International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol. 3, Issue. 4, Jul - Aug. 2013 pp-2139-2146 ISSN: 2249-6645

Applying Classification Technique using DID3 Algorithm to improve Decision Support System under Uncertain Situations

Ahmed Bahgat El Seddawy[1], Prof. Dr. Turky Sultan[2], Dr. Ayman Khedr[3]
[1]*Department of Business Information System, Arab Academy for Science and Technology, Egypt*
[2, 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 will support top level management to make a good decision in any time under any uncertain environment using classification technique by DID3algorithm.*

Key words: *DSS, DM, MIS, CLUSTERING, CLASSIFICATION, ASSOCIATION RULE, K-MEAN, OLAP, MATLAB*

I. 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 four 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. Section three presents the proposed Investing Data Mining System 'IDMS. Section four presents conclusion and finally section five present future works2. Tables, Figures and Equations.

II. BACKGROUND AND RELATED WORK

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 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.

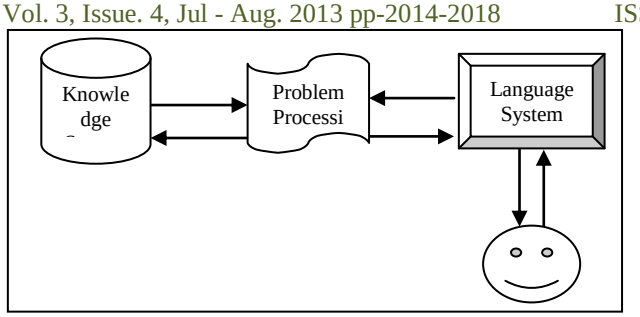

Fig 1: DSS Main Components

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. 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]

2.1 Classification Technique

Classification is to use the model to predict the class of objects whose class label is unknown.

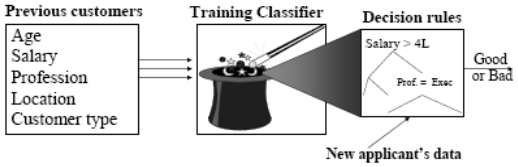

Fig 3: Example for classification procedures

Classification based on Bayes Theorem of classification is "composed of two steps supervised learning of a training set of data to create a model, and then classifying the data according to the model. Some well-known classification algorithms include Bayesian Classification, decision trees, neural networks and back propagation based on neural networks, k-nearest neighbor classifiers based on learning by analogy, and genetic algorithms" [10]. Classification can be used for predicting the class label of data objects. However, in many applications, one may like to predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data, and is often specially referred to as prediction. Although prediction may refer to both data value prediction and class label prediction, it is usually connected to data value prediction and thus is distinct from classification. Also, prediction encompasses the identification of distribution trends based on the available data. Classification and prediction may need to be preceded by relevance analysis which attempts to identify attributes that do not contribute to the classification or prediction process.

2.1.1 Classification Tree

Trees used for regression and trees used for classification have some similarities - but also some differences, such as the procedure used to determine where to split. Some techniques use more than one decision tree for their analysis:

- A Random Forest classifier uses a number of decision trees, in order to improve the classification rate.
- Boosted Trees can be used for regression-type and classification-type problems.
- Rotation forest - in which every decision tree is trained by first applying principal component analysis (PCA) on a random subset of the input features.

There are many specific decision-tree algorithms such as, ID3 algorithm, C4.5 algorithm, CHi-squared Automatic Interaction Detector 'CHAID'. Performs multi-level splits when computing classification trees. And MARS extends decision trees to better handle numerical data.

2.1.1.1 ID3 Algorithm

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan (1983) [74]. The basic idea of ID3 algorithm is to create a decision tree of given set, by using top-down greedy search to check each attribute at every tree node [75].

