

Interactive analysis of genomic data: the Meta Clustering and Consensus approach ¹

R. Tagliaferri ², I. Bifulco, F. Napolitano and G. Raiconi

NEuRoNe Lab, DMI Università di Salerno, Fisciano (Sa), e-mail: {rtagliaferri, ibifulco, fnapolitano, gianni}@unisa.it

Abstract: Clustering of real-world datasets is a complex problem. Optimization models seeking to maximize a fitness function assume that the solution corresponding to the global optimum is the best clustering solution. Unfortunately, this is not always the case, mainly because of noise or intrinsic ambiguity in the data. Here we present a set of tools implementing classical and novel techniques to approach clustering in a systematic way, with applications to complex biological datasets. The tools deal with the problem of generating multiple clustering solutions, performing cluster analysis on such clusterings (i.e. Meta Clustering) and reducing the final number of clusterings by the appropriate application of different Consensus techniques. A subsequent crossing of prior knowledge to the obtained clusters helps the user in better understanding its meaning and validates the solutions. A collection of visualizations and interactive tools makes possible the challenge of managing such huge amount of information, allowing the user to extensively investigate data and extract knowledge from it.

Keywords: clustering, global optimization, k-means, EM, Consensus techniques, Meta Clustering

1. Introduction

In the last years machine learning methodologies has had new impulses inside Statistical Pattern Recognition area, due to the need of knowledge discovery from very large data sets, through both supervised and unsupervised approaches. Between unsupervised methods, clustering is one of the most used. It is the process of grouping sets of items in such a way that items belonging to the same group are similar and items belonging to different groups are not similar. Although it is very simple in definition, it becomes very complex in real-world applications. Clustering techniques have been applied to a wide variety of data mining problems. For example, in the field of medicine, clustering diseases and cures for diseases gives a key to understand relations between them. Furthermore, due to the exponential growth of genetic data availability, clustering proved to be a useful technique in searching for meaning in such huge quantity of information. On the other hand, as the amount of analyzed data grows, often also both noise intensity and the number of candidate clusters grows: in such a situation the weakness of the most used clustering approaches becomes more evident. Usually, clustering techniques aim at maximizing a

¹The present paper has been realized under the collaboration of the *Gruppo Interdipartimentale di Bioinformatica, Università "Federico II" di Napoli e Università di Salerno*

²Address of correspondence: DMI Università di Salerno, via Ponte don Melillo, 84084, Fisciano (Sa), Italia, rtagliaferri@unisa.it <http://www.neuronelab.dmi.unisa.it>

function indicating the quality of the formed clusters, but are usually able to find only local minima of such function. In real-world applications, the only way to attempt to find a good approximation of the global minimum is to resort to an algorithm specifically designed for Global Optimization (GO) (Di Gesú *et al.* (2005); Bifulco *et al.* (2008)). Unfortunately, even if an algorithm able to always find an approximation of the global optima existed, it would not ultimately solve the clustering problem. This is because of noise and ambiguity. Noise is due to experimental conditions and induces perturbations in the data that can force quality measures towards solutions that are not necessarily the best (some discussions about this point can be found in (Hu and Hu (2007); Agarwal and Mustafa (2004); Kaukoranta *et al.* (1996)). Ambiguity is an intrinsic property of data that can be inherently clustered in more than one way, with little or no clues about which solution is the best. The number of clusters (k), for example, in many situations just acts as the level of granularity the researcher chooses to analyze data. In other cases, once fixed k , more than one stable solution may exist.

In real cases several local minima could explain in a convincing manner the data distribution. In such cases sticking to the global optimum, even if it was possible to find it, would not be the right method. We approach the clustering as a multi-step exploration problem of possible solutions.

2. The proposed approach to clustering

As discussed in the introduction, when clustering complex data sets, the existence of a unique optimal solution is usually questionable and the problem of extrapolating a small number of good different solutions becomes crucial. Our approach to clustering includes the generation of a large number of solutions, the clustering of such solutions, their reduction through consensus clustering applied to the clusters of solutions and the analysis of the obtained final solutions supported by prior knowledge. In the following, some details are given for each of the steps.

1. **Generation of multiple clustering solutions.** The generation of multiple clustering solutions is performed exploiting a Global Optimization approach. However, in many applications the function of interest is multi-modal and possibly not differentiable and it is with this view in mind that the Controlled Random Search (CRS) algorithm (Price (1983)) was initially developed. Some CRS algorithms also include genetic techniques (Brachetti *et al.* (1997); Bresco *et al.* (2005)). We applied such technique to cluster analysis. We used a Price based Global Optimization algorithm to explore local minima of the K-means or Expectation Maximization (EM) objective functions. Instead of sticking to the best found solution, we collect all the solutions corresponding to local minima.

2. **Clustering the clustering solutions.** The first step towards the analysis of the solutions obtained as previously shown consists in grouping together the similar ones, that is in turn a clustering problem. This process is called Meta Clustering (Caruana *et al.* (2006)). The subset of clusterings can be evaluated by the domain expert or by quality indices.

3. **Reducing the number of solutions.** Consensus clustering, also known in literature as clustering ensembles or clustering aggregation, is the process of extrapolating a single clustering solution from a collection, in such a way to maximize a measure of agreement. We implemented software modules for six different Consensus techniques, by integrating the software released by the respective authors of the methods into our interactive framework. They include: Furthest Consensus Algorithm (Gionis *et al.* (2007)) and BallClust

(Gionis *et al.* (2007)), based on pairwise similarity; Best (Gionis *et al.* (2007)), based on an objective function; Cluster-based Similarity Partitioning (Strehl and Ghosh (2002)); Iterative Voting Consensus (Nguyen and Caruana (2007)) and Iterative Probabilistic Voting Consensus (Nguyen and Caruana (2007)) based on an iterative approach.

