Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2018, 6, 11-17
http://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Research on Initialization on EM Algorithm Based on Gaussian Mixture Model

Ye Li, Yiyan Chen*

UK Systems Science Association, London, UK
Email: *townjam_sovietnia@163.com

How to cite this paper: Li, Y. and Chen, Y.Y. (2018) Research on Initialization on EM Algorithm Based on Gaussian Mixture Model. *Journal of Applied Mathematics and Physics*, **6**, 11-17.
https://doi.org/10.4236/jamp.2018.61002

Received: October 17, 2017
Accepted: January 2, 2018
Published: January 5, 2018

Abstract

The EM algorithm is a very popular maximum likelihood estimation method, the iterative algorithm for solving the maximum likelihood estimator when the observation data is the incomplete data, but also is very effective algorithm to estimate the finite mixture model parameters. However, EM algorithm can not guarantee to find the global optimal solution, and often easy to fall into local optimal solution, so it is sensitive to the determination of initial value to iteration. Traditional EM algorithm select the initial value at random, we propose an improved method of selection of initial value. First, we use the k-nearest-neighbor method to delete outliers. Second, use the k-means to initialize the EM algorithm. Compare this method with the original random initial value method, numerical experiments show that the parameter estimation effect of the initialization of the EM algorithm is significantly better than the effect of the original EM algorithm.

Keywords

EM Algorithm, Gaussian Mixture Model, K-Nearest Neighbor, K-Means Algorithm, Initialization

1. Introduction

Assessment of this performance of an algorithm generally relates to its efficiency, ease of operation and operation results. We are concerned about the efficiency of the iteration, and one of the factors that affect the efficiency of the iteration is the selection of the initial value. EM algorithm has a very important application in the Gaussian mixture model (*GMM*). In simple terms, if we don't know neither the parameters of the mixed model nor the classification of the observed data, EM algorithm is a popular algorithm for estimating the parameters of finite mixture model. However, sometimes its performance is poor. This algorithm has an obvious shortcoming: it is very sensitive to the initial value. Therefore, in order to get the parameter estimation of the closest to the true value, we have to find a method to initialize the EM algorithm. We can list several usual methods with initialization: random center, hierarchical clustering, k-means algorithm and so on [1]. As a result of the k-means clustering algorithm is also a kind of dynamic iterative algorithm and decides the classification number by subjective factors. Further, it is accordant with EM algorithm for parameter estimation of finite mixture model. Hence we can use outlier detection based on proximity to remove the outliers in order to reduce the influence of noise for the parameter estimation. Then, a rough grouping of the rest of the mixed data is given by k-means clustering. Finally, a rough estimate of parameters is given based on packet data.

2. Gaussian Mixture Modeling

The mixture model is a useful tool for density estimation, and can be viewed as a kind of kernel method [2]. If the d-dimensional random vector has a finite mixture normal distribution, its probability density function is as follows:

$$p(x\,|\,\theta)=\sum_{i=1}^{k}\alpha_i p_i(x\,|\,\theta_i)=\sum_{i=1}^{k}\alpha_i(2\pi)^{-\frac{d}{2}}\left|\Sigma_i\right|^{-\frac{1}{2}}\exp(-\frac{1}{2}(x-\mu_i)^T\,\Sigma_i^{-1}(x-\mu_i)) \quad (1)$$

Also there is:

$$p_i(x\,|\,\theta_i)=(2\pi)^{-\frac{d}{2}}\left|\Sigma_i\right|^{-\frac{1}{2}}\exp(-\frac{1}{2}(x-\mu_i)^T\,\Sigma_i^{-1}(x-\mu_i)),i=1,2,\cdots,k \qquad (2)$$

This is the probability density function of the *i*-th branch, which has a mean μ_i, a covariance Σ_i, and mixing proportion α_i, and $\sum_{i=1}^{k}\alpha_i=1$, $\theta_i=(\mu_i^T,\Sigma_i)^T$, $\theta=(\theta_1^T,\theta_2^T,\cdots,\theta_k^T,\alpha_1,\alpha_2,\cdots,\alpha_k)^T$ is a vector corresponding to all unknown parameters.

