IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-4, Issue-11 E-ISSN: 2347-2693

A Role of decision Tree classification data Mining Technique in Diagnosing Thyroid disease

G. Rasitha Banu

FPHTM, Dept. of HIM&T, Jazan University, Jazan, KSA

E-mail Id:rashidabanu76@gmail.com

Available online at: www.ijcseonline.org

Received: Oct/26/2016 Revised: Nov/05/2016 Accepted: Nov/24/2016 Published: Nov/30/2016

ABSTRACT- Thyroid disease is one of the common diseases to be found in human beings. The disease of thyroid gland varies from the low production as well as high production of the thyroid hormone, respectively. However, it is always recommended to diagnose the disease at an earlier stage in order to prevent further harmful effects and to provide the treatment to keep the thyroid hormone at normal level. Data Mining is playing vital role in health care applications. It is used to analyze the large volumes of data. One of the important task in data mining is predicting disease in earlier stage, which assist physician to give better treatment to the patients. Classification is one of the most significant data mining technique. It is supervised learning and used to classify predefined data sets. Data mining technique is mainly used in healthcare organizations for decision making, diagnosing diseases and giving better treatment to the patients. The data set used for this study on hypothyroid is taken from University of California Irvine (UCI) data repository. The entire research work is to be carried out with Waikato Environment in Knowledge Analysis (WEKA) open source software under Windows 7 environment. An experimental study is to be carried out using data mining techniques such as J48 and Decision stump tree. The data records are classified as negative, compensated, primary and secondary hypothyroid. As a result, the performance will be evaluated for both classification techniques and their accuracy will be compared through confusion matrix. It has been concluded that J48 gives better accuracy than the decision stump tree technique.

Keywords: Hypothyroid, Data Mining, Classification, Decision Tree.

1. Introduction

Thyroid disease is one of the common diseases to be found in human beings. The disease of thyroid gland varies from the low production as well as high production of the thyroid hormone, respectively. In this study, we have focused on the low production of the thyroid hormones which is known as Hypothyroidism. Hypothyroidism (underactive thyroid or low thyroid) is a condition in which the thyroid gland doesn't produce enough amount of certain important hormones. It is a relatively common problem worldwide often with insidious onset and is relatively asymptomatic. The main functions of thyroid hormone is to "run the body's metabolism," and it is understandable that people with this condition will have symptoms associated with a slow metabolism. The signs and symptoms such as weight gain, decreased appetite, constipation, fatigue, weakness, coarse, dry hair with hair loss, cold intolerance, depression, puffy face and muscle cramps and frequent muscle aches are commonly observed in all the patient suffering from hypothyroid condition. Women, especially those older than age 60, are more likely to have hypothyroidism. Hypothyroidism upsets the normal balance of chemical reactions in a human body. It seldom causes symptoms in the early stages, but, over time, untreated hypothyroidism can cause a number of health problems, such as obesity, joint pain, infertility and heart disease. [1]

Data Mining is the process of semi-automatically analyzing large databases to find patterns. Classification is a data mining (machine learning) technique used to predict group membership for data instances.[2] In our research, J48, decision stump Algorithm is used to predicate thyroid disease.A data set with 29 features downloaded from UCI repository site is used for the experimental purpose. [3] The entire work is carried out with WEKA open source software under Windows 7environment [4]. The performance of classifiers are evaluated through confusion matrix. The rest of the paper is organized as follows. Section II contains data collection. Section III explains the methodology. Section IV represents the performance of classifier. Section V represents Discussion and Section VI contains conclusions and future scope.

II. Data Set Description

The hypothyroid dataset used in this work is collected from the website . The hypothyroid dataset consists of 3772 instances from which 3481 instances belongs to category negative,194 instances belongs to category compensated hypothyroid , 95 instances belongs to primary hypothyroid category while 2 instances belongs to category secondary hypothyroid. There are totally 30 attributes. In our research work we have taken only 12 attributes which will be used to classify the data. The hypothyroid data set is given below.

Data Description SN	Attribute Name	Value Type
1	class	Negative, Compensated,primarythyroid,secondary thyroid
2	On thyroxine	F,T
3	Pregnant	F,T
4	TSH measured	F,T
5	TSH	continuous
6	Goiter	F,T
7	T3	continuous
8	TT4 measured	F,T
9	TT4	continuous
10	Query hypothyroid	F,T
11	Thyroid Surgery	F,T
12	FTI	continuous

Table 1: Hypothyroid Data Set

III. Methodology

a. Preprocessing

Data preprocessing is a data mining technique. It is used to reduce the volume of data. There are many data reduction techniques are available such as data compression, numerosity reduction, dimensionality reduction and discretisation. In our work, we have used dimensionality reduction to select the subset of attributes from original data.

b. Classification

Classification is one of the data mining Technique. It is used to group the instances which belong to same class. It is a supervised learning, in which predefined training data is available. Most popular data mining classification techniques are decision trees and neural networks.

c. Decision tree

Decision tree is one of the classification technique in data mining. It is tree-like graph. [5] The internal node denotes a test on attribute, each branch represents an outcome of the test, and the leaf node represent classes. It is a graphical representation of possible solutions based on condition from these solutions optimum course of action is carried out. In our work, we have used two decision tree classifier such as decision stump and J48 to classify the hypothyroid data set.

