LOCAL OPTIMIZATION METHOD FOR THE GENERALIZED WARD CLUSTERING PROBLEM *

Vladimir Batagelj †
University of Ljubljana, Department of Mathematics
Jadranska 19, 61 111 Ljubljana, Yugoslavia

Signum to LaTeX: May 29 1989

Abstract

In the original Ward clustering problem the criterion function is based on the squared Euclidean distance, which is in the generalized Ward clustering problem replaced by any dissimilarity measure. In the paper it is shown that the local optimization procedures based on the matrix representation of dissimilarity can be used also for solving the generalized Ward clustering problem.

Key words : local optimization, generalized Ward clustering problem

Math. Subj. Class. (1985): 62 H 30, 90 C 48 .

1 Generalized Ward clustering problem

The generalized Ward clustering problem was introduced in the paper [2] in an attempt to legalize "extended" uses of Ward and related agglomerative methods. Replacing the squared Euclidean distance in original Ward clustering problem by any dissimilarity d we get the generalized Ward clustering problem. The aim of this paper is to show that local optimization clustering methods can also be used for solving the generalized Ward clustering problem.

*Paper presented at MAJSKI SKUP'88 SEKCIJE ZA KLASIFIKACIJE SSDJ-a, Mostar, 27.-28. may 1988

†Supported in part by the Research Council of Slovenia, Yugoslavia.

The *Ward clustering problem* can be posed as follows:

Determine the clustering $\mathcal{C}^* \in P_k$, for which

$$P(\mathcal{C}^*) = \min_{\mathcal{C} \in P_k} P(\mathcal{C})$$

where

$$P_k = \{\mathcal{C} : \mathcal{C} \text{ is partition of set of units } E \subseteq R^m \text{ and } \operatorname{card}(\mathcal{C}) = k\}$$

and the Ward criterion function $P(\mathcal{C})$ has the form

$$P(\mathcal{C}) = \sum_{C \in \mathcal{C}} p(C)$$

and

$$p(C) = \sum_{X \in C} d_2^2(X, \bar{C})$$

where $\bar{C}$ is the center of gravity of the cluster C:

$$\bar{C} = \frac{1}{n_C} \sum_{X \in C} X, \qquad n_C = \operatorname{card}(C)$$

and d_2 is the Euclidean distance.

To generalize the Ward clustering problem we proceed as follows: Let $E \subseteq \mathcal{E}$, where $\mathcal{E}$ is the space of units (set of all possible units; the set of descriptions of units is not neccessary a subset of R^m), be a finite set,

$$d : \mathcal{E} \times \mathcal{E} \to R_0^+$$

be a dissimilarity between units and $w : \mathcal{E} \to R_0^+$ be a weight of units, which is extended to clusters by:

$$\forall X \in \mathcal{E} : w(\{X\}) = w(X)$$

$$C_u \cap C_v = \emptyset \;\Rightarrow w(C_u \cup C_v) = w(C_u) + w(C_v)$$

To obtain the *generalized Ward clustering problem* we must appropriately replace the expression for $p(C)$. Relying on the equality:

$$p(C) = \sum_{X \in C} d_2^2(X, \bar{C}) = \frac{1}{2n_C} \sum_{X,Y \in C} d_2^2(X, Y)$$

we define

$$p(C) = \frac{1}{2w(C)} \sum_{X,Y \in C} w(X) \cdot w(Y) \cdot d(X, Y)$$

Note that d in this definition can be any dissimilarity on $\mathcal{E}$ and not only the squared Euclidean distance.

From the definition we can easily derive the following equality: If $C_u \cap C_v = \emptyset$ then

$$w(C_u \cup C_v) \cdot p(C_u \cup C_v) = w(C_u) \cdot p(C_u) + w(C_v) \cdot p(C_v) + \sum_{X \in C_u, Y \in C_v} w(X) \cdot w(Y) \cdot d(X, Y)$$

In [2] it is also shown how to replace $\bar{C}$ by a generalized, possibly imaginary (with descriptions not neccessary in the same set as $\mathcal{E}$), central element in the way to preserve the properties characteristic for Ward clustering problem.

2 Local optimization methods in general

Often for a given optimization problem (Φ, P) there exist rules which relate to each element of the set Φ some elements of Φ. We call them *local transformations*. They are the basis of the local optimization procedures which starting in an element of Φ repeat moving to an element determined by local transformation which has better value of the criterion function until no such element exists.

