Decision Tree: Data Mining Techniques

Seema
Department of Computer Science Engineering
U.I.E.T (MDU), Rohtak, Haryana, India

Monika Rathi
Department of Computer Science Engineering
U.I.E.T (MDU), Rohtak, Haryana, India

Mamta
Department of Computer Science Engineering
U.I.E.T (MDU), Rohtak, Haryana, India

Abstract: Classification of data objects based on a predefined knowledge of objects is a data mining. There are many classification algorithms available but decision tree is the most commonly used. In this paper we will review the different decision tree techniques are explored with weakness and strengths in construction of decision tree in the field of data mining.

Keywords: Decision tree, tree pruning, data mining

I. INTRODUCTION

Decision tree is one of the classification technique used in decision support system and machine learning process. A decision tree is a predictive modeling technique that used in classification, clustering and predictive task. Decision tree uses a divide-conquer technique to split the problem search space into subsets. The most important feature of decision tree classifier is their ability to break down a complex decision making process into collection of simpler decision, thus providing solution which is easier to interpret.

II. DECISION TREE

A Decision tree is a tree where root and each internal node are labeled with question. The arcs emanating from each node represent each possible answer to the associated question. Each leaf node represents a prediction of a solution to the problem under consideration. The constructions of decision tree classifier don't require any domain knowledge or parameter setting and therefore is appropriate for exploratory knowledge discovery. Decision tree can handle high dimensional data. Their representation of acquired knowledge in tree free from is intuitive and generally easy to assimilate by humans.

