A Data Mining Method For Refining Groups In Data Using Dynamic Model Based Clustering

Tayfun Servi
Department of Elementary Education
Mathematics Education Section
Faculty of Education
Adıyaman University
02040 Adıyaman – Turkey.
e-mail: tservi@adiyaman.edu.tr

Hamza Erol
Department of Software Engineering
Faculty of Computer Sciences
Abdullah Gül University
Erciyes Teknopark No: 4
38039 Melikgazi / Kayseri – Turkey.
e-mail: hamza.erol@agu.edu.tr

Abstract—**A new data mining method is proposed for determining the number and structure of clusters, and refining groups in multivariate heterogeneous data set including groups, partly and completely overlapped group structures by using dynamic model based clustering. It is called dynamic model based clustering since the structure of model changes at each stage of refinement process dynamically. The proposed data mining method works without data reduction for high dimensional data in which some of variables including completely overlapped situations.**

Keywords-Data mining; dynamic model based clustering; refining groups in data.

I. INTRODUCTION

It is very difficult to find the number and structure of clusters when number of groups and/or variables in multivariate data ser gets higher and some of groups in data overlaps partly or completely [1]. There are several methods used for clustering multivariate data set [2][3][4]. Mixture model based clustering is one of these methods. It is used for partitioning of p -dimensional heterogeneous multivariate data into meaningful subgroups [5]. Each component of the mixture model corresponds to a cluster or group in heterogeneous multivariate data set. Mixture models of multivariate normal densities are used for modeling a wide variety of random phenomena [6].

Mixture model of multivariate normal densities for clustering assumes a set of n p-dimensional vectors $\boldsymbol{x}_1,...,\boldsymbol{x}_n$ of observations from g groups or clusters each with some unknown proportion $\pi_1,...,\pi_g$. It is assumed that the mixture of multivariate normal densities of the j th data point $\boldsymbol{x}_j$ for $j = 1,...,n$ can be written as

$$f(\boldsymbol{x}_j;\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{x}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) \quad (1)$$

where π_i denotes the mixing proportion of the data points in group or cluster i such that $0 < \pi_i < 1$ and $\sum_{i=1}^{g} \pi_i = 1$. The group conditional densities $f_i(\boldsymbol{x}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ depend on unknown parameter vectors $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ [7]. $f_i(\boldsymbol{x}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$'s are assumed to be multivariate normal group conditional densities, with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$, of the form

$$f_i(\boldsymbol{x}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)\right\} \quad (2)$$

where superscript T denotes the transpose.

Working principle of dynamic model based clustering for determining the number and structure of clusters in multivariate heterogeneous data set will be explained on glass identification data [8]. The estimation of parameters for mixture model at each stage of refinement process is computed by using MIXMOD program [9]. Parameters in mixture model are estimated by the method of maximum likelihood together using with EM algorithm [10]. The best dynamic model based clustering is obtained by an information criterion. There are several information criterions for model selection. Bayesian information criterion (BIC) is used for model selection in this study [11]. BIC is the most popular and frequently used information criterion in model selection [12].

II. THE METHOD FOR FINDING MIXTURE STRUCTURE IN HETEROGENEOUS MULTIVARIATE DATA SET

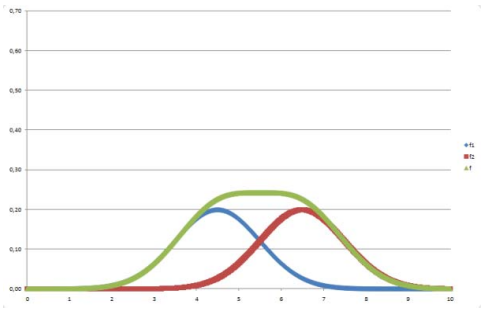
A. *Partly And Completely Overlapped Group Structures In Heteregenous Multivariate Data Set*

Partly and completely overlapped group structures make it difficult to determine the number and structure of clusters in multivariate data set. The cases of partly and completely overlapped group structures in heterogeneous data are shown in Figure 1. It shows also mixture structure in data.

