Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 6 (2017) pp. 1731-1743

http://www.ripublication.com

Comparative Study of J48, Naive Bayes and One-R Classification Technique for Credit Card Fraud Detection using WEKA

Farhad Alam[1] and Sanjay Pachauri[2]

[1]*Research Scholar, Himalayan University, Arunachal Pradesh, India.*

[2]*Assistant Professor, IMS Unison University, Dehradun, Uttarakhand, India.*

Abstract

Breakthroughs in information and communication technology establish payment-collection technologies like Debit and Credit card systems at the level where their rapid penetration in commercial market have led to an ever-larger share of the world payment system and leading to a higher rate of stolen account numbers and subsequent losses by banks. Improved fraud detection thus has become essential to maintain the viability of the payment system, especially for e-commerce. This tremendous growth in databases, fraud detection and payment security systems has spawned a pressing need for new techniques and tools that can intelligently and automatically transform data into useful information and knowledge. In recent years, data-mining (DM) has become one of the most valuable tools for extracting and manipulating data and for establishing patterns in order to produce useful information for decision-making. This paper introduces three important data mining techniques J48, Naive Bayes and One-R classifier algorithm using weka work bench to achieve classification response for frand detection dataset. Beside the basic description, paper compares these classifiers over different parameters and helps the e-commerce companies to select optimal classification algorithm.

Keywords: Data Mining Tools; J48; Navie Bayes; One-R; Classification Methods.

INTRODUCTION

Financial fraud is becoming an increasingly serious problem in online shopping and e-businesses. With an increasing number of transactions older data handling technology can no longer control and secure all of them. Credit cards are one of the most famous targets of fraud. A number of algorithms, process and preventive mechanism help banking systems and credit card companies to stop credit card fraud and reduce financial risks. But to perform highly automated and sophisticated screenings of incoming transactions and flagging suspicious transactions recent studies proposed Data Mining techniques as an indispensible technique. Data mining is the process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify interesting patterns in databases and subsequently gain knowledge that can then be used in decision making. (Bose and Mahapatra, 2001, Turban et al, 2005, Kantardzic, 2002)

Weka is open source software for data mining under the GNU General public license. This system is developed at the University of Waikato in New Zealand. "Weka" stands for the Waikato Environment for knowledge analysis. Weka provides implementation of state-of-the-art data mining and machine learning algorithm. User can perform association, filtering, classification, clustering, visualization, regression etc. by using weka tool. This paper presents discussion about Navie Bayes, J48 and One R classifier. Naive Bayesian (NB) algorithm is simple and very effective in many real world data sets because it can give better predictive accuracy than well known methods like J48 and Back Propagation algorithms (Domingos P and Pazzani M, 1996; Elkan, 2001). J48 can help not only to make accurate predictions from the data but also to explain the patterns in it. It deals with the problems of the numeric attributes, missing values, pruning, estimating error rates, complexity of decision tree induction, and generating rules from trees (Witten and Frank, 1999). OneR, short for "One Rule", is a simple, yet accurate, classification algorithm that generates one rule for each predictor in the data, and then selects the rule with the smallest total error as its "one rule" (Berry M and Linoff, 2000). This study implements and compares three machine learning algorithm Bayes Network, J48 Decision tree and OneR Algorithm for credit card fraud detection.

WEKA: DATA MINING TOOL

Weka is a widely accepted machine learning toolkit in the domain of computer vision, image interpretation and data mining (Frank et al., 2004) implemented in Java. Graphical User Interface to go on different Weka applications is depicted in the figure. The weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality.

Figure 1. WEKA GUI Chooser.

The workbench includes methods for all the standard Data Mining problems: regression, classification, clustering, association rule mining, and attribute selection. All algorithms and methods take their input in the form of a single relational table, which can be read from a file or generated by a database query. In this study, we made use of the Naive Bayes, J48 and One R classifier algorithm over a credit card dataset and compare the efficiency of these three classifiers over different weka attributes.

Data Used

The current study uses German Credit fraud data proposed by Professor Dr. Hans Hofmann from the Institute for Statistics Hamburg having 1000 instances with categorical/symbolic attributes. This data file has been edited and several indicator variables added to make it suitable for algorithms which cannot cope with categorical variables. Several attributes that are ordered categorical have been coded as integer.

