Available online at www.sciencedirect.com

ScienceDirect

Journal of Statistical Planning and Inference 137 (2007) 3722–3740

journal of statistical planning and inference

www.elsevier.com/locate/jspi

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

Dingxi Qiu, Ajit C. Tamhane*

Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL 60208, USA

Available online 8 May 2007

Abstract

This paper gives a comparative study of the *K*-means algorithm and the mixture model (MM) method for clustering normal data. The EM algorithm is used to compute the maximum likelihood estimators (MLEs) of the parameters of the MM model. These parameters include mixing proportions, which may be thought of as the prior probabilities of different clusters; the maximum posterior (Bayes) rule is used for clustering. Hence, asymptotically the MM method approaches the Bayes rule for known parameters, which is optimal in terms of minimizing the expected misclassification rate (EMCR).

The paper gives a thorough analytic comparison of the two methods for the univariate case under both homoscedasticity and heteroscedasticity. Simulation results are given to compare the two methods for a range of sample sizes. The comparison, which is limited to two clusters, shows that the MM method has substantially lower EMCR particularly when the mixing proportions are unbalanced. The two methods have asymptotically the same EMCR under homoscedasticity (resp., heteroscedasticity) when the mixing proportions of the two clusters are equal (resp., unequal), but for small samples the MM method sometimes performs slightly worse because of the errors in estimating unknown parameters.

MSC: 62H30; 62F10

Keywords: Bayes rule; Clustering; Data mining; EM algorithm; *K*-means algorithm; Misclassification rate; Mixture model; Prior and posterior probabilities

1. Introduction

The *K*-means algorithm (MacQueen, 1967) is one of the most popular methods for clustering multivariate quantitative data. This algorithm is non-parametric in nature as it does not assume any probability model for the data. Given a fixed number of clusters, it determines an assignment of the data vectors (observations) to the clusters so as to minimize the total of the squared distances between the observations assigned to the same cluster and summed over all clusters. See Everitt (1993) for a review of clustering methods.

The mixture model (MM) method provides a parametric approach to the clustering problem. The EM algorithm (Dempster et al., 1977) is a natural method for obtaining the maximum likelihood estimators (MLEs) of the unknown parameters of the MM. The parameters include the mixing proportions or the prior probabilities of the clusters since the

* Corresponding author. Tel.: +1 8474913577; fax: +1 8474918005.
E-mail address: ajit@iems.northwestern.edu (A.C. Tamhane).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.03.045

true cluster memberships of the observations are unobserved. Clustering is done by applying the maximum posterior (Bayes) rule.

The *K*-means algorithm makes "hard" (deterministic) assignments of the observations to the clusters, i.e., each observation is assigned to exactly one cluster. On the other hand, the MM method computes posterior probabilities (called *responsibilities*) of belonging to different clusters for individual observations. Hastie et al. (2002, p. 463) note that the MM method is a "soft" version of the *K*-means algorithm in that if the data from each cluster is assumed to be multivariate normal (MVN) with the mean vector depending on the cluster and a common covariance matrix $\sigma^2\boldsymbol{I}$, then as $\sigma^2 \to 0$, the MM method based on the EM algorithm converges to the *K*-means algorithm. Thus, as in the *K*-means algorithm, asymptotically the MM method assigns each observation to that cluster whose estimated mean is closest to the observation.

Although there is asymptotic convergence of the MM method and the *K*-means algorithm, it is under very restrictive conditions of homoscedasticity not only among the clusters, but also among the measured variables. More crucially, it assumes independence among the variables. These assumptions underlie the *K*-means algorithm, which ignores correlations and heteroscedasticity among the variables by using the simple Euclidean distance measure. Therefore, it is of interest to compare the performances of the two methods under the practical conditions of small samples, correlated responses and heteroscedasticity. In this paper we initiate this study by focusing on the univariate case for $K = 2$ under homoscedasticity and heteroscedasticity. The MVN case, which allows the study of how correlations between measured variables affect the performance of the competing algorithms, will be considered in a future paper. Surprisingly, even the univariate case has not been studied in this context to the best of our knowledge.

The outline of the paper is as follows. In Section 2 we formulate the problem and define the notation. In Section 3 we review the *K*-means algorithm and the MM method with the associated EM algorithm. The discussions in both sections are framed in the general setting of MVN data, but in the remainder of the paper we focus exclusively on univariate normal data. In Section 4 we give analytical results for comparing the two methods in the homoscedastic case. In Section 5 we extend these results to the heteroscedastic case. In Section 6 we present simulation results on misclassification rates of the two methods. Finally, Section 7 gives a discussion and conclusions.

2. Problem formulation and notation

Suppose that there are N subjects on each of whom m variables are measured resulting in observations $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})'$ $(1 \leqslant i \leqslant N)$. The goal of clustering is to group these N subjects into $K < N$ clusters, C_k $(1 \leqslant k \leqslant K)$, so that similar subjects are grouped into the same cluster and dissimilar subjects are grouped into different clusters. We will assume that K is the true known number of clusters and is fixed. (In practice, of course, K is not known. The problem of determining the optimal K will not be addressed here.) Let N_k denote the true number of subjects belonging to cluster C_k where $\sum_{k=1}^{K} N_k = N$. A clustering rule (denoted by R) is a many-to-one mapping, $R(\boldsymbol{x}_i) = C_k$ $(1 \leqslant i \leqslant N, 1 \leqslant k \leqslant K)$.

We will assume that the observations $\boldsymbol{x}_i$ are mutually independent and the observations from different clusters have MVN distributions with different mean vectors; the covariance matrices may be equal (homoscedastic) or unequal (heteroscedastic). Specifically, if subject i belongs to cluster C_k (denoted by $i \in C_k$) then

$$\boldsymbol{X}_i | i \in C_k \sim \mathrm{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.1}$$

where $\boldsymbol{X}_i$ denotes the random variable (r.v.) corresponding to the observation $\boldsymbol{x}_i$, and $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote the mean vector and covariance matrix of the MVN distribution for cluster C_k.

3. Review of two clustering methods

3.1. K-means algorithm

The *K*-means algorithm uses the Euclidean squared distance measure:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2 = \sum_{j=1}^{m} (x_{ij} - x_{i'j})^2,$$