An Optimization K-Modes Clustering Algorithm with Elephant Herding Optimization Algorithm for Crime Clustering

***Farhad Soleimanian Gharehchopogh*[1], *Sevda Haggi*[2]**
1,2- Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, IRAN.
(bonab.farhad@gmail.com)

Received (2019-10-05) Accepted (2020-04-26)

Abstract: *The detection and prevention of crime, in the past few decades, required several years of research and analysis. However, today, thanks to smart systems based on data mining techniques, it is possible to detect and prevent crime in a considerably less time. Classification and clustering-based smart techniques can classify and cluster the crime-related samples. The most important factor in the clustering technique is to find the centrality of the clusters and the distance between the samples of each cluster and the center of the cluster. The problem with clustering techniques, such as k-modes, is the failure to precisely detect the centrality of clusters. Therefore, in this paper, Elephant Herding Optimization (EHO) Algorithm and k-modes are used for clustering and detecting the crime by means of detecting the similarity of crime with each other. 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 used. Second, k-modes are used to find the clusters of crimes with close similarity criteria based on distance. The proposed model was evaluated on the Community and Crime dataset consisting of 1994 samples with 128 characteristics. The results showed that purity accuracy of the proposed model is equal to 91.45% for 400 replicates.*

Keywords: *Crime Clustering, Clustering, Elephant Herding Optimization Algorithm, K-modes.*

How to cite this article:
Farhad Soleimanian Gharehchopogh, Sevda Haggi. An Optimization K-Modes Clustering Algorithm with Elephant Herding Optimization Algorithm for Crime Clustering. J. ADV COMP ENG TECHNOL, 6(2) Spring 2020 : 79-90

I.INTRODUCTION

Currently, crimes are growing as social problem. Undoubtedly, in condition of human society, it is impossible to come up with a phenomenon called crime, humans always have need to knowledge of crime analysis and their discovery. Crime analysis is the use of systematic approaches to identify, detect, and crime prediction. With the development of database systems and the high volume of data stored in these systems, a tool is needed to process this data and provide information to users. One of the most important tools for data discovery is data mining [1].

Increasing crime is the problem that security guards are facing with it. With the enormous volume of crime-related data and information in police departments and the complexity of communication between these crimes, statistical and manual models are not able to detect and predict. There are two major flaws to crime traditional statistics methods that often employed by experienced crime scene experts. First, they require a great deal of time and human expense. Second, because of the high rate of human involvement in decision making, they are unable to incorporate all of the influential factors in a crime and the relationships between them and slow down the steps of discovery process. Such situations make it increasingly clear the need to employ an intelligent approach to crime detection and analysis. Developing a model for the police system for crime analysis and discovery based on machine learning techniques can be the solution to these two challenges[2-4].

The enormous volume of crimes and criminal's data on the one hand and the existence of complex and intangible semantic relationships between this information on the other hand, criminology has made to one of the most important areas of data mining [5]. By extracting the properties of the crime, it will be possible to make a first step for any analysis of the properties of a crime. For this purpose, in this paper, we use a hybrid model based on K-Modes [6] and Elephant Herding Optimization (EHO) [7] algorithm to analyze crime clustering in order to detect similarity of crimes.

Clustering is one of the branches of unsupervised learning and is an automated process in which the samples are divided into groups whose members are similar to each other, these are called clusters [8]. Thus, the cluster is a set of samples that are similar in each cluster and are not similar to those in the other clusters. For similarity, different criteria can be considered, for example, the distance criterion can be used for clustering and samples that are closest to each other can be considered as one cluster.

The K-Modes [6] algorithm based on K-Means algorithm is generalized, which is applicable to big datasets[9]. The K-Modes algorithm is the most primitive algorithm in the batch data clustering process. This algorithm produces denser clusters than other algorithms. This method is faster than other clustering methods. In the K-Modes method, the final answer depends on the choice of the initial centers of the clusters and there is no specific procedure for calculating the initial centers of the clusters.

Swarm Intelligence algorithms are very useful for complex problems and other fields such as clustering and classification. The EHO Algorithm [7] is one of the swarm intelligence algorithms inspired by elephant life behavior. The elephant population is made up of several groups, each class has a leader who manages the elephants. In each generation, a fixed number of elephants leave the bunch and start a new life and form a group. In the EHO algorithm, each elephant is a solution that must have the ideas and behavior of the team leader.

Lack of awareness of clustering algorithms about the purpose of clustering and data structure causes these methods to work without any knowledge of the problem and with particular assumptions of trying to solve the problem. This will produce inappropriate answers if there is a discrepancy between the assumption and purpose of the clustering or the actual structure of the data. However, a wide variety of clustering algorithms in real-world problems have attempted to solve the problem with assumptions about the purpose of clustering and data structure. In this paper, we use an elephant batch optimization algorithm to improve K-Modes. In the hybrid model, we use the EHO algorithm to find the cluster center and optimize the K-Modes algorithm.