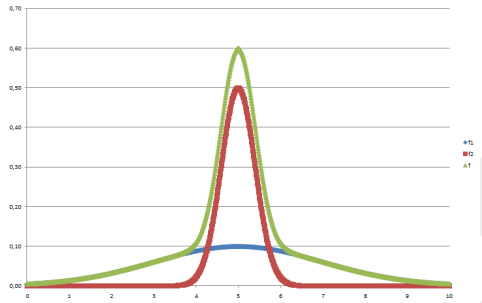
The first case is partly group overlapping, in this case group means are different but very close to each other. Group variances are the same as shown in Figure 1(a). The second case is completely group overlapping, in this case group means are the same. Group variances are different from each

978-1-4799-0661-1 / 13 / $31.00 © 2013 IEEE

other as shown in Figure 1(b). The horizontal axis denotes the values of independent variable and the vertical axis denotes the values of dependent variable with respect to the values of independent variable having heterogeneous structure thus having mixture structure in Figure 1. There is only one independent variable having two subgroups in Figure 1.



(a)



(b)

Figure 1. (a) The case of partly overlapping: group means are different but very close to each other. Group variances are the same. (b) The case of completely overlapping: group means are the same. Group variances are different from each other.

B. *Algorithm For Diagnosing Mixture Structure In Multivariate Heteregoneous Data Set And Refining Groups*

Model selection methods based on information criterions is applied for diagnosing the mixture structure in heterogeneous multivariate data set. The algorithm given in this section is a new algorithm. The steps of the algorithm proposed for refining groups in data using dynamic model based clustering is given below:

1. Apply model based clustering for multivariate data set by assuming that multivariate data have mixture structure thus data comes from a mixture of multivariate normal densities. This step shows either multivariate data is homogeneous thus there is no mixture structure or multivariate data have a group structure thus there is a mixture structure.
2. If multivariate data have mixture structure or multivariate data contains groups then find these groups in multivariate data. Test each group found in multivariate data for homogeneity. In other words check each group found for further mixture structure or subgroups.
3. If the data in a subgroup is homogeneous then determine that subgroup as a cluster in multivariate data and construct multivariate normal model for the data in that group. If the data in a subgroup is not homogeneous then go back to the step 1 and repeat the process for the data in that heterogeneous subgroup.
4. Repeat these processes in previous steps until homogeneity is achieved for all subgroups.

The mixture model of multivariate normal densities in (1) is established for whole data at the beginning and then each subgroup of data is obtained at the successive stages until homogeneity is achieved for all subgroups. The parameters π_i, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ of the mixture model of multivariate normal densities in (1) should be estimated for establishing the mixture model and finding the groups in heterogeneous multivariate data. The parameters π_i, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ can iteratively be estimated by maximizing the likelihood function using the Expectation-Maximization (EM) algorithm [12] in terms of the following updates:

$$z_j^{i(k)} = \frac{\boldsymbol{\pi}_i^{(k)} f_i(\boldsymbol{x}_j; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)})}{\sum_{i=1}^{g} \boldsymbol{\pi}_i^{(k)} f_i(\boldsymbol{x}_j; \boldsymbol{\mu}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)})} \tag{3}$$

$$\boldsymbol{\pi}_i^{(k+1)} = \frac{\sum_{j=1}^{n} z_j^{i(k)}}{n} \tag{4}$$

$$\boldsymbol{\mu}_i^{(k+1)} = \frac{\sum_{j=1}^{n} (z_j^{i(k)} \boldsymbol{x}_j)}{\sum_{j=1}^{n} z_j^{i(k)}} \tag{5}$$

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^{n} z_j^{i(k)} (\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{x}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^{n} z_j^{i(k)}} \tag{6}$$

where $z_j^{i(k)}$ denotes the posterior probability of j-th observation belongs to i-th group at k-th iteration; $\boldsymbol{\pi}_i^{(k+1)}$ denotes the mixture proportion of i-th component at $k+1$-th iteration; $\boldsymbol{\mu}_i^{(k+1)}$ denotes mean vector of i-th component at $k+1$-th iteration and $\boldsymbol{\Sigma}_i^{(k+1)}$ denotes variance-covariance matrix of i-th component at $k+1$-th iteration.

III. The Data Mining Method For Refining Groups In Heterogeneous Multivariate Data Using Dynamic Model Based Clustering

A. *Properties Of Heteregenous Multivariate Data Set*

The working principle of data mining method proposed for refining groups in multivariate data set using dynamic model based clustering will be explained on glass identification data set [8]. The study of classification of types of glass was

motivated by criminological investigation. Glass identification data is consists of 214 data points and nine variables: variable 1 - RI: refractive index, variable 2 - Na: Sodium, variable 3 - Mg: Magnesium, variable 4 - Al: Aluminum, variable 5 - Si: Silicon, variable 6 - K: Potassium, variable 7 - Ca: Calcium, variable 8 - Ba: Barium, variable 9 - Fe: Iron. All variables except first are in unit measurement: weight percent in corresponding oxide. Type of glass: building windows float processed, building windows non float processed, vehicle windows float processed, vehicle windows non float processed, containers, tableware and headlamps.

