Bangladesh J. Sci. Res. **28**(1): 1-7, 2015 (June)

ELLIPSOIDAL MIXTURE MODELS FOR CLUSTERING: AN APPLICATION TO WEB-BASED EDUCATION DATA

Saujanna Jafreen, Taslim S. Mallick[1*] and Jafar A. Khan[1]

Department of Natural Sciences, Daffodil International University, Dhaka-1207, Bangladesh

Abstract

This study considers the Gaussian mixture models for clustering. Since spherical and diagonal models occur very rarely in practice and analysis can be simplified when these models are implemented, we focus on the ellipsoidal models. EM algorithm is used to fit these models to a real data set related to an adaptive educational electronic course. Misclassification rates and Bayesian Information Criteria (BIC) values are used for comparison.

Key words: Model-based clustering, Gaussian mixture models, EM algorithm

Introduction

Suppose, we have n multivariate observations $x_1, x_2, ..., x_n$, where $x_i = (x_{i1}, x_{i2}, ..., x_{id})'$, $i = 1, 2, ..., n$, is the ith observation of a d-dimensional continuous random vector $(X_1, X_2, ..., X_d)'$. Our problem is to divide the n observations into K clusters (subsets) $C_1, C_2, ..., C_K$ in such a way that observations in the same cluster are more similar to one another than the observations assigned to other clusters. Usually, the number of clusters K is unknown and one requires that it be estimated from the data.

There are different types of clustering algorithms available in the literature. Partition-based methods represent each cluster by its centre (e.g. mean vector). These algorithms choose K initial centres, and then iteratively assign the observations to the nearest centres and updates the centres until the assignments do not change. Examples of partition-based algorithms are K-means (Lloyd 1957 and 1982, MacQueen 1967, Gersho and Gray 1992) and K-medoids (Kaufman and Rousseeuw 1990).

Hierarchical methods build a hierarchy of nested clusters by either agglomerative or divisive approaches. Agglomerative or 'bottom up' approaches (Ward 1963, Fernandez and Gomez 2008) start with each object as a cluster and recursively merges two clusters with the most similarity. Divisive or 'top down' approaches (Chavent *et al.* 2007, Zhong 2008) start with all observations as one cluster and at each step divides the cluster with the most dissimilar observations.

Partition-based methods have no guarantee of convergence to the global minimum and the value of K has to be specified by users. On the other hand, hierarchical methods require intensive computation. This is why model-based methods have been proposed in the literature which assume that the data are generated by a mixture of K underlying probability distributions. The most popular model is the Gaussian mixture model (Banfield and Raftery 1993) where each cluster C_l is modeled by a multivariate normal distribution with mean vector μ_l and variance-covariance matrix Σ_i.

* Corresponding author: <tsmallick@yahoo.com>. [1]Department of Statistics, Biostatistics and Informatics, University of Dhaka, Dhaka-1000, Bangladesh.

Several spherical, diagonal and ellipsoidal models have been proposed in this general frame work (Banfield and Raftery 1993, Celeux and Govaert 1995) and implemented by EM algorithm (Fraley and Raftery 1998). Pernkopf and Bouchaffra (2005) proposed genetic-based EM algorithm for learning Gaussian mixture models from multivariate data which is capable of selecting the number of components of the model using the minimum description length criterion. Dongbing (2008) used a distributed EM algorithm, a stochastic approximation to the standard EM algorithm, for Gaussian mixtures in sensor networks.

The objective of this study is to compare the performance of several Gaussians mixture models. EM algorithm is used to fit these models to a real data set related to an Adaptive Educational Electronic Course obtained from UCI machine learning repository. Misclassification rates and Bayesian Information Criteria (BIC) values are used for comparison.

Gaussian mixture models

Mixture models for clustering assume that the data are generated by a finite mixture of underlying probability distributions. Suppose the data set $\times$ consists of n independent multivariate observations x_i to be divided into K clusters $C_1, C_2, \ldots, C_K$. The likelihood for the mixture model is

$$L(\theta_1, \theta_2, \cdots, \theta_K \mid x) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(x_i \mid \theta_k),$$

where f_k and θ_k are the density and parameter-vector of the k th cluster, and π_k is the probability that an observation belongs to the k th cluster.

