Journal of Statistical Planning and Inference 140 (2010) 1701–1711

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A comparative study of the *K*-means algorithm and the normal mixture model for clustering: Bivariate homoscedastic case

Dingxi Qiu

Department of Industrial Engineering, University of Miami, Coral Gables, FL 33146, USA

ARTICLE INFO

Article history:
Received 23 May 2009
Accepted 15 December 2009
Available online 4 January 2010

Keywords:
Clustering
Data mining
Mixture model
K-means algorithm
EM algorithm
Elongation
Mixing proportion
Misclassification rate

ABSTRACT

The *K*-means algorithm and the normal mixture model method are two common clustering methods. The *K*-means algorithm is a popular heuristic approach which gives reasonable clustering results if the component clusters are ball-shaped. Currently, there are no analytical results for this algorithm if the component distributions deviate from the ball-shape. This paper analytically studies how the *K*-means algorithm changes its classification rule as the normal component distributions become more elongated under the homoscedastic assumption and compares this rule with that of the Bayes rule from the mixture model method. We show that the classification rules of both methods are linear, but the slopes of the two classification lines change in the opposite direction as the component distributions become more elongated. The classification performance of the *K*-means algorithm is then compared to that of the mixture model method via simulation. The comparison, which is limited to two clusters, shows that the *K*-means algorithm provides poor classification performances consistently as the component distributions become more elongated while the mixture model method can potentially, but not necessarily, take advantage of this change and provide a much better classification performance.

1. Introduction

Cluster analysis is a common unsupervised learning technique for statistical data analysis, which seeks to group objects of a similar kind into separate categories. It is widely used in different fields, including social sciences, database marketing and bioinformatics. Cluster analysis encompasses a number of heuristic and model-based methods, including the *K*-means algorithm (MacQueen, 1967) and the normal mixture model (MM) method (Day, 1969; Titterington et al., 1985).

A general question facing researchers and practitioners is which method to use in practice. The *K*-means algorithm is a nonparametric approach that aims to classify objects into *K* mutually exclusive clusters by minimizing the expected squared distance of an object from its nearest center. It is generally known to be a fast algorithm, but the main limitation is its convergence reliability. Everitt (1993, p. 98) showed by an example that the *K*-means algorithm can have extremely poor performance on a two-component mixture model if the variables are highly correlated, and this phenomenon has been quoted in many places as a practical warning. For example, Cheung (2003) mentions that the *K*-means algorithm assumes that the data clusters are ball-shaped, and works poorly for elliptical clusters. Various efforts have been made to address the elongation problem. Art et al. (1982) proposed an iterative algorithm for estimating the shapes of the component distributions and suggested the Mahalanobis distance as the appropriate measure. Other distance metrics, such

E-mail address: dingxi@miami.edu

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.12.025

as the $L1$ distance and the sample correlation between objects, can also be used for the K-means algorithm. These variants of K-means algorithm have been implemented in commercial software such as Matlab and SAS. Considering the fact that the standard K-means algorithm is the most popular clustering tool, and its naive use can, in some cases, lead to nonsensical results, we would like to analytically study its performance as a function of component cluster elongation.

The normal MM method is another approach to cluster analysis, where each cluster is modeled by a component distribution from the Gaussian distribution family. The expectation–maximization (EM) algorithm of Dempster et al. (1977) is often used to estimate the associated parameters. The EM algorithm maximizes the incomplete log-likelihood function through maximizing a sequence of complete log-likelihood functions. When the covariance matrices Σ_k $(1 \le k \le K)$ of different clusters are equal to $\sigma^2 I$, it reduces to the K-means algorithm as $\sigma^2 \to 0$ (Hastie et al., 2002; Steinley, 2006); however, the mixture component means do not model the cluster means from the K-means algorithm in general. The reliability of the MM method on normal mixture data is a function of the sample size and can degenerate drastically as the sample size becomes small due to parameter estimation errors; even though in theory the MM method has a potential to provide the best classification performance.

Qiu and Tamhane (2007) provided a relatively thorough comparison of K-means algorithm and the MM method based on data generated from a two-component univariate normal mixture model. The comparison was based on a rigorous treatment of the asymptotic behavior of the two clustering methods. Simulation results were given to compare the two methods for a range of sample sizes. In this paper, we extend the study to the bivariate homoscedastic case where the component clusters have common covariance matrices. We plan to provide a theoretical justification of why the K-means algorithm performs poorly on elliptical data. Its classification performances are compared to those of the MM method under various conditions.

The outline of the paper is as follows. In Section 2, we formulate the problem and define the notations. We also translate the homoscedastic normal mixture model into a mixture model of clusters with independent variables. In Section 3, we review the K-means algorithm and the MM method. In Section 4, we examine the behavior of the classification rules as a function of the elongation measure of the component clusters. The examination was conducted under asymptotic assumption where the sample size $n \to \infty$. In Section 5, a simulation study with finite sample sizes are discussed to verify the analytical results for both methods, which is followed by a discussion and future research directions in Section 6.

2. Homoscedastic bivariate normal mixture model

Let us begin with a graphic representation of the homoscedastic bivariate normal mixture model, consisting of two component clusters with a common covariance structure. In Fig. 1, the two plotted variables, Y_1 and Y_2, are correlated in both clusters. Suppose $\boldsymbol{Y} = (Y_1, Y_2)'$ follows a bivariate (more generally, a multivariate) normal distribution with covariance matrix $\boldsymbol{\Omega}$ in each component, then there exists a unitary matrix $\mathbf{P}$, whose rows are orthonormal eigenvectors of $\boldsymbol{\Omega}$, such that $\boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\Omega}\mathbf{P}'$ is a diagonal matrix. In other words, even if Y_1 and Y_2 are correlated, they can always be transformed to independent random variables X_1 and X_2 with unequal variances by $\boldsymbol{X} = (X_1, X_2)' = \mathbf{P}\boldsymbol{Y}$. The classification performances of the K-means algorithm and the MM method are invariant to rotations and scaler transformations. Thus, the problem of studying the effect of correlation can be transformed to that of the effect of the ratio σ_2/σ_1, where σ_1^2 and σ_2^2 are the variances of the two independent random variables. We define the ratio $\varsigma = \sigma_2/\sigma_1$ as *elongation* which measures the deviation of the contour line of the bivariate distribution from the ball shape. The following proposition shows that the absolute correlation, $|\rho|$, between the correlated variables in each component cluster increases as the elongation, ς, increases.

Proposition 1. *Define X_1 and X_2 as two independent random variables with common mean 0 and variances σ_1^2 and σ_2^2 $(\sigma_2^2 > \sigma_1^2)$. Let $(Y_1, Y_2)'$ be an orthogonal one-to-one transformation of $(X_1, X_2)'$ given by*

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \cos v & \sin v \\ -\sin v & \cos v \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \tag{2.1}$$

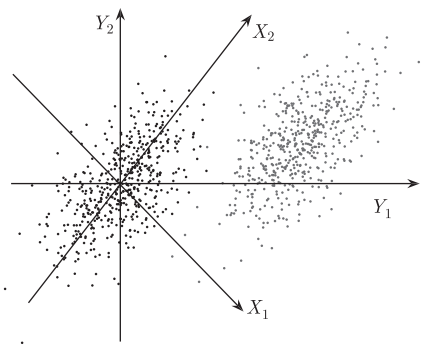

Fig. 1. Simulated data from a mixture of two bivariate normal distributions.