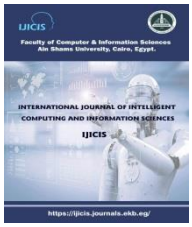
IJICIS, Vol.21, No.1, 19-32
DOI: 10.21608/ijicis.2021.148607

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



INTELLIGENT CLUSTERING TECHNIQUE BASED ON GENETIC ALGORITHM

Shaymaa Adnan Abdulrahman*

Department of computer Engineering, Imam Ja'afar Al-Sadiq University, Baghdad, Iraq

Shaymaaa416@gmail.com

Mohamed Roushdy

Faculty of Computers &Information Technology, Future University in Egypt, New Cairo, Egypt.

Mohamed.Roushdy@fue.edu.eg

Abdel-Badeeh M. Salem
Department of Computer&Information Science, Ain Shams University, Cairo, Egypt

absalem@cis.asu.edu.eg

Received 2021-1-12; Revised 2021-2-4; Accepted 2021-02-11
Available online 2021-02-14

Abstract : *This paper focuses on the problems of data clustering where the similarity between different objects is estimated with the use of the Euclidean distance metric. Also, K-Means is used to remove data noise, genetic algorithms are used for finding the optimal set of features and the Support Vector, Machine (SVM) is used as a classifier. The experimental results prove that the proposed model has attained an accuracy of 94.79 % when using three datasets taken from the UCI repository.*

Keywords*: data mining, clustering, genetic algorithms, feature extraction, K-means*

1.INTRODUCTION

Clustering can be specified as the organization after the collection of un-labeled patterns. For instance, the organization of vectors related to the measurement or point in multi-dimensional space into clusters depending on their similarity [1]. It is widely utilized in plenty of applications including information retrieval, pattern recognition, computer vision, data mining, and image processing. There are three kinds of clustering techniques: such as Overlapping; Hierarchical, and Partitioning [2]. Usually, clustering is an operation that is performed on unsupervised data. Additionally, clustering facilitates getting or finding relevant information at a faster speed. A very popular clustering method, the K-means algorithm, is most notably used for data clustering [3]. Even though these methods are made for efficiency, ease of convergence and simplicity, the

* Corresponding author: Shaymaa Adnan Abdulrahman
Department of computer Engineering, Imam Ja'afar Al-Sadiq University, Baghdad, Iraq
E-mail address: Shaymaaa416@gmail.com

majority of them .suffer from the disadvantage of requiring a great number of clusters in addition to their. initial centers and can have poor clustering quality. Many clustering algorithms have been developed, such as the Classical clustering algorithms that have been enumerated as "k-means", nearest neighbor clustering, spectral clustering, and fuzzy c-mean clustering [4][5].Among them , a genetic algorithm (GA) was proposed early in 1989 [6]. Concerning the genetic algorithm (GA), this study will apply a model that relates to real-life natural selection, in which the solutions' initial population referred to as individuals will be generated in a random manner [4]. The algorithm will produce novel solutions regarding populations through certain genetic operations, like mutations, cross-over, and re-production. The new generations include the potential survivors with maximum fitness scores. Also, the new individuals will be determined from the preceding population with the use of genetic operations [7] .This paper focuses on genetic algorithms and K-mean and presents a way for the automatic generation of those parameters, at the same time enhancing the algorithm's accuracy and speed. It also uses classification with SVM to determine accuracy . Genetic Algorithms deal with handling heterogeneous populations. It operates with a variety of structure types[8] . Typically, using certain parameters like cross-over as well as mutation probabilities that, adapt to the advancement of the algorithm is a suitable choice. High population diversity is necessary to prevent the algorithm from falling in the local minima.[9]. To accelerate the process of the GA and increase the initial population's diversity, that population is generated in 2 ways. The 1st sub-population has been generated in a deterministic manner, and the 2nd sub-population has been generated randomly. The increase in the diversity of the population allows for achieving higher quality. The remaining parts of this study will be structured in the following ways:

related works will be specified in section 2, Preliminaries (clustering methods (GAs and k-means) will be presented in Section 3, and Section 4 describes proposed method experimental results and discusses in Section 5, Section 6 will provide future work and conclusion

2.RELATED WORK

The major challenge with using the (K-Means) method is determining the number of clusters in addition to their initial centers. The random determination of the initial centers doesn't guarantee sufficient clustering quality, because multiple K-Means running on the same data-set can result in different offspring. Additionally, such a procedure might be converging to sub-optimal partitions.

