International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S3, September 2019

Constructing a System for Analysis of Machine Learning Techniques for Early Detection of Thyroid

Sayyad Rasheeduddin, Kurra Rajasekhar Rao

Abstract: ***Thyroid is an unending and complex infection caused by unedifying levels of TSH (Thyroid Simulation Hormone) or by thyroid organ problems themselves. Hashimoto's thyroid is the most widely recognized cause of hypothyroidism. The body makes anticorps that pulverize the thyroid organ in an auto-safe condition. It offers machine learning algorithms in the system proposed to predict thyroid disease in disease-intensive societies effectively. This is a serious concern for public health even though it is massively increasing in many countries. This shows that the problem must be predicted as urgently as possible to overcome the shortcomings of previously existing clinical decision-making tools with low precision. This paper examines numerous machine learning strategies for osteoporosis prediction. The paper examines and assesses the use of the strategy of feature selection combined with classification techniques. WEKA's classification techniques are used to measure an osteoporosis data set. The results are calculated by means of various test options, including 10-fold cross-validation, training sets and the percentage divided with and without the selection method. The results are compared with correctly classified instances, runtime, kappa and absolute mean values for experiments with and without feature selection techniques.***

Keywords: Classification, Data mining, Machine Learning, Decision Tree

I. INTRODUCTION

A doctor's examination or a number of blood tests is traditional ways of diagnosing thyroid disorder. A doctor's evaluation or a number of blood tests is traditional ways of diagnosing thyroid disorder. The primary task is to make disease diagnosis more accurate in the early stages. Data mining plays an important role for disease diagnosis in the medical field. It provides many classification methods for predicting the accuracy of the disease. Over the years, hospitals and clinics have collected a large number of patient data. Health prevention is a continuous concern of the doctors, and due to the implied risk, the right diagnosis at the right time is crucial for a patient. In the recent past, a supplementary report can be accompanied by a decision support system or other advanced symptom-based diagnostic techniques. They focus on the use of classification methods and the identification of the best algorithm for thyroid classification disorders. Machine learning seems to be an area for the researchof artificial intelligence using different data classification methods.

Manuscript published on 30 September 2019.
***** Correspondence Author (s)
Sayyad Rasheeduddin, Department of Computer Science and Engineering.
Dr.Kurra Rajasekhar Rao, Department of Computer Science and Engineering.

In clinical settings, severe engineering strategies are being used to prevent disease and have demonstrated higher accuracy to be diagnosed than conventional methods [9]. Machine learning was widely used to support vector machines, random forests (RFs) and artificial neural networks (ANNs)[9]. this study evolves osteoporosis prediction models using a number of WEKA machine learning methods, including SVMs, RFs and ANNs. In order to determine the accuracy of correctly classified instances of the data set, the paper calculates the outcomes of numerous learning techniques for the machine. The research findings show that the data set contains some irrelevant and redundant features. The paper therefore examines various feature selection technologies available at WEKA in order to select the most promising features of the benchmark data set and attempts to increase the accuracy of the outcomes from slower runtime. In terms of execution time and accuracy, the results with and without the function selection technique are compared to the benefit of using the most promising features. There are several different methods of machine learning and feature selection [10].

The main contributions and organization of this paper are summarized as follows: In section 2 we describe background details of different machine learningtechniques schemes. The section 3 proposed work. The section 4 Results and discussion work. Finally in section 5 we concluded the paper.

II. BACKGROUND WORK

In this paper [1], the authors describes phenotypes and treatment of patients as an under-used source of data with a significantly higher scientific research capability than presently noticed and explained by the clinical data. Mining electronic health records (EHRs) have the ability to develop new principles as well for patient homogeneity and to learn about unknown correlations between diseases. The combination of EHR data with genetic data will also give a better appreciation of the personal relationships between genotype and phenotype. The systemic deposition of these data in EHRs and their mining is currently hindered by a broad range of ethical, law and technology reasons. The potential to advance medical re-search and clinical care by using EHR data and the challenges to be surmount. In this paper [2], researchers will explain how artificial intelligence was used for cost effective treatment in the medical field. For that purpose, they use a neighbor's algorithm k and use UCI machine learning datasets to check the accuracy of the algorithm. They had to generate patient input and diagnostic test data. They use real data from patients.

Add additional training sets to classify more chronic conditions with a minimum of no algorithm changes.

In this paper [3] the distributed computer environment on the basis of map reduction is used to process the large volume of a data. The classification is used to find the accuracy of a patient data. This paper focuses more on finding the closest precision of a classifier. For the data and accuracy of the classifier, the CART model and random forest are built. The closest precision to the prediction can be found rather than using the random forest algorithm. The analysis of the prediction helps doctors to identify the hospital admissions of the patient. Predictive model that uses a scalable random forest classification that can accurately yield the risk rate.

In this paper [4] introduces data mining and the big data in the healthcare sector. The algorithm for machine learning was used to study data on healthcare. The steady increase in healthcare data. Various countries spend a lot of resources, scientists address the problem of data storage and data organization, which helps the complexity of data and finds out the new results that this article is based on the use of Data Mining and Big Data in the health care sector.

In this paper [5] the application by EMC 'S from the ambulatory department and the algorithm is based on DNN AND DBDT, a high UAR to predict a future stroke prediction can be achieved. It offers a number of benefits, such as high precision, rapid prediction and consistency of results. A smaller amount of data is also required by the DNN algorithm. By using a lower amount of patient data than the GDBT algorithm, the DNN algorithm can produce optimal results.

In this paper [6] they use a Naive Bayes and a decision tree algorithm for prediction of cardiovascular disease. After reducing the size of the data sets, the PCA reduces the number of attributes; SVM can outperform a Naive Bayes and Decision tree. SVM can also be used for heart disease prediction. The main aim of this paper is to accurately predict the disease of diabetics. Use a data mining tool from WEKA. Data mining is very particularly useful strategies for classifying diseases used by the healthcare sector. The objective of this paper is to study the supervised algorithm for the prediction of cardiovascular diseases.

III. PROPOSED FRAMEWORK

In this paper, we implemented a music recognition system, called based System as shown in figure.1.The important phases of the proposed framework are described as follows:

Stage 1. Dataset Selection Stage

In this phase the benchmark data set was selected. Data were selected for 1,000 patients, along with family history, calcium blood, vitamin D, thyroid, kidney function and phosphorus, with nine attributes. These features were the most important features for osteoporosis prediction.

Stage 2. Pre-processing and Transformation Stage

This stage involved the preprocessing of the data set to remove noise and scale

Stage 3. Feature Selection Stage

In this phase, highly promising features were selected using the feature selection strategy implemented by WEKA. This contributed to a reduction in the number of features, which would in turn reduced training and prediction time. Correlationship-based sub-set evaluator and greedy step-by-step search techniques have been used. Four characteristics for the osteoporosis assessment were selected: family history, blood calcium, vitamin D and phosphorus.

Stage 4. Model Generation Stage

This phase involved model generation by selecting several training options, and testing for 10-fold cross validation, a training set and percentage division using the WEKA osteoporosis data set tool. There were two sets of experiments: with and without selection of the feature.

Stage 5. Performance Evaluation Stage

This stage of development consisted mainly of evaluating results for the classification of thyroid data set in corrected classified instances from a variety of classification techniques, namely J48, ZeroR, Naive - Bayes, and OneR.

Stage 6. Precision of Diseases

Unknown instances in osteoporosis and non - osteoporosis classes had been predicted by the trained model of machine techniques. The identified data set was divided into training and test data sets according to the choices available in WEKA in this set of experiments.

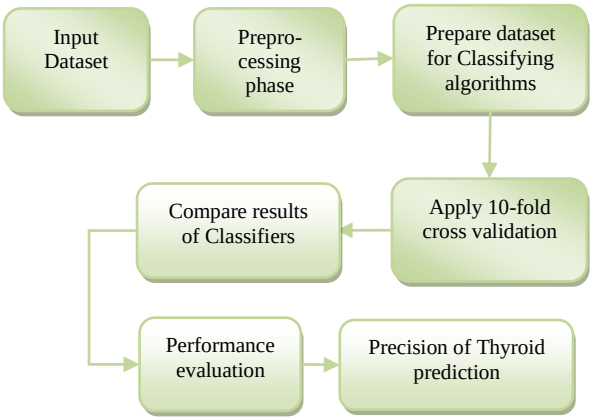

Fig. 1: Thyroid recognition system

Improved J48 decision tree:

J48 is an ID3 extension possessing of unique features of J48 include missing values, decision-making processes, continuous value ranges, rule derivation, etc. J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. The WEKA tool offers a number of tree pruning options. In the event of potential over mounting, tapping can be used as a precision tool. The classification is done recursively in other algorithms until each leaf is clean, that is, the classification of the data should be as perfect as possible. This algorithm produces the rules that give rise to a particular identity of the data. The aim is to gradually generalize the decision tree until flexibility and accuracy are balanced.

Basic Steps in the Algorithm: [7]

(i) In case the instances are of the same class, the tree is a leaf, so that the leaf is returned by the same class labeling.

(ii) For each attribute given by an attribute test, the potential information is calculated. The specific information gain is then calculated as a result of a test of the attribute.

(iii) The best attribute is then determined by the current selection criterion and the branching attribute.

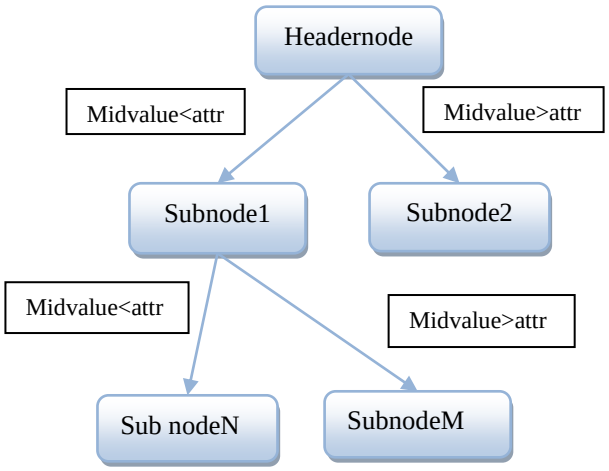

Fig 2: Structure of the modified J48 decision tree

The modified J48 decision tree classifier investigates the standard information gain resulting from selecting an attribute to split data. The attribute with the highest standardized information gain is being used to decide. So the machine learning reappears on smaller subsets. If all instances in a subset be-long to the same class, the splitting procedure stops. Then the decision tree creates a leaf node telling us to choose whatever class. The modified J48 decision tree classifier produces a decision node higher in the tree using the class's expected value. If the LSB value generated in CADL and attribute features are the same then the disease is marked as Class'0' therefore the disease advised for Class'1'. Figure.2 shows the modified J48 decision tree structure. Tree's first level is a single header node. It's just a kid's pointer node. The tree's second level has 2 sub-trees labeled 1-2.

IV. RESULTS AND DISCUSSIONS

The thyroid data set is taken for highly experimental assessment from the UCI learning machine repository, which consists of 300 samples with no missing values. The samples of the data set include 148 hypothyroid, 41 hyperthyroid and 111 records of negative patients. The dataset comprises of 28 attributes, along with 26 conditional attributes for thyroid discomfort symptoms and blood tests and two Class & Referral Source (CRS) decision attributes supporting all three types of thyroid disorder and the recommended health center for even more treatment. The six continuous attributes are preprocessed using the WEKA Tool discretionary Equal width binning technique. The six continuous attributes are preprocessed using the WEKA Tool discretionary Equal width binning technique. As discretization enables a quicker and more accurate learning process, the accuracy of the graders is enhanced for the discreted data set.

Table.1 Classifiers types

Classifiers	Correctly classified instances using 10-fold cross-validation (%)
ZeroR	59.88
J48	68.03
Naïve bayes	41.30
OneR	63.62

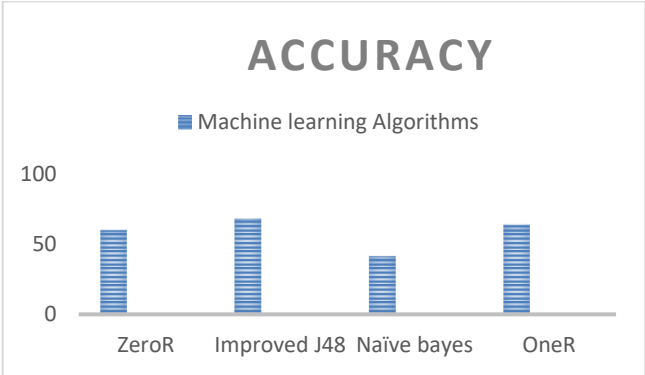

Fig. 3: Comparison of classifiers accuracy

Table.2 Classifiers types and its MAE

Classifiers	Mean Absolute error
ZeroR	0.0551
J48	0.0023
Naïve bayes	0.0598
OneR	0.0153

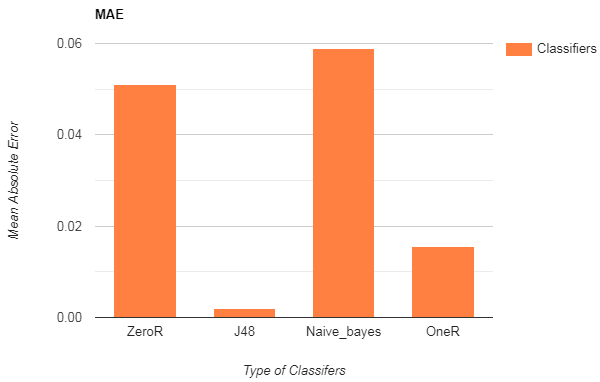

Fig.4: Comparison of classifiers formean of absolute error

Figure.4 shows that the mean absolute error for ZeroR, OneR and Naïve algorithms is higher when compared to J48.

