2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)
Zeal Education Society, Pune, India, Feb 24-26, 2017

Smart heart disease prediction system using Improved K-Means and ID3 on Big Data

Ms. Tejaswini U. Mane
Department of Computer Engineering,
Zeal College of Engineering & Research, Narhe,
Pune, India
tejaswinimane18@gmail.com

Abstract— **The term Big Data is becoming global today. The Big data is huge amount of variety of data, and the data is increasing very rapidly according to the time. So there is need to process that data and instead of just storing that data need to extract some meaningful information or knowledge from that data applying some clustering and classification techniques of data mining. There are various era available in the Big Data so that decided the medical field first. And after that there are various diseases available to work on them or gain some knowledge or predict for help we decided the Heart disease. Heart disease is one of the disease due to that death will occurred mostly, and according to the world health organization the percentage is more for that. So Heart disease is decided for the big Data approach, and as Big Data is considered so used Hadoop Map reduce platform. For clustering Improved K-Means and for the classification purpose decision tree algorithm i.e. ID3 is used in the hybrid approach. As we know the taking second opinion is too increased, the system is very useful for the helping in prediction, basis on the some parameters like chest pain, cholesterol, age, resting Bp, Thalac and many more. Due to this system clinical decision making will be improved as well as being fast. It's also will impact on the improving the treatment process. In such way it will be very useful in the prediction of the heart disease.**

Keywords—Big Data; Clustering; Classification; Data mining; Improved K-Means; ID3; Map reduce.

I. INTRODUCTION

The digital data is increasing very tremendously according the time, to get the meaningful information from that need to process it and apply some data mining techniques [1]. As the era is medical, in that heart disease; we know that the clinical determination is taken basis on the maturity and the awareness. But most of the time due to the lack of knowledge and the bad decision lots of people suffered. Now days we can see that in lots of the hospital patient's information and histories of them are stored, but those are going to increase the storage size instead of using it for the some good kind of knowledge [2]. The heart disease is the major diseases due to these peoples have to spend large amount of money starting from analysis to the medication, but if we can analysis it basis on the some parameter then it will reduce the peoples complication of several of testing's and will save their money.

The world health organization said that 12 million deaths occurred worldwide because of heart disease per year [3]. Heart disease is related to the heart problem and its related vessels. Heart is the main organ of the human body which pumps the blood, if it can't work properly then brain will suffered and they both altogether will stop the working and within minutes death will occurs [4]. So the heart is too much important for the human body and we have to take care of it. Now days due to the changing of life style and the too much stress of works on the people it also lead for the heart disease [5]. So we have to avoid all those things which will be made stress to us. And need to keep calm and cool our heart then we will also live long happily.

In the System the idea is if any hospital has various branches located various geographically locations. And their main hospital branch playing the main Server role, because it is generating the data from other branches. Patient or User directly asks the query basis on his parameters related to health Like B.P, Age and all. And From the Server he or she got the reply about the disease status.

In this our contribution can be summarized as follows:

1) As Big Data is Considered So We are using the Standard Repository Database of the Heart disease information. With the help of that data we have decided some parameters for our system Like: Age, Sex, Chest Pain, Resting B.P, Cholesterol, Fasting BS, Resting ECG, Thalac, Exercise Induced Angina, Old Peak, Slope, Thal, Ca, and Diagnosis [7].

2) First off all there is need to clusterised the database according to the parameters those passed by the user for the diagnosis purpose. So for that here used improved K-means Algorithm Instead of just Simple K-Means for the Accuracy of the Clustering Centroids.

3) As the geographically distributed data at different locations it is in vast amount available on the centralized main server. And then second algorithm ID3 in the category of decision tree we used to build the tree on the result of Improve K–Means. It is used Hybrid approach.

4) By using the Standard Database and algorithm decision tree is build and due to that when user pass the parameters system will show the status of the heart disease.

978-1-5090-4083-4/17/$31.00 ©2017 IEEE

The rest of the paper is arranged as follows: Section 2 presents the literature survey over the related work. In section 3, Implementation Details are given; It also includes the System Framework With the Used Algorithm and the Mathematical Model. Section 4 covers the Comparative results of the system. Finally, the section 5 concludes the paper.

