Knowledge-Based Systems 70 (2014) 154–166

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Structural max-margin discriminant analysis for feature extraction

Xiaobo Chen [a,*], Yan Xiao [b], Yinfeng Cai [a], Long Chen [a]

[a] Automotive Engineering Research Institute, Jiangsu University, Zhenjiang 212013, PR China
[b] School of Chemistry and Chemical Engineering, Jiangsu University, Zhenjiang 212013, PR China

ARTICLE INFO

Article history:
Received 13 August 2013
Received in revised form 30 May 2014
Accepted 14 June 2014
Available online 24 June 2014

Keywords:
Discriminant analysis
Max-margin principle
Quadratic programming
Constrained concave–convex procedure
Column generation

ABSTRACT

Subclass discriminant analysis (SDA) is a recently developed dimensionality reduction technique which takes into consideration the intrinsic structure information lurking in data by approximating unknown distribution of each class with multiple Gaussian distributions, namely, subclasses. However, in SDA, the separability between heterogeneous subclasses, i.e. those from different classes, is measured by the between-subclass scatter calculated as average distance between the means of these subclasses. In this paper, in the view of maximum margin principle, we propose a novel feature extraction method coined structural max-margin discriminant analysis (SMDA), in order to enhance the performance of SDA. Specifically, SMDA targets at finding an orthogonal linear embedded subspace in which the margin, defined as the minimum pairwise between-subclasses distance, is maximized and simultaneously the within-subclasses scatter is minimized. The concrete formulation of the resulting model boils down to a nonconvex optimization problem that can be solved by combining the constrained concave–convex procedure with the column generation technique. We evaluate the proposed SMDA on several benchmark datasets and the experimental results confirm the effectiveness of the proposed method.

1. Introduction

In exploratory data analysis problems, most of real-world data have a large number of input variables, many of which are noisy and filling with redundant information. As a promising way to reduce the number of variables, and thus the dimensionality of problems, feature extraction has become a crucial preprocessing step to analyze these data for two reasons. First, the storage requirement and computational cost for the subsequent machine learning tasks such as classification and visualization, will be cut down remarkably on the refined data. Second, the generalization ability of learning algorithms can be vastly improved due to the removal of noisy or irrelevant features. A category of popular and widely used feature extraction approach known as subspace learning aims at seeking a low-dimensional feature space in which the projected data can be well reconstructed or separated according to the specific purpose.

Principle component analysis (PCA) [1], multidimensional scaling (MDS) [2] and linear discriminant analysis (LDA) [3] are three representative subspace learning methods. PCA tries to find an orthonormal subspace in such a way that the projected data result in the largest variance. MDS attempts to find low-dimensional embeddings of data which preserve the pairwise distance between original data as accurately as possible. Actually, PCA can be viewed as a special case of MDS when the Euclidean distance is used. However, PCA and MDS fail to incorporate the available supervised information into their learning algorithms, thus may lead to poor classification performance. On the contrary to PCA and MDS, LDA exploits the class label information to find a subspace best discriminating different classes. This is achieved by maximizing the Fisher criterion, that is, the ratio of between-class scatter to within-class scatter. Therefore, it is generally believed that the features extracted by LDA are more reliable than those captured by PCA in pattern classification scenarios [3].

As is well-known, the subspace derived from LDA is optimal for binary-class problems in the sense that the Bayes error reaches its minimum as long as each class follows multivariate normal distributions having a common covariance matrix but different class means [4]. This assumption, however, is too strict to satisfy in many real-world applications because the data often own certain inherent structure which can described by neither within-class nor between-class scatter matrices. For instance, when the underlying data within each class are multimodally distributed, LDA may suffer severely due to the violation of above assumption. During the last decades, many extensions to the basic LDA have been proposed to cope with this limitation. A popular and widely used approach is reformulating the within-class and between-class

* Corresponding author. Tel.: +86 18362880812.
E-mail addresses: xbchen82@163.com, xbchen82@gmail.com (X. Chen).

http://dx.doi.org/10.1016/j.knosys.2014.06.020
0950-7051/

scatter matrices such that as much prior structure information as possible can be incorporated into the framework of LDA. Generally speaking, the possible ways to discover inner-structure of data can be divided into two categories: manifold-based and subclass-based approaches.

Manifold-based approaches tries to discover local geometric structure, the researchers have developed many manifold learning algorithms for dimensionality reduction, including locally linear embedding (LLE) [5], ISOMAP [6], etc. Unfortunately, these methods cannot yield an explicit mapping function from the original samples to their low-dimensional representations, thus making them cannot be applied to the samples which are not in the training set. He et al. utilized a linearization procedure and proposed locality preserving projection (LPP) [7] which can build explicit mapping between the input space and the reduced space. Some other methods using similar idea also have been developed [8,9]. However, LPP and those related methods only attempt to preserve local neighborhood relationship without considering the valuable class label information. Consequently, similar to PCA, the features they produce may not be reliable enough in terms of pattern classification due to the fact that the embeddings of inter-class neighbors may congregate in the reduced space. To preserve local geometric structure of the data and simultaneously exploit class information, a lot of discriminant manifold learning approaches have been proposed [10–15], most of which can be characterized by the graph embedding framework [10]. Among these methods, a representative one is coined as local fisher discriminant analysis (LFDA) [13] which combines the ideas derived from LDA and LPP. Through taking into consideration the local information of data which is often characterized by K nearest neighbor relations, LFDA replaces traditional within-class and between-class scatter matrices used in LDA with their weighted counterparts. Consequently, LFDA is capable of revealing the within-class multimodal structures, thus becoming a useful tool for extracting features from multimodal data.

