IJCTA, 9(27), 2016, pp. 321-327 ISSN: 0974-5572
© International Science Press

Efficient Clustering Techniques in Medical Diagnostics of Heart Diseases

Sampath Premkumar M.* and Hari Ganesh S.**

ABSTRACT

Data Mining (DM) techniques analyzes large datasets for exhibiting the underlying information that are useful and previously unknown. Clustering is one of most prevalent data mining techniques that groups objects of same type. The applications of clustering are enormous in fields like medical, business and education especially in medical diagnostics such as grouping of patients at same stage, having common diseases and so on. The objective of this paper is to survey the top clustering techniques that are appropriate for processing medicinal datasets and their limits and enhancements. Moreover, an experimental study has been presented to analyze the performance of traditional clustering algorithms in terms of time and accuracy of clustering. The results are interpreted and presented for motivating future researches.

Keywords: clustering, data mining, connectivity, centroid, density and distribution models.

1. INTRODUCTION

Clustering is categorized as one of the data descriptive analysis technique that builds clusters of data objects in such a way that objects in a cluster are closer to each other than the objects of other clusters [1]. Clustering techniques groups the objects of same characteristics and helps in analyzing the present state so as to predict the future state according to the dispersal of data objects. Clustering can be implemented in various domains like pattern recognition, artificial intelligence, statistics and neural networks.

The term Medical Data Mining explores the veiled exact patterns of datasets in the health care domain. Interpretations of medical data help in the structuring of patient treatment. But the real difficulties involved in analyzing the medical datasets are the close relevancy of data objects and the data generated by the medical centers are immense and complex which could potentially degrade the performances of clustering techniques. Hence, the challenge is to identify the best clustering algorithms that could always yield consistent results in terms of time and accuracy.

Heart disease is a dangerous cause of death in the past few decades as the food and working habits have been changed unvaryingly. Many researches have been proposed to help the professionals in the diagnosis of heart diseases. This paper surveys the clustering efforts that have been made so farin analyzing the heart disease along with their merits and demerits. Moreover, an comparative study has also been conducted to study the performance of traditional clustering models with heart diseases.

2. REVIEW OF LITERATURE

Shouman et al.[2] have presented a heart prediction system by collaborating Naïve Bayes and k-means clustering algorithm. The authors have initially clustered the training data using k-means algorithm which is then passed on to the naïve bayes algorithm for calculating the prior probability of the target attribute and conditional probability of the remaining attributes. The result of this process helps in predicting the outcome of test data. The authors have investigated the implementation of various methods for initial selection such as range, inlier, outlier, random attribute values and random row methods with k-means for the diagnosis of heart disease. Among them random attributes

* Research Scholar, Department of Computer Applications, Bishop Thorp College, Dharapuram-638657.

** Asst. Professor, Department Computer Science, H.H. The Rajah's College, Pudukottai-622 001.

and random row method have achieved higher accuracy with 84.5% than the other experimented methods. The authors have claimed that k-means suffers from initial centroid selection.

Wilson et al. [3] have used k-means and weighted association rule for eliminating the manual process for extracting the data directly from electronic records by transferring into a secure electronic system of medical records used for saving the lives of people and decreases the cost of the healthcare services. The authors have employed K-means inlier clustering through which the prediction is made through decision tree classifier. The results of the experiment have demonstrated that k-means with decision tree makes the prediction more accurate. The authors have stated the initial centroid problem infers the output.

Banu et al. [4] have presented an approach for fragmenting and extracting substantial forms from heart problem dataset for the prediction of heart attack. The authors have combined association rule mining, clustering and decision tree algorithms to analyze different kinds of heart problems. The authors have claimed that the accuracy of the proposed approach is up to 94%. The authors have intended to carry out the same task with real time health datasets.

Kumar [5] has stated that the heart disease prediction system suffers from the problem of missing data which may bias the results of prediction algorithms and implemented the Expectation Maximization algorithm as a preprocessing technique to calculate the missing values in the dataset. The author has alsoproposed a non-negative matrix factorisation with hierarchical clustering methods for extracting the significant patterns of the heart disease which would ease the prediction process. The author has claimed that proposed EM method is effective in finding out the missing values.

Kaur [6] has presented a recommender system for the prediction of heart disease using if then else association with fuzzy c means clustering and genetic clustering. Initially when the patient record is inputted to the system, an if then else rules are executed for predicting the two classes of patients such as positive or negative. The proposed system then predicts the internal status of the patient by applying fuzzy c means clustering along with the probability of crossover and mutation ratio. Finally, the accuracy, time, specificity and sensitivity of the prediction results are computed to evaluate the results. The author has proclaimed that the achieved accuracy of the proposed work is 86.6%. The author has also stated the accuracy of the system can further be improved.

3. EXPERIMENTATION

This paper presents experimentation on the performance of the traditional clustering models such as connectivity, centroid, distributive and density based over heart disease dataset. The dataset has been taken from UCI machine

Table 1
Heart Disease Dataset Description

S. No	*Attribute Name*	*Description*
1	Age	Patient's Age
2	Sex	Male=1; Female=0;
3	Cp	Chest Pain Type: Distinctive angina=1;Adistinctive angina=2; Non-anginal pain=3;Asymptomatic=4;
4	Trestbps	Resting blood pressure
5	Chol	Serum cholesterol
6	fbs	Fasting blood sugar>120 mg/dl: True=1;False=0;
7	Restecg	Resting electrocardiographic results: Normal=0;ST – T Wave anomaly =1;displaypossible or sure left ventricular hypertrophy by Estes' criteria=2;
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina: Yes=1;No=0;
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	slope of the peak exercise ST segment: upsloping=1;flat=2;downsloping=3;
12	ca	number of major vessels (0-3) colored by fluoroscopy
13	Thal	Normal=3;Fixed defect=6;Reversible defect=7;
14	Num	Diagnosis of the heart disease; Diameter narrowing<50%=0; Diameter narrowing>50%=1;

learning repository [7]. The dataset is created by the collaborated effort of Hungarian institute of cardiology, university hospital of zurich, university hospital of Basel, V.A medical center long beach and Cleveland clinic foundation. However the dataset which contains seventy six attributes with two ninety seven instances, only fourteen of the most important attributes including class attribute are taken for all published experiments so far. Mostly commonly the Cleveland dataset has been the top most dataset used by the ML researches. The class attribute of the dataset refers to the presence of heart problems in the patient. The value 0 in class attribute denotes the absence of heart problem and the values ranging from 1 to 4 denotes the severity of heart problem.The dataset is executed through a popular data mining tool Weka 3.6 for data analysis [8]. The data set has originally contained 303 instances with which the occurrence of instances with missing values are six which are discarded from the execution. The description of the dataset is given in table.1.

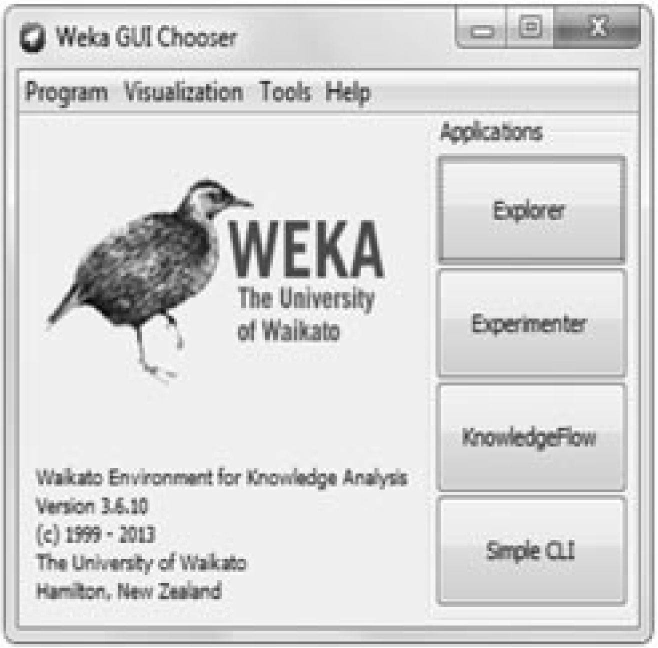
The step by step explanation of the experimentation is given as procedure in fig. 1.

1. *Start*
2. *Open Weka3.6 tool*
3. *Click on Explorer tab on Weka 3.6*
4. *Open the file tab on the explore window and choose the location of the heart disease dataset*
5. *Select Edit tab to view the instances of the dataset*
6. *Discard the instances that contains missing values*
7. *Save the modified dataset*
8. *Select All tab for choosing the attributes of the dataset*
9. *Click on the cluster tab for generating clusters of similar instances*
10. *Select the clustering algorithm through clusterer tab*
11. *Input number of clusters by clicking on the right text box on the clusterer tab*
12. *Click start tab for performing the clustering operation*
13. *Right click on the cluster result list to visualize the cluster assignments*
14. *Repeat steps 8 to 11 until all clustering algorithms are executed*
15. *Interpret the results of the cluster assignments*
16. *Stop*

Figure 1: Procedure for Weka Execution

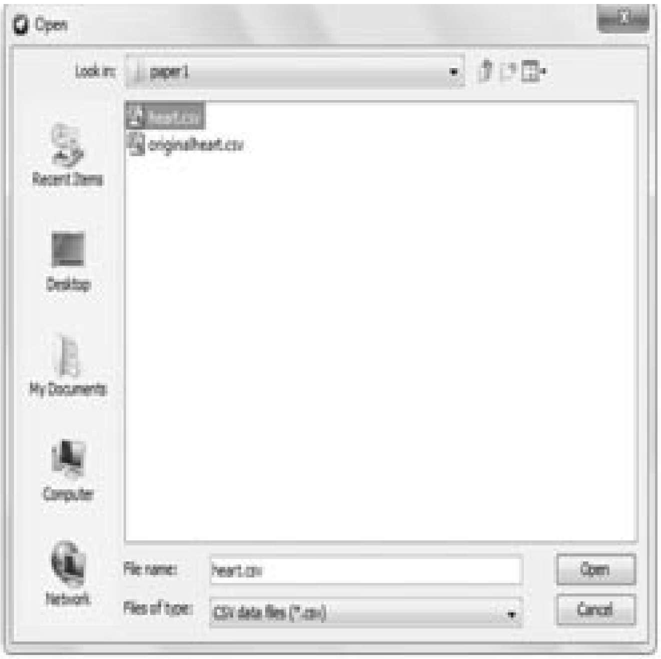
Fig.2.a. represents the procedures for opening explorer window on weka 3.6 tool. Fig.2.b. represents the inputting of heart disease on to the tool. Fig.2.c. denotes the attributes of the dataset that selected for clustering, finally fig.2.d.represents the selection of clustering algorithm for executing the dataset. Here hierarchical clustering algorithm is selected for execution connectivity model.

This paper encompasses the comparative analysis of four different types of clustering models such as connectivity, centroid, distributive and density. Hierarchical clustering is the classic example of connectivity model, k-means clustering is for centroid, Expectation Maximization(EM) is for distributive and Density Based Spatial Clustering of Applications tools with Noise (DBSCAN) is for density models respectively. Fig.3.a. represents the selection of linked connectivityclustering model for the execution heart dataset inclusive of number of groups and linkage types. As the number of major divisions of heart diseases is2, the numbers of clusters to be formed is also set to 2. Once when the selection of clustering algorithm and number of clusters are over, the algorithm is executed by clicking the start button.



Waikato Environment for Knowledge Analysis
Version 3.6.10
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

