D-METRIC SPACES IN AGGLOMERATIVE CLUSTERING

[1]MRS. M. SREEBALA , [2] MR. G. NAGESWARA RAO, [3] DR.S.SAI SATYANARAYANA REDDY

Research Scholar Supervisor

[1] Asst. Professor, Dept. of CSE, LBR College of of Engineering, Mylavaram, Krishna Dist., A.P.

[2] Asst. Professor, Dept. of CSE LBR College of of Engineering, Mylavaram, Krishna Dist., A.P. and Research Scholar, Dravidian University, Kuppam, A.P.

[3]Professor & HOD, Dept. of CSE, LBR Collge of Engineering, Mylavaram, Krishna Dist., A.P.

Email [1]malladisreebala9@gmail.com, [2]garlapati.nag@gmail.com, [3]saisn90@gmail.com

ABSTRACT

Hierarchical agglomerative clustering merges the clusters basing on their distance similarity. In this paper we present a new mathematical method called D metric spaces by Indian mathematician B.C.Dhage who has submitted his thesis in 1984 at Maratwada university. We present an algorithm for hierarchical clustering using D metric concept instead of Euclidean distance which reduces the total number of iterations ,complexity and computation time. Here we showed an example that contains the results of both techniques and also comparisons.

Keywords: *Agglomerative Clustering, Single Linkage Clustering, D Metrics Spaces.*

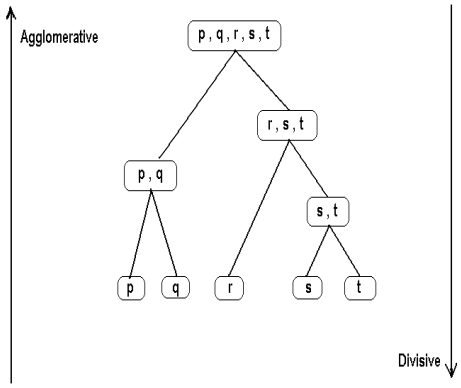
1. INTRODUCTION

Clustering (12-25) also called as data segmentation has variety of goals, all relates to grouping or segmenting a collection of objects (also called as observations, individuals, cases or data rows) into subsets or clusters such that those with in each cluster are more closely related to one another than objects assigned to different clusters. Central to all other goals of clustering is the notion of degree of similarity or dissimilarity between individual objects being clustered.

So the basic principle of clustering hinges on a concept of distance metric or similarity metric. Since the data are invariably real numbers for statistical applications and pattern recognition, a large class of matrices exists and one can define one's own metric depending on a specific requirement. The main emphasis of this is to cluster with a high accuracy as possible, while keeping the I/O costs high. Thus it is not relevant to apply the classical clustering algorithms in the context of data mining and it is necessary to investigate the principle of clustering is to devise efficient algorithms, which meets the specific requirements of minimizing the I/O operations.

2. HIERARCHICAL CLUSTERING

Here the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to n clusters each containing a single object. Hierarchical Clustering is subdivided into *agglomerative* methods, which proceed by series of fusions of the **n** objects into groups, and *divisive* methods, which separate **n** objects successively into finer groups. (12-25).



The hierarchical agglomerative algorithm can be written as

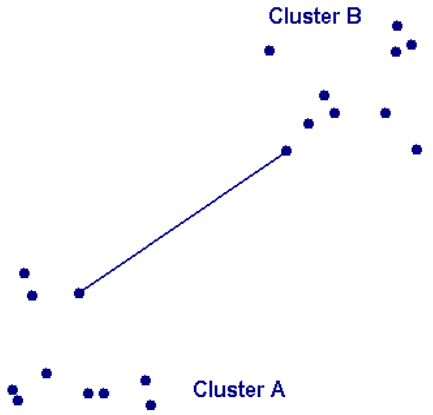
Given a set of N items to be clustered, and an **N** x **N** distance (or similarity) matrix, the basic process of hierarchical clustering is this:

1. Start by assigning each item to its own cluster, so that if you have **N** items, you now have **N** clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of
5. size **N**.

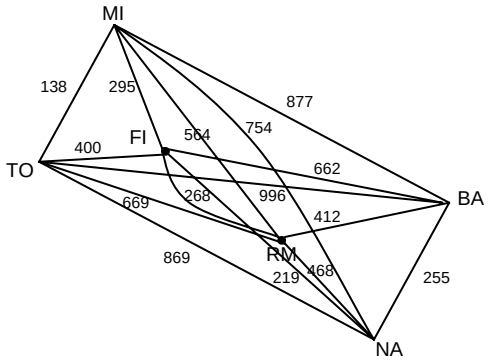
Step 3 can be done in different ways, which is what distinguishes *single-link* from *complete-link* and *average-link* clustering.

1. Single linkage method : Here the similarity of two clusters is the similarity of their *most similar* members This single-link merge criterion is *local*. Here we consider where the two clusters come closest to each other. Other, more distant parts of the cluster and the clusters' overall structure are not taken into account.

Graphically it can be represented as ,….



Algorithm that illustrates single linkage clustering with the following example:



The matrix is follows.

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

Input distance matrix (L = 0 for all the clusters):

1. Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.
2. Find the least dissimilar 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 in this way: d[(k),(r,s)] = min d[(k),(r)], d[(k),(s)] If all objects are in one cluster , stop. Else, go to step 2.

now see a simple example: a hierarchical clustering of distances in kilometers between some Italian cities. The method used is single-linkage.

The nearest pair of cities is MI and TO, at distance 138. These are merged into a single cluster called "MI/TO". The level of the new cluster is L(MI/TO) = 138 and the new sequence number is m = 1.

Then we compute the distance from this new compound object to all other objects. In single link clustering the rule is that the distance from the compound object to another object is equal to the shortest distance from any member of the cluster to the outside object. So the distance from "MI/TO" to RM is chosen to be 564, which is the distance from MI to RM, and so on.

After merging MI with TO we obtain the following matrix:

	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	877	295	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0

min d(i,j) = d(NA,RM) = 219 => merge NA and RM into a new cluster called NA/RM L(NA/RM)=219 and m = 2

	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	295	268
MI/TO	877	295	0	564
NA/RM	255	268	564	0

min d(i,j) = d(BA,NA/RM) = 255 => merge BA and NA/RM into a new cluster called BA/NA/RM
L(BA/NA/RM)=255 m = 3

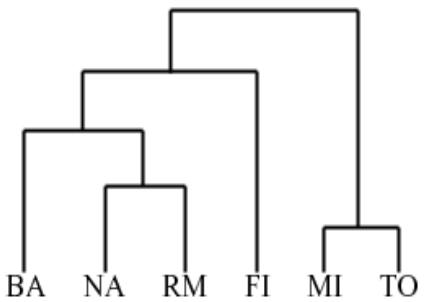
	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	295
MI/TO	564	295	0

min d(i,j) = d(BA/NA/RM,FI) = 268 => merge BA/NA/RM and FI into a new cluster called BA/FI/NA/RM L(BA/FI/NA/RM)=268 m = 4

	BA/FI/NA/RM	MI/TO
BA/FI/NA/RM	0	295
MI/TO	295	0

Finally, we merge the last two clusters at level 295.

The process is summarized by the following hierarchical tree:



The above example shows the single linkage algorithm which requires more number of iterations , computation time and complex. So to get fast results than the above we have introduced a new method of fixed point theorem in D metric spaces from applied mathematics by Indian mathematician B.C.Dhage who has submitted his thesis to Maratwada University in 1984 which is used to calculate the perimeter of the triangle.(1-11)

3. DEFINITION:

Let X be a non-empty set. R denotes the real numbers.A real valued function d on X * X * X is said to be a D-metric on X if

1. To each pair of distinct points x,y in X, there exists a point z in X such that d(x,y,z) ≠ 0.
2. d(x,y,z) = 0 when at least two of x,y,z are equal.(coincidence).
3. d(x,y,z)=d(y,x,z)=d(y,z,x) for all x,y,z in X symmetry and
4. d(x,y,z)≤d(x,y,w)+d(x,w,z)+d(w,y,z) for all x,y,z ,w in X.

If *X* is a nonempty set and *d* is a *D*-metric on *X*, then the ordered pair (*X*,*d*) is called a *D*-metric space.When the *D*-metric is understood, *X* itself is called a *D*-metric space.(1-11)

Here we present triple linkage clustering algorithm followed by an example instead of using single linkage

1. Begin with disjoint clustering having level L(0) =0 and sequence number m=0
2. Find the least dissimilar triplet of clusters in the current clustering, say triplet (r),(s),(t), according to d[(r),(s),(t)]=min d[i,j,k], Where the minimum is over all triplets of clusters in the current clustering.
3. Increment the sequence number; m=m+1. Merge clusters (r),(s) and (t) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) =d[(r),(s),(t)].
4. Update the values d (i,j,k) by considering the newly formed compound cluster in step 3. The proximity between the new cluster, denoted (r,s,t) and old clusters (k) and (l) is defined in this way d[(k),(l),(r,s,t)]= min{[d[(k),(l),(r)],d[(k),(l),(s)], (k),(l),(t)]}.
5. If all objects are in one cluster, stop. Else, go to step 2.
6. If there are only two clusters, namely (r) and (s), then we use usual Euclidian metric d[(r),(s)] to merge the two clusters (r) and (s).

4. AN EXAMPLE

Let us consider a simple example in which a hierarchical clustering of distance in kilometers between some Italian cities.

In this method, we use triple-linkage clustering using D-metric spaces

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

Input distance matrix

The nearest triplet of cities is MI, TO and FI, at distance (perimeter of the triangle joining the three cities) 833. These are merged into a single cluster "MI/TO/FI". The level of the new cluster is L(MI/TO/FI)=833 and the new sequence number is m=1.

Then we compute the distance (perimeter) from this new compound object to all other pairs of objects. In triple-linkage clustering the rule is that the perimeter of the triangle formed by the compound object with another pair of objects is equal to the smallest perimeter values of the triangles formed by each member of the compound cluster with the pair of outside objects. Here the perimeter of the triangle formed with MI/TO/FI and the pair RM and NA is chosen to be 955, which is the perimeter of the triangle formed with FI, RM and NA. Similarly, we can compute the perimeters of the triangles formed with the compound cluster and with other pairs of clusters.

Perimeters of the Triangles with MI/TO/FI is one vertex:

	(RM, NA)	(RM, BA)	(NA, BA)
MI/TO/FI	955	1342	1385

Perimeters of the Triangles with RM is one vertex.

	MI/TO/FI, BA	MI/TO/FI, NA	(NA, BA)
RM	1342	955	886

Perimeters of the Triangles with NA is one vertex:

	MI/TO/FI, RM	MI/TO/FI, BA	(RM,BA)
NA	955	1385	886

Perimeters of the Triangles with BA is one vertex:

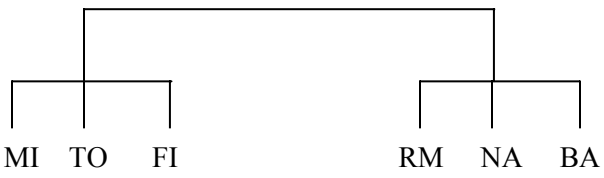
	MI/TO/FI, RM	MI/TO/FI, NA	(RM,NA)
BA	1342	1385	886

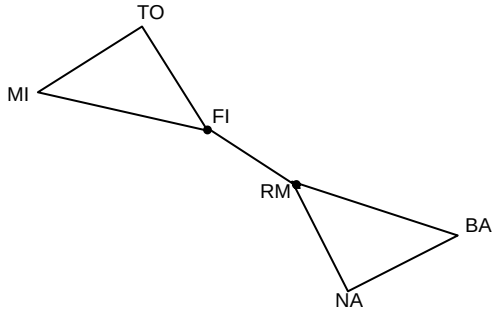
We can observe that the perimeter of the triangle formed with vertices RM,NA,BA is the smallest one. Hence we merge these three clusters into one compound cluster namely 'RM/NA/BA'. L(RM/NA/BA)=886 and m=2. Now we have only two (compound) clusters. We merge these two clusters using Euclidian metric.

D(MI/TO/FI,RM/NA/BA)=d(FI/RM)=268.

Merge MI/TO/FI with RM/NA/BA into a new cluster MI/TO/FI/RM/NA/BA.

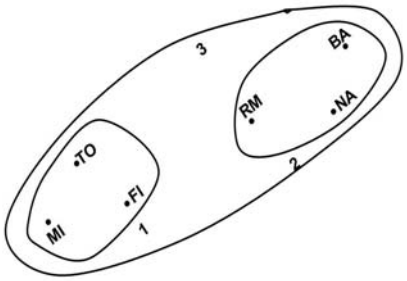
L(MI/TO/FI/RM/NA/BA)=268 and m=3.



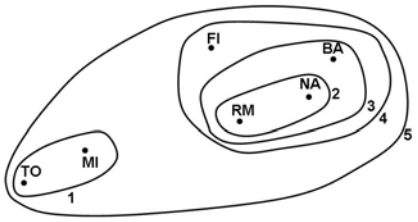


6. COMPARISON OF RESULTS:

Now we compare the results (clustering of six cities) of the same example using both methods one is single linkage and other one is triple linkage of D metric spaces. Following figures show the result of applying the single link ,triple link to our example data of six points . D-metric spaces is much better (faster) than the Euclidian metric (existing method) and triple linkage is faster than single linkage algorithm.. We are giving two types of results (1) Clustering of ellipses and (2) Dendrograms



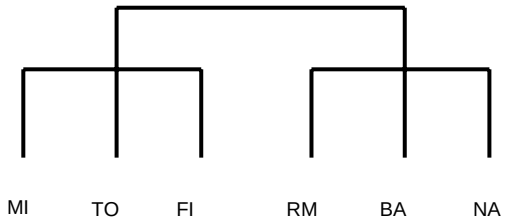
clustering using D Metric



7. CLUSTERING USING EUCLIDIAN METRIC

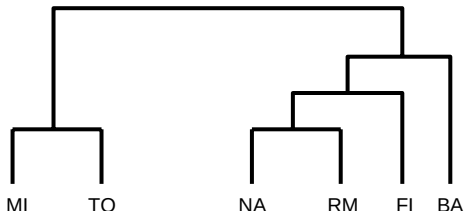
A hierarchical clustering is often displayed graphically using a tree-like diagram (Hierarchical tree) called a Dendrogram, which displays both the cluster-sub cluster relationships and the order in which the clusters were merged (agglomerative view). A hierarchical clustering can also be graphically represented using a nested cluster diagram.

. The Figures of above show the nested clusters as a sequence of nested ellipses, where the numbers associated with the ellipses indicate the order of the clustering. Figures below show the same information, but as **Dendrograms.**



Dendrogram using EUCLIDEAN

In view of the above diagrams, it is clear that the clustering D-metric takes less number of steps or iterations than the clustering using Euclidian metric and requires less computation time.



Dendrogram using Dmetric spaces

8. CONCLUSION

In this paper using the D–metric concepts, we designed a triple linkage algorithm in agglomerative clustering can be done much faster than the previous techniques.

Fixed point theorem in D metric spaces concept we have used and observed that it is a very use full concept in clustering. Many clustering algorithms (algorithms that are designed using the Euclidian metric) can be modified using this and can get the results much faster.

REFERENCES:

[1].A general existence principle for fixed point theorems for Dmetric spaces. B.C.Dhage ,A.M.Pathan and B.E.Rhoades. Internat. Journal of math&math scie. vol.19.no.3(1996) 457-460.

[2].On convergent sequences and fixed point theorems in D-metric spaces by SVR Naidu,KPR Rao,NS Rao. www.downloads. hindawi.com/journals /ijmms/2005/37687.pdf

[3].A fixed point theorem in generalized D metric spaces by Y.J.Cho and R.Saadati a bulletin of Iranian mathematical society vol 32.no. 2(2006)pp 13-19. www.ims.ir/publications/ bulletin/3202.pdf

[4].Fixed point theorems in generalized metric spaces by Han goo jung Department of mathematics and statics Graduate school Changwon national University.

[5].A Fixed point theorem in Generalized metric spaces by pratulananda das and lakshmi kanta dey. Soochow Journal of Mathematics vol 33 no.1 PP 33-39 Jan2007.

[6].Some fixed point theorems in T metric spaces Nabiolla Shobkolaci Dept of Mathematics Islamic azad University Iran. Proceedings of 5th Asian Mathematical Conference Maleysia 2009.

[7].Variational Principles,minimization theorems and fixed point theorem on generalized metric spaces J.S.UME2 communicated my M.J.Balas. Journal of optimization theory and applications Vol 118,No 3 pp 619-633 sept.2003. Introduction to metric fixed point theory M.A.Khamsi .International workshop on nonlinear Functional analysis and its applications shahid beheshti University Jan 20-24 2002.

[8].Fixed point theorem in Metric spaces by Andrew F sound B.S California. 8.On fixed point theorems in D metric spaces Seong – Hoon Cho. International Journal of Math .Analysis Vol 1.2007,no 22 1059-1065.

[9].Commentationes,Mathematicae Universitatis Carolinae Jun iti Nagata Vol.29(1988).No.4 715-722Materializability, generalized Metric spaces and g functions.

[10]. A new approach to generalized metric spaces Zead Mustafa & Brailey Sims Journal of nonlinear and convex nanlysis.vol-7 no.2 2006 289-297.

[11]. On generalized metric spaces and topological structures II B.C.Dhage .Pure and Applied Mathematika Sciences. Vol xxxx no 1-2 Sept 1994.

[12]. Clustering Gene Expression Data. The good,the bad, and the misinterpreted. Elizabeth,Garrett Mayer April 19,2004.

[13]. Fisher D: Iterative optimization and simplification of hierarchal clustering Technical Report CS-95-01, Department of Computer Science Vanderbilt University, 1995

[14]. Hong J and Mao. C : Incremental Discovery of rules and structure by Hierarchal and parallel clustering, Knowledge discovery I databases, pp. 177-193. Cambridge, MA; AAAI/MIT Press

[15]. Karypis. G, Han.E.H. and Kumar R.V : A hierarchical clustering algorithm using dynamic modeling, computer,32, pp.68-75,1999.

[16]. S.C Johnson (1967): "Hierarchical Clustering Schemes" *Psychometrika*, 2:241-254

[17]. R.D'andrade (1978): "U-Statistic Hierarchical Clustering" *Psychometrika*, 4:58-67

[18]. Andrew Moore: "K-means and Hierarchical Clustering–Tutorial Slides" http://www-.cs.cmu.edu/~awm/tutorials/ kmeans.html

[19]. Osmar R.Zaïane: "Principles of Knowledge Discovery in Databases – Chapter 8: Data Clustering"http://www.cs.ualberta.ca/~zaiane/cources/cmut690/slides/Chapter8/index.html

[20]. Stephen P.Borgatti: "how to explain hierarchical clustering"http://www.analytictech.com/networks/hiclus.html

[21]. Agarwal C., Procopiuc C., Wolf J.L., Yu P.S., and Park J.S A frame work for finding projected clusters in high dimensional spaces. In *proceedings* of *ACM SIGMOD International Conferences on Management of data,* 1999

[22]. Ankerst M., Breunig M., H.-P. Kriegel, and Sander J. OPTICS: Ordering points for identify the clustering structure. In *Proceedings ACM SIGMOD International Conference on Management of Data*, Philadelphia, PA, 1999

[23]. Han Eui-Hong (Sam), Karypis George, Kumar Vipin and Mobasher B.. Clustering based on association rule hypergraphs. In *SIGMOD'97 Workshop on research Issues on Data Mining and knowledge Discovery*, 1997

[24]. Han E.H., Karyapis G., Kumar V., and Mobasher B. Clustering in a high-dimensional space using hypergraph models. *Technical Report 97-019*, Department of Computer Science, University of Minnesota.

[25]. Li C., and Biswas G. Conceptual clustering with numeric and nominal mixed data – A new similarity based system. *IEEE Transactions on Knowledge and Data Engineering*, in review, July 1996.