Volume 3, Issue 9, September 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Proposed Work for Classification and Selection of Best Saving Service for Banking Using Decision tree Algorithms

Hardeep Kaur*
M. Tech CSE (Student),SBBSIET,
Punjab, India

Harpreet Kaur
Assistant Professor (CSE), SBBSIET,
Punjab, India

Abstract— Data mining is a class of database application that finds hidden patterns in a large amount of data. These hidden patterns are useful for analysis and can easily predict the future behavior. This paper presents earlier literature of algorithms used in field of data mining such as ID3, CART and C4.5. These algorithms are decision tree classifiers that are used for classification of data. Decision trees are attractive because they can generate more understandable results as compared to other data mining techniques. Then a method is proposed that classifies and select the appropriate and beneficial service for a customer by analyzing banking data. In proposed work, decision tree classification approach is used to classifying the large amount of data.

Keywords— Data mining, Classification, Decision tree, C4.5 algorithm, ID3 algorithm, CART algorithm.

I. INTRODUCTION

Data mining is a knowledge discovery process in which analysis of the data store in very large repositories is done using different perspectives and useful information is obtained from result. Data mining refers to finding the hidden patterns from large amount of data. Data mining is the concept which can be applied in various fields such as banking, insurance, medicine, real estate. In Insurance and banking field data mining is used to detect frauds, identify the loyal customers, increase sales and enhance research. There are various data mining tools such as traditional data mining tools, dashboards and text mining. Traditional data mining tools are used when a number of complex algorithms and techniques are used by companies to find data patterns and trends. A dashboard reflects data changes and update on the screen. Dashboards are installed on computer and are used to monitor the database information. Text data mining tools mines data from different kind of text. These tools change the format of text that is compatible with database tool.

There are various data mining techniques such as clustering, classification, association. Clustering refers to the process in which similar objects are grouped to form multiple classes. Association rules are in the form of if/then statements. It analyzes data that is seemingly unrelated and discovers relationships from large dataset. Data classification is used to classify large data using certain rules in order to gain knowledge. The data mining process is an iterative process which typically consists of following phases:

1. Problem Definition: The process of data mining starts with understanding the business problem. Data mining experts, business experts and domain experts define objectives and requirements. Then objectives translated into a data mining definition.
2. Data exploration: The meaning of metadata understood by domain experts. Data is collected, described and explored by domain experts.
3. Data preparation: In this phase a data model is built by domain experts for modeling process. Data is collected, cleansed and formatted by domain experts.
4. Modeling: Various data mining functions can be selected and applied by domain experts on data. Different data mining functions can be used for same type of data mining problem.
5. Evaluation: The model is evaluated by data mining experts. If the model does not meet their expectations then process again started from modeling phase to rebuild the model until they are finally satisfied with the model. Then at the end of evaluation phase, a decision is made by data mining experts how to data mining results can be used.
6. Deployment: Deployment is the process in which data mining uses mining results and results are exported into tables or into other applications.

