IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 2, Ver. I (Mar.-Apr. 2017), PP 39-43
www.iosrjournals.org

Credit Card Fraud Detection Using Split Criteria in Classification

Manisha[1], Neena Madan[2]
[1,2]*Department of Computer Science & Engg. GNDU Regional Campus , Jalandhar, India*

Abstract: *With an increase usage of credit cards for online purchases as well as regular purchases causes a credit card fraud. In the mode of electronic payment system, fraud transactions are rising on the regular basis. The Modern techniques based on the Data Mining, Genetic Programming etc. has used in detecting fraudulent transactions. Currently; data mining is a popular way to battle frauds because of its effectiveness. Data mining is a well defined procedure that takes data as input and produces output in the forms of models or patterns. In this paper, we focus on Decision Tree technique; basically it provides a system which is supposed to classify a current transaction into fraud or non-fraud using split criteria.*

Keywords: *KDD, SVM, Index Ratio, Gain Ratio, Gini Index .*

I. Introduction

Data mining is an Extraction of hidden, predictive information from large databases. It is also called Knowledge Discovery from Databases (KDD). It performs an Identification and evaluation of hidden patterns in database. It is powerful technology with great potential to help organizations to locate and generate information from their data warehouses. Data mining tools predict future trends and behaviors [1]. Data mining can be conducted on any kind of data as long as the data are meaningful for a target application, such as database data, data warehouse data, transactional data, and advanced data types [2].

Data Mining Techniques

There are four basic approaches of data mining.

1. Classification

It is the organization of data in given classes. Also known as supervised classification, classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. Common techniques for classification are decision tree, neural networks, SVM etc.

2. Clustering

It is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar amongst them and dissimilar compared to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details, but achieve simplification. It represents many data objects by few clusters, and hence, it models data by its clusters [4]. Some algorithms are Model Based algo, Density Based algo etc.

3. Regression or Prediction

Regression technique can be used for prediction. Regression can also be used to describe the relationship between two or more independent and dependent variables. In data mining independent variables are those attributes whose values are already known and response variables are those which we want to predict. For example, sale volumes, stock prices, and product failure rates are all quite difficult to predict because they involve complex interactions of multiple prediction variables.

4. Association Rule

Association rule mining, one of the most important and well researched techniques of data mining. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Various association mining techniques and algorithms will be briefly introduced and compared later. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database [5].Some Association rule mining algorithms are Apriori**,** Apriori hybrid.

II. Literature Survey

Anika Nahar, Sharmishta Roy, and Syedashabnam Hasan, January 2016, there is an survey on different techniques used in credit card fraud detection such as Neural Network,Bayesian Network, Decision Trees, Blast-Saha Algorithm, Fuzzy System with Neural Network, Fuzzy Darwinian System, Hidden Markov Model, Support Vector Machines, Meta Learning, and Genetic Algorithm is demonstrated.

Snehal Patil, Harshada Somavanshi, Jyoti Gaikwad, Amruta Deshmane and Rinku Badgujar in April 2015 their research is totally concerned with credit card application fraud detection by performing the process of asking security queries to the persons intricate with the transactions and as well as by eliminating real time data faults. Ashish Thakur, Bushra Shaikh, Vinita Jain, and A.M Magar in April 2015 recommended system which is an application of Hidden Markov Model in Anomaly or fraud detection. The diverse steps in credit card transaction handling are represented as the essential method of an HMM. The system implemented takes all the user information and deals with the data carefully to detect online frauds. It has also been described how they can detect whether an inbound transaction is fraudulent or not. Additional security features like MAC address detection and also shipping address verification are provided for enhanced security and better detection of fraud transaction.

Sahil Hak, Suraj Singh, and Varun Purohit, April 2015, in this paper, they present a credit card fraud detection model which takes into account present as well as the past behaviour. The detection model comprises of card validation via Luhn"s algorithm, two initial probability assignments based on address mismatch and spending pattern, an advanced combination heuristic, spending-pattern database and Bayes theorem. Advanced combination heuristic is an improvement that eliminates the conflict of existing Dempster-Shafer theory.

