ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

# Survey on Thyroid Diagnosis using Data Mining Techniques

S. Sathya Priya[1], Dr. D. Anitha[2]

Research Scholar, Dept. Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore[1]

Assistant Professor, Dept. Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore[2]

**Abstract:** Recently, thyroid diseases are more and more spread worldwide. India, for example, one of eight women suffers from hypothyroidism, hyperthyroidism or thyroid cancer. Factors that affect the thyroid function are: stress, infection, trauma, toxins, low-calorie diet, certain medication etc. It is very important to prevent such diseases rather than cure them, because the majority of treatments consist in long term medication or in chirurgical intervention. The current study refers to thyroid disease classification in two of the most common thyroid dysfunctions (hyperthyroidism and hypothyroidism) among the population. The authors analyzed and compared four classification models: Naive Bayes, Decision Tree, Multilayer Perceptron and Radial Basis Function Network. The results indicate a significant accuracy for all the classification models mentioned above, the best classification rate being that of the Decision Tree model.

**Keywords:** Data Mining, Classification Model, Thyroid Diseases, Neural Network, Decision Tree, Naïve Bayes, Chi Square

## I. INTRODUCTION

Thyroid Disease diagnosis is one of the very difficult and deadly tasks, because it needs lots of experience and knowledge.The traditional ways for diagnosis thyroid disease is doctor’s examination or a number of blood tests. Mainly task is to provide disease diagnosis at early stages with higher accuracy. Data mining plays a vital role in medical field for disease diagnosis. It offers lot of classification techniques to predict the disease accuracy. Hospitals and clinics gathered a large amount of patient data over the years.. Prevention in health care is a continuous concern for the doctors and the correct diagnostic at the right time for a patient is crucial, due to the implied risk. Recently, the usual medical report can be accompanied by an additional report given by a decision support system or other advanced diagnosis techniques based on symptoms. Questions such as: “what are the most important factors that affect thyroid?”, “which is the category of the population predisposed to goiter disease?”, “what is the most adequate treatment for a certain disease?” etc. may find answers in applying data mining techniques. Health care data can be processed and after rigorous usage can provide knowledge used in decision making, diagnosing diseases more rapidly and accurately, offering better medication for patients and minimizing the death risk. The authors focus their work on using classification methods and identifying the best algorithm for classification thyroid disorders. Thyroid is a butterfly-shaped gland, which is located at the bottom of the throat responsible for producing two active thyroid hormones, levothyroxine (T4) and triiodothyronine (T3) that affect some functions of the body such as: stabilizing body temperature, blood pressure, regulating the heart rate etc. Reverse T3 (RT3) is manufactured from thyroxine (T4), and its role is to block the action of T3. An abnormal function of the thyroid implies the occurrence of hyperthyroidism and hypothyroidism, two of the common thyroid affections. Hypothyroidism (underactive thyroid or low thyroid) means that the thyroid gland doesn't produce enough of certain important hormones. Without an adequate treatment, hypothyroidism can cause various health problems such as: obesity, joint pain, infertility and heart disease. Hyperthyroidism (overactive thyroid) refers to a condition in which the thyroid gland produces too much of the hormone thyroxin. In this case, the body's metabolism is accelerating significantly, causing sudden weight loss, a rapid or irregular heartbeat, BRAIN. Broad Research in Artificial sweating, and nervousness or irritability (eMedonline, 2016). In Figure 1 are presented the main factors that affect the thyroid function. It is obvious that factors such as stress, infection, toxins, trauma and certain medication are directly responsible for the improper production of thyroid hormones. Symptoms identification and the early detection of abnormal values of thyroid hormones after clinical investigation will help in establishing the proper diagnostic and to prescribe the right medication. The patient must periodically evaluate his clinical state in order to receive the treatment as long as he needs it.

## II. LITERATURE REVIEW

Thyroid function diagnosis is an important classification issue. Proper interpretation of the thyroid data, along with clinical examination and investigation, is a considerable problem in the detection as well as diagnosis of thyroid disease. In this literature[9], the author considers the Thyroid data set with multi class and proposes the classification for thyroidism in a separate layer. In this work, a multi classification approach for detecting thyroid attacks is designed to achieve higher efficiency and to

