Mining the Financial Multi-Relationship with Accurate Models

C.Durka[1], Kerana Hanirex.D[2]

[1] *PG Student, Computer science department,*
Bharath University
Chennai
[2] *Assistant Professor, Computer science department,*
Bharath University
Chennai

[1]durkamtech@gmail.com
[2] keranarobinson@gmail.com

***Abstract*- In order to overcome the difficulty of cognitive and mining for large-scale multi-relationship system, a classification method or algorithm is proposed. The development of data-mining applications such as classification or clustering has shown the need for machine learning algorithms to be applied to large scale data. In this paper we present the comparison of different classification techniques using Waikato Environment for Knowledge Analysis or in short, WEKA. WEKA is an open source software which consists of a collection of machine learning algorithms for data mining tasks. The algorithm or methods tested are Bayes Net, Logistic, Decision Table, Random Tree, JRip, Decision Stump and J48. Finally, the experiments in financial multi-relationship dataset demonstrate the efficiency of the methods.**

Keywords: **Classification, Bayes Net, Logistic, Decision Table, Random Tree, JRip, and J48.**

I. INTRODUCTION

As a new knowledge representation and system modelling method, FCM simulates the dynamic behaviour of system through the causal link between nodes of the entire map and plays a significant role in qualitative reasoning compared with other models. FCM model is inadequate in dealing with multi-relationship system, because there are a large collection of objects or concepts and intricate links between them in these systems so as to form tens of thousands of nodes and complex links between the nodes. The traditional FCMs are difficult to be cognitive and processed for such large scale and complex internal contacts. Thus, classification method are proposed to simulate complex multi-relationship system.

The aim of our work is to investigate the performance of different classification methods using WEKA for financial dataset. WEKA is a collection of open source of many data mining and machine learning algorithms, including: pre-processing on data, classification, clustering, association rule extraction.

A major problem in financial data analysis is in attaining the correct accuracy of certain important information. For the function of multi-relationship data set, normally, many tests generally involve the clustering or classification of large scale data. However, on the other hand, too many tests could complicate the multi-relationship data set and lead to the difficulty in obtaining the end results, particularly in the case where many tests are performed. This kind of difficulty could be resolved with the aid of machine learning which could be used directly to obtain the end result with the aid of several artificial intelligent algorithms which perform the role as classifiers.

Machine learning covers such a broad range of processes that it is difficult to define precisely. A dictionary definition includes phrases such as to gain knowledge or understanding of or skill by studying the instruction or experience and modification of a behavioral tendency. The extraction of important information from a large pile of data and its correlations is often the advantage of using machine learning. New knowledge about tasks is constantly being discovered by humans and vocabulary changes. There is a constant stream of

new events in the world and continuing redesign of Artificial Intelligent systems to conform to new knowledge is impractical but machine learning methods might be able to track much of it.

There is a substantial amount of research with machine learning algorithm such as Bayes Network, Logistic, Decision Table, Random Tree, JRip, and J48.

II. METHODS

A. Bayes Network Classifier

A sequence Bayesian networks are a powerful probabilistic representation, and their use for classification has received considerable attention. This classifier learns from training data the conditional probability of each attribute Ai given the class label C . Classification is then done by applying Bayes rule to compute the probability of C given the particular instances of A1…..An and then predicting the class with the highest posterior probability. The goal of classification is to correctly predict the value of a designated discrete class variable given a vector of predictors or attributes. In particular, the naive Bayes classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent.

B. Logistic Regression

The term regression is defined as an analyzing or measuring the relation between a dependent variable and one or more independent variable. Regression can be defined by two types: Linear regression and logistic regression. Logistic regression is a generalization of linear regression. It is basically used for estimating binary or multi-class dependent variables and the response variable is discrete, it cannot be modeled directly by linear regression i.e. discrete variable changed into continuous value. Logistic regression basically is used to classify the low dimensional data having non-linear boundaries. It also provides the difference in the percentage of dependent variable and provides the rank of individual variable according to its importance. So, the main motto of Logistic regression is to determine the result of each variable correctly. Logistic regression is also known as logistic model/ logit model that provide categorical variable for target variable with status.

