Journal of the Indian Society of Remote Sensing
https://doi.org/10.1007/s12524-018-0808-9

RESEARCH ARTICLE

A New Semi-supervised Classification Method Based on Mixture Model Clustering for Classification of Multispectral Data

Maruf Gogebakan[1] · Hamza Erol[2]

Received: 13 February 2017 / Accepted: 29 May 2018

Abstract
A new method for semi-supervised classification of remotely-sensed multispectral image data is developed in this study. It consists of unsupervised-clustering for data labelling and supervised-classification of clusters in multispectral image data (MID) using spectral signatures. Mixture model clustering, based on model selection, is proposed for finding the number and determining the structures of clusters in MID. The best mixture model, for the best clustering of data, finds the number and determines the structure of clusters in MID. The number of elements in the best mixture model fits to the number of clusters in MID. The elements of the best mixture model fits to the structure of clusters in MID. Clusters in MID is supervised-classified using spectral signatures. Euclidean distance is used as the discrimination function for the supervised-classification method. The values of Euclidean distances are used as decision rule for the supervised-classification method.

Keywords Mixture model clustering · Model selection · MID · Supervised-classification · Unsupervised-clustering · Variable data segmentation

Introduction

Gaussian mixture model clustering is one of the clustering methods for partitioning of p-dimensional multivariate/MID into meaningful subgroups (Fraley and Raftery 1998). Each component in the Gaussian mixture model corresponds to a cluster in multivariate/MID (McLachlan and Chang 2004). Gaussian mixture model clustering assumes a set of n p-dimensional vectors $x_1, \ldots, x_n$ of observations from a finite mixture model of g groups or clusters each with some unknown proportions $\pi_1, \ldots, \pi_g$. It is assumed that the Gaussian mixture model of the jth data point $\boldsymbol{x}_j$ for $j = 1, \ldots, n$ can be written as

$$f(\boldsymbol{x}_j;\boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{x}_j;\boldsymbol{\psi}_i) \tag{1}$$

where π_i shows the weights/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 or cluster conditional densities $f_i(\boldsymbol{x}_j;\boldsymbol{\psi}_i)$ depend on an unknown parameter vector $\boldsymbol{\psi}_i$. $f_i(\boldsymbol{x}_j;\boldsymbol{\psi}_i)$'s are assumed to be Gaussian group or cluster conditional densities, with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$, of the form

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

where $\boldsymbol{\psi}_i$'s are the vectors of parameters in the component densities, thus $\boldsymbol{\psi}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. The superscript T shows the transpose. So $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_g)$ shows the vector of all unknown parameters in the Gaussian mixture model densities over the parameter space Ω.

Genetic algorithm based on variable data segmentation and information criteria, which are developed for model based clustering of RS-MID using normal mixture clustering (Fraley et al. 2012), were used to determine the best clustering model. Meaningful subgroups of each variable

✉ Maruf Gogebakan
maruf.gogebakan@agu.edu.tr

Hamza Erol
herol@mersin.edu.tr

[1] Department of Applied Mathematics, Faculty of Computer Sciences, Abdullah Gul University, Kayseri, Turkey

[2] Department of Computer Engineering, Faculty of Engineering, Mersin University, Mersin, Turkey

Published online: 20 July 2018

determine the clusters in MID (Erol 2013). Among all the mixture models generated by the subgroups of variables, the appropriate mixture models were selected and the inappropriate mixture models were removed so that candidate mixture models suitable for the data cluster structure were determined. For the best model to be chosen based on the statistical information criteria, the unknown parameters to be used in normal mixture models are estimated from the sample using variable data segmentation.

The subfields in the remotely-sensed multispectral satellite image data correspond to meaningful subgroups in the model-based clustering analysis. Today's equipped satellites are collecting as much data as required and there is a need for supervised or unsupervised methods which are applicable, known for performance and cost less than other methods for the analysis of obtained data (Erol and Akdeniz 2005).

Method

Gaussian mixture model clustering based model selection using variable segmentations will be explained on remotely-sensed multispectral satellite image data.

RS-MID

The proposed Gaussian mixture model clustering based model selection method was applied to clustering the RS-MID of an agricultural area. A Landsat Thematic mapper image of an agricultural area of the Seyhan plain ($\approx$ 37°N, 36°E) in Adana having a size of 198 $\times$ 200 (in total 39,600) pixels was used as the MID (Erol and Akdeniz 2005). The data were collected on 27 March 1992 (Path 175—Row 34). Landsat Thematic Mapper bands 3, 4, and 5 were used for the model based clustering. Satellite image of working agricultural area is illustrated in Fig. 1.