(Fahim AM, et al) [10] implemented a straightforward approach for enhancing the effectiveness related to k-means clustering when using three datasets: Wind, Abalone, and Letter Image[10]. They also used the k-means method with the CLARA algorithm to compare the execution time of the three different implementations of the datasets. (Lin Wei-Chao, et al) [11] applied two strategies for clustering techniques and used the "DIARETDB0" data-set in their work. The initial approach utilized cluster centers to represent the majority class, while the other approach utilized nearest neighbors related to cluster centers. The dataset consists of eighty-nine color images that have been obtained in (Kuopio university hospital). Additionally, a decision tree was used as a classifier to find the optimal performance between both small and large scale data sets.

Big data clustering has been utilized through (Shrivastava Puja, et al) [12]. An estimated MapReduce framework was established by using the steps of the genetic approach. They solved the problem of local optimal clusters with big data K-means using the concept of GA . While (Chen Huihui, et al) [13] utilized novel gene selection techniques that depended on clustering using eight genes as a data set. It presented a new method that is referred to as the Kernel-Based Clustering method for Gene Selection (KBCGS). SVM and KNN have been used to classify the accuracy when comparing six features of the cancer dataset. Tao Lei, et al [16] relied on the morphological

reconstruction of the fuzzy c.means clustering FCM method to ensure noise immunity and the modification of membership. The partitioning process here depended on the distance between pixels in the local spatial neighbors and cluster centers. In his experiment,. two 256x256 images were utilized, the first image containing 3 values of intensity (3 classes), being (0.85 & 170), and the 2nd image containing 4 intensity values (4 classes), which are(0, 85 & 255). Their work proved that the (FCM) algorithm is fast and powerful especially in the process of image segmentation . (W. Cai, D. Zhang and S. Chen), [17] suggested that fast generalized fuzzy c. means (.FG FCM) algorithm using new factors as a local measure of similarity that aims to reduce noise for images and eliminates empirically altered parameters (α) which are required in the (En-FCM) algorithms . It finally carries out the clustering on the grey levels histograms. (Chen Yewang et al) [18] proposed a clustering technique to improve "DBSCAN" called (novel local neighborhood searching technique – "NQ-DBSCAN"). One of the benefits of this technique is to reduce the numerous needless calculations of distance. The results of this technique were NQ-DBSCAN that averagely runs in O(n log(n)). The optimal case was O(n). Two type of data sets were used one was (synthetic data) and the second was (real-time data) .There is a technique that reverses the nearest neighbor called (RNN-DBSCAN) which was proposed by (Bryant Avory and Cios Krzysztof) [19]. They preferred to use this technique for several reasons such as an improved ability to deal with large variations and reducing complexity problems while using single parameters. The Genetic-based clustering technique by (Varshali , et al)[20] using colored image segmentation . This technique proved to be more efficient and less time –consuming .

3.PRELIMINARIES

Clustering could be considered to be the process of data segmenting or partitioning applied to groups of the same type . Clustering is usually achieved by specifying similarity among data on predefined attribute . K-means might be defined as the main algorithm that is extensively applied with clustering. GA is also utilized in clustering, combined with k-means or separate [14].

3.1 K-means

This algorithm has the goal of partitioning a set of data points in unique clusters in a way so that inertia in each cluster is minimalized. Inertia is specified as the sum regarding the distances between all data points in clusters and centroids (or representatives)[15]. The K-means algorithm divides experimental dataset (X) into (k)clusters in which k has been provided via the user. Initially, objects might be assigned randomly to k clusters. Throughout the consecutive iterations cluster. centers X_j will be re-distributed over data space in such. a way that. the objects, that are in one cluster, will be similar to each. other in comparison to the objects in other clusters. Similarities between the objects can be determined through Euclidean distance [16]. The K-means algorithm has main disadvantage:

. *it is implying that data clusters might be ball-shaped since it is performing clustering depending on just the Euclidean distance*

. *As indicated in the study of Xu et al.[17]. dead-unit problem exists in a way that when certain units might be initialized far from input dataset when compared with the other units* **,** *then they will be instantly dead with no learning chance in the overall process of learning.*

3.2 Genetic algorithms(GA*)*

GAs are robust algorithms of stochastic optimizations. Genetic algorithms are utilized for solving many different issues. They have been advanced to gain a better understanding of natural processes like adaptation, which is analogous to a certain search method type that mimics natural selection's principle for developing solutions regarding large optimization problems[18]. GAs are utilized with: fitness evaluation, representation of the chromosomes, stating the initial population, selections, and reproduction (i.e. the mutation and the cross-over). GAs operate through the manipulation as well as the maintenance of the population. Chromosomes are regarded as possible solutions for a variety of adaptation problems. Usually each chromosome has a fitness value, which is a qualitative measure regarding the effectiveness of the solution that has been encoded within it. The process of the GA begins with the determination of an initial solution which is made up of a group of the chromosomes (i.e. the initial population), followed by the iterative application of the operators of the reproduction (selections, mutation, and cross-over) to the point of reaching a specific parameter of the quality or a pre-defined amount of iterations. Using a fitness function guides the stochastic selections regarding the chromosomes that are utilized for generating new candidate solutions via mutation and crossover. A Straightforward outline regarding the process of GA is indicated in Figure 1.

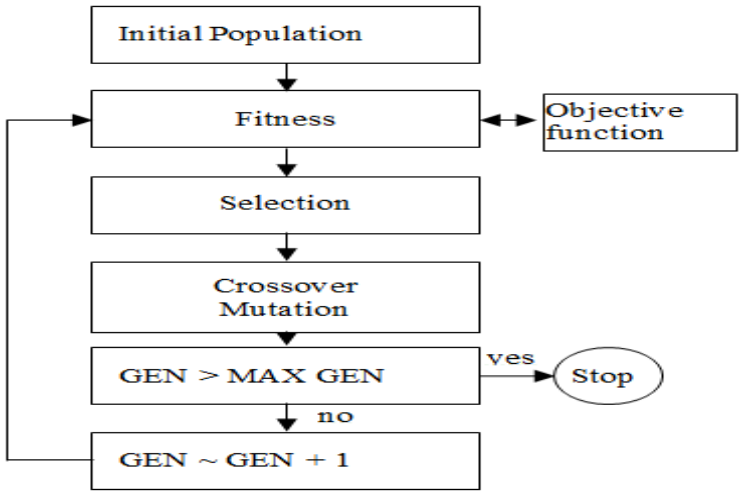


Figure 1. Structure of the (GAs algorithms)[7]

Genetic algorithm includes characteristics like population size, encoding data, objective function, crossover, and mutation. Typically, the objective function is assigning a fitness value for each one of the individuals in a population. Individuals with high fitness indicate an effective solution to the problem when compared to individuals with low fitness values [19]. Whereas the encoding represents an independent variable for the objective function. Crossover represents a procedure where a high fitting chromosome is provided the possibility of reproduction through exchanging pieces regarding its genetic information with other high fitting chromosomes. The operation also utilized crossover through randomly changing certain genes of the individual parents. This is referred to as a mutation. The number of individuals in a population indicates this population's

size. If the population size is larger then there is a higher possibility that an excellent solution is going to be found.

3.2.1 Chromosomes representation

The representation of chromosomes is a problem that is solved by utilizing GAs throughout a variety of significant steps. Encoding both number of cluster and their centers .Our propose a chromosome framework that applies a real coded genetic algorithm to the clustering problem. Crossover and mutation operators are utilized directly to real parameter values [7].A chromosome in a subpopulation may consist of two, three, and up to (N-max1) genes where the (N-max1) value is the maximum number of possible clusters. Those genes hold data that corresponds and are related to cluster of centers. Also, the number of genes in the chromosome indicates the number of clusters. The Value of N will be in the interval [2, Nmax1]. Furthermore, each one of the solutions, i, will be a strong value with fixed-length specified through cluster centers (.Cij ; j = 2;…, Nmax1) . Finally , the solution, i, of the population is represented using a vector.

