Comparative Study on Classification of Thyroid Diseases

Suman Pandey[1], Deepak Kumar Gour[2], Vivek Sharma[3]

[1]M-tech Scholar, Computer Science & Engineering, RGPV Bhopal, India
[2]Assistant Professor, Computer Science & Engineering, RGPV Bhopal, India
[3]Head of Department, Computer Science & Engineering, RGPV Bhopal, India

Abstract— *Diagnosis of health conditions is a very challenging task in field of medical science. Diagnoses of health conditions are based on the physician experience. Data mining technique plays role to diagnosis of diseases of patients. Classification is one of the important data mining applications for classification of data. In this work, our main purpose is to propose a robust classifier and compared with other existing classifier which is developed by various authors. We have developed classifications models and its ensemble model for classification of thyroid data. Feature selection is also applied to improve the classification accuracy and increases performance. This paper has used an ensemble of C4.5 and Random Forest which improve classification accuracy of model compared to individual models as well as existing techniques.*

Keywords— *Thyroid diseases, Feature Selection, C4.5, Random Forest, Multilayer Perceptron (MLP), Bayesian Net.*

I. INTRODUCTION

In today's world modern health comprise not only doctors, patients and medical staff but also various process including patient's treatment. In this decade many modern techniques and computational system have been emerged in order to facilitate their operations. Today Health care organization generates a large amount of medical data in databases and data warehouse. The basic data include only primary information about patients such as name, age, address, blood type, etc. The more advanced ones let the medical staff record patient's visits and store detailed information concerning their health condition. Thyroid [7] is one of the most common diseases found in human being, it is not a deadly disease, but it is chronic disease which can give rise to other diseases. Thyroid is a butterfly-shaped gland, which is located at the bottom of the throat responsible for producing to active thyroid hormones, T3 and T4 that affect some function of the body [3]. Triiodothyronine (T3) and Thyroxin which can sometimes be referred to as Tetraiodothyronine (T4) are hormones which regulate the rate of metabolism and affect the growth and rate of function of many other systems in the body. Iodine deficiency has multiple adverse effects on the growth of the body.

This paper proposed a robust decision support system for diagnostic of thyroid. This research focus on compared to our developed model with existing model. The performances of models are compared with other various existing models like accuracy, sensitivity, specificity.

II. RELATED WORK

Various authors have worked in the field of various data mining techniques. Farhad Soleimanian Gharehchopogh et al. [2] have proposed Multilayer Perceptron(MLP) for classification of thyroid diseases and achieved 98.6% of accuracy.

M.R. Nazari Kousarrizi et al. [3] have suggested support vector machine (SVM) for classification of data. They have used two data sets, the first dataset is collected from UCI repository and the second data is the real data which has been gathered is collected by Intelligent System Laboratory. The suggested algorithm gives 98.62% of accuracy with 3 numbers of features in case of first data set.

Ali keles et.al [4] aims at developing an expert system for thyroid diagnosis that is known as Expert System for Thyroid Diseases Diagnosis (ESTDD). In this expert system authors have used neuro fuzzy rules which could diagnose thyroid diseases with 95.33% of accuracy.

Esin Dogantekin [5] [6] have proposed two hybrid method for thyroid disease diagnosis. One method is based on [5] principal component analysis and least square support vector machine and has produced 97.67% accuracy. The other method [6] is based on Generalized Discriminate Analysis and wavelet support vector machine and this method has achieved 91.86% of accuracy. For both these studies thyroid dataset has been downloaded from UCI machine learning repository.

Shivani Pandey et.al [7] have used various statistical and data mining methods such as Bayes Net, C4.5, Classification and Regression Tree (CART), Multi-layer Perception (MLP), Radial basis function network (RBFN), Reduced error pruning (REP) Tree etc. on UCI machine learning Thyroid dataset and found that C4.5 yields highest accuracy among all methods.

D. Kerana Hanirex et.al [9] have developed a multi-layer thyroid detection system to get the higher efficiency. The proposed system consists of 2 stages for attacked detection and classification. Experimental results show that the proposed layer model can result in better prediction.

E. Zoulias et.al [10] have developed a new classifier based on Decision Trees is proposed to assist the task of thyroid malignancy diagnosis. The purpose of the study was twofold: (a) to introduce a classification scheme for medical applications, and (b) to validate the FNA examination. The

former was established through the comparative analysis with various options of splitting criteria showing marginal superiority of the proposed algorithm. The latter was established through its overall classification accuracy compared with the FNA accuracy in discriminating malignant from non malignant cases with respect to the verified diagnosis.

