Classification and Selection of Best Saving Service for Potential Investors using Decision Tree – Data Mining Algorithms

Amritpal Kaur, Sandeep Singh

Abstract— This research delineates a comprehensive and successful application of decision tree induction to large banking data set of different banking services obtained by numbers of customers. Complex interaction effects among banking services that lead to increased policy variability have been detected. The extracted information has been confirmed by the database managers, and used to improve the decision process. The research suggests that decision tree induction may be particularly useful when data is multidimensional, and the various process parameters and highly complex interactions.

In order to classify and identify effective and beneficial saving service and design the appropriate criteria for selecting the right scheme for different persons having different taste, this study developed a data mining framework for analyzing banks and post office data, in which suitable[1] technique is employed to extract rules between present saving schemes. In other words, the objectives of this thesis are

- ***To classify the available saving services of banks and post office to good, medium and bad level.***
- ***To select the best saving service according the investor's choice and its preference.***
- ***To guide the potential investor to invest his money in the particular scheme so as to get more benefits.***
- ***To help to take the right decision for investment.***
- ***To reduce the time to take particular decision as there will be no need to analyze each and every available investment scheme thoroughly.***

Index Terms—, CHAID, C4.5, , cluster, data mining, decision tree induction, ID3 .

I. INTRODUCTION

A database is an organized and typically large collection of detailed facts concerning some domain in the outside world. The aim of Data Mining is to examine this database for regularities that may lead to a better understanding of the domain described by the database. In Data Mining we generally assume that the database consists of a collection of individuals. Depending on the domain, individuals can be anything from customers of a bank to molecular [2] compounds or books in a library. For each individual, the database gives us detailed information concerning the different characteristics of the individual, such as the name and address of a customer of a bank, or the accounts owned.

Manuscript received on April, 2013.

Amritpal Kaur received her MSc degree in IT from Khalsa College, Patiala, Punjab in 2010. Currently pursuing MTech degree in Computer Science and Engineering at Lovely Professional University (LPU), Phagwara, Punjab, India.

Sandeep Singh, Assistant Professor, Computer Science and Engineering at Lovely Professional University, Punjab. He received his B.TECH (computer science) degree from NIT Kurukshetra and M.TECH(computer science) degree from NIT Kurukshetra.

Decision tree is one of the important analysis methods in classification. It builds its optimal tree model by selecting important association features. While selection of test attribute and partition of sample sets are two crucial parts in building trees. Different decision tree methods will adopt different technologies to settle these problems. Traditional algorithms include ID3, CART, SPRINT, SLIQ etc. ID3 is the representation of decision tree method. It is easy to understand and has fast classified speed which is applicable to large datasets. Many decision tree algorithms are improved based on it, like CART, SLIQ. [3]But these algorithms more or less have some problems in selection of test features, type of samples, memory utilization of data and the pruning of trees etc. Presently, researchers have present many improvements.

II. DATA MINING TECHNIQUES

The term data mining refers to the broad spectrum of mathematical modeling techniques and software tools that are used to find patterns in data and uses to build models. Classical data mining techniques include classification of users, finding associations between different product items or customer behavior and clustering of users.

Cluster analysis: the process in which similar objects are grouped into multiple classes. Cluster analysis helps to discover the different customer groups and also help to analyze the characteristics of each group. This is quite helpful technique for the market analyst.

Classification: the classification approach includes mining processes intended to discover rules that define whether the technique involves two sub processes, building a model and predicting classification. Item belong to a particular subset or class of data.

Association Analysis: association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis.

III. DECISION TREES

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown on the right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable.[4] Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is

repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees[7] (TDIDT) is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data, but it is not the only strategy. In fact, some approaches have been developed recently allowing tree induction to be performed in a bottom-up fashion.

CHAID stands for Chi-squared Automatic Interaction Detector. Although it can be used for regression problems, in this paper I will only build classification trees [1]. The "Chi-squared" part of the name arises because the technique essentially involves automatically constructing many cross-tabs, and working out statistical significance of the proportions. The most significant relationships are used to control the structure of a tree diagram. Because the goal of classification trees is to predict or explain responses on a categorical dependent variable, the technique has much in common with the techniques used in the more traditional methods of Discriminant Analysis, Cluster Analysis, Nonparametric Statistics, and Nonlinear Estimation. The flexibility of classification trees makes them a very attractive analysis option, but this is not to say that their use is recommended to the exclusion of more traditional methods. Indeed, when the typically more stringent theoretical and distributional assumptions of more traditional methods are met, the traditional methods may be preferable.

