Validation Indicies

Samantha Bothwell | Last Updated: May 17th, 2021

Cluster Validation indicies

Cluster Validation is a method for evaluating the accuracy of clustering algorithm results.

Internal Cluster Validation

Internal cluster validation is used when we do not know the true clusters. This could be useful for estimating the number of clusters without any external data. Internal validation is based on the following [1, 2]:

- **Compactness** : A measure of how close objects within the same cluster are. A lower within-cluster variance of the measures indicates better compactness. Measures of compactness include max/avg pairwise distance and max/avg center-based distance.
- **Separation** : A measure of how well-separated a cluster is from other clusters. Measures of separation include the distance between cluster centers and the pairwise minimum distance b/w objects in different clusters.
- **Connectivity** : A measure of the extent to which items are placed in the same cluster as their nearest neighbors in the data space. The connectivity has a value between 0 and infinity and should be minimized.

Some common internal cluster validation indicies are :

- **Silhoutte Index** : The Silhoutte index is a normalized sum index that measures cohesion based on the distance between all points in the same cluster and separation based on the nearest neighbour distance. The Silhouette index is calculated as [3] :

$$Sil(C) = \frac{1}{N}\sum_{c_k \in C}\sum_{x_i \in c_k}\frac{b(x_i, c_k) - a(x_i, c_k)}{max\{a(x_i, c_k), b(x_i, c_k)\}}$$

where
$a(x_i, c_k) = \frac{1}{|c_k|}\sum_{x_j \in c_k} dist_{eucl}(x_i, x_j)$,
$b(x_i, c_k) = \min_{c_l \in C/c_k}\left[\frac{1}{|c_l|}\sum_{x_j \in c_l} dist_{eucl}(x_i, x_j)\right]$

- **Calinski-Harabasz Index** : The Calinski-Harabasz Index is a ratio-type index that measures cohesion based on the distances from cluster points to the cluster centroid and separation based on the distance from the cluster centroids to the global centroid [3] :

$$CH(C) = \frac{N-K}{K-1}\frac{\sum_{c_k \in C}|c_k| dist_{eucl}(\bar{c_k}, \bar{X})}{\sum_{c_k \in C}\sum_{x_i \in c_k} dist_{eucl}(x_i, \bar{c_k})}$$

- **Davies-Bouldin Index** : The DB index is a ratio between "within-cluster" and "between-cluster" distances, where a smaller value is better [3, 4] :

$$\frac{1}{k}\sum_{i=1}^{k}\max_{1 \leq j \leq k, j \neq i}\left(\frac{diam(c_i) + diam(c_j)}{dist(c_i, c_j)}\right)$$

.

- **Dunn Index** : The Dunn index identifies clusters that are compact and separated. The Dunn Index is calculated as $D = \frac{\text{min inter-cluster separation}}{\text{max intra-cluster diameter}}$ [1]. More formally, let c_i represent the i-cluster then [3, 4] :

$$D = \min_{1 \leq i \leq k} \left(\min_{i+1 \leq j \leq k} \left(\frac{dist(c_i, c_j)}{\max_{1 \leq l \leq k} diam(c_l)} \right) \right)$$

  where $dist(c_i, c_j)$ is the distance between clusters c_i and c_j where $dist(c_i, c_j) = \min_{x_i \in c_i, x_j \in c_j} d(x_i, x_j)$,
  $d(x_i, x_j)$ is the distance between data points $x_i \in c_i$ and $x_j \in c_j$,
  $diam(c_l)$ is the diameter of cluster c_l where $diam(c_l) = \max_{x_{l_1}, x_{l_2} \in c_l} d(x_{l_1}, x_{l_2})$

External Cluster Validation

External cluster validation is used when we know the true classification of the data. This approach is often used for selecting which clustering algorithm is best given the specific data set.

- **Rand Index** : The Rand index can take on values between 0 and 1, where 0 indicates no agreement and 1 indicates perfect agreement. The Rand Index is essentially calculated as $RI = \frac{TP+TN}{TP+FP+FN+TN}$ [5].

- **Adjusted Rand Index** : The adjusted rand index is an extension of the rand index that adjusts for the chance grouping of elements. Given that $TP + TN$ can be simplified to a linear transformation of $\sum_{i,j} \binom{n_{ij}}{2}$, where n_{ij} is the number of elements with true classification u_i with cluster assignment v_j, the adjusted Rand index can be simplified to [6, 7]:

$$\frac{\sum_{i,j} \binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2} \right] / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2} \right] - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2} \right] / \binom{n}{2}}$$

  where $n_{i.}$ and $n_{.j}$ represent the number of elements with true classification u_i and cluster assignment v_j respectively.

- **Jaccard Index** : The Jaccard Index is a similarity measure between a true cluster, A, and the assigned cluster, B. The Jaccard index is then calculated as [8]:

$$J(A, B) = \frac{A \cap B}{A \cup B}$$

- **Fowlkes-Mallows Index** : The Fowlkes-Mallows Index (FM) is calculated as [9]:

$$FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}$$

- **Variation of Information** : The variation of information (VI) criteria is a measure of the extent to which information is lost or gained by changing elements from the true clustering C to the assigned clustering C'. Formally, VI is calculated as [10]:

$$VI(C, C') = H(C) + H(C') - 2I(C, C')$$

  where $H(C)$ is the entropy associated with cluster C and $I(C, C')$ is the mutual information.
  $H(C)$ takes a value of 0 when there is no uncertainty (i.e. there is only one cluster). $I(C, C')$ is the reduction in uncertainty, averaged over all points.

References

[1] https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/

[2] http://datamining.rutgers.edu/publication/internalmeasures.pdf

[3] https://ccc.inaoep.mx/~ariel/2013/An%20extensive%20comparative%20study%20of%20cluster%20validity%20indices.pdf

[4] https://arxiv.org/ftp/arxiv/papers/1507/1507.03340.pdf

[5] https://en.wikipedia.org/wiki/Rand_index

[6] https://www.researchgate.net/profile/Ka-Yee-Yeung/publication/239537124_Details_of_the_Adjusted_Rand_index_and_Clustering_algorithms_Supplement_to_the_paper_An_empirical_study_on_Principal_Component_Analysis_for_clustering_gene_expression_data_to_appear_in_Bioinformatics/links/543b62a30cf24a6ddb976fd3/Details-of-the-Adjusted-Rand-index-and-Clustering-algorithms-Supplement-to-the-paper-An-empirical-study-on-Principal-Component-Analysis-for-clustering-gene-expression-data-to-appear-in-Bioinformatics.pdf

[7] https://link.springer.com/article/10.1007/BF01908075

[8] Hwang, C.-M., Yang, M.-S., & Hung, W.-L. (2018). New similarity measures of intuitionistic fuzzy sets based on the Jaccard index with its application to clustering. International Journal of Intelligent Systems, 33(8), 1672–1688. https://doi.org/10.1002/int.21990

[9] https://en.wikipedia.org/wiki/Fowlkes%E2%80%93Mallows_index

[10] Meilă, M. (2007). Comparing clusterings—An information based distance. Journal of Multivariate Analysis, 98(5), 873–895. https://doi.org/10.1016/j.jmva.2006.11.013