"FINANCIAL BANKING DATASET COMPARISION BY USING RULE-BASED CLASSIFICATION METHODS IN MACHINE LEARNING TOOL"

Er. Sangita , Er. Parminder Singh, Dr.Naveen DhillionT

Abstract

Data mining is the process is to extract information from a data set and transform it into an understandable structure. There are several major data mining techniques *have been developing and using in data mining projects recently including* classification, clustering, prediction, sequential patterns *and* decision tree. *With the huge amount of information available online, the World Wide Web is a fertile area for data mining research. The data sets which contain marketing data can be used for two different business goals. Prediction of the results of the marketing campaign for each customer and clarification of the factors which affect the campaign results, and second are finding out customers segments, using data for the customers. In order to optimize the marketing campaigns with the help of data sets, we can use these steps, initially import data from the data sets and perform high level analysis, clean the irrelevant data and predict data by using machine learning techniques. Classification is a major technique in data mining and widely used in various fields. Four rule based classification algorithm considered are Decision Table, One R, PART and Zero R. For comparing the four algorithm three performance parameters number of correct or incorrect instances, error rate and execution time are considered. This research work also shows that which algorithm is most suitable for predicting the performance of the selected algorithms. Our work shows the process of WEKA analysis of file converts and selection of attributes to be mined and comparison with Knowledge Extraction of Evolutionary Learning not only analysis the data mining classifications but also the genetic, evolutionary algorithms is the best efficient tool in learning.*

.

Key Words *–Data Mining, Classification, Decision Table, One R, Part, Zero R and WEKA.*

1. INTRODUCTION

In this era of digital age and with the improvement in computer technology, many organizations usually gather large volumes of data from operational activities and after which are left to waste in data repositories. Any tool that will help in the analysis of these large volumes of data that is being generated daily by many organizations is an answered prayer. With the aid of improved technology in recent years, large volumes of data are usually accumulated by many organizations and such data are usually left to waste in various data repositories [1].

Data Mining is all about the analysis of large amount of data usually found in data repositories in many organizations. Its application is growing in leaps and bounds and has touched every aspect of human life ranging from science, engineering to business applications [2]. Data mining can handle different kinds of data ranging from ordinary textand numeric data to image and voice data. It is a multidisciplinary field that has applied techniques from other fields especially statistics, database.

The growth of the internet has created a vastnew arena for information generation. There is huge amount of data available in Information Industry. Databases today can range in size of the terabytes or more bytes of data. To address these issues, researchers turned to a new research area called.

Data mining in the field of computer science is ananswered prayer to the demand of this digital age. It is used to unravel hidden information from large volumes of data usually kept in data repositories to help improve management decision making[3].

2. CLASSIFICATION

Classification is the act of looking for a model that describes a class label in such a way that such a model can be used to predict an unknown class label [8]. Thus, classification is usually used to predict an unknown class labels. For instance, a classification model can be used to classify bank loans as either safe or unsafe. Classification applies some methods like decision tree, Bayesian method and rule induction in building its models [9]. Classification process involves two steps. The first step is the learning stage which involves building the models while the second stage involves using the model to predict the class labels.