II. RELATED WORK

A. Big Data

Big Data applications wherever knowledge assortment has grownup enormously and is on the far side the flexibility of ordinarily used software tools to control, take and method through a tolerable period. The foremost elementary test for Big knowledge applications is to explore the big volumes of data and extract helpful data or information for planned actions[1,2].Big data has the various characteristic such as volume, velocity and verity, And the Big data is composite, self determining and geographically assigned and having or evolving the complex relationship[1,7].

B. Big Data mining

For Big Data mining we have several platforms are available. Hadoop Map reduce platform is having good approach with the HDFS file system over the distributed file system. Big data processing refers to the activity of researching vast knowledge sets to seem for relevant information [12]. Great knowledge samples area unit on the market in physics, field science, social networking sites, bioscience, life sciences, government knowledge, natural disaster and capability management, mobile phones, web logs, sensing element networks, research and telecommunications. By match with the results previously unoriginal from drilling the typical datasets, unveiling the massive volume of interconnected composite great knowledge has the potential to maximize our information and insights within the target domain[13]. However, this brings a series of recent test to the search community. To overwhelm the challenges will reshape the long run of the info Drilling technology, resulting in a spectrum of groundbreaking knowledge and mining techniques and algorithms. One possible approach is to boost already available techniques and algorithms by apply massively parallel computing architectures.

III. IMPLEMENTATION DETAILS

A. System Framework

The System Framework diagram is shown below, and by looking that it gives us the clear idea about the system and its working. In the System there are the main key terms are given bellow:
-Server (Centralized Control)
-Admin (Add Manage the Database)
-Client (Client side or Patients side
-Data Nodes (Distributed Data Nodes)
-HDFS (Hadoop Distributed File System)
-DB (Database)
All those are the playing important role in the system, to made work it properly

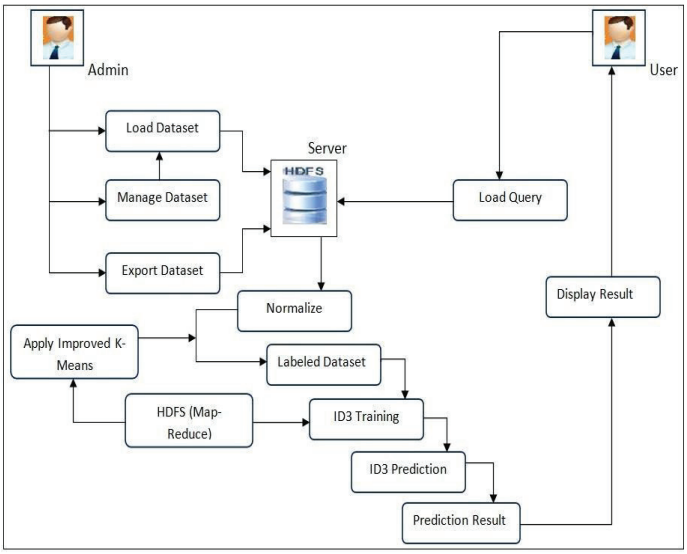


Fig1: System Framework

Near 25 percent of deaths in the age group of 25 to 69 years occur because of heart diseases[15].By seeing the above diagram of the system its cleared the system working flow.

Module 1: Admin module
The Admin Module is consisting with the following main operations those are listed below:

- Login to the system:

The admin have to login in the system first with valid id and password to perform the further functionalities in the system.

- Load the Main Database csv file:

After the successful login of the admin, he will load the main database csv file which is used for the learning and the building the test database file, which is important for the training purpose to the system. This file is having 303 records and having various entries details with the parameters those are used in the system for the heart disease analysis.

- Export database: Uploading the testing csv file:

In this Exporting phase here admin can upload the test database file for the training procedure. This file also having the parameters entries but those are not repeated, in this file complexity is reduced and focused on the actual and exact accuracy of the entries with various combination. So that in the training phase it will be helpful to train the system by all features.

