Original Research Article Bioinformatics

International Journal of Pharma and Bio Sciences

ISSN
0975-6299

A PREDICTIVE MODEL FOR HEART DISEASE USING CLUSTERING TECHNIQUES

A.SOWMITH[*1], V.SUCHARITA[2], P.SOWJANYA[3] AND B.GEETHA KRISHNA[4]

[1,2,3,4]Department of Computer Science,K L University, Guntur,Andhra Pradesh, India.

ABSTRACT

Data mining is the area of computer and information science with large perspective of knowledge discovery from large database. Now a days people are dependent on furious timetable and garbage sustenance which impact the heart basically. Prediction of Heart diseasae utilizing different clustering algorithms are implemented in this paper. Grouping of data into related groups is known as clustering.In this work we have implemented K-Means, Hierarchical method and DBSCAN.And compared the results which are genereated by the above algorithms and declaring which is the best algorithm for the prediction of the heart disease.

KEYWORDS: *Clustering, Heart Disease, Data Mining, K Means, Hierarchical, DBSCAN.*

A.SOWMITH*
Department of Computer Science,K L University, Guntur,Andhra Pradesh, India.

Received on : 1/25/2017
Revised and Accepted on : 6/13/2017
DOI: http://dx.doi.org/10.22376/ijpbs.2017.8.3.b529-534

INTRODUCTION

In prosperity organizations, data information mining is more prominent and key for all the restorative applications. It contains different information, however these information have not been utilized. This information will be changed over into the some huge one by utilizing data mining systems. Data mining in remedial organizations is a making field of high importance for desires and a more critical impression of restorative information. Restorative administrations Data mining endeavors to deal with certifiable success issues in the acknowledgment and treatment of surgeries. Specialists are utilizing data mining methodology in the healing assurance of a couple ailment, for example, diabetes, stroke, risk, and coronary issues[12]. Coronary sickness is a run of the mill name for a wide course of action of torments, issue and conditions that effect the heart and once in a while the veins as well. Bigger piece of men and women pass on due to heart diseases. Disarranges of the Coronary issues depend on kind of coronary sickness. Fundamental Syndrome is Chest pain.The data mining is the course toward finding the secured information from the data base or some other information storage facilities. The basic motivation driving the prosperity business is to overhauling the method for human organizations information by decreasing the missing qualities and expelling the commotion in the data base[1]. A few data burrowing systems are used for examination of heart diseases, for example, guileless bayes, decision tree, and neural structure, divide thickness, pressing estimation, k-mean gathering and reinforce vector machine showing unmistakable levels of exactness. K-means is the most prominently used gathering strategy. It is used as a piece of various applications thus of its basic and all around grounded frameworks. Several inspectors have seen that age, heartbeat and cholesterol are fundamental hazard parts related with coronary malady.Clustering algorithms are fundamentally unsubstantiated learning process. Cluster analysis is an element which is used to study the internal structure of a dataset, which can be defined through mean and co-variance[2,3]. The drawback of clustering is it divides the data set into parts, so that each element within cluster is similar to other cluster[4]. The popular clustering techniques are: hierarchical clustering techniques and partitioning clustering techniques. Various clustering techniques are alienated into independent classes, such as density based method, grid based methods, etc. Some methods are described below.

METHODOLOGY

The heart disease data set is applied on different algorithms of clustering like k-means,agloremative and DBSCAN. And results are compared based on the number of clusters, cluster instances , unclustered instances and time taken by the algorithm.And declaring which is the suitable algorithm for the prediction of the heart disease.

K-Means Clustering

K Means clustering main objective is to divide n observations into k clusters. K Means clustering is as shown in the figure 1.The finest form of cluster analysis is partitioning technique, which partitions the substantial dataset into clusters. It appoints points into predefined clusters, and repeats the procedure for the improvement of clusters[5]. The least complex calculation is K-means. The k-means calculation is portrayed as takes after:

1) Randomly select cluster centers.
2) Calculate the distance between data point and center of cluster.
3) Calculate new center by taking mean value of object.
4) Repeat 2 & 3 until mean errors converge.

