Early heart disease detection using data mining techniques with hadoop map reduce Early Heart Disease Detection Using Data Mining Techniques with Hadoop Map Reduce

S. Bagavathy, V. Gomathy, S. Sheeba Rani, Monica Murugesan, K. Sujatha, M. K. Bhuvana

Early heart disease detection using data mining techniques with hadoop map reduce

Dr V Gomathy
Sri Krishna College of Engineering and Technology

International Journal of Pure and Applied Mathematics
Volume 119 No. 12 2018, 1915-1920
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue
ijpam.eu

Early Heart Disease Detection Using Data Mining Techniques with Hadoop Map Reduce

S.Bagavathy[1],V.Gomathy[2], S.Sheeba Rani[3], Sujatha.K[4], Bhuvana.M.K[5], Monica.Murugesan[6]

[1,2,3]Associate professor ,Department of Electrical and Electronics Engineering

[4]Professor,[5]PG Students Department of Computer Science and Engineering

Sri Krishna College of Engineering and Technology

Kuniyamuthur,Coimbatore-641008.

Abstract

Heart and other organs are important parts in human body. As per World Health Organisation(WHO)'s statistics, the cause of death in all over world is mostly due to cardiovascular diseases. The reason behind this are sedentary lifestyle which may lead to obesity, increase in cholesterol level, high blood pressure and hypertension. In this paper, by using various data mining techniques, such as Naive Bayes(NB), Decision Tree(DT), Artificial Intelligence (AI), Neural Network (NN) and clustering algorithms such as Association Rules. Support Vector Machine (SVM) and K-NN algorithms are used to extract the Knowledge from the large number of data set. The generated reports help doctors and nurses to identify about disease and their levels with which they can provide a better treatment to the patient. Text Mining is most commonly used mining technique in health care industry. In this paper we compare K-means clustering algorithm with Map Reduce Algorithm's implementation efficiency in parallel and distributed systems.

Keywords- Artificial Intelligence(AI), Decision Tree(DT) ,Naive Baye's(NB), Neural Networks(NN),K-Means Clustering,World Health Organization(WHO),Map Reduce .

1. Introduction:

Heart diseases are also known as cardiovascular diseases which occur due to unhealthy lifestyle, smoking, alcohol, high intake of fats which may cause strokes, hypertension, high blood pressure, diabetics. Sometimes these diseases may also being caused by genetic reasons and some other reasons. Symptoms of Heart diseases are Fluttering in chest, Racing heartbeat, Chest pain, Dizziness, Light-headedness, Fainting. Sometimes some people may suffer from heart infection which is caused due to some virus, bacteria. In the year 2017, WHO estimated that because of cardiovascular diseases , almost 17.7 million people were dead . The death percentage is overall 31% and 13% death from overall deaths was due to coronary heart disease and 6.7 million was due to stroke. Recent days all the people from the age group of 25-40 are advised to consult a heart specialist twice a year to know about their heart health. For that people were advised to undergo certain test conducted by well equipped heart speciality hospitals like blood tests, exercise stress test, cardiac catheterization radionuclide ventriculography, chest X-Rays, Magnetic Resonance Imaging (MRI), electro cardiogram, echo cardiograph or multiple gated acquisition scanning that results important information. The reports will help the doctors to view about the risk level of organs in human body. In this paper we are using Artificial Neural Networks and K-mean clustering for disease prediction and provides 100% accuracy by comparing Naïve Bayes and Decision Tree. This is cost effective approach for healthcare organizations for making analysis of financial and clinical data decisions. Not only the experienced Doctor can do the decision , new practitioner also can able to generate the report and make treatment decision based on disease stage of the heart.

2. Literature Survey:

Rucha Shinde et al have used k-means algorithm for clustering and Naïve Baye's algorithm for disease prediction. This is done by taking a large number of raw data as input and produces useful information from the data. Serialization technique is used to convert the data into bytestream and it is stored in database. Input Data are converted into Comma Separated Vector files (CSV) where these files contain the attributes of patients (Age,Gender,Blood Pressure, Sugar Level etc) and it is grouped by using k-means algorithm and prediction is applied by using Naïve Baye's algorithm[1].

Kaan Uyara et al have used Genetic Algorithm which works based on Recurrent Fuzzy Neural Networks idea about the overall estimation of heart disease faced by people in the world for heart disease detection which uses overall 297 patient dataset from them 252 was considered as training dataset and 45 was considered as testing dataset and overall accurate result came out to be 97.78% from testing dataset [2]. It uses Root Mean Square Error(RMSE).Accuracy of overall performance is analyzed using factors like sensitivity, F-score, probability of miss classification error, precision and specificity. A comparative study made between Artificial Neural Network and Genetic Algorithm based Recurrence Fuzzy Neural Network and accuracy is more in Genetic Algorithm based Recurrence Fuzzy NN[2].

Jyoti Soni et al used Tanagra tool to compare the classifications of different types of data mining algorithms such as Naive Baye's, Decision Tree, K-Nearest Neighbour ,Artificial Neural Network and Genetic Algorithm. Artificial Neural Network has been implemented in .NET framework. By comparing all these algorithm, Decision Tree has more accuracy. Association Rule Mining is also used with one additional parameter lift along with support and confidence to get specificity and sensitivity values[3].

G. Purusothaman et al used two data models-single models and combined models,combined models are also called as hybrid data models they have used both classification and clustering techniques such as K-NN,Decision Tree, SVM,ANN,Association Rule,Hybrid Approach and Naïve Baye's Approach.On comparing both these models hybrid models are best models to be used and Hybrid Approach is said to be the best approach to be used as it can provide better result for diagnosis[4].

Mirpouya Mirmozaffari et al have used WEKA tool for analyzing the dataset and discovering knowledge and predicting the patterns from dataset. The process is carried out in two steps Multilayer filter preprocess and Evaluation in classification. In order to get balanced dataset from imbalanced noisy dataset , they applied filters to remove redundant data and noise from the dataset. In Evaluation Classification three different steps are carried out:1. Training set 2. Fold cross validation 3. Percentage split[5].

Vikas Chaurasia et al used WEKA tool to compare several attributes for prediction of heart diseases with by using different types of data mining algorithms like CART(Classification and Regression Tree) ,Decision Table and ID3
Amongst these CART has highest accuracy and ID3 has highest error rate[6].

Data Mining Methodology:

Data Mining is a technique of finding interesting patterns from the existing data as per different conditions with which we can make the data into useful information. Here we use patient's dataset and retrieve output as a useful information with which doctor needs to diagnose the patient or not. Since raw dataset may contain some noise, hence data mining techniques are followed. Data Pre-processing ,Transforming then we use Data Mining Techniques like Classification or Clustering out of which we will be using K-means Clustering and Evaluate the Result based upon that a knowledge is represented.

K-Means Clustering Algorithm:

K-means clustering algorithm, which is an unsupervised type of learning algorithm which means clustering process is done on data and these data are being categorized into similar groups. The procedure is very simple it takes input dataset which contains values of patients training dataset which is with respect to attributes like age, sex, sugar levels etc, based upon these dataset we retrieve some information about heart disease and we form k-clusters with respect to diseases. Since main idea is to place these clusters in distant manner so we place these clusters within some distance. In next iteration we place few clusters in such a way as they are somewhat related to each other. Further iterations we proceed calculating k-centroids from previous step and then binding is then between same set of data set and nearest centroids. These k-centroids change their locations at every iterations until no more calculations are performed.

Steps performed in K-means clustering algorithm:

1.Assume A={a_1,a_2,....a_n} be the attributes of datasets and B={b_1,b_2,.....b_n} be set of centre points.

2. Randomly choose the c_i cluster points.

3.Calculate distance between data points[attributes(a)] and centre points(b) using formula:

$$D(V)=\sum\sum(|a_i-b_i|)^2$$

Where the range is between i=1 to c_i.

4. Assign the data point to the cluster points whose distance from the cluster point to the data point is minimum of all.

5.Re-calculate the cluster point.

Map Reduce Algorithm:

Map Reduce 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. It is implemented by open source Apache Hadoop.Map Reduce is programmed using Java for high reliability and scalability.

Hadoop Distributed File System(HDFS) comprises of Hbase and Hadoop Map Reduce. It analyzes large data.It has two nodes: Master and Slave Nodes. Master Node comprises of Job Tracker and Name Node. Slave Node comprises of Task Tracker and Data Node.

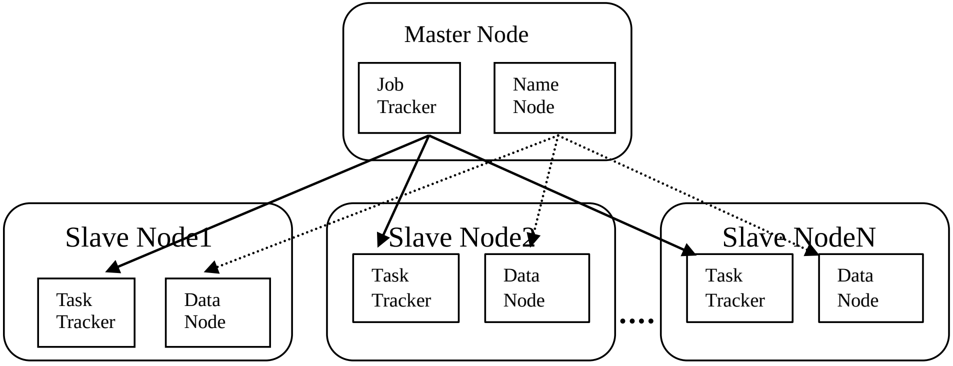

Fig 1: Components of Hadoop

Steps involved in Map Reduce Algorithm:

1.**Input** : Large Dataset is taken.
2.**Splitting**: Splits Dataset into number of small
3. **Mapping** : Generates Key-value pairs in each dataset.
4: **Shuffling**: Key-Value pairs are separated based on their association.
5: **Reduce**: Reduces the Single Key-Value pair with its count.
6: **Result**: Final Result are reduced and stored in database.

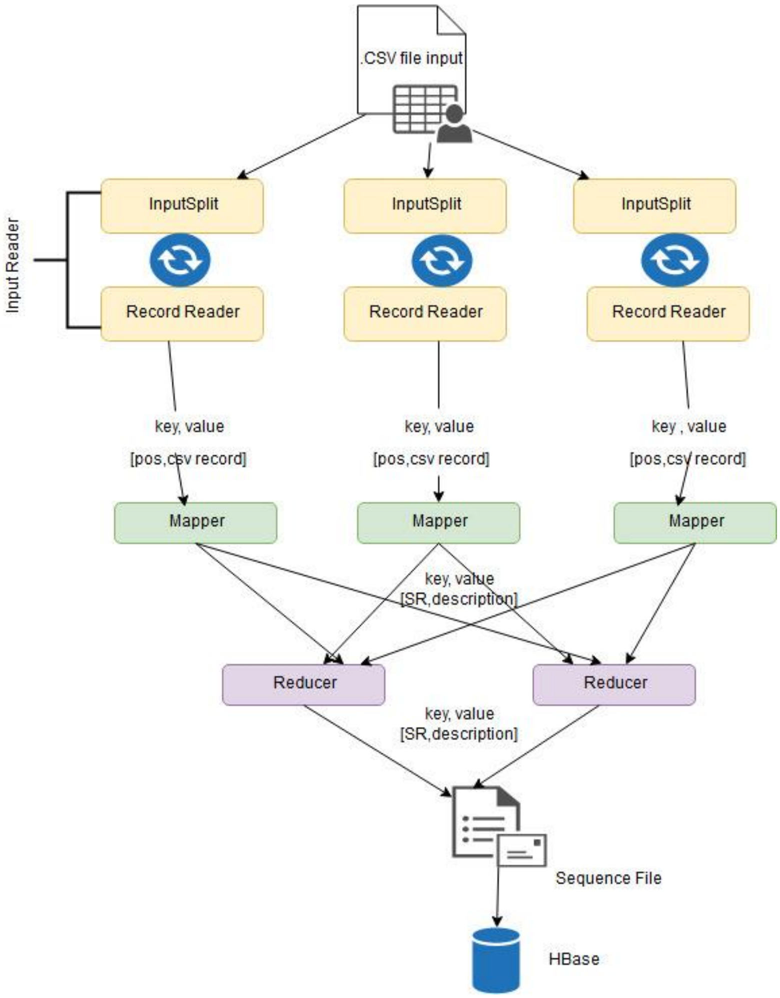

Fig.2: Working of Map Reduce Algorithm.

Conclusion:

In this paper we have studied various data mining techniques used to obtain information from the large dataset. We have studied about K-Means Clustering Algorithm and compared with Map Reduce Algorithm for detecting heart diseases. The output accuracy of Map Reduce Algorithm is better than K-Means Clustering Algorithm because of dynamic schema and linear scaling. We use Hbase 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. Shinde, R., Arjun, S., Patil, P., & Waghmare, J. (2015). An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm. *IJCSIT) International Journal of Computer Science and Information Technologies*, *6*(1), 637-639..
2. Uyar, K., & İlhan, A. (2017). Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. *Procedia Computer Science*, *120*, 588-593.
3. Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), 43-48.
4. Krishnamoorthy, Sujatha, Shalini Punithavathani, and Jaya K. Priya. "Extraction of well-exposed pixels for image fusion with a sub-banding technique for high dynamic range images." *International Journal of Image and Data Fusion* 8.1 (2017): 54-72.

5. Mirmozaffari, M., Alinezhad, A., & Gilanpour, A. (2017). Data Mining Classification Algorithms for Heart Disease Prediction. *Int'l Journal of Computing, Communications & Instrumentation Engg (IJCCIE)*, *4*(1).
6. Gomathy.V, Sheeba Rani.S.2018."Real time intruder detection using passive infrared sensors" in *Int.Journal of Pure and Applied Mathematics* , Volume.118, No.20, 2018, pp 1219-1224
7. Rajalakshmi, K., & Nirmala, K. (2016). Heart disease prediction with mapreduce by using weighted association classifier and k-means. *Indian Journal of Science and Technology*, *9*(19).
8. Zolfaghar, K., Meadem, N., Teredesai, A., Roy, S. B., Chin, S. C., & Muckian, B. (2013, October). Big data solutions for predicting risk-of-readmission for congestive heart failure patients. In *Big Data, 2013 IEEE International Conference on* (pp. 64-71). IEEE.
9. Nalavade, J., Gavali, M., Gohil, N., & Jamale, S. (2014). Impelling Heart Attack Prediction System using Data Mining and Artificial Neural Network. *International Journal of Current Engineering and Technology*, *4*(3), 1-5.
10. Yadwad, P. P. K. S. A., & Tejaswi, V. V. D. L. A Data Mining Technique for Prediction of Heart Disease using Hadoop Mapreduce.
11. Punithavathani, D. Shalini, K. Sujatha, and J. Mark Jain. "Surveillance of anomaly and misuse in critical networks to counter insider threats using computational intelligence." *Cluster Computing* 18.1 (2015): 435-451.