Normal Mixture Model-Based Clustering of Data Using Genetic Algorithm

Maruf Gogebakan[1(✉)] and Hamza Erol[2]

[1] Department of Applied Mathematics, Faculty of Computer Science, Abdullah GUL University, Kayseri, Turkey
maruf.gogebakan@agu.edu.tr
[2] Department of Computer Engineering, Faculty of Engineering, Mersin University, Mersin, Turkey
herol@mersin.edu.tr

Abstract. In this study, a new algorithm was developed for clustering multivariate big data. Normal mixture distributions are used to determine the partitions of variables. Normal mixture models obtained from the partitions of variables are generated using Genetic Algorithms (GA). Each partition in the variables corresponds to a clustering center in the normal mixture model. The best model that fits the data structure from normal mixture models is obtained by using the information criteria obtained from normal mixture distributions.

Keywords: Genetic Algorithm · Gaussian mixture models · Model based clustering · Information criteria

1 Introduction

Components in multivariate normal mixture distributions match a fragmentation in heterogeneous data [2]. Mixture model of normal distributions is defined as

$$f(x_j;\theta) = \sum_{i=1}^{g} \pi_i f_i(x_j;\psi_i) \tag{1}$$

where π_i represents probability weights for i. cluster centers $0 < \pi_i < 1$ and $\sum_{i=1}^{g} \pi_i = 1$ and $\psi_i = (\mu_i, \Sigma_i)$ denotes μ_i mean vectors and Σ_i covariance matrix. $f_i(x_j;\mu_i,\Sigma_i)$ multivariate Gaussian densities shown as,

$$f_i(x_j;\mu_i,\Sigma_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_i|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(x_j-\mu_i)^T\Sigma_i^{-1}(x_j-\mu_i)\right\} \tag{2}$$

thus, in the equation of (1), $\theta = \pi_1,\ldots,\pi_g,\psi_1,\ldots,\psi_g$ vectors are the multivariate parameter vector of normal mixture distributions in the Ω space vector [1].

© Springer Nature Switzerland AG 2020
D. J. Hemanth and U. Kose (Eds.): ICAIAME 2019, LNDECT 43, pp. 539–543, 2020.
https://doi.org/10.1007/978-3-030-36178-5_43

2 Data Set and Method

The use of Genetic Algorithms for model-based clustering by using fragmentation of variables in multivariable data is described on synthetic big data.

2.1 Variable Data Fragmentation in Multivariable Data

In multivariable data, the fragmentation of the variable is determined using univariate normal mixture distributions [7]. Univariate normal mixture distribution define as

$$f(x;\theta) = \sum_{i=1}^{g} \pi_i f_i(x;\mu_i,\sigma_i) \tag{3}$$

where $(x;\theta)$, g and π_i represents normal mixture distribution density function, number of component in distribution and mixed probability weight respectively.

The number of fragmentation of variables is obtained by optimization with information criteria using Gaussian mixture distributions [7]. In multivariable data, fragmentation numbers can be estimated approximately by looking at the graphs of the variables (Histogram and Normal Probability Plots) [6].

2.2 Determining the Observations of Fragmentation in Variable by k-Means Algorithm

In multivariable data, the fragmentation of the variables is determined with the help of normal mixed models, and the values of the observations falling into the fragmentation in the heterogeneous variables are assigned to the subgroups in the data. The determination of the observations to the subgroups is calculated using the Euclid distance with the help of the following equation (Table 1).

Table 1. Determining the fragmentations of variables in the multivariate data set based on normal mixture models.

Variables	X_1	X_2	$X_3X_4X_5 \ldots X_{15}$
Fragments	2	2	1 1 1 … 1

$$argmin_s \sum_{i=1}^{k} \sum_{j \in s_i} \| x_j - \mu_i \|^2 \tag{4}$$

2.3 Determining the Number of Cluster and Location of Normal Mixture Models Using Genetic Algorithms

When creating normal mixed models, the number of sets, the location of the cluster centers, the structure of the sets and the volume are determined by heterogeneous (fragmented) variables [3]. Genetic Algorithms are used to determine the number of

clusters and mixture models. Lower and upper limit of the numbers cluster depending on the fragmentation of variables in the mixture model, $s = 1, \ldots, p$ and k_s are obtained as follows to show disintegration on the variable.

$$C_{max} = \prod_{s=1}^{p} k_s \; ve \; C_{min} = max\{k_s\} \tag{5}$$

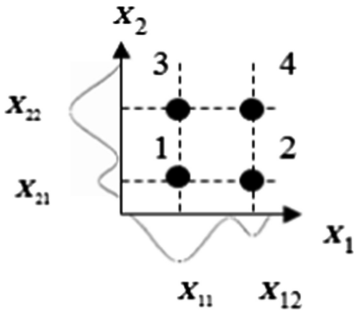

Fig. 1. Cluster centers corresponding to heterogeneous variables in multivariate data

It is obtained in the form of $C_{max} = 2.2.1. \ldots 1.1 = 4$ and $C_{min} = max\{2, 2, 1, \ldots 1, , 1\} = 2$. corresponding to the fragmentation in the data set (Fig. 1).

