# Analysis of Classification Techniques for Efficient Disease Prediction

N. Sandhya, PhD
Professor
VNRVJIET
Hyderabad

M. M. Sharanya
M.Tech CSE
VNRVJIET
Hyderabad

## ABSTRACT

Data mining plays an important role in processing large volumes of data. It refers to the process of obtaining knowledge from raw data. Classification is the most widely used data mining techniques, which employs some set of pre-classified samples to develop a model called a classifier. Many researches showed that C4.5 algorithm need to be improvised to maximize accuracy, handle large amounts of data, where C5.0 is the improved version. The major goal of the classification technique is to predict the target class accurately for each case in the data. The main objective of this research work is to predict diseases using classification algorithms such as Decision trees, C5.0 and Bayesian Networks. The performance of classification algorithms is compared using the datasets, Breast cancer and Heart disease. The experimental results are compared based on different performance parameters like dataset scalability, accuracy and error rate values. The research shows that in terms of scalability Bayesian networks algorithm was proved to have more accuracy rate and less error rate than the C5.0 algorithm.

## General Terms

Data Processing, Classification Algorithms

## Keywords

Classification, C5.0, Bayesian Networks, Decision tree, Disease rules, Disease Prediction

## 1. INTRODUCTION

With the advancement of computer technology, there has been brought a significant emergence of huge volumes of data. These days, major field of research lies in creating knowledge and managing large amounts of heterogeneous data, which is called data mining. Data Mining [1] is the process of recognizing valid, undiscovered patterns in data [2]. Data mining techniques categorized into supervised and unsupervised learning techniques. Classification [3] is the supervised learning technique used for developing models which are called classifiers. It is the data analysis process used to classify and predict the classes from categorical data, where data can be partitioned into training and testing phases. Training dataset is used to categorize the data to develop the model whereas the performance of the classifier can be determined by using testing dataset. Accuracy is the performance measure for a classifier to find how the records are correctly classified. It can be observed by using the confusion matrix from the classified data. Therefore, classification techniques in data mining can be applied to healthcare dataset which makes valuable predictions and conclusions. Accuracy may be varied based on the conditions like the size of the dataset, number of attributes, type of attributes, etc.

This paper gives the accuracy of classification algorithms C5.0 and Bayesian Belief Networks when applied on the datasets differing in size and number of attributes. Proposed work is organized as follows:

Section 2 deals with the related work determining the performance of classifiers and an insight to classification algorithms. Section 3 describes the classification algorithms used in this research. Section 4 shows analysis of datasets used and its experimental results. Section 5 concludes the research results.

## 2. RELATED WORK / LITERATURE REVIEW

Different data mining techniques have been developed by researchers for future prediction. Many classification techniques were proposed for medical data analysis [4] and disease prediction. Analysis can be carried out using both statistical and non-statistical methods of classification.

Rough set Theory [5, 6] is the first non-statistical data analysis approach concerned with the classification for determining imprecise or incomplete information from data. This theory determines the uncertainty in data by using the terms lower and upper approximations of a set. Lower bound determines the member of set, whereas every non-member can be excluded from the set which is given by the upper bound. It is given by a three-valued function with its values as: yes, no, perhaps. Therefore, rough sets can be merged with other methods such as clustering, classification and rule induction.

Fuzzy set [7] approach is the methodology to represent and process uncertainty. This approach not only deals with uncertain data but also used to develop certain models of data that provide better performance than traditional systems. Therefore, these approaches can be combined with statistical techniques to develop a classifier model for future prediction.

Lakshmi.K.R et al **[8]** analyzed Logical Regression, Artificial Neural Networks and Decision tree supervised machine learning algorithms. They used a data mining tool named Tanagra for the classification process for kidney dialysis. Solanki [9] performed analysis by comparing Random tree and J48 algorithms using WEKA data mining tool, gave a predicted model with respect to person’s age of different blood group types. This study showed that random tree algorithm produces more depth decision tree than J48. Milan Kumari, Sunila Godara [10] compared classification models on the basis of Sensitivity, Specificity, True positive rate, False positive rate, Accuracy, Error rate. The study proved that Support vector machine is the best classifier for cardiovascular disease prediction.

