Clustering QC

Mikhail Dozmorov
Fall 2017

Assess cluster fit and stability

- Most often ignored.
- Cluster structure is treated as reliable and precise
- BUT! Clustering is generally VERY sensitive to noise and to outliers
- Measure cluster quality based on how “tight” the clusters are.
- Do genes in a cluster appear more similar to each other than genes in other clusters?

Clustering evaluation methods

- Sum of squares
- Homogeneity and Separation
- Cluster Silhouettes and Silhouette coefficient: how similar genes within a cluster are to genes in other clusters
- Rand index
- Gap statistics
- Cross-validation

Sum of squares

- A good clustering yields clusters where genes have small within-cluster sum-of-squares (and high between-cluster sum-of-squares).

Homogeneity

- **Homogeneity** is calculated as the average distance between each gene expression profile and the center of the cluster it belongs to

$$H_k = \frac{1}{N_g} \sum_{i \in k} d(X_i, C(X_i))$$

N_g - total number of genes in the cluster

Separation

- **Separation** is calculated as the weighted average distance between cluster centers

$$S_{ave} = \frac{1}{\sum_{k \neq l} N_k N_l} \sum_{k \neq l} N_k N_l d(C_k, C_l)$$

Homogeneity and Separation

- Homogeneity reflects the compactness of the clusters while Separation reflects the overall distance between clusters
- Decreasing Homogeneity or increasing Separation suggest an improvement in the clustering results

Variance Ratio Criterion (VCR)

$$VRC_k = (SS_B/(K-1))/(SS_W/(N-K))$$

- SS_B – between-cluster variation
- SS_W – within-cluster variation

The goal is to maximize VRC_k over the clusters

$$\kappa_k = (VRC_{k+1} - VRC_k) - (VRC_k - VRC_{k-1})$$

- Select K to minimize the value of $kappa_k$
- Calinski & Harabasz (1974) http://www.tandfonline.com/doi/abs/10.1080/03610927408827101

Silhouette

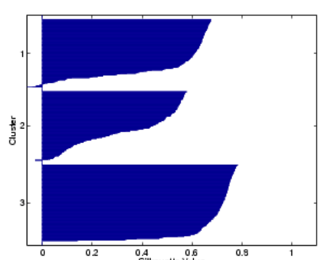
- Good clusters are those where the genes are close to each other compared to their next closest cluster.

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

- $b(i) = min(AVGD_{BETWEEN}(i, k))$
- $a(i) = AVGD_{WITHIN}(i)$
- How well observation i matches the cluster assignment. Ranges $-1 < s(i) < 1$
- Overall silhouette: $SC = \frac{1}{N_g} \sum_{i=1}^{N_g} s(i)$
- Rousseeuw, Peter J. "**Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.**" Journal of Computational and Applied Mathematics 1987 http://www.sciencedirect.com/science/article/pii/0377042787901257

Silhouette plot

- The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters.
- Silhouette width near +1 indicates points that are very distant from neighboring clusters
- Silhouette width near 0 indicate points that are not distinctly in one cluster or another
- Negative width indicates points are probably assigned to the wrong cluster.



Rand index

Cluster multiple times

- Clustering A: 1, 2, 2, 1, 1
- Clustering B: 2, 1, 2, 1, 1

Compare pairs

- $a := \; and \; =$, the number of pairs assigned to the same cluster in A and in B
- $b : \neq \; and \; \neq$, … different clusters in A and in B
- $c : \neq \; and \; =$, … same in A, different in B
- $d := \; and \; \neq$, … same in B, different in A

Rand index

$$R = \frac{a + b}{a + b + c + d}$$

- Adjust the Rand index to make it vary between -1 and 1 (negative if less than expected)
- $AdjRand = (Rand - expect(Rand))/(max(Rand) - expect(Rand))$

Rand index

$$RI = (a + b)/\binom{N}{2}$$

where a is the number of pairs that belong to the same true subtype and are clustered together, b is the number of pairs that belong to different true subtypes and are not clustered together, and N is the number of possible pairs that can be formed from the N samples.

Intuitively, RI is the fraction of pairs that are grouped in the same way (either together or not) in the two partitions compared (e.g. 0.9 means 90% of pairs are grouped in the same way).