Table 1. Attribute description for German

Attribute 1	**(qualitative)**
	Status of existing checking account A11 : ... < 0 DM A12 : 0 <= ... < 200 DM A13 : ... >= 200 DM / salary assignments for at least 1 year A14 : no checking account
Attribute 2	**(numerical)**
	Duration in month
Attribute 3	**(qualitative)**
	Credit history A30 : no credits taken/ all credits paid back duly A31 : all credits at this bank paid back duly A32 : existing credits paid back duly till now A33 : delay in paying off in the past

	A34 : critical account/ other credits existing (not at this bank)
Attribute 4	**(qualitative)**
	Purpose A40 : car (new) A41 : car (used) A42 : furniture/equipment A43 : radio/television A44 : domestic appliances A45 : repairs A46 : education A47 : (vacation - does not exist?) A48 : retraining A49 : business A410: others
Attribute 5	**(numerical)**
	Credit amount
Attribute 6	**(qualitative)**
	Savings account/bonds A61 : ... < 100 DM A62 : 100 <= ... < 500 DM A63 : 500 <= ... < 1000 DM A64 : .. >= 1000 DM A65 : unknown/ no savings account
Attribute 7	**(qualitative)**
	Present employment since A71 : unemployed A72 : ... < 1 year A73 : 1 <= ... < 4 years A74 : 4 <= ... < 7 years A75 : .. >= 7 years
Attribute 8	**(numerical)**
	Installment rate in percentage of disposable income
Attribute 9	**(qualitative)**
	Personal status and sex A91 : male : divorced/separated A92 : female :divorced/separated/married A93 : male : single A94 : male : married/widowed A95 : female : single
Attribute 10	**(qualitative)**
	Other debtors / guarantors A101 : none A102 : co-applicant A103 : guarantor
Attribute 11	**(numerical)**
	Present residence since
Attribute 12	**(qualitative)**
	Property A121 : real estate A122 : if not A121 : building society savings agreement/life insurance A123 : if not A121/A122 : car or other, not in attribute 6 A124 : unknown / no property
Attribute 13	**(numerical)**
	cc_age in months
Attribute 14	**(qualitative)**
	Other installment plans A141 : bank A142 : stores A143 : none
Attribute 15	**(qualitative)**

	Housing A151 : rent A152 : own A153 : for free
Attribute 16	**(numerical)**
	Number of existing credits at this bank
Attribute 17	**(qualitative)**
	Job A171 : unemployed/ unskilled - non-resident A172 : unskilled - resident A173 : skilled employee / official A174 : management/ self-employed/ highly qualified employee/ officer
Attribute 18	**(numerical)**
	Number of people being liable to provide maintenance for
Attribute 19	**(qualitative)**
	Telephone A191 : none A192 : yes, registered under the customer's name
Attribute 20	**(qualitative)**
	foreign worker A201 : yes A202 : no