The K-means algorithm is used when mean is declared [6]. The drawback of k-means there is rejection respond for calculating the less number of clusters. One approach is to evaluate the multiple runs with various clusters[7]. *K*-means minimizes within-cluster point scatter as shown in equation below:

$$W(C)=\frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(j)=k}\left\|x_i-x_j\right\|^2=\sum_{k=1}^{K}N_k\sum_{C(i)=k}\left\|x_i-m_k\right\|^2$$

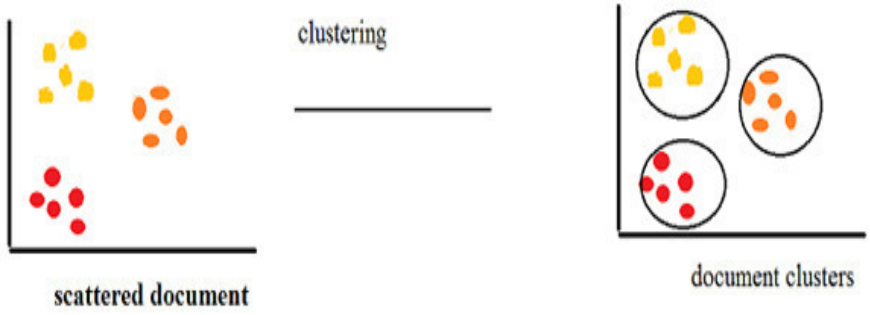

Figure 1
K-Means Clustering

Hierarchical Method

Tree like arrangements can be observed in the hierarchical clustering and each cluster will represent data nodes or demonstration of small level cluster. This is grouped into two classes: agglomerative and divisive. The implementation of Hierarchical clustering is shown in figure 2.

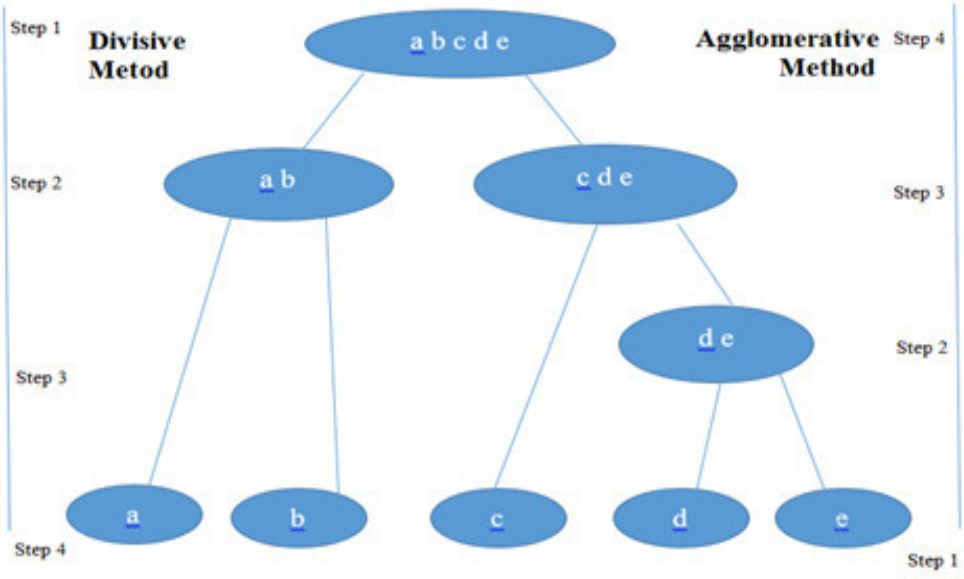

Figure 2
Hierarchical Clustering

Agglomerative: It takes bottom-up approach. It begins from every entity which creates its own particular cluster and it blends group into larger clusters. This process is repeated through multiple iterations until all entities are named under some cluster. The single group gets to be base of order. For combining, entities discoveries the group which is nearest to it, and consolidates the two as one group. Divisive: it works inverse to agglomerative strategy i.e. top-bottom approach. In this strategy, all items put into single cluster. At that point it separates parent cluster into a few little clusters, and by means of repeated procedures the root cluster isolate into little child cluster. It also guarantees that procedure will proceed until every cluster at most reduced level can never again be isolated[11].