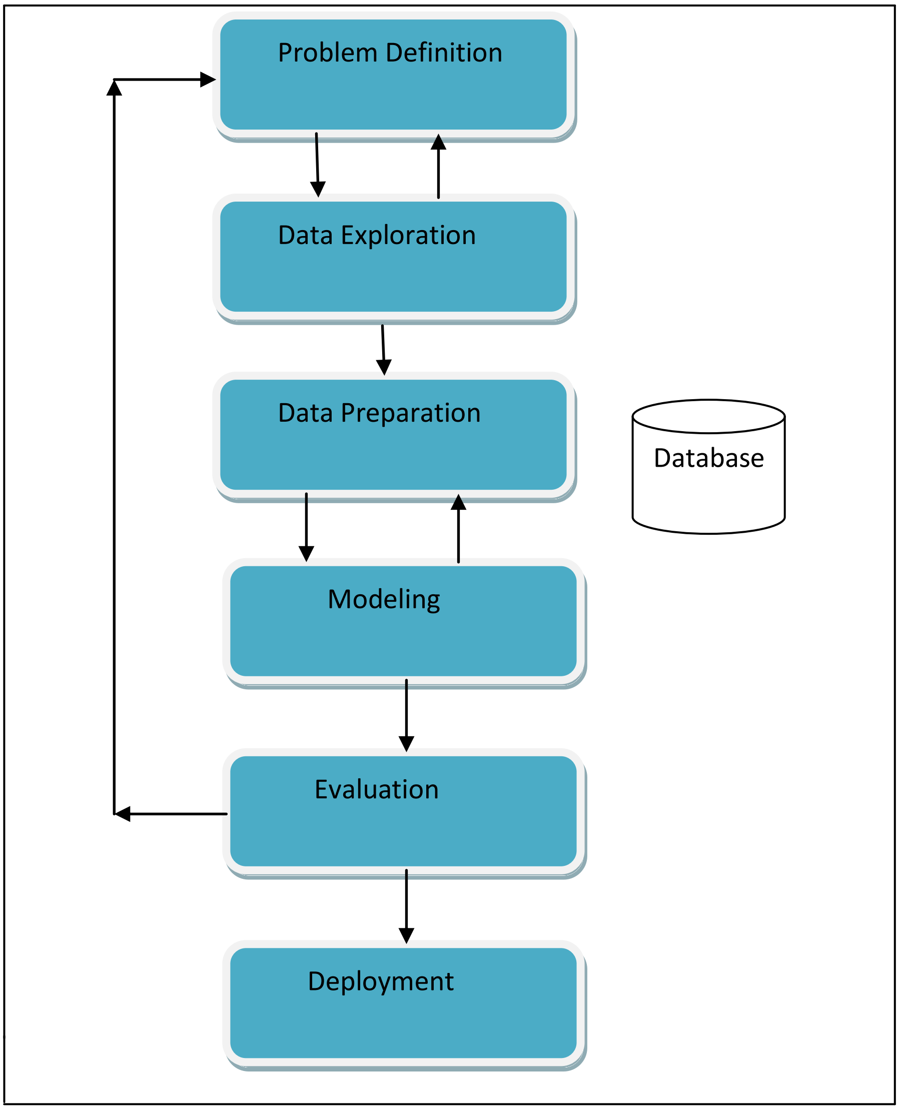
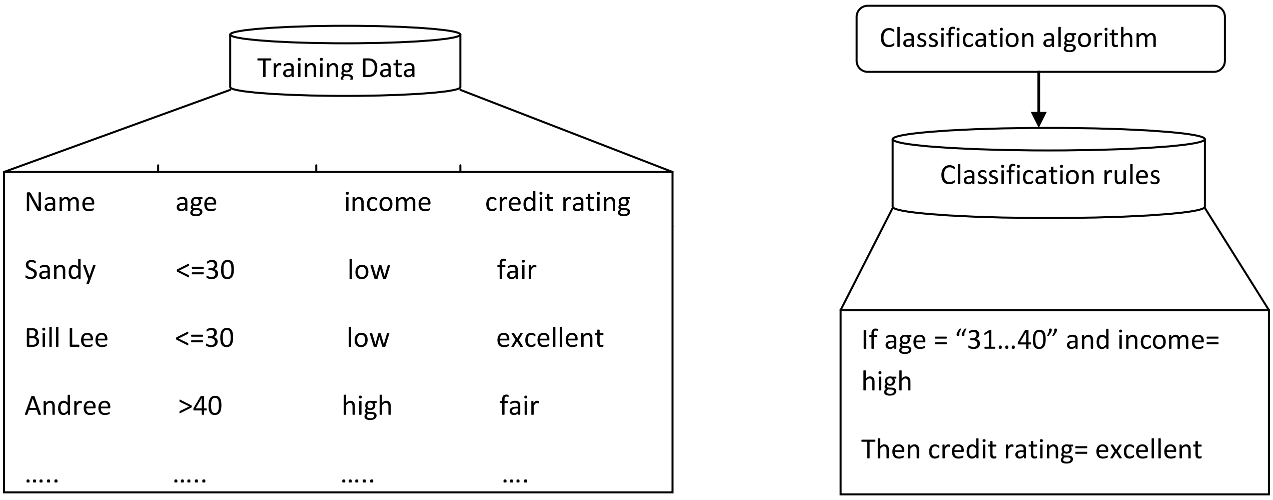


Fig.1 Data Mining Process Model

II. DATA CLASSIFICATION

Data classification is a two step process which involves building a classifier and classification. The first step is building the classifier which is also known as learning step (Training phase). In this step training data analysis is done by a classification algorithm. The second step is known as classification. In this step, test data are used to estimate the accuracy of classification rules.

