Neurocomputing 000 (2018) 1–12

Contents lists available at ScienceDirect

Neurocomputing

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

# Adaptive Neighborhood MinMax Projections

Haifeng Zhao [a], Zheng Wang [b], Feiping Nie [b,*]

[a] Key Lab of Intelligent Computing and Signal Processing of MOE & School of Computer and Technology, Anhui University, HeFei 230039, PR China
[b] School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shanxi, PR China

ARTICLE INFO

*Article history:*
Received 18 July 2017
Revised 19 March 2018
Accepted 11 June 2018
Available online xxx

Communicated by Yiming Ying

*Keywords:*
Supervised dimensionality reduction
Feature extraction
Adaptive neighbors
Optimal subspace

ABSTRACT

Dimensionality reduction as one of most attractive topics in machine learning research area has aroused extensive attentions in recent years. In order to preserve the local structure of data, most of dimensionality reduction methods consider constructing the relationships among each sample and its $k$ nearest neighbors, and they find the neighbors in original space by using Euclidean distance. Since the data in original space contain some noises and redundant features, finding the neighbors in original space is incorrect and may degrade the subsequent performance. Therefore, how to find the optimal $k$ nearest neighbors for each sample is the key point to improve the robustness of model. In this paper, we propose a novel dimensionality reduction method, named Adaptive Neighborhood MinMax Projections (ANMMP) which finds the neighbors in the optimal subspace by solving Trace Ratio problem in which the noises and redundant features have been removed already. Meanwhile, the samples within same class are pulled together while the samples between different classes are pushed far away in such learned subspace. Besides, proposed model is a general approach which can be implemented easily and applied on other methods to improve the robustness. Extensive experiments conducted on several synthetic data and real-world data sets and achieve some encouraging performance with comparison to metric learning and feature extraction methods, which demonstrates the efficiency of our method.



## 1. Introduction

In the research areas of machine learning and pattern recognition, dimensionality reduction is a crucial technique whose goal is to embed high-dimensional data sets to a low-dimensional space so that most of discriminative information can be preserved while the useless information are discarded simultaneously. It can make the analysis of discriminative structure for complex data sets more efficient. For the purpose of solving various practical problem such as face recognition, text recognition, biometric recognition etc, many supervised and unsupervised dimensionality reduction methods have been proposed in the past several decades. In this paper, we focus on the supervised dimensionality reduction methods whose label information are taken into consideration in the training. Linear Discriminant Analysis (LDA) [1] is the one of most famous supervised global dimensionality reduction methods, which pulls the samples with same label to the class mean and enlarges the distance of samples in different classes. However, it has several drawbacks in real applications. First, it suffers from the *Small Sample Size* (SSS) problem, in other words, in order to calculate the inverse of within-class scatter matrix, the number of input data samples should larger than the number of features, which limits its popularization in practical applications. Numerous variants of LDA have been presented to solve this problem such as two-stage LDA [2], Orthogonal LDA (OLDA) [3], Null space LDA (NLDA) [4], Direct LDA (DLDA) [5] and so on.

Another drawback is the assumption of Gaussian distribution. It is well-known that the data in real world distribute multimodally, which means the data lie on several different underlying manifolds while LDA is incompetent to tackle them. As a consequence, the idea of $k$ nearest neighbors is naturally incorporated to several locality preserved metric learning methods to handle multimodal data such as Large Margin Nearest Neighbors classifier (LMNN) [6], Local Discriminative Distance Metrics (LDDM) [7], etc. The objective function of LMNN has two terms, pull energy term and push energy term. The former is the sum of distances between a data point $\mathbf{x}_i$ and $k$ homogeneous neighborhoods of $\mathbf{x}_i$ rather than all pairwise points in the same class, and the latter is the sum of distances between each point and all other points which do not have same label. Moreover, it also requires the margins between the data points from different classes at least are one, and some extensions of LMNN [8–11] have been proposed to solve some real applications. However, it is prone to be

