Volume 5, Issue 6, June 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Analysis of Factors Which Contribute to Thyroid Disease Using Data Mining Techniques

Shweta Taneja, Charu Gupta, Tripti Grover, Diksha Tripathi
CSE Dept., Bhagwan Parshuram Institute of Technology,
Guru Gobind Singh Indraprastha University,
New Delhi, India

Abstract— Thyroid disease is one of the most crucial diseases that affect the people of all ages now a days. It includes conditions associated with excessive release of thyroid hormones (Hyperthyroidism) and those associated with thyroid hormone deficiency (Hypothyroidism).This research investigates parameters like gender and the age group that are most likely to be affected by thyroid disease. It also investigates the sick and healthy factors which contribute to the thyroid disease in males and females. Association rule mining and Clustering, two intelligent data mining approaches, are implemented to identify these factors. The UCI Cleveland dataset for thyroid is taken in our study. In our results, females are seen to have more chance of thyroid disease than males. Similarly old people are detected to have more chance of thyroid disease as compared to the young ones. Also, the attributes indicating healthy and sick conditions were identified.

Keywords— TSH, Association Rule Mining, K-means Clustering, Chi Square Method,

I. INTRODUCTION

A thyroid disease is a medical condition that impairs the function of thyroid gland. Imbalance in production of thyroid hormones arises due to improper functioning of the thyroid gland itself, the pituitary gland, which produces thyroid-stimulating hormone (TSH), or the hypothalamus, which regulates the pituitary gland via thyrotropin-releasing hormone(TRH).As described in [7] the thyroid produces two major hormones called T3 (triiodothyronine) and T4 (thyroxine) that helps to control our body's functions. In case of under activity of thyroid hormone, our body functions get slow down, this condition is called hypothyroidism. If the increased amount of thyroid hormones seen in our blood, our body functions will speed up. This condition is called hyperthyroidism. About 200 million people are suffering from thyroid disease .Their early diagnosis is very important. So, this research uses the computational intelligence approach. The utilization of KDD and data mining is increasing in medical informatics. The usage of data mining tools with advanced algorithms is popular for pattern discovery in biological data [8].

Particularly, this research presents rule extraction and analysis experiments on thyroid disease data using two different data mining algorithms – Association rule mining and Simple k-means Clustering Algorithm.

Cleveland UCI dataset [15], a publicly available dataset and widely popular among data mining researchers, has been used and using the Chi-square method, the pre-processing of the data has been done. The main purpose of using Chi Square here is feature selection so as to reduce the number of features while maintaining acceptable classification accuracy.

The "expected frequency" for any cell is thus calculated as: Say, N observations are divided among n cells.

$$E_i = \frac{N}{n}$$

……Eq(1)

The value of test statistic is:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

……..Eq(2)

This research will make extensive use of Association rule algorithms. [2]. It aims to extract interesting correlations, frequent patterns or associations among sets of items in the transnational databases or other data management systems. As stated in [1], in its basic structure, every association rule fulfilling the minimum support and confidence are extracted. A rule obtained from the association rule mining algorithm is having the form 'LHS (left hand side) => RHS (right hand side)', where in this LHS as well as the RHS are the disjoint sets.

In this research, two techniques have been used: Apriori, Predictive Apriori.

In the algorithmic process of Apriori, an item set Y of length k is frequent if every subset of Y, having length k _ 1, are also frequent.

Predictive Apriori an algorithm motivated by Apriori, maximises the expected accuracy of an association rule. While Apriori ranks the rules based on confidence only, Predictive Apriori considers both the confidence and support in ranking the rules.

This research also makes use of the partition based clustering method called as k- Means clustering. K-Means algorithm is one of the popular partitioning algorithms. As described in [5] the idea is to classify the data into k clusters where k is the input parameter specified in advance through iterative relocation technique which converges to local minimum.

The paper is organized as follows: In section II, related work in thyroid disease has been discussed. In section III, we have described our proposed work and some related concepts and algorithms. The obtained experimental results and analysis are given in section IV. Finally, Section V presents the conclusions.

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 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.

III. PROPOSED WORK

The research has been organised into four parts:

A. Data Pre-processing

In the first stage feature selection technique, Pearson's Chi Square Test is used to select the appropriate features used in the thyroid prediction. The main motive of feature selection is to lessen the number of features that have been used in classification and maintaining acceptable classification accuracy as well. Less important features are removed, giving a subset of the original features which retains sufficient important information.

B. Rule extraction

In the Association Rule Mining, all sick individuals were regarded to be in one class and healthy individuals to be in another class. Two popular association rule mining algorithms, Apriori, Predictive Apriori, were used in this experiment. As there can be many such rules, only the rules that contain the 'sick' or 'negative' class were considered.

