A tool for interactive analysis of bioinformatics data

ID - 103

Tagliaferri Roberto[1], Cocozza Sergio[2], Iorio Francesco[1,3], Miele Gennaro[4], Napolitano Francesco[1], Pinelli Michele[2], Raiconi Giancarlo[1]

[1]DMI, University of Salerno, I-84084, via Ponte don Melillo, Fisciano (SA)

[2]CDGU, Department of Cellular and Molecular Biology and Pathology "L. Califano" University of Naples "Federico II", Napoli

[3]Telethon Institute of Genetics and Medicine, Via P. Castellino 111, 80131, Napoli

[4]Dipartimento Scienze Fisiche, University of Naples "Federico II", I-80136, via Cintia 6, Napoli

Motivation

We propose a scientific data exploration software environment that permits to obtain both data clustering and visualization. The approach is based on a pipeline, from input to cluster analysis and assessment with each stage supported by visualization and interaction tools. The proposed approach enables the user to answer the following questions: Is a dataset really clusterizable? does it contain localized uniform groups and are these groups well separable? How many cluster should we produce? How much is the clusterization reliable? If we reassign some point to a different cluster, how much does the total reliability change? In which clusters do the points of this subset lie? How can we use a priori partial information to validate clusterization?

Methods

1. Parameter estimation The first step of the pipeline implemented by our tools consists of a procedure for testing the stability of clusters for any value on a prescribed range of parameters of the algorithm and is based on the Model Explorer approach. 2. Clustering and visualization The same procedure can be used to obtain a first raw clusterization of the dataset.

In our tool K-Means, EM, SOM, PPS algorithms are provided, which can be used both directly as clustering methods and in the parameter estimation process. The clusterization can be inspected by a module offering the visualization of clusters convex hulls as they appear in projected 2-3D spaces through PCA and MDS. The user can select each cluster and each point in a cluster checking all its features. 3. Cluster reliability The reliability of each cluster in a clusterization is obtained by first producing several perturbed versions of the dataset, and then a clusterization over each of these new datasets and their similarity matrices. The Fuzzy Similarity Matrix over these clusterizations ia build summing over the similarity matrices. 4. Fuzzy Membership analysis Starting from the fuzzy similarity matrix, we can compute the value of each cluster membership function for each pattern and quantify how much single points belong to a fixed group of other points over different trials of clustering on the perturbed versions of the dataset. The fuzzy membership is represented through a 2-3D visualization of the points of the dataset. The convex hulls of the clusters composed by the points for which the membership values are greater than a given threshold are shown. The user is allowed to change the threshold and save the corresponding clusters. Basing on the values of every membership function for each outlier, the system suggests to manually reassign a point to a different cluster. The user can accept such suggestions or explore alternative clusterizations.

5. Interactive Agglomerative Clustering Since rarely the number of clusters is known in advance, we adopt an interactive agglomerative technique on the clusters obtained in the previous phase. The user can choose different partitioning of the data on the base of a threshold value. The result is a dendrogram in which each leaf is associated with a cluster from the previous phase. This step could employ the Negentropy distance or any hierarchical clustering.

6. Visualization of prior knowledge A priori information can be used to: - Validate a clustering result.
- Infer new knowledge.

The validation of a clustering can be obtained comparing the prior knowledge and the knowledge obtained from the clustering itself, obtaining a confidence degree. The prior knowledge can also be used to produce new knowledge inferred by the presence or absence of objects of a certain class in a certain cluster. Different hypotheses can be made depending on the relations between the prior knowledge and the features used to cluster the objects. Prior knowledge can be visually shown and permits to visualize distance and cluster information together with the prior knowledge.

Results

Comparing the results of our cluster analysis with those of hierarchical clustering by Whitfield et al., it appears that the genes classified into the third NEC cluster are distributed into three different Whitfield clusters. Two of these clusters were reported as associated with DNA replication function, whereas the third were composed by various, non-classified genes. It seems that the NEC cluster is more enriched in genes

belonging the same functional class (DNA replication = 9.7%; Fisher exact test p = 3.3*10^-11) than two of the three different Whitfield clusters (Whitfield 1, DNA replication = 9.5%; Fisher exact test p=4.4*10^-9 - Whitfield 2, DNA replication = 5.3 %, p = 7.5*10^-3). Concerning the third Whitfield cluster (composed by non-classified genes) the hierarchical clustering appears to misclassify this set of 77 genes. The analysis of their function showed that they are highly associated with the DNA replication function (DNA replication = 11.6%, Fisher exact test p = 8.4*10^-6). In conclusion, at least in several cases, the proposed method of clustering appears to be more useful.

Email: rtagliaferri@unisa.it