II. RELATED WORKS

Recently, several investigations have been made in the field of criminology detection to accelerate discovery and crime clustering. Data mining techniques have attracted the attention of many researchers because of their learning and data training [10-14]. These are methods that express logical patterns and relationships with the least user intervention and automatically. This section suggests different mining methods and models and that each of these algorithms is used in what application.

A hybrid model based on the K-Means and K-Medoid algorithms for crime clustering has been proposed [15]. In the K-Means algorithm, at first points are randomly selected the based on number of clusters required. Then, the data are attributed to one of these clusters with respect to proximity (similarity) and thus new clusters are obtained. By repeating the same procedure, it is calculated new centers for each iteration by averaging the data and reassign the data to new clusters. This process will continue until there is no change in the data. In the hybrid model, the objective is to optimize the K-Medoid algorithm using K-Means. The K-Means model identifies optimal points for clustering, and the data are assigned to the cluster head based on distance. The most important element in clustering is finding centrality. If the center of the cluster is correctly detected, all the cluster head data will be the same. The evaluation of the hybrid model is performed on 613 data points. The results show that the hybrid model has high accuracy in the number of clusters.

Self-Organizing Map (SOM) artificial neural network (ANN) model [16] has been proposed for the detection of crime and the similarity between different crime groups. SOM is a type of ANN that is trained through unsupervised learning to generate low dimensional state space from input space. Self-organized mappings are differing from other ANNs that SOM use a neighborhood function to maintain the input space property. This feature makes these mappings suitable for giving intuitive data sets of a high dimensionality. In the SOM, the competitive learning method is used for training and is developed based on specific characteristics of the human brain. The competitive learning used in these networks of networks is that in each learning step, the units compete to activate each other, at the end of a competition stage only one unit wins, whose weights are compared to the weights of the other units. It will be changed differently. SOM-ANN based on data training and testing tries to discover similar patterns. Evaluation was performed on 862 crimes with 16 features. The results show that SOM model was able to cluster all the features.

Fuzzy association rules based on Apriori algorithm for clustering crime have been proposed [17]. The hybrid model uses a fuzzy classification for crime ratings and Apriori algorithm for generating rules. Apriori is a classic algorithm for learning the rules of dependency. The input of this algorithm is a set of examples. The algorithm tries to find subsets of samples that are at least shared by C in the sample set. Apriori is a bottom-up algorithm, as one instance is added to the repeated subsets at each step. Evaluation is done on the Communities and Crimes dataset with 40 features. The results show that the accuracy of detection in the hybrid model is 60%.

Data mining techniques such as linear regression, decision tree are used for clustering crimes [18]. Evaluation was performed on the Communities and Crime Un-normalized dataset with 2215 crime samples in the WEKA environment. Weka software includes a set of machine learning algorithms and data preprocessing tools. The software is designed to quickly test existing methods flexibly on datasets. Linear regression is used to model the value of a quantitative dependent variable that linear relationship with one or more predictors is established. The results show that the accuracy of linear regression detection is higher than the decision tree model and the relative absolute error value is lower than the decision tree.

The Apriori algorithms and FP-Growth algorithms have been used to cluster detention crimes [19]. The FP-Growth algorithm is one of the association rule mining algorithms. This algorithm stores the data in a tree called FP-tree and then discovers the samples by constructing a recursive FP-tree. Evaluation was performed on 72 samples. The results showed that the FP-Growth model had higher accuracy, precision and F-Measure criteria than the Apriori.

The crime clustering was performed using the K-Means algorithm[20]. Evaluation was performed on the England and Wales dataset from 1990 to 2011 in the Rapid-Miner environment. In the K-means algorithm, k members (k is the number of clusters) are randomly selected from n members as the centers of the clusters. Then the remaining n-k members are assigned to the nearest cluster. After assigning all members to the cluster centers, they are recalculated and assigned to the clusters with the new centers, and this will continue until the cluster centers remain constant. The results on England and Wales dataset show that in 2002 the highest number of crimes was committed.

Crime clustering was performed using a combination of Genetic Algorithm (GA) and K-means [21]. Evaluation on England and Wales dataset was performed in the Rapid-Miner environment between 1990 and 2011. GA is used to optimize K-means. The purpose of the hybrid model of GA and K-Means is escape from local optimal points. Many researchers have used the GA to escape local optimal locations. In this model, a new method is proposed for crossover and mutation operators. The logic of the hybrid model is that if shift operators are performed in a predefined finite area, rather than randomly applied to the entire response space, they will achieve better solutions. The results showed that the criterion of accuracy in non-optimal condition is 85.74% and in optimum condition is 91.64%. As a result, the GA has been able to increase the accuracy of clustering.

A model based on multilayer ANN has been proposed for crime detection and clustering [22]. In the multilayer ANN, there is an input layer that receives the information, there are some middle layers that take the information from the previous layers, and finally there is an output layer where the result of the computation goes to this layer and the answers are placed. The multilayer ANN model uses data training and testing for crime detection. The DBH dataset for residential burglary was used to evaluate the multilayer ANN model.