S. Yasodha et al.[8] have proposed CACC-SVM techniques which is hybridization of class-Attribute Contingency Coefficient (CACC) and support vector machine(SVM) for classification of thyroid data. The proposed model achieved better accuracy compared to other traditional models.

Alfonso Bastias et.al [11] have focused on developing an AIS-based machine learning classifier for medical diagnosis and investigating the capability of the proposed classifier. The proposed classifier successfully improved the identification process of thyroid gland disease.

Sheetal Gaikwad et.al [19] have proposed majorly focuses on hypothyroid medical diseases caused by underactive thyroid glands. The dataset used for the study on hypothyroid is taken from UCI repository. Classification of this thyroid disease is a considerable task. An experimental study is carried out using filter method and wrapped method has helped in removing irrelevant as well as useless features from the data set and to achieve better accuracy.

Gurmeet kaur et al. [20] has proposed an efficient neural network training model for thyroid disease diagnosis. It presents general model for diagnosing any disease. The objective of this paper is to diagnose thyroid disease by using three different neural network algorithms which have different architecture and characteristics.

III.MATERIALS AND METHODS

A. *Material*

The data which is required for this study is the thyroid dataset .This dataset has been downloaded from the University of California at Irvine (UCI) machine learning repository [19] to demonstrate the technique. The dataset contain 30 attributes in thyroid which 29 features are considered as input and the 30^{th} feature is considered as output. In this data set contains 7547 records in which 776 belong to thyroid and 6771 belong to non- thyroid data. Thyroid data consist both hypothyroid and hyperthyroid data. This data set is binary class either thyroid our non class.

B. *Methods*

C4.5 [1] is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees and rule derivation.C4.5 is classification algorithm that can classify records that have unknown attribute values by estimating the probability of various possible results unlike CART, which generates a binary decision tree. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. C4.5 has additional features such as handling missing values, categorization of continuous attributes, and pruning of decision trees, rule derivation and others.

Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. Random forests are often used when we have very large training datasets and a very large number of input variables. Random Forests are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. The basic principle is that a group of "weak learners" can come together to form a "strong learner". Random Forests are a wonderful tool for making predictions considering they do not over fit because of the law of large numbers.

Feature selection [13] is a technique of detecting the principal feature subset from actual set of features, according to some defined feature selection criterion, without feature construction or transformation. Feature selection (also known as subset selection) is a technique that is frequently used in machine learning; where in a subset of the features available from the data are selected for application of a learning algorithm. The primary subset contains the less number of features that most subscribe to accuracy; we drop the remaining, unimportant dimensions.

An ensemble model is a combination of two or more models to avoid the drawbacks of individual models and to achieve high accuracy. Bagging and boosting [17] are two techniques that use a combination of series of k learned models (classifiers), M1, M2,…..Mk, with the aim of creating an improved composite model, M. Both bagging and boosting can be used for classification. In the proposed model we have used voting scheme related to bagging ensemble model for classification of thyroid data.

IV.EVALUATION CRITERIA

The performance of the individual models are calculated using various statistical measures; accuracy, sensitivity, specificity etc. These measures are interpreted using true positive, true negative, false positive and false negative. A true positive [21][22]decision occurs when the positive prophecy of the system matches with a positive prediction of a thyroid. A true negative [21][22] decision occurs when both the system and the physician predict the absence of positive predictions. False positive occurs when a system marks a healthy case; a negative one as a positive case. Finally, false negative exists when the system marks a positive case as negative.

Following are some of the performance measurements:
Accuracy: Accuracy [21] can be defined as the no of instances classified correctly by the total number of cases

Accuracy = (TP + TN) ……. (1)

$$\frac{}{(P + N)}$$

Sensitivity: Sensitivity measures the possibility of positive classification of instances i.e. TP to the sum of TP and FN

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \quad \ldots\ldots\ldots\ldots\ldots (2)$$

Specificity: Specificity measures the possibility of negative classification of instances i.e. TN to the sum of TN and FP

$$\text{Specificity} = \frac{TN}{(TN + FP)} \quad \ldots\ldots\ldots\ldots\ldots (3)$$

