IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Information Retrieval From Thyroid Database Through Data Mining

[1*]N. Vijayalakshmi, [2]P. Nithya

[1,2]Department of Computer Science, Shrimati Indira Gandhi College, Bharathidasan University, Tiruchirappalli, Tamilnadu, India

[*]*Corresponding Author: nvijimca@gmail.com Tel: +91 9965779358*

Available online at: www.ijcseonline.org

Accepted: 19/July/2018, Published: 31/July/2018

Abstract: Thyroid disorders occur due to dysfunction of the thyroid gland or pituitary gland, iodine deficiency, cancer in some parts of the body, or due to side-effects from other medications. Hyperthyroidism, Hypothyroidism, Goitre, and Thyroid cancer are some of the ailments that result due to thyroid disorders. Some other reasons like pregnancy, or medications for other illnesses may also show abnormal levels of thyroid hormones This research study aims to identify conditions based on which we could predict the type of thyroid disorder in patients. This could help in further diagnosis and treatment. We study various attributes commonly found in patients with thyroid disorders to identify those attributes that may specifically describe the type of thyroid disorder in a person. Moreover we analyze six different classes of thyroid disorders, their symptoms and try to classify what kind of disorder a person has based on the symptoms. Totally 1535 records with 29 attributes are taken for the study. Statistical techniques are used to analyze the frequency of occurrence of various factors towards each type of the disease and test for significance of factors is also done. The results are used to build a data model that helps to predict the occurrence of specific type of thyroid disorder in a patient based on significant symptoms. The results emphasize that age, sex, values of hormones like TSH, T3, T4 and FTI of a patient play a predominant role in classifying and determining the type of thyroid disorder in the person. We also classify the given dataset using various decision tree techniques in different ways and compare the results.

Keywords: Data mining, Thyroid disorder, Classification, Prediction

I. INTRODUCTION

Today we are in the knowledge and information era. Knowledge is a crucial organizational resource. It is used for better decision making that gives competitive advantage. Knowledge acquisition and management has become the foremost activities of all organizations. Valuable information hides in historical data. We should use the right tools to analyze the past data to find new facts that could help in taking better decisions to improve the present status or to use the new knowledge to take new steps. Data mining is a process to extract useful and interesting facts or knowledge from existing large datasets.

The thyroid is a small gland located below the Adam's apple in your neck. It releases hormones, *thyroxine* (T4) and *triiodothyronine* (T3), which increase the amount of oxygen your body uses and stimulate your cells to produce new proteins. By controlling the release of these hormones, the thyroid determines the metabolic rate of most of your body's organs. The thyroid gland is regulated by *thyroid-stimulating hormone* (TSH), which is made by the pituitary gland in the brain. Normally, when thyroid hormone levels in the body are high, they will "switch off" the production of TSH, which in turn stops the thyroid from making more T4 and T3. Problems occur when the thyroid gland becomes either underactive (*hypothyroidism*) or overactive (*hyperthyroidism*). Thyroid problems are more common in women than men. Cancer may also develop in the thyroid gland.

Thyroid diseases sometimes result from inappropriate TSH levels, or may be caused by problems in the thyroid gland itself. Hypothyroidism results in low levels of T4 and T3 in the blood. Hyperthyroidism results in high levels of T4 and T3 circulating in the blood. The usual treatment for hypothyroidism is thyroid hormone replacement therapy. Hyperthyroidism can be treated with iodine (including radioactive iodine), anti-thyroid medications or surgery.

This research study uses data mining processes and tools to mine new knowledge from record sets of patients suffering from thyroid disorders. Several factors leading to thyroid disorders have been taken for the study and significant factors out of all factors were identified through statistical analysis and through subset evaluators using WEKA. Decision rules for prediction of thyroid disorders were also generated.

This paper is organized as follows: Section I gives an introduction to the study. Section II briefs on what other people have done in related areas of study. Section III outlines the methods and techniques used for this study. Section IV produces the results obtained through this study. Section V discusses about the outcome of the study. Section VI gives concluding remarks and limitations of our study.

