ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

A Study of Data Mining Techniques to Detect Thyroid Disease

Roshan Banu D.[1], K.C.Sharmili[2]

PG Scholar, Department of Computer Science, Stella Maris College, Chennai, India[1]

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

ABSTRACT: Thyroid disease is across the board around 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 like growth, energy balance, body temperature and heart rate which is in charge of delivering wide-range of thyroid hormones. Data mining technique is mainly used in healthcare organizations for decision making, diagnosing disease and giving better treatment to the patients. classification of this thyroid disease is an important task. the algorithms used in this paper are CART,LDA, classification, clustering, decision tree and k-fold cross validation

KEYWORDS: LDA, decision tree, CART , neural network, hyperthyroid, DBSCAN, Thyroid disease,etc.

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 Table 1 and Table2.. 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.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

1. Weight loss 2. Heart intolerance 3. Nervousness

Table 1:lists of symptoms or signs of Hyperthyroidism [11]

1. Slower heartbeat 2. Hair loss 3. Depression 4. More Hypertension

Table 2:lists of symptoms or signs of Hypothyroidism [11]

II. LITERATURE SURVEY

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.

LDA:

In [5], Linear discriminant analysis (LDA) is a method used in statistics, pattern recognition and machine learning to find a linear combination of features that separates two or more classes of objects or events. LDA algorithm is used to detect thyroid diseases.

K – Fold cross –Validation:

The process carried out in this is the original sample is randomly partitioned into k equal size subsamples, among the k sample a simple is retained as the validation data for testing the model, now the remaining k-1 subsamples are used as training data. This cross validation process is carried out k folds with each of the k samples used one time of the data to validate. The obtained k results are combined to produce a single estimation. The dominance considered in this method over repeated random sub-sampling is that all observations observed are used for both training and validation, and each

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

observation is used for validation exactly the data is loaded into weka software(in this paper they use this weka open source software to carry out their experiment). After pre-processing, various data mining classification techniques are applied to develop the predictive models on the data set. And then system is trained using the training set. After that it is trained and tested using up to 10 fold cross validation method. Evaluation is performed using certain performance measures.

In [5]This LDA gives better accuracy and is measured as 99.62 with cross validation k=6. The carried work using LDA algorithm has the main disadvantage that it has unbalanced design(i.e. the number of objects in various classes are highly different) it is also not applicable for non-linear problems.

DECISION TREE:

Decision trees are widely used method in data mining. These technology follow divide and conquer policy to divide the space for decision areas. The first node designed is called the root node and the attributes which have split from it are called as leaf nodes.

[1] This paper specifies the diagnosis of thyroid disorders using decision tree attribute splitting rules. As because decision trees are used to follow one decision.

It also helps to divide the data in dataset according to described disorders. This method provides five different splitting conditions for the construction of decision tree. The conditions are Information Gain, Gain Ratio, Gini Index, Likelihood Ratio Chi-Squared Statistics, and Distance Measure. Among, the above splitting conditions three belongs to Impurity based splitting condition and other two are Normalized Impurity based splitting criteria. As a result, the decision tree classifies the thyroid data-set into three classes of disorders. When the data-set is small, splitting criteria will make it. But for large data-set, it is difficult to produce more accurate decision tree. Various splitting rule for decision tree attribute selection had been analyzed and compared. This helps to diagnosis the thyroid diseases through the extracted rules. From this experiment, it is clear, that normalized based splitting rules have high accuracy and sensitivity or true positive rate. This work can also be extended for any medical datasets. Further enhancement can be made by using various optimization algorithms or rule extraction algorithms.

NEURAL NETWORK:

In [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 largely used algorithm in this area. Artificial Neural Network(ANN) has been developed based on back propagation of error used for earlier prediction of thyroid disease. The proposed experiment selects the back propagation algorithm for prediction analysis of Thyroid disease. Back propagation is a neural network learning algorithm. Neural network is a set of connected input/output units in which each connection has a weight associated with it. There are many different kinds of neural networks and neural network algorithms. The back propagation algorithm performs learning on a multilayer feed-forward neural network. The back propagation algorithm works in two phases: propagation and weight update.

With error back propagation in training the neural network in combination with gradient based training methods, from this experiments we conclude that Levenberg Marquardt method has shown a better performance in comparison with simple gradient descent algorithm.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

Also gradient descent which is first order method has shown a poor convergence in comparison to Levenberg Marquardt algorithm. In addition, our observations conclude that error accuracy limit achieved by Levenberg Marquardt method is superior and it trains the models to accuracy level of the order of 10-5 for the applied data set. Research can be extended at different angles like analysis and effects of varying network parameters like number of layers and number of neurons in hidden layers, learning rate and other network parameters

Farhad Soleimanian Gharehchopogh et al. performed a case study in diagnosis of thyroid disease using artificial neural network . The advantage of using ANNs to diagnose disease is to increase the accuracy of performance. The appropriate selection of Artificial Neural Network architecture affects the network performance effectively to reach the high accuracy. By selecting a hidden layer and log-sigmoid activation function for hidden layer and 6 neurons in the hidden layer, we can reach the classification accuracy for Thyroid disease to 98.6%. The proposed method in this paper can be a solution to increase the performance of ANN. So, it can be generalized to the other disease diagnoses systems of ANN.