The elements which can be obtained by local transformations from a given element are called neighbours – local transformations determine the *neighbourhood relation* $S \subseteq \Phi \times \Phi$ in the set Φ. The *neighbourhood* of element $X \in \Phi$ is called the set $S(X) = \{Y : XSY\}$. The element $X \in \Phi$ is a *local minimum* for the neighbourhood structure (Φ, S) iff

$$\forall Y \in S(X) : P(X) \leq P(Y)$$

The basic scheme of local optimization procedure is therefore very simple:

determine the initial element $X_0 \in \Phi,\ X := X_0$;
while $\exists Y \in S(X) : P(Y) < P(X)$ **repeat** $X := Y$

To obtain a "good" solution and an impression about its quality we repeat the procedure with different (random) X_0.

To build a local optimization procedure we must select an appropriate neighbourhood structure and determine into which among neighbours to move. There are several possibilities to do this:

- choose the first among neighbours which satisfies the conditions;
- choose the neighbour for which the decrease of the value of criterion function is maximal/minimal;
- choose a neighbour at random;
- choose a neighbour which has some additional properties.

Note also that "rich" neighbourhood structure gives better results, but it is also more time consuming.

3 Local optimization clustering methods

Although the basic scheme of local optimization clustering procedure is very simple we have to make some important decisions in its further developement.

To obtain an efficient algorithm we have to balance among:

- rich neighbourhood structure increases the possibility to reach the global minimum. The extreme case represents the case in which all units are neighbours;
- rich neighbourhood structure increases the time spent by each step of the algorithm;
- an efficient algorithm for generating neighbours should exist.

Usually the neighbourhood relation in local optimization clustering procedures is determined by the following two transformations:

- clustering $\mathcal{C}'$ is obtained from the clustering $\mathcal{C}$ by *moving* a unit X_k from cluster C_p to cluster C_q (*transition*):

$$\mathcal{C}' = (\mathcal{C} \setminus \{C_p, C_q\}) \cup \{C_p \setminus \{X_k\}, C_q \cup \{X_k\}\}$$

- clustering $\mathcal{C}'$ is obtained from the clustering $\mathcal{C}$ by *interchanging* units X_u and X_v from different clusters C_p and C_q (*transposition*):

$$\mathcal{C}' = (\mathcal{C} \setminus \{C_p, C_q\}) \cup \{(C_p \setminus \{X_u\}) \cup \{X_v\}, (C_q \setminus \{X_v\}) \cup \{X_u\}\}$$

In both cases only two clusters are changed. Therefore it is useful to introduce for criterion functions of the form

$$P(\mathcal{C}) = \sum_{C \in \mathcal{C}} p(C)$$

the quantity

$$\Delta P(\mathcal{C}, \mathcal{C}') = P(\mathcal{C}) - P(\mathcal{C}') = p(C_p) + p(C_q) - p(C_p') - p(C_q')$$

which allows quicker tests of the condition $P(\mathcal{C}') < P(\mathcal{C})$.

The next important decision is the selection of the representation of dissimilarity d. We can store a dissimilarity matrix or we store the descriptions of units and calculate the dissimilarity between units each time it is needed. The details can be found in any book on clustering [3, 4].

The advantages of the second representation are the following:

- in the case when units are described by few variables we can cluster relatively big sets of units in the main memory;

- in the clustering procedure we can consider, besides dissimilarities, also some other properties of (sets of) units.

The main drawback of this approach is the fact that the repeated computations of the same dissimilarities take a lot of time. For this reason for sets up to 200 units the procedures based on the dissimilarity matrix are mostly used. On the other side the decision for dissimilarity matrix reduce the selection of criterion functions for which ΔP can be efficiently computed. In the following we shall show that this can be done in the case of generalized Ward criterion function.

For this purpose it is useful to introduce the quantity

$$a(C_u, C_v) = \sum_{X \in C_u, Y \in C_v} w(X).w(Y).d(X, Y)$$

Using the quantity $a(C_u, C_v)$ we can express $p(C)$ in the form:

$$p(C) = \frac{a(C, C)}{2w(C)}$$

and the equality mentioned in the introduction of the generalized Ward clustering problem: If $C_u \cap C_v = \emptyset$ then

$$w(C_u \cup C_v).p(C_u \cup C_v) = w(C_u).p(C_u) + w(C_v).p(C_v) + a(C_u, C_v)$$

Let us first analyze the transition of a unit X_k from cluster C_p to cluster C_q: We have $C_p' = C_p \setminus \{X_k\}$, $C_q' = C_q \cup \{X_k\}$,

$$w(C_p).p(C_p) = w(C_p').p(C_p') + a(X_k, C_p') = (w(C_p) - w(X_k)).p(C_p') + a(X_k, C_p')$$

and

$$w(C_q').p(C_q') = w(C_q).p(C_q) + a(X_k, C_q)$$

From $d(X_k, X_k) = 0$ it follows $a(X_k, C_p) = a(X_k, C_p')$. Therefore

$$p(C_p') = \frac{w(C_p).p(C_p) - a(X_k, C_p)}{w(C_p) - w(X_k)} \qquad p(C_q') = \frac{w(C_q).p(C_q) + a(X_k, C_q)}{w(C_q) + w(X_k)}$$

and finally

$$\Delta P(\mathcal{C}, \mathcal{C}') = p(C_p) + p(C_q) - p(C_p') - p(C_q') =$$

$$= \frac{w(X_k).p(C_q) - a(X_k, C_q)}{w(C_q) + w(X_k)} - \frac{w(X_k).p(C_p) - a(X_k, C_p)}{w(C_p) - w(X_k)}$$

In the case when d is the squared Euclidean distance it is possible to derive also expression for corrections of centers [4, p. 69-70].

Now let us analyze the transposition – interchange of units X_u and X_v from different clusters C_p and C_q : We have

$$C_p' = (C_p \setminus \{X_u\}) \cup \{X_v\}, \quad C_q' = (C_q \setminus \{X_v\}) \cup \{X_u\}$$

and

$$\mathcal{C}' = (\mathcal{C} \setminus \{C_p, C_q\}) \cup \{C_p', C_q'\}$$

Because of the symmetry in the expression $\Delta P(\mathcal{C}, \mathcal{C}') = p(C_p) + p(C_q) - p(C_p') - p(C_q')$ it is sufficient to analyze the term $p(C_p) - p(C_p')$. From $p(\{X_v\}) = 0$ it follows:

$$w(C_p').p(C_p') = (w(C_p) + w(X_v) - w(X_u)).p(C_p')$$

$$= (w(C_p) - w(X_u)).p(C_p \setminus \{X_u\}) + a(X_v, C_p \setminus \{X_u\})$$

and from $C_p = (C_p \setminus \{X_u\}) \cup \{X_u\}$ also

$$w(C_p).p(C_p) = (w(C_p) - w(X_u)).p(C_p \setminus \{X_u\}) + a(X_u, C_p \setminus \{X_u\})$$

Combining both equalities we get

$$w(C_p).p(C_p) = (w(C_p) + w(X_v) - w(X_u)).p(C_p') - a(X_v, C_p \setminus \{X_u\}) + a(X_u, C_p \setminus \{X_u\})$$

or in other form

$$p(C_p') = \frac{w(C_p).p(C_p) + a(X_v, C_p \setminus \{X_u\}) - a(X_u, C_p \setminus \{X_u\})}{w(C_p) + w(X_v) - w(X_u)}$$

Therefore

$$p(C_p) - p(C'_p) = \frac{(w(X_v) - w(X_u)).p(C_p) + a(X_u, C_p \setminus \{X_u\}) - a(X_v, C_p \setminus \{X_u\})}{w(C_p) + w(X_v) - w(X_u)}$$

and symmetrically

$$p(C_q) - p(C'_q) = \frac{(w(X_u) - w(X_v)).p(C_q) + a(X_v, C_q \setminus \{X_v\}) - a(X_u, C_q \setminus \{X_v\})}{w(C_q) + w(X_u) - w(X_v)}$$

Both expressions get much simpler form if all units are of the same weight

$$p(C_p) - p(C'_p) = \frac{1}{w(C_p)}(a(X_u, C_p \setminus \{X_u\}) - a(X_v, C_p \setminus \{X_u\}))$$

In the previous lines we have shown that for the generalized Ward clustering problem we can in both cases (transitions and transpositions) efficiently test for the condition $P(\mathcal{C}') < P(\mathcal{C})$. Therefore the local optimization methods based on the dissimilarity matrix are an efficient approach for solving the generalized Ward clustering problem.

References

[1] Batagelj, V., Algorithmic aspects of clustering problem. Proceedings of 7th International Symposium "Computer at the University", Cavtat, 1985. Zagreb. SRCE, 1985, 502 1 - 15.

[2] Batagelj, V., Generalized Ward and related clustering problems. Proceedings of the First Conference of the IFCS, Aachen, 29 June - 1 July, 1987. North-Holland, Amsterdam, 1988, 67-74.

[3] Hartigan, J.A., Clustering Algorithms. John-Wiley, New York, 1975.

[4] Späth, H., Cluster Analyse Algorithmen zur Objekt-Klassifizierung und Datenreduction. R. Oldenbourg Verlag, Munchen, 1977.