2.1. The EM Algorithm

The classical and natural method for computing the maximum-likelihood estimates (*MLEs*) for mixture distributions is the EM algorithm (*Dempster et al.*, 1977), which is known to possess good convergence properties. In other words, the EM algorithm is a kind of the iterative algorithm to solve the maximum likelihood estimate when the observation data is incomplete data, which has good application value as well. The process of parameter estimation of the EM algorithm is given in reference [3]: Y is generally assumed to be the observed data, Z is potential data, θ is unknown parameters, a posteriori distribution density function based on the observed data Y of parameter θ denoted by $p(\theta\,|\,Y)$, called observed posterior distribution, $p(\theta\,|\,Y,Z)$ represents that the posterior distribution density function of parameter θ is obtained after the addition of data *Z*, called addition posterior distribution, $p(Z\,|\,\theta,Y)$ represents the conditional distribution density function of the potential data Z in given θ

and observation data *Y*. The goal is to estimate the parameter values of the observed posterior distribution $\theta^{(t)}$, hence the EM algorithm is carried out as follows: Let $\theta^{(t)}$ be the estimate values of the posterior mode at the beginning of the $t+1$ iteration. The two step of the $t+1$ iteration is:

Expectation Step: calculate mathematical expectation of $p(\theta|Y,Z)$ or $\log p(\theta|Y,Z)$ on the conditional distribution of Z, namely:

$$Q(\theta|\theta^{(t)},Y)=E_z[\log p(\theta|Y,Z)|\theta^{(t)},Y]=\int_z \log[p(\theta|Y,Z)]p(Z|\theta^{(t)},Y)\mathrm{d}Z \tag{3}$$

Maximization Step: maximize the $Q(\theta|\theta^{(t)},Y)$, that is to find a point $\theta^{(t+1)}$, then:

$$Q(\theta^{(t+1)}|\theta^{(t)},Y)=\max_{\theta} Q(\theta|\theta^{(t)},Y) \tag{4}$$

after getting $\theta^{(t+1)}$, this forms an iteration $\theta^{(t)}\to\theta^{(t+1)}$, iterate Expectation step and Maximization step until $\left\|\theta^{(t+1)}-\theta^{(t)}\right\|$ or $\left\|Q(\theta^{(t+1)}|\theta^{(t)},Y)-Q(\theta^{(t)}|\theta^{(t)},Y)\right\|$ is sufficiently small.

2.2. Outlier Detection Based on Proximity

Outliers are data objects that are inconsistent with behavior or model of most of the data in the entire data set [4]. An object is abnormal if it is far away from most points. This method is more general and easier to use than the statistical method, because it is easier to determine meaningful proximity metrics than statistical distribution for data sets.

One of the easiest ways to measure whether an object is far away from most points is using the distance to the k-nearest neighbor. The outlier score of an object is given by the distance between the object and its the k-nearest neighbor. The minimum outlier score is 0, and the maximum value is the maximum value of the distance function: it is usually infinite.

Outlier score is highly sensitive to the value of *k*. If the *k* is too small, then a small amount of adjacent outliers can lead to a low outlier score. If *k* is too large, all objects in the clusters with points less than *k* are likely to be outliers. In order to make the scheme more robust for the selection of *k*, we can use the average distance of the first *k* nearest neighbors.

2.3. K-Means Algorithm

The K-means algorithm is one of the most popular iterative descent clustering methods [5]. This algorithm takes k as the parameter, then n objects are divided into k clusters, in the same cluster objects in a high similarity, while objects in different clusters in the greater dissimilarity. First, selecting k objects at random, each object represents a cluster center. For the rest of each object, which assigned to the most similar class according to the distance between the object and the cluster centers. And then calculate the new center of each cluster. This process is iterated until convergence. The method is based upon the following steps:

Step 1: Choose samples m randomly: $\overline{x}_1^{(k)},\overline{x}_2^{(k)},\cdots,\overline{x}_m^{(k)}$ as the cluster centers of mixed data.

Step 2: For the remaining $n-m$ data, noted:

$$x_1, x_2, \cdots, x_j, \cdots, x_{n-m}, d_{ji}^{(k)} = \left| x_j - \overline{x}_i^{(k)} \right| \tag{5}$$

This function represents the distance from x_j to $\overline{x}_i^{(k)}$, $j = 1, 2, \cdots, n-m$, $i = 1, 2, \cdots, m$. For fixed x_j, if $\left| x_j - \overline{x}_i^{(k)} \right| = \min d_{ji}^{(k)}, i = 1, 2, \cdots, m$. Then x_j is classified into the category i. Thus, we divide the data into m classes: $C_1^{(k)}, C_2^{(k)}, \cdots, C_m^{(k)}$.

Step 3: Use formula:

$$\overline{x}_i^{k+1} = \frac{1}{n_i^{(k)}} \sum_{j=1}^{n_i^{(k)}} x_{ji}, x_{ji} \in C_i^{(k)} \tag{6}$$

$$n_1^{(k)} + n_2^{(k)} + \cdots + n_i^{(k)} = n, i = 1, 2, \cdots, m$$

Find out the sample mean of the data of each class in the second step as a new cluster center.

Step 4: For a given number $\varepsilon > 0$, if $\sum_{i=1}^{m} \left| \overline{x}_i^{(k+1)} - \overline{x}_i^{(k)} \right| < \varepsilon$, then stop iteration, output value.

Otherwise, the new clustering center in the third step into the first step to replace the old clustering center, repeat the second step and third step operation.

2.4. The K-Means Algorithm Is Used to Initialize EM Algorithm after Deleting Outliers

First, we compute the distance between each observation data and its the k-nearest neighbor (*this distance is called the k-NN distance*), deleting those points which have relatively large the k-NN distances, namely abnormal observations that is far from most points. Second, we give a rough grouping of mixed data using k-means cluster method for the rest of data. Then according to the packet data, a rough estimate of parameters, as the initial value of the iterative algorithm, are given. Finally, execute EM algorithm until convergence to estimate the parameters of gaussian mixture modeling.

2.5. Simulation Studies

We compare the effect of parameter estimation about the improved initial value method and original method on the simulation data in this section.

First, we generate a one-dimensional data set; its histogram is shown in **Figure 1**. Simulated example with two classes, the data in first class are generated from one-dimensional normal population $N(3,1)$, and the data in second class are generated from one-dimensional normal population $N(-2, 2^2)$. The numbers of the data points of the 2 classes are 2000 and 3000. After iteration of the EM algorithm, the parameter estimation results are shown in later.

The original random initial value method of EM algorithm is very unstable, sometimes it need to be iterated nearly 100 times, and usually can not be very

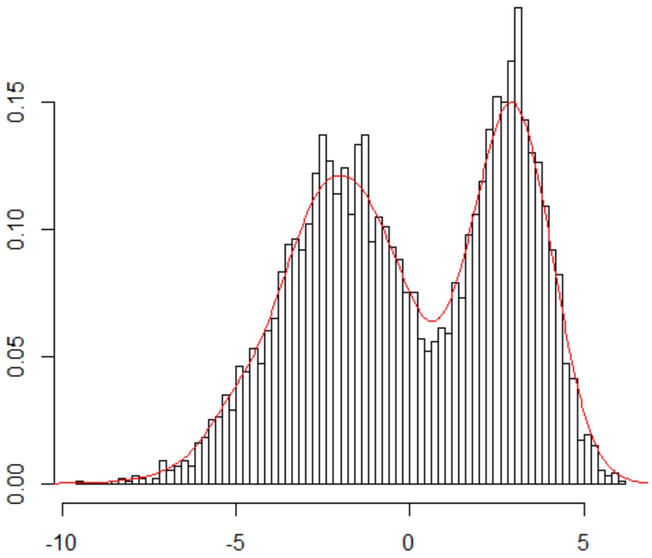

Figure 1. Histogram and density curve of the original data.

close the true value of the parameters. However, number of iterations of EM algorithm is greatly reduced by the improved initial value method which is pretty stable. And the final parameter estimates are closer to the true value compared to original method. Here, we select a test result to draw a line graph (see **Figure 2** in next page).

The average iteration times of the original EM algorithm is approximately 33, and the final iteration result is also highly unstable. Then we compare the parameter estimation results of the original method and the improved method on some realization. We can draw some conclusion from **Figure 2**, after 22 iterations, the final output parameter estimates of the original method are:

$$\begin{gathered}\hat{\mu}_1 = 2.829, \hat{\mu}_2 = -2.233 \\ \hat{\sigma}_1 = 1.112, \hat{\sigma}_2 = 1.795 \\ \hat{\alpha}_1 = 0.443, \hat{\alpha}_2 = 0.556\end{gathered} \tag{7}$$

The outliers are removed by the distances between each sample and its 10-nearest neighbor in improved initial value method. Then use the k-means method to classify the remaining points. The class center of each class is used as the initial value of the mean, the sample variance of each class is used as the initial value of the variance. The proportion of each class is used as the initial value of the coefficient. After 7 iterations, the change of the parameter is less than 10^{-2}. The results of parameter estimation are as follows

$$\begin{gathered}\hat{\mu}_1 = 2.949, \hat{\mu}_2 = -2.034 \\ \hat{\sigma}_1 = 1.011, \hat{\sigma}_2 = 1.992 \\ \hat{\alpha}_1 = 0.404, \hat{\alpha}_2 = 0.596\end{gathered} \tag{8}$$

3. Conclusion

Obviously, the performance of the improved method outperforms that of the original method [6]. The improved method has less number of iterations and

