An Experimental Survey on Single Linkage Clustering

Krishna K. Mohbey
Maulana Azad National Institute of Technology, Bhopal

G. S. Thakur
Maulana Azad National Institute of Technology, Bhopal

ABSTRACT

Clusters are useful to identify required object from the huge amount of datasets. There are lots of clustering methods, used to create clusters. Single linkage clustering method is an example of hierarchical agglomerative clustering which is used to merge objects in a cluster, based on minimum distance. In this paper we performed an experiment on two dimensional spaces where multiple objects are available and combine in clusters by Euclidean distance. In this paper, MATLAB is used to calculate the distance between two objects and constructing distance matrix. After completing the whole single linage clustering method dendogram has been prepared. This dendogram is similar to minimum spanning tree because it is prepared using minimum distance of objects. These prepared clusters and dendogram are useful for finding different knowledge from the huge data.

General Terms

Clustering, Euclidean distance, Distance matrix

Keywords

Single Link clustering, Similarity measurements, Dendogram.

1. INTRODUCTION

Today the whole world depends on the data because it is our need. Every one store the large amount of data for different purpose and uses for future analysis and management[3]. Peoples have the requirement to categorize these data into different sets according to their need, is called clustering. This categorization process is mostly done by the people on the bases of similarities or dissimilarities based on the some rules or standards. The process of data classification may be supervised or unsupervised it depends on whether they assign new inputs to one of a finite number of discrete supervised classes or unsupervised categories, respectively [11] [12] [4]. In supervised classification, the mapping from a set of input data vectors to a finite set of discrete class labels is modeled in terms of some mathematical function f=y(A,B), where B is a vector of adjustable parameters. The values of these parameters are determined (optimized) by an inductive learning algorithm (also termed inducer), whose aim is to minimize an empirical risk functional (related to an inductive principle) on a finite data set of input–output examples[11]. Clustering or exploratory is an unsupervised classification data analysis process in which no labeled data are available [7][9]. The goal of clustering is to separate a finite unlabeled data set into a finite and discrete set of objects rather than provide an accurate characterization of unobserved samples generated from the same probability distribution [13] [12]. Fig.1 shows the clustering process which is applied on datasets. Backer and Jain [10] defined that “in clustering process an document or group of object is divided into a number of more or less similar subgroups on the bases of similarity measurement”. There are lots of clustering algorithms available today which are used to partition data into a certain number of clusters/groups/categories. Most researchers describe a cluster by considering the object similarity and the external separation [8], [9], i.e. patterns in the same cluster should be similar to other while patterns in different clusters or groups will be dissimilar. Both the similarity and the dissimilarity should be examinable in a clear and meaningful way. Clusters are useful to search relevant document from the available groups and it provides the searching efficiency because it is applied on the smaller collection instead of whole document collections. There are different clustering techniques each have their own working procedure, performance and issues. Some Clustering techniques include on [3] [4][6][7][9] references.

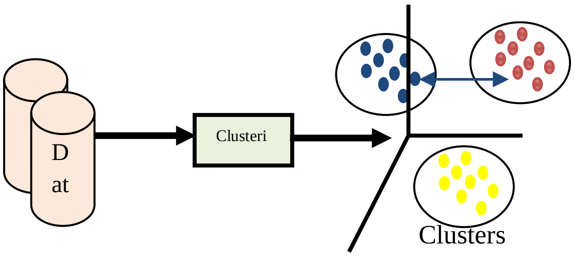

Fig 1: Clustering process

The purpose of this paper is to provide a comprehensive and systematic description about the clustering algorithms which is mostly useful in statistics, computer science, machine learning and other areas. In this paper, we mainly focus on the single linkage clustering technique which is the part of hierarchical clustering. Before applying clustering method it is important to use data mining preprocessing activities such as stemming and removing stop words [16]. For preprocessing of data we can also use predefined algorithms.

2. CLUSTERING METHODS

Clustering is a process of dividing objects into multiple groups using their similarity or dissimilarity. There are various methods for data clustering, each have their own significances. Some important clustering methods are as follows.

2.1 Distance and Similarity Measures

It is important to find out the distance or similarity of objects for clustering processes. There are different formulas for distance calculations some are given below.

2.1.1 Minkowski distance

$$\boldsymbol{dist} = \left(\sum_{\boldsymbol{k}=1}^{\boldsymbol{n}} |\boldsymbol{p_k} - \boldsymbol{q_k}|^{\boldsymbol{r}}\right)^{\frac{1}{\boldsymbol{r}}}$$

2.1.2 Euclidean distanc

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

2.1.3 Pearson correlation

$$\rho(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

2.1.4 Cosine similarity

$$\cos(x, y) = \frac{x \cdot y}{|x||y|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

2.2 Hierarchical Clustering

These methods can be divided as follows.

