

RESEARCH ARTICLE

A Novel Approach to Increase the Efficiency of Filter-Based Feature Selection Methods in High-Dimensional Datasets With Strong Correlation Structure

SERKAN AKOGUL 

Department of Statistics, Faculty of Science, Pamukkale University, 20160 Denizli, Turkey

e-mail: sakogul@pau.edu.tr


ABSTRACT Nowadays, data dimensions have increased depending on the developments in information and measurement technologies. Due to the high dimensionality, it is necessary to use pre-analysis data reduction methods for many analyzes such as classification and regression analysis. In the solution of high-dimensionality, filter feature selection methods based on statistical criteria are widely used in terms of simplicity and efficiency. One of the important problems with filter feature selection methods is the selection of multiple features carrying the same information unnecessarily when strong correlations exist between features. In this study, a novel approach is proposed to solve this problem of filter feature selection methods. In addition, with the proposed new approach, the question of how many appropriate features will be included is also solved. The performance of the proposed approach is demonstrated on high-dimensional reflectance data with high correlations between features. The results obtained revealed that the proposed approach improves the classification performance of filter feature selection methods in mixture discriminant analysis in terms of classification accuracy and entropy criteria.

INDEX TERMS Feature selection, filter feature selection, Gaussian mixture model (GMM), Gaussian mixture discriminant analysis (GMDA).

I. INTRODUCTION

Technological developments in recent years have enabled the creation of large databases in many fields, and the amount of data stored has further increased. This situation has led to the emergence of the term high-dimensional data or big data. Many data mining methods have been developed because of the inadequacy of traditional methods in analyzing high-dimensional data.

Data mining [1] can be defined as a multi-stage process that aims to reveal the relationships, patterns and information hidden in high-dimensional data using different tools and technologies. One of the stages of this process is the feature selection process.

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson .

Feature selection methods involve selecting a subset of the most useful features that produce results compatible with the entire original dataset [2], [3]. The methods are generally categorized into three groups: filter methods based only on statistical information, wrapper methods that perform searches on features, and embedded methods based on finding the best divisor criterion [4], [5], [6].

Among the feature selection methods, filter feature selection methods are widely used in the literature because of their ease of calculation and speed. Filter feature selection methods make feature selection with the help of functions based on statistical criteria such as distance, information, dependency, and consistency measurements [6], [7]. One of the important problems with these methods is the selection of multiple features carrying the same information when strong correlations exist between features unnecessarily. For

example, high-dimensional and strongly correlated data, such as reflectance, image, text or DNA microarray data, can hinder the learning process, especially in classification [6], [8], [9], [10], [11].

The focus of this study is to determine the features that will give the highest classification accuracy in the classification of high-dimensional data. When the literature on this subject is examined, there are many studies under general titles such as feature selection for classification [2], [5], [12], [13], [14], [15], [16], [17], feature selection for high-dimensional data [6], [18], [19], [20], [21], [22], [23], [24], [25], [26], and feature selection for dimensionality reduction [27], [28], [29], [30], [31], [32], [33], [34], [35].

In this study, a novel approach is proposed to increase the efficiency of filter feature selection methods based on the clustering of features using gaussian mixture models. The proposed approach aims to select the most appropriate feature by bringing together features with similar characteristics with clustering of features, especially in data with high-dimensional and strong correlation structure. Thus, it is desirable to avoid the problem of unnecessarily selecting multiple features carrying the same information. In addition, with the proposed new approach, the question of how many appropriate features will be included is also solved, and an algorithm for calculating the initial cluster centers has been proposed to improve the performance of the approach. The performance of the proposed approach is demonstrated on high-dimensional reflectance data with high correlations between features. The results obtained from the study revealed that the proposed approach improves the classification performance of filter feature selection methods in mixture discriminant analysis in terms of classification accuracy and entropy criteria.

The paper is organized as follows. In Section II, we present the filter feature selection methods and the proposed feature selection approach. In Section III, we not only introduce the dataset but also present the model-based clustering analysis, model-based discriminant analysis, and performance criteria. In addition, we report the experimental results to support the proposed approach and provide a comparison table with the filter-based feature selection methods. Finally, in Section IV, we summarize this study and draw some conclusions.

II. MATERIALS AND METHODS

A. FILTER FEATURE SELECTION METHODS

Feature selection can be defined as the process of identifying relevant features and discarding irrelevant ones in order to obtain a subset of features that will best represent the data with minimal performance degradation [2], [3], [5].

Filter feature selection method, which is one of the feature selection methods, is widely used in the literature because of its ease of calculation and speed. Filter feature selection methods make feature selection with the help of statistical criteria such as distance, information, dependency and consistency measurements. Basically, these methods calculate a

score for each feature found in the dataset through a function calculated according to the statistical criterion determined. With this calculated score, the features are ranked in order of importance, and subsets are created by selecting the desired number of features. These methods, which are less complex and have lower computational cost than other methods, are more suitable for high-dimensional data because they give fast results [4], [6], [7].

In this section, the most popular filters are described, which will be used throughout this paper.

