Pattern Recognition 45 (2012) 3950–3961

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

A robust EM clustering algorithm for Gaussian mixture models

Miin-Shen Yang*, Chien-Yo Lai, Chih-Ying Lin

Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li 32023, Taiwan

ARTICLE INFO

Article history:
Received 23 November 2011
Received in revised form
18 March 2012
Accepted 24 April 2012
Available online 14 May 2012

Keywords:
Cluster analysis
EM algorithm
Gaussian mixture model
Robust EM
Initialization
Number of clusters

ABSTRACT

Clustering is a useful tool for finding structure in a data set. The mixture likelihood approach to clustering is a popular clustering method, in which the EM algorithm is the most used method. However, the EM algorithm for Gaussian mixture models is quite sensitive to initial values and the number of its components needs to be given a priori. To resolve these drawbacks of the EM, we develop a robust EM clustering algorithm for Gaussian mixture models, first creating a new way to solve these initialization problems. We then construct a schema to automatically obtain an optimal number of clusters. Therefore, the proposed robust EM algorithm is robust to initialization and also different cluster volumes with automatically obtaining an optimal number of clusters. Some experimental examples are used to compare our robust EM algorithm with existing clustering methods. The results demonstrate the superiority and usefulness of our proposed method.

1. Introduction

Data analysis is a science for analyzing data in real world, and cluster analysis is a useful tool for data analysis. Cluster analysis is a method for finding clusters within a data set characterized by the greatest similarity within the same cluster and the greatest dissimilarity between different clusters. Hierarchical clustering was the earliest clustering method used by biologists and social scientists, whereas cluster analysis became a branch of statistical multivariate analysis. Many theories and methods for cluster analysis have been presented in the literature [1–3]. In general, learning and recognition mostly start from clustering, so that cluster analysis becomes a type of unsupervised learning in pattern recognition and has been widely applied in various areas [4].

From the statistical point of view, clustering methods may be divided into probability model-based approaches and nonparametric approaches. The probability model-based approach assumes that the data set follows a mixture model of probability distributions so that a mixture likelihood approach to clustering may be used [2]. For a mixture model, the expectation and maximization (EM) algorithm [5] is commonly used. For a nonparametric approach, clustering methods may be based on an objective function of similarity or dissimilarity measures, and these can be divided into hierarchical and partitional methods. A hierarchical clustering method is a procedure for transforming a data set into a diagram, known as a *dendrogram*, based on the similarity or dissimilarity matrix of the data set. Most partitional methods suppose that the data set can be represented by finite cluster prototypes with their own objective functions. Therefore, defining the dissimilarity (or distance) between a point and a cluster prototype is essential for partition methods. The most popular partition methods with cluster prototypes are *k*-means [6,7], trimmed *k*-means [8,9], fuzzy c-means (FCM) [10,11], and mean shift [12,13].

In this paper we focus on clustering based on probability models, and in particular, we propose a robust type of EM algorithm for Gaussian mixture models. We know that the EM algorithm is quite sensitive to initial values, in which the number of components needs to be given a priori. In this paper we present a robust EM clustering algorithm which will be robust to initials and different cluster volumes with automatically obtaining an optimal number of clusters. Although some authors have considered the initial problems for the EM algorithm [14,15] and some have considered estimation of the number of components [15,16], there has been less consideration about robustness to initial values associated with the number of components for the EM algorithm. Since this robustness property is very important for the EM, we present a new means of solving these initial problems by automatically finding an optimal number of components. We first propose a new objective function based on mixture distributions and then create new update equations for the EM algorithm. We also construct a learning schema to automatically obtain an optimal number of components.

The rest of the paper is organized as follows. In Section 2, we briefly review the EM algorithm. In Section 3, we propose a robust

* Corresponding author.
E-mail address: msyang@math.cycu.edu.tw (M.-S. Yang).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.04.031

EM clustering algorithm. In Section 4, we use our algorithm with some artificial datasets and real datasets to demonstrate that this algorithm is effective in Gaussian mixture models. Finally, we state conclusions in Section 5.