Table 2: Strengths and weakness of ID3 algorithm

Strengths	Weakness
• Understandable prediction rules are created from the training data. • Builds the fastest tree. • Builds a short tree. • Only need to test enough attributes until all data is classified. • Finding leaf nodes enables test data to be pruned, reducing number of tests. • Whole dataset is searched to create tree.	• Data may be over-fitted or over-classified, if a small sample is tested. • Only one attribute at a time is tested for making a decision. Classifying continuous data may be computationally expensive, as many trees must be generated to see where to break the continuum.

For building ID3 algorithm decision tree consists of nodes and arcs or sweeps which connect nodes. To make a decision, one starts at the root node, and asks questions to determine which arc to follow, until one reaches a leaf node and the decision is made. This basic structure is shown in figure 4.6

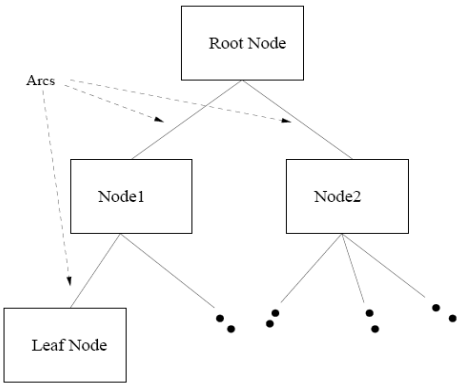

Fig 4: Basic decision tree structure

The main ideas behind the ID3 algorithm are:

Step 1: Each non-leaf node of a decision tree corresponds to an input attribute, and each arc to a possible value of that attribute. A leaf node corresponds to the expected value of the output attribute when the input attributes are described by the path from the root node to that leaf node.

Step 2: In a "good" decision tree, each non-leaf node should correspond to the input attribute which is the most informative about the output attribute amongst all the input attributes not yet considered in the path from the root node to that node. This is because we would like to predict the output attribute using the smallest possible number of questions on average.

Step 3: Entropy is used to determine how informative a particular input attribute is about the output attribute for a subset of the training data. Entropy is a measure of uncertainty in communication systems introduced by Shannon (1948). It is fundamental in modern information theory.

```
function ID3 (I, O, T) {
/*  I is the set of input attributes
*   O is the output attribute
*   T is a set of training data
*
*   function ID3 returns a decision tree
*/
    if (T is empty) {
        return a single node with the value "Failure";
    }
    if (all records in T have the same value for O) {
        return a single node with that value;
    }
    if (I is empty) {
        return a single node with the value of the most frequent value of
        O in T;
        /* Note: some elements in this node will be incorrectly classified */
    }

    /* now handle the case where we can't return a single node */
    compute the information gain for each attribute in I relative to T;
    let X be the attribute with largest Gain(X, T) of the attributes in I;
    let {x_j| j=1,2, .., m} be the values of X;
    let {T_j| j=1,2, .., m} be the subsets of T when T is partitioned
        according the value of X;
        return a tree with the root node labelled X and
        arcs labelled x_1, x_2, .., x_m, where the arcs go to the
        trees ID3(I-{X}, O, T_1), ID3(I-{X}, O, T_2), .., ID3(I-{X}, O, T_m);
}
```

Fig 5: Pseudo Cod of ID3 Algorithm

Enhanced ID3 to 'DID3'

After taking data from clustering technique by 'E_K-m' algorithm, and before inserting to next technique association rules need to classify data by ID3 algorithm, but before use it need to decompose data to several parts to can be managed, to be more accurate and faster that by enhancing ID3 using adding sample step called 'Decompose Data' to equal sized subsets. It shows also in next steps of algorithm.

1. Insert data from (E_K-m) to ID3,
2. Dividing data sets to 'K clusters',

Decompose data to K classification;

3. Run original DID3,
4. Create nods based on tree classification;

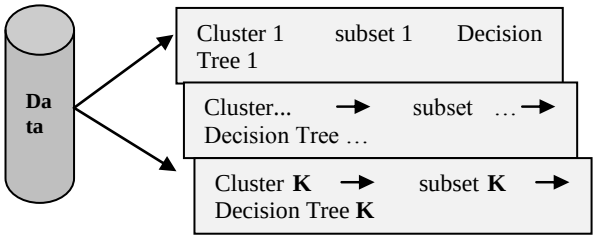

Fig 6: Example of decomposition data using ID3 Algorithm

3. Implementation

IDMS execution done via several techniques started with clustering technique using K-M and enhanced k-mean algorithm - it published in last paper [18], classification technique using ID3 algorithm that discussed in this paper. Next section will discuss the results of execution for second technique classification by DID3 algorithm.

