training data to be misclassified, and form a penalty term in the optimization function of typical linearly separable SVMs. A parameter is used to control the trade-off between the slack variable penalty and the margin. Researches show that considering the class distribution in the penalty term of SVM improves its classification performance [14–16]. Each training sample is assigned a fuzzy membership which represents its contribution to the learning of decision surface, and SVM is reformulated into fuzzy SVM (FSVM) [14]. An extension of FSVM is proposed in [15], in which two membership values are assigned to each training sample. Based on the density distribution information of the training data, a density-induced margin SVM (DMSVM) is presented [16]. A relative density degree is assigned to each data point as the relative margin, and is presented in the constraints of the optimization function to reflect the density distribution of data in SVMs.

Inspired by the idea of taking into account the local manifold structure of data in the margin of the optimization function of SVMs [4] or the class distribution in the penalty term [14,15], we propose a novel learning algorithm named as Locality Correlation Preserving SVM (LCPSVM) in which the locality correlation within each class is explicitly considered in the margin and the penalty term. Many locality-preserving approaches have been proposed to implement dimensionality reduction while keeping unchanged the manifold structure of the training data, they also have been used commonly to deal with nonlinear classification problems. These approaches include locally linear embedding (LLE) [17], Locality preserving projection (LPP) [10], and Isomap [18]. They take into account the local neighborhood structure of data to discover the low dimensional manifold structure embedded in the original high dimensional space. Canonical Correlation Analysis (CCA) [19] is used to reveal the hidden correlations between two datasets, and by integrating local information into the objective function of CCA, a Correlation Analysis model which implements Locality preserving (LCA) [20] is proposed. In LCA, global means of data are replaced by local means when revealing hidden correlations between two datasets. Inspired by the idea used in LCA, we replace the class mean sample vectors with particular local mean sample vectors in the within-class scatter matrix, and propose a locality correlation preserving within-class scatter matrix. This substitution has the property of keeping the locality correlation.

The rest of the paper is organized as follows. Section 2 briefly reviews the SVM learning theory and discusses the related works, and Section 3 presents the linear case of Locality Correlation Preserving SVM (LCPSVM). The relationship of LCPSVM with the related existing approaches is discussed in Section 4. Section 5 presents the small sample size case when the number of the training samples is less than the feature dimensionality of samples, and the problems under nonlinear conditions are solved in Section 6. Section 7 reports the experimental results, and Section 8 concludes the paper.

2. Related works

We will briefly introduce SVM, MCVSVM, and MCLPV_SVM in this section. Only two class classification problems are discussed in this paper, and multiclass classification problems can be solved by constructing multiple separate SVMs. Given a training dataset $\{(\mathbf{x}_i, t_i)\}_{i=1}^n$ of n samples with input data $\mathbf{x}_i \in \Re^m$ and corresponding binary class labels $t_i \in \{+1, -1\}$, samples with class label $+1$ belong to X_+, and all the others belong to X_-. For the soft margin SVM, the primal optimization problem used to construct a linear decision surface can be expressed as follows:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i, \qquad \text{s.t.} \quad t_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ i = 1, 2, \ldots, n \tag{1}$$

where $\mathbf{w}$ is a weight vector and $\mathbf{w} \in \Re^m$, $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)$ is the vector of the slack variables, and C is a parameter that controls the trade-off between the slack variable penalty and the margin. Nonlinear decision surfaces can also be constructed by applying a nonlinear function $\boldsymbol{\varphi}$ to map the samples in the original space to a very high dimensional feature space ($\mathbf{x}_i$ is replaced with $\boldsymbol{\varphi}(\mathbf{x}_i)$ in (1)). In the feature space, samples are linearly or near-linearly separable, and a maximum margin hyperplane can be found. The classifier takes the form $\mathrm{sgn}(\mathbf{w}^T\mathbf{x} + b)$ for a linear decision surface or $\mathrm{sgn}(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}) + b)$ for a nonlinear decision surface.

2.1. Minimum class variance support vector machine (MCVSVM)

MCVSVM [3] is a modified class of SVM. It takes into account the class distribution of the training data when constructing a separable surface, that means the constructed surface is not only exclusively determined by the support vectors (SVs), but also is influenced by all other non-SVs, thus its solution is more robust than that of typical SVMs. MCVSVM uses the following optimum function with soft margin to learn a linear decision surface:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad \mathbf{w}^T\mathbf{S}_w\mathbf{w} + C\sum_i \xi_i, \qquad \text{s.t.} \quad t_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ i = 1, 2, \ldots, n \tag{2}$$

where the within-class scatter matrix $\mathbf{S}_w$ is defined by

$$\mathbf{S}_w = \sum_{\mathbf{x}_i \in X_+} (\mathbf{x}_i - \mathbf{u}_+)^T(\mathbf{x}_i - \mathbf{u}_+) + \sum_{\mathbf{x}_i \in X_-} (\mathbf{x}_i - \mathbf{u}_-)^T(\mathbf{x}_i - \mathbf{u}_-) \tag{3}$$

in (3), $\mathbf{u}_+$ and $\mathbf{u}_-$ are the mean sample vectors for class X_+ and X_- respectively.

When $\mathbf{S}_w$ is nonsingular, in order to solve the optimization problem (2), a transformation of the problem to its Wolfe dual problem using a Lagrangian formulation is made. But when the number of training data is smaller than the dimensionality of the training data, $\mathbf{S}_w$ may be singular, the solution of (2) in such case can be obtained through PCA dimensionality reduction. The optimization problem of the nonlinear MCVSVM decision surfaces can be found similarly.

2.2. Minimum class locality preserving variance support vector machine (MCLPV_SVM)

MCLPV_SVM [4] takes into consideration the intrinsic manifold structure of data in each class. LPP [10] and Isomap [18] have been widely used in face recognition [11,12] to preserve the intrinsic geometry of data and local structure, and MCLPV_SVM merges LPP with SVM and inherits the characteristics of SVM and MCVSVM. For each vector $\mathbf{x}_i$ in the training set, $ne(\mathbf{x}_i)$ denotes its k nearest neighbors with the same class label, a weight matrix $\mathbf{V}$ is calculated with its ith row entries associated with $\mathbf{x}_i$ being defined by the Gaussian kernel as follows:

$$v_{ij} = \begin{cases} e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma} & \text{if } \mathbf{x}_j \in ne(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in ne(\mathbf{x}_j) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\|\mathbf{x}\|^2 = \sum_{i=1}^m x_i^2$ denotes the Euclidean norm in $\Re^m$, and σ is the heat kernel parameter greater than 0. Let D be a $M \times M$ diagonal matrix with its diagonal entries $d_{ij} = \sum_{j=1}^M v_{ij}$ (M denotes the number of samples), given the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{V}$, the locality preserving scatter matrix for the positive class data and the negative class data is defined by $\mathbf{Z}_+ = \mathbf{X}_+\mathbf{L}_+(\mathbf{X}_+)^T$ and $\mathbf{Z}_- = \mathbf{X}_-\mathbf{L}_-(\mathbf{X}_-)^T$ separately, where $\mathbf{L}_+$ is the Laplacian matrix