International Conference on Advancements in Computing Technologies - ICACT 2018 ISSN: 2454-4248
Volume: 4 Issue: 2 85 – 87

Classification Model Using Random Forest and SVM to Predict Thyroid Disease

Roshan Banu D [1], Sharmeli K C [2]

P.G. Student, Department of Computer Science, Stella Maris College, Tamilnadu, India[1]

Assistant Professor, Department of Computer Science, Stella Maris College, Tamilnadu, India[2]

Abstract— This paper presents a better classification model to predict thyroid disease in early stage. Thyroid disease is all over the world. The thyroid gland is an organ at the lower section of the human neck that produces the hormones that helps to regulate many process which is in charge of delivering wide-range of thyroid hormones. The thyroid gland secretes two types of hormones T3 triiodothyronine and T4 thyroxine.In this paper Data Mining Techniques like classification algorithms used are Random forest algorithm and support vector machine algorithm to get the better accuracy combined than one algorithm.

Keywords— *LDA, decision tree, random forest,c4.5,Hypothyroid Neural network, KNN,SVM.*

*****

I. INTRODUCTION

Most tedious and challenging task is to provide disease diagnosis at early stage with higher accuracy in the medical science field. The disease prediction plays an important role in data mining. Data mining is a process of analyzing and extracting hidden information from large data sets to find some patterns. These patterns are useful in prediction method. Clinics and hospitals collect a large amount of patient data over the years. These data provides a basis for the analysis of risk factors for many disease. There are various types of diseases predicted in data mining namely lung cancer, liver disorder, breast cancer, thyroid disease, diabetics etc. Predicting thyroid disease is analyzed in this paper. Thyroid gland will stow thyroid hormones to maintain the body's metabolic rate. Thyroid disorders are caused due to the malfunction of thyroid hormones. Thyroid or thyroid gland releases triodothyronine(T3) and thyroxin(T4) into the blood stream as the vital hormones. Thyroid hormones function are to regulate the rate of metabolism and effect the growth. There are two most common problems of the thyroid disorder or thyroid disease they are hyperthyroidism and hypothyroidism. Hyperthyroidism releases too much thyroid hormone into the blood due to over active of thyroid. This can stimulate your body's metabolism significantly, symptoms like slower heart beat, hair loss, depression and more hypertension. Hypothyroidism is when the thyroid is not active and releases too low thyroid hormone into the blood. This upsets the normal balance of chemical reactions in your 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. Correct explanation of the thyroid disease disorder dataset, also clinical analysis is an important problem in the diagnosis of thyroid. The thyroid prediction techniques will help to reduce the attributes used in classifying thyroid disease.

Several classification algorithms are discussed in this paper like LDA,C4.5,KNN,Neural network,random forest to classify the hypothyroid datasets to early predict the thyroid disease. These algorithms give the accuracy level and some combined with kfold cross validation gives the exact accuracy

II. LITERATURE REVIEW

In this paper provides us a comparative study on various data mining techniques to predict thyroid disease.

[1] In this paper various data mining techniques like Bayes net, Multilayer perceptron, RBF network ,C4.5, CART, REP tree, decision stump are used to develop classifier for diagnosis of hypothyroid disease. K-fold validation is also performed for each technique. Results reflected that a model with k=6 is performing better than others, accuracy in this case is obtained as 99.60% which is acceptable in range for diagnosis thyroid disease

[2] Hypothyroidism and Hyperthyroidism are the two levels of thyroid malfunction. In data mining, Support Vector Machine (SVM) and KNearest Neighbor (KNN) are the two important modes applied to the prediction of hypothyroid. This paper discusses that predictions of Hypothyroid using K-Nearest Neighbor better than the Support Vector Machine.

[3] This paper presents a systematic approach for earlier diagnosis of Thyroid disease using back propagation algorithm used in neural network. Back propagation algorithm is a widely used algorithm in this field. ANN has been developed based on back propagation of error used for earlier prediction of disease. ANN was subsequently trained with experimental data and testing is carried out using data that was not used during training process. Results show that outcome of ANN is in good agreement with experimental data; this indicates that developed neural network can be used as an alternative for earlier prediction of a disease. While training the neural network with error back propagation in conjunction with gradient based training methods, from our experiments we conclude that Levenberg Marquardt method has shown a

 ISSN: 2454-4248
85 – 87

better performance in comparison with simple gradient descent algorithm.

[4] Classification of this thyroid disease is a considerable task. In this paper an experimental study is carried out using Linear Discriminant Analysis (LDA) to achieve better accuracy. There are many data mining classification Algorithms such as CART, REP Tree, and J48 and so on. The LDA Algorithm gives accuracy is 99.62% with cross validation k=6.