DBSCAN (Density-based spartial clustering of applications with noise) :

In [11] This is a density based clustering algorithm under data clustering algorithm: A space is considered with given set of points it groups together points that are closely packed together i.e., points with many closely neighbors. DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature. It is opposite to k-means, using an R*tree. In DBSCAN algorithm this chooses an unassigned object p from the data set. Hierarchical multiple classifier classification method to classify the dataset. In this way it as an efficient to classify an data provide accurate information, this reduces time and cost. In this proposed system we can identify the prediction result based on DBSCAN and Hierarchical multiple classifier to classify and clustering the information in efficient result. It displays the result positive when the user symptoms matches the thyroid symptoms, otherwise it shows negative. Future enhancement can support other probabilities top-K queries to improve DBSCAN and Hierarchical multiple classifier.

CART(Classification and regression tree) :

[4][3]Proposed by breiman et al. CART is a non-parametric decision tree machine learning techniques depending on the variable CART produces either classification or regression tree. In this paper it has recursive portioning method to split the data into with segments and similar output field values CART choosing the splitting attribute. All splits are binary. Pruning is also a method. The separation among the probability distributions of the target attribute's value are measured by cost complexity pruning.
[4]CART, NBTree, BFTree, ADTree, REPTree, Random Tree, Random Forest, LMT, FT, DS algorithms were used for the experiment. The WEKA framework was used in order to implement these algorithms.

In [Table values [4]] It is seen that the pair wise comparison between DS-LMT ($p=0.019<0.05$), DS-BF Tree ($p=0.012<0.05$), DS-LAD Tree ($p=0.002<0.05$), DS-NB Tree ($p=0.000<0.05$), CART-LMT ($p=0.037<0.05$),

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

CART-BF Tree (p=0.024<0.05), CART-LAD Tree (p=0.040<0.05), CART-NB Tree (p=0.000<0.05), Random Tree-LAD Tree (p=0.012<0.05), Random Tree-NB Tree (p=0.001<0.05), FT-LAD Tree (p=0.045<0.05), FT-NB Tree (p=0.007<0.05), and REP Tree NB Tree (p=0.012<0.05) are significant. 40% of these significant comparisons include NB Tree.

III. CONCLUSION

This study suggests the practice of few data mining approaches for classification of thyroid disease. Disease diagnosis plays a vital role and it is indispensable for any top clinician. Thyroid disease is one major disease and prediction is the very difficult task. This model gives with classification and clustering accuracy with less number of features compared to other existing developed model. Various splitting rule for decision tree attribute selection had been analyzed and compared.

REFERENCES

[1] J. Jacqulin Margret, "Diagnosis of Thyroid Disorders using Decision Tree Splitting Rules", 2012.
[2] Shivanee Pandey, "Diagnosis And Classification Of Hypothyroid Disease Using Data Mining techniques",IJERT volume 2,issue 6 , June, 2013.
[3] Mr. Brijain R Patel, Mr. Kushik K Rana" A Survey on Decision Tree Algorithm for classification", IJEDR volume 2 issue 1,2014.
[4] Dr.G.Rasitha Banu "A study on thyroid disease using data mining algorithms",Ijtra volume 3 issue 4 (july- august 2015)
[5] Suman Pandey1, Deepak Kumar Gour2, Vivek Sharma, "Comparative study onclassification of thyroid diseases" (IJETT) – Volume 28 Number 9 - October 2015
[6] Prerana1, Parveen Sehgal2, Khushboo Taneja3 "Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network,April 2015.
[7] G. Rasitha Banu,., — Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data MiningTechnique, IJTRA
[8] Komal Agrawal1, Mradul Dhakar2"Thyroid prediction system using auto associative neural network ",IJAREEIE ,volume 6, issue 4, april 2017
[9] Ebru Turanoglu-Bekar1, Gozde Ulutagay1, Suzan Kantarcı-Savas2 "Classification of Thyroid Disease by Using Data Mining Models: A Comparison of Decision Tree Algorithms OJIDS volume 2016,issue 2.
International Conference on Modern Trends in Engineering, Science and Technology ICMTEST-17
[10] Ammulu K., Venugopal T., "Thyroid data prediction using data classification algorithm", IJIRST Volume 4 Issue 2 July 2017.
[11] Dr.G.Rasitha Banu, M.Baviya"Predicting thyroid disease using data mining techniques",IJMTER