ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.10

Applying Classification Algorithms to Predict Thyroid Disease

S.Umadevi[1], Dr.K.S.JeenMarseline[2]
Research Scholar[1], Principal[2]
Department of Computer Science
Sri Krishna Arts and Science College, Coimbatore, India

Abstract:
Thyroid disease is a major cause of concern in medical diagnosis and the prediction of which is a difficult proposition in medical research. The machine learning plays a vital role in the process of disease prediction and this paper handles the analysis of the classification of the thyroid disease based on the information gathered from the UCI machine learning repository. In data mining, Artificial Neural Network and K-Nearest Neighbor are the two important modes applied to the prediction of thyroid disease. This work aims at enriching the dataset quality by concentrating on the feature subset selection and the process of prediction with the integration of fuzzy logic and artificial neural network. The feature selection process is performed to remove the useless information and eliminated the redundant information from the raw dataset. Once the features are selected for classification it is applied to the fuzzy based ANN for predicting the type of thyroid disease in the earlier stage itself. The result shows the most prominent result in the field of predicting the thyroid disease very precisely. This paper discusses that prediction of thyroid using fuzzy with artificial neural network is better than the above said approaches

Keywords: K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Fuzzy ANN

I. INTRODUCTION

Disease diagnosis is a very complex and tedious task; as it requires lots of experience and knowledge. One of the traditional ways for diagnosis is doctor's examination or a number of blood tests. The main task is to provide disease diagnosis at early stages with higher accuracy. Data mining plays a vital role in medical field for disease diagnosis. According to an analysis while one in ten adults in India's people is suffering from hypothyroidism. This estimation is found on the premise of an analysis conduct by Indian thyroid society. The study also alert for thyroid and thyroid is 9th ranked in comparison to other type common disease like asthma, cholesterol, depression, diabetes etc. medical practitioner say that thyroid are same as other disorders, however, the investigation population are alert to thyroid disorders, know that there are diagnostic tests for finding of this disease[2].

Thyroid gland is broken into two section (1) Normal category of thyroid gland (2) this category the gland create abnormal type of thyroid hormones like as hypothyroid and hyperthyroid. Hypothyroidism (underactive thyroid or low thyroid) that is called the thyroid hormones are not generating as much as necessary of certain important hormones [8]. Hypothyroidism can justification various health problems such as: heaviness, joint pain, unfruitfulness and heart disease. Hyperthyroidism (overactive thyroid) belongs to a position is the thyroid gland delivers a lack of the hormone thyroxin. For this situation, the body's digestion system is quickening essentially, bringing about sudden weight reduction, a fast or irregular heartbeat. Predictive system is a group of accurate systems from machine learning, & information mining that consider present and reliable reality to make opportunity concerning future or generally difficult to understand occasions [12]

II. LITERATURE REVIEW

Many theoretical works have been proposed for thyroid disease diagnosis. Jacqulin Margret et al proposed a decision tree splitting rule for this disease diagnosis [1]. S.B.Patel worked to predict the diagnosis of heart disease using classification mining techniques [2]. Chang et al adopted seven feature extraction technique viz., co-occurrence matrix, grey level run-length matrix, laws textures energy measures, wavelet and Fourier features based on local Fourier coefficients[3]. Savelonas et al in 2005 proposed variable background active contour model for detection of thyroid nodules in USG images [4,5] Isa et al has experimented for several activation functions such as sigmoid, Hyperbolic tangent, Neuronal, Logarthmic and Sine activation function for the MLP Neural Network and determined the most suitable function to classify the thyroid disease as Hypothyroid and Hyperthyroid. Senol et al proposed a hybrid structure in which Neural Networks and Fuzzy logic are combined to diagnose the thyroid disease [6]. AnupamShukla et al in 2009, has presented the diagnosis of thyroid disorders using ANN's [7] Eystratios et al proposed an USG image analysis technique for boundary detection of thyroid nodule [8].

III. DATASET DESCRIPTION

The original thyroid disease (ANN-thyroid) dataset from UCI machine learning repository is a classification dataset, which is suited for training ANNs. It has 3772 training instances and 3428 testing instances. There are 21 features, 15 of them are categorical and 6 of them are real attributes. The problem is to determine whether a patient referred to the clinic is hypothyroid. Therefore three classes are built to decide whether a patient has thyroid over--, normal-- or under function. The class probability for the normal function is rather high 92.6% and there are only few samples for disfunctions of the thyroid. Thus, it is a very hard classification problem.

TOOL USED

MATLAB (“ MATrixL A Boratory”) is a tool for numerical computation and visualization. MATLAB is an interactive system whose basic data element is an array that does not require dimensioning. This allows us to solve many technical computing problems, especially those with matrix and vector formulations. It also includes tools for developing, managing, debugging, and profiling M-files, MATLAB's applications. All the algorithms in this work will be implemented using MATLAB.MATLAB 2010a version will used for implementation purpose.