1) CHI-SQUARED

This method is a univariate filter that evaluates each feature based on chi-square (χ^2) statistics independently by class. Given the number of intervals (V), the number of classes (K), and the total number of instances (N); the value of chi-squared for a feature is calculated as

$$\chi^2 = \sum_{i=1}^V \sum_{j=1}^K \frac{\left[A_{ij} - \frac{R_i * B_j}{N} \right]^2}{\frac{R_i * B_j}{N}} \quad (1)$$

where R_i denotes the number of instances in the range i th, B_j stands for the number of instances in class j th, and A_{ij} indicates the number of instances in the range i th and class j th [6], [36].

2) INFORMATION GAIN

This method [37] evaluates features according to their information gain and considers a single feature at a time. The entropy measure is considered as a measure to rank variables. The entropy of class feature Y is

$$H(Y) = - \sum p(y) \log_2 p(y) \quad (2)$$

where $p(y)$ is the marginal probability density function for the random variable Y . When calculating the entropy value, if the values in the Y property are grouped according to the X property, the entropy value of Y will be higher than the entropy value of the data grouped according to X . According to the observation of the X property, the entropy value of the Y property is calculated as follows

$$H(Y|X) = \sum p(x) \sum p(y|x) \log_2 p(y|x) \quad (3)$$

where $p(y|x)$ is the conditional probability of y given x . The measure of information gain (IG), which is an indicator of the dependence between X and Y , is calculated as

$$IG = H(Y) - H(Y|X) \quad (4)$$

3) GAIN RATIO

The gain ratio (GR) method [38] is a normalized version of the IG . Normalization is performed by dividing the IG by the entropy of the quality by class; consequently, it reduces the bias of the IG algorithm. The GR formula is as follows:

$$GR = \frac{IG}{H(X)} \quad (5)$$

The *GR* takes values between 0 and 1. When *GR* is equal to 1, it indicates that the information *X* can predict all information *Y*, and when it is equal to 0, there is no relationship between *Y* and *X*.

4) SYMMETRICAL UNCERTAINTY

The Symmetrical Uncertainty (*SU*) [39] compensates for the bias of the *IG* by dividing it by the sum of the entropy of *X* and *Y*. The *SU* is calculated as

$$SU = \frac{2 * IG}{H(X) + H(Y)} \quad (6)$$

where $H(X)$ and $H(Y)$ are the entropy of *X* and *Y*. The *SU* takes values between 0 and 1. The *SU* is interpreted similarly to the *GR*.

5) RELIEF

The main purpose of the Relief algorithm [40] is to predict the quality of the features according to how well their values are distinguished between samples that are close to each other. The algorithm searches for two nearest neighbors, with a randomly selected number of samples *R*. The first of these is defined as *H*, which refers to the closest neighbor from the same class, and *M*, which refers to the closest neighbor from a different class. The estimated $W[A]$ weighting coefficient for the *A* feature based on *R*, *M* and *H* values are updated and run *t* times to find the weight coefficients. The $W[A]$ formula is as follows

$$W_{new}[A] = W_{old}[A] - \frac{diff(A, R, H)}{t} + \frac{diff(A, R, M)}{t} \quad (7)$$

6) CORRELATION-BASED FILTER

The algorithm finds weights of continuous attributes based on their correlation with the continuous class attribute. In general, a feature is good if it is relevant to the class concept but is not redundant to any of the other relevant features. According to whether the relationship is parametric or nonparametric, the relationship scores of the features are calculated using Pearson or Spearman correlation coefficients. The attributes with a high absolute value relationship score are ranked in order of importance [41], [42].

7) RANDOM FOREST

The random forests (RF) method uses a collection of decision tree classifiers, where each tree in the forest has been trained using a bootstrap sample of individuals from the data, and each split attribute in the tree is chosen from among a random subset of attributes. The classification of individuals is based on aggregate voting over all trees in the forest. Feature selection using the RF method starts with determining the threshold value to give a boundary between the features to be selected and the features that will be eliminated. Then all features will be sorted by Gini importance score from the smallest to the largest. Furthermore, features with the Gini importance score below the threshold value will be eliminated [43], [44].

8) ONE-WAY ANOVA

The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of independent $k > 2$ groups. The larger *F* test statistics value in one-way ANOVA indicates a statistically significant difference between groups. The *F* test statistic is obtained by

$$F = \frac{(n - k) \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2}{(k - 1) \sum_{i=1}^k \sum_{j=1}^{n_j} (x_{ij} - \bar{x}_i)^2} \quad (8)$$

where *n* is total sample size, *k* is number of groups, \bar{x}_i is the *i*th group mean for the relevant feature, \bar{x} is the general average. Dimension reduction is performed by selecting those with higher *F* test statistic values among the *F* test statistic values obtained for each feature [45].

