Theoretical framework of cluster analysis

Identifying groups of individuals or objects that are similar to each other but different from individuals in other groups, cluster analysis is a statistical method both intellectually satisfying and profitable. In this chapter, theoretical frameworks are constructed and applications of cluster analysis, especially to marketing research, are reviewed. Alternative methods of cluster analysis are presented and evaluated in terms of recent empirical work. A discussion about cluster analysis methodology is presented. Issues and problems related to the use and validation of cluster analytic methods are discussed, followed by issues and problems related to the use and validation of these cluster analytic methods.

2. 1 A Short Review

Previously, various statisticians have attempted to define cluster analysis. For example, Cormack (1971) and Gordon (1980) define cluster analysis using properties of internal cohesion and external isolation. Everitt (1993) gives a graphical presentation about this idea for different situations, as can be seen in Figure 2. 1. External isolation refers to that a discontinuity should be observable between classes (Rogers et al. 1967). Internal cohesion refers to that an individual should be accepted into a cluster if its smallest correlation with any member is greater than some threshold (Cattell, 1944). Note that external isolation and internal cohesion are not limited to two dimensions. A formal definition of cluster analysis has been given by statisticians two decades ago. Kaufman and Rousseeuw (1990) discuss that cluster analysis is the classification of similar objects into groups, where the numbers of groups, as well as their forms are unknown. Here, the “ form of a

group” refers to the parameters of cluster such as cluster-specific means, variances and covariances.

Figure 2. 1 Clusters with external isolation (left plot) and internal cohesion (right plot)

(Reproduced from Everitt, 1993)

Then, cluster analysis has become a common tool for statisticians and social sciences researchers. Both the academic researcher and the application researchers rely on this technique to develop empirical groupings of persons, products, or occasions which may serve as the basis for further analysis.

Despite the frequent use of cluster analysis, little literature is given about the comparison of the available clustering methods and how clustering methods should be employed. Punj and Stewart (1983) discuss that the first evidence of lack of specificity of clustering algorithm is that numerous social sciences researchers fail to specify what clustering method is being used in their work. Another such indicator is the tendency of some authors to differentiate among methods which actually vary only in names. This could be indicated by a table in appendix 1 which summarizes the primary name and alternative name of clustering methods from Punj and Stewart (1983).

The lack of specificity about the clustering algorithm is an indicator of the problems associated with the use of cluster analysis. It tends to impede social science researchers who might seek a suitable method of clustering. Thus a sound review of clustering methodology is required.

Generally speaking, three sets of issues confront statisticians and those social science researchers seeking to use cluster analysis. One set of issues includes measures of similarity. These issues are summarized by Cormack (1971), and thus are not addressed here. Another set of issues is the theoretical properties of particular algorithms. These issues are considered in the literature on cluster analysis (Anderberg 1973). In this dissertation, the comparisons between these algorithms are addressed. The third set of issues is how to analyse data using these clustering algorithms. These issues are the foci of chapter 2 and chapter 5, in which we review applications of clustering methodology, provide a theoretical framework for the clustering options, and use both theoretical and empirical findings to suggest which clustering options may be most useful for particular data analysis problems.

2. 2 Clustering Algorithms

Basically speaking, the main procedures of different clustering methods in data analysis are similar. First, a number of cases which want to be subdivided into homogeneous groups are required. Then, choose the variables on which the clustered groups are supposed to be similar. Next, decide whether to standardize the variables in some way so that they all contribute equally to the distance or similarity between cases. Finally, determine which clustering procedure to use, perhaps based on the number of cases and types of variables that are used for producing clusters. The last step is also one of the most controversial issues of clustering research, because there are numerous ways in which clusters can be formed.

Punj and Stewart (1983) summarize a description of the common clustering algorithms and their various alternative names by which the algorithms are

known, also a brief discussion of how clusters are formed by each of these methods is provided. Descriptions of these clustering algorithms could be seen in appendix 1. Fraley and Raftery (1998) discuss that clustering methods range from those that are largely heuristic to more formal procedures based on statistical models. Hierarchical methods proceed by stages producing a sequence of partitions, each corresponding to a different number of clusters. They can be either ‘ agglomerative’, meaning that groups are merged, or ‘ divisive’, in which one or more groups are split at each stage. On the other hand, relocation methods move observations iteratively from one group to another, starting from an initial partition. The number of groups has to be specified in advance and typically does not change during the course of the iteration.

