Predicting Thryroid disease using Machine learning methods

Ayisha khanum[#1], Chethan M.S[*2]

[#]*PG Student, CSE Department , Sridevi institute of technology and management, Visveswaraya Technological University*

* *Assistant Professor, CSE Department, Sridevi institute of technology and management, Visveswaraya Technological University*

ayishakhanum.cs@gmail.com, mailtochethanms@gmail.com

Abstract - Hypothyroidism or hyperthyroidism is a major disease in India which arises due to malfunctioning of thyroid hormones. Medical industry has enormous quantity of data, but the bulk of this data is not processed. For proper diagnosis data must be processed accurately. For accurate processing intelligent Machine Learning techniques are widely used. In this paper an attempt is made to analyze Logistic regression and Support Vector Machine (SVM) for multiclass classification of thyroid dataset.Performance of these techniques ison basis of Precision, Recall, F measure, ROC, RMS Error and accuracy. Our analysis shows that logistic regression is more efficient than SVM for multiclass classification of thyroid dataset.

Keywords — Thyroid, Machine learning, Hypothyroidism, hyperthyroidism, prediction

I. INTRODUCTION

The thyroid is a little gland in the neck that produces thyroid hormones. It may produce too much or too small of these hormones. Hypothyroidism is a situation in which thyroid gland is not able to produce sufficient thyroid hormones. These hormones regulate metabolism of the body and further affects how the body uses energy. Lacking the accurate amount of thyroid hormones, body's normal functions start to slow down and body faces changes each day (hello, mood swings, happy,sadfatigue, depression, constipation, feeling cold, weight gain, muscle weakness, dry, thinning hair,slowed heart rate). Hyperthyroidism is a condition when thyroid gland produces too much thyroid hormones[2]. Symptoms of hyperthyroidism are nervousness, restlessness, inability to concentrate ,increased appetite, difficulty sleeping, itching, hair loss, nausea and vomiting. For diagnosis entire medical history and physical tests (free T4, T3Test, Cholesterol test, TSH Test)are required. As these test produces large amount of data and MLcan be used forfinding important features from large amount of data. Due to this specialty of ML can be usedin combination with medical science for the accurate diagnosis of hypo thyroidismdisease[1]. A number of ML techniqueshave been evolved and in order to achieve best accuracy of a model ensembles are widely used[7].

A nonparametric test exposed accurately massive contrasts between FV-PTC and ordinary thyroid using both parameters (p <; 0.05). These preliminary results advised that phantom based quantitative ultrasound imaging using machine learning is valuable for the administration of thyroid tumour [4]. The outcomes demonstrates that the JET model providedexact depictions of thyroid knobs when contrasted with LD and copes with the confinements of the past thyroid depiction approaches [5].Polat considered artificial immune- recognition system (AIRS) for thyroid diagnosis, and found 81% accuracy [6]. Keles proposed an expert system using Neuro-Fuzzy classification method for thyroid diagnosis and found an accuracy of 95.33% [8].Temurtas did thyroid disease diagnosis with the help of Multi Layer Perception (MLP) with Levenberg Marquardt-LM algorithm was done and found accuracy of 93.19%[9]. Wavelet used Support Vector Machine (WSVM) and Generalized Discriminant Analysis (GDA) methods for thyroid diagnosis and got 91.86% classification accuracy[10]. Chen did optimization using particle swarm optimization for thyroid disease, and found accuracy of 97.49% [11].Chen,Hui-Ling proposed an expert system, called Fisher Score Particle Swarm Optimization Support Vector Machines (FS-PSO-SVM) and was evaluated on thyroid disease dataset [12].Binary logistic regression, naïve Bayes classifier,support vector machine (SVM), and radial basis function neural network (RBFNN) were analyzed forthyroid diagnosis[13–14]. Analyzed various ML techniques for medical diagnosis[15]. Cure of disease is a regular concern for the health care practitioners, and the errorless diagnostic at the right time for a patient is very important. Recently, by some advanced diagnosis methods, the common medical report can be generated with an additional report based on symptoms. The different questions like ‘‘what are the causes for affecting the thyroid?’’, ‘‘Which age group of people are affected due to thyroid?’’, ‘‘what is the relevant treatment for a disease?’’, etc. may find answers on implementing machine learning methods.

The large amount of data can be handled using the machine learning techniques. Classification models are well suited for the classification and distinction of the data classes. The handling of both numerical and categorical values can be done by the classification processes. Classification is a two-step classification model in the step one, based on some training data, a model is constructed, and in step two, an unknown tuple is given to the model to classify into a class label [6]. In human life, the classification has a great influence. The comparison of different classification techniques is a non-trivial and has a great dependency on the data set properties. In the statistics community, logistic regression, decision tree and k-nearest neighbor have got an esteemed position for classification problems [7]. Based on the research works and literature review, very little work has been done in the classification methods of patients pruned by the thyroid disease. The methods of classification used are the well-known methods. To focus on the above-discussed issues, this paper explains the use of three classification machine learning algorithms: logistic regression classification, decision tree classification and nearest neighbors classification to classify the people pruned by thyroid disease using the thyroid disease database. The paper explain in detail about the preparation, training and testing of the data, step-by-step description of each of the techniques used, and a comparison of the accuracy of the methods used in the prediction.

