Diagnosis of Various Thyroid Ailments using Data Mining Classification Techniques

A.Nagaratnam*1, B.Deepika2, Shiek Ameer3, T.Sharoon4, CH.Ajay5

*1Department of CSE, BITS Vizag ,Visakhapatnam ,AndhraPradesh ,India
2Department of CSE, BITS Vizag ,Visakhapatnam ,AndhraPradesh ,India
3Department of CSE, BITS Vizag ,Visakhapatnam ,AndhraPradesh ,India
4Department of CSE, BITS Vizag ,Visakhapatnam ,AndhraPradesh ,India
5Department of CSE, BITS Vizag ,Visakhapatnam ,AndhraPradesh ,India

DOI: 10.29322/IJSRP.10.05.2020.p101117
http://dx.doi.org/10.29322/IJSRP.10.05.2020.p101117

Abstract- Classification is one of the most considerable supervised learning data mining technique used to classify predefined data sets the classification is mainly used in healthcare sectors for making decisions, diagnosis system and giving better treatment to the patients. In this work, the data set used is taken from one of recognized lab of Kashmir. The entire research work is to be carried out with ANACONDA3-5.2.0 an open source platform under Windows 10 environment. An experimental study is to be carried out using classification techniques such as k nearest neighbors, Support vector machine, Decision tree and Naïve bayes. The Decision Tree obtained highest accuracy of 98.89% over other classification techniques.

Index Terms- Thyroid disease, K-Nearest Neighbor, Support Vector Machine, Decision Tree, Naïve Bayes.

I. Introduction

Classification techniques play a vital as well as major role in analyzing survivability of diseases and providing facilities to reduce the cost to the patients. Now-a-days, Disease diagnosis has become very crucial because of occurrence of so many diseases every year.

People from all over the world have been suffering from various health issues like diabetes, heart disease, typhoid, tuberculosis, kidney disease etc [16] [17].Beside these health issues , thyroid disease have also been detected worldwide and thus become a serious endocrine health problem and an issue of concern. It is expected that in India about 42 million people suffer from thyroid disorders [2]. As per resent studies, women are 5 to 8 times more prone to thyroid disorders than men worldwide. It is caused by the improper secretion of thyroid hormones released from the thyroid gland which is one of the important organlocated in the front of the neck and below the Adam's apple of our body. The secretion of thyroid hormones from the thyroid gland are of two types i.e. 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 [18]. Thyroid disease occurs when thyroid gland stop to functioning properly and are mainly divided into hypothyroidism and hyperthyroidism [4].

The excess and deficient secretion amount of thyroid hormone causes hyperthyroidism and hypothyroidism respectively. The common symptoms of hyperthyroidism are sudden weight loss, rapid heartbeat, nervousness, etc. and hypothyroidism has weight gain, tiredness, weakness, feeling cold etc. One of the most common cause occurred due to hyperthyroidism is graves' disease [3]. The underestimated thyroid disease causes thyroid diseases. The Decision Tree outperformed over other techniques.

The rest of the paper is followed as: - Section 2 represents related work in diagnosis of thyroid storm and myxedema which may lead to death [12].

In this research work, a classification model is trained using classification algorithms like K- nearest neighbor (KNN),support vector machine (SVM), decision tree (DT) and Naïve bayes (NB) for the diagnosis of thyroid diseases. Section 3 contains dataset and methods. Section 4 represents the results and discussion. Section 5 contains conclusion and at last references are mentioned.

II. Materials and Methods

2.1 Data Collection

The data used for this work was collected from Ibadan Synoptic Airport through the Nigerian Meteorological Agency, Oyo State office. The case data covered the period of 120 months, that is, January 2000 to December 2009. The following procedures were adopted at this stage of the research: Data Cleaning, Data Selection, Data Transformation and Data Mining.

2.2 Data Cleaning

In this stage, a consistent format for the data model was developed which took care of missing data, finding duplicated data, and weeding out of bad data. Finally, the cleaned data were transformed into a format suitable for data mining.

2.3 Data Selection

At this stage, data relevant to the analysis was decided on and retrieved from the dataset. The meteorological dataset had ten (10) attributes, their type and description is presented in Table 1, while an analysis of the numeric values are presented in Table 2. Due to the nature of the Cloud Form data where all the values are the same and the high percentage of missing values in the sunshine data both were not used in the analysis

2.4 Data Transformation

This is also known as data consolidation. It is the stage in which the selected data is transformed into forms appropriate for data mining. The data file was saved in Commas Separated Value (CVS) file format and the datasets were normalized to reduce the effect of scaling on the data.

2.5 Data Mining Stage

The data mining stage was divided into three phases. At each phase all the algorithms were used to analyze the meteorological datasets. The testing method adopted for this research was percentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Thereafter interesting patterns representing knowledge were identified.

