International Journal of Pure and Applied Mathematics
Volume 119 No. 12 2018, 13881-13890
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Classification of Thyroid Disease Using Data Mining Techniques

Sumathi A, Nithya G and Meganathan S

Department of Computer Science and Engineering, SRC

SASTRA Deemed to be University, Tamil Nadu, India.

***Corresponding Author: sumathi@src.sastra.edu**

Abstract:

Disease is increasing day by day due to lifestyle. Especially, thyroid disease is a very common disease among humans. Thyroid hormone regulates numerous metabolic mechanism throughout the body. Female is more affected than male due to thyroid disease. Two types of thyroid diseases are i) Hyperthyroidism-produces a lot of thyroid hormone in the blood and ii) Hypothyroidism-produce less thyroid hormone in the blood. Hypothyroidism is a condition which underlies not only chronic degenerative diseases but also hormone irregularities and results in a weakened immune system. People taking the right dosage of thyroid medication and have normal thyroid levels, but still suffering from hypothyroidism. Subtype classification is necessary for better diagnosis. Some additional tests such as RT3 and basal metabolic temperature will improve the efficiency of diagnosis. Hence this work proposes to classify the types of Thyroid disease and its subtypes in an efficient way using various data mining approach.

Keywords: *Preprocessing, Feature selection, EM Clustering, J48 Classification.*

Introduction

The thyroid[1] is the most familiar disease among the humans and it is not a deadly disease. The thyroid gland shapes look likes butterflies and that is located at the bottom of the neck. Higher or lower construction of thyroid hormones, triiodothyronine, and levothyroxine are produces form the thyroid gland. T3 and T4 are controlled by TSH and they are generated by the brain. Hypothyroidism[2] is underactive thyroid that the glands can't generate enough thyroid hormones and Hyperthyroidism[3] is an overactive thyroid that the gland can produce more thyroid hormone. Hypothyroid symptoms include constipation, weight gain, slowed heart rate, fatigue, depression, dry skin, puffy face, muscle weakness, thinning hair and hyperthyroid symptoms include weight loss, irregular heartbeat, nervousness, difficulty sleeping, hair loss, shortness of breath. Disease diagnosis is a very complex and challenging task. To identify the disease on correct time helps to give the proper treatment. In earlier days the disease is diagnosed by the symptoms. But now doctors are diagnosing the disease by

various intelligent systems. People taking the correct dosage of thyroid medication and have normal thyroid levels, but still suffering from hypothyroidism.

Diagnosis of the subtypes of hypothyroidism are necessary to give accurate result to the patients. Triiodothyronine and levothyroxine levels are low and TSH level is high means overt (primary) hypothyroidism[4]. And, high TSH levels and normal T4 levels are called subclinical hypothyroidism[5]. Also, normal TSH levels and flat T4 levels are called central hypothyroidism.

Some additional tests RT3 and basal metabolic temperature will improve the efficiency of diagnosis. This paper proposes to analyze the various classification algorithms which are used to diagnose the accurate result of thyroid disease and to evaluate their efficiency, Precision, and Error-Rate.

Preprocessing:

In a data mining preprocessing is an essential task to change the incomplete data into a readable format. Raw Data may be inconsistent, noisy and missing values. It affects the results. Before applying for any data mining mechanism, data must apply preprocessing that will helps to improve mining process and efficiency For thyroid disease dataset the missing values will be replaced by mean value. For example
Mean (x) = sum of all the values $\sum$ (x) / number of values (n).
Let x= 145,136,250,356,318,104 n=6

Mean =(145+136+250+356+318+104) / 6 = (1309) / 6 =218.17

T3	T4	TSH
1.5	61	14.8
1.6	82	15
2.2	83	19
2.0	95	7.8
2.2	113	8.8
?	70	9.4
2.2	74	6.1
1.7	75	16
2.2	117	9
2.4	99	11.4

In this example for handling missing values using hypothyroid dataset.