C. Decision Table

Decision Tree and Pruning A decision tree partitions the input space of a data set into mutually exclusive regions, each of which is assigned a label, a value or an action to characterize its data points. The decision tree mechanism is transparent and we can follow a tree structure easily to see how the decision is made. A decision tree is a tree structure consisting of internal and external nodes connected by branches. An internal node is a decision making unit that evaluates a decision function to determine which child node to visit next. The external node, on the other hand, has no child nodes and is associated with a label or value that characterizes the given data that leads to its being visited. However, many decision tree construction algorithms involve a two - step process. First, a very large decision tree is grown. Then, to reduce large size and overfiting the data, in the second step, the given tree is pruned. The pruned decision tree that is used for classification purposes is called the classification tree.

D. Random Tree

The Random Forest method is based on bagging (bootstrap aggregation, see definition of bagging) models built using the Random Tree method, in which classification trees are grown on a random subset of descriptors. The Random Tree method can be viewed as an implementation of the Random Subspace method for the case of classification trees. Combining two ensemble learning approaches, bagging and random space method, makes the Random Forest method very effective approach to build highly predictive classification models.

E. JRip

JRip (RIPPER) is one of the basic and most popular algorithms. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced error JRip (RIPPER) proceeds by treating all the examples of a particular judgment in the training data as a class, and finding a set of rules that cover all the members of that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered.

F. J48

The C4.5 technique is one of the decision tree families that can produce both decision tree and rule-sets; and construct a tree. Besides that, C4.5 models are easy to understand as the rules that are derived from the technique have a very straightforward interpretation. J48 classifier is among the most popular and powerful decision tree classifiers. C5.0 and J48 are the improved versions of C4.5

algorithms. WEKA toolkit package has its own version known as J48. J48 is an optimized implementation of C4.5

III.EXPERIMENTAL RESULTS AND ANALYSIS

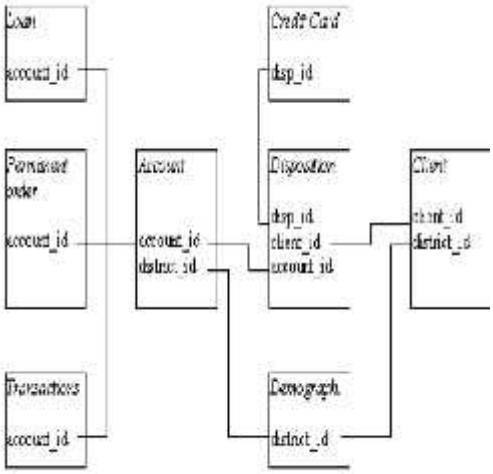
In this section, we test the implementation accuracy of algorithm and compare with multi-relationship financial data set. Weka tool is used to select the attributes from the dataset.

A. Dataset

a) Dataset Information

The experiment is established on financial data set of PKDD CUP 1999, which is adjusted at some parts: to delete 90% tuples of table Transaction because of too much tuples about 1056320, to delete some positive tuples of table Loan at random to make it more balanced. As a result, the whole database has 8 tables, 15 feature attributes and 75982 relationship tuples, where Loan(L) is target table and "status" is class or target attribute[1].

Figure 1.Financial multi-relationship



3.1.2 Attribute Information:

The data about the clients and their accounts consist of following relations:[1]

1.relation account (4500 objects in the file ACCOUNT.ASC) - each record describes static characteristics of an account,

2.relation client (5369 objects in the file CLIENT.ASC) - each record describes characteristics of a client,

3.relation disposition (5369 objects in the file DISP.ASC) - each record relates together a client with an account i.e. this relation describes the rights of clients to operate accounts,

4.relation permanent order (6471 objects in the file ORDER.ASC) - each record describes characteristics of a payment order

Table 1. Background of Financial multi-relationship dataset

Table	Attributes
Loan	includes: amount,
Account	duration,
Order	payments
Trans	includes:
Disp	account_id,
District	district_id
Client	includes: amount
	includes:
	account_id,
	amount
	includes:
	account_id,
	client_id
	includes: A3
	includes:
	client_id,
	account_id

5.relation transaction (1056320 objects in the file TRANS.ASC) - each record describes one transaction on an account,

6.relation loan (682 objects in the file LOAN.ASC) - each record describes a loan granted for a given account,

7.relation credit card (892 objects in the file CARD.ASC) - each record describes a credit card issued to an account,

8.relation demographic data (77 objects in the file DISTRICT.ASC) - each record describes demographic characteristics of a district.