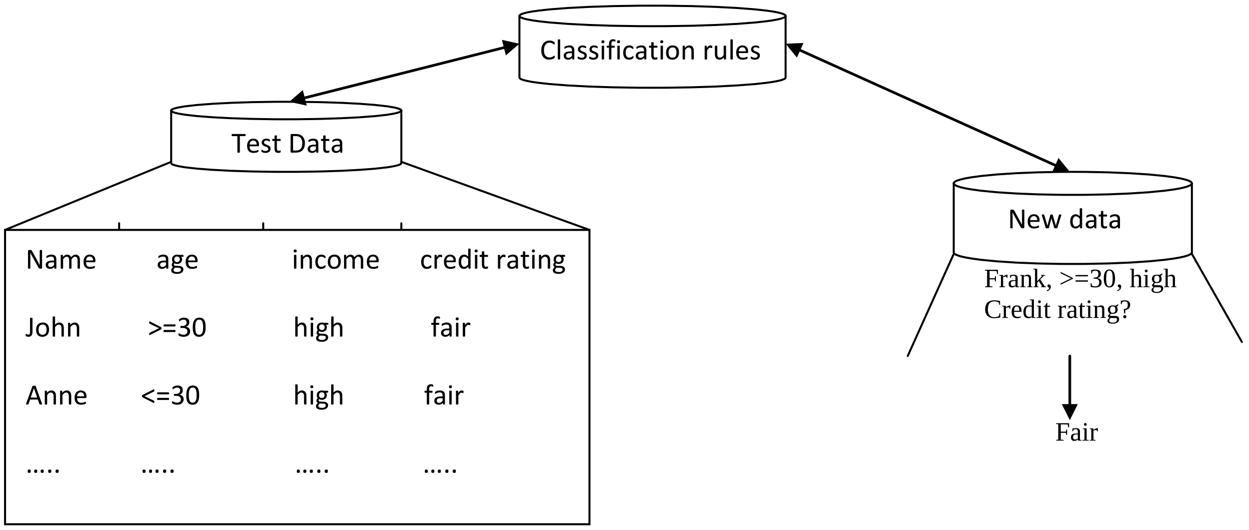


Name	age	income	credit rating
Sandy	<=30	low	fair
Bill Lee	<=30	low	excellent
Andree	>40	high	fair
…..	…..	…..	….

If age = "31…40" and income= high

Then credit rating= excellent

Fig.2 Learning Phase of Data Classification



Name	age	income	credit rating
John	>=30	high	fair
Anne	<=30	high	fair
…..	…..	…..	…..

Fig.3 Data Classification

There are two areas of classification: Decision Tree Induction and Neural Induction. Decision tree induction is a classification method in which a tree is generated by recursively applying a set of rules on a dataset. These set of rules represents the model of different classes for a given datasets. Decision tree induction creates a decision tree that helps to find out valuable information from a large dataset.

III. DECISION TREE

A decision tree is a flow chart like tree structure which consists of internal node, branch and leaf node. The internal node represents a test on the attribute. Internal node is non leaf node. Branch of the decision tree represents the outcome of test. The leaf node represents class label. The leaf node is also known as terminal node. The topmost node of the tree represents root node of tree. The following diagram shows a decision tree for the concept of buy_computer. The tree indicates whether a customer is likely to purchase a computer or not.