Jyoti R.Gaikwad, Amruta B. Deshmane, Harshada V. Somavanshi, Snehal V. Patil and Rinku A. Badgujar in November 2014, presents data mining technology, classification models based on ID3 decision trees and visual cryptography are applied on credit card fraud detection problem. Thus by the implementation of this approach in fraud detection systems, financial losses due to fraudulent transactions can be decreased more.

Vaibhav P.Vasani and Rajendra D. Gawali in March 2014 presents Classification of the data collected from students of polytechnic institute has been discussed. The students are then classified into different categories like brilliant, average, weak using decision tree and naïve Bayesian algorithms. The processing is done using WEKA data mining tool. This paper also compares results of classification with respect to different performance parameters.

Renu and Suman in Feb 2014 present a survey of current techniques used in credit card fraud detection and telecommunication fraud. That paper provides a comprehensive review of different techniques to detect fraud. Yusuf Sachin and Serol Bulkan and Ekrem Duman in 2013 shows thatcost-sensitive decision tree algorithm outperforms the existing well-known methods on the given problem set with respect to the well-known performance metrics such as accuracy and true positive rate, but also a newly defined cost-sensitive metric specific to credit card fraud detection domain. Accordingly, financial losses due to fraudulent transactions can be decreased more by the implementation of this approach in fraud detection systems.

Dr R.Dhanapal and Gayathiri.P in Sept 2012 Presents a Credit Card Fraud Detection using effective algorithm for Decision Tree Learning. In their paper estimating the best split of Purity Measures of Gini, Entropy and Information Gain Ratio to test the best classifier Attribute. In this Technique we simply find out the Fraudulent Customer/Merchant through Tracing Fake Mail and IP Address. Customer /merchant are suspicious if the mail is fake they are traced all information about the owner/sender through IP Address.

III. Credit Card Fraud

Fraud can be defined as the undesired activities taking place in an operational system. The credit card is a small plastic card issued to users as a system of payment. It allows its cardholder to buy goods and services based on the cardholder's promise to pay for these goods and services. Credit card security relies on the physical security of the plastic card as well as the privacy of the credit card number. Globalization and increased use of the internet for online shopping has resulted in a considerable proliferation of credit card transactions throughout the world. Thus a rapid growth in the number of credit card transactions has led to a substantial rise in fraudulent activities. Credit card fraud is a wide-ranging term for theft and fraud committed using a credit card as a fraudulent source of funds in a given transaction. Credit card fraudsters employ a large number of techniques to commit fraud. To combat the credit card fraud effectively, it is important to first understand the mechanisms of identifying a credit card fraud. Over the years credit card fraud has stabilized much due to various credit card fraud detection and prevention mechanisms.

IV. Implementation

A. Decision Tree

A decision tree is flowchart like structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The topmost node in the tree is the root node. The construction of decision tree classifiers does not require any domain knowledge discovery. Decision trees can handle multi-dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. Decision tree induction algorithms have been used for classification in many application areas such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology.

B. Dataset

ID	CustName	Mail Type	IP Address	Trans Amt	Class: Trans Type
1	E	Cust	117.204.23.162	Low	Fraud
2	E	Cust	117.204.23.162	High	Fraud
3	R	Cust	117.204.23.162	Low	Fraud
4	V	Cust	117.204.23.162	Low	Legal
5	V	Mrchnt	61.16.173.243	Low	Legal
6	V	Mrchnt	117.204.23.162	Low	Legal
7	R	Mrchnt	61.16.173.243	High	Fraud
8	E	Mrchnt	117.204.23.162	Low	Legal
9	E	Mrchnt	61.16.173.243	Low	Fraud
10	V	Cust	61.16.173.243	Low	Legal
11	E	Cust	117.204.23.162	High	Legal
12	R	Cust	117.204.23.162	High	Legal

Table: dataset of credit card fraud

In this paper we use the dataset of [14].

C. Attribute Selection Measures

The attribute selection measure provides a ranking for each attribute describing the given training tuples. The attribute having the best score for the measures chosen as the splitting attribute for the given tuples.

There are three attribute selection measures- Information Gain, Gain Ratio, and Gini Index [2].

1. Information Gain

ID3 uses information gain as its attribute selection measure. This measure is used to select among the candidate attribute at each step while growing the tree.

$$Info_{(D)} = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j).$$

The term $\frac{|D_j|}{|D|}$ act as the weight of the jth partition. $Info_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A. The smaller the expected information required, the greater the purity of the partitions.