9) NEIGHBORHOOD COMPONENT ANALYSIS

Neighborhood component analysis (NCA) [46] is a non-parametric feature selection algorithm that maximizes the correct classification probability of classification algorithms. In neighborhood component analysis, features are weighted to maximize the probability of correct classification. Feature selection is performed using the weights obtained.

p_{ij} is defined as the probability that data point x_i chooses x_j as its neighbor and can be calculated as follows.

$$p_{ij} = \frac{\exp(-d_{Aij}^2)}{\sum_{k \in N_i} \exp(-d_{Aik}^2)} \quad (9)$$

where N_i denotes the set of neighbors of x_i . The purpose of NCA is to learn a linear transform *A* that maximizes log probability by selecting each data point as a neighbor with the same labels as itself after the transformation [21].

B. THE PROPOSED APPROACH

The biggest problem in filtering-based feature selection methods is the unnecessary selection of features that carry the same information in the case of high correlation between features. Another important problem of the filter-based feature selection method is that the number of features to be selected is unknown. The framework of the classical feature selection methods is shown in Figure 1.

Figure 1 indicates that correlated features are selected when the filter-based feature selection method is applied for many related features. In this study, a novel algorithm is proposed to increase the efficiency of filter-based feature selection methods by focusing on these two problems. The basis of the proposed algorithm is based on the clustering of features. The features are clustered by mixture cluster analysis based on the gaussian mixture model (GMM) in the proposed algorithm. In the clustering with the GMM of features, the appropriate number of clusters is determined using the Bayesian information criterion (BIC). The best feature is determined by applying the feature selection method to

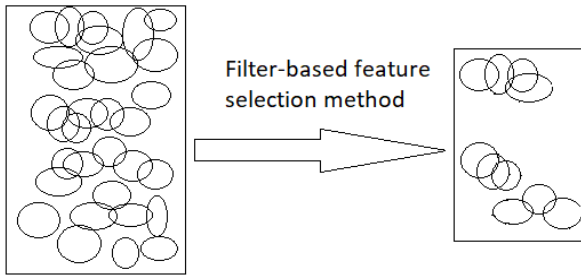


FIGURE 1. Framework of classical filter-based feature selection methods.

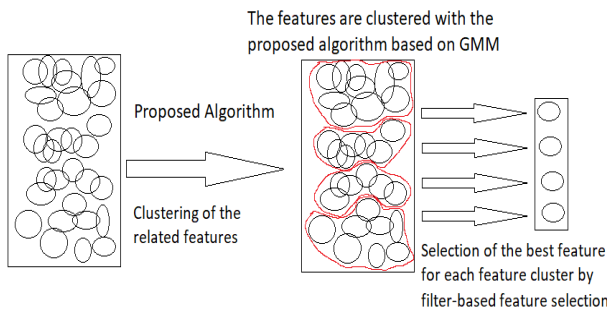


FIGURE 2. Use of the filter-based feature selection method with the proposed algorithm.

each of the created feature clusters. The reduced dataset is created by using the class means for each feature to increase the efficiency of the proposed algorithm in clustering the features. In the proposed algorithm, the effect of randomness is avoided by creating an algorithm based on the range for the initial cluster memberships in the mixture cluster analysis. The use of the filter-based feature selection method with the proposed algorithm is shown in Figure 2.

The items of the proposed algorithm are as follows:

III. EXPERIMENTAL RESULTS

A. REAL DATASETS

In this study, data containing spectrometric measurements of the critical nitrogen nutrient content of peach leaves were provided from the literature [10]. The data includes the training ($N_1 = 84$) and test ($N_2 = 96$) data with $K = 3$ class and $P = 601$ feature. The features measured for each observation were hyperspectral wavelength reflections of 400-1000 nm. The peach trees were divided into three classified as deficient, sufficient and excess, according to the amount of nitrogen nutrient, which play an important role in the development of the plant, applied to the plant.

B. MODEL-BASED CLUSTERING ANALYSIS

Model-based clustering assumes that each observation emerges from a finite mixture of G probability distributions, each representing a different set or group. The general form

Algorithm The Proposed Algorithm

Input: $X_{N \times P} \rightarrow D(F_1, F_2, \dots, F_P, C)$ # The dataset with N observations, P features and K classes.

Output: $S_{N \times G}$ # Selected G features.

Part 1: Processing the dataset for clustering of features

- 1) $m_{1 \times P}^k$ # Calculate the vectors of cluster centers in terms of the feature P of the k th class.

$$2) M_{K \times P} = \begin{bmatrix} m_{1 \times P}^1 \\ m_{1 \times P}^2 \\ \vdots \\ m_{1 \times P}^K \end{bmatrix} \# \text{ Reduced dataset consisting of } K \text{ class center vectors.}$$

- 3) $Y_{P \times K} = (M_{K \times P})^T$ # Transformed dataset, which is the transpose of matrix M .

Part 2: A new algorithm for initial cluster memberships in clustering

- Minimum (*Min*) and maximum (*Max*) values are determined for each of the K dimensions.
- The *Range* = (*Max* – *Min*) is calculated.
- The $1 \times K$ vector consisting of minimum values is chosen as the 1th cluster center.
- The increment amount for each dimension is determined as $H = \frac{\text{Range}}{g-1}$; $g = 2, 3, \dots$
- The origin centers for the g cluster are determined by adding H to the 1th cluster center.
- Initial cluster memberships are created by the k -means clustering using initial cluster centers.

