A COMPARISON ANALYSIS OF THYROID DISEASE DETECTION USING DATA MINING TECHNIQUES AND CLASSIFIERS

[1]Parimala.S, [2]Dr.Senthil Vadivu

[1]Assistant Professor, [2]Head
[1]Department of Computer Science[2]Department of Computer Applications,
[1]SRM Institute of Science and Technology, Chennai,
[2]Hindustan College of Arts and Science, Coimbatore.

Abstract: Thyroid hormone produces the thyroid glands which influences the metabolic activities in our body. Abnormal production of hormone produces thyroid disorders. When the production of thyroid hormones are more then we call that as hyperthyroidism On the other hand when it is less then we call that as hypothyroidism. Data mining classification techniques plays a vital role in predicting thyroid diseases. The initial step in data mining process is that the datasets are sorted in order. The next step is you identify the patterns and relationships. In this paper a comparison analysis is done based on the thyroid dataset and classification techniques available in data mining. The classification techniques used are k nearest neighbors, Support vector machine, j48, Bagging and Naïve Bayes

Index Terms -.

Hypothyroidism, K nearest neighbors, Support vector machine, j48, Bagging, Naïve Bayes.

I. INTRODUCTION

In health care services, classification of data mining plays a major role in data mining. This is one of the significant and the most challenging task to diagnose the various health conditions and to give the proper treatment in the early stage of the disease.

People from all over the world have been suffering from various health issues like diabetes, heart disease, typhoid, tuberculosis, kidney disease etc. [1] [2]. For example, the thyroid disease can cue diagnosed and detected through some clinical investigations and blood tests. Data mining techniques detect the thyroid disease in an early stage and also at lower cost.

From the thyroid gland the hormones are triggered by the indecorous excretion of thyroid hormones released which is one of the important organ situated in the front of the neck and below the Adam's apple. The two types of thyroid glands are levothyroxine or T4 and triiodothyronine or T3.These hormones help in production of balanced amount of proteins, regulating the temperature of body, and maintaining overall production of energy [3]. Thyroid disease occurs when thyroid gland stop to functioning properly and are mainly divided into hypothyroidism and hyperthyroidism [4].

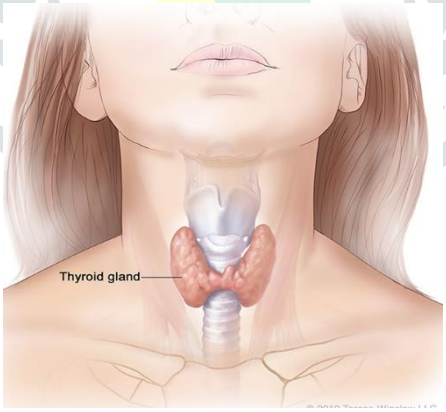


Fig1: Woman affected with thyroid disease

The surplus and poor discharge volume of thyroid hormone will cause hyperthyroidism and hypothyroidism respectively. The common symptoms of hyperthyroidism are sudden weight loss, rapid heartbeat, nervousness, etc. and hypothyroidism has weight gain, fatigue, weakness, shivering etc. One of the most common reason occurred due to hyperthyroidism is Graves' disease [5].

The second part of the paper enlightens the related work in diagnosis of thyroid storm and myxedema which may lead to death [6]. In this research work, a trained classification prototype is like K- nearest neighbor (KNN), support vector machine (SVM), decision tree (DT) and Naïve Bayes (NB) for the diagnosis of thyroid diseases. The third part encompasses the respective dataset and methods. The fourth part signifies the results and discussion based on the research.

II.RELATED WORK:

K. Rajam et.al observed data mining supervised functionalities Naïve Bayes, decision tree, back propagation, Support vector machine classifies thyroid disease at earlier stage. Results assessed based on parameters accuracy, performance, speed, and cost initiate effective for treatment of the patient [7].

Liyong Ma, Chengkuan Ma et.al suggested capable way using convolutional neural network to notice disease illnesses on SPECT datasets. Suggested approach result worked better than existing methods [8].

Eystraints G[9]have proposed a computer-aided diagnosis(CAD) system model named as TND(Thyroid Nodule Detector).It is used for the discovery of nodular tissue in ultrasound(US)thyroid images and videos acquired during thyroid US examinations.

Marissa Lourdes De Ataide et.al applied a two multilayer perceptron classifier for classifying thyroid diseases into three main classes as euthyroid, hyperthyroid and hypothyroid and to categorize hypothyroid disease into primary, secondary and tertiary hypothyroid with dedicated on extreme accuracy in minimum time. This classifier gave good accuracy of classification [10].To reduce problem [11], [12], [13], [14], applied neural network for analyzing thyroid problem.

Fatemeh Saiti et.al, applied Genetic Algorithms Using Support Vector Machine for thyroid verdict [15]. G. Rasitha Banu predicted problem using Linear Discriminant Analysis (LDA) - Data Mining approach [16].

.

III. MATERIALS AND METHODS

The dataset for this research work are collected from kaggle data explorer (Thyroid dataset). A data set is usually described as the group of data. Predominantly it is associated with the alike to a database table in which each and every column determines the distinct variable and every row denotes the member of the data set. The values for the entire variable will be distributed in the list.

In this dataset more than 150 patients are tested with hypothyroid with all age groups and with various attributes.

Age:	continuous.
Sex:	M, F.
On thyroxin:	f, t.
Query on thyroxin:	f, t.
On ant thyroid medication:	f, t.
Sick:	f, t.
Pregnant:	f, t.
Thyroid surgery:	f, t.
I131 treatment:	f, t.
Query hypothyroid:	f, t.
Query hyperthyroid:	f, t.
Lithium:	f, t.
Goiter:	f, t.
Tumor:	f, t.
hypo pituitary:	f, t.
Psych:	f, t.
TSH measured:	f, t.
TSH:	continuous.
T3 measured:	f, t.
T3:	continuous.
TT4 measured:	f, t.
TT4:	continuous.
T4U measured:	f, t.
T4U:	continuous.
FTI measured:	f, t.
FTI:	continuous.
TBG measured:	f, t.
TBG:	continuous.
Referral source:	WEST, STMW, SVHC, SVI, SVHD, other.

Table1: Attributes and Values in Thyroid dataset

3.1 K nearest Neighbor (KNN) Classifiers:

K-nearest neighbor (KNN) is a supervised method as well as non-parametric. The input feed depends on the K closest instances existing in the feature. The output which is generated depends on whether KNN is Classification or regression methods [17] [18].When there is requirement for prediction for the data that are undetected instances, the KNN algorithm will search through the training data instances for the k-most similar instances. The prediction attribute of the most similar instances is summarized and returned as the prediction for the undetected instance [19].

KNN Classifiers are based on learning by similarity, it is being compared with a given test tuple with training tuples that are alike to it. Each tuple signifies a point in an n-dimensional space. When an unknown tuple is given as input, a k nearest neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. Created on these k training tuples are the k "nearest neighbors" of the tuple which is unknown which is categorized by a mainstream vote of its neighbors, and allocated to the class most mutual amongst its k-nearest neighbors. When a training tuple k-Nearest Neighbor is assumed it stores and pauses until it is specified a test tuple.

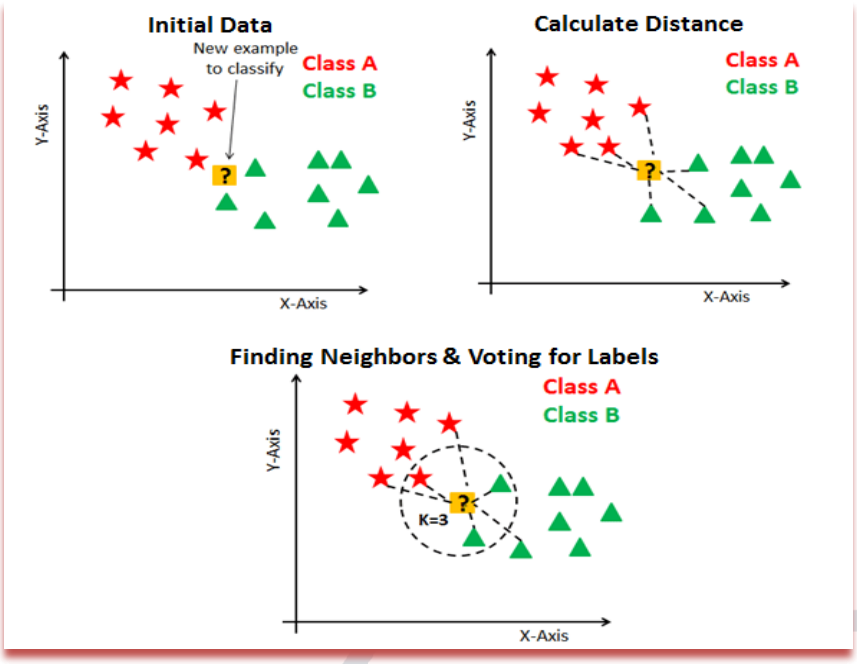
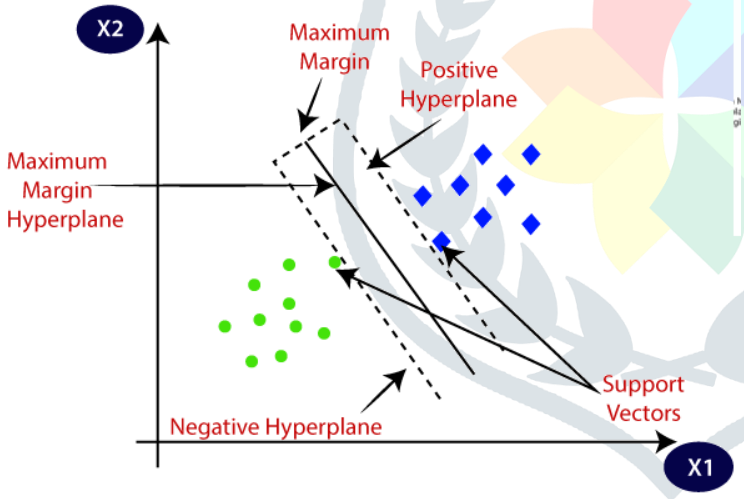


Fig2: K nearest neighbors

3.2 Support Vector Machine

Support vector machine (SVM) is a also a supervised learning classification technique that are utilized to inspect the data for regression and classification methods [20].It builds an best hyper plane in a high- or infinite- dimensional space in which new instances are allocated to one group or the other one[21].The division of data is attained by the hyper plane is generally done, that has largest distance to the neighboring training data point of any class since in general the greater the margin the lesser the generality error of the classifier [22].



J48 Algorithm

Correctly Classified Instances	3765	99.8144 %
Incorrectly Classified Instances	7	0.1856 %
Kappa statistic	0.9872	
Mean absolute error	0.002	
Root mean squared error	0.0288	
Relative absolute error	2.7772 %	
Root relative squared error	15.1421 %	
Total Number of Instances	3772	

3.3 Naïve Bayes Algorithm:

Naïve is one of the most accessible and skilful algorithm in data mining techniques. The Naïve Bayes is also called as eager learner since they have the ability of constructing a model straightaway after a training set is given. Naïve Bayes classifier is based on Bayes theorem based on conditional probability. Naive Bayes theorem is stated as

P (Cj/ d) = P (d/ Cj) P Cj /P (d)

P(Cj|d) represents the chance of instance „d" being in class Cj. P(d|Cj) represents the probability of generating instance „d" given a class Cj. P(d) means the probability of instance "d" happening. The main benefit of Naive Bayes is that it trains and categorizes occurrences faster and is not sensitive

Naïve Bayes Classifier :

Correctly Classified Instances	3594	95.281 %
Incorrectly Classified Instances	178	4.719 %
Kappa statistic		0.6008
Mean absolute error		0.0357
Root mean squared error		0.1382
Relative absolute error		48.9161 %
Root relative squared error		72.5471 %
To irrelevant features.		
Total Number of Instances	3772	

Bagging Classifier:

Correctly Classified Instances	3760	99.6819 %
Incorrectly Classified Instances	12	0.3181 %
Kappa statistic		0.9782
Mean absolute error		0.0035
Root mean squared error		0.0364
Relative absolute error		4.7406 %
Root relative squared error		19.1213 %
Total Number of Instances	3772	

IV. CONCLUSION AND FUTURE WORK:

The experimental results shows that the J48 classifier has the highest accuracy of 99.8% other classifiers like Bagging and Naïve Bayes has the accuracy of 99.6% and 95.2% respectively. The research has been carried out using classification data mining techniques for the analysis of thyroid disease. For this drive, K nearest neighbor, Support vector machine, Decision tree and Naive Bayes classifiers have been utilized. In our future work merging some of the algorithms and classifiers like bagging, boosting ,j48 and ensemble algorithm can be done and investigate the results for better results with the thyroid dataset.

REFERENCES

[1]Sehgal MSB, Gondal I (2014) K-ranked covariance based missing values estimation for microarray data classification. In: IEEE, 2004

[2] Bonner A (2004) Comparison of discrimination methods for peptide classification in tandem mass spectrometry. In: IEEE, 2004

[3] HalifeKodaz et al. Medical application of information gain based artificial immune recognition system (AIRS): Diagnosis of thyroid disease.

[4] G. Zhang, L.V. Berardi, An investigation of neural networks in thyroid function diagnosis, Health Care Manage. Sci. (1998)

[5]http://www.ncbi.nlm.nih.gov/pmc/articles/ PMC3169866/ (accessed Dec 2015)

[6] F. Temurtas, "A comparative study on thyroid disease diagnosis using neural networks," Expert Systems with Applications, vol. 36, 2009, pp. 944-949.

[7] K. Rajam, R. Jemina Priyadarsini, "A Survey on Diagnosis of Thyroid Disease Using Data Mining Techniques", International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 5, Issue. 5, May 2016, pg.354 – 358

[8] Liyong Ma,Chengkuan Ma,Yuejun Liu,Xuguang Wang,"Thyroid Diagnosis from SPECT Images Using Convolutional Neural Network with Optimization "Computational Intelligence andNeuroscience,Volume2019,https://doi.org/10.1155/2019/6212759.

[9] Eystraints G,"TND: A thyroid Nodule Detection System for analysis of Ultrasound Image and Videos", Springer Science and Business Media, LLC 2010

[10]Marissa Lourdes De Ataide1, Amita Dessai2Thyroid Disease Detection using Soft Computing Techniques, , International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072.

[11]. Jamil Ahmed and M.Abdul Rehman Soomrani,"TDTD: Thyroid disease type diagnosis"2016 international conference on intelligent systems engineering (ICISE), pp.44- 50, IEEE, 2016,
DOI:10.1109/INTELSE.2016.7475160

[12]. Anupam Shukla , Prabhdeep Kaur ,"Diagnosis of thyroid disorders using ANN" International advance computing, conference , IEEE 2009.

[13] Gurmeet Kaur and Er. Brahmaleen Kaur Sidhu, "Proposing Efficient Neural Network Training Model for Thyroid Disease Diagnosis." International Journal For Technological Research In Engineering Volume 1, Issue 11, ISSN (Online): 2347 - 4718, pp. 1383-1386, July-2014.

[14] Prerana, Parveen Sehgal, and Khushboo Taneja, "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, No. 2, April 2015.

[15]. Fatemeh Saiti,Afsaneh Alavi and Naini Mahdi Aliyari, Shoorehdeli," Thyroid Disease Diagnosis Based on Genetic Algorithms Using PNN and SVM", 3rd international conference on bioinformatics and biomedical engineering, 2009.

[16]. G. Rasitha Banu ," Predicting Thyroid Disease using LinearDiscriminant Analysis (LDA) Data Mining Technique ",Communications on Applied Electronics (CAE) – ISSN : 2394- 4714 Foundation of Computer Science FCS, New York, USA Volume 4– No12, January 2016.

[17] Roychowdhury S (2014) DREAM: diabetic retinopathy analysis using machine learning. In: IEEE, 2014

[18] Chetty N, Vaisla KS, Patil N (2015) an improved method for disease prediction using fuzzy approach. In: IEEE, 2015

[19] Joel Jacob et al. "Diagnosis of Liver Disease Using Machine Learning Techniques". (IRJET) Volume: 05 Issue: 04 | Apr-2018

[20]S. Sathiya Keerthi, Olivier Chapelle, Dennis DeCoste "Building Support Vector Machines with Reduced Classifier Complexity" Journal of Machine Learning Research, Vol: 7, PP 1493- 515, January - (2006).

[21] Shen X, Lin Y (2004) Gene expression data classification using SVM-KNN classifier". In: IEEE, 2004

[22]Joel Jacob et al. "Diagnosis of Liver Disease Using Machine Learning Techniques". (IRJET) Volume: 05 Issue: 04 | Apr-2018

[23]Xia C, Hsu W (2006) BORDER: efficient computation of boundary points. In: IEEE, 2006 Available from: http://en.wikipedia.org.Last accessed on Dec24].

[24] Apte & S.M. Weiss, Data Mining with Decision Trees and Decision Rules, T.J. Watson Research Center, http://www.research.ibm.com/dar/papers/pdf/f gcsaptewe issue_with_cover.pdf, (1997).