Applied Mathematics and Computation 256 (2015) 591–601

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Clusterability assessment for Gaussian mixture models

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

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

ARTICLE INFO

Keywords:
Clusterability
Gaussian mixture models
Fisher's discriminant
Principal component analysis

ABSTRACT

There are numerous measures designed to evaluate quality of a given data grouping in terms of its distinctness and between-cluster separation. However, there seems to be no efficient method to assess distinctness of the intrinsic structure within data (clusterability) before actual clustering is determined. Based on recent findings, we propose such measure in terms of covariance matrix decomposition for appropriately transformed data. The data is assumed to come from a Gaussian mixture model. The transformation reshapes the data so that unsupervised technique of principal component analysis is able to uncover information directly indicative of the data clusterability characteristics. In this work we propose the measure and explain the motivation as well as the relation to supervised structure distinctness coefficients. We also show how the measure can be applied for number of clusters and feature selection tasks.

1. Introduction

1.1. State-of-the-art

Literature on clusterability and related distinctness assessment problems suffers from inconsistent and ambiguous terminology. What can be currently found under the term of clusterability is gathered in [1]. The indices analyzed there though, are meant to assess a given data partition, while the term clusterability should rather be confined to measures that are partition-independent. It should refer to assessing the possibility to cluster efficiently rather than to quality evaluation of a previously determined solution. In what follows, we will follow this distinction and refer to measures based on certain partition as structure distinctness coefficients. In other words, for structure distinctness coefficients we will require the model parameters to be known or estimated. The term clusterability we will use exclusively to refer to measures that do not need this kind of information and therefore can operate on raw data, without the knowledge of partition in particular. Based on some of the indices of [1], partition-independent measures can be obtained, as it was done for instance in [2]. Somewhat unfortunately, this heuristic method tends to fail in actual applications when the number of dimensions increases. More formal inspiration for clusterability analysis is provided by considerations on the number of clusters (main references include [3–6]) and component overlap analysis for mixtures of normal distributions (see [7–10]). However, both approaches originally assume the underlying partition to be already determined.

Direct inspiration for this work comes from a series of works on learning mixture parameters in a selected subspace. It started with the challenge of random projections considered first in one-dimensional space – by Kalai et al. [11] for two and

[*] Corresponding author.
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.amc.2014.12.038

Moitra and Valiant [12] for arbitrary number of clusters. Then [13] suggested random projections to substantially lower potentially more than one-dimensional subspace based on Johnson–Lindenstrauss (concentration) theorem. In [14] some of the distributional assumptions were relaxed but as the concentration theorem was still used, the assumption of high initial cluster separation had to be maintained. This was only relaxed by Brand and Huang [15], where random projections were replaced with spectral approach. These results were further applied and developed in [16–18]. The key insight for this work comes from [19], where a preliminary data transformation was used to enhance the unknown structure in data in order to improve performance of a parameter learning algorithm. This proved that it is possible to use the structure in data without actually knowing it and inspired [20], which in turn became the basis for the clusterability assessment method proposed in this work.

1.2. Model and notation

We assume that the data $X = (x_1, \ldots, x_n)^T,\ X \in \mathbb{R}^{n\times d}$ – consisting of n observations – comes 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 refer to each $f_l(\mu_l, \mathbf{\Sigma}_l),\ l = 1, \ldots, k$ as a component of the mixture (see [21] or [22] for model details and [23] or [24] for example alternatives). We call $\pi_l,\ l = 1, \ldots, k$ a mixing factor of the corresponding component. We assume equal mixing factors for all the components $\pi_1 = \cdots = \pi_k = \frac{1}{k}$, however we allow different covariance matrices $\mathbf{\Sigma}_l$. Additionally we assume the space dimension to be large with respect to the number of components $d > k - 1$. We also assume the number of observations is large with respect to d or k, that is $n \gg d$. We take the number of components k as known, which puts no constrains on our considerations they can easily be repeated for all k within the range of interest.

We assume the cluster centers to be independent in terms of point independence. Also, we assume covariance matrix to be of full 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 $T_X = \mathbf{I}$. We recall the known fact that $T_X = W_X + B_X$, which decomposes the total scatter to its within (W_X) and between (B_X) cluster components. For notation ease we will often assume the data is centered and the origin and indicate that with a zero subscript X_0 or Z_0.

A cluster – or a class – is understood as a subset of observations that are similar (close in the space) to one another but different (far in the space) from other observations. A grouping that divides observations into clusters is called a *cluster solution* or a *cluster structure*. To set up a link between the theoretical model and the data, we assume that each mixture component corresponds to one cluster.

By $PC(k-1)$ we denote the 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). 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

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

where $v_1, \ldots, v_{k-1}$ is the orthonormal basis for S. Details are given in [25], while the concept is discussed in [21]. In the course of the work we will use the known result that S^* may be represented as a solution of an eigenproblem defined by $T_X^{-1}B_X$. The proof under the assumed model can be found in [26].

Following [20], we define *structure distinctness coefficient* in terms of Fisher's discriminant as

$$\bar{\lambda}^X = \frac{1}{k-1}\sum_{j=1}^{k-1} \lambda_j^{T_X^{-1}B_X}, \tag{2}$$

which is the average eigenvalue over $k-1$ largest eigenvalues of the $T_X^{-1}B_X$ eigenproblem and the mean variability in the Fisher's subspace at the same time. As [26] shows, (2) well reflects the behavior of the generic integral overlap measure (also referred to as MLE-misclassification rate) while possessing the advantage of being analytically tractable at the same time. It allows it to be included in formal derivations and the analysis of the system behavior.

1.3. Concept and related results

The idea of clusterability assessment proposed here is based on the results obtained and presented in [20]. In this work a data transformation is derived that preserves structure distinctness – as defined by (2) – up to a negligible error. The transformation consists of two steps – isotropic transformation and weighting. The first step de-correlates the data – it can easily be shown (see for instance [20]) that the following transformation brings the data to isotropic position so $\mu_Y = \mathbf{0}$ and $T_Y = \mathbf{I}$