International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 4, June - 2012

Comparison of C5.0 & CART Classification algorithms using pruning technique

Prof. Nilima Patil
Computer Department
K.C.college of Engineering
Thane,(E) , Mumbai, India

Prof. Rekha Lathi
Computer Department
Pillai Institute of Technology
New Panvel,Navi Mumbai, India

Prof. Vidya Chitre
Computer Department
Bharti Vidyapith college of Engg.
Kharghar, Navi Mumbai, India

***Abstract*—Data mining is the useful tool to discovering the knowledge from large data. Different methods & algorithms are available in data mining. Classification is most common method used for finding the mine rule from the large database. Decision tree method generally used for the Classification, because it is the simple hierarchical structure for the user understanding & decision making. Various data mining algorithms available for classification based on Artificial Neural Network, Nearest Neighbour Rule & Baysen classifiers but decision tree mining is simple one. The objective of this paper is to provide the way for Decision making process of Customer for recommended the membership card. Here C5.0 & CART algorithms applied on customer database for classification. Both algorithms first applied on training dataset & created the decision tree, pruning method used for reducing the complexity then rule set are derived from decision tree. Same rules then applied on evaluation data set. Comparing the results of both algorithms & recommended the card to the new customer those having similar characteristics.**

***Keywords*- Data mining, classification algorithm, decision tree, Regression tree, membership card**

I. INTRODUCTION

In commercial operation, using the membership card service is the most superior method to help the businessmen to accumulate the customers' information. On the one hand they may obtain customers' basic information to maintain long-term contact with them. On the special day, like the holiday or the customer's birthday, they can by delivering a warm blessing, promote customers' satisfaction. On the other hand, they may through the customer's transaction information, like the purchase volume, the purchase frequency, analyze what is the value of the customer to the enterprise and analyze the characteristics of each kind card customers to help the enterprise with a clear goal to recommend the card to a new customer. The classification analysis [6] is by analyzing the data in the demonstration database, to make the accurate description or establish the accurate model or mine the classifying rule for

each category, generated classifying rule used to classify records in other databases.

II. DATA MINING TECHNOLOGY

Data mining [2] is the extraction of hidden predictive information from large databases. It is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

III. CLASSIFICATION ALGORITHMS

Decision Tree [7] is an important model to realize the classification. It was a learning system—CLS builded by Hunt etc, when they researched on human concept modeling early in the 1960s. To the late 70s, J. Ross .Quinlan put forward the ID3 algorithm [3]-[4]. In 1993, Quinlan developed C4.5 algorithm on basis of the ID3 algorithm & C5.0 classifier is successor of C4.5. Below is the detailed explanation of the C5.0 algorithm and the CART algorithm [1] which are used in this paper.

A.C5.0 Algorithm:

C5.0 algorithm is an extension of C4.5 algorithm. C5.0 is the classification algorithm which applies in big data set. C5.0 is better than C4.5 on the efficiency and the memory. C5.0 model works by splitting the sample based on the field that provides the maximum information gain. The C5.0 model can split samples on basis of the biggest information gain field..The sample subset that is get from the former split will be split afterward. The process will continue until the sample subset cannot be split and is usually according to another field. Finally, examine the lowest level split, those sample subsets that don't have remarkable contribution to the model will be rejected.

Information Gain:

Gain [8] is computed to estimate the gain produced by a split over an attribute

Let S be the sample:

- C_i is Class I; i = 1,2,…,m

  $I(s_1,s_2,..,s_m)= - \Sigma\, p_i \log_2 (p_i)$

- Si is the no. of samples in class i

  Pi = Si /S, log2 is the binary logarithm

- Let Attribute A have v distinct values.

- Entropy = E(A) is

$\Sigma\{(S1j+S2j+..+Smj)/S\}*I(s1j,..smj)$
$j=1$

- Where Sij is samples in Class i and subset j of Attribute A.

$I(S1j,S2j,..Smj)= - \Sigma\, p_{ij} \log_2 (p_{ij})$

- $Gain(A)=I(s_1,s_2,..,s_m) - E(A)$

Gain ratio then chooses, from among the tests with at least average gain,

The Gain Ratio= P(A)

$$\sum_i^t \frac{S_i}{S} \log\left(\frac{S_i}{S}\right).$$

Gain Ratio(A)= Gain(A)/P(A)

***B.** CART Algorithm*

Classification and Regression Trees (CART) is a flexible method to describe how the variable Y distributes after assigning the forecast vector X. This model uses the binary tree to divide the forecast space into certain subsets on which Y distribution is continuously even. Tree's leaf nodes correspond to different division areas which are determined by Splitting Rules relating to each internal node. By moving from the tree root to the leaf node, a forecast sample will be given an only leaf node, and Y distribution on this node also be determined.

Splitting criteria:

CART uses GINI Index to determine in which attribute the branch should be generated. The strategy is to choose the attribute whose GINI Index is minimum after splitting.

