Enhanced K-mean Algorithm to Improve Decision Support System under Uncertain Situations

Ahmed Bahgat El Seddawy[1]
[1]Department of Business Information System, Arab Academy for Science and Technology, Egypt

Turky Sultan[2]
[2]Department of Information System, Helwan University, Egypt

Ayman Khedr[3]
[3]Department of Information system ,Helwan University, Egypt

Abstract:
Decision Support System (DSS) is equivalent synonym as management information systems (MIS). Most of imported data are being used in solutions like data mining (DM). Decision supporting systems include also decisions made upon individual data from external sources, management feeling, and various other data sources not included in business intelligence. Successfully supporting managerial decision-making is critically dependent upon the availability of integrated, high quality information organized and presented in a timely and easily understood manner. Data mining have emerged to meet this need. They serve as an integrated repository for internal and external data-intelligence critical to understanding and evaluating the business within its environmental context. With the addition of models, analytic tools, and user interfaces, they have the potential to provide actionable information that supports effective problem and opportunity identification, critical decision-making, and strategy formulation, implementation, and evaluation. The proposed system Investment Data Mining System (IDMS) will support top level management to make a good decision in any time under any uncertain environment and on another hand using enhancing K-mean algorithm.

Keywords:
dss, dm, mis, clustering, classification, association rule, k-mean, olap, matlab

1. Introduction

Decision Support System (DSS) is equivalent synonym as management information systems (MIS). Most of imported data are being used in solutions like data mining (DM). Decision supporting systems include also decisions made upon individual data from external sources, management feeling, and various other data sources not included in business intelligence. Successfully supporting managerial decision-making is critically dependent upon the availability of integrated, high quality information organized and presented in a timely and easily understood manner. Data mining have emerged to meet this need. They serve as an integrated repository for internal and external data-intelligence critical to understanding and evaluating the business within its environmental context. With the addition of models, analytic tools, and user interfaces, they have the potential to provide actionable information that supports effective problem and opportunity identification, critical decision-making, and strategy formulation, implementation, and evaluation. The proposed system will support top level management to make a good decision in any time under any uncertain environment [4]. This study aim to investigate the adoption process of decision making under uncertain situations or highly risk environments effecting in decision of investing stoke cash of bank. This applied for two types of usage investment - direct or indirect - or credit and any sector of investment will be highly or moderate or low risk. And select which one of this sectors risk 'rejected' or un-risk 'accepted' all that under uncertain environments such as; political, economical, marketing, operational, internal policies and natural crises, all that using the contribution of this study enhancing k-mean algorithm to improve the results and comparing results between original algorithm and enhanced algorithm.

The paper is divided into six sections; section two is a background and related work it is divided into two parts, part one is about DSS, part two is about DM including K-m algorithm and enhancing in K-m algorithm. Section three presents the proposed Investing Data Mining System 'IDMS. Section four presents IDMS experiments, implementations and results using original and enhanced k-m algorithm. Section five presents conclusion and finally section six presents future work.

2. Background and related work

2.1 Decision Support System (DSS)

DSS includes a body of knowledge that describes some aspects of the decision maker's world that specifies how to accomplish various tasks, that indicates what conclusions are valid in different circumstances [4].The expected benefits of DSS that discovered are higher decision quality, improved communication, cost reduction, increased productivity, time savings, improved customer satisfaction

Manuscript received July 5, 2013
Manuscript revised July 20, 2013

and improved employee satisfaction. DSS is a computer-based system consisting of three main interacting components:

- **A language system**: a mechanism to provide communication between the user and other components of the DSS.
- **A knowledge system**: A repository of problem domain knowledge embodied in DSS as either data or procedures.
- **A problem processing system**: a link between the other two components, containing one or more of the general problem manipulation capabilities required for decision-making.

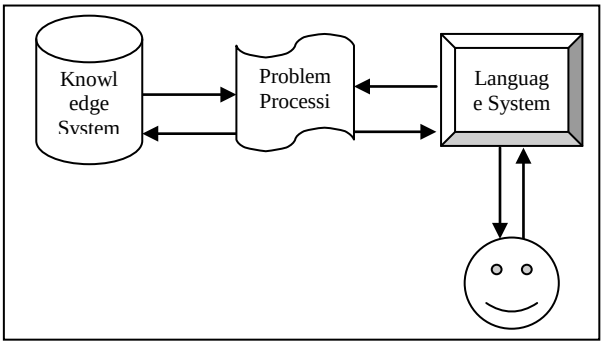

Fig 1: DSS Main Components

After surveying multiple decision support systems, it is concluded that decision support systems are categorized into the following [5]:

- **File drawer systems**: This category of DSS provides access to data items.
- **Data analysis systems**: Those support the manipulation of data by computerized tools tailored to a specific task or by more general tools and operators.
- **Analytical information systems**: Those provide access to a series of decision-oriented databases.
- **Accounting and financial models**: those calculate the consequences of possible actions.
- **Representational models**: those estimate the consequences of actions based on simulation models that include relationships that are causal as well as accounting definitions.
- **Optimization models**: those provide guidelines for actions by generating an optimal solution consistent with a series of constraints.
- **Suggestion models**: those perform the logical processing leading to a specific suggested decision or a fairly structured or well understood task.

This section describes the approaches and techniques mostly used when developing data warehousing systems that data warehousing approaches such as; Online Analytical Processing 'OLAP', Data Mining 'DM' and Artificial Intelligence 'AI'. And in this paper will using DM as approach and technique.

2.2 Data Mining Techniques (DM)

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information [10]. DM techniques are the result of a long process of research and product development [10]. There are several processes for applying DM:

1. Definition of the business objective and expected operational environment.
2. Data selection is required to identify meaningful sample of data.
3. Data transformation that involves data representation in an appropriate format for mining algorithm.
4. Selection and implementation of data mining algorithm depends on the mining objective.
5. Analysis of the discovered outcomes is needed to formulate business outcomes.
6. Representing valuable business outcomes.

DM techniques usually fall into two categories, predictive or descriptive. Predictive DM uses historical data to infer something about future events. Predictive mining tasks use data to build a model to make predictions on unseen future events. Descriptive DM aims to find patterns in the data that provide some information about internal hidden relationships. Descriptive mining tasks characterize the general properties of the data and represent it in a meaningful way. Figure2 shows the classification of DM techniques.

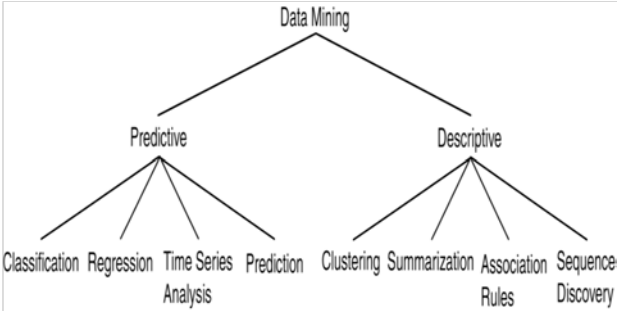

Fig 2: DM Techniques [5]

2.2.1 Clustering Technique

Clustering can be considered the most important learning problem; like every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".
A cluster is therefore a collection of objects which are "similar" between them while are "dissimilar" to the objects belonging to other clusters as shown in figure 3. [6] Define clustering by saying "Clustering involves identifying a finite set of categories or segments 'clusters' to describe the data according to a certain metric. The clusters can be mutually exclusive, hierarchical or overlapping". [6] Defines clustering as follows: "clustering enables to find specific discriminative factors or attributes for the studied data. Each member of a cluster should be very similar to other members in its cluster and very dissimilar to other clusters. When a new data is introduced, it is classified into the most similar clusters. Techniques for creating clusters include partitioning methods as in k-means algorithm, and hierarchical methods as decision trees, and density-based methods".

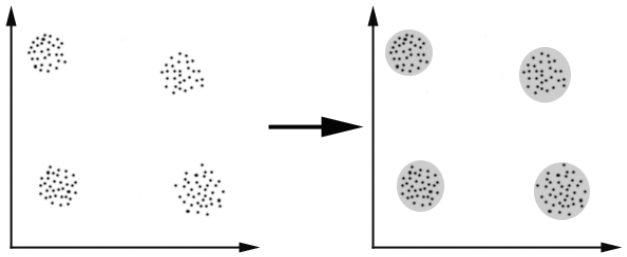

Fig 3: Simple graphical for clustering data [6]

The goal of clustering is [6] to determine the basic grouping in a set of unlabeled data, to find representatives for homogeneous groups which are called "data reduction", to find 'natural clusters' and describe their unknown properties these clusters are called 'natural' data types, to find useful and suitable groupings this can be called 'useful' data classes' and to find unusual data objects outlier detection. There are several advantages for the clustering technique use it. Among these advantages are recognizing the number of clusters, grouping similar members together, identifying the discriminate attributes, ranking the discriminate attributes, and recognizing the discriminate attributes of one cluster and representing them within the business context. A major disadvantage of clustering is that it is suitable for static data but not for dynamic data where its value changes over time or due to any other factor. Another disadvantage is that sometimes the generated clusters may not have a practical meaning. Finally, it is possible not to spot the cluster sometimes as there is no exact idea of what to look for or there is a lack of collected data [7].