[5]Linear Discriminant Analysis (LDA) data mining technique is used to predict the thyroid disorder. In this proposed work, the random forest approach is utilized to predict the hypothyroid disorder by collecting the dataset from UCI repository. The performance measure is calculated from the confusion matrix with the accuracy. In this paper, hypothyroid disorder is predicted using the random forest approach from data mining technique. The experimental result provides improved accuracy, precision, recall and F-measure by comparing the random forest with LDA algorithm. Future

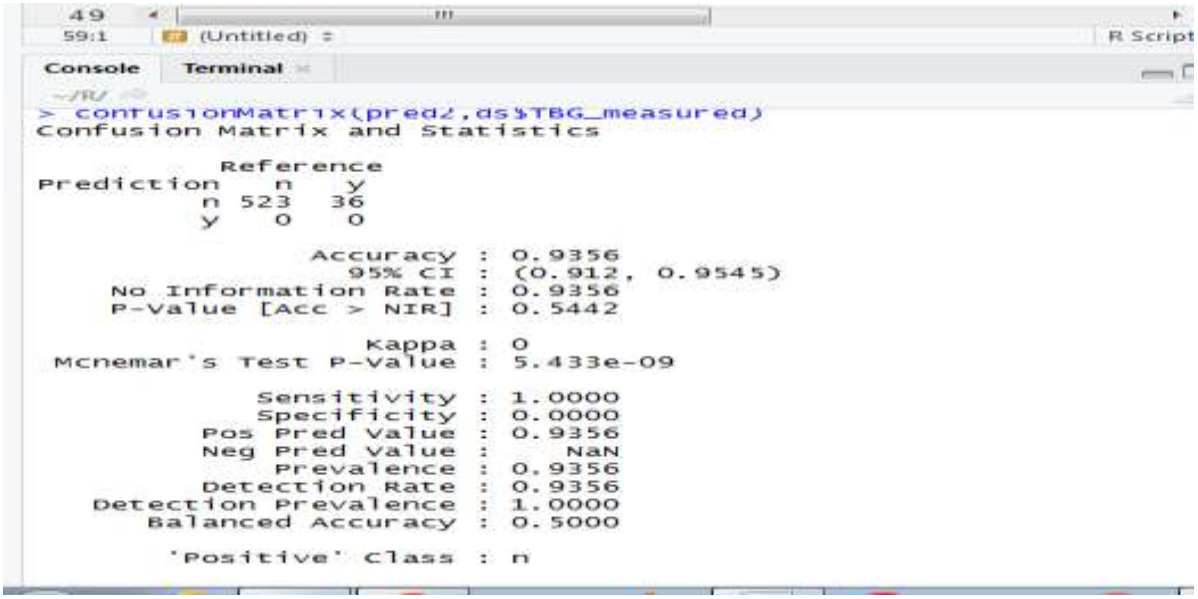
III. Proposed Method

In this paper have made use of random forest algorithm with support vector machine algorithm.Random Forest is a group of tree predictors . A random vector is used. It is sampled independently by using the same distribution θkis handled from the old vectors θ1, θ2, ... , θk−1. X is defined as an input vector. The construction of the tree is handled on the training set by using the random vector θk. The resulting is defined with h (X, θk). If a large numbers of trees are generated, they are voted in order to find the most popular class. The procedure is called as random forests. It is a classifier. Each tree has a cost as a vote for the class selected the most popular at input X.SVM algorithm is Support Vector Machine is one type of learning system algorithm, which is used to perform classification more accurately.SVM used for two class classifier. The essence of SVM is hyper plane also known as "Decision boundary or decision surface". This hyper plane separates the positive and negative of training data sample.

IV. Performance Analysis

V. In this paper we have predicted thyroid disease using two classification models they are random forest and Support Vector Machine. Random Forest algorithm gives more accuracy to predict thyroid disease. When only random forest algorithm alone is being first implemented the accuracy found were 70% by taking the TBG_MEASURED attribute from the thyroid dataset.

VI.



We use Thyroxine-Binding Globulin (TBG) as our predicting value and we proceed by collecting the attributes like TSH,T3 value,T4U_measured.
This gives better accuracy when compared to only random forest algorithm result. So the classification model when combined preferably the accuracy level is high and we found out this using the confusion matrix function.

VII. CONCLUSION

As the medical reports show serious thyroid dysfunctions among the population, more As the medical reports show serious thyroid dysfunctions among the population, more affected being women, thyroid classification is a very important subject for researchers in medical science. In literature are mentioned various research works in the field of thyroid classification based on different data mining techniques used to build robust classifier. In this paper we have discussed about random forest and SVM algorithms. The future work will focus on the identification of factors that affect the thyroid diseases and on testing more data mining techniques for the classification of different diseases (diabetes, heart diseases etc.).

ISSN: 2454-4248
85 – 87

REFERENCES

[1] Pandey, S., Miri, R., & Tandan, S. R. (2013). Diagnosis and classification of hypothyroid disease using data mining techniques. *IJERT, ISSN*, 2278-0181.

[2] Kumar, K. S., & Chezian, D. R. M. (2014). Support Vector Machine And K-Nearest Neighbor Based Analysis For The Prediction Of Hypothyroid. *International Journal of Pharma and Bio Sciences.*, *0ct*.

[3] Prerana, P. S., & Taneja, K. (2015). Predictive data mining for diagnosis of thyroid disease using neural network. *International Journal of Research in Management, Science & Technology*, *3*(2), 75-80.

[4] Banu, G. R. Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data Mining Technique.

[5] Ammulu, V. (2017). Thyroid Data Prediction Using Data Classification Algorithm. *InternationalJournal*, *4*,208-212