Information gain is defined as the difference between the original information requirement and the new requirement.

$$Gain(A) = Info(D) - Info_A(D).$$

The Class labeled attribute Trans Type has distinct two values namely Legal and Fraud. To find the splitting criterion for these tuples, we must compute information gain of each attribute

We first compute the expected information needed to classify a tuple in D:

$$Info(D) = -\frac{8}{12}\log_2\frac{8}{12} - \frac{4}{12}\log_2\frac{4}{12} = 0.9183$$

There are eight tuples belonging to the class Legal and the remaining four tuples belonging to the class Fraud.

Next we need to compute the expected information requirement for each attribute.

$$Info_{\text{custname}}(D) = 0.675$$

Gain(Custname) = 0.9183 - 0.675 = 0.2433

Similarly, we can compute

Gain(Mail Type) = 0.0103

Gain(IP Address) = 0.5849

Gain(TransAmt) = 0.0116

Because IP Address has the highest information gain among the attributes, it is selected as the splitting attribute.

2. Gain Ratio

It applies a kind of normalization to information gain using a "split information" value defined analogously with Info(D) as

$$\text{SplitInfo}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2\left(\frac{|D_j|}{|D|}\right).$$

It differs from information gain, which measures the information with respect to classification that is acquired based on the same partitioning. The gain ratio is defined as

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}$$

The attribute with the maximum gain ratio is selected as the splitting attribute.

For the attribute Custname

SplitInfo = 1.5546 GainRatio = 0.1565

For the attribute IP Address

SplitInfo = 2.3083 GainRatio = 0.2534

For the attribute MailType

SplitInfo = 3.2479 GainRatio = 0.0032

For the attribute TransAmt

SplitInfo(D) = 2.3083 GainRatio = 0.005

The IP Address has maximum gain ratio and it is selected as the splitting attribute.

3. Gini Index

The Gini index considers a binary split for each attribute. We compute a weighted sum of the impurity of each resulting partition. For example, if a binary split on A partitions D into D_1 and D_2, the Gini index of D given that partitioning is

$$\text{Gini}_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

The reduction in impurity that would be incurred by a binary split on a discrete –or continuous-valued attribute A is

$$\text{^Gini}(A) = \text{Gini}(D) - \text{Gini}_A(D).$$

The attribute that maximizes the reduction in impurity is selected as the splitting attribute. This attribute and either is splitting subset or split point together forms the splitting criterion.

There are eight tuples belonging to the class Legal and the remaining four tuples belonging to the class Fraud.

$$\text{Gini}(D) = 1 - (\frac{8}{12})^2 - (\frac{4}{12})^2 = 0.4444$$

$\text{Gini}_{\text{custname}\in\{E,R\}(D)} = 0.1875$

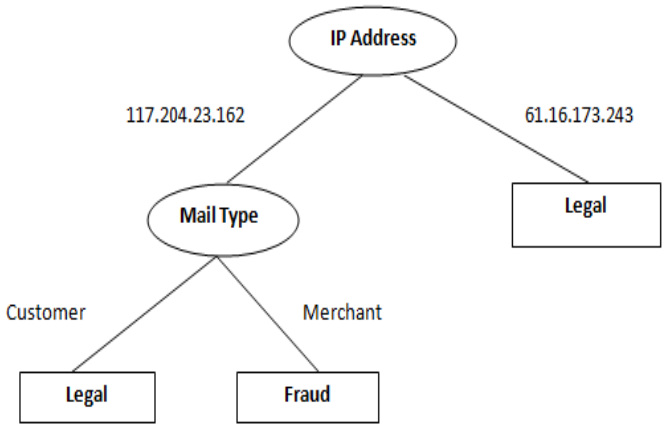
$\text{Gini}_{\text{custname}\in\{R,V\}(D)} = 0.3429$

$\text{Gini}_{\text{custname}\in\{V,E\}(D)} = 0.3704$

Gini index of attribute Mail type is 0.4381. Gini index of IP Address is 0.3333. Gini index value of Trans Amt is 0.4375.