* Corresponding author.
*E-mail address:* feipingnie@gmail.com (F. Nie).

https://doi.org/10.1016/j.neucom.2018.06.045
0925-2312/

over-fitting sometimes and sensitive to noise and redundant features when it selects neighbors for each point by using Euclidean distance as the metric. Besides LMNN learns a single metric on all data points which causes that it can not deal with the data contain noisy points [7]. Yang et al. propose Local Discriminative Distance Metrics algorithm (LDDM) which learns a set of local discriminative distance metrics from each training sample and aligns those metrics via ensemble learning to built local classifiers. Although LDDM can handle the data with multimodal distribution, it is time consuming in training.

The aforementioned approaches are metric learning methods based on $k$ nearest neighbors which do not possess the ability of feature extraction. In Neighborhood Component Analysis (NCA) [12], it defines a variable $p_{ij}$ which is a probability of point $\mathbf{x}_i$ is the neighborhood of $\mathbf{x}_j$. The goal of NCA is to maximize the probability that the point $\mathbf{x}_i$ is correctly classified in the projection space, which results in NCA having to solve a non-convex problem for obtaining a local optimal solution while suffering from high computational cost. To overcome this deficiency, Yang et al. [13] propose a Fast Neighborhood Component Analysis (FNCA) which builds a local probability distribution model based on $k$ nearest neighbors within same class and between different classes in subspace. However, although the computational efficiency of FNCA is faster than NCA, the subspace generated by FNCA does not contain any discriminant information, so as to the neighbors selected in such subspace may be not optimal neighbors. Globerson et al. propose Maximally Collapsing Metric Learning algorithm (MCML) [14] which establishes on an ideal case that all points in the same class can be mapped to a single point and different classes points are infinitely far. Then the optimization problem solved in this model is to minimize a KL divergence between the neighbor probability $p_{ij}$ and ideal case, moreover it is a convex problem unlike NCA. Besides, Average Neighborhood Margin Maximization algorithm (ANMM) [15] is another classical local supervised feature extraction method. It defines two types of neighborhoods, homogeneous neighborhoods and heterogenous neighborhoods to calculate the total average neighborhood margin $\gamma$, and the optimization problem of ANMM is to maximize the $\gamma$ so that the data points with same labels can be pulled together and the data points with different labels will be pushed far away in projection space. Additionally, there are some other supervised local feature extraction methods such as Local Fisher Discriminant Analysis (LFDA) [16], Discriminant Neighborhood Embedding (DNE) [17] and some approaches use boosting technique to learn a Mahalanobis distance metric, such as BOOSTMETRIC [18], BOOSTMDM [19], etc.

In this paper, we propose a novel linear supervised dimensionality reduction method, namely Adaptive Neighborhood MinMax Projections (ANMMP), which inspired by Neighborhood MinMax Projections (NMMP) [20]. NMMP focuses on the local pairwise points whose goal is to pull the pairwise points in the same class as close as possible and push the points in different classes as far as possible. Besides, it formulates the task as a constraint optimization problem in which the global optimal solution can be effectively and efficiently obtained. However, all aforementioned approaches find the $k$ nearest neighbors in original space where contains some noises and redundant features, which leads to wrongly selecting the neighbors for each point. The ideal case is that we select $k$ nearest neighbors in the optimal subspace, where the noises and redundant features have been removed already. However, the optimal subspace is absent at first, therefore, the acquirement of optimal subspace is a challenge of proposed method. To address this issue, we propose a novel algorithm which iteratively updates subspace by solving a trace ratio problem and assigns the $k$ nearest neighbors for each point in the optimal subspace adaptively at the end of iterations. Compared with LDA, some local feature extraction and metric learning methods based on the idea of $k$ nearest neighbors, our method has better performance and much more efficient on synthetic data and on several real world benchmark data sets. Compared to FNCA, the subspace generated by our method equips with much more discriminative information so that the within class samples can be pulled together while those in different classes are pushed far away.

