Gene expression

Mosclust: a software library for discovering significant structures in bio-molecular data

Giorgio Valentini

DSI, Dipartimento di Scienze dell’Informazione, Università degli Studi di Milano, Via Comelico 39, Italy

Received on September 10, 2006; revised on November 15, 2006; accepted on November 21, 2006

Advance Access publication November 24, 2006
Associate Editor: David Rocke

Summary: The R package mosclust (model order selection for clustering problems) implements algorithms based on the concept of stability for discovering significant structures in bio-molecular data. The software library provides stability indices obtained through different data perturbations methods (resampling, random projections, noise injection), as well as statistical tests to assess the significance of multi-level structures singled out from the data.

Availability: http://homes.dsi.unimi.it/~valenti/SW/mosclust/download/mosclust_1.0.tar.gz
Contact: valenti@dsi.unimi.it
Supplementary information: http://homes.dsi.unimi.it/~valenti/SW/mosclust

The identification and validation of structures underlying complex bio-molecular data is a fundamental step in post-genomic data analysis. Different clustering validation techniques (see Handl et al., 2005 for a recent review), and software tools implementing classical validity indices (such as the Dunn’s index and the Silhouette index) have been recently proposed (Bolshakova et al., 2005).

The mosclust software library implements methods based on the concept of stability to estimate the ‘natural’ number of clusters, or to select a clustering solution from a set generated by a range of different algorithms. Moreover the library may be used to detect multiple structures simultaneously present in complex bio-molecular data, such as sub-clusters inside higher level clusters or multi-level hierarchical structures discovered in gene expression data. In this conceptual framework multiple clusterings are obtained by introducing perturbations into the original data, and a clustering is considered reliable if it is ‘stable’, that is if it is approximately maintained across multiple perturbations. The main logical steps of a high-level algorithm for a stability-based evaluation of clustering solutions may be summarized as follows:

1. Perform multiple random perturbations of the data.
2. Supply the perturbed data to a given clustering algorithm.
3. Apply a given clustering similarity measure to multiple pairs of k-clusterings obtained according to steps (1) and (2).
4. Use appropriate stability indices, based on the similarity measures applied at step (3), to score the reliability of a given k-clustering.
5. Repeat steps (1) to (4) for different numbers k of clusters and elect the most stable clustering(s) as the most reliable.

Step (1) may be implemented through resampling techniques (Monti et al., 2003) or random noise injection into the data (McShane et al., 2002), or random projections into lower dimensional subspaces (Valentini, 2006). To estimate the similarity between clusterings (step 3), classical measures, such as the Jaccard or the Fowlkes and Mallows coefficients may be applied (Jain and Dubes, 1988). For step (4), several stability indices for model order selection have been proposed in the literature: very schematically they can be divided into indices that use statistics of the similarity measures (McShane et al., 2002; Valentini, 2006) or their overall empirical distribution (Ben-Hur et al., 2002). The last step is usually approached by choosing the best scored clustering (according to the chosen stability index), but a major problem is represented by the estimate of the statistical significance of the discovered solutions.

The proposed library provides similarity measures, stability indices and different random perturbations procedures to assess the reliability of clustering solutions. Moreover statistical tests to assess the significance of the solutions, and to discover multiple structures simultaneously present in the data are also implemented.

More precisely, let $S_k$ ($0 \leq S_k \leq 1$) be a random variable given by a similarity measure (e.g. Fowlkes and Mallows coefficient) between two k-clusterings obtained from pairs of randomly perturbed data. Using its cumulative distribution function $F_k(s)$, we can obtain the following integral to estimate stability of a given k-clustering:

$$g(k) = \int_0^1 F_k(s) ds$$

Indeed if $g(k)$ is close to 0, all the similarity values are concentrated near 1, and hence the corresponding clustering is stable with respect to a given perturbation, while larger values of $g(k)$ indicate less stability and hence less reliability of the corresponding clustering (see Bertoni and Valentini (2006) and Supplementary information for more details). By estimating Equation (1) for different values of $k$ and sorting the corresponding values we can rank the k-clusterings from the most to the least reliable one. Moreover, if we consider two groups of pairwise clustering similarities values separated by a threshold $\rho$, by defining a binomial random variable $X_h$ that counts how many times $S_k > \rho$, and its associated probability of success $\theta_a$, for a sufficiently large number $n$ of evaluations of $S_k$, we have that the following random variables have respectively the normal and
\( \chi^2 \) distribution:

\[
\frac{X_k - n\theta_k}{\sqrt{n\theta_k(1 - \theta_k)}} \sim N(0, 1) \quad \sum_{k \in K} \frac{(X_k - n\theta_k)^2}{n\theta_k(1 - \theta_k)} \sim \chi^2
\]

From Equation (2) we can obtain a \( \chi^2 \)-based test to find significant clustering solutions (Bertoni and Valentini, 2006). Relaxing the assumption that some probability distributions are normal, we implemented also another statistical test based on the classical Bernstein inequality. In all cases, using these statistical tests, we are able to detect statistically significant structures, comprising also multiple structures (e.g. hierarchical structures) simultaneously present in bio-molecular data.

The R package mosclust implements stability methods for unsupervised structure discovery in bio-molecular data through a set of functionalities that may be summarized as follows:

1. Data perturbation through random projections, bootstrap resampling techniques or noise injection.
2. Functions to compute similarity measures between pairs of perturbed clusterings.
3. Functions to compute stability indices for different number of clusters.
4. Integrated functions to compute stability indices with widely used clustering algorithms (k-means, hierarchical clustering, Prediction Around Medoids, fuzzy-c-means).
5. Functions to perform \( \chi^2 \)-based and Bernstein inequality-based tests of hypothesis to select k-clustering solutions significant at a given significance level (as well as to estimate the associated P-value).
6. Graphical functions to plot histograms and empirical cumulative distribution functions of the similarity measures for different number of clusters, and to plot the p-values for different tests of hypothesis.

As an applicative example of the mosclust R package, we present a stability analysis of the DNA microarray clusters obtained from the well known data set composed by 72 leukemia samples (Golub et al., 1999). With the classical k-means algorithm, using the \( \chi^2 \)-based statistical test, two and three clusters are predicted at \( \alpha = 0.01 \) significance level. The same results are obtained with both subsampling techniques and random projections to lower dimensional subspaces, while using noise injection techniques only two clusters are detected. Figure 1 shows the empirical cumulative distribution functions (ecdfs) of the similarity measures obtained with subsampling and random projections techniques using different numbers of clusters. Note that the ecdfs for two and three clusters are well separated and ‘below’ the others, showing that the corresponding clusterings are more stable and reliable. Even if the two-clustering is more stable than the three-clustering [because \( g(2) < g(3) \), Equation (1)], their ecdf graphs are quite close, showing that that there is not a well-defined ‘winner’: indeed the statistical test does not find any significant difference about the reliability of the two clusterings (at 0.01 significance level), revealing a very likely two-level structure in the data.

Even if stability-based techniques do not suffer from biases towards particular algorithms, as sometimes is the case with traditional validation techniques based on inter or intra-clusters distances, they may nevertheless exhibit other kinds of bias under certain circumstances. Indeed, as shown by Handl et al. (2005), a given clustering may converge to a suboptimal solution owing to the shape of the data manifold and not to the real structure of the data. Even if this is an open problem for stability-based methods, the usage of different clustering algorithms and a comparison with the results obtained with different techniques (Bolshakova et al., 2005) may in part overcome these problems.

The reference manual and the Supplementary information available at the associated web-site (see Supplementary information) describe in detail the usage, the functionalities and the theoretical background underlying the proposed software library.

Conflict of Interest: none declared.

REFERENCES


