ISSN 2229-5518

Single Link clustering on data sets

Ajaya Kushwaha, Manojeet Roy

Abstract— Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Most data-mining methods assume data is in the form of a feature-vector (asingle relational table) and cannot handle multi-relational data. Inductive logic programming is a form of relational data mining that discovers rules in _first-order logic from multi-relational data. This paper discusses the application of SLINK to learning patterns for link discovery. Clustering is among the oldest techniques used in data mining applications. Typical implementations of the hierarchical agglomerative clustering methods (HACM) require an amount of $O(N^2)$-space when there are N data objects, making such algorithms impractical for problems involving large datasets. The well-known clustering algorithm RNN-CLINK requires only $O(N)$-space but $O(N^3)$-time in the worst case, although the average time appears to be $O(N^2 \log N)$.

Index Terms—clustering, nearest neighbor, reciprocal nearest neighbor, complete link, probabilistic analysis.

1 INTRODUCTION

CLUSTER analysis, or clustering, is a multivariate statistical technique which identifies groupings of the data objects based on the inter-object similarities computed by a chosen distance metric. Clustering methods can be divided into two types: partitioned and hierarchical. The partitioned clustering methods start with a single group that contains all the objects, then proceed to divide the objects into a fixed number of clusters. Many early clustering applications used partitional clustering due to its computational efficiency when large datasets need to be processed. However, partitional clustering methods require prior specification of the number of clusters as an input parameter. Also, the partitions obtained are often strongly dependent on the order in which the objects are processed.[7] The hierarchical agglomerative clustering methods (HACM) attempt to cluster the objects into a hierarchy of nested clusters. Starting with the individual objects as separate clusters, the HACMs successively merge the most similar pair of the clusters into a higher level cluster, until a single cluster is left as the root of the hierarchy. HACMs are also known as unsupervised classification, because their ability in automatically discovering the inter-object similarity patterns. Clustering techniques have been applied in a variety of engineering and scientific fields such as biology, psychology, computer vision, data compression, information retrieval, and more recently, data mining.[3 4 8 10] .

The various technical aspect work on

- *Author name is Mr. Ajay Kushwaha, Reader c.s.e Deptt.RCET ,bhilaiM.C.A , Mtech(CS),PhD (CSE) pursuing from CSVTU ,Chhattisgarh Research area – MANET*
  kushwaha.bhilai@gmail.com
  Address : RCET ,KOHKA - KURUD ROAD ,KOHKA , BHILAI -490023
- *Co-Author name is **Mr. Manojeet Roy**, Computer Science Department , CSVTU University/ Rungta College of Engineering And Technology College/ Rungta Group of colleges Organization , City Bhilai Country India, Phone/ Mobile 09993649781., (e-mail: roy.mannu@gmail.com).*

2. DATA CLUSTERING METHODS

ROCEDURE FOR PAPER SUBMISSION

2.1 Review Stage

Detailed DATA CLUSTERING is considered an interesting approach for finding similarities in data and putting similar data into groups. Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups [1]. The idea of data

Grouping, or clustering, is simple in its nature and is close to the human way of thinking; whenever we are presented with a large amount of data, we usually tend to summarize this huge number of data into a small number of groups or categories in order to further facilitate its analysis. Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings. Nevertheless, finding these groupings or trying to categorize the data is not a simple task for humans unless the data is of low dimensionality (two or three dimensions at maximum.) This is why some methods in soft computing have been proposed to solve this kind of problem. Those methods are called "Data Clustering Methods" and they are the subject of this paper. Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction. By finding similarities in data, one can represent similar data with fewer symbols for example. Also if we can find groups of data, we can build a model of the problem based on those groupings.

Clustering techniques are broadly divided into hierarchical clustering and partitional clustering

2.2 Clustering Algorithms

The community of users has played lot emphasis on developing fast algorithms for clustering large data sets [13]. Clustering is a technique by which similar objects are grouped to-

gether. Clustering algorithms can be classified into several categories such partitioning-based clustering, hierarchical algorithms,
density based clustering and grid based clustering.
Now a day's huge amount of data is gathered from remote sensing, medical data, geographic information system, environment etc. So everyday we are left with enormous amount of data that requires proper analysis. Data mining is one of the emerging fields that are used for proper decision-making and utilizing these resources for analysis of data. They are several researches focused on medical decision making [14]

2.3 A. Hierarchical Clustering

"Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity ".
Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) [Jain and Dubes 1988; Kaufman and Rousseeuw 1990]. An agglomerative clustering starts with one-point (singleton clusters. It then recursively merges two or more most appropriate clusters . Fig. 2 provides a simple example of hierarchical clustering

2.4 Figure

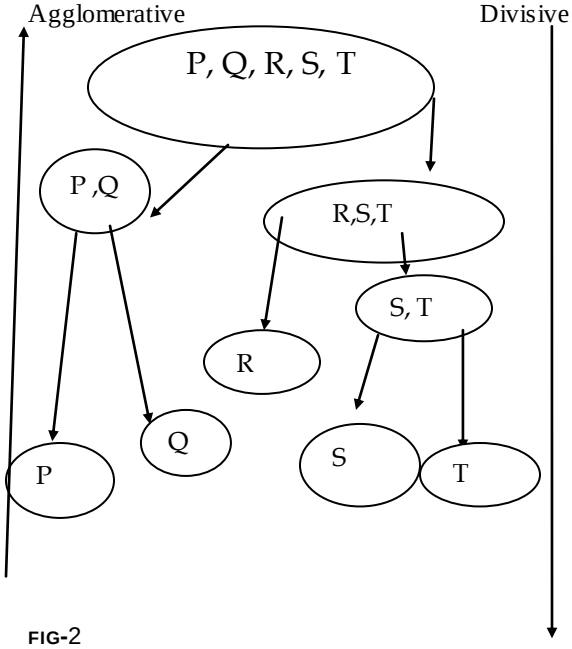

FIG-2

clusters. It then recursively merges two or more most appropriate clusters. Fig. 2 provides a simple example of hierarchical clustering.

Hierarchical Clustering

3 PARTITIONAL CLUSTERING

Instead of a clustering structure a partitional clustering algorithm obtains a single partition of the data [2]. They generate a single partition of the data to recover natural groups present in the data. The proximity matrix among the objects is required by hierarchical clustering techniques. The partitional techniques expect data in the form of a pattern matrix. Partitioning techniques are used frequently in engineering applications where single partitions are important. Partitional clustering methods are especially appropriate for the efficient representation and compression of large databases [10]. The algorithm is typically run multiple times with different starting states. The best configuration obtained from all the runs is used as the output clustering provides a simple example of partitional clustering. Fig. 3 Splitting of a large cluster by Partitional Algorithm [4] Dubes and Jain (1976) emphasize the distinction between clustering methods and clustering algorithms. The K-means is
the simplest and most commonly used algorithm employing a squared error criterion [McQueen 1967]. The K-means algorithm is popular because it is easy to implement. Its time complexity is O (n), where n is the number of partitions. The algorithm is sensitive to the selection of initial partition. It may converge to a local minimum of the criterion function value if the initial partition is not properly chosen

SLINK Algorithm

In cluster analysis, **single linkage**, **nearest neighbour** or **shortest distance** is a method of calculating distances between clusters in hierarchical clustering. In single linkage, the distance between two clusters is computed as the distance between the two closest elements in the two clusters.

Mathematically, the linkage function – the distance $D(X,Y)$ between clusters X and Y – is described by the expression

$$D(X,Y) = \min_{x \in X, y \in Y} d(x,y),$$

where X and Y are any two sets of elements considered as clusters, and $d(x,y)$ denotes the distance between the two elements x and y.

A drawback of this method is the so-called *chaining phenomenon*: clusters may be forced together due to single elements being close to each other, even though many of the elements in each cluster may be very distant to each other.

NAIVE ALGORITHM

The following algorithm is an agglomerative scheme that erases rows and columns in a proximity matrix as old clusters are merged into new ones. The $N \times N$ proximity matrix D contains all distances $d(i,j)$. The clustering are assigned

sequence numbers 0,1,....., $(n-1)$ and $L(k)$ is the level of the kth clustering. A cluster with sequence number m is denoted (m) and the proximity between clusters (r) and (s) is denoted $d[(r),(s)]$.

The algorithm is composed of the following steps:

1. Begin with the disjoint clustering having level $L(0) = 0$ and sequence number m = 0.
2. Find the most similar pair of clusters in the current clustering, say pair (r), (s), according to $d[(r),(s)] = \min d[(i),(j)]$ where the minimum is over all pairs of clusters in the current clustering.
3. Increment the sequence number: $m = m + 1$. Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to $L(m) = d[(r),(s)]$
4. Update the proximity matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r,s) and old cluster (k) is defined as $d[(k),(r,s)] = \mathbf{min}\ \boldsymbol{d[(k),(r)], d[(k),(s)]}$.
5. If all objects are in one cluster, stop. Else, go to step 2.

.

4. DATA CLUSTERING OVERVIEW

As mentioned earlier, data clustering is concerned with the partitioning of a data set into several groups such that the similarity within a group is larger than that among groups. This implies that the data set to be partitioned has to have an inherent grouping to some extent; otherwise if the data is uniformly distributed, trying to find clusters of data will fail, or will lead to artificially introduced partitions. Another problem that may arise is the overlapping of data groups. Overlapping groupings sometimes reduce the efficiency of the clustering method, and this reduction is proportional to the amount of overlap between groupings.

Usually the techniques presented in this paper are used in conjunction with other sophisticated neural or fuzzy models. In particular, most of these techniques can be used as preprocessors for determining the initial locations for radial basis functions or fuzzy ifthen rules. The common approach of all the clustering techniques presented here is to find cluster centers that will represent each cluster. A cluster center is a way to tell where the heart of each cluster is located, so that later when presented with an input vector, the system can tell which cluster this vector belongs to by measuring a similarity metric between the input vector and al the cluster centers, and determining which cluster is the nearest or most similar one.Some of the clustering techniques rely on knowing the number of clusters apriori. In that case the algorithm tries to partition the data into the given number of clusters. K-means and Fuzzy C-means clustering are of that type. In other casesit is not necessary to have the number of clusters known from the beginning; instead the algorithm starts by finding the first large cluster, and then goes to find the second, and so on. Mountain and Subtractive clustering are of that type. In both cases a problem of known cluster numbers can be applied; however if the number of clusters is not known, K-means and Fuzzy C-means clustering cannot be used.

Another aspect of clustering algorithms is their ability to be implemented in on-line or offline mode. On-line clustering is a process in which each input vector is used to update the cluster centers according to this vector position. The system in this case learns where the cluster centers are by introducing new input every time. In off-line mode, the system is presented with a training data set, which is used to find the cluster centers by analyzing all the input vectors in the training set. Once the cluster centers are found they are fixed, and they are used later to classify new input vectors. The techniques presented here are of the off-line type. A brief overview of the four techniques is presented here. Full detailed discussion will follow in the next section.

The first technique is K-means clustering [6] (or Hard C-means clustering, as compared to Fuzzy C-means clustering.) This technique has been applied to a variety of areas, including image and speech data compression, [11,12] data preprocessing for system modeling using radial basis function networks, and task decomposition in heterogeneous neural network architectures [6]. This algorithm relies on finding cluster centers by trying to minimize a cost function of dissimilarity (or distance) measure. The second technique is Fuzzy C-means clustering, which was proposed by Bezdek in 1973 [5] as an improvement over earlier Hard C means clustering. In this technique each data point belongs to a cluster to a degree specified by a membership grade. As in K-means clustering, Fuzzy C-means clustering relies on minimizing a cost function of dissimilarity measure. The third technique is Mountain clustering, proposed by Yager and Filev [5]. This technique builds calculates a mountain function (density function) at every possible position in the data space, and chooses the position with the greatest density value as the center of the first cluster. It then destructs the effect of the first cluster mountain function and finds the second cluster center. This process is repeated until the desired number of clusters have been found.

The fourth technique is Subtractive clustering, proposed by Chiu [5]. This technique is similar to mountain clustering, except that instead of calculating the density function at every possible position in the data space, it uses the positions of the data points to calculate the density function, thus reducing the number of calculations significantly.

1. 5. EXPERIMENT

They are several series of experiments performed in this section to determine relevant pattern detection for medical diagnosis.

Clinical Database II

In this study, the data was taken from SEER datasets which has record of cancer patients from the year 1975 – 2001 from Ref.[11]. The data was again classified into two groups that arespatial and non spatial dataset. Spatial dataset consists of location collected include remotely sensed images, geographical information with spatial attributes such as location, digital sky survey data, mobile phone usage data, and medical data. The five major cancer areas such as lung, kidney, throat, stomach and liver were experimented. After this data mining

algorithms were applied on the data sets such as K-means, SOM and Hierarchical

clustering technique

The K-means method is an efficient technique for clustering large data sets and is used to determine the size of each cluster. After this the HAC (hierarchical agglomerative clustering), is used on our datasets in which we have used tree based partition method in which the results has shown a tree structure and the gap between the nodes has been highlighted in the Table 3. The HAC has proved to have for better results than other clustering methods. The principal component analysis technique has been used to visualize the data. The X, Y coordinates identify the point location of objects. The coordinates were used and the clusters were determined by appropriate attribute value. The mean and standard deviation of each cluster was determined hey were interesting facts that suggests that number of cancer cases that occur during the time interval. Table 1 represents the size of each cluster determined by Kmeans clustering technique for dataset 2 .In Table 2 shows the
number trials generated for the cluster determination.

Table 1. Result of K-means

cluster	Descriprion	size
cluster n°1	c_kmeans_1	8
cluster n°2	c_kmeans_2	17
cluster n°3	c_kmeans_3	2

Table 2 Number of trials

Number of trials	5
Trial	**Ratio explained**
1	0.45731
2	0.48054
3	0.45688
4	0.52069
5	0.20595

6.5 Theorems and Proofs

Table 3 presents HAC (hierarchical agglomerative clustering) in which the cluster were determined with appropriate size. Table 4 represents the best cluster selection in which the gap is defined as the space in between the clusters

Table 3. HAC Cluster size

Clusters	3	
Cluster	**Description**	
cluster n°1	c_hac_1	16
cluster n°2	c_hac_2	2
cluster n°3	c_hac_3	9

Cluster	**BSS ratio**	**Gap**
1	0	0
2	0.4365	3.205
3	0.5817	0.5671
4	0.6753	0.5323
5	0.7205	0.1321
6	0.7536	0.0937
7	0.7783	0.0247
8	0.8007	0.0348
9	0.82	0.028
10	0.8367	0.0176
11	0.8518	0.016
12	0.8655	0.0185
13	0.8775	0.0083
14	0.8887	0.0073
15	0.8993	0.0147

16	0.9085	0.011
17	0.9167	0.0065
18	0.9244	0.0114
19	0.931	0.0107
20	0.9366	0.005

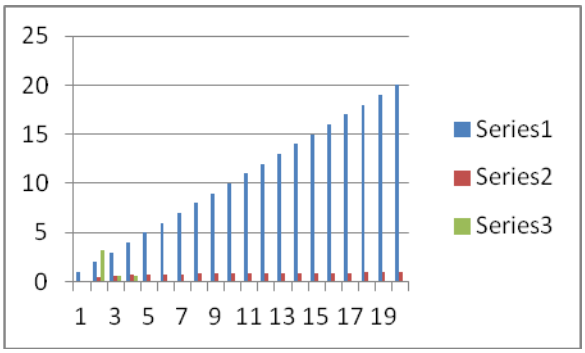

Figure. - Dendrogram

Above Figure represents the dendogram in which the dataset has been partitioned into three clusters with the K-means. The HAC clustering algorithm is applied on K-means to generate the dendogram. In a dendogram, the elements are grouped together in one cluster when they have the closest values of all elements available. In the diagram the cluster2 and cluster 3 are combined. The subdivisions of clusters are then analyzed

4. CONCLUSIONS

This paper focuses on clustering algorithms such as HAC and KMeans in which, HAC is applied on K-means to determine the number of clusters. The quality of cluster is improved, if HAC is applied on K-means. The paper has referenced and discussed the issues on the specified algorithms for the data analysis. The analysis does not include missing records. The application can be used to demonstrate how data mining technique can be combined with medical data sets and can be effectively demonstrated in modifying the clinical research.

This study clearly shows that data mining techniques are promising for clinical datasets. Our future work will be related to missing values and applying various algorithms for the fast implementation of records. In addition, the research would be focusing on spatial data clustering to develop a new spatial data mining algorithm.

.

REFERENCES

[1] Li Zhan, Liu Zhijing, , ' Web Mining Based On Multi-Agents ', COMPUTERSOCIETY,IEEE(2003)

[2] Margaret H. Dunham and Sridhar, Data Mining, Introduction and Advanced Topics, (Prenticce Hall Publication), ISBN 81-7758-785-4, chap nos.1,7, pp.3,4,195-218.

[3] M. R. Anderberg, *Cluster Analysis for Appliations*, Academic Press, New York, 1973.

[4] A. Berson and S. J. Smith, *Data Warehousing, Data Mining, and OLAP*, McGraw-Hill, New York, 1997.

[5] Jang, J.-S. R, Sun, C.-T., Mizutani, E., "Neuro- Fuzzy and Soft Computing – A Computational Approach to Learning and Machine Intelligence," *Prentice Hall.*

[6] Nauck, D., Kruse, R., Klawonn, F., "Foundations of Neuro-Fuzzy Systems," *John Wiley & Sons Ltd., NY, 1997.*

[7] M. S. Chen, J. Han, and P. S. Yu. Data mining: an overview from database perspective. IEEE Trans. On Knowledge and Data Engineering, 5(1):866—883, Dec.1996

[8] W. B. Frakes and R. Baeza-Yates, *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, 1992.

[9] Y. Zhao and G. Karypis. Evaluation of hierarchical clusteringalgorithms for document datasets. In CIKM, 2002.

[10] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, 1988.

[11] Lin, C., Lee, C., "Neural Fuzzy Systems," *Prentice Hall, NJ, 1996*

[12] Tsoukalas, L., Uhrig, R., "Fuzzy and Neural Approaches in Engineering," *John Wiley & Sons,Inc., NY, 1997*

[13] U.M. Fayyad and P. Smyth. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Menlo Park, CA, 1996.

[14] Kaur H, Wasan S K, Al-Hegami A S and Bhatnagar V, A Unified Approach for Discovery of Interesting Association Rules in Medical Databases, *Advances in Data Mining*, Lecture Notes in Artificial Intelligence, Vol. 4065, Springer- Verlag, Berlin, Heidelberg (2006).

[15] Kaur H and Wasan S K, An Integrated Approach in Medical Decision Making for Eliciting Knowledge, Web-based Applications in Health Care & Biomedicine, *Annals of Information Systems (AoIS), ed. A. Lazakidou*, Springer