Proceedings of the 2016 International Conference on Machine Learning and Cybernetics, Jeju, South Korea, 10-13 July, 2016

CRIMINAL PATTERN IDENTIFICATION BASED ON MODIFIED K-MEANS CLUSTERING

TURKI ALJREES, DAMING SHI, DAVID WINDRIDGE, WILLIAM WONG
School of Science and Technology, Middlesex University
E-MAIL: T.aljrees@mdx.ac.uk, D.shi@mdx.ac.uk, D.windridge@mdx.ac.uk, W.wong@mdx.ac.uk

Abstract:

Data mining methods like clustering enable police to get a clearer picture of criminal identification and prediction. Clustering algorithms will help to extracts hidden patterns to identify groups and their similarities. In this paper, a modified k-mean algorithm is proposed. The data point has been allocated to its suitable class or cluster more remarkably. The Modified k-mean algorithm reduces the complex nature of the numerical computation, thereby retaining the effectiveness of applying the k-mean algorithm. Firstly, the data are extracted from the communications and movements record after tracking the park visitors over three days. Then, the original data will be visualised in a graphical format to help make a decision about how many numbers to consider as the K cluster. Secondly, the modified k-means algorithm on the clusters initial centre sensitivity will be performed. This will link similar segments and determine the occurrence of each data point in every segment group rather than partitioning the entire space into various segments and calculating the occurrence of the data point in every segment. Thirdly, result checking and a comparison with the normal k-mean will be performed. The investigation will focus on the movement of people around the park where the crime occurred, and how people move and communicate in the park, how patterns change, and the movement of groups and individuals. The experiments show that the modified K-means algorithm leads to a better way of observing the data to identify groups and their similarities and dissimilarities in the criminal dataset as a specific domain.

Keywords:

Data Mining; K-means Algorithm; Hierarchical Clustering

1. Introduction

The problem of finding a pattern between documents has recently attracted the attention of many scientists and researchers in the data mining field. The presence of too much information can make any task such as data mining quite difficult. The calculation of similarities can be time-consuming, especially for large sets of data that rapidly increase new information when patterns of large groups are identified, including finding how the data has changed and evolved over time. Processing the data manually for analysis has been automated in view of different approaches, for example, association, clustering, classification and decision tree. These approaches can be connected in substantial datasets for data mining. Machine learning has the ability to adapt to a new method for developing algorithms that are more generic in nature; also, it can change behaviours to detect patterns and can be applied to various domain-related problems. The clustering algorithm can cull existing information to highlight patterns and serves as the foundation for artificial intelligence and machine learning. An alternative algorithm method has been proposed that takes into account the nature of the data and the input parameters, keeping in mind the end goal of clustering the dataset. The efficiency of each one of these algorithms relies on the depends on the complexity of the data construction [7].

In the criminal data analysis, it can be much harder to grasp the motivations for the patterns when the pattern has changed to create a better plan for future occasions, as it includes several factors. One of the key procedures in managing these data is to classify or cluster them into an arrangement of types or clusters. The focus point is not just the collection of data but careful analysis of the collected data so that meaningful results can be obtained. There are various ways of handling the massive incoming streams of data. One such way is the clustering of data into compact units. Clustering not only reduces the size of the data but also helps it to be used in a more efficient manner [5].

Cluster analysis is most widely recognised as an unsupervised learning strategy that uses examining data analysis to discover hidden patterns or forms of grouping in the data. The clusters are displayed using a measure of similarity characterised by metrics such as probabilistic or Euclidean distance. Determining the number of clusters in the data is not a very simple job. .

978-1-5090-0390-7/16/$31.00 ©2016 IEEE

To segment the dataset into a particular number of clusters, we need to use the unsupervised learning method known as cluster analysis. Researchers have not yet settled on one definition of this method, and the majority of researches have only defined the cluster as patterns in the same cluster that should be similar to each other, whereas patterns in various clusters should not be. The intention of clustering is to group the information in the viewpoint of their similarities or dissimilarities. There are many famous clustering algorithms, for example, k-means cluster, two-step cluster, and hierarchical cluster. Applying any clustering technique, the data must take into account the data structure and its input factors. The different between the k-means and hierarchical algorithms are that k-means clustering segments the dataset into k distinct clusters based on the distance to the centroid of a cluster, whereas hierarchical clustering uses the method of creating a cluster tree to construct a multilevel hierarchy of clusters. Consequently, either the similarity or the dissimilarity should be meaningful .

Lastly, it's recommended to explain the cluster's' method and the result mathematically in a meaningful way, for instant, given a set of input patterns

$$X = \{x_1, ..., x_j, ..., x_N\} \tag{1}$$

Where

$$x_j = (x_{j1}, x_{j2}, ..., x_{jd})^T \in \mathbb{R}^d \tag{2}$$

Each measure x_{ji} is thought to be a feature, for example, the dimension, attribute, or variable based on the descriptions was in [2].

In this paper, we present how to optimise the original k-mean algorithm by applying the modified k-mean algorithm to the criminal dataset. The reason for using the modified k-mean algorithm is to acquire the benefit of reducing the complexity of the numerical calculation when applying the k- mean algorithm. Instead of calculating the occurrence of each point in segments and dividing the entire space into several segments, our method involves using the modified K-means algorithm based on the initial centre sensitivity of the clusters. The algorithm then merges similar segments and calculates the occurrence of each point in each segment cluster.

The rest of the paper is structured as follows: Section II, we present work that relates to both the k-mean and the modified k-means algorithms. We also described the drawbacks of the k- mean algorithm, which the modified algorithm resolves. In section III, we will present our experimental results and the application we used. In the last section, we will give our conclusion

2. Related Work

Clustering analysis is a very challenging assignment because only observatory research has completed by experts who adjust the initial point of the centroid location of the cluster in datasets. Researchers have not yet found a standard method for several problemsfor example, how the dataset should be normalised, a suitable similarity measurement for the dataset, and the correct way to cluster a high-dimensional dataset efficiently. The following list some of the difficulties of applying clustering algorithms. In the first place, the majority of the clustering methods that have used require a number of iterations. Secondly, for the feature selection or extraction, there are no general guides. Besides the quality of the results, there is not yet a general validation measurement and no standard solution is available. To overcome these drawbacks different clustering algorithms have been proposed. In the next section, we will briey analyse various clustering algorithms.

The clustering procedure sometimes has the same name as the cluster analysis, which also contains a few stages that have a feedback pathway. Firstly, these steps involve the feature determination or extraction and selection, the algorithm design, the algorithm validation and lastly, the results interpretation. These stages are associated with each other and inuence the clusters result, whereas feature extraction uses some transformations to create valuable features based on the original ones, in which the feature selection process selects the distinguishing features from a set of candidates. However, the clustering algorithm design step or procedure is usually joined with the selection of a proximity measure. Patterns clustered according to whether they are similar to each other. The cluster validation procedure comes before the last stage in the results interpretation. Regardless of whether the structure exists or not, every clustering algorithm can simply create a division for a given dataset. The vital objective of clustering is the results interpretation stage, in which meaningful insights will be provided to users from the original dataset.

Following summarised lists are the existing clustering algorithms:

- Distance and Similarity Measures , Primarily concentrate on calculating the semantic comparability among concepts in ontology domain The results of semantic matching degree can be considered as next: exact match, plug in match, subsumes match and fail match. This calculation has dependably been laid a foundation for the examination among different algorithms. [10].

- Graph Theory-Based which is mathematical structures used to model pairwise relations between objects. A net-

work diagram represent consisting of a set of edges and nodes demonstrating the communication among the nodes [4].

- Fuzzy algorithm which real values of variables might be any number between 0 1. also, gives a natural technique to develop systems that copy human making decision processes [6].

- Kernel-Based methods are a class of algorithms for pattern analysis, whose best known member Support Vector clustering (SVC) and Kernel K-means. Kernel methods extremely compelling and have been demonstrated that they can remove non-linear elements giving better recognition results. [3] .

  - Support Vector clustering is Kernal-Based clustering technique. [9]

- Large-Scale Data Sets

  - CLARA, CURE, CLARANS, BIRCH

- Squared Error Based

  - Iterative self-organizing data analysis, K-means, Genetic K-means

- High-dimensional Data and Data visualization

  - Principal Component Analysis, is optimal and provides the optimal center for your clusters. And a widely used statistical technique for unsupervised dimension reduction However if you only want two clusters as and force certain object into one cluster then K-Means is a great way to go. [1]

- Hierarchical Clustering algorithm

  - Agglomerative: Complete, Single, group average, median, and centroid linkage,

  - Divisive: divisive, and monothetic analysis

- Neural Networks-Based:

  - Self-organizing feature map, Learning vector quantization, simplified ART, clustering network, self splitting competitive learning network

3. K-Means Clustering

K-means is a widely known clustering partitioning approach. For applications in which a number of clusters are required, K-mean is very useful; similarly, the k-means clustering technique groups the data together taking into account their closeness. Mainly, k-means is an iterative algorithm and performs two mechanisms: first, the cluster assignment step, and second, the group centroid step, which is sometimes called the move centroid step. The k-mean algorithm must have at least the following steps or procedures:

- The first is a cluster assignment step, which calls cluster centroids by randomly initializing the cluster's centroid vectors into two or more points. This step must always start with k centres

- The second step is a move centroid step, which is sometimes named the update step, in which the cluster centroid points that initialised in the assignment step are moved to the average of the points in the same group. In this step, we must cluster each point with the centre nearest to it.

- To compute the distance from the data vector to the cluster, the next equation was used:

  Where d is the dimension.

$$\mathrm{d}(\mathrm{Z_p}\,\mathrm{M_j}) = \sqrt{\sum_{k=1}^{2}(Z\ p, k - Mj, k)} \tag{3}$$

  Where z_p is the pth data point, and M_J is centroid of i^th cluster. Each data vector calculates the distance between the data vector and each cluster centroid, which will use the minimum data vector to assign that cluster and calculate the distance using the above equation

- The centroid of the cluster is then recalculated using the following equation

$$\mathrm{M_j} = 1/\mathrm{n_j}(\sum\ \mathrm{Z_p})\,\nabla \mathrm{Z_p} \in \mathrm{C_j} \tag{4}$$

  Where n_j is the number of data point in cluster j [8].

- Repeat the previous step until the centres converge, which means that we re-run these steps until we have found no more changes.

We always need to remember that there are two undefined aspects of the algorithm. One is the set of starting centres and the other is the stopping condition. Two inputs must be taken

into consideration in order to apply the k-means algorithm. The first input is parameter K, which indicates the number of clusters we are looking for, and the second input is the k-means, which takes as the input the unlabeled training set of just the Xs, where

$$x^i \in \mathbb{R}^n \tag{5}$$

3.1. K-mean drawback and limitation

Even with the simplicity of K-means implementation, the method has many drawbacks and limitations. It is also not very flexible. The most important disadvantage is the initial centre point and the fact that there is no correct way of doing it. This is one of the reasons why many researchers have presented improvements to k-means algorithms. The other limitation of k-means is that the parameter K does not provide for external constraints. According to Singh and Bhatia in their paper [8], the Initial choice of the cluster number needs to be specified and known by the user. In the next section, we will mention in detail the k-means limitations that the modified k-mean solves.

4. Modified K-means

The simple implementation of the k-means algorithm means it has been broadly used in many fields of machine learning. It does not require more computation. However, there are some obstacle and limitation when applying K-means in the large dataset, as described in [11].

Several researchers have completed work to modify the k-means algorithm, especially in the area of the initialization point of the clusters centroid location. The modified k-means algorithm approach is designed to reduce the complexity of the numerical calculation of applying the k-mean algorithm. The modified k-mean features an alternative way of calculating and computing the frequency of each data point in segments and dividing the entire space into more than a few segments. The partition of the space is used to calculate the frequency of the input vector in every part and to pick the highest k-frequency section.

4.1. Discussion

In paper [8], a novel research on the initial point of the centroid named as the modified k-mean algorithm has completed. The modified k-mean approach is used to optimise the current algorithm. The K-means limitation resolved by running the algorithm for different numbers of k values.

The algorithm was constructed based on the density method of the different regions when the centroid of the cluster located in the first iteration with the maximum density of the data points. The algorithm was applied here to pick the initial centroid, and its pick by pick was arbitrary or by any fixed way. Then, to enhance the initial centroid selection through the use of the most populated space as a centroid of the cluster. Then, the centroid calculated using the mean of each segment and locating the clusters initial centroid. After dividing the space into many parts, we counted the frequency of the data points in each segment. The distance calculated between each cluster's centroid was used to assign the data point to the appropriate cluster's centroid and then, for each centroid, the minimum distance calculated from the remaining centroid, and this was made into a half.

The data point has been assigned to correct the cluster's centroid; the first step involved half of the minimum distance from i^{th} cluster's centroid to the remaining cluster's centroid. The second step involved considering any data point to compute its distance from i^{th} centroid and then comparing it with $d_{C(i)}$. Then, in the third step, if it was equivalent or a smaller amount than $d_{C(i)}$, then the data point was assigned to the i^{th} cluster. Otherwise, we computed the distance from another centroid. We then had to repeat the process until the data point was assigned to any of the remaining clusters. In addition to this, If the data point was not assigned to any of the clusters, the centroid that displayed the minimum distance for a data point became the cluster for that data point. Finally, we repeated the second step phase until the end of the condition was achieved

5. Experimental Results

The outcome will obtain from the record, which was a result of tracking the park visitors information over three days. This will allow us to achieve various hidden predictive patterns by analysing all the observations from large VAST databases

5.1. Background and The Application

One incident occurred at the park during the weekend. A crowd had gathered at the park to honour a local soccer celebrity. Local police appeared on the scene shortly after the park visitors discovered some vandalism. Security guards were questioned to eliminate the possibility of an inside job.

Visitors use park applications to check-in, to go on rides, and to communicate with fellow visitors. If visitors do not have compatible phones, they are provided with loaned devices. Visitors are assigned IDs and must use the app to check into rides and other attractions.

The park equipped with sensor beacons that record movements within the park. The sensors each cover a 5m x 5m grid cell. All the pathways are covered by these sensors including both the park and all the ride check-in locations. Locations are not recording while people are on rides or inside attractions, including restaurants, stores, and restrooms. App users may send text messages to anyone within their preferred group (for example, a family could have their private group). An app user may also make a friend at the park and can send and receive texts if both persons accept the friend invitations.

5.2. The DATA

VAST data has two datasets, Movement and Communication, which captured from the park attendees apps during one weekend (Friday, Saturday, and Sunday).

5.2.1 Movement Data

TABLE 1. Movement data

Timestamp	Id	Type	X	Y
06/06/14 08:00:11	1591741	check-in	63	99

The values are as follows: timestamp, person-id, type of activity (either check-in or movement), X is coordinate, and Y is coordinate. The park area is gridded to assist in specifying locations. The data file contains movement information around the grid, with coordinate locations. This data appears as in Table 1. The movement data size is $103MB$. People either move from grid square to grid square or check in at rides, which means they either get in line or enter a ride. Therefore, as can be seen from the above information, a person with ID 5231584 checked into a ride located at $(63, 99)$ at 8:00:08 AM. Person 93275931 also just moved to location $(64, 98)$ at $8:00:42$. As a person travels through the park, their locations are recorded in this file at a one-second resolution. If there is no record during a particular second of time, it means that the individual has not moved out of their previous grid square. People not tracked after they check into a ride, and they will eventually appear back on the grid when the ride is over.

5.2.2 Communications Data

This the second data file and contains communications information. The values are as follows: timestamp, from (the sender ID), to (the recipient ID), and location (the area where communications occurred). The location can be named as follows: Entry Corridor, Kiddie Land, Tundra Land, Wet Land, or Coaster Alley. The records are shown in Table 2. The communications data size is $180MB$. Graph 1 illustrates the difference in both data. The dimensions of the communications data include having four variables, with movement data being the fifth.

TABLE 2. Communication data example

Timestamp	from	to	location	
2014-06-06 09:21:35.000	790661	1920255	Tundra Land	

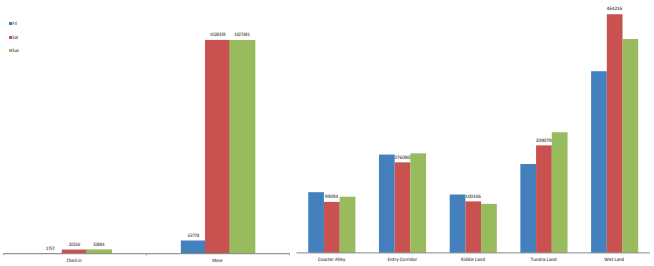

FIGURE 1. Comparison of (a) Movement (b) Communication

5.3. PROBLEMS and CHALLENGES

The scenario was in an amusement park. The simulated park covers a vast geographic space $(approx.500x500m^2)$. Patterns can be found in the movement through the park and in the communications among visitors, including expected regular visit patterns and unexpected patterns.

Our tasks in the investigation were to focus on the movement of people around the park, how people move and communicate in the park, how patterns change and how the data change and evolve over time. As we mentioned earlier, the dataset is large, and each data has a different size and values. Once the investigation is completed, the following questions should be answered: How big is the group type? Where does this type of group like to go in the park? How common is this kind of group? What are the other observations about this type of group? What can you infer about the group? If improvement in the park needed to be made to meet this group's needs, what would it be?

5.4. Patterns Identifying Tasks

The experimental tasks were time-consuming, especially with such a large set of data for identifying the patterns of large groups

5.4.1 Cluster and Visualization Using R

We have used an R environment for both the data miners and the graphics. The original data must always be visualised in a graphical format to help make decisions about how many numbers should be to consider as K clusters. Graph 2 shows all the data before clustering.

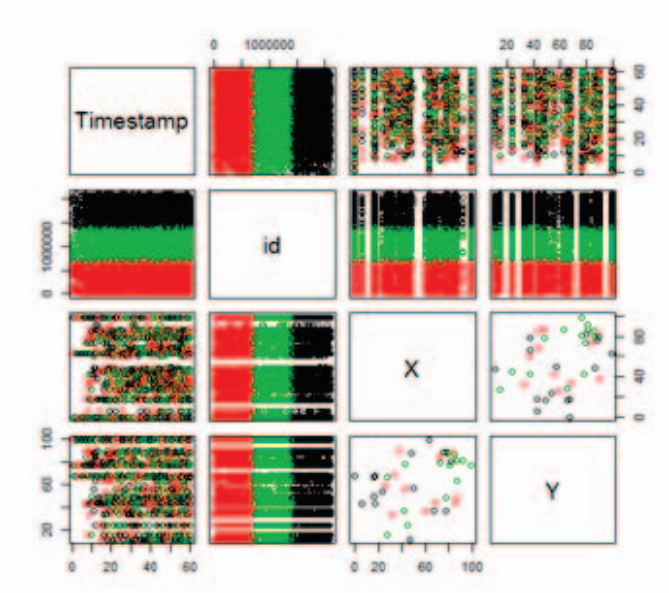

FIGURE 2. All the data before clustring

Visualize the Friday Movement data in their three dimension, plotted it in 3D as follows:

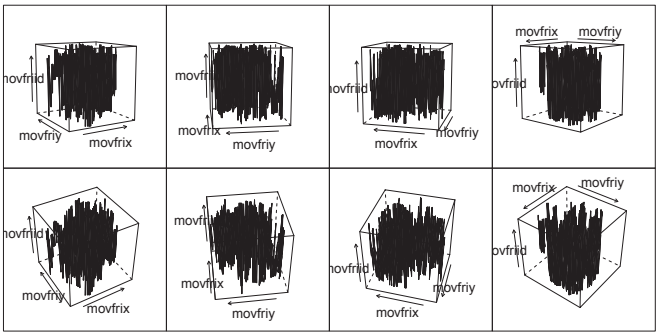

FIGURE 3. plot it in 3D

The Friday movement data is visualised in three dimensions and plotted in 3D, as displayed in figure 3. As we can see in Fig 4, the data point on graph 4 (a) is the location of all the datasets after being plotted. This is the same as the real footpath in 4 (b) for the park map data points.

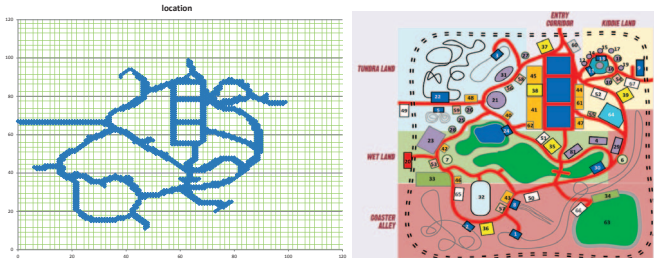

FIGURE 4. (a) Location of all data (b) The real Park map

5.4.2 Determining the Number of Clusters

For the clustering problem, we have been giving unlabelled data sets, and we would like to have an algorithm to group the data into subsets. In this clustering approach, the number of clusters needed to be decided in advance. The best way we found to determine the appropriate number of clusters was to plot the data in the way that we did. The average of the data vector of the cluster then assigned to the cluster that indicates the minimum distance.

Considering all the factors, in particular for the criminal data analysis, is a crucial aspect of making a decision that could lead us to a potential suspect. We could still use one factor at a time in the early stage of the analysing to assist later in determining how many clusters we need. However, we should not only rely on that value as the primary factor, as we still have the ID and location as important values. For example, in figure 5, the plot-

ting of the frequency of the Saturday and Sunday timestamps could easily help us to identify some patterns. The following graph shows how many moves happened at a particular time on Saturday

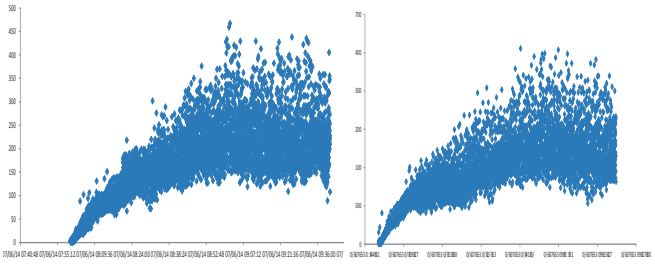

FIGURE 5. (a) Saturday (b) Sunday Timestamps Frequency

5.4.3 K-mean Implementation

In our application, the first step in the loop of the k-mean is the cluster assignment step, in which the cluster will go through each vector giving it a data set, and depending on whether it is closer or less close to the cluster centroid points, it will assign each data point to one of the cluster centroids.

The clustering algorithm will go through the data set, which records the time and id of each customer in the park. Then, when we plot them, we will be able to see each of the data points and how far or close they are to each centroid point they were assigned to. When we have decided that the number of the clustering $= 3$, then$nK = 3$. Therefore, the X and Y are with the $3Cj$ centroid locations. The group we found can be seen in graph 6.

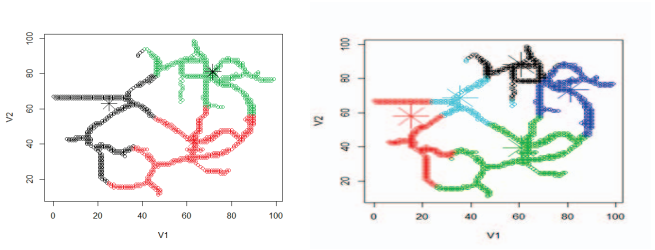

FIGURE 6. Clusters (a) 3 = K-means (b) 5 = K-means

5.4.4 Modified K-means Implementation

We ran the algorithm for different numbers of k values, as in figure 6, for $k = 3$ and $k = 5$.

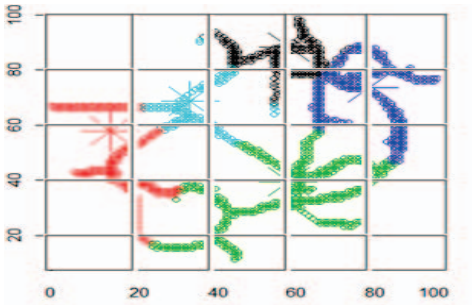

FIGURE 7. K-Means results Partitioned into different segments

We then used the modified k-mean algorithm based on the density. As we defined the best value of k as equalling 5, the space will be partitioned into $5multipliedby5$, which is the same as the $k * k$ segments, as shown in figure 7. We then have to count the frequency of the data points in each section, as shown in the Table III.

Then we have count the frequency of data point in each section, as shown in following table III . There are many segments

TABLE 3. Number of data points in different segment

Segment	Number
((0,0),(20,20))	0
((0,20),(36, 67))	7280
((21,40),(15,93))	11074
((41,60),(20,40))	11842
((0,60),(20,80))	0
((61,80),(23,99))	24520
((0,80),(20,100))	0
((81,100),(38,88))	11032
((80,0),(100,20))	0

that show the highest frequency for a given k. Therefore, there is no merge for the segments due to the importance of each segment. Consequently, the centroids of the cluster are measured in all the segments. The centroid will be calculated using the mean of each segment. The cluster centroid then initialized, and we then assign each data point to each cluster's centroid by calculating the distance between each cluster's centroid and each data point. We then calculate the distance from any of the points from the centroid to compare them with what has calculated with each cluster's centroid and each data point. We then recalculate all the data points of each cluster individually. The clusters' centroids will change for each iteration. This phase

must reiterate until the end of this condition is achieved.

6. Conclusion

K-Means is a very important method for data clustering, but it is not efficient enough to handle the criminal data analysis, for which all factors need to be considered. Nevertheless, an improved version of this method is needed. The modified k-means algorithm approach has improved the ability to overcome the limitations of the k-mean algorithm, which included running the algorithm for different k-value numbers and adjusting the initial point of the centroid location. By applying the modified k-means cluster, we can quickly identify how many k-mean clusters are needed to initialize the centroid point that will lead to better clustering. Moreover, by applying such an approach, we could clearly identify the right factor to be considered in the cluster. Future research will be considered to learn how to predict a new pattern using a new method to prevent possible crimes in the future.

References

[1] C. Ding and X. He. K-means clustering via principal component analysis. In Proceedings of the Twenty-first International Conference on Machine Learning, ICML 04, pages 29, New York, NY, USA, 2004. ACM.

[2] P. Hansen and B. Jaumard. Cluster analysis and mathematical programming. Math. Program., 1997.

[3] C.-K. Hsieh and Y.-C. Chen. Kernel-based pose invariant face recognition. In Multimedia and Expo, 2007 IEEE International Conference on, pages 987990, July 2007.

[4] N. Nader, C. Marque, M. Hassan, N. Nader, W. Falou, A. Diab, and M. Khalil. Pregnancy monitoring using graph theory based analysis. In Advances in Biomedical Engineering (ICABME), 2015 International Conference on, pages 7376, Sept 2015.

[5] D. Pandove and S. Goel. A comprehensive study on clustering approaches for big data mining. In Electronics and Communication Systems (ICECS), 2015 2nd International Conference on, pages 13331338, Feb 2015.

[6] D. Ramot, M. Friedman, G. Langholz, and A. Kandel. Complex fuzzy logic. Fuzzy Systems, IEEE Transactions on, 11(4):450 461, Aug 2003.

[7] K. Singh. Article: A new structural approach for mining frequent itemset. International Journal of Computer Applications, 71(9):3942, June 2013. Full text available.

[8] R. Singh and M. Bhatia. Data clustering with modified k-means algorithm. In Recent Trends in Information Technology (ICR- TIT), 2011 International Conference on, pages 717721, June 2011.

[9] L.-Y. Wu and J.-S. Wang. A simplified support vector clustering algorithm. In Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on, pages 12591264, Oct 2008.

[10] H. Yang, P. Fu, B. Yin, M. Ma, and Y. Tang. A semantic similarity measure between web services based on google distance. In Computer Software and Applications Conference (COMPSAC), 2011 IEEE 35th Annual, pages 1419, July 2011.

[11] B. Yi, H. Qiao, F. Yang, and C. Xu. An improved initialization center algorithm for k-means clustering. In Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on, pages 14, Dec 2010.