ISSN:0975-9646

Thyroid Classification using Ensemble Model with Feature Selection

Suman Pandey[1], Anshu Tiwari [2], Akhilesh Kumar Shrivas[3] and Vivek Sharma[4]

[1] *MTech (CSE), Dept of Computer Science and Engineering*
TIT College Bhopal (M. P.)

[2-4]*Dept of Computer Science and Engineering*
TIT College Bhopal (M. P.)

[3]*DLS, PG College Bilaspur (C.G.)*

Abstract: **Now days, diagnosis of health conditions is a very critical and challenging task in field of medical science. Medical history data comprises of a number of tests essential to diagnose a particular disease and the diagnoses are based on the physician experience. The thyroid gland faced by physician which is one of the important organs in the body and also increases cellular activity. Data mining technique can greatly deal the diseases of patients. Classification is one of the most important decision making techniques in many real world problem. In this paper, the main objective is to classify the data as thyroid or non thyroid and improve the classification accuracy. We have used various classifications techniques and its ensemble model for classification of thyroid data. Feature selection technique is also important role to improve the classification accuracy and increases performance. An ensemble of C4.5 and Random forest gives better accuracy 99.47% with 5 features.**

Keywords: Thyroid, Feature Selection, Classification, Info Gain.

I. INTRODUCTION

To rapid increasing of population in the world, various challenges face by the medical science. Due to rapid development of technology and information science, experts are capable to diagnosis of diseases and also achieved better performance. Many authors have worked in field of thyroid classification. Farhad Soleimanian Gharehchopogh et al. [1] have proposed Multilayer Perceptron(MLP) for classification of thyroid diseases and achieved 98.6% of accuracy. M. R. Nazari Kousarrizi et al. [2] have suggested support vector machine (SVM) for classification of data. They have used two data set, The first dataset is collected from UCI repository and the second data set is the real data which has been gathered is collected by Intelligent System Laboratory of K. N. Toosi University of Technology from Imam Khomeini hospital. The suggested algorithm gives 98.62% of accuracy with 3 numbers of features in case of first data set. S. Yasodha et al. [3] 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.

II. TECHNIQUES

C4.5 [7] 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.

Random forest (or RF) (R. Parimala et al. 2011) 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 forest (or RF) [8] 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.

MLP [7] is a development from the simple perceptron in which extra hidden layers (layers additional to the input and output layers, not connected externally) are added. More than one hidden layer can be used. The network topology is constrained to be feedforward, i.e., loop-free. Generally, connections are allowed from the input layer to the first (and possible only) hidden layer, from the first hidden layer to the second and so on, until the last hidden layer to the output layer. The presence of these layers allows an ANN to approximate a variety of non-linear functions. The actual construction of network, as well as the determination of the number of hidden layers and determination of the overall number of units, is sometimes of a trial-and-error process, determined by the nature of the problem at hand. The transfer function generally a sigmoid function. Multilayer perceptron is a neural network that trains using back propagation.

Bayesian Net [5] is statistical classifiers which can predict class membership probabilities, such as the probability that a given tuple belong to a particular class. Let X is a data sample whose class label is unknown. Let H be some hypothesis, such as that the data sample X belongs to a specified class C. For classification problems, we want to determine P(H|X), the probability that the hypothesis H holds given the observed data sample X. P (H|X) is the posterior probability, or a posteriori probability, of H conditioned on X.

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 [5] are two techniques that use a combination of models. Each combines a 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.

Feature subset selection [4] is an important problem in knowledge discovery, not only for the insight gained from determining relevant modeling variables, but also for the improved International Journal of Decision Science & Information Technology, understandability, scalability, and, possibly, accuracy of the resulting models. In the Feature selection the main goal is to find a feature subset that produces higher classification accuracy.

III. DATA SET

Thyroid data set is collected from UCI repository[9] database. The data set contains 7547 records in which 776 belong to thyroid and 6771 belong to non- thyroid data. Thyrid data consist both hypothyroid and hyperthyroid data. The data set contents 29 features and 1 class. This data set is binary class either thyroid our non thyroid class. The features _id and feature name are shown in Table I.

Table I: Features of thyroid data set

Id	Feature _Name
1	age
2	sex
3	on thyroxine
4	query on thyroxine
5	on antithyroid medication
6	sick
7	pregnant
8	thyroid surgery
9	I131 treatment
10	query hypothyroid
11	query hyperthyroid
12	lithium
13	goitre
14	tumor:
15	hypopituitary
16	psych
17	TSH measured
18	TSH
19	T3 measured
20	T3
21	TT4 measured
22	TT4
23	T4U measured
24	T4U
25	FTI measured
26	FTI
27	TBG measured
28	TBG
29	referral source
Class	Thyroid or No-Thyroid

IV. PERFORMANCE MEASURES

Performance [6] of each individual classifier and its ensemble models can be evaluated by using some very well-known statistical measures: classification accuracy, sensitivity and specificity. These measures are defined by true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

Table II: Confusion matrix

Actual Vs. Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The above Table shows that confusion matrix. Various performance measures like sensitivity, specificity and accuracy are calculated using this matrix.