4. Results

First step in classification results that get classification for all sectors related to every cluster which clustered in last step using enhanced K-M. Second step classification for uncertain situations may be appearing in IDMS for decision making.

4.1 Classification sectors and fields

Table 5.4 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 the results of the simulation are shown in Table 5.4 for classify fields based on sectors.

Table 3: Simulation result of ID3 algorithm

Total Number of Instances	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (value)	Time Taken (seconds)	Kappa Statistic
32	89.7143 % (157)	10.2857 % (18)	3.19	0.7858

In WEKA, all data is considered as instances and features in the data are known as attributes. The simulation results are partitioned into several sub items for easier analysis and evaluation. On the first part, correctly and incorrectly classified instances will be partitioned in numeric and percentage value and subsequently Kappa statistic, mean absolute error and root mean squared error will be in numeric value only. Also show the relative absolute error and root relative squared error in percentage for references and evaluation. The results of the simulation are shown in Table 5.5.

Table 4: Simulation errors

Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
0.1062	0.3217	22.2878	65.1135

These graphs present classifications results of classified sectors that are used in investment department to find a good usage for cash in bank based on given data using ID3 algorithm through WEKA tools. The results have shows as next figures.

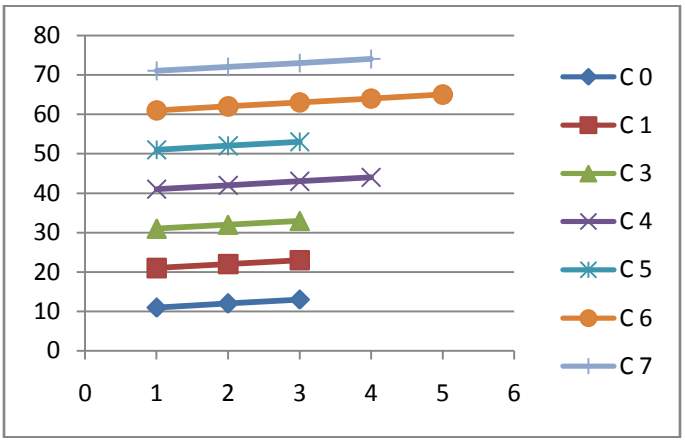

Fig 7: Distribution segments of sectors in testing set of data for a 7 classifications

Figure 7 describes the results of testing data. The results show that classification 1 of the data exist in cluster 0 refer to agriculture field and contain followed sectors, classification 2 of the data exist in cluster 1 refer to industry field and contain followed sectors, classification 3 of the data exist in cluster 2 refer to securities field and contain followed sectors, classification 4 of the data exist in cluster 3 refer to trading field and contain followed sectors, classification 5 of the data exist in cluster 4 refer to tourism field and contain followed sectors, classification 6 of the data exist in cluster 5 refer to petrochemicals field and contain followed sectors and classification 7 of the data exist in cluster 6 refer to technologies field and contain followed sectors.

Fig 8: Classification trees in testing set of data for a 7 classifications of fields

Figure 8 describes the results of testing data. These graph present clustering of sectors are used in investment sector to cluster sectors based on risk level.

4.2 Classification Uncertain Situations

Table 4 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 the results of the simulation are.

Table 5: Simulation result of ID3 algorithm

Total Number of Instances	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (value)	Time Taken (seconds)	Kappa Statistic
28	89.7121 % (157)	10.2871 % (18)	60.19	0.7858

Table 6: Simulation errors

Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error
0.1062	0.211	10.809 %	89.191 %

Figure 9 present classifications results of classified sectors that are used in investment department to find a good usage for cash in bank based on given data using DID3 algorithm through WEKA tools. The results have shows as next figures.