Density-based method

By using hierarchical and partition methods, it is easy to find spherical-shaped cluster[8]. But it is very difficult to recognize the arbitrary fashioned cluster such as the oval shape and “S” shape clusters. By utilizing density based technique, it is so simple to discover irregular molded clusters, so we can configure clusters as dense region in data space, isolated by meager locales[9]. The vital part of density based is that to discover other than round shape cluster DBSCAN (Density-based Spatial Clustering of Application with Noise) might have been recommended that entry thickness connectivity to taking care of the irregular formed group. DENCLUE (Density-based clustering) is an dissemination built algorithm, which fill in viably with respect to huge dataset which hold large amount noise, Anyhow different way, it meets expectations for secondary velocity contrast with DBSCAN, other than this, it holds vast number for parameters, thus it may be finer during Contributing those irregular state clusters, Yet because of non-linear complexity, it could pertinent best around little or average level datasets[10].

RESULTS

Table 1 shows the method of performing K Means, Hierarchical clustering(AGNES) and DBSCAN. Figure 3 represents implementation of the K Means clustering algorithm. The output of K-means is shown in figure 4, The implementation of the Hierarchical clustering is shown in figure 5. Figure 6 represents the implementation of the DBSCAN.

Table 1
Similarity between the different Clustering algorithms

Fields	Partition method	Hierarchical method	Density-based method
Main	In this we utilize mean or mediods to characterize cluster	Hierarchical implies tree like structure	It characterizes separates between closest focuses
functioning phase	Find usually solely Shape cluster. Through multiple iteration every entity needs to discover its cluster	Here both agglomerative and divisive strategy are used.	Clusters are hard of every object that are separated by low-density regions.
Pros	For this technique information is not required. It is strong straightforward.	In this we don't require what number of clusters are required and input parameters.	It finds arbitrary shape cluster.
Cons	Fixed number cluster will make hard to break down the clusters.	May not adjust well. No computerization to discover best cluster	The estimation of calculation relies on upon the separation measure. Can't work effectively with huge data set.

These techniques based on the performance analysis give the information of the patient is suffering from heart disease or not depending on the data set. So that the results generated using K Means clustering. The accuracy of K Means is 39.2%, 3.7%, and 31.3% respectively.