# 3. CLASSIFICATION ALGORITHMS

Classification is a supervised learning technique [11], which can be used to develop a model that classifies larger data sets Classifiers define the classes based on the dataset attribute values. It describes those classes according to the characteristics of the data which are known to belong to classes. Data can be categorized as training and testing samples, where training algorithm uses the pre-defined samples to develop a model. The testing sample instances can be used to determine the performance and accuracy of a classifier.

## 3.1 C5.0 Decision tree algorithm

The C5.0 classification algorithm [12] produces the classifiers which can be expressed either as decision trees or rule sets. C5.0 shows the best accuracy for the attributes with missing values and also supports continuous attributes. The entropy and the information gain are used for pruning the tree. The entropy is a measure which determines the extent of similarity among the sets. The Information Gain [13] determines the percentage of given attribute used to partition the training dataset. The Entropy [14] for a set S can be calculated as:

$$\boldsymbol{Entropy(s) = \sum_{i=1}^{n} p_i \, log_2 \, p_i}$$

Where, n = number of classes

$P_i$ = Probability of S belonging to class i

Information Gain is calculated as :

$$\boldsymbol{Gain(A) = Entropy(s) - \sum_{k=1}^{m} \frac{|s_k|}{|s|} Entropy(S_k)}$$

Therefore, Information Gain determines which attribute lies as the root node and includes all the attributes which have more importance in decision making.

## 3.2 Bayesian Networks:

Bayesian algorithm is the statistical classification algorithm which is based on Bayes theorem [15]. It assumes the class conditional independencies between the variables, but there exist dependencies when large health care datasets are used. Bayesian networks or Probabilistic networks represents dependencies between variables and gives the joint conditional probability distribution. The network model is represented using Directed Acyclic Graph (DAG) [16] on a variable values and their conditional dependencies. For each variable it generates Conditional Probability Table (CPT) [17] for each variable which indicates all the possible combinational values of its parent.

Bayesian network model generates a mathematical structure which can be used for modeling very complicated relations among the random variables. The algorithm represents the following steps:

Selection of Rule generation attributes using Conditional Probability Set Theory:

Step 1: It can be done by using the formula: Conditional Support

CS=P (Condition Attribute value | Decision Attribute value)

Step 2: It classifies the value of decision attribute into yes or no and computes the conditional support for each attribute.

Step 3: Hereafter determining a root attribute, Probabilities can be given by:

CS=P (Condition Attribute value | P(Root Attribute | Decision Attribute value))

Therefore, a probability for each attribute is calculated and uses the above algorithm steps for Bayesian Network. Each node in the graph is assigned with weights representing the possible ranges for each attribute. The network shows the qualitative dependencies among the variables, whereas the quantitative information is given by the probabilistic distributions which determine the strength of the shown dependencies.

# 4. EXPERIMENTAL RESULTS

## 4.1 Dataset Used

In this work, datasets related to Breast cancer and Heart diseases were collected from UCI machine learning repository [18]. Cancer dataset consists of 11 attributes and 699 instances. Heart disease dataset consists of 14 attributes with 270 instances. Attribute values can be continuous and categorical where classification algorithms were applied to determine the accuracy and the performance of the classifier for disease prediction.

The data set is partitioned into Training and Testing data, each with an equal number of instances.

**Table : I Dataset Description**

| S.No | DATASET | ATTRIBUTES | INSTANCES | TYPE |
|---|---|---|---|---|
| 1 | Breast cancer | 11 | 699 | Numeric & Nominal |
| 2 | Heart Disease | 14 | 270 | Numeric & Nominal |

## 4.2 Classification using C5.0

### *a) Rule set*

Initially, 50% of instances were selected from both the datasets as Training dataset It can be used to obtain the classification rule sets.

.