In this chapter, the hierarchical clustering algorithm, the k-means algorithm which is the most common relocation method and a relatively new clustering method, the two-step clustering algorithm, are discussed. For hierarchical clustering, the first step is to choose a statistic that quantifies how far apart (or similar) two cases are. Summary of these measures of similarity could be found in Cormack (1971)’s work. Then a clustering method is selected to form the clusters. Since it is possible to construct as many clusters as the number of cases, (for example the maximum number of clusters of a 100 respondents’ dataset is 100), the last step is to determine how many clusters is needed to represent the underlying structure of the data. This could be done by checking how similar these clusters are when create additional clusters or collapse existing ones. But still, choosing the number of clusters is subjective. Therefore in the later part of his chapter, several methods of

how to estimate the number of clusters are reviewed and compared. In terms of k-means clustering, the number of clusters should be fixed in advance. This algorithm iteratively estimates the cluster means and assigns each case to the cluster for which its distance to the cluster mean is the smallest, until all the distances reach minimum. In two-step clustering algorithms, in order to make large dataset analysing available, firstly, cases are assigned to “ pre-clusters”. Then, the pre-clusters are clustered using the hierarchical clustering algorithm. Unlike the other two methods, there are several statistical criteria available during clustering procedures. Therefore it is easier to specify the number of clusters objectively. An alternative way is let the algorithm to determine the number of clusters automatically.

2. 2. 1 Hierarchical clustering

Hierarchical clustering is one of the most common and straightforward clustering methods. It could be either agglomerative or divisive. Suppose there are n cases, agglomerative hierarchical clustering starts with every case being a cluster. In the next step, the two clusters that have the smallest value for the distance measure (or largest value if using similarity measure) are joined into a single cluster or two separate cases are combined into a new cluster. In the following steps, either an individual case are added to an existing cluster, two cases are combined, or two existing clusters are combined. The algorithm ends with all the cases in one cluster. This idea could be illustrated by a 6 cases example in Figure 2. 2.

At the start of this procedure, when each cluster contains only one case, the smallest distance between cases in two clusters is clear. Once start forming clusters with more than one case, a distance between pairs of clusters need

to be defined. For example, if cluster A has cases 1, 2 and 3 while cluster B has cases 4, 5, and 6, we need a measure of how different or similar the two clusters are. There are many available definitions of the distance between two clusters. The most frequently used methods are summarized by Cormack (1971).

Step 1

Step 2

Step 3

Step 4

Figure 2. 2 6 Cases Example of Agglomerative Hierarchical Clustering

Although hierarchical clustering is one of the most popular clustering algorithms, it still has some limitations. The first limitation is sample size. Fraley and Raftery (1998) discuss that hierarchical clustering algorithm has storage and time requirements for computer and it grows faster than linear rate relative to the size of the initial partition. Thus when applied to large data sets, hierarchical clustering performs very slow.

On the other hand, divisive clustering begins with every case in one cluster and ends up with everyone in individual clusters. This idea could be illustrated by Figure 2. 3. Obviously, neither the first step nor the last step is a worthwhile solution, because both one cluster and n clusters are useless classifying results for n cases in practice.

Step 1

Step 1

Step 3

Step 4

Figure 2. 3 6 Cases Example of Divisive Hierarchical Clustering

In terms of when to stop cluster formation, in other words, how many clusters we need to represent the data, clustering coefficients could provide the value of the distance (or similarity) statistic. We may stop cluster formation when the significant increase of coefficient (for distance measures) or decrease (for similarity measures) stops. This idea will be further discussed in the appropriate part in this chapter.

2. 2. 2 K-Means Clustering

Since hierarchical clustering requires a distance or similarity matrix between all pairs of cases, a huge matrix would be generated when it comes to a large dataset and this huge matrix could leads to inefficient clustering. An alternative clustering method is k-means clustering, which does not require calculation of distance matrix. It also differs from hierarchical clustering in other ways. For example, the number of clusters is needed to be specified in advance. Moreover, k-means algorithm repeatedly reassigns cases to clusters, so the same case could moves from one cluster to another cluster during the clustering procedure. However, in agglomerative hierarchical clustering, cases are only added to existing clusters, in other words, once a case has been assigned to a cluster, it is forever trapped in that cluster, with a widening family of neighbors.

The procedure of k-means clustering is not complicated either. It could be demonstrated by Figure 2. 4. A 12 cases example is represented in step 1 and the number of clusters is supposed to be four. Start with an initial set of means and classify cases based on their distances to the centers (centers are denoted by dashed circle in Figure 2. 4). This stage of k-means clustering is illustrated by step 2 in Figure 2. 4. Then calculate the cluster means again, using the cases that are assigned to the cluster. Reclassify all cases based on the new set of means, which can be seen from step 3. Keep repeating this step until cluster means will not change much in the successive steps. Finally, calculate the means of the clusters again and assign the cases to their permanent clusters, as indicated in step 4.

Step 1

Step 2

Step 3

Step 4

Figure 2. 4 K-Means Clustering of 12 cases example

Since the number of clusters has been specified in advance, usually the clustering procedure of k-means algorithm is much faster than hierarchical algorithm. That is also the reason for k-means algorithm sometimes been called “ fast clustering” (Chih, et. al 2000). Compared with hierarchical algorithm, k-means could be efficiently applied large data set. Also the initial starting point could be specified before clustering, using results of other cluster analysis. This is a superior advantage that other algorithms do not

possess. However, the disadvantages of k-means clustering algorithm are also obvious. First, k-means requires researchers set the number of clusters in advance, which is not very convenient in practice. Second, unlike hierarchical algorithm, k-means could only cluster cases, but not variables. In other words, the usage is restricted. Third, all the variables used in k-means clustering algorithm should be continuous variables but this requirement is not often satisfied in social sciences researches.

2. 2. 3 Two-Step Clustering

Given the above discussion, we know that hierarchical clustering requires a matrix of distances between all pairs of cases, and k-means requires shuffling cases in and out of clusters and knowing the number of clusters in advance. When it comes to a large data set and need a clustering procedure on the basis of either categorical or continuous data, neither of the two procedures mentioned above could fills the bill. Thus there is a need to introduce another clustering algorithm, two-step clustering. NoruÅ¡is (2009) discusses that two-step clustering algorithm produces solutions based on mixtures of continuous and categorical variables and for varying numbers of clusters.

Step 1: Pre-clustering

The first stage of the two-step clustering is the formation of pre-clusters. The aim of pre-clustering is to reduce the size of the distance matrix. Pre-clusters are just clusters of the original cases that are used in place of the raw data in the hierarchical clustering. The algorithm decides if the current case should be merged with a previously formed pre-cluster or starts a new pre-cluster. When pre-clustering is complete, all cases in the same pre-cluster are

treated as a single entity. The size of the distance matrix is no longer depends on the number of cases but on the number of pre-clusters.

Step 2: Hierarchical Clustering of Pre-clusters

In the second stage, standard hierarchical clustering algorithm was implemented on the pre-clusters. Forming clusters hierarchically enables exploring a range of solutions with different numbers of clusters become available. During this process, the number of clusters to be formed could be specified in advance, or the algorithm could automatically select the optimal number based on either the Schwarz Bayesian Criterion or the Akaike information criterion which will be explained in this chapter 3. These procedure could be graphically represented by Figure 2. 5, which is's work.

Average Linkage or Ward's Minimum Variance Method

Preliminary Cluster Solution

Select Candidate Number of Clusters

Obtain Centro of Clusters

Eliminate Outliers

Iterative Partitioning Algorithm using Cluster Centro of Preliminary Analysis as Starting Points (Outliers are not included)

Final Cluster Solution

Stage 1

Stage 2

Figure 2. 5 Two-step Clustering Algorithm

(reproduced from Punj and Stewart, 1983)

Compared with hierarchical clustering and k-means algorithm, two-step method has several pros and cons. First, unlike k-means algorithm, the continuous requirement of the variables need to be classified is not necessary for two-step algorithm. Both categorical and metrical variables could be clustered using this method. This relaxation does improve the practicability of two-step algorithm used in social sciences, because most variables of social sciences are categorical. Second, given the special procedure of this method, two-step clustering is also a fast algorithm without taking much space of computer's memory. This is a significant advantage compared with hierarchical clustering, because for cluster analysis, even 1000 could be accounted for large sample size. Third, two-step algorithm uses statistic as distance measure to cluster cases or variables, and could automatically estimate optimal number of clusters. All these features make the results of two-step clustering more stable. Therefore this relatively new clustering algorithm has become a more and more popular topic in cluster analysis.

2. 2. 4 Comparison between the Algorithms

Cluster analysis is an exploratory data analysis method, and different datasets require different clustering algorithms. It is impossible to give a definitive answer for which algorithm is completely superior. But the rule of thumb is that the method which maximizes homogeneity within cluster and difference between clusters is the best choice. Although several differences

between these algorithms have been discussed separately, there is still a need to conclude the difference systematically. Table 2. 1 summarizes this discussion.

Table 2. 1 Usage of three algorithms

	Hierarchical	k-means	two-step
Types of Clusters			
cases need to be classified	Y	Y	Y
variables need to be classified	Y		
Sample size			
small sample size (<100)	Y	Y	Y

medium sample size (100~1000)

Y

Y

Y

large sample size (> 1000)

Y

Y

Type of variables

continuous

Y

Y

Y

categorical

Y

Whether number of clusters need to be specified in advance

Yes

Y

Y

No

Y

Y

In terms of types of variable, if it is the case that need to be classified, then all the three algorithms are equally suitable; if variable that need to be classified, then hierarchical clustering is the best choice.

In terms of sample size, there is no clear distinction between these algorithms. However, for small size dataset, which is usually less than 100, although three methods are both available, hierarchical clustering would be the best choice, because the dendrogram generated during clustering procedure could visualize the outcomes. This graph provides a convenient way to explain the results in social sciences research. Moreover, the profusion of distance measure in hierarchical clustering very much exceeds the other two. On the contrary, for large dataset, k-means clustering and two-step would be better solution. For medium dataset, all the three methods could be applied in theory, but perhaps too tough to observe the results through dendrogram.

In terms of the types of variables need to be classified, if continuous, then all the three algorithms could be considered; if these variables contain categorical variable, two-step clustering would be better. An alternative way to solve this problem is to continuous these variables before clustering.

In terms of whether need to set the number of clusters in advance, two-step algorithm automatically generates the estimation according to some criteria,

hierarchical algorithm could generate results with a certain range number of clusters and k-means algorithm requires the number of clusters to be specified in advance.

2. 3 Estimating the Number of Clusters

K-means algorithm aims to classify cases into a given number of clusters while hierarchical clustering algorithm aims to generate the structure of clusters. Both these two clustering algorithms require researchers to specify the optimal number of clusters. However, the estimation of optimal number of clusters has always been a major challenge in cluster analysis. Various strategies for simultaneous determination of the number of clusters have been proposed (Fraley and Raftery, 1998). A review and comparison of these methods is presented in this chapter and further application could be seen in chapter 5.

2. 3. 1 “ Elbow “ method

A typical and practical method to determine the number of clusters is examining the heuristic. Suppose data {}, i= 1, 2,…n, j= 1, 2,…p, consist of p featured on n independent observations. Let denote the distance between observations i and. Let

= (2. 1)

Be the sum of the pairwise distances for all points in cluster r, and set

= (3. 2)

Figure 2. 5 shows an error measure (the within cluster dispersion) for a clustering procedure versus the number of cluster k. The error measure

decreases monotonically as the number of clusters k increases, but from some k onwards the decreases flattens significantly. Statisticians use that ‘ elbow’ to indicate the appropriate number of clusters. But this method has not been formalized.

2 4 6 8

Number of clusters k

600 400 200 0

Within sum of squares Wk

Figure 2. 6 within sum of squares function

2. 3. 2 “ Gap” statistic method

Tibshirani et al (2001) discuss that a gap statistic could formalize the heuristic. The basic idea of this approach is to standardize the graph of log () by comparing it with its expectation under an appropriate null reference distribution of the data. An estimate of the optimal number of clusters is then the value of k for which log () falls the farthest below this reference curve. Define

= {log()} -log () (3. 3)

where denotes expectation under a sample of size n from the reference distribution. The estimate will be the value maximizing after taking sampling distribution into account. This idea could be illustrated in Figure 2. 7 (reproduced from Tibshirani et al. 2000). And Figure 2. 8 (reproduced from Tibshirani et al. 2000) indicates a gap curve obtained for Equation 3. 3.

Additionally, a simulation study shows that the gap statistic usually outperforms other methods that have been proposed in the literature. Also, this estimate is very general, applicable to any clustering method and distance measure. Therefore, it is preferable than the traditional “ elbow” method.

2 4 6 8

Number of clusters

6 5 4 3

Obs and exp log (Wk)

E

E

E

E

E

E

E

Figure 2. 7 Functions log() (O) and {log() (E)

(reproduced from Tibshirani et al. 2000)

2 4 6 8

Number of clusters

1. 5 1. 0 0. 5 0

Gap

Figure 2. 8 Gap curve

(Reproduced from Tibshirani et al. 2000)

2. 4 Review of cluster analysis application

Cluster analysis has a rich history using in lots of disciplines such as psychiatry, psychology, archaeology, geology, geography, and marketing. The importance and interdisciplinary nature of clustering is evident through its vast literature.

2. 4. 1 Cluster Analysis in Medicine

Examining patients with a diagnosis of depression is available if distinct subgroups can be identified based on a symptom checklist and results from psychological tests. McLachlan and Basford (1988) discuss that classification is useful in Medical prognosis, where the group in the classification scheme correspond to the possible outcome of a medical condition, injury, for example. Data are recorded individually after the injury, and a prediction of the outcome is required in advance in order to guide the clinician as to whether this particular treatment is appropriate. It also provides a suitable basis to suggest the chance of recovery.

2. 4. 2 Cluster Analysis in Psychiatry

Disease of the mind is much trickier than disease of the body, hence there is increasing interest in psychiatry using multivariate analysis. Everitt et al (1971) proposed that cluster analysis is a more suitable method to the problem of taxonomy in psychiatry than other multivariate techniques such as factor analysis, because cluster analysis produces groups of cases with signs and symptoms in common, whereas factor analysis produces groups of variables. In other words, compared with other classifying method, cluster analysis is more useful for clinical diagnosis.

Moreover, cluster analysis could also be applied to identify individuals who attempt suicide. The grouping was meaningful in terms of clinical interpretation and had both therapeutic and prognostic implications. Kurtz et al (1987) studied 485 patients, consecutively admitted to the poisoning treatment unit of a large university hospital in Munich. Eight cases of drug experimentation were included, and the patients were representative of suicide attempters in Western European cities. Information was gathered on demographic characteristics, psychiatric history, pre-suicide situation, circumstances of the suicide attempt and psychiatric diagnosis. Results could be summarized as follow: Group A: Patients were characterized by low mean age, more males, more previous suicide attempts, more suicidal pre-meditation and more hostility. Group B: Patients were distinguished by above-average mean age, more males, and more precautions against discovery, more severe intoxications, low interpersonal and high auto aggressive motivation, high depression and low hostility scores. Group C: Patients were low mean age, more females, low percentage of previous

suicidal episodes, low rates of recent suicidal intent and of concealment, less severe intoxications and high interpersonal and low self-directed motivation. Also differences between groups were statistically significant on all variables. Cluster analysis provides a useful evidence for clinical diagnostic from statistical perspective and this topic need to be further studied.

2. 4. 3 Cluster Analysis in Marketing

Market segmentation is the main use of cluster analysis in marketing research. Punj and Stewart(1983) discuss all segmentation research, regardless of the method used, is designed to identify groups of entities (people, markets, organizations) that share certain common characteristics (attitudes, purchase propensities, media habits, etc.). Stripped of the specific data employed and the details of the purposes of a particular study, segmentation research becomes a grouping task. Reviewed of literature of market segmentation research and methodology, Wind (1978) discusses both the impact of cluster analysis as a fundamental of marketing tool and some significant problem areas.

Another important use of cluster analysis is seeking an underlying pattern of buyer behaviors by identifying homogeneous groups of buyers. Even though the importance of cluster analysis has been recognized in marketing, research methodologists have paid very little attention to classify groups of buyers using statistical clustering approach. There is clearly a need for better classification of relevant buyer characteristics. For example, Bettman (1979) calls for the further study of taxonomies of both consumer choice task and individual difference characteristics. And cluster analysis is one possible way

to develop such taxonomies. In chapter 5, an application in buyer behavior classifying has been discussed.

Punj and Stewart (1983) also discuss another two possible ways of cluster analysis using in marketing research, which are developing potential new product opportunities and testing market selection respectively. The former is determining competitive sets within the large market structure by clustering brands or products. And the latter refers to identifying relatively homogeneous sets of test markets which may become interchangeable in test market studies.

In brief, cluster analysis is a practical statistical approach which could be used in a plethora of studies. Therefore it should be further investigated by both statisticians and social sciences researchers.