The International Journal Of Engineering And Science (IJES)
||Volume||2 ||Issue|| 6 ||Pages||01-04||2013||
ISSN(e): 2319 – 1813 ISSN(p): 2319 – 1805

Increasing the Efficiency of Credit Card Fraud Deduction using Attribute Reduction

Geetha Mary A, Arun Kodnani, Harshit Singhal, Swati Kothari
School of Computing Science and Engineering, VIT University, Vellore, Tamil Nadu, India

ABSTRACT

The detection of fraudulent credit card usage is of immense importance for banks as well as the card users and requires highly efficient techniques to increase the chance of correctly classifying each transaction as fraud or genuine. One of the techniques used to perform this classification is decision tree. Attribute reduction is used to increase the efficiency of the technique, which is decided based on entropy.

***INDEX TERMS**:- Data Mining, Decision tree, data cleaning, Attribute Reduction, Entropy*

Date Of Submission: 15 May 2013 Date Of Publication: 10,June.2013

I. INTRODUCTION

In today's economy, credit cards hold utmost importance in many sectors. Credit card fraud detection is a topic which is applicable to many industries including banking and financial sectors, insurance, government agencies, etc. Fraudulent transactions are a significant problem, one that will grow in importance as the number of access points grow. Certainly, all transactions which deal with known illegal use are not authorised. Nevertheless, there are transactions which appear to be valid but experienced people can tell that they are probably misused, caused by stolen cards or fake merchants. So, the task is to avoid fraud by a credit card transaction before it is known as illegal.

The paper deals with the problem specific to this special data mining application and tries to solve them by doing data cleaning[3]using attribute reduction and then applying decision tree technique to achieve improved efficiency of output.

II. METHODOLOGY

Decision trees are the methodologies to classify data into discrete ones using the tree structured algorithms. The main purpose of the decision tree is to expose the structural information contained in the data.

Decision tree is made by tentatively selecting an attribute to place on the root node and make one branch for each possible value of that attribute [1]. Thus, the data set at the root node split and moves into daughter nodes producing a partial tree. Then an assessment is made of the quality of the split. This process is repeated with all the attributes. Each attribute chosen for splitting produces a partial tree. Depending on the quality of the partial tree, one partial tree is selected. This virtually means selecting an attribute for splitting. The process is repeated for the data in each daughter node of the selected partial tree. If at any time, all instances at the node have the same classification, stop developing that part of the tree.

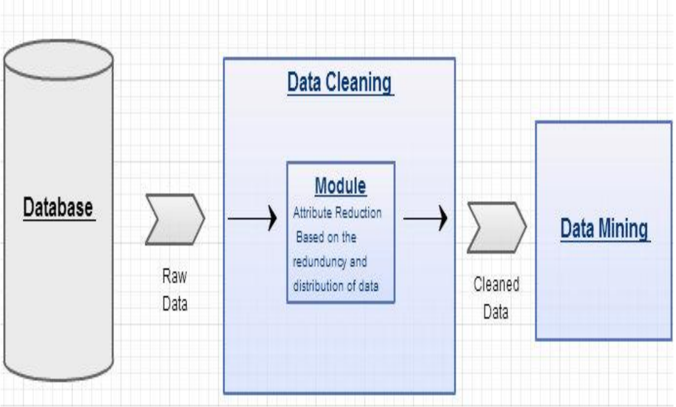
The assessment is made depending upon the purity of the daughter nodes produced; the most widely used measure to do this is called ***information entropy***[1].

Entropy is a concept originated in thermodynamics but later found its way to information theory. In the decision tree construction process, definition of entropy as a measure of disorder suits well. If the class values of the data in a node are equally divided among possible values of the class value, it can be said, that entropy is maximum. If the class values of the data in a node are same for all records, then entropy is minimum.

Through splitting, pure nodes have to be achieved, as far as it is possible. This corresponds to reducing the entropy of the system. However, this may not be as simple as it sounds for there is no way to stop the training set from containing two examples with identical sets of attributes but different classes.To overcome the problem mentioned above, the concept of information gain is used. In addition to information entropy, information ***gain***[1] also takes into consideration the factor of the number and size of daughter nodes into which attribute splits the data set.

