Classification of Import Market by using CART Algorithm

Wint Thinzar Aung, Pearl
Computer University, Pathein
wtzaung@gmail.com, pearl417@gmail.com

Abstract

The goal of this paper is to classify Import Market data to support the decision makers for Import Market. The main task performed in this system is used Classification and Regression Tree (CART) Algorithm. In this paper, there are two mains phases. In the training phase, attributes are analyzed by impurity function (Gini index) method. The impurity function (Gini index) measured method is used which computes the maximally reduce impurity of each attribute for the training data set. By using the training data, this system will construct the model or classifier to generate the form of classification rules. In the testing phase, calculate the accuracy of the classification rules by using the test data. This system will construct the decision tree according to the new product data and produces the result that is decision for how to investment.

Keywords: Classification, Decision Tree, CART (Classification and Regression Tree), Impurity Function (Gini Index)

1. Introduction

The Decision Tree is one of the most popular classification algorithms in current use in the Data Mining. Decision tree classifiers are found the widest applicability in the large-scale data mining environments. Classification rules represent the classification knowledge as IF-THEN rules and are easier to understand for human users [11].

Decision tree is one of the most common techniques to solve the classification problem [2, 6]. It consists of nodes, branches, leaf nodes, and a root. To classify an instance, one starts at the root and finds the branch corresponding to the value of that attribute observed in the instance. This process is repeated at the sub tree rooted at that branch until a leaf node is reached. The resulting classification is the class label on the leaf. The main objective of a decision tree construction algorithm is to create a tree such that the classification accuracy of the tree.

The objective of classification is to build a model of the classifying attribute based upon the other attributes. Applications of classification arise in diverse fields, such as retail target marketing, customer retention, fraud detection and diagnostic processes.

Classification is an important data mining problem and can be described as follows. The input data, also called the training set, consists of multiple examples. Each example is tagged with a special class label. An attribute that best partitions the training data is chosen as the splitting attribute for the root, and the training data are then partitioned into disjoint subsets satisfying the values of the splitting attribute. For each subset, the algorithm proceeds recursively until all instances in a subset belong to the same class.

Classification is one of the most frequent decision making tasks performed by human. The branching decision at each node is determined by a certain splitting criterion that generates a minimal tree, meaning that the tree has a minimum number of branches.

A classification technique (or classifier) is a systematic approach to building classification models from an input data set. Examples include decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and Naive Bayes classifiers. Each technique employs a learning algorithm to identify a model that best fits the relationships between the attribute set and class label of the input data [7, 8].

The rest of the paper is organized as follows. Section 2 summarizes the related work. In section 3 we describe necessary background theory. Section 4 presents the classification system design. Finally, we conclude the paper in Section 5.

2. Related Work

This section deals with data mining and related researches and focuses on current research work on data mining. Madigan EA Curet OL [5] employed the data mining approach CART (Classification And Regression Tree) to determine the drivers of home healthcare service outcomes (discharge destination and length of stay and examine the applicability of induction through data mining to home healthcare data.

CART methodology was developed in 80s by Breiman, Freidman, Olshen, Stone in their paper "Classification and Regression Trees" (1984) [3]. For building decision trees, CART uses so-called learning sample - a set of historical data with pre-assigned classes for all observations.

Decision trees are represented by a set of questions which splits the learning sample into smaller and smaller parts. CART asks only yes/no questions. A possible question could be: "Is age greater than 50?" or "Is sex male?", CART algorithm will search for all possible variables and all possible values in order to find the best split – the question that splits the data into two parts with maximum homogeneity. The process is then repeated for each of the resulting data fragments. Classification tree used by San Diego Medical Center for classification of their patients to different levels of risk.

A number of popular classifiers construct decision trees to generate class models. They addressed the problem of constructing "simple" decision trees with few nodes that are easy for humans to interpret. By permitting users to specify constraints on tree size or accuracy, and then building the "best" tree that satisfies the constraints, and ensured that the final tree is both easy to understand and has good accuracy.

3. Classification of Decision Tree

Data classification is a two-steps process. In the first step, a model is built to describe a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples describes by attributes. The data tuples analyzed to build the model collectively form the training data set. In the second step, a model is used for classification and the predicatively accuracy of the model is estimated. Classification and prediction have numerous applications including credit approval, medical diagnosis, performance prediction, and selective marking. Classification has many ways:

- Classification by Decision Tree
- Bayesian Classification
- Classification by Back propagation

Decision trees are powerful and popular tools for classification and prediction. Decision trees are often used in data mining and classification system because they are easily interpreted, accurate and fast. One important step in constructing a decision tree is the selection of node attributes so that the tree has a minimum number of branches. Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute. This process is then repeated for the sub tree rooted at the new node. Above these is the principle in building tree.

