Set Partition Principles Revisited

Ventzeslav Valev
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad.G.Bontchev St., Bl.8, 1113 Sofia, BULGARIA
vvalev@bgearn.acad.bg

Abstract

Two principles for partitioning a set into groups are revisited in the paper. In addition to the well-known cluster analysis principle, two other set partition principles are considered: the similarity principle and the anticluster principle. In similarity principle the initial set is partitioned into groups, so that each group possesses property similar to the property of the initial set. In anticluster principle, the initial set is partitioned into groups in such a way, that elements belonging to each group are dissimilar but the groups are similar. If a criterial function for quality of partitioning is defined on the set of all possible partitions, then the set partitioning problem is to construct such a partition, for which the criterial function is extremal. Optimization procedures are suggested for both partitioning principles.

Key words: Set Partition Principles, Cluster Analysis Principle, Similarity Principle, Anticluster Principle, Discrete Optimization Procedures

1 Introduction

Let us consider the set $X = \{x_1, \ldots, x_n\}$ with a metric ρ. Let k be an integer, $k < n$. It is well-known that the solution of the cluster analysis problem consists of partitioning the set X into k subsets (groups, clusters, taxons) in such a way that each element $x \in X$ belongs to one and only one subset, so that the elements belonging to one and the same subset are similar, and elements belonging to different subsets are dissimilar, see for example [1]. The metric ρ is used as a quantitative measure of similarity or dissimilarity. Obviously, the result of partitioning essentially depends on the metric ρ.

In the present paper two principles for partitioning the set X into disjoint subsets (groups) are revisited [2]. The similarity principle requires that each of the obtained groups (similarity clusters) possesses property similar to the property of the set X. The anticluster principle requires that the elements of each of the obtained groups (anticlusters, antitaxons) are dissimilar but the groups are similar. The metric ρ is used as a quantitative measure for similarity or dissimilarity in both principles.

If a criterial function for quality of partitioning is defined on the set of all possible partitions, then the problem is to construct such a partitioning for which the criterial function is extremal. Accordingly, the partitioning problem becomes a well-posed problem of discrete optimization.

A recent review by P.Arabie and L.J.Hubert referred to H. Späth's works as "interesting and novel development" on anticlustering neglecting the published origins of this concept ([5], p.15). Similarity and anticluster principles were first reported by V.Valev in 1982 [2]. The idea of the anticluster principle was reported again by H. Späth in 1986 [3], [4].

2 Similarity Principle

In this section, we consider the problem for partitioning the set X into similarity clusters. Let the number of similarity clusters k be given, where $k < n$.

2.1 Consecutive Procedure

Let m be the mean vector of the set X and m_j be the mean vector of the similarity cluster $P_j, j = 1, \ldots, k$. We consider a consecutive procedure for partitioning the set X into similarity clusters in which the mean vector m is used as a similarity property:

1. The mean vector m is calculated for the set X.

2. Applying a certain rule, possibly in a random way, k elements from the set X are chosen and are assigned to similarity clusters $P_1, P_2, \ldots, P_k$ as initial values.

3. An element x_i, which is not yet assigned to any similarity cluster, is chosen from the set X. Mean vectors $m_1, m_2, \ldots, m_k$ of similarity clusters $P_1, P_2, \ldots, P_k$ are calculated with the element x_i sequentially included in them.

4. The element x_i is assigned to the similarity cluster P_j if

$$\rho(m_j, m) = \min_{1 \le l \le k} \rho(m_l, m).$$

   The new mean vector m_j is calculated.

5. If all elements $x \in X$ are assigned to similarity clusters then stop; otherwise go to step 3.

The described procedure could be repeated with different initial partitions to obtaining a steady partitioning.

2.2 Parallel Procedures

Let the following criterial function for quality of partitioning be given:

$$J_m = \sum_{i=1}^{k} \|m_i - m\|^2,$$

where

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$m_i = \frac{1}{n_i}\sum_{x_j \in P_i} x_j,$$

and n_i is the number of elements in the similarity cluster $P_i, i = 1, 2, \ldots, k$.

The partitioning of the set X into similarity clusters for which the minimum of the criterial function J_m is obtained is called optimal partition.

