Journal of Advances in Computer Research
Quarterly ISSN: 2008-6148
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 11, No. 3, August 2020), Pages: 61-82
www.jacr.iausari.ac.ir

An Improved K-Means with Artificial Bee Colony Algorithm for Clustering Crimes

Mohammad Karimi, Farhad Soleimanian Gharehchopogh✉
Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran
mk2016@chmail.ir; bonab.farhad@gmail.com

Received: 2020/03/09; Accepted: 2020/10/03

Abstract

Crime detection is one of the major issues in the field of criminology. In fact, criminology includes knowing the details of a crime and its intangible relations with the offender. In spite of the enormous amount of data on offenses and offenders, and the complex and intangible semantic relationships between this information, criminology has become one of the most important areas in the field of clustering. With the development of computer systems and the development of clustering algorithms, it has been possible to interpret mass data and extract knowledge from them. There are different types of attribute in the mass data set, each of which can be suitable for crime detection. By clustering, different groups of crime can be identified and also the percentage of their occurrence. In this paper, a K-Means improved by Artificial Bee Colony (ABC) algorithm is proposed for crime clustering. In the proposed model, an ABC algorithm has been used to improve cluster centers and increase the accuracy of clustering and assignment of samples to appropriate clusters. The main motivation is to exploit the search ability of ABC algorithm and to avoid the original limitation of falling into locally optimal values of the K-Means. Evaluation has done on data set with 1994 samples and 128 features. The results show that the accuracy of the proposed model is higher than K-Means, and the Purity value of the proposed model with 500 iterations is 0.943.

Keywords: *Crime Detection, Clustering, K-Means, Artificial Bee Colony*

1. Introduction

Human social conditions are such that exposure to a phenomenon called crime is inevitable, and humans always need knowledge of crime analysis. Crime analysis is the use of a smart way to detect, explore and anticipate crimes. By expanding the database servers and the large amount of data stored on these servers, it needs a tool to process these data and to classify and analyze the data [1]. With massive amounts of data and information related to crimes in police centers and the complexity of communications between these crimes, other statistical and manual models are not responsive to detection and prediction [2].

In traditional statistical methods of crime, often skillful detectives and experienced crime scene experts are employed, with two major defects [3]. First of all, it requires a lot of human time and cost. Secondly, due to the high level of human factor involvement in decision making, it is not able to account for all the factors affecting a crime and communication between them and slow down the stages of crime detection (detection, detection and prevention). Such a situation further expands the necessity of using an

intelligent method to detect and analyze crimes. The preparation and presentation of a model for the police system for crime analysis and detection, based on machine learning methods, can be a solution to these two challenges[4, 5]. For this purpose, we use hybrid algorithms of K-Means [6] and Artificial Bee Colony (ABC) [7] for analyzing and clustering crime in order to detect the similarity of offenses.

Data Clustering [8] is one of the uncontrolled learning branches and is a process of automation, during which samples are divided into categories whose members are similar to each other. These categories are called clusters. Therefore, a cluster is a set of samples that are similar in each cluster of samples and are not compatible with the samples in the other clusters [9]. For similarity, we can take different measures, for example, we can use the distance criterion for clustering and consider instances closer to each other as a cluster [10].

The K-Means algorithm [6] is the most widely used algorithm for detecting related data in different clusters. This algorithm attributes the data to a cluster according to the proximity (similarity), and thus each cluster is created with different data [11]. By repeating the same procedure, it can be calculated new clusters for them by averaging data in each iteration, and again it can be assigned data to new clusters. This process continues as long as there is no change in the data.

An ABC algorithm [7] starts with the initial population of random vectors. How to do this is that, in each iteration of the algorithm, the artificial bees find new solutions by performing random searches around the answers obtained in the previous iterations. Obviously, these new answers will not necessarily all be better than the answers given in previous iterations. After that each of the working artificial bees found a new answer, all of them returning to the hive again, and decide on the next path to move. Therefore, the optimality of the answers obtained by the bees is calculated, and then the answer that has more fit is chosen as a search path in the next iteration. Therefore, the regions surrounding the optimal responses are searched by the number of more bees in the subsequent iterations. The search process takes place as long as the required condition is fulfilled to complete the program execution.

Finding a comprehensive clustering algorithm that can solve all problems is a difficult task. The difficulty of clustering can be attributed to factors such as unsupervised nature and different goals of clustering in different sciences and applications [12]. The lack of awareness of clustering algorithms from the clustering purpose and the structure of the data causes these methods to solve problem without any knowledge of the problem and taking into account specific assumptions. And this will result in inappropriate responses in the event of mismatch between the assumption and the goal of clustering or the actual structure of the data. Nonetheless, a broad spectrum of clustering algorithms in real-world affairs, with presuppositions on the purpose of clustering and data structure, is an attempt to solve the problem. So far, different methods have been used to combine ABC and k-means [13-16]. In this paper we use an ABC algorithm to improve K-Means. The main contributions of this paper are as follows:

- Hybrid clustering approach based on K-means and ABC algorithm is proposed for optimal cluster analysis.
- In this work, proposed model is used for data clustering on Communities and Crime dataset and the performance of ABC algorithm is compared with k-means.
- ABC is proposed to overcome local optima traps of K-Means clustering.

- The ABC algorithm is applied to obtain the optimal clusters center.

The overall structure of this paper is organized as follows: In Section 2, we will explain the research done. In the Section 3, we explain the ABC algorithm. In Section 4, we describe the proposed model and its steps. In the Section 5, we will explain the evaluation and the results of the proposed model and compare it with other models, and finally, in the Section 6, we will draw conclusions and future work.

2. Related Works

What is important in clustering is to minimize data that has not been properly written, or to classify data in each class that has the closest similarity to each other. Several investigations have recently been launched to accelerate the discovery and clustering of crime in the field of criminology. Data mining technicians have attracted the attention of many researchers due to learning and teaching data. These methods, with minimal user interference, automatically express logical patterns and relationships [17]. This section addresses the various data mining techniques and patterns and how each of these algorithms is used.

Fuzzy association rules have been proposed using the Apriori algorithm for clustering crimes [18]. The fuzzy hybrid model is used for grading crimes and uses the Apriori algorithm to create rules. The most important thing in exploring dependency rules is to find a collection of repetitive items. One of the most basic algorithms in the field of exploring repetitive items is the Apriori algorithm. The hybrid model tries to find subsets of the samples, which are at least between the C sample collections. Apriori is an up-and-down algorithm, as each instance adds a sample to repeated subsets. The evaluation on Communities and Crimes data collection has been done with 40 features. The results show that the diagnostic accuracy in the hybrid model is 60%. The most important advantage of this model is that it uses fuzzy rules to explore the relationship between samples for clustering. The important advantage of the Apriori algorithm is to find dependencies between different sets of data. Also, the fundamental disadvantage of this model is a lot of computational time to explore the relationship between different fuzzy rules.

Data mining methods such as Linear Regression (LR), decision trees for clustering offenses have been used [19]. The evaluation on the Community and Crime dataset has been normalized with 2215 instances of crime and in the vice environment. Weka software includes a set of machine learning algorithms and data preprocessing tools. This software is designed to be able to quickly test existing methods in a flexible way on the data set. The LR is used to model the value of a dependent variable, whose linear relationship is based on one or more predictors. The results based on the efficiency coefficients, Mean Absolute Error, Root Mean Squared Error, Absolute Relative Error, and Relative Error of Root Square showed that the accuracy of LR detection is higher than the decision tree model and also the mean value of the absolute error in it compared to the tree decisions are less. The main disadvantage of this model is that it uses data mining methods unchanged in their parameters and calls for the usual mode of algorithms in the WEKA software. The advantage of this model is that it has used an incremental regression for crime analysis, which this regression model can separate the data more accurately.

Apriori algorithms and FP-Growth have been used to explore association rules on the prison dataset [20]. The FP-Growth algorithm is one of the search algorithms for associative rules. This algorithm stores the data in the dataset in a compact form in a tree

called FP-tree, and then retrieves samples using FP-tree returns. The evaluation was done on 72 samples. The results show that the accuracy, call and F-Measure criteria of FP-Growth algorithm are more than the Apriori. The most important advantage of the hybrid model is that the FP-Growth algorithm uses a strategy to split and overcome to explore the rules. The main problem with this model is that with a variety of features, the size of the tree will be very large, and so the processing time will be high.

The clustering of crimes was done using the K-Means algorithm [21]. The evaluation has been carried out in WEKA environment on the datasets of England and Wales between 1990 and 2011. In the K-Means algorithm, the k-member (k is the number of clusters) is randomly selected from among the n members as cluster centers. Then the n-k remaining members are assigned to the nearest cluster. After assigning all members, the cluster centers are recalculated and assigned to the clusters according to the new centers, and this continues until the centers of the clusters stagnate. The results show that in 2002 the highest number of crimes was committed. The disadvantages of this model are that clustering process has been done using the Rapid Miner software, and so the center of clusters is not at all paid attention. The advantage of this model is that normalization operations and ambiguous values for features are replaced.

The clustering of crimes was done using the hybrid of Genetic Algorithm (GA) and K-means [22]. The evaluation was made on the data collection between England and Wales between 1990 and 2011 in the Rapid Miner environment. The GA is used to optimize K-means. The characteristics that their iterations are too high are assigned a weight, and the initial population of the GA is based on the initial weight of the characteristics. The goal of the hybrid algorithm is K-Means and GA for leaving local optimum points. Using the GA to identify the important points among the samples. In this model, a new method for crossover and mutation operators is presented. The logic of the hybrid model is based on the assumption that the crossover and mutation operators can detect similar properties in a finely defined region based on the weights, instead of being randomly applied throughout the whole response. The results showed that the accuracy criterion in non-optimal state was 85.74% and in optimal mode, it was 91.64%. As a result, the GA has been able to increase the accuracy of clustering. One of the main advantages of this model is the operation of a GA that simultaneously detects several centers of the cluster and then locates the clusters on the basis of the distance between the centers for assigning the samples.

Based on the identification of the characteristics and characters of the perpetrators, the Apriori hybrid model and K-Nearest Neighbor have been proposed for prediction of crime [23]. To assess, the INSCR dataset, which includes 1966 samples of murders and violence in India has been used. The results show that the prediction of the hybrid model for the year of 2009 has decreased when it is compared with the data for the years 2001 to 2008. The advantage of this model is that the features that are effective in determining the accuracy of the selection are selected, and KNN algorithm assigns each instance to a similar category based on them. The advantage of the KNN algorithm in this model is that a group of K samples are selected from the set of training samples and then they are decided on the category of test samples based on their grade or label. The disadvantage of this model is that choosing the K value in this way is very important and key. If the K value is too small, the algorithm becomes sensitive to noise. In fact, the noise makes the clustering process not done accurately. If a very large K value is selected, there may be records from other clusters among the closest neighbors.

Models of Support Vector Machine (SVM), J48, Artificial Neural Networks (ANN), and KNN have been proposed for classifying and identifying risky points of crime [24]. The performance of each model is evaluated using F-Measure and accuracy criteria for different months. Training the artificial neural network with the SVM is relatively simpler. It also works well for high-dimensional data. Only the most important factor in a SVM is a proper learning function. The design of the J48 decision tree is more difficult compared to the ANN and SVM. Among the above-mentioned algorithms, the KNN algorithm has a simpler computational structure.

The Fuzzy C-Means (FCM) model is proposed to cluster crime patterns obtained from USA with 70000 stealing samples in 2015 in San Francisco [25]. In cluster patterns, the number of crimes has been evaluated for different days. The results show that the FCM model has high accuracy in clustering and the similarity of data between each cluster. The main advantage of this model is that it has a minimum distance between clusters. The distance between centers of clusters in each iteration is evaluated and therefore cluster headings are created that are less distant than other neighbors.

The hybrid model of the Ant Colony Optimization (ACO) and K-Means algorithm [26] is proposed for the detection of crimes. The ACO algorithm is used to find the nearest edge between clusters. Using the ACO algorithm, the ants detect a shorter path based on pheromone and find the optimum distance between the clusters. And K-Means are used for the similarity between the data of each cluster. The results are modeled and the advantages can be deduced from the low computational time. The disadvantage of this model is that if the parameters of the ACO algorithm such as pheromone grazing and sight are not exact, the answer to the problem is not optimal, and therefore the accuracy of the clustering will be low. The advantage of this model is that no computational time is high, and also the accuracy of the k-means algorithm will be higher because similar samples are almost discovered.

Models of Naïve Bayes (NB) and back-propagation ANN have been tested and implemented to categorize crimes on 2000 samples of crime [27]. The NB model is one of the algorithms for classification and the back-propagation ANN of one of the ANNs models. In the ANN model, back-propagation training and testing of data are repeated in several steps up to reach optimal value. In the NB model, Euclidean distance is used and in the model of ANN, the back-propagation of weight is used for classification. The results show that the NB model is more accurate than the back-propagation ANN. The accuracy value for the NB model is 90.22% and 94.08%, respectively. The most important disadvantage of this model is that the amount of neuron in the ANN must be accurate to obtain the correct results. In this study, the amount of neuronal weight was not investigated, and they were also given directly as an input to the ANN. The advantage of this model is in the NB Algorithm; in the case that even if the number of teaching samples is low, the NB Algorithm can map the relationships between the samples Base of probability.

A new hybrid method of Enhanced ABC algorithm and K -means (EABCK) in order to improve the performance of K -means clustering has been proposed [13]. The best solution is updated by K -means in each iteration for data clustering. Evaluate the performance of EABCK on eleven benchmark datasets were used. The obtained results showed that EABCK has had a good performance. In [14] is proposed a method based on K-Means and ABC algorithm for segmentation. Obtained results demonstrated on satellite images that the performance of the proposed method is better than the other methods.

Elephant Herding Optimization (EHO) Algorithm and k-modes has been proposed used for clustering and detecting the crime [28]. The proposed model consists of two basic steps: First, the cluster centrality should be detected for optimized clustering; in this regard, the EHO Algorithm is applied. Second, k-modes are applied to find the clusters of crimes with close similarity criteria based on distance. The results showed that purity of the proposed model is equal to 0.9145 for 400 iterations.

In Table (1), the comparison of proposed models by researchers is shown on the basis of important factors.

Table 1. Comparison of Proposed Models by Researchers to Crime Detection

Refs	Models	Datasets	# Records	Feature Selection
[18]	Fuzzy Apriori	Communities and Crime	1994	√
[19]	Linear Regression	Communities and Crime	1994	X
	Additive Regression			
	Decision Stump			
[20]	FP-Growth	Rosewood crimes data	72	X
	Apriori			
[21]	K-Means	England and Wales	600	X
[22]	K-Means+ GA	England and Wales	600	√
[23]	KNN	INSCR	1966	X
	Apriori			
[24]	SVM	city's police department	2000	√
	J48			
	ANN			
	KNN			
[25]	FCM	robberies in San Francisco	70000	X
[26]	K-Means-ACO	Crime INDIA	-	X
[27]	NB	Communities and Crime	1994	X
[13]	EABCK	Different benchmarks	1000	X
[14]	K-Means-ABC	Satellite Images	100	X
[28]	EHO- k-modes	Communities and Crime	1994	√

3. Artificial Bee Colony Algorithm

In the ABC algorithm [7] with an initial population of random answers, the search begins (first step). Then, using of a process iteration, it tries to improve random responses. The ABC algorithm consists of the following phases: initialization, employed bee phase, computation of the probabilities for onlooker bees, onlooker bee phase and scout bee phase, as described below[29, 30].