3.2.2 Fitness function (Ff)

Usually fitness function (Ff) represents a measure of success during optimization. It is considered to be an objective function that is utilized for determining how close a certain solution is to achieve aims. The fitness function has a significant impact on GA's success [20]. The Measure of fitness helps us develop good solutions and implement natural selection. After defining the Fitness function (Ff), various parameters of the genetic algorithm's operators are fixed. the fitness of a chromosome is computed by using (Mean Square Error (MSE)).

$$MSE := \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \dots\dots\dots (1)$$

$$Fitness := \frac{1}{MSE} \dots\dots\dots.. (2)$$

3.2.3 Stage of Generation of the Initial - Population

To accelerate the process of the GAs and to increase the initial population's individual varieties meticulous care must be taken in regard to initial population generation .The generation of this initial population is done through the use of (*two*) methods. The 1st one includes a deterministic production way [21][23] whereas the other one includes random generation. 20% of the initial population is

obtained from the 1st way and 80% is obtained from the 2nd way. From the input dataset, a sorted dataset is produced. After that, algorithm- 1 is applied to the sorted dataset (as can be seen in the algorithm below) [24][7]. Doing this increases the variety of results which enhances genetic reproduction

Algorithm-1: generation of 20% of the population [7]

1-Sorting data-set;
2-For (N1=2, N1<:=N_{Max1} -1; N1++) {
- Dividing sorted data to N equal segments: {SN};
- Modal step: "taking every segment k's modal value as centre" CN;
- Means step: taking mean value of every one of the segments k as centre CN:
- Min1_Max1 step:
For (ii=0, ii<k, N1++){
Calculating the C = (max1-min1)/2
Considering C as a i^{th} center.

3.2.4 Reproduction

A Genetic algorithm help look for the better solutions from several possible solution Reproduction could be applied in a form based algorithmically on reproduction operators which are better than the encoding scheme [2][25]. The main aim of reproduction operators is to be ensure of variety in a population where solution that are the fittest could be derived and contemplated through the process of evolution . Such evolutionary processes are composed of both crossover and mutation [7] .

A. Crossover

In the process of crossover, the children's chromosomes are generated in accordance with the fitness function value of parents. In this case, it is necessary to define the operator of the crossover in a way that it would accept the parent's chromosomes with multiple numbers of genes [13]. For each of the two parents selected , crossover is implemented and produces two offspring [7].

B. Mutation

The mutation is designed to prevent all the population's solutions from falling in to the same solved problems, and it enhances obtained chromosomes . Mutation takes place with a lesser probability than that of the crossover [22]. Mutation consists of two different operators, the First operator, topological mutation, is designed to either delete or add genes from an offspring [23]. The second operator, gene mutation, is intended to edit or modify the selected gene from the offspring. Usually, mutation is implemented by selecting a gene from an offspring randomly and replacing it with a data point which is also randomly extracted from a certain dataset [24]. Afterwards, a check is performed to avoid the occurrence of redundant centers and for keeping cluster number in [*2, Nmax*]. When there is a need for finding the closest cluster for all points, the K-means operator must be used. This is enumerated "Classical K-means". [25] Concerning such a condition, data

points will be assigned to their clusters for all the new chromosomes using the K-means algorithm to compute the fitness level.

3.3. SUPPORT VECTOR MACHINE (SVM)

SVM is a classifier that performs classification tasks by constructing hyper.-planes in a multi-dimensional space separating the variety of the class label cases. SVM's create a hyper -.plane that separates (*two*) groups of data. The data is further pushed away from each other and the hyperplane creates a space between the two groups named the "margin". As shown in figure 2.

