European Journal of Business & Social Sciences

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

Thyroid Disease Prediction by Artificial Neural Network Technique

M.Shyamala
M.Phil Scholar,
Department of Computer Science,
National College, Tiruchirappalli – 620 002
Dr.P.S.S.Akilashri.,
Assistant Professor, PG & Research
Departments of Computer Science
National College, Tiruchirappalli – 620 002

ABSTRACT

Because of the huge information in biomedical and healthcare communities, correct study of medical data benefits early disease detection, community services and patient care. The exactness of study is reduced when the value of medical data is incomplete. Moreover, various regions exhibit unique appearances of particular regional diseases, those results in weakening the prediction of disease outbreaks. In the proposed system, it provides machine learning algorithms for effective prediction of thyroid disease occurrences in disease-frequent societies. It experiment the changed models over real-life hospital data collected. To overcome the difficulty of incomplete data, it uses a latent factor model to rebuild the missing data. It experiment on a thyroid diseases using structured and unstructured data from hospital it use Naïve Bayes and MLP algorithm. Compared to numerous approximation algorithms, the calculation exactness of our proposed MLP algorithm reaches 98.8% with a convergence speed which is faster than that of the Naïve Bayes algorithm on disease risk prediction on thyroid using Weka tool.

Keywords: Data mining, Machine Learning, Artificial Neural Network.

I. INTRODUCTION

A process in which the interesting patterns and the knowledge are extracted from the large dataset is called as data mining. Many techniques are used to discover this kind of knowledge, mostly extracted from the machine learning and statistics. The Highlighted part of these approaches is to find the accurate knowledge from the discover data. In the data mining the tasks performed is depends on what sort of knowledge someone needs to mine.

Recently, thyroid diseases spread more in today's world. In India, for example, one of eight women suffers from hypothyroidism, hyperthyroidism or thyroid. Various research studies estimate that about 30% of India is diagnosed with endemic goiter. Reasons that affect the thyroid function are: low-calorie diet, infection, trauma, toxins, stress, certain medication etc. It is to prevent such diseases rather than cure them, because the majority of treatments consist in long term medication or in chirurgical intervention. This paper will help to predict the thyroid diseases through the machine learning and MLP algorithm.

II LITERATURE SURVEY

In this paper [1] The phenotypes and treatment of patients represents an underused data source that has much greater research potential than is currently realized and explains by the Clinical data. Mining of electronic health records (EHRs) has the ability to establish a new patient-stratification principles and for knowing unknown disease correlations. Combining the EHR data with genetic data will also give a finer understanding of genotype-phenotype relationships. A broad range of ethical, legal and technical reasons currently

hinder the systematic deposition of these data in EHRs and their mining. Here, the potential for furthering medical research and clinical care using EHR data and the challenges that must be overcome.

In this paper [2] the researchers present how artificial intelligence applied to medical field for the efficient diagnosis. For that purpose they use a k nearest neighbor's algorithm and they check the accuracy of the algorithm with the help of UCI machine learning repository datasets. They had to generate patient's input and test data for diagnosis. They use a real patient data. They add additional training sets allow more medical conditions to be classified with the minimal no of changes to the algorithm.

In this paper [3] distributed computing environment processing the large volume of a data is done based on Map Reduce. To find the accuracy of a patient data the classification is used. In this paper more focused on find out the nearest accuracy of a classifiers. The CART model and random forest is built for the data and accuracy of the classifier is found. By using the random forest algorithm they can found the more nearest accuracy of the prediction. The prediction analysis helps to the doctors to identify the patient's admissions on to the hospital. Predictive model using scalable random forest classification which can accurately give the result rate of risk.

In this paper [4] the data mining and the big data in the healthcare sector is introduced. Machine learning algorithm has been used to study the healthcare data. The continuous increase of data in a healthcare. Several countries are spending a Lot of resources, scientist leads to fix the problem of storage and organization of data the data mining will help exploitation complexity of the data and find out the new result this paper is based on the use of data mining and big data in the healthcare sector.

In this paper[5] they applying a machine learning techniques by using EMC'S from outpatients department and the algorithm are based on a DNN AND DBDT, It can be achieve a high UAR for predicting the future stroke prediction. It provides a several advantages like high accuracy, fastest prediction, and consistency of results. DNN algorithm also requires a lesser amount of data. DNN algorithm can achieves optimal results by using a lesser amount of a patient data than compared to the GDBT algorithm.

