A Review of a Novel Decision Tree Based Classifier for Accurate Multi Disease Prediction

Sagar S. Mane1 Dhaval Patel2
1P.G Scholar, 2Assistant Professor
1,2Department of Computer Engineering
Hasmukh Goswami College of Engineering, Vehlal

Abstract—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 synopsis, we are presenting a novel classification based disease prediction system. It uses the concept of classification using the decision tree. Our proposed technique uses greedy approach to select the best attribute for construction of the decision tree. The modified information gain is used. The attribute with highest information gain is selected as the root of the tree. The experimental results have shown that the proposed algorithms classify the data sets more accurately and efficiently. In this synopsis, we have also presented an overview of existing data classification algorithms. These algorithms are described more or less on their own. Data Classification is a very popular and computationally expensive task. We have also elaborated the fundamentals of data classification. From a large number of available algorithms that have been developed we will compare the most important ones.

Key words: data mining techniques, disease prediction, decision tree

I. INTRODUCTION

A. Decision Tree:

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. 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.

B. Data Mining:

Data mining techniques have been applied to medical services in several areas, including prediction of effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data. 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. To build a model for classifying the inhabitants based on disease hit.

C. Disease Prediction:

The main intention of this these are to presenting a novel classification based disease prediction system. It uses the concept of classification using the decision tree. Our proposed technique uses greedy approach to select the best attribute for construction of the decision tree. The modified information gain is used. The attribute with highest information gain is selected as the root of the tree. The proposed algorithms classify the data sets more accurately and efficiently. In this synopsis, we have also presented an overview of existing data classification algorithms. These algorithms are described more or less on their own. Data Classification is a very popular and computationally expensive task. We have also elaborated the fundamentals of data classification. From a large number of available algorithms that have been developed we will compare the most important ones.

D. Various Application of Data Mining:

Data mining can be applied to following different fields.

E. Health Care:

Drug development–to help uncover less expensive but equally effective drug treatments.

- Medical diagnostics—imaging, real-time monitoring (e.g., predicting women at high risk for emergency C-section).
- Insurance claims analysis—identify customers likely to buy new policies; define behavior patterns of risky customers.

F. Business and Finance:

- Banks—to detect which customers are using which products so they can offer the right mix of products and services to better meet customer needs—cross sell and up sell.
- Credit card companies—to assist in mailing promotional materials to people who are most likely to respond.
- Lenders—to determine which applicants are most likely to default on a loan.

G. Sports and Gambling:

Sports teams—to analyze data to determine favorable player match ups and call the best plays.
II. METHODOLOGY

A. Decision-Tree Method:
The decision tree induction algorithm has been used broadly for several years.

It is an approximation discrete Function method and can yield lots of useful expressions. It is one of the most important methods for classification.

This algorithm’s terms follow the “tree” metaphor. It has a root, which is the first split point of the data attribute for building a decision tree. It also has leaves, so that every path from root to leaf will form a rule that is easily understood. Since the decision tree is built by given data, the data value and character will be more important.

B. Induction Of Decision Trees Algorithm, ID3:
ID3 is a typical decision-tree algorithm. It introduces information entropy as the splitting attribute’s choosing measure. It trains a tree from root to leaf, a top-down sequence. Each path from that form is a decision rule.

C. Definition 1:
D is defined as a training data set whose attributes are divided into two parts: non-target and target. The non-target attribute is named as Q (Q1,…,Qm), where each attribute Qi(1≤i≤m) takes Ki values {1, ..., ik ai a}. The target attribute (usually just one attribute) is named as C. Suppose it has l values; thus we get l classes; C={C1, …,Cl}. Let Djbe a subset in D whose class is Cjand |D| be the number of elements in D. The information entropy of data set Dis defined as:

\[ E(D) = \sum_{j=1}^{l} \left( P_j \log \frac{1}{P_j} \right) \]

Where \( P_j \) is the proportion of D belonging to class j

\[ P_j = \frac{|D_j|}{|D|} \]

D. Definition 2:
The measure of the impurity in a collection of training examples is defined as information gain, Gain (D, Qi), of attribute Qi:

\[ \text{Gain}(D, Q_i) = E(D) - \sum_{j=1}^{l} \left( P_j \cdot E(D_j) \right) \]

Where, \( D_j \) is the obtained jth subset which is divided by attribute Qi on D, and

\[ P_j = \frac{|D_j|}{\sum_{i=1}^{k} |D_i|} \]

III. PROPOSED METHOD

A. Flow Chart of the Proposed Algorithm:

Fig. 1: Flow chart of the proposed algorithm

IV. LITERATURE REVIEW

A. T Zung-I Tan, Gang Zheng:
The data They study is collected from patients with coronary heart disease. In this paper system-reconstruction method to weight it. they use decision tree algorithms, such as induction of decision , (ID3), classification and regression tree (C4.5), classification and regression tree (CART), Chi-square automatic interaction detector(CHAID), and exhausted CHAID. They use the results to compare the correction rate, leaf number, and tree depth of different decision-tree algorithms. According to the experiments, They know that weighted data can improve the correction rate of coronary heart disease data but has little effect on the tree depth and leaf number.