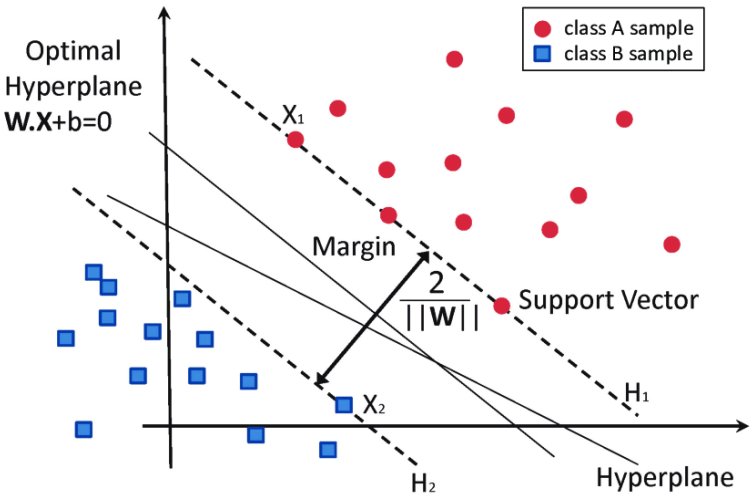


Figure 2: support vector machine – hyper-plane and margin

Support Vectors can be described as a subset of the data samples that are utilized for defining the hyper-plane. The SVMs support the classification as well as the regression tasks and are capable of handling numerous categorical and continuous variables. There are two SVM types. First is the linear SVMs, which are utilized for the separation of data points with the use of linear decision boundaries. And secondly, there's the Nonlinear SVMs, which separate data points with the use of a non-linear boundary of the decision. The conventional approaches of the SVMs need the package of the quadratic programming (QP).

The main problem of using quadratic programming is that it is very time consuming and needs massive amount of memory in addition to detailed numerical analysis knowledge.

4. PROPOSED METHOD

The basis for the proposed work is shown in figure 3. This figure consists of four steps, first: we begin with collecting the data set, second: data cleaning is done to replace missing values with means. Third, we perform clustering by using the K-Means algorithm to remove and delete outliers and use genetic algorithms (GA) to select optimal features. Finally , the last step is the classification when using SVM as a classifier to achieve better accuracy . Additionally, a 10–fold cross-validation technique is applied to increase the reliability of the Classifier's performance .