An example of cybercrime has been done on 1638 samples using the K-Means algorithm [23]. Evaluation was performed in SPSS and Rapid-Miner software. Samples are divided into three categories based on clustering: the first cluster equals 1479 samples, the second cluster equals 133 and the third cluster equals 26, and the overall clustering accuracy is 90.33%.

In 2012, KAUR et al. [24] proposed C4.5 model for cybercrime on 265 samples. They have used Weka software for evaluation and results. Algorithm C4.5 extends the classification range in addition to deductive attributes in a variety of numerical attributes. The C4.5 basically selects the attribute that has the highest degree of separation between the categories and makes the decision tree based on it. The results show that the detection accuracy of the C4.5 model is 94.67% and the kappa coefficient is 0.8976. The kappa coefficient is a numerical measure of between -1 and +1, which is closer to +1 indicating that there is fit more. Clustering is also done in three categories: soft, hard and none.

Researchers based on Support Vector Machine (SVM) and Adaboost [25]have evaluated and tested on 2700 samples to detect cybercrime, such as Facebook. 2430 samples were used for testing and 270 samples for training. Using this method for high-dimensional datasets and low data volumes is very useful. The dataset contains 6 features that have been tested in Weka. The results show that the accuracy of the Adaboost model is 98.70% and is higher than the SVM.

In 2016, researchers proposed a model based on a fuzzy inference system for detecting cybercrime [26]. Fuzzy inference system is a system that formulates a mapping from input to output using fuzzy logic. The fuzzy inference system is also called the rule-based system. Because this system is made up of a number of "if-then" terms. Words like Terrorists, Crime and Killers have been dealt with in the documents. Each word is assigned a weight based on its number and then a triangular membership function is used to fuzzification. The results show that the detection accuracy of the fuzzy inference system is higher than the SVM.

III. PROPOSED MODEL

The proposed model is derived from the combination of the EHO algorithm with the k-modes algorithm. The innovation of the proposed model relates to the phase formation of the center of the cluster to form optimal clusters on the samples. In the proposed model, EHO algorithm is used to find the center of the samples. The EHO algorithm helps the k-modes algorithm to find the center of the clusters. Each elephant, as a representative on the basis of different iterations, finds the best center of the cluster. From the n data , it is selected the number of k data as the center of the cluster using the EHO algorithm. While the center of the cluster is found, then the clustering of the samples is performed by the k-modes algorithm. By the EHO algorithm, the distance between the samples of the individual clusters is maximized and the distance between the samples is minimized using k-modes. The points discovered by the EHO algorithm as the centers of the clusters and in fact the average points belonging to each cluster. Each sample is assigned to a cluster that has the least distance to the center of that cluster. In Figure (1), the general flowchart of the proposed model is shown.

In the proposed model, at first the dataset is read and then the normalization is performed on data. The number of cluster centers is determined to discover the cluster centers at the beginning of the program operation. Based on the number of centralities, the search for the environment is performed by groups of elephants. Each elephant searches for the best point based on the position of the best leader in each group. The assignment of samples to each cluster is based on Euclidean distance. Each sample is assigned to the nearest center of the cluster based on the distance.

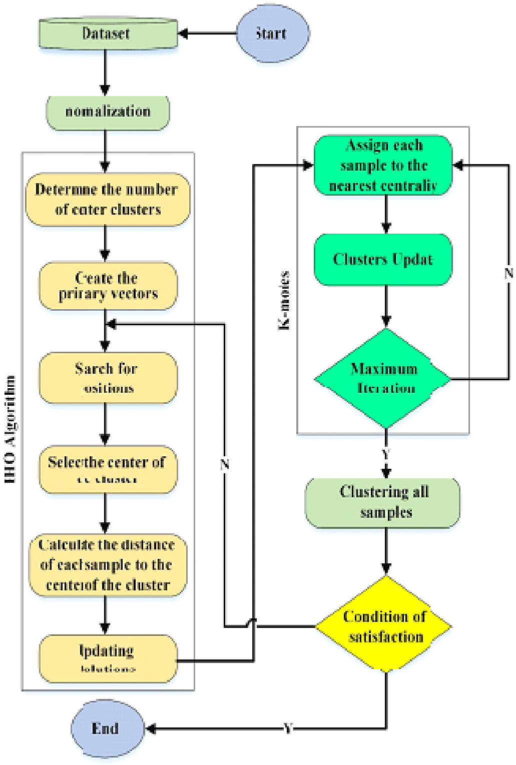
Dataset
Start
nomalization
IHO Algorithm
Determne the number of cnter clusters
Ceate the prirary vectors
Sarch for ositions
Selecthe center of
Calculte the distance of cacl sample to the centerof the cluster
Ipdating solutions
N
K-moies
Assign each sample to the nearest centraliy
Clusters Updat
Maximum Iteration
N
Y
Clustering all samples
Condition of satisfaction
Y
End

Figure 1: Flowchart of the proposed model

3.1. Normalization of data

Normalization is used to calibration data. Using normalization, the data are adjusted in similar and close range. This action is done because the algorithm was not faced with values that are in a different range or range. Large-scale features are likely to increase the cost of running the fitness function over low-value features. This problem will be solved by normalizing the features so that their values are in the same domain. Eq.(1) is used to normalize the samples[27, 28].

$$x_i(t) = \frac{\tilde{x}_i(t) - \mu_{s,i}}{\sigma_{s,i}} \qquad (1)$$

In Eq.(1), $\tilde{x}_i(t)$ is the actual value of the sample *ith* , the parameters $\mu_{s,i}$ and $\sigma_{s,i}$ are mean and standard deviation. The purpose of Eq.(1) is to standardize the value of the samples to a real value.

