Heart Disease Prediction using Data Mining with Mapreduce Algorithm

T.Nagamani, S.Logeswari, B.Gomathy

Abstract- The World Health Organization (WHO) estimated that cardiovascular diseases (CVD) are the major cause of mortality globally, as well as in India. They are caused by disorders of the heart and blood vessels, and includes coronary heart disease (heart attacks), Data mining acts as a major role in the construction of an intellectual prediction model for healthcare systems to detect Heart Disease (HD) using patient data sets, which support doctors in diminishing mortality rate due to heart disease. Several researches have been carried out for building model using individually or by combining the Data Mining with computational techniques involving Decision tree (DT), Naïve bayes (NB) along with Meta-heuristics approach, Trained Neural Network (NN), Machine intelligence or AI and unsupervised learningalgorithms like KNN and Support vector machine (SVM). In the proposed system, large set of medical instances are taken as input. From this medical dataset, it is aimed to extract the needed information from the record of heart patients using Mapreduce technique. The performance of the proposed Mapreduce Algorithm's implementation in parallel and distributed systems was evaluated by using Cleveland dataset and compared with that of the predictable ANN method. The trial results verify that the projected method could achieve an average prediction accuracy of 98%, which is greater than the conventional recurrent fuzzy neural network. In addition, this Mapreduce technique also had better performance than previous methods that reported prediction accuracies in the range of 95–98%. These findings suggest that the Mapreduce technique could be used to accurately predict HD risks in the clinic.

Keywords:Data Mining, cardiovascular diseases (CVD), accuracy, prediction, heart disease (HD), recurrent fuzzy neural network(RFNN), Mapreduce, world health organization (WHO)

I.INTRODUCTION

Heart diseases are also known as cardiovascular diseases which occur due to unhealthy lifestyle, smoking, alcohol and high intake of fats which may cause hypertension, high blood pressure, diabetics and strokes. The World Health Organization (WHO) [1] analysed that thirteen millions of death worldwide due to the reason of Heart diseasesin2017. based model of diagnosis. It has been proven in studies, that A good life style as well as an early detection is one of the alternatives for the prevention of heart disease. One alternative to early detection can be done with a computer the use of the computer is able to provide health service improvements.

Revised Manuscript Received on 10 January 2019.

T.Nagamani, Department of Computer Science and Engineering,Bannari Amman Institute of Technology, Sathyamangalam, Tamil nadu, India

Dr. S.Logeswari, Department of Computer Science and Engineering,Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India

Dr. B.Gomathy, Department of Computer Science and Engineering,Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India

The electronic health record accumulates large amount of health information which are necessary to be mined for finding unknown information for effective decision making. Due to the daunting disease such as heart disease the mortality rate is increased every year. As the data is very huge, researchers feel tough to extract data. The data mining techniques are applied to sort this problem. With the help of patient's Electro cardiogram (EKG or ECG), Echocardiography (ECHO) test reports and doctor's practice, Diagnosis is being done. Medical diagnosis is yet challenging and complicated task that needs to be done efficiently and accurately in addition with patient's Electrocardiogram (EKG or ECG), Echocardiography (ECHO) test reports. A suitable computer based information and decision support should be supported for decreasing the cost while doing the process of clinical test reports.

Data mining is a technique for extracting knowledge which is used to identify patterns and potential useful information from the large volumes of data by using different algorithms. The large volume of data can be handled by big data analysis. Mapreduce Algorithm basically uses parallel programming to process large dataset. It reduces issues from distributed and parallel programming, such as network performance, fault tolerance, load balancing. Mapreduce is programmed using Java for high reliability and scalability.[4]

II. LITERATURE REVIEW

Most of the research works have been implemented with several knowledge discovery techniques for predicting coronary disease diagnosis along with the fuzzy logic, neuro fuzzy, deep learning algorithms and artificial neural network. Syed Umar et al. [5] used Backpropogation algorithm for learning and testing neural network. The neural network weights were weighted with the help of optimization technique called genetic algorithm. So this multi layered network had twelve input nodes, two output nodes and ten hidden nodes. The heart disease risk factor is based upon the number of input layer. Weight and Bias is updated and recorded with the help of network training function.Matlab.