Let us assume that element $\hat{x}$ from similarity cluster P_i is moved to similarity cluster P_j. Then m_j is changed to:

$$m_j^* = m_j + \frac{\hat{x} - m_j}{n_j + 1},$$

and m_i is changed to:

$$m_i^* = m_i - \frac{\hat{x} - m_i}{n_i - 1}.$$

We thereby assume that $n_i \neq 1$ for $i = 1, 2, \ldots, k$. Shifting the element $\hat{x}$ from P_i to P_j leads to decreasing the criterial function iff

$$(\|m_j - m\|^2 + \|m_i - m\|^2) > (\|m_j^* - m\|^2 + \|m_i^* - m\|^2).$$

Substituting m_i^* and m_j^* and rearranging, we obtain the condition:

$$\frac{2\langle m_j - m, \hat{x} - m_j\rangle}{n_j + 1} - \frac{2\langle m_i - m, \hat{x} - m_j\rangle}{n_i - 1} +$$

$$+\frac{\|\hat{x} - m_j\|^2}{(n_j + 1)^2} + \frac{\|\hat{x} - m_i\|^2}{(n_i - 1)^2} < 0.$$

This leads to the following iterative procedure for minimizing the criterial function J_m:

1. An initial partitioning of set X into k similarity clusters is chosen.
2. The next element $\hat{x}$ for shifting is chosen. Let $\hat{x} \in P_i$.
3. If $n_i = 1$ then go to step 2; otherwise calculate the value:

$$e_j = \frac{2\langle m_j - m, \hat{x} - m_j\rangle}{n_j + 1} - \frac{2\langle m_i - m, \hat{x} - m_j\rangle}{n_i - 1} +$$

$$+\frac{\|\hat{x} - m_j\|^2}{(n_j + 1)^2} + \frac{\|\hat{x} - m_i\|^2}{(n_i - 1)^2},$$

for $j = 1, 2, \ldots, k$. All indices j, $j \neq i$, such that $e_j < 0$ are recorded.

4. The element $\hat{x}$ is shifted into P_j if

$$e_j = \min_u e_u,$$

where u runs the set of recorded indices.

5. The values of J_m, m_i and m_j are calculated.

6. If the value of J_m is not changed after n iterations then stop; otherwise go to step 2.

In the last procedure for the element $\hat{x} \in P_i$ chosen in advance, a similarity cluster P_j is found, so that shifting the element $\hat{x}$ into P_j leads to minimizing the criterial function J_m. This procedure can be modified if at each iteration the most prospective similarity clusters P_i and P_j ($\hat{x}$ is shifted from P_i to P_j) is found using the condition:

$$d_{ij} = \min_{1 \le p,q \le k} \langle m_p - m, m_q - m \rangle.$$

As another property for measuring similarity, the dispersion of set X is used:

$$s = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho(x_i, x_j).$$

The mean dispersion of the set X is calculated as:

$$\bar{s} = \frac{s}{N},$$

where $N = n(n-1)/2$.

Now we will consider another procedure for partitioning the set X into similarity clusters. In this procedure the mean dispersion $\bar{s}$ is used as a property for measuring similarity. Let $\bar{s}_j$ be the mean dispersion of similarity cluster P_j, $j = 1, 2, \ldots, k$. The procedure of similarity is as follows:

1. For set X, $\bar{s}$ is calculated.

2. From set X, k elements are chosen, possibly in a random way, as initial clusters and assigned to similarity clusters $P_1, P_2, \ldots, P_k$.

3. A next element $x_i \in X$ is chosen which is not yet assigned to any of similarity clusters. The mean dispersion $\bar{s}_j$ of similarity cluster P_j, $j = 1, 2, \ldots, k$ with the element x_i included in them are calculated sequentially. The mean dispersions are recorded.

4. The element x_i is assigned to the similarity cluster P_j if

$$|\bar{s} - \bar{s}_j| = \min_{1 \le l \le k} |\bar{s} - \bar{s}_l|.$$

5. If all elements $x \in X$ are assigned to similarity clusters then stop; otherwise go to step 3.

2.3 Hierarchical Procedure

As an hierarchical procedure for partitioning the set X into similarity clusters we will consider the following aglomerative (bottom-up) procedure:

1. All elements $x_i \in X$ are assigned into n similarity clusters $P_1, P_2, \ldots, P_n$, where $P_i = \{x_i\}, i = 1, 2, \ldots, n$.
2. A pair of similarity clusters P_i and P_j, $i \neq j$ with property similar to the property of the set X is found.
3. Similarity clusters P_i and P_j are merged, P_j is deleted and the number of the similarity clusters is diminished by one.
4. If the number of the similarity clusters is equal to k then stop; otherwise go to step 2.

If a distance between two matrices is defined, then the covariance matrix may be used as another property for similarity measuring. In the similarity procedures, as well as in all cluster procedures where the initial partitioning is chosen randomly, in a general case a local extremum of the criterial function for quality of partitioning is obtained.