2.2.1.1 K-Mean Algorithm

[8] Defines K-means as one of the simplest unsupervised learning algorithms that solve the well known clustering problem. Build to classify or grouping objects based on features into 'k' number of group. K is positive integer number and the grouping is done by mining the sum of squares of distance between data and the corresponding cluster centriod. The cluster centroid is the average point in the multidimensional space defined by the dimensions [9]. There are a lot of applications of the K-mean clustering, range from unsupervised learning of neural network, Pattern recognitions, Classification analysis, Artificial intelligent, image processing, machine vision, etc. In principle, we have several objects and each object have several attributes and we want to classify the objects based on the attributes, then we can apply this algorithm. There are commonly four steps followed for K-mean idea [10];

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centers.
2. Assign each object to the group that has the closest centered.
3. When all objects have been assigned, recalculate the positions of the K centered.
4. Repeat Steps 2 and 3 until the centered no longer change. This produces a separation of the objects into groups from which the metric to be minimized can be calculated as follows.

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left\|x_i^j - c_j\right\|^2$$

K: number of clusters
nj: number of points in jth cluster
xij: ith point in jth cluster

```
Input:
    D = {t1, t2, ..., tn}    // Set of elements
    A      // Adjacency matrix showing distance between elements.
    k      // Number of desired clusters.
Output:
    K      // Set of clusters.
K-Means Algorithm:
    assign initial values for means m1, m2, ..., mk ;
    repeat
        assign each item ti to the cluster which has the closest mean
        calculate new mean for each cluster;
    until convergence criteria is met;
```

Fig 4: Procedures code for k-means algorithm [10]

2.2.1.2 Enhanced K-means 'E_KM'

Before looking to the enhancements that are applied on K-mean algorithm, one should give the impression of pros, cons and problems of k-mean algorithm, then presents the enhanced k-mean algorithm and presents a way for evaluating this enhancement. There are several strengths and weakness shown in table 5.1 (Arthur, 2006).

Table 1: Strengths and weakness of K-men algorithm (Arthur, 2006)

Strengths	Weakness
1. Vectors can flexibly change clusters during the process. 2. Always converges to a local optimum 3. Quite fast for most application	1. Quality of the output depends on the initial point. 2. Global optimum solution not guaranteed.

There are some problems in k-mean algorithm such as (Arthur, 2006):

1. Non globular clusters (overlapping in data between clusters)
2. Assume wrong number of clusters.
3. Find empty clusters.
4. Bad initialization to centriod point
5. Choosing the number of clusters

The most common measure to evaluate k-men algorithm is Sum of Squared Error 'SSE' with another words called, Sum of Squared Distances 'SSD' (Wenyan Li, 2009).

- For each point, the error is the distance to nearest cluster.
- To get SSE or SSD, we square these error and sum them

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

x is a data point in cluster Ci .
mi is the representative point for cluster Ci.

- 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.

After investigating the most common algorithms for data clustering, there is enhancement for the most familiar algorithm which called k-means trying to solving some problems of k-mean algorithm. More features will be added to enhanced algorithm such as; change centriod point from random points to center points for data and adding step to avoid empty clusters when visualize data it shows also in steps of algorithm. Figure 5.31 shows the flowchart of enhanced K-men algorithm.

1. Assume number of cluster "K".
2. Calculate 'K' at center points of data set

H → Y →Rows

W → X → Columns

Mid P = (X,Y)

P = ($\frac{W}{2}$, $\frac{H}{2}$) % for every cluster

Ex,

H= 400, W= 30

Mid P = ($\frac{30}{2}$, $\frac{400}{2}$) P = (15, 200) ← ***Center point for Cluster***

Fig 6: Pseudo code of calculating 'k' center point

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

```
Loop through each cluster
    If all items inside cluster equal 0 ,
delete cluster
    Else
    do nothing;
```

