



A Novel Decision Tree Based Classifier for Accurate Heart Disease Prediction

Shikha Sharma, Pallavi Jain

Shri Vaishnav institute of Science & Technology,
Department Of Computer Science Engg.
SVITS, Indore, M.P., India

Abstract: In this paper, we have also presented an overview of existing data classification algorithms. Data Classification is a very popular and computationally expensive task. We have presented a comprehensive survey of the modern data classification cum disease prediction techniques. Most of these data classification techniques are based on the concept of decision trees. Many researchers have worked on the disease prediction systems using the data mining techniques. Some of the systems are for predicting a single disease and some for the predicting the multiple diseases. Still there is scope to improve the efficiency of the disease prediction. In this paper, we are presenting a survey of some most popular classification techniques extensively used for the disease prediction. They generally use the concept of classification using the decision tree.

Keywords:- Data mining, ID3 algorithm, Decision Tree, Classification Algorithm

I. INTRODUCTION

Data mining concepts:-Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful understandable patterns in large databases. The patterns must be actionable so that they may be used in an enterprise's decision making process. It is usually used by business intelligence organization and financial analysts, but it is increasingly used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods [6].

A typical example for a data mining scenario may be "In the context of a super market, if a mining analysis observes that people who buy pen tend to buy pencil too, then for better business results the seller can place pens and pencils together."

Data mining strategies can be grouped as follows:-

Classification- Here the given data instance has to be classified into one of the target classes which are already known or defined [19, 20]. One of the examples can be whether a customer has to be classified as a trustworthy customer or a defaulter in a credit card transaction data base, given his various demographic and previous purchase characteristics.

- Estimation- Like classification, the purpose of an estimation model is to determine a value for an unknown output attribute. However, unlike classification, the output attribute for an estimation problem are numerical rather than categorical. An example can be "Estimate the salary of an individual who owns a sports car?"

- Prediction- It is not easy to differentiate prediction from classification or estimation. The only difference is that rather than determining the current behavior, the predictive model predicts a future outcome. The output attribute can be categorical or numeric. An example can be "Predict next week's closing price for the Dow Jones Industrial Average". [53, 54] explains the construction of a decision tree and its predictive applications.

- Association rule mining - Here interesting hidden rules called association rules in a large transactional data base is mined out. For e.g. the rule {milk, butter->biscuit} provides the information that whenever milk and butter are purchased together biscuit is also purchased, such that these items can be placed together for sales to increase the overall sales of each of the items [46].

- Clustering- Clustering is a special type of classification in which the target classes are unknown. For e.g. given 100 customers they have to be classified based on certain similarity criteria and it is not preconceived which are those classes to which the customers should finally be grouped into.

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. An example is shown on the right. 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]

(A) Decision tree advantages:-

Amongst other data mining methods, decision trees have various advantages:

- (a) **Simple to understand and interpret.** People are able to understand decision tree models after a brief explanation.
- (b) **Requires little data preparation.** Other techniques often require data normalization, dummy need to be created and blank values to be removed.
- (c) **Able to handle both numerical and categorical data.** Other techniques are usually specialized in analyzing datasets that have only one type of variable. Ex: relation rules can be used only with nominal variables while neural networks can be used only with numerical variables.
- (d) **Uses a white box model.** If a given situation is observable in a model the explanation for the condition is easily explained by Boolean logic. An example of a black box model is an artificial neural network since the explanation for the results is difficult to understand.
- (e) **Possible to validate a model using statistical tests.** That makes it possible to account for the reliability of the model.
- (f) **Robust.** Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.
- (g) **Performs well with large data in a short time.** Large amounts of data can be analyzed using standard computing resource.

II. LITERATURE SURVEY

Early methods of identifying patterns in data include Bays' theorem and regression analysis. The proliferation, ubiquity and increasing power of computer technology has increased data collection and storage. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, clustering [1], genetic algorithms, decision trees and support vector machines. Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns. [2,3,13]. Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can be easily converted to classification rules. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast with good accuracy. 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 unvaried 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 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. PROPOSED APPROACH

In decision tree classifiers, the criteria used for the attribute selection is as follows: First information gain of each attribute is computed then the attribute having maximum information gain is chosen.

This means that an attribute with maximum values is chosen for splitting the tree. But in most of the cases, it is not necessary that an attribute with maximum values will be the best. Also ID3 algorithm uses the concept of information gain for selecting an attribute. The information gain is based on the concept of the probability. Probability based method is suitable for stochastic problems. But it cannot be the common criteria for attribute selection.

For solving this problem, we propose a more accurate decision tree based classifier. The proposed solution will use new attribute selection criteria. It will give more weight to attributes with less value but more importance. Also it will reduce the weight of attribute with more values and less importance.

ATTRIBUTE SELECTION: The proposed methodology uses a modified gain based greedy approach to select the best attribute, which will be used the attribute with highest information gain. But we have modified the formulae of information gain for partitioning the training data set into smaller partitions. Similar to ID3, The proposed algorithm also chosen. The modified formulae contain utility value of each attribute. In this the selection criteria has improved, which ultimately will result is more classification and prediction. Entropy measures the amount of information in an attribute [5].

IV. CONCLUSION

In this paper, we also surveyed the existing data classification techniques. We restricted ourselves to the classic classification problem. The most of the techniques discussed in this survey paper are decision tree based classification techniques. It is observed that there is a scope to improve the accuracy of these methods. In next paper, we will present a more accurate algorithm for classification.

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 Brahman, L., Friedman, J. H., Olsen, R. A., and Stone, C. J. "Classification and Regression Trees". Wadsworth International Group. Belmont, CA: The Wadsworth Statistics/Probability Series 1984.
- [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 Brazdil P.B. (ed.), *Machine Learning: ECML-93*; 1993, pp. 153-164. Springer, Berlin/Heidelberg-New York/Tokyo.
- [7] Witten, I. & Frank, E., "Data Mining: Practical machine learning tool sand 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-Bays 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, Version 1, 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 improved 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.