Mean (T3) = (1.5+1.6+2.2+2.0+2.2+2.2+1.7+2.2+2.4) / 9 = (18) / 9 = 2

T3	T4	TSH
1.5	61	14.8
1.6	82	15
2.2	83	19
2.0	95	7.8

2.2	113	8.8
2	70	9.4
2.2	74	6.1
1.7	75	16
2.2	117	9
2.4	99	11.4

Feature Selection:

In the machine learning feature selection is otherwise called attribute selection[6]. It is the technique of choosing the relevant features and simplify the models to understand the users. It is used to take shorter training time. It is different from feature extraction. In this thyroid dataset to decrease the dimension using correlation attribute evaluation method. It is used to compute the relation between each attribute. It is work in ranker search method. Select the relevant attributes that are moderate to negative or positive correlation and remove those attribute values closer to zero. Finally, to take the selected attributes.

Feature Construction:

Feature construction is a mechanism that frames intermediate features from the initial dataset. The aim of this is to build more efficient features for data mining task. In thyroid dataset by adding RT3 and Basel metabolic temperature attributes which are used to diagnose the hypothyroidism and its subtypes in an efficient way[7].

Reverse Triiodothyronine:

Reverse triiodothyronine (RT3) is measured by blood test. The liver can regularly convert the T4 hormone to RT3. It means 40% of T4 convert into T3 then 20% of the T4 convert to RT3[8]. To calculate the RT3 ratio FT3 and RT3 levels is essential. If the result of ft3 is smaller number means to multiply that value by 100. If the ratio is > 20 means no problem, otherwise it makes the RT3 problem.

Example:1

In this example to assume that ft3 value is 3.25 and rt3 value is 22.

So, we get ft3 value is 3.25 * 100 = 325.

Then we get the new ft3 value by the rt3 value.

Hence, simply get 325 divides by 22 =14.77

This is RT3 ratio and it is below 20. So, there is an RT3 problem.

Example: 2

In this example to assume that ft3 value is 3.25 and rt3 value is 14.

So, we get ft3 value is 3.25 * 100 = 325.

Then we get the new ft3 value by the rt3 value.

Hence, simply get 325 divides by 14 =23.21

This is RT3 ratio and value is greater than 20. So, RT3 problem.

Basal Metabolic Temperature Test:

To diagnose the hypothyroid disease using basal body temperature[9]. Mercury thermometer is used to take the body temperature in four consecutive days. If the body temperature is less than 36.5 degrees Celsius means it indicates hypothyroidism and TSH level is coming back as 2.0 m U/L or higher.

Example:1

Change the body temperature from Celsius to Fahrenheit by the following formula

°F = (9/5 °C) + 32°

Converting 35.5 degrees Celsius to Fahrenheit:

1) 35.5 x 9 = 319.5

2) 319 divided by 5 = 63.8

3) 63.8 + 32 = 95.78

Clustering:

Clustering is the concept of grouping the set of similar object in one group and dissimilar objects belonging to the other group. It is considered by the unsupervised learning problem. This will helps to grouping the parameter in the dataset. Expectation-Maximization (EM)[11] is the mixture based algorithm that is used to discover the maximum estimate attributes in probabilistic models. It is used to cluster the given thyroid dataset. Cllustering techniques are used to improve the accuracy of classifier.

Step 1: First initialize the attributes to some random values.
Step 2: Calculate the probability for each possible valus.
Step 3: Then use the calculated values to compute better estimate for the attributes.
Step 4: Repeat step 2 and 3 until the convergence.

Classification:

Classification is a machine learning task that predicts the classes according to some constraints. Supervised learning is a classification algorithm in data mining. The main desire of the classification issue is to diagnose the class for new data. Various classification algorithms are used to diagnose the classes. In this thyroid dataset using C4.5[2] classification algorithm it's give the better accuracy for the results.

Decision Tree:

Decision tree is a method in data mining which is also called classification tree. The aim is to predict the target value for the variable based on some input. It is used for the decision making and easy for understanding of the users and to split the source set into the subset based on the attribute. Each of the internal node has been labeled with input feature. Information gain ratio is applied for each subset and it will display the results.

