Preface

Michele Ceccarelli

Department of Biological and Environmental Sciences,
University of Sannio,
Viale Traiano 1, Benevento, Italy
and
Bioinformatics Core, BIOGEM,
Ariano Irpino, Italy
E-mail: ceccarelli@unisannio.it

Roberto Tagliaferri*

NeuRoNe Lab, DMI, University of Salerno,
via Ponte don Melillo, 84084 Fisciano, (SA) Italy
E-mail: rtagliaferri@unisa.it
*Corresponding author

Modern applied sciences are all based on evidence coming from observations and the revolution carried by current technology allows to gather more and more information and hypothesis from observed data. The way we observe biological systems, for example, both at the molecular and at the ecological and social levels, is always increasing in its degree of detail and are fundamentally changing the way of reaching scientific discoveries. Many applied sciences such as biology and geology, for example, are witnessing a transition becoming physical sciences relying of mathematics and engineering. This transition is due to a twofold and self-influencing effect: the former due to the fact that most of the work now is devoted to the analysis and interpretation of data coming from observations, and the latter because there is a strong pressure in developing better and new data analysis techniques which allow to extract structure and semantics from observed data, i.e., the captured data need to be converted into information and knowledge in order to become useful. 'Unsupervised clustering and explorative data analysis' are the basic tools of using computing power to apply mathematical and statistical methodologies, including new techniques, to structure extraction and knowledge discovery from data. It is also often referred as a branch of 'data mining'. It has the ambition of providing to non-expert users tools and algorithms for data analysis. However, without a strong knowledge of the methodologies and continuous relationships between domain scientists and statisticians and computer scientists can result in false-positives, no useful results and worst of all, results that are misleading and/or misinterpreted.

The development of unsupervised clustering and learning dates back to mid-fifties and sixties mainly in the field of perception modelling, image analysis and compression, if we think at the complexity and efficiency performed by visual systems of mammals in interpreting visual stimuli, then it is obvious that this has been a source of inspiration and an application arena for the evaluation of methodologies. Within this context,

self-organisation has been modelled as a process for describing the distribution of input stimuli through competitive learning. This process is aimed at finding structure and regularity in data. At the lower level of structure discovering from data there is the process of grouping or clustering. Clustering is useful because the underlying labels can be more meaningful than the numerical values and can lead to a better description of data. In addition, clusters representatives can be used for a compressed and more compact representation of data.

Many clustering algorithms have been proposed in literature and daily applied in hundreds of applications in economy, biology, geology, chemistry and others. This is also an active area of investigation as many researchers are faced with problems of improving current techniques or adapting them to specific problems. Some of the current topics of research in 'Unsupervised clustering and exploratory data analysis' are deepened in this issue. The main research topics can be summarised as follows:

- *New unsupervised learning techniques*. Novel techniques are being developed for facing some of the classical problems in unsupervised learning such as use of a priori knowledge during the inference process or how to model complex geometrical relationships among observations in very high-dimensional spaces etc.
- *New cluster validation and validity measures*. Many novel clustering algorithms are insufficiently evaluated, such that users remain unaware of their relative strengths and weaknesses. A more thorough use of quantitative, reproducible and objective cluster-validation techniques would permit users to alleviate this uncertainty, thus assisting the distinction between more and less useful methods and encouraging the acceptance of novel advanced clustering techniques.
- *Novel applications of unsupervised clustering*. There are several recent works trying to improve the ways in which unsupervised methods can help users into discovering novel knowledge in fields such as functional genomics, earth observations, analysis of social systems and so on.

The above mentioned problems are described in the papers of this issue with a deep level of details. In particular, the first two papers report approaches for the problem of multi-clustering with two different methods: in the former, 'Multiple data structure discovery through global optimisation, meta clustering and consensus methods' by Ida Bifulco, Carmine Fedullo, Francesco Napolitano, Giancarlo Raiconi and Roberto Tagliaferri, the authors propose a systematic approach to clustering, including the generation of a number of good solutions through global optimisation, the analysis of such solutions through meta clustering and the final construction of a small set of solutions through consensus clustering, all supported by a visual and interactive tool called MIDA. In the latter, 'A stability-based algorithm to validate hierarchical clusters of genes' by Roberto Avogadri, Matteo Re, Giorgio Valentini, Matteo Brioschi, Alessandro Beghini and Fulvia Ferrazzi, stability-based methods have been successfully applied to the validation of gene clusters discovered in gene expression data of patients affected by human myeloid leukaemia to discover significant clusters in hierarchical clusterings with a large number of examples and clusters. The third paper is also related to the stability analysis of the results of clustering, in particular the paper 'Concordance indices for comparing fuzzy, possibilistic, rough and grey partitions' by Michele Ceccarelli and Antonio Maratea, faces the problem of modelling concordance indices in the presence of uncertainty. Indeed, crisp partitions however cannot model ambiguity, vagueness or

uncertainty in class definition and thus are not suitable to model all cases where information lacks, terms definitions are intrinsically imprecise or the classification results from a human expert knowledge representation. In presence of vagueness, it is not obvious how to quantify overlap or agreement of two different partitions of the same data, and many facets of vagueness have emerged in literature through complimentary theories. The aim of the paper is to give simple numerical indices to quantify partitions agreement in the fuzzy, possibilistic, rough and grey frameworks.

In the fourth paper, 'Normalised compression distance and evolutionary distance of genomic sequences: comparison of clustering results' by Massimo La Rosa, Salvatore Gaglio, Riccardo Rizzo and Alfonso Urso, the clustering and the mapping obtained using a SOM-like algorithm, with the traditional evolutionary distance and the compression distance are compared in order to understand if the two distances sets are similar to compare genomic strings.

Exploratory analysis of genomic data sets using unsupervised clustering techniques is often affected by problems due to the small cardinality and high dimensionality of data sets. A way to alleviate those problems lies in performing clustering in an embedding space where each data point is represented by a vector of its memberships to fuzzy sets characterised by a set of prototypes selected from the data set. In the fifth paper, 'Clustering in the membership embedding space' by Maurizio Filippone, Francesco Masulli and Stefano Rovetta, the authors propose a constructive technique based on simulated annealing able to select sets of prototypes of small cardinality and supporting high quality clustering solutions.

The last paper, 'A one class KNN for signal identification: a biological case study' by Vito Di Gesú, Giosué Lo Bosco and Luca Pinello, describes an application of a one-class KNN to identify different signal patterns embedded in a noisy structured background. The problem becomes harder whenever only one pattern is well represented in the signal, in such cases one class classifier techniques are more indicated. The classification phase is applied after a preprocessing phase based on a multi layer model (MLM) that provides preliminary signal segmentation in an interval feature space. The one-class KNN has been tested, with a good recognition rate, on synthetic data simulating microarray data for the identification of nucleosomes and linker regions across DNA.