International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Machine Learning Models for Relevant Feature Identification and Classification of Thyroid Data

S. Nandhinidevi, S. Poorani, P. Gokila Brindha

Abstract: ***Inappropriate creation of thyroid glands turn into prime subject of concern amongst Indian women. The two key chaos of thyroid which ought to be taken care at the earliest are hypothyroidism and hyperthyroidism. The improper secretion of thyroid may leads to obesity, fertility related problems, feeling depressed, etc,. Most of the thyroid problems can be managed if it has been properly treated. In recent years a number of models have been developed to investigate the thyroid-disorder. The laboratory tests are conducted to find the levels of the hormones and some of the physical examinations are used to identify the presences of thyroid. These examinations and test results are taken as the feature for developing the model. The feature importance can promote the performance of ML algorithms. The core intention of this study is to improve the classification performance by identifying the relevant features before classification. In this work, random-forest model is considered for identifying the important features and KNN algorithm is implemented for multiclass-classification to envisage the kind of thyroid chaos. Applying KNN after the feature selection improves the prediction accuracy. The developed model can be used to predict the presence of thyroid so that it can be treated accordingly.***

Keywords: ***Feature selection, classification, Multiclass classification, KNN, Random Forest, Machine Learning, Thyroid***

I. INTRODUCTION

In general, Women are embellished by thyroid mayhem, which set the base for an assortment of fitness nuisance like hormonal difference, increase in weight, decrease in weight and others. Common statistics reports that women may have a chance of thyroid up to eight times higher than men. In India nearly 42 million people have thyroid disorder. Hyperthyroidism exists mainly in the west and south zones and hypothyroidism exists mostly in the North zone of India[1]. Various ML methods such as Naïve-Bayes, K-nearest neighbour, and SVM were used in thyroid-disorder prediction. The KNN provides improved performance than Naïve-Bayes[2]. Random forest is a nonparametric system which constructs an ensemble-model with decision trees by using arbitrary subsets of traits and bagged trials of training data[3]. The arbitrary investigation of features in the Random Forest leads to good feature selection process [4]. The aim of this study is to improve the performance of multiclass classification in the analysis of thyroid data with feature selection process. The KNN method provides the best result in thyroid disease classification. First, we implement the KNN classifier without feature selection and find the accuracy and then we implement the same KNN in the same dataset after feature selection using Random Forest Algorithm.

II. LITERATURE SURVEY

In the digital era, dealing with the huge information is a difficult assignment among the scientists. Since the information are amassed through different information procurement procedures, strategies, and gadgets. In common, the high dimensional information contains unessential and the repetitive features. Subsequently, these issues can be handled by the feature selection. Due to presence of noisy, redundant and irrelevant dimensions, they can not only make learning algorithms very low and even reduce the performance of learning process. Feature selections are capable of choosing a small subset of related features from the original ones by removing noisy, irrelevant and repetitive features.

Genetic algorithm based Random Forest (GARF) [5] strategy was proposed for feature selection from the positron emission tomography (PET) images and clinical data. A new multi-parametric fitness function is combined with genetic algorithm. GARF perform the feature selection in 2 steps. In first step correlated attributes were eliminated and in second step feature selection was done using genetic algorithm along with RA.GARF improves the outcome prediction compared to other tested methods.

ICFS [6]method was proposed to automate the seizure detection by selecting features using random forest classifier. The efficiency of this method is improved in the performance of classification problems. BSRF [7] was proposed to find the credit risk accurately and make decisions. This Bolasso (Bootstrap-Lasso) selects most relevant features from pool of features by using random forest classifier to yield better classification accuracy. Modified RELIEF formulation [8] was proposed for segmenting latent finger prints to distinguish between ridge and non-ridge patterns. It performed the feature selection and classification using Random Decision Forest(RDF).RDF used repetitive random sub-sampling strategy to provide strong and faster results for overlapping features

Revised Manuscript Received on March 30, 2020.
* Correspondence Author
S. Nandhinidevi*, Department of Computer Technology-UG ,Kongu Engineering College,Perundurai,Erode,TamilNadu,India.
Email:nandhinidevi@kongu.ac.in
S. Poorani, Department of Computer Technology-UG ,Kongu Engineering College,Perundurai,Erode,TamilNadu,India.
Email:poorani@kongu.ac.in
P. Gokila Brindha, Department of Computer Technology-UG ,Kongu Engineering College,Perundurai,Erode,TamilNadu,India.
Email: brindha@kongu.ac.in

SRF method was proposed for feature subspace selection for high dimensional data[9]. It increased the strength and continue the diversity of the trees in random forest.

This approach recognized two groups of features: strong and weak features. And also, this model reduces the occurrence of generalization error and increase the classification accuracy with minimum computational cost.

Random Forest classifier was used in IOT based health care system to classify the E-health data with optimal features[10]. Better classification accuracy was acquired and compared with Gaussian mixture model and logistic regression.

clustering and classification were combined for feature selection in microarray data. This, random approach eradicate correlations between decision trees to improve the accuracy of the method, such as the ensemble method characteristics.