Table III: Various Performance Measures

Accuracy	TP+TN)/(TP+FP+TN+FN)
Sensitivity	TP/ (TP+FN)
Specificity	TN/ (TN +FP)

V. EXPERIMENTAL WORK

The experiment carried out using WEKA environment which is open source data mining tools. These experiments have used thyroid data set which is collected from UCI repository data source. This data set applied in different data mining techniques for classification of thyroid and non thyroid diseases.

In this experiment we have applied different partitions of data set in different data mining techniques like C4.5, Random Forest MLP and Bays Net for classification of thyroid data. First we have applied this data set into various individuals' data mining techniques and calculated the accuracy of models. Second we have ensemble the two models for classificationof thyroid data. We have also ensemble C4.5 and Random Forest for classification of this data which gives higher accuracy compared to each individual's models. Partitions of data plays very important role for accuracy of model. From one partition to other partitions accuracy is varying and our proposed ensemble C4.5 and Random Forest gives high classification testing accuracy 99.20% in case of 90-10% as training- testing partitions. Table IV shows that accuracy of various individuals and ensemble model with three data partitions. Figure 1 also shows that classification accuracy of different individuals and ensemble models.

Table IV: Accuracy of Different model with different partitions

Models	75-25% Partition	85-15% Partition	90-10% Partition
	Testing		Testing
C4.5	98.88	99.02	99.07
Random Forest	98.46	98.76	98.94
MultilayerPerceptron	89.03	95.22	94.70
Bayes Net	96.97	97.08	97.61
C4.5+Random Forest	**98.99**	**99.11**	**99.20**

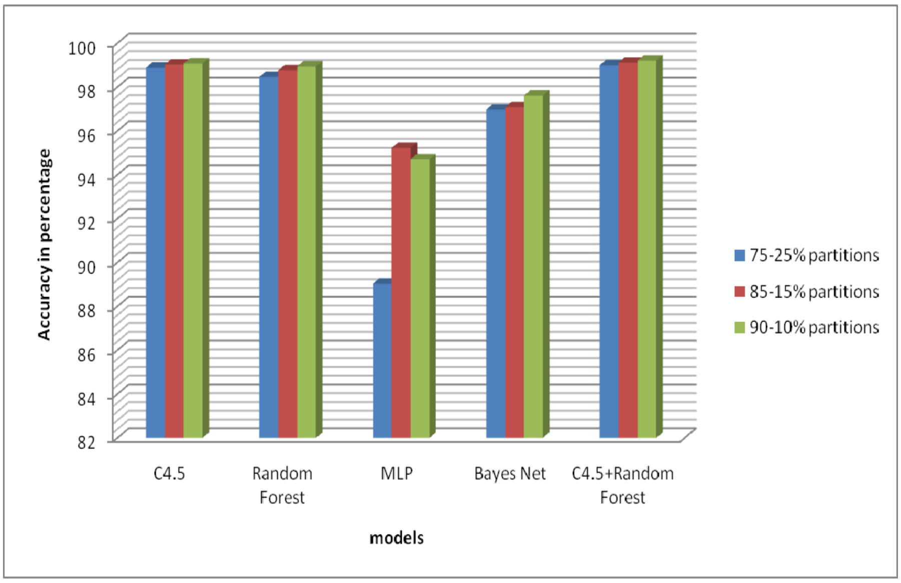

Figure 1: Classification accuracy of individuals and ensemble models

Feature selection is one of the important role for improve the classification accuracy. In this experiment also applied ranking based Info Gain feature selection technique on thyroid data set.The ranking of feature id from higher to lower as given below:

18,26,22,20,29,3,11,24,2,16,17,21,10,27,13,25,23,8,6,19,14,12,9,4,5,7,15,28,1

This experiment used feature selection techniques on best ensemble of C4.5 and Random Forest model with 90-10% training-testing partition. We have eliminated features one by one and applied on proposed model with different data partitions. Table V shows that accuracy of proposed model with reduced feature subsets with different data partitions. In 75-25% and 85-15% training-testing data partitions our model gives 99.15% and 99.29% accuracy respectively with reduced 20 features. Similarly, our proposed model gives highest accuracy 99.47% in case of reduced 5 features with 90-10% training testing partitions. Finally, our feature selection technique is helpful to improve classification accuracy in each data partition, but partitions of data play very important role for varying accuracy of model. Our model gives highest accuracy 99.47% in case of 90-10% training testing partition. Table VI shows that confusion matrix of best model with different partitions in case of reduced feature subsets. Confusion matrix also shows that correct classification and misclassification of data. The other performance measures like sensitivity, specificity and accuracy are calculated using confusion matrix shown in Table VII. Our model gives 99.47%, 99.70% and 97.33% accuracy, sensitivity and specificity respectively in case of 90-10% training-testing data partition as best performance measures .Figure 2 shows that various performance measures with different partitions.

