BCB420Y Project Documentation

Xiao Wang
998816456

April 9, 2015

Input Data

The input format of this section is a straight forward list of gene pairs and association scores from each track. It follows the following format:

	Gene ID 1	Gene ID 2	iRef	GEO	GO	pathW
1	1	2	0.15	0.98	0.12	0.32

Note that the gene ids may differ depending on the convention used.

Algorithm

After the input has been read into memory, the graphs must be preprocessed so that Markov Clustering algorithm can be applied to the input data set.

Normalization

The first step in the prepossessing procedure is to normalize the given data. Since this is currently the initial stage of this study, the distribution of data for each track of analysis is unknown. In order to account for this uncertainty, the data in this section is normalizing using the empirical cumulative distribution function, defined as:

$$F_n(x) = \frac{1}{n}(\#x_i \leq x) \tag{1}$$

Due to the rank based nature of this normalization procedure, any outliers contained in the original data will not pose a significant impact. However, one potential problem with this approach is that the lower bound of the normalized data is not fixed, since one always has to count the data point itself in the calculation of ecdf. Further strategies are needed in order to have this problem resolved. As it only influences one entry for every track, the impact of this issue should not greatly influence the result of this analysis.

Fusion

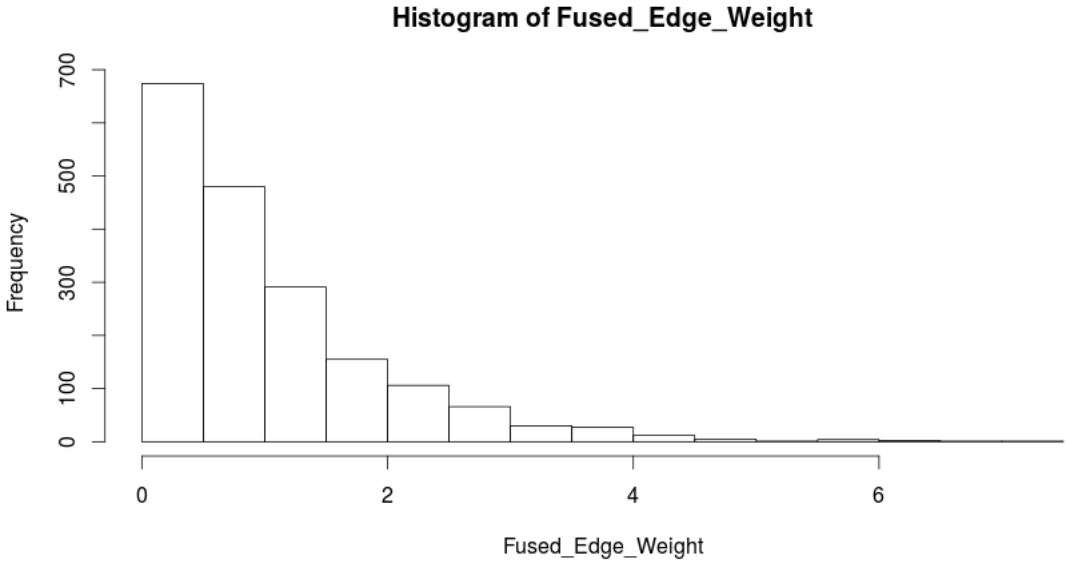
Since Markov Clustering Algorithm cannot be performed on multigraphs, the association scores from each track must be compiled into one summary statistic. In order to correctly emphasize the association between each gene pair, the summary statistic must obey the following properties:

1. In a given pair, if the score of one of the tracks is high, then the statistic must have a high value.
2. In a given pair, if the score of all of the tracks are high, the the statistic must have a high value.
3. If all values in a given column are approximately the same, then the summary statistic should correlate with the arithmetic mean of the scores

To accommodate the above factors, the following summary measure was proposed and the rationale explained below:

$$S = \frac{4}{\sum_{i=1}^{4} \frac{1}{\log(F_{n,i}(x_i))}} \tag{2}$$

In here $F_{n,i}(x_i)$ is simply the normalized score for ith track. This measure computes the harmonic mean of the log of the entries.
The reason a logarithm of each entry is used is due to the fact that the logarithm function is asymptotic close to zero. This property makes this function ideal for down-scaling values that are small, while its monotonicity keeps the high values high. Since each entry smaller or equal to 1 as a result of previous normalization, the scores may take full advantage of this construct.
A histogram of fused edge weights is provided below.



MCL Parameter Optimization

After the graphs have been fused, the final graph will be provided as input to the Markov Clustering Algorithm. The Markov Clustering Algorithm takes two parameters, the expansion parameter and the inflation parameter (For detailed introduction about the parameters please refer to the Markov Clustering Algorithm paper). Although no rigorous analysis was done in this study, it must be noted that if the parameters of MCL were not set properly, the algorithm will simply place all genes in one cluster. After some experimentation, the optimal parameter values was determined to be expansion: 2.5 and inflation 1.75. These values were roughly established since they introduce the greatest spread while keeping the number of clusters reasonable. It is important to note that these values may differ as input data changes. Further analysis is needed in this area.

Output Format

The out put format is rather straight forward. It consist of a list of gene ids and corresponds to each a number indicating which cluster it is in. An example below:

	gen-id	Cluster
1	1	35