ConsensusClustering Documentation

Description: Resampling-based clustering
Author: Stefano Monti (Broad Institute) gp-help@broadinstitute.org

Summary

Given a set of items to be clustered (items can be either genes or chips/experiments), Consensus clustering provides for a method to represent the consensus across multiple runs of a clustering algorithm and to assess the stability of the discovered clusters. To this end, perturbations of the original data are simulated by resampling techniques. The clustering algorithm of choice is applied to each of the perturbed data sets, and the agreement, or *consensus*, among the multiple runs is assessed and summarized in a *consensus matrix.* Each matrix entry is indexed by an item pair that measures the proportion of times the pair's items are clustered together across the resampling iterations (ideally, always, or never). A distinct consensus matrix is generated for each of the number of clusters considered (e.g., if kmax=5, consensus matrices corresponding to 2, 3, 4, and 5 clusters will be generated). Visual inspection of the consensus matrices, and of the corresponding summary statistics can be used to determine the best number of clusters (see reference for more details).

References

- S. Monti, et al. "Consensus Clustering: A resampling-based method for class discovery and visualization of gene expression microarray data", *Machine Learning Journal*, 52(1-2):91-118, 2003.

Parameters:

Name	**Description**
input filename	The data to be clustered (.gct, .res, .odf)
kmax	Try K=2,3,...,kmax clusters (must be > 1)
resampling iterations	Number of resampling iterations
seed value	Random number generator seed
clustering algorithm	Type of clustering algorithm
cluster by	Whether to cluster by rows/genes or columns/experiments
distance measure	Distance measure
resample	resampling scheme (one of 'subsample[ratio]', 'features[nfeat]', 'nosampling') For example to specify The default is subsampling with a proportion of 0.80 (80%). To specify a subsample, simply type 'subsampleN', where N is a number strictly between 0 and 1 (without square brackets, which indicate an optional argument). For example, to specify a 0.90 proportion, type 'subsample0.9'. To use features, simply type featuresN, where N is a number less than the number of features in your dataset.
merge type	Ignored when algorithm other than hierarchical selected
descent iterations	Number of SOM/NMF iterations
output stub	Stub pre-pended to all the output file names

normalize type	row-wise, column-wise, both
normalization iterations	number of row/column normalization iterations (supercedes normalize.type)
create heat map	Whether to create heatmaps (one for each cluster number)
heat map size	point size of a consensus matrix's heat map (between 1 and 20)

Output Files:

1. <output.stub>.<sampleid>.<k>.clu, is a text file listing the items belonging to each cluster. (<sampleid> indicates the type of resampling scheme, and <k> denotes the number of clusters).
2. <output.stub>.<sampleid>.<k>.gct is the consensus matrix for <k> clusters, with the entries sorted as in the input data.
3. <output.stub>.<sampleid>.srt.<k>.gct is the consensus matrix for <k> clusters, with the entries sorted so as to have items clustering together adjacent to each other.
4. <output.stub>.<sampleid>.srt.<k>.gif is the heat map corresponding to the sorted consensus matrix.
5. <output.stub>.<sampleid>.statistics.pdf includes a series of plots of statistics (Lorenz curve, Gini index, Consensus CDF) that can be used to determine the best number of clusters.

Platform dependencies:

Module type:	Clustering
CPU type:	any
OS:	any
Java JVM level:	1.5
Language:	Java, R 2.5

GenePattern Version Notes:

Date	Version	Description
02/19/08	5	Updated for R 2.5
08/08/12	6	Corrected default parameter values in manifest, Upgraded Jama jar file
08/28/12	7	Fixed bugs in classpath (changed Jama-1.0.1.jar to Jama-1.0.2.jar and acme.jar to Acme.jar)
11/16/12	8	Improved description of resample parameter