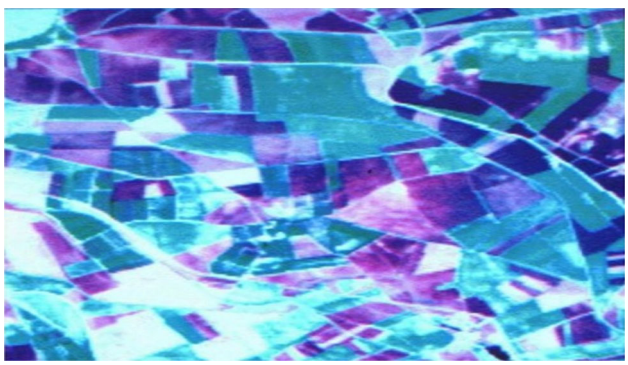

Fig. 1 198 rows and 200 columns Landsat Thematic Mapper satellite image data of agricultural area in Seyhan Plain in Adana—Turkey

New Method for Unsupervised-Clustering of RS-MID

In this section the number of partitions in heterogeneous variables for MID using variable data segmentation is determined. The number of cluster centers and structures thus, locations and orientations of clusters are determined.

The values of the 3rd, 4th, and 5th bands in the digitized data of the remotely-sensed multispectral satellite image are taken as variable values in this study. X_1, X_2, and X_3 variables are obtained from 198 $\times$ 200 = 39,600 observation values at 3rd, 4th, and 5th bands, respectively. There are 39600 $\times$ 3 = 118,800 observations in data set in total.

It is investigated whether these three variables are homogeneous or heterogeneous, i.e. whether the distributions are normal or mixture of normal distributions. Since each band in the data set represents a variable, the structures or segmentations of the variables obtained from these bands are used for the model generation (Gögebakan and Erol 2016). In homogeneous variables, clustering does not occur in the model because there is no segmentation. For all that segmentations in heterogeneous variables constitute the cluster centers of the model. Hence, homogeneous variables are eliminated and heterogeneous variables have formed the structure of models.

The elimination of homogeneous variables and the use of heterogeneous variables to construct a model correspond to the variable selection in the literature. Both the calculation method and the graphical methods are used while determining the segmentation in heterogeneous variables. By looking at the histogram and P–P graphs of heterogeneous variables, the number of segments of the variable was estimated. The segmentations of the X_1 (3.band), X_2 (4.band), and X_3 (5.band) variables in the multivariate/MID was first examined by using graphical methods on histogram and P–P graphs as illustrated in Fig. 2. As a result of this examination, it was determined that each variable was heterogeneous and each of them had 3 meaningful subgroups. In the case of three variable satellite image data, the calculation method used to determine the number of segmentation in the variables was determined using the univariate normal mixture model (Erol and Erol 2016).

Mixture of univariate normal distributions is defined as

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

where $f(x)$ shows probability density function of mixture of univariate normal distributions, k shows the number of components in the mixture of normal distributions, π_i shows weights/mixing proportions, and $f_i(x;\mu_i,\sigma_i)$ shows component probability density functions of the form

