A Review on clustering and visualization methodologies for Genomic data analysis

Roberto Tagliaferri[1], Alberto Bertoni[2], Francesco Iorio[1,3], Gennaro Miele[4], Francesco Napolitano[1], Giancarlo Raiconi[1] and Giorgio Valentini[2]

[1] DMI, Università degli Studi di Salerno - Fisciano (Sa), Italy
[2] DSI, Università degli Studi di Milano - Milano, Italy
[3] Telethon Institute of Genetics and Medicine - Napoli, Italy
[4] Dipartimento di Scienze Fisiche, Università degli Studi di Napoli "Federico II" - Napoli, Italy

This abstract presents a survey on the aims, the problems and the methods concerning Cluster Analysis and its applications in genomic data analysis. With the term *Cluster Analysis* we refer to a data exploration tool whose goal is grouping objects of similar kind into their respective categories without a priori information on their classes. We can look at cluster analysis as a classification problem with no labeled samples, or without any a priori knowledge about the way the objects have to be put together. There are several and heterogeneous problems linked to the cluster analysis and several times they are treated separately. In this work we examine these problems, and we illustrate the different approaches and their applications to Computational Biology and Bioinformatics. The problems related to Cluster Analysis in the context of high-dimensional genomic data analysis can be summarized as shown in figure 1 . In this figure each node represents an item of the data exploration problem via cluster analysis or computational methods used in this kind of data analysis. The edges of this graph can be mono-directional or bi-directional and, following a path (according the edge directions), one can see a sequence of steps toward the final goal of cluster analysis and the relationships between different problems and computational methods involved in unsupervised genomic data analysis.

As we can see in the figure, the problems tackled with cluster analysis are particular cases of a more general class of problems: the partitioning problems. In this class of problems, given a set of objects N and a set of K functions $f = (f_1, \ldots, f_K)$ from the set N to the real numbers, the aim is to find a partition $A = (A_1, \ldots, A_K)$ of the set N that minimizes or maximizes an objective function $g(f_1(A_1), \ldots, f_K(A_K))$. In the case of cluster analysis the function defined on the subsets A_i of N is the same for every i and usually it is the sum of the pairwise similarity between the elements of A_i (intra-cluster similarity) or the ratio between intra-cluster similarity and this sum of the similarity between clusters (inter-cluster similarity).

The similarity can be seen as the inverse of the distance between objects. The function g is usually a sum and it should be minimized (if we use distances

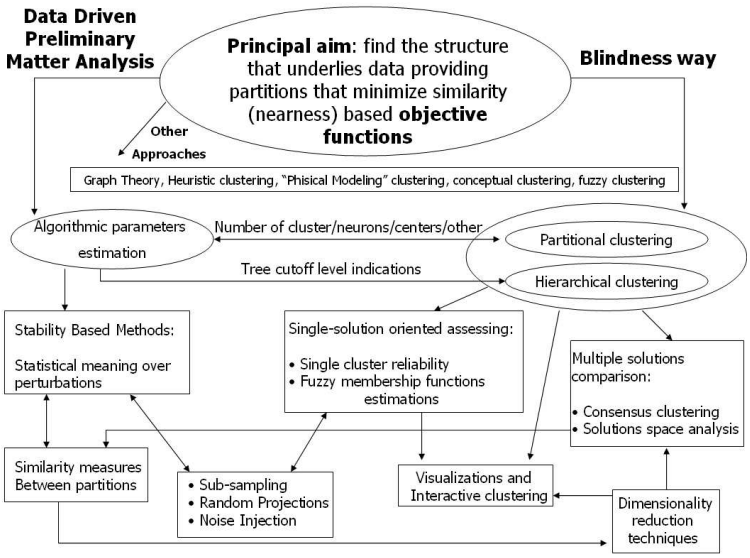

Fig. 1.

