International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume: 3 | Issue: 3 | Mar-Apr 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 - 6470

# Decision Tree Models for Medical Diagnosis

**Aung Nway Oo, Thin Naing**

University of Information Technology, Myanmar

***How to cite this paper***: Aung Nway Oo | Thin Naing "Decision Tree Models for Medical Diagnosis" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-3, April 2019, pp.1697-1699, URL: https://www.ijtsrd.com/papers/ijtsrd23510.pdf

IJTSRD23510

**ABSTRACT**

Data mining techniques are rapidly developed for many applications. In recent year, Data mining in healthcare is an emerging field research and development of intelligent medical diagnosis system. Classification is the major research topic in data mining. Decision trees are popular methods for classification. In this paper many decision tree classifiers are used for diagnosis of medical datasets. AD Tree, J48, NB Tree, Random Tree and Random Forest algorithms are used for analysis of medical dataset. Heart disease dataset, Diabetes dataset and Hepatitis disorder dataset are used to test the decision tree models.

***KEYWORDS:*** *Data mining, Classification, Decision tree*

## 1. INTRODUCTION

At present, Data mining has had a significant impact on the information industry, due to the wide availability of huge datasets, which are stored in databases of various types. Data mining is presence place into apply and considered for databases, along with relational databases, object relational databases and object oriented databases, data warehouses, transactional databases, unstructured and partially structured repositories, spatial databases, multimedia databases, time-series databases and textual databases [6].

Different methods of data mining use different purpose of uses. The methods contribute some of its own advantages and disadvantages. In data mining, classification plays a crucial role in order to analyses the supervised information. Classification is a supervised learning method and its objectives are predefined [1]. The role of classification is important in real world applications including medical field. Decision trees play a vital role in the field of medical diagnosis to diagnose the problem of a patient. In this paper various decision tree classifiers are used to analyses the medical datasets.

The rest of the paper is organized as follows. Section 2 provides the related work and section 3 presents the overview of classification algorithms. The experimental results are discussed in section 4. Finally, conclusion of this study was provided in section 5.

## 2. RELATED WORKS

Many papers are proposed the performance evaluation of decision tree classifiers. G. Sujatha [7] presented the performance of decision tree induction algorithms on tumor medical data sets in terms of Accuracy and time complexities are analyzed. In the paper of T.Karthikeyan [8] mainly deals with various classification algorithms namely, Bayes. NaiveBayes, Bayes. BayesNet, Bayes. NaiveBayes Updatable, J48, Random forest, and Multi Layer Perceptron. It analyzes the hepatitis patients from the UC Irvine machine learning repository. T. Swapna [9] proposed the analysis of classification algorithms for Parkinson's disease classification. In this paper a comparative study on different classification methods is carried out to this dataset and the accuracy analysis to come up with the best classification rule. In the research work of [10], training and test diabetic data sets are used to predict the diabetic mellitus using various classification techniques. And compared the data by applying the material to the conventional techniques of Bayesian statistical classification, J48 Decision tree and SVM to form a prediction model. E. Venkatesan [1] proposed the performance analysis of decision Tree algorithms for breast cancer classification. The paper of Anju Jain [11] reviewed the use of machine learning algorithms like decision tree, support vector machine, random forest, evolutionary algorithms and swarm intelligence for accurate medical diagnosis. Anju Jain etal. [11] proposed medical diagnosis system using machine learning techniques. In recent year, various paper are proposed for medical diagnosis using data mining and machine learning methods.

## 3. DECISION TREE CLASSIFIERS

Decision tree learning uses a decision tree to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees

Yoav Freund and Llew Mason introduced Alternating Decision Tree (ADTree), a machine learning method for classification, which generalizes decision tree and data structure. This tree predicts the nodes in the leaves and roots. The classification is done by traversing through all paths for all decision nodes. The binary classification trees are distinct and the AD Tree is different among that [1].

J48 is an extension of ID3 algorithm. J48 is a tree based learning approach. It is developed by Ross Quinlan which is based on iterative dichtomiser (ID3) algorithm. J48 uses divide-and-conquer algorithm to split a root node into a subset of two partitions till leaf node (target node) occur in tree. Given a set T of total instances the following steps are used to construct the tree structure [2].

NB-Tree is a hybrid algorithm with Decision Tree and Naïve-Bayes. In this algorithm the basic concept of recursive partitioning of the schemes remains the same but here the difference is that the leaf nodes are naïve Bayes categorizers and will not have nodes predicting a single class [3].

Random Tree (RT) is an efficient algorithm for constructing a tree with K random features at each node. Random tree is a tree which drawn at random from a set of possible trees. Random trees can be generated efficiently and the combination of large sets of random trees generally leads to accurate models [4].

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [5].

## 4. EXPERIMENTAL RESULTS

Heart disease dataset, diabetes dataset and liver disease dataset from UCI machining learning repository are used for classification task. 66 % of dataset is used for training and remaining 34 % is used for testing.

Heart disease dataset contains 270 observations and 2 classes: the presence and absence of heart disease. There are 150 patient records without suffer heart disease and 120 records for patient with heart disease. The results of classifiers are showed in table 1.