Part 3: Features are clustered with the GMM

- 4) The $Y_{P \times K}$ transformed dataset is modeled for $g = 2, 3, \dots$, with the GMM according to the initial cluster centers obtained by proposed algorithm.
- 5) The appropriate number of clusters G is determined by the BIC.
- 6) The features are divided into G clusters and subsets are determined as $X_{N \times P} = [X_1 X_2 \dots X_g \dots X_G]$ where X_g of size $N \times R_i$ ($\sum_{i=1}^G R_i = P$) are subsets of X dataset.

Part 4: The best feature is determined for each feature cluster

- 7) The best feature for each feature cluster is determined with respect to the related feature selection method. The reduced dataset is determined by combining the best features selected for each feature set. The reduced dataset is given as

$$S_{N \times G} = [S_1 S_2 \dots S_g \dots S_G]$$

where S_g is $N \times 1$ dimensional vector consisting of the best feature selected from X_g .

of the finite mixture distribution is specified as follows

$$f(x_i; \psi) = \sum_{g=1}^G \pi_g f_g(x_i, \theta_g) \tag{10}$$

where the data $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ in \mathbf{R}^{np} , the π_g are the mixing probabilities ($0 < \pi_g < 1$ and $\sum_{g=1}^G \pi_g = 1$), θ_g is the parameter set corresponding to component g and $\psi = (\pi_1, \dots, \pi_G, \theta_1, \dots, \theta_G)$ denotes the set of all parameters of the mixture. The Gaussian mixture model (GMM), is obtained by taking $f_g(\mathbf{x}_i, \theta_g)$ as a multivariate Gaussian density in Equation (10). The component densities fully characterize the group structure of the data, and each observation belongs to the corresponding cluster according to a latent cluster membership indicator variable z_i , such that $z_{ig} = 1$ if \mathbf{x}_i arises from the g th subpopulation [47], [48].

Maximum likelihood estimations of parameters are usually estimated using the expectation-maximization (EM) algorithm [49], [50]. The EM algorithm maximizes the conditional log-likelihood

$$\log L(\psi) = \sum_{g=1}^G \sum_{i=1}^n z_{ig} \{ \log \pi_g + \log f_g(\mathbf{x}_i, \theta_g) \} \quad (11)$$

After the parameters have been estimated, each observation is assigned to g th cluster, which has the highest posterior probability. Posterior probabilities $\tau_{ig} = Pr(z_{ig} = 1 | \mathbf{x}_i)$ of observing cluster g given the data point i , and the mixing probabilities are estimated as follows [51]:

$$\hat{\tau}_{ig} = \frac{\pi_g f_g(\mathbf{x}_i, \theta_g)}{f(\mathbf{x}_i; \psi)} \quad (12)$$

$$\hat{\pi}_g = \sum_{i=1}^n \frac{\hat{\tau}_{ig}}{n} \quad (13)$$

In determining the number of components G , information criteria based on the log-likelihood function are generally used, and the Bayes Information Criteria (BIC) [52] is the most popular. The BIC is given following

$$\text{BIC} = -2 \log L(\hat{\psi}) + d \log(n) \quad (14)$$

where \log is the loglikelihood function and d is the number of free parameters. The appropriate number of clusters is the first g value at which the $\text{BIC}(g) \leq \text{BIC}(g+1)$ inequality is satisfied.

C. MODEL-BASED DISCRIMINANT ANALYSIS

Suppose we have K classes in the data and the number of subclasses in each class is $G_k, k = 1, \dots, K$. The probability density function in the mixture discriminant model based on the multivariate Gaussian mixture model (Gaussian mixture discriminant analysis - GMDA)

$$f(\mathbf{x}, \psi) = \sum_{k=1}^K \pi_k f_k(\mathbf{x}, \theta_k) \quad (15)$$

here \mathbf{x} is $1 \times p$ dimensional observation vector and π_k is prior probability of k th class ($0 < \pi_k < 1$ and $\sum_{k=1}^K \pi_k = 1$). $\psi = (\pi_1, \dots, \pi_K, \theta_1, \dots, \theta_K)$ is the vector containing all unknown parameters of the mixture discriminant model based

on the GMM. $f_k(\mathbf{x}, \theta_k)$ is probability density function of the GMM for k th class and $f_k(\mathbf{x}, \theta_k)$ is defined

$$f_k(\mathbf{x}, \theta_k) = \sum_{g=1}^{G_k} \pi_{kg} \frac{1}{(2\pi)^{P/2} |\Sigma_{kg}|^{1/2}} \times \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_{kg}) \Sigma_{kg}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{kg})^T \right\} \quad (16)$$

where $\theta_k = (\pi_{k1}, \dots, \pi_{kG_k}, \boldsymbol{\mu}_{k1}, \dots, \boldsymbol{\mu}_{kG_k}, \Sigma_{k1}, \dots, \Sigma_{kG_k})$ is the vector containing unknown parameters of the GMM for k th class. Where π_{kg} is mixture rate of g th subclasses in the k th class ($0 < \pi_{kg} < 1$ and $\sum_{g=1}^{G_k} \pi_{kg} = 1$). In the k th class, $\boldsymbol{\mu}_{kg}$ and Σ_{kg} are denoted mixture rate, mean vector and covariance matrix of g th subclass, respectively [47], [53], [54].