III. APPROACH

There are also cases in which an attribute claims its position as the root node of a certain partial tree during the building of the decision tree on the basis of both information gain and entropy. There may exist such an attribute which has same value for majority of the records. Although, it may be an appropriate candidate for that position but since its value doesn't vary much over the entire dataset, considering it for classification would reduce the efficiency of the algorithm. Including these attributes results in undesirable efficiency reduction. Thus, such unnecessary attributes are reduced first, and then the decision tree algorithm is applied. This can be shown through the following analysis.


Database
Raw Data
Data Cleaning
Module
Attribute Reduction Based on the redundancy and distribution of data
Cleaned Data
Data Mining

Architecture Diagram

(Module here represents our approach of attribute reduction)

IV. ANALYSIS USING WEKA AND ORANGE

1) *Dataset description:*

The experimental data considered for analysis contains 1000 records, each consisting of values for attributes such as over draft, credit usage, existing credits, no. of dependents, employment of the user and more and finally there is a class label which classifies the transaction based on these values into two classes namely, good and bad, indicating whether the transaction was legal or illegal.A symbolic field can contain as low as two values(e.g. the kind of credit card) up to several hundred thousand values (as the code) [2].

Note that the dataset used is a sample dataset, as actual dataset for credit card cannot be accessed since it is a protected property of the banks or any other concerned financial organization.

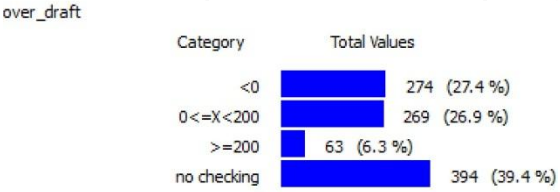
The idea proposed in this paper is valid for the entire classification algorithm but for the sake of analysis, only J48 (i.e. BasedC4.5) is used to show the result.

ATTRIBUTE	VAL	MAX (%)	ATTRIBUTE	VAL	MAX (%)
over_draft	4	39.4	Other _payment _plans	3	81.4
avg_credit_balance	5	60.3	Employem--ent	5	33.9
other parties	3	90.7	Credit_ History	5	53
personal_ status	5	54.8	Property _ magnitude	4	33.2
Housing	3	71.3	Purpose	11	28
own_ telephone	2	59.6	Job	4	63
foreign_ worker	2	96.3			

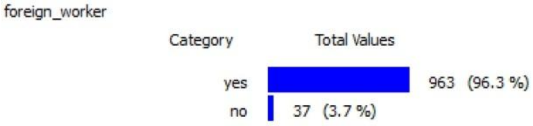
(Table 1)

VAL: number of the distinct values of the attribute
MAX: the percentage of maximum occurrence of a value in a particular attribute

Snapshot of attribute statistics for the attribute over_draft using ORANGE:



Snapshot of attribute statistics for the attribute foreign_worker using ORANGE:



Using the attribute statistics widget in ORANGE software the maximum percentage of occurrence of a value in an attribute is obtained. For instance, in case of over_draft the MAX% is very low, hence it should not be removed, whereas in case of foreign_worker, the value of VAL is very low and that of MAX is very high,hence, this attribute will decrease efficiency of output and should be removed.

2) Comparison of different classification results done on the dataset:

All the following tests were performed with the help of WEKA software:

Test1- In this test the J48 algorithm has been implemented (i.e. extension of C4.5) on the dataset without making any changes to any of its attributes.

Test2- In this test the *foreign_worker* attribute has been removed (as it has the highest % in the MAX column) and then the same algorithm has been applied to compute the result.

Test3- In this test the *other_payment_plans*attribute has been removed from the original dataset (note that this test is exclusive of the previous test i.e.*foreign_worker*attribute has not been removed) and then the same algorithmhas been applied to compute the result.

Test4- Similarly, in this test the *other_parties* attribute has been removed (without changing any other attribute) and then the same algorithm has been applied to compute the result.