4. CART Algorithm

Classification and Regression Tree (CART) is primarily used as a classification, where the objective is to classify an object into two or more populations. This algorithm is analyzed both continuous and categorical variables. CART is constructing a tree that will separate the by finding binary splits on variables; finding the best splitting variable and the best splitting point at each stage with minimizing diversity.

The most important objective of decision trees is to seek accurate and small models. There are many criteria to measure node impurity. For example, the CART (Breiman et al.., 1984) [2] uses Gini index and C4.5 (Quinlan 1986) [9] uses information. In this CART algorithm, Gini index is used to aim an attribute to split on that can maximally reduce impurity. To measure purity, we can use its opposite, impurity. Gini index tends to favor tests that result in equal-sized partitions and purity in both partitions [7].

CART algorithm is performed tasks:

In growing phase, in a top down approach, create the tree by splitting recursively the nodes.

1. At the parent node, search all the possible splits for each predictor

2. Choose the best split using smallest impurity criterion among all possible predictors
3. Split
4. If each child node is the terminal node, then maximal tree
5. Else, let the child node be the parent node, go to 1
6. Classify each terminal node using plurality rule
7. If each child node is terminal node, then labeled with majority class in data partition.

CART handles missing values automatically using "surrogate splits". Uses any combination of continuous/discrete variables. Very nice feature: ability to automatically bin massively categorical variables into a few categories; eg: zip code, business class, etc…. Discovers "interactions" among variables is good for "rules" search.

4.1. Impurity Function (Gini Index) Measure

The Gini index is used in CART. To measure the impurity of D, a data partition or set of training tuples, as

$$GINI(D) = 1 - \sum_{j=1}^{k} p_i^{\,2}$$

where $p_i =$ the probability that a tuple in D belongs to class C_i.
The sum is compute over k classes.

we compute a weighted sum of the impurity of each resulting partitions. For each attribute, if a binary split on A partitions D into D_1 and D_2, the gini index of D given that partitioning is

The sum is compute j over k classes

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

The attribute A that maximizes the reduction in impurity is

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

The attribute that maximizes the reduction in impurity (or, equivalently, has the minimum Gini index) is selected as the splitting attribute.

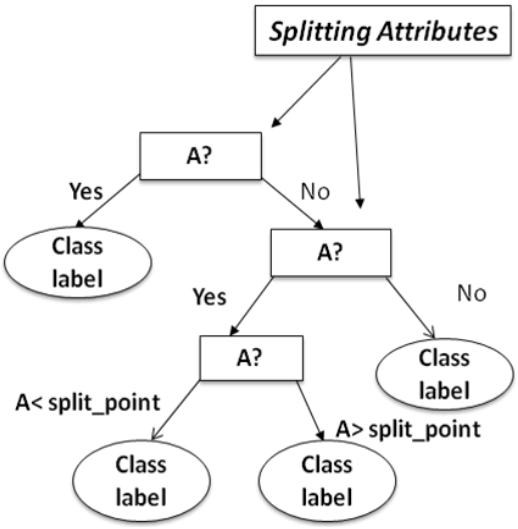


Figure 1: Decision Tree

Then, the algorithm computes Gini index of each attribute. The attribute with minimum Gini index is chosen as the test attribute for the given set. A node is created and labeled with the test attribute. Below node, branches are created for each value of the test attribute and samples are partitioned until leaf node is reached.

5. System Design

In Figure 2, the required data are collected store in dataset. And then, two third of dataset is used for training data and remaining one third is used for testing data. Training data is applied CART algorithm to generate decision tree for this system. CART algorithm uses Gini index to select the test attribute at each in the tree. The dataset is then split on the different attributes. Gini index is used to aim an attribute to split on that can maximally reduce impurity. Then it is calculate a weighted sum of the impurity of each resulting partitions for the split. The resulting value is subtracted from the impurity value before the split. The result is maximizes the reduction in impurity or minimum Gini index. The attribute that yields the smallest Gini index is chosen for the decision node. This system will describe decision tree and extract rules from decision tree. Testing data are used to evaluate accuracy of this system. New user who used this system chooses data of all attributes firstly. The selected data is compared decision tree that appeared from applying CART algorithm and produces result (Investment) to the user.