II. RELATED WORK

Dr.G.Rasitha Banu and M.Baviya , in their study use the DBSCAN algorithm used for predicting the thyroid disease with related symptoms. Hierarchical multiple classifier scheme is used for classification[1][7]. Irina and Liviu in their work use a data set from the UCI machine learning repository with data from Romania. They use several factors that affect thyroid function like stress, infection, trauma, toxins, low-calorie diet, medication etc. They have analyzed and compared four classification models: Naïve Bayes, Decision tree, multilayer perceptron and radial basis function network and obtained a significant accuracy for all. Tools used were KNIME Analytics Platform and Weka. They were able to classify the records into three classes normal, hyper thyroidism and hypothyroidism[2].

Prerana et al. in their paper present a systematic approach for early diagnosis of Thyroid disease using back propagation algorithm used in neural network. 29 attributes were taken. They found that Levenberg Marquardt method was better in performance to simple gradient descent algorithm. MATLAB NN Toolbox was used for mining[3]. In another work, the authors use decision tree algorithms to classify types of thyroid disease and compared their performance based on six performance metrics Accuracy, Mean absolute error, prediction, recall, and Kappa statistic to understand how to apply decision tree algorithms[4].

Dr. Srinivasan et al. and K.Rajam et al. have used Decision tree, Naïve Bayes, Backpropagation neural network, support vector machines to predict diagnosis of thyroid disease[5] [6]. Zoya Khalid et al., in their study, scan the thyroid genome to hit molecular targets which are highly associated with thyroid cancer. The results reveal GPM6A as a novel associated gene marker [8]. B.Jothi et al., use the K-means clustering algorithm in Hadoop framework by finding the data centroid among nearest data node using MapReduce framework for training and analyzing thyroid related data sets to say whether a patient suffers from hypo or hyper thyroidism based on symptoms[9].

K. Saravana kumar et al, apply the model of deception of thyroid dataset and use apriori algorithm to generate the rules. The rules are used to test the thyroid dataset as deceptive or not. They try to track important information that mark the occurrence of thyroid[10]. Limin Wang et al. in their study use k-dependence causal forest model to generate a series of sub models in the framework of maximum spanning tree and demonstrate stronger dependence representation. Friedman test on 12 UCI datasets shows that KCF has classification accuracy advantage over the other BNCs, such as Naïve Bayes, tree augmented NB, and k-dependence Bayesian classifier[11]. Kevin M.Pantalone et al. reports that the assessment of the serum FTI may be helpful in making the diagnosis of central hypothyroidism in the appropriate clinical setting and when free T4 is in the low-normal range, particularly in patients with multiple anterior pituitary hormone deficiencies and /or with symptoms suggestive of hypothyroidism[12].

III. METHODOLOGY

We have taken a thyroid disease database from the UCI Repository left by Ross Quinlan in 1987 at the University of California, Irvine. The database we have taken consists of 1535 records selected at random from a bigger database consisting of 9172 records. Totally 29 attributes are available in the data repository. Through statistical analysis and subset evaluation, we shall select only those attributes that are significant enough to determine the presence of thyroid disorders in patients.

The attributes given in the database are: Age, Sex, On Thyroxine, Query(Thyroxine), Anti-thyroid medication, sickness, pregnancy, thyroid surgery, I131 treatment, Query(Hypothyroid), Query(Hyperthyroid), Lithium, Goitre, Tumor, Hypo-pituitary, Psychological disorder, TSH, TSH value, T3, T3 value, FT4,FT4 value, T4U, T4u value, FTI, FTI value, TBG, TBG value, Referral source, Predicted Disease.

Thyroid diseases are classified into six different classes as follows:

Table 1 Classification of thyroid disorders

S. NO	CLASS	CLASS NAME	REPRESENTATION	DISEASE
1	Class 1	Hyperthyroid	A,B,C, D	Hyperthyroid T3 Toxic Toxic Goitre Secondary toxic
2	Class 2	Hypothyroid	E,F,G, H	Hypothyroid Primary hypothyroid Compensated Hypothyroid Secondary Hypothyroid
3	Class 3	Binding Protein	I,J	Increased binding protein Decreased binding protein

4	Class 4	General health	K	Concurrent non-thyroidal illness
5	Class 5	Replacement Therapy	L,M,N	Consistent with replacement therapy Under replaced Over replaced
6	Class 6	Anti-thyroid treatment	O,P,Q	Anti-thyroid drugs I131 treatment Surgery
S. NO	CLASS	CLASS NAME	REPRESENTATION	DISEASE
7	Class 7	Miscellaneous	R,S,T	Discordant assay results Elevated TBG(S) Elevated thyroid hormones