2. The EM clustering algorithm

Let the data set $\{X_1, X_2,\ldots,X_n\}$ be a random sample of size n from the d-variate mixture model

$$f(x;\alpha,\theta)=\sum_{k=1}^{c}\alpha_k f(x;\theta_k) \tag{1}$$

where $\alpha_k>0$ denotes mixing proportions with the constraint $\sum_{k=1}^{c}\alpha_k=1$ and $f(x;\theta_k)$ denotes the density of x from kth class with corresponding parameters θ_k. Let $Z=\{Z_1, Z_2,\ldots,Z_n\}$ be the missing data in which $Z_i\in\{1, 2,\ldots,c\}$. If $Z_i=k$, it means that the ith data point belongs to the kth class. Thus, the joint pdf of the complete data $\{X_1, X_2,\ldots, X_n, Z_1, Z_2,\ldots,Z_n\}$ becomes

$$f(x_1,\ldots,x_n,z_1,\ldots,z_n;\alpha,\theta)=\prod_{i=1}^{n}\prod_{k=1}^{c}[\alpha_k f(x_i;\theta_k)]^{z_{ki}} \tag{2}$$

where $z_{ki}=\begin{cases}1, & \text{if } Z_i=k\\ 0, & \text{if } Z_i\neq k\end{cases}$. The log likelihood function is obtained as follows:

$$L(\alpha,\theta;x_1,\ldots,x_n,z_1,\ldots,z_n)=\sum_{i=1}^{n}\sum_{k=1}^{c}z_{ki}\ln[\alpha_k f(x_i;\theta_k)] \tag{3}$$

E-step: Since the latent variables z_{ki} are unknown, according to Dempster et al. [5], the conditional expected value $E(Z_{ki}|x_i;\alpha,\theta)$ is substituted for z_{ki}. By Baye's Theorem, we have

$$\hat{z}_{ki}=E(Z_{ki}|x_i;\alpha,\theta)=\frac{\alpha_k f(x_i;\theta_k)}{\sum_{s=1}^{c}\alpha_s f(x_i;\theta_s)} \tag{4}$$

M-step: Under the constraint $\sum_{k=1}^{c}\alpha_k=1$, to maximize

$$\tilde{L}(\alpha,\theta;x_1,\ldots,x_n)=\sum_{i=1}^{n}\sum_{k=1}^{c}\hat{z}_{ki}\ln[\alpha_k f(x_i;\theta_k)] \tag{5}$$

We can obtain the updated equation for mixing proportions with

$$\alpha_k=\frac{\sum_{i=1}^{n}\hat{z}_{ki}}{n} \tag{6}$$

We now consider the d-variate Gaussian mixture model

$$\begin{aligned}f(x;\alpha,\theta)&=\sum_{k=1}^{c}\alpha_k f(x;\theta_k)\\&=\sum_{k=1}^{c}\alpha_k(2\pi)^{-(d/2)}|\Sigma_k|^{-(1/2)}e^{-(1/2)(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)}\end{aligned} \tag{7}$$

The parameter θ_k consists of a mean vector μ_k and a covariance matrix Σ_k. Then the update equations for those parameters are as follows:

$$\mu_k=\frac{\sum_{i=1}^{n}\hat{z}_{ki}x_i}{\sum_{i=1}^{n}\hat{z}_{ki}} \tag{8}$$

$$\Sigma_k=\frac{\sum_{i=1}^{n}\hat{z}_{ki}(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{n}\hat{z}_{ki}} \tag{9}$$

Thus, the EM clustering algorithm can be summarized as follows.

EM clustering algorithm for normal mixtures

Step 1: Fix $2\le c\le n$ and fix any $\varepsilon>0$.
Give initials $\hat{z}^{(0)}=(\hat{z}_1^{(0)},\ldots,\hat{z}_c^{(0)})$ and let $s=1$.
Step 2: Compute $\alpha^{(s)}$ and $\mu^{(s)}$ with $\hat{z}^{(s-1)}$ using (6) and (8).
Step 3: Compute $\sum^{(s)}$ with $\hat{z}^{(s-1)}$ and $\mu^{(s)}$ using (9).
Step 3: Update $\hat{z}^{(s)}$ with $(\alpha^{(s)}, \mu^{(s)}, \sum^{(s)})$ using (4).
Step 4: Compare $\hat{z}^{(s)}$ to $\hat{z}^{(s-1)}$ in a convenient matrix norm $\|\cdot\|$.
IF $\|\hat{z}^{(s)}-\hat{z}^{(s-1)}\|<\varepsilon$, STOP
ELSE $s=s+1$ and return to step 2.

We mention that the convergence properties of the EM algorithm had been well discussed in Wu [17]. Afterwards, Xu and Jordan [18] considered more convergence properties of the EM algorithm for Gaussian mixtures. Ma et al. [19] considered the convergence rate of the EM algorithm for Gaussian mixtures. Since the EM algorithm is quite sensitive to initialization, in which the cluster numbers need to be given a priori, Figueiredo and Jain [20] proposed an algorithm to deal simultaneously with the number of clusters and also the estimates of parameters for mixture models by using the particular form of a minimum message length (MML) criterion. This criterion is the minimization of the following cost function via EM estimators

$$\begin{aligned}K(\alpha,\theta;x_1,\ldots,x_n)=&\frac{P}{2}\sum_{m:\alpha_m>0}\ln\left(\frac{n\alpha_m}{12}\right)+\frac{c_{nz}}{2}\ln\left(\frac{n}{12}\right)+\frac{c_{nz}(P+1)}{2}\\&-\sum_{i=1}^{n}\ln\left[\sum_{k=1}^{c}\alpha_k f(x_i;\theta_k)\right]\end{aligned} \tag{10}$$

where P is the number of parameters specifying each component and c_{nz} denotes the number of non-zero-probability components. Then the update equation for the proportion is as follows:

$$\frac{\alpha_k=\max\{0,\sum_{i=1}^{n}\hat{z}_{ki}-\frac{P}{2}\}}{\sum_{s=1}^{c}\max\{0,\sum_{i=1}^{n}\hat{z}_{si}-\frac{P}{2}\}} \tag{11}$$

In the d-variate Gaussian mixture model, $\theta_k=(\mu_k,\Sigma_k)$, $P=d+(d(d+1)/2)$ and the update equations of $\hat{z}_{ki}$, μ_k, and Σ_k are the same as the formulas (4), (8) and (9), respectively. For updating parameters, Figueiredo and Jain [20] used the component-wise EM method proposed by Celeux et al. [21], in which the parameters are sequentially updated. The algorithm proposed by Figueiredo and Jain [20] performs first by inputting larger cluster numbers and then by using the formula (11) to eliminate these smaller clusters to reduce the cluster number. After that, they use the criterion (10) to find the clustering that best minimizes the criterion. However, using random initial conditions for the EM in Figueiredo and Jain [20] still has an initialization problem in which using larger initial cluster numbers only makes this initialization problem lighter. We give an example to illustrate it as follows.

Example 1. In this example we use a data set, as shown in Fig. 1(a), generated from a two-component Gaussian mixture distribution with a sample size 800 and the parameters

$$\alpha_1=\alpha_2=0.5,\quad \mu_1=(0\quad 0)^T,\quad \mu_2=(20\quad 0)^T,$$
$$\Sigma_1=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix},\quad \text{and}\quad \Sigma_2=\begin{pmatrix}9 & 0\\ 0 & 9\end{pmatrix}.$$

We use the starting cluster number $c_{initial}=30$ and 100 different random initial conditions for the algorithm of Figueiredo and Jain [20]. Finally we have 78 of 100 with the results of $c^*=2$, as shown in Fig. 1(b), another 11 of 100 with the results of $c^*=3$, and the other 11 of 100 with the results of $c^*>3$. Fig. 1(c) demonstrates an incorrect clustering result of $c^*=3$. In fact, if a data set with a larger cluster number is considered, then the initialization problem for the algorithm of Figueiredo and Jain [20] will become more serious. In next section, we will compare the algorithm in Figueiredo and Jain [20] with our proposed robust EM clustering algorithm.