The prior probability π_k can be estimated from the training data or other sources [55]. Estimation of the prior probability π_k from the training data is defined by

$$\hat{\pi}_k = \frac{n_k}{n} \quad (17)$$

where n_k is the number of observations for the k th class of the training data. The maximum likelihood estimations of parameters $\pi_{kg}, \boldsymbol{\mu}_{kg}$ and Σ_{kg} used in Equation (16) for k th class can be estimated from training data using the EM algorithm [49], [50]. The EM algorithm maximizes the conditional log-likelihood, $\log L(\theta_k)$, for the k th class of the training data

$$\log L(\theta_k) = \sum_{g=1}^{G_k} \sum_{i=1}^{n_k} P_k(G = g | \mathbf{X}_k = \mathbf{x}_{ki}) \times \{ \log \pi_{kg} + \log f_{kg}(\mathbf{x}_{ki}, \theta_{kg}) \} \quad (18)$$

where \mathbf{x}_{ki} is $1 \times P$ dimensional observation vector of i th in the k th class of the training data. An observation is classified into k th class, which has the highest posterior probability, based on the Bayes rule, denoted by

$$P(K = k | \mathbf{X} = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x}, \theta_k)}{f(\mathbf{x}, \psi)}. \quad (19)$$

D. PERFORMANCE CRITERIA

The efficiency of the proposed algorithm will be evaluated according to the classification accuracy and entropy criteria in the mixture discriminant analysis.

Classification accuracy (CA), which is one of the most widely used criteria in the measurement of classification performance, can be stated as:

$$\text{CA} = \frac{\sum_{g=1}^G f_{gg}}{N} \quad (20)$$

where N is the number of observations, G is the number of classes and f_{gg} is the number of units assigned to class g due to the classification analysis whose actual class membership is g . If CA is close to 1, it indicates that the classification performance is successful.

TABLE 1. CA and entropy values according to the classical and proposed approach.

Methods	Selected wavelengths (features)							CA		Entropy		
	S ₁	S ₂	S ₃	S ₄	S ₅	S ₆	S ₇	Training	Test	Training	Test	
The proposed approach	F	21	237	294	314	327	486	590	0.976	0.823	9.675	9.769
	IG	12	116	295	325	338	372	523	0.929	0.760	12.001	15.312
	GR	1	105	284	314	327	361	509	0.976	0.792	2.747	4.090
	SU	1	105	284	314	327	361	509	0.976	0.792	2.747	4.090
	Chi-2	32	145	300	326	328	440	533	0.940	0.750	9.825	12.863
	NCA	1	157	313	314	327	503	600	0.929	0.813	7.830	9.186
	R	21	105	284	326	360	486	590	0.976	0.781	7.098	8.302
	Relief	22	105	284	326	327	361	584	0.988	0.792	5.107	6.706
	RF	22	182	290	326	331	361	584	0.976	0.792	9.078	9.467
	Average								0.963	0.788	7.345	8.865
The classical approach	F	294	293	295	21	296	292	297	0.952	0.708	15.613	17.067
	IG	12	1	2	3	4	5	6	0.798	0.677	35.907	38.667
	GR	1	2	3	4	5	6	7	0.821	0.698	31.399	35.557
	SU	1	2	3	4	5	6	7	0.821	0.698	31.399	35.557
	Chi-2	300	145	142	301	306	307	148	0.929	0.729	15.986	18.551
	NCA	314	313	312	315	311	316	310	0.929	0.760	15.936	17.980
	R	21	22	23	25	26	24	20	0.940	0.719	11.346	14.523
	Relief	32	31	23	24	22	21	25	0.917	0.667	9.171	10.254
	RF	160	22	200	155	330	149	162	0.929	0.729	11.664	11.366
	Average								0.893	0.709	19.825	22.169
Wilcoxon signed-rank test (p-value)								0.006	0.004	0.004	0.004	

F: F-Statistics, IG: Information Gain, GR: Gain Ratio, SU: Symmetrical Uncertainty, Chi-2: Chi-Square, NCA: Neighborhood component analysis, R: Correlation-Based Filter, RF: Random Forests.

Another criterion used in the measurement of classification performance is entropy, which is a measure of classification uncertainty. The entropy measure is calculated as follows

$$En(\tau) = - \sum_{i=1}^N \sum_{g=1}^G \tau_{ig} \ln(\tau_{ig}) \quad (21)$$

The smaller the entropy measure, the more successful the classification.

E. APPLICATION OF THE PROPOSED APPROACH

In this section, feature selection was performed for the training dataset using both the classical and the proposed approach with filter-based feature selection methods. Then, the dataset was modeled using the GMDA with the selected features. The classification of the test dataset was performed using the model obtained by the GMDA.

