Neurocomputing 000 (2018) 1–12

Contents lists available at ScienceDirect

Neurocomputing

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

A new two-layer mixture of factor analyzers with joint factor loading model for the classification of small dataset problems

Xi Yang[a], Kaizhu Huang[a,*], Rui Zhang[b], John Y. Goulermas[c], Amir Hussain[d]

[a] Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, SIP, Suzhou 215123, PR China
[b] Department of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, SIP, Suzhou 215123, PR China
[c] Department of Computer Science, Ashton Building, University of Liverpool, Liverpool L69 3BX, UK
[d] Division of Computing Science and Maths, School of Natural Sciences, University of Stirling, Stirling FK9 4LA, UK

ARTICLE INFO

Article history:
Received 26 October 2017
Revised 9 March 2018
Accepted 25 May 2018
Available online xxx

Communicated by Dr Zhu Jianke

Keywords:
Factor analyzer
Joint learning
Classification
Dimensionality reduction

ABSTRACT

Dimensionality Reduction (DR) is a fundamental topic of pattern classification and machine learning. For classification tasks, DR is typically employed as a pre-processing step, succeeded by an independent classifier training stage. However, such independent operation of the two stages often limits the final classification performance notably, as the generated subspace may not be maximally beneficial or appropriate to the learning task at hand. This problem is further accentuated for high-dimensional data classification in situations of the limited number of samples. To address this problem, we develop a novel joint learning model for classification, referred to as two-layer mixture of factor analyzers with joint factor loading (2L-MJFA). Specifically, the model adopts a special two-layer mixture or a mixture of mixtures structure, where each component represents each specific class as a mixture of factor analyzers (MFA). Importantly, all the involved factor analyzers are intentionally designed so that they share the same loading matrix. This, apart from operating as the DR matrix, largely reduces the parameters and makes the proposed algorithm very suitable to small dataset situations. Additionally, we propose a modified expectation maximization algorithm to train the proposed model. A series of simulation experiments demonstrate that what we propose significantly outperforms other state-of-the-art algorithms on various benchmark datasets. Finally, since factor analyzers are closely linked with Auto-encoder networks, the proposed idea could be of particular utility to the community of neural networks.

1. Introduction

Dimensionality reduction (DR) is a very important topic of pattern recognition and machine learning that has been studied intensely in the relevant literature. Its objective is the finding of a subspace to effectively reduce the computational time while improving the performance of the learning task [1,2]. Traditionally, DR is performed as a pre-processing step to remove noise and compact the representation. Subsequently, the reduced features can be fed to various models for accurately learning a classification task. A typical example of this workflow, includes a Gaussian mixture model (GMM) classifier applied after a linear DR method, such as principal component analysis (PCA), linear discriminant analysis (LDA), factor analyzer (FA) [3–5], or a method from the recently proposed [6–9]. Besides linear methods, there are other DR techniques that achieve nonlinear projections of the data [10–12].

While the independent realization of DR and classification can be easily implemented, it may notably diminish the final performance [2,13] as the two tasks do not necessarily interact with each other, and the optimal subspace obtained by the DR may not be maximally beneficial to the learning task. This is particularly the case for the small sample size (S3) problem [14,15], where the data patterns are high-dimensional but of low cardinality. In such problems, the subspace derived by the independent DR may even significantly deteriorate the classification performance.

Motivated from the above issues, we propose within an FA framework, a novel model referred to as the two-layer mixture of factor analyzers with joint factor loading (2L-MJFA). This relies upon a mixture of mixtures structure, used to better capture the complex properties of each class and realize efficiently the joint learning requirements. An important characteristic of 2L-MJFA is that all of its involved latent factors are designed to share the same loading matrix. This has a dual purpose, in the sense that, on the one hand, it operates as the driving DR structure, and on the other

* Corresponding author.
E-mail addresses: xi.yang@xjtlu.edu.cn (X. Yang), kaizhu.huang@xjtlu.edu.cn (K. Huang), rui.zhang02@xjtlu.edu.cn (R. Zhang), j.y.goulermas@liverpool.ac.uk (J.Y. Goulermas), ahu@cs.stir.ac.uk (A. Hussain).

https://doi.org/10.1016/j.neucom.2018.05.085
0925-2312/

hand it significantly reduces the number of parameters. The latter accelerates training while mitigates the negative effect caused by the limited number of per class samples.