The Algorithm of J48 and decision stump is given below.

Algorithms

1. J48 Algorithm

J48 is a tree based learning approach. It is developed by Ross Quinlan which is based on Iterative Dichotomiser 3 (ID3) algorithm.[5] 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.

Step 1: If all the instances in T belong to the same group class or T is having fewer instances, than the tree is leaf labeled with the most frequent class in T.

Step 2: If step 1 does not occur then select a test based on a single attribute with at least two or greater possible outcomes. Then consider this test as a root node of the tree with one branch of each outcome of the test, partition T into corresponding T1 , T2 , T3 ....., according to the result for each respective cases, and the same may be applied in recursive way to each sub node.

Step 3: Information gain and default gain ratio are ranked using two heuristic criteria by algorithm J48.

2. Decision Stump

A **decision stump** is a machine learning model consisting of a one-level decision tree. That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature. Sometimes they are also called 1-rule. [5]

IV. Experiments with Weka

The open source software Waikato Environment for knowledge Analysis 3.7(WEKA) is used for experiment. It is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, feature selection and visualization. Weka can downloaded from the website.

Performance Measure of Classifications

In our experiment data is supplied to classifier of J48 Algorithm and decision stump to classify the data. The classifiers performance is evaluated through Confusion Matrix.

a. Confusion Matrix

It is used for measuring the performance of classifiers. In the confusion matrix, correctly classified instances are calculated by sum of diagonal elements TP (True Positive) and TN (True Negative) and others as well as FP (false positive) and FN (False Negative) are called incorrectly classified instances.

b. Accuracy

It is defined as the ratio of correctly classified instances to total number of instances in the dataset.

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

Result Analysis

There are totally 3772 records in the hypo thyroid dataset. All the records are classified as negative, compensated hypothyroid, primary hypothyroid or secondary hypothyroid. The following Table 2 representsconfusion matrix for decision stump Algorithm.

Target class	Negative	Compensated hypothyroid	Primary hypothyroid	Secondary hypothyroid
Negative	**3404**	77	0	0
Compensated hypothyroid	0	**194**	0	0
Primary hypothyroid	0	95	**0**	0
Secondary hypothyroid	2	0	0	**0**

Table 2: Confusion matrix for decision stump

In decision stump classifier, the correctly identified instances are 3598 and incorrectly identified instances are 174**.**

The following Table 3 represents confusion matrix for J48Algorithm.

Target class	Negative	Compensated hypothyroid	Primary hypothyroid	Secondary hypothyroid
Negative	**3476**	3	2	0
Compensated hypothyroid	0	**192**	6	0
Primary hypothyroid	2	5	**88**	0
Secondary hypothyroid	2	0	0	**0**

Table3: Confusion matrix for J48 Algorithm

In J48 classifier, the correctly identified instances are 3756 and incorrectly identified instances are 16.

The following Table 4 depicts detailed accuracy for J48 and decision stump algorithm.

Classifier	Accuracy
Decision Stump	95.38%
J48	99.57%

Table 4: Accuracy of classifier

The following chart1 shows the Accuracy of classifiers.

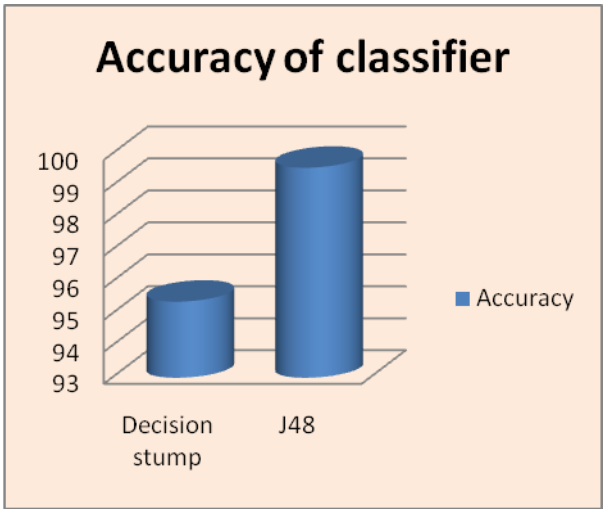

Chart1: Accuracy of classifiers

In this chart, X axis represent the algorithm and Y axis represent the Accuracy. It shows that the Accuracy of decision stump is 95.38% and the Accuracy of J48 is 99.57% which is more than decision stump.

The Misclassification error rate is calculated by the following formula

Misclassification error rate = 1-Accuracy

The Misclassification error rate of classifier is shown below

Table 5: Error rate of classifier

The following chart2 shows the error rate of the classifier.

