A Comprehensive Study of Clustering Algorithms to Analyze Medical Datasets

1 Madhuri Potnuru, [2]Dr G Lavanya Devi, [3]N Naresh

1. Dept of Computer Science and Sytems Engineering, AU College of Engineering, Visakhapatnam
madhuripotnuru248@gmail.com
2. Assistant Professor, Dept of Computer Science and Systems Engineering, AU College of Engineering, Visakhapatnam, lavanyadevig@yahoo.co.in
3. Research Scholar, Dept of Computer Science and Systems Engineering, AU College of Engineering, Visakhapatnam
naresh855@gmail.com

Abstract:

The objective of this research work is focused on the ethical cluster creation of heart disease data and analyzed the performance of partition based algorithms. This research work would help the doctors to identify the stages of heart disease and also enhances the medical care. One of the most difficult jobs in medicine is to diagnose a disease. The recognition of heart disease from diverse features or signs is a major issue which is not free from false presumptions often accompanied by unpredictable effects. Unfortunately, the huge amount of data about the heart diseases provided by the health care industries are not useful to give information for effective diagnosing. Increase in these stats data which will be a look for the researchers to dig into these medical databases for useful information. As there is an increase in the volume of stored data, as well as to find the patterns and to extract the knowledge for providing better patient care and to provide effective capabilities for diagnosis, this can be done using the data mining techniques. Predictions for the Heart disease goes wrong highly due to missing of data, due to which stats goes wrong which results in approximate results, which are ineffective in diagnostic procedures. Imputation is one the solution for this problem. This imputation method will help to replace the missing attributes from the datasets by the 13 medical attributes which are from the Cleveland heart disease database. Most of the researchers analyzed the heart disease dataset using algorithms to find the cluster among the small cell or non-small cell heart disease in various stages. The very famous two partition based algorithms namely K-Means and Hierarchical clustering are implemented. A comparative analysis of clustering algorithms is also carried out using two different datasets. Data clustering is a process which is a collection of similar data considering them as a group. A clustering algorithm divides a data set into several groups such that the similarity within a group is larger than among groups. This thesis analyze three types of clustering techniques- k-means, hierarchical and dbscan clustering algorithms. The performance and various other attributes of the three techniques are presented and compared.

Keywords — Data clustering, K-Means Clustering, Hierarchical Clustering, Agglomerative clustering, Data Mining, Heart diseases, Attribute Selection.

1.Introduction

Health Care Data mining is an emerging field which provides medical data for deeper understanding and for prognosis. It attempts to solve real world health problems in the diagnosis and treatment of diseases [1]. Several diseases like diabetes [2], stroke [3], cancer [4], and heart disease [5] are being diagnosed by the Researchers using data mining techniques. Heart disease is a general name for a wide variety of the diseases, disorders and conditions that affect the heart and sometimes the blood vessels as well. Symptoms of heart disease vary depending on the specific type of heart disease. A classic symptoms of heart disease is chest pain. However, with some forms of heart disease, such as atherosclerosis, there may be no symptoms in some people until life-threatening complications develop. There are number of conditions that can affect the heart. The data mining is the process of finding the hidden knowledge from the data base or any other information repositories. The main purpose of the health care industry is to improving the quality of healthcare data by reducing the missing values and removing the noise in the data base. Several data mining techniques are used in the diagnosis of heart disease such as naive bayes, decision tree, and neural network, kernel density, bagging algorithm, kmean clustering, support vector machine and agglomerative hierarchical clustering showing different levels of accuracies. Clustering is a data mining technique to group the similar data into a cluster and dissimilar data into different clusters. Clustering is the unsupervised classification of patterns into groups. A clustering algorithm divides a data set into several groups such that the similarity within a group is larger than among groups. Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings. Clustering is the process of arranging a similar member of objects into groups. A cluster is a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. For the compression of data and to construct a model, the data need to be organized and categorized which can be done extensively using clustering algorithms. Clustering methods can be broadly classified as Hierarchical clustering and Partition (nonhierarchical) Clustering. A Hierarchical clustering is a nested sequence of partitions. This method is further improved for the quality, as agglomerative hierarchical clustering (Bottom-up) approach and divisive hierarchical clustering (Top-down) approach. Starting from an initial partitioning, instances are relocated from one cluster to another cluster by these partitioning methods. Such methods typically require the number of clusters which will be pre-set by the user. In order to achieve global optimality in partitioned based clustering, considering all aspects of all possible partitions is required. Reason is this is not realistic, many greedy heuristics are used in the form of iterative optimization. Namely, a relocation method iteratively relocates points between the k clusters. DBSCAN has a crucial place in finding non-linear shapes structure on the basis of density. Density-Based Spatial Clustering of Applications with Noise is the most widely used density based algorithm. It uses the concept of density reachability and density connectivity. Comparing the performance of the DBSCAN algorithm with a proven segmentation algorithm that utilizes k-means clustering demonstrated that the DBSCAN algorithm had a higher sensitivity and correctly segmented more swallows. When comparing its performance with a threshold-based algorithm that used the quadratic variation of the signals showed, that the DBSCAN algorithm offered no rapid increase in performance.

2. K-MEANS Clustering

In the K-means algorithm, it randomly selects K initial centroids where K is a user defined number of desired clusters (any positive number). Each point is then assigned to a closest centroid and the collection of points close to a centroid form a cluster. This leads to partitioning of the data space into Verona cells. K-Means (Macqueen, 1967) is one of the simplest unsupervised learning algorithms that solved many clustering problems. In this procedure given data set is group into certain number of clusters (assume k Clusters). The main idea is to define k centroids, one for each cluster. The challenging task is, that these centroids must be positioned in an artful way because of different location causes different result. So, the good choice is to place them as much as possible far away from each other. Further step is to take each point belonging to a given data set and united it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. Recalculation has to be done for k new centroid as bar centers of the clusters resulting from the previous step. Once we have these k new centroids, another binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. There are k number of centroids which can change their location step by step until there are no more changes to be done and this can be noticed as a resultant of this loop. In other words centroids do not move any more. The K-Means algorithm is one of the simplest unsupervised learning algorithms that answer the well-known clustering problem. In this procedure given data set is group into certain number of clusters (assume k Clusters). The K-Means algorithm can be run multiple times to decrease the complexity of grouping data. The K-Means is a simple algorithm that has been modified to many problem areas and it is a noble candidate to work for a randomly generated data points.

This algorithm is flows as:

Step 1: Allocate k points into the space represented by the objects that are being clustered. These points represent initial group centroids.

Step 2: Allocate each item to the group that has the closest centroid.

Step 3: Once all objects items have been allocated, now recheck the positions of the k centroids.

Step 4: Continue to repeat Steps 2 and 3 until the centroids no longer move.

The algorithm is also significantly sensitive to the initial randomly selected cluster centers. This is proved by more than a few times in recent as well as in the past research; recurring problem has to do with the initialization of the algorithm. The K- Means has been adapted in many domains for ease.

3. Hierarchical Clustering

Hierarchical Clustering Analysis also called HCA is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for Hierarchical clustering generally fall into two types: Agglomerative Hierarchical clustering and Divisive Hierarchical clustering.

a) Agglomerative Hierarchical Clustering

This algorithm produces sequence of clustering of decreasing number of clusters at each step. The clusters produced at each step results from the previous step, by merging two clusters into one. The clusters are merged by computing the distance between each pair of clusters. For n samples, agglomerative algorithms [1] begin with n clusters and each cluster contains a single sample or a point. Then two clusters will merge so that the similarity between them is the closest until the number of clusters becomes 1 or as specified by the user [4] [7] [14].

The algorithm is composed of the following steps:

1. Start with n clusters, and a single sample indicates one cluster.
2. Find the most similar clusters Ci and Cj then merge them into one cluster.
3. Repeat step 2 until the number of cluster becomes one or as specified by the user. The pair of clusters are computed considering their distances, the pair with less distance between have the highest chances of merging, choose such a pair.

There are several ways to calculate the distances between the clusters Ci and Cj.

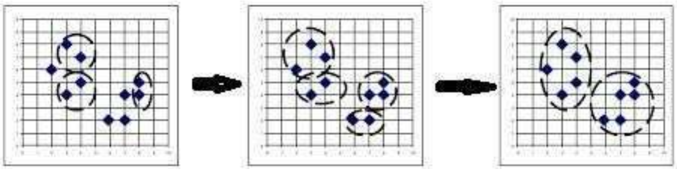

Figure 1: Agglomerative Hierarchical Clustering

b) Divisive Hierarchical Clustering

Divisive Hierarchical clustering algorithm acts quite reverse to Agglomerative Hierarchical clustering that is, this method produces clustering sequence of increasing number of cluster at each step. The clusters produced at each step results from the previous step by splitting a single cluster into two. Divisive algorithms start off with only one cluster which contains all sample data. Then, the single cluster splits into 2 or more clusters that have higher dissimilarity between them until the number of clusters becomes number of samples or as specified by the user. The following algorithm is one kind of divisive algorithms using splinter party method.

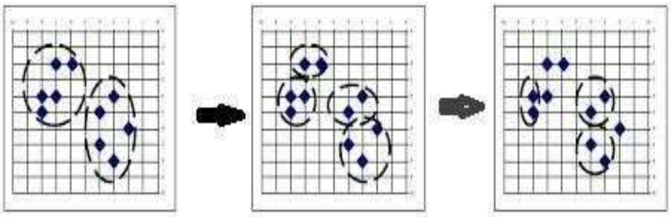

Figure 2: Divisive Hierarchical Clustering

4. DBScan Clustering

DBSCAN was proposed by Martin Ester et al in 1996. It is one of the most common clustering algorithms [8]. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. This algorithm is based on connecting points within certain distance limits similar to linkage based clustering. However, it only connects points that satisfy a density criterion (minimum number of objects within radius). An arbitrary shape cluster is formed which consists of all density connected objects. DBSCAN classifies data points into three categories.

- Hub points: Points that are at the interior of a cluster (Centre).
- Edge points: Falls within the neighborhood of a hub point which is not a hub point.
- Noise points: Any point that is not a hub point or an edge point.

For detecting a cluster, DBSCAN starts retrieving all instance of the data set (D) including an arbitrary instance (p) with respect to epsilon (Eps) and minimum points (minPts). minPts defined as the minimum number of points required to exist in a neighborhood to be declared a cluster, and Eps defined as the radius of the neighborhood of a point based on a distance (Euclidean, Manhattan or Minkowski) metric. To locate points withi the core points of the

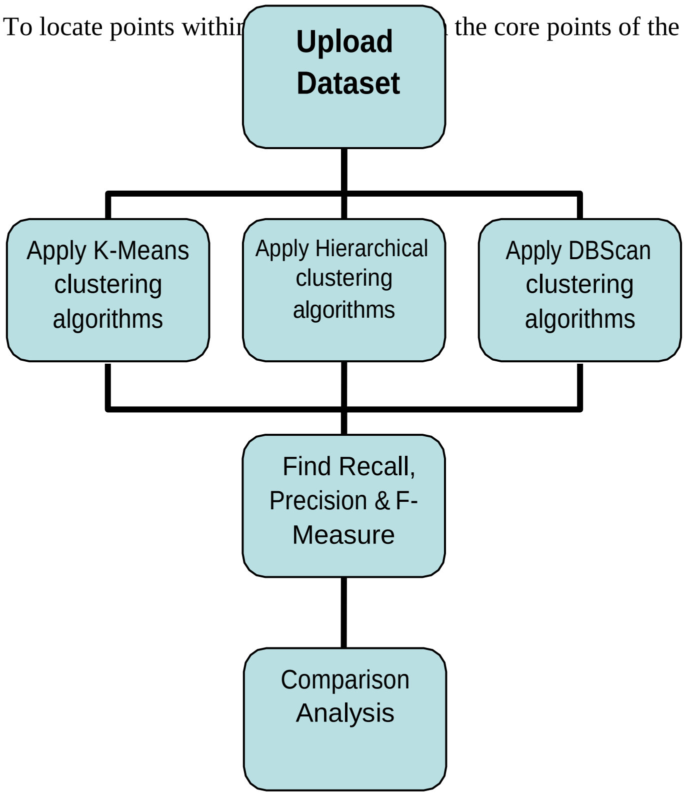
Upload Dataset
Apply K-Means clustering algorithms
Apply Hierarchical clustering algorithms
Apply DBScan clustering algorithms
Find Recall, Precision & F-Measure
Comparison Analysis

Figure 3: Flowchart of Comparison Analysis.

5. Comparison Analysis

Properties	**K-Means**	**Hierarchical Clustering**	**DBScan Clustering**
Definition	K-Means Clustering generates a specific number of disjoint, flat (nonhierarchical) clusters.	Hierarchical Clustering Method construct a hierarchy of Clustering, not just a single partition of objects.	DBScan algorithm finds a number of clusters starting from the estimated density distribution of corresponding nodes based on connecting points within certain distance thresholds similar to linkage based clustering.
Clustering Criteria	It is well suited for generating globular Clusters.	Use a Distance matrix as clustering criteria. A Termination condition can be used, Example-A number of Clusters.	It uses hub points, edge points and noise point. Find cluster completely surrounded by different clusters. Robust towards outlier detection.
Performance	The performance of K-Means algorithm is better than Hierarchical clustering and DBScan.	The performance of Hierarchical clustering is less than KMeans clustering	The performance of dbscan clustering is less than K-Means clustering and Hierarchical clustering.
Cluster	There are always k.	The number of clusters k is not required as an input.	The number of clusters k is not required as an input.
Data Set	K-Means algorithm is good for large dataset.	Hierarchical algorithm is good for small dataset.	DBScan algorithm is good for small and medium dataset
Precision	0.8056	0.1238	0.2438
Recall	0.3257	0.3559	0.3427
F-Score	0.4575	0.1776	0.2153

Table 1: Comparisons Analysis is Based on Heart Disease

6. Conclusion

Main advantage of K - mean algorithm is increase in the data sets makes increase in the better performance. But these conditions apply when researchers use is limited to numeric values. Hierarchical algorithm was accepted for certain type data, and due to its complication, a new method for assigning rank value to each categorical attribute using K- means can be used in which categorical data is first converted into numeric by assigning rank. Generally, the time taken will vary from processor to processor. The algorithms K-Means and Hierarchical algorithm are have been implemented here. This work was intended in grouping the requirements where a large number of requirements are decomposed into small groups which can be easily analyzed and further grouped. The performance of the partitioning based algorithms was analyzed using the only selected attributes from the total number of attributes of input dataset. It is very evident from the results that the computational complexity of the K-Means algorithm with Heart diseases dataset is better than that of Hierarchical algorithm for both of the dataset. The K-Means algorithm is efficient for Heart diseases dataset with arff or CSV format. It is well suited for requirement clustering of Heart diseases related medical applications.

References

[1] Sung Young Jung, and Taek-SooKim, "An Agglomerative Hierarchical Clustering Using Partial Maximum Array and Incremental Similarity Computation Method", Proceedings of the 2001 IEEE International Conference on Data Mining, p.265-272, November 29-December 02, 2001.

[2] R.J. Gil-Garcia; J.M. Badia-Contelles, "A General Framework for Agglomerative Hierarchical Clustering Algorithms A Pons-Porrata Pattern Recognition, 2006. ICPR 2006. 18th International Conference on Volume 2, p.569 – 572, 2006.

[3] K.P.Soman, ShyamDiwakar, and V.Ajay, "Insight into Data Mining- Theory and Practice", Eastern Economy Edition, Prentice Hall of India Pvt. Ltd, New Delhi, 2006.

[4] "Measuring Association d12 Between Clusters 1 and 2" in http:// www.stat.psu.edu/online/courses/stat505/18_cluster/05_cluster_be tween. html.

[5] Margaret H.Dunham "Data Mining Introductory and Advance Topics", Low price Edition – Pearson Education, Delhi, 2003.

[6] "Euclidean Distance" in http://people.revoledu.com/kardi/tutorial l/Similarity /EuclideanDistance.html.

[7] "Cluster analysis" in http://en.wikipedia.org/wiki/Cluster_ analysis.

[8] "Levenhtein_Distance" http://en.wikipedia.org/wiki/levenshtein_ Distance.

[9] "Similarity Metrics" in http://www.dcs.shef.ac.uk/~sam /stringmetrics.html# hamming.

[10] "Levenshtein_Distance" in http://www.dcs.shef.ac.uk/~sam /stringmetrics.html# Levenshtein.

[11] "Tsunami victimlist" http://www.ems.narenthorn.thaigov. net/ tsunami_e/tsunamilist.php.

[12] "Euclidean distance" inhttp://en.wikipedia.org/wiki/Euclidean_ distance#Onedimensional_distance.

[13] "Distance" in http://en.wikipedia.Org/wiki/ Distance# Mathematics.

[14] "Hierarchical Clustering Algorithms"in http://home.dei. polimi.it/matteucc/Clustering/tutorial_html/hierarchical.ht ml.

[15] "Hierarchical Clustering Algorithms"in http://home.dei. polimi.it/matteucc/Clustering/tutorial_html/hierarchical.ht ml.

[16] Hui-Chuan Lin "Survey and Implementation of Clustering Algorithms" an Unpublished master's thesis for master's degree, Hsinchu, Taiwan, Republic of China (2009).

[17] Jinxin Gao, David B. Hitchcock ―James-Stein Shrinkage to Improve KMeans Cluster Analysis, University of South Carolina, Department of Statistics November 30, 2009.

[18] Schilham A, Prokop M, Van Ginneken B. Computer analysis computed tomography scans of the lung: a survey. IEEE TransMedicalImaging.25(4):38405doi:10.1109/TMI .2005.8 62753.: 2006

[19] Velmurugan T. Efficiency of K-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points. International Journal of Computer Technology & Applications. 2012; 3(5):1758–64.

[20] Velmurugan T. Performance based analysis between k-Means and fuzzy C-means clustering algorithms for connection oriented telecommunication data. Appl Soft Comput. 2014; 19:134–46.

[21] Zhong W, Altun G, Harrison R, Tai PC, Pan Y. Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property. IEEE Transactions on NanoBioscience. 4(3):255–65, 2005.