B. Computing Minimum And Maximum Number Of Clusters In Heterogeneous Multivariate Data Set

At the beginning of data mining method for refining groups in data using dynamic model based clustering process an interval for the number groups in heterogeneous multivariate data set should be constructed.

TABLE I. THE RESULTS OF COMPUTATIONS FOR GLASS IDENTIFICATION DATA SET

n	p	x_s	k_s	$C_{\min}$ and $C_{\max}$ values	g
214	9	x_1	2	$C_{\min} = 2$ $C_{\max} = 32$	6
		x_2	2		
		x_3	1		
		x_4	2		
		x_5	2		
		x_6	1		
		x_7	2		
		x_8	1		
		x_9	1		

The minimum number of clusters denoted by $C_{\min}$ and the maximum number of clusters denoted by $C_{\max}$ in heterogeneous multivariate data set. These can be computed by the method proposed by Servi and Erol [13] as

$$C_{\min} = \max\left\{k_1, k_2, ..., k_p\right\} \tag{7}$$

$$C_{\max} = \prod_{s=1}^{p} k_s \tag{8}$$

In (7) and (8), p denotes the number of variables and k_s denotes the number of partitions in each variable in the heterogeneous multivariate data. k_s can be computer by using k-means algorithm. The results of computations for glass identification data set are given in Table I. According to results in Table I we expect the number of clusters for glass identification data between 2 and 32 according to the method proposed by Servi and Erol [13]. We need this knowledge for processing data using the data mining method proposed in this study. The variables x_1, x_2, x_4, x_5 and x_7 have mixture structure and they form the group structure in whole heterogeneous multivariate data that is why they are chosen among the variables x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 and x_9 in glass identification data.

C. Stages Of Data Mining Method For Refining Groups In Data Using Dynamic Model Based Clustering Process

We will use the variables x_1, x_2, x_4, x_5 and x_7 in glass identification data since these variables have mixture structure. We assume that the variance-covariance matrices of component densities are in one of forms ellipsoidal pk_Lk_Ck, diagonal pk_Lk_Bk and spherical pk_Lk_I [9]. These forms of the variance-covariance matrices of component densities makes model based clustering dynamic.

Log-likelihood function values should be computed for each mixture model of multivariate normal densities. Likelihood function for the mixture model of multivariate normal densities in (1) is

$$L(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{j=1}^{n} f(\boldsymbol{x}_j; \boldsymbol{\theta}) = \prod_{j=1}^{n} \sum_{i=1}^{g} \boldsymbol{\pi}_i f_i(\boldsymbol{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{9}$$

and log-likelihood function for the mixture model of multivariate normal densities in (1) is

$$\log L(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{g} \boldsymbol{\pi}_i f_i(\boldsymbol{x}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\right) \tag{10}$$

Log-likelihood function values for the mixture model of multivariate normal densities computed at *k-th* iteration by using the estimated values of $\boldsymbol{\pi}_i^{(k+1)}$, $\boldsymbol{\mu}_i^{(k+1)}$ and $\boldsymbol{\Sigma}_i^{(k+1)}$ in (4), (5) and (6) respectively.

Bayesian information criteria (BIC) values should be computed for each mixture model of multivariate normal densities. BIC values will be used as criterion for selecting the best model. BIC can be computed as

$$BIC = -2\log L(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) + d \log n \tag{11}$$

where $\log L(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ is the value of log-likelihood function for mixture model of multivariate normal densities computed at *k-th* iteration by using the estimated values of $\boldsymbol{\pi}_i^{(k+1)}$, $\boldsymbol{\mu}_i^{(k+1)}$ and $\boldsymbol{\Sigma}_i^{(k+1)}$ in (4), (5) and (6) respectively. d is the number of parameters in mixture model of multivariate normal densities and n is the number of observation.

Stage 0: Let's call the group containing all glass identification data as G_0. Then the mixture model of multivariate normal densities in (1) is established for G_0 by using MIXMOD program [9]. The structure of the models and BIC values corresponding to these models for Stage 0 is given in Table II.