Out of the 1535 records taken, 422 were found to be having thyroid disorders. The rest 1113 records did not have any problems. The following table gives information on the number and percentage of persons diagnosed with each type of thyroid disorder found in the 422 patients:

TYPES	A	B	C	D	F	G
COUNT	43	4	2	3	57	39
%	10%	1%	0.5%	0.7%	13.5%	9.2%
TYPES	H	I	J	K	L	M
COUNT	1	44	3	94	20	22
%	0.2%	10.4%	0.7%	22.3%	4.7%	5.2%
TYPES	N	O	P	Q	R	S
COUNT	35	4	2	1	31	17
%	8.3%	0.9%	0.5%	0.2%	7.3%	4.0%

Statistical Analysis of the database was done using Pivot Tables in Excel. Using this, significant factors were determined. This was confirmed using Subset Evaluation in Weka. The given database was also classified using decision tree classifiers using Weka. Based on the frequency of occurrence of the significant factors, decision rules were derived for prediction of the type of thyroid disorder in a patient.

IV. RESULTS

Statistical analysis of various factors / attributes given in the thyroid dataset produced the following information:

1. Most of the people with thyroid disorder are prevalent in the age group 26 - 75 (77%).
2. Thyroid disorder is more prevalent among women than men (68%)
3. Out of 422 diagnosed with some thyroid disorder, only 9% have hypothyroid. Only 3% have queried on having thyroid. But 22% are taking thyroxine supplement. However none have hypo pituitary problems.
4. Sickness, Pregnancy, Psychotherapy and Surgery: Occurrence of these four attributes is found to be very rare in the given dataset.
5. Out of the 422 cases diagnosed with thyroid disorders, only 8% have hyperthyroid. None have Goitre or Lithium. 2% have tumor and 3% have taken I131 treatment. Even among those who have not been diagnosed with any thyroid disorders, 3% have tumors and have taken I131 treatment. But this may not be connected with thyroid disorder and may be linked with other types of tumors metastased to the thyroid gland.
6. Almost all of the patients (90%) have taken all thyroid related tests like test for TSH, FT4, T3, FTI etc. Only few have tested for TBG.
7. Out of those diagnosed with thyroid disorder 50% have TSH values within normal range while 50% do not. Out of the whole, only 20% have values within normal range, out of which 14% have thyroid disorder. **It is interesting to find that 92% of those without thyroid disorder and 67% of the entire population have elevated TSH levels.**
8. T3, FT4, FTI values are 90% - 100% outside the range, irrespective of whether the person is diagnosed with thyroid or not. This shows that many people have elevated T3 and T4 levels, i.e., signs of hyper thyroid. TBG values appear to be good for people without any thyroid disorder. However enough data is not available about those with thyroid disorder.

Statistical analysis of the significant factors identified, produced the following decision rules:

- If T3 is low and T4 and FTI is high then hyperthyroid.
- If High TSH, T4, FTI and low T3 then hypothyroid.
- If normal TSH. Very low T3 and T4, and a high FTI then secondary hypothyroid.
- If normal TSH and T4, low T3 and very high FTI then Increased binding protein
- If normal TSH, very low T3, lower T4 and higher FTI then Decreased binding protein
- If normal TSH, T4 but high FTI then Concurrent non-thyroid illness
- If on thyroxine and normal TSH and T4 and very high FTI could be consistent with replacement therapy or overreplaced.

- If on thyroxine, high TSH, low T3, normal T4 and little high FTI, under replaced.
- Low T4, high TSH and FTI, with surgery or I131 treatment, could be class 5
- If everything else false with high FT4 and FTI then discordant assay
- If everything else false with high TBG, then elevated TBG.

Classification of the given knowledge base is done in three ways. The results produced (in each case) are given below:

Table 3 All 1535 records with 29 attributes

Classifier	Correctly Classified Instances	Relative Absolute Error	Roc Area Weighted Avg	TP Rate Weighted Avg	FP Rate Weighted Avg
RANDOM FOREST	92.7%	32.2%	0.988	0.927	0.111
RANDOM TREE	87.8%	27.9%	0.900	0.878	0.119
J48 PRUNED TREE	**93.16%**	**21.09%**	**0.953**	**0.932**	**0.058**
NAÏVE BAYES	83.32%	41.16%	0.948	0.833	0.113