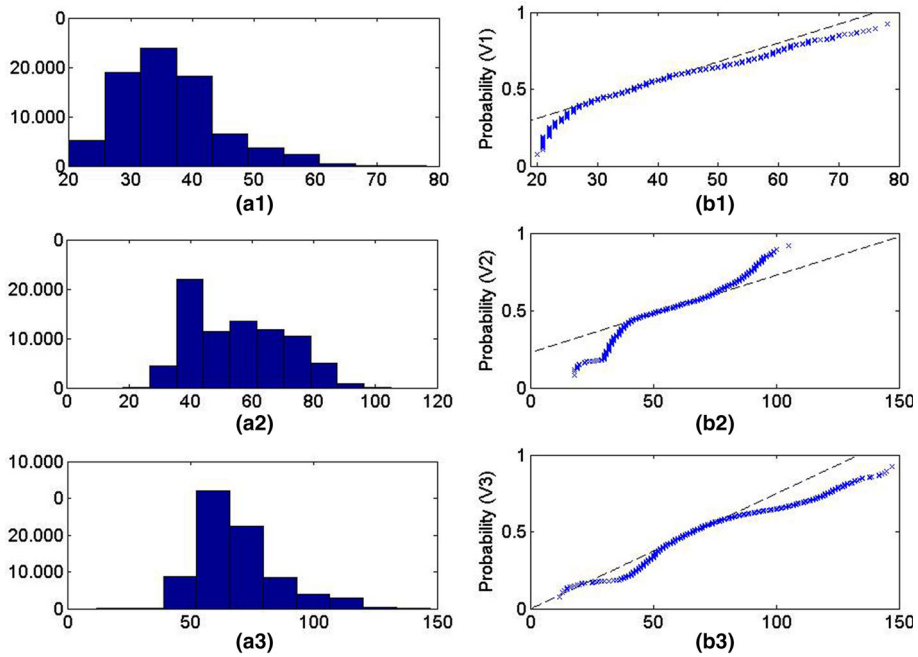

Fig. 2 The segmentations of the X_1 (3.band), X_2 (4.band), and X_3 (5.band) variables in the multiband satellite image data, **a** histogram and **b** P–P plots, respectively

$$f_i(x;\boldsymbol{\mu}_i,\sigma_i)=\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{1}{2}\left(\frac{x-\boldsymbol{\mu}_i}{\sigma_i}\right)^2\right\}, \quad (4)$$

where μ_i and σ_i shows respectively mean vector and standard deviations, respectively. In order to reveal segmentations in each variable log-likelihood function values, AIC and BIC values are examined in mixture of univariate normal distributions. The number of segmentation is obtained as the optimal number of segments, which is obtained from the univariate normal mixture model that has the largest log-likelihood value and the smallest AIC and BIC values.

The log-likelihood, AIC and BIC values calculated using estimated parameter values for mixing weights π, means μ and variances σ^2 of univariate mixture models for variables X_1, X_2 and X_3 of RS-MID to determine the appropriate number of variable data segmentation are given in Tables 1, 2 and 3 respectively. k shows the number of components in univariate mixture models.

Table 1 The log-likelihood, AIC and BIC values calculated using estimated parameter values for mixing weights π, means μ and variances σ^2 of univariate mixture models for variable X_1 of RS-MID to determine the appropriate number of variable data segmentation

K	Log-L	AIC	BIC
k = 1	− 13,947	27,895	27,897
k = 2	− 13,767	27,534	27,538
k = 3*	**− 13,567**	**27,235**	**27,242**
k = 4	− 13,617	27,236	27,246
k = 5	− 13,617	27,237	27,249

Table 2 The log-likelihood, AIC and BIC values calculated using estimated parameter values for mixing weights π, means μ and variances σ^2 of univariate mixture models for variable X_2 of RS-MID to determine the appropriate number of variable data segmentation

k	Lofi-L	AIC	BIC
k = 1	− 16,418	32,837	32,838
k = 2	− 15,912	31,825	31,829
k = 3*	**− 15,858**	**31,717**	**31,724**
k = 4	− 15,857	31,715	31,725
k = 5	− 15,857	31,716	31,728

Table 3 The log-likelihood, AIC and BIC values calculated using estimated parameter values for mixing weights π, means μ and variances σ^2 of univariate mixture models for variables X_3 of RS-MID to determine the appropriate number of variable data segmentation

k	Log-L	AIC	BIC
k = 1	− 16,664	33,329	33,331
k = 2	− 16,169	32,340	32,344
k = 3*	**− 16,126**	**32,254**	**32,261**
k-4	− 16,134	32,330	32,339
k = 5	− 16,138	32,340	32,352

According to the results of Table 1, 2 and 3, the number of suitable segmentations for the variables X_1, X_2 and X_3 is three. For the variables X_1, X_2, and X_3 of RS-MID, the mixture of univariate normal distributions yielding the appropriate number of segmentations where log-likelihood function value is largest and AIC and BIC values are smallest is obtained as $k_1 = k_2 = k_3 = 3$ for each variable as given in bold with an asterisk.

It is determined that the number of segmentation in each variable is 3, thus $k_1 = 3$, $k_2 = 3$ and $k_3 = 3$. Observations are assigned to the segments in variables with using the k-means algorithm. Segmentations of X_1, X_2, X_3 variables' are described as (X_{11}, X_{12}, X_{13}), (X_{21}, X_{22}, X_{23}) and (X_{31}, X_{32}, X_{33}) respectively. The number of observations in each segmentation of the variables is given in Table 4.

The minimum and maximum numbers of the cluster center numbers ($C_{\min}$ and $C_{\max}$) corresponding to the segments in the variables can be evaluated by the following relation developed by Servi and Erol (2007) as

$$C_{\min} = \max\{k_1, k_2, k_3\} \quad \text{and} \quad C_{\max} = \prod_{s=1}^{p} k_s \tag{5}$$

where p shows the number of variables and k_s shows the number of partitions in each variable $k_1 = k_2 = k_3 = 3$ for X_1, X_2, and X_3. $n \times 3$ data matrix for X is of the form $X = [X_1 \quad X_2 \quad X_3]$.

Partitions of X_1 variables in n_1 elements is of the form $X_1 = \begin{bmatrix} X_{11} \\ X_{12} \\ X_{13} \end{bmatrix}$ where X_{11}, X_{12}, and X_{13} partitions have n_{11}, n_{12}, and n_{13} elements, respectively. Thus, $n_1 = n_{11} + n_{12} + n_{13}$. Partitions of X_2 variables in n_2 elements is of the form $X_2 = \begin{bmatrix} X_{21} \\ X_{22} \\ X_{23} \end{bmatrix}$ where X_{21}, X_{22}, and X_{23} partitions have n_{21}, n_{22}, and n_{23} elements, respectively. Thus, $n_2 = n_{21} + n_{22} + n_{23}$. Partitions of X_3 variables in n_3 elements is of the form $X_3 = \begin{bmatrix} X_{31} \\ X_{32} \\ X_{33} \end{bmatrix}$ where X_{31}, X_{32}, and X_{33} partitions have n_{31}, n_{32}, and n_{33} elements, respectively. Thus, $n_3 = n_{31} + n_{32} + n_{33}$.

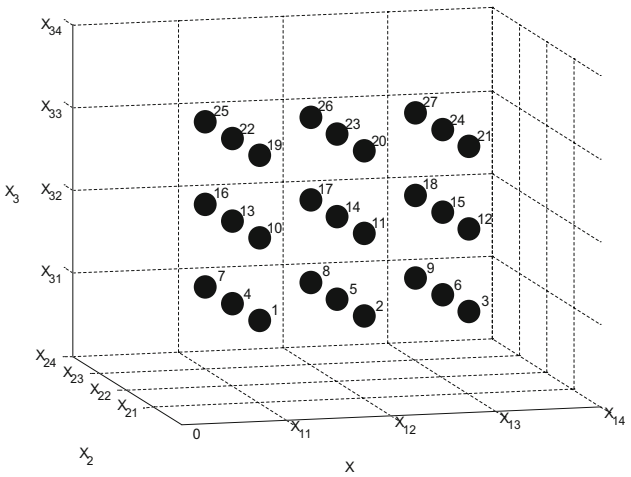

Fig. 3 The number and location of cluster centers corresponding to segments in variables in the multispectral satellite image data

For the case considered, $C_{\min} = \max\{3, 3, 3\} = 3$ and $C_{\max} = k_1 k_2 k_3 = 3.3.3 = 27$. Thus, the smallest and biggest number of cluster centers are 3 and 27 respectively. The number and location of candidate cluster centers are illustrated in Fig. 3.

Partitions of variables X_1, X_2 and X_3 of RS-MID are listed in Table 5.

Determination of the Total Number and Structure of Mixture Models

The total number of models that can be generated by the mixture models of three variable normal distributions, denoted by M_{Total}, for the cluster centers that occur in the X_1, X_2 and X_3 variables in the RS-MID:

$$M_{Total} = 2^{C_{\max}} - 1 \tag{6}$$

Table 4 The number of observations in each segmentation of the variables in satellite image data

Variable	Variable segments	Number of observations in segments	Total
X_1	X_{11}	4291	39,600
	X_{12}	17,241	
	X_{13}	18,068	
X_2	X_{21}	10,279	39,600
	X_{22}	12,697	
	X_{23}	16,624	
X_3	X_{31}	17,929	39,600
	X_{32}	4999	
	X_{33}	16,672	

Table 5 Components and number of clustering centers corresponding to segments in variables in the multispectral satellite image data

Number of cluster center	Components of clusters	Number of cluster center	Components of clusters
1.	(X_{11}, X_{21}, X_{31})	15.	(X_{13}, X_{22}, X_{32})
2.	(X_{12}, X_{21}, X_{31})	16.	(X_{11}, X_{23}, X_{32})
3.	(X_{13}, X_{21}, X_{31})	17.	(X_{12}, X_{23}, X_{32})
4.	(X_{11}, X_{22}, X_{31})	18.	(X_{13}, X_{23}, X_{32})
5.	(X_{12}, X_{22}, X_{31})	19.	(X_{11}, X_{21}, X_{33})
6.	(X_{13}, X_{22}, X_{31})	20.	(X_{12}, X_{21}, X_{33})
7.	(X_{11}, X_{23}, X_{31})	21.	(X_{13}, X_{21}, X_{33})
8.	(X_{12}, X_{23}, X_{31})	22.	(X_{11}, X_{22}, X_{33})
9.	(X_{13}, X_{23}, X_{31})	23.	(X_{12}, X_{22}, X_{33})
10.	(X_{11}, X_{21}, X_{32})	24.	(X_{13}, X_{22}, X_{33})
11.	(X_{12}, X_{21}, X_{32})	25.	(X_{11}, X_{23}, X_{33})
12.	(X_{13}, X_{21}, X_{32})	26.	(X_{12}, X_{23}, X_{33})
13.	(X_{11}, X_{22}, X_{32})	27.	(X_{13}, X_{23}, X_{33})
14.	(X_{12}, X_{22}, X_{32})		

is obtained above equation (Servi and Erol 2007). For RS-MID is calculated as: $M_{Total} = 2^{3.3.3} - 1 = 2^{27} - 1 = 134.217.727$. The subtracted model is empty that has no cluster center.

The number of total models and the corresponding cluster centers are obtained from the segmentations of variables. When normal mixture models are constructed, the number of possible candidate models are investigated in such a way that each segment in the variables correspond to at least one cluster center. These candidate models are derived from the assumption that there will be at least one cluster in each dimension over the 27 centers specified in Table 5 and Fig. 3.

It is assumed that, the cluster centers formed by the segmentation of the variables in Fig. 3 have at least 3 and at most 27 clusters obtained from the $C_{\min}$ and $C_{\max}$ equations. Models that meet the assumption are called the possible candidate model. The number of possible candidate models is calculated based on the number of variables and the number of segments in the variables. The number of models, that match the assumption that there is at least one center in each dimension (band) can be obtained by the combination calculation according to the position in the dimension in which each center is located. Since the number of variables and the number of observations in the RS-MID are big data, the calculation is obtained by the developed algorithm. The number of centers, the total number of models and the number of possible candidate models, which occur after the segmentation of the variables, are given in Table 6.

Since the algorithm developed above works with the Brute-Force method, the processing intensity increases greatly in big data. For this reason, the equation obtained by Cheballah et al. (2015), which gives the number of candidate models in 2D matrices, was developed by Gögebakan and Erol (2016) in order to calculate the number of candidate models in multi-dimensional big data with less cost:

$$\begin{aligned} f(n,m,s,k) &= \sum_{i=0}^{n} (-1)^i \binom{n}{i} \sum_{j=0}^{m} (-1)^j \binom{m}{j} \\ &\quad \sum_{t=0}^{s} (-1)^t \binom{s}{t} \binom{(n-i)(m-j)(s-t)}{k} \\ &= \sum_{i,j,t=0}^{n,m,s} (-1)^{i+j+t} \binom{n}{i} \binom{m}{j} \binom{s}{t} \\ &\quad \binom{(n-i) \quad (m-j) \quad s-t}{k} \end{aligned} \tag{7}$$

where n, m and s are the segment numbers in variables X_1, X_2 and X_3 respectively. Indices i, j and t shows the number of clusters. k is the cases for the number of cluster centers in mixture models.

Generating Candidate Normal Mixture Models Based on Segmentation of Variables

In multispectral satellite image data, candidate normal mixture models, mixing weights, mean vectors and covariance matrices, which are composed of segmentations in variables, are estimated on the basis of sampling (Erol 2013). The obtained parameters are used to calculate the mixture of probability density functions of the possible candidate models in Table 6.

In Table 6, string representations consisting of 0 and/or 1, and 27 characters are made using center numbers for the possible each candidate models and in Table 7, full model representation has been given. In these string representations, "1" is written in response to the center forming the

Table 6 The number of centers, the total number of models and the number of possible candidate models, which occur after the segmentation of the variables

Number of cluster centers	Number of total models	Number of candidate models
1	27	0
2	351	0
3	2925	36
4	17,550	1890
5	80,730	24,300
6	296,010	153,828
7	888,030	623,106
8	2,220,075	1,839,672
9	4,686,825	4,255,194
10	8,436,285	8,044,245
11	13,037,895	12,751,803
12	17,383,860	17,216,811
13	20,058,300	19,981,143
14	20,058,300	20,030,760
15	17,383,860	17,376,516
16	13,037,895	13,036,518
18	8,436,285	8,436,123
18	4,686,825	4,686,816
19	2,220,075	2,220,075
20	888,030	888,030
21	296,010	296,010
22	80,730	80,730
23	17,550	17,550
24	2925	2925
25	351	351
26	27	27
27	1	1
Total	134,217,728	131,964,460

model and "0" is written in the case of not forming. String representation developed for the possible models, allows to make calculations in the models.