The decision-tree algorithm is one of the most effective classification methods. The data used in the paper were collected directly from clinical diagnoses. From the data They get the conclusion that data weighted by the system-reconstruction method can get a higher correction rate but will have little effect on the leaf number and tree depth of the decision tree. They will study principle component analysis, rough set, feature selection to reduce the attributes of the coronary heart disease data.
searchers are using other. In some papers it is attack with the data processing; traditional statistical methods a dramatic difference in the range of techniques is presents. They use the weight motivation for this paper is to classify the heart disease with values to diagnose heart based on an explicit similarity measure with biomedical test propagation algorithm and k selection by using multilayer perceptron with back D.

B. HninWintKhaing:

Firstly heart disease database is clustered using the K-means clustering algorithm, which will extract the data relevant to heart attack from the database. This approach allows mastering the number of fragments through its k parameter. Subsequently the frequent patterns are mined from the extracted data, relevant to heart disease, using the MAFIA (Maximal Frequent Item set Algorithm) algorithm. The machine learning algorithm is trained with the selected significant patterns for the effective prediction of heart attack. They have employed the ID3 algorithm as the training algorithm to show level of heart attack with the decision tree. The results showed that the designed prediction system is capable of predicting the heart attack effectively. The heart disease database is preprocessed successfully by removing duplicate records and supplying missing values as shown in table I. There fined heart disease data set, resultant from preprocessing, is then clustered using K-means algorithm with K value . One cluster consists of the data relevant to the heart diseases and the other contains the remaining.

C. Hianchyekoh and Gerald Tan:

The following methodology for data mining: business understanding understanding and preparation, modeling, evaluation, and deployment. The modeling stage is the actual data analysis. Most data mining software include online analytical processing; traditional statistical methods, such as cluster analysis, discriminant analysis and regression analysis; and non-traditional statistical analysis, such as neural networks, decision trees, link analysis and association analysis. This extensive range of techniques is not surprising in light of the fact that data mining has been viewed as the offspring of three different disciplines, namely database management, statistics, and computer science, including artificial intelligence and machine learning. The evaluation stage enables the comparison of models.

D. KittipolWisaeng:–

In this paper, the prediction of heart disease based on feature selection by using multilayer perceptron with back-propagation algorithm and k-nearest neighbor algorithm based on an explicit similarity measure with biomedical test values to diagnose heart disease is presented. The main motivation for this paper is to classify the heart disease with reduced number of attributes. They use the weight information by a multilayer perceptron to determine the attributes which reduces the number of attributes which is needed to be taken from original datasets (13 attribute is reduced to 8 attributes). Afterward, they used k-nearest neighbor algorithm to predict the diagnosis of heart disease after the reduction of a number of attributes. The accuracy differs between 13 attributes and 8 attributes in testing data set is 93% and 90%, respectively. The experimental results show that our propose classification help in the best prediction of heart disease which even helps doctors in their diagnosis decisions.

E. Hlaudi Daniel Masethe:

The heart disease accounts to be the leading cause of death worldwide. It is difficult for medical practitioners to predict the heart attack as it is a complex task that requires experience and knowledge. The health sector today contains hidden information that can be important in making decisions. Data mining algorithms such as J48, Naïve Bayes, REPTREE, CART, and Bayes Net are applied in this research for predicting heart attacks. The research result shows prediction accuracy of 99%. Data mining enable the health sector to predict patterns in the dataset.

The research undertook an experiment on application of various data mining algorithms to predict the heart attacks and to compare the best method of prediction. The research results do not presents a dramatic difference in the prediction when using different classification algorithms in data mining. The experiment can serve as an important tool for physicians to predict risky cases in the practice and advise accordingly. The model from the classification will be able to answer more complex queries in the prediction of heart attack diseases. The predictive accuracy determined by J48, REPTREE and SIMPLE CART algorithms suggests that parameters used are reliable indicators to predict.

