International Journal of Engineering and Techniques - Volume 2 Issue 3, May – June 2016

RESEARCH ARTICLE OPEN ACCESS

Classification of Thyroid Disease with Feature Selection Technique

Amit Kumar Dewangan[1], Akhilesh Kumar Shrivas[2], Prem Kumar[3]

[1,2,3]Dr. C.V.Raman University,Kota,
Bilaspur, Chhattisgarh,India

Abstract:

In medical science, Thyroid classification is one the important role for classification of thyroid diseases. Diagnosis of health condition is very challenging task for every human being because life is directly related to health condition. Data mining based classification is one of the important applications for classification of data. In this research work, we have used various classification techniques for classification of thyroid data. CART gives highest accuracy 99.47% as best model. Feature selection plays very important role to computationally efficient and increase the performance of model. This research work focus on Info Gain and Gain Ratio feature selection technique to reduce the irrelevant features from original data set and computationally increase the performance of model. We have applied both the feature selection techniques on best model i. e. CART. Our proposed CART-Info Gain and CART-Gain Ratio gives 99.47% and 99.20% accuracy with 25 and 3 feature respectively.

Keywords—**Thyroid, Classification, Feature Selection.**

I. INTRODUCTION

Now a day's, data is increasing every day in every organization. Today, healthcare is one of the most important field where every day lots amount of patient data are storing. Data mining based techniques are playing very important role for classification of data. Classification is one of the important data mining applications that can apply for classification of medical data. In this research work, we have used various data mining based classification technique for classification of thyroid data. There are various authors have worked in the field of classification of data. F. S Gharehchopogh, et al. [1] have proposed Multilayer Perceptron (MLP) for thyroid diseases classification and given 98.6% of accuracy. M. R. Nazari Kousarrizi, et al. [2] have suggested support vector machine (SVM) for classification of thyroid diseases with two data set, the first dataset is collected from UCI repository and the second data set is the real data which has been gathered by Intelligent System Laboratory of K. N. Toosi University of Technology from Imam Khomeini hospital. The proposed technique given 98.62% of accuracy with 3 features in case of first data set. S. Gaikwad, et. al. [3] have suggested random forest for classification of thyroid data. The suggested model given 96.63% of accuracy. D. Snthikumar, et.al. [4] have suggested various classification techniques like Naïve Bays(NB) , k-Nearest Neighbor (K-NN).S. Panday, et. al. [5] have used various classifiers like C4.5, Random Forest, Multilayer perceptron and Bayes Net for classification of thyroid data. The classifier C4.5 given 99.47% of accuracy with 5 feature subset as robust classifier. A. Upadhyay, et al. [6] have used Tree (CT) ,Clark & Nilbert2(CN2) for classification of thyroid diseases. Clark and Nilbert2 (CN2) given better result two decision tree classifier as C4.5 and C5.0 for classification of thyroid data. C5.0 model given 95% of accuracy which is better than C4.5 classifier. D . Kerana Hanirex, et al. [7] have suggested NNge model for classification of thyroid data. NNge classifier given 96.44% of accuracy with reduced number of features. Md F . Kabiret al. [8] have suggested naïve bayes classifier for classification of thyroid diseases. The proposed model given 94.13% of accuracy as best classifier. D. Lavanya, et al. [9] have suggested CART classifier and compared with other decision tree classifier as C4.5 and ID3 for classification of thyroid data. The CART achieved highest accuracy as 94.68% as best model.

II. METHODS AND MATERIALS

This section elaborates the various data mining based classification techniques as well as feature selection technique for classification of thyroid data. In this section ,we have also described about data set and performance measures which play major role in this research work.

A C4.5

C4.5 [10] is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees and rule derivation. In building a decision tree, we can deal with training sets that have records with unknown attributes values by evaluating the gain, or the gain ratio, for an attribute values are available. We can classify the records that have unknown attribute value by estimating the probability of the various possible results. Unlike, CART, which generates a binary decision tree, C4.5 produces tree with variable branches per node. When a discrete variable is

chosen as the splitting attribute in C4.5, there will be one branch for each value of the attribute.

B. CART

CART (Classification and Regression Tree) [10] is one of the popular methods of building decision tree in the machine learning community. CART builds a binary decision tree by splitting the record at each node, according to a function of a single attribute. CART uses the gini index for determining the best split. The initial split produces the nodes, each of which we now attempt to split in the same manner as the root node. Once again, we examine the entire input field to find the candidate splitters. If no split can be found then significantly decreases the diversity of a given node, we label it as a leaf node. Eventually, only leaf nodes remain and we have grown the full decision tree. The full tree may generally not be the tree that does the best job of classifying a new set of records, because of overfitting.

C. Bayesian Classification

Bayesian classifiers [11], are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem. Classification algorithms have found a simple Bayesian classifier known as the Naive Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

D. Multilayer Perceptron (MLP)

MLP [10] 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 on hidden layer can be used. The network topology is constrained to be feed forward, 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.