Test5- In this test the *housing* attribute has been removed (without changing any other attribute) and then the same algorithm has been applied to compute the result.

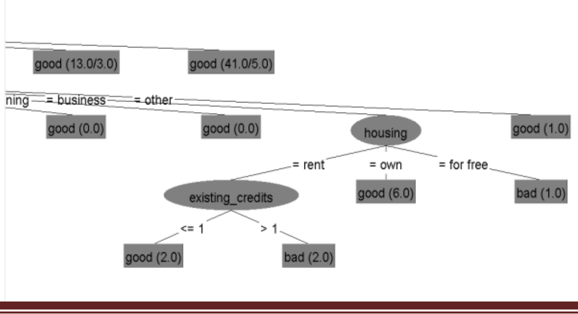
Test 6- In this test the *foreign_worker, other_payement_plans, other_parties and housing* attribute has been removed (without changing any other attribute) and then the same algorithm has been applied to compute the result.

After each test is done, the decision tree algorithm is applied; the following table contains the six outcomes of the six tests respectively.

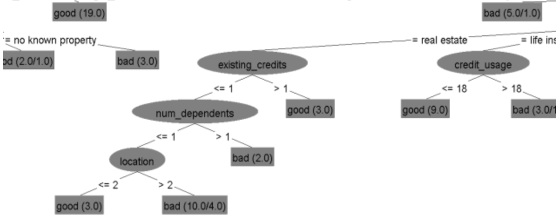
Tests	% of correctly classified instances using J48 Algorithm
Test 1	70.5
Test 2	72
Test 3	71
Test 4	72.1
Test 5	71
Test 6	72.9

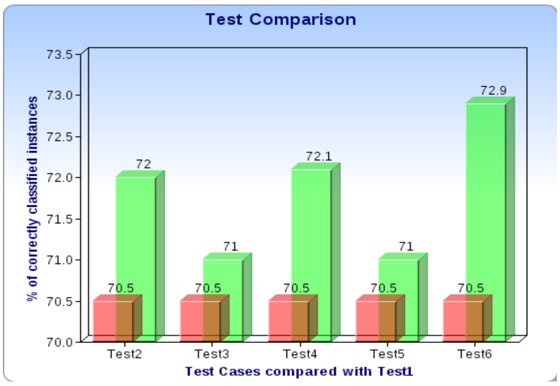
(Table 2)

Snapshot of partial decision tree after test 1 using weka:



Snapshot of partial decision tree after test 6 using weka:





(Comparison of Test Case1 with all the Test Cases)

V. CONCLUSION

From the results shown in the table 2, it can be seen that the percentage of the correctly classified instances of J48 Algorithm is increasing whenever an attribute with a very high percentage of occurrence of a single value (i.e. high value of MAX in table 1), and with very low number of distinct values for that particular attribute (i.e. low value for VAL in table 1) throughout the records is being removed. This means that even if the value of MAX is considerably high, the attribute should not be removed if the value of VAL is also high. In the last test when all the 4 attributes satisfying this criteria were deleted the percentage of the correctly classified instances jumped to 72.9% from 70.5% (result in case of normal classification done without deleting any attribute). This showed an increase of 2.4% whichis a very prominent increase, considering the sensitivity of its application and its impact on saving major monetary loss.

Hence by removing these kind of attributes the efficiency of the classification can be increased.Thus resulting in a better and efficient way to identify the fraudulent transactions amongst all the credit card transactions specified in the dataset.

REFERENCES

[1].

[2]. K.P. Soman, ShyamDiwaka, V. Ajay, “Insight into datamining theory and practice”, prentice-hall of india, 2006.

[3]. R. Brause, T. Langsdorf, M. Hepp, “Neural Data Mining for Credit Card Fraud Detection”, Frankfurt, Germany.

[4]. Dipti Thakur, Shalini Bhatia, ”Distributive Data Mining approach to Credit Card Fraud detection”, SPIT-IEEE Colloquium and International Conference, Mumbai, India