Generalized Ward and Related Clustering Problems *

Vladimir BATAGELJ
Department of mathematics, Edvard Kardelj University,
Jadranska 19, 61 000 Ljubljana, Yugoslavia

Abstract

In the paper an attempt to legalize the use of Ward and related clustering methods also for dissimilarities different from squared Euclidean distance is presented.

1 Introduction

"The Ward method is often successfully used for solving clustering problems over dissimilarity matrices which do not consist of squared Euclidean distances between units" [2, p. 145]. "While use of the centroid and incremental sum of squares strategies has been regarded as dubious with other than distance measures and squared Euclidean distance, it is argued that both can be viewed as formulated algebraically and then retain their established space-distortion properties". "There is, however, some empirical evidence that use of the ISS strategy with non-metrics does produce readily interpreted classifications" [1, p. 439-440].

In this paper an attempt to legalize such "extended"uses of Ward and related methods is presented. We start with an overview of the properties of the ordinary Ward criterion function (2-11). Replacing the squared Euclidean distance in (5) by **any** dissimilarity d we get the ***generalized Ward clustering problem*** (1,2,13). To preserve the analogy with the ordinary problem and for notational convenience we introduce the notion of generalized center of cluster (15,16). In general there is no evident interpretation of generalized centers. By the analogy with the ordinary cases, where the dissimilarity d is the squared Euclidean distance, we generalize also the Gower-Bock dissimilarity between clusters (28) and the dissimilarities based on inertia, variance and weighted increase of variance (32-34). For all these generalized dissimilarities we obtain the **same** coefficients in the Lance-Williams-Jambu formula (29,35-38). Therefore the corresponding agglomerative clustering methods can be used for any dissimilarity d and not only for the squared Euclidean distance. At the end the generalized Huyghens theorem and some properties of generalized dissimilarities are given.

*Published in: *Classification and Related Methods of Data Analysis*. H.H. Bock (editor). North-Holland, Amsterdam, 1988. p. 67-74.

2 Ward clustering problem

Let us first repeat some basic facts about ***Ward clustering problem***. It can be posed as follows:

Determine the clustering $\mathbf{C}^\star \in \Pi_k$, for which

$$P(\mathbf{C}^\star) = \min_{\mathbf{C}\in\Pi_k} P(\mathbf{C}) \tag{1}$$

where

$$\Pi_k = \{\mathbf{C} : \mathbf{C} \text{ is partition of set of units } E \text{ and } \text{card}(\mathbf{C}) = k\}$$

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

$$P(\mathbf{C}) = \sum_{C\in\mathbf{C}} p(C) \tag{2}$$

and

$$p(C) = \sum_{X\in C} d_2^2(X, \bar{C}) \tag{3}$$

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

$$[\bar{C}] = \frac{1}{n_C} \sum_{X\in C} [X], \quad n_C = \text{card}(C), \quad [X] \in \mathbb{R}^m \tag{4}$$

and d_2 is the Euclidean distance. We make a distinction between the (name of) the unit X and its value (description) $[X]$.

In the French literature the quantity $p(C)$ (error sum of squares of the cluster C) is often called ***inertia*** of the cluster C [4, p. 30] and denoted $I(C)$.

Let us list some well known equalities relating the above quantities [11, p. 50]:

$$p(C) = \frac{1}{2.n_C} \sum_{X,Y\in C} d_2^2(X, Y) \tag{5}$$

$$\sum_{X\in C} d_2^2(X, U) = p(C) + n_C \cdot d_2^2(U, \bar{C}) \tag{6}$$

$$d_2^2(U, \bar{C}) = \frac{1}{n_C} \sum_{X\in C} (d_2^2(X, U) - d_2^2(X, \bar{C})) \tag{7}$$

$$\bar{C} = \text{argmin}_U \sum_{X\in C} d_2^2(X, U) \tag{8}$$

If $C_u \cap C_v = \emptyset$ then

$$(n_u + n_v) \cdot p(C_u \cup C_v) = n_u \cdot p(C_u) + n_v \cdot p(C_v) + \sum_{X\in C_u, Y\in C_v} d_2^2(X, Y) \tag{9}$$

and

$$p(C_u \cup C_v) = p(C_u) + p(C_v) + \frac{n_u \cdot n_v}{n_u + n_v} d_2^2(\bar{C}_u, \bar{C}_v) \tag{10}$$

The last term in (10)

$$d_W(C_u, C_v) = \frac{n_u \cdot n_v}{n_u + n_v} d_2^2(\bar{C}_u, \bar{C}_v) \tag{11}$$

is also called ***Ward dissimilarity*** between clusters C_u and C_v.

3 The generalized Ward clustering problem

To justify the use of Ward and related agglomerative clustering methods also for dissimilarities different from squared Euclidean distance we have to appropriately generalize the Ward clustering problem. Afterward we shall show that we obtain for generalized problems the same coefficients in Lance-Williams-Jambu formula as for the ordinary problems (based on the squared Euclidean distance). Therefore the agglomerative (hierarchical) procedures corresponding to these problems can be used for any dissimilarity.

To generalize the Ward clustering problem we proceed as follows:

Let $E \subset \mathcal{E}$, where $\mathcal{E}$ is the space of units (set of all possible units; the set $[\mathcal{E}]$ of descriptions of units is not neccessary a subset of $\mathbb{R}^m$), be a finite set,

$$d : \mathcal{E} \times \mathcal{E} \to \mathbb{R}_0^+$$

be a ***dissimilarity*** between units and

$$w : \mathcal{E} \to \mathbb{R}^+$$

be a ***weight*** of units, which is extended to clusters by:

$$\begin{aligned} &\forall X \in \mathcal{E} : w(\{X\}) = w(X) \\ &C_u \cap C_v = \emptyset \quad \Rightarrow \quad w(C_u \cup C_v) = w(C_u) + w(C_v) \end{aligned} \tag{12}$$

To obtain the generalized Ward clustering problem we must appropriately replace formula (3). We define, relying on (5):

$$p(C) = \frac{1}{2 \cdot w(C)} \sum_{X,Y \in C} w(X) \cdot w(Y) \cdot d(X,Y) \tag{13}$$

Note that d in (13) can be **any** dissimilarity on $\mathcal{E}$ and not only the squared Euclidean distance.

From (13) we can easily derive the following generalization of (9) : If $C_u \cap C_v = \emptyset$ then

$$\begin{aligned} w(C_u \cup C_v) \cdot p(C_u \cup C_v) \quad = \quad & w(C_u) \cdot p(C_u) + w(C_v) \cdot p(C_v) + \\ & \sum_{X \in C_u, Y \in C_v} w(X) \cdot w(Y) \cdot d(X,Y) \end{aligned} \tag{14}$$

All other properties are expressed using the notion of center $\bar{C}$ of cluster C and for dissimilarities different from squared Euclidean distance these properties usually do not hold.

There is another possibility: to replace $\bar{C}$ by a generalized, possibly imaginary (with descriptions not neccessary in the set $[\mathcal{E}]$) central element and try to preserve the properties characteristic for Ward problem. For this purpose let us introduce the ***extended space of units***:

$$\mathcal{E}^\star = \{\tilde{C} : C \subset \mathcal{E}, 0 < \mathrm{card}(C) < \infty\} \cup \mathcal{E} \tag{15}$$

The ***generalized center*** of cluster C is called an (abstract) element $\tilde{C} \in \mathbf{C}^\star$ for which the dissimilarity between it and any $U \in \mathcal{E}^\star$ is determined by

$$d(U, \tilde{C}) = d(\tilde{C}, U) = \frac{1}{w(C)} \left(\sum_{X \in C} w(X) \cdot d(X,U) - p(C) \right) \tag{16}$$

When for all units $w(X) = 1$, the right part of (16) can be read: average dissimilarity between unit/center U and cluster C diminished by average radius of cluster C.

It is easy to verify that for every $X \in calE$ and $U \in \mathcal{E}^\star$

$$C = \{X\} \Rightarrow d(U, \tilde{C}) = d(U, X) \tag{17}$$

Note that the equality (16) tells us only how to determine the dissimilarity between two units/centers; but does not provide us with an explicit expression for $\tilde{C}$. Therefore the generalized centers can be viewed as a notationaly convenient notion generalizing our intuition relaying on the squared euclidean distance; although in some cases it may be possible to find a representation for them. It would be interesting to obtain some results in this direction. Can the Young-Householder theorem [9] be used for this purpose?

Multiplying (16) by $w(U)$ summing on U running over C we get

$$p(C) = \sum_{X \in C} w(X) \cdot d(X, \tilde{C}) \tag{18}$$

which generalizes the original definition of $p(C)$ given by formula (3).

Substituting (18) in (16) we obtain generalized (7):

$$d(U, \tilde{C}) = \frac{1}{w(C)} \sum_{X \in C} w(X) \cdot (d(X, U) - d(X, \tilde{C})) \tag{19}$$

which gives for $U = \tilde{C} : d(\tilde{C}, \tilde{C}) = 0$.

Again, applying (16) twice we get:

$$d(\tilde{C}_u, \tilde{C}_v) = \frac{1}{w(C_u) \cdot w(C_v)} \sum_{\substack{X \in C_u \\ Y \in C_v}} w(X) \cdot w(Y) \cdot d(X, Y) - \frac{p(C_u)}{w(C_u)} - \frac{p(C_v)}{w(C_v)} \tag{20}$$

Multiplying (20) by $w(C_u) \cdot w(C_v)$, rearanging and substituting so obtained double sum into (14) we get for $C_u \cap C_v = \emptyset$:

$$p(C_u \cup C_v) = p(C_u) + p(C_v) + \frac{w(C_u) \cdot w(C_v)}{w(C_u \cup C_v)} d(\tilde{C}_u, \tilde{C}_v) \tag{21}$$

which generalizes (10).

Therefore we can define the ***generalized Ward dissimilarity*** between clusters C_u and C_v as:

$$D^W(C_u, C_v) = \frac{w(C_u) \cdot w(C_v)}{w(C_u \cup C_v)} d(\tilde{C}_u, \tilde{C}_v) \tag{22}$$

4 Agglomerative methods for Ward and related clustering problems

It is well known that there are no efficient (exact) algorithms for solving the clustering problem (1), except for some special criterion functions; and there are strong arguments

that no such algorithm exists [5]. Therefore approximative methods such as local optimization and hierarchical (agglomerative and divisive) methods should be used.

The equality (21) gives rise to the following reasoning:

Suppose that the criterion function $P(\mathbf{C})$ takes the form

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} p_D(C) \tag{23}$$

where

$$p_D(C_u \cup C_v) = p_D(C_u) + p_D(C_v) + D(C_u, C_v) \tag{24}$$

and $D(C_u, C_v)$ is a dissimilarity between clusters C_u and C_v. Some examples of such dissimilarities will be given in the continuation.

Let be $\mathbf{C}' = (\mathbf{C} \setminus \{C_u, C_v\}) \cup \{C_u \cup C_v\}$ then

$$P(\mathbf{C}') = P(\mathbf{C}) + D(C_u, C_v) \tag{25}$$

We can found on the last equality the following ***greedy heuristic***:

> if we start with $\mathbf{C}_0 = \{\{X\} : X \in E\}$ and at each step we join the most similar pair of clusters, we probably obtain clusterings which are (almost) optimal.

This heuristic is the basis for the agglomerative methods for solving the clustering problem. It is useful to slightly generalize ([7], [3]) the equality (24) to:

$$p_D(C) = \min_{\emptyset \subset C' \subset C} (p_D(C') + p_D(C \setminus C') + D(C', C \setminus C')) \tag{26}$$

which still leads to the same heuristic.

Therefore we can, for a given dissimilarity between clusters $D(C_u, C_v)$, look at the agglomerative methods as the approximative methods for solving the clustering problem (1) for the criterion function defined by (23,26) and $p_D(\{X\}) = 0$.

Very popular scheme of agglomerative clustering algorithms is the scheme based on the ***Lance-Williams-Jambu*** formula for updating dissimilarities between clusters [10, 6]:

$$\begin{aligned} D(C_s, C_p \cup C_q) \;=\; & \alpha_1 \cdot D(C_s, C_p) + \alpha_2 \cdot D(C_s, C_q) + \beta \cdot D(C_p, C_q) + \\ & \gamma \cdot |D(C_s, C_p) - D(C_s, C_q)| + \delta_1 \cdot f(C_p) + \delta_2 \cdot f(C_q) + \delta_3 \cdot f(C_s) \end{aligned} \tag{27}$$

where $f(C)$ is a value of cluster C. In the following (29,35-38) we shall show that the (generalized) Ward and some related dissimilarities satisfy this formula.