E. Radial Basis Function (RBF) Network

The Radical Basis Function (RBF) network [12] is popular several times. The popularity of this network arises from the two basic facts. The first one is that unlike most supervised learning neural network algorithms, it is able to find global optimum. For comparison, using a feed forward neural network with the back propagation learning rule usually finds only the local optimum. The second fact is that training time for RBF network is short compared with the other neural network, most notably when using the back propagation rule for adjustment of the weights. In addition, the topology of the RBF network is very simple to set up, and requires no guessing as with back propagation.

F. Feature selection

Feature selection (FS) [12] is an optimization process in which one tries to find the best feature subset from the fixed set of the original features, according to a given processing goal and feature selection criteria. In this research work we have used two feature selection techniques: Info Gain and Gain ratio and compared to computationally improve the performance of model. Information gain [12] is attribute selection measure based on its ranking. This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or "information content" of messages. The Information gain[12] measures prefers to select attributes having a large number of values . The extension to information gain known as gain ratio based on ranking, which attempts to overcome bias.

G. Data Set

Thyroid data set is collected from UCI repository [13] database. The data set contains 7547 records in which 776 belong to thyroid and 6771 belong to non- thyroid data. Thyrid data consists 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.

H. Performance Measures

Performance [14] of model can be evaluated various performance measures: classification accuracy, sensitivity and specificity. These measures are evaluated using true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

TABLE I.

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 II.

Various Performance Measures	
Accuracy	(TP+TN)/(TP+FP+TN+FN)
Sensitivity	TP/ (TP+FN)
Specificity	TN/ (TN +FP)

III. EXPERIMENTAL RESULTS

This research work carried out using WEKA data mining software in window environment. In this work, we have used various classification techniques like C4.5, CART, Bayesian Net, MLP and RBFN for classification of thyroid data. We have collected thyroid data set from UCI repository and applied on various classification techniques with different data partitions like 70-30%, 80-20% and 90-10% training –testing partitions as shown in TABLE III. TABLE III show that the accuracy of model is varying from partition to partition. In most of partitions CART gives highest accuracy and indicates the robustness of model. TABLE III shows that CART gives highest accuracy i.e. 99.47% of accuracy in case of 90-10% training –testing partitions. Fig.1 shows that accuracy of models with different partitions.

TABLE III.

Accuracy of Different Models with Different Partitions			
Model	**70-30% Partition**	**80-20% Partition**	**90-10% Partition**
C4.5	98.40	98.87	99.07
CART	98.23	98.87	**99.47**
Bayesian Net	97.21	97.21	97.61
MLP	92.57	95.16	94.83
RBFN	93.77	93.70	94.17

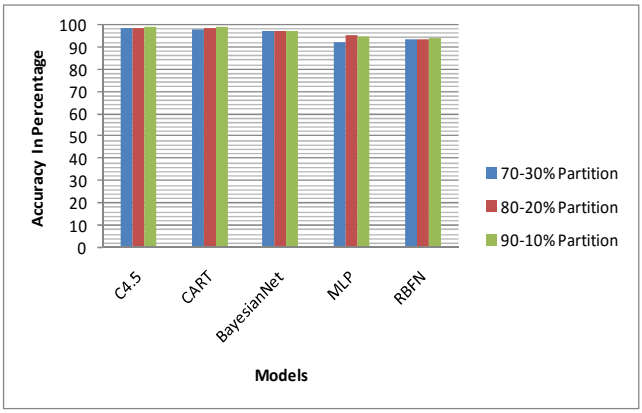

Fig 1. Accuracy of models with different partitions.

Feature selection plays very important role to computationally increase the performance of model. In this research work, we have used two ranking based feature selection technique like Info Gain and Gain Ratio to rank the features of thyroid data set. TABLE IV shows that ranking of features in descending order.

In this research work, we have proposed CART-Info Gain and CART- Gain Ratio feature selection technique for classification of thyroid disease. We have applied Info Gain and Gain Ratio feature selection technique in best model as CART to computationally increase the performance of model. TABLE V shows that accuracy of CART model with different feature subset in case of Info Gain and Gain ratio feature selection technique. In both feature selection technique, the CART gives 99.47% of accuracy with 25 feature subset which accuracy is same as all features. Our Model Info Gain and

CART- Gain Ratio feature selection gives 99.20% of accuracy in case of 3 features which is satisfactory for classification of thyroid disease. The other various performance measures like sensitivity and specificity are given in TABLE VI and TABLE VII in case of Info Gain and Gain ratio FS technique respectively.

TABLE IV.

Ranking of features using Info Gain feature selection technique	
Features Selection Technique	**Ranking of Features in descending order**
Info Gain	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
Gain Ratio	18,26,22,20,13,11,16,24,3,21,27,10,17,8,29,2,14,6,25,23,12,9,19,4,5,7,15,28,1

TABLE V.