III. METHODOLOGY

This work implements two types of methods. The first method does not include any feature selection method. The data classification is done with all the features. The second method includes the feature selection process before classification. Finally the performance of two methods are compared. The algorithm-1 and algorithm-2 describes the step by step process of first method and second method used in this work respectively.

Algorithm-1:
Step1: Read the Input
Step 2: Train the KNN classifier
Step 3: Test the KNN classifier
Step 4: Evaluate the performance

Algorithm-2:
Step1: Read the Input
Step 2: Apply random forest method for feature selection
Step 2: Train the KNN classifier
Step 3: Test the KNN classifier
Step 4: Evaluate the performance

Dataset

The dataset is taken from UCI library, which includes 21 variables. But, not all the variables are important. So we applied feature selection and then only 11 variables are considered for classification.

Feature Identification and Random-Forest

Random forest is an ensemble Machine Learning Technique. Many applications implements Random Forest to classify the dataset like Network intrusion detection, Email spam detection, gene classification, Credit card fraud detection, and Text classification. Random forests can be used to rank the significance of variables in a regression or classification problem in a likely way. Large number of predictors can readily be handled by Random forest algorithm and also it is more interpretable. It provides estimates of what variables are important in the classification .So, We implement the Random Forest for feature selection on Thyroid data.

KNN for multiclass classification

Multiclass classification is a fast developing area of machine learning. Text classification is the main application area of multi-label classification techniques. However, relevant works are found in areas like bioinformatics, medical diagnosis, scene classification and music categorization.we implement the KNN on thyroid medical data set for multiclass classification after the feature selection.

K-Nearest-Neighbour is one of the most widely used simple and straight forward data mining techniques in classification problems. Its simplicity and relatively high convergence speed make it a popular choice. It plays a vital role in multi class classification rather than Binary classification.

IV. RESULTS AND DISCUSSION

Our feature selection method provides the top 11 features which is given in figure.1&2. Among those 11 features only 6 features having the highest mean value. So these six features are considered for classification. First the classification is done with all traits and it provides the accuracy value as 99.86 shown in Table 1 , then the selected features(top six) are considered for classification. Finally ,the classification model with selected features shows better result as 100% accuracy which is shown in Table 2.

S.No	Feature	MeanImp	Decision
1	TSH	74.921303	Confirmed
2	On_thyroxine	35.387856	Confirmed
3	FTI	27.335598	Confirmed
4	T3	23.753211	Confirmed
5	TT4	20.027001	Confirmed
6	T4U	10.248280	Confirmed
7	Thyroid_surgery	8.636580	Confirmed
8	On_antithyroid_medication	7.205765	Confirmed
9	Psych	3.470349	Confirmed
10	Sex	3.242307	Confirmed
11	X.inputs.Age	1.986602	Confirmed

Fig.1 Top 11 features from the dataset

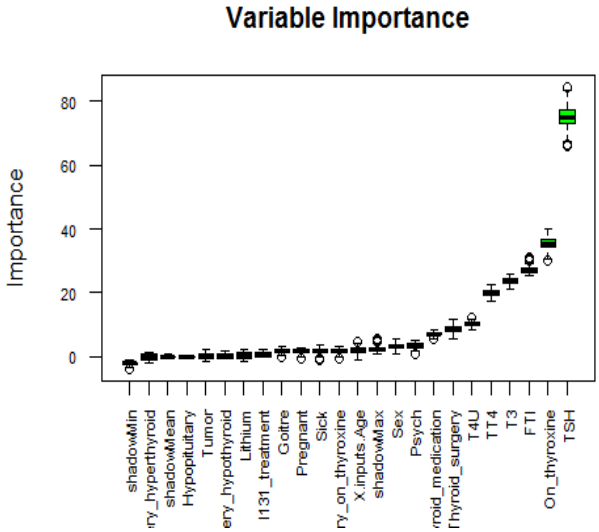

Fig. 2. variable importance derived from Random Forest

The table I shows the confusion matrix and accuracy of the classification before the implementation of Feature selection

Table –I: Confusion Matrix of KNN Classification before feature selection

Disease - Category / Prediction	1	2	3
1	15	0	0
2	0	25	0
3	0	1	679
Accuracy = 99.86111 %			

The Table II shows the confusion matrix and accuracy of the classification after the implementation of Feature selection using Random Forest technique

Disease - Category / Prediction	1	2	3
1	19	0	0
2	0	42	0
3	0	0	659
Accuracy = 100 %			

Table I and Table II depicts that Random Forest plays a vital role in improving the classification accuracy.

V. CONCLUSION

Accuracy is the one of the performance assessment criteria. In medical data set, classification plays vital role in prediction of disease. In this paper we have taken the dataset with 21 features, extracted the important features and developed the model using KNN. We found that the classifier's accuracy is increased after selecting the important features We made 100% accuracy with feature identification in multiclass classification of thyroid data. We can conclude that prediction of any disease can be done accurately by applying classification models with feature selection.