IV. RESULTS

To gauge and investigate the performance on the selected classification methods or algorithms namely Bayes Network classifier, Logistic Regression, Decision Table, Random Tree, JRip and J48. we use the same experiment procedure as suggested by WEKA. The 75% data is used for training and the remaining is for testing purposes. 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 precision and recall will be in numeric value only. We 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 Tables 2 and3 below. Table 2 mainly summarizes the result based on accuracy and time taken for each simulation. Meanwhile, Table 23shows the result based on execution time of each simulation.

Table 3. Comparison of accuracy

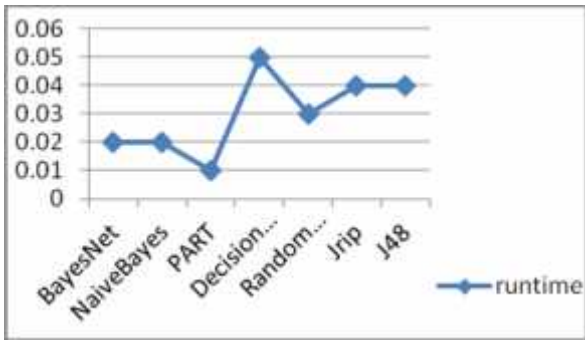
	After normalize	
Method	**correctly classified**	**IN correctly classified**
bayes.BayesNet	95.98%	4.02%
bayes.NaiveBayes	95.18%	4.82%
functions.logistic	96.12%	3.88%
rules.PART	98.78%	1.22%
rules.DecisionTable	96%	4%
rules.Jrip	96.24%	3.76%
rules.ZeroR	86.36%	13.64%
trees.DecisionStump	94.42%	5.52%
trees.J48	97.18%	2.82%

Based on the above and Table 1 and 2, we can clearly see that the highest accuracy is 98.78% and the lowest is 86.36%. The other algorithm yields an average accuracy of around 96%. In fact, the highest accuracy belongs to the PART, followed by J48 with a percentage of 97.18% and subsequently decision tree with pruning and single conjunctive rule learner. The total time required to build the model is also a crucial parameter in comparing the classification algorithm.

Algorithm	**Runtime (s)**
Bayes Net	0.02
Naïve Bayes	0.02
Logistic	0.02
Decision Table	0.01
JRip	0
J48	0
J48	0.04

Table 4.Comparison of execution time

Figure 2: The runtime in different classifier algorithm to financial multi-relationship



CONCLUSIONS

As a conclusion, we have met our objective which is to evaluate and investigate five selected classification algorithms based on Weka. The best algorithm based on the financial data set is PART with an accuracy of 96.24% and the total time taken to build the model is at 0.02 seconds. Logistic has the lowest average error compared to others. These results suggest that among the machine learning algorithm tested, Bayes network Logistic classifier has the potential to significantly improve the conventional classification methods for use in financial multi-relationship data set.

REFERENCES

[1] Comparison of Different Classification Techniques Using WEKA for Breast Cancer Mohd Fauzi bin Othman, Thomas Moh Shan Yau Control and Instrumentation Department, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Skudai, Malaysia.

[2] International Conference & Workshop on Recent Trends in Technology, (TCET) 2012 Proceedings published in International Journal of Computer Applications® (IJCA) 27 Analysis of Machine Learning Algorithms using WEKA. Aaditya Desai Ph.D. Scholar, NMIMS University. TCET, Mumbai and Dr. Sunil Rai Ph.D. Guide, NMIMS University.

[3] WEKA at http://www.cs.waikato.ac.nz/~ml/weka.

[4] C. H. Lee, D. G. Shin (1999), "A multi strategy approach to classification learning in database", Data Knowledge Engg. 31, 67-93.
[5] WEKA Tutorial http://www.cs.utexas.edu/users/ml/tutorials/Weka-tut/