Each binary string representation of candidate mixture model fits to one of 41,503 possible models. Mixture model with k elements, binary strings of them are given below as:

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{k} \pi_i^{(u)} f_i(x;\mu_i^{(u)},\Sigma_i^{(u)}) \quad \text{for } u = 1,\ldots,131964460 \tag{8}$$

where mixing proportions for element normal density function is in the form

$$\pi_i^{(u)} = \frac{\pi_i}{\sum_{l=1}^{k} \pi_l} \quad \text{for } u = 1,\ldots,131964460 \tag{9}$$

mean vector for element normal density function is in the form:

$$\mu_i^{(u)} = \begin{bmatrix} \mu_{1p}^{(u)} \\ \mu_{2q}^{(u)} \\ \mu_{3r}^{(u)} \end{bmatrix} \quad \text{for } u = 1,\ldots,131964460 \text{ and } p,q,r = 1,2,3 \tag{10}$$

covariance matrices for element normal density function is in the form:

$$\Sigma_i^{(u)} = \begin{bmatrix} (\sigma_{1p}^{(u)})^2 & \rho_{1p,2q}^{(u)}\sigma_{1p}^{(u)}\sigma_{2q}^{(u)} & \rho_{1p,3r}^{(u)}\sigma_{1p}^{(u)}\sigma_{3r}^{(u)} \\ \rho_{2q,1p}^{(u)}\sigma_{2q}^{(u)}\sigma_{1p}^{(u)} & (\sigma_{2q}^{(u)})^2 & \rho_{2q,3r}^{(u)}\sigma_{2q}^{(u)}\sigma_{3r}^{(u)} \\ \rho_{3r,1p}^{(u)}\sigma_{3r}^{(u)}\sigma_{1p}^{(u)} & \rho_{3r,2q}^{(u)}\sigma_{3r}^{(u)}\sigma_{2q}^{(u)} & (\sigma_{3r}^{(u)})^2 \end{bmatrix} \quad \text{for } u = 1,\ldots,131964460 \tag{11}$$

In the multispectral satellite image data, the mixing weights, mean vectors and covariance matrices for the

Table 7 String representations consisting of 0 and/or 1, and 27 characters are made using center numbers for the possible each candidate models

1	2	3	4	5	6	7	8	9
1	1	1	1	1	1	1	1	1
10	11	12	13	14	15	16	17	18
1	1	1	1	1	1	1	1	1
19	20	21	22	23	24	25	26	27
1	1	1	1	1	1	1	1	1

cluster centers corresponding to the segments in the variables are estimated from the sample. Information complexity is less than other clustering methods since EM algorithm is not preferred. For $i = 1,2, \ldots, 27$ mixing weights, mean vectors and covariance matrices are respectively: $\widehat{\pi}_i^{(u)} = \frac{n_i}{\sum_{l=1}^{k} n_l}$ for $u = 1, \ldots, 131964460$ and $l = 1, 2 \ldots, k$,

$$\widehat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_{1p}^{(u)} \\ \bar{x}_{2q}^{(u)} \\ \bar{x}_{3r}^{(u)} \end{bmatrix}$$

and

$$\widehat{\Sigma}_i^{(u)} = \begin{bmatrix} (S_{1p}^{(u)})^2 & r_{1p,2q}^{(u)} S_{1p}^{(u)} S_{2q}^{(u)} & r_{1p,3r}^{(u)} S_{1p}^{(u)} S_{3r}^{(u)} \\ r_{2q,1p}^{(u)} S_{2q}^{(u)} S_{1p}^{(u)} & (S_{2q}^{(u)})^2 & r_{2q,3r}^{(u)} S_{2q}^{(u)} S_{3r}^{(u)} \\ r_{3r,1p}^{(u)} S_{3r}^{(u)} S_{1p}^{(u)} & r_{3r,2q}^{(u)} S_{3r}^{(u)} S_{2q}^{(u)} & (S_{3r}^{(u)})^2 \end{bmatrix}.$$

Where, for $u = 1, \ldots, 131964460$, $l = 1, 2 \ldots, k$ and $p, q, r = 1, 2, 3$ segmentations. $\rho_i = \boldsymbol{Corr}(\boldsymbol{X}_{1p}, \boldsymbol{X}_{2q}, \boldsymbol{X}_{3r})$ shows Pearson correlation coefficients. Each cluster center has, $N(x; \mu_i, \Sigma_i)$ normal distribution. When normal mixture models are generated with cluster centers corresponding to the variable segmentations in three variable data, mixing weights, mean vectors, and covariance matrices are used. The Pearson correlation coefficient is used to determine the direction and degree of the relationship between the components of the covariance matrices of the cluster centers corresponding to the segments in the variables. For the covariance matrices of the symmetric structure of type 3×3 corresponding to the segmentation of each variable in the data set, the correlation developed by Gögebakan and Erol (2016) and the $k_1k_2 + k_1k_3 + k_2k_3 = 3.3 + 3.3 + 3.3 = 27$ different correlation coefficients are obtained.

