Heart Disease Prediction using Cluster Based MapReduce Paradigm

J.Sukanya,[*] Dr.K.Rajiv Gandhi,[**] Dr. V. Palanisamy[***]

[*] Research Scholar, Alagappa University,Karaikudi
[**]Department of Computer Science, Alagappa University Model Constituent College of Arts and Science, Paramakudi,
[***]Department of Computer Applications, Alagappa University, Karaikudi
Email: [1]sukanrajam76@gmail.com

DOI: 10.29322/IJSRP.11.03.2021.p11167
http://dx.doi.org/10.29322/IJSRP.11.03.2021.p11167

Abstract

The healthcare industry contains a large amount of information that is difficult to process by means of manual methods. Big data is too valuable to extract information and to form relationships in the area of large data sets. All the clustering techniques are utilized to gather things that are like each other. In this paper, we propose an equal k-implies grouping calculation dependent on MapReduce, which is a straightforward yet amazing equal programming procedure with enormous information examination. The test results show that the proposed calculation can be proficiently measure huge datasets on item equipment.

I. INTRODUCTION

Clustering [1][2][3][4] is the task of finding groups and structures in the data that are in some way or another "similar", without using predefined class labels. Clustering Analysis or Clustering is a method of grouping data into different groups (i.e.) set of objects, so that the data in each group share similar trends and pattern. A clustering technique will deliver top notch groups with high intra-bunch similitude and low between group comparability. The nature of an outcome created by bunching relies upon both the likeness measure utilized by the technique and its execution.

In this work, the MapReduce based equal k-implies grouping calculation is proposed for foreseeing the coronary illness. MapReduce has two steps that are Map step and Reduce step, it is a parallel task executed on many different computers and one of those tasks that a developer can design is the map process. There are different map processes that are introduced into a MapReduce job, each of those map processes works on data independently from the enormous set of data that you've got as a base set. After the map process is done, Hadoop shuffles data around to several reduce processes. Those operate in parallel in conjunction with the map processes.

II. RELATED WORK

There is several research works carried out on the basis heart disease prediction system using clustering algorithm in past decades. They are discussed below

In [5] they have been proposed MapReduce based k-means for diagnosing diabetic patients. They have used Prima Indians diabetics' database of National Institute of Diabetes and Digestive and Kidney Diseases with only 8 attributes and 768 instances. The k-means bunching sort of information mining with MapReduce has been connected to the dataset where the class variable result has been punched into two gatherings in particular cluster1 (Diabetic – 268 instances) and cluster 2 (Non-Diabetic – 500 instances). They have been followed the given steps for clustering, step 1: read the bunch focuses into memory from a sequential file, step 2: iterate over every bunch community for every key/value pair, step 3: measure the separations and recovery the closest focus which has the most minimal separation to the vector. The elapsed time for clustering is 15.842 sec.

In [6] authors proposed a hybrid approach (k-means and artificial neural network) for predicting the risk levels of heart diseases. They have used 14 attributes named as age, se, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting cardio graphic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, thalassemia, number of major vessels and angiographic disease status. They have been used k-means for obtaining the appropriate number of data groups. The Euclidean distances were calculated for different attributes. From their observation, it has been identified that if the mean value of the patient is closest to the sample

mean value, the patient more likely to be affected by heart diseases. Then the data set was divided into two part 70% for training and 30% for testing. They have obtained higher accuracy rate of 97%.

In this paper [7] authors proposed clustering based classification techniques for predicting the patients of heart diseases. They have proposed the system with two steps. The first step is clustering which will cluster the similar and dissimilar type of data. The second step consists of classification which will classify the clustered data for the prediction analysis. Similarly clustering consists of two steps, in the first step, the mean of the dataset is computed for identifying the centroid point, in the second step, Euclidean distance from the centroid point and the object is computed that represents the similarity between the data objects. They have used the normalization techniques (min-max) for reducing the complexity of the large datasets. The proposed clustering based classification and normalization technique are implemented in matlab and increased the accuracy from the existing technology.

In [8] have been established an advanced k-means clustering algorithm based on compressed sensing theory in combination with k-SVD method. They have used large ECG dataset including 668486 beats. They have also demonstrated the proposed algorithm with a collaboration of PCA as a dimensionality reduction method. The obtained accuracy and sensitivity are 99.98% and 99.92% respectively. The proposed algorithm also reduced 13% clustering energy consumption compared to existing clustering algorithm. They also suggested that the proposed algorithm has many practical applications including wireless ECG systems, Holter monitoring, electronic health and mobile health.