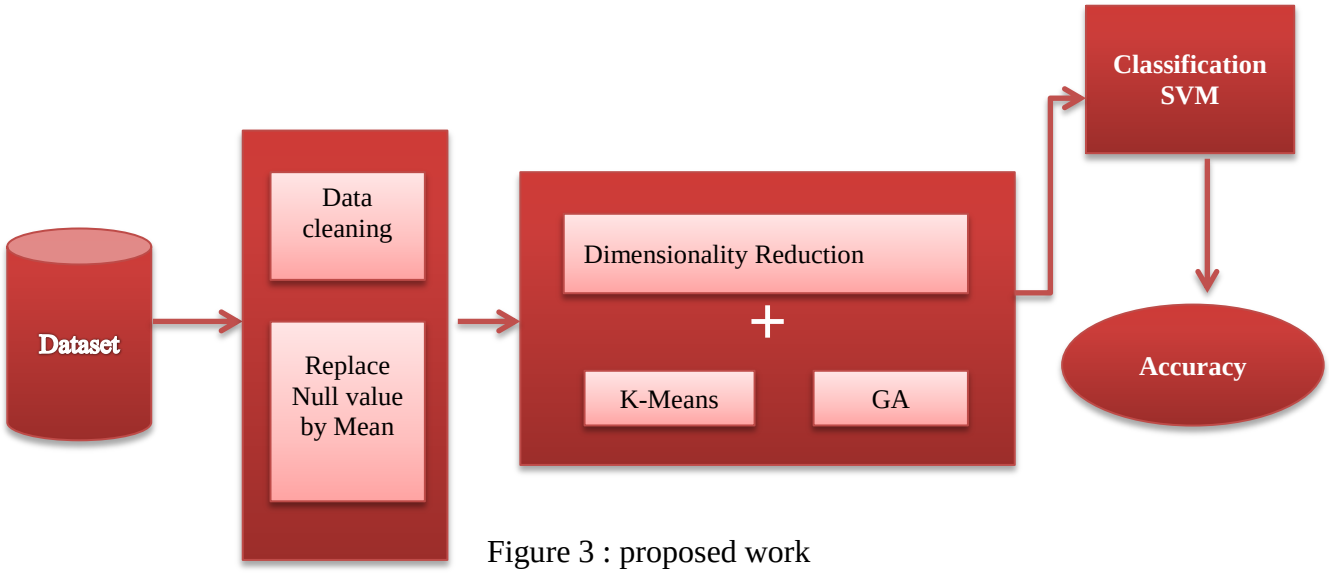


Figure 3 : proposed work

4.1. Evaluation Metrics

To visualize the efficiency of the supervised algorithms of ML, a confusion matrix has been utilized. The 4 indices of classification performance present in the matrix of the confusion are illustrated in Figure 4. The utilized metrics of the evaluation were: Positively Predicted, Specificity, Sensitivity, and Negatively Predicted values. Specificity and Sensitivity represent the ratio of the actual positives and negatives that are identified correctly as such. The Negative and Positive predictive values represent the predicted result proportions.

$$\text{Sensitivity (\%)} = \frac{TP}{TP.+FN.} \times 100 \quad \ldots\ldots\ldots(3)$$

$$\text{Specificity (\%)} = \frac{TN}{TN.+FP.} \times 100 \quad \ldots\ldots\ldots(4)$$

$$\text{Positively Predicted (\%)} = \frac{TP}{TP.+FP.} \times 100 \quad \ldots\ldots\ldots(5)$$

$$\text{Negatively Predicted value (\%)} = \frac{TN}{TN.+FN} \times 100 \quad \ldots\ldots\ldots(6)$$

		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Figure 4 : Confusion Matrix

5. Experimental Results and Discussions

In this research, we used 3 different data-sets (abalone, wine, and network) from the UCI repository. The datasets chosen are the same ones used by the research paper [10]. Each dataset has different numbers of attributes, sizes, types, and clusters as shown in Table 1

Table 1: Characteristics of data set

Dataset	Number of attributes	Number of records	Type
Abalone	8	4177	categorical, real
Wine	15	6574	integer, real
Network	21	53413	integer, real

At this stage, after using the mean to replace null values, some of the values became outliers. And, the way to do away with those noisy, inconsistent values, the simple K-Means algorithm was used. After removing the outliers, the process of feature selection begins. The genetic algorithm has been utilized as a tool for feature selection . The different stages in which this process happens have been mentioned above. It is important to note that the GA selects the features from the original data set.

The last stage of the experiment is classification. Results are given to the SVMs to use a (ten.-fold cross.- validation) technique for the classification. Throughout each one of the runs, the genetic algorithms choose a variety of characteristics from the original group of features and the accuracy of the classification is recorded. Table 2 shows that the experimentation has been repeated (20) times with an average SVM classifier accuracy of 94.79%.

Table 2: Experimental Outcome when 20 runs

Test run	Attribute No	SVM (%)	Test run	Attribute No	SVM (%)
1	2	94.45	11	3	94.46
2	4	94.15	12	3	94.24
3	4	94.76	13	4	94.32
4	3	94.98	14	2	94.11
5	2	94.56	15	5	94.73
6	5	94.59	16	4	94.05
7	3	94.87	17	3	94.64
8	3	94.08	18	4	94.58
9	5	94.65	19	2	94.77
10	2	94.57	20	3	94.19

While Table 3 refer to the evaluation metrics obtained when applied SVM techniques to Classifying the reduced dataset. The resulting accuracy improved on the previous literature shown in Table 4.

Table 3 :Evaluation Metrics from classification

(Performance. Measures)	Reduced Dataset
#(No. of Attributes Used)	4
(Sensitivity (%))	94.40
(Specificity (%))	94.55
(Positively Predicted value) (%)	94.98
(Negatively Predicted. values (%)	94.36
SVM- Accuracy	94.79

Table 4: Comparison between. proposed work and ,existing works in terms of Accuracy different techniques