Accuracy of CART-Info Gain and CART-Gain Ratio model with different feature subsets			
Number of features	**Accuracy in case of Info Gain FS**	**Number of features**	**Accuracy in case of Gain Ratio FS**
25	99.47	25	99.47
18	99.33	14	99.33
8	99.07	8	99.07
3	99.20	3	99.20

TABLE VI.

Performance measures in case of CART-Info Gain FS with different feature subsets				
Performance measures	**25 Features**	**18 Features**	**8 Features**	**3 Features**
Accuracy	99.47	99.33	**99.07**	**99.20**
Sensitivity	99.70	99.70	**99.26**	**99.41**
Specificity	97.33	96.00	**97.33**	**97.33**

TABLE VII.

Performance measures in case CART- Gain Ratio				
Performance measures	**25 Features**	**14 Features**	**8 Features**	**3 Features**
Accuracy	99.47	99.33	**99.07**	**99.20**
Sensitivity	99.70	99.70	**99.26**	**99.41**
Specificity	97.33	96.00	**97.33**	**97.33**

IV. COMPARATIVE RESULT

TABLE VIII shows that comparative analysis of various existing model with proposed model for classification of thyroid disease. There are various authors have worked in the field of classification of thyroid data but our proposed model gives high classification accuracy compare to other existing model.

TABLE VIII.

Comparative Analysis of Various Models with Proposed Model		
Author's	**Techniques**	**Accuracy**
Lavanya, D., et (2011)	CART	94.68%
Md Faisal Kabiret al.(2011)	Naive bayes	94.13%
Nazari Kousarrizi, M. R. et al.(2012)	Support Vector Machine (MLP)	98.62%
Kerana hanirex, D . et al.(2013)	NNge	96.44%
Gharehchopogh, F. S et al. (2013)	Multilayer Perceptron (MLP)	98.6%
Gaikwad S. et.al.(2014)	Random Forest	96.63%
Amit et al. (2016)	Proposed CART-Info Gain and CART-Gain Ratio	**99.47%**

V. CONCLUSION

Classification of data is very important role to diagnosis of diseases in medical science. This research work focus specially features selection technique to develop computationally efficient model. In this research work, we have applied Info Gain and Gain Ratio feature selection on CART for classification of thyroid disease. The proposed CART-Info Gain and CART-Gain Ratio are computationally efficient and recommended for classification of thyroid disease.

REFERENCES

1. F. S Gharehchopogh, M. Molany, and F. D Mokri, " Using artificial neural network in diagnosis of thyroid disease : A case study", International Journal on Computational Sciences & Applications (IJCSA), Vol.3, No.4, 2013.
2. 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, 2012.
3. S. Gaikwad, and N. Pise, "An experimental study on hypothyroid using rotation forest ",International Journal of Data Mining & Knowledge Management Process(IJDKP),vol.4,No.6,pp.36-37, 2014.
4. D. Senthilkumar, N. Sheelarani, and S. Paulraj, " Classification of multi-dimensional thyroid dataset using data mining techniques: comparison study", Advances in Natural and Applied Sciences, 9(6) Special , pp. 24-28,2015.
5. S. Panday, A. Tiwari, A. K. Shrivas, and V. Sharma, "Thyroid classification using ensamble model with feature selection",International journal of Computer Science & Information Technologies(IJCSIT), vol. 6(3),pp.2395-2298,2015.
6. A. Upadhyay, S. Shukla, and S. kumar, "Empirical comparison by data mining classification algorithms(C4.5 & C5.0) for thyroid cancer data set", International Journal of Computer Science & Communication Networks, vol. 3(1),pp.64-68.
7. D. Kerana Hanirex , and K.P. Kaliyamurthie," Multi-classification approach for detecting thyroid attacks",International journal of Pharma and Bio sciences, vol.4(3),pp.1246-1251,2013.
8. Md. F. Kabir ,C. M. Rahman, A. Hossain, and K. Dahal, "Enhancement classification accuracy of naïve bayes data mining models ",International Journal of Computer Application (IJCA),vol.28(3),2011.
9. D. Lavanya, and D. K. Usha Rani, "Performance evalution of decision tree classifiers on medical datasets", International Journal of Computer Application vol.26(4),2011.
10. A. K. Pujari, "Data mining techniques", 4th edition, Universities Press (India) Private Limited,2001.
11. J. Han, and K. Micheline, "Data mining: concepts and techniques", Morgan Kaufmann Publisher,2006.
12. K. J. Cios, W. Pedrycz , and , R. W. , Swiniarski " Data mining methods for knowledge discovery", Kluwer Academic Publishers, 3rd ed., ISBN: 0-7923-8252-8, 1998.
13. Web source: http://www.archive.ics.uci.edu/ml/datasets.html (last accessed on Feb 2016)
14. H. S. Hota, A. K. Shrivas, S. K. Singhai , "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, 2011