Table 4 All 1535 records with only 9 important attributes

Classifier	Correctly Classified Instances	Relative Absolute Error	Roc Area Weighted Avg	TP Rate Weighted Avg	FP Rate Weighted Avg
RANDOM FOREST	81.7%	46.8%	0.975	0.818	0.418
RANDOM TREE	80.3%	46.3%	0.791	0.803	0.344
J48 PRUNED TREE	87.3%	42.4%	0.934	0.873	0.140
NAÏVE BAYES	83.8%	44.2%	0.948	0.838	0.137

Table 5 422 records of those diagnosed with thyroid disorder with all 29 attributes

Classifier	Correctly Classified Instances	Relative Absolute Error	Roc Area Weighted Avg	TP Rate Weighted Avg	FP Rate Weighted Avg
RANDOM FOREST	88.63%	31.43%	0.990	0.886	0.014
RANDOM TREE	77.26%	26.69%	0.887	0.773	0.026
J48 PRUNED TREE	88.39%	18.33%	0.955	0.884	0.012
NAÏVE BAYES	76.78%	27.28%	0.963	0.768	0.025

V. DISCUSSION

Statistical analysis of the given dataset using pivot tables indicated that age, sex, values of the hormones TSH, T3, T4, and the metric FTI are good attributes that could help in classification and description of the dataset. Moreover, people at an age from 26 to 75 were more prone to this disorder and females were more susceptible than males. So analysis was made using Weka tool to get an accurate idea about the dataset. Subset evaluation using CFS subset evaluator has provided us with 11 significant attributes in one case and 9 significant attributes in another.

It was also found that elevated TSH levels are very prevalent irrespective of thyroid disorders. Elevated T3 and T4 levels are indicative of thyroid disorders due to several reasons even though we see that tumors, goiter, and hypo-pituitary attributes show very low occurrences in the sample population. Therefore there is very good reason to believe that TSH, T3, FT4, FTI and TBG are good candidates for further analysis, besides, age and sex of the respondent.

Classification was carried out using various decision tree algorithms like Naïve Bayes, Random trees, Random forest and J48 techniques. In each case, Accuracy of classification, average error rate, ROC area, average True Positive rate and average False Positive rate for each classifier is compared against others. It is found that J48 pruned tree is the best classifier as it yields more accurate results in two cases.

The results show that taking only 9 important attributes is not a good idea. Analysis of all 29 attributes with all the 1535 records has provided us with the best possible result. Highest Accuracy of classification 93.16% has been achieved in this case and weighted average of True Positive rate is also found to be high at 0.932.

Important decision rules obtained through analysis and classification is used to build a model for prediction of the type of thyroid disorder among new patients using their values for the significant factors mined.

Table 6 Rules generated from statistical analysis:

ON THYROXINE	TSH	T3	T4	FTI	TBG	CLASS / TYPE
		LOW	HIGH	HIGH		A,B, C,D
	HIGH	LOW	HIGH	HIGH		F,G
	NORMAL	LOW	LOW	HIGH		H
	NORMAL	LOW	NORMAL	V. HIGH		I
	NORMAL	V LOW	LOW	HIGH		J

	NOR MAL		NOR MAL	HIGH		K
YES	NOR MAL		NOR MAL	V. HIGH		L,N
YES	HIGH	LOW	NOR MAL	HIGH		M
	HIGH		LOW	HIGH		O,P,Q
			HIGH	HIGH		R
					HIGH	S

VI. CONCLUSION

This research study has successfully mined enough knowledge from 1535 data records of people tested for thyroid disorders. A total of 29 attributes were taken for the study. These were related to the occurrence of thyroid disorders. Out of 1535 records, 422 records belonged to people with some type of thyroid disorder. Totally 6 classes of thyroid disorders exist. Totally 20 types of thyroid disorders exist under various classes.

Using Pivot tables, every one of the 29 attributes for all the records were analyzed and the statistics indicated that only a few attributes like age, sex, TSH, T3, T4 and FTI values were significant indicators for the occurrence of thyroid disorders. But it was very difficult to generate rules from these statistics as many were not predominant enough to arrive at any decision. However, classification using different classifiers with different sets of records and different sets of attributes could throw some light on the significance of certain factors like TSH, T3, T4 and FTI values.

CFS Subset evaluator was used to confirm the attributes that were guessed to be significant through statistical analysis. The trees produced by random tree, random forest and J48 pruned tree classifiers were very large. A few rules could be derived from these trees, but a comprehensive set of rules for each type of disorder was derived from statistical analysis. This could be used for accurate prediction of the type of thyroid disorder for new patients. Thus data mining was used to accurately classify and predict thyroid disorders.