**Table : II Sample Ruleset**

| Dataset | No. of Rules | Decision rules |
|---|---|---|
| Cancer Dataset | 8 | Uni_of_cell_size <=3 and Bare_Nuclei <=2 and Normal_Nuclei<=2 then **2.0** |
| | | Uni_of_cell_size <=3 and Uni_of_cell_shape <=3 and Bare_Nuclei <=2 then **2.0** |
| | | Uni_of_cell_size >3 then **4.0** |
| Heart Disease dataset | 7 | Chol <=215 and old peak<=2.40 and num <=0 then **1.0** |
| | | CP <=3 and old peak <=2.40 then **1.0** |
| | | Old peak > 2.40 then **2.0** |
| | | CP>3 and num<=0 and thal >6 and chol >215 then **2.0** |

### *b) Testing Phase & Performance Analysis*

To determine the classifier's accuracy, the generated classification rules were applied to the testing data. From this, the actual and predicted values will be generated from which confusion matrix is obtained**.**

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + FP + TN + FN}}$$

Error rate for a classifier can be calculated

**Table : III Confusion Matrix Representation**

| | **A** | **B** |
|---|---|---|
| **A** | True Positive | False Negative |
| **B** | False Positive | True Negative |

$$\boldsymbol{Error\ rate} = \frac{\boldsymbol{FP + FN}}{\boldsymbol{TP + TN + FP + FN}}$$

Confusion matrix can be used to analyze the accuracy, which is the performance measure of a classifier.

**Table : IV C5.0 Classification Results**

| **Dataset** | **Accuracy** | | **Build time** | **No. Attributes used** | **Error rate** | |
|---|---|---|---|---|---|---|
| | **Training set (%)** | **Testing set (%)** | | | **Training set (%)** | **Testing set (%)** |
| **Cancer dataset** | 98.584 | 95.376 | < 1 min | 5 out of 10 | 0.01 | 0.03 |
| **Heart Disease dataset** | 92.248 | 76.596 | < 1min | 8 out of 13 | 0.07 | 0.23 |

## 4.3 Classification using Bayesian Networks

### *a) Bayesian Network structure*

Network structure can be obtained by using a target variable "class" for both the datasets. The following structure shows the dependencies between the attributes.

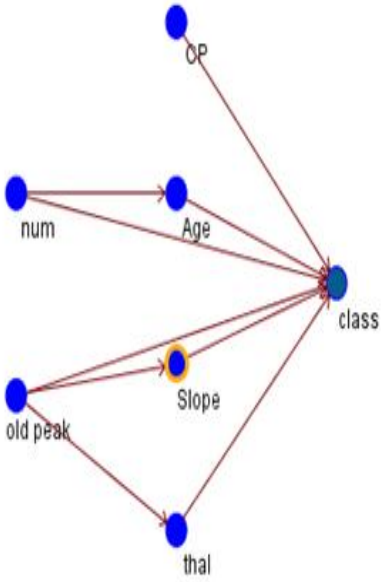

**Fig 1: Bayesian Network for Heart disease Dataset**

### *b) CPT for Attribute- Slope*

The following table shows the sample CPT's

**Conditional Probabilities for Slope:**

Target variable -> class – 1,2

P(Slope| P(Old Peak/ Class)

**Table : V Sample probabilities for Slope and its parent old peak**

| Parents | Probability | | |
|---|---|---|---|
| **Old peak** | **<=1.6** | **1.6-2.4** | **>2.4** |
| **<=2.4** | 0.67 | 0.30 | 0.03 |
| **1.24-2.48** | 0.40 | 0.47 | 0.13 |
| **2.48-3.72** | 0.06 | 0.72 | 0.22 |
| **3.72-4.96** | 0.00 | 1.00 | 0.00 |
| **>4.96** | 0.00 | 0.00 | 1.00 |

**Table: VI Bayesian Networks Classification Results**

| Dataset | Accuracy | | Build time | No. Attributes used | Error rate | |
|---|---|---|---|---|---|---|
| | Training set (%) | Testing set (%) | | | Training set (%) | Testing set (%) |
| **Cancer dataset** | 99.43 | 85.55 | < 1 min | 10 | 0.005 | 0.02 |
| **Heart Disease dataset** | 91.47 | 70.92 | < 1min | 13 | 0.08 | 0.21 |

## 5. CONCLUSION

The experimental results on two datasets showed that the training and testing data sets accuracies May differ. Also, it is concluded that the accuracy of a classifier may vary based on the number of instances in the dataset, the number and type of attributes etc. The results here shows that Bayesian Networks classification algorithm is more scalable compared to C5.0 algorithm which perform analysis for efficient disease prediction on high dimensional datasets, with higher accuracy rate and less error rate. This assumption can be used in data analysis to determine a classification technique based on the data set.