We can generalize the ***Gower-Bock dissimilarity*** by defining

$$D^G(C_u, C_v) = d(\tilde{C}_u, \tilde{C}_v) \tag{28}$$

It is not difficult to verify, using (14) and (21), the Gower-Bock equality (generalized median theorem, [4, p. 57]):

$$D^G(C_s, C_p \cup C_q) = \frac{w_p}{w_{pq}} D^G(C_s, C_p) + \frac{w_q}{w_{pq}} D^G(C_s, C_q) - \frac{w_p \cdot w_q}{w_{pq}^2} D^G(C_p, C_q) \tag{29}$$

For this purpose it is useful to introduce the quantity

$$a(C_u, C_v) = \frac{1}{w_u \cdot w_v} \sum_{\substack{X \in C_u \\ Y \in C_v}} w(X) \cdot w(Y) \cdot d(X, Y) \tag{30}$$

for which the relations

$$\begin{aligned} w_{uv} \cdot a(C_t, C_u \cup C_v) &= w_u \cdot a(C_t, C_u) + w_v \cdot a(C_t, C_v) \\ w_{uv} \cdot p(C_u \cup C_v) &= w_u \cdot p(C_u) + w_v \cdot p(C_v) + w_u \cdot w_v \cdot a(C_u, C_v) \\ D^G(C_u, C_v) &= d(\tilde{C}_u, \tilde{C}_v) = a(C_u, C_v) - \frac{p(C_u)}{w_u} - \frac{p(C_v)}{w_v} \end{aligned} \tag{31}$$

hold.

Using (29) and relations among the following dissimilarities [6, 10, p. 127]:

***Inertia*:**

$$D^I(C_u, C_v) = p(C_u \cup C_v) \tag{32}$$

***Variance*:**

$$D^V(C_u, C_v) = \mathrm{var}(C_u \cup C_v) = \frac{1}{w_{uv}} p(C_u \cup C_v) \tag{33}$$

***Weighted increase of variance*:**

$$D^v(C_u, C_v) = \mathrm{var}(C_u \cup C_v) - \frac{1}{w_{uv}}(w_u \cdot \mathrm{var}(C_u) + w_v \cdot \mathrm{var}(C_v)) = \frac{1}{w_{uv}} D^W(C_u, C_v) \tag{34}$$

we derive the well known equalities, $w = w(C_p \cup C_q \cup C_s)$:

$$D^W(C_s, C_p \cup C_q) = \frac{w_{ps}}{w} D^W(C_s, C_p) + \frac{w_{qs}}{w} D^W(C_s, C_q) - \frac{w_s}{w} D^W(C_p, C_q) \tag{35}$$

$$\begin{aligned} D^I(C_s, C_p \cup C_q) &= \frac{w_{ps}}{w} D^I(C_s, C_p) + \frac{w_{qs}}{w} D^I(C_s, C_q) + \frac{w_{pq}}{w} D^I(C_p, C_q) - \\ &\quad \frac{w_p}{w} p(C_p) - \frac{w_q}{w} p(C_q) - \frac{w_s}{w} p(C_s) \end{aligned} \tag{36}$$

$$\begin{aligned} D^V(C_s, C_p \cup C_q) &= (\frac{w_{ps}}{w})^2 D^V(C_s, C_p) + (\frac{w_{qs}}{w})^2 D^V(C_s, C_q) + (\frac{w_{pq}}{w})^2 D^V(C_p, C_q) - \\ &\quad \frac{w_s}{w^2} p(C_p) - \frac{w_q}{w^2} p(C_q) - \frac{w_s}{w^2} p(C_s) \end{aligned} \tag{37}$$

$$D^v(C_s, C_p \cup C_q) = (\frac{w_{ps}}{w})^2 D^v(C_s, C_p) + (\frac{w_{qs}}{w})^2 D^v(C_s, C_q) - \frac{w_s \cdot w_{pq}}{w^2} D^v(C_p, C_q) \tag{38}$$

which therefore hold also in the case when we take for $p(C)$ in relations (27-29) the generalized cluster error function (13). So, we can conclude that the coefficients in the Lance-Williams-Jambu formula for Ward and related agglomerative clustering methods are valid for **any** dissimilarity d.

5 The generalized Huyghens theorem

Taking in (20) $C_u = C$ and $C_v = E$, multiplying it with $w(C) \cdot w(E)$ and summing it over all $C \in \mathbf{C}$, we obtain:

$$p(E) = \sum_{C \in \mathbf{C}} (w(C) \cdot d(\tilde{C}, \tilde{E}) + p(C)) \tag{39}$$

If we denote [4, p. 50]:

$$I = p(E), \qquad I_W = \sum_{C \in \mathbf{C}} p(C), \qquad I_B = \sum_{C \in \mathbf{C}} w(C) \cdot d(\tilde{C}, \tilde{E}) \tag{40}$$

we can express (39) in the form (***generalized Huyghens theorem***):

$$I = I_W + I_B \tag{41}$$

6 Some properties of the generalized dissimilarities

For dissimilarities different from d_2^2 the extended dissimilarity $d(U, \tilde{C})$ is not neccessarily positive quantity for every $U \in \mathcal{E}^\star$; so do $D^G(C_u, C_v)$ and $D^W(C_u, C_v)$.

To obtain a partial answer to this problem we substitute (13) into (14) giving

$$\begin{aligned} d(U, \tilde{C}) &= \frac{1}{2 \cdot w(C)^2} \sum_{X,Y \in C} w(X) \cdot w(Y) \cdot (2 \cdot d(X, U) - d(X, Y)) \\ &= \frac{1}{2 \cdot w(C)^2} \sum_{X,Y \in C} w(X) \cdot w(Y) \cdot (d(X, U) + d(U, Y) - d(X, Y)) \end{aligned} \tag{42}$$

Therefore: if the dissimilarity d satisfies the triangle inequality then for each $U \in \mathcal{E}$ it holds

$$d(U, \tilde{C}) \geq 0 \tag{43}$$

The property (43) has another important consequence. Considering it in (16) we get: for each $U \in \mathcal{E}$

$$\sum_{X \in C} w(X) \cdot d(X, U) \geq p(C) \tag{44}$$

which in some sense generalizes (8).

Note that (42) does not imply $d(\tilde{C}_u, \tilde{C}_v) \geq 0$, because we do not know that

$$d(X, \tilde{C}) + d(\tilde{C}, Y) \geq d(X, Y) \tag{45}$$

but it is easy to verify that: if d satisfies the triangle inequality then also

$$\begin{aligned} d(\tilde{C}, Z) + d(Z, X) &\geq d(\tilde{C}, X) \\ d(\tilde{C}_u, Z) + d(Z, \tilde{C}_v) &\geq d(\tilde{C}_u, \tilde{C}_v) \end{aligned} \tag{46}$$

for $X, Z \in \mathcal{E}$.

So, it is still possible that there exists a center $\tilde{C}_1$ for which

$$\sum_{X \in C} w(X) \cdot d(X, \tilde{C}_1) < p(C) \tag{47}$$

Note also that d_2^2 does not satisfy the triangle inequality.

The main open question remains: for which dissimilarities d it holds $d(U, V) \geq 0$ for every $U, V \in \mathcal{E}^\star$? This inequality combined with (16) gives us the generalized (8).

References

[1] Abel, D.J., Williams, W.T., A re-examination of four classificatory fusion strategies. The Computer Journal, **28**(1985), 439-443.

[2] Anderberg, M.R., Cluster Analysis for Applications. Academic Press, New York, 1973.

[3] Batagelj, V., Agglomerative methods in clustering with constraints. Submitted, 1984.

[4] Bouroche, J-M., and Saporta, G., L'analyse des donnees, Que sais-je? Presses Universitaires de France, Paris, 1980.

[5] Brucker, P., On the complexity of clustering problems, in: Henn, R., Korte, B. and Oettli, W., (eds.), Optimization and Operations Research; Proceedings in Lecture Notes in Economics and Mathematical Systems, Vol. 157. Springer-Verlag, Berlin, 1978.

[6] Diday, E., Inversions en classification hierarchique: application a la construction adaptive d'indices d'agregation. Revue de Statistique Appliquee, **31**(1983), 45-62.

[7] Ferligoj, A. and Batagelj, V., Clustering with relational constraint. Psychometrika, 47(1982), 413-426.

[8] Gower, J.C., Legendre, P., Metric and Euclidean Properties of Dissimilarity Coefficients. Journal of Classification, **3**(1986), 5-48.

[9] Gregson, A.M.R, Psychometrics of Similarity. Academic Press, New York, 1975.

[10] Jambu, M., Lebeaux, M-O., Cluster analysis and data analysis. North-Holland, Amsterdam, 1983.

[11] Späth, H., Cluster analyse algorithmen. R. Oldenbourg Verlag, München, 1977.