Information Sciences 330 (2016) 74–87

Contents lists available at ScienceDirect

Information Sciences

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

Dimensionality reduction for data of unknown cluster structure

Ewa Nowakowska [a,*], Jacek Koronacki [a], Stan Lipovetsky [b]

[a] Institute of Comuter Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, Warszawa 01-248, Poland
[b] GfK Custom Research North America, Marketing & Data Sciences, 8401 Golden Valley Rd., Minneapolis, MN 55427, USA

ARTICLE INFO

Article history:
Received 31 July 2014
Revised 17 August 2015
Accepted 6 October 2015
Available online 16 October 2015

MSC:
62H25
62H30

Keywords:
Dimensionality reduction
Gaussian mixture model
Fisher's subspace
Principal component analysis

ABSTRACT

Dimensionality reduction that preserves certain characteristics of data is needed for numerous reasons. In this work we focus on data coming from a mixture of Gaussian distributions and we propose a method that preserves the distinctness of the clustering structure, although this structure is assumed to be yet unknown. The rationale behind the method is the following: (i) had one known the clusters (classes) within the data, one could facilitate further analysis and reduce space dimensionality by projecting the data to the Fisher's linear subspace, which — by definition — best preserves the structure of the given classes; (ii) under some reasonable assumptions, this can be done, albeit approximately, without prior knowledge of the clusters (classes). In this paper, we show how this approach works. We present a method of preliminary data transformation that brings the directions of largest overall variability close to the directions of the best between-class separation. Hence, for the transformed data, simple PCA provides an approximation to the Fisher's subspace. We show that the transformation preserves the distinctness of the unknown structure in the data to a great extent.

1. Introduction

1.1. State-of-the-art

Dimensionality reduction techniques, also referred to as feature extraction algorithms, are a common way of reducing intrinsic complexity of data and thus facilitates further analysis. It is typically expected that certain characteristics of data will be preserved in the process. In particular, for data exhibiting a clustering structure, the structure is expected to be preserved to the largest possible extent. Frequently, it is captured in terms of distances between observations as in [8], where one of the first methods for suitable linear feature extraction is described. Another line of research starts with [32], where a transformation for continuous data that lowers the dimensionality without increasing the probabilities of misclassification is proposed. The approach is further developed in [10,37,44]. Among more recent works, a method of dimensionality reduction that preserves clustering structure is proposed in [11], where, however, the assumption of known cluster assignments has been made. Finally, an interesting overview of methods suitable for a pattern recognition task is provided in [40], while in [4] challenges of feature selection in the context of big data are presented. In [14,20,39] recent advances in the area of feature selection are described, detailing approaches via stratification, grouping and information theory, respectively. In [5] the focus is directly on selecting attributes of largest power of class discrimination.

[*] Corresponding author. Tel.: +48 223800500.
E-mail addresses: ewa.nowakowska@ipipan.waw.pl (E. Nowakowska), jacek.koronacki@ipipan.waw.pl (J. Koronacki), stan.lipovetsky@gfk.com (S. Lipovetsky).

http://dx.doi.org/10.1016/j.ins.2015.10.009
0020-0255/© 2015 Elsevier Inc. All rights reserved.

An approach to dimensionality reduction which directly aims at preserving the distinctness of the structure has been originated in a series of works on learning mixture parameters in an appropriate subspace. In [22], one-dimensional random projections were considered and generalized to arbitrary number of clusters in [29]. Based on Johnson–Lindenstrauss (concentration) theorem, random projections to substantially lower — but in general — more than one-dimensional subspace were suggested in [9]. In [3], distributional assumptions were relaxed, however the main assumption of high initial cluster separation, intrinsic for concentration theorem, was retained. Only in [6], were random projections replaced with a spectral approach, making substantial progress in relaxing the requirement of initial cluster separation. It was first applied in [38] and the results were improved in [1,23]. A breakthrough was achieved in [7]. The authors presented an affine invariant parameter learning algorithm where a preliminary data transformation was used to enhance the distinctness of the clustering structure, thereby further relaxing the separability assumptions. From our perspective, it suggested that it is possible to sharpen the clustering structure without actually knowing it. This significant discovery has become the major inspiration for the method proposed in the following sections.

The novelty of the proposed approach is reflected in its ability to reduce dimensionality and preserve cluster structure to the largest possible extent without explicitly knowing it. As described above, methods that aim at preserving structure typically assume knowledge of the classes. On the other hand, those that do not assume such knowledge, do not evaluate (let alone maximize) the clustering structure.