- Training the System:

In this phase admin did the system training on server side. In that system used the test database file and basis on that by

using algorithm it will be train the system for the accurate prediction of the heart disease.

Module 2: Client module
The Client Module is consisting with the following main operations those are listed below:

- Enter the parameter of single entry to analysis:

In this phase user will do the entry of the clinical parameters like age, weight, cholesterol, resting bp, resting ECG, Old peak + slope, exercise induced angina, thalac and there are other parameters entry is done. User also sees here the prediction analysis of the system and also with the graph.

- Load the CSV file for the multiple entry analysis:

Here if user has the already filled entry csv file with multiple user's parameters then he can done such analysis here with graph prediction.

Module 3: Heart disease server module

- Making database proper in to the required numerical format:

Here as there is in the system Hadoop platform is used the server coding in system is done with the map reduce functionalities. For building the accuracy clusters data is checked first with their valid entries and converted into the appropriate numerical format for the cluster building.

- Normalize the all database with checking each record of the test database:

Database is having various different entries in that each record is very precisely checked first and then normalization can performed on it. The data is send into the chunk to server for the processing.

- Apply the improved K-Means clustering output to the ID3 algorithm:

After the all normalization processing done on the database the clusterised output is sending to the classifier for the further processing and analysis of the heart disease.

- For the training purpose apply the classifier:

Here in this phase classifier ID3 will generate the tree with all the available parameters with the using output of the normalization and classifies all attributes with its information gain factor till the end to show the prediction of the heart disease.

B. Algorithms Used

There various algorithms are available for the Data Mining and Big Data. We can choose as per the requirement of our database, and requiring approach.

a. Improved K-Means Algorithm

In the system used 2 necessary algorithms; first is Improved K-Means for the clump and second is ID3 for Classification call tree[7]. One among the famed clump algorithms is K Means. K-means is one among the best unattended learning algorithms acknowledged for its speed and ease. However, this algorithmic rule suffers from 2 major limitations. First, the clusters made area unit sensitive to the choice of initial centroids (cluster centers)[8]. Second, the algorithmic rule needs worth of variety of clusters to be made (K) as input. For the system we used changed K-means algorithmic rule that classifies the computer file set into applicable clusters while not taking variety of clusters K as input, because it was needed within the case of K-means. The projected algorithmic rule doesn't need the quantity of clusters K as input. And conjointly compared the time complexness and accuracy of the clusters made thereupon of the initial K-means algorithmic rule.

Algorithm:
Input:
D: The set of n records with attributes Al, A2,…, Am
Where m = no. of attributes. All attributes are
numeric. Output:
Suitable number of clusters with n records distributed properly
Method:
1. Compute sum of the attribute values of each record (to find the points in the data set which are farthest apart) and take records with minimum and maximum values of the sum as initial centroid's.
2. Create initial partitions (clusters) using Euclidean distance between every record and the initial centroid's. And then find distance of every record from the centroid in both the initial partitions. Take d=minimum of all distances. (Other than zero)
3. Compute new means (centroids) for the partitions created in step 2.Compute Euclidean distance of every tuple from the new means (cluster centers) and find the outlier's depending on the following objective function: If Distance of the tuple
from the cluster mean < d then not an Outlier.
4. Compute new centroid's of the clusters. Calculate Euclidean distance of every outlier from the new cluster centroid's and find the outlier's not satisfying the objective function in step 3.
5. Let B=Yl, Y2,…,Y p) be the set of outlier's obtained in step 4 (value of k depends on number of outlier's). Repeat unti I (B==<D)
a) Create a new cluster for the set B, by taking mean value of its members as centroid.
b) Find the outlier's of this cluster, depending on the objective function in step 3.
c) If no. of outlier's = p then
i) Create a new cluster with one of the outliers as its member and test every other outlier for the objective function as in step 3.
ii) Find the outlier's if any
d) Calculate the distance of every outlier from the centroid of the existing clusters and adjust the outlier's in the existing which satisfy the objective function in step3.
e) B= ZI, Z2,...., Zq, be the new set of outlier's. (Value of q depends on number of outlier's)[9]

b. ID3 Algorithm

ID3 may be an easy call tree learning algorithmic rule developed by Ross Quinlan (1983). The essential plan of ID3 algorithmic rule is to construct the choice tree by using a top- down, greedy search through the given sets to check every attribute at each tree node. ID3 may be a non progressive algorithmic rule, which means it derives its categories from a hard and fast set of coaching instances. A decision tree consists of nodes and arcs that connect nodes. To form a choice, one starts at the basis node, and asks inquiries to confirm that arc to follow, until one reaches a leaf node and therefore the call is formed [10].

