Oriental Journal of Computer Science & Technology Vol. **2**(1), 95-99 (2009)

Decision Tree Approach For Predicting Customer's Credit Risk

K. VENKAT RAO[1], T. JYOTHIRMAYI[2] and M.V. BASAVESWARA RAO[3]

[1]Department of Computer Science & Systems Engineering, College of Engineering,
Andhra University, Visakhapatnam - 530 003 (India).
[2]Department of Computer Science, GITAM University, Visakhapatnam - 530 013 (India).
[3]Sadineni Chowdaraiah College of Atrs and Science, Maddirala, Chilakaluripet - 522 611(India).

(Received: March 12, 2009; Accepted: April 04, 2009)

ABSTRACT

This paper aims at constructing the customer data warehouse which adopts an improved ID3 decision tree algorithm to implement data mining in order to predict the risk class of the customer. The obtained results are compared with experimental results in order to verify the validity and accuracy of the developed model.

Keywords: Decision Tree, ID3, classification, association rules.

INTRODUCTION

Credit risk analysis and management is important to financial institutions which provide loans to businesses and individuals. Credit risk prediction is one of the critical parts of a bank's loan approval decision process for efficient functioning of the banking system. Statistical predictive analytic techniques are one of the conventional methods applied by many researchers to analyze or to determine risk levels involved in credits, finances, and loans[1]. Siddiqi, Naeem developed score card methodology for predicting credit risk. The scorecard is a statistically based model for attributing a number (*score*) to a customer (or an account) which indicates the predicted probability that the customer will exhibit a certain behavior[2]. Sara C. Madeira *et al.,*[3], described a *data mining* approach to credit risk evaluation for a Portuguese telecommunication company .

Prediction model was used for predicting churn in the home loans of ZBANK[4]. But the regression methods do not give correct prediction results for complex data. So the decision tree approach which overcomes the problems of regression method is used. To construct the decision tree the bank customer database is considered as case study.

Data Warehouse

The bank needs different types of information in order to manage risk through capital allocation for Value at Risk coverage. The bank is concerned about the customer's details and his credit risk. While constructing the data warehouse bank may collect various information about the customer, but only a part of it is used to predict the risk. So only the relevant columns have been selected in this paper to implement the algorithm. A sample of 20 records from the data warehouse was chosen for the decision analytical purpose. In our case study the attributes owns home with data yes or no, marital status with values married or unmarried, gender male or female, education bachelor or masters and income (a<25000, 25000<b<50000,c>50000) are considered. The output of classification is assumed as two credit risk categories low c1 and high c2.

Table 1: Sample records

S.No	Own home	Married	Gender	Education	Income	Risk
1	no	Yes	male	bachelors	a	c2
2	yes	Yes	male	bachelors	c	c1
3	yes	Yes	female	masters	b	c1
4	no	No	male	bachelors	a	C2
5	yes	Yes	male	masters	a	c2
6	no	No	female	masters	c	c2
7	yes	Yes	male	masters	a	C2
8	no	No	male	bachelors	c	c2
9	yes	No	female	masters	c	c2
10	no	Yes	male	bachelors	b	c1
11	yes	No	male	bachelors	a	c1
12	no	no	male	masters	b	C1
13	no	no	female	bachelors	a	c2
14	yes	no	female	masters	c	c2
15	yes	Yes	female	bachelors	a	C2
16	yes	yes	female	bachelors	c	c1
17	yes	No	male	masters	b	c2
18	yes	No	male	masters	b	c2
19	no	Yes	female	bachelors	b	c2
20	Yes	No	male	bachelors	c	c1

Model Used

The researchers are using multiple or logistic regression, neural networks and decision trees for prediction. The regression methods are not working well for complex modeling when data have severe skews and they are producing random predictions. Hence traditional financial models for credit risk prediction are not adequate for describing today's complex relationship between the financial health and potential bankruptcy of a company. Classification approach which separates objects into classes is used. If the classes are created with out looking at data, the classification is called apriori classification else if classes are created by looking at data then the classification is called posteriori classification. There are various techniques of data mining for prediction and classification. The common methods include decision tree, Bayesian, rule based.

Decision Tree

A decision tree is a flow chart like tree structure, where each internal node denotes a test on attribute, each branch represents an outcome of the test, and the leaf nodes represent classes. To classify a unknown sample the attribute values of the sample are tested against the tree. A path is traced from the root to leaf node that holds the class prediction for that sample. The Decision Trees can be easily converted to rules.

