

Cluster Detection through Model Based Clustering Approach

Vikas Singh*¹ and Mukesh Kumar²

Department of Statistics,
Central University of South Bihar,
Gaya-824236, Bihar.

*¹Email : singhvikasbhu@gmail.com

²Email : Mukseh.kumar@cub.ac.in

Abstract

Cluster analysis is an important technique for creating homogeneous group of objects, items. These groupings are sometimes essential for the proper analysis of data. Therefore, the technique of clustering is very useful in data analysis, especially in case of large data-sets. In clustering, the groups or clusters are found based on some characteristics of interest. Based on the chosen characteristics of interest similarity measures/distance matrices are calculated to perform the cluster analysis through heuristic clustering techniques. However, a more scientific approach of clustering is model based clustering approach. This approach is basically based on the distributional setups of observations. In this method, the data points are supposed to be drawn from a mixture distribution. The mixture distribution of multivariate normal gives some exciting and interesting result as compared to the other component distributions. In this paper, we have made an attempt to find clusters through a mixture of multivariate normal as components for the real dataset of student enrollment in higher education in India for clustering. The datasets is taken from All India Survey on Higher Education (AISHE) report for the year 2017-18. BIC is used for reporting the optimal no. of cluster and adequacy of the model.

Keywords: Model based cluster, EM-algorithm, Datasets, Bayesian Information Criterion.

Mathematics Subject Classification (MSC): 62H30, 68T10, 62Hxx, 62H12

1. INTRODUCTION

Clustering is a statistical procedure, which was primarily used for biological grouping/classification in 1960s. Cluster analysis getting popularity day by day in the present era of abundant data for extracting the useful information from the huge amount of data. The common heuristic procedures of cluster analysis are based on measures of similarity or dissimilarity among the data points. For clustering purposes, the dissimilarity matrices are calculated based on standard distances such as Euclidean distance, City-block or Manhattan distance, etc. for quantitative measurements while the similarity coefficients are defined for qualitative studies.

The clusters are created in a manner so that more similar observations are placed within the groups/clusters and dissimilar observations make the parts of different groups. The application of cluster analysis includes market research, psychology, bioinformatics, social science, criminology, pattern recognition, data mining, etc. In the literature (Johnsons (2007), Sharma (1996), etc.), there is a number

of well-known clustering methods or algorithms like single linkage (nearest neighbour), complete linkage, ward's method, etc. But these heuristic algorithms free from distributional forms of data. A better approach is model-based clustering approach. The method of model-based probability is based on some statistical probability distribution and its associated inference. The literature in the fields of cluster analysis includes Hartigan (1975), Everitt et al. (2011), Everitt (1993), Anderberg (1973), Symons (1981), Bensmail et al. (1997) and many others.

2. MODEL BASED CLUSTER ANALYSIS

Model-based probability clustering procedure is more scientific than the well-known heuristic clustering (distance based) techniques in the sense that model-based methods are based on mixture of probability density. In the theory of model-based probability grouping, it is supposed that the observations are coming from a mixture distribution. Particularly, finite mixture distribution is used in model based clustering setup. Multivariate normal distributions is used as a mixture in the literature as (McLachlan & Basford, 1988, Melnykov & Maitra, 2010) because of many interesting results might be obtained through this form. Some basics regarding this method of clustering is discussed below.

Let there is a set of n , p -dimensional vectors $y_1, y_2, \dots, y_n$ of observations coming from a mixture of g -component density and each component having some unknown mixing proportions or weights $\pi_1, \pi_2, \dots, \pi_g$ (McLachlan & Peel, 2004). Assuming that there is only finite no. of clusters in the data and the data belonging to different component distributions are forming a separate cluster. The probability density function (PDF) of finite mixture density (see Raftery & Dean, 2006 and Celeux & Govaert, 1995) is given by

$$f(y_j; \psi) = \sum_{i=1}^g \pi_i f_i(y_j; \theta_i) \quad (1)$$

Where $f_i(y_j; \theta_i)$ is the PDF of component distributions in the finite mixture model. Here $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^g \pi_i = 1$ for all $i = 1, 2, \dots, g$ and $j = 1, 2, \dots, n$. The vector $\psi = (\pi, \theta)$ of parameter includes all the parameters of finite mixture density. The no. of components or cluster forming the mixture is g .

