International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume 2, Issue 6, November - 2014

Performance Tuning of K-Mean Clustering Algorithm a Step towards Efficient DSS

Ayman E. Khedr, Ahmed I. El Seddawy, Amira M. Idrees

***Abstract*— This research is the first step in building an efficient Decision Support System (DSS) which employs Data Mining (DM) predictive, classification, clustering, and association rules techniques. This step considers finding groups of members in the dataset that are very different from each other, and whose members are very similar to each other, therefore one DM task is applied which is clustering task. The main objective of the proposed research is to enhance the performance of one of the most well-known popular clustering algorithms (K-mean) to produce near-optimal decisions for telcos churn prediction and retention problems. Due to its performance in clustering massive data sets. The final clustering result of the k-mean clustering algorithm greatly depends upon the correctness of the initial centroids, which are selected randomly. This research will be followed by a serious of researches targeting the main objective which is an efficient DSS which will be applied on customer banking data. In this research a new method is proposed for finding the better initial centroids to provide an efficient way of assigning the data points to suitable clusters with reduced time complexity. The proposed algorithm is successfully developed an applied on customer banking data, and the evaluation results are presented.**

***Index Terms* — Data Mining, Classification, K-Mean, Business Information, Data Envelopment Analysis, Artificial Neural Network, Rough set Theory**

I. INTRODUCTION

DM is one of the new concepts that support the corporations in an uncertain world. Its systems are still complex and costly, but when they are designed, built, and operate in the right way, they can deliver a competitive advantage. The aim of the DM and DSS is to provide information at the right time, in the right place and in the right form. This is usually important and invaluable for decision makers.

Manuscript received November 25, 2014.
Ayman E. Khedr, I.S Department Helwan University, Egypt
e-mail- Ayman_Khedr@helwan.edu.eg
Ahmed I. El Seddawy, BIS Department, AAST, Egypt.
e-mail- Ahmed.ALseddawy@aast.edu
Amira M. Idrees, I.S. Department, Fayoum University, Egypt
e-mail- Ami04@fayoum.edu.eg

DM is a process, not a product, for assembling and managing data from various sources for the purpose of gaining a single, detailed view of a part or all of a business. DSS and DM is a computing environment where users can find strategic information. It is a user centric environment where users are put directly in touch with the data to help them make better decisions. The DM includes extracting data from source systems, to be transformed and stored in order to provide user interfaces for easy access. DM offers a variety of advanced data processing techniques that may beneficially be applied for DSS purposes, also analyzing the customers' data through DM techniques can lead to churners' prevention.

The main approach to churn prediction is to model individual customers and derive their likelihood of churn using a predictive model. A predictive model is used to define the hidden patterns in a given data set therefore; it can identify the customers who are aiming to churn. Using a model for prediction, may help in developing effective customer retention programs for telcos. DM offers a variety of advanced data processing techniques that may beneficially be applied for BI purposes. The comprehensive process of applying DSS for a business problem is referred to as the Knowledge Discovery in Databases (KDD) process and is vital for successful DM implementations with DSS applications [5].As our final target is to provide a DSS system for solving CRM problems, to reach this target, different steps will be developed. This research will be followed by a serious of researches to complete our main objective.

This research focus on the first step which to apply an efficient clustering algorithm for categorizing the customers' data. Clustering is the process of organizing data objects into a set of disjoint classes called clusters. Cluster analysis seeks to partition a given dataset into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups [6],

[9]. Numerous methods have been proposed to solve clustering problem. One of the most popular clustering methods is k-mean clustering algorithm developed by Mac Queen in 1967.
The easiness of k-mean clustering algorithm made this algorithm used in several fields. The k-mean clustering algorithm is a partitioning clustering method that separates data into k groups [1], [2], [4], [5], [7], [9]. The k-mean clustering algorithm is more prominent since its intelligence to cluster massive data rapidly and efficiently. However, k-mean algorithm is highly precarious in initial cluster centers. Because of the initial cluster centers produced arbitrarily, k-mean algorithm does not promise to produce the peculiar clustering results. Efficiency of the original k-mean algorithm heavily rely on the initial centroids [2], [5]. Initial centroids also have an influence on the number of iterations required while running the original k-mean algorithm. The computational complexity of the original k-mean algorithm is very high, specifically for massive data sets [2]. Various methods have been proposed in the literature to enhance the accuracy and efficiency of the k-mean clustering algorithm.