Fig 7: Pseudo code of avoiding empty clusters

8. Perform visualization

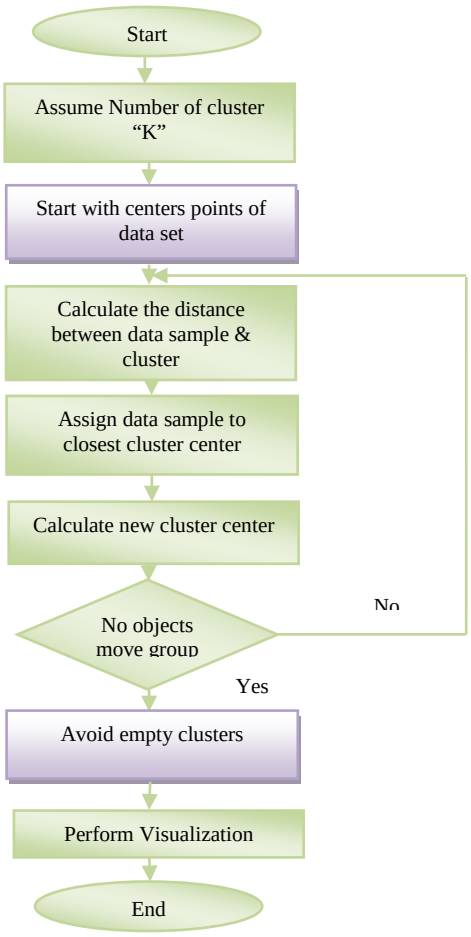

Fig 8: E_K-mean algorithm flowcharts

3. IDMS Implementation

Investment Data Mining System 'IDMS' aims to build a data mining system for investment in the banking sector. IDMS consists of several components; data gathering, preparing data to discover knowledge, data preprocessing, using data mining techniques in sequences steps start with classification data, clustering data especially using K-mean algorithm and enhanced K-mean algorithm to set which best result and then set and run association rules to solve problem, post processing and finally get result and visualize result to create best decision to take a good decision for investment under uncertain situations [18].

3.1 Experiments

3.1.1 Data Collected

The following data is taken from a bank is located in Egypt and has several branches. The bank serves more than 100000 customers per year, contracting with more than 50000 organizations. There are some basics found through interviews such as; explain overall process in bank for usage and resources of cash in bank. IDMS will focus in depth of investment department. All customers and departments data's is stored electronically using SQL database. The database stores values in six fields: Customer name, Customer number or ID, Previous commitments, Paperwork, Type of investment, Sector of the field investment and Previous debt, shown in table 2. The received data is in the form of SQL database converted to excel sheets.

Table 2: Collected 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	Samed	20520	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 3: Format for each field for data

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

3.2 Preprocessing

The data collected undergoes four preprocessing steps and the data matrix is reduced from 8 columns to 7 columns. The first step converts data from textual values to numeric ones in order to deal with identification numbers as in table 4.

Table 4: Example of data after converting to numeric sheet

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
…	….	….	….	….	….	….

In the second step, the interesting selected attributes are Request data, Customer ID, Type of Investment and Sector of investment as shown in table 5. Elements with values indicate that the type and sector of investment.

Table 5: Example of data after selecting main segments

Request Date	Customer ID	Type of investment	Sector of investment
2010	20123	03	032
2010	20520	03	032
2010	73002	01	011
2011	20135	03	031
2012	20033	01	011
….	….	….	….

IDMS execution done via several techniques started with clustering technique using K-M and enhanced k-mean algorithm, classification technique using ID3 algorithm and association rules technique using apriori algorithm. Next section will discuss the results of execution for fist technique clustering.

4. Clustering Results

4.1 Clustering sectors

First: Original K-mean Algorithm results

Table 6 presents the set of data used in the implementation experiments for training data to set a best result and choosing an effective number of clusters and percentage of data set to apply the technique for the system. These data are divided among 10 clusters representing the different percentage of the set of data used using K-Mean algorithm.