3.2. Clustering

In the proposed model, the best solution obtained in the EHO algorithm is used to estimate the center of the cluster. In the proposed model, the number k is found by the EHO algorithm and is not predetermined. This method leads to faster execution of the algorithm due to reduced computation and linear search in the target population and also the amount of output of the EHO algorithm is converging more rapidly. Convergence means approaching the optimal solution. The optimal solution is to find the best centrality for each cluster based on learning in the problem space.

In the ci cluster, the next position to choose centrality depends on the position of the group leader. That is, the best particle in the group must have the best position for centrality. For elephant j in the ci group, the method is updated according to Eq. (2). In Eq. (2), are the new and old positions for the j-element in the ci cluster, respectively. The parameter is the best leader position in the whole cluster. This parameter detects the best position when searching. Also, α parameter is in the range [0, 1] and is a scale factor for group leader effects to change positions. And the r parameter in the range [0, 1] is used for uniform distribution.

$$x_{new,ci,j} = x_{ci,j} + \alpha \times (x_{best,ci} - x_{ci,j}) \times r \quad (2)$$

The positon of leader of the group as the best elephant in the group is updated according to Eq. (3). The update is used to detect further positions in the ci cluster and then selected the best location as the center of the cluster based on the comparison between solutions. In Eq. (3), the parameter β is in the range [0, 1] and is an effective factor for controlling $xcenter, ci$.

$$x_{new,ci} = \beta \times x_{center,ci} \qquad (3)$$

Figure (2) shows the search to find the center of the clusters. Each elephant in each cluster changes its position based on the leader's position toward the center.

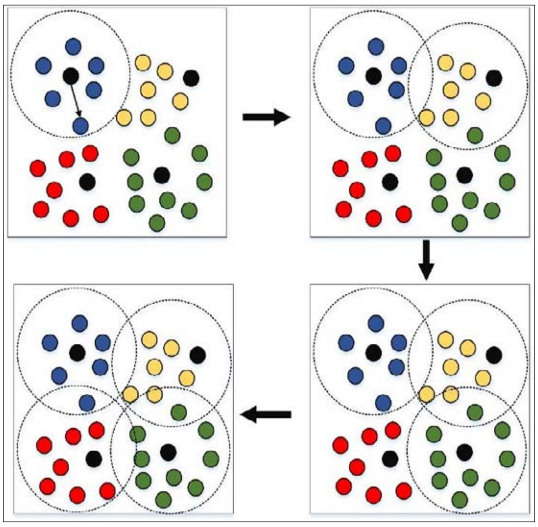

Figure 2: Finding the center of clusters based on learning

The distance between the clusters should be minimal. Eq. (4) is used to calculate the inter-cluster distance. In Eq. (4), Np is the number of samples in the dataset, for example $S = \{N1, N2, \ldots, Np\}$, zp is the sample data in the cluster Ck, mk is the center of the cluster kth, and K is the number of clusters.

$$\text{intra} = \frac{1}{N_p}\sum_{k=1}^{K}\sum_{z\in C_k}||z_p - m_k||^2 \qquad (4)$$

Figure (3) shows the inter-cluster distance. In cluster analysis, usually the p feature is measured on n samples and a matrix $n \times p$ of primary samples is formed. The primary samples matrix is converted to the similarity or distance matrix and the samples are grouped by similarity using the clustering method

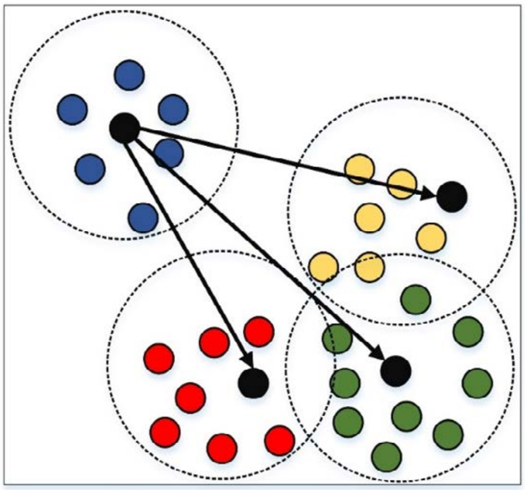

Figure 3: Calculate the inter-cluster distance

Cluster centrality and distance between samples to cluster center are two important factors in clustering. If these two factors are well recognized, all samples are added to their own clusters. In the proposed model, k-modes algorithm is used to discover the centrality of the EHO algorithm and to assign the samples to the center of the cluster. The steps of the proposed model are as follows:

1) Determine the number of center clusters

2) Create basic solutions based on basic vectors

$$x_{ij} = \begin{cases} 0, & if \quad U(0,1) \leq 0.5 \\ 1, & if \quad U(0,1) \geq 0.5 \end{cases} \qquad (5)$$

3) Repeat cycle

4) For each xi solution is performed the following operations

Select the center of the cluster if the value is x = 1.