This paper presents an enhanced method for finding the better initial centroids and to provide an efficient way of assigning the data points to suitable clusters with reduced time complexity.

This paper is organized as follows. Section 2 presents an overview of k-mean algorithm and a short analysis of the existing clustering methods. Section 3 introduces proposed method. Section 4 describes the proposed algorithm. Section 5 experimentally demonstrates the performance of the proposed approach. And the final Section 6describes the conclusion and future work.

II. THE K-MEAN ALGORITHM

One of the most popular clustering methods is k-mean clustering algorithm. It generates k points as initial centroid arbitrarily, where k is a user specified parameter. Each point is then assigned to the cluster with the closest centroid [3], [4],[10]. Then the centroid of each cluster is updated by taking theme an of the data points of each cluster. Some data points may move from one cluster to other cluster. Again we calculate new centroids and assign the data points to the suitable clusters. We repeat the assignment and update the centroids, until convergence criteria is met i.e., no point changes clusters, or equivalently, until the centroids remain the same. In this algorithm mostly Euclidean distance is used to find distance between data points and centroids [2]. Pseudo code for the k-mean clustering algorithm is described in Algorithm

Algorithm 1: The k-mean clustering algorithm [2]
Require:

1. D = {d1, d2, d3, ... di, ..., dn } // Set of n data points.
2. k // Number of desired clusters
3. Ensure: A set of k clusters.
4. Steps:
5. Arbitrarily choose k data points from D as initial centroids;
6. Repeat
7. Assign each point di to the cluster which has the closest centroid;
8. Calculate the new mean for each cluster;
9. Until convergence criteria is met.

Although k-mean has the great advantage of being easy to implement, the quality of the final clustering results of the k-mean algorithm highly depends on the arbitrary selection of the initial centroids. In the original k-mean algorithm, the initial centroids are chosen randomly and hence we get different clusters for different runs for the same input data [10]. Therefore we can determine the drawbacks of k-mean as (1) Its performance depends highly on initial cluster centers, (2) The number of clusters must be previously known and fixed, and (3) The algorithm contains the dead-unit problem which results in empty clusters. Random k-mean initialization generally leads k-mean to converge to local minima i.e. inacceptable clustering results are produced.

III. RELATED WORK

The original k-mean algorithm is very impressionable to the initial starting points. So, it is quite crucial for k-mean to have refined initial cluster centers. Several methods have been proposed in the literature for finding the better initial centroids. And some methods were proposed to improve both the accuracy and efficiency of the k-mean clustering algorithm. In this paper, some of the more recent proposals are reviewed [1-5], [8]. A. M. Fahim et al. [1] proposed an enhanced method for assigning data points to the suitable clusters. In the original k-means algorithm in each iteration the distance is calculated between each data element to all centroids and the required computational time of this algorithm is depends on the number of data elements, number of clusters and number of iterations, so it is computationally expensive. In Fahim approach the required computational time is reduced when assigning the data elements to the appropriate clusters. But in this method the initial centroids are selected randomly. So

this method is very sensitive to the initial starting points and it does not promise to produce the unique clustering results. K. A. Abdul Nazeer et al. [2] proposed an enhanced algorithm to improve the accuracy and efficiency of the k-means clustering algorithm. In this algorithm two methods are used one method for finding the better initial centroids. And another method for an efficient way for assigning data points to appropriate clusters with reduced time complexity. This algorithm produces good clusters in less amount of computational time. Zhang Chen et al. [3] proposed the initial centroids algorithm based on k-mean that have avoided alternative randomness of initial center. Fang Yuan [4] proposed the initial centroids algorithm. The standard k-mean algorithm selects k-objects randomly from the given data set as the initial centroids. If different initial values are given for the centroids, the accuracy output by the standard k-mean algorithm can be affected. In Yuan's method the initial centroids are calculated systematically. Koheri Arai et al. [5] proposed an algorithm for centroids initialization for k-mean. In this algorithm both k-mean and hierarchical algorithms are used. This method utilizes all the clustering results of k-mean in certain times. Then, the result transformed by combining with Hierarchical algorithm in order to find the better initial cluster centers for k-mean clustering algorithm. A. Bhattacharya et al. [8] proposed a novel clustering algorithm, called Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes. DCCA is able to produce clusters, without taking the initial centroids and the value of k, the number of desired clusters as an input. The time complexity of the algorithm is high and the cost for repairing from any misplacement is also high.

