IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

An Overview of Classification Algorithm in Data mining

S. Venkata Krishna Kumar[1], P. Kiruthika[2]

Associate Professor, Dept of Computer Science, PSG College of Arts and Science, Coimbatore, Tamilnadu, India[1]

Researcher Scholar & Assistant Professor, Department of Computer Science, Dr.N.G.P Arts and Science College, Coimbatore, Tamilnadu, India[2]

Abstract: Data mining is the process of extracting hidden analytical information from large databases using multiple algorithms and techniques. Classification technique can be solving several problems in different fields like medicine, industry, business, and science. Basically it involves finding rules that categorize the data into disjoint groups. In this paper we present the basic classification techniques. Several major kinds of classification method including decision tree, CHID, ID3, C4.5 and C5.5algorithms. The goal of this paper is to provide a review of different classification techniques in data mining.

Keywords: CHAID, CART, ID3, C4.5, C5.0, Hunt's algorithm.

I. INTRODUCTION

In the present scenario there is enormous amount of data being collected and stored in databases everywhere across the world. It is not difficult to find the repositories with Terabytes of data in organizations and research fields. There is huge collection of data present and it is very difficult to extract important pieces of information out of it and without automatic extraction methods this information is practically impossible to mine. Year after year many algorithms were created to extract important information from large sets of data. There are different methodologies to approach this problem like classification rule, association rule, clustering, etc.

II. CLASSIFICATION

Classification is the process of finding a model (or function) that describes and distinguishes 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. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). Classification and Prediction are two forms of data analysis that can be used to extract the models describing important data classes or to predict the future data trends. Classification predicts categorical labels and prediction models used continuous valued functions.

2.1 DECISION TREES

Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. Researchers have developed various decision tree algorithms over a period of time with enhancement in performance and ability to handle various types of data. A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test and each leaf node holds a class label. The topmost node in the tree is the root node. The following decision tree is for the concept buy_computer that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute. Each leaf node represents a class.

The benefits of having a decision tree are as follows

- It does not require any domain knowledge.
- It is easy to comprehend.
- The learning and classification steps of a decision tree are simple and fast.

2.2 CHID:

CHAID (CHi-squared Automatic Interaction Detector) is a fundamental decision tree learning algorithm. It was developed by Gordon V Kass in 1980. CHAID is easy to interpret, easy to handle and can be used for classification and detection of interaction between variables. CHID is an extension of the AID (Automatic Interaction Detector) and THAID (Theta Automatic Interaction Detector) procedures. It works on principal of adjusted significance testing. After detection of interaction between variables it selects the best attribute for splitting the node which made a child node as a collection of homogeneous values of the selected attribute. The method can handle missing values. It does not imply any pruning method.

2.3 CART:

Classification and regression tree (CART) proposed by Breiman et al. constructs binary trees which is also refer as Hierarchical Optimal Discriminate Analysis (HODA). CART is a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively. The word binary implies that a node in a decision tree can only be split into two groups. CART uses gini index as impurity measure for selecting attribute. The attribute with the largest

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

reduction in impurity is used for splitting the node's records. CART accepts data with numerical or categorical values and also handles missing attribute values. It uses cost-complexity and also generate regression trees.

2.4 ID3:

ID3 (Iterative Dichotomiser 3) decision tree algorithm is developed by Quinlan. In the decision tree method, information gain approach is generally used to determine suitable property for each node of a generated decision tree. Thus, we can select the attribute with the highest information gain (entropy reduction in the level of maximum) as the test attribute of current node. In this way, the information needed to classify the training sample subset obtained from later on partitioning will be the smallest. That is to say, the use of this property to partition the sample set contained in current node will make the mixture degree of different types for all generated sample subsets reduce to a minimum. Therefore, the use of such an information theory approach will effectively reduce the required dividing number of object classification.

2.5 C4.5:

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm.

The decision trees generated by C4.5 can be used for classification, and for this reason C4.5 is often referred to as a statistical classifier. C4.5 algorithm uses information gain as splitting criteria. It can accept data with categorical or numerical values. To handle continuous values it generates threshold and then divides attributes with values above the threshold and values equal to or below the threshold. C4.5algorithm can easily handle missing values. As missing attribute values are not utilized in gain calculations by C4.5.

