Performance Improvement with Multiple Approaches to Predict Disorders Caused by Thyroid Disease

[1]Dr.D.Anitha, Mrs.S.SathyaPriya[2]

[1]*Assistant Professor Department of Computer Science Sri Ramakrishna College of Arts and Science for women Coimbatore anithasuresh2003@gmail.com anithacs@srcw.ac.in*
[2]*Research Scholar Department of Computer Science Sri Ramakrishna College of Arts and Science for women Coimbatore sathyamitra@gmail.com*

Abstract: *Thyroid diseases are widespread worldwide. India there is a generic problems caused due to thyroid diseases. research studies emphasis that about 43 million people in India suffer from thyroid diseases. This work focused on developed of a predictive system for thyroid detection such as hypothyroidism, hyperthyroidism, sick people and normal people.Classification is one of the most significant data mining techniques. The supervised learning used to classify predefined data sets. Data mining technique is mainly used in healthcare organizations for decision making, diseases diagonosis and for better treatment to the patients. .It is a process to extract the data information from the huge amount of the data set classification is the analyze, processing time, sensitivity, error rate and classification accuracy. This paper describes to recognize the better method classification. As a result, the performance will be concluded for both classification techniques and their accuracy will be compared through confusion matrix. It has been concluded that J48 gives better accuracy than the other decision tree technique.*

Keywords: *Classification, Thyroid, Hyper Thyroid, Hypo Thyroid, Data Mining (DM), Data Pre-processing (DP).*

I. Introduction

According to an analysis "DAILY TIME OF INDIA" while one in ten adults in India's people is suffering from hypothyroidism. This estimation is found on the premise of an analysis conduct by India thyroid society. The study also alert for thyroid and thyroid is 9th ranked in comparison to other type common disease like asthma, cholesterol, depression, diabetes etc. medical practitioner say that thyroid are same as other disorders, however, the investigation population are alert to thyroid disorders, know that there are diagnostic tests for finding of this disease.The thyroid is a 2-inch-long, butterfly-shaped gland weighing less than 1 ounce. Located in the front of the neck below the larynx, or voice box, it has two lobes, one on either side of the windpipe. The thyroid is one of the glands that make up the endocrine system. The glands of the endocrine system produce, store, and release hormones into the bloodstream. The hormones then travel through the body and direct the activity of the body's cells.Thyroid gland is broken into two section (1) Normal category of thyroid gland (2) this category the gland create abnormal type of thyroid hormones like as hypothyroid and hyperthyroid.

Hypothyroidism (underactive thyroid or low thyroid) that is called the thyroid hormones are not generating as much as necessary of certain important hormones. Hypothyroidism can justification various health problems such as: heaviness, joint pain, unfruitfulness and heart disease. Hyperthyroidism (overactive thyroid) belongs to a position is the thyroid gland delivers a lack of the hormone thyroxin. For this situation, the body's digestion system is quickening essentially, bringing about sudden weight reduction, a fast or irregular heartbeat. Predictive system is a group of accurate systems from machine learning, & information mining that consider present and reliable reality to make opportunity concerning future or generally difficult to understand occasions. Disease determination is not easy task it requires lots of experience and knowledge.Thyroid hormone production is regulated by thyroid-stimulating hormone (TSH), which is made by the pituitary gland in the brain. When thyroid hormone levels in the blood are low, the pituitary releases more TSH. When thyroid hormone levels are high, the pituitary responds by decreasing TSH production. The unique task is to give problem ending at before time stages with higher correctness.Data mining is a unique part in medicinal field for disease finding. Show predictive detection system of thyroid or illness finding systems important part of data mining.

II. Overview Of Thyroid

Thyroid is worldwide. In India also, the thyroid gland is positioned at the lower section of the human neck that produce the hormones that help regulates body processes, including growth, energy balance, and heart

rate, which is produce two kind of thyroid hormones, levothyroxine (T4) and triidothyronine(T3) that hormones involve some function of the body such as stabilizing body temperature, blood pressure, changeable the heart rate Thyroid gland disease are different from other disease in conditions of their ease of diagnosis, accessibility, of health conduct, and the virtual visibility, In thyroid hormones is categorized into two common types Hyperthyroidism or hypothyroidism .