Advantages of Decision Trees

- They are easy to use and efficient.
- Rules can easily be generated and are easy to interpret.
- They are suitable for large databases also.
- Each tuple in the data has to be filtered through the tree. This takes the time proportional to height of the tree.

Disadvantages also exist with the Decision Trees. They do not easily handle continuous data. Over fitting may occur.

The issues with Decision Tree algorithms are

- Choosing the splitting attribute.
- Ordering of the splitting attribute
- Splits

- Tree structure
- Stopping criteria
- Training data
- Pruning.

For ID3 decision tree, concept used to quantify information is called entropy. Entropy is used to measure the amount of uncertainty in a set of data. When all data in a set belongs to a single class the entropy is zero that is there is no uncertainty.

Given the probabilities $p_1,p_2,\ldots..p_n$ where $\sum_{i=1}^{n} p_i=1$, Entropy I is calculated as

$$I(p1,p2,\ldots pn)= \sum_{i=1}^{n} (p_i \log(1/p_i) \qquad ...(1)$$

$$Gain\ (D,S)=I(D)- \sum p(Di)I(Di) \qquad ...(2)$$

ID3 Algorithm

Considering the advantages and disadvantages of ID3 decision tree and the customer risk prediction we chose the ID3 algorithm.

• Advantages of ID3 algorithm

1. Every discrete classification function can be represented by a decision tree. It cannot happen that ID3 will search an incomplete hypothesis space.
2. Instead of making decisions based on individual training examples, ID3 uses statistical properties of all examples (information gain) resulting search is much less sensitive to errors in individual training examples.

• Disadvantages of ID3 algorithm

1. ID3 determines a single hypothesis, not a space of consistent hypotheses ID3 cannot determine how many different decision trees are consistent with the available training data.
2. ID3 grows the tree to perfectly classify the training examples without performing a backtracking in its search ID3 may overfit the training data and converge to locally optimal solution that is not globally optimal.
3. ID3 algorithm does not handle noise nodes.

Algorithm

To overcome limitation of noise nodes in ID3 an improved version of ID3 is adopted. In this current algorithm once a node is identified as noise node it is marked as bad node and added to a set. This node is not further considered for splitting.

Input: 20 Sample Records with attributes owns home, gender, married, education, and income.

```
PROCEDURE BuildTree(Data, ATTRIBUTE)
{
    Build(Data);
    IF (all Risk Class values of sample data in Data are the same)
    THEN Return N as a leaf node;
    ELSE
    {
        FOR (each attribute in ATTIBUTE)
        {
            IF (the attribute of the node hasn't been used to be a classification attribute before) THEN
                Compute the information gain of the attribute of the node;
        }
        IF ( the attribute whose information gain is the biggest (>0) is marked as ATT)THEN
        {
            Mark node as the node which needs to be divided next step according to ATT ;
            Divide N into Nk, and generate each branch of the node N;
        }
        ELSE
        {
            Mark the node as a bad node;
            Return the node as a leaf node;
        }
        FOR (each branch Nk) BuildTree(N, ATTIBUTE);
    }
}
```

Rules Generated From Decision Tree:

1. if income= a and married= no and owns home= no then risk=high
2. if income = a and married= no and owns home= yes then risk=low
3. if income = a and married= yes then risk=high
4. income = b and gender= male and education= bachelors then risk=low
5. if income = b and gender= male and education= masters then risk=high
6. if income = b and gender= female then risk=low
7. if income = c then risk=high