Table 6: Testing data on a 10 clusters model using K-M Algorithm

C 0	C 1	C 2	C 3	C 4	C 5	C 6	C 7	C 8	C 9
Agriculture	Trading	Tourism	Securities	Industry	Petrochemicals	Non	Non	Technologies	Non
23 %	19 %	17 %	3 %	8.16%	14%	0%	7.09 %	9%	0 %

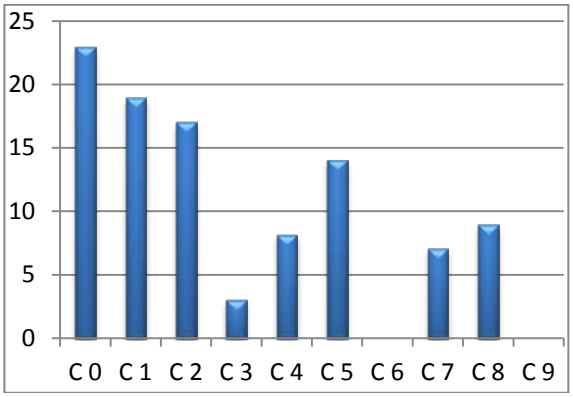

Fig 10 : Distribution percentages of sectors in testing set of data for a 10 cluster model

These graph present clustering of sectors are used in investment sector to find a good usage for cash in bank. The results which are appearing from this algorithm describe the distribution percentage of sectors in testing set of data for 10 clusters by K-M, this enhance consider effect on data results on IDMS.

Second: Enhanced K-mean Algorithm results

After applying enhanced k-mean algorithm to avoid empty clusters and re clustering data more accurate using change center point for 'K', it gives the research more accurate results appear in next section. Table 7 presents the set of data used in the implementation experiments for training data to set a best result and choosing an effective number of clusters and percentage of data set to apply the technique for the system. These data are divided among 7 clusters representing the different percentage of the set of data used using enhanced K-Mean algorithm E_K-M.

Table 7: Testing data on a 7 clusters model using E_K-M Algorithm

C 0	C 1	C 2	C 3	C 4	C 5	C 6
Agriculture	Trading	Tourism	Securities	Industry	Petrochemicals	Technologies
28.57%	4.08%	8.16%	2.04%	8.16%	6.12%	10.20%

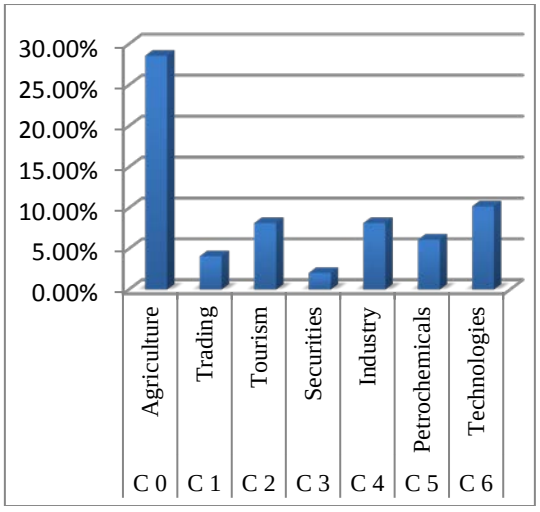

Fig 11: Distribution percentages of sectors in testing set of data for a 10 cluster model

These graph present clustering of sectors are used in investment sector to find a good usage for cash in bank. The results which are appearing from this algorithm describe the distribution percentage of sectors in testing set of data for 7 clusters by E_K-M, this enhance consider effect on data results on IDMS.

4.2 Clustering risks based on sectors using E_K-M

Table 8 presents the set of data used in the implementation experiments for training data to set a best result and choosing an effective number of clusters and percentage of data set to apply the technique for the system. These data are divided among 3 clusters representing the different percentage of the set of data used and how distribute risk of sectors after clustering them to seven clusters using in IDMS.

Table 8: Testing data on a 3 cluster model using E_K-M Algorithm

C 0 Low Risk	C 1 Mid Risk	C 2 High Risk
Agriculture (1) Petrochemicals (6)	Trading (4) Technologies (7) Industry (2)	Tourism (5) Securities (3)
48.57%	40.08%	11.35%

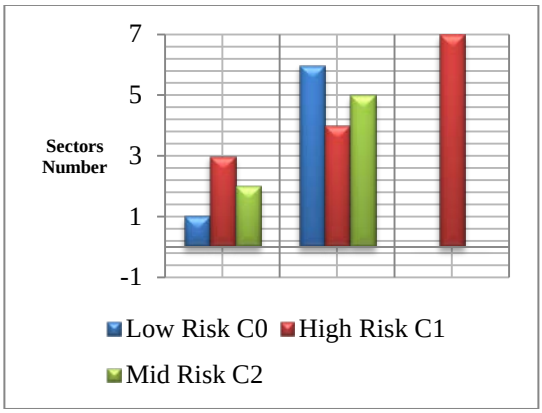

Fig 12: Distribution sectors numbers based on risk level

Figure 12 represent distribution of sectors number based on risk level. Every level of risk has colure such as, red color for high risk to invest in this sector, green color for medium level of risk to invest in this sector and blue color for low risk to invest in this sector.

4.3 Result of Comparison between K-M and E_K-M

There are some similarities and differences between algorithms and their enhancement which describe the utility, efficiency, applicability, accuracy, popularity, flexibility and visualization of the two algorithms. A comparison between the two algorithms is shown in the following table 9.

Based on comparison table finding that, first traditional K-mean algorithm as the size of the data set decreases and

the number of clustering increases, the amount of overlapping elevates and the level of accuracy declines. Second the efficiency is shown in E_K-M and original K-M in two aspects. Time efficiency in original K-M is less than time efficiency in E_K-M because of adding two steps of enhancement, so time will increase and it may be accepted when raise the quality of results. Size efficiency in original K-M is less than space of saving data efficiency in E_K-M because there are steps added to K-M to avoid empty clusters. On the other hand, using the enhanced K-mean algorithm decreases the size of the data set and increases the number of clusters. Accordingly, the amount of overlaps decreases and accuracy increases. Generally, the enhanced K-mean algorithm showed a better performance compared to traditional K-mean algorithm.

Table 9: Comparison between KM and E_K-M algorithms

Approach	K-Mean Algorithm	E_K-mean Algorithm
1. Theory	K-mean start with assume number of clusters, picking 'K' randomly, calculate distance and visualize.	E_K-mean start with assume number of clusters, picking 'K' at center point, calculate distance and avoid empty clusters then visualize.
2. Efficiency (Time and Space)	**Time** : Less efficient **Space** : More Efficient	**Time :** More efficient **Space :** Less Efficient
3. Applicability	Applicable	Applicable
4. Accuracy	Less accurate	More accurate
5. Popularity	Popular	Under Testing
6. Flexibility	Same level	Same level
7. Visualization data	Use visualization	Use visualization
8. Generalization	Widely used	Limited use
9. Number of cluster	Use small and huge numbers	Use small and huge numbers
10. Type of data source	Use all type of data source	Use all type of data source
11. Size of data source	Huge and small data	Huge and small data

5. Conclusions

This paper represents applying clustering technique by enhancing K-M algorithm for DSS in banking sector especially in investment department under uncertain situations which has been rarely addressed before. IDMS is a new proposed system which is simple, straightforward with low computation needs. The proposed preprocessing component is an aggregation of several known steps. The post processing component is an optional one that eases the interpretation of the investment results. The banking is planning a set of actions in accordance of IDMS outcomes for decision making in investment sector. The investment department in the banking is starting to analyze the approached investment sector, to introduce a good decision under uncertain situation after enhance K-M algorithm to give high accurate and high quality data, that shown in comparison table.

Future work

In next step of this study implementing this proposed approach using classification and association techniques to give full image and best result for high level of management with a good decision and high accurate results.

ACKNOWLEDGMENT

I want to express my deepest gratitude for my professor's supervisors Prof. Dr. Turky Sultan and Dr. Ayman Khedr for their major help and support through all the phases of research and development.

REFERENCE

[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] Pascal Poncelet, Florent Masseglia and Maguelonne Teisseire (Editors). Information Science Reference. Data Mining Patterns New Methods and Applications, ISBN 978 1599041629, October 2007.
[13] Thearling K, Exchange Applications White Paper, Inc. increasing customer value by integrating data mining and campaign management software, 1998.
[14] Noah Gans, Spring. Service Operations Management, Vol. 5, No. 2, 2003.
[15] Joun Mack. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. An Efficient k-Means Clustering Algorithm, Analysis and Implementation, VOL. 24, NO. 7, JULY 2002.
[16] Andrew Moore and Brian T. Luke. Tutorial Slides, K-means and Hierarchical Clustering and K-Means Clustering, Slide 15, 2003.
[17] E. Turban, J. E. Aronson, T. Liang, and R. Sharda, Decision Support and Business Intelligence Systems, eighth edition. Prentice Hall, 2007.
[18] Ahmed El Seddawy, Ayman Khedr, Turky Sultan, “Adapted Framework for Data Mining Technique to Improve Decision Support System in an Uncertain Situation”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.2, Jun 2012.

Ahmed B. El Seddawy, M.S. degrees in Information System from Arab Academy for Science and Technology in 2009. He now with AAST Egypt Teacher Assisting.