Table V: Accuracy of best ensemble model with different partitions using Info Gain FS

Data Partition (Training-Testing)	No. of Features	Feature ID	**Accuracy**
75-25%	20	18,26,22,20,29,3,11,24,2,16,17,21,10,27,13,25,23,8,6,19	**99.15**
85-15%	20	18,26,22,20,29,3,11,24,2,16,17,21,10,27,13,25,23,8,6,19	**99.29**
90-10%	5	18,26,22,20,29	**99.47**

Table VI: Confusion matrix of best accuracy with best feature subsets

Actual Vs. Predicted	75-25% Partition		85-15% Partition		90-10% Partition	
	Non Thyroid	Thyroid	Non Thyroid	Thyroid	Non Thyroid	Thyroid
Non-Thyroid	1693	3	1012	2	678	2
Thyroid	13	178	6	112	2	73

Table VII: Performance measures in case of Info Gain FS

Performance measures	75-25% Partition	85-15% Partition	90-10% Partition
Accuracy	99.15	99.29	**99.47**
Sensitivity	99.82	99.80	**99.70**
Specificity	93.19	94.91	**97.33**

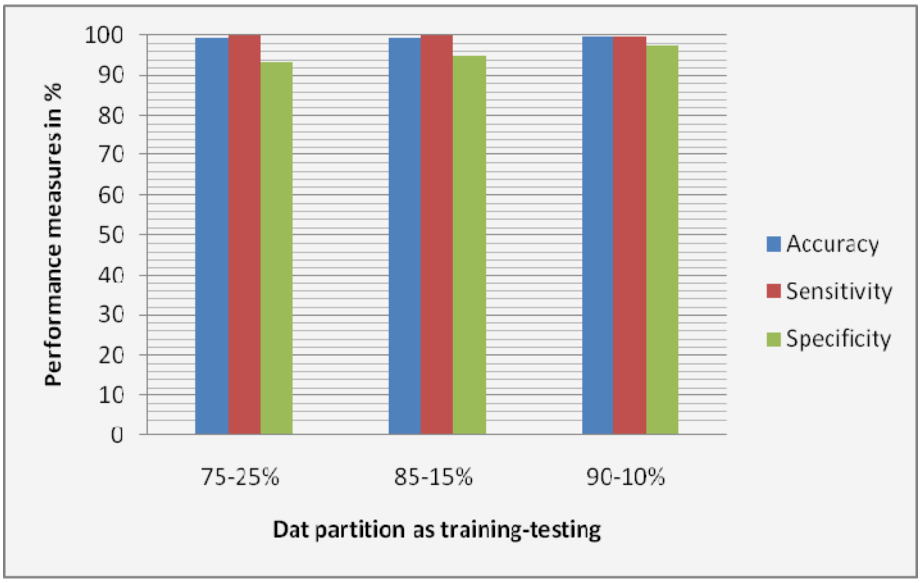

Figure2: Performance measures of proposed model

VI. Conclusion

Diagnosis of thyroid deceases is very challenging and critical issues in medical science. Data mining based classification techniques play very important role to diagnosis of thyroid diseases. In this research work we have used various classification models to classify the thyroid deceases. Features selection is also play one of the important techniques to improve the performance of model. An ensemble of C4.5 and Random forest model give 99.47% of accuracy in case of Info gain feature selection technique for classification of thyroid deceases.

References

[1] Farhad Soleimanian Gharehchopogh, Maryam Molanyand and Freshte Dabaghchi (2013).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.

[2] M. R. Nazari Kousarrizi, F.Seiti, and M. Teshnehlab (2012). 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.

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

[4] Wang, J. (2003). Data Mining: opportunities and challenge, Idea Group, USA.

[5] Han,J.,& Micheline, K. (2006). Data mining: Concepts and Techniques, Morgan Kaufmann Publisher.

[6] H. S. Hota, Akhilesh Kumar Shrivas, S. K. Singhai (2011).An Ensemble Classification Model for Intrusion Detection System with Feature Selection,International Journal of Decision Science of Information Technology, Vol. 3, No. 1, 2011, pp.13-24.

[7] Pujari, A. K. (2001). Data mining techniques, 4th edition, Universities Press (India) Private Limited.

[8] R., Parimala et al. 2011. A study of spam E-mail classification using feature selection package, Global General of computer science and technology, vol. 11, ISSN 0975-4172.

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