1.2. Model and notation

We consider a data set $X = (x_1, \dots, x_n)^T,\ X \in \mathbb{R}^{n\times d}$ of n observations coming from a mixture of k d-dimensional normal distributions

$$f(x) = \pi_1 f_1(\mu_1, \mathbf{\Sigma}_1)(x) + \cdots + \pi_k f_k(\mu_k, \mathbf{\Sigma}_k)(x),$$

where

$$f_l(\mu_l, \mathbf{\Sigma}_l)(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det \mathbf{\Sigma}_l}} e^{-\frac{1}{2}(x-\mu_l)^T \mathbf{\Sigma}_l^{-1}(x-\mu_l)}.$$

We call each $f_l(\mu_l, \mathbf{\Sigma}_l)$, $l = 1, \dots, k$ a component of the mixture and each π_l, $l = 1, \dots, k$ a mixing factor of the corresponding component (see [17] or [28] and [26] or [27] for alternatives). We assume that for all the components equal mixing factors are assigned, $\pi_1 = \cdots = \pi_k = \frac{1}{k}$. However, we allow different covariance matrices $\mathbf{\Sigma}_l$. Additionally, we assume large space dimensionality with respect to the number of components, $d > k - 1$, to leave room for dimensionality reduction. We also assume large number of observations with respect to d, that is $n \gg d$. We take the number of components k as known. This puts no constraints on our considerations as the procedure may easily be repeated for all k within the range of interest. The parameters of the mixture are given by $\mu = (1/k)\sum_{l=1}^{k} \mu_l,\ \mu \in \mathbb{R}^d$ and $\mathbf{\Sigma} = (1/k)\sum_{l=1}^{k} \mathbf{\Sigma}_l + (1/k)\sum_{l=1}^{k} (\mu_l - \mu)(\mu_l - \mu)^T,\ \mathbf{\Sigma} \in \mathbb{R}^{d\times d}$. The latter constitutes the covariance decomposition to the within and between cluster component (see [28]).

We assume that each mixture component corresponds to one cluster. A grouping that divides observations into clusters is called a clustering solution or a clustering structure. Note that heterogeneity of covariance matrices allows for varied cluster shapes, while equal mixing factors imply balanced cluster sizes.

Let $\mu_X \in \mathbb{R}^d$ and $\mathbf{\Sigma}_X \in \mathbb{R}^{d\times d}$ refer to the empirical estimates of the mixture parameters. We assume the covariance matrix to be of full rank, $\mathrm{rank}(\mathbf{\Sigma}_X) = d$. Let $T_X = n\mathbf{\Sigma}_X$ be the *total scatter matrix* for X. We say that data is in *isotropic position* if $\mu_X = \mathbf{0}$ and $T_X = \mathbf{I}$.

For symmetric $C \in \mathbb{R}^{d\times d}$ let $C = A_C L_C A_C^T$ be the *spectral decomposition (eigenproblem solution)* for matrix C, where $L_C = \mathrm{diag}(\lambda_1^C, \dots, \lambda_d^C)$, $\lambda_1^C \geq \cdots \geq \lambda_d^C$) is a matrix of eigenvalues for C in a non-decreasing order and $A_C = (a_1^C, \dots, a_d^C)$ is a matrix of the corresponding column eigenvectors. Alternatively, when considering the eigenproblem for different data sets, we will use the data set as a subscript or superscript (e.g. $C_X = A_X L_X A_X^T$). By $PC(k-1)$ we denote the *principal component subspace* spanned by the first $k - 1$ principal components (i.e. $k - 1$ eigenvectors of the matrix $\mathbf{\Sigma}_X$ corresponding to its $k - 1$ largest eigenvalues, see more in [17],[25] or [28]and references therein for possible extensions).

By S^* we denote the *Fisher's discriminant (Fisher's subspace)*, which is a $(k-1)$-dimensional subspace that best discriminates k given classes as

$$S^* = \underset{\substack{S \subset \mathbb{R}^d \\ \dim(S) = k-1}}{\mathrm{argmax}} \frac{\sum_{j=1}^{k-1} v_j^T B_X v_j}{\sum_{j=1}^{k-1} v_j^T T_X v_j},$$

where $B_X = \sum_{l=1}^{k} n_l(\mu_{X,l} - \mu_X)(\mu_{X,l} - \mu_X)^T$ is the between cluster component of the total scatter matrix for X with $\mu_{X,l}$ denoting the empirical mean of lth cluster, $l = 1, \dots, k$, and $v_1, \dots, v_{k-1}$ is the orthonormal basis for S. Details of this specific definition are given in [13], while the general concept is discussed in [28].

It is well known that S^* is the subspace spanned by $k - 1$ eigenvectors corresponding to the non-zero eigenvalues of a generalized eigenproblem defined by B_X and T_X matrices,

$$B_X v = \lambda T_X v, \tag{1}$$