REFERENCES

1. https://economictimes.indiatimes.com/magazines/panache/over-30-indians-suffering-from-thyroid-disorder-survey/articleshow/58840602.cms May 25, 2017.
2. Khushboo Chandel, Veenita Kunwar, Sai Sabitha, Tanupriya Choudhury & Saurabh Mukherjee, "A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques", CSIT, December 2016 ,pp.313–319.
3. Thanh-Tung Nguyen, oshua Zhexue Huang and Thuy Thi Nguyen, "Unbiased Feature Selection in Learning Random Forests for High-Dimensional Data", volume 2015 |Article id 471371 | 18 pages | https://doi.org/10.1155/2015/471371.
4. Jeremy Rogers, Steve Gunn, "Identifying Feature Relevance Using a Random Forest", International Statistical and Optimization Perspectives Workshop "Subspace, Latent Structure and Feature Selection" SLSFS 2005: Subspace, Latent Structure and Feature Selection pp 173-184, https://link.springer.com/chapter/10.1007/11752790_12
5. Desbordes Paula,b, Ruan Sua, Modzelewski Romaina,c, Vauclin Sébastienb,Vera Pierrea,c, Gardin Isabelle, "Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier, Computerized Medical Imaging and Graphics" ,Volume 60, September 2017, Pages 42-49, https://doi.org/10.1016/j.compmedimag.2016.12.002
6. M Mursalin, Y Zhang, Y Chen, NV Chawla , "Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier", Neurocomputing, 2017 -Volume 241, 7 June 2017, Pages 204-214- Elsevier
7. Nisha Arora Pankaj DeepKaur , "A Bolasso Based Consistent Feature Selection Enabled Random Forest Classification Algorithm: An Application to Credit Risk Assessment"(2019)Applied Soft Computing-Volume 86, January 2020, 105936
8. Anush Sankaran, Aayush Jain, Tarun Vashisth, Mayank Vatsa , Richa Singh,"Adaptive latent fingerprint segmentation using feature selection and random decision forest classification" , Information Fusion Volume 34, March 2017, Pages 1-15
9. Yunming Ye a,e,n, QingyaoWua,e, Joshua Zhexue Huang b,d, MichaelK.Ng c, XutaoLi a,e ,"Stratified sampling for feature subspace selection in random forests for high dimensional data" , Pattern Recognition Volume 46, Issue 3, March 2013, Pages 769-787
10. Lakshmanaprabu, S.K., Shankar, K., Ilayaraja, M. *et al.* "Random forest for big data classification in the internet of things using optimal features". *Int. J. Mach. Learn. & Cyber.* **10,** 2609–2618 (2019). https://doi.org/10.1007/s13042-018-00916-z
11. Lei Ma, Tengyu Fu , Thomas Blaschke , Manchun Li , Dirk Tiede, Zhenjin Zhou, Xiaoxue Ma and Deliang Chen, "Evaluation of Feature Selection Methods for Object-Based Land Cover Mapping of Unmanned Aerial Vehicle Imagery Using Random Forest and Support Vector Machine Classifiers", *ISPRS Int. J. Geo-Inf.* **2017**, *6*(2), 51; https://doi.org/10.3390/ijgi6020051

AUTHORS PROFILE

Ms.S.Nandhinidevi, received M.Sc degree at Navarasam Arts and science college for women, Bharathiar University from the Department of Computer Science, India in 2004 and also received M.Phil Degree from Bharathiar University, in 2008. She has 14 years of teaching experience. She published 4 articles in International journals. She has also presented papers in National Conferences. she is also currently working as an Assistant Professor in the department of Computer Technology-UG, Kongu Engineering College, affiliated to Anna University Tamilnadu, India.

Ms. S. Poorani, received M.Sc degree at Sri Vasavi College, Bharathiar University from the Department of Computer Science, India, in 2004. She has 12 years of teaching experience. She published 4 articles in International journals. She has also presented papers in National Conferences. Her area of interest includes data mining and big data. She is currently pursuing the Ph.D. degree working with Dr. P. Balasubramanie. Simultaneously she is also currently working as an Assistant Professor in the department of Computer Technology-UG, Kongu Engineering College, affiliated to Anna University Tamilnadu, India.

Ms. P.Gokial Brindha, completed M.Sc (Information Technology) from Anna University in the year 2007. She has 11 years of teaching experience. She published 2 research papers in International journals and presented 2 papers in the conferences. Presented seminars on the "R Tool", "Tableau – Visualization Tool" and "Data Analytics using R". She is working as an Assistant Professor in the department of Computer Technology-UG, Kongu Engineering College affiliated to Anna University, Chennai,Tamilnadu. Perusing PhD under Anna University in the area of Data Mining and Machine Learning.