Employed bees: At this point, artificial bees, looking around the food at point xi, look for better food supplies in new vi situations. Determining the new position of the food source is carried out using Eq. (1).

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \tag{1}$$

Where $k \in \{1,2,\ldots,SN\}$ and $j \in \{1,2,\ldots,D\}$ are randomly chosen while *SN* is the number of food sources and *D* denotes the problem dimension. Based on Eq. (1) k has to be different from i, also ϕ_{ij} is a uniform random number in the range of [-1, 1]. If the new location value v_{ij} is better than x_{ij} the position is updated and replaced with the older value. In Eq. (1), the vector of the new position of the bees, the vector of the position of the bee, is an integer random number in the interval [1, SN], where *SN* is equal to the number of artificial bees. Equal to the dimensions of the solutions. The parameter is a random number with uniform distribution in the interval [-1, 1]. The choice of the random number x_i of the problem domain is made using Eq. (2).

$$x_{ij} = L_j + rand\ (0,1) \times (U_j - L_j) \tag{2}$$

In Eq. (2), U_i and L_i are the upper and lower bounds of the variable x_i, and rand () is also a function of random numbers in the interval (0, 1). After determining the position of the new food source, should be calculated its optimal amount. For this purpose, the fitness vector x_i is defined according to Eq. (3), and finally, greedy selection is made between the current data and its neighbor.

$$fit_i = \begin{cases} \dfrac{1}{1+f_i} & f_i \geq 0 \\ 1 + abs(f_i) & f_i < 0 \end{cases} \tag{3}$$

In Eq. (3), f_i is the value of the target function for the *ith* data sample. The work bees bring information about the specimens in the hive through the dance to the onlooker bees, which this action is done in the algorithm by assigning the probability to each instance using Eq. (4).

Onlooker bees: At this stage, each onlooker bee uses an Eq. (1) with a probability-based choice to search around food sources. And the neighbors are looking for the best. (Onlooker bees make their choice based on the probable values of the working bees). Therefore, the probability of choosing the food source by the onlooker bees is carried out using Eq. (4). A higher probability sample has more chances to choose. After selecting a neighbor, they compare it with the current sample if it is more fitting. In the vector, the population is replaced by i. If the counter is replaced, it is zero and otherwise, one unit is added to the counter.

$$p_i = \frac{fit_i}{\sum_{j=1}^{SN} fit_i} \tag{4}$$

After that the all onlooker bees have selected food supplies using Eq. (4), the neighboring food source's position is calculated using the new food source position, and this process continues until the condition is fulfilled to complete the program's execution.

Scout Bees: In an ABC algorithm to increase the efficiency of the search process and not to fall into the optimal local trap, if there is no optimal answer after a predetermined number of repetitions (the sample counter exceeds the specified value), A number of bees have set aside their solutions, and by turning into scout bees, they again search for randomly in the domain of the problem using Eq. (2). And they remove the current sample, replace it with a zero counter and a new sample. The implementation of the stage of the scout bees can dramatically increase the likelihood of a global optimal answer.

4. Proposed Model

The proposed model is a hybrid of K-Means and ABC algorithm that is used for ABC to optimize K-Means clustering. The K-Means algorithm has weaknesses such as finding the center of the cluster, which uses ABC algorithm to try to fix it and then optimize it for crime clustering by applying changes to the combined algorithm. In this paper, ABC algorithm is used to cluster n samples in a d-dimensional space to k clusters. Initially, the data is read and become ready for its preprocessing stage. Preprocessing is done on the Communities and Crime [31] data set. In the preprocessing step, the normalization of the data is done for compatibility. Then data clustering is performed based on a hybrid model. In Figure (1), the flowchart of the proposed model is shown. The proposed model consists of two parts, ABC and k-means. In the ABC section, the best value is found in each vector as the center of the cluster for k-means. Iterations of the ABC algorithm are effective in finding the best center of the clusters.

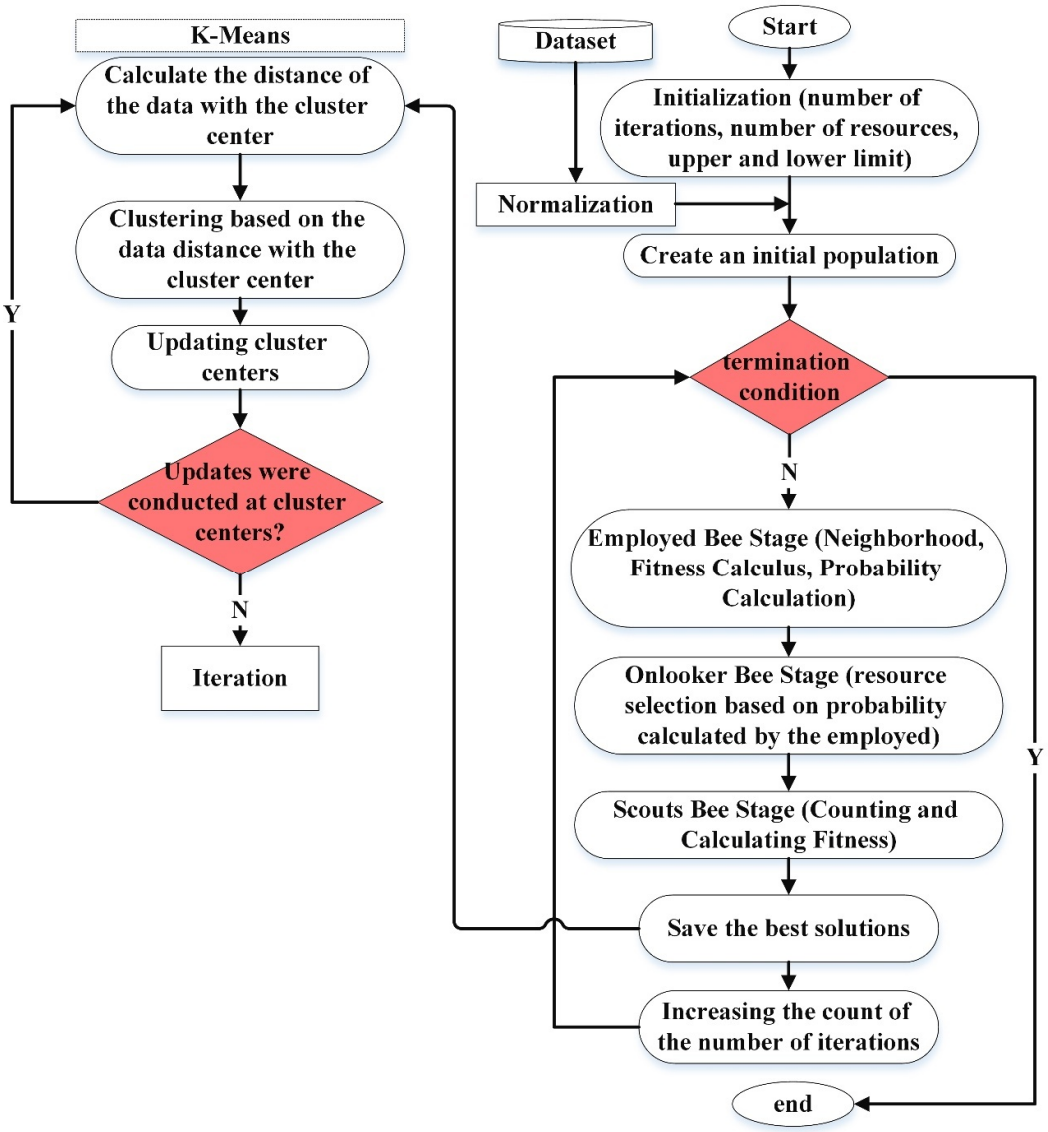

Figure 1. Flowchart of the proposed model

The pseudocode of the proposed model is as follows:

- Dataset
- Normalization
- Creating a Initialize Population
- Formation of early vectors
- Allocation of food supplies to bees
- Repeat the program until the termination condition has been established.
- Employed bees: Search among food sources
- Onlooker bees: Search among food sources
- Scouts: Search among food sources
- Save the best solutions
- Increase the count

- Discover the bees that have the best k value.
- Bees that have discovered a point and have more neighbors around it.
- Find the best points and calculate them based on the distance k-means
- End

4.1. Normalization

The lack of preprocessing in bulk data leads to serious bugs in assessing the identification and discovery of hidden relationships and patterns. The main preprocessing operation is to clear data that is used to eliminate uncertainty and includes estimating unnecessary values in the data set, eliminating noise in the data, removing irregular data, and eliminating incompatibility in the data. The better this step is to be, the output of the algorithm will have a higher quality. It is also possible that all data is not needed, and only a fraction of the data that needs to be processed should be processed.

Since the properties of the dataset have different values, for the purpose of unification, the value of the attributes is converted to a distance between 0 and 1. In the normalization process, the smallest number associated with each attribute is displayed with min and the largest number with max. And x is the value of each attribute. The normalization operation is carried out according to Eq. (5). The new_max and new_min values are 1 and 0 respectively [32].

$$y=\frac{x-\min}{\max-\min}\times[new\_\max-new\_\min]+new\_\min \tag{5}$$

4.2. Hybrid Model

One of the most suitable methods for data clustering is the K-Means algorithm. This algorithm specifies clusters based on minimizing the sum of square squares, the least distance of points to the centers of the clusters[33]. In this manner, the algorithm has different clustering results based on the number of clusters and early cluster centers. This algorithm works on clustering based on minimizing and searching space of states, but it does not have the ability to search optimally and globally in the cluster state space. The k-means algorithm has the following disadvantages:

- The dependence of the results on the selection of cluster centers
- Stagnation in a local minimum
- Determine of K (number of clusters) that the user should determine at the beginning of the algorithm's process, and the appropriate and precise method for determining it is not specified.
- It is sensitive to noise and data that is far from the cluster centers. Because the data can easily change in cluster centers and give poor results to the algorithm.

Therefore, in this paper we use the ABC algorithm to improve K-Means. The K-Means algorithm starts from k random point in state space, and updates the cluster centers with duplicate data to the nearest attribution cluster. Continue this operation to the point where no improvement is possible.

In the proposed model, every food source in the search space is displayed as a cluster center. This source is likely to be selected as a solution for the cluster center.

$$x_i = (C_{i1,} C_{i2},..., C_{ik})$$

(6)

In Eq. (6), k is the number of clusters discovered by the bees, C_{ik} means the i[th] center of the k-food source. Each bee is rooted around food sources and compares them to nearby neighbors. A resource that has more neighbors and fewer spaces is chosen as k. The fitness function of the bees is calculated according to Eq. (7).

$$x_i = \frac{\sum_{j=1}^{k} \sum_{x_{ij} \in c_{ij}} d(x_p, c_{ij}) / \left| c_{ij} \right|}{k}$$

(7)

In Eq. (7), x_p is the location of the bees in the vector of characteristics, c_{ij} is the center of the cluster discovered by the bees, d is the distance and k is the number of clusters. The number of clusters was detected by the bees and therefore the position of each bee is evaluated as the center of the cluster. By Eq. (7), the distance between cluster centers is also determined, and the distance below indicates that the best centers have been selected.

The hybrid model is a new algorithm for finding the best cluster centers in K-Means, in the ABC algorithm everywhere in the problem space as a food source, is evaluated. Scouting bees randomly evaluate the response space and report the quality of the visited positions by the fitness function. Then the answers are ranked. The other category, which is the onlooker bees and is considered to be a motivating stimulus, searches for the space of responses around the superior answers of the worker's bees, with a local search to improve the answers.

The onlooker bees search for a food source adjacent to the selected food source to the previous stage according to Eq. (1). If a source search is over, the worker bee leaves the source and becomes a scout and begins to search for a food source randomly according to Eq. (8). This means that the bees are trapped in a local optimal way, and therefore that point is eliminated, and a new point is randomly generated between the current cluster center and the best point near it.

$$x_i = x_{\min} + \phi\ (x_{\max} - x_{\min}), where \rightarrow \phi \in [0,1]$$

(8)

The algorithm selectively searches the search space for finding the optimal point to converge to the best points and find the best centers.

4.3. Application of Bees in Clustering

1) At first, the food resource positions are randomly determined and then using the K-Means algorithm, clustering is performed for each of the situations (data) and according to Eq. (9), the cluster fitness is calculated.

In Eq. (5), the parameter k is the number of points in the cluster centers and n the number of attributes. Each sample is attributed to the cluster whose center of the cluster has the smallest distance to that data. And || || is the criterion of the distance from the point x to c_j (center of clustering jth). The best method of clustering is to minimize the total similarity between the cluster center and all cluster members and minimize the total similarity between cluster centers.

$$E = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \tag{9}$$

2) Using employed bees, food sources are searched, and the location of the resources is updated. Then, the K-Means algorithm is applied and by calculating the bee fitness according to Eq. (7), the best source is introduced as the observer bee.
3) Calculate the probable values of food sources and update their position based on the values obtained from the onlooker bees.
4) Count food sources and produce a new food source in search space, and update each resource.
5) Repeat the steps from step two until the end condition is completed.

The purpose of the hybrid of K-Means algorithms and ABC is to solve the problem of the k-number for the K-means algorithm and the dependence of the K-means result on the initial values of the cluster centers. However, considering that the purpose of this paper is to improve the results of the crime clustering with the two mentioned algorithms, therefore, we need to make some changes to obtain a better result than the hybrid of the two algorithms.

- For each sample of the dataset, the target vector is created by the ABC algorithm and distributed as a random source at the beginning of the hybrid model in the problem space.
- For each source, a pointer is considered to represent the type of food source.
- Based on the ABC, food sources are calculated based on the fitness function and the finer amount is stored as the best value. By doing so, the center of the cluster is essentially determined for the K-Means algorithm.
- By finding the center of the cluster, then based on the Euclidean distance, the position of each source is calculated to be assigned to the clusters and each food source is clustered in its own cluster.
- The proposed model aims to provide an optimal solution and solve the following problems:
- The problem of the dependence of the result of k-means on the initial values of cluster centers
- The problem of determining the centrality of the k-means algorithm indirectly.

4.4. Evaluation criteria

To evaluate the proposed model, validation criteria should be used. Eq. (10) measures the purity of the clusters [34]. The purity criterion is used to measure clusters and distribute them.

$Purity = \frac{1}{c}\sum_{i=1}^{c}\frac{C_i^d}{C_i}$	(10)
$P(i, j) = \frac{n_{ij}}{n_j}$	(11)
$R(i, j) = \frac{n_{ij}}{ni}$	(12)
$F - Measure = \frac{2 * P * R}{P + R}$	(13)
$RI = \frac{a + d}{a + b + c + d}$	(14)

In Eq. (10), c is the number of clusters, C_i is the number of samples in the cluster ith, the number of data that is properly grouped in the ith cluster. In the defined equations, the parameters of ni, nj and nij are the number of class i data, the number of j cluster data (obtained by the clustering algorithm), and the number of class i data in the j cluster [35], respectively. In Eq. (14), the parameters of a, b, c, d are equal to the number of correctly positive samples, wrongly positive, wrongly negative and correctly negative [36]. The RI (Row Index) has a value between 0 and 1 and is close to 1 in the optimal range.