Calculate the distance of each sample to the center of the cluster

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \qquad (6)$$

5) Updating solutions

6) Repeat steps (3) to (5) until iteration maximum

7) Using k-modes to assign each sample to the best center

8) Assign each sample to the nearest cluster

9) Production of new clusters

10) Repeat steps (2) to (9) to discover the maximum number of centralities

The k-modes algorithm has disadvantages such as the inability to manage outliers and noise data in the clustering process, and the final answer depends on the choice of primary cluster centers, and there is no specific procedure for calculating the cluster centers initially. To solve this problem, EHO algorithm is used, and then obtained results is giving to the k-modes algorithm to clustering again. This leads to that noisy data in the k-modes algorithm is managed. The choice of the initial centers of the clusters is also made by the obtained results from the EHO algorithm.

3.3. Evaluation Criteria

On order to evaluate the proposed model should validation criteria for the results is used. Eq. (7) measures the purity of the clusters [29]. The purity criterion is used to measure clusters and their distribution. The F-Measure criterion is derived from a combination of precision and recall criteria, and is used where the two precision and recall criteria do not differ in distinguishing accuracy. The precision criterion shows the accuracy of the cluster *i* with respect to all samples where the label *i* for the sample under by clustering model has proposed. In other words, the precision criterion is basically based on the accuracy of the prediction model. The recall criterion the accuracy of the clustering of the cluster *i* showing based on total number of samples labeled *i*.

$$Purity = \frac{1}{c}\sum_{i=1}^{c}\frac{C_i^d}{C_i} \qquad (7)$$

$$P(i,j) = \frac{n_{ij}}{n_i} \qquad (8)$$

$$R(i,j) = \frac{n_{ij}}{n_j} \qquad (9)$$

$$F-Measure = \frac{2*P*R}{P+R} \qquad (10)$$

$$RI = \frac{a+d}{a+b+c+d} \qquad (11)$$

In Eq. (7), c is the number of clusters, Ci is the number of samples in cluster *ith*, C_i^d the number of data that is correctly classified in cluster *ith*. In the defined equations, the parameters n_i , n_j and n_{ij} are the number of class i data, the number of cluster data j (obtained by the clustering algorithm) and the number of class i data in the j cluster, respectively [27]. In Eq. (11), the parameters a, b, c, d are equal to the number of samples True Positive, False Positive, False Negative and True Negative, respectively [27]. The Rand Index (RI) has a value between 0 and 1 and a value close to 1 is the optimal criterion.

IV. EVALUATION AND RESULTS

The Communities and Crime dataset was used to evaluate the proposed model [30]. Parameters such as the number of elephants, α, and β are also used to run the program, which are 50, 0.5 and 0.6 respectively. The value of the parameters is determined based on iterations and repeat testing of the program. The number of k in the k-modes algorithm is determined using the EHO algorithm, and the number of k is equal to the center of the clusters.

4.1 Number of Iterations

In this section, the proposed model is evaluated based on the number of iterations of EHO algorithm. The results of Table (1) show that the precision of the clustering in the proposed model has increased with increasing number of iterations. The purity values in the proposed model for iterations 100 and 200 are 80.26 and 82.62, respectively.

TABLE1: EVALUATION OF THE PROPOSED MODEL BASED ON THE NUMBER OF ITERATIONS

Criteria	Iteration=100		Iteration=200	
	K-Modes	Proposed Model	K-Modes	Proposed Model
Purity	77.25	80.26	79.52	82.62
Precision	76.59	78.26	77.30	80.55
Recall	76.90	79.05	78.49	80.92
F-Measure	76.74	78.65	77.89	80.73
RI	0.7891	0.8032	0.7986	0.8126

In Figure (4), a comparison chart of the proposed model with K-Modes based on 100 iterations is shown. The proposed model has been able to improve the clustering centers in the K-Modes model and to place similar samples in close proximity to similar clusters.

