Predicting the Value of a Target Attribute Using Data Mining

Ravijeet Singh Chauhan

Abstract— In this paper, the short coming of ID3's inclining to choose attributes with many values is discussed, and then a new decision tree algorithm which is improved version of ID3. Our proposed methodology uses greedy approach to select the best attribute. To do so the information gain is used. The attribute with highest information gain is selected. If information gain is not good then again divide attributes values into groups. These steps are done until we get good classification/misclassification ratio. The proposed algorithms classify the data sets more accurately and efficiently.

Index Terms— Classification, Decision tree, ID3, Prediction, Clustering.

I. INTRODUCTION

Decision tree learning, used in statistics, data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are **classification trees** or **regression trees**. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. This page deals with decision trees in data mining.

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. This process of Top-Down Induction of Decision Trees (TDIDT) is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data, but it is not the only strategy. In fact, some approaches have been developed recently allowing tree induction to be performed in a bottom-up fashion. [2]

Manuscript received May , 2013.

Ravijeet Singh Chauhan, Computer Science Department, Samrat Ashok Technological Institute, Vidisha, Madhya Pradesh, India.

II. RELATED WORK

Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology. Tree-based learning methods are widely used for machine learning and data mining applications. These methods have a long tradition and are commonly known since the works of [2, 3 and 4]. They are conceptually simple yet powerful. The most common way to build decision trees is by top down partitioning, starting with the full training set and recursively finding a univariate split that maximizes some local criterion (e.g. gain ratio) until the class distributions the leaf partitions are sufficiently pure Pessimistic Error Pruning [4] uses statistically motivated heuristics to determine this utility, while Reduced Error Pruning estimates it by testing the alternatives on separate independent pruning set. In a decision tree learner named NB Tree is introduced that has Naive Bayes classifiers as leaf nodes and uses a split criterion that is based directly on the performance of Naive Bayes classifiers in all first-level child nodes (evaluated by cross-validation) an extremely expensive procedure[8]. In [7, 11] a decision tree learner is described that computes new attributes as linear, quadratic or logistic discriminate functions of attributes at each node; these are then also passed down the tree. The leaf nodes are still basically majority classifiers, although the class probability distributions on the path from the root are taken into account. A recursive Bayesian classifier is introduced in [7]. Lots of improvement is already done on decision tree induction method for 100 % accuracy and many of them achieved the goal also but main problem on these improved methods is that they required lots of time and complex extracted rules. The main idea is to split the data recursively into partitions where the conditional independence assumption holds. A decision tree is a mapping from observations about an item to conclusions about its target value [9, 10, 11, 12 and 13]. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Another use of decision trees is as a descriptive means for calculating conditional probabilities. A decision tree (or tree diagram) is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility [14]. Decision tree Induction Method has been successfully used in expert systems in capturing knowledge. Decision tree induction Method is good for multiple attribute Data sets.

III PROBLEM DEFINITION

1. Analyze the database for the creation of an unsupervised model to identify the most significant parameters of affected area and to

predict the chances of hitting the disease using the supervised classifier model.
2. To build a model for classifying the inhabitants based on disease hit.

IV. PROPOSED SOLUTION

Classification is a form of data analysis that extracts models describing important data classes. These models also called as classifiers are used to predict categorical (discrete, unordered) class labels. This analysis can help us for better understanding of large data. Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, credit risk and medical diagnosis. Data Classification is a two-step process. They are: Learning Step and Classification Step

Learning Step: In this step classification model is constructed. A classifier is built describing a predetermined set of data classes or concepts. In learning step or training phase, where classification algorithm builds the classifier by analyzing or "learning from" a training set made up of database tuples and their associated class labels.

This step is also known as supervised learning as the class label of each training tuple is provided. This learning of the classifier is "supervised" by telling to which class each training tuple belongs. In unsupervised learning or clustering, the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance.

Classification Step: In this step, the model is used to predict class labels for given data and it is used for classification. First, the predictive accuracy of the classifier is estimated. To measure the classifiers accuracy, if we use the training set it would be optimistic, because the classifier tends to over fit the data i.e., during learning it may incorporate some particular anomalies of the training data that are not present in the general data set. Therefore, a test set is used, made up of the test tuples and their associated class labels. They are independent of the training tuples, from which the classifier cannot be constructed. The accuracy of a classifier on a given test set is the percentage of test tuples that are correctly classified by the classifier. The associated class label of each test tuple is compared with the learned classifier's class prediction for the tuple. If the accuracy of the model or classifier is considered acceptable, the model can be used to classify future data tuples or objects for which the class label is not known.

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 topmost node in a tree is the root node.

Given a tuple K, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class predicate for that tuple. Decision trees are easily converted to classification rules. The construction of decision does not require any domain knowledge or parameter setting. It can handle high dimension data. The learning and classification steps are simple and fast. It has good accuracy. Decision tree Induction algorithm can be used in many applications like medicine, manufacturing and production etc.

