Implementation of Classification Rules Mining by Using Decision Tree Induction for Decision Making System

Phyo Phyo Ei, Renu, Nilar Myint
Computer University, Myeik, Myanmar
phyoeimoe@gmail.com, renushi@gmail.com

Abstract

In many areas, large quantities of data are generated and collected everyday, such as supermarket, business, astronomy, geography, phone call records etc. These data arrive too fast to be analyzed or mined in time. The goal of this paper is to extract data easily and rapidly. The main task performed in this system is using decision tree induction methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. Test attributes are selected on the basis of Information gain measure. In this paper, the classification of training data is proposed in which the resulting classifier is a decision tree induction. Tree method for classification is exploited to identify user's data. The information gain measure method is used which computes the highest information gain of each attribute for the training data set.

Keywords: Classification, data mining, decision tree, attribute, entropy.

1. Introduction

Data is a fact. People collect more and more data making their original reason for collecting the data harder to accomplish. The problem is that it becomes harder to use the data to derive the knowledge. Data mining is based on automated methods for pattern discovery and general learning from data. Data mining's promise is the ability to discover interesting patterns that are hidden in large amounts of data [2].

Data Mining is the process of discovering interesting knowledge from large amount of data stored either in database, data warehouse or other information repositories. Data Mining is the search for relationships and patterns that exist in large database but are hidden among the vast amount of data. Classification is the process of finding the common properties among different entities and classification of object is based on the set of data features used and strongly affects classifier design [2].

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 [9].

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. The model generated by a learning algorithm should both bit the input data well and correctly predict the class labels of records it has never seen before [10].

The main task performed in this system is using decision tree induction methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. Test attributes are selected on the basis of a heuristic or statistical measure (e.g. Information gain measure). Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. This paper examines the decision tree algorithm and implements using Visual Basic development technology [12].

2. Related Work

M. Garofalakis, et.al, [6] proposed that classification is an important problem in data mining. 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.

M. Kamber, et.al, [5] proposed that the main task performed in these systems is using inductive methods to the given values of attributes of an un-known object to determine appropriate classification according to decision tree rules. They used the decision tree learning algorithm ID3. They first implemented basic ID3 in which they dealt with the target function that has discrete output values.

Myo Myo Than Naing, Tin Htar Nwe [7] described expert system, in which decision making process is applied human's intelligence in machines. When using decision tree induction, the decision making process itself can be easily validated. Their system reduced time consuming and costs and uses easily without requiring much computer-skill.

3. Motivation

People have many difficulties to handle massive amount of data stored in databases. So, we need to find the ways to automatically analyze such data. Among them classification rule mining is becoming the useful rule mining approach. Classification was the process of finding the common properties among different entities and classifying them into classes. The classification of object was based on the set of data features used and strongly effects classifier design. A decision tree is a classifier which can work efficiently over large volumes of data. It is a flow-chart-like tree structure, where each internal; node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The top-most node in a tree is the root node. In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node that holds the class prediction for that sample.

In this paper, Decision tree algorithm is commonly used for gaining information for decision making. The paper presents Women's clothes to describe classification rule and decision tree induction algorithm. In this algorithm, an entropy based attributes selection measure is used to select the test attribute at each node in the tree. This system intends to be easily used the user. It is a user friendly system. Moreover, to assist people who interested in clothes as a reference.

4. Classification of Data Mining

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

Classification Based on Concepts from Association Rule Mining [7].

4.1. Classification by Decision Tree Induction

A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The top-most node in a tree is the root node. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals.

Unknown samples can be classified by testing attributes against the tree. The path traced from root to leaf holds the class prediction for that sample. The basic algorithm for inducing a decision tree from the learning sample set is as follows: [2]

- Initially the decision tree is a single node representing the entire training set.
- If all samples are in the same class, this node becomes a leaf and is labeled with that class label.
- Otherwise, an entropy-based measure, “information gain”, is used as a heuristic for selecting the attribute which best separates the samples into individual classes.
- A branch is created for each value of the test attribute and samples are partitioned accordingly.
- The algorithm advances recursively to form the decision tree for the sub-sample set at each partition. Once an attribute has been used, it is not considered in descendent nodes.
- The algorithm stops when all samples for a given node belong to the same class or when there are no remaining attributes (or game other stopping condition).