R2012a, Global Optimization Toolbox and the Neural Network Toolbox were used for implementation.100 patient's risk components were composed and the accuracy results for training set was obtained with 96.2 and accuracy results for testing set with 89% [5].

Hlaudi Daniel Masethe et al. (2014) used knowledge extraction techniques such as ID3, Naïve bayes, REPTree Simple Cart and Bayes Net were utilized for determining the occurrences of "heart attacks".

The data set was collected from hospital and doctors who were practitioners in South Africa. Eleven attributes were considered from the data set.
That were patient Id, cholesterol, tobacco consumption, Gender, cardiogram, Blood Pressure, rating of heartbeat, fasting sugar, age, chest pain, Thalac. The performance execution had been done with the tool called WEKA in the analysis of heart disease occurrences. WEKA tool was used in determining, analysing and identifying the patterns. The accuracy results obtained were J48 with 99.0741, REPTREE with 99.222, Naïve Bayes with 98.148, Bayes Net with 98.57 and simple CART algorithm with 99.0741. Among these, Bayes Net algorithm produced best results when compare with the Naïve Bayes algorithm [6].

Purusothama et al. (2015) applied various classification algorithms for disease prediction model in diagnosing the HD. The two types of models were used and compared i.e primary model is single model and secondary model is the combined model which is called as hybrid model and the both models are used to train the data. Data analysis was done using these two models. For both the single model and combined model, authors have considered the classification techniques only. The following are the results attained by comparing the algorithms such as decision method, association rule, KNN, ANN, N Bayes, hybrid approach with the accuracy of 76%, 58%, 86%, 69% and 96% respectively. The author recommended that hybrid data mining algorithms performs well and promising accuracy results were attained in heart disease diagnosis [7].

Kathleen Miao et al. (2016) analyzed learning classification and prediction models and applied to four different data sets that are Cleveland Clinic Foundation-CCF, Hungarian Institute of Cardiology-HIC, Long Beach Medical Centre-LBMC, and Switzerland University Hospital-SUH for coronary coronary disease diagnosis. This model achieved the accuracies of 80.14% using CCF, obtained accuracy 89.12% using HIC, 77.78% obtained using LBMC, and 96.72% results produced using SUH. It was concluded that the results obtained here was exceeding the accuracies of previously published research [8].

RovinaDbritto et al. (2016) developed an effective intelligent medical decision support system based on data mining techniques. The authors considered to emphasize on finding the appropriate classifier that has the potential to give better accuracy by applying data mining algorithms viz. Naïve Bayes, Support Vector machine and Logistic Regression and the accuracy with 75%, 80% and 79% respectively [9].

KaanUyar et al (2017) proposed natural selection algorithm based trained neural networks for detecting the occurrence of heart diseases. The authors used totally 297 instances of patient data. Among these 252 were used for training and 45 of them were chosen for testing. The authors compared this RFNN approach with ANN-Fuzzy-AHP approach. By analysing the testing set 97.78% accuracy was attained as the outcome in the above said algorithm [10].

Uyar [9] proposed that meta-heuristic approach with trained fuzzy neural networks approach is used for training the data set. Aliev et al suggested that the prediction of coronary disease can be done with Mapreduce algorithm. The rest of this paper is organized as follows: methodologies used, outcome of this study and conclusion along with the future work.

III. METHODOLOGY

A. Dataset

In this study, University of California Irvine (UCI) machine learning repository data set was used. UCI heart disease dataset [10] consists of four separate databases collected from four various medical hospitals. Totally, there were 303 patient reports in the Cleveland dataset out of which there were 6 omitted values.

S. no.	Name of the Attributes	Description
1.	Age	Age(years)
2.	Sex	Man=1, women=0
3.	Cp	Chest pain type
4.	Rbp	Resting Blood pressure upon hospital admission
5.	Chol	Serum Cholesterol in mg/dl
6.	Fbs	blood sugar during fasting >120 mg/dl true=1 and false=0
7.	Resting ECG	Resting electrocardiographic Results
8.	Thalach	Maximum Heart Rate
9.	Induced Angina	Does the patient experience angina as a result of exercise (value 1: yes, value 0: no)
10.	Old Peak	ST elevation during rest
11.	Slope	Heart rate slope
12.	Thal	Value 3: Normal ,value 6:fixed defect, value 7: reversible defect
13.	CA	Count of major vessels (value 0-3)
14.	Num	Heart disease Diagnosis (0 = healthy; 1 = low; 2 =medium 3 = high; 4= very high)

Table 1 UCI dataset attributes

In this experiment, six missing values of instances in the Cleveland dataset were cleaned and transformed by data preprocessing. So only 297 instances were taken for this study. The attribute num is the heart diseasediagnosis attribute. It was classified as presence and absence. If it is presence, then value of num would be low or medium or high or very high. If it is absence, then value of num would be zero. The dataset itself (297 instances) divided as 250 training instances and 47 as the testing instances.

B. Meta-heuristic with prepared RFNN

The recurrent fuzzy neural network had 13 input layer, seven hidden layer and one output layer as shown in figure 1. 64 bits long genes were used as weights for recurrent fuzzy neural network. The genetic algorithm parameters

were the probability 0.05 of mutation, 0.25 of multipoint crossover and the population size was 100. Figure 2 shows the genetic algorithm based neural network.

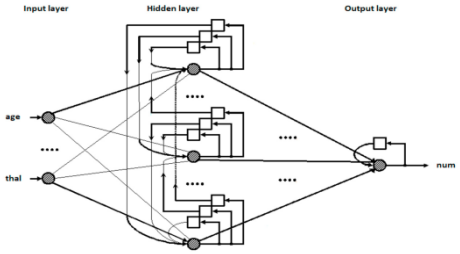

Fig. 1 The structure of RFNN

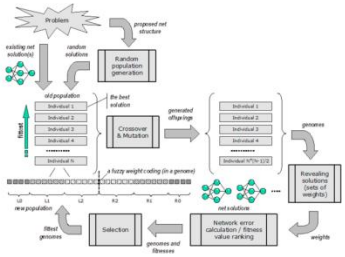

Fig. 2 Meta-heuristic with training of RFNN network

C. Evaluation criteria

The output and estimation of the analysis of coronary artery disease using meta-heuristic approach along with trained neural network and Mapreduce algorithm is presented. The results were obtained with the hardware arrangement of Intel i7 CPU with the capacity of 16GB RAM, LINUX system with Java. Table 2 presents the estimation analysis of the genetic algorithm along with trained neural network and Mapreduce algorithm. Heart disease attribute (num) is used to predict the coronary illness. True Positive (TP) is the condition where the number of instances classified as true while they were actually true. False Positive (FP) is the condition where the number of instances classified as true while they were actually false. False Negative (FN) is the condition here the number of instances classified as false while they were actually true. True Negative(TN) is the condition where the number of records classified as false while they were actually false.The accuracy is calculated as

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

With the advent of Hadoop distributed computing platform and the Mapreduce programming model[11], it has become easy for all the machine learning algorithms to process the data in parallel. This makes it easy to transform the machine learning algorithms to be transformed to Mapreduce paradigm by using Hadoop Distributed File System HDFS. Heuristic approach with prepared RFNN approach and ANN-Fuzzy_AHP which were proved to be efficient algorithms so that could be used and programmed by incorporating Mapreduce and we can evaluate the prediction with highest accuracy. We can make sure there is considerable amount of increase in the speed of processing time by increasing nodes count in the cluster. While Mapreduce model is running across multiple nodes than single node, it could save good amount of time and without compromising in accuracy.

Generally Mapreduce version performs extremely and efficiently well, while dealing with large masses of data. The training data was processed and optimized over several nodes through distribution and genetic algorithm with trained neural network approach can work in parallel on Mapreduce algorithm and one can reduce the training time. Meta-heuristic approach with prepared neural network approach is a dominant method for classification and regression. The computing and storage requirements of genetic algorithm with trained neural network approach increase with the number of training vectors which can be addressed. By distributing, processing and optimizing the subsets of the training data and directing them across several participating nodes we can achieve parallelization in Meta-heuristic approach with neural network approach. The parallel Meta-heuristic approach with prepared neural network approach based on Mapreduce algorithm reduces the training time significantly.