V. PROPOSED METHOD:

Input:

1. A Training Data Set With Class Labels
2. List Of Attributes
3. Feature Selection Criteria. It Is Used For Splitting

Output:
A Decision Tree
Procedure:
Step 1: If All The Tuples Of Dataset Contains Yes Then Create A Yes Node And Stop
Else If All The Tuples Of Dataset Contains No Then Create A No Node And Stop
Else Select A Feature And Create A Decision Node
Step 2: Split The Data Set D Into Smaller Data Sets D1, D2… Dn According To Step 1 Criteria
Step 3: Apply The Algorithm Recursively On Each Data Set D1, D2… Dn.

VI. ATTRIBUTE SELECTION

Our proposed methodology yeses greedy approach to select the best attribute. To do so the information gain is used. The attribute with highest information gain is selected. Entropy measures the amount of information in an attribute. Given a collection D of c outcomes

Entropy (D) = D [-p(J) log2 p(J)]

Where; p (J) is the proportion of D belonging to class J. D is over c. Log2 is log base 2. Here; D is not an attribute but the entire sample set.

If the Entropy is 0 then all members of D belong to the same class i.e., the data is perfectly classified. If the Entropy is 1 then all members of D are totally random. The range of entropy lies between 0 to 1.

Gain (D, A) is information gain of example set D on attribute A is defined as

Gain (D, A) = Entropy (D) - S ((|Da| / |D|) *Entropy (Da))

Where: S is each value v of all possible values of attribute A
Da= subset of D for which attribute A has value a.
| Da| = number of elements in Da
|D| = number of elements in D.

VII. CONCLUSION

In this paper we have presented a more accurate algorithm for classification. Our proposed methodology uses greedy approach to select the best attribute. To do so the information gain is used. The attribute with highest information gain is selected. In this way accuracy has improved.

REFERENCES

[1] Singh Vijendra. Efficient Clustering For High Dimensional Data: Subspace Based Clustering and Density Based Clustering, *Information Technology Journal; 2011*, 10(6), pp. 1092-1105.

[2] D Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J."Classification and Regression Trees". Wadsworth International Group. Belmont, CA: The Wadsworth Statistics/Probability Series1984.

[3] Quinlan, J. R. "Induction of Decision Trees". *Machine Learning;* 1986,pp. 81-106.
[4] Quinlan, J. R. Simplifying "Decision Trees. International Journal of Man-Machine Studies" ;1987, 27:pp. 221-234.
[5] Gama, J. and Brazdil, P. "*Linear Tree. Intelligent Data*Analysis",1999,.3(1): pp. 1-22.
[6] Langley, P. "*Induction of Recursive Bayesian Classifiers*". In BrazdilP.B. (ed.), Machine Learning: ECML-93;1993, pp. 153-164. Springer,Berlin/Heidelberg~lew York/Tokyo.
[7] Witten, I. & Frank, E,"*Data Mining: Practical machine learning toolsand* techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.ch. 3,4, pp 45-100.
[8] Yang, Y., Webb, G. "*On Why Discretization Works for Naive-Bayes*Classifiers*"*, Lecture Notes in Computer Science, vol. 2003, pp. 440 –452.
[9] H. Zantema and H. L. Bodlaender, "Finding Small Equivalent Decision Trees is Hard", *International Journal of Foundations of Computer Science*; 2000, 11(2):343-354.
[10] Huang Ming, Niu Wenying and Liang Xu , "An improved Decision Tree classification algorithm based on ID3 and the application in score analysis", *Software Technol. Inst*., Dalian Jiao Tong Univ., Dalian, China, June 2009.
[11] Chai Rui-min and Wang Miao, "A more efficient classification scheme for ID3",Sch. of Electron. & Inf. Eng., Liaoning Tech. Univ., Huludao, China; 2010,Version1, pp. 329-345.
[12] Iu Yuxun and Xie Niuniu "Improved ID3 algorithm",*Coll. of Inf. Sci. & Eng*., Henan Univ. of Technol., Zhengzhou, China;2010,pp. ;465-573.
[13] Chen Jin, Luo De-lin and Mu Fen-xiang," An im pr oved ID3 decision tree algorithm",Sch. of Inf. Sci. & Technol., Xiamen Univ., Xiamen, China, page; 2009, pp. 127-134.
[14] Jiawei Han and Micheline Kamber**, "***Data Mining: Concepts and Techniques*", 2nd edition, Morgan Kaufmann, 2006, ch-3, pp. 102-130.

Ravijeet Singh Chauhan in an M.Tech. (Computer Science & Engineering) student of Samrat Ashok Technological Institute, Vidisha, Madhya Pradesh, INDIA. His research interest includes Data Mining, Clustering, Soft Computing and Cloud Computing. He is the member of ISTE India student chapter.