TABLE II. THE STRUCTURE OF THE MODELS AND BIC VALUES CORRESPONDING TO THESE MODELS FOR STAGE 0

BIC Values and Model Type	Forms of Variance-Covariance matrix		
	Ellipsoidal pk_Lk_Ck	Diagonal pk_Lk_Bk	Spherical pk_Lk_I
BIC values of mixture model with two components for G_0	**-776.932**	-340.200	1963.862

According to the results in Table II, the group G_0 have two subgroups called G_{11} and G_{12}. The number of data in G_0 is 214. The number of data classified in G_{11} and G_{12} is 140 and 74 respectively.

Stage 1: The mixture models of multivariate normal densities in (1) are established for G_{11} and G_{12} by using MIXMOD program [9]. The structure of the models and BIC values corresponding to these models for Stage 1 is given in Table III.

TABLE III. THE STRUCTURE OF THE MODELS AND BIC VALUES CORRESPONDING TO THESE MODELS FOR STAGE 1

BIC Values and Model Type	Forms of Variance-Covariance matrix		
	Ellipsoidal pk_Lk_Ck	Diagonal pk_Lk_Bk	Spherical pk_Lk_I
BIC values of mixture model with two components for G_{11}	**-1362.698**	-1054.407	425.495
BIC values of mixture model with three components for G_{12}	**164.125**	323.187	1038.236

According to the results in Table III, the group G_{11} have two subgroups called G_{111} and G_{112}. The number of data in G_{11} is 140. The number of data classified in G_{111} and G_{112} is 83 and 57 respectively. The group G_{12} have three subgroups called G_{121}, G_{122} and G_{123}. The number of data in G_{12} is 74. The number of data classified in G_{121}, G_{122} and G_{123} is 28, 28 and 18 respectively.

Stage 2: The mixture models of multivariate normal densities in (1) are established for G_{111}, G_{112}, G_{121}, G_{122} and G_{123} by using MIXMOD program [9]. The structure of the models and BIC values corresponding to these models for Stage 2 is given in Table IV.

TABLE IV. THE STRUCTURE OF THE MODELS AND BIC VALUES CORRESPONDING TO THESE MODELS FOR STAGE 2

BIC Values and Model Type	Forms of Variance-Covariance matrix		
	Ellipsoidal pk_Lk_Ck	Diagonal pk_Lk_Bk	Spherical pk_Lk_I
BIC values of mixture model with two components for G_{111}	**-1093.863**	-975.916	-34.640
BIC values of mixture model with three components for G_{112}	**-491.792**	-320.488	298.880
BIC values of mixture model with two components for G_{121}	**70.1632**	133.0939	472.680
BIC values of mixture model with two components for G_{122}	**-131.540**	-43.899	246.304
BIC values of mixture model with two components for G_{123}	**41.080**	90.028	286.032

According to the results in Table IV, the group G_{111} have two subgroups called G_{1111} and G_{1112}. The number of data in G_{111} is 83. The number of data classified in G_{1111} and G_{1112} is 37 and 46 respectively. The group G_{112} have three subgroups called G_{1121}, G_{1122} and G_{1123}. The number of data in G_{112} is 57. The number of data classified in G_{1121}, G_{1122} and G_{1123} is 26, 18 and 13 respectively. The group G_{121} have two subgroups called G_{1211} and G_{1212}. The number of data in G_{121} is 28. The number of data classified in G_{1211} and G_{1212} is 20 and 8 respectively. The group G_{122} have two subgroups called G_{1221} and G_{1222}. The number of data in G_{122} is 28. The number of data classified in G_{1221} and G_{1222} is 21 and 7 respectively. The group G_{123} have two subgroups called G_{1231} and G_{1232}. The number of data in G_{123} is 18. The number of data classified in G_{1231} and G_{1232} is 7 and 11 respectively.

Stage 3: The mixture models of multivariate normal densities in (1) are established for G_{1111}, G_{1112}, G_{1121}, G_{1122}, G_{1123}, G_{1211}, G_{1212}, G_{1221}, G_{1222}, G_{1231} and G_{1232} by using MIXMOD program [9]. The structure of the models and BIC values corresponding to these models for Stage 3 is given in Table V.

