An Overlapping Cluster Scheme

Kazumasa Ozawa
Department of Engineering Informatics,
Osaka Electro-Communication University,
Neyagawa, Osaka, 572-8530, Japan

Abstract

Classification plays an important role in understanding a set of findings in archaeology. Some of traditional cluster schemes have been employed for classification not only in archaeology, but also in other areas. Almost all of traditional cluster schemes provide non-overlapping partition of a set as seen in dendrograms. In archaeology, such non-overlapping partition might cause misunderstanding of a set of findings. Suppose a situation that three classes A, B, and C of a family of artefacts have already been defined. When an artefact belonging to the family is newly found, we immediately face at a problem to classify it into which class among A, B, and C. Sometimes, it might happen that we could find no clear attributes of the artefact to be classified into a single class, either A or B. In such case, a reasonable solution is that the artefact should be classified into two classes A and B; i.e. kind of overlapping scheme should be introduced. This paper theoretically presents an overlapping cluster scheme to provide overlapping classes of a set of findings. A consideration on an artificially defined set of patterns is also made to illustrate the proposed scheme.

Introduction

Clustering has been carried out in many fields. In one word, clustering is aimed at understanding of a set of individuals; i.e. a set of artefacts or findings in archaeology. Understanding of a set of individuals could be followed by abstracting new knowledge, building models and theories, structuring the set or defining new types. Some of traditional cluster schemes have been employed for understanding of a given set of individuals not only in archaeology, but also in other fields. Almost all of traditional cluster schemes provide non-overlapping partition of a set as seen in tree-diagrams (Jardin 1971, Van Ryzin 1977). In archaeology, such non-overlapping partition might cause misunderstanding of a set of findings. Suppose an artefact be newly found. Then we immediately face at a problem to classify it into which of existing types defined previously. It might sometimes happen that we could find no clear evidence to classify the artefact into a single type. In such case, one of reasonable solutions should be to classify the artefact into more than two types; i.e. kind of overlapping cluster scheme should be introduced (Jardin 1971, Matula 1977, Ozawa 1985). This paper presents a theoretical base of an overlapping cluster scheme that could be employed in archaeology.

Two Types of Cluster Schemes

A cluster is a subset of a set. Here a set means a collection of individuals such as artefacts or findings. Mathematical description of a set X and its n clusters v_1 to vn that cover X are written by

$X = v_1 \cup v_2 \cup ... \cup v_n$

Note that all clusters $v_1, ..., v_n$ are not needed to separate with each other: It is no problem that clusters contain the same individuals in common, i.e. clusters are overlapped. In contrast, if all clusters separate with each other, set X is to be partitioned into non-overlapped n clusters.

In theory, we can have two types of cluster schemes; one is non-overlapping and another overlapping. The first type is very conventional. In fact, almost all of popular cluster schemes have been non-overlapping. On the other hand, we face at some practical problems that would not be fit for such non-overlapping schemes. The real world has sometimes overlapping structures: For instance, there are many people who have plural nationalities. They belong to more than two clusters, in terms of nationality, of a set of individuals. We also have similar problems in archaeology, which should be handled by the overlapping cluster scheme.

On the other hand, another approach to the clustering problem has been well known as fuzzy clustering (Zadeh 1977). In short, with fuzzy clustering, membership grades of plural clusters are given to an individual. Then it looks to provide kind of overlapping clustering. However, when making a practical decision which cluster an individual strongly belongs to, i.e. when determining a threshold value to cut

Name	Attribute A	Attribute B
1	4	10
2	2	7
3	4	5
4	6	8
5	11	7
6	16	8
7	18	10
8	19	7

Table 1 Example set of eight individuals

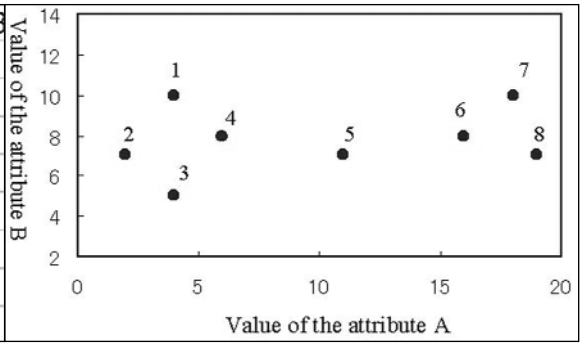

Figure 1 Distribution map of the example set of eight individuals

membership grades, you will find there nothing but non-overlapping clustering. In theory, overlapping clustering is given by excluding the transitive merging of clusters, described later, in either case of fuzzy or non-fuzzy clustering.

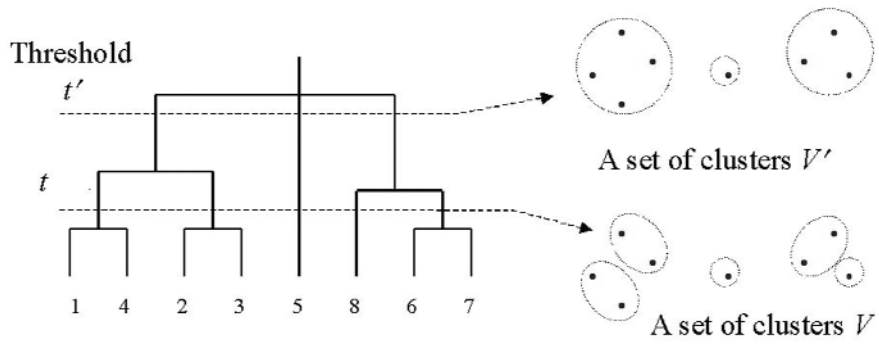

Figure 2 Tree-diagram representation of non-overlapping clustering of the example set

Quantitative Attributes and Distance Measures

To discuss quantitative clustering, an example set of eight individuals named 1, 2,..., 8 is shown in Table 1, which is associated with two quantitative attributes A and B. The table presents two values of A and B for all individuals. Figure 1 presents a two-dimensional distribution map of the eight individuals. By visual inspection on the map, we can have three possible ways of clustering: The first is non-overlapping clustering such that we have three clusters {1, 2, 3, 4}, {5} and {6, 7, 8}. The second is also non-overlapping; we have two clusters such as {1, 2, 3, 4, 5} and {6, 7, 8}, or {1, 2, 3, 4} and {5, 6, 7, 8}. In this case, individual 5 is merged into either of two big clusters in the left hand and right hand sides. The third is overlapping clustering such that individual 5 belongs to both two big clusters; i.e. we have {1, 2, 3, 4, 5} and {5, 6, 7, 8}. To discuss this theoretically, we have to evaluate distances or dissimilarities between individuals. Let d(*i, j*) be the distance between two individuals *i* and *j*. And, ai is value of attribute A for individual i. Similarly b_j is a value of B for j. A matrix of Euclidean distances between eight individuals in the example set is given as follows:

$$D = \begin{bmatrix} 0 & 3.6 & 5.0 & 2.8 & 7.6 & 12.2 & 14.0 & 15.3 \\ 3.6 & 0 & 2.8 & 4.1 & 9.0 & 14.0 & 16.3 & 17.0 \\ 5.0 & 2.8 & 0 & 3.6 & 7.3 & 12.4 & 14.9 & 15.1 \\ 2.8 & 4.1 & 3.6 & 0 & 5.1 & 10.0 & 12.2 & 13.0 \\ 7.6 & 9.0 & 7.3 & 5.1 & 0 & 5.1 & 7.7 & 8.0 \\ 12.2 & 14.0 & 12.4 & 10.0 & 5.1 & 0 & 2.8 & 3.1 \\ 14.0 & 16.3 & 14.9 & 12.2 & 7.7 & 2.8 & 0 & 3.1 \\ 15.3 & 17.0 & 15.1 & 13.0 & 8.0 & 3.1 & 3.1 & 0 \end{bmatrix}$$

Where distance d(*i, j*) between *i* and *j* is written at the cross of *i*-th row and *j*-th column.

Here we introduce the between-cluster distance. Suppose there be two clusters *v* and *v′*. Then between-cluster distance of *v* and *v′*, written m(*v, v′*), has been given in several ways: One is the *single link* that defines it as the distance between the nearest two individuals each of which belongs to either cluster. Symbolically, we have $m(v,v') = \min_{i \in v-v',\ i' \in v'-v} d(i,$

Another is the complete link that defines it as the distance between the farthest individuals. Other definitions could also been employed.

Procedures for Merging Clusters

Clustering can be regarded as a process of merging clusters. Here we discuss two types of merging procedures; i.e. transitive merging and strictly linked merging. We have the following definition of the transitive merging that is closely associated with non-overlapping clustering:

Definition 1 (*Transitive merging*) Now suppose there be three clusters *v, v′* and *v′′*. For a given threshold value *t*, if the distance between *v* and *v′* is smaller than *t* and the distance between *v′* and *v′′* is also smaller than *t*, then let the distance between *v* and *v′′* be smaller than *t* automatically and merge the three clusters *v*, *v′* and *v′′* into one cluster.

Metaphorically speaking, transitive merging means that if you are my friend and he is your friend, then he should automatically be my friend. This is obviously unrealistic. But almost all existing clustering schemes are based on this transitive theory associated with non-overlapping property. Another type of merging procedure is the strictly linked merging defined as follows:

Definition 2 (Strictly linked merging) For a given threshold value *t*, if and only if any pair of clusters selected from the three clusters *v, v′* and *v′′* are closer than t, then they are merged into one cluster. We call such a situation that every pair of clusters is close as ´strictly linked´.

In other words, definition 2 says that if and only if you are my friend, he is your friend and he is my friend, then we all are friends. Note that such strictly linked merging allows clusters to overlap.

Figure 2 presents a well-known tree-diagram to illustrate a non-overlapping clustering process by transitive merging. At a stage on threshold *t*, five clusters are given. At the next stage, on threshold *t′*, they are merged into three clusters. The central cluster is a singleton that contains only single individual. The three clusters are merged into only one cluster at the final stage. In this example, single link has been employed for between-cluster distance. Other between-cluster distances would offer different tree-diagrams. On the other hand, when you introduce the strictly linked merging, you could meet with overlapped clusters. Figure 3 shows a process of strictly linked merging of clusters. In this case, we begin with an initial set of singletons and get to the final stage by increasing the threshold value step-by-step. At the third stage, we can meet with two overlapped clusters {1, 2, 3, 4, 5} and {5, 6, 7, 8}.

Statistic and Quantitative Methods

Conclusion

In this paper, we have discussed an overlapping cluster scheme associated with strictly linked merging. For simplicity, we have limited our discussion within an example set of eight individuals. Nevertheless, our results will have good generality in clustering every set of individuals in archaeology. It should be noted that the conventional transitive merging never offer overlapped clusters. In archaeology, overlapping clustering has never been employed for practical data analysis. The author is hoping it will be applied to understanding of sets of archaeological findings.

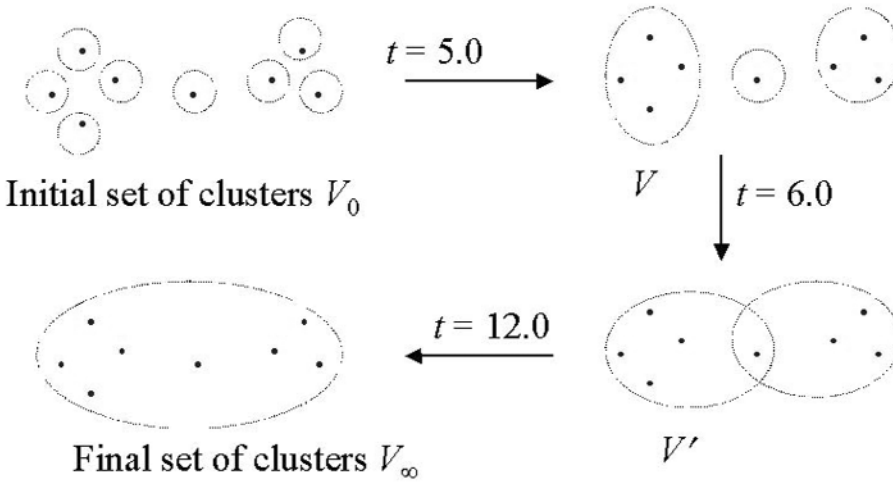

Figure 3 A process of strictly linked merging of clusters

References

JARDIN, N. and SIBSON, R., 1971. Mathematical Taxonomy. Wiley.

MATULA, D.W., 1971. Graph Theoretic Techniques for Cluster Analysis Algorithms. In Van Ryzin, J. (ed.), Classification and Clustering, Academic Press.

OZAWA, K., 1985. A Stratificational Overlapping Cluster Scheme. Pattern Recognition Vol.18, Nos.3/4:279-286.

VAN RYZIN, J., 1977. Classification and Clustering. Academic Press.

ZADEH, L.A., 1971. Fuzzy Sets and Their Application to Pattern Classification and Clustering Analysis. In Van Ryzin, J., (ed.), Classification and Clustering, Academic Press.