Table1. Prediction results of heart disease dataset

| | ADTree | J48 | NBTree | Random Forest | Random Tree |
|---|---|---|---|---|---|
| Correctly Classified Instances | 77 | 70 | 73 | 74 | 70 |
| Incorrectly Classified Instances | 15 | 22 | 19 | 18 | 22 |
| Kappa statistic | 0.6758 | 0.5317 | 0.5927 | 0.6125 | 0.5278 |
| Accuracy | 83.7% | 76.1% | 79.3% | 80.4% | 76.1% |

Diabetes dataset contains 768 instances and 2 classes: the presence and absence of diabetes. There are 500 patient records without suffer diabetes and 268 records for patient with diabete. The results of classifiers are showed in table 2.

Table2. Prediction results of diabetes dataset

| | ADTree | J48 | NBTree | Random Forest | Random Tree |
|---|---|---|---|---|---|
| Correctly Classified Instances | 198 | 199 | 204 | 205 | 189 |
| Incorrectly Classified Instances | 63 | 62 | 57 | 56 | 72 |
| Kappa statistic | 0.4592 | 0.4342 | 0.4916 | 0.4889 | 0.3951 |
| Accuracy | 75.9% | 76.2% | 78.2% | 78.5% | 72.4% |

Hepatitis disease dataset contains 155 instances and 2 classes: stating the life prognosis yes (or) no.. There are 123 patient records for life prognosis yes and 32 records for patient with no. The results of classifiers are showed in table 3.

Table3. Prediction results of hepatitis dataset

| | ADTree | J48 | NBTree | Random Forest | Random Tree |
|---|---|---|---|---|---|
| Correctly Classified Instances | 40 | 42 | 43 | 46 | 44 |
| Incorrectly Classified Instances | 13 | 11 | 10 | 7 | 9 |
| Kappa statistic | 0.1669 | 0.2299 | 0.4313 | 0.5099 | 0.3057 |
| Accuracy | 75.5% | 79.2% | 81.1% | 86.8% | 83% |

The following Fig. 1 visualizes the accuracy results of decision tree classifiers on various medical datasets.

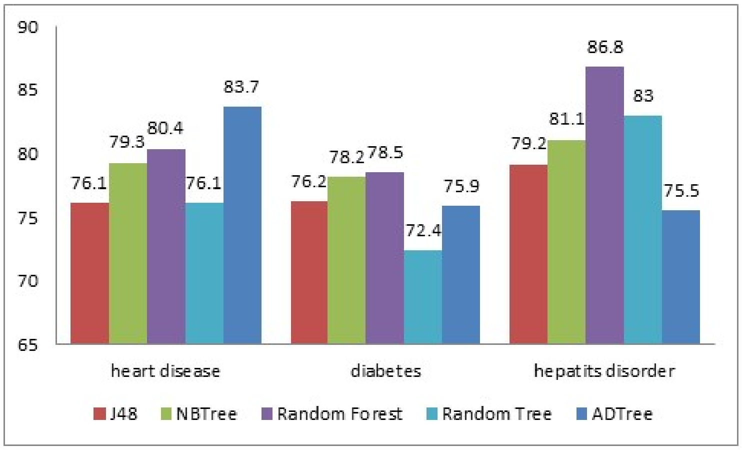

Fig. 1 Accuracy results of classifiers

## 5. CONCLUSION

In this paper, data mining algorithms are used for medical diagnosis. The focus of this paper is to use the different decision tree models for disease prediction in medical diagnosis and work evaluate the performances in terms of classification accuracy of decision tree classifiers. In the future, a new optimized intelligent system can be designed for medical field by using data mining approach and algorithms.

## REFERENCES

[1] E. Venkatesan etal., “Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification”, Indian Journal of Science and Technology, Vol 8(29), DOI: 10.17485/ijst/2015/v8i29/84646, November 2015

[2] D.L.Gupta etal.,” Performance Analysis of Classification Tree Learning Algorithms”, International Journal of Computer Applications (0975 –8887)Volume 55–No.6, October 2012

[3] R. Kohavi. “Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid” Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.

[4] B. Rebecca Jeya Vadhanam retal., “Performance Comparison of Various Decision Tree Algorithms for Classification of Advertisement and Non Advertisement Videos”, Indian Journal of Science and Technology, Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/102098, December 2016

[5] https://en.wikipedia.org/wiki/Random_forest

[6] Osmar R.; Zaine. (1999): Introduction to DataMining, CMPUT690 Principles of Knowledge Discovery in Databases, University of Alberta, Department of Computing Science.

[7] G. Sujatha, Dr. K. Usha Rani:” Evaluation of Decision Tree Classifiers onTumor Datasets”, nternational Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 2, Issue 4, July –August 2013

[8] T.Karthikeyan, P.Thangaraju : “Analysis of Classification Algorithms Applied to Hepatitis Patients”, International Journal of Computer Applications (0975 – 8887) Volume 62–No.15, January 2013

[9] T.Swapna, Y.Sravani Devi: “Performance Analysis of Classification algorithms onParkinson’s Dataset with Voice Attributes”, nternational Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 2 (2019)

[10] T.Nithyapriya, S.Dhinakaran: “Analysis of various data mining classification techniques to predict diabetes mellitus”, Volume5, Issue 4, IJEDR, 2017

[11] Anju Jain: “Machine Learning Techniques for Medical Diagnosis: A Review”, 2nd International Conference on Science, Technology and Management, 2015

[12] Anju Jain etal.:”Medical Diagnosis System Using Machine Learning Technique”, volume 7, Number 1, 2016