2.4 Number and Structure of Normal Mixed Models

In the case of multivariate data, the fragmentation of the heterogeneous variables X_1 and X_2 the homogeneous structure of the other variables, the number of models indicated by the total number of M_{Total} is calculated as follows.

$$M_{Toplam} = 2^{C_{Max}} - 1 = 2^4 - 1 = 15 \tag{6}$$

2.5 Number of Candidate Normal Mixture Models and Genetic Algorithms

In multivariable data, the number of normal mixture models is determined depend on the number of cluster centers and subgroups of heterogeneous variables. In normal mixture models, at least one subset (cluster) corresponds to the fragmentation of heterogeneous variables, whereas models with centers that do not correspond to subgroups are invalid models. As a result of the calculation, the locations and model numbers of the centers are given in Table 2.

Genetic Algorithm determines the components of normal mixed models calculated by fragmentation. While mixture models are represented by DNA sequence, centers with clustering in the sequence are shown as 1, non-clustering centers are 0 and models are generated with Genetic Algorithm. Gene sequences and model numbers corresponding to the assumption mixture models are given in Table 3.

Table 2. Cluster numbers, clusters' positions, and model numbers for candidate models matching the hypothesis among mixture models.

Cluster numbers	Location of centers	# of model
2	1st Column 1 and 2nd Column 1 Cluster	2
3	1st Column 2 and 2nd Column 1 Cluster	4
4	1st Column 2 and 2nd Column 2 Cluster	1
# of model		15
# of candidate model		7

Table 3. The number of normal mixture models, the number of suitable models and the DNA sequence in data with fifteen variables.

# of cluster	# of models	# of suitable models	DNA sequence
			1 2 3 4 …14 15
1 Cluster model	4	–	–
2 Cluster models	6	2	100100…0
			011000…0
3 Cluster models	4	4	111000…0
			110100…0
			101100…0
			011100…0
4 Cluster models	1	1	111100…0

2.6 Parameter Estimation of Normal Mixture Models Using Genetic Algorithm

Component weights, mean vectors and covariance matrices are calculated from the sample for normal mixture distribution models in multivariable data. Fragmentations in heterogeneous variables constitute mixed models using the number of clusters and the positions of the genes. While applying the genetic algorithm on the model, the 0 and 1 genes in the DNA sequence representing the model determine the presence or absence of clustering in the position corresponding to the mixture distribution.

3 Conclusion

In this study, the method for calculating the normal mixed models on the big data for clustering using genetic algorithm has been proposed. In the method, fragmentation in the appropriate variables of the data was found based on univariate normal mixture distributions. Cluster centers corresponding to subgroups of variables were determined and calculation method using Genetic Algorithms was proposed to determine the number of models. Normal mixture models have been created by applying genetic algorithms and DNA sequences corresponding to the obtained models are shown. The calculation of the parameters in the mixture model was used to calculate the

information criteria for each model using the genetic algorithm and to obtain the best normal mixture model.

Acknowledgment. The authors also would like to thank to editor(s) for his support to adapt the study content to Springer chapter requirements.

References

1. Fraley, C., Raftery, A.E.: How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput. J. **41**, 578–588 (1998)
2. McLachlan, G.J., Chang, S.U.: Mixture modelling for cluster analysis. Stat. Methods Med. Res. **13**, 347–361 (2004)
3. Erol, H.: A model selection algorithm for mixture model clustering of heterogeneous multivariate data. In: 2013 IEEE International Symposium on Innovations in Intelligent Systems and Applications Innovations in Intelligent Systems and Applications, Albena, pp. 1–7 (2013). https://doi.org/10.1109/inista.2013.6577617
4. Servi, T., Erol, H.: On total number of candidate component cluster centers and total number of candidate mixture models in model based clustering. Selçuk J. Appl. Math. **8**(2), 57–69 (2007)
5. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19** (6), 716–723 (1974)
6. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
7. Gogebakan, M., Erol, H.: A new semi-supervised classification method based on mixture model clustering for classification of multispectral data. J. Indian Soc. Remote Sens. **46**(8), 1323–1331 (2018)
8. Erol, H., Gogebakan, M., Erol, R.: Grid structures and orientations of clusters using discretization of variables in big data. In: Proceedings of International Conference on Engineering, Technology, and Applied Science ICETA 2017, pp. 16–31 (2017). ISSN 2411-9318