Silhouette (clustering)

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.[1]

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.

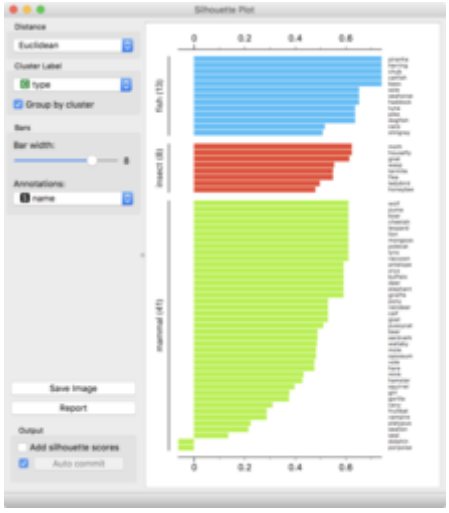
Definition

Assume the data have been clustered via any technique, such as k-means, into k clusters.

For data point $i \in C_i$ (data point i in the cluster C_i), let

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

be the mean distance between i and all other data points in the same cluster, where $d(i, j)$ is the distance between data points i and j in the cluster C_i (we divide by $|C_i| - 1$ because we do not include the distance $d(i, i)$ in the sum). We can interpret $a(i)$ as a measure of how well i is assigned to its cluster (the smaller the value, the better the assignment).



A plot showing silhouette scores from three types of animals from the Zoo dataset as rendered by Orange data mining suite. At the bottom of the plot, silhouette identifies dolphin and porpoise as outliers in the group of mammals.

We then define the mean dissimilarity of point i to some cluster C_k as the mean of the distance from i to all points in C_k (where $C_k \neq C_i$).

For each data point $i \in C_i$, we now define

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

to be the *smallest* (hence the $\min$ operator in the formula) mean distance of i to all points in any other cluster, of which i is not a member. The cluster with this smallest mean dissimilarity is said to be the "neighboring cluster" of i because it is the next best fit cluster for point i.

We now define a *silhouette* (value) of one data point i

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

and

$$s(i) = 0, \text{ if } |C_i| = 1$$

Which can be also written as:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

From the above definition it is clear that

$$-1 \leq s(i) \leq 1$$

Also, note that score is 0 for clusters with size = 1. This constraint is added to prevent the number of clusters from increasing significantly.

For $s(i)$ to be close to 1 we require $a(i) \ll b(i)$. As $a(i)$ is a measure of how dissimilar i is to its own cluster, a small value means it is well matched. Furthermore, a large $b(i)$ implies that i is badly matched to its neighbouring cluster. Thus an $s(i)$ close to one means that the data is appropriately clustered. If $s(i)$ is close to negative one, then by the same logic we see that i would be more appropriate if it was clustered in its neighbouring cluster. An $s(i)$ near zero means that the datum is on the border of two natural clusters.

The mean $s(i)$ over all points of a cluster is a measure of how tightly grouped all the points in the cluster are. Thus the mean $s(i)$ over all data of the entire dataset is a measure of how appropriately the data have been clustered. If there are too many or too few clusters, as may occur when a poor choice of k is used in the clustering algorithm (e.g.: k-means), some of the clusters will typically display much narrower silhouettes than the rest. Thus silhouette plots and means may be used to determine the natural number of clusters within a dataset. One can also increase the likelihood of the silhouette being maximized at the correct number of clusters by re-scaling the data using feature weights that are cluster specific.[2]

Kaufman et al. introduced the term *silhouette coefficient* for the maximum value of the mean $s(i)$ over all data of the entire dataset. [3]

$$SC = \max_k \tilde{s}\,(k)$$

Where $\tilde{s}\,(k)$ represents the mean $s(i)$ over all data of the entire dataset for a specific number of clusters k.

See also

- k-medoids
- Determining the number of clusters in a data set

References

1. Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics*. **20**: 53–65. doi:10.1016/0377-0427(87)90125-7 (https://doi.org/10.1016%2F0377-0427%2887%2990125-7).
2. R.C. de Amorim, C. Hennig (2015). "Recovering the number of clusters in data sets with noise features using feature rescaling factors". *Information Sciences*. **324**: 126–145.

arXiv:1602.06989 (https://arxiv.org/abs/1602.06989). doi:10.1016/j.ins.2015.06.039 (https://doi.org/10.1016%2Fj.ins.2015.06.039).

3. Leonard Kaufman; Peter J. Rousseeuw (1990). *Finding groups in data : An introduction to cluster analysis* (https://archive.org/details/findinggroupsind00kauf/page/87). Hoboken, NJ: Wiley-Interscience. p. 87 (https://archive.org/details/findinggroupsind00kauf/page/87). doi:10.1002/9780470316801 (https://doi.org/10.1002%2F9780470316801). ISBN 9780471878766.