In this paper [6] for heart disease prediction they use a Neavi Bayes and Decision tree algorithm. They used a PCA to reduce the no of attributes, after reducing the size of the datasets; SVM can outperform a Neavi Bayes and Decision tree. SVM can also be used for prediction of hearts disease. The main goal of this paper is to predict the diabetics' disease. Using a WEKA data mining tools. Data mining is very useful techniques used by health care sector for classification of disease. The aim of this paper is to study supervised machine learning algorithm to predict the heart disease.

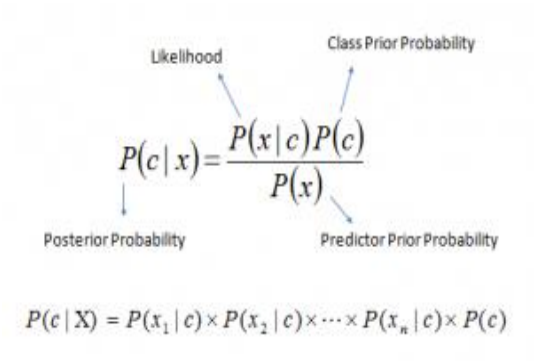
III METHODOLOGY

3.1 NAÏVE BAYES ALGORITHM

The method that is easy to built the classifier is the Naive Bayes. In this model that will assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There are more classifier algorithm but a family of algorithms based on a general rule: all naive Bayes classifiers believe that the value of a exacting characteristic is independent of the assessment of other feature, given the class variable. It is a categorization method based on Bayes' Theorem with a supposition of self-rule among predictors. Finally, a Naive Bayes classifier assumes that the attendance of a exacting characteristic in a class is unrelated to the occurrence of any other feature.

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

Naive Bayes model brings flexibility to build and mainly useful for huge data sets. Along with ease, Naive Bayes is recognized to break even extremely complicated categorization methods. Bayes theorem provides a way of calculating posterior chance P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



Above,
$P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
$P(c)$ is the prior probability of *class*.
$P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
$P(x)$ is the prior probability of *predictor*.

3.2 MLP ALGORITHM

A Multilayer Perceptron (MLP) is a partition of feedforward ANN. A MLP contains three layers of nodes: an input layer, a hidden layer and an output layer. Apart from the input nodes, each node is a neuron that uses a nonlinear launch utility. MLP occupy a supervised learning technique called back propagation for training. Its several layers and non-linear establishment differentiate MLP from a linear perceptron. It can discriminate data that is not linearly divisible.

If a multilayer perceptron has a linear creation occupation in all neurons, that is, a linear function that maps the subjective inputs to the output of each neuron, then linear algebra shows that any number of layers can be reduced to a two-layer input-output model. In MLPs a few neurons use a nonlinear foundation occupation that was developed to model the occurrence of action potentials, or firing, of biological neurons. The two ordinary foundation functions are both sigmoids, and are described by

$$y(v_i) = \tanh(v_i) \text{ and } y(v_i) = (1 + e^{-v_i})^{-1}$$

The first is a hyperbolic tangent that ranges from -1 to 1, while the other is the logistic function, which is similar in shape but ranges from 0 to 1. Here y_i is the output of the *i*th node (neuron) and v_i is the weighted sum of the input connections. Alternative activation functions have been proposed, including the rectifier and soft plus functions. More specialized activation functions include radial basis functions.

3.3 WEKA TOOL

Weka is a familiar sight system in the history of the machine learning research communities and data mining, because it is the only toolkit that has gained such extensive acceptance and survived for an extensive period of time. It provides lots of diverse algorithms for data mining and machine learning. Weka is open source and freely available. It is also platform-independent.

Advantages
As weka is fully implemented in java programming languages, it is platform independent & portable.
It is freely available under GNU General Public License.
Weka s/w contain very graphical user interface, so the system is very easy to access.
There is very large collection of different data mining algorithms.
Disadvantages
Lack of possibilities to interface with other software
Performance is often sacrificed in favour of portability, design transparency, etc.
Memory limitation, because the data has to be loaded into main memory completely

IV RESULTS AND DISCUSSION

The experiment uses thyroid disease Analysis dataset obtained from UCI machine learning repository. The dataset is taken from the UC Irvine Machine Learning Repository. The Hypothyroid dataset is used in this research and
Development department for experimental purposes. The dataset contains 3090 instances. In this 149 data comes under thyroid and 2941 data is negative cases.
The attributes are shown in the table below.