5. Evaluation and Results

The evaluation and results of the proposed model have been done in the Visual C#.NET 2017. Experiments is done on the Communities and Crime dataset [31]. To simulate, K-Means functions such as finding cluster centers, sample clustering and clustering of ABC for optimization of cluster centers and similarity between samples have been used. Determining the number of parameters in achieving optimal solution is very important. Therefore, the number of initial population and the number of iterations in the ABC is 50 and 200, respectively.

5.1. Number of Iteration

In this section, the proposed model is evaluated based on the number of iterations of the ABC algorithm. The results of Table (2) show that the purity of the clustering in the proposed model has increased with increasing number of iterations. Bees are more accurately evaluating the discovery of food sources and neighbors of each food source in 200 iterations. The purity value in the proposed model for iterations of 100 and 200 is 0.8516 and 0.8630, respectively. Table (2) shows a comparison of the proposed model with k-means and Fuzzy C-means (FCM) [37]. FCM clustering is one of the segmentation methods applied for analysis of data and information. FCM algorithm acts by classified membership to each of the data point based on each cluster center. The results of Table (2) show that the purity values in the EHO-K-modes [28] for 100 and 200 iterations are 0.8026 and 0.8262, respectively.

Table 2. Evaluation of the proposed model based on the number of iterations

Iterations	Models	Purity	P	R	F-Measure	RI
100	k-means	0.7831	75.31	76.04	75.61	81.30
	FCM	0.8171	75.92	77.15	76.53	82.68
	EHO-K-modes [28]	0.8026	78.28	79.05	78.65	80.32
	Proposed Model	**0.8534**	76.10	79.61	77.82	**85.40**
200	k-means	0.7831	75.31	76.04	75.61	81.30
	FCM	0.8171	75.92	77.15	76.53	82.68
	EHO-K-modes [28]	0.8262	80.55	80.92	80.73	81.26
	Proposed Model	**0.8637**	81.37	84.31	82.81	**87.90**

In Figure (2), the comparison diagram of the proposed model with the K-Means, FCM, and EHO-K-modes is shown based on 100 iterations. In Figure (2), it is clear that the proposed model has been able to increase the P, RI in comparison with k-means, FCM, and EHO-K-modes. Because the proposed model has been able to improve the clustering centers in the K-Means model using repeatability and make it more precise and closer to finding similar samples.

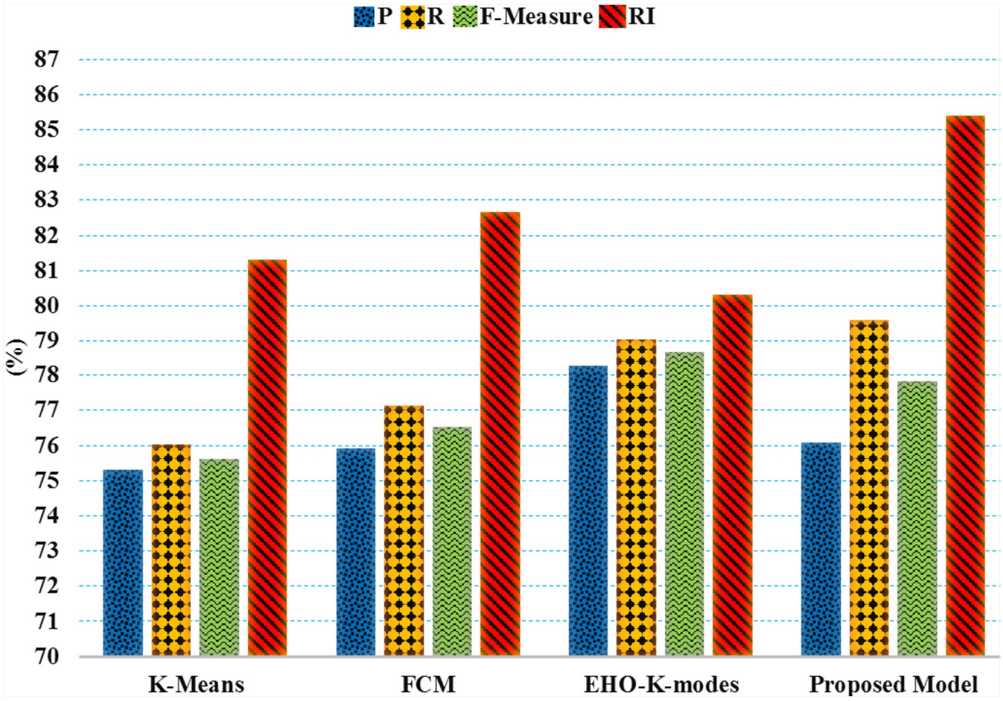

Figure 2. Comparison of the proposed model with the K-Means and FCM based on 100 iterations

In Figure (3), the comparison of the proposed model with the K-Means, FCM, and EHO-K-modes based on 200 iterations is shown. In Figure (3), it is clear that the proposed model has been able to increase the P, R, F-Measure, and RI. With 200 iterations, there are a number of more comparisons to detect data relative to the center of the cluster, and similar data are more likely to be grouped into similar clusters.

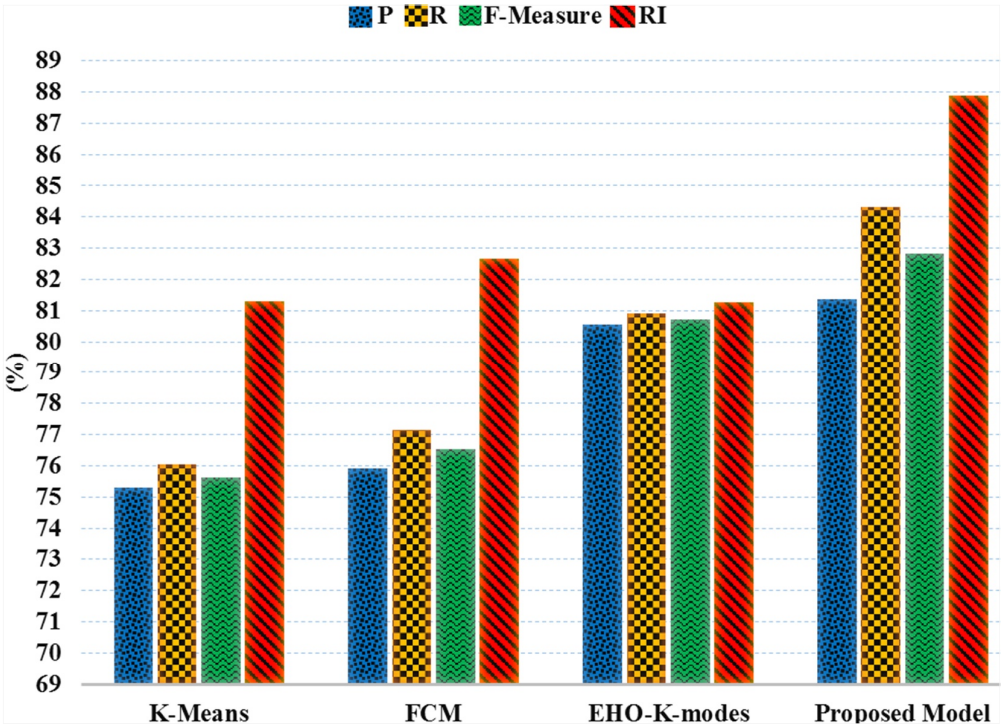

Figure 3. Comparison of the proposed model with the K-Means model based on 200 iterations

In Figure (4), the comparison diagram of the proposed model is based on 100 and 200 iterations has been shown. In Figure (4), comparisons based on different criteria show that the proposed model with 200 iterations has better accuracy.