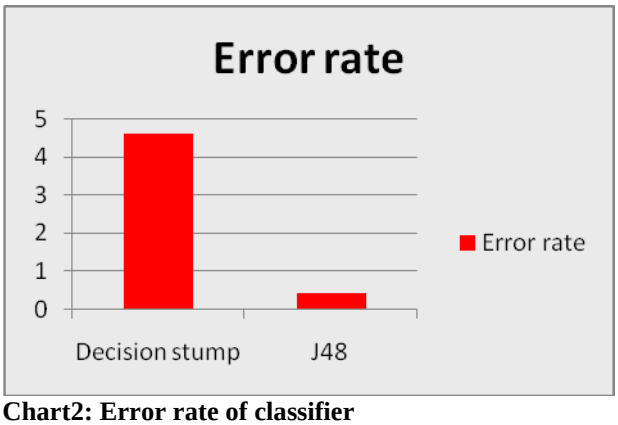

Chart2: Error rate of classifier

In the above chart, X axis represent the classifier and Y axis represent the error rate of classifiers. This chart shows that the J48 classifier is having minimum error rate than Decision stump classifier.

V. Discussion:

In WEKA, there are many classification techniques available. These classification techniques are used to diagnose the thyroid diseases and some other clinical diagnosis issues. Studies showed that many researchers used different methods to diagnose the thyroid disease and achieved the high accuracy of classifiers for the dataset is taken from UCI machine learning repository. [K.Saravana Kumar et al 2014], proposed KNN and SVM classification Algorithm on diagnosing the thyroid disease. They showed that the prediction accuracy of SVM is 94.4336%. However, KNN accuracy is 96.3430%. [Pandy et al 2015], proposed c4.5 and random forest classification Algorithm which gives prediction accuracy of 99.47%.

In our study, the hypothyroid dataset is taken from the website UCI machine learning repository. The hypothyroid dataset consists of 3772 instances from which 3481 instances belongs to category negative, 194 instances belongs to category compensated hypothyroid , 95 instances belongs to primary hypothyroid category while 2 instances belongs to category secondary hypothyroid. There are totally 29 attributes. In our research work we have taken only 12 attributes which will be used to classify the data. We have used dimensionality reduction to select the subset of attributes from original data to improve the performance of classifier. In the original dataset there are 29 attributes. We have used ranking method to select the subset of attributes 12 out of 29 to improve the accuracy performance. We proposed two classification algorithms namely J48 and decision stump to diagnose the thyroid disease. The

Classifier	Error rate
Decision stump	4.61
J48	.424

performances of classifiers are evaluated through the confusion matrix. In J48 classifier**,** the correctly identified instances are 3756 and incorrectly identified instances are 16.But in decision stump classifier, the correctly identified instances are 3598 and incorrectly identified instances are 174**.**

In this experiment, the classifier J48 is giving accuracy of 99.58% and minimum error rate of .424 which is better than decision stump.

VI. Conclusion and Future Scope

Diagnosis of disease is a very challenging task in the field of health care. Many data mining techniques are used in decision making process. In our work, we have used dimensionality reduction to select the subset of attributes from original data and we have applied J48and decision stump data mining classification techniques which are used to classify the hypothyroid disease. The performance of classifiers are evaluated through the confusion matrix in terms of accuracy and error rate. The J48 Algorithm gives 99.58% which is providing better Accuracy than decision stump tree accuracy and also J48 Algorithm gives very minimum error rate than Decision stump. As a future work the same technique is used to apply for other disease datasets such as heart disease, breast cancer, Lung cancer and so on.

References

[1] Available from: http:// www.mayoclinic.org/ diseases conditions/ hypothyroidism/ symptoms-causes/ dxc-20155382.[Last accessed on Dec24].

[2] Jiawei Han, Kamber Micheline (2009). Datamining: Concepts and Techniques, Morgan Kaufmann Publisher.

[3] "UCI Machine Learning Repository of machine learning database", University of California, school of Information and Computer Science, Irvine. C.A. Available from: http://www.ics.uci.edu/.
[4] Available from: http:// www.cs.waikato.ac.nz /ml/weka/. [Last accessed on Dec24].
[5] Availablefrom: http://en.wikipedia.org.[Last accessed on Dec24].
[6] Dr.G.Rasitha Banu, Baviya, "A study on Thyroid disease using Data Mining Technique". IJTRA Journal, Volume -3,Issue- 4,page no- (376-379),August 2015.
[7] Dr.G.Rasitha Banu, Baviya , "predicting Thyroid disease using Data Mining Technique", IJMTER journal, Volume -2,Issue -3,page no- (666-670),March 2015.
[8] K.Saravana Kumar, Dr. R. ManickaChezian, "Support Vector Machine and K- Nearest Neighbor Based Analysis for the Prediction of Hypothyroid. International Journal of Pharma and Bio Sciences",volume – 2,Issue - 5,page no-(447-453),2014 .
[9] Suman Pandey et al, "Thyroid Classification using Ensemble Model with Feature Selection", (IJCSIT) International Journal of Computer Science and Information Technologies, volume – 2,Issue- 6,page no - (2395-2398),2015.

Author Profile

Dr. G. Rasitha Banu, Assistant Professor, Department of Health Information Management and Technology in Jazan University, KSA. She is having 19 years of teaching experience and 10 years of research experience. She has published more than 20 papers in national and International Research journals. She has presented many Technical papers in national and International conferences. Her research area includes Data Mining, Bio-informatics and Cloud computing etc.