2.6 C5.0/Sec 5:

C5.0 algorithm is an extension of C4.5 algorithm which is also extension of ID3. It is the classification algorithm which applies in big data set. It is better than C4.5 on the speed, memory and the efficiency. 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. C5.0 is easily handled the multi value attribute and missing attribute from data set.

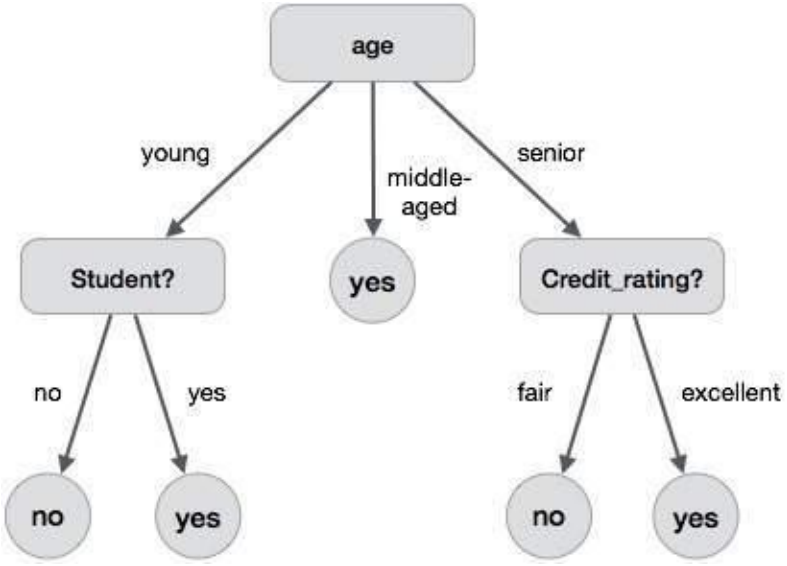

Table 1 Comparisons between different Decision Tree Algorithms

Algorithms	ID3	C4.5	C5.0	CART
Type of data	Categorical	Continuous and Categorical	Continuous and Categorical, dates, times, timestamps	continuous and nominal attributes data
Speed	Low	Faster than ID3	Highest	Average
Pruning	No	Pre-pruning	Pre-pruning	Post pruning
Boosting	Not supported	Not supported	Supported	Supported
Missing Values	Can't deal with	Can't deal with	Can deal with	Can deal with
Formula	Use information entropy and information Gain	Use split info and gain ratio	Same as C4.5	Use Gini diversity index

DOI 10.17148/IJARCCE.2015.41259

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

2.7 Hunt's Algorithm:

Hunt's algorithm generates a Decision tree by top-down or divides and conquers approach. The sample/row data contains more than one class, use an attribute test to split the data into smaller subsets. Hunt's algorithm maintains optimal split for every stage according to some threshold value as greedy fashion.

3. CONCLUSION

In this article provided an overview on six different algorithms decision tree, CHID, ID3, C4.5 and C5.5algorithms and their drawbacks which would be helpful to find the new solution for the problems found in these algorithms and also presents a comparison between different classification algorithms.

REFERENCES

1. Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques: concepts and techniques. Elsevier, 2011.
2. Fayyad, Usama M., et al. "Advances in knowledge discovery and data mining." (1996).
3. Michalski, Ryszard S., Jaime G. Carbonell, and Tom M. Mitchell, eds. Machine learning: An artificial intelligence approach. Springer Science & Business Media, 2013.
4. Kass, Gordon V. "An exploratory technique for investigating large quantities of categorical data." Applied statistics (1980): 119-127.
5. Breiman, Leo. "Bagging predictors." Machine learning 24.2 (1996): 123-140.
6. Quinlan, J. Ross. "Simplifying decision trees." International journal of man-machine studies 27.3 (1987): 221-234.
7. Quinlan, J. Ross. "Induction of decision trees." Machine learning 1.1 (1986): 81-106.
8. PANG, Su-lin, and Ji-zhang GONG. "C5. 0 classification algorithm and application on individual credit evaluation of banks." Systems Engineering-Theory & Practice 29.12 (2009): 94-104.

DOI 10.17148/IJARCCE.2015.41259