4. **Validating the final solutions.** For this step we propose an exhaustive enrichment analysis based on prior knowledge. This information can be exploited in at least two ways: to validate a clustering result (see Bifulco *et al.* (2009) as an example) and to infer new knowledge. The latter is the approach embraced here, where we seek for clusters (in the final solutions set) that are the most enriched with respect to some prior knowledge about the data. When a cluster is highly characterized by the presence of a given type of points, hypotheses can be made about the points that belong to different classes or that are not known to belong to any class. We perform an exhaustive search for such enriched clusters: consensus clustering is applied to all the nodes of the meta-clustering tree and each obtained solution is searched for enriched clusters. Of course, this step is done in a non-interactive way, while the results are shown visually as explained in section 3.. For example, this approach is a generalization of that proposed in (Ciaramella *et al.* (2008)).

3. Visual and interactive tools

The human eyes and brain together make a formidable pattern detection tool, but, for them to work, the data must be represented in a low-dimensional space, usually of two or three-dimensions. Our tools enable the user to go through the entire clustering process with the support of ad hoc visualizations and interactive tools. The software we developed runs in the MATLAB environment and is mainly integrated in the MIDA (MIDA-Software (2009)) project.

In the above step 4 we described the validation step, that is a non-visual process. Here we describe how the results of such process can be visualized. We are interested in analyzing mainly two aspects of the validation phase: where is located the consensus solution containing the most enriched clusters on the meta-clustering dendrogram and where the points belonging to such clusters are localized in the data space. Our tools visualize such information in a simple way. For the former we highlight the node of the meta-clustering tree that is the root of the best subtree. This gives important insight on the usefulness of the whole process: if the best solution was found on the root of the tree, than meta clustering would not be necessary (the best solution is obtained applying consensus to all the solutions); if the best solution was found on a leaf of the tree, than both meta clustering and consensus would not be necessary (the best solution is just one of the starting solutions). In our experience both cases are quite rare. Information about localization of the most enriched clusters inside the data space is shown through a simple 2D Multi-Dimensional Scaling (MDS) visualization: we plot all the points using a different color for those belonging to the cluster and a different shape for the points belonging to the class whose enrichment is high for that cluster.

4. Conclusions

In this paper we showed how Global Optimization, Meta Clustering, Consensus Clustering and prior knowledge can be exploited to extract structured groups from data. We

showed how a systematic approach to clustering can help in extensively searching the space of clustering solutions towards a better understanding of data structure. Prior knowledge (possibly partial) can help in both validating clusters and formulate hypotheses about missing information. The entire process is simplified and sped up by using Matlab visual and interactive tools developed by us. Most of the tools have been integrated in a modular environment called MIDA.

References

- Agarwal P. and Mustafa N. (2004) k-means projective clustering, in: *Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ACM Press, New York – NY, 155–165.
- Bifulco I., Fedullo C., Napolitano F., Raiconi G. and R. T. (2009) Global optimization, meta clustering and consensus clustering for class prediction, in: *Proceedings of IJCNN 2009*, IEEE Press, (in press).
- Bifulco I., Murino L., Napolitano F., Raiconi G. and R. T. (2008) Using global optimization to explore multiple solutions of clustering problems, *Lecture Notes in Computer Science*, 5179, 724–731.
- Brachetti P., De Felice Ciccoli M., Di Pillo G. and Lucidi S. (1997) A new version of the price’s algorithm for global optimization, *Journal of Global Optimization*, 10, 165–184.
- Bresco M., Raiconi G., Barone F., De Rosa R. and Milano L. (2005) Genetic approach helps to speed classical price algorithm for global optimization, *Soft Computing Journal*, 9, 525–535.
- Caruana R., Nam Nguyen M. and Smith C. (2006) Meta clustering, in: *Proceedings of the Sixth International Conference on Data Mining (ICDM’06)*, IEEE Press, New York – NY, 107–118.
- Ciaramella A., Coccozza S., Iorio F., Miele G., Napolitano F., Pinelli M., Raiconi G. and Tagliaferri R. (2008) Interactive data analysis and clustering of genomic data, *Neural Networks*, 21, 368–378.
- Di Gesù V., Giancarlo R., Lo Bosco G., Raimondi A. and Scaturro D. (2005) Genclust: A genetic algorithm for clustering gene expression data, *BMC Bioinformatics*, 6, 289–300.
- Gionis A., Mannila H. and Tsaparas P. (2007) Clustering aggregation, *ACM Trans. Knowl. Discov. Data*, 1.
- Hu Y. and Hu Y. (2007) Global optimization in clustering using hyperbolic cross points, *Pattern Recognition*, 40, 1722–1733.
- Kaukoranta T., Franti P. and Nevalainen O. (1996) Reallocation of gla codevectors for evading local minima, *Electronics Letters*, 32, 1563–1564.
- MIDA-Software (2009) *NeuRoNe lab*, *DMI*, *University of Salerno*, <http://www.neuronelab.dmi.unisa.it>.
- Nguyen N. and Caruana R. (2007) Consensus clustering, in: *ICDM 2007*, IEEE Press, 607–612.
- Price W.L. (1983) Global optimization by controlled random search, *Journal of Optimization Theory and Applications*, 55, 333–348.
- Strehl A. and Ghosh J. (2002) Cluster ensembles - a knowledge reuse framework for combining multiple partitions, *Journal of Machine Learning Research*, 3.