In this paper [9], have proposed explored the K-means clustering algorithm for prediction of various diseases (heart, liver, diabetics and cancer).The assessment of disease, data mining techniques and accuracy is estimated, in which k-means algorithm forecast more accurate than all other techniques. The k-means clustering method for prediction reduces the human effects and is cost effective one.

III. PROBLEM DEFINITION

Clustering algorithms partition data into a certain number of clusters (groups, subsets, or categories). The idea is clusters are used to identify the heart disease of the patients with fraction of seconds as well as the proposed method works smoothly when the dataset is large.

A.MapReduce Implementation

An input to a MapReduce job is divided into fixed-size pieces called input splits Input split is a chunk of the input that is consumed by a single map. The MapReduce processes are depicted in Fig. 1.

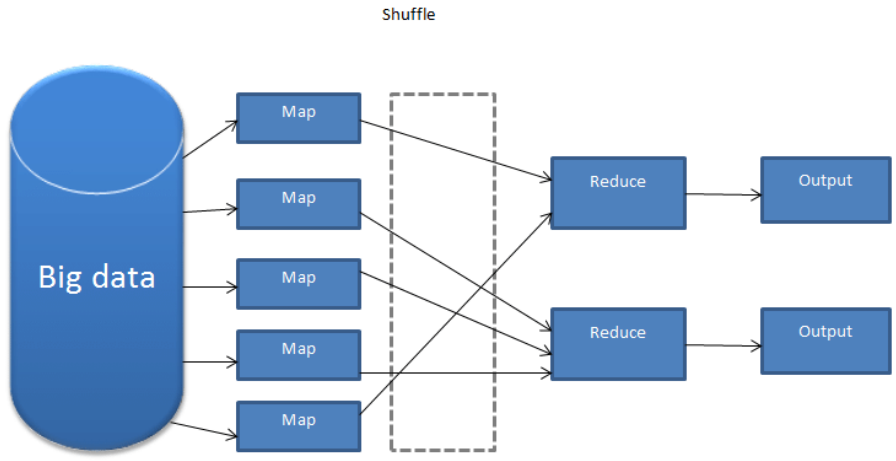

Fig. 1. MapReduce Architecture

The MapReduce calculation contains two significant undertakings, specifically Map and Reduce.

1. The map task is finished by methods for Mapper Class
2. The reduce task is finished by methods for Reducer Class.

Mapper class takes the information, tokenizes it, guides and sorts it. The yield of Mapper class is utilized as contribution by Reducer class, which thus look through coordinating with sets and diminishes them.

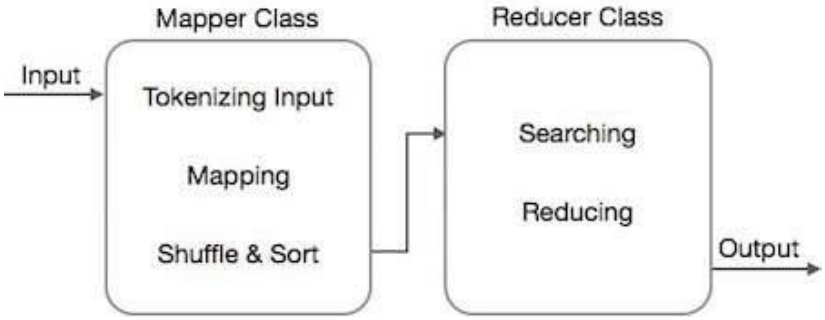

Fig. 2. Block diagram for MapReduce Function

MapReduce executes different numerical calculations to partition an undertaking into little parts and dole out them to numerous frameworks. In specialized terms, MapReduce calculation helps in sending the Map and Reduce undertakings to proper workers in a cluster. The MapReduce algorithm process descried in Fig.2.

B. Parallel k-means clustering algorithm

The parallel k-means algorithm is proposed for efficiently and accurately clustering the heart patients data under MapReduce framework on the hadoop environment. The proposed algorithm follows traditional k-means steps in parallel fashion. It selects k of objects as initial cluster mean, where k represents the number of clusters.

Every one of the excess articles is appointed to the group dependent on the closeness and the distance between the items and the bunch centroids. The centroid for each group is figured. This progression proceeds until there is no component to part and no adjustments in the centroid of the ach group. Generally k-implies calculation depends on the distance measures and instatement. The distance between the objects and cluster means are calculated by Euclidean distance. The Euclidean distance between two points, a and b, with k dimensions is calculated as[10][11]: Euclidean (and squared Euclidean) distances are usually computed from raw data, and not from standardized data. One advantage of this method is that the distance between any two objects is not affected by the addition of new objects to the analysis, which may be outliers [11]. However, the distances can be highly affected by differences in scale among the dimensions from which the distances are computed. The following formula is used to calculate the Euclidean formula.

$$\text{Distance} = \sqrt{\sum_{i=1}^{n}(qi - pi)^2}$$

The works of the parallel k-means clustering are composed of three tasks such as (i) Compute the distance between the object and the cluster mean. (ii) Assign each object to its closest centroid. (iii) Re-compute new cluster mean for each cluster.

C. Implementation Details

The symptom of heart disease includes feeling gripping and tight usually on the chest but spread to shoulders up to the stomach. The types of angina are atypical angina, typical angina, asymptomatic and non-anginal pain [12].

The prediction of heart disease is made with 303 samples with seven independent features like age, chest pain type, blood pressure, blood glucose level, ECG in rest, heart rate and four types of chest pain and the habitual of physical exercise. Age is considered as the main risk factor for heart diseases as coronary fatty streaks develops in the adolescence stage. Male are at higher risk of coronary diseases than females [13]. High blood pressure is one of the major causes of heart disease as it damages arteries. Only 14 attributes are considered as input for the proposed system.

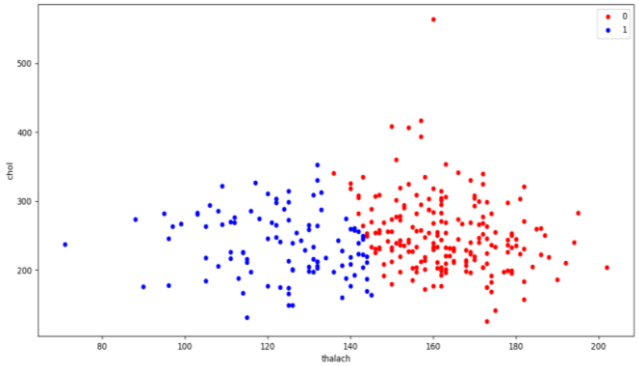

Fig. 3. Visualization of clustered data (thalach and chol attributes without centroids)

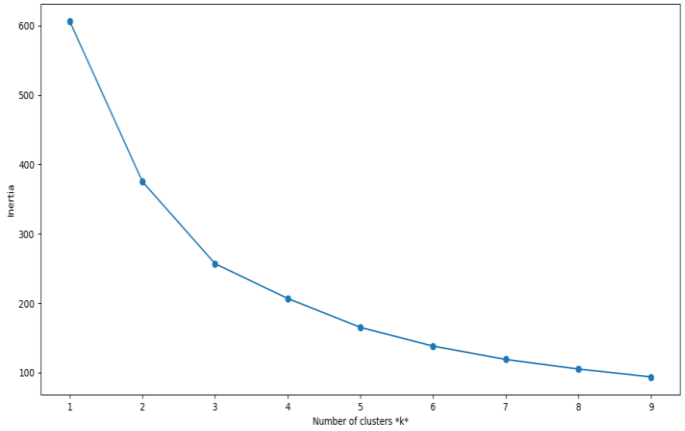

Fig. 4. Inertia curve of the proposed clustering algorithm

Fig. 3 represents the visualization of clustered data with chol and age attributes with cluster centroids The number 1 indicates there is chance of the heart disease. It also represents the visualization of clustered data thalach and chol attributes without cluster centroids. Fig. 3 represents inertia curve of the proposed clustering algorithm. The Fig. 4 depicts when the no of the clusters increases the time complexity and computational cost also increases.

D. Comparative Analysis

The proposed parallel method also compared with bench-mark clustering algorithms such as canopy, cobweb, EM (Expectation Maximization), Farthest first, filtered cluster, hierarchical cluster, density based cluster and simple k-means.