Considering the multivariate normal distributions as component distribution which is widely used in literature and includes (Melnykov, 2010, McLachlan, 2011, Fraley, 1988) and some others. The PDF of the i^{th} component density can be written as

$$f_i(y_j; \theta_i) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{\{-\frac{1}{2}(y_j - \mu_i)^T \Sigma_i^{-1} (y_j - \mu_i)\}} \quad (2)$$

Where μ_i & Σ_i denote the mean and the covariance matrix vector, respectively and $\theta_i = (\mu_i, \Sigma_i)$ for $i = 1, 2, \dots, g$.

Banfield & Raftery, (1993) proposed that when the component distribution chosen as a multivariate normal in model-based clustering, the relationship between data points can be obtained by variance-covariance matrix and is decomposed in terms of eigenvalue as

$$\Sigma_i = \lambda_i O_i D_i O_i^T$$

Where $\lambda_i = |\Sigma_i|^{-\frac{1}{p}}$ is the scalar which is the highest among all eigenvalue of Σ_i . O_i is the orthogonal matrix of eigenvectors and orientation is obtained by O_i , see (Banfield and Raftery, 1993) D_i is diagonal matrix such that $|D_i| = 1$, with the normalize eigenvalue of Σ_i on the diagonal in decreasing order and it determines the shape of the cluster. λ_i measure the volume of corresponding i^{th} cluster. In the method of mixture of multivariate normal distribution of clustering given by (Banfield and Raftery, 1993), the clustering problem reduces to parameter estimation problem. Mclust (Fraley and Raftery (1999)) programme is used for conducting the analysis using R-software.

3. MAXIMUM LIKELIHOOD ESTIMATION

The parameter estimation in mixture model may be done through likelihood method of estimation. Let us suppose that the random sample (observation) $y_1, y_2, \dots, y_n$ drawn from the given finite mixture density (1). The likelihood function takes the following form

$$L(\psi) = \prod_{j=1}^n \sum_{i=1}^g \pi_i f_i(y_j | \theta_i) \quad (3)$$

$$L(\psi) = \prod_{j=1}^n f(y_j; \psi) \quad (\text{from (1)})$$

Taking log both the side of above equation we get,

$$\begin{aligned} \log L(\psi) &= \log \prod_{j=1}^n f(y_j; \psi) \\ l(\psi) &= \sum_{j=1}^n \log f(y_j; \psi) \end{aligned} \quad (4)$$

Where $l(\psi) = \log L(\psi)$

We solve the following likelihood equation (5) to estimate the parameters of the distribution,

$$\frac{\partial l(\psi)}{\partial(\psi)} = 0 \quad (5)$$

Where $\psi = (\pi, \theta)$.

The computational complexity in the parameter estimation of under the mixture of multivariate normal is quite high. Therefore, such cases may be handled with the EM algorithm see (Dempster et al., 1977 and Melnykov & Maitra, 2010, McLachlan & Krishnan 2008) and others. According to Everitt and Hand (1981), the final equations (estimates) can be obtained in the following form:

$$\begin{aligned} \widehat{\pi}_k &= \frac{1}{n} \sum_{i=1}^n \widehat{Pr}(k | x_i), & k &= 1, 2, \dots, g-1, \\ \widehat{\mu}_k &= \frac{1}{n\widehat{\pi}_k} \sum_{i=1}^n \widehat{Pr}(k | x_i) x_i, & k &= 1, 2, \dots, g, \\ \widehat{\Sigma}_k &= \frac{1}{n\widehat{\pi}_k} \sum_{i=1}^n \widehat{Pr}(k | x_i) (x_i - \widehat{\mu}_k)(x_i - \widehat{\mu}_k)', & k &= 1, \dots, g. \end{aligned}$$

The above equations do not provide the estimators explicitly; instead they must be solved by the iterative procedure. So this type of problem can be explained and solved by EM-algorithm.