#	Authors	Techniques	Data set	Years	Accuracy
1	Ahmad, F. et al [10]	Improved (GA) algorithm	Three dataset	2013	80.4 %
2	Y.S. Thakare and S. B. Bagal[24]	k-means algorithm	Wine data set	2017	79.74
3	Zeinab et al [21]	Genetic algorithm	- Z- Alizadeh Sani dataset	2017	93.85%,
4	Chen Huihui ,et al [13]	KBCGS technique	Eight cancer dataset (Lung) (http://datam.i2r.astar.edu.sg/datasets/krbd/)	2016	87.00 %
7	Chen Yewang et al [18]	NQ-DBSCAN clustering	Two dataset	2018	93.85%
8	Nihat Yilmaz et al [26]	K-Means + SVM	-	2017	93.65%

10	**Proposed method**	GA+ K-means +SVM	Three dataset	-	**94.79%**

Finally, the results of the study are:

1- The minimal numbers of the attributes, that have been chosen with, the use of the Genetic Algorithms techniques. was (*two*) and the maximal was (*five*) .
2- The minimal and maximal, accuracy of classification of the. SVM classifier was 94.79%.
3- Through experience there were 511 instances. The k-.means selected (318) samples as correctly while (193) as outliers.
4- The outlier detection percentage is (33.46)

6. Conclusion and future work

One of the most important issues to tackle in the field of machine learning is the reduction of dimensionality to achieve higher accuracies and to make the work more efficient. The goal of this study was to improve on data cleaning, dimensionality reduction, and classification. By using GAs and K-means algorithms this paper achieved a high accuracy rating of 94.79% which is notably higher than the results of previous literature shown in Table 4 . Future works should concentrate on, applying box-plots to detect outliers. It should also concentrate on using standard deviation to find the replacement of missing values, algorithms that include PCA/ Nonlinear SVMs for feature selection, and experimenting with classifiers from the tree, statistical, c-fuzzy, and neural families for enhancing precision.

References

[1] Shahrivari S, Jalili S " Single-pass and linear-time k-means clustering based on MapReduce". pp 1–12 , 2016.

[2] V Chittu, N Sumathi. "A Modified Genetic Algorithm Initializing K-Means Clustering".

[3] Thakare, YS and Bagal, SB ," Performance evaluation of K-means clustering algorithm with various distance metrics" International Journal of Computer Applications, vol 110, number 11 ,pp 12-16 , 2015

[4] A.Y.Ng, M.I. Jordan, and Y. "Weiss On spectral clustering: Analysis and an algorithm". pp 849-856, .

[5] J. C. Bezdek. Fuzzy c-means cluster analysis. Scholarpedia, 6(7):2057, 2017.

[6] Shaymaa Adnan Abdulrahman, Mohamed Roushdy, Abdel-Badeeh M. Salem "Support Vector Machine approach for human identification based on EEG signals , journal of mechanics of continua and mathematical sciences , ISSN (Online) : 2454 -7190 Vol.-15, No.-2, pp 270-280 ISSN 0973-8975-February-2020 https://doi.org/10.26782/jmcms.2020.02.00023

[7] Amina Bedboudi , Cherif Bouras , Mohamed Tahar Kimour," An Heterogeneous Population-Based Genetic Algorithm for Data Clustering " Indonesian Journal of Electrical Engineering and Informatics (IJEEI), Vol. 5, No. 3, pp. 275-284, 2017

[8] Eduardo R. Hruschka and Nelson F.F. Ebecken "A genetic algorithm for cluster Analysis "intelligent Data analysis vol 7, pp15-25, 2017.

[9] Bedboudi Amina , Bouras Cherif , Kimour Mohamed Tahar " An heterogeneous population-based genetic algorithm for data clustering " Indonesian Journal of Electrical Engineering and Informatics (IJEEI), vol 5, No 3 , pp 275-284 ,2017

[10] Fahim AM, Salem AM , Torkey F Af ,Ramadan MA," An efficient enhanced k-means clustering algorithm " Journal of Zhejiang University-Science A ,Springer ,vol 7,No 10 , 2006

[11] Lin Wei-Chao, Tsai Chih-Fong, Hu Ya-Han , Jhang Jing-Shang," Clustering-based under sampling in class-imbalanced data",vol 409,pp17-26, 2017.

[12] Shrivastava Puja, Sahoo Laxman , Pandey Manjusha , Agrawal Sandeep," AKM—Augmentation of K-Means Clustering Algorithm for Big Data" Springer, pp 103-109 ,2018

[13] Chen Huihui , Zhang Yusen , Gutman Ivan," A kernel-based clustering method for gene selection with gene expression data " Journal of Biomedical Informatics, vol 62, pp 12-20, 2016

[14] Lu Huijuan , Chen Junying , Yan Ke Jin, Qun Xue, Yu Gao Zhigang "A hybrid feature selection algorithm for gene expression data classification " Neurocomputing , vol 256, pp 56-62 ,2017

[15] Huang, Xiaohui and Ye, Yunming and Xiong, Liyan and Lau, Raymond YK and Jiang, Nan and Wang, Shaokai , " Time series k-means: A new k-means type smooth subspace clustering for time series data" Information Sciences , vol 367, pp 1-13, 2016

[16] Lei, Tao and Jia, Xiaohong and Zhang, Yanning and He, Lifeng and Meng, Hongying and Nandi, Asoke K (2018) Significantly fast and robust fuzzy c-means clustering algorithm based on morphological reconstruction and membership filtering vol 26, number 5 , pp 3027-3041 ,2017

[17] W. Cai, S. Chen, and D. Zhang, “Fast and robust fuzzy c-means clustering algorithm incorporating local information for image segmentation,” Pattern Recognition ., vol. 40, no. 3, pp 825-838 ,2017

[18] Chen Yewang , Tang Shengyu , Bouguila, Nizar Wang and Cheng Du and Jixiang and Li HaiLin(2018) A fast clustering algorithm based on pruning unnecessary distance Computations in DBSCAN for high - dimensional data, Pattern Recognition, vol 83PP 375-387,2017

[19] Bryant, Avory and Cios, Krzysztof , RNN-DBSCAN(2018) A density-based clustering algorithm using reverse nearest neighbor density estimates ,vol 30 , number 6 , pp 1109 1121

[20] Jaiswal Varshali , Sharma Varsha , Varma Sunita," An implementation of novel genetic Based clustering algorithm for color image Vol 17, number 2 , pp 1461-1467,2019

[21] Arabasadi Zeinab , Alizadehsani Roohallah , Roshanzamir Mohamad , Moosaei Hossein Yarifard Ali Asghar" Computer aided decision making for heart disease Detection using hybrid neural network-Genetic algorithm" Computer methods and Programs in biomedicine , vol 141,pp 19-26 , 2017

[22] Lu Huijuan Chen Junying , Yan Ke , Jin Qun , Xue Yu , Gao Zhigang," A hybrid feature selection algorithm for gene expression data classification" Elsevier, vol 256,pp 56-62 , 2017

[23] Kabir Mir Md Jahangir,Xu Shuxiang ,Kang Byeong Ho, Zhao Zongyuan," A new multiple seeds based genetic algorithm for discovering a set of interesting Boolean association rules, vol74,pp 55-69, 2017

[24] Y. S. Thakare and S. B. Bagal, " Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics. " Vol 110 – No. 11

[25] Shaymaa Adnan Abdulrahman, Wael Khalifa, Mohamed Roushdy, Abdel-Badeeh M. Salem" (2020) Comparative Study for 8 Computational Intelligence Algorithms for Human Identification" Journal , Computer Science Review

[26] Nihat Yilmaz, Onur Inan, Mustafa Serter Uzer. A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases. Springer: Transaction Processing Systems:J Med Syst; April 2014; 38:48.

[23] Kabir Mir Md Jahangir,Xu Shuxiang ,Kang Byeong Ho, Zhao Zongyuan," A new multiple seeds based genetic algorithm for discovering a set of interesting Boolean association rules, vol74,pp 55-69, 2017

[24] Y. S. Thakare and S. B. Bagal, " Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics. " Vol 110 – No. 11

[25] Shaymaa Adnan Abdulrahman, Wael Khalifa, Mohamed Roushdy, Abdel-Badeeh M. Salem" (2020) Comparative Study for 8 Computational Intelligence Algorithms for Human Identification" Journal , Computer Science Review

[26] Nihat Yilmaz, Onur Inan, Mustafa Serter Uzer. A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases. Springer: Transaction Processing Systems:J Med Syst; April 2014; 38:48.