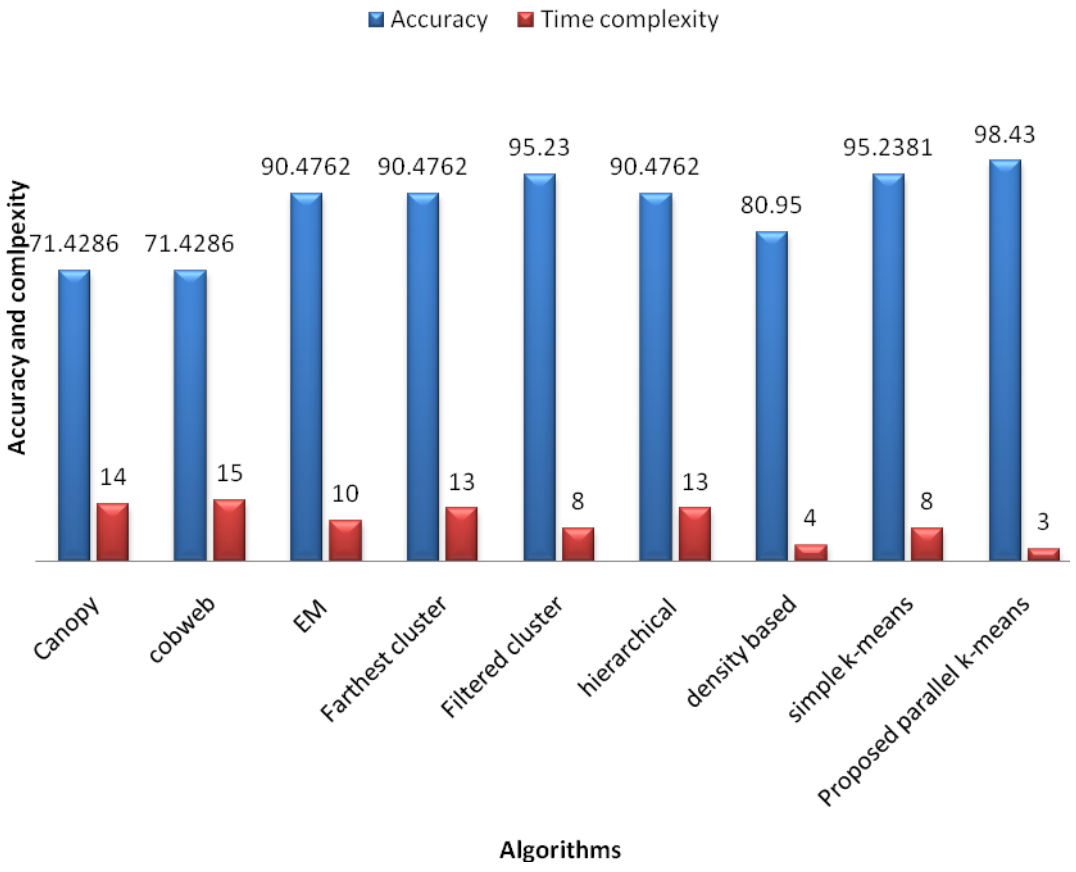

Fig. 5. Comparative Analysis

Fig. 5 represents the accuracy and time complexity of all the bench mark clustering algorithms. Among 303 instances there are 298 instances are correctly identified in 3 seconds.

IV. CONCLUSION

The main objective of the work is to comparative analysis methods can be employed in heart disease predictions. From the observation, it has been identified that MapReduce based parallel clustering algorithm takes less amount of time as well as gives more accuracy compared with bench-mark algorithms and disease can be prevented by taking certain precautions wither for a particular age group or a specific gender or for a particular region people. This may lead to the next generation of the health care treatments.

References

[1] Pavel Berkhin, "A Survey of Clustering Data Mining Techniques", pp.25-71, 2002.

[2] Arun K Pujari " Data Mining Techniques" pg. 42-67 and pg. 114- 149,2006.

[3] Pradeep Rai, Shubha Singh" A Survey of Clustering Techniques" International Journal of Computer Applications, October 2010.

[4] M.Vijayalakshmi, M.Renuka Devi, "A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets" , International Journal of Advanced Research in Computer Science and Software Engineering, pp.305-307, 2012.

[5] Sharmila, K., and Manickam, S. A. V., "Diagnosing Diabetic dataset using hadoop and k-means clustering techniques", Indian Journal of Science and Technology, Vol. 9(40), 2016.

[6] Malav, A., Kadam, K., and Kamat, P., "Prediction of the heart disease using k-means and artificial neural network as hybrid approach to improve accuracy", International Journal of Engineering and Technology (IJET), Vol. 9(4), 2014.

[7] Singh, R., and Rajesh, E., "Prediction of heart diseases by clustering and classification techniques, International Journal of Computer Sciences and Engineering, Vol. 7(5), 2019.

[8] Bolouchestani, M., and Krishnan, S., "Advanced k-means clustering algorithm for large ECG datasets based o a collaboration of compressed sensing theory and k-svd approach", SIVip, Springer,Vol. 10(1), 2016.

[9]K.Rajalakshmi., Dr.S.S.Dhenakaran., and N. Roobin., "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, Issue 7, July 2015

[10] Elena Deza , Michel Marie Deza, "Encyclopedia of Distances", Springer, 2009, page 94.

[11] http://www.statsoft.com/textbook/cluster-analysis/, March 2, 2011

[12] Indrakumar, R., Poongodi, T., and Jem, S. R., "Heart disease prediction using exploratory data analysis", International Conference on smart sustainable intelligent computing and applications under ICITET 2020, Procedia computer science, Vol. 173, 2020.

[13] UCI Machine learning Repository https://archive.ics.uci.edu/ml/index.php