In the Gaussian mixture model, each cluster C_K is modeled by the multivariate normal distribution with mean vector μ_k and covariance matrix Σ_k. That is,

$$f_k(x_i \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|} \exp\left[-\frac{1}{2}(x_i - \mu_k)'\Sigma_k^{-1}(x_i - \mu_k)\right].$$

The shape, volume and orientation of each cluster C_k is determined by the covariance matrix Σ_k. Banfield and Raftery (1993) separated by the above mentioned geometric features of a cluster by representing the covariance matrix in terms of its Eigen value decomposition as follows:

$$\Sigma_k = \lambda_k D_k A_k D'_k,$$

where D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix whose diagonal elements are proportional to the eigenvalues of Σ_k, and λ_k is a scalar. The matrix D_k determines the orientation of cluster, A_k determines the shape, and λ_k determines its volume. Table 1 presents a list of Gaussians mixture models that can be obtained by varying one or more parameters. Clearly, *EII* is the most constrained model because it restricts $D_k A_k D'_k$ to the identity matrix I and assumes $\lambda_k = \lambda$ for all clusters. The unconstrained model *VVV* allows all of D_k, A_k and λ_k to vary between clusters. The unconstrained model has the advantage that it is the most general model, but the

number of parameters (that need to be estimated) is the largest. The other ellipsoidal models have fewer parameters.

Table 1. List of Gaussian mixture models

Name	Model	Distribution	Volume	Shape	Orientation
EII	λI	Spherical	Equal	Equal	NA
VII	$\lambda_k I$	"	Variable	Equal	NA
EEII	λA	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k A$	"	Variable	Equal	Coordinate axes
EVI	λA_k	"	Equal	Variable	Coordinate axes
VVI	$\lambda_k A_k$	"	Variable	Variable	Coordinate axes
EEE	$\lambda DAD'$	Ellipsoidal	Equal	Equal	Equal
EEV	$\lambda D_k A D'_k$	"	Equal	Equal	Variable
VEV	$\lambda_k D_k A D'_k$	"	Variable	Equal	Variable
VVV	$\lambda_k D_k A_k D'_k$	"	Variable	Variable	Variable

Fig. 1 shows the four ellipsoidal models in the case of four clusters. The model *EEE,* for example, considers that each cluster is ellipsoidal, but all the four ellipsoids (representing the four clusters) have equal volume, shape and orientation.

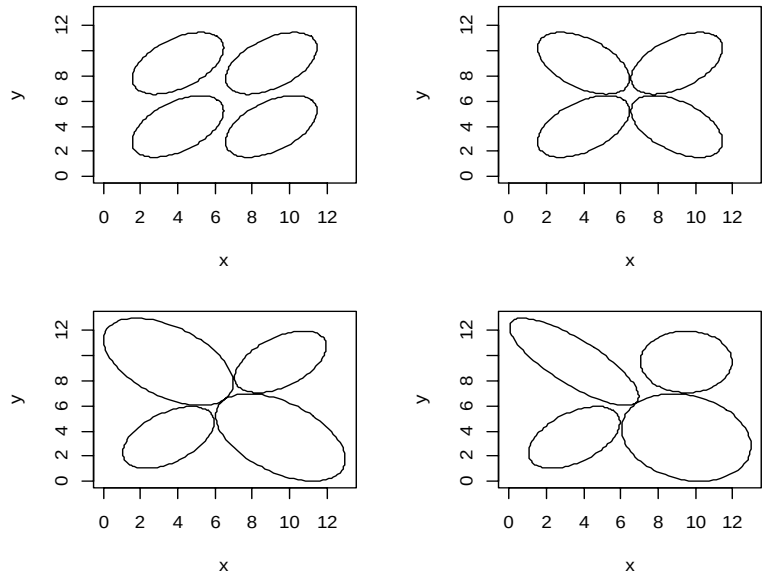

Fig. 1. The four ellipsoidal models in the case of four clusters.

Since spherical and diagonal models occur very rarely in practice and analysis can be very simplified when these models are implemented, we focus on the ellipsoidal models in our study.

The ellipsoidal mixture models can be fitted to clustered data by the Expectation-Maximization (EM) algorithm (Fraley and Raftery 1998) which is presented in the next section.

The EM algorithm

The Expectation-Maximization (EM) algorithm is usually employed to fit a model when data contain missing values. The algorithm first replaces the missing vlaues by some 'initial values' and fits the model to the 'complete data'. Then the algorithm alternates between the E-step (where the missing values are replaced by values expected under the current model) and the M-step (where the model is fitted again by maximizing the likelihood). The algorithm stops when the estimates of two successive iterations are very close.

In the clustering problem, we do not know which cluster an observation belongs to. In that sense, clustering data contain missing vlaues, where all the values of the clustering variable are 'missing'. Therefore, the EM algorithm can be used to obtain the estimates of these missing values.

In order to fit a particular Gaussian Mixture Model, the desired number of clusters K is specified. Then the model parameters π_k, $\boldsymbol{\mu}_k$, Σ_k, $k = 1, 2, ..., K,$ are estimated by the EM algorithm. To implement the EM algorithm, we require the initial estimates of the model parameters which are obtained either arbitrarily or by implementing a hierarchical clustering algorithm. Then the Expectation step (E-step) and the Maximization step (M-step) alternate until convergence. The E-step estimates the conditional probability of each observation belonging to each cluster, given the current parameter estimates. The M-step estimates the model parameters given the current conditional probabilities of the class occurrence. The two steps are elaborated below:

1. E-step: Compute the conditional probabilities $\pi_{ik} = P(C = K \mid x_i)$, which is proportional to $P(C = k) f(x_i \mid C = k) = \pi_k f_k(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \Sigma_k)$.

2. M-step: Compute new estimates of π_k, $\boldsymbol{\mu}_k$, Σ_k as follows:

$$\pi_k = \sum_i \pi_{ik}$$

$$\boldsymbol{\mu}_k = \frac{1}{\pi_k} \sum_i \pi_{ik} \boldsymbol{x}_i$$

$$\boldsymbol{\Sigma}_k = \frac{1}{\pi_k} \sum_i \pi_{ik} \boldsymbol{x}_i \boldsymbol{x}_i'.$$

In order to select the number of clusters K and the covariance structure, the Bayesian Information Criteria (BIC) (Schwarz 1978) is used. BIC is the value of the maximized log-likelihood with a penalty for the number of parameters in the model. The value of K is varied from 1 to 9 and the model with the largest BIC score is selected.

In the next section, we fit different mixture models to domain-dependent data of students' educational activities and classify their knowledge level using the EM algorithm.

Application: Web-based education

One of the main goals in machine learning is to develop an alrogithm that can best classify a user into one among different possible categories. User-models are commonly used in interactive systems, where the system adapts its behavior according to users specific needs. Based on the aquired information from the user, the user-model classifies the user as one of the K possible categories according to which an interactive system adapts its behavior. In web-based adaptive courses, a user-model collects students' data and use them to predict (classify) their knowledge level about the course.

In this paper, we analyse students data (n = 258) related to an Adaptive Educational Electronic Course (AEEC) obtained from the machine learning repository of the University of California Irvine (http://archive.ics.uci.edu/ml). Kahraman *et al.* (2013) proposed a user-model where *K* NN classifier is used combined with weights of the predictors to classify users knowledge status of the course. The objective of this paper is to use a classifier that fits different mixture models using EM algorithm and compare the predicted knowledge levels of students enrolled in AEEC with that of Kahraman *et al.* (2013). Similar to Kahraman *et al.* (2013), we use five predictors, namely degree of study time for AEEC (STG), degree of repetition number (SCG), performance in exams (PEG), degree of study time in prerequisite objects (STR) and learning status of prerequisite objects (LPR), where STG, SCG and PEG are the features about learning objects and others are the features about the prerequisite objects. The domain dependent data on these five features are obtained by real-valued functions over the range of 0 to 1. The response of the user-model is the student's current knowledge level which can be any one of 4 levels, high, medium, low and very low.