GINI index:

Assuming training set T includes n samples, the target property has m values, among them, the ith value show in T with a probability Pi, so T's GINI Index can be described as below:

$$\mathrm{GINI}(T) = 1 - \sum_{i=1}^{m} P_i^2.$$

Assuming the A be divided to q subsets, {T1, T2, · · · Tq}, among them, Ti's sample number is ni, so the GINI Index divided according to property A can be described below:

$$\mathrm{GINI}(T) = 1 - \sum_{i=1}^{q} \frac{n_i}{n} \mathrm{GINI}(T_i).$$

CART divides the property which leads a minimum value after the division.

IV. METHODOLOGY USED

In commercial operation, using the membership card service is the most superior method to help the businessmen to accumulate the customers' information. The Membership card system management which is helpful not only to accumulate the customer's information but also to offer corresponding service for different card-rank users. From this way we can enhance customers' loyalty to the store.

Therefore, so as to recommend corresponding card to the appropriate customer, we want to obtain different card-rank customers' characteristics and which is the most important factor that affects the customers to choose this kind of card not that kind.

There are various data mining classification algorithm. For this application we can make use these data mining techniques in order to find some rules from large database. Firstly on training data set, I am using two well known decision tree classification algorithms are C5.0 and CART. Figure 1: Block diagram shown below.

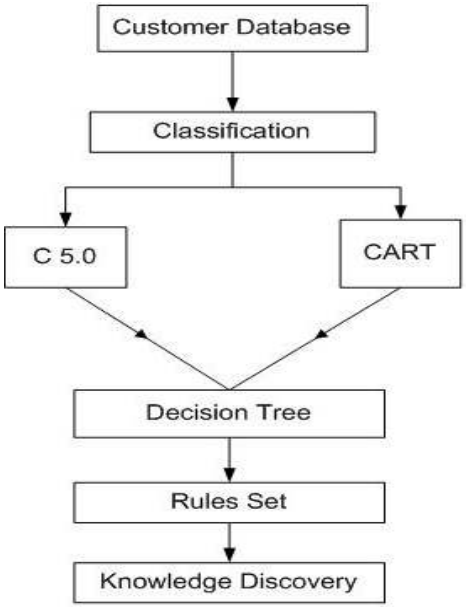

Figure 1: Block Diagram

Before implementing the algorithms, big dataset of customer information is stored in the database. Through the analysis of the customer basic information table's attributes, we need to use the target column as a membership card. In the data understanding stage, by means of analyzing the characteristics of the primary data, we can make further understanding of the data distribution and the affect of various factors to membership card type. Through the preliminary analysis, we try to find out the statistics of customer records in the database, means the percentage of golden card customer's account, the silver card customers, the bronze card customers & the normal card customer.

Methodology used in this project:-

- First consider the Data source as a training data of customer for algorithms processing. 5000 records used in training dataset & 2000 records used in Evaluation set.
- Select the membership card as a output, & other attributes used as a input.
- Apply the two algorithms on the database for finding the root & leaf of the tree by using splitting criteria.
- Perform the same operations to each attribute up to last split & create decision tree
- Pruning method is used on both the decision trees for better accuracy.

Decision tree Pruning:

Pruning [5] is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. The dual goal of pruning is reduced complexity of the final classifier as well as better predictive accuracy by the reduction of over_fitting and removal of sections of a classifier that may be based on noisy or erroneous data.
In C5.0 used the post pruning method by Binomial Confidence Limit and CART algorithm used pre_pruning method using Cost complexity model to carry on the classification for decision tree & rule set formation.

V. COMPARATIVE STUDY

In customer membership card model two algorithms will be used and comparative study of both of them will be done.

C5.0 algorithm:

In customer membership card model C5.o algorithm is used to split up data set & find out the result in the form of decision tree or rule set.

1) Splitting criteria used in c5.0 algorithm is information gain. The C5.0 model can split samples on basis of the biggest information gain.

2) Test criteria is decision tree have any number of branches available not fixed branches like CART.

3) Pruning method performed after creating decision tree i.e. post pruning single pass based on binomial confidence limits.

4) Speed of c5.0 algorithm is significantly faster than c4.5 & more accurate.

CART algorithm:

CART, short for Classification And Regression Tree, When the value of the target attribute is ordered, it is called regression tree; when the value is discrete, it is called classification tree.

1) Selection criterion used in CART algorithm is Gini index (diversity index).

2) The decision-tree generated by CART algorithm is a simple structured binary tree.

3) The same with algorithm C4.5, CART also builds the tree before pruning cross-validated using cost-complexity model.

VI. RESULTS & RULE SUMMERY

After algorithm processing, decision trees are generated for both algorithms. Pruned C5.0 & CART Decision trees are given below.