II. THYROID DETECTION USING ML

Logistic Regression Logistic regression is a ML technique used to allocate records into discrete set of classes. Linear regression produces continuous number values as output. Linear Regression can predicts the student’s test attain on a scale of 0 - 100 (predictions are continuous as range is required). Logistic Regression[1] can be used to predict whether the student is pass or fail. Logistic regression predictions are discrete (only exact values or categories are permissible). Binary logistic regression: this will take two values 0 or 1. Binary logistic regression: this will take two values 0 or 1

$Y=b0+b1X+e$

To map predicted values to probabilities, sigmoid function is used. This function maps any real value into another value between 0 and 1. In machine learning, sigmoid is used to map predictions to probabilities.

$\sigma t=11+e-t$

The output of this is the estimated probability whichtells how confident can predicted value be actual value when X is given as an input.

Support Vector Machine SVM[12] is used for the classification of both linear and non-linear data. This technique is derived from statistical learning theory given by Vipnik in 1992. SVM technique solvesthe problem by finding out the hyper-plane with maximum margin.For nonlinearlyseparable data, it transformsthe training data into a higher dimension space by doing non linear mapping. By transforming it into high dimensional space, it searches for linear optimal separating hyper-plane. This transformation technique into high dimension always helps in searching for an optimal hyper-plane using support vectors and margins[13]. SVM achieved classification by finding optimal MMHand minimizing the classification errors.

Decision Tree

Total serum thyroxin and total serum triiodothyronine are selected as the feature names for making the decisions. The class that the output produce will be class 0 (having thyroid) and class 1 (normal). To prepare the model, data set is divided into training set (70%), validation set (15%) and test set (15%). On evaluating the performance of the algorithm, it shows validation misclassification percentage of 12.5% and test misclassification percentage of 3.125%. The confusion matrix is drawn here for calculating the accuracy of the model is shown in Fig. 1d. The accuracy of this matrix can be calculated using the Eq. (1). Here, putting the values in the above equation

Accuracy = 10+12/(10+0)+(1+21)

Accuracy = 31/32 = 0.968

So, the accuracy calculated here is 96.8%.

kNN : While applying the algorithm at random chosen a point [4.2 1.2] as query point. The true class of the query point is 0. On applying the algorithm, the nearest neighbors of the query point are: ([4.2 1.2] [4.2 0.7] [4.7 1.1] [3.6 1.5] [4.7 1.8]), classes of the nearest neighbors are: ([1] [0] [0] [0] [0]) and predicted class for query point is also 0. The visualization of working of kNN. On evaluating the performance of the k-NN classifier, the test misclassification percentage = 3.125%. For calculating the accuracy of the matrix, used. Here, putting the values from the matrix,

Accuracy = 6 + 22/(6+3)+(1+22)

Accuracy = 28/32 = 0.875

So, the accuracy calculated here is 87.5%. From our research work, it is shown that how can thyroid disease be predicted and give an intution how to apply the logistic regression, decision tree classification and kNN algorithms.

The efficiency of an algorithm depends upon the data set and its features selected for the prediction. Some papers written during 2018–2020 have less accuracy than proposed algorithms, and some algorithms have a better accuracy which is due to the data set they have chosen. The paper given in below in Ref. [13] has shown less accuracy in case of decision tree, while in case of kNN they have better accuracy shown in Table 2: compare with previous work. The UCI thyroid repository itself contains many data sets for thyroid disease. For proposed work, ‘‘new-thyroid’’ data set has been taken [8]. The paper authors [13] might have taken different data set of the same UCI thyroid repository. This is the reason of variation of result.

Rafikhan et al. [14] has used a clinical data of Kashmir of 807 patients and UCI thyroid repository of ‘‘new thyroid’’ has only 215 instances. Proposed method has not taken this data set for thyroid prediction; it will consider in future work and measure accuracy using decision tree and kNN. Hence, according to the data set which is used in this work, the accuracy obtained is satisfactory. The current scenario is of the developing of the models that help in the various sectors of life using the machine learning. The availability of data and its generation day by day increased a chance for the computer scientists to make prediction and analysis on such data sets that make the human life better and comfort. This study is concern with this motivation. The prediction and classification of any data depends on the data set itself and the various algorithms that are used. If anyone organizes a better data set of real time and applies various other machine leaning and deep learning algorithms such as SVM, Naı¨ve Bayes, auto encoders, ANNs and CNNs then further better results may be achieved.

III. DATASET

To detect Thyroid Disease, dataset wastaken from UCI repository. The Thyroid dataset has 30 attributes and 3772 records.

In Thyroid dataset,Class is nominal variable having four different values. From above Figure we can find that age, TSH, T3,TT4,T4U,FTI,TBG,Referral source are numeric variables.Sex, on thyroxine, query on thyroxine,on antithyroid medication,sick, pregnant,thyroid surgery,I131 treatment, query hypothyroid, query hyperthyroid,lithium,goitre and all the remaining attributes are nominal having two values.