F. Beantkaur, Williamjeet Singh:

The objective of our work is to provide a study of different data mining techniques that can be employed in automated heart disease prediction systems. Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient and effective heart disease diagnosis. The analysis shows that different technologies are used in all the papers with taking different number of attributes. So, different technologies used shown the different accuracy to each other. In some papers it is shown that neural networks given the accuracy of 100% in prediction of heart disease. On the other hand, this is also given that Decision Tree. Has also performed well with 99.62% accuracy by using 15 attributes. So, different technologies used shown the different accuracy depends upon number of attributes taken and tool used for implementation. Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amounts of data, researchers are using data mining techniques in the diagnosis of heart disease. Although applying data mining techniques to help health care professionals in the diagnosis of heart disease is having some success, the use of data mining techniques to identify a suitable treatment for heart disease patients has received less attention.
G. Sivagowry, Dr. Durairaj:
Existing literature shows that Classification task in Data Mining plays a vital role in heart disease prediction when compared with Clustering, Association Rule and Regression. In Classification Decision Tree outperforms in some cases, where Neural Network and Naïve Bayes outperforms in some other cases. Each technique has its own merits and demerits. When combined with each other or with Fuzzy logic, Heart disease prediction with Data Mining techniques will become most successful with less number of attributes. Text mining the medical data is another extension found in predicting the health care data.

H. Aswathy Wilson, Gloria Wilson:
In this paper They are using the Data mining techniques like K Means and Weighted Association rule for the elimination of manual tasks and easier extraction of data directly from electronic records, transferring onto secure electronic system of medical records which will save lives and decrease the cost of the healthcare services. K-means clustering is a usually used data clustering for unsupervised learning tasks. Decision tree is used to prediction process. The WAC has been used to get the significant rule instead of flooded with insignificant relation and the Apriori algorithm is used to find out the frequent item set from the patient database. The weighted association classifier is a idea of weighted association rule for classification. Weighted Association Rule Mining uses Weighted Support and Confidence Framework to extract Association rule from data warehouse. Classification rule mining takes a training data set and generates a small set of rules to organize future data. The k-means clustering is the technique to cluster the attributes from the patient record. The decision tree with K-means clustering can enhance the classifier’s performance in diagnosing heart disease. The initial centroid selection technique among the five K-means clustering techniques is used here because it can provide better performance over heart disease prediction. The experiments shows that K-means with decision tree technique make the system more accurate and efficient when compared to the weighted association rule with Apriori algorithm.

I. V. Manikanta, S. Latha:
The data classification is based on MAFIA algorithms which result in accuracy, the data is evaluated using entropy based cross validations and partition techniques and the results are compared. Here using the C4.5 algorithm as the training algorithm to show rank of heart attack with the decision tree. Finally, the heart disease database is clustered using the K-means clustering algorithm, which will remove the data applicable to heart attack from the database. The results showed that the medicinal prescription and designed prediction system is capable of prophesying the heart attack successfully. In this paper They have presented an efficient approach for fragmenting and extracting substantial forms from the heart attack data warehouses for the efficient prediction of heart attack.

<table>
<thead>
<tr>
<th>Algorithm used</th>
<th>Accuracy</th>
<th>Time taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>52.33%</td>
<td>609ms</td>
</tr>
<tr>
<td>Decision List</td>
<td>52%</td>
<td>719ms</td>
</tr>
<tr>
<td>KNN</td>
<td>45.67%</td>
<td>1000ms</td>
</tr>
</tbody>
</table>

J. B. Venkatalakshmi:
This research paper intends to use data mining Classification Modeling Techniques, namely, Decision Trees, Naïve Bayes and Neural Network, along with weighted association Apriori algorithm and MAFIA algorithm in Heart Disease Prediction. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting heart disease. A total of 909 records were obtained from the Cleveland Heart Disease database. It has been observed during the analysis that Naïve Bayes appears to be most effective as it has the highest percentage of correct predictions (86.53%) for patients with heart disease, followed by Neural Network (85.53%) and Decision Trees. Decision Trees, however, appears to be most effective in case of predicting patients with no heart disease.

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>86.53%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>89%</td>
</tr>
<tr>
<td>ANN</td>
<td>85.53%</td>
</tr>
</tbody>
</table>

V. Conclusion
We propose a more accurate decision tree based classifier. Our proposed solution will use a new attribute selection criteria. It will give more weight to attributes with less values but more importance. Also it will reduce the weight of attribute with more values and less importance.
We have also elaborated the fundamentals of data classification. From a large number of available algorithms that have been developed we will compare the most important ones. We have also investigated the strengths and weaknesses of some modern decision tree data classification algorithms.
VI. REFERENCES