Algorithm:

1) Establish Classification Attribute (Let Table R which is the set of records containing the all parameters (Database))
2) Compute Classification Entropy.
3) For each criticize in R, jump to a conclusion Information Gain via classification attribute.
4) Select Attribute mutually the highest win to be the while later Node in the tree (starting from the Root node).
5) Remove Node Attribute, creating drained table R.
6) Repeat steps 3-5 until bodily attributes have been secondhand, or the related classification worth remains for the most part rows in the all table.

c. Mathematical Model

In the Smart Heart Disease Prediction system, first database is loaded then applied data mining algorithms Improved K-Means and ID3 respectively for clustering and classification in the hybrid approach. INPUT: Parameters entered by the users
OUTPUT: Diagnosis or Prediction of Heart Disease.
Where,

The system is represented as:

S = {C;N; S; P;R; MD; TD; ob1; ob2}

Where,

C= Set of Client,
N= Set of Cluster nodes,
S= Server,
P= set of Parameters,
R= Record from database,
MD= Main Database,
TD= Training Database,
ob1= Euclidean distance

R Є (MD; TD)

P Є (R)

ob2= used for classification

ob2 Є (H(S); H (G)g)

1. Client: have his parameters values related to parameters set. P={Age; Gender;RestingECG; FastingBs; Cholesterol; Ca; ExersiceInducedAngina; Thal; Old peak; slope; Thalac; Chestpain}
2. Clustering the Database: For the clustering here used the Improved K-Means so the use of object function one which uses the Euclidean distance formula:

$$d_{euc}(x, y) = \sum_{i=1}^{n} \sqrt{(xi - yi)^2} \quad \ldots\ldots. (1)$$

3. Classification:

For the classification here ID3 algorithm is used and its always work with the entropy calculation H(S) and by computing the information gain (IG).

$$H(S) = -\sum_{x \in X} p(x) log2 p(x) \quad \ldots\ldots\ldots (2)$$

This is used for the Entropy calculating.
Where,

- S= The current (data) set for which entropy is being calculated (changes every iteration of the ID3 algorithm)
- X = Set of classes in S.
- P(x) = the proportion of the number of elements in class to the number of elements in set.
- When H(S) = the set is perfectly classified (i.e. all elements are of the same class).

In ID3, entropy is calculated for each remaining attribute. The attribute with the smallest entropy is used to split the set on this iteration. The higher the entropy, the higher the potential to improve the classification here. Information gain IG (A) is the measure of the difference in entropy from before to after the set S is split on an attribute A. In other words, how much uncertainty in S was reduced after splitting set S on attribute A. Information Gain:

$$IG(A, S) = H(S) - \sum_{t \in T} p(t) H(t) \quad \ldots\ldots (3)$$

Where,

- H(s)= Entropy of set S
- T= The subsets created from splitting set S by attribute A such that S = Ut2T t
- P (t) = the proportion of the number of elements in to the number of elements in set S.
- H (t) = Entropy of subset t.

In ID3, information gain can be calculated (instead of entropy) for each remaining attribute. The attribute with the largest information gain is used to split the set on this iteration.