Figure 2: a) Weka 3.6 Explorer Window



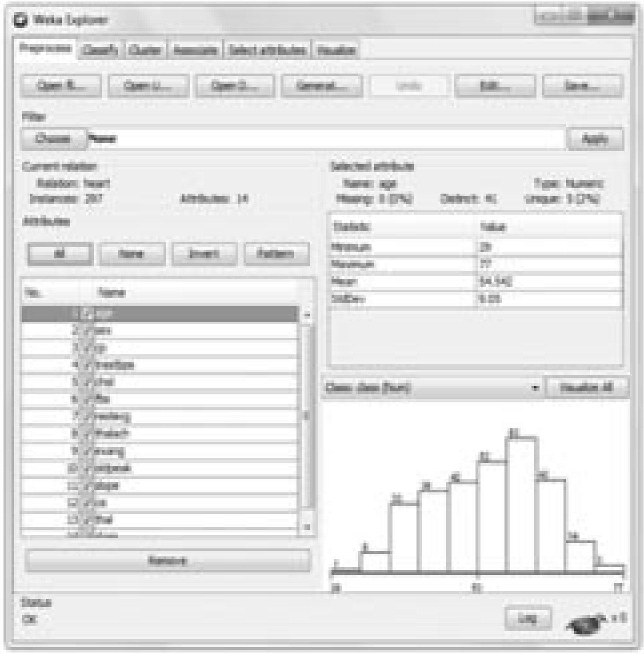
Look in: paper1

heart.csv
originalheart.csv

File name: heart.csv

Files of type: CSV data files (*.csv)

Figure 2: b) Loading of Heart disease Dataset


Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Choose
Apply
All None Invert Pattern
Remove
Visualize All
Log
x 0

Filter: None

Current relation
Relation: heart
Instances: 297
Attributes: 14

Selected attribute
Name: age
Type: Numeric
Missing: 0 (0%)
Distinct: 41
Unique: 5 (2%)

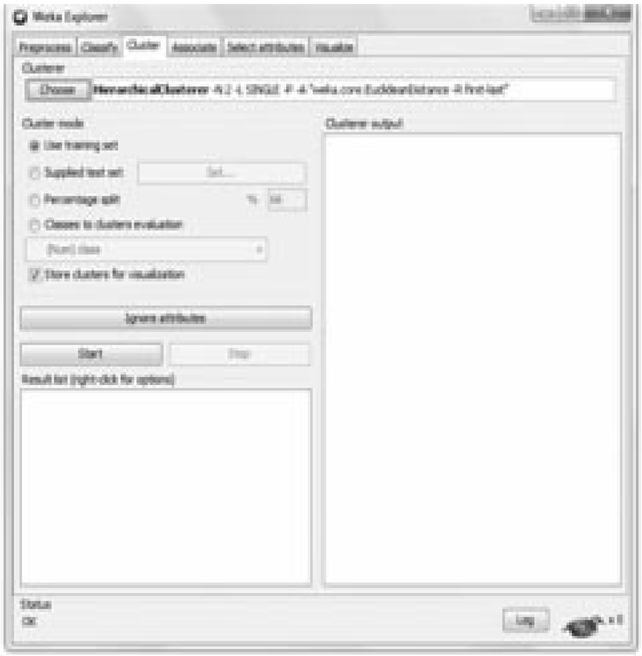
Statistic	Value
Minimum	29
Maximum	77
Mean	54.542
StdDev	9.05

No.	Name
1	age
2	sex
3	cp
4	trestbps
5	chol
6	fbs
7	restecg
8	thalach
9	exang
10	oldpeak
11	slope
12	ca
13	thal

Class: class (Num)

Status
OK

Figure 2: c) Selection of Attributes


Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Choose
Set...
Ignore attributes
Start
Stop
Log
x 0

Clusterer: HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

Cluster mode
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) class
☑ Store clusters for visualization

Result list (right-click for options)