The features were clustered by the GMM and the appropriate number of clusters was determined as $G = 7$ according to the BIC values obtained for $g = 2, 3, \dots, 10$. Thus, the features containing the same information were clustered and the answer to the question ‘‘How many features should be selected?’’ has been answered. A graph of BIC values according to the number of clusters is shown in Figure 3.

In the GMDA application, each nitrogen class was modeled by dividing it into three subgroups. In the GMM and the GMDA applications, 10^{-12} was taken as the regularization value for the variance covariance matrix. Both the classical approach and the proposed algorithm were applied to the

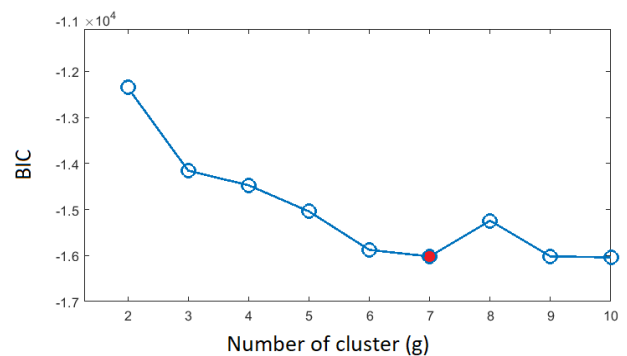


FIGURE 3. BIC values according to the number of clusters.

training data, and the best features ($S = 7$) were selected. The CA and entropy values obtained as a result of the GMDA of the training and test data using these selected features are given in Table 1. As seen in Table 1, feature selection algorithms in the classical approach tend to select sequential, i.e., correlated features. This situation was overcome using the proposed approach.

According to Table 1, the proposed approach produced better results than the classical approach in terms of both training and test data. This approach increased the CA and decreased the entropy values in all feature selection methods. To demonstrate the validity of the proposed approach, the Wilcoxon signed-rank test was also used, and the p-values

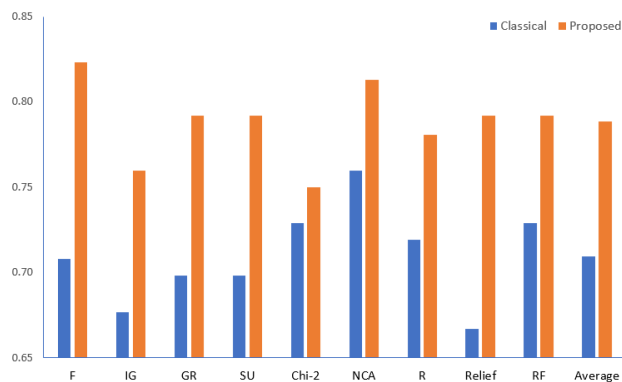


FIGURE 4. CA values of the test data obtained using the classical and proposed approach of feature selection methods.

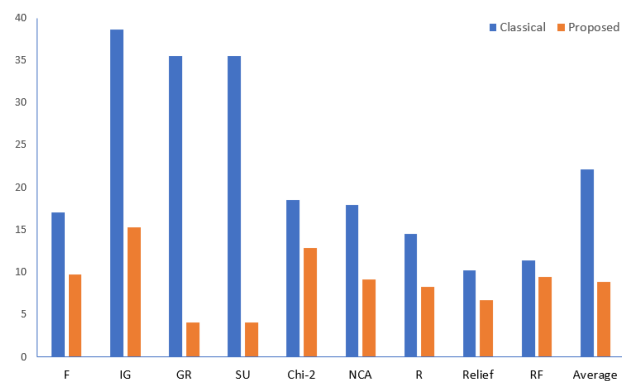


FIGURE 5. Entropy values of the test data obtained using the classical and proposed approach of feature selection methods.

from the one-tailed, paired difference test are given in Table 1. As seen in the table, all the p-values are at the 99% confidence level. These results confirm that the proposed approach increases the success of all feature selection methods.

The CA values of the test data obtained with the classical and proposed approach of the feature selection methods are shown in Figure 4, and the entropy values are shown in Figure 5.

IV. CONCLUSION

In this study, we presented a new feature selection approach based on the clustering of variables for data with high-dimensional and correlated features (variables). Using this approach, dimension reduction was made using the cluster centers to the data with class K and the variables were assigned to the appropriate G clusters by the GMM. The feature with the highest classification success was selected by the feature selection methods from each cluster, and the most suitable G features to represent the data were determined. Thus, the selection of features that have the same information has been prevented, and a suitable number of features has been decided as the proposed approach. The results obtained from the study revealed that the proposed approach improved the classification accuracy and entropy criteria compared with the classical approach.