TABLE V. THE STRUCTURE OF THE MODELS AND BIC VALUES CORRESPONDING TO THESE MODELS FOR STAGE 3

BIC Values and Model Type	Forms of Variance-Covariance matrix		
	Ellipsoidal pk_Lk_Ck	Diagonal pk_Lk_Bk	Spherical pk_Lk_I
BIC values of mixture model with two components for G_{1111}	**-545.419**	-500.599	-50.240
BIC values of mixture model with three components for G_{1112}	**-694.525**	-597.743	-40.716
BIC values of model for G_{1121}	**-243.290**	-158.627	121.205
BIC values of model for G_{1122}	**-204.082**	-157.577	63.146
BIC values of model for G_{1123}	**-156.963**	-108.814	37.356
BIC values of mixture model with three components for G_{1211}	**21.649**	84.661	306.726
BIC values of model for G_{1212}	**18.814**	39.595	132.421
BIC values of model for G_{1221}	**-147.724**	-45.088	190.240
BIC values of model for G_{1222}	**-52.117**	38.895	115.068
BIC values of model for G_{1231}	**-49.315**	27.519	112.303
BIC values of model for G_{1232}	**18.248**	56.918	186.041

According to the results in Table V, the group G_{1111} have two subgroups called G_{11111} and G_{11112}. The number of data in G_{1111} is 37. The number of data classified in G_{11111} and G_{11112} is 10 and 27 respectively. The group G_{1112} have two subgroups called G_{11121} and G_{11122}. The number of data in G_{1112} is 46. The number of data classified in G_{11121} and G_{11122} is 15 and 31 respectively. The group G_{1121} have no subgroup. The number of data in G_{1121} is 26. The group G_{1122} have no subgroup. The number of data in G_{1122} is 18. The group G_{1123} have no subgroup. The number of data in G_{1123} is 13. The group G_{1211} have three subgroups called G_{12111}, G_{12112} and G_{12113}. The number of data in G_{1211} is 20. The number of data classified in G_{12111}, G_{12112} and G_{12113} is 6, 8 and 6 respectively. The group G_{1212} have no subgroup. The number of data in G_{1212} is 8. The group G_{1221} have no subgroup. The number of data in G_{1221} is 21. The group G_{1222} have no subgroup. The number of data in G_{1222} is 7. The group G_{1231} have no subgroup. The number of data in G_{1231} is 7. The group G_{1232} have no subgroup. The number of data in G_{1232} is 11.

Stage 4: The mixture models of multivariate normal densities in (1) are established for G_{11111}, G_{11112}, G_{11121}, G_{11122}, G_{12111}, G_{12112} and G_{12113} by using MIXMOD program [9]. The structure of the models and BIC values corresponding to these models for Stage 4 is given in Table VI.

TABLE VI. THE STRUCTURE OF THE MODELS AND BIC VALUES CORRESPONDING TO THESE MODELS FOR STAGE 3

BIC Values and Model Type	Forms of Variance-Covariance matrix		
	Ellipsoidal pk_Lk_Ck	Diagonal pk_Lk_Bk	Spherical pk_Lk_I
BIC values of model for G_{11111}	-153.015	**-157.471**	-28.202
BIC values of model for G_{11112}	**-441.829**	-422.343	-59.472
BIC values of model for G_{11121}	**-228.259**	-206.997	-11.084
BIC values of model for G_{11122}	**-527.689**	-443.599	-29.466
BIC values of model for G_{12111}	**-33.378**	8.648	89.521
BIC values of model for G_{12112}	-8.428	92.881	**-49.907**
BIC values of model for G_{12113}	23.953	109.889	**-31.282**

According to the results in Table VI, the group G_{11111} have no subgroup. The number of data in G_{11111} is 10. The group G_{11112} have no subgroup. The number of data in G_{11112} is 27. The group G_{11121} have no subgroup. The number of data in G_{11121} is 15. The group G_{11122} have no subgroup. The number of data in G_{11122} is 31. The group G_{12111} have no subgroup. The number of data in G_{12111} is 6. The group G_{12111} have no subgroup. The number of data in G_{12111} is 6. The group G_{12112} have no subgroup. The number of data in G_{12112} is 8. The group G_{12113} have no subgroup. The number of data in G_{12113} is 6.

D. Classification Tree For Data Mining Method Proposed For Refining Groups In Multivariate Data Set Using Dynamic Model Based Clustering

A classification tree is obtained at the end of data mining method, proposed for refining groups in multivariate data set using dynamic model based clustering, processed for glass identification data set. The classification tree consists of groups at levels or stages. There is whole data group at level 0 or stage 0. There are homogeneous groups at leafs of the classification tree. The depth of the classification tree is determined by the last stage of data mining method, proposed for refining groups in multivariate data set using dynamic model based clustering. The depth of the classification tree for glass identification data set is 4.