IV. RESULTS

Logistic regression and SVM machine learning techniques are used to analyze Thyroid dataset using Weka version 3.6. Initially dataset had 30 attributes and 3772 records. Logistic regression and Support Vector Machine are compared on basis of Precision,Recall,F measure,ROC and RMS Error.Figure 1. shows confusion matrix obtained by logistic regression. Figure 2. shows results obtained usinglogistic regression..Figure 3 shows confusion matrix obtained by SVM.Figure 4 shows results obtained using

SVM Precision	Recall	F measure	ROC	RMS Error	
Logistic regression	.968	.968	.967	.979	.112
SVM	.888	.936	.91	.594	.3213

Table 1. depicts the comparison between Logistic regression and Support Vector Machine (SVM) on basis of Precision, Recall, F measure, ROC and RMS Error.Figure 5 shows the Graphical representation of Logistic regression and Support Vector Machine (SVM) on basis of Precision,Recall,F measure,ROC and RMS Error. Figure 5 showsthat logistic regression is performed better than Support Vector Machine in allthe parameters like Precision,Recall,F measure,RMS Error.Figure 6 and Figure 7 shows the ROC curve of Logistic regression and Support Vector Machine (SVM). Logistic

Regression outperformed Support Vector Machine as shown in ROC curve. Table 2 exhibits the comparison between Logistic regression and Support Vector Machine (SVM) on basis of Accuracy.Figure 8 shows the Graphical representation of Logistic regression and Support Vector Machine (SVM) on basis of accuracy. Figure 8 shows that logistic regression is performed better than Support Vector Machine on the basis of accuracy. No doubt, SVM is more stable technique than Logistic Regression for binary classification. But in case of multiclass classification Logistic Regression outperforms SVM.Hence, it can conclude that SVM performance deteriorates as number of classes increases.

V. CONCLUSION

ML techniquescan be used for Thyroid detection. In this paper logistic regression and SVM are usedto predictThyroid.These techniques are compared on the basis of Precision,Recall,F measure,ROC, RMS Error and accuracy. This paper showed that instead of SVM,logistic regression turns out to be best classifier for Thyroid detection when number of classes increases. The idea for thyroid disease diagnosis and therapy is represented by the functional behavior of the thyroid disease and is the key in most thyroid diseases. The basis of classification of thyroid disease is euthyroidism, hyperthyroidism and hypothyroidism which are denoting normal, excessive or defective levels of thyroid hormones. The state euthyroidism depicts the normal production of thyroid hormones and normal levels at the cellular level by the thyroid gland. The state hyperthyroidism is clinical symptom due to excessive circulation and intracellular thyroid hormones. The state hypothyroidism is most of due to the lack of thyroid hormone generation and poor alternate therapy

REFERENCES

[1] Chen Ling, Li Xue, Sheng Quan Z, Peng W-C (2016) Mining health examination records—a graph-based approach. IEEE Trans Knowl Discov Eng 28:2423–2437

[2] Temurtas F (2009) A comparative study on thyroid disease Ulutagay G (2012) Modeling of thyroid disease: a fuzzy inference system approach. Wulfenia J 19(1):346–357

[3] Monaco Fabrizio (2003) Classification of thyroid diseases: suggestions for a revision. J Clin Endocrinol Metab 88:1428–1432

[4] Ionita I, Ionita L (2016) Prediction of thyroid disease using data mining techniques. Broad Res Artif Intell Neurosci 7(3):115–124

[5] Gorade SM, Deo A, Purohit P (2017) A study of some data mining classification technique. Int Res J Eng Technol 4(4):3112–3115

[6] Bichler M, Kiss C (2004) A comparison of logistic regression, knearest neighbor, and decision tree induction for campaign management. In: Proceedings of the tenth Americas conference on information systems, New York

[7] http://archive.ics.uci.edu/ml/machine-learning-databases/thyroiddisease/

[8] Peng CYJ, Lee KL, Ingersoll GM (2002) An introduction to logistic regression analysis and reporting. J Educ Res 96(1):3–14

[9] Mesaric´ J, Sebalj D (2016) Decision trees for predicting the academic success of students. Croat Oper Res Rev 7:367–388

[10] Patel BN, Prajapati SG, Lakhtaria K (2012) Efficient classification of data using decision tree. Bonfring Int J Data Min 2(1):6–12

[11] Introduction to machine learning edition 2, by Ethem Alpaydin.

[12] https://kkpatel7.files.wordpress.com/2015/04/alppaydin_machinelearning_2010.pdf

[13] Tyagi A, Mehra R (2018) Interactive thyroid disease prediction system using machine learning technique. In: 5th IEEE international conference on parallel, distributed and grid computing (PDGC-2018), 20–22 Dec, Solan, India

[14] Sidiq U, Aaqib SM, Khan RA (2019) Diagnosis of various thyroid ailments using data mining classification techniques. Int J Sci Res Comput Sci Eng Inf Technol 5(1):2456–3307.