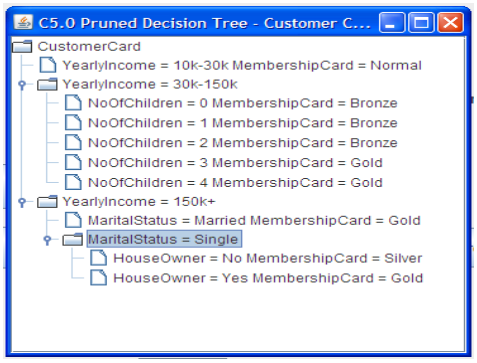

Figure 2: C5.0 Decision tree

CART decision tree

Figure 3: CART decision trees

Rule sets are formed from decision tree of both algorithms it shows that which factors are mainly affected for customer card classification & take those attributes as the main standards to recommend card to a future customer. Rule set result of both the algorithms are given below.

Classification result of C5.0 Algorithm.

Rule summary: Customer whose yearly income is from 10,000 to 30,000 is the normal card customer; Customer whose yearly income is from 30,000 to 150,000 and having less than 3 children is the bronze card customer, having more than 3 children is the golden card customer; Customer whose yearly income is more than 150,000, and unmarried is the silver card customer, married is the golden card customer.

Classification result of CART algorithm:

Rule summary: Customer whose yearly income is from 10,000 to 30,000 is the normal card customer; Customer whose yearly income is from 30,000 to 150,000, and having less than 3 children is the bronze card customer; Customer whose yearly income is more than 30,000, and having more than 3 children is the golden card customer; Customer whose yearly income is more than 150,000, and marital status single is the silver card customer.

Comparing the two classification algorithm we can reach such a conclusion: The main factors that have a big influence on customer ranks are the income and child number. Normal card customers, bronze card customer and golden card customers that are obtained from the two methods are basically the same.

Statistical analysis of both algorithms is shown in Figure 4.

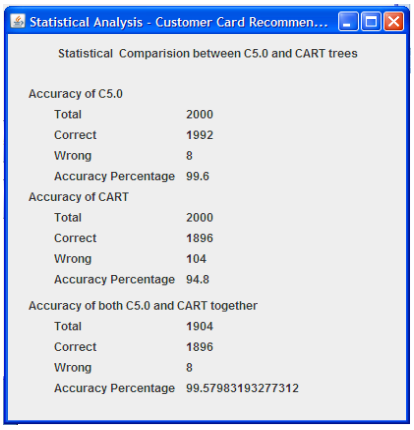

Figure 4: Statistical Analysis

Evaluation dataset of 2000 records are used for result verification. Comparing the decision tree rule set with evaluation records then checked the result.

From statistical analysis accuracy of c5.0 is 99.6 & CAR T 94.8. The following graph shows the accuracy of C5.0 is greater than CART algorithm.. In the graph on X axis consider the algorithms and Y axis shows percentage of accuracy.

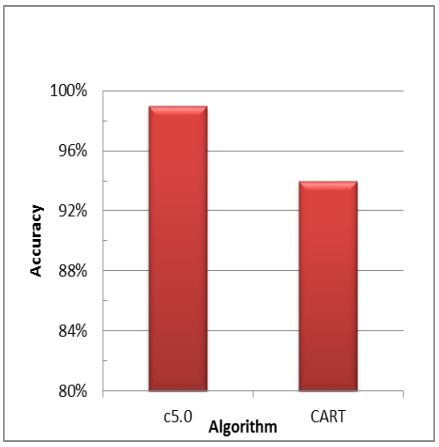

Figure 5: Comparison of algorithms

VII. CONCLUSION

Applying data mining classification algorithm in the customer membership card classification model can help to understand Children number & income level factors are affecting on the card ranks. Therefore, we are able to these two attributes as the main standards to recommend card to a new customer. Moreover, we can refer other attributes to help judge which card to recommend to future customers. On the basis of resulting factors the enterprise with a clear goal to recommend the corresponding membership card to the customer in order to provide the special service for each kind of card users, and is helpful to enterprise's development.

REFERENCES

[1] Lin Zhang, Yan Chen, Yan Liang "Application of data mining classification algorithm for Customer membership Card Model" College of Transportation and Management China 2008.

[2] Jiawei Han, Michelin Kamber,"Data Mining Concepts and Techniques" [M], Morgan Kaufmann publishers, USA, 2001, 70-181.

[3] Zhixian Niu, Lili Zong, Qingwei Yan, Zhenxing Zhao College of Computer and Software, " Auto-Recognizing DBMS Workload Based on C5.0 Algorithm" Taiyuan University of Technology, 2009.

[4] Quinlan J R, "Induction of Decision Trees", [J], Machine Learning, 1986.

[5] Ron kohavi, Ross Quinlan "Decision tree Discovery" school of computer science & engineering, New South Wales University, October 2000.

[6] Yue He† , Jingsi Liu , Lirui Lin "Research on application of data mining." School of Business and administration, Sichuan University May 2010

[7] S.B. Kotsiantis, Supervised Machine Learning: "A Review of Classification Techniques", Informatica 31,2007,pp. 249-26.

[8] Zhu Xiaoliang, Wang Jian, Wu Shangzhuo, Yan Hongcan "Research and Application of the improved Algorithm C4.5 on Decision Tree" Hebei Polytechnic University, International Conference on Test and Measurement 2009.