CLUSTERING ANALYSIS

Romli*

ABSTRACT

CLUSTERING ANALYSIS. Cluster analysis is the name of a group of multivariate techniques whose principal purpose is to distinguish similar entities from the characteristics they process. To study this analysis, there are several algorithms that can be used. Therefore, this topic focuses to discuss the algorithms, such as, similarity measures, and hierarchical clustering which includes single linkage, complete linkage, and average linkage methods. Also, non-hierarchical clustering method, which is popular name "K-mean method", will be discussed. Finally, this paper will be described the advantages and disadvantages of every methods.

ABSTRAK

Analisa pengelompokan adalah salah satu bagian dari teknik multivariat, yang tujuan utamanya adalah menentukan variabel-variabel yang mempunyai sifat-sifat similar untuk pemrosesan. Untuk mempelajarinya, ada beberapa algoritma yang digunakan. Untuk itu, tulisan ini memfokuskan pada algoritma yang digunakan, seperti pengukuran similar, dan pengelompokan yang hirarki yang meliputi hubungan tunggal, hubungan komplet, dan hubungan rata-rata. Pengelompokan yang tidak hirarki dengan nama yang popular metoda "K-mean" juga diuraikan di tulisan ini. Akhirnya, didiskusikan beberapa kelebihan dan kekurangan pada setiap metoda untuk penggunaannya.

INTRODUCTION

The name of a group of multivariate techniques whose principal objective is to identify similar entities from the characteristics they process, called cluster analysis. They identify and classify objects or variables so that each object is very similar to the others in the cluster which features some resolved acceptance criterion (Aldenderfer and Blashfield, 1984; Everitt, 1980; Hartigan, 1975; Johnson and Wichern, 1992; Manly, 1986).

This is a helpful tool for data analysis in many dissimilar situations (Johnson and Wichern, 1992). For example, a researcher who has composed data by means of a questionnaire may be faced with a large number of observations that are incomprehensible unless classified into manageable groups. Cluster analysis can be

* UPT-LAGG, BPP Teknologi

used to achieve this data reduction objectively by reducing information from the entire population or set of information about specific smaller subgroups (Arabie, Carroll and DeSarbo, 1987).

Cluster analysis is also useful when a researcher wishes to develop hypotheses concerning the nature of the data or to examine previously stated hypotheses (Anzai, 1992). For example, a researcher may believe that attitudes toward the consumption of diet versus regular coca cola could be used to separate coca cola consumers into logical segments or groups. Cluster analysis can be used to classify coca cola consumers by their attitudes about diet versus regular coca cola, and the resulting clusters, if any, can be profiled for demographic similarities or differences (Sinsich, 1993; Weimer, 1993).

Cluster analysis as defined by several authors is the generic name for a wide variety of procedures that can be used to create a classification.

In addition to this, many proposed definitions contain statements such as a cluster is a set of entities that are alike, and entities from different clusters are not alike (Kodratoff, 1988). Others suggest that entities within a cluster are more similar to each other than the entities from other clusters.

Also, the terms of cluster, type, group, class, etc., have been used in an intuitive sense without attempting any formal definition. However, several attempts have been made to define these terms: for instance in their Dictionary of Statistical Terms, Kendell and Buckland define the term cluster as follows :

> Cluster - a group of contiguous elements of a statistical population; for example, a group of people living in a single house, a consecutive run of observations in an ordered series, or a set of adjacent plots in one part of a field.

A description of cluster that most closely agrees with our intuitive understanding of the term is given by considering entities as points in a p-dimensional space, with each of the p variables being represented by one of the axis of this space. For variable values for each entity now define a p-dimensional coordinate in this space. Clusters may now be described as continuous regions of this space containing a relatively low density of points. Clusters described in this way are sometimes referred to as natural clusters. This description matches the way we detect clusters visually in two or three dimensions (Arabie, Carroll and DeSarbo, 1987; Jain, 1991).

To study this topic, four principal goals need addressing as follows: (1) development of a typology or classification, (2) investigation of useful conceptual schemes for grouping entities that are used for this topic, (3) hypothesis generation through data exploration, and (4) hypothesis testing, or the attempt to determine if types defined through other procedure are infect present in a data set (Lauriere, 1990; Manly, 1986).

ALGORITHMS THAT CAN BE USED

The technique of clustering produces the classification from unclassified data. It is difficult to analyse if the data are multi variable and multi purpose. So, when the data is classified, it is easy to measure or to analyse.