J48:

J48(C4.5)[2] algorithm is used to develope decision tree. It is the continuation of ID3 algorithm. This algorithm is mainly used for classification purpose. Handling the training dataset with missing values that is calculate by the gain value.

$$GainRatio\,(A) = \frac{Gain(A)}{SplitInfo\,(A)}$$

Gain Ratio is utilized for splitting the dataset. To select the highest gain ratio for splitting attribute.

$$SplitInfo_{A}(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{D})$$

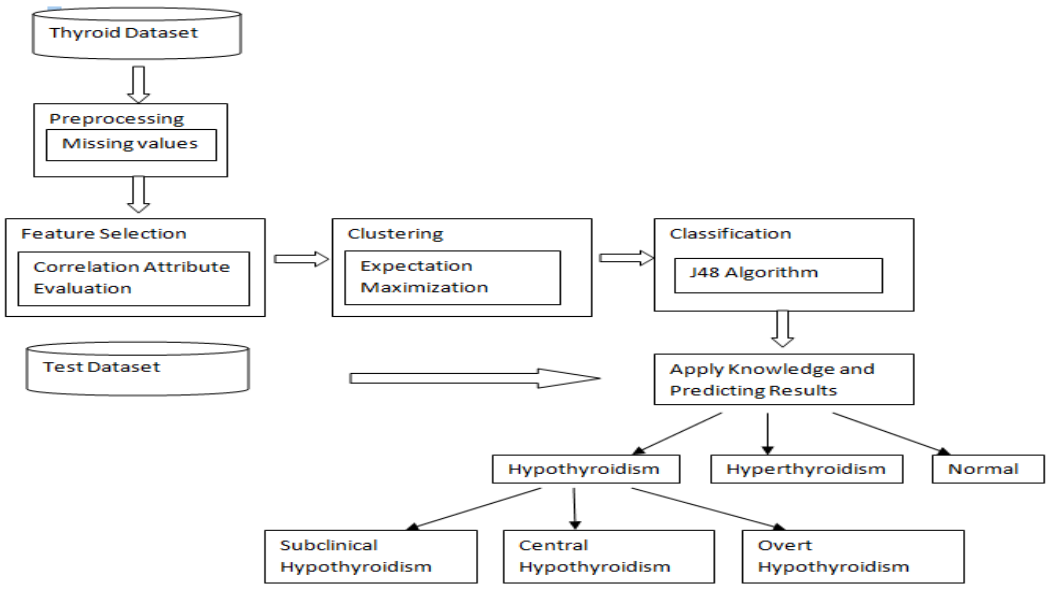
Step 1: Let T be the the training dataset. i.e.,(Thyroid disease dataset)

Step 2: Compute information gain for all parameter and split on it.

Step 3: Let T_best be the parameter with the topmost normalized information gain.

Step 4:Compute a node that divides on T_best.

Step 5: Repeat splitting on T_best, and add to the children node[10].



Experimental Results:

In this section the result will be reported. Thyroid dataset is obtained from the UCI repository and it contains 7200 records. The prepocessing and feature selection techniques are applied to that dataset. The processed data will be given to an input to the clustering technique. This method is used to give the better accuracy of classifier. Applying J48 algorithm for the thyroid dataset is used to diagnose the disease. To analyse the data results visually and show it in the following figures.