The New Semi-supervised Classification Method

The new semi-supervised classification method consists of unsupervised clustering of RS-MID and then supervised classification of clusters in RS-MID using spectral signatures. To determine the true mixture model based clustering of RS-MID; the log-likelihood function, AIC and BIC values for candidate normal mixture models for suitable states based on variable data segmentation are computed for each candidate mixture model.

The log-likelihood function values are calculated as the first criterion in order to determine the best clustering structure of the heterogeneous dataset using mixture models of Gaussian distributions. The values of the log-likelihood functions for the candidate mixture models were used as a criterion for selecting the best mixture model. Likelihood function for the mixture of normal densities is defined as

$$L(\Psi) = \prod_{i=1}^{n} f(x_i; \Psi) = \prod_{i=1}^{n} \left[\sum_{j=1}^{k} f(x_i; \theta_j) \right], \tag{12}$$

$f(x_i; \theta_j)$ shows mixture probability density functions. Log-likelihood function for the mixture of normal densities is computed as follows

$$\log L(\Psi) = \sum_{j=1}^{n} \log(f(x_i; \Psi)) = \sum_{j=1}^{n} \log \left(\sum_{i=1}^{k} \pi_i f_i(x_j; \theta_i) \right) \quad i = 1, 2, \ldots n \text{ and } j = 1, 2, \ldots k \tag{13}$$

Log-likelihood function values for possible mixture of normal densities are evaluated by the estimated values of $\widehat{\pi}_i^{(u)}$, $\widehat{\mu}_i^{(u)}$ and $\widehat{\Sigma}_i^{(u)}$.

Akaike's information criterion (AIC) (Akaike 1974) can be obtained by:

$$AIC = -2 \ln L(\Psi) + 2d \tag{14}$$

Bayesian information criterion (BIC) (Schwarz 1978) can be obtained by:

$$BIC = -2 \ln L(\Psi) + d \log n, \tag{15}$$

where $\ln L(\Psi)$ is the value of log-likelihood function for possible mixture of normal densities, d is the number of free parameters in possible mixture of normal densities and n is the number of observation. The number of free parameters in possible mixture of normal densities d can be computed by:

Table 8 Log-l, AIC and BIC values of the best model for model-based clustering of normal mixture distributions in multispectral satellite image data

Log-l	− 105,752,585,4
AIC	211,505,668,8433
BIC	211,507,806,9028
String representation of best model	1101111111111101111111111111

Table 9 The contingency table or classification error matrix

	W	P	VG	C	BS	RT	UA
W	12,989	235	33	0	273	13,530	0.96
P	502	3724	79	24	312	4641	0.80
VG	0	27	11,194	0	86	11,307	0.99
C	188	0	0	1379	0	1567	0.88
BS	191	0	151	0	8213	8555	0.96
CT	13,870	3986	11,457	1403	8884	39,600	
PA	0.94	0.93	0.98	0.98	0.92		0.95

W, wheat; P, potato; VG, vegetable garden; C, citrus; BS, bare soil; RT, row total; CT, column total; UA, user's accuracy; PA, producer's accuracy

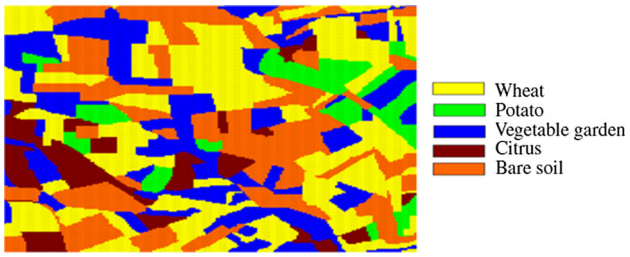

Fig. 4 Categorized or semi-supervised classified (false color) map of the remotely sensed multispectral satellite image data

$$d = (K-1) + (Kp) + \left(Kp\frac{(p+1)}{2}\right) \tag{16}$$

where k is the number of components, p is the number of variables or dimension in mixture model (Bozdogan 1984). Log-likelihood, Akaike and Bayesian information criteria values are obtained from partitions of variables using mean vectors and covariance matrices. Log-likelihood function, AIC and BIC values will be used as criteria for selecting the best mixture model of normal mixtures densities. All calculations are performed using MATLAB software.

