Compliance between partitioning and distance information: An alternative measure of cluster quality is an estimation of the degree of distance information preserved from the original datasets in clusters. This measure uses the cophenetic matrix C that is symmetric of size $N \times N$, and N is the number of samples in the dataset. Each element $C(i,j)$ of the matrix C acts as an indicator if a pair of samples is assigned to a common cluster. For the evaluation of a hierarchical clustering, the cophenetic matrix can also be constructed to reflect the level within the dendrogram. Here, an entry $C(i,j)$ represents the level within the dendrogram at which the two samples i and j are first assigned to the same cluster.

Several methods have been proposed that capture the correlation between the cophenetic matrix and the original dissimilarity matrix to assess the preservation of distances under different distance functions and within different feature spaces or to compute the dendrograms obtained for different algorithms.

9.4.2 Performance Evaluation Using Validity Indices

A great deal of research is focused on finding the correct or optimal number of partitions. Cluster validity indices help address this problem by estimating the correct number of clusters and finding the quality clusters (Halkidi et al. 2001). The most commonly used validity indices have been described below (Azuaje and Bolshakova 2002).

9.4.2.1 Silhouette Index (SI)

The computation of the silhouette index is described by the following steps:

1. For a given cluster, $X_j (j = 1,...,c)$, the silhouette technique assigns a silhouette width, $s(i)(i = 1,...,m)$, to the ith sample of X_j. This value is defined as

$$s(i) = (b(i) - a(i))/\max\{a(i), b(i)\},$$

   where $a(i)$ is the average distance between the ith sample and all of the samples included in X_j, and $b(i)$ is the minimum average distance between the ith sample and all of the samples clustered in $X_k (k = 1,...,c; k \neq j)$. $s(i)$ lies between −1 and 1.
2. When the value of $s(i)$ is near 1, it can be assumed that the ith sample has been assigned to an appropriate cluster.
3. When $s(i)$ is near zero, it can be assumed that the ith sample can be assigned to the nearest neighboring cluster.
4. When $s(i)$ is near −1, it can be assumed that the ith sample has been misclassified (Rousseeuw 1987).

A global silhouette value or silhouette index, GS_u, can be used as a validity index for a partition U. This measure can be determined using Equation 9.8, as shown below, which helps estimate the "correct" number of clusters for partition U (Rousseeuw 1987). Thus, a high value of silhouette index indicates that partition U is a better or optimal cluster. This method can be represented as

$$GS_u = \frac{1}{c}\sum_{j=1}^{c} S_j \tag{9.8}$$

9.4.2.2 Davies-Bouldin and Dunn's Index

Unlike the SI, the Davies-Bouldin (DB) index is defined as the ratio of the sum of the within-cluster scatter to the between-cluster scatter (Davies and Bouldin 1979). A small DB value indicates a compact cluster. Mathematically, such a reading can be defined as

$$DB = \frac{1}{n}\sum_{i=1,i\neq j}^{n} \max\left(\frac{\sigma_i + \sigma_j}{\sigma(c_i, c_j)}\right), \tag{9.9}$$

where n is number of clusters, σ_i is the average distance of all patterns in cluster i to their cluster center c_i, σ_j is the average distance of all patterns in cluster j to their cluster center c_j, and $d(c_i, c_j)$ is the distance of cluster centers c_i and c_j.

Similarly, the Dunn index (D) is defined as the ratio of the minimum intracluster distance to the maximum intercluster distance. The Dunn index lies within the range of 0 to 1, and values approaching 1 correspond to good clusters. The index is given by

$$D = d_{\min}/d_{\max}, \tag{9.10}$$

where $d_{\min}$ is the minimum distance between two objects from different clusters, and $d_{\max}$ is the maximum distance of two objects from the same cluster.

9.4.2.3 Calinski Harabasz (CH) Index

The Calinski Harabasz (CH) index, proposed by Maulik and Bandopadhyay (2002), is computed as

$$(race(b)/(k-1)/(trace(w)/(n-k)), \tag{9.11}$$

where b and w represent the between- and within-cluster scatter matrices, respectively, and k and n represent the cluster and data points, respectively.

The trace for the between-cluster scatter matrix B can be written as

$$Trace(b) = \sum_{k=1}^{k} nk \,||\, zk - z \,||^2, \tag{9.12}$$

where nk is the number of points in cluster k and z is the centroid of the entire dataset. The trace of the within-cluster scatter matrix W can be written as $trace(W)$,

$$Trace(w) = \sum_{k=1}^{k} \sum_{i=1}^{nk} (xi - zk)^2, \tag{9.13}$$

9.4.2.4 Rand Index

A Rand index determines the similarity between two partitions with respect to positive and negative agreements and can be used to assess the degree of agreement between two clusters (Rand 1971; Youness and Saporta 2010). The Rand index ranges in value from 0 to 1; a higher Rand index value indicates a higher similarity between two partitions. This index is defined as the ratio of the number of agreements between two partitions divided by the total number of objects (Hubert and Arabie 1985).

9.5 Conclusion

This chapter provides an explanation of computation techniques used to validate and benchmark results obtained using either clustering or classification techniques on datasets. Moreover, it should be noted that these techniques are used for hypothesis testing in bioinformatics.

References

Abeel, T., Y. Van de Peer, and Y. Saeys. Toward a gold standard for promoter prediction evaluation. *Bioinformatics* 25 (2009): i313–i320.

Azuaje, F., and N. Bolshakova. Clustering genome expression data: Design and evaluation principles. In D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *Understanding and using microarray analysis techniques: A practical guide*. London: Springer Verlag, 2002, 230–245.