Table.3 Classifiers types and its RMSE

Classifiers	Root mean squared error
ZeroR	0.1658
J48	0.0385
Naïve bayes	0.1551
OneR	0.1237

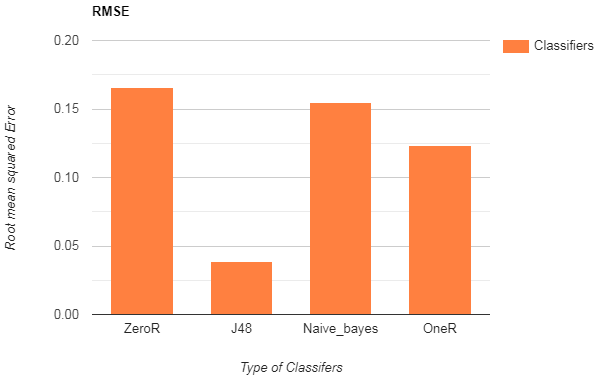

Fig.5: Comparison of classifiers forRoot mean squared error

Figure.5 shows that the Root meansquared error for ZeroR, OneR and Naïve algorithms the squared error is higher than J48.

Table.4 Classifiers types and its classification error

Classifiers	**Classification error (%)**
ZeroR	7.27
J48	0.41
Naïve bayes	5.63
OneR	3.82

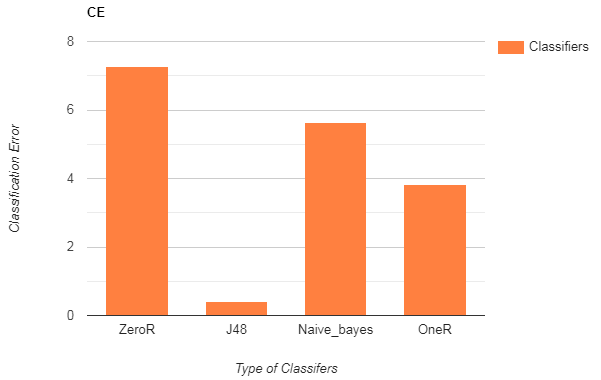

Fig.6: Comparison of classifiers forRoot mean squared error

Figure.6 indicates the hypothyroid attribute classification error (as shown by WEKA) for all algorithms. However these values were obtained after each cross - validation run averaged the documented values. It is clear that J48 has less error than ZeroR, OneR and Naïve Bayes in classification.

V. CONCLUSION

Thyroid is one of the most serious diseases of pregnancy and menopause for many women. Thyroid disorder is frequently misdiagnosed and left untreated for a long time needs urgent thyroid hormone injection. This paper examines various machine learning techniques for osteoporosis prediction. The paper examines and examines the effects of the combination of the method of feature selection with the classification technique. This paper employs classification techniques for the use of WEKA tools and their tests for benchmark data sets for osteoporosis, in combination with, and without the use of, feature selection methods, 10-cross validation, training sets and percentage split methods. The outcomes were directly compared for correctly classified instances, run-time, kappa statistics and mean absolute value for and without feature selection experiments. The results were compared. Overall improved J48 decision tree having more accuracy compared to all other schemes.

ACKNOWLEDGEMENT

Authors would like to express sincere gratitude to management and principal ofcollege for their support and encouragement to carry out the research work.

REFERENCES

1. P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: towards better research applications and clinical care.
2. HahabTayeb*, MatinPirouz*, Johann Sun1, Kaylee Hall1, Andrew Chang1, Jessica Li1, Connor Song1, Apoorva Chauhan2, Michael Ferra3, Theresa Sager3, Justin Zhan*, ShahramLatifi,Toward Predicting Med-ical Conditions Using k-Nearest Neighbours, 2017 IEEE International Conference on Big Data.
3. ReekanthRallapalli Faculty of computing Botho University Gaborone, Botswana predicting the Risk of Diabetes in Big Data Electronic Health Records by using Scalable Random Forest Classification Algorithm, 2016 IEEE.
4. OubidaAlaouiMdaghri, Mourad El Yadari, AbdelillahBenyoussef, Ab-dellah El Kenz Faculty of Science Rabat Morocco, Rabat, Study and analysis of Data Mining for Healthcare, 2016 IEEE.
5. Hen-Ying Hung, Wei-Chen Chen, Po-Tsun Lai, Ching-Heng Lin, and Chi-Chun Lee, Comparing Deep Neural Network and Other Machine Learning Algorithms for Stroke Prediction in a Large-scale Population-Based Electronic Medical Claims Database, 2017 IEEE.
6. Rof. DhomseKanchan B. Assistant Professor of IT department METS BKC IOE, Nasik Nasik, India kdhomse@gmail.com , Mr. MahaleKishor M. Technical Assistant of IT department METS BKC IOE, Nasik, India kishu2006.kishor@gmail.com, Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analy-sis,2016 IEEE.
7. Korting, Thales Sehn. "C4. 5 algorithm and Multivariate Decision Trees." Image Processing Division, National Institute for Space Research--INPE.
8.