CLASSIFICATION ALGORITHMS

Classification is the process of converting the data records into set of classes. It is divided into Supervised classification and unsupervised classification. In supervised classification, the data that are to be classified is previously known based on few assumptions. In Unsupervised classification, the set of cases were not predicted by the users. By some assumption it is the job of the user to classify the given data and try to assign the name for those cases [10]. This type of classification is known as clustering. Some examples of popular data mining classification algorithms include Decision tree, neural networks, Support vector machine, Naïve Bayes and K-Nearest neighbor. This work only focuses on K-Nearest neighbor, Artificial Neural Network and Fuzzy ANN

KNN

When given a training tuple K-Nearest Neighbor simply stores it and waits until it is given a test tuple. Hence it is a “lazy learner” as it stores the training tuples or the “instances”, they are also known as “Instance- Based Learners”. [9] Thus, k is a positive integer and decides how many neighbors influence the classification. “Closeness” is defined in terms of a distance metric such as “Euclidean Distance” or “Manhattan Distance”.

ARTIFICIAL NEURAL NETWORK

Neural networks (NNs), more accurately called Artificial Neural Networks (ANNs). It is expressed in terms of biological neuron system. It consists of number of separate units. The individual units are communicated to each other by sending signals. It is similar to the brain composed of many processing components. It is organized as a directed graph which contains nodes and the edges connecting each node. The edges are the interconnections between each node

FUZZY BASED ANN

In the field of artificial intelligence, neuro-fuzzyrefers to combinations of artificial neural networks and fuzzy logic. Neuro-fuzzy hybridization results in a hybrid intelligent system that synergizes these two techniques by combining the human-like reasoning style of fuzzy systems with the learning and connectionist structure of neural networks. The term “Neuro-Fuzzy” can be associated with hybrid systems which act on two distinct sub problems: a neural network is utilized in the first sub problem (e.g., in signal processing) and a fuzzy logic system is utilized in the second sub problem (e.g., in reasoning task).

III. FEATURE SELECTION

In this work four different feature selection techniques are used for finding the best subset of features from the whole feature set of the thyroid dataset. They are Mutual Information, Random subset feature selection, sequential forward selection, statistical dependency. Each of these feature selection is explained in the following subsections

Mutual Information: It has been used as a criterion for feature selection and feature transformations in machine learning. It can be used to characterize both the relevance and redundancy of variables, such as the minimum redundancy feature selection

Random Subset Feature Selection: The basic goal of Random Subset Feature Selection (RSFS) is to find a subset of features that are beneficial in the given classification problem. These features are obtained by repetitively classifying the data with a KNN classifier while using randomly chosen subsets of all possible features and adjusting the relevance of each feature according to the classification performance of the subset that the feature participates in.

Sequential Forward Selection: It is the simplest greedy search algorithm n Starting from the empty set, sequentially add the feature x^+ that results in the highest objective function $J(Yk+x^+)$ when combined with the features Yk that have already been selected. SFS performs best when the optimal subset has a small number of features.

Statistical Dependency: It is a technical issue that is based on the concept of probability. In Bayesian theory probability is related to beliefs, which gives us a somewhat less technical interpretation of dependencies.

IV. RESULTS

The prediction of type of thyroid disease is simulated using MATLAB. The thyroid dataset is collected from the UCI machine learning repository. This work consists of two different stages. In the first stage feature subset selection is performed by adapting mutual information and prediction of the thyroid dataset is done using fuzzy ANN.

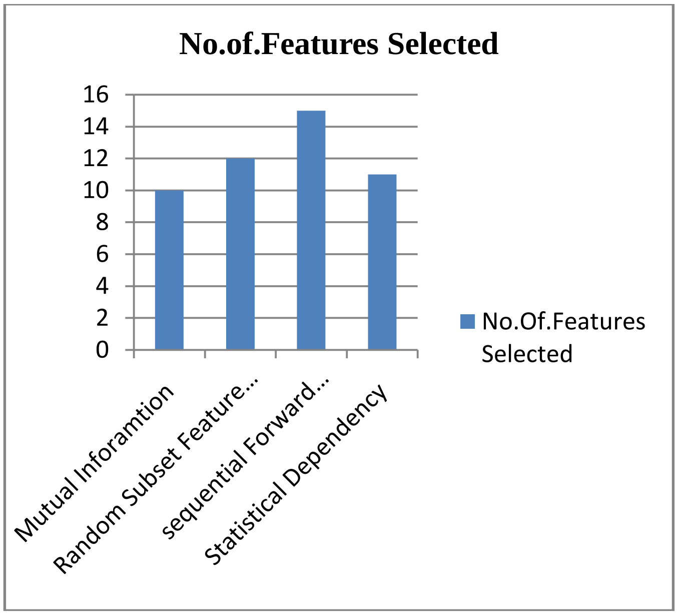

Figure.1. Feature Subset Selection

The figure shows the four different feature subselection technique. The result shows that the proposed mutual information based feature selection method have minimum number of attributes while comparing the other three feature selection technique.