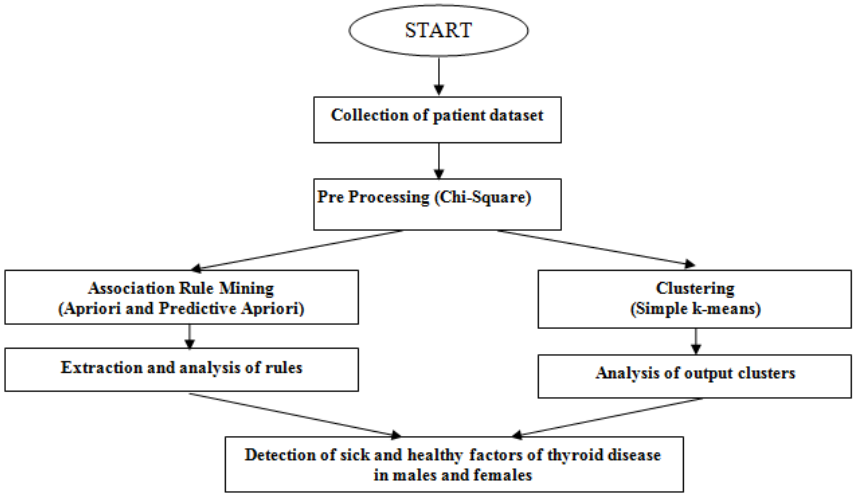

Fig. 1 Flowchart of the proposed work.

C. Clustering

In the Clustering method, Simple k-means clustering is done with two clusters (k=2) between hormones T3 and TT4, T3 and TT4 on X and Y axis respectively to obtain well defined clusters.

D. Analysis of Rules and Clusters

The gender and age of a person has been found to be an important factor influencing thyroid disease. Rules are separated based on sex and then age and clusters (k=2) are studied to determine the gender and the age that is most likely to be affected and unaffected by the thyroid disease. The rules based on gender are discovered to analyse the factors that contribute to sick and healthy conditions in that particular gender.

IV. EXPERIMENTAL ANALYSIS AND RESULTS

A. Dataset

As mentioned earlier, we use the publicly available UCI thyroid disease dataset [15] in our research. 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, 10 attributes were selected out of the 29 attributes in the original dataset in which first attribute is class. The dataset is then filtered by applying the unsupervised discretised filter on the attributes to convert the continuous values into nominal. These 10 attributes are as follows:

TABLE I SELECTED ATTRIBUTES

Class	Sick, Negative
age	young, middle, young-old, old
Sex	M,F
On thyroxine	f, t
Sick	f, t
Goitre	f, t
TSH	Nominal
T3	Nominal
TT4	Nominal
T4U	Nominal

B. 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, Apriori, Predictive Apriori were 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. Results of the experiment are shown in Tables 2-3.

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.

TABLE II RULE EXTRACTION FOR SICK AND NEGATIVE CLASS USING APRIORI ALGORITHM.

Algorithm	**Rules**
Apriori	**Sick Rules:**
	age=old sex=F on thyroxine=f TT4='(78.4-97.2]' 6 ==> class=sick. 6 acc:(0.96951)
	age=old sex=F sick=f TT4='(78.4-97.2]' 7 ==> class=sick. 7 acc:(0.97479)
	age=old sex=F goitre=f TT4='(78.4-97.2]' 7 ==> class=sick. 7 acc:(0.97479)
	sex=F TSH='(-inf-14.5045]' T4U='(0.724-0.821]' 9 ==> class=sick. 9 acc:(0.98135)
	age=old goitre=f TSH='(-inf-14.5045]' T4U='(0.724-0.821]' 7 ==> class=sick. 7 acc:(0.97479)
	Healthy Rules:
	age=young sex=M TT4='(78.4-97.2]' 8 ==> class=negative. 8 acc:(0.97856)
	age=young on thyroxine=f TT4='(78.4-97.2]' 8 ==> class=negative. 8 acc:(0.97856)

age=young sick=f TT4='(78.4-97.2]' 8 ==> class=negative. 8 acc:(0.97856)
age=middle sex=M on thyroxine=f TT4='(97.2-116]' 4 ==> class=negative. 4 acc:(0.94972)
sex=M on thyroxine=f goitre=f TT4='(78.4-97.2]' T4U='(0.724-0.821]' 5 ==> class=negative. 5 acc:(0.96177)
sex=M on thyroxine=f TSH='(-inf-14.5045]' TT4='(78.4-97.2]' T4U='(0.724-0.821]' 5 ==> class=negative. 5 acc:(0.96177)
sex=F TSH='(-inf-14.5045]' TT4='(97.2-116]' T4U='(0.918-1.015]' 3 ==> class=negative. 3 acc:(0.92946)

TABLE III RULE EXTRACTION FOR SICK AND NEGATIVE CLASS USING PREDICTIVE APRIORI ALGORITHM.