Meta-heuristic approach with prepared neural network approach is considered as powerful tool for classification and regression. By using parallel algorithms like Mapreduce efficiently, the scalability and performance requirements are achieved for large scale data mining along with big data analytics and training time is also become less.

The author Samuel[12] separated the Cleveland heart disease dataset as three subsets which include 193 samples as training set, 59 instances as validating set and 45 samples as testing set. This author compared the results with the conventional artificial neural network and author's approach had better accuracy results when compared with previous work. Table 2 represents the testing set performance of genetic algorithm based trained recurrent fuzzy neural network, conventional fuzzy artificial neural network and map educe algorithm.

This shows that Inverted index algorithm gives best accuracy than the meta-heuristic algorithm based trained neural network and conventional fuzzy artificial neural network.

Table 2 Performance comparison

Investigator	Algorithm used	True Negative	False Negative	TruePositive	False Positive	Total	Accuracy (%)
Samuel (2017)	Artificial neural network.	20	0	21	4	45	91.1
KaanUyar (2017)	Genetic alg.with RFNN	20	0	24	1	45	97.78
Proposed System	Mapreduce algorithm	18	0	26	1	45	98.12

IV. CONCLUSION

This study uses Mapreduce algorithm by comparing meta-heuristic approach along trained persistent fuzzy neural network on UCI machine learning repository dataset for predicting heart disease. The results of this study achieved 98.12 % accuracy for the 45 instances of testing set, when compared with meta-heuristic approach along with prepared neural network and conventional fuzzy artificial neural network. The output accuracy of proposed Mapreduce algorithm is distinctively better, because of dynamic schema and linear scaling. Here Hbase is used for storing resultant data. It has some latency due to batch processing, so in future batch processing is reduced then we can get more accurate output on comparing with other data mining techniques.

REFERENCES

1. WHOCardiovascularDiseases.http://www.who.int/cardiovascular_diseases.
2. A Statistical Update 2018 Report from the American Heart Association (AHA),http://circ.ahajournals.org/content/137/12/e67,(https://doi.org/10.1161/CIR.0000000000000558), March 2018.
3. AnkitaDewan, Meghna Sharma, "Prediction of Heart Diseases using a hybrid technique in Data Mining Classifiscation", 2nd International Conference on Computing for Sustainable Global Development (INDIACom),2015.
4. S.Bhagavathy, V.Gomathy ,S.Sheeba Rani, Sujatha.K, "Early Heart Disease Detection Using Data Mining Techniques with HadoopMapreduce", International Journal of Pure and Applied Mathematics, 119(12), 1915-1920, 2018.
5. Syed Umar Amin, Agarwal, K., & Beg, R., "Genetic neural network based data mining in prediction of heart disease using risk factors", IEEE Conference in Information & Communication Technologies (ICT), 1227-1231, 2013.
6. S. Shilaskar and A. Ghatol, "Feature selection for medical diagnosis : Evaluation for cardiovascular diseases," Expert Syst. Appl., vol. 40, no. 10, pp. 4146–4153, Aug. 2013.
7. Purusothaman, G., &Krishnakumari, P. "A survey of data mining techniques on risk prediction: Heart disease", Indian Journal of Science and Technology, 8(12), 1, 2015.
8. Miao, K. H., Miao, J. H., & Miao, G. J. (2016), "Diagnosing Coronary Heart Disease Using Ensemble Machine Learning", International Journal of Advanced Computer Science and Applications, 7(10), 30-39, 2016.
9. Dbritto, R., Srinivasaraghavan, A., & Joseph, V., "Comparative Analysis of Accuracy on Heart Disease Prediction using Classification Methods", International Journal of Applied Information Systems (IJAIS)–ISSN, 2249-0868, 11(2), 2016.
10. KaanUyar ,Ahmetİlhan, " Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks ", 9th International Conference on Theory and Application of Soft Computing, Procedia
11. UCI Machine learning Repository fromhttp://archive.ics.uci.edu/ml/datasets/Heart+Disease.
12. Mapreduce,https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
13. Samuel, O.W., Asogbon , G.M., Sangaiah, A.K., Fang, P. , Li, G., "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction", Expert Systems with Applications 68. 163–172, 2017.