BITS :: Call for Abstracts 2019 - Oral communication

Type	Oral communication
Session	Algorithms in Bioinformatics
Title	Nested partitions from hierarchical clustering statistical validation
All Authors	Christian Bongiorno (1), Salvatore Micciche`(2), and Rosario N. Mantegna (2,3,4)

Affiliation

(1) Laboratoire de Mathe´matiques et Informatique pour les Syste`mes Complexes, CentraleSupe´lec, Universite´ Paris Saclay, 3 rue Joliot-Curie, 91192, Gif-sur-Yvette, France
(2) Dipartimento di Fisica e Chimica, Universita` di Palermo, Viale delle Scienze, Ed. 18, I-90128, Palermo, Italy
(3) Complexity Science Hub Vienna, Josefsta¨dter Strasse 39, 1080, Vienna, Austria
(4) Computer Science Department, University College London, 66 Gower Street, WC1E 6BT, London, UK

Motivation

Currently, the big data revolution is changing the way many disciplines perform a large amount of experimental measures and their interpretation and modeling. In fact, due to the successes of information technology revolution and the advances in robotics many scientific experiments have today an observational character rather then a design investigating a fully controlled set up. Examples are space-time records of particles dynamics, climatological monitoring of large scale regions, earthquake investigations, brain activities, gene expressions, dynamics of social and financial systems. All these types of complex systems present datasets that are genuinely multivariate and that are recorded in the presence of sources of uncertainty (modeled as noise). Their interpretation and modeling with statistically validated data mining tools require the characterization of the hierarchical sub-units present in them.
A traditional unsupervised tool for the characterization of sub-units of a complex system is hierarchical clustering. In spite of the effectiveness and simplicity of this approach the extraction of a hierarchically nested partition from a hierarchical tree is still today an open problem. The most widely used approach for cluster detection used in the scientific literature is an approach originally proposed in phylogenetics and today implemented by the algorithm called Pvclust. This algorithm is widely used in many disciplines and especially in genomics. It is the standard reference in the literature but present two serious limits. The first limit concerns computational time and scalability with system size. The algorithm is relatively slow and has a limited scalability and therefore it is not appropriate for very large datasets. The second limit (partly overlapping with the previous one) is related to the open problem of how to deal with the so-called familywise error. This type of error is a source of statistical errors occurring when a large number of statistical tests is performed in parallel in a system. This type of errors originates naturally in very large datasets.

Methods

In this work, we propose a greedy algorithm based on bootstrap resampling that associates a p-value at each clade of a hierarchical tree. Our algorithm gives good results when applied to benchmarks mimicking the complexity of hierarchically nested complex systems. We call our algorithm statistically validated hierarchical clustering (SVHC). Specifically, for each pair of parent and children nodes in the hierarchical tree, we test the difference between the proximity measure (in our approach a dissimilarity) associated with a clade h and the dissimilarity measure associated with the clade defined by its parent node in the genealogy of the dendrogram. The statistical test we perform consider as a null hypothesis that the dissimilarity of the parent node is larger than the dissimilarity of the children node. Our tests are performed by considering multiple hypothesis test correction. In fact, we always apply the control of false discovery rate. By selecting those clades that reject our null hypothesis, we identify a hierarchically nested partition involving a certain number of elements of the investigated systems. In order to evaluate the performance of our method, we test it with some benchmarks obtained by using a hierarchical factor model

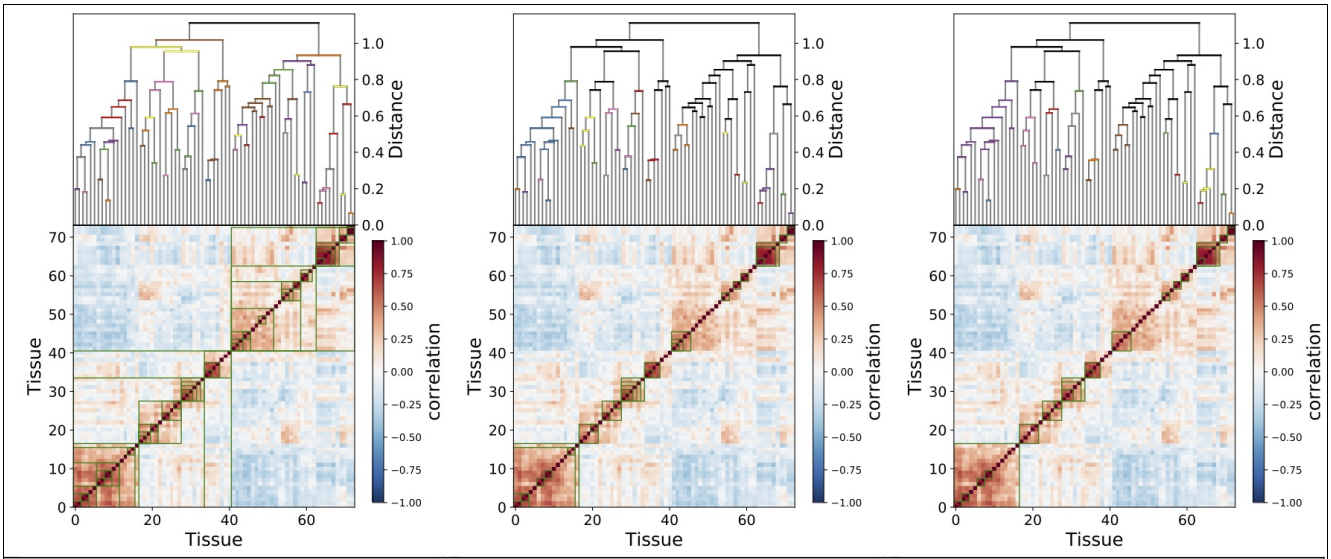
Results

By performing numerical experiments on a representative benchmark and on a reference empirical dataset, we show that our algorithm is quite accurate and much faster and scalable than the state of the art algorithm (Pvclust). Moreover, it shed light on the role and limits of the presence or absence of a procedure for the multiple hypothesis test correction. For these reasons, we believe the new algorithm will be of interest for those scholars working with large multivariate datasets in biology, computer science, neuroscience, physics, sociology, and other disciplines dealing with large scale multivariate data.

Info

-

Figure



Hierarchical trees (average HC) and correlation matrices of lung tissues dataset. The dataset was originally collected in [1] and it was used to provide an illustrative example of Pvclust performance in [2], In the correlation matrices we highlight with boxes hierarchically nested clusters detected by different algorithms. (*a*) SVHC, (*b*) Pvclust without multiple test correction, (*c*) Pvclust with "FDR" multiple test correction.

[1] Garber, M. E. *et al.* Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci.* 98, 13784–13789 (2001).
[2] Suzuki, R. & Shimodaira, H. Pvclust: an r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22, 1540–1542 (2006).

Availability	-

Corresponding Author

Name, Surname	Salvatore, Micciche'
Email	salvatore.micciche@unipa.it
Submitted on	25.04.2019