DOI 10.17148/IJARCCE

improve the detection and classification accuracy. This method finds that the method NNge provides higher efficiency to classify the thyroid attacks. Further in the literature [10] several methods of feature selection and classification for thyroid disease diagnosis, which is one of the most important classification problems, was proposed. In literature [11] there is an introduction of Bayesian association rule mining algorithm (BAR) that uses the Apriori association rule mining algorithm with Bayesian networks. Two interesting-ness measures of association rules: Bayesian confidence (BC) and Bayesian lift (BL). In literature [12] the author has proposed that Support Vector Machine (SVM) and K nearest Neighbour (KNN) are the two important modes applied to the prediction of hypothyroid. This paper discusses those predictions of Hypothyroid using K- Nearest Neighbour better than the Support Vector Machine. In literature[13].there is an introduction of Algorithms work on USG, SPECT images and planar scintigraphy .Pre-processing step, segmentation step, feature extraction step, feature selection step, classification step for thyroid disease diagnosis are used. In the past few years, numerous image processing algorithms have been proposed for efficient and effective detection of thyroid nodules. So Fuzzy cognitive map based decision support system and other recently proposed methods are presented in the paper. Texture representation via noise resistant image features is used. The author in literature [14], has proposed a new hybrid structure in which Neural Network and Fuzzy Logic are combined and its algorithm is developed.A. Data Mining

Simply stated, data mining refers to “extracting” or “mining” knowledge from large number of data, which is used to process the inconsistence data automatically and find the best data. Data Mining having two categories, they are Data Mining in descriptive - summarize the general properties of the data in database, Data Mining in Predictive – to predict the inference of the present data.

## 2.1 Classification

Classification is the process of predicting output based on some given input data. The goal of classification is to accurately predict the target class for each case in the data [5].In order to predict the data, it processes the training set and predictive set. It first develop relationships between the attributes of training data set .Then it is provided with the predictive data set, which contains similar attributes but with different data values, Then it analyze the given data and produce prediction by placing the different data sets in different classes based on the relationship of attributes.

### 2.1.1 Decision trees

Decision tree is similar to flow chart in which every non-leaf node denote a test on a particular attribute and every branch represent a outcome of the test. Root node is the topmost node in the decision tree. For example, with the help of readmission tree, we can decide whether a patient needs to be readmitted or not. Using Decision Tree, a decision maker can choose best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain [8] [9]. Decision tree are self explanatory and easy to follow. Set of rules can also be constructed with the help of decision tree. Decision Tree can be considered as nonparametric method because there is no need to make assumptions regarding distribution of space and structure of classifier. Decision tree have several disadvantages. These are: Most of the algorithm like ID# and C4.5 require target attributes to have discrete values as decision tree use divide and conquer strategy. More the complex relationship among attributes lesser is the performance.

### 2.1.2 Support Vector Machines

Vladimir Vapnik first introduced idea of Support Vector Machine [10]. Its accuracy is better than all other available techniques. It was first introduced for binary classification problems; but it can be further extended to multi class problems. It creates hyper-planes to separate data points[11].

It can be implemented in 2 ways:

1. Mathematical programming
2. Using kernel functions

With the help of training data sets, non linear functions can be easily mapped to high dimensional space. This can only be possible using kernel functions like Gaussian, sigmoid etc.

### 2.1.3 Neural Network

It was developed in 20th century. Neural network was regarded as the best classification algorithm before the introduction of decision tree and SVM which has far better results. This was the reason that encouraged NN as the most widely used classification algorithm in various bio-medicine and health care fields. For Example NN has been used as the algorithm supporting the diagnosis of diseases like cancer and predict outcomes. In NN, basic elements are nodes or neurons. These neurons are interconnected and within the network they work together to produce the output functions. They are fault tolerant as they are capable of producing new observations from the existing observations in those situations where some neurons within the network fail. An activation number is associated with each neuron and a weight is assigned to each edge with in the NN. The basic property of NN is that it can minimize the error by adjusting its weights and by making changes in its structure as it is adaptive in its nature. One major advantage of NN is that it can properly handle noisy data for training and can reasonably classify new type of data which is different from training data. There are also various disadvantages of NN. First, it require many parameters including the optimum no of hidden layer nodes that are empirically determined and its classification performance is very sensitive to parameters selected.

Second, its training or learning process is very slow and expensive.

### 2.2 Image Mining

Image Mining uses different algorithms and techniques to process the digital image data. The most common image process steps are,

- image preprocessing
- Segmentation
- Feature extraction
- Feature selection and classification

For example, thyroid US image is taken. Image preprocessing which helps to remove the noisy and inconsistence data. Segmentation process helps to segment the image used to apply object and boundaries. Each segment is characteristics by a color, texture, and intensity. Classification process use different techniques to predict the data like Decision tree, SVM, naïve bayse. Thyroid Diagnosis is based on signs, symptoms and physical examination of patient. Predictive data mining plays a vital role in disease diagnosis. This paper shown survey on thyroid disorder from various papers and gives the idea for the future work. The rest of this paper organized as below: section II as different thyroid disorders and their symptoms, section III as Literature survey, section IV as Classification Techniques in data mining and section V as conclusion and reference.