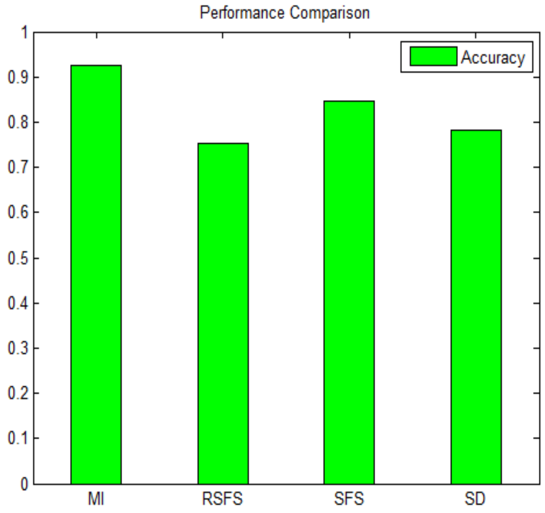

Figure.2. Performance Comparison of Feature Subset Selection

The accuracy of the classification of the thyroid dataset based on the features selected by each feature selection algorithms is depicted int the figure. From their performance it is observed that the mutual information based feature selection is consistently producing more accuracy than the other three exisitng technqiues and the size of the subset is also very less. So for the classificaiton technique the attributes generated by mutual inforamtion is taken

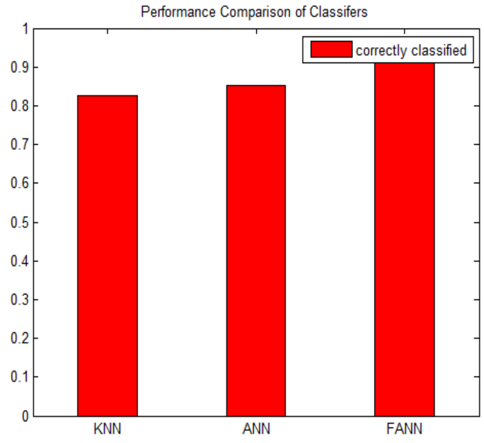

Figure. 3. Performance Comparisons of Classifiers

After performing the feature subset selection the validation process is done with the help of classification algorithms. Here three different classification techniques are used for finding the thyroid disease. K-nearest neighbor, artificial neural network and fuzzy artificial neural network. The result shows that the highest accuracy is obtained using fuzzy ANN which holds the more than 90% of accuracy. The next algorithm ANN performs nearer to 85% whereas KNN obtains the last ranking with the percentage equal to 80%.

IV. CONCLUSION

Thyroid gland is one of the most important gland and largest gland of the endocrine system. The thyroid gland produces two thyroid hormones T3 and T4; these hormones are very helpful to control the body's metabolism. The trial of 21 parameters is used. In KNN the prediction accuracy is 80%, ANN the accuracy is 85%. However fuzzy ANN the prediction accuracy is 90%. Therefore comparatively, Fuzzy ANN performs better than other two classification algorithms

V. REFERENCES

[1]. J.Margret, B. Lakshmipathi and S.A.Kumar, "Diagnosis of Thyroid Disorders using Decision Tree Splitting Rules", International Journal of Computer Applications (0975 – 8887) Volume 44– No.8, April 2012.

[2]. S.B Patel, P. K Yadav, Dr. D. P.Shukla," Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques", (IOSR-JAVS), e-ISSN: 2319- 2380, p-ISSN: 2319-2372. Volume 4, Issue 2 (Jul. - Aug. 2013), Pg.no 61-64.

[3]. Chuan-Yu Chang, Ming-Fang Tsai and Shao-Jer Chen," Classification of the Thyroid Nodules Using Support Vector Machines" International joint conference on Neural Networks 2008, pp 3093- 3098.

[4]. M. Savelonas, D. Maroulis, D. Iakovidis and S. Karkanis, "Variable Background Active Contour Model for Automatic Detection of Thyroid Nodules in Ultrasound Images ", 2005.

[5]. I.S.Isa, Z.Saad, S.Omar, M.K.Osman,K.A.Ahma ,"Suitable MLP Network Activation Functions for Breast Cancer and Thyroid Disease Detection" ,Second international conference on computational intelligence ,modelling and simulation,2010 IEEE, pp 39 – 44.

[6]. Canon SENOL, Tulay YILDIRIM, "Thyroid and Breast Cancer Disease Diagnosis using Fuzzy-Neural Network", International Conference on Electrical and Electronics Engineering, 2009. ELECO 2009.

[7]. Anupama Shukla, PrabhdeepKaur, "Diagnosis of thyroid disorders using artificial neural networks", 2009 IEEE International Advance computing Conference (IACC 2009) – Patiala, India, 2009, pp 1016-1020.

[8]. Eystratios G.Keramidas, DimitrisK. Iakovidis, Dimitris Maroulis and Stavros Karkanis "Efficient and effective image analysis scheme for thyroid nodule detection ".

[9]. PratikshaChalekar, Shanu Shroff, Siddhi Pise, SujaPanicker, "Use of k-Nearest Neighbor in Thyroid disease classification", International Journal of Current Engineering and Scientific Research

[10]. S.Umadevi, K.S.Jeen Marseline, "Data Mining Classification Algorithms"