Decision support system can overcome the issues with personnel attributes and specifications. Personnel specifications have greatest impact on total efficiency. They can enhance total efficiency of critical personnel attributes. This study presents an intelligent integrated decision support system (DSS) for forecasting and optimization of complex personnel efficiency. DSS assesses the impact of personnel efficiency by data envelopment analysis (DEA), artificial neural network (ANN), rough set theory (RST), and K-Means clustering algorithm. DEA has two roles in this study. It provides data to ANN and finally it selects the best redact through ANN results. Redact is described as a minimum subset of features, completely discriminating all objects in a data set. The redact selection is achieved by RST. ANN has two roles in the integrated algorithm. ANN results are basis for selecting the best redact and it is used for forecasting total efficiency. Finally, K-Means algorithm is used to develop the DSS. A procedure is proposed to develop the DSS with stated tools and completed rule base. The DSS could help managers to forecast and optimize efficiencies by selected attributes and grouping inferred efficiency. Also, it is an ideal tool for careful forecasting and planning. The proposed DSS is applied to an actual banking system and its superiorities and advantages are discussed.

IV. ENHANCED K-MEAN ALGORITHM

In k-mean clustering algorithm, the initial centroids are selected randomly. So this method is very sensitive to the initial starting points and it does not promise to produce the unique clustering results. In the research [2] we proposed an enhanced algorithm to improve the accuracy and efficiency of the k-mean clustering algorithm. The following points summarize the problems in k-mean algorithm, they are:

1. Quality of the output depends on the initial point.
2. Global optimum solution not guaranteed.
3. Non globular clusters (overlapping in data between clusters)
4. Assume wrong number of clusters. (you did not solve this problem)
5. Find empty clusters.
6. Bad initialization to centroid point
7. Choosing the number of clusters

The most common measure to evaluate k-mean algorithm is Sum of Squared Error 'SSE' with another words called, Sum of Squared Distances 'SSD' [31].

- For each point, the error is the distance to nearest cluster.
- To get SSE or SSD, we square these error and sum them, equation 1 present the formula used for calculating SSE.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- X is a data point in cluster Ci .
- Miis the representative point for cluster Ci.

Equation 1: Calculate Sum of Squared Error

- There is one way to reduce SSE, is to increase K 'Number of clusters'.
- The good clustering with smaller K can have slower SSE than a poor clustering with higher K.

"Enhanced K-mean" applies a modified approach in order to be able to overcome the above mentioned problems of K-mean. The modified approach will include the following:

1. change centriod point from random points to center points (Mid P) for data
2. adding step while calculate distance between data sample and cluster through inserting several center points to increase time of processing
3. adding last step to avoid empty clusters before visualize data

The main idea of "enhanced k-mean" is to change the method of determining the cluster point, "enhanced K-mean" starts with determining a centroid cluster point instead of providing a random cluster point in K-mean. This modification proved to be more accurate in distributing the clusters as will be shown in comparing the results. The second modification applied by the "enhanced K-mean" algorithm considering the determination of the cluster elements. The problem arise when one element can belong to more than one cluster with difference in percentage, the "K-mean" algorithm assign the element to one of these clusters randomly, however, using "Enhanced K-mean", it determines the degree of closeness between the element and its related clusters, and then it selects the most relevant cluster, and assign the element to this selected cluster. The degree of closeness is determined by determining the degree of closeness between the element into consideration and all the elements that are already assigned to the cluster, and the cluster that has his elements most related to the element in consideration is selected to be the most related cluster and then assign the element into consideration to this selected clusters.

The third modification is removing the empty clusters and avoids presenting it to the user. As the number of clusters is determined randomly in the first step, there is a possibility that this number is more than the discovered clusters, and therefore this may cause that some clusters do not have any assigned elements. While "K-mean" algorithm presents all the clusters whether they contain elements or not, the "enhanced K-mean" avoids this inconvenient representation and remove the entire empty clusters before presenting them to the user. We have presented the proposed algorithm in three different methods, figure 1 presents the pseudo code of the main steps for the "enhanced K-mean", while Figure 2 shows the flowchart of enhanced K-M algorithm.