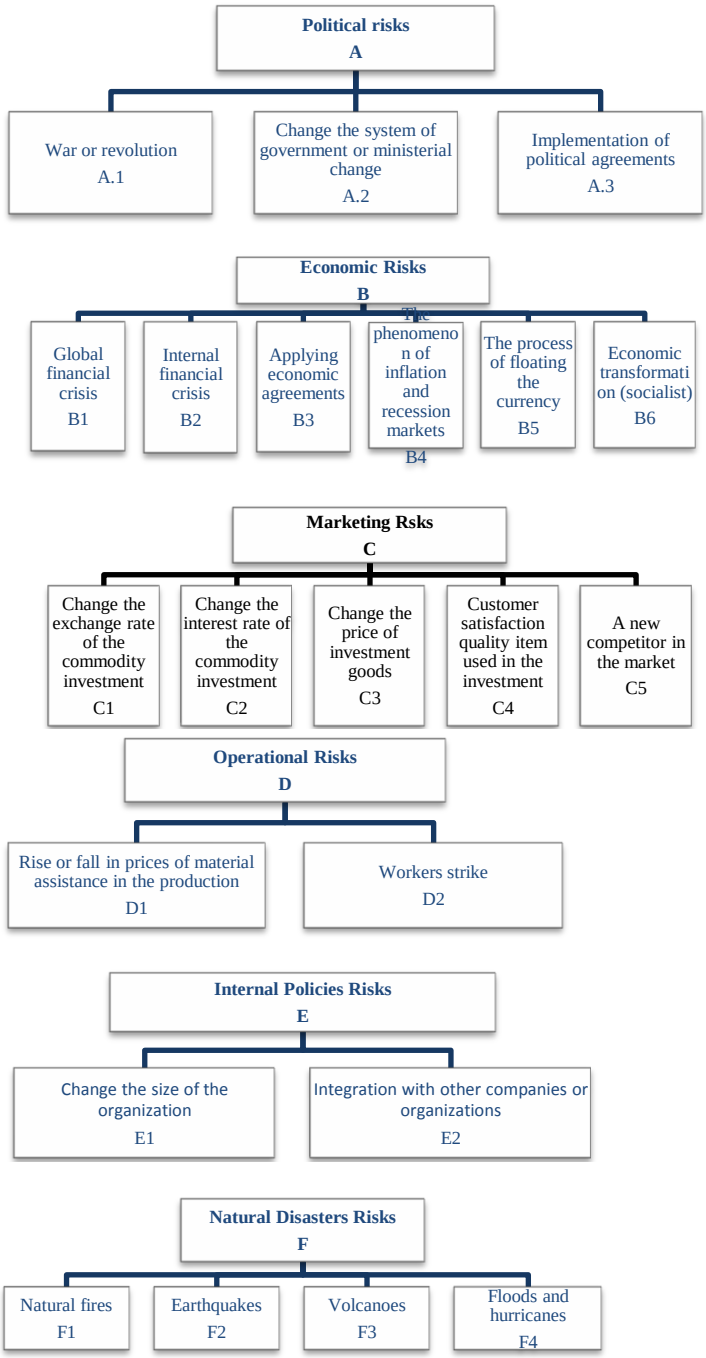

Fig 9: Classification trees in testing set of data for a 6 classifications of uncertain situations

Figure 9 describes the results of testing data. These graph present classification of fields based on sectors to use in IDMS.

III. CONCLUSIONS

This paper represents applying DM using classification technique using enhanced ID3 algorithm to "DID3" algorithm for DSS in banking sector especially in investment department 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.

IV. FUTURE WORK

In next step of this study implementing this proposed approach using association technique using apriori algorithm to give us a best result and support 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, "Enhanced K-mean Algorithm to Improve Decision Support System under Uncertain Situation", International Journal of Modern Engineering Research (IJMER) Vol.2, No.3, Aug 2012.

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