The proposed approach was tested on peach plant data, which has a higher number of variables than the number

of observations and a high correlation between variables. The approach was applied to the training and test data, respectively. The training data were first reduced by using cluster centers, and then the features of the training data were clustered by the GMM. Using filter-based feature selection methods, a feature with the highest classification performance from each cluster was selected and the features that would represent the dataset were determined. The training data was modeled by the GMDA using the determined features. The model performance indicators for training and test data were calculated using the obtained model. According to the results in the training data, while the average CA and entropy values of the classical approach are 0.893 and 19.825, respectively; those of the proposed approach are 0.963 and 7.345. According to the results in the test data, while the average CA and entropy values of the classical approach are 0.709 and 22.169, respectively; these values of the proposed approach are 0.788 and 8.865. Thus, the proposed approach has given better results than the classical approach in terms of all feature selection algorithms in both training and test data.

On the other hand, principal components and multidimensional scaling analysis can be used as an alternative for dimensional reduction of datasets with unknown class memberships. In addition, different clustering algorithms can be used instead of the GMM used in the clustering of features.

It is obvious that this study contributes to the literature in terms of dimensional reduction, clustering of features, selection of cluster initial center and determining the appropriate number of features.

REFERENCES

- [1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.
- [2] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, nos. 1–4, pp. 131–156, 1997.
- [3] L. Ladha and T. Deepa, "Feature selection methods and algorithms," *Int. J. Comput. Sci. Eng.*, vol. 3, no. 5, pp. 1787–1797, 2011.
- [4] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, May 2003.
- [5] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," in *Data Classification: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, 2014, pp. 37–64.
- [6] V. Bolón-Canedo, N. Sánchez-Marroño, and A. Alonso-Betanzos, *Feature Selection for High-Dimensional Data*. Cham, Switzerland: Springer, 2015.
- [7] N. Sánchez-Marroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection—A comparative study," in *Proc. Int. Conf. Intell. Data Eng. Automated Learn.* Berlin, Germany: Springer, 2007, pp. 178–187.
- [8] L.-Y. Chuang, C.-H. Yang, K.-C. Wu, and C.-H. Yang, "A hybrid feature selection method for DNA microarray data," *Comput. Biol. Med.*, vol. 41, no. 4, pp. 228–237, Apr. 2011.
- [9] P. K. Ammu and V. Preeja, "Review on feature selection techniques of DNA microarray data," *Int. J. Comput. Appl.*, vol. 61, no. 12, pp. 39–44, Jan. 2013.
- [10] M. Dedeoglu, "Estimation of critical nitrogen contents in peach orchards using visible-near infrared spectral mixture analysis," *J. Near Infr. Spectrosc.*, vol. 28, nos. 5–6, pp. 315–327, Oct. 2020.
- [11] A. Ridok, N. Widodo, W. F. Mahmudy, and M. Rifai, "FC- SVM: DNA binding proteins prediction with average blocks (AB) descriptors using SVM with FC feature selection," in *Proc. Int. Conf. Sustain. Inf. Eng. Technol. (SIET)*, Lombok, Indonesia, Sep. 2019, pp. 22–27.
- [12] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," *IEEE Trans. Neural Netw.*, vol. 13, no. 1, pp. 143–159, Jan. 2002.