Similarity Measures

The following equations are used in this methods:
The similarity distance between points x and y denoted by d(x , y) is defined as

$$d(x, y) = (a + b) / N \qquad (1)$$

where : a is the number of matches in term of 1-1
b is the number of matches in term of 0-0
N is the number of entities.

The Euclidean distance can be used which has a specific rule in regards to binary variable :

$$(x_{ij} - x_{kj})^2 = 0, \text{ if } x_{ij} = x_{kj} = 1 \text{ or } x_{ij} = x_{kj} = 0 \qquad (2)$$
$$= 1, \text{ elsewhere;}$$

where : x_{ij} is the score (0 or 1) of the j^{th} binary variable on the i^{th} item.
x_{kj} is the score (0 or 1) of the j^{th} binary variable on the j^{th} item.
$j = 1,2,3, \ldots p.$

The algorithm of this method is as follows :

1. Compute the distance for every variable of the data set with using Euclidean equation (equation 1).
2. Find the maximum distance.
3. Compute the similarity distance (equation 2).
4. Alternate steps 1 to 3 until no data points.

Example :

ind-1		0	0	1	1	1
ind-2	1	1	1	0	1	0
ind-3	0	1	0	1	1	0
ind-4	0	0	1	0	1	0
ind-5	1	1	1	0	0	0

1. Find the similarity distances using equation 1.:

ind-1	0	0	0	1	1	1
ind-2	1	1	1	0	1	0

d(ind-1 , ind-2) = (1+0)/6 = 1/6

Similarly, we calculate the others.

	ind-1	ind-2	ind-3	ind-4	ind-5
ind-1	0				
ind-2	1/6	0			
ind-3	4/6	3/6	0		
ind-4	3/6	4/6	3/6	0	
ind-5	0	**(5/6)**	2/6	3/6	0

2. Find maximum distance

d(ind-5 , ind-2) is the maximum distance.

3. Find similarity value using equation 2. :

ind-2	1	1	1	0	1	0
ind-5	1	1	1	0	0	0
ind-2-5	0	0	0	0	1	0

4. In a similar way, we conclude:

	ind-1	ind-3	ind-4	ind-2-5
ind-1	0			
ind-3	4/6	0		
ind-4	3/6	3/6	0	
ind-2-5	4/6	4/6	**(5/6)**	0

	ind-1	ind-3	ind-2-5-4
ind-1	0		
ind-3	**(4/6)**	0	
ind-2-5-4	2/6	2/6	0

	ind-1-3	ind-2-5-4
ind-1-3	0	
ind-2-5-4	**(3/6)**	0

5. Result :

The most similar measure	0.83333	(ind-2 , ind-5)
The most similar measure	0.83333	(ind-2 , ind-5) ind-4
The most similar measure	0.66667	(ind-1 , ind-3)
The most similar measure	0.50000	(ind-1, ind-3) (ind-2, ind-5, ind-4)

Hierarchical Clustering Methods

To study this technique, there are three methods that can be used :

1. Single Linkage or The Nearest Neighbour Method
2. Complete Linkage or The Furthest Neighbour Method
3. Average Linkage Method.

Single Linkage Method

The equation used in this method is the distance between the two groups d_{uw} and d_{vw} denoted by $d_{(uv)w}$ is defined as :

$$d_{(uv)w} = \min\{d_{uw}, d_{vw}\} \tag{3}$$

where : d_{uw} is the distance between two points u and w.
d_{vw} is the distance between two points v and w.

The algorithm of this method is as follows :

1. Find the minimum distance.
2. Compute the distance (equation 3).
3. Alternate steps 1 to 2 until no data points.

Example :

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	**(2)**	8	0

1. Find minimum distance.

d(3 , 5) is the minimum distance.

2. Calculate the distance using equation 3:

$$d(35)1 = \min \{d(31) , d(51)\}$$
$$= 3$$

3. In a similar way, we can calculate the others.

	1	2	4	35
1	0			
2	9	0		
4	6	5	0	
35	**(3)**	7	8	0

	2	4	351
2	0		
4	**(5)**	0	
351	7	6	0

	24	351
24	0	
351	**(6)**	0

4. Result

Minimum distance	2	(3 5)
Minimum distance	3	(3 5) 1
Minimum distance	5	(2 4)
Minimum distance	6	(2 4) (3 5 1)

Complete Linkage Method

The equation used in this method is the distance between the two groups d_{uw} and d_{vw} denoted by $d_{(uv)w}$ is defined as:

$$d_{(uv)w} = \max\{d_{uw} , d_{vw}\} \quad (4)$$

where: d_{uw} is the distance between two points u and w.
d_{vw} is the distance between two points v and w.

The algorithm of this method is as follows:

1. Find the minimum distance
2. Compute the distance (equation 4)
3. Alternate steps 1 to 2 until no data points

Example :

	1	2	3	4	5
1	0				
2	9	0			
3	3	7.	0		
4	6	5	9	0	
5	11	10	**(2)**	8	0

1. Find minimum distance.

d(3 , 5) is the minimum distance.

2. Calculate the distance using equation 4:

$$d(35)1 = \max \{d(31), d(51)\}$$
$$= 11$$

3. In a similar way, we can calculate the others.

	1	2	4	35
1	0			
2	9	0		
4	6	**(5)**	0	
35	11	10	9	0

	1	24	35
1	0		
24	**(9)**	0	
35	11	10	0

	241	35
241	0	
35	**(11)**	0

4. Result

Maximum distance	2	(3 5)
Maximum distance	5	(2 4)
Maximum distance	9	(2 4) 1
Maximum distance	11	(2 4 1) (3 5)

Average Linkage Method

The equation used in this method is the distance between the two groups d_{uw} and d_{vw} denoted by $d_{(uv)w}$ is defined as:

$$d_{(uv)w} = \frac{\sum\sum d_{ik}}{N_{(uv)} N_w} \quad (5)$$

where : d_{ik} is the distance between object i in the cluster (UV) and k in the cluster W.

$N_{(uv)}$ is the number of items in the cluster (UV).

N_w is the number of items in the cluster W.

The algorithm of this method is:

1. Find the minimum distance.
2. Compute the distance (equation 5).
3. Alternate steps 1 to 2 until no data points.

Example :

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	**(2)**	8	0

1. Find minimum distance.

d(3 , 5) is the minimum distance.

2. Calculate the distance using equation 5

d(35)1 = {d(31) + d(51)}/{2.1}

= 7

3. In a similar way, we can calculate the others.

	1	2	4	35
1	0			
2	9	0		
4	6	**(5)**	0	
35	7	8.5	8.5	0

	1	24	35
1	0		
24	7.5	0	
35	(7)	8.5	0

	24	351
24	0	
351	**(8.16667)**	0

4. Result

Average distance	2.00000	(3 5)
Average distance	5.00000	(2 4)
Average distance	7.00000	(3 5) 1
Average distance	8.16667	(2 4) (3 5 1)

Non-Hierarchical Clustering Methods

The most popular method, which is proposed by Mac Queen, is also called "K-means methods". The procedure or algorithm of this method is as follows:

1. Begin with an initial partition of the data set into some specified number of clusters.
2. Compute the distance for every variable of the data set using Euclidean equation.
3. Find the minimum distance.
4. Compute the centroids of clusters.
5. Allocate each data point to the cluster that has the nearest centroid.

6. Compute the new centroids of the clusters; clusters are not updated until there has been a complete pass through the data.
7. Alternate steps 3 to 6 until no data points.

Example :

Program	CPU time	Disk I/O
A	2	4
B	3	5
C	1	6
D	4	3
E	5	2

1. Find distance using Euclidean formula :

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$$

where : x_n is the value nth variable for the x entity
y_n is the value nth variable for the y entity

$$d(A, B) = \sqrt{(2 - 3)^2 + (4 - 5)^2}$$
$$= \sqrt{2}$$

In a similar way, we can calculate the others.

	A	B	C	D	E
A	0				
B	(√2)	0			
C	√5	√5	0		
D	√5	√5	√18	0	
E	√13	√13	√32	(√2)	0

2. Find the minimum distance
d(A , B) and d(D , E) are the minimum distances

3. Calculate centroid in regard of minimum distance
Centroid AB = ((2 + 3) : 2 , (4 + 5) : 2)
= (2.5 , 4.5)
Centroid DE = ((4 + 5) : 2 , (3 + 2) : 2)
= (4.5 , 2.5)

4. In a similar way, we conclude :

	AB	C	DE
AB	0		
C	(√4.5)	0	
DE	√8	√24.5	0

	ABC	DE
ABC	0	
DE	(√12.5)	0

5. Result

Minimum distance	√2 = 1.414214	(A B)
Minimum distance	√2 = 1.414214	(D E)
Minimum distance	√4.5 = 2.121320	(A B) C
Minimum distance	√12.5 = 3.535534	(A B C) (D E)

DISCUSSION

Hierarchical clustering techniques were more popular. Hierarchical procedures do have the advantage of being fast and therefore taking less computer time. But they can be misleading because undesirable early combinations may persist throughout the analysis and lead to artifial results. Of specific concern is the substantial impact of outliers on hierarchical methods, especially with the complete linkage method. To reduce this possibility, the analyst may wish to cluster - analyse the data several times, each time deleting problem observations or outliers.

Non-hierarchical methods have gained increased acceptability and are found in an increasing number of applications. Their use, however, is dependent on the ability of the researcher to select the seed points based on some practical, objective, and theoretical basis. In these instances, non-hierarchical methods have several advantages over hierarchical techniques. The results are less susceptible to outliers in the data, the distance measure used, and the inclusion of irrelevant or inappropriate variables. These benefits are only realised, however, with the use of non-random (e.g. specified) seed points, and the use of non-hierarchical technique with random seed points is markedly inferior to the hierarchical techniques.

Another approach is to use both methods to gain the benefits of each. First, a hierarchical technique can be used to establish the number of cluster, profile the cluster centers, and identify any obvious outliers. After autliers are eliminated, the remaining observations are then clustered with a non-hierarchical method, using the

cluster centers from the hierarchical results as the initial seed points. In this way, the advantages of the hierarchical methods are complemented by the ability of the non-hierarchical methods to "fine-tune" the results by allowing the switching of cluster membership.

CONCLUSION

Cluster analysis can be a very useful data reduction technique. But since its application is more an art than a science, it can be easily be used (misapplied) by the analyst. Different interject measures and different algorithms can and do affect the results. The analyst needs to consider these problems and, if possible, replicate the analysis under varying conditions. If the analyst is cautious, cluster analysis can be very helpful in identifying latent patterns suggesting useful groupings (clusters) of objects.

REFERENCES

1. ALDENDERFER, MARK S. and BLASHFIELD, R.K.. "Cluster Analysis" in Quantitative Applications in the Social Sciences 44. Newbury Park: Sage publications Inc., (1984)
2. ANZAI, YUCHIRO. Pattern Recognition and Machine Learning. San Diego: Academic Press, Inc., 1992.
3. ARABIE, P., CARROLL J.D. and DESARBO, W.S.. "Three-way Scaling and Clustering" in Quantitative Applications in the Social Sciences 65. Newbury Park: Sage Publications Inc., (1987)
4. EVERITT, BRIAN. "Cluster Analysis", New York: John Wiley & Sons, Inc., (1980)
5. HARTIGAN, JOHN A.. "Clustering Algorithms", New York: John Wiley & Sons Inc., (1975)
6. JAIN, RAY. "The Art of Computer Systems Performance Analysis", New York: John Wiley & Sons Inc., (1991)
7. JONSON, R.A., and WICHERN D.W. "Applied Multivariate Statistical Analysis", Englewood Cliffs: Prentice Hall Inc., (1992)
8. KODRATOFF, YVES. "Introduction to Machine Learning" London: Pitman Publishing, (1988)

9. LAURIERE, JEAN-LOUIS. "Problem-Solving and Artificial Intelligence", New York: Prentice Hall, (1990)

10. MANLY, B.F.J. "Multivariate Statistical Methods a Primer", London: Chapman and Hall, (1986)

11. SINSICH, T. "Statistics by Example", New York: Macmillan Publishing Company, (1993)

12. WEIMER, R.C. "Statistics", Belmont: Wim C. Browm Publishers, (1993)

DISKUSI

M. SYAMSA A.

1. Aplikasi apa yang dapat dilakukan pada *clustering analysis* dengan pengelompokan hirarki ini?
2. Mengapa tidak dijelaskan peranan metode *K-mean* dalam pengelompokan non-hirarki?

ROMLI

1. Aplikasi yang sudah dan sedang dilakukan adalah *Command* di *UNIX* dan optimasi *Novell Netware*. Aplikasi yang lain dapat dilakukan apabila sudah diketahui variabel-variabelnya.
2. Pada prinsipnya pengelompokan non-hirarki sama dengan metode *K-mean*. Pada waktu yang akan datang akan dicoba metode tersebut, dibandingkan dengan metode pada *expert system*, kemudian baru akan dibahas peranannya.

KARSONO

Apakah metode ini dapat diaplikasikan pada data tidak bulat (bukan *binary system*)?

ROMLI

Metode hirarki dapat diaplikasikan pada data tidak bulat asal diketahui jaraknya, dan metode non-hirarki merupakan alternatif yang baik. Walaupun dalam makalah ini diberikan contoh bilangan bulat, tapi dapat diterapkan untuk bilangan riil. Sudah dilakukan untuk *command* pada *Unix System*. Metode ini tidak dapat diterapkan untuk *similary measures*, dan untuk keperluan ini disarankan untuk menggunakan *Fuzzy logic*.