DISEASE RISK PREDICTION USING WEIGHTED ENSEMBLE NEURAL NETWORK IN HEALTHCARE FOR DATA ANALYTICS

Gakwaya Nkundimana Joel[1], *Dr. S.Manju Priya*[2]

ABSTRACT

As the big data are growing in biomedical and healthcare communities, so are precise analyses of medical data aids, premature disease identification, patient care as well as community services. On the other hand, the accuracy of disease-analysis will be affected, if the medical data quality is imperfect. As a result, the choice of features from the dataset turns out to be an extremely significant task. Feature selection has exposed its efficiency in numerous applications by means of constructing modest and more comprehensive models, enlightening learning performance and preparing clean and clear data. Random forest tree is one of methods used to select the feature in big data. This technique has certain drawbacks compared to other relative algorithms. This Random Forest tree has been used in disease risk prediction toward malaria and it shows the lack of real-time prediction. In other word., it gives good results, but due to congestion with small trees, some sample are not showing the same improved accuracy. To overcome this unreal - time prediction and lack of accuracy, the Weighted Ensemble based Neural Network for Multimodal Risk Disease Prediction has been introduced and shown to be more effective on large dataset without having any lack of accuracy. This research attempts to find ways to predict and have fast and accurate result. The environment used to classify further big data-storage further is matlab and Hadoop ecosystem.

Keywords: Big Data, Hadoop, Random forest, Neural Network

I. INTRODUCTION

Every living creature needs to be in good health and free from illness or injury. Scientists, dealing with computers, data and those from other fields try everything they can to help people with prevention, diagnosis, and treatment of diseases they are prone to. The role of computer scientists is to provide an easy way of visualizing, analyzing and predicting the data using different techniques which give different results. Some of these techniques are classification and clustering of the data according to their categories. To achieve the task of analysis and prediction with high accuracy, it requires high techniques which can produce the best results with minimum resources. Not long ago, different technologies have been created and they are still being created, for the same purpose. Data that form bigdata, are so complex that the traditional data processing soft wares are not sufficient and capable of dealing with them. Healthcare is one of the big data services, where real-time health monitors have gathered, shared and utilized data for personalized healthcare [1-2]. Use of mobile devices provide the data availability and advancement in analytical

[1]Research Scholar, Dept of Computer Science, Karpagam Academy of Higher Education

[2]Associate Professor, Dept. of CS, CA & IT, Karpagam Academy of Higher Education

methods and technologies, leading to the use of mHealth [3-5]. Electronic Health Records, and Electronic Medical Records capture patient's data and stored them in clinic databases and on local level databases. Furthermore, some systems use internet of things to automate data generation and storage. The contribution of this research paper is to provide a new type of understanding how random forest tree algorithm works and how it is incapable of giving real-time prediction accuracy. It also talks about the use of weighted ensemble-based neural network for multimodal risk disease prediction as solution. Random forest tree is good to deal with large problem including nonlinear. The paper has been arranged as follows: section 2 represents related works. Section 3 discusses important of Random Forest Tree algorithm and its inadequacy. Section 4. demonstrates the advantages of weighted ensemble neural network-based multimodal disease risk prediction. Comparative study and result is described in section 5 and, finally section 6 gives the conclusion and future work of the paper.

2. Related works

Many studies have been conducted to analyze healthcare data in order to provide the right medication, based on better prediction.Data generation sources in healthcare domain are increased,and it requires advanced big data tools and techniques to process such huge volume of data. It is observed thatvarious improvements have been made in the day-to-day clinical system. This advancement is used to develop knowledgefrom huge clinical records and improve business insights. Recently, many research works have been done to reduce the overallcost and improve the disease diagnosis in healthcare [6-9]. Moreover, the impact of big data and cloud computing hasbeen noticeable [10-14]. In addition, there is a need to provide security and privacy in healthcare big data analytics.Bates et al. have developed six use cases to reduce the overall cost of healthcare [15]. These use cases are applied to thefollowing domains to reduce the cost for patients, health record management, triage, readmissions, disease diagnosis, drugrecommendation and healing optimization. Hermon and Williams have identi?ed four uses cases of big data in healthcareit includes patient administration and healthcare delivery, medical decision support system, medical support services andcustomer behavior [16]. Moreover, numerous big data analytical solutions are developed to reduce the cost of treatment path, drug recommendation and healthcare delivery. In [17-18], Gakwaya Nkundimana Joel and S.Manju Priya reviewed big data in healthcare where they have demonstrated the big data to be analaysed, and provided the relavent features during data analyasis and prediction.

3. The Importance of Random Forest Algorithm and its inadequacy

Random forest Tree works as a large collection of de-correlated decision trees. Forest means that we use a lot of decision trees. It uses huge trees to create forest for classification. It is based on bagging e.g., an average noisy and unbiased model in order to create a model with lower variance. Training samples are used to classify the different subsamples.

$$S = \begin{bmatrix} f_{A1} & f_{B1} & f_{C1} & C_1 \\ \vdots & & \vdots & \\ f_{AN} & f_{BN} & f_{CN} & C_N \end{bmatrix}$$

S is a training sample to create classification, where

fA1 is first feature and fBN is feature of nth sample. As mentioned above, random forest is made of combined decision trees. Trees are grouped into subsets. Various subsets form decision trees where S1 form SM and other subset like S2 create another SM1, it keeps growing until a real forest is created.

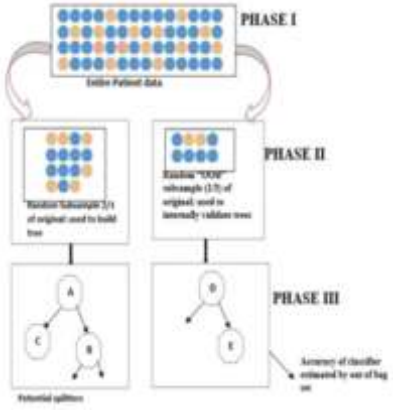
The major work, is done under class prediction, where each class is passing voting stage under subclass to check a high voted class. in the figure below class 1 shows that it is highly voted to be the most predicted.

$$S_1 = \begin{bmatrix} f_{A12} & f_{B12} & f_{C12} & C_{12} \\ f_{A15} & f_{B15} & f_{C15} & C_{15} \\ \vdots & & \vdots & \\ f_{A35} & f_{B35} & f_{C35} & C_{35} \end{bmatrix} \quad S_2 = \begin{bmatrix} f_{A2} & f_{B2} & f_{C2} & C_2 \\ f_{A6} & f_{B6} & f_{C6} & C_6 \\ \vdots & & \vdots & \\ f_{A20} & f_{B20} & f_{C20} & C_{20} \end{bmatrix}$$

$$S_M = \begin{bmatrix} f_{A4} & f_{B4} & f_{C4} & C_4 \\ f_{A9} & f_{B9} & f_{C9} & C_9 \\ \vdots & & \vdots & \\ f_{A12} & f_{B12} & f_{C12} & C_{12} \end{bmatrix}$$

Figure 1: Class Prediction

Phase 1 deals with loading whole dataset into matlab, phase 2 describes the splitting of data into classes, finally phase 3 deals with accuracy of classifier estimated



Random forest algorithm is weak to provide the exact prediction in real-time. As above class prediction shows, it repeats each phases until it gives the same accuracy, due to the congestion with huge trees. Random forest has some limitations. For example, a large number of trees can cause the algorithm to slow down and gives some uncertainty results during online prediction. Whereas on the other hand, they are quick to train, but slow to produce prediction once they are trained.

4. Weighted Ensemble Neural Network

Ensemble-centered diverse collection of individual model into single unit to help the predictive accuracy. Combining ensemble with Neural Network helps to deal with large number of decisions which are pending. This method, reduces the time to predict because it works on top of Hadoop framework. The below algorithm shows how Hadoop is utilized to solve big data problem with help of MapReduce.

Weighted Ensemble Based Neural Network Algorithm (WEBNN):
Input:
number of classes M, number of boosting, iterations N, dataset D
Algorithm:
Init: set f(x) = 0
for t = 1 to N do
compute weight w
train a network to optimize g
calculate optimal coefficient c
update the f(x)=f(x)+(c*g)
end for
Output: Predictor f(x)
MapReduce - Weighted Ensemble Based Neural Network Algorithm
Begin
MAPPER
Create training datasets D based on balanced bootstrapping from the original training dataset
Apply Weighted Ensemble Based Neural Network Algorithm
End for
REDUCER
for each training data record
Sum the results
endfor
end

5. Comparative study and Results

The purpose of this section is to examine the performance of two algorithms namely random forest and weighted ensemble neural network. Both algorithm deals with large dataset, and features are selected with help of improved ant colony features selection methods. The below tables summarize the work of each algorithm:

Algorithm/ Parameters	RFT	WENN
Precision	94.5	97.05
Accuracy	6.4	7.30
Error-rate	83.2	92.70
F1-measure	91.7	95.8

The performance assessment which is being measured in term of precision, accuracy and error-rate, is calculated in the formula given below:

Accuracy = (TP+TN)/(TP+FP+TN+FN)

Precision= (TP)/(TP+FP)

F1-measure= (2*Precision* Recall)/ (Precision+ Recall)

here the F1-Measure is known as the weighted harmonic mean of the precision

True Positive, True Negative, False Positive, False Negative are represented as TP, TN, FP, and FN respectively

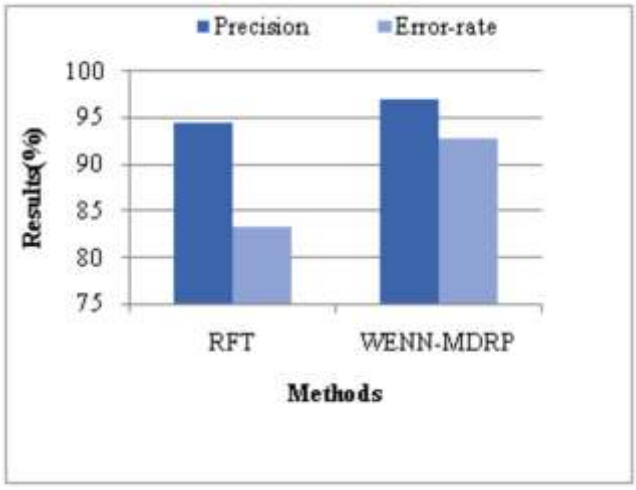


Figure-2- measure of precision and error-rate

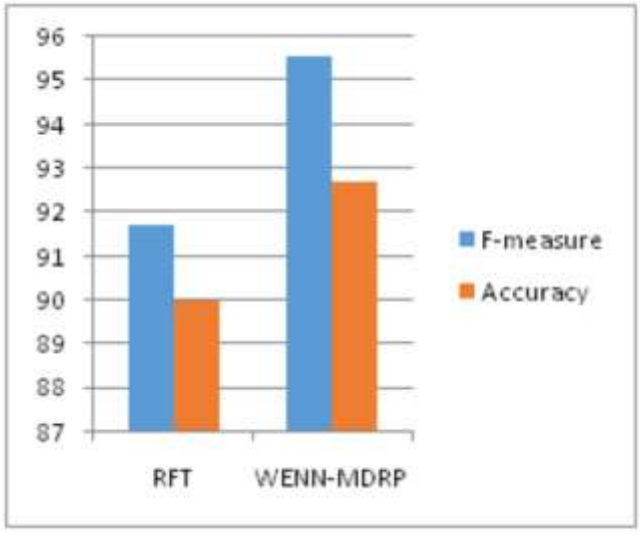


Figure-3. Measure of accuracy

Table 1. Summary of random forest and weighted ensemble neural network algorithm

Objective	Random Forest	Weighted Ensemble Neural Network
Main objective	Processing/analyzing large dataset	Processing/analyzing most all ingested dataset
Size of data	Large data	Large data
Latency	High (minute/ hour)	Low(millisecond /few minutes)

6. Conclusion

This research paper describes a novel approach for ensemble neural network for a high performance in the predictive task. It also describes the pitfalls of random forest tree algorithm, when the data are congested. Both algorithms are good towards a large number of data. When it comes to real-time process and analysis, weighted ensemble neural network shows high performance. Both algorithms are used along with ant colony feature selection algorithm in order to enhance the selection of useful parameter during prediction and analysis. Based on the accuracy and precision of the result, it has been proved that weighted ensemble neural network is useful and reliable. Future work is to implement these algorithms in different fields to improved their performance.