At the end of refining groups in multivariate data set using dynamic model based clustering, 15 homogeneous groups found in glass identification data set.

The classification tree of data mining method, proposed for refining groups in multivariate data set using dynamic model based clustering, processed for glass identification data set is given in Figure 2. The classification tree in Figure 2 represents the grouping structure obtained at the end of the data mining method proposed applied for glass identification data set.

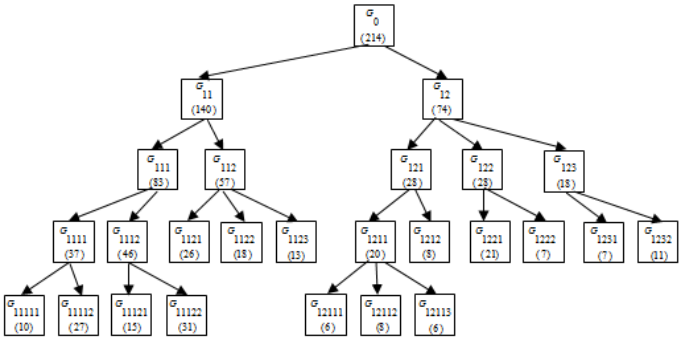


Figure 2. The classification tree of data mining method proposed for refining groups in multivariate data set using dynamic model based clustering, processed for glass identification data set.

IV. Discussions And Conclusions

The data mining method proposed for determining the number and structure of clusters, and refining groups in multivariate heterogeneous data set can be very useful for model based classification of heterogeneous multivariate data. This is a new study. The theoretical basis and an application denoting the working principle of the new data mining method are given in this study. The comparisons of the performances of the proposed data mining method with other techniques in the literature will be subject of another study.

The method proposed can separate groups, partly and completely overlapped group structures by using dynamic model based clustering without any data reduction. It is called dynamic model based clustering since the structure of model changes at each stage of refinement process dynamically.

Model selection methods based on information criterions is applied for diagnosing the mixture structure in heterogeneous multivariate data set. The steps of the algorithm proposed for refining groups in data using dynamic model based clustering is repeated until homogeneity is achieved for all subgroups.

The estimates of parameter values for models in each stage are not given. But the number of observations falling or classified to the related subgroups are given in paranthesis in classification tree in Figure 2.

A classification tree is obtained at the end of data mining method, proposed for refining groups in multivariate data set using dynamic model based clustering. The classification tree consists of groups at levels or stages. There is whole data group at level 0 or stage 0. There are homogeneous groups at leafs of the classification tree. The depth of the classification tree is determined by the last stage of data mining method, proposed for refining groups in multivariate data set using dynamic model based clustering. The depth of the classification tree for glass identification data set is obtained as 4.

References

[1] Fraley, C. (1998). Algorithms for model-based Gaussian hierarchical clustering. SIAM Journal on Scientific Computing. (20). 270-281.

[2] Jain, A.K. And Dubes, R.C. (1988). Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ.

[3] Kaufman, L., And Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons, Inc.

[4] Witten, I. H., Frank, E. and Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Burlington USA: Morgan Kaufmann Publishers.

[5] Fraley, C. and Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association, 97, 611-631.

[6] McLachlan, G. J. and Chang, S. U. (2004). Mixture Modelling for Cluster Analysis. Statistical Methods in Medical Research 13, 347-361.

[7] Fraley, C. and Raftery, A. E. (1998). How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. The Computer Journal, 41, 578-588.

[8] Halbe, Z. And Aladjem, M. (2005). Model-based mixture discriminant analysis: an experimental study. Pattern Recognition. 38, 437-440.

[9] Biernacki, C., Celeux, G., Govaert, G. And Langrognet, F. (2006). Model-based cluster analysis and discriminant analysis with the MIXMOD software. Computational Statistics and Data Analysis, 51, 587-600.

[10] Mclachlan, G.J. And Krishnan, T. (1997). The EM Algorihm and Extensions. John Wiley & Sons, Inc. New York.

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

[12] Mclachlan, G. And Peel, D. (2000). Finite Mixture Models. John Wiley & Sons, Inc. New York.

[13] Servi, T. and Erol, H. (2007). On Total Number Of Candidate Component Cluster Centers And Total Number of Candidate Mixture Models In Model Based Clustering. Selçuk Journal of Applied Mathematics Vol.8. No.2. pp. 57 – 69.