The Ward cluster algorithm of Viscovery SOMine

Viscovery SOMine implements two cluster algorithms that are based on the classical hierarchical agglomerative cluster method of Ward: The first method is basically the unmodified Ward method, the second modifies it in such a way that the topological neighborhood as defined by the SOM influences the cluster merge steps. Both methods use the nodes as data points instead of the original data from which the map was trained. The SOM is effectivly regarded as the compressed original data. By this method cluster methods can be applied to data sets of virtually unlimited size, which would not be possible otherwise because data sets become quickly intractable by these methods as the data volume grows.

Ward clustering in general

The classical cluster method of Ward belongs to the hierarchical agglomerative cluster algorithms which are characterized as follows: Starting with a clustering, where each single node forms a cluster by itself, in each step of the algorithm the two clusters with minimal distance (according to a distance measure characterizing the specific algorithm) are merged. This minimal distance is called *distance niveau* of the step.

The distance measure characterizing Ward's method is based on the variance criterion (goal: small variance within each cluster and large variance between the clusters). In each step the two clusters are merged whose merging contributes least to the variance criterion which is increasing in each step. This distance measure is called the Ward distance and is defined as:

$$d_{rs} := \frac{n_r \cdot n_s}{n_r + n_s} \cdot \left\| \bar{x}_r - \bar{x}_s \right\|^2$$

where r and s denote two specific clusters, n_r and n_s denote the number of data points in the two clusters, and $\bar{x}_r$ and $\bar{x}_s$ denote the centers of gravity of the clusters; $\|.\|$ is the Euclidean norm.

Starting from the full distance matrix (lower triangle matrix as the distance measure is commutative), at every step a row and a column is stripped (and a different row and column is updated) until the matrix is completely cleared and only one cluster remains.

The mean and cardinality of the new cluster built as product of the merge step is computed as follows:

$$\bar{x}_r^{(new)} := \frac{1}{n_r + n_s} \cdot (n_r \cdot \bar{x}_r + n_s \cdot \bar{x}_s) \,,$$

$$n_r^{(new)} := n_r + n_s$$

Ward clustering in Viscovery SOMine

As a specialty, the distance matrix is initialized in a manner that takes into account the number of data records matching to the nodes of the map. Nodes with many matching data records are weighted stronger than nodes with fewer matching records.

As distance measure we have to use a modified Ward distance because it is likely that the SOM contains "empty" nodes:

Let r and s be the two nodes for which we want to compute the distance.

Let further be n_r , n_s the number of data records that match the nodes r and s[1] and $\bar{x}_r$ and $\bar{x}_s$ their node vectors.

Then the distance d_{rs} is defined as follows:

$$d_{rs} := \begin{cases} 0 & \text{if } n_r = n_s = 0, \\ \dfrac{n_r \cdot n_s}{n_r + n_s} \cdot \|\bar{x}_r - \bar{x}_s\|^2 & \text{otherwise.} \end{cases}$$

This definition ensures that during the first merge steps only nodes (and in the sequel clusters) with $n_r = 0$ ("*empty clusters*") are merged until only clusters with $n_r > 0$ remain. Note that if there is at least one empty cluster, there exist many entries in the distance matrix with $d_{rs} = 0$, which are all candidates for the next merge step (since all these Ward distances 0 are minimal). Our implementation chooses among them those (empty or non-empty) clusters which are Euclidean-nearest.[2]

SOM-Ward clusters

For the SOM-Ward, we re-define the distance measure:

$$d'_{rs} := \begin{cases} d_{rs} & \text{if clusters } r \text{ and } s \text{ are adjacent in the SOM,} \\ \infty & \text{otherwise.} \end{cases}$$

Thus, the SOM-Ward distance observes the topological location of the clusters. In particular, two clusters that are not adjacent in the SOM are never considered to be merged.

Cluster indicator

Viscovery SOMine computes an indicator for each element of this hierarchical sequence of clusterings, which indicates a quality measure for each cluster count. It is computed as follows: The series of Ward distances is computed for all possible numbers of clusters. Then the distance niveau is normalized with an exponential function. The ratio

[1] The classic Ward method would set $n_r = 1$.

[2] Note that although empty clusters have a zero Ward distance d_{rs} to all other clusters, they still have a measurable Euclidean distance! Also, our implementation does not necessarily choose the Euclidean-nearest clusters in this situation, but it ensures that all empty nodes are merged with their Euclidean-nearest non-empty node for all clusterings that do not contain empty clusters.

of the distance niveaus of two neighbored clusterings makes up the cluster indicator. The cluster indicator is merely a means to help find an initial clustering. A high indicator points to a possibly good clustering.

Let C be the number of non-empty nodes in the SOM.[3]

Here are the exact formulas for the indicator $\mathrm{I}(c)$ of c clusters:

$$\mathrm{I}(c) := \max(0, \mathrm{I}'(c)) \cdot 100$$

where

$$\mathrm{I}'(c) := \frac{m(c)}{m(c+1)} - 1$$

and

$$m(c) := d(c) \cdot c^{b}$$

and $d(c)$ is the Ward distance niveau that was used to merge c clusters into $c-1$ clusters; and $3 \le c < C$.

As stated above the $d(c)$ "behaves" like c^{-b}. $-b$ is the linear regression coefficient for the "data points" $[\ln(c), \ln(d(c))]$ (where $2 \le c \le C$):

$$-b := \frac{s_{gd} - \bar{g} \cdot \bar{d}}{s_{gg} - \bar{g}^{2}}, \text{ where } g := \ln c \text{ and } d := \ln d(c).$$

Further we define $\mathrm{I}(1) := 0$ and $\mathrm{I}(2) := 0$. And for SOM-Ward clusters we further define $\mathrm{I}(c) := 0$ for inversions at c clusters, i.e. when $d(c) < d(c+1)$.

Note that for the Ward method the distances $d(c)$ are monotonically decreasing (with c increasing), but that is not true for SOM-Ward. Also note that $\mathrm{I}'(c)$ may be negative, but such clusterings are uninteresting, and we set the displayed indicator to zero.

The idea behind this is that when $d(c)$ is high, but $d(c+1)$ is low, c clusters is a good clustering because the next merge step (resulting in $c-1$ clusters) would result in a high variance within the clusters.

[3] This is the number of clusters that is obtained when the last empty cluster is merged to a non-empty node.