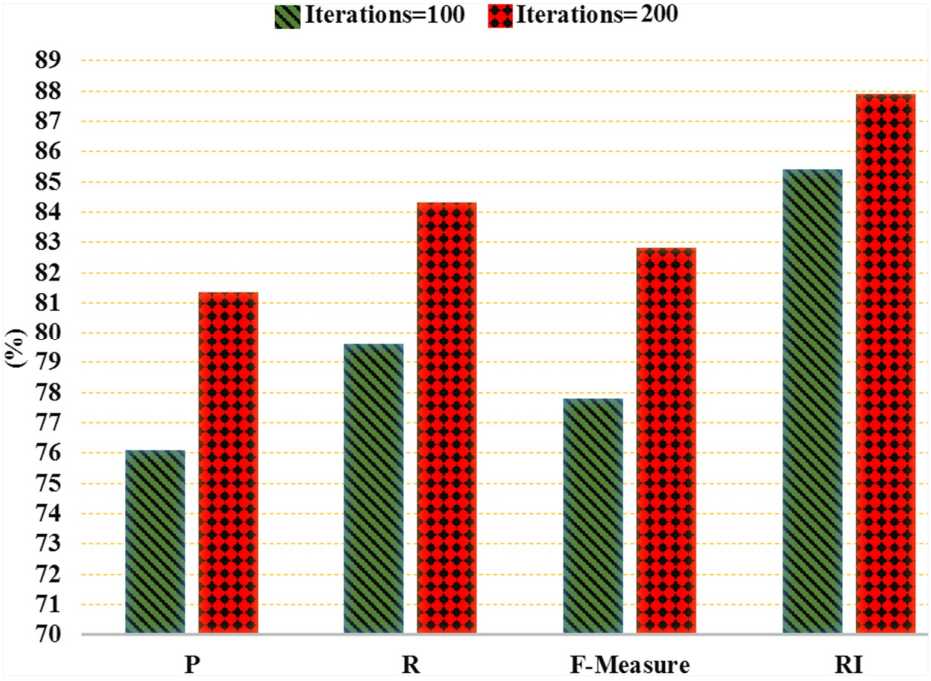

Figure 4. Comparison of the proposed model based on the number of iterations

Table (3) shows the results of the proposed model based on the iteration and different values of the parameter φ. The parameter φ represents the distribution rate of the bees in the search space. In Table (3) it is clear that the proposed model with 500 iterations and the distribution rate of 0.4 has a better purity value than other scenarios. If the distribution rate of the bees is higher, they can better search for optimum spots in the search space and are intended to discover cluster centers. The highest RI values for 100, 200, and 500 iterations are 89.8, 90.3, and 93.4, respectively.

Table 3. The results of the proposed model based on the iteration and distribution rate of the bees

Parameters and criteria						
Iterations	**φ**	**Purity**	**P**	**R**	**F-Measure**	**RI**
100	0.1	0.8410	80.32	82.31	81.30	76.0
	0.2	0.8563	81.59	82.55	82.07	87.1
	0.4	**0.9611**	83.09	83.98	83.49	89.7
	0.6	0.8310	82.45	82.88	82.66	88.3
	0.8	0.8642	82.64	83.06	82.85	86.5
	1	0.8549	81.11	82.54	81.82	89.3
200	0.1	0.8830	85.34	86.31	85.82	89.3
	0.2	0.8935	86.91	87.95	87.43	89.8
	0.4	0.8960	88.16	89.63	88.89	90.3
	0.6	**0.9080**	89.52	89.67	89.59	89.6
	0.8	0.8712	86.30	87.41	86.85	88.1
	1	0.8901	88.17	89.15	88.66	89.4
500	0.1	0.9132	89.62	90.31	89.96	90.3
	0.2	0.9259	86.31	86.97	86.64	92.6
	0.4	0.9340	89.05	90.63	89.83	93.4
	0.6	**0.9431**	88.31	89.52	88.91	92.1
	0.8	0.9284	90.67	90.72	90.69	92.3
	1	0.9200	89.15	90.25	89.80	91.8

Figure (5) shows the comparison diagram of the proposed model based on the iteration and the different values of the parameter φ. In Figure (5), it is clear that the amount of purity with 500 iterations is higher than the other cases. Because the search space is well searched by the bees to find the center of the clusters.

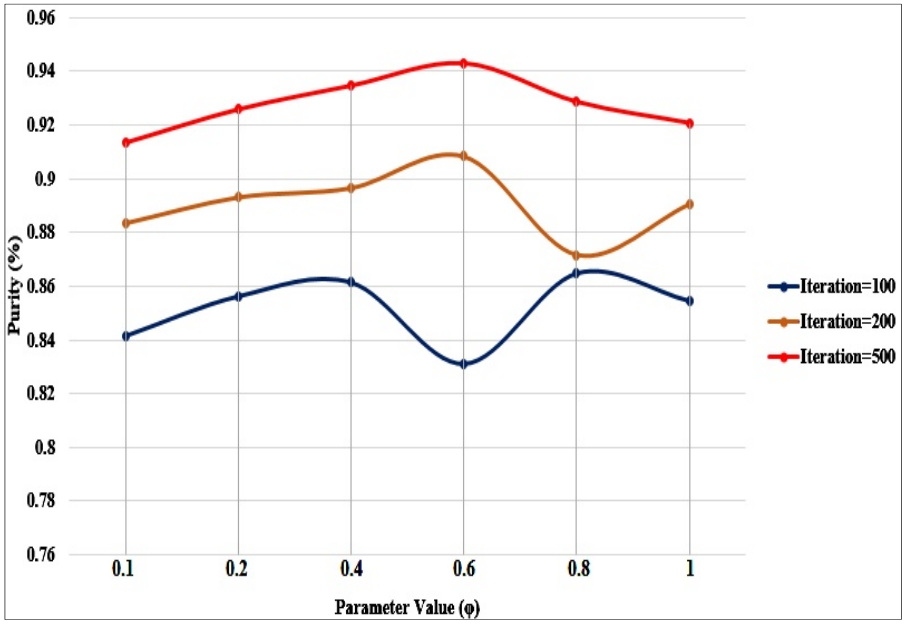

Figure 5: Comparison of the proposed model based on the repetition and different values of the distribution parameter of the bees

5.2. Evaluation of Features

In this section, the proposed model is based on the Feature Selection (FS). The FS has been done in three steps. The purpose of this section is to evaluate the impact of the number of features on the proposed model. The ABC algorithm performs better in a more limited search space. Therefore, it is examined this factor with the number of features and see that if the number of features is less, the value of Purity will be more. The results of Table 4 show that when the number of attributes is 40, the purity of the proposed model is 0.9436, and if the number of features is equal to 80 and 120, the purity value is equal to 0.9230 and 0.9065, respectively. If there are fewer features, searching in the sample space is done better.

Table 4. Evaluation of the proposed model based on the number of attributes

criteria	Features=40		Features=80		Features=120	
	K-Means	Proposed Model	K-Means	Proposed Model	K-Means	Proposed Model
Purity	0.9121	**0.9436**	0.9015	**0.9230**	0.8906	**0.9065**
P	87.47	89.48	86.79	87.69	83.94	85.14
R	89.16	90.15	87.00	89.37	84.15	87.80
F-Measure	88.31	89.99	86.89	88.52	84.04	86.45
RI	90.60	92.80	89.10	90.80	86.10	88.11

In Figure (6), the chart of comparison of the proposed model with the k-means is shown based on the number of features. Comparisons in Figure (6) show that if the number of features is less, the proposed model is more accurate. Because it can optimize

the optimum points for sample clustering, and as a result, clusters are more likely to contain similar data.

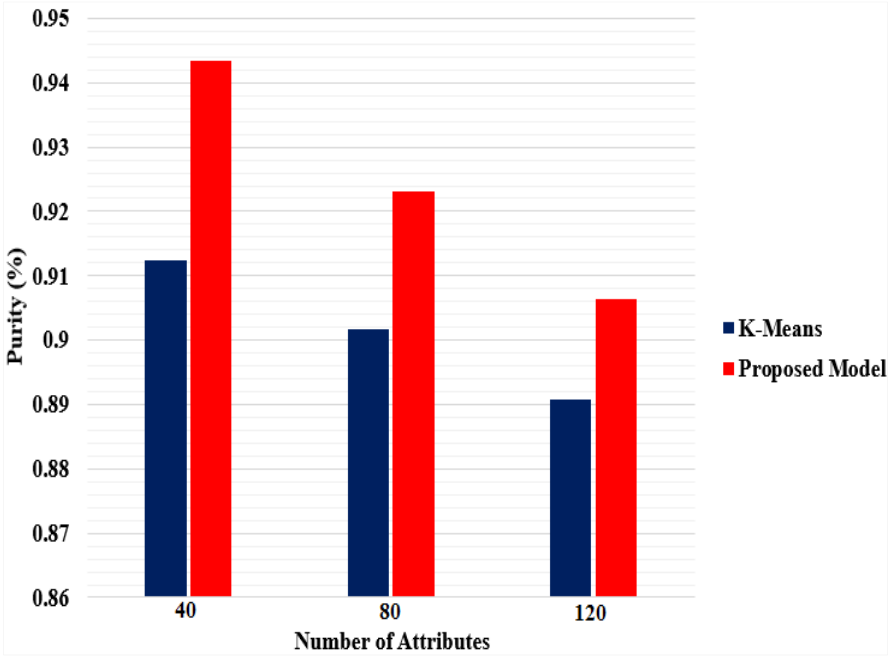

Figure 6. Comparison of the Proposed Model based on the Number of Attributes

5.3. Evaluating the Number of Clusters

In Table (5), the results of the proposed model are compared based on the number of iterations and the number of clusters with the k-means model. In the proposed model, the bees based on searches in the food resources space, the most frequent cases in which the program was discovered for iteration of the k-value were 10, 15 and 20. Therefore, these three values are found for *k* and the results are based on these three values.
The proposed model is better than k-means in the repetitions of 100, 150 and 200. The highest purity in the proposed model for repetitions of 100, 150 and 200 is 0.8532, 0.8937 and 0.9411, respectively. Also, the purity value in the k-means model for repetitions of 100, 150 and 200 is equal to 0.7913, 0.8632 and 0.8920 respectively. Also, the RI is a measure of optimality. In most cases, its value is higher in the proposed model, indicating a correlation between the points of the clusters and the similarity between the cluster dormancy. If the number of clusters is increasing, the performance of the models is harder to detect the data, and as a result, the accuracy of the diagnosis decreases. Because the points become different and the difference between the data increases.

Table 5. Evaluation of the proposed model based on the number of clusters

Iterations	K	Models	Purity	P	R	F-Measure	RI
100	10	K-Means	**0.7913**	80.14	81.54	80.83	83.1
	15		0.7832	75.31	76.04	75.61	81.3
	20		0.7621	74.31	77.15	75.70	79.3
	10	Proposed Model	0.8010	81.04	83.17	82.09	85.9
	15		**0.8532**	76.10	79.61	77.82	85.4
	20		0.8061	83.17	84.91	84.03	84.1
150	10	K-Means	0.8306	80.11	83.52	81.87	86.3
	15		**0.8632**	82.39	84.96	83.66	85.3
	20		0.8515	79.30	81.37	80.32	84.9
	10	Proposed Model	0.8639	83.67	85.11	84.38	89.3
	15		0.8801	84.31	86.81	85.54	90.5
	20		**0.8937**	80.97	81.23	81.10	89.0
200	10	K-Means	0.8618	81.20	85.90	83.48	86.1
	15		0.7832	75.31	76.04	75.61	81.3
	20		**0.8920**	83.66	83.90	83.78	84.0
	10	Proposed Model	**0.9411**	89.11	92.46	90.75	90.6
	15		0.8630	81.37	84.31	82.81	87.9
	20		0.8905	86.20	89.15	87.65	89.3

5.4. Comparison and Evaluation

In this section, the comparison of the proposed model with other models is based on the purity criterion. As you can see in Table (6), the proposed model is more accurate than other models. The purity of the proposed model is based on different metrics. In order to accurate comparisons of iterations and similar runs is used. Naïve Bayes (NB) and Artificial Neural Network (ANN) are prediction algorithms for detecting samples into separate categories. NB is a simple probabilistic classifier based on applying Bayes' theorem with strong (Naïve) independence assumption.

Table 6. Comparison of the proposed model with other models

Refs	Models	Fitness Index	Purity
[18]	Fuzzy Apriori	Run 1	0.8117
		Run 2	0.7083
		Run 3	0.6647
		Run 4	0.6747
[19]	Decision Stump	Iteration=100	0.7802
[27]	NB	Run 1	0.9223
		Run 2	0.9406
	ANN	Run 1	0.6594
		Run 2	0.6592
[28]	EHO-k-modes	Iteration=400	0.9145
-	Proposed Model	Iteration=100	0.8611
		Iteration=200	0.9080
		Iteration=500	0.9431
		Feature=40	**0.9436**
		Feature=80	0.9230
		Feature=120	0.9065

Table (6) shows the purity of the models based on different performances, iterations, and the number of features. In Table (6), the NB is more accurate than other models, and its value in both cases is equal to 0.9223 and 0.9406. In the proposed model, the highest value is 0.9436 and the lowest value is equal to 0.8611. Also, the purity in the Fuzzy Apriori and Decision fuzzy models is 0.8111 and 0.7803 respectively. the obtained purity by ANN is 0.6594; by contrast, the obtained best solution by the proposed model is 0.9436. The experiments showed that the purity and RI values in the EHO-K-modes model were 0.9145 and 91.06, respectively.

6. Conclusions and Future Works

With the growing amount of data and their redundancy, they need to have the technology to group together the same data and their amount is available for evaluation. Among the data mining technicians, clustering is one of the most popular techniques for discovering clandestine data among a wealth of data. On the other hand, due to the massive growth of crimes and their relationship with each other, statistical tools have to compare all traits in the data set for the discovery of knowledge. Clustering as an unclassified method assigns data based on the attribute value to a cluster and separation takes done. In this paper, the proposed model based on K-Means and the ABC algorithm was proposed for clustering of crimes. The results indicated with 200 iterations the purity in the K-Means, FCM, EHO-K-modes, the proposed model was equal to 0.7831, 0.8171, 0.8262, and 0.8637 respectively. By selecting 40 features, the purity value for the proposed model was 0.9436. Clustering algorithms, in spite of their potential advantages in use for detection, are faced with problems such as determining the number of clusters by default and finding the center of the cluster. As a result, we must use meta- heuristic algorithms and data mining to evaluate cluster-related points. Meta-heuristic algorithms can find the similarity between them by searching between points and help centralizing the clustering.

The performance of the proposed model is evaluated in terms of the purity and RI over Communities and Crime dataset. Its performance is compared with the K-means, FCM, and the other four clustering algorithms from the literature. The experimental results confirm the effectiveness of the proposed model and show that it can successfully be applied to data clustering. We are working on the new meta-heuristic algorithms, and our future work focuses on presenting new clustering algorithms by using hybrid algorithms such as Grey Wolf Optimization, Ant Lion Optimization, and Farmland Fertility Algorithm. The goal is to solve the problems of the k-means and distance-based algorithms for clustering.

References

[1] H. K. R. ToppiReddy, B. Saini, and G. Mahajan, "Crime Prediction & Monitoring Framework Based on Spatial Analysis," *Procedia Computer Science*, vol. 132, pp. 696-705, 2018/01/01/ 2018.

[2] Q. Wang, G. Jin, X. Zhao, Y. Feng, and J. Huang, "CSAN: A neural network benchmark model for crime forecasting in spatio-temporal scale," *Knowledge-Based Systems*, vol. 189, p. 105120, 2020/02/15/ 2020.

[3] C. Catlett, E. Cesario, D. Talia, and A. Vinci, "Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments," *Pervasive and Mobile Computing*, vol. 53, pp. 62-74, 2019/02/01/ 2019.

[4] F. S. Gharehchopogh, H. Shayanfar, and H. Gholizadeh, "A comprehensive survey on symbiotic organisms search algorithms," *Artificial Intelligence Review,* pp. 1-48, 2019.

[5] F. S. Gharehchopogh and H. Gholizadeh, "A comprehensive survey: Whale Optimization Algorithm and its applications," *Swarm and Evolutionary Computation,* vol. 48, pp. 1-24, 2019.

[6] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Berkeley, Calif., 1967, pp. 281-297.

[7] D. Karaboga, "an idea based on honeybee swarm for numerical optimization," 2005.

[8] A. Allahverdipour and F. Soleimanian Gharehchopogh, "A New Hybrid Model of K-Means and Naïve Bayes Algorithms for Feature Selection in Text Documents Categorization," *Journal of Advances in Computer Research,* vol. 8, pp. 73-86, 2017.

[9] R. Jafari Jabal Kandi and F. Soleimanian Gharehchopogh, "An improved opposition-based Crow Search Algorithm for Data Clustering," *Journal of Advances in Computer Research,* vol. 11, pp. 1-10, 2020.

[10] N. Rahnema and F. S. Gharehchopogh, "An improved artificial bee colony algorithm based on whale optimization algorithm for data clustering," *Multimedia Tools and Applications,* 2020/08/26 2020.

[11] H. Rabani and F. Soleimanian Gharehchopogh, "An Optimized Firefly Algorithm based on Cellular Learning Automata for Community Detection in Social Networks," *Journal of Advances in Computer Research,* vol. 10, pp. 13-30, 2019.

[12] M. Lashkari and M. Moattar, "Improved COA with Chaotic Initialization and Intelligent Migration for Data Clustering," *Journal of AI and Data Mining,* vol. 5, pp. 293-305, 2017.

[13] D. C. Tran, Z. Wu, Z. Wang, and C. Deng, "A Novel Hybrid Data Clustering Algorithm Based on Artificial Bee Colony Algorithm and K-Means," *Chinese Journal of Electronics,* vol. 24, pp. 694-701, 2015.

[14] M. Venkata Dasu, P. V. N. Reddy, and S. Chandra Mohan Reddy, "Classification of Remote Sensing Images Based on K-Means Clustering and Artificial Bee Colony Optimization," in *Advances in Cybernetics, Cognition, and Machine Learning for Communication Technologies*, V. K. Gunjan, S. Senatore, A. Kumar, X.-Z. Gao, and S. Merugu, Eds., ed Singapore: Springer Singapore, 2020, pp. 57-65.

[15] K. Orkphol and W. Yang, "Sentiment Analysis on Microblogging with K-Means Clustering and Artificial Bee Colony," *International Journal of Computational Intelligence and Applications,* vol. 18, p. 1950017, 2019/09/01 2019.

[16] P. Das, D. K. Das, and S. Dey, "A modified Bee Colony Optimization (MBCO) and its hybridization with k-means for an application to data clustering," *Applied Soft Computing,* vol. 70, pp. 590-603, 2018/09/01/ 2018.

[17] A. Allahverdipour and F. Soleimanian Gharehchopogh, "An Improved K-Nearest Neighbor with Crow Search Algorithm for Feature Selection in Text Documents Classification," *Journal of Advances in Computer Research,* vol. 9, pp. 37-48, 2018.

[18] A. L. Buczak and C. M. Gifford, "Fuzzy association rule mining for community crime pattern discovery," presented at the ACM SIGKDD Workshop on Intelligence and Security Informatics, Washington, D.C., 2010.

[19] L. McClendon and N. Meghanathan, "Using Machine Learning Algorithms to Analyze Crime Data," *Machine Learning and Applications: An International Journal (MLAIJ),* vol. 2, pp. 1-12, 2015.

[20] R. Lawpanom and W. Songpan, "Association Rule Discovery for Rosewood Crime Arrest Planning," in *Information Science and Applications (ICISA) 2016*, Singapore, 2016, pp. 1025-1032.

[21] J. Agarwal, R. Nagpal, and R. Sehgal, "Crime Analysis using K-Means Clustering," *International Journal of Computer Applications,* vol. 83, pp. 1-4, 2013.

[22] R. Kiani, S. Mahdavi, and A. Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification," *International Journal of Advanced Research in Artificial Intelligence (IJARAI),* vol. 4, pp. 11-17, 2015.

[23] A. Malathi and S. S. Baboo, "An Enhanced Algorithm to Predict a Future Crime using Data Mining," *International Journal of Computer Applications,* vol. 21, pp. 1-6, 2011.

[24] C. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime Forecasting Using Data Mining Techniques," in *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 779-786.

[25] D. Mayorga, M. A. Melgarejo, and N. Obregon, "A Fuzzy Clustering based method for the spatiotemporal analysis of criminal patterns," in *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2016, pp. 738-744.

[26] N. Tomar and A. K. Manjhvar, "An improved optimized clustering technique for crime detection," in *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, 2016, pp. 1-5.

[27] A. Babakura, M. N. Sulaiman, and M. A. Yusuf, "Improved method of classification algorithms for crime prediction," in *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*, 2014, pp. 250-255.

[28] F. Soleimanian Gharehchopogh and S. Haggi, "An Optimization K-Modes Clustering Algorithm with Elephant Herding Optimization Algorithm for Crime Clustering," *Journal of Advances in Computer Engineering and Technology,* vol. 6, pp. 79-90, 2020.

[29] F. S. Gharehchopogh, I. Maleki, and A. Talebi, "Using hybrid model of artificial bee colony and genetic algorithms in software cost estimation," in *2015 9th International Conference on Application of Information and Communication Technologies (AICT)*, 2015, pp. 102-106.

[30] F. S. Gharehchopogh, I. Maleki, A. Kamalinia, and H. M. Zadeh, "Artificial bee colony based constructive cost model for software cost estimation," *J. Sci. Res. Dev,* vol. 1, pp. 44-51, 2014.

[31] dataset1, "http://mlr.cs.umass.edu/ml/datasets/Communities+and+Crime."

[32] C. Jin and S.-W. Jin, "Prediction approach of software fault-proneness based on hybrid artificial neural network and quantum particle swarm optimization," *Applied Soft Computing,* vol. 35, pp. 717-725, 2015/10/01/ 2015.

[33] F. S. Gharehchopogh, N. Jabbari, and Z. G. Azar, "Evaluation of fuzzy k-means and k-means clustering algorithms in intrusion detection systems," *International Journal of Scientific & Technology Research,* vol. 1, 2012.

[34] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of Internal Clustering Validation Measures," in *2010 IEEE International Conference on Data Mining*, 2010, pp. 911-916.

[35] L. Jing, M. K. Ng, and J. Z. Huang, "An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data," *IEEE Transactions on Knowledge and Data Engineering,* vol. 19, pp. 1026-1041, 2007.

[36] E. Hullermeier, M. Rifqi, S. Henzgen, and R. Senge, "Comparing Fuzzy Partitions: A Generalization of the Rand Index and Related Measures," *IEEE Transactions on Fuzzy Systems,* vol. 20, pp. 546-556, 2012.

[37] A. Gosain and S. Dahiya, "Performance Analysis of Various Fuzzy Clustering Algorithms: A Review," *Procedia Computer Science,* vol. 79, pp. 100-111, 2016/01/01/ 2016.