V. ALGORITHM FOR PROPOSED MODEL

Step 1: Collected data set from UCI repository.
Step 2: Apply data sets on Decision tree techniques and calculate the performance of individual model using equation 1 to 3.
Step 3: Ensemble the C4.5 and Random forest model using voting scheme and calculate the performance measures using equation 1 to 3.
Step 4: Compare measures of ensemble model with individual model and ensemble model gives better performance.
Step 5: Apply the feature selection on best ensemble model (C4.5+RF) with different feature subsets.
Step 6: Achieved high classification accuracy with reduced 5 feature subset using ensemble of C4.5 and random forest of with thyroid data set.
Step 7: Recommended ensemble of C4.5 and random forest model for thyroid data set.
Step 8: end

The below figure shows that process to developed our proposed model.

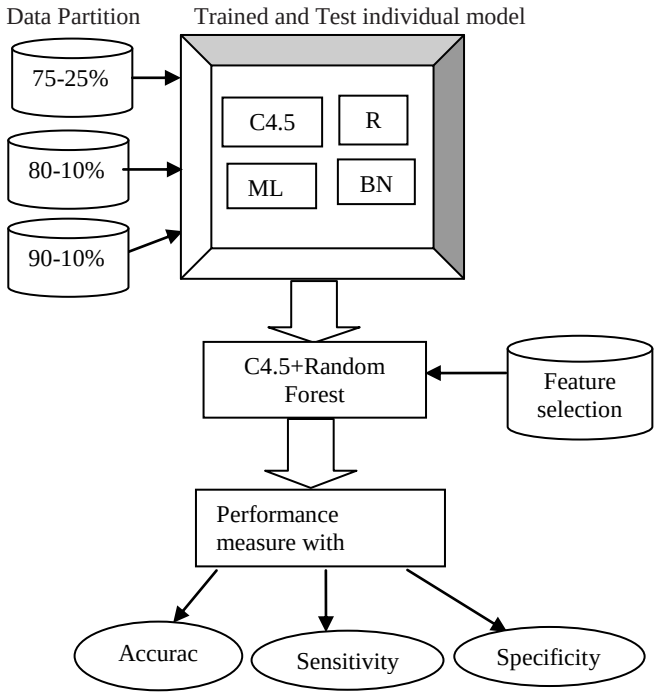

Fig.1 Developed a proposed model

VI. COMPARATIVE CHART OF PROPOSED MODEL WITH EXISTING MODEL

The below table 1 shows that comparative of accuracy with our proposed model with existing developed model. Various authors have worked in the field of thyroid classification and used various techniques and compared with proposed one. The below table shows that our proposed model given high accuracy compared to other existing model.

TABLE1: ACCURACY OF VARIOUS MODELS WITH VARIOUS AUTHORS

Authors	Year	Model	ACR%
Esin Dogantekin	2010	Least square SVM	97.67
Ms. Nikita Singh	2012	SVM, KNN and Bayesian	84.62
M. R. Nazari Kousarrizi	2012	SVM	98.62
D. Kerena Hanirex	2013	Multi classification approach	96.74
Anurag Upadhyay		Decision tree C4.5& C5.0	95
Ali Keles		Neuro fuzzy rule	95.33
Farhad Soleimanian	2013	ANN	98.6
Our proposed Model	2015	Ensemble of C4.5 and Random Forest	99.47

VII. CONCLUSIONS

Thyroid classification is very important in the field of medical science. In this research work, we have developed new classification model like ensemble of C4.5 and random forest for classification of thyroid data. Feature selection is very important role to improve the performance of our developed model. Our proposed model gives high classification accuracy with less number of features compare to other existing developed model.

REFERENCES

[1] Arun K. Pujari, "*Data Mining Techniques*", 4th edition, Universities Press (India) Private Limited, 2001.

[2] Farhad Soleimanian Gharehchopogh, Maryam Molanyand and Freshte Dabaghchi, "*Using artificial neural network in diagnosis of thyroid deceases*" A case Study, International Journal on Computational Sciences & Applications (IJCSA) Vol. 3, No.4, pp. 49-61, 2013 .

[3] M. R. Nazari Kousarrizi, F.Seiti, and M. Teshnehlab, "*An Experimental Comparative Study on Thyroid Disease Diagnosis Based on Feature Subset Selection and classification*", International Journal of Electrical & Computer Sciences IJECS-IJENS Vol: 12 No: 01, pp. 13-19, February 2012.

[4] Ali Keles, Ayturk Keles "*ESTDD: Expert system for thyroid diseases diagnosis*", Expert system with Apllications, 34,242-246,200.

[5] Esin Dogantekin,Asif Dogantekin,Derya Avci. "*An automatic diagnosis system based on thyroid gland:ADSTG*",Expert system with applications, 37(9) ,6368-6372, September 2010.

[6] Esin Dogantekin,Asif Dogantekin,Derya Avci. "*An expert system based on generalized discriminant analysis and wavelet support vector machine for diagnosis of thyroid diseases*," Expert system with Application, 38((1), 146-150, January 2011

[7] Shivanee Pandey, Rohit Miri, S.R. Tandan, "*Diagnosis and Classification of Hypothyroid Disease Using Data Mining Techniques*", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 2 Issue 6, June 2013.

[8] S.Yasodha and P. S.Prakash, "*Data Mining Classification Technique for Talent Management using SVM*", International Conference on Computing, Electronics and Electrical Technologies (ICCEET), 978-1-4673-02l0-4/12, pp. 959-963, 2012.

[9] D.Kerana Hanirex and DR.K.P.Kaliyamurthie, "*Multi-classification approach for detecting Thyroid attacks",* Int J Pharm Bio Sci, pp. - 4(3): 1246 – 1251, July 2013.

[10] E. Zoulias,P.A. Asvestas, G.K. Matsopoulos, N. Uzunoglu, S. Tseleni-Balafouta, H. Gakiopoulou, **"***A data mining approach for classifying FNA thyroid data*", School of Electrical and Computer Engineering, National Technical University of Athens, Greece. Department of Pathology, Medical School, University of Athens, Greece.

[11] Alfonso Bastias, Ph.D., Eleonora Horvath, M.D., Felipe Baesler, Ph.D., and Claudio Silva, M.D., "*Predictive model based on neural networks to assist the diagnosis of malignancy of thyroid nodules*", Proceedings of the 41st International Conference on Computers & Industrial Engineering.

[12] Anurag Upadhayay (M. Tech Scholar), Suneet Shukla, Sudsanshu Kumar, "*Empirical Comparison by data mining Classification algorithms (C 4.5 & C 5.0) for thyroid cancer data set*", International Journal of Computer Science & Communication Networks,Vol 3(1), pp.- 64-68.

[13] R., Parimala, "*A studt of spam E-mail classification using feature selection package*", Global General of computer science and technology, vol. 11, ISSN 0975-4175, 2011.

[14] H.S.Hota, "*Diagnosis of Breast Cancer Using Intelligent Techniques*", International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue 3, .January 2013.

[15] UCI Repository of Machine Learning Databases, University of California at Irvine, Department of Computer Science. Available: http://www.ics.uci.edu/~mlearn/databases/thyroid-disease/newthyroid.data (Accessed: 12 Jan 2015).

[16] Wang, J., "Data Mining: opportunities and challenge, Idea Group, USA, 2003.

[17] Han, J. & Micheline, K., "Data Mining: Concept and Techniques, Morgan Kaufmann publisher", 2006.

[18] Ms. Nikita Singh, Mrs. Alka Jindal "*A segmentation method and classification of diagnosis for thyroid nodules*" , IOSR Journal of Computer Engineering (IOSRJCE) ISSN : 2278-0661 Volume 1, Issue 6, PP 22-27 , July-Aug 2012.

[19] Sheetal Gaikwad and Nitin Pise "*An Experimental Stydy on Hypothyroid Using Rotation Forest*", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.6, November 2014.

[20] Gurmeet Kaur, Er.Brahmaleen Kaur Sidhu, "*Proposing Efficient Neural Network Training Model for Thyroid Disease Diagnosis*", International Journal for Technological Research in Engineering Volume 1, Issue 11, July-2014.

[21] Bruno Fernandes Chimieski1, Rubem Dutra Ribeiro Fagundes, "Association *and Classification Data Mining Algorithms Comparison over Medical Datasets*", J. Health Inform. Abril-Junho*; 5(2): 44-5, 2013.*

[22] *Electronic Decision Support for Austraila's Health Sector*, National electronic decision support taskforce, 2002.

[23] http://www.cs.waikato.ac.nz/ml/weka/