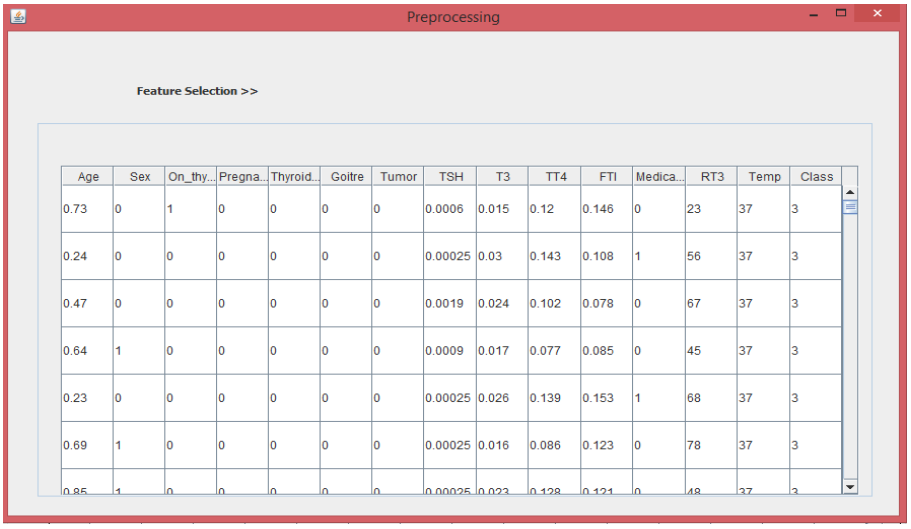


Figure 1: Prerocessing

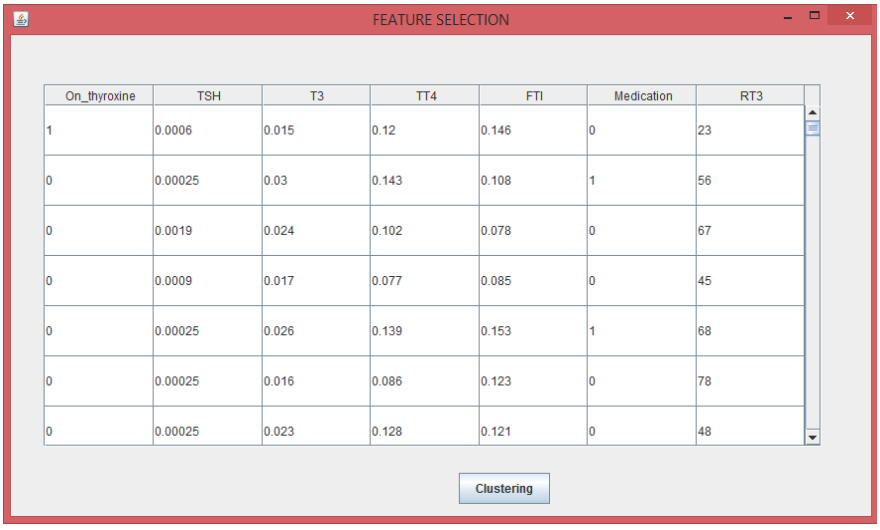


Figure 2: Feature selection

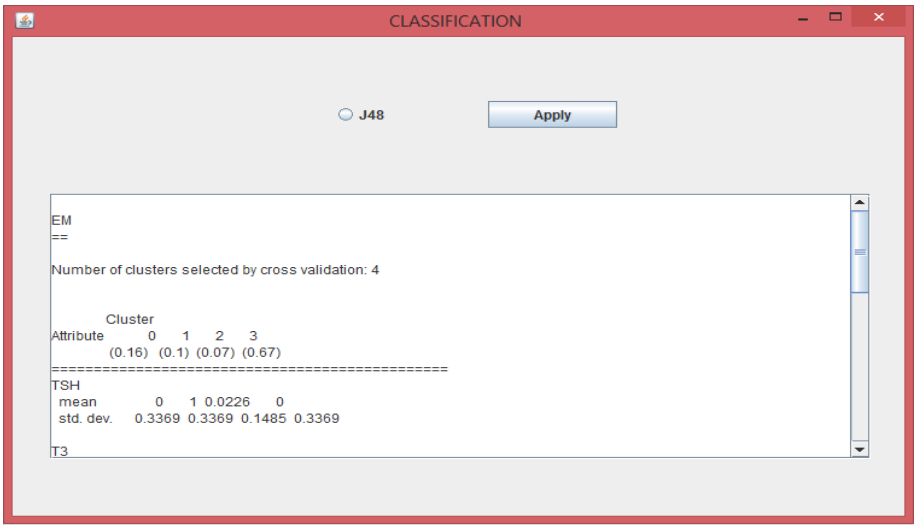


Figure 3: Clustering

Figure 4: Classification

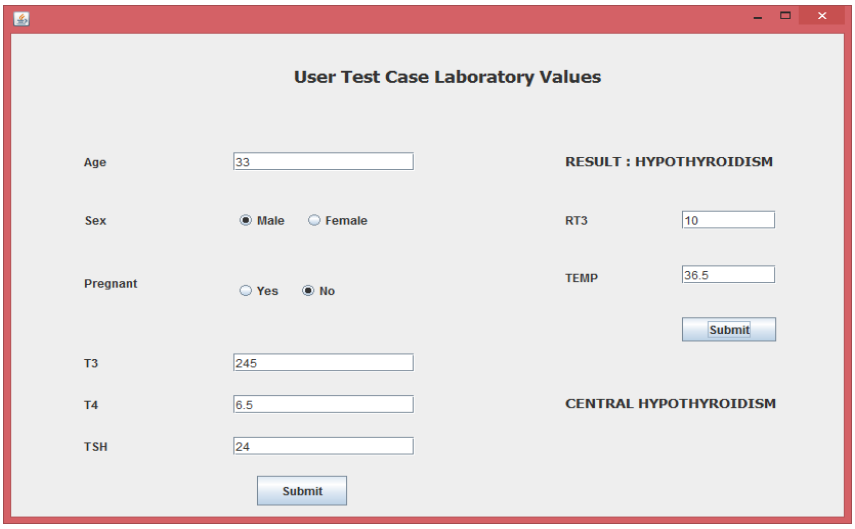


Figure 5:Diagnosis Result

Conclusion:

To analyse the medical dataset, various data mining and machine learning techniques are available. An important aspect of medical data mining is to increase the accuracy and efficiency of disease diagnosis. In this work EM clustering algorithm and J48 classification algorithm are used to classify the thyroid diseases and its subtypes in an efficient way. The additional attributes RT3 and Basel Metabolic temparatue are used to diagnose the subtypes of hypothyroidism. The proposed technique gives better precision, recall and classification accuracy(99.25%) for the given dataset.

References:

[1] Y. Hayashi, "Synergy effects between grafting and subdivision in Re-RX with J48graft for the diagnosis of thyroid disease," *Knowledge-Based Syst.*, vol. 131, pp. 170–182, 2017.