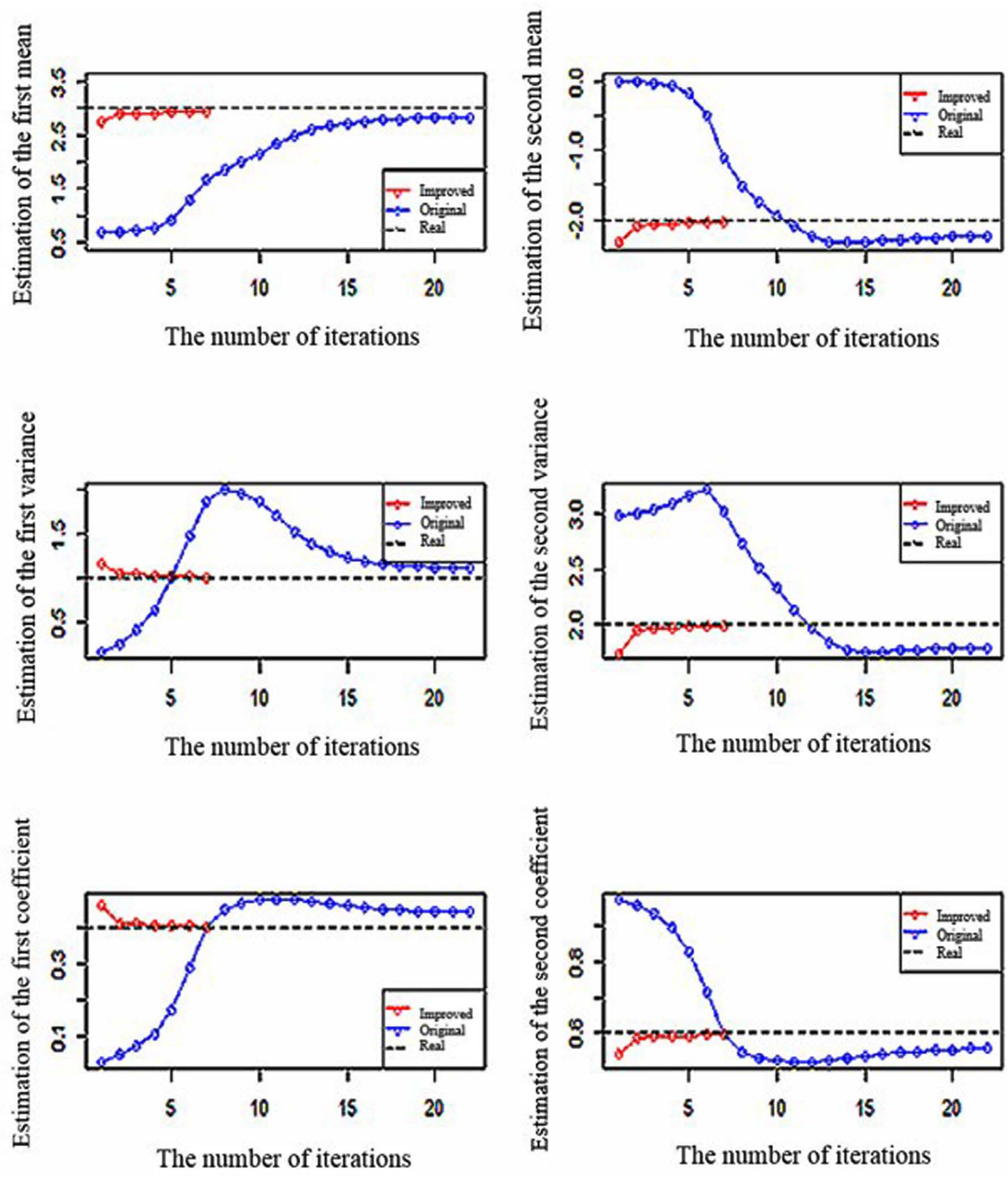

Figure 2. Effect comparison between improved initial value method and original method for parameter estimation.

better parameter estimation results. What is more, its statistical meaning is easy to understand. Then, the improved initial value method can be used not only for one-dimensional Gaussian mixture model, but also for multidimensional Gaussian mixture model.

However, it is need to further study how to choose the k when using the k-NN distance to remove outliers. At the same time, if we can further optimize the process of initial value selection (reducing complexity of deleting outliers and k-means clustering), it will bring greater use value.

References

[1] Wang, X. (2012) Gaussian Mixture Model Based k-Means to Initialize the EM Algorithm. *Journal of Shangqiu Normal University*, **28**, 11-14.

[2] Wang, J.K. and Gai, J.Y. (1995) Mixture Distribution and Its Application. *Journal of Biomathematics*, **10**, 87-92.

[3] Trevor, H., Robert, T. and Jerome, F. (2001) The Elements of Statistical Learning, Springer-Verlag, New York.

[4] Zhang, Z.P., Xu, X.Y. and Wang, P. (2011) Spatial Outlier Mining Algorithm Based on KNN Graph. *Computer Engineering*, 3737-3739.

[5] Zhu, J.Y. (2013) Research and Application of K-Means Algorithm. Dalian University of Technology.

[6] Zhai, S.D. (2009) Research on Clustering Algorithm Based on Mixtured Model. Northwest University.