2.2.1 Agglomerative Hierarchical Clustering

This is a bottom up approach. In this method all objects are initially represents to own cluster. Then clusters are merged until desire cluster structure is obtained.

2.2.2 Divisive Hierarchical Clustering

This is a top down approach. In this method all objects initially belongs to one cluster. Then clusters are divided into their own sub clusters. This process continue until desire cluster structure is obtained.

2.3 Partitioning based Methods

In these methods objects are relocates by moving them from one cluster to another. In these methods it is require that the number of clusters will be pre-set by the user. The examples of such type of clustering are : K-medoids, K-mean, Probabilistic and Density based.

3. SINGLE LINKAGE CLUSTERING

It is a type of agglomerative hierarchical clustering method. It works on bottom-up strategy in which each point compared with others. Each object is placed in a separate cluster, and at each step we merge the closest pair of clusters, until certain termination conditions are satisfied. Single linkage clustering method can find arbitrary shaped cluster in many applications such as image segmentation, spatial data mining, geological mapping etc. It builds a dendogram where each level represents a clustering of the dataset[2]. For the single link, the nearest distance of two clusters is defined as the minimum of the distance between any two points in the two clusters. By the graph terminology, if we start with all points, each one a separate cluster on its own called a singleton cluster and then add links between all points one at a time using shortest links first, and then these single links combines the points into clusters. (i.e. the points with the shortest distance between each other are combined into a cluster first, and then the next shortest distances are combined, and so on). Fig. 3 to 5 shows the different distance values between objects.

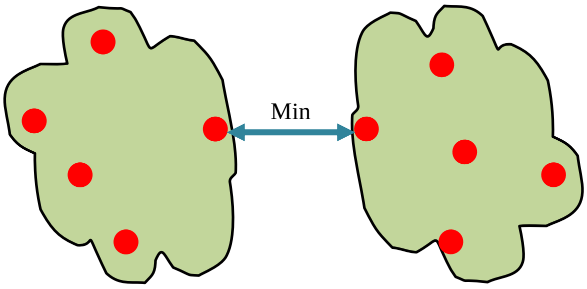

Fig 2: Minimum distance of two clusters for single linkage clustering

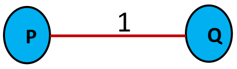

Fig 3: Distances or thresholds=1

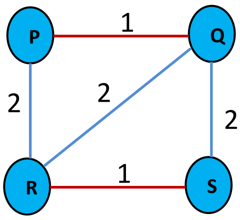

Fig 4: Distances or thresholds=2

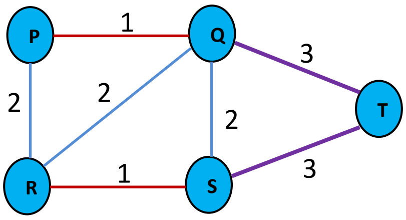

Fig 5: Distances or thresholds=3

3.1 Dendogram

It is used to show the same information as the graph, however distance or threshold are in vertical and points or objects are at the horizontal axis [5]. The height at which two clusters are merged in the dendogram reflects the distance of two clusters.

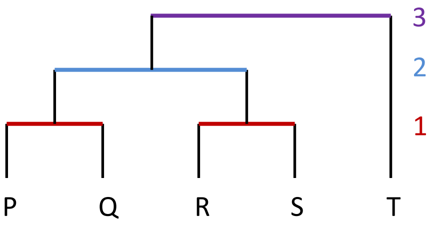

Fig 6: Dendogram example

3.2 Distance Matrix

A distance matrix is a matrix or two-dimensional array contains the distances of two pairs from a set of points. This matrix will have a size of N×N where N is the number of points in a graph [19].

3.3 Euclidean Distance Matrix

An N×N matrix D = [d_{ij}] is called a Euclidean distance matrix [18], if there exists a set of n vectors, say {$x_1 \ldots x_n$}, in a finite dimensional inner product space such that

$$d_{ij} = |x_i - x_j|^2$$

4. ILLUSTRATIVE EXAMPLE

Single linkage clustering prepare clusters by calculating minimum distance between data points. Here we have taken 7 data points on the 2 dimensional space and prepare clusters. These data points are p1,p2,p3..p7 are shown in table 1. We have calculated all distances using MATLAB. The complete procedure for the single linkage clustering has describe below.

Table 1. Dataset with 7 objects

Object	X	Y
p1	0.40	0.50
p2	0.22	0.35
p3	0.45	0.40
p4	0.20	0.25
p5	0.42	0.10
p6	0.30	0.55
P7	0.45	0.32

4.1 Plotting Objects in 2D space

Here we plot attributes x and y in n dimensional space. In above dataset p1, p2, p3..p7 are 7 objects and x and y shows 2 dimensional space.

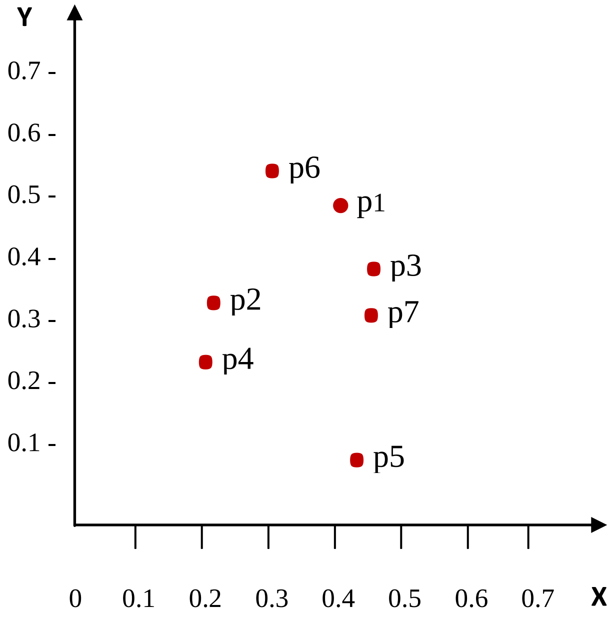

Fig 7: Objects in 2 Dimensional space

4.2 Calculating Distance Matrix

In this step we calculate a distance matrix. This distance matrix is prepared by calculating the distance from each object to all other using Euclidean distance measure.

Distance between two points i and j is

where x_{i1} is the value of attribute 1 for i and x_{j1} is

$$d\,(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j1}|^2 + \ldots + |x_{in} - x_{jn}|^2}$$

$$d\,(p1, p2) = \sqrt{|x_{p1} - x_{p2}|^2 + |y_{p1} - y_{p2}|^2}$$

$$= \sqrt{|0.40 - 0.22|^2 + |0.50 - 0.35|^2}$$

$$= \sqrt{|0.18|^2 + |0.15|^2}$$

$$= \sqrt{0.0324 + 0.0225} \quad = \sqrt{0.0549}$$

$$= \mathbf{0.2343}$$

Similarly we calculate the distance between all points. Here we have calculated distance matrix using MATLAB which is shown in Table 2.

Table 2. Distance matrix

	p1	p2	p3	p4	p5	p6	p7
p1	0						
p2	0.2343	0					
p3	0.1118	0.2354	0				
p4	0.3202	0.1020	0.2915	0			
p5	0.4005	0.3202	0.3015	0.2663	0		
p6	0.1118	0.2154	0.2121	0.3162	0.4657	0	
p7	0.1868	0.2319	**0.0800**	0.2596	0.2220	0.2746	0

4.3 Assigning Objects into cluster

Now we select two objects with the shortest distance in the matrix and then merge them together.

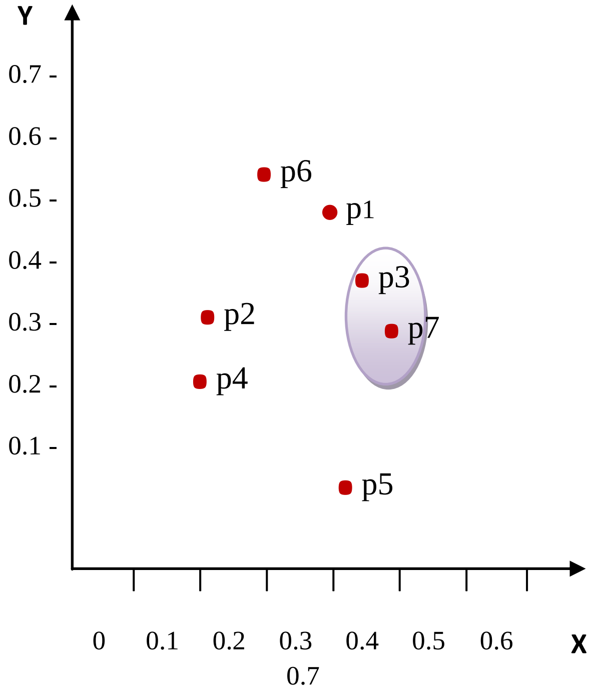

Fig 8: Objects in 2 Dimensional space with one cluster

In Table 2 shortest distance points are p3 and p7. So that we merge these two points in one cluster as shown in Fig. 8. By combining (p3, p7) together in a single cluster it became one entry. We again recalculate the distance from each point to new cluster (p3, p7). In single link method the proximity of two cluster is defined as the minimum distance between two clusters therefore the distance of (p3,p7) and p1 will calculated as-

dis {(p3,p7),p1}=Min{distance(p3,p1),distance(p7,p1)}

=Min {0.1118, 0.1868}

=**0.1118**

Similarly we calculate the distance between (p3, p7) to p2, p4, p5 and p6.

Table 3. Distance matrix for (p3, p7) to all points

	p1	p2	(p3,p7)	p4	p5	p6
p1	0					
p2	0.2343	0				
(p3,p7)	0.1118	0.2319	0			
p4	0.3202	**0.1020**	0.2596	0		
p5	0.4005	0.3202	0.2220	0.2663	0	
p6	0.1118	0.2154	0.2121	0.3162	0.4657	0

Similarly we repeat above steps until all the objects are not assigned into appropriate clusters. In above matrix the smallest distance is **0.1020** between p2 and p4, so these points are merging together.

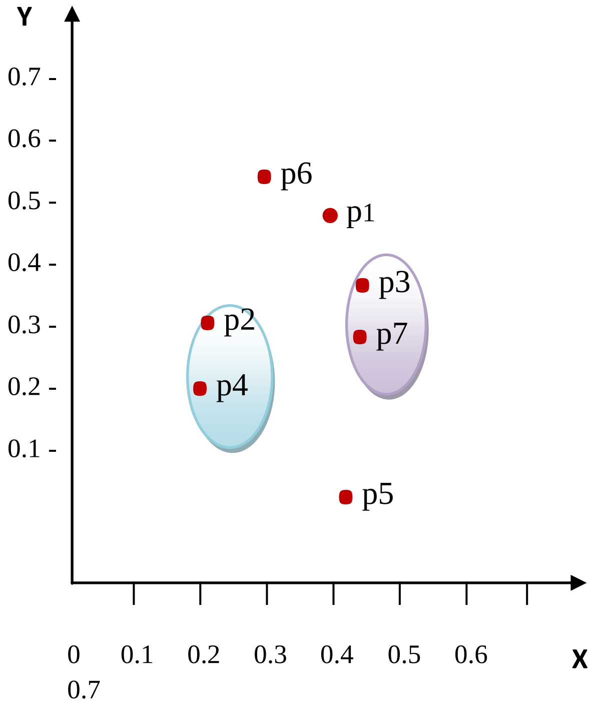

Fig 9: Objects in 2 Dimensional space with two clusters

Similarly the points p2 and p4 are merged together and prepare a cluster which is shown in Fig. 9. This whole process of recalculating distance matrix and combining objects into clusters are repeated. Thus we find final clusters as shown in Fig. 10.

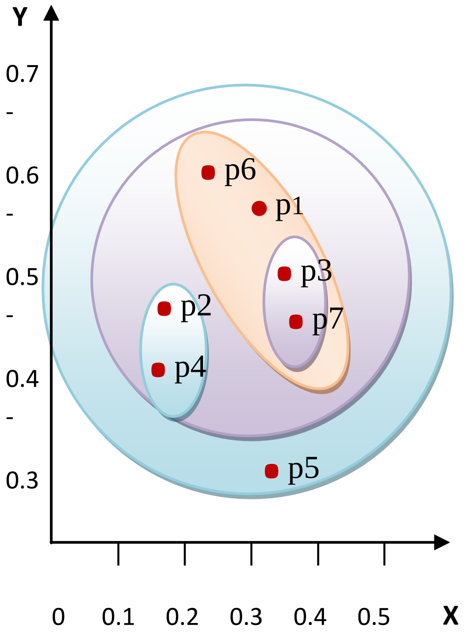

Fig 10: Objects in 2 Dimensional space with final clusters

The various clusters prepared by above single linkage clustering method can also be displayed as a dendogram which is shown in Fig.11.

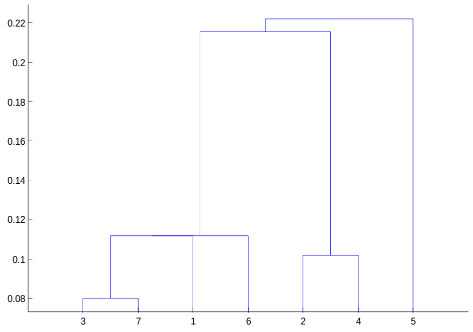

Fig 11: Dendogram of the obtained clusters

Here we would like the data partitioned into several clusters for unsupervised learning. Therefore the process required to stop clustering at some point – either the user will specify the number of clusters he would like to have, or the process has to make a decision on its own. In above example if we stop at the threshold (distance) 0.12 then we have {(p3, p7, p1, p6), (p2, p4), (p5)} that means 3 clusters. But if we stop at threshold 0.1, we have only cluster (p3, p7) and (p2, p4).

5. CONCLUSION

This paper describes the process of making clusters from the given datasets using single link hierarchical agglomerative clustering. Distance calculations are preformed using MATLAB, which provide the fast result. After applying whole method we finally obtain number of clusters and dendogram which can be used as a minimum spanning tree for searching or other purposes. In the future research we plan to design a framework for fast searching that reduces the search time and complexity. Further we can apply this clustering method on the different kind of datasets.

6. REFERENCES

[1] Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE, "Survey of Clustering Algorithms", IEEE Transactions on Neural Networks Vol. 16, No. 3, May 2005.

[2] Bidyut kr. Patra, Sukumar Nandi, P.Viswanath, "A distance based clustering method for arbitrary shaped clusters in large datasets", Pattern Recognition, 44 (2011), 2862-2870.

[3] M. Anderberg, "Cluster Analysis for Applications", New York: Academic,1973.

[4] R. Duda, P. Hart, and D. Stork, "Pattern Classification", 2nd ed. NewYork: Wiley, 2001.

[5] Jin Chen, Alan M. MacEachren and Donna J. Peuquet, "Constructing Overview + Detail Dendrogram-Matrix Views ", IEEE Transactions on Visualization and Computer Graphics, Vol .15, No.6 , Nov 2009.

[6] B. Duran and P. Odell, "Cluster Analysis: A Survey", New York: Springer-Verlag, 1974.

[7] B. Everitt, S. Landau and M. Leese, "Cluster Analysis", London: Arnold, 2001.

[8] P. Hansen and B. Jaumard, "Cluster analysis and Mathematical programming", Math. Program., vol. 79, pp. 191–215, 1997.

[9] A. Jain and R. Dubes, "Algorithms for Clustering Data", Englewood Cliffs, NJ: Prentice-Hall, 1988.

[10] E. Backer and A. Jain, "A clustering performance measure based on fuzzy set decomposition", IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-3, no. 1, pp. 66–75, Jan. 1981.

[11] C. Bishop, "Neural Networks for Pattern Recognition", New York: Oxford Univ. Press, 1995.

[12] V. Cherkassky and F. Mulier, "Learning From Data: Concepts, Theory, and Methods", New York: Wiley, 1998.

[13] A. Baraldi and E. Alpaydin, "Constructive feedforward ART clustering networks" -Part I and II, IEEE Trans. Neural Network , vol. 13, no. 3, pp. 645–677, May 2002.

[14] D.S Rajput, R.S. Thakur, G.S. Thakur, "Rule Generation from Textual Data by using Graph Based Approach", Published in International Journal of Computer Application (IJCA) 0975 – 8887, New york USA, ISBN: 978-93-80865-11-8, Volume 31– No.9, October 2011.

[15] D.S Rajput, R.S. Thakur, Neeraj Sahu, G.S. Thakur, "Analysis of Social networking sites using K-mean Clustering algorithm", Presented in International Conference on Computer Science and Information Technology (CSIT 2012) March 3rd, 2012

[16] Ghanshyam Thakur, Rekha Thakur and R.C. Jain, "Association Rule Generation from Textual Document", International Journal of Soft Computing, 2: 2007, pp. 346-348.

[17] R.S. Thakur, R. C. Jain, K.R. Pardasani , "Graph Theoretic Based Alogorihtm for mining frequent Pattern" International Joint Conference on Neural Networks (IJCNN 2008), pp 628-632.

[18] R. Balaji And R.B. Bapat, "Block Distance Matrices" , Electronic Journal of Linear Algebra ISSN 1081-3810 A publication of the International Linear Algebra Society Volume 16, pp. 435-443, December 2007.

[19] http://en.wikipedia.org/wiki/Distance_matrix.