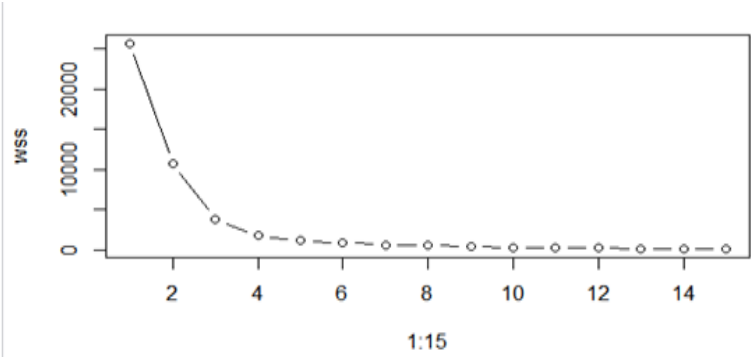

Figure 3
Implementation of K-Means algorithm

```
K-means clustering with 3 clusters of sizes 81, 90, 38

Cluster means:
  rest_bpress max_heart_rate
1    131.0123       161.4938
2    127.7111       129.2222
3    153.3947       106.3684

Clustering vector:
  [1] 2 1 1 1 2 2 2 3 2 3 1 2 1 2 1 2 1 2 2 3 3 1 2 2 2 2 2 1 3 2 2 3 3 1 2 2 1
 [38] 1 2 2 3 1 3 1 1 2 2 2 2 1 2 1 2 3 2 1 1 2 2 1 1 2 2 2 1 3 2 1 1 2 2 1 2 1
 [75] 1 1 1 1 2 3 1 1 1 1 2 3 1 3 2 2 2 2 1 1 3 1 1 1 1 3 3 1 3 1 1 2 3 1 2 3 1
[112] 2 2 1 2 1 1 1 2 3 2 2 2 1 1 1 2 2 1 3 2 1 2 2 2 3 1 2 2 2 1 1 1 2 1 1 3 2
[149] 2 2 3 3 2 2 2 2 1 3 2 2 1 3 3 2 2 1 1 3 3 1 2 1 1 2 1 1 2 2 2 2 2 2 3 1 1
[186] 1 3 2 1 2 1 3 3 2 3 1 2 2 1 3 2 1 2 1 3 2 2 1 1

Within cluster sum of squares by cluster:
[1] 29963.23 24304.04 19347.92
 (between_SS / total_SS =  59.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
> |
```

Figure 4
K-Means Output

The accuracy of Hierarchical clustering is 28%, 23%, 31%, 12%, 7% respectively.

```
Input=read.csv(file.choose())
D<-dist(input,method="Euclidean")
Fit<-hclust(d,method="ward.D")
Groups<-cutree(fit,k=5)
Rect.hclust(fit,k=5,border="red")
Plot(fit)
```

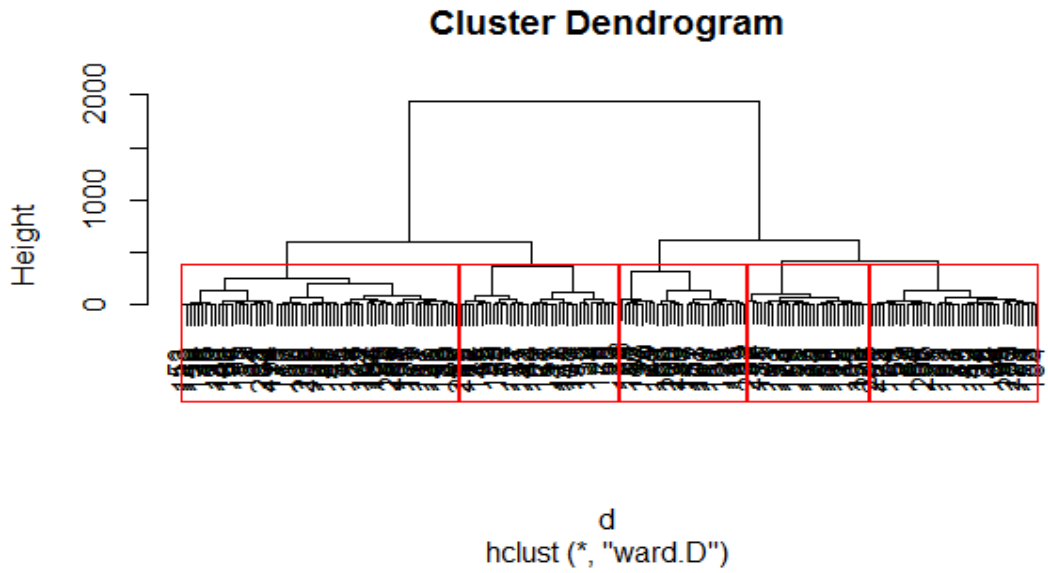

Figure 5
Implementation of Hierarchal Clustering

	82	87	91	92	94	96	98	99	100	102	103	105	106	110	112	114	115	116	118	119	120	121
0	1	1	1	3	2	3	2	2	6	1	1	1	1	6	3	1	5	2	4	1	7	1

	122	124	125	126	128	130	132	134	135	136	137	138	140	142	143	144	145	146	148	150
0	5	3	7	1	2	10	2	4	5	2	3	4	16	4	1	2	4	2	1	21

	152	153	154	155	158	160	162	165	166	167	168	170	172	174	175	178	180	184	185	188
0	2	1	1	5	2	9	1	2	1	1	3	10	3	2	3	1	3	2	2	1