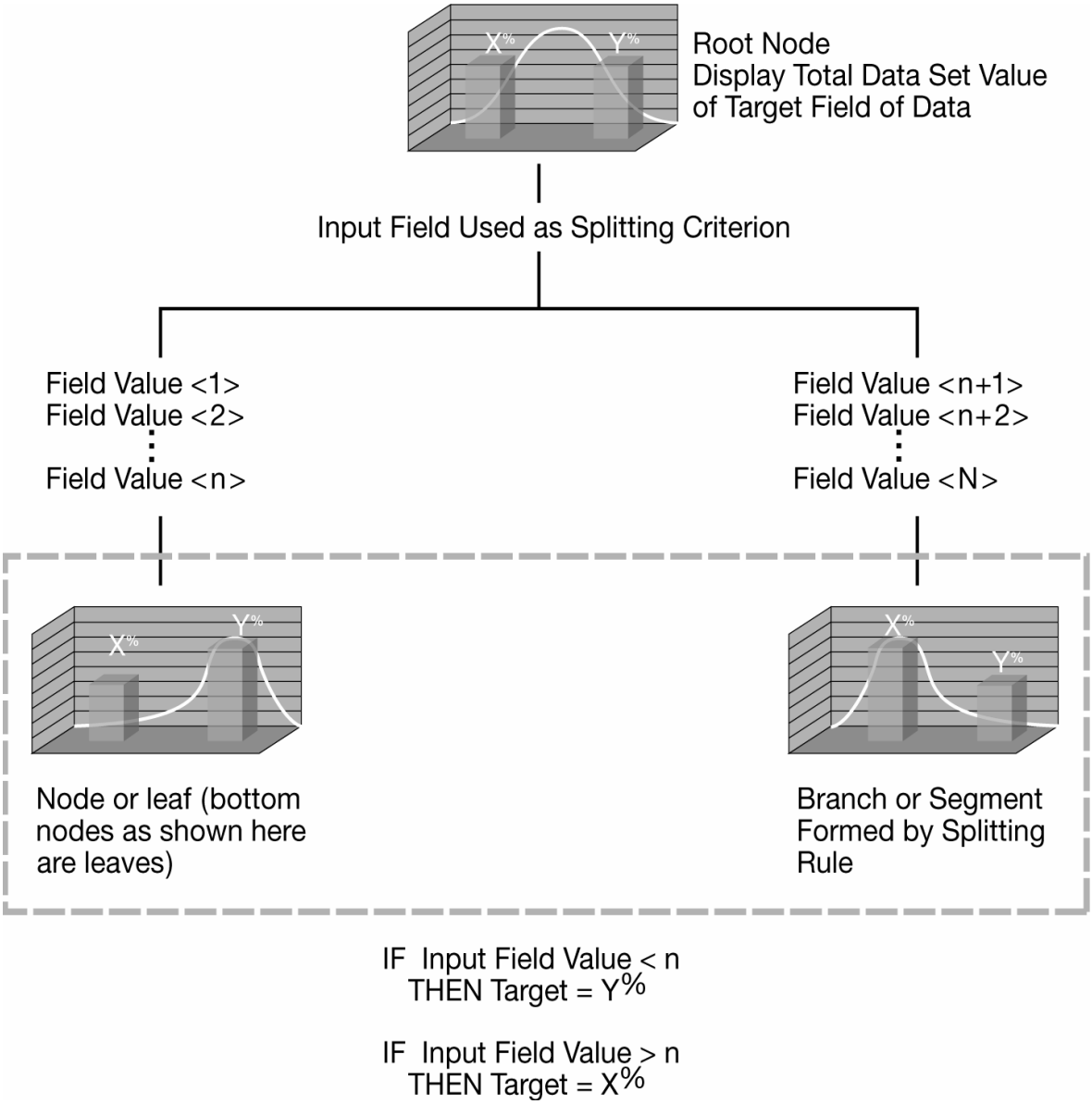

Figure 1.1: Illustration of the Decision Tree

The many benefits in data mining that decision trees offer:

• Self-explanatory and easy to follow when compacted

• Able to handle a variety of input data: nominal, numeric and textual

• Able to process datasets that may have errors or missing values

• High predictive performance for a relatively small computational effort

• Available in many data mining packages over a variety of platforms

• Useful for various tasks, such as classification, regression, clustering and feature selection.

The construction of the decision tree involves the following three main phases.

A. Construction Phase: The initial decision tree is constructed in this phased, based on the entire training data-set. It requires recursively partitioning the training set into two or more, sub-partition using a splitting criterion, until a stopping criteria is met.
B. Pruning Phase: the tree constructed in the previous phase may not result in the best possible set of rules due to over-fitting. The pruning phase removes some of the lower branches and nodes to improve its performance.
C. Processing the Pruned tree: to improve understandability.