between objects) or maximized (if we use similarities). In other words, the main aim of this class of problems is the production of a partition in which data points belonging to the same subset (cluster) are as similar as possible. So, the first observation we can make is that the ability to quantify a similarity (or distance) between two objects is fundamental to clustering algorithms. Hence we need to choose a similarity measure that allows the set of objects to be embedded in a metric space. Usually classical distance metrics (e.g. Euclidean, correlation, cosine, etc.) may be applied, as well as more specific distance or similarity measures [13]. Most of the clustering approaches can be divided in two major classes: *hierarchical clustering algorithms* and *partitional clustering algorithms* [13], even if other approaches such as graph-based methods, or probabilistic and mixture models based clusterings have been recently applied to genomic cluster analysis [14, 17, 15].

Hierarchical clustering algorithms build (*agglomerative algorithms*), or breaks up (*divisive algorithms*), a hierarchy of clusters. The traditional representation of this hierarchy is a tree (called a dendrogram), with individual elements as leaves and a single cluster containing every element as root. Agglomerative algorithms begin from the leaves of the tree, whereas divisive algorithms begin at the root.

Partitional clustering algorithms attempt to directly decompose the data set into a set of disjoint clusters. In this class of algorithms the value of the input parameters (like i.e. the value of K in K-means or the map dimension in Self Organizing Map approach) plays a key role and in many cases it determines the

final number of clusters. In hierarchical clustering algorithms the same role is played by the choice of the dendrogram cutting threshold.

A major problem related to cluster analysis is the proper choice of the number of clusters. To this end we may perform a preliminary statistical analysis on the set we want to cluster instead of blindly make clustering on it. Moreover, this first analysis can check the "effective clusterizability" of a set, in other words, it checks the presence of well localized and well separable homogeneous (by the similarity point of view) object groups in the set. Several approaches have been proposed in the literature: for a recent review in the context of genomic and post-genomic data analysis see, e.g. [12]. In this work we focus on methods based on the concept of stability, as recently several works showed their effectiveness in the analysis of genomic data [1][2][3][4][10].

In these methods many clusterings are obtained by introducing perturbations into the original set, and the candidate clustering is considered reliable if its structure is approximately reflected by the clusterings obtained on the perturbed instances of the data. Informally, the stability of a given clustering is a measure that quantifies the change the clustering is affected by, after a perturbation of the original data set. Different procedures have been introduced to randomly perturb the data, ranging from bootstrapping techniques [1], to noise injection into the data [16] or random projections into lower dimensional subspaces [3].

The underlying idea of stability-based methods is described in the following. Cluster analysis is based on the (almost philosophical) assumption that the phenomenon that generated the data we want to analyze can be modeled by a statistical point of view. Can we say how much the statistical model underlying data has been discovered by a clustering? Can we say how much an assumption on the artificial model (the clustering) is close to the real model? If we know the underlying data statistical model then answers to these questions are obtained in a quite simple way: we generate different data set samples from the same statistical distribution and then we cluster each sample. If the clusterings obtained are similar then we can look to each of them as a slightly modified version of a general stable clustering in which the statistical model underlying data is well detected. So this approach is useful to test the correctness of some assumptions on the artificial model (for our purpose, the number of clusters, input values for parametric clustering algorithms and so on). The "different samples" of the data set can be obtained simulating the underlying statistical model via different perturbation techniques.

We need similarity measures between clusterings to test how two different clusterings are similar, and several classical measures can be applied [13][1]. Recently a novel measure to test similarity between partition on the same set has been introduced [5][6]. It is based on the entropy of the confusion matrix between the partitions and its parametric version quantifies also how much two partitions are in conflict each other.