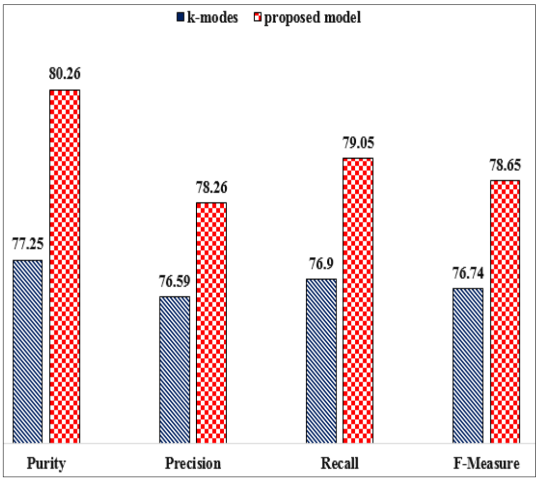

Figure 4: Comparison diagram of proposed model with K-Modes model based on 100 iterations

Figure (5) shows a comparison chart of the proposed model with a K-Modes model based on 200 iterations. In Figure (5), it is clear that the proposed model has been able to increase the accuracy of the purity to 200 iterations. There are number of comparisons more to detect data relative to the center of the cluster heads with 200 iterations, and similar data are more likely to be grouped into similar clusters.

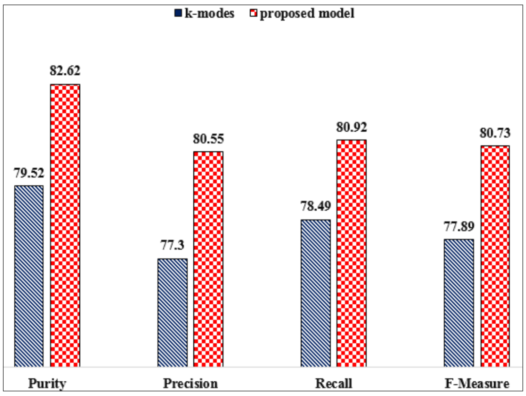

Figure 5: Comparison diagram of the proposed model with K-Modes based on 200 iterations

In Figure (6), a comparison chart of the proposed model based on 100 and 200 iterations is shown. In Figure (6), comparisons based on different criteria show that the proposed model has better precision with 200 iterations.

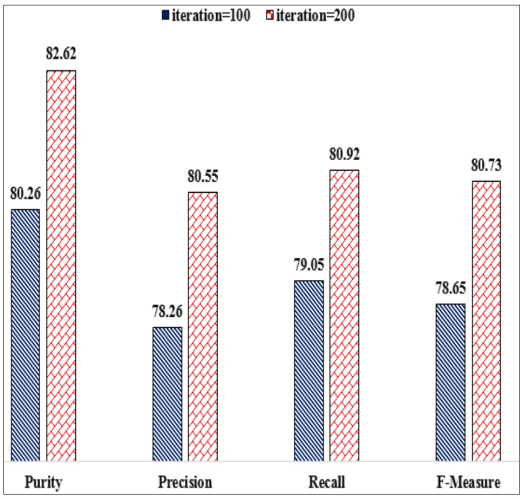

Figure 6: Comparison chart of the proposed model based on the number of iterations

4.2. Feature Evaluation

In this section, the proposed model is evaluated based on Feature Selection (FS) with 200 iterations. The FS is done in four steps. The results of Table (2) show that when the number of features is 40 the purity value in the proposed model is 92.46, if the number of features is 60 the purity value in the proposed model is 92.17 and 120 the purity value is 91.26% and 91.07% respectively.

TABLE 2: EVALUATION OF PROPOSED MODEL BASED ON NUMBER OF FEATURES

	FS=40		FS=60		FS=120	
Criteria	K-Modes	Proposed Model	K-Modes	Proposed Model	K-Modes	Proposed Model
Purity	90.61	92.46	90.23	92.17	88.21	91.07
Precision	88.25	90.63	87.76	89.82	86.43	88.26
Recall	89.26	91.72	88.06	90.64	87.26	88.90
F-Measure	88.75	91.17	87.91	90.23	86.84	88.58
RI	0.9036	0.9384	0.8921	0.9248	0.8762	0.9012

In Figure (7), a comparison diagram of the proposed model based on the number of features is shown. The comparisons in Figure (7) show that the proposed model is more accurate if the number of features is less. Because it can find more optimal points for the center of the clusters, the clusters contain more similar data.

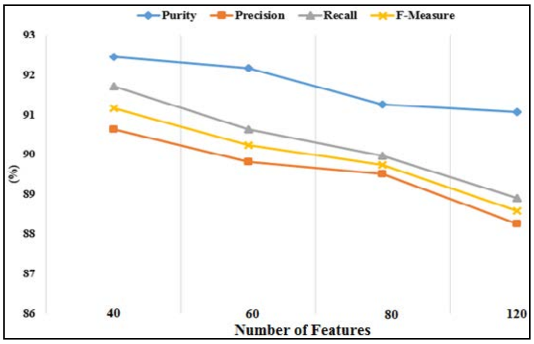

Figure 7: Comparison chart of the proposed model based on the number of iterations

Table (3) shows the results of the proposed model based on the iteration and different values of parameter α and β. α parameter is in the range [0, 1] and is a scale factor for group leader effects to change positions. The β parameter is in the range [0, 1] and is an effective factor for controlling centrality. It is clear in Table (3) that the proposed model with 400 iterations has a better purity value than the other cases. The maximum values of the boundary index in the 200 and 400 iterations are 89.15 and 91.49, respectively. The 200 and 400 iterations were chosen because the proposed model is more accurate in these iterations.