III. DATASET AND METHODS

The dataset used in this investigate work is a clinical dataset. The dataset was taken from one of the leading diagnostic lab in Kashmir. The dataset contains the record of 807 patients of almost all age groups. Out of 807 patients (224 Males and 583 Females) 553 belongs to normal, 218 belongs to hypothyroidism and 36 belongs to hyperthyroidism. The dataset has 6 attributes as: age, gender, TSH, T3, T4 and added classification attribute for indication of normal or hyperthyroidism or hypothyroidism.

Table 1 shows the description of dataset.

Serial No.	Attribute name	descriptio n	Value
1	Age	Age in years	Numeric
2	Gender	M-Male F-Female	Nominal
3	TSH	Continuo us	Numeric
4	T3	Continuo us	Numeric
5	T4	Continuo us	Numeric
6	Results	Normal Hyperthy roidism Hypothyr oidism	Nominal

K Nearest Neighbour (KNN)

K-nearest neighbour (KNN) is a supervised technique as well as non-parametric in nature. The input of K-NN depends on the K closest instances present in the feature space. The generated output depends on whether KNN is Classification or regression methods [7] [8].When prediction is required for undetected data 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].

Support Vector Machine

Support vector machine (SVM) is a supervised learning classification technique that are used to analyzethe data for regression and classification methods [9].It constructs an optimal hyper plane in a high- or infinite- dimensional space in which new examples are assigned to one group or the other one[10].The separation of data is achieved by the hyper plane is generally done, that has largest distance to the closest training data point of any class (so-called functional margin), since in general the greater the margin the smaller the generalization error of the classifier [19].

Decision tree

Decision Tree (DT) is tree like graph known as one of the most admired classification data mining technique that splits the dataset into parts on the decisions [5]. In decision tree, each internal node or non-leaf node represents a test on a particular attribute, each branch denotes the outcome of that test, and each leaf node has a class label. The paths through which a particular test data is to classify from root to leaf represent classification rules based on maximum information gain [6].

Naïve Bayes

Naïve Bayes (NB) is a simple classification algorithm for predictive modeling with clear semantics, representing and the probabilistic learning method based on Bayesian theorem [13]. Naive Bayes classifier assumes that the value of one attribute is not dependent on the value of any other attribute, and it assumes that the presence or absence of particular attribute doesn't affect the prediction process. Suppose there are m classes say C1, C2….Cn having a unidentified data sample X, Naive Bayesian classifier will predict an unknown sample X to the class Ci on the basis of class having highest probability [20].

$P(Ci |X > P(Cj |X)$ for $1 \leq j \leq m$, $j \neq i$ **Anaconda**

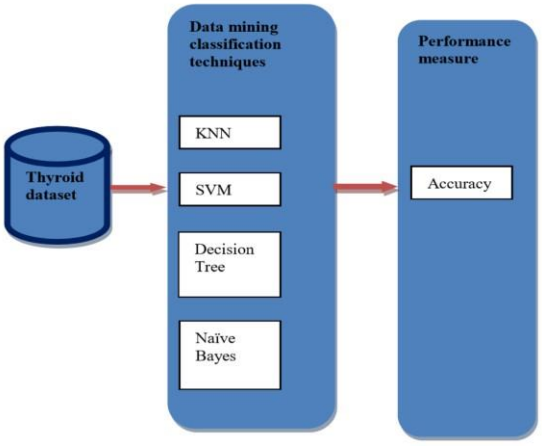
For implementation of our methods we used ANACONDA3-5.2.0 64 bit a free and open source platform distribution of python and R programming language with number of modules, packages and libraries that provides multiple ways of achieving classification problems. ANACONDA can be downloaded from the website [11].

K- Fold cross- validation

In k-fold validation, the whole dataset is divided into K equal size subsets and one of the subset K is taken as test data and the remaining K-1 folds acts as training data. Thus different test results exist on each iteration and at last average of these results gives the test accuracy of the algorithm [12]. In this study 10-fold cross-validation is used to find out the classification accuracy by the classification methods.

System implementation

The thyroid dataset has three output categories. First one is Normal, second one is hypothyroidism and third one is hyperthyroidism. The thyroid dataset of five input attributes Age, Gender, TSH, T3 and T4 is supplied to the classifiers of KNN, SVM, Decision Tree and Naïve bayes in PYTHON to classify the data. The performance of each classifier is evaluated in terms of accuracy as shown in figure 1.



In this work, Out of four classifiers, The Decision Tree shows highest test accuracy of 98.89% over other classifiers. The accuracy comparison among classifiers is shown in figure 2. The highest accuracy of our method is higher than some of the methods available in literature as shown in table 2.