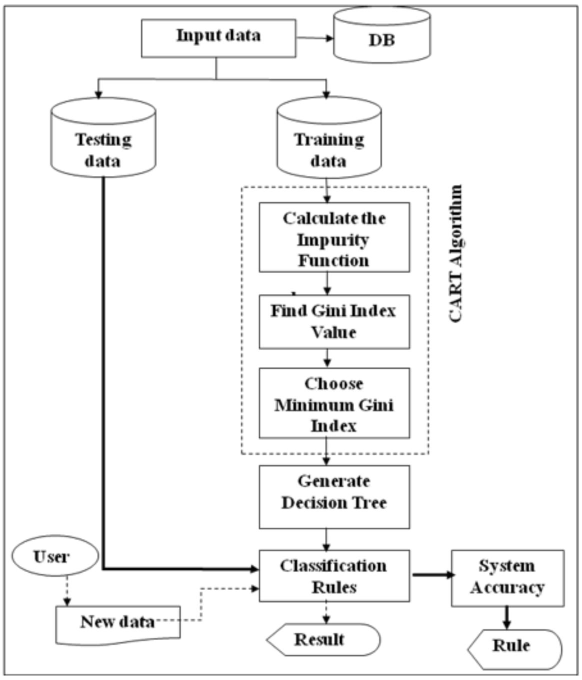


Figure 2: System flow Diagram

Firstly, this system will be used ladies wears dataset from the stored database as shown in Table 1.

Table 1: Sample Training Data

No	Name	Size	Price	Quality	Whole sales	Retail sales	Class
1	butterfly	S	3000	L	40000	40000	M
2	butterfly	M	4000	M	34000	15000	M
3	butterfly	L	5000	H	50000	32000	M
4	butterfly	XL	3000	L	15000	15000	M
5	flora baby pink	S	4000	L	25000	15000	M
6	flora baby pink	M	5000	M	22000	34000	M
7	flora baby pink	L	6000	H	15000	40000	H
8	flora baby pink	XL	5000	H	20000	30000	H
9	Ladies blue	L	5000	L	20000	18000	H

5.1. Classifier Accuracy

The primary metric for evaluating classifier performance is classification accuracy-the percentage of test samples that are correctly classified. The other important matrices are classification time. The ideal goal for a decision tree classifier is to produce compact, accurate tree in a shortest classification time.

Estimating classifier accuracy is important because it determines to evaluate how accurately a given classifier is. And accuracy also help in the comparison of different classifiers. Accuracy is measured using a test set of object for which the class labels are known. Accuracy is estimated as the number of correct class predictions, divided by the total number of test samples.

To prevent over-fitting, CART employs a pruning method which is called minimal cost-complexity pruning. To build a right sized tree by estimating the true misclassification cost. In each sub-tree, misclassification cost and cost-complexity value are calculated using k-fold cross validation.

The k-fold cross validation avoids overlapping test sets. First step, split data into k subsets of equal size. Second step, use each subset in turn for testing, the remainder for training. After the subsets are stratified before the cross-validation is performed. The error estimates are averaged to yield an overall error estimate. Standard method for evaluation : stratified 10-fold cross-validation.

Sensitivity = t_pos / pos
Specificity = t_neg / neg

Accuracy = sensitivity pos / (pos + neg) +
specificity neg / (pos + neg)

- pos is the number of positive samples.
- t_pos is the number of true positives (correct classification of positive cases)
- f_pos is the number of false positives (incorrectly classification of positive cases)
- neg is the number of negative samples.
- t_neg is the number of true negatives (correct classification of negative cases)
- f_neg is the number of false negatives (incorrectly classification of negative cases)

5.2. Experimental Result

This paper presents decision tree construction by using CART algorithm for ladies wear import market. In this algorithm, an impurity function based attributes selection measure is used to select the test attribute at each node in the tree. Thus, node N is labeled and branches are grown for each of the attribute’s value. The tuples are then partitioned to the same class. This process continues until all partition branches have the pure class. Then, decision tree is constructed for Blouse as followed in Figure 3.