Table: thyroid

Attribute Name	*Value type*
age	*continuous,?.*
sex	*M,F,?.*
on_thyroxine	*f,t.*
query_on_thyroxine	*f,t.*
on_antithyroid_medication	*f,t.*
thyroid_surgery	*f,t.*
query_hypothyroid	*f,t.*
query_hyperthyroid	*f,t.*
pregnant	*f,t.*
sick	*f,t.*
tumor	*f,t.*
lithium	*f,t.*
goitre	*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,?.*

A set of dataset are obtained by applying Navie Bayes & MLP Algorithm. The data giving different support and confidence values that are analyzed and obtained different number of rules. During analysis it found that MLP is much faster for large number of

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

transactions as compare to Naïve Bayes. The time taken for generating disease prediction is very low. For this paper, thyroid disease has been taken. Analysis carried out on 3090 transactions. The obtained results are collected from Pentium Dual core processor with 1. 73GHz speed and 1 - GB RAM.

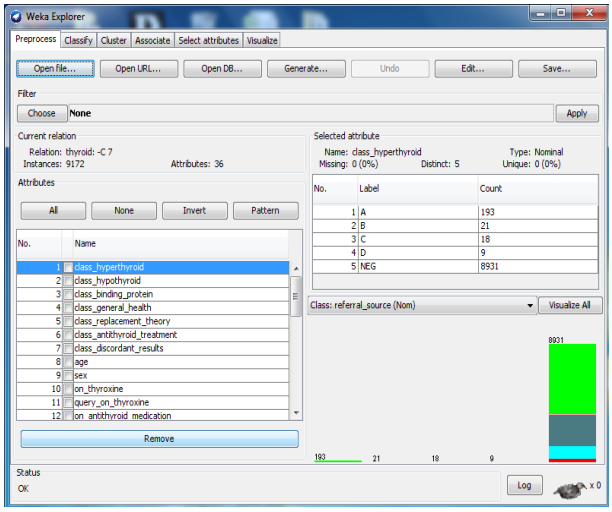


Fig: 4.1.1 Upload Dataset

