EFFECTIVE PREDICTION OF HYPOTHYROID USING VARIOUS DATA MINING TECHNIQUES

Mrs.K.Sindhya
Assistant Professor, Department of Information Technology, Nirmala College for Women, Coimbatore.

ABSTRACT
Thyroid is a common disease around the world. So there is an emerge to identify the disease for treatment. In older days, prediction process were very slow and the accuracy level was too poor. Data mining is one of the effective area where researchers doing implementations rapidly. Data mining in healthcare is providing a great result in various disease prediction. The accuracy of prediction using data mining techniques are high and the cost of prediction is low. The other main advantage is time taken for prediction is very less. In this paper, I implemented classification algorithms on thyroid data set and produced result.

KEYWORDS: Data mining, Classification, Thyroid, J48, Naive Bayes, Random Forest.

INTRODUCTION
Data mining takes a major role in Research which provides effective results in various fields and especially in health care also. It is more important to diagnosis diseases in early stage. Data mining techniques are efficient to predict the disease with high accuracy. These results are more useful to the patients who can start immediate treatment. Classification is used to assign the collected data items to the target class. These classification techniques are widely used because of its effectiveness in the prediction of diseases in healthcare domain.

Thyroid disease are one of the deadly disease which affect the human life worldwide. Thyroid disease is a medical condition that affects the function of thyroid gland which is located in the front neck of the human body. The main effect of this thyroid is to produce hormones which spreads through the bloodstream to all the organs of human body to affect the growth of the organs[1]. The shape of thyroid gland is like butterfly[2].

CLASSIFICATION ON THYROID DISEASE
The thyroid Diseases are majorly classified into two types.

Fig.No:1 Classification of Thyroid Disease
1. Hypothyroid

Hypothyroid is a kind of thyroid which means the insufficient production of hormones of thyroid gland. This insufficient growth can cause lot of problems in human body like too much of weight gain, dry skin and dry hair, heavy menstrual periods etc. These symptoms are different from one person to another.

<table>
<thead>
<tr>
<th>Hypothyroid</th>
<th>Hyperthyroid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory loss</td>
<td>Irritability</td>
</tr>
<tr>
<td>Dry hair</td>
<td>Hair loss</td>
</tr>
<tr>
<td>Muscle aches</td>
<td>Muscle weakness</td>
</tr>
<tr>
<td>Slow Heart beat</td>
<td>Rapid Heart beat</td>
</tr>
<tr>
<td>Heavy menstrual periods</td>
<td>Short Menstrual periods</td>
</tr>
<tr>
<td>Weight gain</td>
<td>Weight loss</td>
</tr>
<tr>
<td>Constipation</td>
<td>Sleeping difficulties</td>
</tr>
</tbody>
</table>

Table No: 1 Symptom of Hypothyroid and Hyperthyroid

2. Hyperthyroid

The term hyper refers that too much or heavy. So hyperthyroid indicates that the massive growth of hormones of thyroid glands. This hyperthyroid is often called as overactive thyroid. This may cause too much of weight loss even though the sufficient intake of food, increased heart rate, shorter menstrual periods etc.[3]. Table 1 shows the symptoms of Hypothyroid and Hyperthyroid.

RESEARCH METHODOLOGY

(i) Data set Description

The dataset was taken from UCI machine learning repository. The name of the dataset is hypothyroid. This dataset contains 3772 instances. In this dataset 3481 instances are in negative category and 194 instances are in compensated hypothyroid category, 95 instances are in primary hypothyroid category and 2 instances are in secondary hypothyroid category.
Methodology used
Classification is the most popular technique of data mining which produces the best results of prediction. For this work, J48, naïve Bayes and Random Forest algorithms were taken for implementation.

- **J48 Classification Model:**
  This model is a tree based model which used Divide-and-Conquer method for prediction. By using this model, we can get a tree structure which contains one root node and many leaf nodes.

- **Naïve Bayes Model:**
  This Naïve Bayes is a collective model of various classification algorithms based on Bayes theorem. This model is one of the powerful model in classification which is used for high accuracy in prediction.

- **Random Forest Model:**
  Random forest is one of the effective classification model which contains collection of decision trees. This Random Forest was trained using Bagging method which produces number of decision trees to produce a forest like model.

RESULTS & DISCUSSION
The implementation was done with the help of WEKA tool. The dataset was taken from UCI repository.

1. Preprocessing
   Preprocessing is the initial step of every data mining process. By using preprocessing we can identify the missing values, irrelevant data in the dataset and redundancy of the dataset. By removing the noisy data we can get preprocessed dataset for implantation. After that the Dataset was loaded into WEKA tool.
2. Classification

(i) Naïve Bayes:

By Using Naïve Bayes model, we can achieve 95% of accuracy. Time taken to build this model is 0.04 seconds.
(ii) J48 Model:
By using J48 model, the accuracy achieved is 99% and the time taken to build this model is 0.2 seconds.

![Image of J48 model implementation]

**Fig.No:4 Implementation using J48 model**

**Tree View of J48 model**

![Image of J48 model tree view]

**Fig.No:5 Tree View of J48 Model**

(iii) Random Forest Model:
By using this model, the highest accuracy is reached. The Accuracy is 99.3% and time taken to construct this model is 1.17 second.
Comparison of Results:
The efficiency of used model is evaluated by the accuracy the reached and the time taken to reach.

1. Accuracy Comparison
Based on the Prediction accuracy, the Random Forest Model produced high accuracy than the other two models used.

<table>
<thead>
<tr>
<th>Model Used</th>
<th>Accuracy Reached</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>95%</td>
</tr>
<tr>
<td>J48</td>
<td>99%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>99.3%</td>
</tr>
</tbody>
</table>

2. Time Comparison:
Based on the time consumption to construct a model, Naïve Bayes model has taken very short period of time for prediction than the other two models.

<table>
<thead>
<tr>
<th>Model used</th>
<th>Time Taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.04 sec</td>
</tr>
<tr>
<td>J48</td>
<td>0.2 sec</td>
</tr>
<tr>
<td>Random Forest</td>
<td>1.17 sec</td>
</tr>
</tbody>
</table>
CONCLUSION

The effectiveness of a model is basically evaluated by two aspects. One is prediction accuracy and the second is time taken for the prediction. Based on our results, Naïve Bayes had taken only 0.04 sec for prediction. But the accuracy level is lower than J48 and Random Forest. If we had considered the prediction accuracy, Random Forest model produced 99.3% accuracy. But the time taken to build the model is high than the other two models. So we can conclude that J48 is the best model for hypothyroid prediction because the accuracy it produced is 99% which is also one of the highest accuracy and the time taken for this model is 0.2 sec which is comparatively lower than the Random Forest model.

REFERENCES