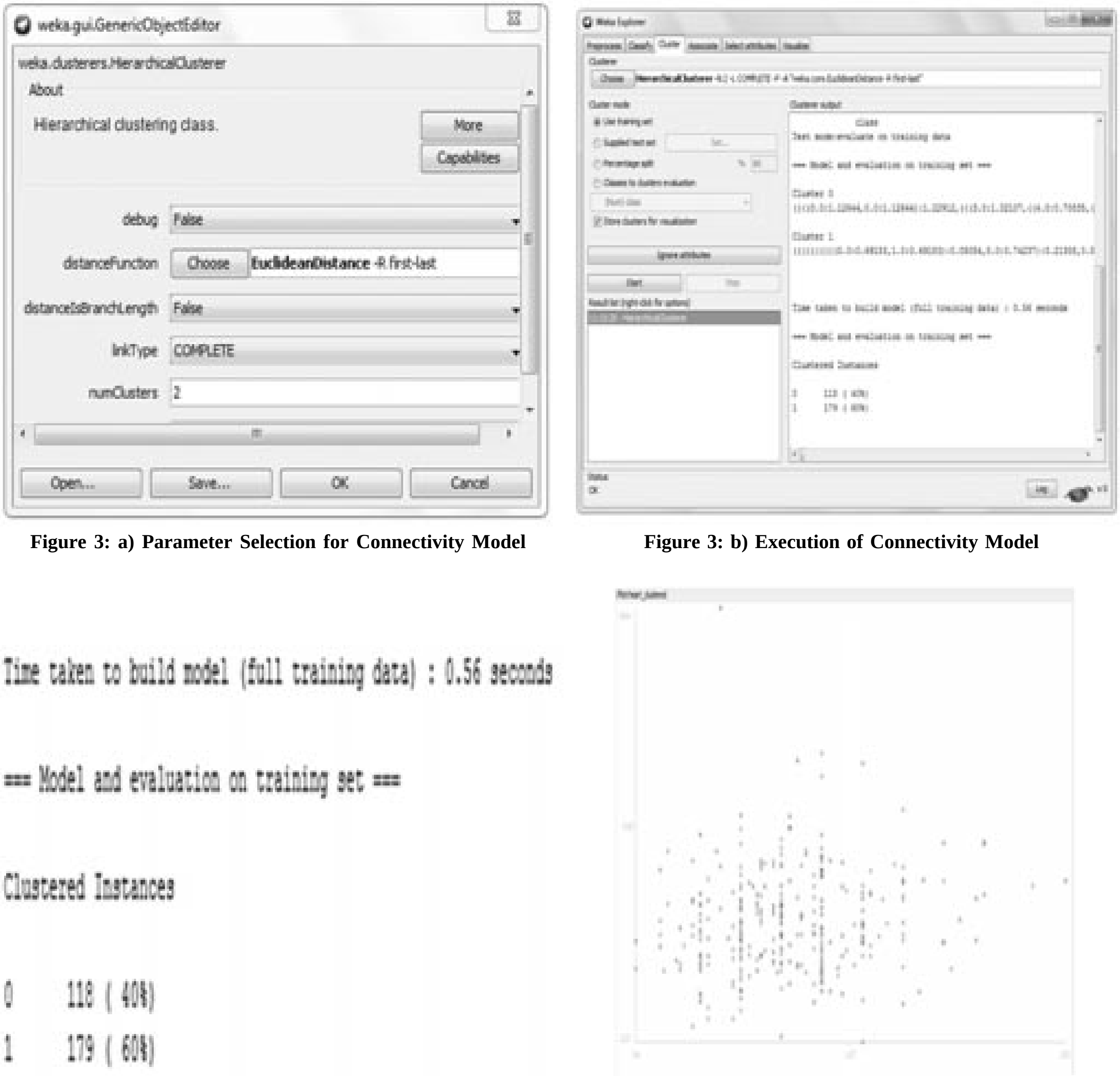
Clusterer output

Status
OK

Figure 2: d) Selection of Clustering Algorithm

Fig.3.b. shows the execution of connectivity clustering model. The results of connectivity clustering model are depicted in fig.3.c. ,which shows the time taken to build clusters and the percentage of cluster assignments along with the number of instances that are hold by each cluster. Fig.3.d. represents the visualization of cluster assignments which represented in x and y axis of selected attributes. The same process has been repeated for all four experimenting algorithms and the results are depicted in Table 2which consists of the time taken to build clusters of the experimenting algorithms and the accuracy obtained by each clustering in terms of random measure.