1. Assume number of cluster "K".
2. Calculate 'K' at center points of data set
3. Calculate the distance between a data sample and clusters.
4. Assign a data sample to closest cluster center.
5. Calculate new cluster center.
6. Repeat step 3.4 and 5 until no objects move group.
7. Avoid empty clusters
8. Perform visualization

Fig.1: Enhanced K-mean pseudo code

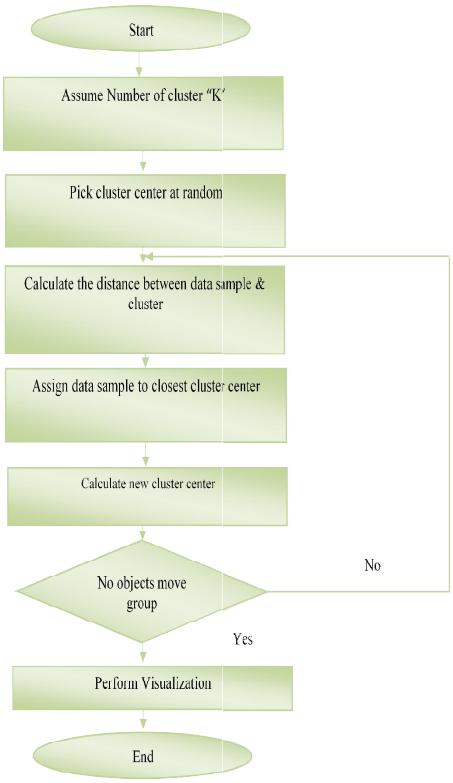

Fig. 2: E_K-Mean algorithm flowcharts

V. COMPARISON BETWEEN K-MEAN ALGORITHM AND ENHANCED K-MEAN ALGORITHM

In this section, we will provide an extensive comparison between "K-mean" and "enhanced K-mean" algorithms

VI. COMPARISON BETWEEN PROCESSING STEPS

First comparison is through the method of processing, figures 3, and 4 provide the main steps of both algorithms, which illustrates the main difference between the processing steps between both algorithms.

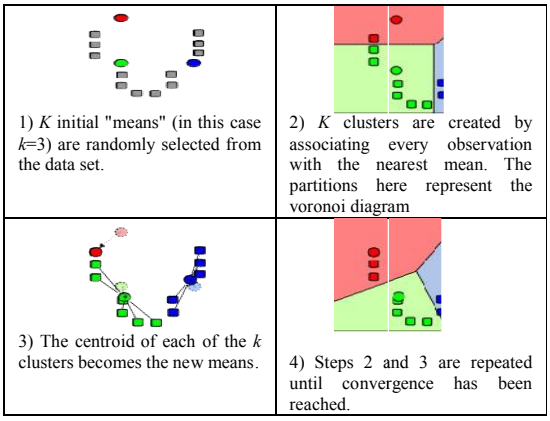

Fig. 3: Example of the original K-mean algorithm processing [30]

VII. Experimental Evaluation

In this section, we present an evaluation of the proposed enhancing algorithm by applying both algorithms on customer investment banking data. As data mining tasks needs pre-steps before applying the data mining algorithms, therefore, the following subsections will demonstrate the main steps we performed for data preparation (section 4.1, 4.2).And in the last subsection (section 4.3), we will describe the proposed enhanced k-mean algorithm.

VIII. Data Gathering Step

In data gathering step, two methods of data gathering are applied; the first method is interviewing experts in bank field who are specialized in investment. Second method is using samples of investment data captured from database of investment department; this data represents some of the transactions for investment department in bank. In this section, we demonstrate a sample of data which was taken from a bank located in Egypt and has several branches distributed among many districts in Egypt. The bank serves more than 100000 customers per year, contracting with more than 50000 organizations. There are some basics that were discussed through the interviews including the overall cash process for usage and resources in the bank. Figure 5 shows the legal relations between the resources and their usage through the bank. As shown in figure 4.2, different resources of the bank are involved, they are:

- Members (Customers) through their accounts including savings accounts, current accounts, savings certificates, deposits accounts and payroll accounts.
- Companies through deposits and current accounts.
- Other Banks through deposits (short term) and current accounts.
- Property Rights resource using equation (capital + reserves + profit stage).