Rand index

The Adjusted Rand Index (ARI) is the corrected-for-chance version of the Rand Index. The ARI takes values from -1 to 1, with the ARI expected to be 0 for a random subtyping.

- Rand index and adjusted Rand index, https://davetang.org/muse/2017/09/21/the-rand-index/, https://davetang.org/muse/2017/09/21/adjusted-rand-index/

Gap statistics

- Cluster the observed data, varying the total number of clusters $k = 1, 2, \ldots K$
- For each cluster, calculate the sum of the pairwise distances for all points

$$D_r = \sum_{i,i' \in C_r} d_{ii'}$$

- Calculate within-cluster dispersion measures

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

Gap statistics

1. Cluster the observed data, varying the total number of clusters from $k = 1, 2, \ldots K$, giving within dispersion measures $W_k, k = 1, 2, \ldots K$.
2. Generate B reference datasets, using the uniform prescription (a) or (b) above, and cluster each one giving within dispersion measures W^*_{kb}, $b = 1, 2, \ldots B$, $k = 1, 2, \ldots K$. Compute the (estimated) Gap statistic:

$$\text{Gap}(k) = (1/B) \sum_b \log(W^*_{kb}) - \log(W_k)$$

3. Let $\bar{l} = (1/B) \sum_b \log(W^*_{kb})$, compute the standard deviation $\text{sd}_k = [(1/B) \sum_b (\log(W^*_{kb}) - \bar{l})^2]^{1/2}$, and define $s_k = \text{sd}_k \sqrt{1 + 1/B}$. Finally choose the number of clusters via

$$\hat{k} = \text{smallest } k \text{ such that } \text{Gap}(k) \geq \text{Gap}(k+1) - s_{k+1}$$

Cross-validation approaches

- Cluster while leave-out k experiments (or genes)
- Measure how well cluster groups are preserved in left out experiment(s)
- Or, measure agreement between test and training set

Clustering validity

- Hypothesis: if the clustering is valid, the linking of objects in the cluster tree should have a strong correlation with the distances between objects in the distance vector

Suppose that the original data $\{X_i\}$ have been modeled using a cluster method to produce a dendrogram $\{T_i\}$; that is, a simplified model in which data that are "close" have been grouped into a hierarchical tree. Define the following distance measures.

- $x(i, j) = | X_i - X_j |$, the ordinary Euclidean distance between the ith and jth observations.
- $t(i, j)$ = the dendrogrammatic distance between the model points T_i and T_j. This distance is the height of the node at which these two points are first joined together.

Then, letting $\bar{x}$ be the average of the $x(i, j)$, and letting $\bar{t}$ be the average of the $t(i, j)$, the cophenetic correlation coefficient c is given by[4]

$$c = \frac{\sum_{i<j}(x(i,j) - \bar{x})(t(i,j) - \bar{t})}{\sqrt{[\sum_{i<j}(x(i,j) - \bar{x})^2][\sum_{i<j}(t(i,j) - \bar{t})^2]}}.$$

WADP - robustness of clustering

- If the input data deviate slightly from their current value, will we get the same clustering?
- Important in Microarray expression data analysis because of constant noise

Bittner M. et.al. "**Molecular classification of cutaneous malignant melanoma by gene expression profiling**" Nature 2000
http://www.nature.com/nature/journal/v406/n6795/full/406536A0.html

WADP - robustness of clustering

- Perturb each original gene expression profile by $N(0, 0.01)$
- Re-normalize the data, cluster
- Cluster-specific discrepancy rate: D/M. That is, for the M pairs of genes in an original cluster, count the number of gene pairs, D, that do not remain together in the clustering of the perturbed data, and take their ratio.
- The overall discrepancy ratio is the weighted average of the cluster-specific discrepancy rates.

WADP - robustness of clustering

- If there were originally m_j genes in the cluster j, then there are $M_j = m_j(m_j - 1)/2$ pairs of genes
- In the new clustering, identify how many of these paris (D_j) still remain in the cluster
- Calculate D_j/M_j

$$WADP = \frac{\sum_{j=1}^{k} m_j D_j / M_j}{\sum_{j=1}^{k} m_j}$$

Summary

Clustering pitfalls

- Any data – even noise – can be clustered
- It is quite possible for there to be several different classifications of the same set of objects.
- It should be clear that any clustering produced should be related to the features in which the investigator in interested.