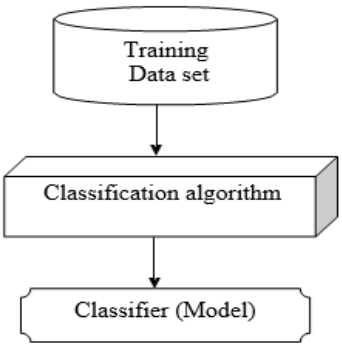

Figure1: Step 1 Construction of amodel

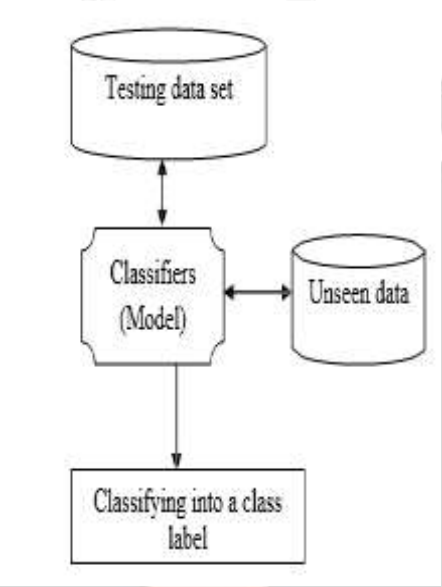

Figure2: Step 2 - Model used for unknown tuple

3. RULE BASED CLASSIFICATION

Rule based classification systems have been widely applied in various expert systems, such as fault diagnosis for aerospace and manufacturing, medical diagnosis, highly interactive or conversational Q&A system, mortgage expert systems etc [13] . Rules are represented in the logic form as IF-THEN statements, e.g. a commonly used rule can be expressed as follows:

IF condition THEN conclusion.

Research Road Map:

In this paper, we apply four well known rule-based classification techniques namely Decision Table, ONE R,PART and ZERO R based on tuple-id propagation techniques . The objective of this paper is to compare the performance of four different rule based classifiers across multiple database relations using tuple-id propagation technique based on the following criteria: number of tuples, number of relations, classification accuracy and runtime. To implement these algorithms, the whole work is divided into three phases:

1. **Class propagation**–Class propagation elementpropagate vital information from the target relation to the background relations based on the foreign key links using tuple id propagation technique. In this way, each resulting relation contains efficient and various information which then enables a propositional learner to efficiently learn the target concept.

2. **Rule Generation**- Classification algorithm builds theclassifier by learning from a training set made up of database tuples and their associated class labels. In this step, the learning model is represented as a set of If-then rules.

3. **Classification and Result analysis**–Next, the model isused for classification and a test data are used to estimate the accuracy of the model and then results are analyzed.

Rule based classification algorithm also known as separate-and-conquer method is an iterative process consisting in first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set [14]. This process is repeated iteratively until there are no examples left to cover. Following are the rule based algorithms considered for study:

(I)Decision Tablealgorithm builds and using a simpledecision table majority classifier as proposed by Kohavi. It summarizes the dataset with a decision table which contains the same number of attributes as the original dataset. Then, a new data item is assigned a category by finding the line in the decision table that matches the non-class values of the data item. Decision Table employs the wrapper method to find a good subset of attributes for inclusion in the table. By eliminating attributes that contribute little or nothing to a model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller and condensed decision table [12].

(II)OneRor "One Rule" is a simple algorithm proposed byHolt. The OneR builds one rule for each attribute in the training data and then selects the rule with the smallest error rate as its one rule. The algorithm is based on ranking all the attributes based on the error rate .To create a rule for an attribute, the most frequent class for each attribute value must be determined. The most frequent class is simply the class that appears most often for that attribute value. A rule is simply a set of attribute values bound to their majority class. OneR selects the rule with the lowest error rate. In the event that two or more rules have the same error rate, the rule is chosen at random. The OneR algorithm creates a single rule for each attribute of training data and then picks up the rule with the least error rate.

(III) PARTis a separate-and-conquer rule learner proposedby Eibe and Witten [14]. The algorithm producing sets of rules called decision lists which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in its each iteration and makes the best leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning.

(IV) ZEROR is the simplest classification method which relies on the target and ignores all predictors. Zero R classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods. It is a simple classification method that works with mode for the prediction of nominal data and mean for the prediction of numeric data. It is usually referred to as majority class method.

4.WEKA

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in java. Weka is free software available under the General Public License [4]. Weka supports several standard data mining tasks, more specifically data preprocessing, clustering, classification, regression, visualization, and feature selection [11]. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes.

Figure3: Weka open source tool

A. Evaluation Metrics

The parameters considered while evaluating the selected classifiers are

(1) Accuracy: This shows the percentage of correctlyclassified instances in each classification model.
(2) TP Rate : Is the statistics that shows correctly classified instances.
(3) FP Rate: Is the report of instances incorrectly labelledas correct instances
(4) Time: Time taken to perform the classification.