Different approaches with random choice of initial partitioning may lead to various solutions and, as a rule, it is never known whether the best solution has been found. A universal approach for choosing the initial partitioning doesn't exist. One of the frequently used approaches consists of the repetition of the procedure with various initial partitions. This may give an evaluation for the steady state of the solution.

3 Anticluster Principle

The anticluster principle consists of partitioning the set X into disjoint subsets (groups, anticlusters, antitaxons), so that the elements belonging to one and the same subset are dissimilar and the groups are similar. As a quantitative measure for similarity or dissimilarity the metric ρ is used again. Let us consider hierarchical and parallel procedures for partitioning the set X into anticlusters. Let the number of anticlusters k be given, where $k < n$.

3.1 Hierarchical Procedure

We will consider the following aglomerative (bottom-up) procedure:

1. All elements $x_i \in X$ are divided into n anticlusters $A_1, A_2, \ldots, A_n$, where $A_i = \{x_i\}, i = 1, 2, \ldots, n$.
2. The pair (A_i, A_j), $i \neq j$ is found for which:
$$\max_{i,j} \min_{x_p \in A_i, x_q \in A_j} \rho(x_p, x_q).$$
3. Anticlusters A_i and A_j are merged, A_j is deleted and the number of anticlusters is diminished by one.
4. If the number of all anticlusters is equal to k then stop; otherwise go to step 2.

3.2 Parallel Procedure

We will consider the following parallel anticluster procedure of type 'Forel-1' [6]:

1. Calculated are
$$m = \frac{1}{n}\sum_{i=1}^{n} x_i,$$
and R_0, where R_0 is the radius of the minimum hypersphere with its center at m, which contains all elements $x_i \in X$. An arbitrarily radius $R < R_0$ is chosen. Let $j = 1$.

2. From an arbitrary element $x_i \in X$ taken as a center, a hypersphere with radius R is constructed. Element x_i is assigned to anticluster A_j. All other elements in the hypersphere are removed.

3. If the remaining set of elements is not empty then go to step 2; otherwise $j := j + 1$.

4. A new set containing all elements which are not yet assigned to any of the anticlusters is constructed. If this set is empty then stop; otherwise go to step 2.

Accordingly, in each anticluster, the distance between any two elements is greater then R. The number of anticlusters depends on the chosen radius R and on the sequence of consideration of the elements $x_i \in X$. Let us assume that the set X is ordered and in step 2 elements are considered consecutively. Performing the procedure for

$$R_t = R_0 - t\delta,$$

where:

$$\delta = \frac{R_0}{N},$$

and $t = 1, 2, \ldots, N-1$, N - integer, $N > 1$, the required number anticlusters k is constructed.

4 Conclusions

In the present paper similarity and anticluster principles for partitioning the set X into disjoint subsets (groups) are revisited. They were first reported in the work [2]. The similarity principle requires that each of obtained groups (similarity clusters) possesses property similar to the property of the set X. The anticluster principle requires that the elements of each of the obtained groups (anticlusters, antitaxons) are dissimilar, but the groups are similar. As a quantitative measure for similarity or dissimilarity in both principles, the given metric ρ is used.

If a criterial function for the quality of partitioning is defined on the set of all possible partitions, then the problem is to construct such a partitioning for which the criterial function is extremal. Accordingly, the partitioning problem becomes a

well-posed problem of discrete optimization. Optimization procedures are suggested for both partitioning principles.

The similarity principle can be used for partitioning a large set into similar representative subsets. Potential applications of the proposed principles are in the fields of medicine, sociology, psychology.

References

[1] Duran, B., Odell, P.: Cluster analysis: A survey. New York: Springer-Verlag, (1974)

[2] Valev, V.: Set partition principles. Transactions of the Ninth Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes. Ninth Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes, Prague, 1982, Publishing House of the Czechoslovak Academy of Sciences: Academia, Prague (1983) 251–256

[3] Späth, H.: Anticlustering: Maximizing the variance criterion. Control and Cybernetics. **15** (1986) 213–218

[4] Späth, H.: Homogeneous and heterogeneous clusters for distance matrices. In: Classification and related methods of data analysis. Bock, H.H. (Editor) North–Holland (1986) 157–164

[5] Arabie, P., Hubert, L.J.: An overview of combinatorial data analysis. In: Clustering and classification. P.Arabie, P., Hubert, L.J., De Soete, G. (Editors) World Scientific (1996) 5–63

[6] Zagoruiko, N.G., Zaslavskaja, T.I.: Pattern Recognition Methods in Sociological Research. In: Quantitative sociology. Blalok, H.M. at al. (Editors), Academic Press (1975) 429–440