Usually the objective functions that clustering algorithms tries to minimize has multiple local minima. It means that multiple and in some cases very different solutions grant very close optimal values for the objective function. This sug-

gests to analyze the whole solutions space before choose the optimal clustering. In some novel approaches multiple solutions are compared, by the objective function value point of view and by the clusters composition point of view both, and embedded in a viewable map.
Another important path in the graph of the figure 1 crosses the *single solution assessment* node. In this case the main aim is providing reliability scores for each cluster of a clustering and for the membership of each data point to each cluster. In order to obtain these results, tools based on random projection techniques [7][8] have been modeled. Combining these tools with very simple fuzzy logic derived concepts [9], membership functions are provided for each data point and interactive clustering is realized. These tools allow the user to consider only the sub set of points belonging to clusters (or sub-cluster) whose reliability is greater than a fixed threshold value. Manually reassignments for a point whose membership function distribution has a very high entropy are also possible.
In an effective usable environment all the tools implementing these models should be equipped with procedures that allow the user to easily visualize and manipulate data and to this end dimensionality reduction techniques need to be applied. The integration of different tools that explicitly consider the problems of cluster validity assessment, clustering reliability and robustness, discovery of multiple structures underlying the data, as well as data and clustering results visualization, are of paramount importance in bioinformatics and bio-medical applications [11][18].

In the full version of this paper each of the general arguments introduced here will be discussed in detail, and relevant literature about them will be provided.

References

1. A. Ben-Hur, A. Elisseeff, I. Guyon. *A stability based method for discovering structure in clustered data.* Pacific Symposium on Biocomputing, 2002.
2. Shai Ben-David, Ulrike von Luxburg, David Pal. *A Sober Look at Clustering Stability.* David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada
3. A. Bertoni and G. Valentini. *Discovering structures through the Bernstein inequality.* In KES-WIRN 2007, Vietri sul Mare, Italy, 2007
4. M. Smolkin and D. Gosh. *Cluster stability scores for microarray data in cancer studies.* BMC Bioinformatics, 36(4), 2003
5. R.J.J.H. van Son. *A method to quantify the error distribution in confusion matrices.* Institute of Phonetic Sciences, University of Amsterdam, Proceedings 18, 1994
6. F. Iorio and F. Napolitano. *ITACA (Integrated tool for assessing clustering algorithm): un tool integrato per la valutazione del clustering.* Degree thesis in Computer Science. Universitá degli studi di Salerno. 2007
7. B. Stein, S. M. zu Eissen, F. Wibrock. *On Cluster Validity and the Information Need of Users.* 3rd IASTED Int. Conference on Artificial Intelligence and Applications (AIA 03), 2003
8. A. Bertoni, G. Valentini. *Random projections for assessing gene expression cluster stability.* IJCNN 2005, The IEEE-INNS International Joint Conference on Neural Networks, Montreal, 2005

9. J. C. Dunn. *Well separated clusters and fuzzy partitions.* Journal on Cybernetics, 1974
10. A. Bertoni and G. Valentini. Model order selection for bio-molecular data clustering. *BMC Bioinformatics*, 8(Suppl.3), 2007.
11. N. Bolshakova, F. Azuaje, and P. Cunningham. An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics*, 21(4):451–455, 2005.
12. J. Handl, J. Knowles, and D. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3215, 2005.
13. A.K. Jain, M.N. Murty, and P.J. Flynn. Data Clustering: a Review. *ACM Computing Surveys*, 31(3):264–323, 1999.
14. H. Kawaji, Y. Takenaka, and H. Matsuda. Graph-based clustering for finding distant relationships in a large set of protein sequences. *Bioinformatics*, 20(2):243–252, 2004.
15. X. Liu, S. Sivaganesan, K.Y. Yeung, J. Guo, R. E. Bumgarner, and M. Medvedovic. Context-specific infinite mixtures for clustering gene expression profiles across diverse microarray dataset. *Bioinformatics*, 22(14):1737–1744, 2006.
16. L.M. McShane, D. Radmacher, B. Freidlin, R. Yu, M.C. Li, and R. Simon. Method for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(11):1462–1469, 2002.
17. K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, and W.L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.
18. R. Yoshida, T. Higuchi, S. Imoto, and S. Miyano. Arraycluster: an analytic tool for clustering, data visualization and module finder on gene expression profiles. *Bioinformatics*, 22(12):1538–1539, 2006.