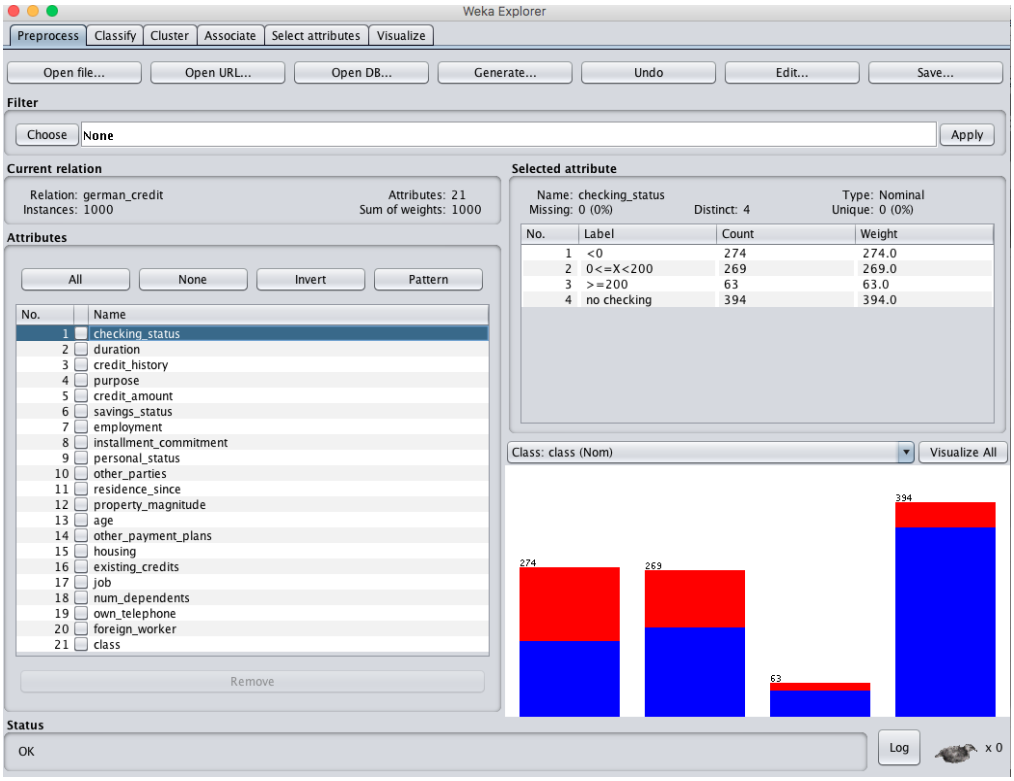

Figure 2. WEKA Explorer with credit card dataset.

Cost Matrix

This dataset requires use of a cost matrix (see below)

	1	2
1	0	1
2	5	0

(1 = Good, 2 = Bad)

The rows represent the actual classification and the columns the predicted classification. It is worse to class a customer as good when they are bad (5), than it is to class a customer as bad when they are good (1).

CLASSIFICATION METHODS

Classification is a classic data mining technique based on the concepts of machine learning. General applications of classification is used it as a tool to categorizes item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, once the software is made that can learn how to classify the data items into groups. Classification—A Two-Step Process

1. Training: describing a set of predetermined classes. Each sample is assumed to belong to a predefined class, as determined by the class label attribute. The set of tuples used for model construction: training set. The model is represented as classification rules, decision trees, or mathematical formula.

2. Classification: for classifying future or unknown objects. Estimate accuracy of the model. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set, otherwise over-fitting will occur.

NAIVE BAYES ALGORITHM

The Naive Bayes algorithm is an intuitive method that uses the conditional probabilities of each attribute belonging to each class to make a prediction. It uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Parameter estimation for naive Bayes models uses the method of maximum likelihood. In spite over-simplified assumptions, it often performs better in many complex real world situations. One of the major advantages of Naïve Bayes theorem is that it requires a small amount of training data to estimate the parameters.

Algorithm:

INPUT

- Set of tuples = D
- Each Tuple is an 'n' dimensional attribute vector
- X : (x1,x2,x3,…. xn)

Let there be 'm' Classes: C1, C2, C3…Cm

Naïve Bayes classifier predicts X belongs to Class Ci iff

- P (Ci/X) > P(Cj/X) for 1<= j <= m , j <> i

  Maximum Posteriori Hypothesis
- P(Ci/X) = P(X/Ci) P(Ci) / P(X)
- Maximize P(X/Ci) P(Ci) as P(X) is constant

With many attributes, it is computationally expensive to evaluate P(X/Ci).

Naïve Assumption of "class conditional independence"

$P(X/Ci) = \prod_{k=l}^{n} P(x_k/Ci)$

P(X/Ci) = P(x1/Ci) * P(x2/Ci) *…* P(xn/ Ci)

DECISION TREE J48 ALGORITHM

J48 is an open source Java implementation of simple C4.5 decision tree algorithm. J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. Being a decision tree classifier J48 uses a predictive machine-learning model which calculates the resultant value of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes; the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final

value (classification) of the dependent variable.

Algorithm:

INPUT

- Training Dataset = D

OUTPUT

- Decision Tree = T

DTBUILD (*D)

- T=φ;
- T= Create root node and label with splitting attribute;
- T= Add arc to root node for each split predicate and label;
- For each arc D= Database created by applying splitting predicate to D;
  - If stopping point reached for this path, then T'= create leaf node and label with appropriate class;
  - Else T'= DTBUILD(D); T= add T' to arc;

1R OR ONE R CLASSIFIER ALGORITHM

The 1R or One R classifier for machine learning classification problems is one of the very simple and most effective classifier algorithms. In comparison to its Zero R Classifier, 1 R does not rely on the frequency of target but induces classification rules based on the value of a single predictor. In order to develop rule set for a predictor, a frequency table corresponds to each predictor against the target is created. It is evident that 1 R produces rules only little less accurate than cutting edge classifiers while producing rules that are simple for humans to interpret.