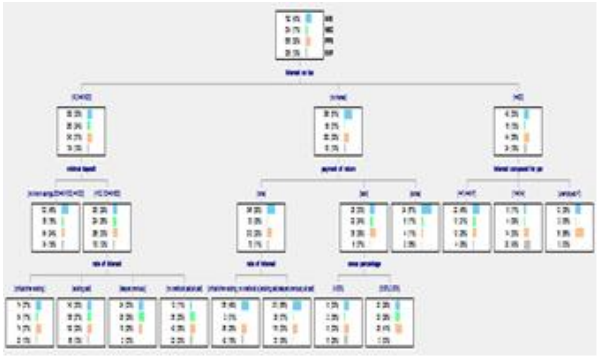

Fig 1 decision tree

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy [6]. The training data is a set $S = s_1, s_2 \ldots$ of already classified samples. Each sample $s_i = x_1, x_2 \ldots$ is a vector where $x_1, x_2 \ldots$ represent attributes or features of the sample. The training data is augmented with a vector $C = c_1, c_2 \ldots$ where c_1, c_2 represent the class to which each sample belongs.

At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub – lists.

IV. PROPOSED WORK

In this work, I propose a technology based on data mining algorithms for the induction of decision trees. It is well suited in our context for various reasons.

1. Presents algorithms CHAID, C4.5 for mining large-scale high dimensional datasets - Classification of large datasets is an important data mining problem. Many classification algorithms have been proposed in the literature, but studies have shown that so far no algorithm uniformly outperforms all other algorithms in terms of quality. Classification tree construction is that separates the scalability aspects of algorithms for constructing a tree from the central features that determine the quality of the tree. The generic algorithm is easy to instantiate with specific split selection methods from the literature (including C4.5, CART, CHAID, FACT, ID3 and extensions, SLIQ.

2. Enhance the efficiency of building decision tree - Propose a series of pruning techniques that can greatly improve the efficiency of the construction of decision trees.

3. Analysis between computation times - A problem is regarded as inherently difficult if its solution requires significant resources, whatever the algorithm used. The theory formalizes this intuition, by introducing mathematical models of computation to study these problems and quantifying the amount of resources needed to solve them, such as time and storage. Other complexity measures are also used, such as the amount of communication (used in communication complexity), the number of gates in a circuit (used in circuit complexity) and the number of processors (used in parallel computing). One of the roles of computational complexity theory is to determine the practical limits on what computers can and cannot do.

4. Eliminate error rate - these algorithms try to reduce the rate of errors made by a predictive model. It is one minus the accuracy.

V. SIPINA OVERVIEW

SIPINA is especially intended to decision trees induction (says also Classification Trees). SIPINA is a Data Mining Software which implements various supervised learning paradigms. SIPINA is an academic tool; it is free for all kind of activities. SIPINA is distributed on the web since 1995, it runs under Windows OS (W95 and later). My main preoccupation is to build an experimentation engine for research activities. I use Sipina for my researches work sharing Sipina on the web; I hope it will help other researchers also.

SIPINA is mainly a Classification Tree Software (specialized on Classification Trees algorithms such as ID3, CHAID, C4.5, ASSISTANT-86, etc....). But, other supervised methods are also available (e.g. k-NN, Multilayer perceptron, Naive Bayes, etc.). We can perform performances comparison and model selection.

SIPINA is a data mining tool. But it is also a machine learning method. It corresponds to an algorithm for the induction of decision graphs. A decision graph is a generalization of a decision tree where we can merge any two terminal nodes of the graph, and not only the leaves issued from the same node.

The SIPINA method is only available under the version 2.5 of SIPINA data mining tool. This version has some drawbacks. Among others, it cannot handle large datasets (higher than 16.383 instances). But it is the only tool which implements the decision graphs algorithm. This is the main reason for which this version is available online to date. If we want to implement a decision tree algorithm such as C4.5 or CHAID, or if we want to create interactively a decision tree ,

it is more advantageous to use the research version (named also version 3.0). The research version is more powerful and it supplies much functionality for the data exploration.

Experiments on the « bank service» dataset. We use the "bank service" dataset in order to evaluate the relevance of the solution implemented into SIPINA. The data file contains 2000 instances and 17 predictive attributes. This is not really a very large dataset. But on a personal computer, handling this kind of data is already a challenge.