IV. DATASET USED IN THE SYSTEM

The Database which is used in the system in that 13 clinical parameters are used for the prediction of heart disease,
These are listed and described bellow in the Table 1:

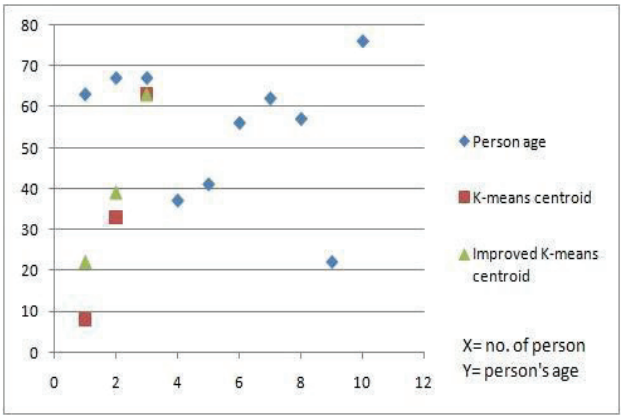
V. COMOPARISON

As using the improved k-means we can compare it with the simple k-means, basis on the centroids selection. Because of this in the improved k-means it will give the better time complexity and the good clustering. There are two graphs are shown below respectively for the parameters

Age and cholesterol those calculated using improved k-means. In that

Graphs Blue color shows the person related parameters, and Red color will show the Simple K-Means centroids, where as Green color will show the Improved K-Means centroids.

1. Over the parameter person's age.



Graph1: Representation of centroids difference for age parameter

In the graph X is representing the number of various people, and Y is representing the person's age.

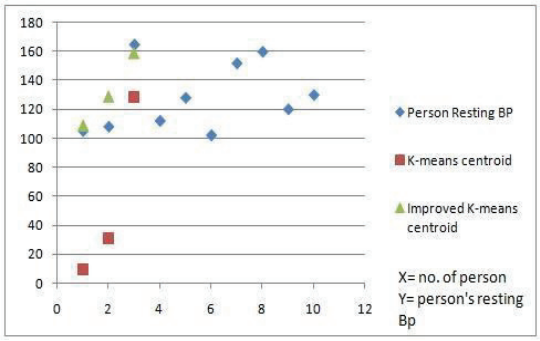
X	1	2	3	4	5	6	7	8	9	10
Y	63	67	67	37	41	56	62	57	22	70

Table 2: X, Y axis values on the graph

Sr.No	K-Means Centroid	Improved K-Means Centroid
1	22	28
2	33	39
3	63	63

Table 3: centroids difference over the parameter person's age

2. Over the parameter person's resting Bp.



Graph 2: Representation of centroids difference for resting Bp parameter

In the graph X is representing the number of various people, and Y is representing the person's Bp.

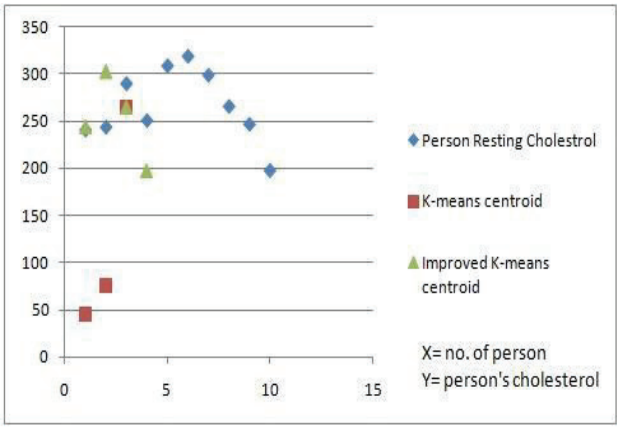
X	1	2	3	4	5	6	7	8	9	10
Y	105	108	165	112	128	102	152	160	120	130

Table 4: X, Y axis values on the graph

Sr.No	K-Means Centroid	Improved K-Means Centroid
1	10	109
2	31	129
3	128	159

Table 5: centroids difference over the parameter person's resting Bp

3. Over the parameter person's resting cholesterol.



Graph 3: Representation of centroids difference for cholesterol parameter

In the graph X is representing the number of various people, and Y is representing the person's cholesterol.