The rest of the paper is organized as follows. Section 2 briefly reviews Neighborhood MinMax Projections algorithm and the Iterative procedure to solve Trace Ratio problem (ITR). In Section 3, we present our method, Adaptive Neighborhood MinMax Projections algorithm. A detailed contrast experiments are proposed in Section 4. Finally, Section 5 concludes this paper.

## 2. Related work

### 2.1. A brief review of Neighborhood MinMax Projections

In this section, we briefly review the Neighborhood MinMax Projections algorithm (NMMP). We use bold upper-case and lower-case fonts to represent matrices and vectors respectively, and regular fonts to represent scalars.

Given a data set that consists of $n$ samples $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{m\times n}$ where $m$ denotes the feature number of each sample and $n$ denotes the number of samples. The goal is to learn a linear transformation matrix $\mathbf{W} \in \mathbb{R}^{m\times d}, \mathbf{W}^T\mathbf{W} = \mathbf{I}$, $\mathbf{I}$ is $d\times d$ identity matrix, which projects high-dimensional data $\mathbf{x} \in \mathbb{R}^m$ into a low-dimensional data $\mathbf{y} \in \mathbb{R}^d$:

$$\mathbf{y} = \mathbf{W}^T\mathbf{x}. \tag{1}$$

In the beginning of NMMP, two kinds of neighborhood for each data are defined as: within-class neighborhood $\mathcal{N}_w(i)$ and between-class neighborhood $\mathcal{N}_b(i)$, in which $\mathcal{N}_w(i)$ is the set of the data's $k_w(i)$ nearest neighbors in the $i$th class and $\mathcal{N}_b(i)$ is the set of the data's $k_b(i)$ nearest neighbors in the class other than $i$. Therefore, the sum of Euclidean distances of the pairwise points within same class after the transformation can be obtained by Eq. (2):

$$s_w = Tr(\mathbf{W}^T\widetilde{\mathbf{S}}_w\mathbf{W}) \tag{2}$$

where $Tr(\cdot)$ is the trace of matrix and

$$\widetilde{\mathbf{S}}_w = \sum_{i,j:\mathbf{x}_i\in\mathcal{N}_w(C_j)\&\mathbf{x}_j\in\mathcal{N}_w(C_i)} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \tag{3}$$

is a positive semidefinite matrix in which $C_j$ is the class label of $\mathbf{x}_j$, $C_i$ is the class label of $\mathbf{x}_i$ and $C_i = C_j$. Analogously, the sum of the Euclidean distances of the considered pairwise points between different classes is:

$$s_b = Tr(\mathbf{W}^T\widetilde{\mathbf{S}}_b\mathbf{W}) \tag{4}$$

where $\widetilde{\mathbf{S}}_b$ is a positive semidefinite matrix defined as:

$$\widetilde{\mathbf{S}}_b = \sum_{i,j:\mathbf{x}_i\in\mathcal{N}_b(C_j)\&\mathbf{x}_j\in\mathcal{N}_b(C_i)} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \tag{5}$$

and $C_i \neq C_j$. In order to achieve the above goal, the $s_w$ should be minimized while the $s_b$ should be maximized. Therefore, the problem is formulated as the following Trace Ratio problem:

$$\mathbf{W}^\star = \arg\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{Tr(\mathbf{W}^T\widetilde{\mathbf{S}}_b\mathbf{W})}{Tr(\mathbf{W}^T\widetilde{\mathbf{S}}_w\mathbf{W})} \tag{6}$$

which is difficult to solve. The work in NMMP transforms the Trace Ratio problem into a constrain optimization problem, in which the global optimum can be obtained [20]. In our paper, a more efficient iterative optimization algorithm is used to solve the Trace Ratio problem which will be introduced in what follows.