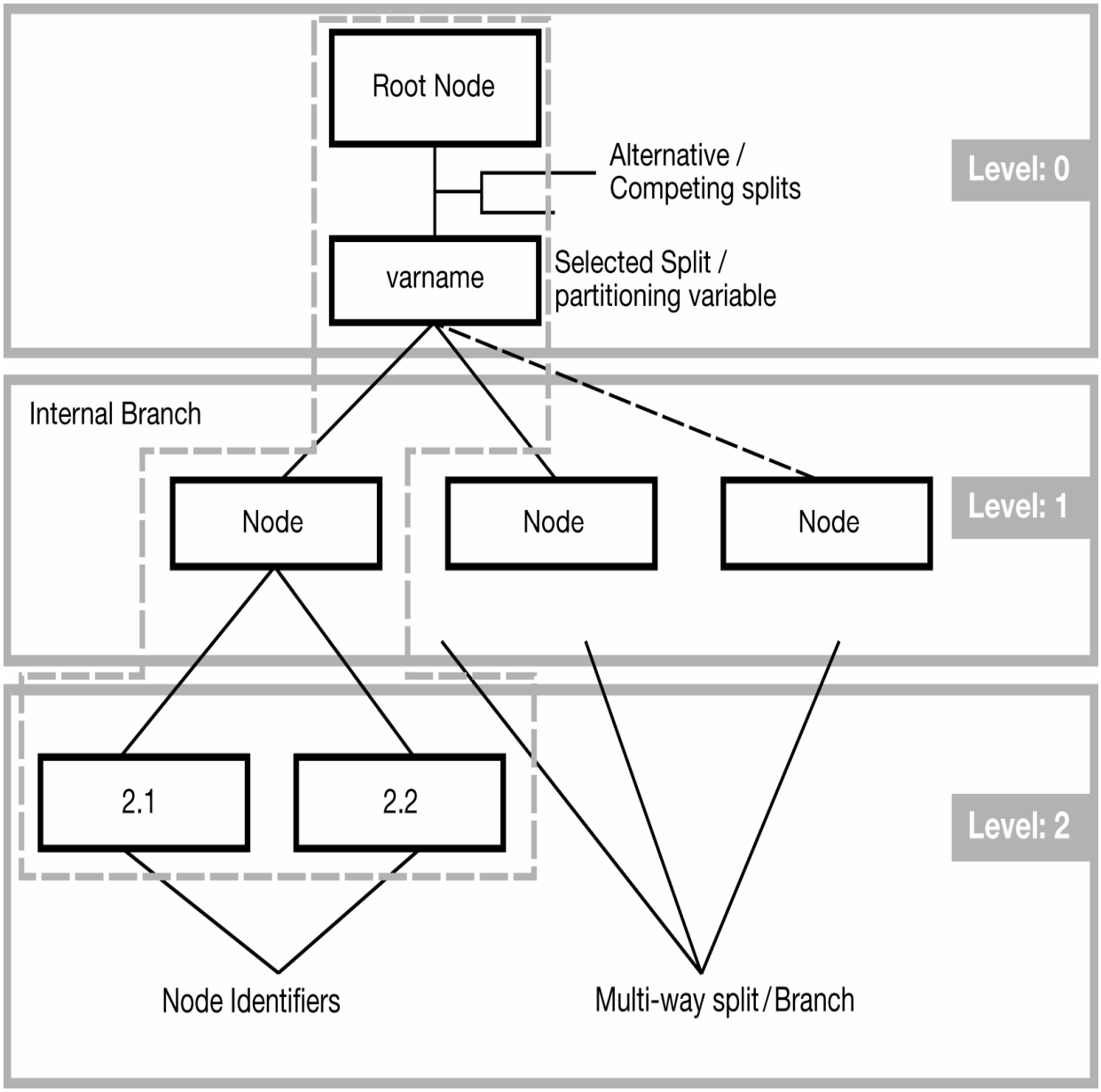

Figure 1.2: Illustration of Decision Tree Nomenclature

III. DECISION TREE ALGORITHM

A decision tree (DT) model is a computational model consisting of three parts:

1) A decision tree is defined.
2) An algorithm to create the tree.
3) An algorithm that applies the tree to data and solves the problem under consideration.

Algorithm:

Input: T// Decision Tree

D// Input Database

Output: M// Model Prediction

DT Proc algorithm:

// simplest algorithm to illustrate prediction technique using DT.

For each t ∈ D do

n = root node of T;

While n not leaf node do

Obtain answer to question on n applied to t;

Indentify arc from t, which contains correct answer;

n = node at end of this arc;

Make prediction for t based on label of n;

Strengths:

a) Decision tree are able to generate understandable rules:- The ability of decision tree to generate rules that can be translated into comprehensive English or SQL is the greatest strength of this technique.
b) Ability to clearly indicate best field:- Decision tree building algorithms put the field that does the best job of splitting the training records at the root node of the tree
c) Decision tree able to handle both continuous and categorical variables:- Continuous variable are equally easy to split by picking a number somewhere in their range of values. Categorical variables, which pose problems for neural networks for statistical techniques, come ready-made with their own splitting criteria: one branch for each category.

Weakness:

a) Decision tree are less appropriate for estimation tasks where the goal is to predict the value of continuous such as income, blood pressure, or interest rate.
b) Decision tree are also problematic for time-series data values a lot of effort is put into presenting the data in such a way that tends and sequential patterns are made visible.
c) Error-prone with too many classes:- Some decision tree algorithm can only deal with binary-valued target classes(yes/no, accept/reject);others are able to assign records to an arbitrary number of classes, but are error-prone when the number of training examples per class gets small.
d) Computationally expensive to train:- The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field must be sorted before its best split can be found. Pruning algorithm can also be expensive since many candidate sub trees must be formed and compared.

IV. ATTRIBUTE SELECTION MEASURES

An attribute selection measure is a heuristic for selecting the splitting criterion that best separates a given data partition, D of class-labeled training tuple into individual classes. It is also known as splitting rules because they determine how the tuple at a given node are to be split. The attribute having the best score for the measure is chosen as splitting attribute for a given tuple.