## III. DIFFERENT THYROID DISEASE AND THEIR SYMPTOMS

Thyroid gland secretes thyroid hormones to control the body's metabolic rate. The malfunction of thyroid hormone will leads to thyroid disorders. The thyroid or the thyroid gland is an endocrine gland. The thyroid gland releases thyroxine (T4) and triiodothyronine (T3) into the blood stream as the principal hormones. The functions of the thyroid hormones are to regulate the rate of metabolism and affect the growth. There are four main types of thyroid diseases hyperthyroidism ( too much thyroid hormone ), hypothyroidism (too little thyroid hormone), benign (noncancerous) thyroid disease and thyroid cancer(malignant). The symptoms of hypothyroidism includes fatigue, mental fogginess and forgetfulness, feeling excessively cold, constipation, dry skin, fluid retention, non specific aches and stiffness in muscles and joints, excessive or prolonged menstrual bleeding (menorrhagia), and depression. Hyperthyroidism can be observed with different signs and symptoms. Common symptoms of hyperthyroidism includes excessive sweating, heat intolerance, increased bowel movements, tremor (usually a fine shake), nervousness, agitation, rapid heart rate, weight loss, fatigue, decreased concentration and irregular and scant menstrual flow.

### 3.1 Data Mining in Health Care

Data mining refers to extracting unknown patterns from an enormous volume of data involving different methods and algorithms which exist at the intersection of fields such as artificial intelligence, machine learning, statistics and database systems (Piatetsky-Shapiro & Parker, 2011). Hospitals, clinics and medical analysis laboratories accumulate a large amount of patient data over the years. These data provide a basis for the analysis of risk factors for many diseases (various types of cancer, heart diseases, diabetes, hepatitis etc.). In literature are mentioned certain applications of data mining techniques in the health domain, some of them being presented in the following paragraphs. The authors have narrowed their research area on thyroid disorders and the examples given below are strictly about the related work described in literature, regarding the application of data mining for these classes of diseases. The majority of examples refer to diagnosing diseases of thyroid using decision trees, artificial neural networks, support vector machine, expert systems etc.

## IV. EXPERIMENTAL ANALYSIS

### 4.1 Dataset

As mentioned earlier, we use the publicly available UCI thyroid disease dataset. The directory contains 6 databases, corresponding test set, and corresponding documentation. We chose one of the datasets that had 29 attributes and applied Feature Selection technique i.e Chi-Square, The dataset is then filtered by applying the unsupervised discredited filter on the attributes to convert the continuous values into nominal. These 10 attributes are as follows

| Class | Sick Negative |
|---|---|
| Age | Young,middle,old |
| Sex | M,F |
| On thyroxine | F,t |
| Sick | F,t |
| Goiter | F,t |
| TSH | Nominal |
| T3 | Nominal |
| TT4 | Nominal |
| T4U | Nominal |

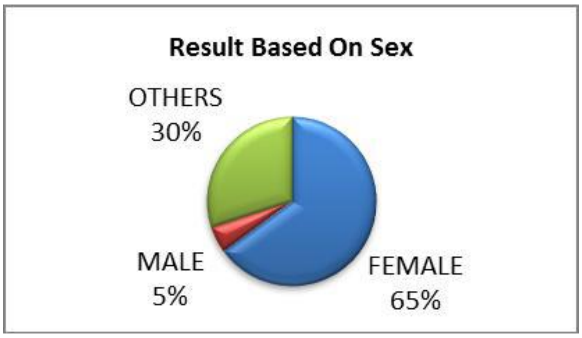

Fig.1 Result Based on Sex from Association Rules

### 4.2 Results of Association Rules

In this experiment, all sick individuals were regarded to be in one class and healthy individuals to be in another class. Two popular association rule mining algorithms, used in this experiment. Rules with confidence levels above 90% are selected having sick and negative classes on the right hand side of the rules

1) Based on Gender: For Apriori , majority of the sick rules were given to female gender indicating that females have more chance of having thyroid disease. Rules mined for sick class on the other hand showed that males have more chance of being free from thyroid disease. As mentioned earlier, in contrast to Apriori technique, that selects rules on the basis of confidence, Predictive Apriori selects rules based on accuracy. Similar to Apriori, most of the rules for sick class were attributed to females and negative class were attributed to males. However, the factors in the LHS varied. Overall females are seen to have more risk of developing thyroid disease.