For model-based clustering, the choice of the best model among the Gaussian mixture models was derived from Log-l, AIC and BIC values of the models. Log-I, AIC and BIC values of candidate models from normal mixture models in multispectral satellite image data are calculated from 131,964,460 possible candidate models with the help of algorithm developed for model based clustering and are given in Table 8. Among the possible candidate models in the model-based clustering, the best model Log-l value is the largest, and the AIC and BIC values are the smallest. The best model, the log-likelihood, AIC and BIC values of all possible candidate models were calculated, with the best model among the twenty-five centered models ranked as the 60th "1101111111111011111111111111" string representation model.

Clusters in data are classified using spectral signatures in the study by Çalış and Erol (2013). Euclidean distance (Messinger et al. 2012) is used as the discrimination function for the supervised classification method. The values of Euclidean distances are used as decision rule for the supervised classification method. The contingency table or classification error matrix is given in Table 9.

The overall accuracy is 0.95 from Table 9. The estimate of Kappa statistics (Congalton 1991) is obtained as 0.93 for the new semi-supervised classification method based on mixture model clustering for classification of RS-MID.

Conclusion

By using variable data segmentation, the structure of clustering in the RS-MID is obtained from mixture model clustering. The clusters are classified using supervised-classification method. In the supervised-classification of clusters in the RS-MID, spectral signatures of control classes worked by Çalış and Erol (2013) were used. A classification overall accuracy of 95% was obtained as a result of the new semi-supervised classification of clusters. The estimate of Kappa statistics (Congalton 1991) were obtained as 0.93 for semi-supervised classification method based on mixture model clustering for classification of RS-MID. Figure 4 shows the categorized or semi-supervised classified (false color) map of the RS-MID. The new semi-supervised classification resulting from the clustering, and the categories represented by each color shown in Fig. 4.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Bozdogan, H. (1984). *Multi-sample cluster analysis as an alternative to multiple comparison procedures (No. Uic/Dqm/A84-3)*. Illinois University at Chicago Circle Department of Quantitative Methods.

Çalış, N., & Erol, H. (2013). A new per-field classification method using mixture discriminant analysis. *Journal of Applied Statistics, 39*(10), 1–12. https://doi.org/10.1080/02664763.2012.702263.

Cheballah, H., Giraudo, S., & Maurice, R. (2015). Hopf algebra structure on packed square matrices. *Journal of Combinatorial Theory, Series A, 133,* 139–182.

Congalton, Russell G. (1991). A review of assessing the accuracy of classification of remotely sensed data. *Remote Sensing of Environment, 37,* 35–46.

Erol, H. (2013). A model selection algorithm for mixture model clustering of heterogeneous multivariate data. In *Innovations in intelligent systems and applications. 2013 IEEE international symposium on innovations in intelligent systems and applications, At Albena, Bulgaria*. (pp. 1–7). https://doi.org/10.1109/inista.2013.6577617.

Erol, H., & Akdeniz, F. (2005). A per-field classification method based on mixture distribution models and an application to Landsat Thematic Mapper data. *International Journal of Remote Sensing, 26*(6), 1229–1244.

Erol, H., & Erol, R. (2016). Logical circuit design using orientations of clusters in multivariate data for decision making predictions: A data mining and artificial intelligence algorithm approach. In *International symposium on innovations in intelligent systems and applications 2–5 August 2016, At Sinaia, Romania* (Vol. 1).

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal, 41,* 578–588.

Fraley, C., Raftery, A. E., & Scrucca, L. (2012). *Normal mixture modeling for model-based clustering, classification, and density estimation*. Department of Statistics, University of Washington. http://cran.r-project.org/web/packages/mclust/index.html. Accessed September 23, 2012.

Gögebakan, M., & Erol, H. (2016). A new approach for mixture model clustering based on selecting the best mixture model among candidate mixture models. In *International conference on information complexity and statistical modeling in high dimensions with applications, At Nevşehir Turkey* (Vol. 1).

Gögebakan, M., & Erol, H. (2016). Mixture model clustering using variable data segmentation. In *Conference: 12th German probability and statistics days, At Bochum, Germany*.

McLachlan, G. J., & Chang, S. U. (2004). Mixture modelling for cluster analysis. *Statistical Methods in Medical Research, 13,* 347–361.

Messinger, D. W., Ziemann, A., Basener, B., & Schlamm, A. (2012). Metrics of spectral image complexity with application to large area search. *Optical Engineering, 51*(3), 036201.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Servi, T., & 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, 8*(2), 57–69.