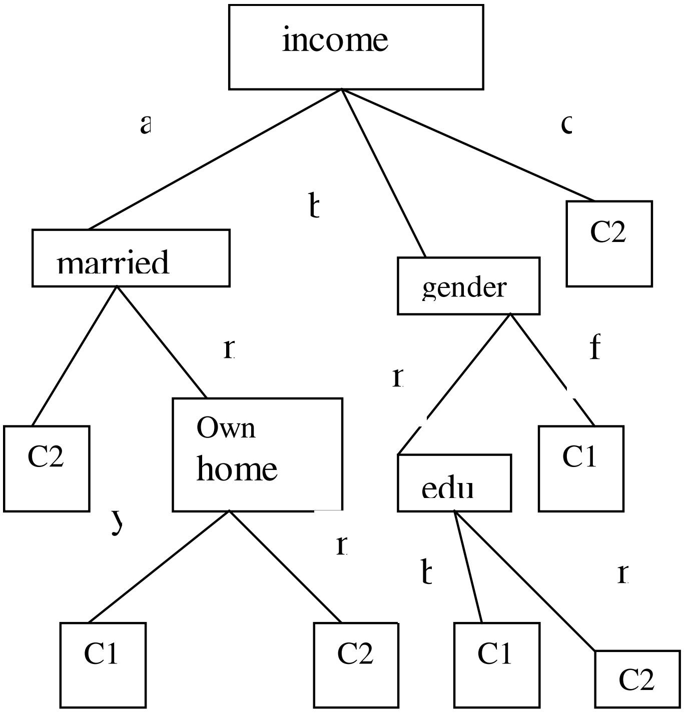

Fig. 1: Decision tree generated for sample records

Verification

Verification of the decision tree

For the same sample data set, decision tree is constructed by considering the gini index as splitting criterion. The association rules generated from the tree are same as that generated by the algorithm adopted in this current paper.

Association Rule

The association rule algorithm identifies the relations between the different attributes of customer. It helps in making decisions about the credit risk level of a customer. The rules generated by the association analysis have support and credibility. The credibility measures the accuracy of the association and support measures the usage scope of the rule.

In the current case study customer's marital status, gender, education, income and whether he owns home or not who belong to credit risk class low are considered as items and support, credibility as 30%, 90%.

The values of owns home,marital status are converted as {o1,o2},{m1,m2}and so on.
The transaction set X generated from the items shown in table 1 is
X={(o1,m1),(o1,g1),(o1,e1),(o1,I3),(m,g1),(m1,e1),(m1,i3),(g1,e1),(g1,i3),(e1,i3),(o1,m1,g1),(o1,m1,e1),(o1,m1,i3),(o1,g1,e1),(o1,g1,i3),(o1,e1,i3),(m1,g1,e1),(m1,g1,i3),(m1,e1,i3),(g1,e1,i3),(o1,m1,g1,e1),(o1,m1,g1,i3),(o1,m1,e1,i3),(m1,g1,e1,i3),(o1,g1,e1,i3),(o1,m2),(m2,g1),(m2,e1),(o1,g1,e1),(m1,e1),(e1,i3),(o1,g1,i3),(o1,e1,i3),(g1,e1,i3),(o1,g1,e1,i3)…….}

X is item combination of all frequent item set who belong to credit risk class c1. Among this set we consider only those item sets which satisfy the min support and min credibility, the following rules for the class type=low are obtained

(1). o1 → m2
(2). o1 → g1
(3). o1 → 1
(4). o1 → 3
(5). (o1 → 1) → 3
(6). m2 → 1

(7). g1 → 1
(8). e1 → 3
(9). (o1 → 1) → 1
(10). m1 → 1

The rule1can be inferred as a customer belongs to class c1 if he owns home and he is married. This rule is equivalent to second rule generated by the current algorithm.

Accuracy

We in this paper used the Holdout method for estimating the accuracy of the method which requires a training set and a test set. The data set is divided into two subsets one the training and other the test set or holdout subset. Once the ID3 algorithm generates the decision tree and the rules then the test set is used to estimate the accuracy.

CONCLUSION

This paper uses an improved ID3 algorithm to construct decision tree and generate the association rules. This algorithm can be extended to implement pruning to the skewed decision tree.

REFERENCES

1. Shu Ling Lin, A new two-stage hybrid approach of credit risk in banking industry, Expert Systems with Applications: An International Journal, v.36 n.4, p.8333-8341, (2009).
2. Credit Risk Scorecards: Developing And Implementing Intelligent Credit Scoring. Series - Wiley And Sas Business Series.
3. A Data Mining Approach to Credit Risk Evaluation and Behaviour, Sara C. Madeira, Arlindo L. Oliveira, Catarina S. Conceição.
4. Data mining and customer relationship marketing in the banking industry. By Leong Gerry, Chan Kin
5. Tian-Shyug Lee, Chih-Chou Chiu,Yu-Chao Chou, Chi-Jie Lu, “Mining the customer credit using classification and regression tree and multivariate adaptive regression splines”, *Computational Statistics & Data Analysis,* **50**, pp.1113 –1130 (2006).
6. Xue Huifeng, Zhang Weiyu, Kou Xiaodong, *Intelligent Data Mining Technology*, northwest industrial university Press (2005).
7. Lan Huang, Chun-guang Zhou, Yu-qin Zhou, Zhe Wang, “Research on Data Mining Algorithms for Automotive Customers’ Behavior Prediction Problem,” icmla, pp.677-681, 2008 Seventh International Conference on Machine Learning and Applications, (2008).