2) Based on Age: On analysing the association rules from both Apriori and Predictive Apriori algorithm; majority of the sick rules were attributed to old indicating that old people have greater chance of having thyroid disease. Similarly in the case of Healthy Rules ; majority of the rules had young age on the L.H.S of the rules indicating that young have least chance of having thyroid disease.

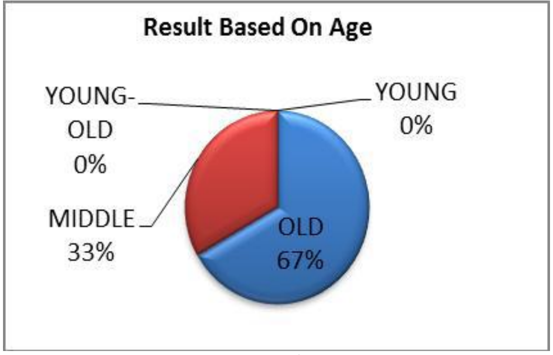

Fig.1 Result Based on Age from Association Rules

## V. CONCLUSION

As the medical reports show serious thyroid dysfunctions among the population, more affected being women, thyroid classification is a very important subject for researchers in medical science. As number of techniques of data mining are being used by various researchers in healthcare sector, so number of publications in this field are increasing. A tool is required to evaluate and summarize all the research work done so far in this particular field. A systematic literature review of all the data mining techniques used in health care is provided in this paper.. A number of data mining techniques such as Decision tree classification, Support Vector Machine classification, Linear regression, Hierarchical clustering are the techniques that are mainly used by researchers as they provide high accuracy and efficiency. Our future research in this direction will try to propose a novel data mining technique that can provide better accuracy in wide variety of disease in comparison to peer available techniques.

## REFERENCES

[1] J. Han and M. Kamber, Data Mining Concepts and Techniques, 2nd ed., Elsevier.

[2] R. Agrawal and R. Srikant., Fast Algorithms for Mining Association Rules, IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120.

[3] T. Scheffer, “Finding Association Rules that Trade Support Optimally Against Confidence”, Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery, Springer-Verlag London, UK 2001

[4] M. Tiwari, R. Singh et al. ”Association–Rule Mining Techniques: A general survey and empirical comparative evaluation” , International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 10, December 2012.

[5] Dr. M.P.S Bhatia and D. Khurana, “Experimental study of Data clustering using k-Means and modified algorithms”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.3, May 2013.

[6] The Weka website. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[7] A. Gopalakrishnan Unnikrishnan and U.V. Menon, “Thyroid disorders in India: An epidemiological perspective”, International Journal of Scientific and Research Publication, 29-Jul-2011, volume 15, issue 6, pg no.(78-81).

[8] Muhamad Hariz, Muhamad Adnan et. al.,” Data Mining for Medical Systems: A Review”, Proc. of the International Conference on Advances in Computer and Information Technology - ACIT 2012.

[9] D. Keranahanirex and Dr .K.P. Kaliamurthi, “Multi class approach for detecting thyroid attack”, International Journal of Pharma and Bio Sciences, ISSN 0975-6299, 2013 July, pg no.(1246 – 1251).

[10] M. R. NazariKousarrizi, F.Seiti, and M. Teshnehlab, “Thyroid Disease Diagnosis Based on Feature Subset Selection and classification”, International Journal of Electrical & Computer Sciences IJECS-IJENS,ISSN 126001-8989, February 2012, Vol: 12.pg no.(13-19)

[11] Tian, D. Gledson ,et.al. , “A Bayesian Association Rule Mining Algorithm, Systems, Man, and Cybernetics (SMC)”, 2013 IEEE International Conference, 13-16 Oct. 2013, Pg no. (3258 – 3264)

[12] K.S. Kumar and R. M. Chezian,” Support vector machine and K-Nearest Neighbour based analysis for the prediction of hypothyroid”, International Journal of Pharma and Bio Sciences, ISSN 0975-6299, 2014 Oct, pg no(447 - 453).

[13] S. W. Mendre, R.D. Raut, ”Thyroid Disease Diagnosis using Image Processing: A Survey”, International Journal of Scientific and Research Publications, ISSN 2250-3153, December 2012 , Volume 2, Issue 12.

[14] Senol,” Thyroid and breast cancer disease diagnosis using fuzzy-neural networks”, Electrical and Electronics Engineering, 2009. ELECO 2009. International Conference on,ieee,E-ISBN 978-9944-89-818-8, 5-8 Nov. 2009,pgno.(390 – 393)