4. EXPECTATION MAXIMIZATION (EM) ALGORITHM

The EM algorithm is proposed by (Dempster et al., 1997) is most widely used method for estimation through maximizing the likelihood when some observations are missing or incomplete. The EM is applied and described by (McLachlan & Krishnan, 2008, Melnykov (2013)). Similar to the K-mean methods, the algorithm first assumes the initial values for the parameters of the distributions. It then improves the estimates in an iterative process. The EM method can also be used to estimate parameters for non-Gaussian distribution as well.

Let $y_1, y_2, y_3, \dots, y_n$ be n independent random sample (observation) from the mixture density as given by the equation number (1) (see Melnykov (2013)). The parameter may be estimated through the technique of maximum likelihood estimation procedure and it depends on the likelihood function $L(\theta|y_1, y_2, \dots, y_n) = \prod_{j=1}^n f(y_j|\theta)$ which is usually not simple one because the complex log-likelihood function and is given in equation (4).

The EM algorithm gives a suitable way to maximize the likelihood through the inclusion of latent (unknown) variable $z_1, z_2, \dots, z_n$. Therefore, log-likelihood function is

$$L'(\theta|z_1, \dots, z_n, y_1, \dots, y_n) = \prod_{j=1}^n \prod_{i=1}^g (\pi_i f_i(y_j|\theta_i))^{I(z_i=g)}$$

where $I(\cdot)$ is an indicator function which takes value 1 for when y_i come from the g^{th} component; otherwise 0. The conditional expectation (mean) of the log-likelihood (L') function, given the observed random sample is the main objective of Expectation (E) step (see Melnykov, 2013). This expectation is represented by Q and is written as

$$Q(\theta|\theta^{(a-1)}, y_1, y_2, \dots, y_n) = \sum_{j=1}^n \sum_{i=1}^g p_{ji}^{(a)} (\log \pi_i + \log f_i(y_j|\theta_i)), \quad (6)$$

Where $a = 1, 2, \dots$ indicates the number of iteration and $p_{ji}^{(a)}$ is the posterior probability see (Melnykov, 2013) that y_j come from the g^{th} distribution obtained by

$$p_{ji}^{(a)} = \frac{\pi_i^{(a-1)} f_i(y_j|\theta_i^{(a-1)})}{\sum_{i'=1}^g \pi_{i'}^{(a-1)} f_{i'}(y_j|\theta_{i'}^{(a-1)})}$$

The objective of Maximization step is to maximize the equation (6) w.r.t θ . From (Melnykov, 2013 and Boyles, 1983) the E and M-steps are repeated till the convergence criteria is achieved at some repetition a . Therefore, we get the maximum likelihood estimate (MLE) as $\hat{\theta} = \theta^{(a)}$ and the ultimate posterior probabilities are $\hat{p}_{ji} = p_{ji}^{(a)}$.

5. FINDING NUMBER OF CLUSTER AND SELECTION OF MODEL

Finding the optimal no. of clusters or group and model selection is always a challenging task in clustering. In Gaussian mixture modeling (GMM), we also need to choose which covariance

parameterization to implement. These two issues, one regarding no. of clusters and model selection can be addressed by the information criteria, namely BIC see (Schwartz, 1978) in the method of model-based clustering. BIC is most popular choice among all in the framework of Gaussian Mixture Modeling see (Fraley & Raftery, 1998), and given as

$$BIC = -2\log L(\hat{\psi}) + d \log n$$

where, $L(\hat{\psi})$ represent the maximum value of log-likelihood which maximises the likelihood with G component, n is the no. of observations (size of sample) in the matrix and d is the no. of estimated parameters from the given density. The maximum value of log-likelihood which minimises the BIC is selected as a best fitted model.

6. DATA ANALYSIS AND RESULT

Model-based probability clustering technique has been applied to the higher education enrollment data (like UG, PG, MPhil, PhD, Diploma and Certificate) of 36 states of India for the year 2017-18. Here we performed cluster analysis using Multivariate normal mixture model through EM- algorithm. There are three clusters have been recorded as first, second and third which includes 15, 16 and 5 states respectively. The mixing probabilities for the three clusters are 0.4166599, 0.4444512 and 0.1388889 respectively. The best (appropriate) model using finite mixture normal model fitted by EM-algorithm is VEV, which is of varying volume, orientation and equal shape. For this fitted model, the BIC value is -4487.412 which is least among all other models. Figure (1) shows the BIC plot of model fitting and optimal number of cluster using BIC. The plot shows that VEV model with three (3) optimal cluster is most appropriate for above dataset. The assignment of cluster using VEV model with 3 clusters shown in figure (2) and figure (3) represent the uncertainty of cluster allocation.

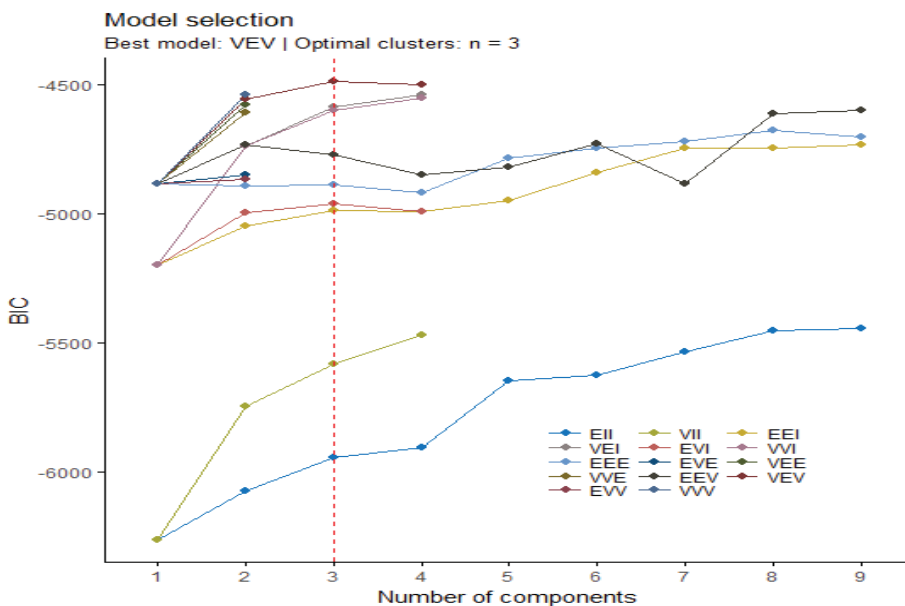


Figure 1. BIC plot for fitted model and no of clusters.