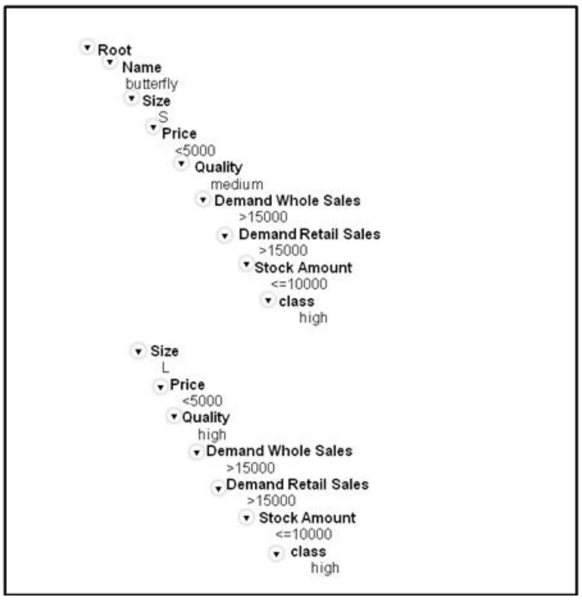
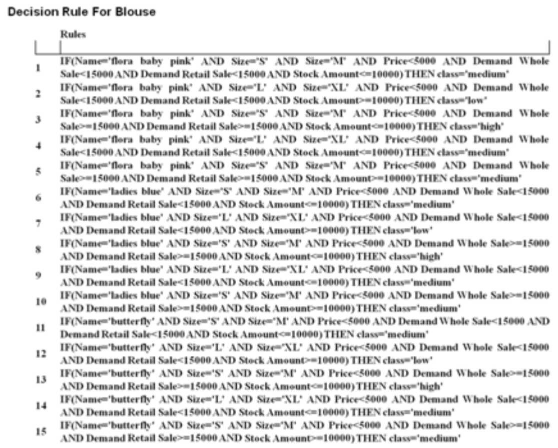


Figure 3: Decision Tree Produced by CART Algorithm

According to the constructed decision tree, system will generate the classification rules as followed in Figure 4.