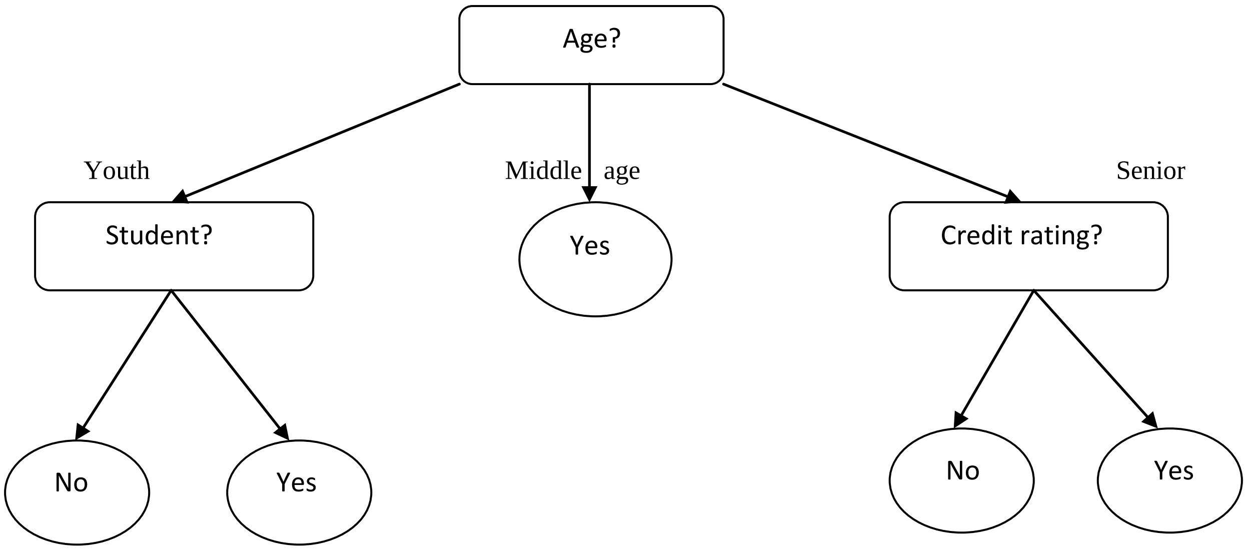


Fig.4 A decision tree for the concept of buy_computer

A decision tree generates understandable rules. They can easily handle both numerical as well as categorical attributes. Decision trees provide a clear induction of fields which is necessity during the prediction or classification. There are various decision tree algorithms that are widely used. Following are some decision tree algorithms:

A. ID3

ID3 is a decision tree algorithm which is introduced by Quinlan Ross [1] in 1986. ID3 uses Hunts algorithm to construct the tree in two phases. Tree building and pruning are the two phases of ID3 algorithm. The splitting criterion of attribute is based on information gain measure. Only categorical attributes are acceptable for building a tree model. If there is noise is present in the data then this algorithm does not give accurate results. So a noise pre-processing technique is used to remove noise.

To construct a decision tree, at each and every attribute information gain is calculated and the attribute having highest information gain is labeled as a root node. The arcs are denoted by the rest possible values. After that all the outcome instances that are possible are examined to find whether they belong to same class or not. If all the instances falls in same class then a single name class is used to denote the node otherwise the splitting attribute is chosen for the classification of the instances. Pruning is not supported by ID3.

B. C4.5

C4.5 algorithm is a successor of ID3 introduced by Quinlan Ross [2]. C4.5 is also based on Hunt's algorithm as ID3 but C4.5 can also handles continuous attributes for decision tree construction. To handle continuous as well as categorical attributes, C4.5 uses the selected threshold to split an attribute into two partitions. The values that are above the threshold are denoted by child and remaining is as another child. C4.5 can also handle the missing attribute values.

To build the decision tree, the attribute selection is based on gain ratio. It can remove information gain biasness if there are many outcome values of an attribute are present. For a given dataset, firstly gain ratio is calculated for each node and the node with maximum gain ratio denoted by root node. The pessimistic pruning approach is used to remove branches that are not necessary in the decision tree. The pessimistic pruning approach is used to improve the accuracy of classification.

C. CART

Classification and Regression Trees (CART) [3] is a classification method in which construction of a decision tree is based on historical data. A splitting rule is used to build a classification tree. The splitting rule splits the learning sample into smaller parts.

Classification trees have classes. Regression trees do not have classes. The splitting rule in regression tree is based on squared residuals minimization algorithm in which the expected sum variances for two resulting nodes should be minimized. CART algorithm is nonparametric and variables are not selected in advance.

IV. BACKGROUND AND RELATED WORK

Data mining techniques can be used in learning process for properly identifying, extracting and evaluating variables related to banking. Kazi Imran Moin and Dr. Qazi Baseer Ahmed [4] provide an overview about data mining and applications of data mining in banking area and some applications of data mining in some core business processes are highlighted to enhance the performance. Mrs. Swati .V. Kulkarni [5] presents a novel technique take input of result obtained by analyzing the data and a set of actions are produced to transform the customers from undesirable class to desirable class. For this purpose decision tree algorithms are used. Pardeep Kumar, Nitin, Vivek Kumar Sehgal and Durg Singh Chauhan [6] predict the accuracy, error rate, training time, classification index, comprehensibility of different classifiers such as C4.5, CHAID, QUEST and K-means on different data sets such as mushroom, Vote, Nursery and Credit.Smith Tsang, Ben Kao, Kevin Y. Yip, Wai- Shing Ho and Sau Dan Lee [7] proposed classifiers that can handle uncertain information. During the process of data collection the values uncertainty can be arises in many applications. So to remove this problem and improve the efficiency a series of pruning methods were proposed.