The following(s) are the basis of several commercial rule induction systems. They are:

- Decision Tree Induction
- Tree Pruning
- Extracting Classification Rules from Decision Trees
- Enhancements to Basic Decision Tree Induction [10]

` The main task performed in this method is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. There are two steps in decision tree induction. They are model construction and model usage [12].

4.1.1. Model Construction

Describe a set of predetermined classes. Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute. The set of tuples used for model construction: training set. The model is represented

as classification rules, decision trees, or mathematical formulae [6].

4.1.2. Model Usage

For classifying future or unknown objects, estimate accuracy of the model. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set, otherwise over-fitting will occur.

Decision tree starts with a root node on which it is for users to take actions. Decision tree is a classifier in the form of a tree structure, where each node is either:

A Leaf Node - Indicates the value of the target attribute (class) of examples, or
A Decision Node - Specifies some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test. Among methods, decision tree learning is attractive for three reasons: [12]

1. Decision tree is a good generalization for unobserved instance, only if the instances are described in terms of features that are correlated with the target concept.
2. The methods are efficient in computation that is proportional to the number of observed training instances.
3. The resulting decision tree provides a representation of the concept those appeals to human because it renders the classification process self-evident [5].

4.1.3. Decision Tree Induction Algorithm

The algorithm for decision tree induction is as follow:

Input : The training samples, samples, represented by discrete-valued attributes; the set of candidate attributes, attributes-list.

Output : A decision tree.

Method :

(1) create a node *N*;
(2) if sample are all of the same class, *C* **then**
(3) return N as a leaf node labeled with the class C_i
(4) if *attribute-list* is empty **then**
(5) return N as a leaf node labeled with the most common class in *samples*;
(6) select *test-attribute*, the attribute among *attribute-list* with the highest information gain;
(7) label node *N* with *test-attribute*;
(8) **for each** known value a_i of *test-attribute* // partition the samples
(9) grow a branch from node *N* for the condition *test-attribute* = a_i ;
(10) let s_i be the set of samples in samples for which *test-attribute* = a_i // a partition
(11) **if** s_i is empty **then**
(12) attach a leaf labeled with the most common class in samples;
(13) else attach the node returned by Generate_decision_tree *(s_i, attribute-Lists-test-attribute);*

5. Proposed System Architecture

The following figure is the architecture for our proposed system. This system uses the training data set to analyze decision tree induction. When the data of the user enters, the system checks whether the data is new data or test data. If the data is test data, it is in the decision tree induction. If the data is new data, this system classifies it. Classification generates the output result that is unknown class.

This system proposes the classification of training data in which the resulting classifier is a decision tree induction. The decision tree induction produces the rules that are needed for the data. The rules support to reduce a lot of conditions. The decision tree induction method decided estimates accuracy for user the system makes decision for the user to buy the data easily and rapidly. The proposed system constructs the decision tree as the top down processing.

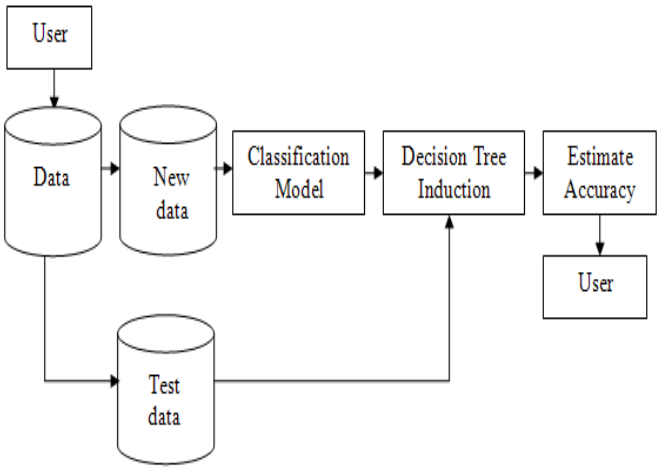

Figure 1. Proposed System Architecture

5.1. Data Flow of the System

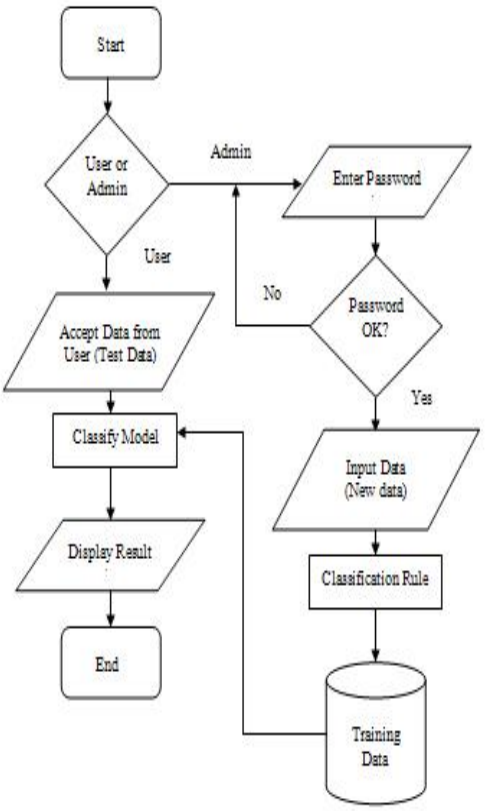

Figure. 2. Data Flow of the System

The data flow of the system is described in the above figure. The system includes two parts which are user and admin. In this system, user is checked whether he/she is administrator of the system who will insert the password. If his / her password is correct, admin adds data into the training data. In this system each user can select the attributed as he/she likes. The user's selected attributes are used to compare the training data set. Based on the decision result of each user, the system makes the user of decision result to buy easily and rapidly.

6. Proposed Method

6.1. Attribute Selection Measure

The information gain measure is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split. The attribute with the highest information gain is chosen as the attributes for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness to "impurity" in these partitions, Entropy is going to be discussed first [4].

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i), \qquad (1)$$

Where,

I=a set consisting of data samples

S_i=number of data samples

P_i=probability that an arbitrary belongs to class C_i

Let attribute A have v distinct values $\{a_1, a_2 \ldots, a_v\}$. Attribute A can be used to partition S into v subset, $\{s_1, s_2, \ldots, s_v\}$, where s_j contains those samples in S that value of a_j of A. A is given by

$$E(A) = \sum_{j=1}^{m} \frac{s_{1j} + \cdots + s_{mj}}{s} I(s_1, s_2, \ldots\ldots, s_m) \qquad (2)$$

$$Gain\ (A) = I(s_1, s_2, \ldots\ldots, s_m) - E(A) \qquad (3)$$

Gain (A) is the expected reduction in entropy caused by knowing the value of attribute (A). The algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given set [4].

6.2. Experimental Evaluation

Table1 shows the training data set for solving classification problems. First, a training set consisting of records whose class labels are known must be provided. The training set is used to build a classification model, which is subsequently applied to the test set, which consists of records with unknown class labels. Most classification algorithms seek models that attain the highest accuracy, or equivalently, the lowest error rate when applied to the test set.

In this paper, this system proposes the classification of training data in which the resulting classifier is a decision tree induction.

The series of requested attributes and their possible results can be organized in the form of a decision tree, which is a hierarchical structure consisting of nodes and directed edges. The tree has three types of nodes: (1) A root node that has no incoming edges and zero or more outgoing edges. (2) Internal nodes, each of which has exactly one incoming edge and two or more outgoing edges.(3)Leaf or terminal nodes, each of which has exactly one incoming edge and no outgoing edges.

Table 1. The Training Data Set

ID	Age	Price	Design	Brand	Size	Color	Customer's Favorite	Class: customer buying status
1	>35	HIGH	MODERN	PIM	XL	WHITE	DISLIKE	NOT BUYING
2	>35	MEDIUM	MODERATE	TRI	L	BLACK	DISLIKE	NOT BUYING
3	21…35	LOW	MODEST	GIO	M	WHITE	LIKE	BUYING
4	<=20	HIGH	MODERN	PIM	S	BLACK	LIKE	BUYING
5	<=20	MEDIUM	MODERATE	TRI	XL	WHITE	LIKE	BUYING
6	<=20	LOW	MODEST	GIO	L	BLACK	LIKE	NOT BUING
7	21…35	HIGH	MODERN	PIM	M	WHITE	LIKE	BUYING
8	>35	MEDIUM	MODERATE	TRI	S	BLACK	LIKE	NOT BUYING
9	>35	LOW	MODEST	GIO	XL	WHITE	LIKE	BUYING
10	<=20	HIGH	MODERN	PIM	L	BLACK	LIKE	BUYING
11	>35	LOW	MODERATE	TRI	M	WHITE	LIKE	BUYING
12	21…35	LOW	MODEST	GIO	S	BLACK	LIKE	BUYING
13	21…35	HIGH	MODERN	PIM	XL	WHITE	LIKE	BUYING
14	<=20	MEDIUM	MODEST	GIO	L	BLACK	DISLIKE	NOT BUYING

In a decision tree, each leaf node is assigned a class label. The non-terminal nodes which include the root and other internal nodes, contain attributes test conditions to separate records that have different characteristics. The root node shown in figure3 uses the attribute age to separate buying and not buying.

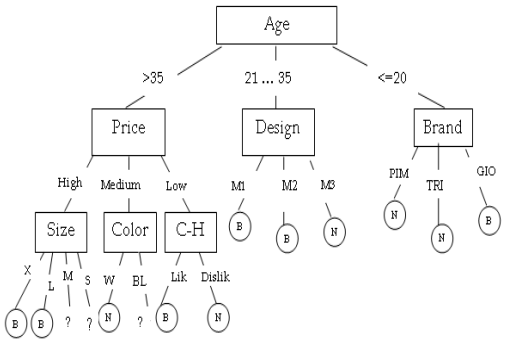

Figure. 3. A Complex Decision Tree Describing Training Data

Where,
XL = Exclusive Large, L = Large
M = Medium, S = Small
M_1 = Modern, M_2 = Moderate, M_3 = Modest
W = White, BL = Black, B = Buying
N = Not Buying, C-H = Customer-Choice

A compact decision tree is preferable since it is more general and its predictive power is often higher than that of a complex decision tree. Figure4 a compact decision tree describing of training data.

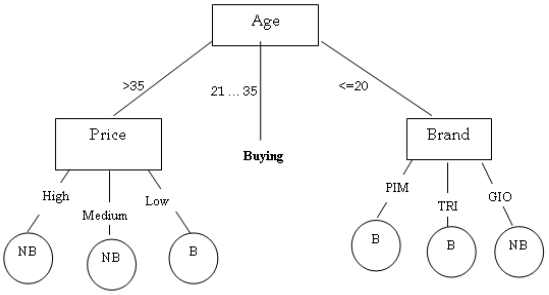

Figure. 4. A Compact Decision Tree Decision of Training Data

The following nine steps compute the information gain of attribute age, prices, brand, design and size. The attribute with the highest information gain is chosen as the test attribute for the given set.

Step 1:

Current node C = root node of the tree.

