Comparison of Data Mining Classification Algorithms, C5.0 and CART for Car Evaluation and Credit Card Information Datasets

Ei Thinzar Win Maung, Dr.Zin May Aye
University of Computer Studies, Yangon
eithinzarwinmaung@gmail.com , zinmayaye@ucsy.edu.mm

Abstract

Data mining is the use of algorithm to discover enormous amount of data automatically by searching hidden information from large data sets using multiple algorithms and techniques. Different methods and algorithms are available in data mining system. Classification and prediction are the most common method used to make out models and predict probable data patterns and can be solving several problems in different domains like education, medicine, business, and science. In the present scenario, as almost everything is becoming computerized, various classifications algorithms have been developed to make the automatic decision process. The Decision Tree is an imperative classification method in data mining classification. This paper provides a comparison between two data mining classification algorithms: C5.0 and CART (classification and regression tree) applied on two different UCI datasets: car evaluation dataset and credit card dataset.

Keywords*: data mining, classification, decision tree, C5.0, CART*

1. Introduction

Now-a-days different types of data around the world are stored in a database. The increasing growth in data and database has generated need for new techniques and tools that can absolutely transform the processed data into meaningful information and knowledge. Hence data mining has become a crucial research for this. In classification, the mega data analyzing is a challenging process, and the need for tools and techniques that are significant in searching huge amounts of data becomes extremely important. The procedures behind this methodology create decision rules as per training and testing individual cases. A number of algorithms have been developed for classification based on data mining. Some of them include Decision Tree, k-Nearest Neighbor, Bayesian classifiers and Artificial Neural Network. At present, the decision tree has become a popular data mining method and is the simple tree structure for the user understanding and easy decision making process. The basic learning approach of decision tree is greedy algorithm, which use the recursive top-down approach of decision tree. Decision rule mining techniques are used to identify relationships between different attributes of dataset in the form of decision rules. Decision rules are more suitable to search rules for highly decision making process. Two essential steps in decision tree algorithm are training and testing. Training data will help to estimate the class label of testing data using the indication of training data that is work out by machine learning. This paper focus on the experimental analysis of the well-known classification algorithms: C5.0 and CART on two UCI datasets. Then, based on practical implementation, the results of both the algorithms have been compared to resolve which one is better in terms of performance accuracy.

2. Related Works

In commercial field, using the membership card service is the most superior method to help entrepreneur to survey the customers' information. They compared and analyzed the performance of two machine learning algorithms, C5.0 and CART applied on customer database. This was done to classify the kinds of customers' membership cards. The data source as a training for this classification process had 5000 records that were calculated from membership card. As a result the output was categorized in four classes such as normal, bronze, gold and silver card. They performed some test cases and make a conclusion that the performance accuracy was 99.6% for C5.0 and 94.8% for CART.

Another similar procedure was done in another research. In comparison of two supervised learning algorithms (C5.0 and CART) were compared using three UCI datasets (Iris flower, Titanic and Pima Indians Diabetes datasets). They evaluated the performance of classification problems under two comparative criteria: classification capacity and

generalization capacity. They carry out a conclusion that pruned trees were reducing the complexity without any loss of its analysis accuracy. It was showed that machine learning algorithm can be used to compare the algorithms for the better classification.

3. Data Collection

Data is collected from statistical websites: UC Irvine Machine Learning Repository for the information of two datasets.

At the point when an individual consider of buying a car, there are many perspectives that could affect the users' choice on which kind of car is interested in. There are different strategic points for buying a car such as PRICE (overall price), TECH (technical indication) and COMFORT (comfort), etc. The total numbers of instances in the dataset are 1728, and there are 6 numbers of attributes. These attributes are:

1. Buying price
2. Price of maintenance
3. Number of doors
4. Capacity in terms of persons to carry
5. Size of luggage boot
6. Estimated safety of the car

This is originally a multi-class classification problem and classifies the instance into 4 types of accessibility of the car. Output class for the car dataset are as follows:

Un-accessed – 1210 (70.023%)
Accessed – 384 (22.222%)
Good – 69 (3.993%)
Very good – 65 (3.762%)

In business field, the bank manager tries hard to minimize the risk and maximize of profit for the bank. When a bank receives a loan application based on the applicant's profile, the bank has to make a decision regarding whether to go ahead with the loan approval or not. Two types of risks are related with the bank's decision. If the applicant is a good credit risk, i.e. is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank. If the applicant is a bad credit risk, i.e. is not likely to repay the loan, then approving the loan to the person results in a financial loss to the bank. The credit card dataset contains data on 13 variables and the classification to consider whether an applicant is a Good or a Bad credit risk for 953 loan applicants. The input variables in the data set are described below:

1. Checking Status
2. Credit History
3. Saving Status
4. Employment
5. Personal Status
6. Other Parties
7. Residence Since
8. Property Magnitude
9. Other Payment Plan
10. Housing
11. Job
12. Own Telephone
13. Foreign Worker

Here are the class distributions (number of instance per class):

Bad – 293 (30.745%)
Good – 660 (69.254%)

4. Background Theory

Data classification is the process of finding a model that describes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. There are many classification algorithms but decision tree is the most commonly used because of its ease of execution when compared to other algorithms. This paper has used the predictive method of data mining classification algorithms: C5.0 and CART to predict the outcomes of collected statistical data.

4.1. C5.0 / See 5

C5.0 is the classification algorithm which is developed by J.Ross.Quinlan in 1994. It is a successor algorithm of C4.5 algorithm which is also extension of ID3. C5.0 builds decision trees from a set of training data in the same formula as ID3, using the concept of information entropy. It offers improved results on C4.5 in terms of speed, memory usage and size of decision tree. C5.0 model calculates the information gain for each attribute and select the maximum gain value as root node or the best splitting attribute. C5.0 is easily handled the multi-valued attributes and missing attributes from data set.

The expected information needed to classify a tuple in D is given by

$$\text{Info(D)} = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad \text{Eq-1}$$

where,
m = the quantity in class label

p_i = probability that an arbitrary tuple in D belongs to class i
Info(D) = the expected information in data D

Attribute A can be used to split D into v partitions or subsets, where D_j contains those tuples in D that have outcome a_j of A. The amount of information needed in order to arrive at an exact classification is measured by

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \quad \text{Eq-2}$$

where,
$Info_A(D)$ = the expected information of each attribute in data D
v = types of the data in that attribute

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

$$Gain(A) = Info(D) - Info_A(D) \quad \text{Eq-3}$$

Gain(A) tells how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A.

4.2. CART

CART is a non-parametric decision tree learning technique and was proposed by group of statisticians, Leo Breiman, Jerome Friedman, Richard Olshen, Charles Stone in 1984. It produces either classification or regression trees, depending on the dependent variable are categorical or numeric. If an outcome variable is continuous, CART produces regression trees; if the variable is categorical, CART produces classification trees. It can construct only binary trees mean that a node in a decision tree can only be split into two groups. CART uses Gini index (diversity index) to find the best splitting attribute, that is, minimum Gini index (or, equivalently, largest reduction in impurity) value attribute among all attributes. CART accepts data with numerical or categorical values and also handles missing attribute values.

To measure the impurity of D, a data partition or set of training tuples, as

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \quad \text{Eq-4}$$

where,
p_i = the probability that a tuple in D belongs to class C_i. The sum is computed over m classes.

For each attribute, if a binary split on A partitions D into D1 and D2, the Gini index of D given that partition is

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \quad \text{Eq-5}$$

The reduction in impurity that would be incurred by a binary split on an attribute A is

$$Gini(A) = Gini(D) - Gini_A(D) \quad \text{Eq-6}$$

The attribute that has the minimum Gini index is selected as the splitting attribute.

5. System Overview

With several respects of the differences between C5.0 and CART, it probably causes one algorithm to outperform the other. The proposed system is coded with C# programming language because C# is one of the most popular programming languages nowadays. Since the system was coded by C#, it is needed Microsoft Visual Studio 2013 for software platform and Microsoft SQL Server Management Studio 2012 to build the database.

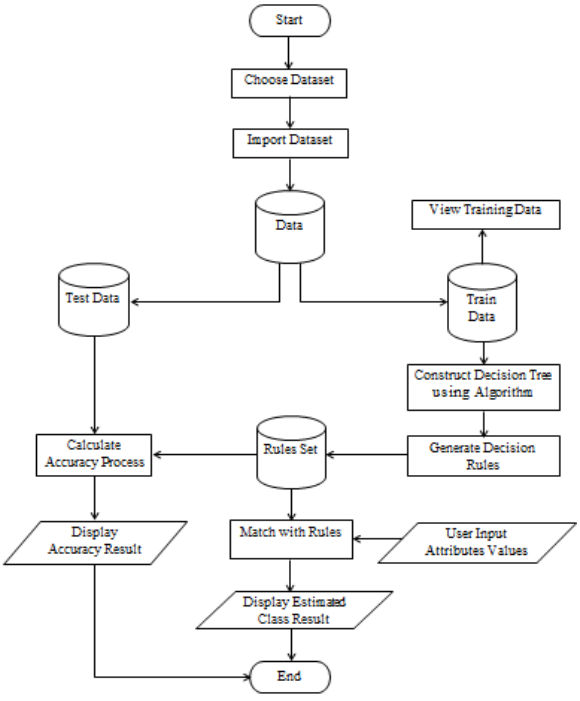

Figure 1. Flow diagram describing the research methodology

The flow diagram can be summarized as follow. In this system, there are two parts, training part to derive the classifier and testing part for accuracy performance. First, the user can choose the database (credit card information or car evaluation). At the training section, the proposed algorithm uses the training dataset and then constructs the decision tree model. After construction the tree, the system produces the appropriate rules according to the decision tree. These producing rules are stored in rule dataset. At the testing section, the user can verify the system accuracy according to the testing dataset and rule dataset by using holdout method. The user can also input the appropriate information and compute the final result for the testing cases.

5.1. Objective of the Proposed System

- To study the characteristics of decision tree algorithms under data mining system
- To know how C5.0 and CART algorithms are applied to the real-world database
- To examine the performance of two classification algorithms, C5.0 and CART
- Based on practical implementation, to determine which algorithm is better in terms of classification accuracy

5.2. Measurements of the System

Results of the training phase and testing phase are used as measurements to compare two algorithms: C5.0 and CART. The total number of leaves and the processing time are used as measurement factors of the training phase.

- The number of leaves of the decision tree represents the number of rules generated by the decision tree model.
- Processing time or run time refers to the time required to build or train a classifier.

As the result of validation phase, the accuracy measure is used to compare the performance of two algorithms: C5.0 and CART. System accuracy is the overall number of correctly classified instances divided by the total number of tuples in the test set.

$$\text{Accuracy} = \frac{number\ of\ correct\ classification}{total\ number\ of\ instances\ in\ the\ dataset} \times 100$$

To evaluate the accuracy of a classifier, holdout method is common technique based on randomly partitions of the given data. The use of this technique to compute accuracy increases the execution time and is useful for model selection. In the holdout method, the given data are partitioned into two independent sets, a training set and an evaluation set. Typically, two-thirds of the data are allocated to the training set, and the remaining is allocated to the test set. The training set is used to derive the model, whose accuracy is estimated with the test set.

6. Comparative Study and Experimental Results

The system is implemented as the windows based system using C# programming language. It is implemented by training records and includes two interfaces: Importing data to train interface and Estimating class interface. The Importing data to train interface is the first step to upload the records to be trained by the system and Estimating class interface is for prediction the class label based on relevant information.

After importing the dataset, the records may change to two parts, training and testing, and then show the home page of the corresponding dataset. In this form consists of four tab panes, Training Data List, C5.0, CART and Report. In the C5.0 and CART menus, there are two sub-menus (Decision Tree and Testing Data) for each buttons. Figure 2 shows the home page for car dataset.

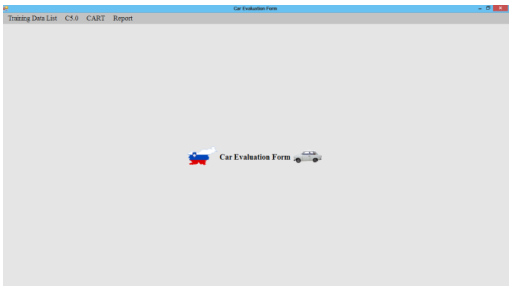

Figure 2. Home Page of the Car Dataset

Car dataset is randomly split the original data (1728 records) into two parts and one-third of data (576 instances) undergo exactly as a testing set. The remaining parts are used as training set to build the decision tree.

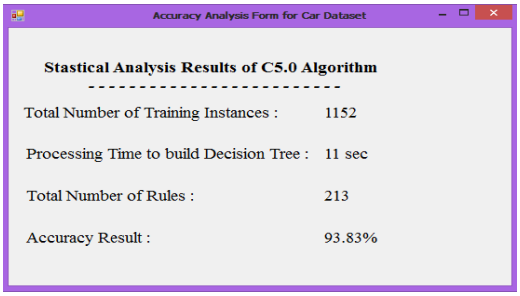

Figure 3. Analysis Report for Car Dataset Using C5.0 Algorithm

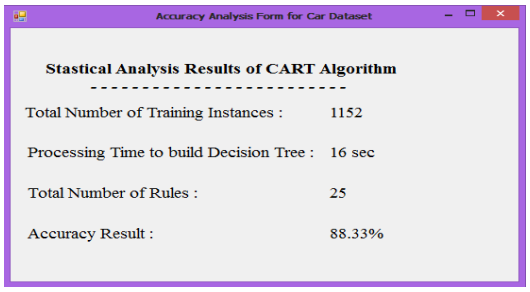

Figure 4. Analysis Report for Car Dataset Using CART Algorithm

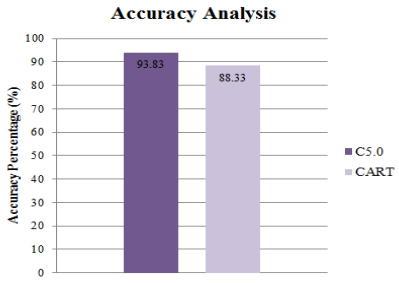

Figure 5. Comparison Chart for Car Evaluation Dataset

Credit card dataset is divided for training and testing data with probability of 0.67 and 0.33 respectively. So, train dataset contains 636 tuples and test dataset contains 317 tuples.

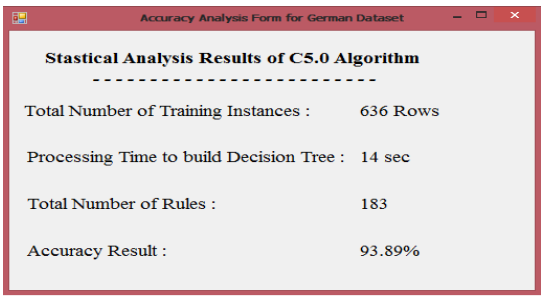

Figure 6. Analysis Report for Credit Card Dataset Using C5.0 Algorithm

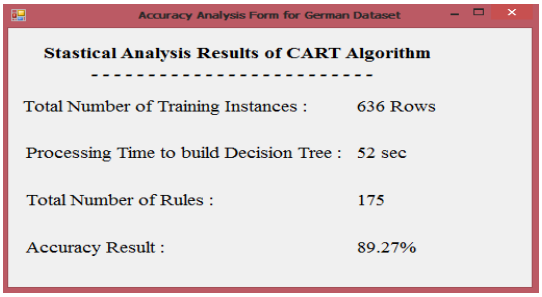

Figure 7. Analysis Report for Credit Card Dataset Using CART Algorithm

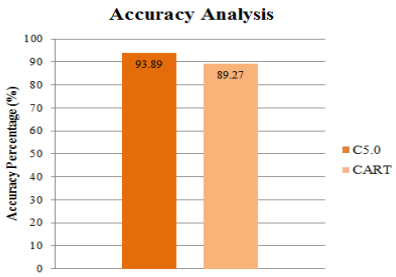

Figure 8. Comparison Chart for Credit Card Dataset

As mentioned above, two different datasets are used with the purpose of comparison between C5.0 and CART algorithms. Results obtained from experiments are shown as follows:

6.1. Processing Time

For Car dataset, Figure 3 and Figure 4 represents that C5.0 algorithm builds decision tree model faster than CART. According to the Figure 6 and Figure 7, C5.0 runs faster than CART algorithm by using German Credit Card data. By comparing training time for all datasets, C5.0 builds decision tree model remarkably faster than CART.

6.2. Number of Decision Rules

The number of leaves of the decision tree model represents the overall number of rules number generated by the model. For all datasets, as the average leaves of the decision tree derived by C5.0 is greater than that derived by CART, C5.0 produces more rule number than CART. This means that the test condition required by C5.0 to classify the whole testing data is more sufficient than that required by CART algorithm.

6.3. Accuracy

Figure 5 shows that error rate for C5.0 is low as compared to CART for Car dataset. C5.0 classifier has identified a number of instances correctly with 93.83%, followed by CART having correct classification rate of 88.33% compared to other classifiers. The resulting performance analysis of the system for Credit Card dataset is shown in Figure 8. The analysis indicates C5.0 classified the most number of correct instances with 93.89%, whereas CART classified showed the least number of correct instances with 89.27%.

7. Conclusion

This system analyzes the implementation of Decision Tree algorithms and compares their classification results based on practical implementation. For two UCI datasets, in training phase, C5.0 produces more rules than CART. So, confidence level of decision rule generated by C5.0 algorithm was more than CART. The decision tree derived by CART has maximum rule length and requires more condition test to classify the whole training set than C5.0. For the analysis of time complexity, C5.0 builds the decision tree within few seconds and CART builds significantly slower than C5.0 because CART calculates impurity of a binary partition for each categorical attribute. In testing phase, C5.0 is more accurate than CART in terms of performance accuracy for given datasets.

8. References

[1] Amit Gupta, Ali Syed, Azeem Mohammad, Malka N. Halgamuge, A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA

[2] S. Venkata Krishna Kumar, P. Kiruthika, An Overview of Classification Algorithm in Data mining

[3] Prof. Nilima Patil, Prof. Rekha Lathi, Prof. Vidya Chitre, Comparison of C5.0 & CART Classification algorithms using pruning technique

[4] J Sharmila Vaiz, Dr M Ramaswami, A Study on Technical Indicators in Stock Price Movement Prediction Using Decision Tree Algorithms

[5] Alvin Nguyen, Comparative Study of C5.0 and CART algorithms

[6] Anurag Upadhayay, Suneet Shukla, Sudsanshu Kumar, Empirical Comparison by data mining Classification algorithms (C 4.5 & C 5.0) for thyroid cancer data set

[7] Quinlan J R, “Induction of Decision Tree Allgorithms”, [J], Machine Learning, 1986