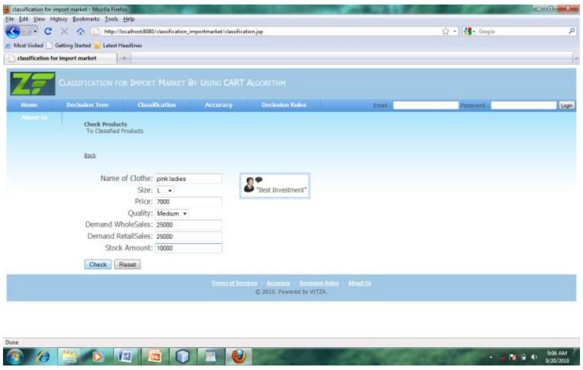


Decision Rule For Blouse

	Rules
1	IF(Name='flora baby pink' AND Size='S' AND Size='M' AND Price<5000 AND Demand Whole Sale<15000 AND Demand Retail Sale<15000 AND Stock Amount<=10000) THEN class='medium'
2	IF(Name='flora baby pink' AND Size='L' AND Size='XL' AND Price<5000 AND Demand Whole Sale<15000 AND Demand Retail Sale<15000 AND Stock Amount>=10000) THEN class='low'
3	IF(Name='flora baby pink' AND Size='S' AND Size='M' AND Price<5000 AND Demand Whole Sale>=15000 AND Demand Retail Sale>=15000 AND Stock Amount<=10000) THEN class='high'
4	IF(Name='flora baby pink' AND Size='L' AND Size='XL' AND Price<5000 AND Demand Whole Sale<15000 AND Demand Retail Sale<15000 AND Stock Amount<=10000) THEN class='medium'
5	IF(Name='flora baby pink' AND Size='S' AND Size='M' AND Price<5000 AND Demand Whole Sale>=15000 AND Demand Retail Sale>=15000 AND Stock Amount>=10000) THEN class='medium'
6	IF(Name='ladies blue' AND Size='S' AND Size='M' AND Price<5000 AND Demand Whole Sale<15000 AND Demand Retail Sale<15000 AND Stock Amount<=10000) THEN class='medium'
7	IF(Name='ladies blue' AND Size='L' AND Size='XL' AND Price<5000 AND Demand Whole Sale<15000 AND Demand Retail Sale<15000 AND Stock Amount>=10000) THEN class='low'
8	IF(Name='ladies blue' AND Size='S' AND Size='M' AND Price<5000 AND Demand Whole Sale>=15000 AND Demand Retail Sale>=15000 AND Stock Amount<=10000) THEN class='high'
9	IF(Name='ladies blue' AND Size='L' AND Size='XL' AND Price<5000 AND Demand Whole Sale<15000 AND Demand Retail Sale<15000 AND Stock Amount<=10000) THEN class='medium'
10	IF(Name='ladies blue' AND Size='S' AND Size='M' AND Price<5000 AND Demand Whole Sale>=15000 AND Demand Retail Sale>=15000 AND Stock Amount>=10000) THEN class='medium'
11	IF(Name='butterfly' AND Size='S' AND Size='M' AND Price<5000 AND Demand Whole Sale<15000 AND Demand Retail Sale<15000 AND Stock Amount<=10000) THEN class='medium'
12	IF(Name='butterfly' AND Size='L' AND Size='XL' AND Price<5000 AND Demand Whole Sale<15000 AND Demand Retail Sale<15000 AND Stock Amount>=10000) THEN class='low'
13	IF(Name='butterfly' AND Size='S' AND Size='M' AND Price<5000 AND Demand Whole Sale>=15000 AND Demand Retail Sale>=15000 AND Stock Amount<=10000) THEN class='high'
14	IF(Name='butterfly' AND Size='L' AND Size='XL' AND Price<5000 AND Demand Whole Sale<15000 AND Demand Retail Sale<15000 AND Stock Amount<=10000) THEN class='medium'
15	IF(Name='butterfly' AND Size='S' AND Size='M' AND Price<5000 AND Demand Whole Sale>=15000 AND Demand Retail Sale>=15000 AND Stock Amount>=10000) THEN class='medium'

Figure 4: Classification Rules

The system will produce the result by using classification rules. The result is implemented as how should we invest for import product.



CLASSIFICATION FOR IMPORT MARKET BY USING CART ALGORITHM

Check Products
To Classified Products

Name of Clothe: pink ladies
Size: L
Price: 7000
Quality: Medium
Demand WholeSales: 25000
Demand RetailSales: 25000
Stock Amount: 10000

"Best Investment"

© 2010. Powered by WTZA.

Figure 5: Result of Investment

This system intends to be easily used the decision maker. Moreover, to assist people who interested in import market as a reference.

6. Conclusion and Further Extension

Classification predicts categorical class labels. It classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data. In this paper, the system can provide for the users who want to know which investment for products for import market. Moreover, this system can support the users to get more precise information. The computer will make the decision instead of the user. This system supports decision marker, business man and other user should be decision for marketing by using CART algorithm.

Furthermore, we are going to decide investment of export market for desire products using CART algorithm.

7. Reference

[1] Applying Decision Trees in Classification Tasks.

[2] L.Breiman et. al. Classification and regression Trees. Wadswort, Belmont, 1984.

[3] L.Breiman, J.Friedman, R.Olshen, C.Stone, Classification and Regression Trees, Wadsworth, Pacific Grove, CA, 1984.

[4] M. Singh, P. K. Wadhwa and P. S. Sandhu, "Human Potein Function Prediction using Decision Tree Induction", Deptt. Of CSE & IT, Guru Nanak Dev Engineering College, Ludhiana, Punjab, INDIA, April, 2007.

[5] Madigan EA Curet OL _ A data mining approach in home healthcare _ outcomes and services use, 2006 Feb 24, 6:18.

[6] Michael Negnevitsky; Artificial Intelligence A Guide to Intelligence System Second Edition, Addison Wesky, 2005.

[7]Phyo Phyo Ei, Renu, **“**Implementation of Classification Rules Mining by Using Decision Tree Induction for decision Making System”, Computer University, Myeik, 2009.

[8] Prof. Dr. Wolfgang H¨ardle, CASE- Center of Applied Statistics and Economics, CART theory and application, Humboldt University, Berlin.

[9] Quinlan. Induction on Decision Trees. Machine learning, 1(1):81_106, 1986.

[10] R. Selvamani and D. Khemani, "Decision Tree Induction with CBR", A.I.&D.B.Lab, Dept. of Computer Science & Engineering I.I.T.Madras, India.

[11] Su Mon Aung, “Developing Decision Making System to Classify Pests of Paddy Infesting”, Computer University, Pathein, 2009.