1. HYPERTHYROIDISM: When hormone is increase then this is generating reason of "hyperthyroidism". In healthcarefield, "hyper" indicates too much. Hyperthyroidism collects when the gland produces overload hormones. This is called as an overactive thyroid; the hormone overload can cause an extensive range of physical changes. Many symptoms have similarities during Hypothyroidism as well as lessening hair, arid skin and high temperature feeling. The symptoms that show the occurrence of hyperthyroidism such as weight failure in bad feeling of a good food eating, an increment of heart rate, high blood pressure, anxiety, large sweating, addition in your neck and shaking hands

2. HYPOTHYROIDISM: When hormone is decrease then this is generating reason of hypothyroidism. In healthcarefield, term hypo means short or not sufficient. For example, hypoglycemia is a term for low blood sugar. Hypothyroidism is a form that the thyroids do not create required hormones. Swelling and break to the gland reason hypothyroidism. Likes as: weight gain, to go down weight despite an accurate weight failure system. Table II shows the list of symptoms of hypothyroid. An increased risk of thyroid happens if there is a family unit history of thyroid like a type I diabetic, over 50 years of age and a stressful life].

1. **TYPES:**

According to modern medicine concept, based upon the etiology of disease, severity of disease and the time of occurrence of hypothyroidism7, it may be classified On the basis of etiology:

1. Primary hypothyroidism (thyroid gland defects)
2. Secondary hypothyroidism (pituitary gland defects) and
3. Tertiary hypothyroidism (hypothalamic defects).

On the basis of time of occurrence:

1. Adult onset hypothyroidism
2. Congenital hypothyroidism

On the basis of severity:

1. Subclinical hypothyroidism
2. Overt hypothyroidism

III. Methodology

The diagram of the proposed system is shown in figure 1 which consists of three main parts:

i.   Data Preprocessing.
ii.   Strong Feature selection.
iii. Predictive model development

DATA PREPROCESSING:

Data pre-processing (DP) is an often neglected but it is main process in the DM. The phrase is "Garbage in, Garbage Out", It is mostly applicable to information mining and machine learning, impossible data combinations (e.g. Gender is male or female, fever is yes or no) to the help recover the feature of the data, as to recover the efficiency and easy to the mining process.DP includes, cleaning, normalization, transformation, feature extraction and selection etc.

Raw data has lower signal to noise ratio, missing values, and inconsistency that affects the DM results, In classify to help improve the feature of the data, as to improve the efficiency and easy to the mining process. DM pre-processing has been following categories:

☐   Data Cleaning
☐   Data Integration
☐   Data Transformation
☐   Data Reduction

B.FEATURE SELECTION

Feature selection (FS) is also well-known as variable selection or attributes selection. It is the usable selection of attributes in dataset that are most important to the predictive modeling FS is the process to select a separation of important features for use in model construction. FS methods include reject attributes present in the dataset without changing them. In this work, we have applied Correlation Based feature selection (CFS) algorithm for attribute selection. CFS algorithm show highest valuation or ability of a separation of features. Thisheuristic

takes useful of features for predicting the class label along with the level of inter correlation among them. The hypotheses on which the heuristic is based can be stated.

C. MODEL EVALUATION:

Model evaluation is a basic step of the prediction system process. It will help to get the best result that provides our data and how to work well the chosen prediction system will work in the future case. Model evolution has two type of evaluating method in data mining1) Cross- validation 2) Hold-out. In this work, Hold out method is using for model development. In thismethod, the mostly large amount dataset is unplanned separated into two parts.

Preprocessing

Data preprocessing is a data mining technique. It is used to reduce the volume of data. There are many data reduction techniques are available such as data compression, numerosity reduction, dimensionality reduction and discretisation. In our work, we have used dimensionality reduction to select the subset of attributes from original data.

Classification

Classification is one of the data mining Technique. It is used to group the instances which belong to same class. It is a supervised learning, in which predefined training data is available. Most popular data mining classification techniques are decision trees and neural networks.

Decision tree

Decision tree is one of the classification technique in data mining. It is tree-like graph. [5] The internal node denotes a test on attribute, each branch represents an outcome of the test, and the leaf node represent classes. It is a graphical representation of possible solutions based on condition from these solutions optimum course of action is carried out. In our work, we have used two decision tree classifier such as decision stump and J48 to classify the hypothyroid data set.

Algorithms
1. J48 Algorithm

J48 is a tree based learning approach. It is developed by Ross Quinlan which is based on Iterative Dichotomiser 3 (ID3) algorithm.[5] J48 uses divide-and-conquer algorithm to split a root node into a subset of two partitions till leaf node (target node) occur in tree. Given a set T of total instances the following steps are used to construct the tree structure.

Step 1: If all the instances in T belong to the same group class or T is having fewer instances, than the tree is leaf labeled with the most frequent class in T.

Step 2: If step 1 does not occur then select a test based on a single attribute with at least two or greater possible outcomes. Then consider this test as a root node of the tree with one branch of each outcome of the test, partition T into corresponding T1 , T2 , T3 ...., according to the result for each respective cases, and the same may be applied in recursive way to each sub node.

Step 3: Information gain and default gain ratio are ranked using two heuristic criteria by algorithm J48.

2. Decision Stump

A **decision stump** is a machine learning model consisting of a one-level decision tree. That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature. Sometimes they are also called 1-rule.

IV. Experiments with Weka

The open source software Waikato Environment for knowledge Analysis 3.7(WEKA) is used for experiment. It is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, feature selection and visualization. Weka can downloaded from the website.

Performance Measure of Classifications

In our experiment data is supplied to classifier of J48 Algorithm and decision stump to classify the data. The classifiers performance is evaluated through Confusion Matrix.

Step-I: divide data into train set and test set

This step thyroid dataset is divided into two part 1) Training part 2) Testing part

In training part contain 60% thyroid dataset or in testing part contain 40 % data dataset shown in figure 2.
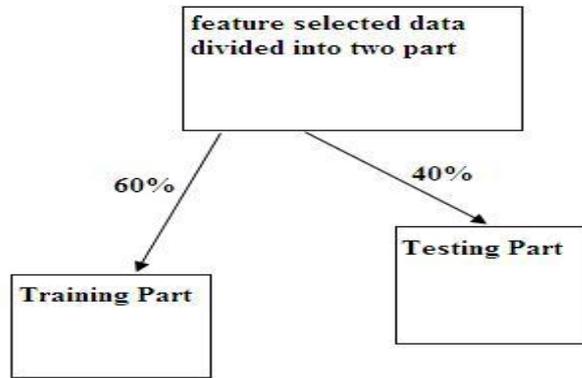
Data division into training and testing set.The hypothyroid dataset used in this work is collected from the website. The hypothyroid dataset consists of 3772 instances from which 3481 instances belongs to category negative,194 instances belongs to category compensated hypothyroid , 95 instances belongs to primary hypothyroid category while 2 instances belongs to category secondary hypothyroid. There are totally 30 attributes. In our research work we have taken only 12 attributes which will be used to classify the data.
The following Table 4 depicts detailed accuracy for J48 and decision stump algorithm

Classifier	Accuracy
Decision Stump	95.38%
J48	99.57%

Table 4: Accuracy of classifier

Accuracy of classifiers

In this chart, X axis represent the algorithm and Y axis represent the Accuracy. It shows that the Accuracy of decision stump is 95.38% and the Accuracy of J48 is 99.57% which is more than decision stump.
The Misclassification error rate is calculated by the following formula
Misclassification error rate = 1-Accuracy
The Misclassification error rate of classifier is shown below
Table 5: Error rate of classifier
The following chart2 shows the error rate of the classifier.

IV. Discussion

In WEKA, there are many classification techniques available. These classification techniques are used to diagnose the thyroid diseases and some other clinical diagnosis issues. Studies showed that many researchers used different methods to diagnose the thyroid disease and achieved the high accuracy of classifiers for the dataset is taken from UCI machine learning repository. [K.Saravana Kumar et al 2014], proposed KNN and SVM classification Algorithm on diagnosing the thyroid disease. They showed that the prediction accuracy of SVM is 94.4336%. However, KNN accuracy is 96.3430%. [Pandy et al 2015], proposed c4.5 and random forest classification Algorithm which gives prediction accuracy of 99.47%.

In our study, the hypothyroid dataset is taken from the website UCI machine learning repository. The hypothyroid dataset consists of 3772 instances from which 3481 instances belongs to

In this experiment, the classifier J48 is giving accuracy of 99.58% and minimum error rate of .424 which is better than decision stump.

V. Conclusion and Future Scope

Diagnosis of disease is a very challenging task in the field of health care. Many data mining techniques are used in decision making process. In our work, we have used dimensionality reduction to select the subset of attributes from original data and we have applied J48 and decision stump data mining classification techniques which are used to classify the hypothyroid disease. The performance of classifiers are evaluated through the confusion matrix in terms of accuracy and error rate. The J48 Algorithm gives 99.58% which is providing better Accuracy than decision stump tree accuracy and also J48 Algorithm gives very minimum error rate than Decision stump. As a future work the same technique is used to apply for other disease datasets such as heart disease, breast cancer, Lung cancer

References

[1]   Jiawei Han, KamberMicheline (2009). Datamining: Concepts and Techniques, Morgan Kaufmann Publisher.
[2]   EBRU TURANOGLU-BEKAR, GOZDE ULUTAGAY, SUZAN KANTARC-SAVAS, "Classification of Thyroid Disease by Using Data Mining Models: A Comparison of Decision Tree Algorithms", Oxford Journal of Intelligent Decision and Data Science, PP: 13-28, 2016.
[3]   N MOHANA SURDARAM, VRENUPRIYA, "Artificial Neural Network Classifiers for Diagnosis of Thyroid Abnormalities", International Conference on Systems, Science, Control, Communication, Engineering and Technology, PP: 206-2011, 2016.
[4]   VIKRAM V HEGDE, DEEPAMALA N, "Automated Prediction of Thyroid Disease using ANN", International Journal of Innovative Research in Science, Engineering and Technology ,Volume 5, Special Issue, PP: 268, 2016.
[5]   SHAIK RAZIA and M. R. NARASINGA RAO, "Machine Learning Techniques for Thyroid Disease Diagnosis - A Review", Indian Journal of Science and Technology, Volume 9, PP: 1- 9, 2016.
[6]   FEYZULLAH TEMURTAS, "A Comparative Study On Thyroid Disease Diagnosis Using Neural Networks", Expert Systems with Applications PP: 944–949, 2016.
[7]   SERAVANA KUMAR, Dr. ADIMULAM YESUBABU, S ANUPKANT, "Autoimmune diseases - a wider perspective", International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 08, PP: 317-323, 2015.
[8]   VADAMODULA PRASAD, THAMADA SRINIVASA RAO, ANKIT KUMAR SURANA, "Standard Cog Exploration on Medicinal Data", International Journal of Computer Applications, Volume 119 – No.10, PP: 34- 38, 2015.
[9]   NILESH BORISAGAR, PROF.AMIT MARU, "ANN in Medical Field Survey Paper", International Journal of Innovative Research in Computer and Communication Engineering, Volume 3, Issue 9, PP: 8210- 8216, 2015.
[10]   PRERANA, PARVEEN SEHGA, KHUSHBOO TANEJA, "Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network", International Journal of Research in Management, Science & Technology, Volume 3, No. 2, ISSN: 2321-3264, PP: 75:80, 2015
[11]   I Md. DENDI MAYASANJAYA, HANUNG ADI NUGROHO, NOOR AKHMAD SETIAWAN, "a comparison of classification methods on diagnosis of thyroid disease", PP: 89-93, 2015.
[12]   K.SARAVANA KUMAR and DR. R. MANICKA CHEZIAN, "Support Vector Machine and K- Nearest Neighbor Based Analysis for The Prediction Of Hypothyroid", International Journal of Parma and Bio Sciences, ISSN 0975-6299, PP: 447-453, 2014.
[13]   K. RAMA LAKSHMI and S.PREM KUMAR," utilization of data mining techniques for prediction and diagnosis of major life threatening Diseases survivability", International Journal of Scientific & Engineering Research, Volume 4, Issue 6, ISSN 2229-5518, PP: 923-932, 2013.
[14]   SHIVANEE PANDEY, ROHIT MIRI, S. R.TANDAN, "Diagnosis and Classification of Hypothyroid Disease Using Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), Volume 2 Issue 6, ISSN: 2278-0181, PP: 3188- 3193 , 2013.
[15]   JASDEEP SINGH BHALLA, ANMOL AGGARWAL, "A Novel Method for Medical Disease Diagnosis Using Artificial Neural Networks Based on Back propagation Algorithm", IEEE, PP: 55-61, 2013.
[16]   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, Volume 12, PP: 13-19, 2012.
[17]   K. SARAVANA KUMAR, Dr. R. MANICKA CHEZIAN, "Analysis on Suspicious Thyroid Reorganization Using Association Rule Mining", Journal of Global Research in Computer Science, 2012.
[18]   FATEMEH SAITI , AFSANEH ALAVI NAINI, MAHDI ALIYARI SHOOREHDELI, MOHAMMAD TESHNEHLAB, "Thyroid Disease Diagnosis Based on Genetic Algorithms using PNN and SVM", IEEE, PP:1-4, 2009
[19]   LALE OZYILMZ, TULAY ULDIRIM, "Diagnosis of Thyroid Disease Using Artificial Neural Network Methods" IEEE In the proceedings of the 9th International Conference on Neural Information Processing (ICONIP'OZ), Volume 4, pp: 2033-2036, 2002.
[20]   https://en.wikipedia.org/wiki/Predictive_analytics.
[21]   http://archive.ics.uci.edu/ml/support/thyroid+Disease.