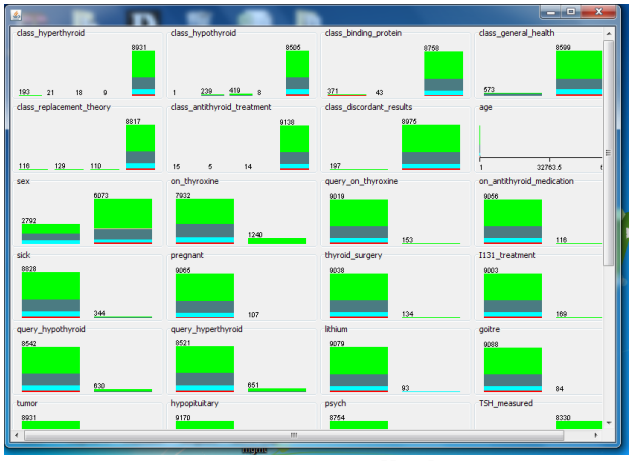


Fig : 4.1.2 Overall Chart

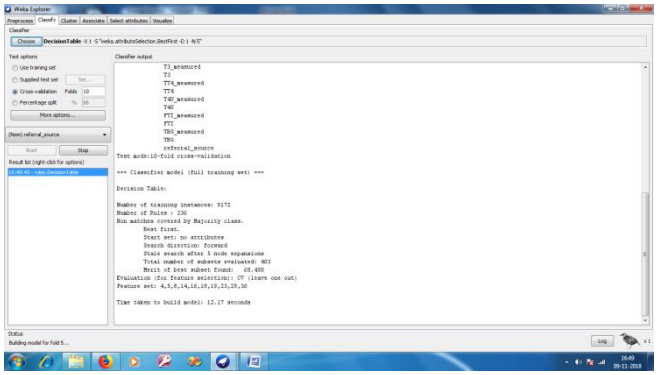


Fig: 4.1.3 Naïve Bayes Execution

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

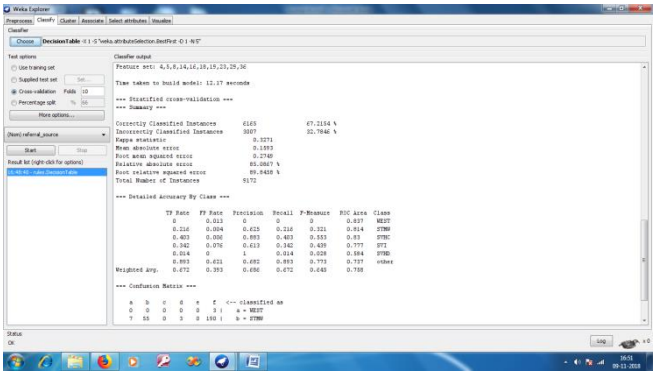


Fig : 4.1.4 MLP Execution

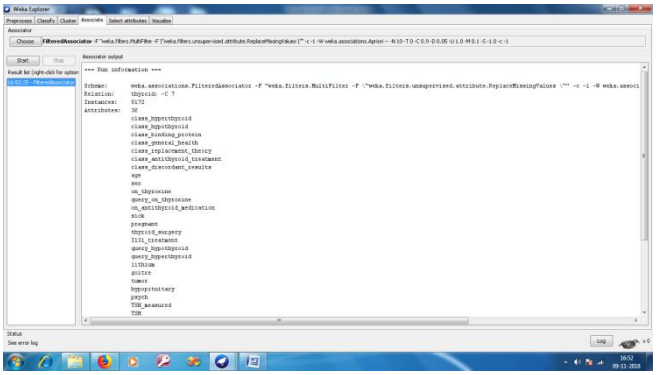


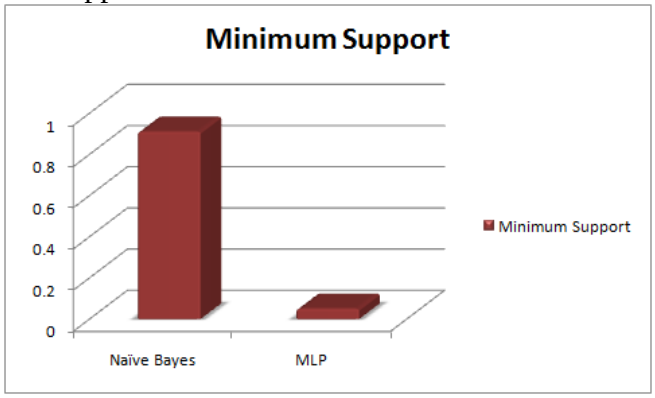
Fig: 4.5 Thyroid Dataset

4.2 MINIMUM SUPPPORT

TABLE 4.2.1 Showing Confidence for thyroid disease Dataset

Thyroid Disease Dataset	
Technique	Minimum Support
Naïve Bayes	0.91
MLP	0.05

CHART 4.2.1 Minimum Support in Dataset



ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

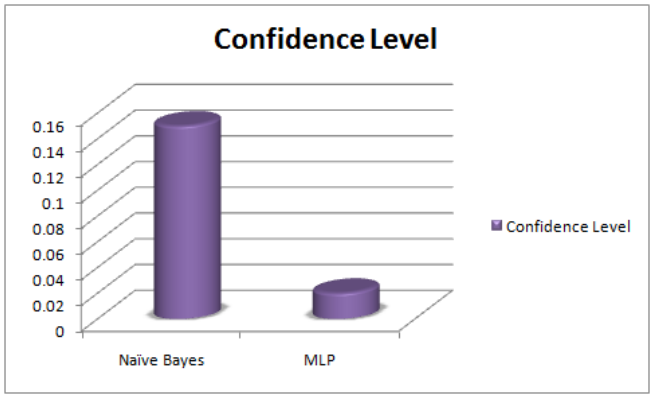
4.2.2 CONFIDENCE

Experiments are performed on the thyroid disease dataset. Machine with configuration of windows Vista 7 system and 2-GB of RAM is used. The results were compared to experiments with Weka Tool implementations of Naïve Bayes and MLP techniques run to ensure that the results are comparable for confidence level.