Algorithm	**Rules**
Predictive Aproiri	**Sick Rules:** sex=F on thyroxine=f T4U='(0.724-0.821]' 9 ==> class=sick. 9 acc:(0.98135) sex=F TSH='(-inf-14.5045]' T4U='(0.724-0.821]' 9 ==> class=sick. 9 acc:(0.98135) age=old sex=F sick=f TT4='(78.4-97.2]' 7 ==> class=sick. 7 acc:(0.97479) sex=M TT4='(40.8-59.6]' 4 ==> on thyroxine=f class=sick. 4 acc:(0.94972) age=old on thyroxine=f goitre=f T4U='(0.724-0.821]' 7 ==> class=sick. 7 acc:(0.97479) age=old goitre=f TSH='(-inf-14.5045]' T4U='(0.724-0.821]' 7 ==> class=sick. 7 acc:(0.97479) **Healthy Rules:** age=young on thyroxine=f TT4='(78.4-97.2]' 8 ==> class=negative. 8 acc:(0.97856) age=young sex=M TT4='(78.4-97.2]' 8 ==> class=negative. 8 acc:(0.97856) sex=M TSH='(-inf-14.5045]' T3='(1.12-1.58]' TT4='(78.4-97.2]' 4 ==> class=negative. 4 acc:(0.94972) sex=M on thyroxine=f goitre=f TT4='(78.4-97.2]' T4U='(0.724-0.821]' 5 ==> class=negative. 5 acc:(0.96177) sex=M on thyroxine=f TSH='(-inf-14.5045]' TT4='(78.4-97.2]' T4U='(0.724-0.821]' 5 ==> class=negative. 5 acc:(0.96177) age=young sex=F sick=f TT4='(97.2-116]' 3 ==> TSH='(-inf-14.5045]' class=negative. 3 acc:(0.92946)

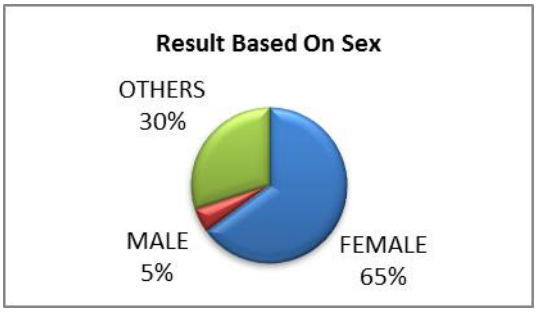

Fig. 2 Result Based on Sex from Association Rules

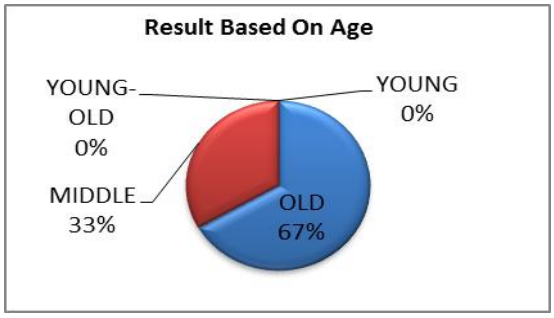

Fig. 3 Result Based on Age from Association Rules

C. Results of Clustering

In Fig. 4 two clusters (k=2) were formed between T3 and TT4 on X and Y axis respectively. Both the clusters with blue and red seeds were analysed along with the outliers. The cluster with blue seeds indicated the healthy cases and was majorly attributed to males and young age. The red cluster belonged to the sick cases and the seeds consisted of majority of females and old. The cases represented by the outliers are attributed to middle aged healthy female and healthy and young female.

Fig 4 Clusters between T3 and TT4

Similarly, Fig.5shows two clusters (k=2) between the hormones T3 and T4U on X and Y-axis respectively. The cluster with blue seeds mainly belonged to the males and young people whereas the red cluster is attributed to females and the old cases. The outliers in this case also indicated the case of middle aged healthy female.