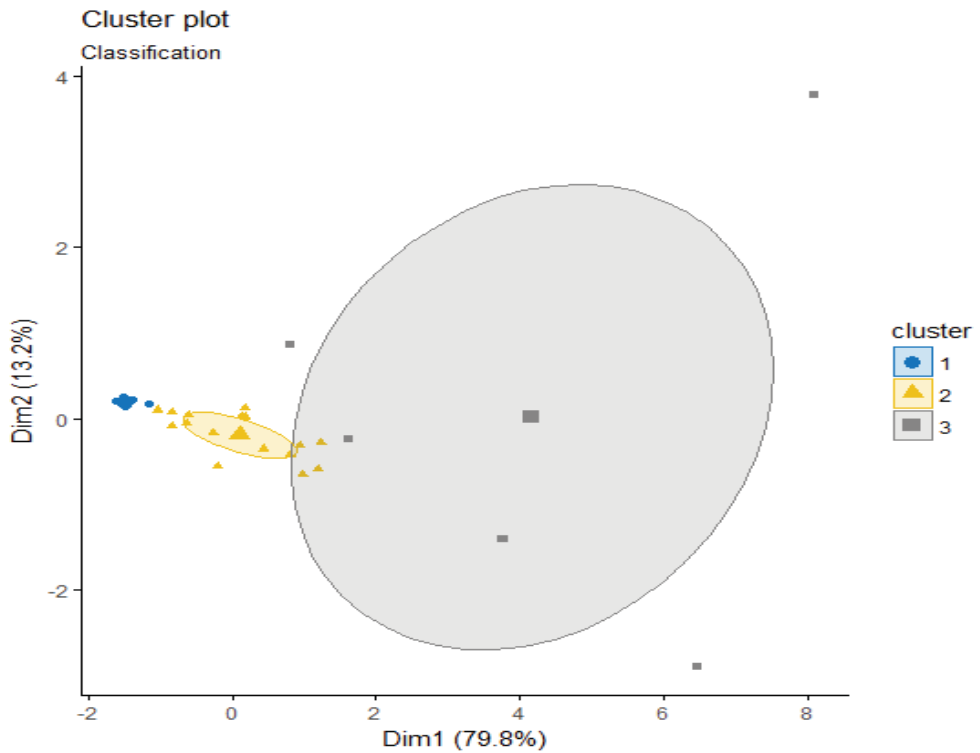


Figure 2. Data assignment in clusters using the VEV model.

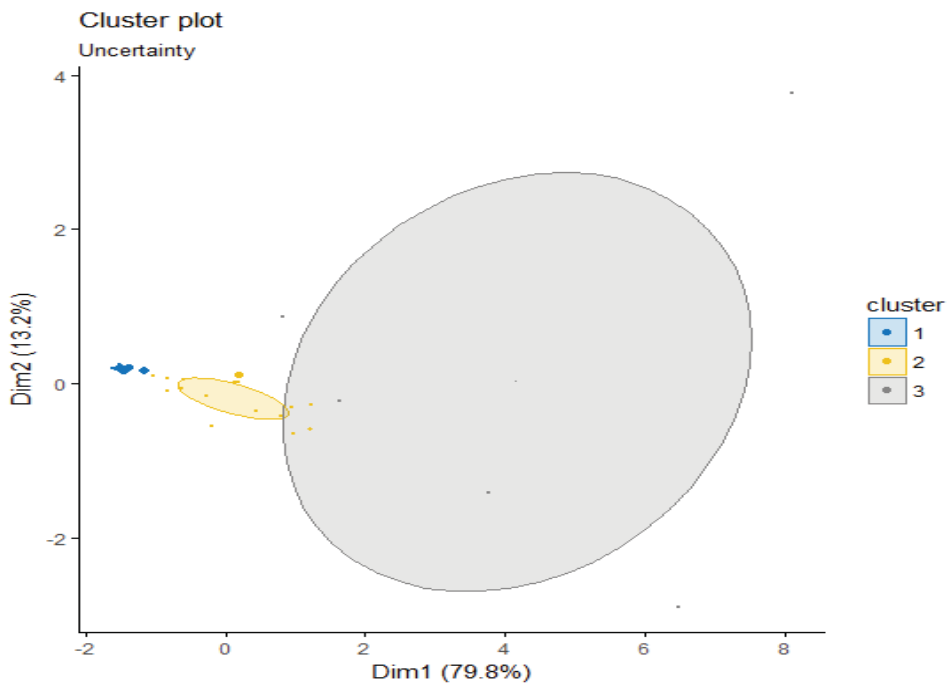


Figure 3. Plot of Uncertainty in respect of allocation of observations in clusters.

7. CONCLUSION

The current study reflects the heterogeneity of enrolment in higher education among the states. The clusters obtained through the concept of finite mixture model approach. This concept is more scientific and well accepted for the purpose of clustering. The above analysis reports that there are three (3) clusters with observations, 15, 16 and 5 respectively. It means 41% of the states lies within cluster 1, 44% of the states lies within cluster 2 and rest of the 13% states lies within cluster 3. Here we conclude that in respect of enrolment in higher education, the Indian education system having three (3) category of high, medium and low enrolment states.