## 6. REFERENCES

[1] Soumen Chakrabarti, Earl Cox, Eibe Frank, Ralf Hartmut Güting, Jaiwei Han, Xia Jiang, Micheline Kamber, Sam S. Lightstone, Thomas P. Nadeau Richard E. Neapolitan, Dorian Pyle, Mamdouh Refaat, Markus Schneider, Toby J. Teorey, Ian H. Witten, "Data Mining-Know it all", Morgan Kaufmann Publishers, 2009

[2] Shomona Gracia Jacob, R.Geetha Ramani, sDiscovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multiclass Categorization of Breast Tissue Data, International Journal of Computer Applications (0975– 8887)Volume 32– No.7, October 2011.

[3]Archana S, Elangovan K.Survey of classification techniques in data mining. International Journal of Computer Science and Mobile Applications. 2014 Feb; 2(2):65–71. ISSN: 2321-8363

[4]Durairaj M, Ranjani V, Data mining applications in healthcare sector a study. Int. J. Sci. Technol. Res. IJSTR, 2(10), 2013.

[5]Zdzislaw Pawlak, Rough Sets, International Journal of Information and Computer Sciences, vol. 11, no. 5, (1982), pp. 341-356.

[6] Pawlak, Z. Granularity of Knowledge, Indiscernibility and Rough Sets, The 1998 IEEE International Conference on Fuzzy Systems Proceedings - IEEE World Congress on Computational Intelligence, (1998) May 4-9, pp. 106-110.

[7] FUZZY SETS AND SYSTEMS, Elsevier An International Journal in Information Science and Engineering

[8] Lakshmi. K.R, Nagesh. Y and VeeraKrishna. M, (2014) Performance Comparison Of Three Data Mining Techniques For Predicting Kidney Dialysis Survivability, International Journal of Advances in Engineering & Technology, Mar., Vol. 7, Issue 1, pg no. 242-254.

[9] SolankiA.V., Data Mining Techniques using WEKA Classification for Sickle Cell Disease, International Journal of Computer Science and Information Technology,5(4): 5857-5860,2014

[10] Milan Kumari, 2Sunila Godara, “Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction”. IJCST Vol. 2, ISSN : 22294333(Print) | ISSN : 0976- 8491(Online) Issue 2, June 2011.

[11] Bhavsar H, Ganatra A. A comparative study of training algorithms for supervised machine learning. IJSCE. 2012 Sep; 2(4). ISSN: 2231-2307.

[12] International Journal of Computer Applications (0975 – 8887) Volume 117 – No. 16, May 2015, C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning.

[13]Informationgain,http://homes.cs.washington.edu/~shapiro/EE596/notes/InfoGain.pdf

[14] International Journal of Information Sciences and Techniques (IJIST) Vol.3, No.1, January 2013, EXTRACTING USEFUL RULES THROUGH IMPROVED DECISION TREE INDUCTION USING INFORMATION ENTROPY.

[15] Choi, J.P., T.H. Han and R.W. Park, 2009. A hybrid bayesian network model for predicting breast cancer prognosis. J. Korean Society Med. Inform., 15: 49-57. DOI: 10.4258/jksmi.2009.15.1.49

[16] Learning Bayesian Network Model Structure from Data Dimitris Margaritis May 2003 CMU-CS-03-153 School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213

[17] Choi, J.P., T.H. Han and R.W. Park, 2009. A hybrid bayesian network model for predicting breast cancer prognosis. J. Korean Society Med. Inform., 15: 49-57. DOI: 10.4258/jksmi.2009.15.1.49

[18]HeartDisease,http://archive.ics.uci.edu/ml/machine/learning-databases/statlog/heart/

[19]International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-11) Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease.