© 2016 IJSRSET | Volume 2 | Issue 6 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Prediction of Hyper Thyroid Disorders using Classifier algorithms in Data Mining

B. Kavitha
Lecturer, Department of Computer Engineering, IRT polytechnic college, Chennai, India

ABSTRACT

Thyroid disorders in women are known as one of the most common diseases. The thyroid gland regulates the metabolism of the body and its development. It also secretes several hormones, such as Calcitonin, Thyroxine (T4), Tri-iodothyronine (T3). Women of any age can be affected by thyroid issues. Women are more likely to have thyroid disease than men. Such symptoms include hypothyroidism, hyperthyroidism, thyroiditis, goitre, thyroid nodules, thyroid cancer. There are also risks if the thyroid condition is untreated and unrecognized. It could be recognized using data mining algorithms,. The proposed work is to build a model that can diagnose the probability of hyperthyroidism with reasonable precision in patients. Naive Bayes, Random Forest, J48, PART classifier algorithms are used to detect the hyper thyroid problem. Simulation studies were performed for experimental data sets sourced from UCI machine learning repository using these classifiers. The performance of these classifiers is analyzed on various performance metrics, such as Precision, Accuracy, F-measure, and Recall. Accuracy measured over true and false classified instances. PART outperforms with the highest accuracy of 97.99% comparatively other classifiers.

Keywords : Hyper Thyroid, Classifiers, Accuracy, Data Mining

I. INTRODUCTION

Classification strategies are broadly used in the medical field for classifying data into different classes according to some constraints comparatively an individual classifier. The method of finding hidden patterns from massive data sets is known as data mining. This is accomplished by integrating statistical and artificial intelligence approaches with database management. Data mining is a technique for analyzing large amounts of behavioral [8].Some of the tools used in Data mining are ANN (Artificial Neural Network)- [3] Nonlinear predictive models with a structure that resembles biological neural networks and learn through preparation.

Decision trees- Structures in the form of trees that reflect sets of decisions. These choices result in classification rules for a dataset.

Rule- Statistical significance is used to extract useful if-then rules from results.

Genetic algorithm- Genetic combination, mutation, and natural selection are all used in optimization techniques.

Nearest neighbor- A classification method that categorizes each record in a historical database based on the records that are most similar to it.

A classification rule is a process for assigning each element of a population set to one of the groups. A classification rule, also known as a classifier, is a function that can be evaluated for any possible value

and will produce a classification that is similar to the data.

II. RELATED WORK

Shivanee Pandey et al. in[1], have used a variety of data mining techniques to build a classifier for hypothyroid disease diagnosis and classification. In addition, k-fold cross validation was used.
The MFHLSCNN algorithm, which is ideal for clustering and classification, Satish N. Kulkarni et al. in [4] has been discussed in this paper. This algorithm is perfect for on-line adaptation because it can learn ill-defined nonlinear cluster boundaries in a few passes.

Orhan E et al. in [5] , used Multilayer neural networks to conduct a study on tuberculosis diagnosis (MLNN). Two separate MLNN structures were used for this purpose. The MLNN with one hidden layer was one of the constructs, while the MLNN with two hidden layers was the other. For the contrast, a general regression neural network (GRNN) was used to achieve tuberculosis diagnosis.

Using multilayer, probabilistic, and learning vector quantization neural networks, a comparative thyroid disease diagnosis was achieved by Feyzullah Temurats in [6] dos Santos et al in [7].Proposed neural network (NN) modeling for Smear negative pulmonary tuberculosis classification, which was able to correctly classify 77% of patients from a test sample

III. METHODOLOGY

Proposed model is given in Figure.1. Classification is the process of grouping data into homogeneous groups based on certain common characteristics found in the data. The steps in the proposed system are: inputting data set, training data set, applying classification algorithm, obtaining performance metrics, comparing performance with other classifiers, and picking the right prediction model with accuracy.

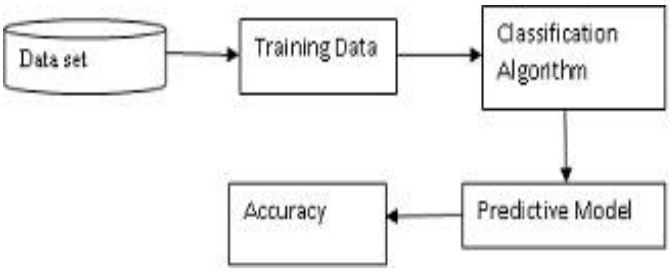

Figure 1: System Model

3.1 Classification Algorithms used

3.1.1 Naïve Bayes Classifier

The Naive Bayes classification algorithm is based on the Bayes Theorem and is used in machine learning. The Bayes Theorem calculates the likelihood of an occurrence happening based on the probability of a previous event. The following equation represents Bayes' theorem mathematically,

$$P(a|b) = \frac{p(b|a)P(a)}{P(b)}$$

P(a|b)=Posterior probability

P(b|a)=Predictor

P(a)= Probability for True

P(b)= Predictors Prior Probability

3.1.2 Random Forest

Random Forest (RF), also known as Random Decision Forest, is a supervised Machine Learning algorithm that uses decision trees to perform classification, regression, and other tasks. From a randomly chosen subset of the training set, the Random forest classifier generates a set of decision trees. It consists of a collection of decision trees (DT) derived from a randomly chosen subset of the training set, which then gathers votes from various decision trees to determine the final prediction. RF is simple to create, forecast and ability to manage data without preprocessing or rescaling. Outliers aren't a problem, and missing values aren't an issue.

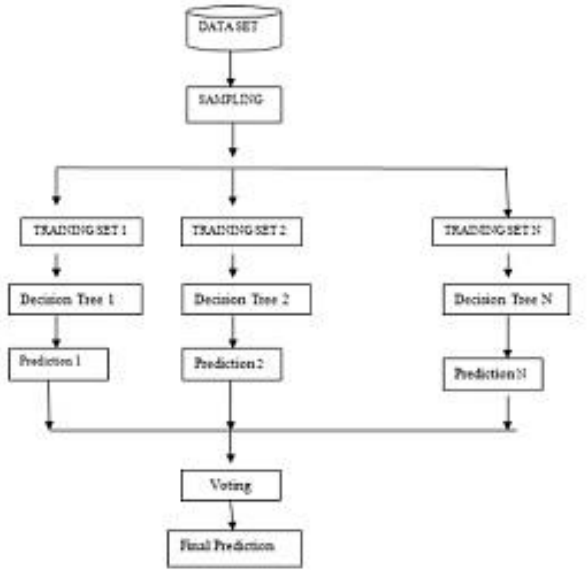

Figure 2: Flow diagram of Random Forest Algorithm

3.1.3 J48 Classifier

J48 classifier is statistical classifier, which is extension of C4.5 algorithm. J48 employs a predictive machine-learning algorithm to determine the resultant value of a new sample based on the available data's various attribute values. The internal nodes of a decision tree represent the various attributes; the divisions between the nodes represent the various potential values for these attributes in the observed samples, while the terminal nodes represent the final value.

3.1.4 PART Algorithm

Projective Adaptive Resonance Theory is abbreviated as PART. The vigilance and distance parameters are used as input for the Component algorithm. [11].

3.2 Cross-Validation

In the proposed model, 10-fold validation has performed. The cross-validation method [9] calculates the average percentage of folds that are correctly classified. Weka runs the learning algorithm 11 times with 10-fold cross-validation, once on each fold of the cross-validation and once more on the entire dataset. Let us consider, Di is the test data set that includes sample xi=(v_i,y_i)and the cross-validation accuracy estimation is defined as:

$$ACC = \frac{1}{n}\sum_{(vi,yi)\epsilon D}^{\infty}\left(\delta(\xi(D\backslash D(i), vi), yi)\right) \quad \text{----(1)}$$

where n is the number of folds.

3.3 Dataset

In this work, WEKA tool is used for performance evaluation of different classifier Waikato Environment for Knowledge Analysis (Weka) [10], developed at the University of Waikato, New Zealand, is free software licensed under the GNU General Public License.

3.4 Attributes used

The proposed system is analyzed on Hyper Thyroid Dataset taken from UCI Repository. This dataset consists of 499 Instances which are of difference aged female and male patients. The dataset comprised of 8 attributes where a class value '0' is predicted as negative for hyperthyroidism and '1' is predicted as positive for hyperthyroidism.

TABLE I

LIST OF ATTRIBUTES USED IN THE DATASET

S.No	Attributes
1	Age
2	Sex
3	TSH
4	T3
5	T4
6	FT3
7	FT4
8	SOURCE

3.5 Performance measures

i) Accuracy - correctly classified

$$\text{Accuracy} = \frac{No.of\ True\ Positive + No.of\ True\ Negative}{Total\ No.of\ Samples}$$

ii) Precision- proportion of positive identifications actually correct

$$\text{Precision} = \frac{No.of\ True\ Positive}{No.of\ True\ Positive + No.of\ False\ Positive}$$

iii)Recall- Proportion of correctly identified actual Positives.

Recall

$$= \frac{No.of\ True\ Positive}{No.of\ True\ Positive + No.of\ False\ Negative}$$

iv)F-measure- measure of a test's accuracy

$$\text{F} - \text{Measure} = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

v)ROC(Receiver Operating Characteristics) area measurement One of the most important values output by Weka. They give an idea of how the classifiers are performing in general.

IV. RESULT AND ANALYSIS

Various performance measures Accuracy, Precision, Recall, F-measure and ROC corresponds to various classifiers are listed in Table-II

TABLE III
PERFORMANCE OVER VARIOUS PERFORMANCE MEASURES

Classification Algorithms	Precision	Recall	F-measure	Accuracy %	ROC
Naïve Bayes	0.981	0.958	0.966	95.24	0.989
Random Forest	0.951	0.973	0.962	97.32	0.948
J48	0.973	0.975	0.974	97.49	0.703
PART	0.977	0.98	0.978	97.99	0.77

Table-III shows performance measures of various classification algorithms. From Table-III, it is analyzed that PART algorithm gives more accuracy as compared with other algorithms. Figure 3, Figure 4,Figure 5,Figure 6 represents the graphical analysis based on various Performance measures .Figure 7,represents ROC area of classifiers taken in the proposed model.

TABLE IIIII
CORRECTLY, INCORRECTLY CLASSIFIED INSTANCES

Total No of Instances	Classification Algorithms	Correctly Classified Instances	Incorrectly Classified Instances
599	Naïve Bayes	574	25
	Random Forest	583	16
	J48	584	15
	PART	587	12