Results obtained can be improved by taking all the 9172 records from the parent dataset. This can be taken as a future work.

REFERENCES

[1]. Dr.G.Rasitha Banu, M.Baviya, "Predicting Thyroid Disease using Data Mining Technique", International Journal of Modern Trends in Engineering and Research, e-ISSN 2349-9745, Pages 666-670, Vol 2(3), March 2015.

[2]. Irina Ionitak Liviu Ionita, "Prediction of Thyroid Disease Using Data Mining Techniques", BRAIN. Broad Research in Artificial Intelligence and Neuroscience. Vol.7. pp.115-124. Nov 18, 2017

[3]. Prerana, Parveen Sehgal, Khushboo Taneha, "Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network", International Journal of Research in Management, Science & Technology, e-ISSN 2321-3264, Vol 3(2), April 2015

[4]. Ebru Turanoglu Bekar, Gozde Ulutagay, Suzan Kantarci Savas, "Classification of Thyroid Disease by Using Data Mining Models: A comparison of Decision Tree algorithms." The Oxford Journal of Intelligent Decision and Data Science, Vol 2016(2), Pages 13-28, doi: 10.5899/2016/ojids-00002,

[5]. K.Rajam, R. Jemina Priyadarsini, "A Survey on Diagnosis of Thyroid Disease using Data Mining Techniques", International Journal of Computer Science and Mobile Computing, ISSN 2320-088x , Vol 5(5), pg 354-358, May 2016

[6]. Dr.B.Srinivasan, K.Pavya, "Diagnosis of Thyroid Disease Using Data Mining Techniques: A study", International Research Journal of Engineering and Technology, e-ISSN 2395-0056, Vol 3(11), Nov 2016.

[7]. Dr. G.Rasitha Banu, M.Baviya, Dr. Murtaza Ali, "A Study on Thyroid Disease using Data Mining Algorithm", International Journal of Technical Research and Applications, e-ISSN 2320-8163, Vol 3(4), PP 376-379, Aug 2015

[8]. Zoya Khalid, Sheema Sameen, Shaukat I Malike, Shehzad S, "Computational Analysis on the Role of GPM6A in human thyroid cancer", Journal of Data Mining in Genome Proteinomics 3:114, doi: 10.4172/2153-0602.1000114, ISSN 2153-0602,Jan 2012.

[9]. B.Jothi, S.KRishnaveni, J. Jeyasudha, "Analysis of thyroid syndrome using K-Means Clustering Algorithm", Journal of Chemical and Pharmaceutical Sciences, ISSN 0974-2115,

[10]. K. Saravana Kumar, Dr. R. Manicka Chezian, "Analysis on Suspicious Thyroid Recognition using Association Rule Mining", Journal of Global research in Computer Science, ISSN 2229-371X, Vol. 3(9), Sep 2012.

[11]. Limin Wang, FangYuan Cao, ShuangCheng Wang, MingHui Sun, LiYan Dong, "Using k-dependence causal forest to mine the most significant dependency relationships among clinical variables for thyroid disease diagnosis", PLOS One, Vol 12(8), e0182070, 2017, doi: 10.1371/journal.pone.0182070

[12]. Kevin M.Pantalone et al., "Measurement of Serum Free Thyroxine Index may provide additional case detection compared to free thyroxine in the diagnosis of central hypothyroidism", Case Reports in Endocrinology, Vol 2015, Dec 8, 2015, doi: 10.1155/2015/965191

Authors Profile

Mrs.N. Vijayalakshmi completed her U.G and P.G in Computer Science at Seethalakshmi Ramaswami College, Trichy, Tamilnadu India in 1993 and did her M.Phil in Computer Science in Manonmaniam Sundaranar University, Tirunelveli. She is serving as Associate Professor in Department of Computer Science, Shrimati Indira Gandhi College, Trichy, for the past 24 years. She has published 12 research papers in reputed international journals indexed in Scopus, ICI, and Google Scholar etc. Her main research area is data mining and knowledge discovery. She is also interested in Network Security. She has 15 years of research experience.

Ms. P. Nithya is currently doing her M.Phil in Computer Science under the guidance of Ms. N. Vijayalakshmi. Her area of specialization is Data Mining.