TABLE 3: RESULTS OF THE PROPOSED MODEL BASED ON THE ITERATION AND PARAMETERS RATE OF EHO ALGORITHM

Parameters			Criteria				
Iterations	α	β	Purity	Precision	Recall	F-Measure	RI
200	0.1	0.1	83.16	82.03	82.61	82.32	0.8278
	0.2	0.2	84.89	83.14	83.92	83.53	0.8314
	0.3	0.3	85.19	84.08	84.79	84.43	0.8562
	0.4	0.4	86.20	85.13	85.72	85.42	0.8647
	0.5	0.5	89.17	87.26	87.36	87.31	0.8711
	0.6	0.6	90.49	89.47	90.25	89.86	0.8915
400	0.1	0.1	84.03	83.29	83.54	83.41	0.8405
	0.2	0.2	85.65	84.15	84.61	84.38	0.8546
	0.3	0.3	86.02	86.93	87.05	86.99	0.8619
	0.4	0.4	89.90	87.02	87.81	87.41	0.8932
	0.5	0.5	90.23	89.41	90.23	89.82	0.9125
	0.6	0.6	91.45	89.76	90.11	89.83	0.9106

The results in Table (3) are shown based on 20 clusters. That is, the number of centers found for clustering is 20. The maximum purity in the proposed model for the 200 and 400 iterations is 90.49 and 91.45, respectively. If the value of α and β is 0.6, the EHO algorithm will be able to detect the center of the clusters.

V. CONCLUSION AND FUTURE WORKS

Security and military centers are face a large amount of criminal data and information, but without a systematic understanding of these phenomena, they cannot be implemented in the field of security. Automated methods and efficient and effective tools needed to extract knowledge in the field of security can greatly help to information and security organizations to monitor as well as adopt useful strategies in the field of security. In this paper, a combination of EHO and k-modes algorithm is used to cluster crime. The EHO algorithm finds the best centrality among the samples and selects the best point as the centrality by updating in each iteration, and then k-modes the sample clustering operation using the nearest distance. The results showed that the purity and RI values in the proposed model were 91.45 and 91.06, respectively. K-modes based on weighting can be used for future work. In this method, each feature is assigned a weight. This is because some features in the dataset may be more important because of the weight gain. The weight a feature must be in the range of 0 to 1.

REFRENCES

1. Tayal, D.K., et al., Crime detection and criminal identification in India using data mining techniques. AI & society, 2015. 30(1): p. 117-127; Available from: https://link.springer.com/article/10.1007/s00146-014-0539-6.

2. Gharehchopogh, F.S., H. Shayanfar, and H. Gholizadeh, A comprehensive survey on symbiotic organisms search algorithms. Artificial Intelligence Review, 2019: p. 1-48; Available from: https://link.springer.com/article/10.1007%2Fs10462-019-09733-4.

3. Shayanfar, H. and F.S. Gharehchopogh, Farmland fertility: A new metaheuristic algorithm for solving continuous optimization problems. Applied Soft Computing, 2018. 71: p. 728-746; Available from: https://www.sciencedirect.com/science/article/abs/pii/S1568494618304216.

4. Gharehchopogh, F.S. and H. Gholizadeh, A comprehensive survey: Whale Optimization Algorithm and its applications. Swarm and Evolutionary Computation, 2019. 48: p. 1-24; Available from: https://www.sciencedirect.com/science/article/abs/pii/S2210650218309350.

5. Chen, P.S., Discovering Investigation Clues through Mining Criminal Databases, in Intelligence and Security Informatics. 2008, Springer. p. 173-198.

6. Ganti, V., J. Gehrke, and R. Ramakrishnan. CACTUS—clustering categorical data using summaries. in Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. 1999.

7. Wang, G.-G., S. Deb, and L.d.S. Coelho. Elephant herding optimization. in 2015 3rd International Symposium on Computational and Business Intelligence (ISCBI). 2015. IEEE.

8. Abedi, M. and F.S. Gharehchopogh, An improved opposition based learning firefly algorithm with dragonfly algorithm for solving continuous optimization problems. Intelligent Data Analysis, 2020. 24(2): p. 309-338; Available from: https://content.iospress.com/articles/intelligent-data-analysis/ida194485.

9. Allahverdipour, A. 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, 2017. 8(4): p. 73-86; Available from: http://jacr.iausari.ac.ir/article_651859.html.

10. Khalandi, S. and F. Soleimanian Gharehchopogh, A New Approach for Text Documents Classification with Invasive Weed Optimization and Naive Bayes Classifier. Journal of Advances in Computer Engineering and Technology, 2018. 4(3): p. 167-184; Available from: http://jacet.srbiau.ac.ir/article_12936.html.