Accuracy of the cluster is calculated through the measure called random measure an external validation measure of clustering techniques that computes how similar the clusters are to benchmark classification [9]. Equ.1. denotes the standard notation of random measure.



weka.gui.GenericObjectEditor

weka.clusterers.HierarchicalClusterer

About

Hierarchical clustering class.

```
debug                  False
distanceFunction       EuclideanDistance -R first-last
distanceIsBranchLength False
linkType               COMPLETE
numClusters            2
```

Figure 3: a) Parameter Selection for Connectivity Model

```
=== Model and evaluation on training set ===

Cluster 0

Cluster 1

=== Model and evaluation on training set ===

Clustered Instances

0      118 ( 40%)
1      179 ( 60%)
```

Figure 3: b) Execution of Connectivity Model

```
Time taken to build model (full training data) : 0.56 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      118 ( 40%)
1      179 ( 60%)
```

Figure 3: c) Clustering Time and Alignment

Figure 3: d) Cluster Visualization of Connectivity Model

Table 2
Performance Analysis of Experimental Algorithms

Metric	*Connectivity Model*	*Centroid Model*	*Distributive Model*	*Density Model*
Time taken to build clusters (in seconds)	0.56 seconds	0.05 Seconds	0.73 Seconds	0.35 seconds
Accuracy (Rand Measure)	83%	86.7%	79.8%	75.2%

$$Random\ Measure = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

Where TP is True Positive i.e. sick people correctly identified as sick; TN is True Negative i.e. healthy people correctly identified as healthy; FP is False Positive i.e. healthy people incorrectly identified as sick; FN is False Negative i.e. sick people incorrectly identified as sick

4. RESULTS AND DISCUSSIONS

The pictorial representation of the time comparison analysis of clustering models is displayed in fig.4.with which the centroid model has taken minimum time limit than the other three models with 0.05 seconds.

The pictorial representation of the accuracy obtained by the experimental algorithms is shown in fig.5. The accuracy obtained by the centroid model is the highest than the other algorithms with 86.7%. From the results it is found from both the analysis of time and accuracy centroid model algorithm stands superior when compared to other clustering models.

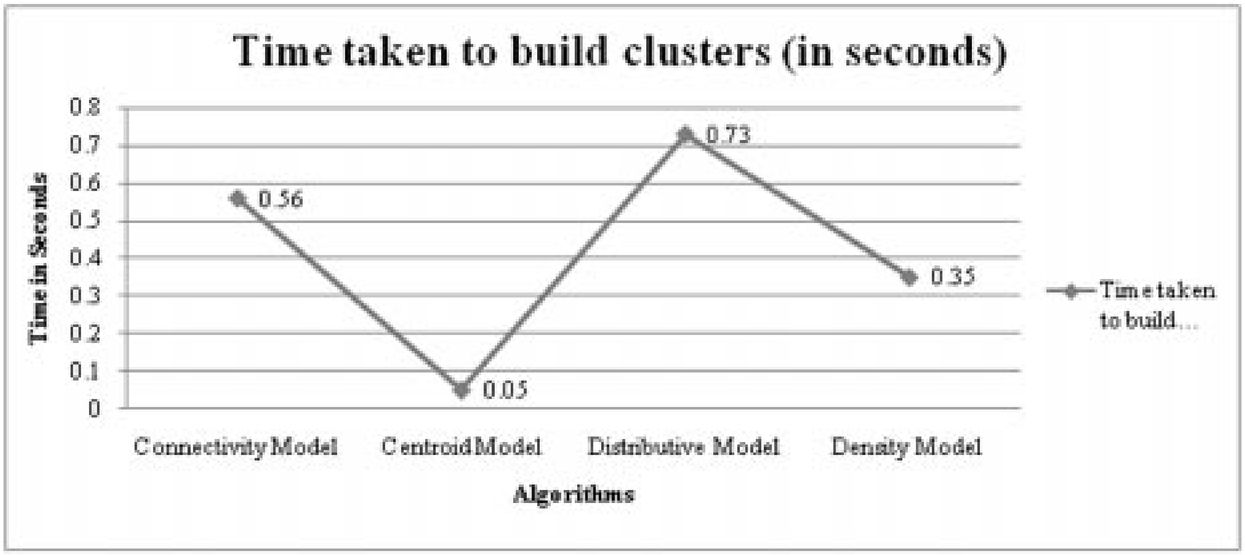


Figure 4: Time Comparison of Clustering Models

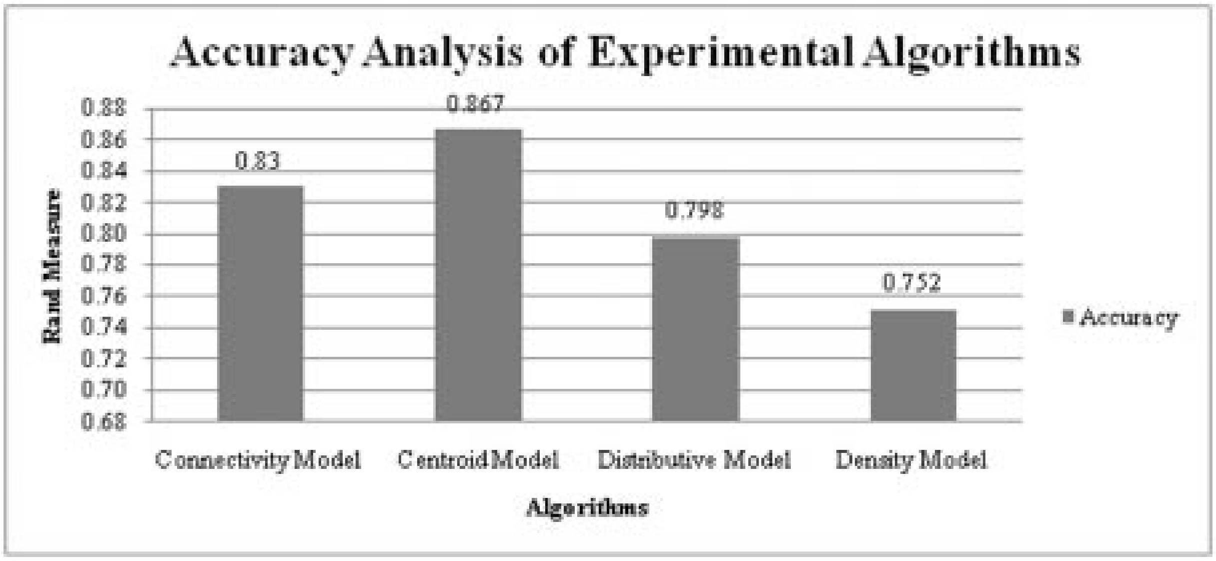


Figure 5: Accuracy Comparison of Clustering Models

5. CONCLUSION

This paper explains the performance on traditional clustering methods in inferring the heart disease. Clustering of heart diseases is highly a challengeable task for the traditional models as the accuracy of the algorithms is below 90% in all experimental algorithms. Moreover, the accuracy may still go lesser if the internal analysis of the severity of heart problems is interpreted as the values of the attributes shares a close commonality between the instances. Hence, there is should be an advanced mechanism that improves the accuracy of patient clustering with enhanced random measure along with the analysis of severity of heart problems.

REFERENCES

[1] A. Joy Christy, S. Hari Ganesh, "Building Numerical Clusters Using Multidimensional Spherical Equation", International Journal of Applied Engineering Research, ISSN 0973-4562, Volume 10, Issue No.82, pp:629-634, 2015.

[2] Shouman, Mai, Tim Turner, and Rob Stocker. "Integrating Naive Bayes and K-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients." CS & IT-CSCP, pp: 125-137, 2012.

[3] Wilson, Aswathy, and Gloria Wilson. "Heart disease prediction using the data mining techniques."International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 1, pp: 84-89, 2014.

[4] Banu, MA Nishara, and B. Gomathy. "Disease Predicting System Using Data Mining Techniques." International Journal of Technical Research and Applications 1.5, pp: 41-45, 2013.

[5] Pandey, Atul Kumar, et al. "DataMining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method." heart disease 14, pp: 16-17, 2013.

[6] LovepreetKaur, "Predicting Heart Disease Symptoms using Fuzzy C-Means Clustering", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 3 Issue 12, pp: 4232-4236, 2014.

[7] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[8] www.cs.waikato.ac.nz/ml/weka/

[9] https://en.wikipedia.org/wiki/Cluster_analysis