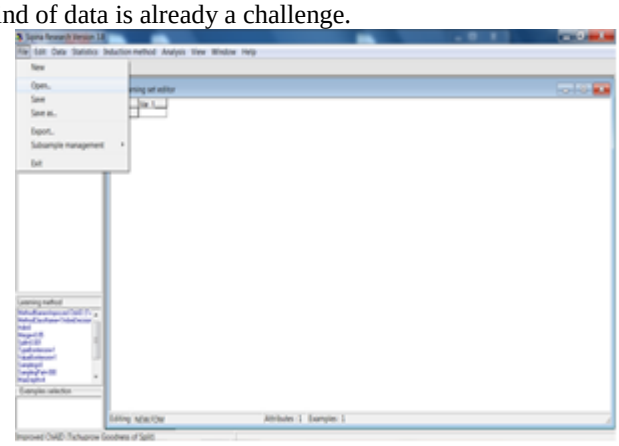

Figure 2: SIPINA 3.8

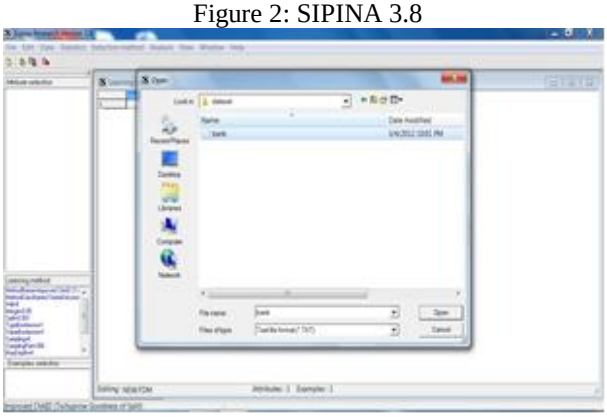

Figure 3: Importing the dataset

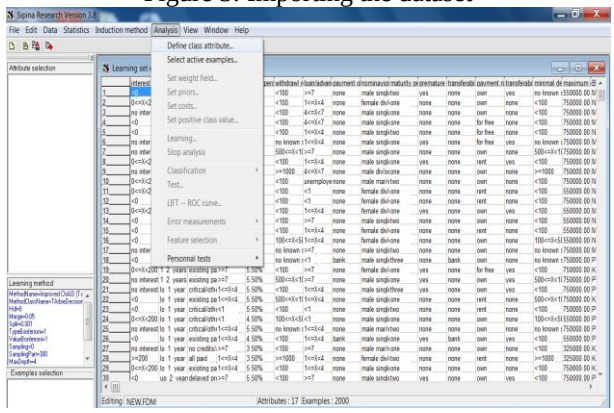

Figure 4: Display 17 attributes with 2000 instances.

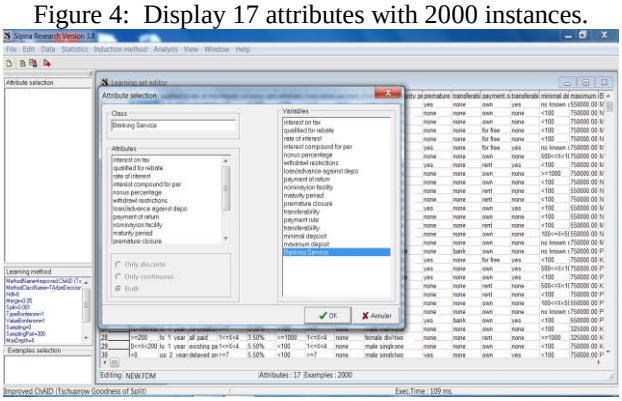

Fig 5: DEFINE CLASS ATTRIBUTE

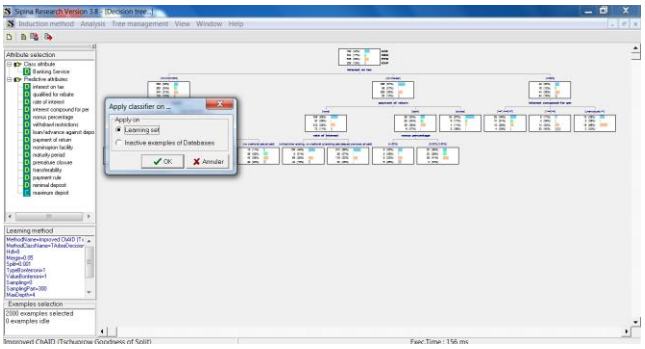

Fig 6: Select the LEARNING SET

VI. CONCLUSION

According to analysis on Banking Data Set, find out best saving service is MIS after implementation of both decision tree techniques.

VII. ACKNOWLEDGMENT

I take this opportunity to express a deep sense of gratitude towards my guide Asst. Prof. Sandeep Singh, for providing excellent guidance, encouragement and inspiration throughout the work. I would like to thank him for his constant assistance and overwhelming interest in the research that has made it successful. I express my deep sense of gratitude and appreciation towards him. Without his invaluable guidance, this work would never have been a successful one.

I would also like to thank my family and friends who have been a source of encouragement and inspiration throughout the duration of this research.

REFERENCES

[1] Alex Berson, *Data Warehousing, Data Mining & OLAP*, pp. 351.
[2] Gu Xiang, *A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree*. New York: Springer-Verlag, 1985, ch. 4.
[3] B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.
[4] E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," *IEEE Trans. Antennas Propagat.*, to be published.
[5] J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," *IEEE J. Quantum Electron.*, submitted for publication.
[6] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
[7] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style)," *IEEE Transl. J. Magn.Jpn.*, vol. 2, Aug. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Magnetics* Japan, 1982, p. 301].
[8] M. Young, *The Techincal Writers Handbook.* Mill Valley, CA: University Science, 1989.
[9] (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). *Title* (edition) [Type of medium]. Volume(issue). Available: http://www.(URL)
[10] J. Jones. (1991, May 10). Networks (2nd ed.) [Online]. Available: http://www.atm.com
[11] (Journal Online Sources style) K. Author. (year, month). Title. *Journal* [Type of medium]. Volume(issue), paging if given. Available: http://www.(URL)
[12] R. J. Vidmar. (1992, August). On the use of atmospheric plasmas as electromagnetic reflectors. *IEEE Trans. Plasma Sci.* [Online]. *21(3).*