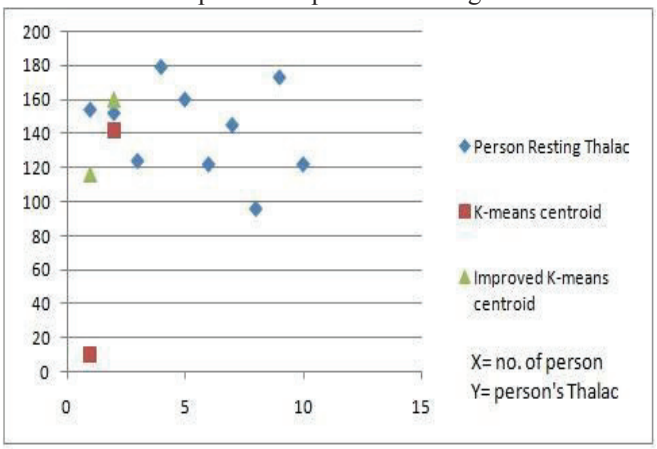
X	1	2	3	4	5	6	7	8	9	10
Y	240	243	289	250	308	318	298	265	246	197

Table 6: X, Y axis values on the graph

Sr.No	K-Means Centroid	Improved K-Means Centroid
1	45	244
2	75	303
3	265	265
4	-	197

Table 7: centroids difference over the parameter person's cholesterol

4. Over the parameter person's resting Thalac.



Graph 4: Representation of centroids difference for Thalac parameter

In the graph X is representing the number of various people, and Y is representing the person's Thalac.

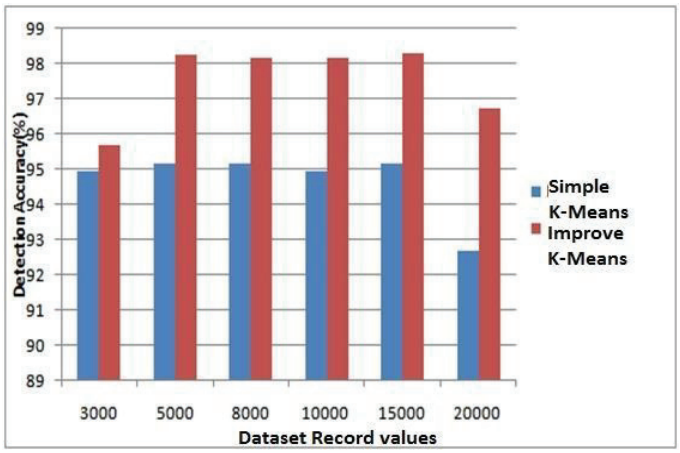
X	1	2	3	4	5	6	7	8	9	10
Y	154	152	124	179	160	122	145	96	173	122

Table 8: X, Y axis values on the graph

Sr.No	K-Means Centroid	Improved K-Means Centroid
1	10	116
2	142	160

Table 9: centroids difference over the parameter person's Thalac

Following graph is showing the accuracy of the Improved K-Means comparing to the simple K-Means algorithm.



Graph. 5: Graph for detection accuracy for different record size.

Different Size of records	K-Means Detection Accuracy (%)	Improved K-Means Detection Accuracy (%)
3000	94.97	95.70
5000	95.18	98.28
8000	95.18	98.18
10000	94.97	98.18
15000	95.18	98.28
20000	92.70	96.73

Table 10: Detection Accuracy for different record size

With comparing some parameters and their different size records, it's cleared that for centroid selection there is more accuracy in the Improved K-Means algorithm, and it shows less in the simple K-Means.

Conclusion

In such way I have learned about the big data and its properties, with its challenges and issues. In the medical field I learn about the various parameters those are affecting to the heart. Improved K-Means is the algorithm which is showing the accuracy in the centroid selection more than the simple K-Means.

Acknowledgment

I have a great pleasure to express my deep regards towards those who have offered their valuable time and guidance in our hour of need. I would like to express our sincere and whole hearted thanks to head of the department Prof. S. M. sangve sir and to my project guide Mr. P. M. Mane for contributing valuable time, knowledge, experience and providing valuable guidance. I am also glad to express my gratitude and thanks to our Principal Dr. A. N. Gaikwad for their constant inspiration and encouragement.