Contrary to the independent approaches, the proposed 2L-MJFA is capable of simultaneously learning the DR matrix as well as the optimal parameters of the classification model. This model is implemented via a GMM for simplicity, but it is straightforward to extend the two-layer mixture approach to the use of other models. Through joint learning, the method achieves efficient DR that not only reduces the computational time for high dimensional data, but more importantly it significantly benefits the final classification stage. Another contribution, is that we also propose a modified expectation-maximization (EM) algorithm that consists of two-layer loops, so that the joint learning is conducted very efficiently. The first layer loop is used to estimate the joint parameters that fit the mixture among different classes, whereas the second one trains the mixture components within each class. The 2L-MJFA is theoretically distinct to other joint learning FA models, such as the FA mixture with common loading (MCFA) [16], the mixture of MCFAs (mMCFA), and the mixture of probabilistic PCA (mPPCA) [17,18]. Further details about these models are presented in the following section. Our experiments show that the proposed method significantly outperforms these existing methods in seven benchmark datasets.

The rest of this paper is organized as follows. Section 2 briefly reviews related work and emphasizes the differences between our proposed approach and existing ones. The baseline model mixture of FAs (MFA) and the MCFA are introduced as preliminaries in Section 3. In Section 4 we introduce the proposed 2L-MJFA model, while Section 5 explains how the model parameters can be estimated by the modified EM algorithm. In Section 6 we present the experimental setup and the classification results with the aid of seven datasets including a synthetic dataset and six real ones. Finally, Section 7 concludes the work. The work presented here is an extension of [19], and is based on redesigning and supplementing the experiments to support evaluations for S3 data cases, and further compare with existing methods with respect to their technical details.

2. Related work

There have been several joint learning FA based approaches [20,21] related to our proposed method. To illustrate the distinction, we present the different alternative structures incorporated in various models in Fig. 1. In particular, the model MFA [3] is the base model for what we propose. It combines DR with clustering and utilizes a subspace metric to guide cluster separation. This work is extended by MCFA [16] which assumes the factor loading of the MFA to be a common matrix that can largely reduce the involved parameters. When MCFA is used for classification, one straightforward way is to regard each class as one component, as shown in Fig. 1(a). Obviously, such a setting is quite basic and not adequately flexible, since data classes may have complex distributions and modalities. Another popular variant that extends MCFA is mMCFA, shown in Fig. 1(b), where the factor loadings $\mathbf{A}_i$ are different for each class. In general, different loading matrices imply independent DR for different classes, and this may be physically impractical. More importantly, mMCFA could be problematic in S3 problems, as the limited number of samples cannot support the accurate learning of the loading matrices. To this end, a non-trivial model is proposed here by sharing one loading matrix for all the classes. The mPPCA method [17,18] extends PCA to a mixture distribution model. As seen in Fig. 1(d), its graphical model is quite similar to MFA with the elements of the common covariance matrix $\mathbf{D} = \sigma^2\mathbf{I}_p$ assumed to be isotropic [22], where $\mathbf{I}_p$ is the p-dimensional identity matrix. For classification, each class is modeled as an mPPCA model. This method is limited due to its poor flexibility and has many redundant parameters for dealing with S3 problems.