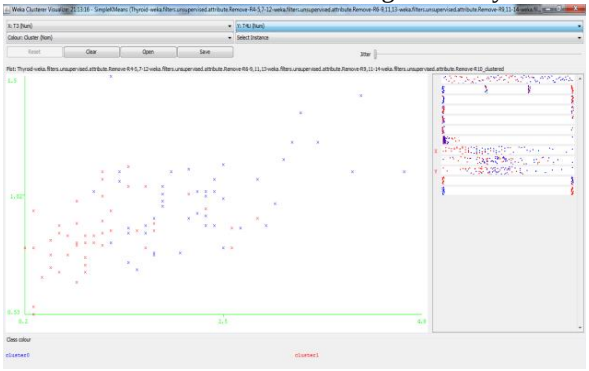

Fig. 5 Clusters between T3 and T4U

D. Analysis

As according to the findings in the previous section, that females are at lower risk of having thyroid disease, are investigated here in more detail. The dataset was split, based on 'males' and 'females' and rules were extracted again for sick and healthy data. The Apriori algorithm was used to extract the rules for males as well as females. The aim was to notice which factors are related to thyroid disease in men and women separately. The results are presented in Tables 4 and 5, listing the rules with above 90% confidence.

The results show that for females the values of hormones T3 in the range (-inf-0.64], TSH in the range (-inf-14.5045] and T4U in the range (0.724-0.821] are the indicators of a female having thyroid disease Whereas a female being young and absence of goitre indicates a lower risk of thyroid disease and a female being healthy.

The significant factors indicating males being healthy include the value of hormones TSH in the range of (1.3135-2.612] , T3 in the range (2.16-2.6] and T4U in the range (0.888-0.945]'.The young age is also a good indicator for a healthy man having less risk of thyroid disease. The value of hormone T3 in the range of (-inf-0.84] indicates the sick condition for males.

TABLE IV RULE EXTRACTION FOR MALES USING APRIORI ALGORITHM FOR SICK AND HEALTHY CONDITIONS.

Algorithm	**Sex**	**Rules**
	Males	**SICK RULES:**
		onthyroxine=f sick=f T3='(-inf-0.84]' 6 ==> class=sick. 6 conf:(1)
		sick=f T3='(-inf-0.84]' 6 ==> class=sick. 6 conf:(1)
		onthyroxine=f T3='(-inf-0.84]' 6 ==> class=sick. 6 conf:(1)
		T3='(-inf-0.84]' 6 ==> class=sick. 6 conf:(1)
		HEALTHY RULES:
		age=young sick=f TSH='(1.3135-2.612]' 6 ==> class=negative. 6 conf:(1)
		age=young on thyroxine=f TSH='(1.3135-2.612]' 6 ==> class=negative. 6 conf:(1)
		age=young TSH='(1.3135-2.612]' 6 ==> class=negative. 6 conf:(1)
		onthyroxine=f sick=f goitre=f T3='(2.16-2.6]' 7 ==> class=negative. 7 conf:(1)
		sick=f goitre=f T3='(2.16-2.6]' 7 ==> class=negative. 7 conf:(1)
		goitre=f T4U='(0.888-0.945]' 11 ==> class=negative. 11 conf:(1)
		sick=f T4U='(0.888-0.945]' 11 ==> class=negative. 11 conf:(1)

TABLE V RULE EXTRACTION FOR FEMALES USING APRIORI ALGORITHM FOR SICK AND HEALTHY CONDITIONS.

Algorithm	**Sex**	**Rules**
	Females	**SICK RULES:**
		goitre=f T3='(-inf-0.64]' 11 ==> class=sick. 11 conf:(1)
		T3='(-inf-0.64]' 11 ==> class=sick. 11 conf:(1)
		onthyroxine=f sick=f goitre=f T3='(-inf-0.64]' 9 ==> class=sick. 9 conf:(1)
		goitre=f TSH='(-inf-14.5045]' T3='(-inf-0.64]' 10 ==> class=sick. 10 conf:(1)
		sick=f TSH='(-inf-14.5045]' T3='(-inf-0.64]' 10 ==> class=sick. 10 conf:(1)
		goitre=f T4U='(0.724-0.821]' 9 ==> class=sick. 9 conf:(1)
		onthyroxine=f T4U='(0.724-0.821]' 9 ==> class=sick. 9 conf:(1)

HEALTHY RULES:

age=young sick=f goitre=f 9 ==> class=negative. 9 conf:(1)
age=young goitre=f 9 ==> class=negative. 9 conf:(1)
age=young sick=f 9 ==> class=negative. 9 conf:(1)

V. CONCLUSION

This research has presented a rule extraction method on thyroid disease data using different rule mining algorithms (Apriori, Predictive Apriori) and k-means clustering technique. Further rule-mining-based analysis and clustering analysis was undertaken by categorising data based on gender, age and significant risk factors for thyroid disease were found for both men and women by splitting the association rule respectively. Interestingly, it is found from the set of sick rules, being 'female' is one of the factors for the thyroid disease condition. In other words, the results indicated females to have more chances of the thyroid disease than males. And old age people are most affected to the thyroid disease and young age the least. The results are predicted results which can be further verified in laboratories.

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)

[15] The weka website. Available: http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