8. REFERENCES

- Akogul, S., Erisoglu, M., 2016, A Comparison of Information Criteria in Clustering Based on Mixture of Multivariate Normal Distributions. *Math. Comput. Appl.* 21-34.
- Akogul, S., Erisoglu, M., 2017, An approach for determining the number of clusters in model based cluster analysis. *Entropy*. **19**(9), 452
- Anderberg, M.R., 1973, *Cluster Analysis for Applications*. New York, Academic Press.
- Polymenis, A., 2019, Some Comments on the Performance of Information Ratios for Assessing Numbers of Components in Finite Mixtures. *International Journal of Applied Mathematics and Statistics*, **58**, 36-43.
- Banfield, J.D., Raftery, A.E., 1993, Model Based Gaussian and Non Gaussian Clustering. *Biometrics*, **49**, 803-821.
- Bhuyan, K.C., 2005, *Multivariate Analysis and Its Application*. New central Book Agency (P) Ltd., Kolkata, India.
- Boyles, R. A., 1983, On the convergence of the EM algorithm. *J. R. Stat. Soc., B*, **45**, 47-50.
- Dasgupta, A., Raftery, A. E., 1998, Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering, *Journal of the American Statistical Association*, **93**, 294-302.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977, Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B*. **39**, 1–38.
- Everitt, B.S., Hand, D.J., 1981, *Finite Mixture Distributions*. Springer Netherlands, Dordrecht.
- Everitt, B.S., 1987, *Introduction to Optimization Methods and their Application in Statistics*. Chapman and Hall, London, New York.
- Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011, *Cluster Analysis*. John Wiley and Sons.
- Fräley, C., 1988, Algorithms for model-based Gaussian hierarchical clustering. *SIAM J. Sci. Comput.* **20**, 270–281.
- Fräley, C., Raftery, AE, 1998, How many clusters? Which clustering method? Answer via model-based cluster analysis. *The Computer Journal*. **41**(8):578–588.
- Fräley, C., Raftery, AE, 1999, MCLUST: Software for model-based cluster analysis. *Journal of Classification*, **16**, 297–206.
- Hartigan, J.A., 1975, *Clustering Algorithms*, Wiley, New York.
- Johnson, R.A., Wichern, D.W., 2015, *Applied Multivariate Statistical Analysis*. Pearson education Inc.

- Kaufman, L., Rousseeuw, P.J., 1990, *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons, Inc.; New York.
- McLachlan, G., Basford, K.E., 1988, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.
- McLachlan, G., Peel, D., 2004, *Finite Mixture Models*. John Wiley & Sons: Hoboken, NJ, USA.
- McLachlan, G., Krishnan, T., 2008, *The EM Algorithm and Extensions*. New York: John Wiley and Sons.
- Melnykov, V., Maitra, R., 2010, Finite mixture models and model-based clustering. *Statistics surv*, **4**, 80-116.
- Melnykov, V., 2013, Challenges in model based clustering. *WIREs Comput Stat*, **5**, 135-148.
- Murtagh, F., Raftery, A. E., 1984, Fitting straight lines to point patterns. *Pattern Recognition*, **17**, 479-483.
- R Core Team, 2018, R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
- Raftery, A.E., Dean, N., 2006, Variable selection for model based clustering. *Journal of American Statistical Association*, **101(473)**, **168-178**.
- Report, 2017-18, *All India Survey on Higher Education (AISHE)*, Ministry of Human Resource Development, Department of Higher Education, New Delhi, India.
- Schwarz, G., 1978, Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464.
- Sharma, S., 1996, *Applied Multivariate Techniques*. John Wiley & Sons, Inc. USA.
- Sneath, H.A., Sokal, R.R., 1973, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, p. 573. Freeman, San Francisco.
- Stahl, D., Sallis, H., 2012, Model-based cluster analysis. *WIREs Comput Stat.* **4**, **341-358**.