B. Datasets

Machine learning algorithms are primarily designed to work with arrays of numbers. This is called tabular or structured data because it is how data looks in a spreadsheet, comprised of rows and columns. Weka prefers to load data in the ARFF format. ARFF is an acronym that stands for Attribute-Relation File Format. It is an extension of the CSV file format where a header is used that provides metadata about the data types in the columns.in this paper we are using Banking data set for execution.

5.RESULTS

1.Number of Classified Instances consisting of number ofcorrectly classified and incorrectly classified instances by four Rule based algorithms using WEKA tool.

Table1: Number of Classified Instances for Banking data set.

Classification	Correct instances	Incorrect instances
Decision Table	4473	791
One R	2175	3089
Part R	4478	786
Zero R	2087	3177

From **Table 1** it is evident that PART algorithm has highest number of correctly classified instances and Zero R algorithm has highest number of incorrectly classified instances.

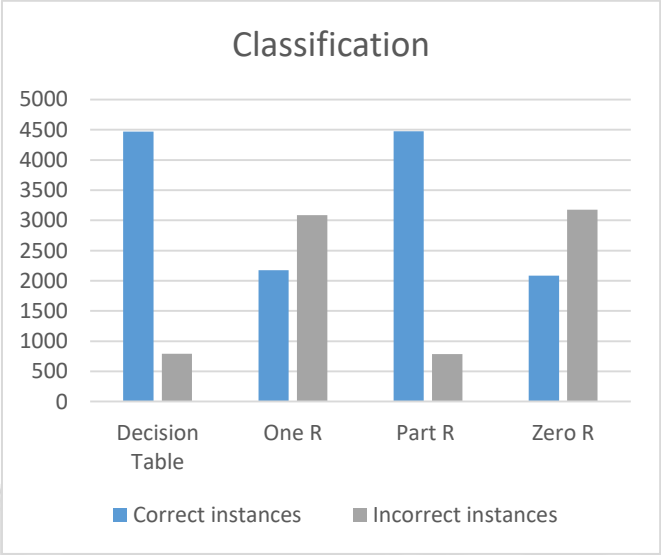

Figure 4: Number of Classified Instance

From **Figure 1** it is evident that PART shows the best performance as compare to other studied algorithms. PART has highest number of correctly classified instances followed by Decision Table and One R. Decision table and Part R has approximately same performance of both categories. Zero R has maximum number of incorrect instances.

2.EXECUTION TIME for Banking Data set fortime measurement four parameters are considered:

Classification	**Execution Time(Seconds)**
Decision Table	0.56
One R	0.02
Part R	0.54
Zero R	00

Table 2 shows the three parameters for evaluating execution time of four rule based algorithms. Zero r has the best performance in execution time. Other hand decision table has worst performance in same parameter with 0.56 second.

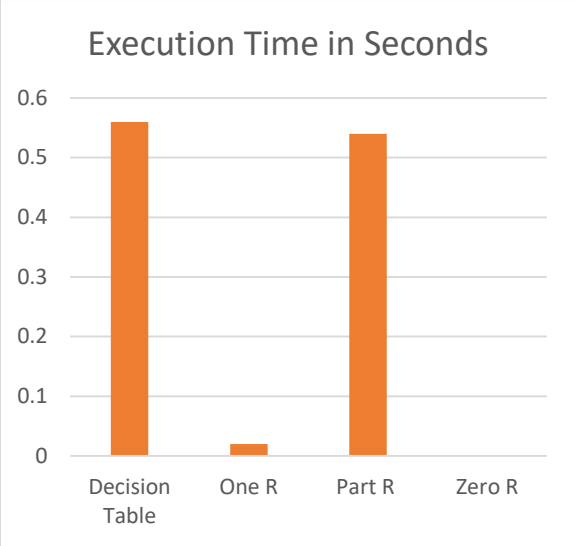

Figure 5 : Execution Time of Rule based classification Algorithms

From **Figure 2** it is evident that Zero R algorithm has the highest performance. PART and decision table both algorithms almost equal in execution time.