Algorithm:

INPUT

- Set of tuples = D_n
- Each Tuple is an 'n' dimensional vector
- Set of Attribute Values = A_j
- Attribute Value is an 'j' dimensional attribute vector

1RBUILD (*D, *A)

For each attribute,

For each value of the attribute,

- Decide Rule: count how often each class appears
- Find the most frequent class
- Assign that class to this attribute-value

Calculate the error rate of the rules

Choose the rules with the smallest error rate

RESULT AND DISCUSSION

In this study we are taking vulnerability dataset of 1000 customers holding credit card and making comparison using three classifiers Naïve Bayes, j48 and oneR. This dataset classifies people described by a set of attributes as good or bad credit risks. It Comes in two formats (one all numeric) with a cost matrix. Algorithms are applied to the dataset the confusion matrix is generated. Experiments are performed on Weka with 10 fold cross validation. Ten fold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier. The first step is to find the number of instances of Credit card dataset using Naïve Bayes , j48 and oneR classification algorithm. In the next step of the experiment we will calculate the classification accuracy and cost analysis. Confusion matrix describes the information about actual and predicted classification, computed in the last. Standard terms defined for confusion matrix are.

1) True positive –if the outcome of prediction is p and the actual value is also p than it is called true positive (TP).
2) False positive-if actual value is n than it is false positive (FP)
3) Precision – precision is measure of exactness and quality-Precision = tp/(tp + fp)
4) Recall- measure of completeness and quantity -Recall = tp / (tp + fn)

J48 is a module for generating a pruned or unpruned C4.5 decision tree. When applying J48 on Credit Card' dataset results are discussed in the figure and const analysis id depicted in the figure3. Naïve Bayes algorithm is applied on Credit Card dataset, we got the result shown as below on figure4. oneR is implemented in the last and results are depicted using figure5. A detailed comparison statistics for these three classification methods are discussed in the table and discussed in the Graph1.