Surjeet Kumar Yadav and Sourabh Pal [8] discussed decision tree classification method and evaluate the performance of students in examination. The evaluation is done with the help of C4.5, ID3 and CART decision tree classifiers on educational data and performance of student in final examination is predicted.

Vivek Bhambri [9] discussed about the role of data mining in baking and financial institution for acquiring new customers, real time fraud detection, customer's purchase patterns analysis, detecting the emerging trends and launching new products and services.

R.R Kabra and R.S.Bichkar [10] predict the use of decision tree in educational data mining. Decision tree algorithms are used to generate a model and performance of students is predicted from the generated model.

Milija Suknovic, Boris Delibasic, Milos Jovanovic, Milan Vukicevic, Dragana Becejski Vujaklija and Zoran Obradovic [11] propose a generic decision tree framework that supports design of reusable components. The decision tree induction algorithms namely ID3, C4.5, CART, CHAID, QUEST, GUIDE, CRUISE and CTREE were analyzed to propose a generic decision tree framework. The proposed generic decision tree framework consists of several sub-problems that are recognized through the analysis of decision tree induction algorithms.

V. PROPOSED WORK

The aim of proposed work is to find efficiency, error rate and execution time of the algorithms ID3, C4.5 and CART. The appropriate analysis will be done by applying decision tree classifier ID3, C4.5 and CART on banking dataset and results are compared for different decision tree classifiers. Then the results of the algorithms will be applied on high dimensional dataset to build a novel method which classify and select the appropriate service for customer.

VI. SOURCES OF DATA SET

Data is collected from two sources one is primary sources and another is secondary source. Data collected from primary source is known as primary or raw data; whereas data collected from secondary source is known as secondary data.

VII. CONCLUSIONS

Classification is one of the interesting topics for researchers to accurately and efficiently classify data for knowledge discovery. In data mining decision trees are very attractive because they can express natural language readily. Decision tree classifiers are used to analyze the data and experiment will be conducted to find efficiency and performance of classifiers on banking data. Decision tree algorithms ID3, C4.5, CART can learn effective predictive model from banking dataset. This work will also help to find the services that can be useful for the customer.

REFERENCES

[1] J. R. Quinlan, '*Introduction of decision tree*', Journal of Machine learning, 1986.

[2] J. R. Quinlan, '*C4.5: Programs for Machine Learning'*, Morgan Kaufmann Publishers, Inc, 1992.

[3] Aman Kumar Sharma and Suruchi Sahni, ''A Comparative Study of Classification Algorithms for Spam Email Data Analysis'', International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 5, pp. 1890-1895, 2011.

[4] Kazi Imran Moin and Dr. Qazi Baseer Ahmed, ''Use of Data Mining in Baking'', International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 2, pp.738-742, 2012.

[5] Mrs. Swati .V. Kulkarni, ''Mining knowledge using Decision Tree Algorithm'', International Journal of Scientific & Engineering Research, Volume 2, Issue 5, 2011.

[6] Pardeep Kumar, Nitin, Vivek Kumar Sehgal and Durg Singh Chauhan, ''A BENCHMARK TO SELECT DATA MINING BASED CLASSIFICATION ALGORITHMS FOR BUSINESS INTELLIGENCE AND DECISION SUPPORT SYSTEMS'', International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.2, No.5, pp. 25-42, 2012.

[7] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai- Shing Ho and Sau Dan Lee, ''Decision Trees for Uncertain Data'', *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, VOL.23, NO. 1, 2011.

[8] Surjeet Kumar Yadav and Sourabh Pal, '' Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification'', World of Computer Science and Information Technology Journal (WCSIT) , Vol. 2, No. 2, pp. 51-56, 2012.

[9] Vivek Bhambri, ''Role of Data Mining in Banking Sector'', International Indexed & Referred Research Journal, VoL.III, ISSUE-33, pp. 70-71, 2012.

[10] R.R Kabra and R.S.Bichkar, ''Performance Prediction of Engineering Students using Decision Trees'', International Journal of Computer Applications, Volume 36, No.11, pp. 8-12, 2011.

[11] Milija Suknovic, Boris Delibasic, Milos Jovanovic, Milan Vukicevic, Dragana Becejski-Vujaklija and Zoran Obradovic, ''Reusable components in decision tree induction algorithms'', Springer, 2011.