[2] P. Durga, V. S. Jebakumari, and D. Shanthi, "Diagnosis and Classification of Parkinsons Disease Using Data Mining Techniques," *ISSNOnline) Int. J. Adv. Res. Trends Eng. Technol.*, vol. 3, no. 14, pp. 2394–3777, 2016.

[3] D. Liping, Y. Lifen, D. Ruihong, and H. Zhihong, "Analysis on the quality of life of the hyperthyroidism patients," *Proc. 2011 Int. Conf. Hum. Heal. Biomed. Eng. HHBE 2011*, pp. 791–794, 2011.

[4] P. Rotman-pikielny, O. Borodin, R. Zissin, and Y. Levy, "Overt Hypothyroidism in Hospitalized Patients : Clinical Characteristics," vol. 2013, no. May, pp. 128–131, 2013.

[5] D. Khandelwal and N. Tandon, "Overt and Subclinical Hypothyroidism.," *Drugs*, vol. 72, no. 1, p. 17–33 17p, 2012.

[6] M. Ashraf, G. Chetty, and D. Tran, "Feature Selection Techniques on Thyroid , Hepatitis , and Breast Cancer Datasets," vol. 3, no. March, pp. 1–8, 2013.

[7] https://www.igi-global.com/dictionary/feature-construction/10959

[8] http://www.holistic-hypothyroidism-solutions.com/rt3-ratio.html

[9] http://www.holistic-hypothyroidism-solutions.com/support-files/basal-metabolic-temperature-test.pdf

[10] https://en.wikipedia.org/wiki/C4.5_algorithm#Algorithm

[11] https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm