DATA MINING AND HIERARCHICAL CLUSTERING TECHNIQUES

Mrs. B. Meena Preethi[1], Aishwarya S.[2], Sabarini[3], Gokul Krishnan M.[4]
[1]*Asst. Professor, Department of MSc Software Systems & Computer Applications*
[2,3,4]*Department of MSc Software Systems & Computer Applications*
Sri Krishna Arts and Science College, Kuniamuthur, Coimbatore

Abstract- **This paper we have made note on a variety of techniques of data mining, approaches clustering method and different areas of usage which are helpful in the important field of data mining Technologies. Bulking in the usage of data in each place of operation may generate large volumes of data. Corporate data usage made the decision makers require access from all such sources and take strategic decisions. To analyze, manage and make a decision of such type of huge amount of data we need techniques called the data mining which will transforming in many fields. Hierarchical clustering their working methods and applications are explained as a major topic.**

I. INTRODUCTION

Each and every human being makes use of large datas which are distributed as vast data and are used in the different fields. They may be in the form of documents, graphical formats, and the video or even may be records of varying array. To analyze these available data and also to take a good decision for maintaining the data Mining technique is used. It is a process of semi-automatically analyzing large databases to find the valid, novel, useful and understandable patterns.). Data mining use algorithms to extract the information and patterns derived by the KDD process**.** It is also known as Knowledge Discovery in Databases (KDD).

II. CLUSTERING

Clustering is defined as process of converting a group of abstract objects into classes of similar objects. Clustering is a process of dividing data into groups of similar objects. It is explorative and hence it does make any distinction between dependent and independent variables. Representing the data by fewer clusters may lose certain fine details, but still it achieves simplification. Cluster analysis is used to identify groups of cases. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. Cluster analysis is also known as segmentation analysis or taxonomy analysis. Clusters correspond to hidden patterns, the search for clusters is independent learning, and the resulting system represents a data concept from a machine learning perspective. The Cluster Analysis is an explorative analysis that tries to identify structures within the data, and includes the spatial database applications, medical diagnostics, CRM, marketing computational biology, Web analysis and many others. More specifically, it tries to identify homogenous groups of cases, i.e., observations, participants, respondents.

III. HIERARCHICAL CLUSTERING

Hierarchical clustering the goal is to produce a hierarchical series of nested clusters, ranging from the bottom clusters of individual points to an all-inclusive cluster at the top. Dendogram, a diagram which graphically represents this hierarchy describes the order in an inverted tree structure in which points are merged in a bottom-up view or a top-down view where clusters are split. Common application of hierarchical techniques is that they correspond to represent taxonomies that are very commonly used in the biological sciences, e.g., species, phylum, kingdom, genus another important feature is that the hierarchical techniques do not assume any particular number of clusters it is inadequate. Instead any desired number of clusters can be obtained by "cutting" the dendogram (tree nodes) at the proper level. . Some of cluster analysis work occurs under the name of "mathematical taxonomy". The given are the steps of clustering n items.

1. Start the process.
2. Assign each item to a cluster, if there are N items, then it denotes N clusters, each containing just one item. The distances

(similarities) between the clusters are same as the distances between the items they contain .The distance are denoted as the similarities.

3. Find the most similar pair of clusters which has the close distance and merge them together into a single cluster.
4. Compute distances between the new cluster and the old clusters.
5. Repeat steps 2 and 3 and 4 until it reach a cluster less stage.

Step 3 can also be done in other ways such as *single-linkage* from *complete-linkage* and *average-linkage* clustering. In *single-linkage* clustering, we consider the distance between one cluster and another cluster is equal to the shortest distance from any member of one cluster to any member of the other cluster. In *complete-linkage* clustering (also called the *diameter* or *maximum* method), the distance between one cluster and another cluster to be equal to the greatest distance from any member of one cluster to any member of the other cluster. In *average-linkage* clustering, we consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster. This method of hierarchical clustering merges the clusters hence it is called *agglomerative.* another method of clustering which reverses the process by dividing the objects from a cluster to individual group , also called as *divisive* hierarchical clustering .Divisive methods are not generally available, and are rarely have been applied.

IV. AGGOLOMERATIVE METHODS

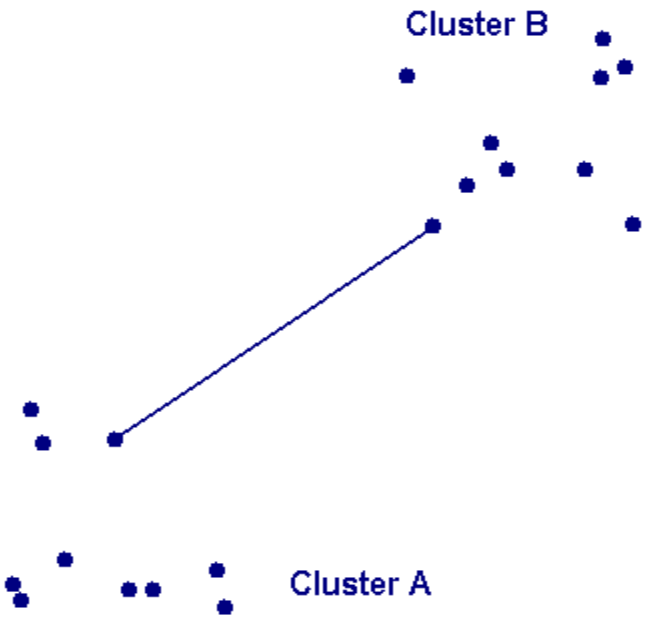
An agglomerative hierarchical clustering procedure produces a series of partitions of the data, P_n, P_{n-1}, ... , P_1. The first part (P_n) consists of n single object clusters, the last(P_1), consists of single group containing all n object cases. During each stage of the method , it joins together the two clusters that has the similarity i.e the close distance .Differences between methods arise due to different ways of defining distance or similarity between clusters. The given are the few methods of aggolomerative method.

4.1SINGLELINKAGE CLUSTERING

Here the distance between the given number of groups is defined as the distance between the closest pair of objects where the object has less distance or similarity, where only pairs consisting of one object from each group are considered. This method is One of the simplest form of agglomerative hierarchical clustering and it is also known as the nearest neighbor technique.

In the single linkage method, **F(r,s)** is computed as

F(r,s) = Min { *f(i,j)* : Where object *i* is in cluster ***r*** and object *j* is cluster ***s*** }



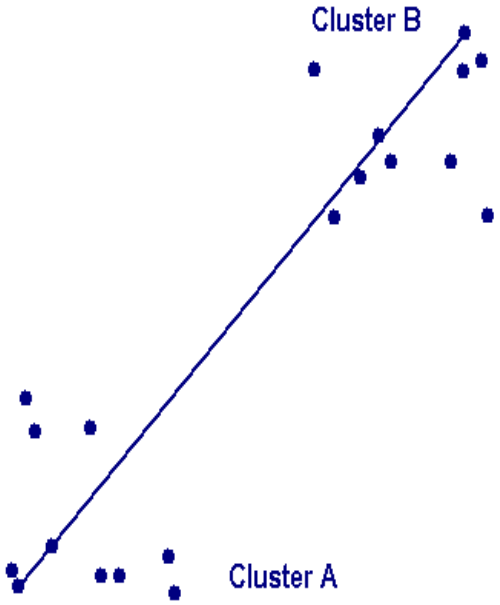
4.2COMPLETE LINKAGE CLUSTERING

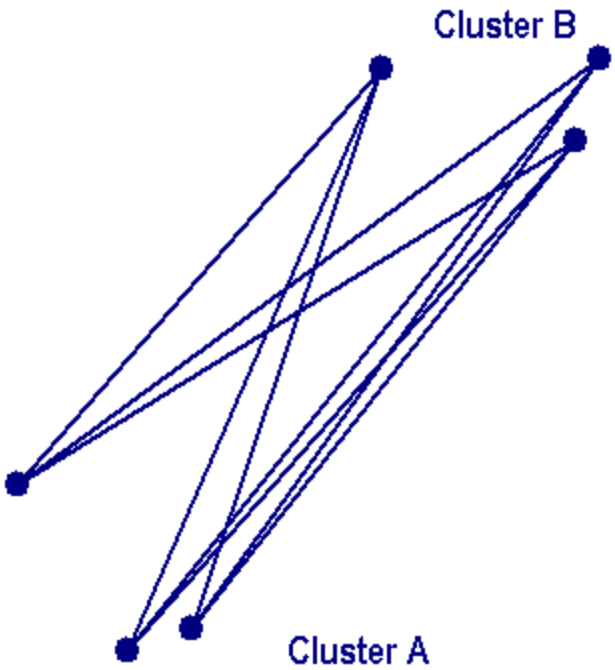
In the complete linkage, also called farthest neighbor, the clustering method is the opposite of single linkage. Distance between groups is now defined as the distance between the most distant pair of objects, one from each group.

In the complete linkage method, **D(r,s)** is computed as

D(r,s) = Max { *d(i,j)* : Where object *i* is in cluster ***r*** and object *j* is cluster ***s*** }

The distance between every possible object pair *(i,j)* is computed, where object *i* is in cluster **r** and object *j* is in cluster **s** and the maximum value of these distances is said to be the distance between clusters **r** and **s.** The distance between two clusters is given by the value of the longest link between the clusters.





4.3AVERAGE LINKAGE CLUSTERING

In this method the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group.

In the average linkage method, **F(r,s)** is computed as

$$F(r,s) = T_{rs} / (N_r * N_s)$$

Where $\mathbf{T_{rs}}$ is the sum of all pairwise distances between cluster **r** and cluster **s.** $\mathbf{N_r}$and $\mathbf{N_s}$ are the sizes of the clusters **r** and **s**, respectively**.**

At each stage of hierarchical clustering, the clusters **r** and **s** , for which **F(r,s)** is the minimum, are merged.

4.4AVERAGE GROUP LINKAGE

Here , the two clusters **r** and **s** are merged together so that the average pair wise distance within the newly formed cluster is minimum. using this type of method, groups once formed are represented by their mean values for each variable .

Then the distance between clusters **r** and **s**, **F(r,s)**, is computed as

F(r,s) = Average { d(i,j) : Where observations i and j are in cluster **t**, the cluster formed by merging clusters **r** and **s** }

At each and every stage of this type of clustering , the clusters **r** and **s** , for which **F(r,s)** is minimum, are merged together. Thus those two clusters are merged together so the newly formed cluster, on average, will have minimum pair wise distances between the points.

V. BASIC AGGOLOMETRIC ALGORITHM

Lance-Williams algorithm is a technique used to express alogglomerative method.

Basic Agglomerative Hierarchical Clustering Algorithm

1) Compute the proximity of the graph, if necessary and sometimes the proximity may be given already.

2) Merge the similar (most closest) two clusters.

3) proximity between the new cluster and the original clusters can be reflected by updating the proximity of the matrix.

4) Repeat steps 3 and 4 until only a single cluster remains and all others get paired.

The main aim of the algorithm is the calculation of the proximity between two clusters, and this is where the various agglomerative hierarchical techniques differ. The for the proximity between clusters *Q* and *R*, where *R* is formed by merging clusters *A* and *B*.

$$p(R, Q) = \alpha_A p(A, Q) + \alpha_B p(B, Q) + \beta p(A, Q) + \gamma |p(A, Q) - p(B, Q)|$$

this formula denotes that merge , *A* and *B* to form cluster *R*, then the distance of the new cluster, *R*, to an existing cluster, *Q*, is a linear function of the distances of *Q* from the original clusters *A* and *B*. Any hierarchical technique that can be phrased in this way does not need the original points, only the proximity matrix, which is updated as clustering occurs

VI. DATAMINING CLASSIFICATION SCHEMES

There are two major classification schemes of data mining. They are

1. Decisions in data mining
2. Data mining tasks.

6.1DECISIONS IN DATA MINING

It deals with the kinds of knowledge to be discovered, database to be mined, kinds of techniques to be utilized and kinds of applications to be adapted

1. Databases to be mind

It deals with object-oriented, transactional, Relational, object-relational, spatial, active, legacy, time-series, multi-media, text, heterogeneous, WWW, etc.

2. **Knowledge to be mined**

It deals with discrimination, Characterization, classification, association, clustering, trend, deviation and outlier analysis, etc.

3. **Techniques utilized**

It is Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.

4. **Applications adapted**

Telecommunication, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, Retail, banking, etc.

6.2DATA MINING TASKS

Few data mining tasks are given as follows

1. Similarity matching

It is way of finding similarities between data according to their characteristic features. Similarity matching or Clustering is a collection of similar kinds of data object. For example, pattern recognition, image processing, city planning.

2. Association rules and variants

Association is the most popular techniques of data mining and one of the fined most frequent item set. This technique is used for the market based analysis in order to identify a set, or sets of products that consumers often purchase at the same time. Because of its basic methodology, association is referred to as "relation technique". Association discovers the patterns in data which are based upon relationships between items in the same transaction

3. Sequence Discovery

This method of data mining is used within the market based analysis in order to identify a set, or sets of products that consumers often prefer or purchase at the same time. For example, scientific experiment, natural disaster and analysis of DNA sequence.

VII. APPLICATION OF DATA MINING

Data mining technologies are used in various fields because of the valuable information and fast access of data and from a huge collection large amount of data. Data mining application includes fraud detection, telecommunication, finance, marketing, and education sector, medical and so on. Some of the main applications used in education are listed below

7.1EDUCATION SECTOR

The new emerging field called the "Education Data Mining" is defined as the process of applying data mining in education sector; Using of this term enhances the performance of student, student behavior, and drop out student, which subject selected in the course. Student's data are used to analyze their learning behavior to predict the result. Data mining in higher education is a recent research which is being carried out.

7.2DATA MINING IN TELECOMMUNICATION

The telecommunications field requires data mining at a huge amount and they implement data mining technology because of telecommunication industry have the large amounts of data and have a very large customer, and they are rapidly changing and highly competitive environment. Telecommunication companies uses data mining technique to improve their marketing efforts, detection of fraud, and better management of telecommunication networks

7.3DATA MINING IN MARKET BASKET ANALYSIS

The stores can use this information by putting these products in close proximity of each other and making them more visible and accessible for customers at the time of shopping .They include shopping database. The main goal of market basket analysis is finding the products that customers frequently purchase together.

7.4DATA MINING IN CLOUD COMPUTING

Data Mining techniques are also used in cloud computing. The implementation of data mining techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage. The data mining technique in Cloud Computing to perform efficient, reliable and secure services for their users.

7.5DATA MINING IN BANKING AND FINANCE

Data mining is also been used extensively in the banking and financial markets. Here, data mining is used to predict credit card fraud, to estimate risk, to analyze the trend and profitability. In the financial markets, data mining technique such as neural networks used in stock forecasting, price prediction and so on.

7.6 DATA MINING IN EARTHQUAKE PREDICTION

Data mining is used in the Prediction of the earthquake from the satellite maps. Earthquake is the sudden movement of the Earth's crust caused by the abrupt release of stress accumulated along a geologic fault in the interior. There are two basic categories of earthquake predictions: forecasts (months to years in advance) and short-term predictions (hours or days in advance).

VIII. APPLICATIONS OF CLUSTERING

8.1SIMILARITY SEARCHING IN MEDICAL IMAGE DATABASE

Similarity searching in medical image database is one of the most important applications of medical image retrieval. In order to detect many diseases like Tumor etc, the scanned pictures or the x-rays are compared with the existing ones and the dissimilarities are recognized. During a complete scanning, we have clusters of images of different parts of the body. For example, the images of the CT scan of brain are kept in one cluster. To further arrange things, the images in which the right side of the brain is damaged are kept in one cluster. The hierarchical clustering is used. The stored images have already been analyzed and a record is associated with each image. In this form a large database of images is maintained using the hierarchical clustering.

When a new image is considered , it is firstly recognized that what particular cluster this image belongs, and then by similarity matching with a healthy image of that specific cluster the main damaged portion or the diseased portion is recognized. Then the image is sent to that specific cluster and matched with all the images in that particular cluster. Now the image, with which the query image has the most similarities, is retrieved and the record associated to that image is also associated to the query image. This means that now the disease of the query image has been detected. Using this technique and some really precise methods for the pattern matching, diseases like really fine tumor can also be detected. So by using clustering an enormous amount of time in finding the exact match from the database is reduced.

8.2ANOTHER APPLICATIONS CLUSTERING IN DATA MINING

Clustering is defined as one of the first steps in data mining analysis . It relates a group of images or objects with an identical group of objects for establishing relationship and finds the similarity. This technique is applicable for the development of population difference models, such as demographic-based customer segmentation. analyses done using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired result , for example a political may need to understand the negative points

of people to gain his/her attention by promising certain needs .This process is done by data mining techniques. But if the system clusters the products that are given less features then only the cluster of such products would have to be checked rather than comparing the sales value of all the products...

IX. CONCLUSION

This paper provides a general idea of data mining, data techniques and data mining applications in various fields. The main objectives of data mining techniques are to discover the knowledge from active data and to abstract data from one source even deep. These applications use classification, Prediction, clustering, Association techniques and so on are been discussed in this paper.

REFERENCE

[1] https://en.wikipedia.org/wiki/Data_mining
[2] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
[3] http://www.thearling.com/text/dmwhite/dmwhite.htm
[4] http://searchsqlserver.techtarget.com/definition/data-mining
[5] http://www.investopedia.com/terms/d/datamining.asp