Step 2 and 3:

Entropy of the node C = E(C) = - (9/14) $\log_2$ (9/14) - (5/14) $\log_2$ (5/14) = 0.940

Step 4:

Entropy of the partial tree based on the age attributes:
For age = ">35''
E (age) = - (2/5) $\log_2$ (2/5)-(3/5) $\log_2$ (3/5) = 0.971
For age = "21 … 35''
E (age) = - (4/4) $\log_2$ (4/4)-(0/4) $\log_2$ (0/4) = 0
For age = "<=20 ''
E (age) = - (3/5) $\log_2$ (3/5)-(2/5) $\log_2$ (2/5) = 0.971
E (age) = (5/14)*E (age) + (4/14)*E (age) + (5/14)*E (age) = 0.694

Step 5:

Information gain due to the partition by the age attributes:
G (age) = E(C) –E (age) = 0.246

Step 6:

Similarly, the information gains due to the partition by the price, design, brand, size and color attributes, respectively, are:
G (price) = 0.195, G (design) = 0.052
G (brand) = 0.052, G (size) = 0.282
G (color) = -0.043

Step 7:

The age attribute is selected as the classifying attribute for the current node C

since its information gain is the largest among all of the attributes.

Step 8:

The resulted partial decision tree is:

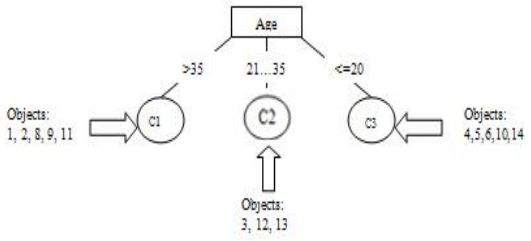

Figure. 5. The Resulted Partial Decision Tree

Step 9:

The analysis continues for the node C1 and C3 until all of the leaf nodes are associated with objects of the same class.

The resulted final decision tree is:

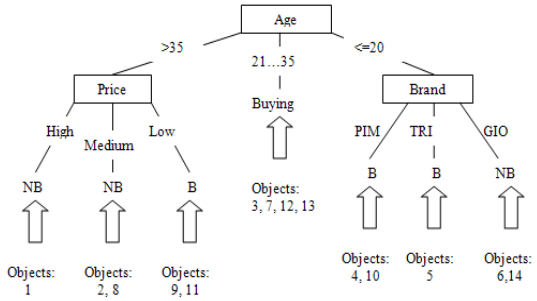

Figure. 6. The Resulted Final Decision Tree

7. Conclusion

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. This paper uses decision tree induction in data mining classification algorithms to get useful information to decision-making out of customer behaviors. It intends to help the selecting and interesting customer for decision making. It reduces time consuming, and can easily choose women's wears. By this means, enterprises can provide special service to different users in order to attract more and more customers.

References

[1] M. Garofalakis, D. Hyun, R. Rastogi, and K. Shim, "Efficient Algorithms for Constructing Decision Trees with Constraints".

[2] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Department of Computer Scienc, University of Illinois at Urbana-Champaign, 2001.

[3] M. Kamber, L. Winstone, W. Gong, S. Cheng, and J. Han, "Generalization and Decision Tree Induction: Efficient Classification in Data Mining", Database Systems Research Laboratory, School of Computing Science, Simon Fraser University, B.C., Canada V5A 1S6.

[4] D. A Kem, "Knowledge Discovery and Data Mining", Newport Beach, USA, 1997.

[5] Khin Kywe Kywe and Aye Aye Thein, "Evaluation of Classification Algorithms", University of Computer Studies, Mandalay, 2008.

[6] Y. Lee, "Classification: Decision Tree Induction", Department of Management Information Systems, College of Management, National Chiayi University, October 1, 2008.

[7] Myo Myo Than Naing and Tin Htar New, "Decision Making System Using Decision Tree Induction Algorithm", University of Computer Studies, Yangon.

[8] Win Mar Oo, "Biomedical Data Analysis for Diabetes using Hybrid Learning Method with Genetic Algorithms and Decision Tree", University of Computer Studies, Yangon.

[9] 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.

[10] 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.

[11] Tan, "Data Mining Classification: Alternative Techniques", Steinbach, Kumar.

[12] Yin Mu Yar Win and Tin Htar New,"Decision Support System for Diagnosis And Treatment Of Pulmonary Tubeculosis (TB) Disease", Computer University, Dawei.

[13] Phyo Min Zaw, "Analysis of Naïve Bayesian and Decision Tree Induction Using Statistical Information", University of Computer Studies, Mandalay, 2008.