Sr. No	Parameters Name	Description
1.	Age	Age in years
2.	Gender	Gender(1=male, 0=female)
3.	Chest pain	Chest pain type Value 1:typical angina Value 2:typical angina Value 3:non-angnal pain Value 4:asymtomatic
4.	Resting Blood Pressure	Mm Hg on admission to hospital
5.	Cholesterol	Serum cholesterol in mg/dl
6.	Fasting Bs	(value 1: >120 mg/dl; value 0: < 120 mg/dl)
7.	Resting ECG	Resting electrographic result Value 0: normal Value 1: having ST-T wave abnormality Value 2: showing probable or definite left ventricular hypertrophy
8.	Thalac	Maximum heart rate achieved in number
9.	Exercise Induced Angina	Value 1: yes Value 2: No
10.	Oldpeak + slope	Oldpeak= ST depression induced by exercise relative to rest Slope= Slope of the peak exercise ST segment (Value 1: unslopig Value 2: flat Value 3: downsloping)
11.	Thal	Value 3: normal Value 6: fixed defeat Value 7: reversible defeat
12.	Ca	Number of major vesels(0-3) colored by flurosopy
13.	Diagnosis	Value 0: normal Value 1: <50% affected Value 2: >50% affected Value 3: <75% affected Value 4: 100% affected

Table 1: Smart Heart Disease diagnosis parameters se

References

[1] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Din, "*Data Mining With Big Data*",IEEE Transactions on Knowledge and Data Engineering, Vol. 26,No. 1, January 2014

[2] Ankita Dewan, Meghna Sharma, "*Prediction of Heart Dis- eases Using A Hybrid Technique In Data Mining Classifiscation*", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom).Jack Galilee, Dr. Ying Zhou, "*A Study on Implementing Iterative Algorithms Using Big Data Frameworks*", University of Sydney, School of Information Technologies,Faculty of Engineering and Information technologies,2014.

[3] Wullianallur Raghupathi, Viju Raghupathi, "*Big data analytics in healthcare: promise and potential*" , Raghupathi and Raghupathi, Health Information Science and Systems, 2:3, 2014.

[4] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, "*Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction*", International Journal of Computer Applications (0975,a 8887), Volume 17, a No.8, March 2011.

[5] XindongWu, Vipin Kumar,J.Ross, Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J, McLachlan, Angus Ng, Bing Liu, Philip S.Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, "*Top 10 algorithms in data mining*", Knowl Inf Syst (2008) 14:137 DOI 10.1007/s10115-007-0114-2

[6] Huang Xiuchang, SU Wei, "An Improved K-means Clustering Algorithm", JOURNAL OF NETWORKS, 161-167, VOL. 9, NO. 1, JANUARY 2014.

[7] Ritu Yadav, Anuradha Sharma, "*Advanced Methods to Improve Performance of K-Means Algorithm: A Review Clustering*", Global Journal of Computer Science and Technology Volume 12, Issue 9, Version 1.0, 46-52, April 2012.

[8] Anupama Chadha, Suresh Kumar,"*An Improved K-Means Clustering Algorithm: A Step For ward for Removal of Dependency on K*", 2014 International Conference on Reliability, Optimization and Information Technology -ICROIT 2014, India, Feb 6-8 2014

[9] Anand Bahety, "Extension and Evaluation of ID3- Decision Tree Algorithm",11-18, ICCCS, 2014, ICCC, 2014.

[10] Vikas Chaurasia, et al, Carib.j., Early Prediction of Heart Diseases Using Data Mining Techniques, SciTech,2013,Vol.1,208-217

[11] Tejaswini U. Mane, Mrs. Asha M. Pawar, "A Survey On Big Data And Its Mining Algorithm", IJIRCCE, Vol. 3, Issue 12, December 2015.

[12] Tejaswini U.Mane,Mrs.Asha M.Pawar, "Big Data Mining Platforms': A Survey", IJIRCCE, Vol. 4, Issue 6, June 2016.

[13] Ms. Tejaswini U. Mane, Mrs.A.M.Pawar, "*Big Data Mining: Problem, Protest and Explanation A Review*"