Another promising approach exploiting structure information is to take into consideration the underlying cluster structure lurking in data [16]. The most representative method is the recently developed subclass discriminant analysis (SDA) [17]. By relaxing the assumption that the data inside each class form a single compact cluster, SDA regards each class can include multiple clusters, where each one can be fitted or approximated by a single Gaussian distribution. Under such an assumption, the clustering technique is first conducted to partition each class into several subclasses while the resulting cluster structure information is then incorporated into the formulation of LDA through maximizing the average distance between the clusters from different classes and at the same time minimizing the scatter within each cluster. It leads to a similar formulation as LDA and the solution can be calculated by solving a generalized eigenvalue problem. Owing to its advantages, SDA has attracted much attention and several extensions have been developed, such as subclass support vector machine [18], subclass-based nonnegative matrix factorization [19] and so forth. It should be pointed out that integrating the cluster structure information into traditional pattern classification algorithms, e.g. SVM, have drawn much attention and some extended algorithms have been developed [20–22]. Generally, most of the feature extraction methods, including PCA, LDA, LFDA, SDA, etc., calculate the discriminant vectors by solving their respective associated generalized eigenvalue problems.

Some recent works [23,24] have demonstrated that the use of margin characterizing the separability between different classes can effectively improve the performance of feature extraction methods. In contrast to conventional between-class scatter representing the average distance between different classes, margin often takes on more discriminant information since it concentrates on the minimum distance between different classes [25,26]. To this end, it is expected that maximizing the margin rather than the between-class scatter can produce the projection direction with better discriminability. In addition, the solution of these methods generally follows from solving a related mathematical programming problem instead of traditional generalized eigenvalue problem. It is worth noting that the margin-based feature extraction algorithms mentioned above are all developed under the framework of LDA, thus ignoring the inherent structure information contained in each class. Therefore, developing feature extraction algorithms which can jointly explore the merits of structure information lurking in data and maximum margin principle is an important problem worthy of study.

Inspired from the above discussion concerning the idea of maximum margin, we propose in this paper a novel feature extraction method, referred to as structural max-margin discriminant analysis (SMDA) to improve SDA [25–28,40,41]. The central idea of SMDA is to seek an orthogonal subspace that maximizes the margin defined as the minimum pairwise distance between clusters (or subclasses) from different classes, while minimizing the total within-cluster scatter. In such a way, the features extracted by SMDA are expected to yield robust performance. It is interesting of our method from the following perspectives:

(1) Different from original SDA [17] which employs the sum of between-cluster distances to characterize the between-cluster separability, our formulation originates from maximum margin principle to guarantee heterogeneous clusters can be well-separated in the reduced subspace.
(2) Different from the methods [23,24] incorporating maximum margin principle, our method takes full account of the intrinsic structure information lurking in each class. Using such kind of information as much as possible is proved to be an effective approach for improving generalization performance [20–22].
(3) An efficient algorithm is developed to solve the nonconvex optimization problem involved in SMDA. On the basis of constrained concave–convex procedure (CCCP) [35], the original problem is converted into a series of convex quadratic programming (QP) problems, each of which can be solved by resorting to the column generation technique [36–38].
(4) Extensive comparisons are made on both artificial and benchmark datasets. The results verify the advantage of SMDA in comparison with other related algorithms.

The rest of this paper is organized as follows. Section 2 briefly reviews the formulation of SDA. Section 3 first introduces the proposed SMDA on the basis of maximum margin principle and then derives a related algorithm. Section 4 reports the experimental results on many datasets. Finally, Section 5 contains some concluding remarks and future works.

2. Brief review of SDA

Suppose we are given a set of D-dimensional input samples $X = \{x_1, x_2, \ldots, x_L\}$ from K known pattern classes. The L_i samples in class i are denoted by X_i, i.e. $|X_i| = L_i$, $\sum_{i=1}^{K} L_i = L$. The main purpose of discriminant feature extraction is to find a low-dimensional subspace where the samples from different classes can be well separated. Subclass discriminant analysis (SDA) first employs a clustering procedure to derive a subclass division of each class, and then incorporates such structure information into the LDA criterion. More specifically, suppose the samples in class i are partitioned into H_i disjoint clusters by using clustering methods, i.e., $X_i = X_{i1} \cup X_{i2} \cup \cdots \cup X_{iH_i}$, $X_{ij} \cap X_{ik} = \Phi$, $\forall j \neq k$ where X_{ij} denotes the j-th cluster in class i, $i = 1, 2, \ldots, K$, $j = 1, 2, \ldots, H_i$. Let $|X_{ij}|$ be the