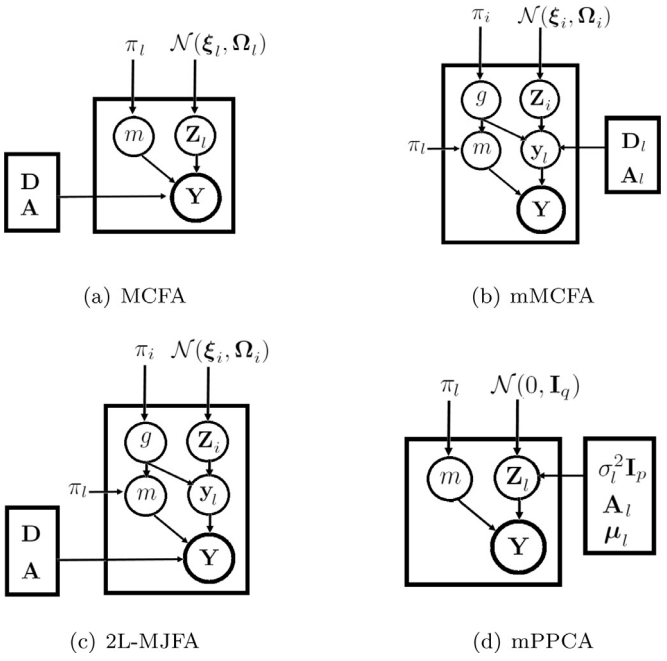

Fig. 1. Comparison of different models. $\mathbf{Y}$ denotes observed data, $\mathbf{Z}_l/\mathbf{Z}_i$ denotes latent factors, and $\mathbf{A}/\mathbf{A}_l$ is (common) factor loadings, where m is the class label and g is the number of mixtures. (a) MCFA which is the fundamental MFA model with a common $\mathbf{A}$. (b) Mixture of MCFAs with each class consisting of a components mixture with individual local factor loadings $\mathbf{A}_l$. (c) The proposed 2L-MJFA with a global factor loading $\mathbf{A}$ shared between and within classes in the 2-layer mixture model. (d) Mixture of probabilistic PCA which is similar to MFA but with isotropic common covariance matrix.

Table 1
Summary of the number of parameters for the main models. The rightmost column shows the simplified number of parameters which makes the total numbers easy to compare.

Model:	Number of parameters:	Approximation:
mPPCA	$m[g+gp+gpq-\frac{gq(q-1)}{2}]$	$(mg+mq)p$
mMCFA	$m[pq-q^2+p+g(1+q+\frac{q(q+1)}{2})]$	$(m+mq)p$
2L-MJFA	$pq-q^2+p+m[g+gq+\frac{gq(q+1)}{2}]$	$(q+1)p$

We now analyse the parameter numbers in the different models, assuming p dimensions, q reduced dimensions from p, and m classes. Setting g mixture components in each class, the covariance matrix of each component has $N = \frac{p(p+1)}{2}$ parameters. Since mPPCA converts the diagonal covariance matrix into an isotropic one as $\boldsymbol{\Sigma}_i = \mathbf{W}_i\mathbf{W}_i^T + \sigma^2\mathbf{I}_p$, where factor loading $\mathbf{W}_i \in \mathbb{R}^{p\times q}$ contains $\frac{q(q-1)}{2}$ constraints, its total number of parameters is

$$N_1 = m\left(g + gp + gpq - \frac{gq(q-1)}{2}\right).$$

If either p or q is large, the number of parameters may not even be manageable with a diagonal covariance. To further reduce the parameters and accelerate training, the component covariance matrices of mMCFA has a factor-analytic representation $\boldsymbol{\Sigma}_i = \mathbf{A}\boldsymbol{\Omega}_i\mathbf{A}^T + \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix, and $\mathbf{A}$ contains the factor loading for all the components [23]. From the orthogonality requirement, $\mathbf{A}$ has $pq - q^2$ constraints. Hence, in mMCFA the total number of parameters is reduced to

$$N_2 = m\left[pq - q^2 + p + g\left(1 + q + \frac{q(q+1)}{2}\right)\right].$$

Table 1 lists the associated parameter numbers for FA models. Since $p >> q$, the order of the number of parameters can be ap-