Consensus Clustering

Pedro Feijao[1] **and Sean La**[2]

1 SFU `pfeijao@sfu.ca`
2 SFU `laseanl@sfu.ca`

Abstract

Consensus clustering is a special case of Correlation clustering. Here, we show how we can use consensus clustering to combine clusterings from SNP, MLST and CNV data, dealing with the fact that the input clusterings might be at different granularities or obtained with different thresholds.

1 Background

1.1 Correlation clustering

Given a distance matrix D of the input elements, d_{ij} representing the distance between element i and j, we define $s_{ij} = T - d_{ij}$, where T is a *distance threshold*, intuitively meaning that if $s_{ij} > 0$, i and j are close and should possibly be in the same cluster, while $s_{ij} < 0$ means i and j should be separate.

The *minimum correlation clustering problem* aims to find a clustering that minimizes the sum of all positive s_{ij} for i, j in different clusters (penalty for separating good pairs) minus the sum of all negative s_{ij} if i, j are in the same cluster (penalty for joining bad pairs).

Defining binary variables x_{ij} such that $x_{ij} = 0$ if i and j are in the same cluster and $x_{ij} = 1$ otherwise, the minimum correlation clustering objective function can be written as

$$f(x) = \sum_{s_{ij}>0} s_{ij}x_{ij} - \sum_{s_{ij}<0} s_{ij}(1 - x_{ij}) = \sum s_{ij}x_{ij} - \sum_{s_{ij}<0} s_{ij}$$

and it is possible to find an optimal clustering with the following integer linear program:

$$\begin{aligned} &\underset{x}{\text{minimize}} && \sum s_{ij}x_{ij} && (1)\\ &\text{s.t.} && x_{ik} \leq x_{ij} + x_{jk} \quad \text{for all } i, j, k \\ &&& x_{ij} \in \{0, 1\} \quad \text{for all } i, j \end{aligned}$$

1.2 Consensus Clustering

Given a set of clusterings and a measure of distance between clusterings, the *consensus clustering problem* aims to find a clustering minimizing the total distance to all input clusterings. A simple distance between two clusterings π_1 and π_2 is the number of elements clustered differently in π_1 and π_2, that is, the number of pairs of elements co-clustered in π_1 but not co-clustered in π_2, plus the number of elements co-clustered in π_2 but not co-clustered in π_1.

If a clustering π is represented by a set with all the pairs that are co-clustered, this distance can be defined as the symmetric different between two clustering sets:

$$d(\pi_1, \pi_2) = |\pi_1 - \pi_2| + |\pi_2 - \pi_1| \qquad (2)$$

The distance can also be defined as an objective function to be minimized, using the binary variables defined above ($x_{ij} = 0$ means i, j are co-clustered, otherwise they are not), as follows:

$$d(x, \pi) = \sum_{\pi_{ij}=1} (1 - x_{ij}) + \sum_{\pi_{ij}=0} x_{ij} \tag{3}$$

or, more generally, with weights w_{ij} between each pair of elements i and j:

$$d(x, \pi) = \sum_{\pi_{ij}=1} w_{ij}(1 - x_{ij}) + \sum_{\pi_{ij}=0} w_{ij} x_{ij} \tag{4}$$

and this can be written as

$$d(x, \pi) = \sum s_{ij} x_{ij} + \sum_{x_{ij}=1} w_{ij} \tag{5}$$

where $s_{ij} = (-1)^{\pi_{ij}} w_{ij}$. Notice the connection with the minimum correlation clustering problem. Therefore, solving the minimum consensus problem for a given set of clusterings $\pi^{(1)}, \ldots, \pi^{(n)}$ is equivalent to solving a minimum correlation clustering problem with the matrix S defined as

$$s_{ij} = \sum_{\{k | \pi_{ij}^{(k)}=0\}} w_{ij}^{(k)} - \sum_{\{k | \pi_{ij}^{(k)}=1\}} w_{ij}^{(k)} = \sum_{k=1}^{n} (-1)^{\pi_{ij}^{(k)}} w_{ij}^{(k)} \tag{6}$$

1.3 Consensus clustering with different granularities

In this setting, we assume that the input clusterings are on different granularities. To avoid penalizing the differences between a finer clustering π_1 and a coarser clustering π_2, we introduce the following non symmetric distance:

$$d(\pi_1, \pi_2) = |\pi_1 - \pi_2| \tag{7}$$

and Eq (3) can be updated to

$$d(x, \pi) = \sum_{\pi_{ij}=0} x_{ij} \tag{8}$$

that is, removing the penalty for pairs that are co-clustered in π but not in x.

Then, given the clusterings $\pi_1, \ldots, \pi_n$ and a subset F of these clusterings, representing the clusterings with finest resolution, the *finest consensus clustering* problem is to find a clustering x that minimizes the total distance between x and all input clusterings, where

$$d(x, \pi) = \begin{cases} \sum_{\pi_{ij}=1} w_{ij}(1 - x_{ij}) + \sum_{\pi_{ij}=0} w_{ij} x_{ij}, & \text{if } \pi \in F \\ \sum_{\pi_{ij}=0} w_{ij} x_{ij}, & \text{otherwise} \end{cases} \tag{9}$$

Then, to solve this problem using the minimum correlation clustering again, we can define the matrix S as

$$s_{ij} = \sum_{\{k | \pi_{ij}^{(k)}=0\}} w_{ij}^{(k)} - \sum_{\{k | \pi_{ij}^{(k)}=1, \pi^{(k)} \in F\}} w_{ij}^{(k)} \tag{10}$$

1.4 Selecting appropriate weights for the consensus clustering problem

There might be many meaningful ways of defining the weights $w_{ij}^{(k)}$ used on the previous equations. If we assume that a clustering π was inferred based on a distance matrix D, normalized such that $0 \leq d_{ij} \leq 1$, we can define w_{ij} as

$$w_{ij} = \begin{cases} d_{ij}, & \text{if } \pi_{ij} = 1 \\ 1 - d_{ij}, & \text{otherwise} \end{cases} \tag{11}$$

The reasoning behind this definition is that if $\pi_{ij} = 1$ (i, j are not co-clustered in π), then the distance d_{ij} should be large, therefore it is a good penalty for co-clustering i, j in x. On the other hand, if $\pi_{ij} = 0$, d_{ij} is small, which means that $1 - d_{ij}$ is a better candidate for the penalty of choosing $x_{ij} = 1$. Therefore, the distance between two clusterings, given on Eq. (4), can be updated to

$$d(x, \pi) = \sum_{\pi_{ij}=1} d_{ij}(1 - x_{ij}) + \sum_{\pi_{ij}=0} (1 - d_{ij})x_{ij} \tag{12}$$

and Eq. (6) is now

$$s_{ij} = \sum_{\{k|\pi_{ij}^{(k)}=0\}} \left(1 - d_{ij}^{(k)}\right) - \sum_{\{k|\pi_{ij}^{(k)}=1\}} d_{ij}^{(k)} = \Pi_{ij} - D_{ij} \tag{13}$$

where $\Pi_{ij} = |\{k|\pi_{ij}^{(k)} = 0\}|$ and $D = \sum_{k=1}^{n} d_{ij}^{(k)}$.

1.5 Weighted consensus clustering with different granularities

In the situation where the clusterings in consideration have different granularities, we may set the weights to that of (11) so that (9) becomes

$$d(x, \pi) = \begin{cases} \sum_{\pi_{ij}=1} d_{ij}(1 - x_{ij}) + \sum_{\pi_{ij}=0}(1 - d_{ij})x_{ij}, & \text{if } \pi \in F \\ \sum_{\pi_{ij}=0}(1 - d_{ij})x_{ij}, & \text{otherwise} \end{cases} \tag{14}$$

and (15) is now given by

$$s_{ij} = \sum_{\{k|\pi_{ij}^{(k)}=0\}} \left(1 - d_{ij}^{(k)}\right) - \sum_{\{k|\pi_{ij}^{(k)}=1, \pi^{(k)} \in F\}} d_{ij}^{(k)} = \Pi_{ij} - D'_{ij} \tag{15}$$

where

$$D'_{ij} = \sum_{\{k|\pi_{ij}^{(k)}=0\}} d_{ij}^{(k)} + \sum_{\{k|\pi_{ij}^{(k)}\pi_{ij}^{(k)}=1,\ \pi^{(k)} \in F\}} d_{ij}^{(k)}$$

and $\Pi_{ij} = |\{k \mid \pi_{ij}^{(k)} = 0\}|$, as before.