- [13] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," in *Proc. New Challenges Feature Selection Data Mining Knowl. Discovery*, 2008, pp. 90–105.
- [14] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1656–1671, Dec. 2013.
- [15] C. R. Kumar and R. Anuradha, "RETRACTED ARTICLE: Feature selection and classification methods for vehicle tracking and detection," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 3, pp. 4269–4279, Mar. 2021.
- [16] X. Liu and J. Tang, "Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method," *IEEE Syst. J.*, vol. 8, no. 3, pp. 910–920, Sep. 2014.
- [17] S. Chaib, Y. Gu, and H. Yao, "An informative feature selection method based on sparse PCA for VHR scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 147–151, Feb. 2016.
- [18] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 856–863.
- [19] J. Biesiada and W. Duch, "Feature selection for high-dimensional data—A Pearson redundancy-based filter," in *Computer Recognition Systems 2*. Berlin, Germany: Springer, 2007, pp. 242–249.
- [20] A. A. Yahya, A. Osman, A. R. Ramli, and A. Balola, "Feature selection for high dimensional data: An evolutionary filter approach," *J. Comput. Sci.*, vol. 7, no. 5, pp. 800–820, May 2011.
- [21] W. Yang, K. Wang, and W. Zuo, "Neighborhood component feature selection for high-dimensional data," *J. Comput.*, vol. 7, no. 1, pp. 161–168, Jan. 2012.
- [22] C. H. Park and S. B. Kim, "Sequential random k-nearest neighbor feature selection for high-dimensional data," *Exp. Syst. Appl.*, vol. 42, no. 5, pp. 2336–2342, Apr. 2015.
- [23] D. P. Ismi, S. Panchoo, and M. Murinto, "K-means clustering based filter feature selection on high dimensional data," *Int. J. Adv. Intell. Inform.*, vol. 2, no. 1, pp. 38–45, 2016.
- [24] A. Ben Brahim and M. Limam, "Ensemble feature selection for high dimensional data: A new method and a comparative study," *Adv. Data Anal. Classification*, vol. 12, no. 4, pp. 937–952, Dec. 2018.
- [25] Z. Manbari, F. AkhlaghianTab, and C. Salavati, "Hybrid fast unsupervised feature selection for high-dimensional data," *Exp. Syst. Appl.*, vol. 124, pp. 97–118, Jun. 2019.
- [26] C. Fahy and S. Yang, "Dynamic feature selection for clustering high dimensional data streams," *IEEE Access*, vol. 7, pp. 127128–127140, 2019.
- [27] Y. Lu, I. Cohen, X. S. Zhou, and Q. Tian, "Feature selection using principal feature analysis," in *Proc. 15th Int. Conf. Multimedia*, Germany, 2007, pp. 301–304.
- [28] D. Mladenović, "Feature selection for dimensionality reduction," in *Proc. Int. Stat. Optim. Perspect. Workshop Subspace, Latent Structure Feature Selection*. Berlin, Germany: Springer, 2005, pp. 84–102.
- [29] H.-L. Wei and S. Billings, "Feature subset selection and ranking for data dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 162–166, Jan. 2007.
- [30] K. M. Faraoun and A. Rabhi, "Data dimensionality reduction based on genetic selection of feature subsets," *J. Comput. Sci.*, vol. 6, no. 3, pp. 36–46, 2007.
- [31] M. Maseali, G. Fung, and J. G. Dy, "From transformation-based dimensionality reduction to feature selection," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 751–758.
- [32] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 2, pp. 56–70, May 2020.
- [33] M. Li, H. Wang, L. Yang, Y. Liang, Z. Shang, and H. Wan, "Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction," *Exp. Syst. Appl.*, vol. 150, Jul. 2020, Art. no. 113277.
- [34] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Inf. Fusion*, vol. 59, pp. 44–58, Jul. 2020.
- [35] M. Gogebakan, "A novel approach for Gaussian mixture model clustering based on soft computing method," *IEEE Access*, vol. 9, pp. 159987–160003, 2021.
- [36] H. Liu and R. Setiono, "chi2: Feature selection and discretization of numeric attributes," in *Proc. 7th IEEE Int. Conf. Tools Artif. Intell.*, Nov. 1995, pp. 388–391.
- [37] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.
- [38] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.
- [39] J. Novaković, P. Štrbac, and D. Bulatović, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugoslav J. Oper. Res.*, vol. 21, no. 1, pp. 119–135, 2011.
- [40] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine Learning Proceedings*. Burlington, MA, USA: Morgan Kaufmann, 1992, pp. 249–256.
- [41] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: A study on high-dimensional spaces," *Knowl. Inf. Syst.*, vol. 12, no. 1, pp. 95–116, May 2007.
- [42] H. Zhou, Z. Deng, Y. Xia, and M. Fu, "A new sampling method in particle filter based on Pearson correlation coefficient," *Neurocomputing*, vol. 216, pp. 208–215, Dec. 2016.
- [43] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016.
- [44] M. Huljanah, Z. Rustam, S. Utama, and T. Siswantining, "Feature selection using random forest classifier for predicting prostate cancer," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 546, no. 5, 2019, Art. no. 052031.
- [45] N. O. F. Elssied, O. Ibrahim, and A. H. Osman, "A novel feature selection based on one-way ANOVA F-test for e-mail spam classification," *Res. J. Appl. Sci., Eng. Technol.*, vol. 7, no. 3, pp. 625–638, Jan. 2014.
- [46] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighborhood component analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17, 2005, pp. 513–520.
- [47] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *J. Amer. Stat. Assoc.*, vol. 97, no. 458, pp. 611–631, Jun. 2002.
- [48] S. Akogul and M. Erisoglu, "An approach for determining the number of clusters in a model-based cluster analysis," *Entropy*, vol. 19, no. 9, p. 452, Aug. 2017.
- [49] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc., Ser. B*, vol. 39, no. 1, pp. 1–22, Sep. 1977.
- [50] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, vol. 382. Hoboken, NJ, USA: Wiley, 2007.
- [51] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, "Finite mixture models," *Annu. Rev. Statist. Appl.*, vol. 6, pp. 355–378, Jan. 2019.
- [52] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.
- [53] T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *J. Roy. Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 155–176, Jan. 1996.
- [54] R. P. Browne and P. D. McNicholas, "Model-based clustering, classification, and discriminant analysis of data with mixed type," *J. Stat. Planning Inference*, vol. 142, no. 11, pp. 2976–2984, Nov. 2012.
- [55] A. H. Strahler, "The use of prior probabilities in maximum likelihood classification of remotely sensed data," *Remote Sens. Environ.*, vol. 10, no. 2, pp. 135–163, Sep. 1980.



SERKAN AKOGUL was born in Adana, Turkey, in 1987. He received the Ph.D. degree in statistics from Selçuk University, in 2018. From 2014 to 2018, he was a Research Assistant with the Statistics Department, Yıldız Teknik University. Since 2018, he has been an Assistant Professor with the Department of Statistics, Pamukkale University. He has published 11 articles and six book chapters on statistics and data analysis. His research interests include applied statistics, classification and clustering analysis, big data, and multivariate statistics applications.

• • •