Reduction in impurity of attribute Custname{Ezhil, Raju}is 0.2569, Custname{ Raju, Viki}is 0.1015, and Custname{Ezhil, Viki} is 0.074.

Reduction in impurity of attribute MailType is 0.0063, Reduction in impurity of attribute IP Address is 0.1111 and Reduction in impurity of attribute TransAmt is 0.0069.



V. Conclusion

In this paper efficient algorithm is used for decision tree learning. This paper describes three popular attribute selection measures – Information Gain, Gain Ratio, and Gini Index. In these tree techniques we find out that IP Address is selected as the splitting attribute. IP Address takes as the root node in decision tree induction. In this technique we simply find out fraudulent customers or merchant from their IP address through tracing fake IP address by using some software.

References

[1]. Arpita M. Hirudkar and Mrs. S.S Sherekar. "*Comparative Analysis of Data Mining Tools and Techniques for Evaluating Performance of Database System*". International Journal of Computer Science and Applications, Vol.6, No.2, Apr 2013.

[2]. Data Mining Concepts and Techniques by Jiawei Han, Micheline kamber, Jian pei.

[3]. Kalpana Rangra and Dr.K.L Bansal. "*Comparative Study of Data Mining Tools". International Journal of Advanced Research in Computer Science and Software Engineering*, Vol.4, Issue 6, June 2014.

[4]. Osama Abu Abbas."*Comparisons Between data Clustering Algorithms*". The International Arab Journal of Information Technology, Vol.5, No.3, July 2008.

[5]. Sotiris Kotsiantis, Dimitris Kanellopoulos."*Association Rules Mining: A Recent Overview*". GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp.71-82.

[6]. http://www.tutorialspoint.com/data_mining/dm_applications_trends.htm

[7]. Anika Nahar, Shamisha Roy, and syeda Shabnam Hasan."*A Survey on Different Approaches used for Credit Card Fraud Detection*". International Journal of Applied Information System (IJAIS), Volume 10-No 4, January 2016.

[8]. Sneha Patil, Harshada Somavanshi, Jyoti Gaikwad, Amruta Deshmane, and RinkuBadgujar, "*Credit Card Fraud Detection Using Decision Tree Induction Algorithm*". International Journal of Computer Science and Mobile Computing (IJCSMC), Vol.4, April 2015, pg.92-95.

[9]. Ashish Thakur, Bushra Shaikh, Vinita Jain, and A.M Magar, "*Credit Card Fraud detection Using Hidden Markov Model and Enhanced Security Features* ", International Journal of Engineering Sciences & Research Technology (IJERST), April 2015.

[10]. Sahil Hak, Suraj Singh, and Varun Purohit, "*Credit Card Fraud Detection Using Advanced Combination Heuristic and Bayes' Theoram*", International Journal of Innovative Research in Computer Science and Communication Engineering, Vol.3, Issue 4, April 2015.

[11]. Jyoti R.Gaikwad, Amruta B. Deshmane, Harshada V. Somavanshi, Snehal V. Patil and Rinku A. Badgujar, "*Credit Card Fraud Detection Using Decision Tree Induction Algorithm*", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-4, Issue-6, November 2014.

[12]. Renu and Suman, "*Analysis on Credit Card Fraud Detection Methods",* International Journal of Computer Trends and Technology (IJCTT), Vol.8, No.1, Feb 2014. Vaibhav P.Vasani and Rajendra D. Gawali, "*Classification and Performance Evaluation Using Data Mining Algorithms*", International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), Vol.3, Issue 3, March 2014.

[13]. Yusuf Sahin, Serol Bulkan, and Ekrem Duman, "*A Cost Sensitive Decision Tree Approach for fraud Detection*", ELSEVIER, Expert Systems with Applications,2013

[14]. Dr. R. Dhanpal and Gayathiri.P, "*Credit Card Fraud Detection using Decision Tree for Tracing Email and IP*", International Journal of Computer Science Issues, IJCSI, Vol.9, Issue 5, No.2, September 2012.

[15]. Bruno Carneiro da Rocha and Rafael Timoteo de souse Junior, "*Identifying Bank Frauds Using CRISP-DM and Decision Trees*", International Journal of Computer Science & Information Technology (IJCSIT), Vol.2, No.5, October 2010.