3. ERROR RATE for Banking Data set formean square error measurementfour parameters are considered:

Classification	**Mean Absolute Error**
Decision Table	23
One R	29
Part R	00
Zero R	34

Table 3 shows the mean square error rate parameters for the evaluation of four Rule based classification algorithms. PART algorithm had the least value for all four parameters .Zero R algorithm had the highest value for mean square error.

Figure 6: Error Rate of Rule based classification Algorithms

From **Figure3** it is evident that PART algorithm have minimum error rate and have highest performance. Decision table algorithm has second minimum error rate and it also have over all good performance. One R and Zero R average error rate and thus average performance. As seen in the graph Zero R have high error rate and have poor performance as compare to other algorithm under study.

6.CONCLUSION

FourClassification rule based algorithms Decision Table, One R, PART and Zero are introduced and experimentally evaluated using Banking data sets. The rule based classification algorithms are experimentally compared based on number of classified instances, accuracy and error rate using WEKA tool. From the result it is evident that PART is best rule based classification algorithm when compared to the other studied rule based algorithms. Zero R algorithm had over all low performance for all the parameters. The overall position is done based on the number of relations, number of tuples, number of attributes, number of foreign keys, classification accuracy and runtime. Based on the experimental results, PART Classifier appears to be superior to Decision table , One R and Zero R.

7.REFERENCES

1.Silwattananusarn, Tipawan , Tuamsuk, Kulthida (2017) " *Data Mining and Its Applications for Knowledge Management : A Literature Review*" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, pp 13-24.

2.Ramamohan, Y., Vasantharao, K., Chakravarti, C. Kalyana, Ratnam, A.S.K. (2016)" *A Study of Data Mining Tools in Knowledge Discovery Process*" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, pp 191-194.

3.Balaji,S., .Srivatsa, S.K. (2016)" *Decision Tree induction based classification for mining Life Insurance Databases*" IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 2, No.3, pp 699-704.

4.Shi, Yong, Meisner, Jerry (2018) "*An Approach to Selecting Proper Dimensions for Noisy Data"* Int' Conf. Data Mining DMIN pp 172-175.

5.Kanellopoulos,Y. , Antonellis, P., Tjortjis,C., Makris,C., Tsirakis, N. (2015) " *k-attactors a partitional clustering algorithm for numeric data analysis*"Applied Artificial Intelligence, 25:97–115, Taylor & Francis Group, LLC pp 97-115.

6. Pranav Patil (2017)" Application for Data Mining and Web Data Mining Challenges"IJCSMC, Issue. 3, March 2017, pg.39 – 44.

7.Ritu, Shiv , Mohit (2015)" Comparative Analysis of Classification Techniques in Data Mining Using Different Datasets"IJCSMC, Vol. 4, Issue. 12, December 2015,pg 125-134.

8.Velmurugan, T.,Santhanam, T. (2015)" *Clustering Mixed Data Points Using Fuzzy CMeans Clustering Algorithm for Performance Analysis*" International Journal on Computer Science and Engineering Vol. 02, No. 09.

9.K Prasanna (2017) "A Study of Classification Techniques of Data Mining Techniques in health related research" IJIRCCE Vol5 july 2017.

10.Li, Xiangyang, Ye, Nong (2014)"*A Supervised Clustering and Classification Algorithm for Mining Data With Mixed Variables*" IEEE Tranactionon systems, man and Cybernetics-part systems and humans, VOL. 36, NO. 2pp 396-406.

11.Lin, Zetao, Ge, Yaozheng, Tao, Guoliang (2016) "*Algorithm for Clustering Analysis of ECG Data"* Proceedings of the IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, pp-3857-3860.

12.Biao , Sunil (2014)" A Rule-Based Classification Algorithm for Uncertain Data" IEEE International Conference on Data Engineering may 2014.

13. Shi Na, Liu Xumin, Guan yong(2017) , "Research on k-means Clustering Algorithm" 978-0-7695-4020-7/10© IEEE DOI 10.1109/IITSI.74.

14. M Thangaraj (2013)" Performance Study on Rule-based Classification Techniques across Multiple Database Relations" International Journal of Applied Information Systems (IJAIS), Volume 5– No.4, March 2013.