ISSN:0975-9646

Classification of Thyroid Disease: A Survey

Prem Kumar[1], Amit Kumar Dewangan[2]

[1,2]*Dept of CSE,*
Dr. C. V. Raman University,
Kota, Bilaspur,Chhattisgarh, India

***Abstract*— Data mining based classification is one of the important role in the field of healthcare. Diagnosis of health conditions is a very important and challenging task in field of medical science. There are various types of diseases are diagnosis in medical science. Thyroid disease is one of critical diseases that is very serious problem and affected the health of human being. Thyroid decease classification is one of the important problems in medical science because it is directly related to health condition of human body, this type of disease can be solve by proper identify and carefully treatment. This research paper focus on the survey of diagnosis of thyroid. There are various authors have worked in the field of thyroid diseases classification and give the classification accuracy with robust model. This research is also focus on the various techniques that is applied for classification of thyroid data.**

Keywords*— *Classification, Thyroid, Healthcare.

I. INTRODUCTION

Data mining based applications are very beneficial and important in healthcare and medical science. In health care, there are large amount of data, and this data has no organizational value until converted into information and knowledge, which can help control costs, increase profits, and maintain high quality of patient care. Thyroid is one of the most serious health challenges in both developing and developed.

One of the most important applications of data mining technique is classification. Classification is one of the most important decision making techniques in many real world problem. In this paper, there are various authors have worked in the field of classification of thyroid data. The authors have used various data mining based classification techniques, soft computing techniques and statistical techniques and they given satisfactory results for classification of thyroid data.

II. LITERATURE REVIEW

There are various authors have worked in the field of medical science in which thyroid disease is very critical case in human. There are various authors have suggested various techniques to classification of thyroid disease.

Farhad Soleimanian Gharehchopogh et al. [1] have suggested Multilayer Perceptron (MLP) for classification of thyroid diseases and achieved 98.6% of accuracy. M. R. Nazari Kousarrizi et al. [2] have proposed support vector machine (SVM) for classification of thyroid data and 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 with of first data set. S. Yasodha et al. [3] have suggested CACC-SVM techniques which is hybridization of class-Attribute Contingency Coefficient (CACC) and support vector machine (SVM) for classification of thyroid data. The suggested model achieved better accuracy compared to other traditional models. Seetal Gaikwad et.al. [11] have proposed random forest for classification of thyroid data set. The proposed model gives 96.63% of accuracy. D.Snthikumar et.al. [12] have used various classification techniques like Naïve Bays(NB) , k-Nearest Neighbor (KNN) , classification Tree(CT) ,Clark & Nilbert2(CN2) for classification of thyroid data. Clark and Nilbert2 (CN2) gives better result compare to other models. Shivanee Panday et.al. [13] have compared different classification model for classification of thyroid data. C4.5 gives 99.60% accuracy as best classifier. Anurag Upadhyay, et al. [14] have compared two decision tree classifier as C4.5 and C5.0 for classification of thyroid data. C5.0 model gives 95% of accuracy which is better than C4.5 classifier. Suman Panday et.al. [15] have used various classifiers like C4.5 ,Random Forest Multilayer perceptron and Bayes Net for classification of thyriod data. The classifier C4.5 gives 99.47% of accuracy as robust model. D . kerana hanirex, et al. [16] have used NNge model for classification of thyroid data. NNge classifier gives 96.44% of accuracy with reduced number of features. Md faisal kabiret al. [17] have proposed naïve bayes classifier for classification of thyroid data. The proposed model gives 94.136% of accuracy as best model. D. Lavanya, (2011) [18] have proposed CART classifier and compared with other decision tree classifier as C4.5 and ID3 for classification of thyroid data. The CART gives highest accuracy as 94.68% as best model. The details of work of some authors are described in below table 1:

Table 1: Findings by different authors with different techniques for thyroid classification

Sr. No.	AUTHOR NAME	METHODS	DATASET	ACCURACY	FEATURE SELECTION
1	Seetal Gaikwad et.al,2015	Rotation forest, filter method	Thyroid dataset from UCI repository	96.63%	Correction method
2	D. Snthikumar et.al,2015	t-set, Naïve Bays(NB) ,k-Nearest Neighbour(KNN) ,classification Tree(CT) ,Clark & Nilbert2 (CN2)	Multimedia thyroid disease dataset	99.135%	Correlation attribute evaluation for dimensionality reduction
3	Shivanee Panday et.al,2013	RBF N/W, C4.5,CART, REP Tree, Decision stump ,Bays net, Multilayer	Thyroid dataset from UCI repository 29 attribute ,4 classes	MLP -94.035% RBF-95.228% BayesNet-98.59% C4.5 - 99.57%	
4	Ahmad Taher Azar et.al,2014	Linguistic Hedges Neural-Fuzzy Classifier with Selected Features (LHNFCSF)	UCI repository	98.604%	
5	Huiping cheng et.al,	Artificial immune system	Thyroid dataset from UCI repository	99.87%	
6	Jacaulin Margret Et.al	Gini index, Chi square, Distance measure	UCI repository	88.34% in distance Measure	Info gain, Gain Ration feature selection
7	D . Kerana Hanirex	Bayesian network,CART		96.442%	Genetic algorithm
8	Md faisal kabir	Naive Bayes	UCI repository	94.136%	
9	Nazari Kousarrizi et.al	Support vectorMachine	UCI and real data	98.62%	

III. TECHNIQUE USED

There are various data mining, statistical and soft computing techniques have applied by various authors to develop the robustness of model. Various techniques are described below:

A. Decision Tree

Decision tree is very important data mining techniques for classification of data. The decision tree [23] can be construct the rule. Decision tree induction is the learning of decision trees from class labelled training tuples. A decision tree is a flow chart like tree structure, where each internal node denote a test on an attribute, each branch represent an outcome of the test, and each leaf node hold a class label. The topmost node in a tree is the root node. Decision tree can handle high dimensional data. Decision tree algorithm is simple and fast. These tree classifiers have good accuracy. Decision tree induction algorithms have been used for classification in many application areas such as medicine, manufacturing, and production, Financial Analysis, astronomy, and molecular Biology. There are various decision tree algorithms have used by different authors for classification of thyroid data. The brief descriptions of decision trees algorithms are describe below:

1) This algorithm is a successor to ID3 (Iterative Dichotomies 3) [23] developed by Quinlan Ross. It is also based on Hunt's algorithm. C4 .5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values.

2) C 5.0 [14] builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S = (s1. S2, …s n) of already classified samples. Each sample consists of a p-dimensional vector (x1, x2,….xn) where the xn represent attributes or features of the sample, as well as the class in which si.

3) CART [14] stands for Classification And Regression Trees introduced by Breiman. It is also based on Hunt's algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values. CART uses Gini Index as an attribute selection measure to build a decision tree .Unlike ID3 and C4.5 algorithms, CART produces binary splits. Hence, it produces binary trees. Gini Index measure does not use probabilistic assumptions like ID3, C4.5. CART uses cost

complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.

4) Random forest (or RF) [15] 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.
5) **Naïve Bayesian** (NB) [17] is one of the most popular data mining techniques for classifying the large dataset. It has been successfully applied to the different problem domains of classification task such as weather forecasting, intrusion detection, image and pattern recognition, medical diagnosis, loan approval and bioinformatics etc. Naive Bayesian classifier also efficiently applied in feature selection and web
6) Multilayer Perceptron (MLP) [15] 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.
7) Support vector machine (SVM) [22] is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features.

IV. CONCLUSIONS

Thyroid classification is very important role for human being in medical science. Various authors have worked in the field of thyroid classification and they have used various data mining techniques to build robust classifier. Our main motive of this research work is analyzing the various models that developed by different authors. Based on this analysis, we can develop a new and robust model which will give better performance and helpful for diagnosis of thyroid disease.

REFERENCES

[1] Paolo Giudici, Silvia Figini**, "**Applied Data Mining for Business and Industry" ,John Wiley & Sons Ltd,. United Kingdom, 2009.

[2] Alessio Pascucci," Toward a PhD Thesis on Pattern Recognition" , 2006.

[3] **V. Vapnik**, "The Nature of Statistical Learning Theory" ,Springer; 2 edition , 1998.

[4] Han, J.,& Micheline, K., "Data mining: Concepts and Techniques", Morgan Kaufmann .Publisher, 2006.

[5] V. N. Vapnik, "Statistical Learning Theory" , New York: John Wiley and Sons, 1998.

[6] Web source: http://www.archive.ics.uci.edu/ml/datasets.html

[7] Hota, H. S., "Diagnosis of Breast Cancer Using Intelligent Techniques", International Journal of Emerging Science and Engineering (IJESE),Vol. 1, pp. 45-53, 2013.

[8] Lavanya, D. and Usha Rani K., "Ensemble decision tree classifier for breast cancer data", International Journal of Information Technology Convergence and Services (IJITCS), vol. 2, pp. 17-24, 2012.

[9] Kharya, S., "Using data mining techniques for diagnosis and prognosis of cancer disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 2, pp. 55-66.

[10] Polat,K.,Gunes,S., "Breast cancer diagnosis using least square support vector machine", Digital Signal Processing,Elsevier, vol. 17, pp. 694-701,2007.

[11] Sheetal Gaikwad and Nitin 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.

[12] 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.

[13] Shivanee Pandey, "Diagnosis And Classification Of Hypothyroid Disease Using Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT),Vol. 2 , 2013.

[14] Anurag Upadhyay,Suneet shukla,Sudanshu 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.

[15] Suman Panday ,Anshu Tiwari, Akhilesh Kumar Srivas & Vivek Sharma ,"Thyroid Classification using Ensamble Model with Feature Selection",International journal of Computer Science & Information Technologies(IJCSIT), vol. 6(3),pp.2395-2298,2015.

[16] D.Kerana Hanirex And Dr.K.P.Kaliyamurthie ," Multi-Classification Approach For Detecting Thyroid Attacks",International journal of Pharma and Bio sciences, vol.4(3),pp.1246-1251,2013.

[17] Md.Faisal Kabir ,Chowdhury Mofizur Rahman,Almgir Hossain and Keshav Dahal, "Enhancement Classification Accuracy of Naïve Bayes Data Mining Models ",International Journal of Computer Application (IJCA),vol.28(3),2011.

[18] D.Lavanya & Dr.K.Usha Rani,"Performance Evalution of Decision Tree Classifiers on Medical Datasets", International Journal of Computer Application vol.26(4),2011.

[19] Saroj Khatiwada,Basanta Gelal, Nirmal and Madhab Lamsal, "Association between Iron Status and Thyroid Function in Nepalese Children ",Bio Med Central ,vol.9(2),2016 .

[20] Suhitha Chittamuri, Vivekanand Bongi,Mythili Ayyugri,dileep Kumar , Regula Subramanyam and A.V. Kandregula"Pregnancy outcomes in Subclinical Hypothroidism and Thyroid autoimmunity",Thyroid Research & Practice,vol. 1,2016.

[21] J. Jacqulin Margret, B. Lakshmipathi and S. Aswani Kumar , " Diagnosis of Thyroid Disorders using Decision Tree Splitting Rules ", *International Journal of Computer Applications ,pp.-0975 – 8887,Vol. 44– No.8, April 2012* .

[22] Manish Kumar Shrivastava, "*Exploring Data Mining Classification Techniques*", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 2 Issue 6, June – 2013.

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