Figure 6
Implementation of Density Based Clustering

Table 2
Results for comparison of clustering techniques

Algorithm	No. of Clusters	Cluster Instances	Un clustered Instances	Time build
K Means	3	39.2 % , 3.7% , 31.3%	0	0.05
Hierarchical Clustering AGNES	5	28% , 23% , 31% , 12% , 7%	0	0.52
DBSCAN	2	98% , 2%	39	0.98

The accuracy of DBSCAN clustering is 98%, 2% respectively. Among all these three K Means perform better when compared with Hierarchical clustering and DBSCAN as shown in table2.

```
Library(fpc)
Ds < - dbscan(input,eps=0.42)
Table(ds$cluster,input$max_heart_rate)
```

CONCLUSION

There are various more techniques that can be implemented on the heart disease data set. In this research paper in the heart disease prediction is done using K-Means, DBSCAN and AGNES algorithms but K-Means clustering is used mainly for improving the efficiency. In our findings based upon the cluster instances we recommend that K-means is better. Therefore K-means is the efficient method of clustering for predicting patients suffering from heart disease. This model gives solution for any kind of query with accuracy and detail information. This paper can be extended to compare more clustering techniques.

ACKNOWLEDGEMENTS

The authors also thankful for the management of KL University , Guntur , A.P , for their support and encouraging this work by providing the facilities in KERG in Computer Science and Engineering.

CONFLICT OF INTEREST

Conflict of interest declared none.

REFERENCES

1. P.Venkateswara Rao and A.Ramamohan Reddy , Shrimp disease detection using Back Propagation Neural networks, International Journal of Pharma and Bio Sciences. 2016;7:3.
2. A. K. Jain and R. C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ. 1988;3:33-6.
3. R C Dubes and A K Jain, Algorithms for Clustering Data, Prentice Hall,1988. p.181.
4. R Xu and D Wunsch, Survey of Clustering Algorithms, IEEE Transactions on Neural Networks, 2005;16:3.
5. Chonghui GUO and Li PENG, A Hybrid Clustering Algorithm Based on Dimensional Reduction and K-Harmonic Means, IEEE. 2008;145(1):68-71.
6. Liang Sun and Shinichi Yoshida, A Novel Support Vector and K-Means based Hybrid Clustering Algorithm, Proceedings of the IEEE International Conference on Information and Automation .2010 ;64(4):20-23..
7. B Zhang, M C Hsu and UmeshwarDayal, K-Harmonic Means-A Data Clustering Algorithm, HPL .1999;81(4):31-4.
8. Suresh Chandra Satapathy, JVR Murthy and Prasada Reddy P.V.G.D, An Efficient Hybrid Algorithm for Data Clustering Us1ing Improved Genetic Algorithm and Nelder Mead Simplex Search, International Conference on Computational Intelligence and Multimedia Applications. 2007;85(4):24-7.
9. R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, 2nd ed., John Wiley & Sons Inc., 2001.
10. A.Y. Ng, M.I. Jordan, and Y. Weiss, On spectral clustering: Analysis and an algorithm, In T. G. Dietterich et al., eds., Proc.of NIPS 14. 2002;44(10):57-9.
11. S.R.Pande, S.S.Sambare and V.M Thakre, Data Clustering using data mining techniques, International journal of advance research in computer and communication engineering. 2012; 142(4):87-91.
12. M.Akhil jabbers, Dr.Prirti Chandra and Dr.B.L Deekshatulu, Heart disease prediction system using associative classification and genetic algorithm, International conference on emerging trends in electrical, Electronics and communication technologies.2012;25(1):78-90.

Reviewers of this article

DR.V.VIJAYARAJA B.E,M.Tech,Ph,D

Professor in CSE Department.
K.L University,Green Field,
Vaddesswaram.
Gundur District.

Asst.Prof.Dr. Sujata Bhattacharya

Assistant Professor, School of Biological and Environmental Sciences, Shoolini University, Solan (HP)-173212, India.

Prof.Dr.K.Suriaprabha

Asst. Editor , International Journal of Pharma and Bio sciences.

Prof.P.Muthuprasanna

Managing Editor , International Journal of Pharma and Bio sciences.

We sincerely thank the above reviewers for peer reviewing the manuscript