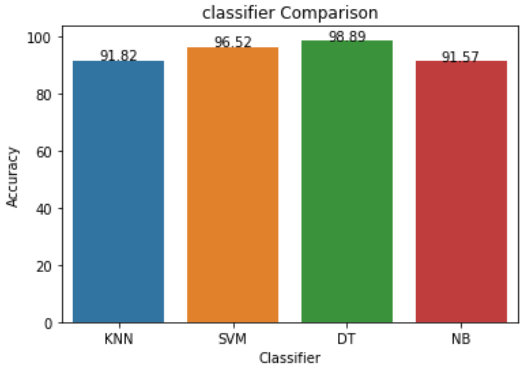


Figure 2. Accuracy comparison of classifiers

Study	Method	Highest Accuracy
Ozyılmaz and Yıldırım [14]	MLPNN with BP (3×FC)	91.14%
MLPNN with FBP (3×FC)		
RBF (3×FC)		
CSFNN (3×FC)		
Polat et al. [15]	AIRS (10×FC)	85.00%
AIRS with Fuzzy (10×FC) weighted preprocessing		
keles et al. [21]	ESTDD	95.33%
F. Temurtas [12]	MLNN with LM (3×FC)	94.81%
PNN (3×FC)		
LVQ (3×FC)		

Table-2 shows the highest classification accuracy obtained by our method over other methods used for the diagnosis of thyroid diseases.

IV. CONCLUSION

The work has been done using classification data mining techniques for the diagnosis of thyroid disease. For this purpose, K nearest neighbor, Support vector machine, Decision tree and Naive Bayes classifiers have been used. The Decision Tree classifier outperformed over other classifiers. However, if we merge it with any other classification technique such as neural network, then the result might be even better as compared to what we got with the current study.

REFERENCES

[1] K.Saravana Kumar, Dr. R. ManickaChezian, "Support Vector Machine and K- Nearest Neighbor Based Analysis for the Prediction of Hypothyroid. International

[2] Journal of Pharma and Bio Sciences",volume - 2,Issue - 5,page no-(447-453),2014

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

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

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

[6] 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/fgcsaptewe issue_with_cover.pdf, (1997).

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

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

[9] 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).

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

[11] www.anaconda.com.

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

[13] G. H. John and P. Langley, "Estimating Continuous Dis- tributions in Bayesian Classifiers," Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, San Francisco, 1995, pp. 338-345.

[14] Ozyılmaz, L., Yıldırım, T. (2002). Diagnosis of thyroid disease using artificial neural network methods. In Proceedings of ICONIP'02 9th international conference on neural information processing (pp. 2033-2036). Singapore: Orchid Country Club.

[15] Polat, K., Sahan, S., & Gunes, S. (2007). A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted preprocessing for thyroid disease diagnosis. Expert Systems with Applications, 32, 1141-1147.

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

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

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

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

[20] K. Pavya et al. "Diagnosis of Thyroid Disease Using Data Mining Techniques: A Study". (IRJET) Volume: 03 Issue: 11 | Nov - 2016.

[21] Keles, A., and Keles, A., ESTDD: Expert system for thyroid diseases diagnosis. Expert Syst. Appl. 34(1):242-246, 2008.

[22] Dogantekin, E., Dogantekin, A., and Avci, D., An expert system based on generalized discriminant analysis and wavelet support vector machine for diagnosis of thyroid diseases. Expert Syst. Appl. 38(1):146- 150, 2011.

[23] M.P.Gopinath "Comparative Study on Classification Algorithm for Thyroid Data Set". International Journal of Pure and Applied Mathematics Volume 117 No. 7 2017, 53-63.Cite this article as : Umar Sidiq, Dr. Syed Mutahar Aaqib, Dr. Rafi Ahmad Khan, "Diagnosis of Various Thyroid Ailments using Data Mining Classification Techniques", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5

[24] Issue 1, pp. 131-136, January-February 2019. Available at doi : https://doi.org/10.32628/CSEIT195119 Journal URL :http://ijsrcseit.com/CSEIT195119

AUTHORS

First Author – A.Nagaratnam, Department of CSE, BITS Vizag ,Visakhapatnam ,AndhraPradesh ,India
Second Author – B.Deepika, Department of CSE, BITS Vizag ,Visakhapatnam ,AndhraPradesh ,India
Third Author – Shiek Ameer, Department of CSE, BITS Vizag ,Visakhapatnam ,AndhraPradesh ,India
Fourth Author – T.Sharoon, Department of CSE, BITS Vizag ,Visakhapatnam ,AndhraPradesh ,India
Fifth Author – CH.Ajay, Department of CSE, BITS Vizag ,Visakhapatnam ,AndhraPradesh ,India