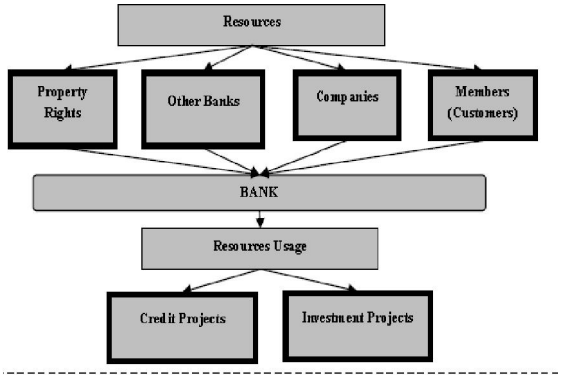

Fig. 5: Resources and usage of cash in bank

The proposed system considered only one of these resources which are the customers' resource. Also as discussed in the interviews, usage bank for cash has several directions, they are Investment Direction, and Credit Direction, figure 6, and 7 presents the types of both directions.

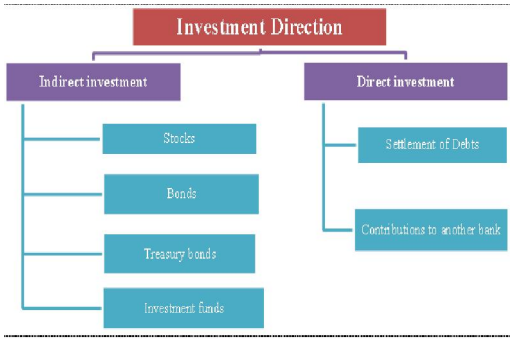

Fig. 6: Investment Direction

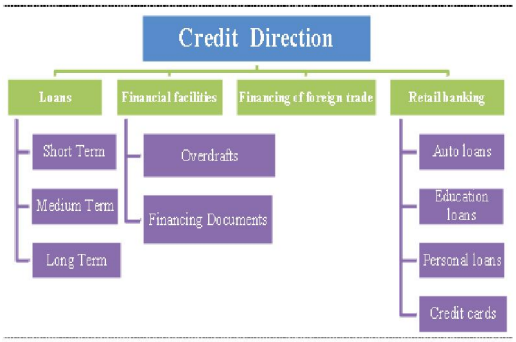

Fig. 7: Credit Direction

IDMS focuses in depth of investment department. All customers' and departments data is stored in SQL database. These data is constructed in the database in six fields; they are Customer name, Customer number or ID, Previous commitments, Paperwork, Type of investment, Sector of the field investment and Previous debt. Table 1 shows a sample of customers' data in SQL database. Then data are converted from SQL format to excel sheets.

Request Date	Request Serial Number	Customer Name	Customer ID	Paperwork	Type of investment	Sector of investment	Previous debt
1-11-2010	1	Ahmed	20123	Ok	Industry	Durable goods	No
5-12-2010	2	Co.	20123	Ok	Securities	Durable goods	No
1-11-2010	3	Samed	20320	Ok	Securities	Cultivation strategies	No
1-11-2010	4	Saaed	20002	Ok	Industry	Import and export	No
1-11-2010	5	Co.	20135	Ok	Agriculture	Cultivation strategies	No
1-11-2010	6	Mohamed	20123	Ok	Industry	Foodstuffs	No
...	....	...	....	....	....	....	....

Table 1: collected data from database

IX. Data Preprocessing Step

In this step, the collected data undergoes three preprocessing steps, as shown in figure 4.1, customer data includes eight attributes. In data preprocessing step, the data is reduced from 8 columns to 7 columns, as we neglected one attribute which is "Customer Name" according to the bank rules of preserving customer data security. Figure 8 shows the proposed steps for preprocessing. In summary, Data set is converted from textual values to numeric values, then selecting main attributes are performed for the system, and finally the selected attributes are categorized into two clusters, they are supervised and unsupervised attributes.

Fig. 8: Data Pre-Processing Step

In the first step, "textual values to numeric values", data is converted from textual values to numeric values in order to deal with identification numbers using Mat lab program, table 2 shows the original format for each filed of the data, and table 3 shows a sample of this data in its original format.

Request Date	Request Serial Number	Customer Name	Customer ID	Paperwork	Type of investment	Sector of the field investment	Previous debt
Date	Number	Text	Number	Text	Text	Text	Text

Table 2: format for each field for data

Request Date	Request Serial Number	Customer Name	Customer ID	Paperwork	Type of investment	Sector of investment	Previous debt
1-11-2010	1	Ahmed	20123	Ok	Industry	Durable goods	No
5-12-2010	2	Co.	20123	Ok	Securities	Durable goods	No
1-11-2010	3	Saved	20520	Ok	Securities	Cultivation strategies	Yes

Table 3: sample of original data

The data is then converted to the required numeric format as shown in the sample of data after conversion in table 4. As shown in table 4, customer name attribute is removed as it will not add any value to the data mining task, however, customer id will play a significant role which will be discussed later.

Presenting some examples of the conversion task, considering the attribute "paperwork", the value "1" means that all papers are ready and accepted, value "0" means that papers are not completed. Another example is the conversion of the 'type of investment" into a corresponding value which can be considered as an "investment type ID".

Request Date	Request Serial Number	Customer ID	Paperwork	Type of investment	Sector of investment	Previous debt
2010	1	20123	1	03	032	1
2010	2	20123	1	03	032	1
2010	3	20520	1	01	011	1
2011	4	20002	1	03	031	1
2012	5	20135	1	01	011	1
2011	6	20123	1	03	033	1
...	....	....	....	....	....	....

Table 4: sample of data after converting to numeric sheet

In the second step, "selecting main attributes", we have conducted interviews with the responsible in the Bank, and the interesting attributes are selected as a

result of these interviews. These interesting attributes are the type of investment and sector of investment.
And finally in the third step "filtering data", we used WEKA program tool in categorizing the attributes is performed to two categories, they are, unsupervised attributes and supervised attributes. One of the supervised attributes is "sector of investment" which is used as a main attribute for the next step "data mining step".

X. Applying Enhanced K-Mean Clustering Algorithm

The aim of this step is to produce set of clusters for banking data based on the investment sector, we propose a comparison between the results of this step using two different algorithms, they are K-mean algorithm, and enhanced K-mean algorithm. By this comparison, we argue that the enhanced K-mean algorithm produce more accurate results for investment banking data which is a step forward for supporting better decision making process. In the following subsections, the results of applying the previously mentioned data mining techniques are proposed. As previously discussed in section 3, we have performed an extensive survey for different clustering algorithms, and according to the strength and weakness of these algorithms, we deduced that the best algorithm for our data is "K-mean" clustering algorithm.

However, "K-mean" algorithm has weakness points which will be discussed at the start of the next sub-section. These weaknesses affect the results when applied on investment banking sectors data. Therefore we proposed an enhancing algorithm "enhanced K-mean" that can overcome these weaknesses to enhance the clustering step results accuracy. We argue that "enhanced K-mean" algorithm is able to perform the clustering process over the investment sector banking data with higher accuracy than the "K-mean" algorithm. The following sections describe the proposed algorithm "enhanced K-mean" and a comparison of the results produced of both algorithms which confirm our arguments.

XI. Comparison between Results of K-Mean and Enhanced K-Mean

Both algorithms are applied to a set of investment banking data; the data set included 5000 records. The results of "K-mean" and "enhanced K-mean" are presented in Tables 5, and 6 respectively. These results show differences in elements distribution among the clusters.

Agriculture	Trading	Tourism	Securities	Industry	Petrochemicals	Non	Non	Technologies	Non
23 %	19 %	17 %	3 %	8.16%	14%	0%	7.09 %	9%	0 %

Table 5: k-mean results

Agriculture	Trading	Tourism	Securities	Industry	Petrochemicals	Technologies
12.57%	14.08%	28.16%	20.71%	8.16%	6.12%	10.20%

Table 6: enhanced k-mean results

The difference between the results is introduced to a number of experts who confirmed the correctness of the "Enhanced K-mean" algorithm. Table 7proposes some samples of elements that are assigned in a cluster using K-mean, while they are assigned to a different cluster using "Enhanced K-mean", this sample demonstrate that enhanced k-mean algorithm is more accurate in defining clusters than the original k-mean algorithm. According to the expert's opinion, they confirmed that the "Enhanced K-man" provided more accurate cluster to these elements. We completed our evaluation by measuring the accuracy percentage of both algorithms by measuring the precision measure, the result showed that "Enhanced K-mean" had precision equal 97.6 % while "K-mean" had precision equal 91.5 %.

Investment Sector	Type of Investment using "K-mean"	Type of Investment using "Enhanced K-mean"
Foodstuffs	Agriculture	Industry
Automakers	Industry	Trading
Shares	Securities	Securities
Import and export	Industry	Industry
Electricity Industry	Agriculture	Trading
Foodstuffs	Agriculture	Industry

Table 7: Examples of clusters for comparison between K-mean &Enhanced K-mean Results

XII. Conclusion

This research proposed "Enhanced K-mean" algorithm, an enhancing algorithm for K-mean clustering algorithm which aims at improving K-mean output accuracy. The results of applying both algorithms on a set of investment banking data which included 5000 records, and these results presented that "Enhanced k-mean" is more accurate than "K-mean" algorithm. This research is the first step

towards building a complete decision support system for investment banking data decision makers.

The targeted system aims to mine investment data in the banking sector to create logical thinking and support decision makers in banking sector to be able to have a high quality decision for investment under uncertain situations and to minimize the risk in cash usage for banker owners. Our future target is to integrate the proposed algorithm with the other components of the system to reach the he determined target. The proposed algorithm can also be applied either in banking sector for other data set, or moreover, it can be applied to other sectors with an expected success for accurate results.

REFERENCES

[1] A. Hunter and S. Parsons, "A review of uncertainty handling formalisms", Applications of Uncertainty Formalisms, LNAI 1455, pp.8-37. Springer -Verlag, 1998.

[2] E. Hernandez and J. Recasens, "A general framework for induction of decision trees under uncertainty", Modelling with Words, LNAI 2873, pp.26–43, Springer-Verlag, 2003.

[3] M. S. Chen, J. Han, and P. S. Yu. IEEE Trans Knowledge and Data Engineering Data mining. An overview from a database perspective, 8:866-883, 1996.

[4] U. Fayyad, G. Piatetsky-Shapiro and W. J. Frawley. AAAI/MIT, Press definition of KDD at KDD96. Knowledge Discovery in Databases, 1991.

[5] Gartner. Evolution of data mining, Gartner Group Advanced Technologies and Applications Research Note, 2/1/95.

[6] International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98), 1995-1998.

[7] R.J. Miller and Y. Yang. Association rules over interval data. SIGMOD'97, 452-461, Tucson, Arizona, 1997.

[8] Zaki, M.J., SPADE An Efficient Algorithm for Mining Frequent Sequences Machine Learning, 42(1) 31-60, 2001.

[9] Osmar R. Zaïane. "Principles of Knowledge Discovery in Databases - Chapter 8 Data Clustering". & Shantanu Godbole data mining Data mining Workshop 9th November 2003.

[10] T.Imielinski and H. Mannila. Communications of ACM. A database perspective on knowledge discovery, 39:58-64, 1996.

[11] BIRCH Zhang, T., Ramakrishnan, R., and Livny, M. SIGMOD '96. BIRCH an efficient data clustering method for very large databases. 1996.

[12] A. M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, "An Efficient enhanced K-Means clustering algorithm", journal of Zhejiang University, 10 (7): 1626 - 1633, 2006.

[13] Chen Zhang and Shixiong Xia, " K-Means Clustering Algorithm with Improved Initial center," in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 790-792, 2009.

[14] F. Yuan, Z. H. Meng, H. X. Zhangz, C. R. Dong, " A New Algorithm to Get the Initial Centroids", proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26-29, August 2004.

[15] Chaturvedi J. C. A, Green P, "K - Modes clustering," Journals of Classification, (18):35–55, 2001.

[16] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, "An Efficient enhanced K-Means clustering algorithm", Journal of Zhejiang University, 10(7):1626–1633, 2006.

[17] K. Thangavel and E.Elayaraja, Salem, Tamilnadu, "Performance Analysis of Enhanced Clustering Algorithm for Gene Expression Data", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011 ISSN (Online): 1694-0814 www.IJCSI.org.