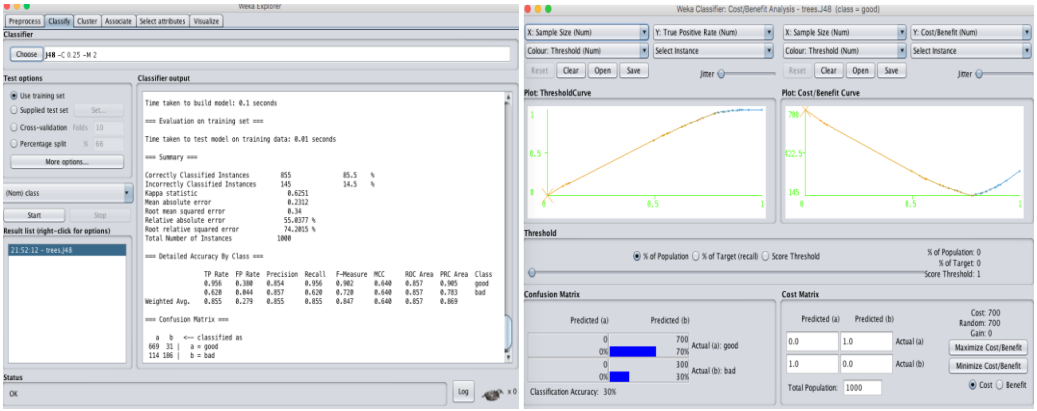

Figure 3. Accuracy and Cost analysis of j48

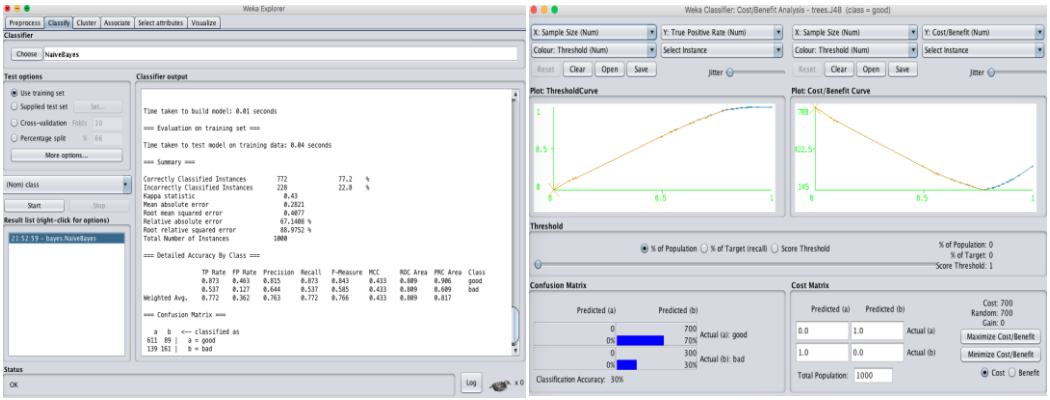

Figure 4. Accuracy and Cost analysis of Naïve Bayes

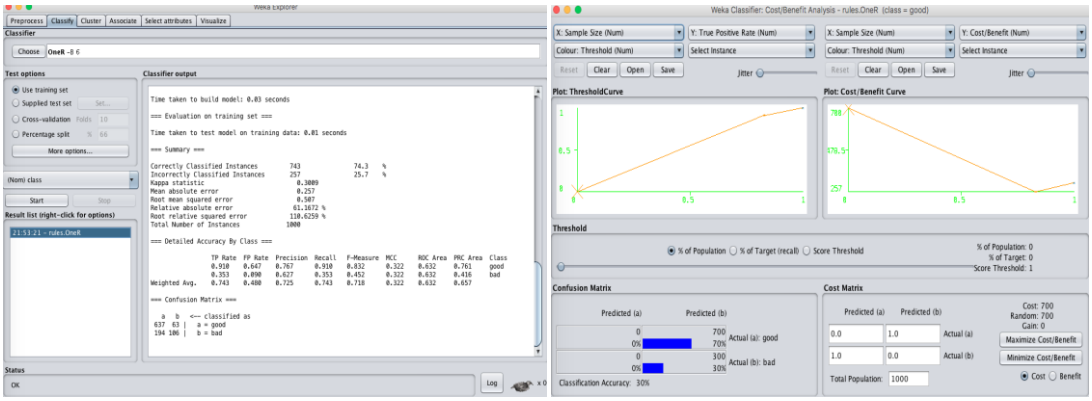

Figure 5. Accuracy and Cost analysis of oneR

Table 2. Performance parameters and their values for selected classification algorithms.

S. No.	Parameters	J48	Navie Bayes	One R
1.	Time to Build Model (in Sec)	0.10	0.01	0.03
2.	Correctly Classified Instances	855	772	743
3.	Incorrectly Classified Instances	145	228	257
4.	Kappa statistic	0.6251	0.4300	0.3009
5.	Mean absolute error	0.2312	0.2821	0.2570
6.	Root mean squared error	0.3400	0.4077	0.5070
7.	Relative absolute error	55.04%	67.14%	61.17%
8.	Root Relative squared error	074.20%	088.98%	110.63%
9.	Prediction Accuracy	85.50%	77.20%	74.30%
10.	Total number of Instances	1000	1000	1000

Graph 1. Comparison different parameters for selected classification algorithms.

CONCLUSION

There are so many benchmarks comparing the performance and accuracy of different classification algorithms but there are still very few experiments carried out on Credit card risk assessment and fraud detection datasets. In this work, we focuses on various classification techniques most frequently used in data mining and compare the performance and the interpretation level of confidence on different classification techniques applied on Credit card datasets in order to determine which one is more suitable.

From the result we see that time to build the model is less when using j48 and correctly classified instances are more and prediction accuracy is also greater in j48 than the other two. Hence it is concluded that j48 performed better on credit card dataset. Beside the credit card dataset this can also be concluded that the different classification algorithms are designed to perform better for certain types of dataset.

REFERENCES

[1]. Chai, K.; H. T. Hn, H. L. Chieu; "Bayesian Online Classifiers for Text Classification and Filtering", Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval, August 2002, pp 97-104

[2]. DATA MINING Concepts and Techniques,Jiawei Han, Micheline Kamber Morgan Kaufman Publishers, 2003

[3]. Domingos P and Pazzani M. "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier", in Proceedings of the 13th Conference on Machine Learning, Bari, Italy, pp105-112, 1996.

[4]. Elkan C. Magical Thinking in Data Mining: Lessons From CoIL Challenge 2000, Department of Computer Science and Engineering, University of California, San Diego, USA, 2001.

[5]. Witten I and Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java, Morgan Kauffman Publishers, California, USA, 1999.

[6]. Berry M and Linoff G. Mastering Data Mining: The Art and Science of Customer Relationship Management, John Wiley and Sons, New York, USA, 2000.

[7]. Bose, I., & Mahapatra, R. K. (2001). Business data mining - A machine learning perspective. Information and Management, 39(3), 211-225. DOI: 10.1016/S0378-7206(01)00091-X

[8]. Turban, E., et al. Decision Support and Intelligent Systems. Upper Saddle River, NJ: Prentice Hall, 2005.

[9]. Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

[10]. Frank, Eibe, Mark Hall, Len Trigg, Geoffrey Holmes, and Ian H. Witten. "Data mining in bioinformatics using Weka." Bioinformatics 20, no. 15 (2004): 2479-2481.