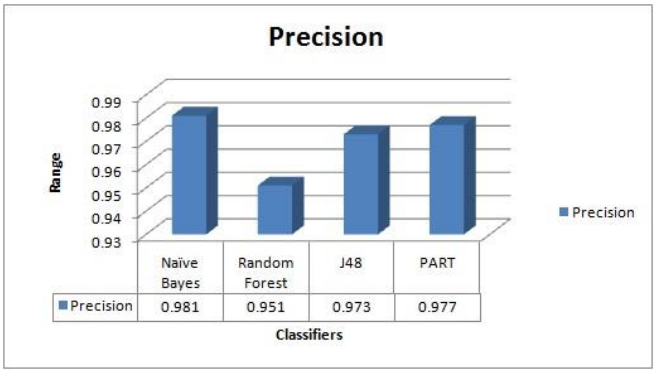

Figure 3: Precision over various classifiers

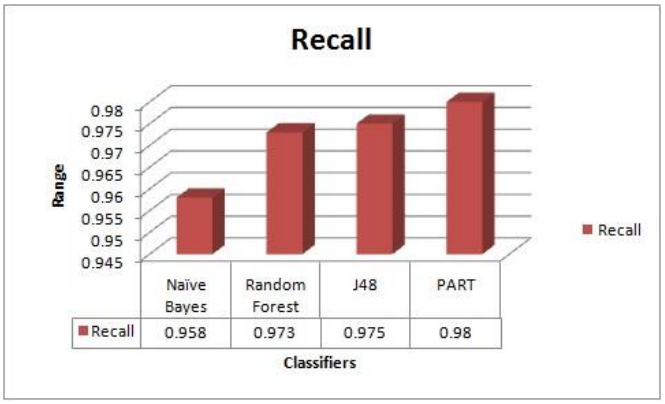

Figure 4 Recall over various classifiers

Figure 5 F-measure over various classifiers

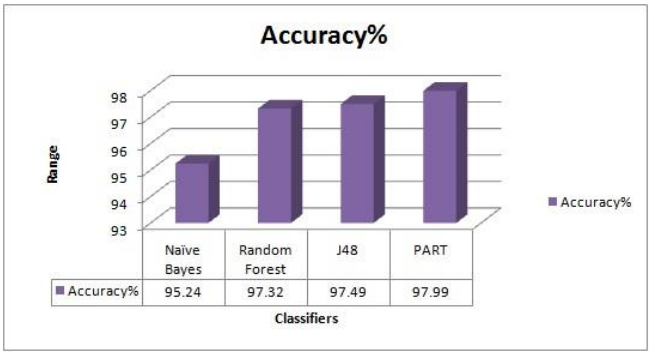

Figure 6 Accuracy over various classifiers

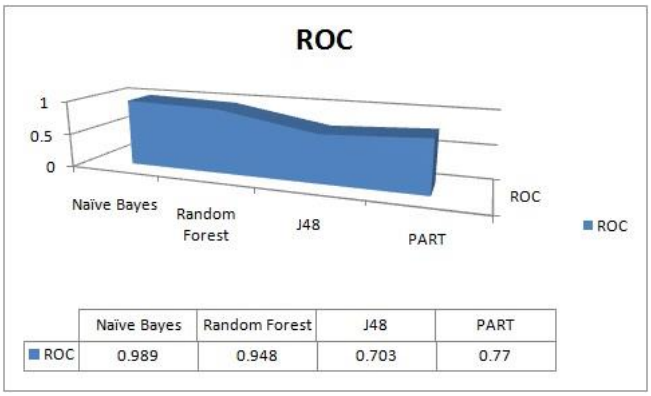

Figure 7 ROC of various classifiers

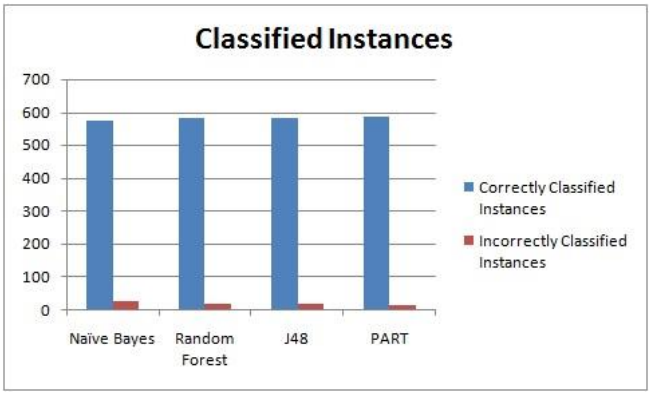

Figure 6 Classified instances over Various Classifiers

V. CONCLUSION

It is important to detect hyperthyroidism early. The aim of this research is to use a systemic approach to predict hyperthyroidism. This paper examines and evaluates four machine learning classifiers based on a number of parameters. Experiments are conducted using a dataset obtained from the UCI repository, which includes eight attributes. Overall, PART's accuracy is 97.99 percent .In future, the designed model can be extended to predict some other diseases.

VI. REFERENCES

[1]. 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- Vol. 2 Issue 6, June – 2013

[2]. Jeffrey W. Seifert, "Data Mining An Overview", CRS Report for Congress.

[3]. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufman, San Mateo, CA, 1993.

[4]. Satish N. Kulkarni, Dr. A. R. Karwankar, Thyroid disease detection using modified fuzzy hyperline segment clustering neural network, International Journal of Computers & Technology, Volume 3 No. 3, Nov-Dec, 2012

[5]. Orhan Er,Feyzullah and A.Cetin Tannkulu, Temurats,Tuberculosis Disease Diagnosis Using Artificial Neural Networks

[6]. Feyzullah Temurats,A comparative study on thyroid disease diagnosis using neural networks, Expert Systems with Applications,Volume 36, Issue 1, January 2009, Pages 944-949

[7]. dos Santos, A. M., Pereira, B. B., de Seixas, J. M., "Neural Networks: An Application for Predicting Smear Negative Pulmonary Tuberculosis", Proceedings of the Statistics in the Health Sciences, March, 2004.

[8]. 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.

[9]. R. Kohavi, A study of cross validation and bootstrap for accuracy estimation and model selection, in: Proceedings of the 14th IJCAI, Morgan Kaufmann, San Francisco, CA, 1995, pp. 338–345.

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

[11]. Yongqiang Cao, Jianhong Wu, "Projective ART for clustering data sets in high dimensional spaces", Elsevier Science Ltd, Neural Networks 15, 2002, pp. 105-120.