Pattern Recognition 45 (2012) 66–79

Contents lists available at ScienceDirect

Pattern Recognition

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

A novel SVM+NDA model for classification with an application to face recognition[☆]

N.M. Khan[a,1], R. Ksantini[b,*], I.S. Ahmad[b], B. Boufama[b]

[a] Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada M5B 1Z2
[b] School of Computer Science, University of Windsor, Windsor, ON, Canada N9B 3P4

ARTICLE INFO

Article history:
Received 10 December 2009
Received in revised form
29 April 2011
Accepted 8 May 2011
Available online 18 May 2011

Keywords:
Linear discriminant analysis (LDA)
Nonparametric discriminant analysis (NDA)
Support vector machines (SVM)
Kernel NDA and SVM
Small sample size problem
Face recognition

ABSTRACT

Support vector machine (SVM) is a powerful classification methodology, where the support vectors fully describe the decision surface by incorporating local information. On the other hand, nonparametric discriminant analysis (NDA) is an improvement over LDA where the normality assumption is relaxed. NDA also detects the dominant normal directions to the decision plane. This paper introduces a novel SVM+NDA model which can be viewed as an extension to the SVM by incorporating some partially global information, especially, discriminatory information in the normal direction to the decision boundary. This can also be considered as an extension to the NDA where the support vectors improve the choice of k-nearest neighbors on the decision boundary by incorporating local information. Being an extension to both SVM and NDA, it can deal with heteroscedastic and non-normal data. It also avoids the small sample size problem. Moreover, it can be reduced to the classical SVM model, so that existing softwares can be used. A kernel extension of the model, called KSVM+KNDA is also proposed to deal with nonlinear problems. We have carried an extensive comparison of the SVM+NDA to the LDA, SVM, heteroscedastic LDA (HLDA), NDA and the combined SVM and LDA on artificial, real and face recognition data sets. Results for KSVM+KNDA have also been presented. These comparisons demonstrate the advantages and superiority of our proposed model.

1. Introduction

In the context of supervised learning, given a training set $\mathcal{X}$ of input vectors $\{x_i\}_{i=1}^N$, where $x_i \in \mathbb{R}^M (M \geq 1)\ \forall i=1,2,\ldots,N$, along with corresponding set $\mathcal{T}$ of tags $\{t_i\}_{i=1}^N$, where $t_i \in \mathbb{Z}\ \forall i=1,2,\ldots,N$, we wish to learn a model of dependency of the targets on the inputs. The final objective is to be able to make accurate predictions of t for unseen values of x. In the case of real world data, the presence of class overlap in classification implies that the principal modeling challenge is to avoid over-fitting of the training set. Typically, we base our predictions upon some function $y(x)$ defined over the input space (or training space) $\mathcal{X}$, and learning is the process of inferring the parameters or weights of this function. We concentrate here on functions of the type corresponding to those implemented by some relevant linear models, such as, the support vector machine (SVM) [32] and the linear discriminant analysis (LDA) [7,26]. The SVM and LDA make predictions based on the function:

$$y(x;w) = \sum_{i=1}^{M} w_i a_i + w_0 = w^T x + w_0, \tag{1}$$

where $x = \{a_i\}_{i=1}^M$ represents one input vector. Each a_i represents one attribute of the underlying class. $w = \{w_i\}_{i=1}^M$ and w_0 represent the unknown weights to compute.

The practical attractiveness of LDA can be explained by its (intrinsically) low model complexity, and its ability to capture the essential characteristics of the data distributions (mean and covariance) from finite training data, and then estimating the decision boundary using these "global" characteristics of the data. The LDA has proven to be powerful and competitive to several linear classifiers [31]. Its main goal is to find linear projections such that the classes are well separated, i.e., maximizing the distance between means of classes and minimizing their intra-class variances. The LDA has successfully been applied in appearance-based methods for object recognition, such as face recognition [24] and mobile robotics [14].

However, the LDA is incapable of dealing explicitly with heteroscedastic data, i.e., data in which classes do not have equal

[☆] N.M. Khan and R. Ksantini contributed equally to this work.

[*] Corresponding author.

E-mail addresses: n77khan@ee.ryerson.ca (N.M. Khan), ksantini@uwindsor.ca (R. Ksantini), imran@uwindsor.ca (I.S. Ahmad), boufama@uwindsor.ca (B. Boufama).

[1] N.M. Khan completed this work when he was still at the School of Computer Science, University of Windsor.

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.05.004

covariance matrices [26]. Moreover, most of the existing LDA-based methods inherit the parametric nature from the traditional LDA approach—the construction of the scatter matrices relies on the underlying assumption that the samples in each class satisfy the Gaussian distribution. Thus, they suffer from performance degradation in cases of non-normal distribution [15]. To overcome the heteroscedasticity problem, Loog and Duin [19] proposed the heteroscedastic LDA (HLDA) which is a heteroscedastic extension of the Fisher criterion and is based on the Chernoff distance. To relax the normality assumption, Fukunaga [15] proposed the nonparametric DA (NDA) which measures the between-class scatter matrix on a local basis in the neighborhood of the decision boundary. This is done based on the observation that the normal vectors on the decision boundary are the most informative for discrimination [5]. In the case of a two-class classification problem, these normal vectors are approximated by the *k-NN*'s from the other class for one point. Therefore, NDA can be considered as a classification method based on the "partially global" characteristics of data which are represented by the *k-NN*'s. Although NDA gets rid of the underlying assumptions of LDA and the results in better classification performance in case of non-Gaussian and heteroscedastic data, it is not always an easy task to find a common and appropriate choice of *k-NN*'s on the decision boundary for all data points to obtain the best linear discrimination.

Support vector machine (SVM) [32] is another powerful method which emphasizes the idea of maximizing the margin or degree of separation in the training data. There are many hyperplanes which can divide the data between two classes for classification. One reasonable choice for the optimal hyperplane is the one which represents the largest separation or margin between the two classes. SVM tries to find the optimal hyperplane using support vectors. The support vectors are the training samples that approximate the optimal separating hyperplane and are the most difficult patterns to classify [25]. In other words, they consist of those data points which are closest to the optimal hyperplane. As SVM deals with a subset of data points (support vectors) which are close to the decision boundary, it can be said that the SVM solution is based on the "local" characteristics of the data. However, SVM does not take into consideration the global or partially global properties of the class distribution on which LDA-based methods (e.g., LDA, HLDA, NDA) are based.

In this paper, we propose an SVM+NDA classification model which takes into account both the partially global characteristics of data distribution represented by NDA and the local characteristics represented by SVM. Being an extension to both SVM and NDA, this classification model does not depend on any global distribution pattern of training data. Therefore, it is capable of dealing with heteroscedastic and non-normal data. Moreover, our method combines the discriminatory information represented by the normal vectors to the decision surface and the support vectors which are crucial for accurate classification. Also, our method avoids the small sample size problem, which is a general problem for LDA-based methods (e.g., LDA, HLDA, NDA) [10]. The small sample size problem arises when the dimension of data is higher compared to the number of training samples. Our method solves this problem by using the regularization matrix [26]. We have particularly targeted the face recognition problem as an application of interest to our proposed model given that it has become one of the most challenging tasks in the pattern recognition area [2]. Furthermore, face recognition is also central to many other applications such as video surveillance and identity retrieval from databases for criminal investigations.

We also extend this linear SVM+NDA model to incorporate nonlinear relations. Like any other linear classifier, the famous *kernel trick* [32] has been used to extend this method. For kernel-based methods, the predictions of class targets are based on the following function:

$$y(x;w) = \sum_{i=1}^{N} w_i \langle \Phi(x),\Phi(x_i)\rangle + w_0 = \sum_{i=1}^{N} w_i \mathcal{K}(x;x_i) + w_0, \quad (2)$$

where $\mathcal{K}(x;x_i) = \langle \Phi(x),\Phi(x_i)\rangle$ represents the inner product of $\Phi(x)$ and $\Phi(x_i)$. The function $\Phi(x):\mathcal{X}\rightarrow\mathcal{F}$ describes a nonlinear mapping from the input space $\mathcal{X}$ to a feature space $\mathcal{F}$ of higher dimensionality. Again, $w=\{w_i\}_{i=1}^{N}$ and w_0 represent the unknown weights to compute in kernel space. Hence, nonlinear classifiers have two stages. First, a fixed nonlinear mapping transforms the data into a feature space $\mathcal{F}$. Then, a linear classifier is used to classify them in $\mathcal{F}$. Indeed, if we compare Eqs. (1) and (2), the major difference we see is between the input vectors used (x vs. $\Phi(x)$). Since kernel support vector machines (KSVM) and kernel nonparametric discriminant analysis (KNDA) [15,26,7] already exist, the philosophy behind our kernel extension (KSVM+KNDA) is similar to that of the linear method (SVM+NDA).

Rest of this paper is organized as follows. Section 2 provides the formulation of the classical SVM and NDA methods along with their kernel extensions. In Section 3, we present the derivation of the novel SVM+NDA model as well as the kernel extended model (KSVM+KNDA). We also show that both the SVM+NDA and the KSVM+KNDA models are variations of the classical SVM, so that the existing SVM softwares can be used. Section 4 provides a comparative evaluation of the SVM+NDA model to the LDA, SVM [32], NDA [15], HLDA [19] and combined SVM and LDA [29], carried out on a collection of benchmark synthetic and real data sets. The KSVM+KNDA model is also compared with the corresponding kernel methods, namely KSVM, KNDA and the kernel Fisher discriminant analysis (KFD) [26] for real data sets. This section also contains results of our experiments on face recognition. Finally, Section 5 provides some conclusion.

2. The SVM and NDA methods

Let $\mathcal{X}_1 = \{x_i\}_{i=1}^{N_1}$ and $\mathcal{X}_2 = \{x_i\}_{i=N_1+1}^{N_1+N_2}$ be two different classes constituting an input space of $N=N_1+N_2$ samples or vectors in $\mathbb{R}^M$. Here, class $\mathcal{X}_1$ contains N_1 samples and class $\mathcal{X}_2$ contains N_2 samples. Let the associated tags with these vectors be represented by $\mathcal{T}=\{t_i\}_{i=1}^{N}$, where $t_i \in \{+1,-1\}\ \forall i=1,2,\ldots,N$. As stated before, the goal of classifiers like SVM or Fisher's discriminant analysis is to construct an optimal linear separating hyperplane from the training data by using the function described in Eq. (1). In case of nonlinearity, the kernel trick is applied where the function Φ maps the classes $\mathcal{X}_1$ and $\mathcal{X}_2$ to higher dimensional $\mathcal{F}_1=\{\Phi(x_i)\}_{i=1}^{N_1}$ and $\mathcal{F}_2=\{\Phi(x_i)\}_{i=N_1}^{N_1+N_2}$ such that $\mathcal{K}(x_i;x_j)=\langle\Phi(x_i),\Phi(x_j)\rangle$, $\forall i,j=1,2,\ldots,N$.

In the linear case, Fisher's discriminant aims at finding linear projections such that the classes are well separated, i.e., maximizing the distance between means of the classes and minimizing their intra-class variances. Finding the most discriminative projectional direction w^* can be described by the following optimization problem:

$$w^* = \underset{w}{\operatorname{argmax}} \frac{w^T S_b w}{w^T S_w w}, \quad (3)$$

where the within-class scatter matrix contains within-class or class independent scatter information and is defined as

$$S_w = \frac{1}{N_1+N_2}(N_1 S_1 + N_2 S_2), \quad (4)$$

where S_1 and S_2 are the covariance matrices for the two classes. The between-class scatter matrix contains between-class scatter