Three popular Attribute Selection Measures

a) Information Gain: - It is defined as the difference between the original information requirement and the new requirement.
   Gain (A) = Info (D) – $Info_A$ (D)
   ; Where D is data partition, be a training set of class-labeled tuple.
   ; Info (D) is Entropy of D; $Info_A$ (D) is expected information required to classify tuple from D based on the partitioning by A.
b) Gain Ratio: - It applies a kind of normalization to information gain using split info valued defined analogously with Info (D).
   Gain Ratio (A) = Gain (A)/ split Info (A); attribute with maximum gain ratio is selected as the splitting attribute.
c) Gini Index: - the Gini Index measures the impurities of D, data partition or set of training tuples. Gini (D) = $1 - \sum_{i=1}^{n} p_i^2$; p_i is probability that an arbitrary tuple in D belongs to class C_i,; $p_i = |c_i,D|/|D|$. The attribute that have maximum the reduction in impurity is selected as the splitting criteria.

V. TREE PRUNING

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of over fitting the data. There are two common approaches to tree pruning.

i. Pre-pruning:- In the pre-pruning approach, a tree is "pruned" by halting its construction early. Upon halting, the node becomes leaf. The leaf may hold most frequent class among the subsets tuples or the probability distribution of those tuples. When constructing a tree, attribute selection measures such as statistical significance, information gain, gini index and so on can be used to access the goodness of a split. If partitioning the tuple at node would result in a split that falls below a prespecified threshold, then further partitioning of a given subset is halted. There are difficulties, however in choosing an appropriate threshold. High threshold could result in oversimplified trees; whereas low thresholds could result in very little simplification.
ii. Post-pruning:-Post-pruning removes sub trees from a "fully grown" tree. A sub tree at a given node is pruned by removing its branches and replacing it with leaf. The leaf is labeled with the most frequent class among the sub tree being replaced. The "best" pruned tree is one that minimizes the encoding bits. This method adopts MDL (Minimum Description Length) principle. The basic idea is that the simplest solution is preferred. Alternatively, pre-pruning and post-pruning may be interleaved for a combined approach. Post-pruning requires more computation than pre-pruning, yet generally leads to a more reliable tree. Although pruned tree tends to be more compact than their unpruned counterparts, they may still be rather large and complex. Decision tree can suffer from repetition and replication.

VI. CONCLUSIONS

Decision tree is one of the classification techniques used in decision support system. With decision tree technique the training data set is recursively partitioned using depth-first or breadth-first greedy technique until each partition is pure or belongs to the same class/leaf node. In this study we focus on literature, decision tree are constructed in two phase: tree growth and tree pruning phase. In future we will perform experimental analysis of commonly used decision tree techniques.

REFERENCES

[1] Berry, Michael J.A, and Gordon Linoff. ”Data Mining Techniques for marketing, sales and customer support”. N.P.: John Wiley & sons, Inc. 1997.
[2] gautam.lis.illinois.edu/monkmiddleware/public/analytics/decisiontree.html
[3] http://support.sas.com/publishing/pubcat/chaps/57587.pdf
[4] Lior Rokach and Oded Maimon “DATA MINING WITH DECISION TREES Theory and Applications” a e-book.
[5] DATA MINING WITH DECISION TREES - Theory and Applications© World Scientific Publishing Co. Pte. Ltd. http://www.worldscibooks.com/compsci/6604.htm
[6] Han, J. and M. Kamber(2000). Data Mining: concepts and techniques, Morgan Kaufmann.
[7] Springer-Verlag Berlin Heidelberg (2011) Data Mining Concepts model and Techniques “Florin Gorunescu”
[8] Simarjit Kaur, Asmita “Data Mining:Decision tree techniques” (RTCMC-2012), ISBN 978-81-9234446-0-7.
[9] Seema, Monika Rathi. Anamika “Data Mining Analytics for Business Intelligence and decision Support” (RTCMC-2012), ISBN 978-81-923446-0-7.
[10] Fayyad, U., G. Piatetsky-Shapiro, et al., Eds. (1995). Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press.
[11] Fayyad, U. and R. Uthurusamy, Eds. (August 2002). Communications of the CACM – Evolving data mining into solutions for insights, ACM Press, New York, NY.
[12] Fine, S. and K. Scheinberg (2002). "Efficient SVM Training Using Low-Rank Kernel Representation." Journal of Machine Learning Research 2(2): 243-264.