While we fitted four different mixture models to the data, only *EEE* identified that the data has 4 clusters. All of the three other models indicate the data to have only 3 clusters. To compare the results with that of Kahraman *et al.* (2013), we compute percentage-mismatch. Suppose for a given x_i, $\boldsymbol{x}_i$, C_{ij}^{KNN} and C_{ij}^{EM} denote that the i th unit has been classified into cluster j by the weighted *K*NN method of Kahraman *et al.* (2013) and EM estimation method, respectively. We say that a mismatch for i th unit occurs if $C_{ij}^{KNN} \neq C_{ij}^{EM}$. Thus the percentage-mismatch is computed as

$$\textbf{Mismatch}\ (\%) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{K} I\left(C_{ij}^{KNN} \neq C_{ij}^{EM}\right) \times 100}{n}.$$

Table 2 presents the percentage-mismatches observed for four different models fitted by the EM algorithm. *EEV* has the smallest mismatch percentage, whereas all other percentages are approximately the same. The table also shows the Bayesian Information Criteria (BIC) values for the four fitted models. According to the BIC values, *EEE* is the best model for the AEEC data.

Table 2. A comparison of the ellipsoidal mixture models.

Criteria	Model			
	EEE	*EEV*	*VEV*	*VVV*
Mismatch (%)	31.78	30.62	31.78	31.40
BIC values	151.57	86.87	78.24	46.47

Conclusion

This paper investigates the performance of several Gaussian Mixture Models. The spherical and diagonal models are not interesting from the practical point of view. Therefore, we focus on the ellipsoidal models *EEE*, *EEV*, *VEV* and *VVV*. The models are applied to a real data set related to an Adaptive Educational Electronic Course obtained from UCI machine learning repository. Only *EEE* identifies correctly that the data have 4 clusters, and the BIC value for this model is the largest. On the other hand, *EEV* is the best model with respect to the misclassification rate. This indicates that model selection based on BIC scores does not guarantee the best model with respect to other criteria.

References

Banfield, J.D. and A.E. Raftery. 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**(3): 803-821.

Celeux, G. and G. Govaert. 1995. Gaussian Parsimonious Clustering Models. *Pattern Recognition* **28**(5): 781-793.

Chavent, M., O. Briant and Y. Lechevallier. 2007. DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis* **52**(2): 687-701.

Dongbing, G. 2008. Distributed EM Algorithm for Gaussian Mixtures in Sensor Networks. *Neural Networks* **19**(7): 1154-66.

Fernandez, A. and S. Gomez. 2008. Solving Non-Uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms. *Journal of Classification* **25**(1): 43-65.

Fraley, C. and A.E. Raftery. 1998. How Many Clusters? Which Clustering Method? – Answers via Model-based Cluster Analysis. *Computer Journal* **41**: 578-588.

Gersho, A. and R.M. Gray. 1992. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers. Boston.

Kaufman, L. and P.J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

Lloyd, S. P. 1957 and 1982. Least squares quantization in PCM. *Technical Note, Bell Laboratories*. Published in 1982 in IEEE Transactions on Information Theory **28**: 128-137.

MacQueen, J.B. 1967. Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* **1**: 281-297.

Pernkopf, F. and Bouchaffra, D. 2005. Genetic-based EM algorithm for learning Gaussian mixture models. *Pattern Analysis and Machine Intelligence* **27**(8): 1344-48

Schwarz, G.E. 1978. Estimating the dimension of a model. *Annals of Statistics* **6**(2): 461-464.

Ward, J.H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**: 236-244.

Zhong, C., D. Miao, R. Wang and X. Zhou. 2008. DIVFRP: An automatic divisive hierarchical clustering method based on the furthest reference points. *Pattern Recognition Letters* **29**(16): 2067-2077.

(Manuscript received on 14 September, 2014; revised on 27 April, 2015)