References :

[1] P.K. Drain, et al., Diagnostic point-of-care tests in resource-limited settings, Lancet Infect. Dis. 14 (3) (2014) 239-249.

[2] R. Peeling, Bringing diagnostics to developing countries: an interview with Rosanna Peeling, Expert Rev. Mol. Diagn. 15 (9) (2015) 1107-1110.

[3] M. Urdea, et al., Requirements for high impact diagnostics in the developing world, Nature (2006) 73-79.

[4] C.C.Y. Poon, Yuan-Ting Zhang, Shu-Di Bao, A novel biometrics method to secure wireless body area sensor networks for telemedicine and m-health,IEEE Commun. Mag. 44 (4) (2006) 73-81.

[5] M. Fiordelli, N. Diviani, P. Schulz, Mapping mHealth research: a decade of evolution, J. Med. Int. (2013).

[6] Adler-Milstein J. America's health IT transformation: Translating the promise of electronic health records into better care. Mar 17, 2015. [June 5, 2015]. (Paper presented at U.S. Senate Committee on Health, Education, Labor and Pensions)

[7] Adler-Milstein J, DesRoches CM, Furukawa MF, Worzala C, Charles D, Kralovec P, Stalley S, Jha AK. More than half of U.S. hospitals have at least a basic EHR, but stage 2 criteria remain challenging for most. Health Affairs (Millwood). 2014;33(9):1664-1671.

[8] AHIMA (American Health Information Management Association). Appropriate use of the copy and paste functionality in electronic health records. 2014. [March 27, 2015].

[9] AHRQ. Hospital survey on patient safety culture: 2014 user comparative database report: Chapter 5. 2014a. [February 25, 2014]. (Overall results). www.ahrq.gov/professionals/quality-patient-safety/patientsafetyculture/hospital/2014/hosp14ch5.html.

[10] F. A. Alaba, M. Othman, I. A. T. Hashem,and F. Alotaibi. Internet of things security:A survey. Journal of Network and ComputerApplications, 88:10 {28, 2017. ISSN 10848045.doi: http://doi.org/10.1016/j.jnca.2017.04.002.URL http://www.sciencedirect.com/science/article/pii/S1084804517301455.

[11] H. Alemdar and C. Ersoy. Wireless sensor networksfor healthcare: A survey. Computer Networks, 54(15):2688 {2710, 2010. ISSN 13891286.doi: http://doi.org/10.1016/j.comnet.2010.05.003.UR

http://www.sciencedirect.com/science/article/pii /S1389128610001398.

[12] S. Allen. New prostheses and orthoses step up theirgame: Motorized knees, robotic hands, and exosuitsmark advances in rehabilitation technology. IEEE Pulse, 7(3):6{11, May 2016. ISSN 2154-2287. doi:10.1109 / MPUL.2016.2539759.

[13] L. Alpay, P. Toussaint, and B. Zwetsloot-Schonk.Supporting healthcare communication enabled by informationand communication technology: Can hci and related cognitive aspects help? In Proceedingsof the Conference on Dutch Directions in HCI, Dutch HCI '04, pages 12{, New York, NY,USA, 2004. ACM. ISBN 1-58113-944-6. doi: 10.1145/1005220.1005236. URL http://doi.acm.org/10.1145/1005220.1005236.

[15] J. Sun, C.K. Reddy, Big data analytics for healthcare, in: Proceedings of the19th ACM SIGKDD International Conference on Knowledge Discovery andData Mining, ACM, 2013 1525-1525.

[16] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, S. Iyengar, Computational health informaticsin the big data age: a survey, ACM Comput. Surv. (CSUR) 49 (2016)

[17] Gakwaya Nkundimana Joel, S. Manju Priya, "Big Data Analytics in Healthcare and Delve Bioinformatics Data Space for Health Amelioration," International Journal of Computer Applications 180(9):43-45, January 2018.

[18] Gakwaya Nkundimana Joel, S. Manju Priya, "Improved Ant Colony Based Feature Selection And Weighted Ensembled Base Neural Network Based Multimodal Disease Risk Prediction Classifier For Disease Prediction Over Big Data " International Journa of Engineering and Technology", vol.7 no,3.27(2018)