11. Parlar, T., S. Ozel, and F. Song, Analysis of data pre-processing methods for sentiment analysis of reviews. Computer Science, 2019. 20; Available from: http://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-c8123943-cf0e-46d6-acd2-8b42784e4235.

12. Allahverdipour, A. 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, 2018. 9(2): p. 37-48; Available from: http://jacr.iausari.ac.ir/article_655529.html.

13. Aci, M., C. İnan, and M. Avci, A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm. Expert Systems with Applications, 2010. 37(7): p. 5061-5067; Available from: https://www.sciencedirect.com/science/article/abs/pii/S0957417409010501.

14. Majidpour, H. and F. Soleimanian Gharehchopogh, An improved flower pollination algorithm with AdaBoost algorithm for feature selection in text documents classification. Journal of Advances in Computer Research, 2018. 9(1): p. 29-40; Available from: http://jacr.iausari.ac.ir/article_653945.html.

15. Grubesic, T.H., On the application of fuzzy clustering for crime hot spot detection. Journal of Quantitative Criminology, 2006. 22(1): p. 77; Available from: https://link.springer.com/article/10.1007/s10940-005-9003-6.

16. Wang, W.B., et al. Detecting criminal relationships through som visual analytics. in 2015 19th International Conference on Information Visualisation. 2015. IEEE.

17. Buczak, A.L. and C.M. Gifford. Fuzzy association rule mining for community crime pattern discovery. in ACM SIGKDD Workshop on Intelligence and Security Informatics. 2010.

18. McClendon, L. and N. Meghanathan, Using machine learning algorithms to analyze crime data. Machine Learning and Applications: An International Journal (MLAIJ), 2015. 2(1): p. 1-12.

19. Lawpanom, R. and W. Songpan, Association Rule Discovery for Rosewood Crime Arrest Planning, in Information Science and Applications (ICISA) 2016. 2016, Springer. p. 1025-1032.

20. Agarwal, J., R. Nagpal, and R. Sehgal, Crime analysis using k-means clustering. International Journal of Computer Applications, 2013. 83(4); Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.1621&rep=rep1&type=pdf.

21. Kiani, R., S. Mahdavi, and A. Keshavarzi, Analysis and prediction of crimes by clustering and classification. International Journal of Advanced Research in Artificial Intelligence, 2015. 4(8): p. 11-17.

22. Keyvanpour, M.R., M. Javideh, and M.R. Ebrahimi, Detecting and investigating crime by means of data mining: a general crime matching framework. Procedia Computer Science, 2011. 3: p. 872-880; Available from: https://www.sciencedirect.com/science/article/pii/S1877050910005181.

23. Zulfadhilah, M., Y. Prayudi, and I. Riadi, Cyber profiling using log analysis and k-means clustering. International Journal of Advanced Computer Science and Applications, 2016. 7(7): p. 430-435; Available from:https://www.researchgate.net/profile/Yudi_Prayudi/publication/305737193_Cyber_Profiling_using_Log_Analysis_and_K-Means_Clustering_A_

Case Study Higher Education in Indonesia/links/579eeaa608ae6a2882f5479a.pdf.

24. Kaur, M., S. Vashisht, and K. Saurabh, Adaptive algorithm for cyber crime detection. International Journal of Computer Science and Information Technologies (IJCSIT), 2012. 3(3): p. 4381-4384; Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.1130&rep=rep1&type=pdf.

25. Deylami, H.-M. and Y.P. Singh, Adaboost and SVM based cybercrime detection and prevention model. Artif. Intell. Research, 2012. 1(2): p. 117-130.

26. Vadivel, A. and S. Shaila, Event Pattern Analysis and Prediction at Sentence Level using Neuro-Fuzzy Model for Crime Event Detection. Pattern Analysis and Applications, 2016. 19(3): p. 679-698; Available from: https://link.springer.com/article/10.1007/s10044-014-0421-7.

27. Hasanluo, M. and F. Soleimanian Gharehchopogh, Software cost estimation by a new hybrid model of particle swarm optimization and k-nearest neighbor algorithms. Journal of Electrical and Computer Engineering Innovations (JECEI), 2016. 4(1): p. 49-55; Available from: http://jecei.sru.ac.ir/article_556.html.

28. Asghari Agcheh Dizaj, S. and F. Soleimanian Gharehchopogh, A New Approach to Software Cost Estimation by Improving Genetic Algorithm with Bat Algorithm. Journal of Computer & Robotics, 2018. 11(2): p. 17-30; Available from: http://www.qjcr.ir/article_543464_115388.html.

29. Huang, X., et al., DSKmeans: a new kmeans-type approach to discriminative subspace clustering. Knowledge-Based Systems, 2014. 70: p. 293-300; Available from: https://www.sciencedirect.com/science/article/abs/pii/S0950705114002664.

30. Reaves, B.A. and A.L. Goldberg, Law enforcement management and administrative statistics, 1997: Data for individual state and local agencies with 100 or more officers. 1999: DIANE Publishing.