An integrated algorithmic procedure for the assessment and discovery of clusters in DNA microarray data.

Avogadri R[(1)], Valentini G[(1)], Bertoni A[(1)]

[(1)] Computer Science Department, University of Milano, Milano

Mo

Clustering methods are used in various bioinformatics fields, among them the analysis of gene expression data. One of the main problems related to this unsupervised approach is the fact the quality assessment of the results is an ill-posed problem. With regard to this, different methods to check the "validity" of the clustering solutions have been proposed; the classical methods are generally based on information-theory concepts and statistical approaches, but new techniques based on the concept of "stability" have been recently proposed. Two of the most relevant properties used to evaluate the quality of the clustering algorithms are respectively the robustness of the solutions and the the ability to detect the "natural" number of groups underlying the data. Indeed the selection of the correct number of clusters, and the assessment of the reliability of clustering solutions are a central problem in bioinformatics and in bio-medical applications. A further characteristic of data generated through high-throughput biotechnologies is represented by their high dimensionality, and several approaches have been proposed to reduce the dimension of the available data. In recent years different techniques have been developed to deal with the above mentioned issues, but only a few efforts have been dedicated for their integration in a unified general algorithmic procedure.

M

Main objective of the proposed work is to set up a a pipeline-based integrated approach to jointly consider the issues described in the previous section: data dimensional reduction, choice of a suitable number of clusters, robustness of the clustering solutions, and clustering solution validation. The first step is performed using the random projection technique. By means of this method, the dimension of a data set is reduced by using random maps; this step can be performed efficiently through randomized linear function. The most suitable projected space dimension can be evaluated according to Johnson-Lindenstrauss lemma. In this way it is guaranteed that, with a bounded error, the projected data preserve approximately their original structure, in terms of distance among the examples. The effectiveness of this approach has been tested in literature in particular on DNA-microrray data sets.
In the second step, the projected data sets are used to assess the most reliable number(s) of clusters, by applying a chi-square or a Bernstein inequality-based statistical test. An implementation can be found in the mosclust R library (http://homes.dsi.unimi.it/~valenti/SW/mosclust/). The ensemble aggregation of multiple clustering is performed by using the number(s) of clusters estimated at the previous step. In particular the consensus matrix is obtained by an averaged sum of the pairwise similarity matrix associated to each base clustering, and the consensus clustering is achieved by applying a suitable clustering algorithm to the rows of the consensus matrix. The last validation step applies a stability index. It is based on multiple perturbations of the available data. It is implemented in the clusterv R library (http://homes.dsi.unimi.it/~valenti/SW/clusterv/), and it allows to evaluate the reliability of each cluster discovered by the nsemble clustering algorithm, as well as the reliability of the overall consensus clustering.

Re sult

Preliminary results with DNA microarray data, show that the proposed integrated method can permit to estimate the unknown number of clusters in the data, to discover biologically meaningful clusters, and to assess the reliability of each cluster discovered by the ensemble clustering algorithm. In our experiments we used for both the base and the consensus clustering a classical hierarchical algorithm, but other clustering methods can be in principle applied.

C **:** avogadri@dsi.unimi.it