TABLE 4.2.2 Confidence in All Dataset

Thyroid Disease Dataset	
Technique	Confidence Level
Naïve Bayes	0.15
MLP	0.02

CHART 4.2.2 Confidence in All Dataset

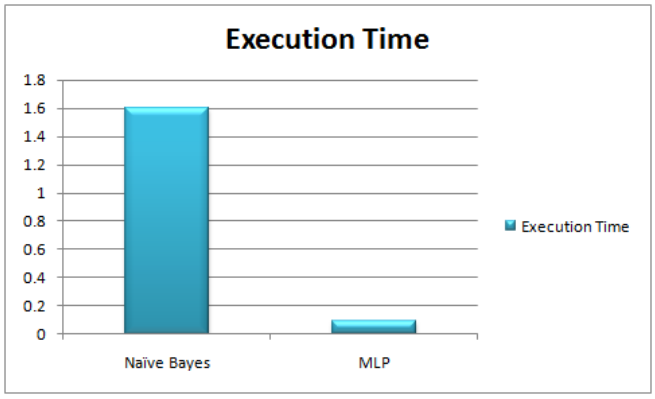


4.2.3 EXECUTION TIME SECOND

Table 4.2.3 Execution Time Second

Thyroid Disease Dataset	
Technique	Execution Time
Naïve Bayes	2.60
MLP	0.09

CHART 4.2.3 Execution Time in All Dataset



V. CONCLUSION

The thyroid is the chief and biggest gland in the endocrine system. Data mining method is used on the hypothyroid dataset for influential the positive and the negative cases from the entire dataset. The cataloging of dataset is used to give improved treatment, choice making, identify disease. This thesis is a challenging Machine learning data mining problem for finding the thyroid prediction method. The explained method is very simple and efficient one. This is successfully tested for large data sets. The results given in this paper are accurate and appropriate. However, a more wide-ranging empirical valuation of the proposed method will be the objective for future research. In this thesis, it is analyzed that the MLP algorithm takes more time to compute association rules then Naïve Bayes algorithm by using the same number of data set. MLP is much faster than Naïve Bayes because it uses compact data structure, and it eliminates the repeated transaction scan.

REFERENCES

[1] P. B. Jensen, L. J. Jensen, and S. Brunak,"Mining electronic health records: towards better research applications and clinical care.

[2] hahab Tayeb*, Matin Pirouz*, Johann Sun1, Kaylee Hall1, Andrew Chang1, Jessica Li1, Connor Song1, Apoorva Chauhan2, Michael Ferra3, Theresa Sager3, Justin Zhan*, Shahram Latifi, Toward Predicting Med-ical Conditions Using k-Nearest Neighbours, 2017 IEEE International Conference on Big Data.

[3] reekanth Rallapalli Faculty of computing Botho University Gaborone, Botswana redicting the Risk of Diabetes in Big Data Electronic Health Records by using Scalable Random Forest Classification Algorithm, 2016 IEEE.

[4] oubida Alaoui Mdaghri, Mourad El Yadari, Abdelillah Benyoussef, 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.

ISSN: 2235-767X
Volume 07 Issue 07
June 2019
Special Issue

[6] Prof. Dhomse Kanchan B. Assistant Professor of IT department METS BKC IOE, Nasik Nasik, India kdhomse@gmail.com , Mr. Mahale Kishor
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] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining," Boston, U.S.A.: Pearson Education Inc., 2006, ch. 4, pp. 151-154.

[8] R. Polikar, "Ensemble Based Systems in Decision Making," IEEE Circuits and Systems Magazine, vol. 6, pp. 21-45, Third Quarter, 2006.

[9] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 20, no.1, pp 832-844, August 1998

[10] Hossam M. Zawbaa, Maryam Hazman, Mona Abbass, Aboul Ella Hassanien, "Automatic fruit classification using random forest algorithm", 14th International Conference on Hybrid Intelligent Systems,pp. 164 – 168, 2014.

[11] Jiawei Hanl, Yanheng Liu, Xin Sun, "A Scalable Random Forest Algorithm Based on MapReduce", IEEE 4th International Conference on Software Engineering and Service Science, pp. 849 – 852, 2013.

[12] B.V.S Dheeraj Reddy, Mounika Booreddy, "Classification And Clustering Medical Datasets By Using Artificial Neural Network Models", Publications Of Problems & Application In Engineering Research – Paper, Vol 04, Special Issue 01, 2013.