IJCSS
International Journal of Complex Systems in Science

Int. J. Complex Systems in Science
vol. **1** (2011), pp. 21–24

Data clustering using community detection algorithms

Clara Granell[1,†], Sergio Gómez[1] and Alex Arenas[1]

[1] *Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili*

Abstract. One of the most important problems in science is that of inferring knowledge from data. The most challenging issue is the unsupervised classification of patterns (observations, measurements, or feature vectors) into groups (clusters) according to their similarity. The quantification of similarity is usually performed in terms of distances or correlations between pairs. The resulting similarity matrix is a weighted complete graph. In this work we investigate the adaptation and performance of modularity-based algorithms to analyze the structure of the similarity matrix. Modularity is a quality function that allows comparing different partitions of a given graph, rewarding those partitions that are more internally cohesive than externally. In our problem cohesiveness is the representation of the similarity between members of the same group. The modularity criterion, however, has a drawback, the impossibility to find clusters below a certain size, known as the resolution limit, which depends on the topology of the graph. This is overcome by applying multi-resolution analysis. Using the multi-resolution approach for modularity-based algorithms we automatically classify typical benchmarks of unsupervised clustering with considerable success. These results open the door to the applicability of community detection algorithms in complex networks to the classification of real data sets.

Keywords: Clustering, networks, community structure, multi-resolution.
MSC 2000: 62H30, 05C82

1. Introduction

Unsupervised classification (or clustering) stands for the process of grouping data according to a certain distance. Generally speaking, the matrix of distances between any pair of data (similarity matrix) can be represented as a graph (or network) [1]. Our idea is to confront the problem of clustering using techniques developed in the field of complex networks.

Complex networks are graphs representative of the intricate connections between elements in many natural and artificial systems, whose description in

terms of statistical properties have been largely developed looking for a universal classification of them. However, when the networks are locally analyzed, some characteristics that become partially hidden in the global statistical description emerge. The most relevant is perhaps the discovery in many of them of community structure, meaning the existence of densely (or strongly) connected groups of nodes, with sparse (or weak) connections between these groups. Very often networks are defined from correlation data (or distances in a certain space) between elements. Our goal is to study the use of community detection algorithms for unsupervised data classification.

2. A complex networks approach to the unsupervised classification of data

The methodology we devise consist in to analyze the similarity matrix using a community detection algorithm based on modularity. The first idea is to propose a problem-specific similarity measure such that the resulting clusterization problem will be reduced to that of finding the most densely connected groups.

We will use the algorithms we generated to detect community structure in networks [2] to discover clusters in the distance matrix obtained from the IRIS data set. This set is composed of (4th dimensional) patterns of width and length of petals and sepals of three different classes of flowers [3]. The idea is to define some distance between patterns and analyze the subsequent network using our methods. We will screen the distance matrix using the multi-resolution scheme proposed in [4]. Summarizing, the proposed scheme to classify the data is as follows:

- Variable selection: decide which variables are the most adequate for the classification problem using multivariate statistics.
- Construct the similarity matrix, in such a way that the distances between pairs of data are willing to be mapped as a link with a certain strength.
- Apply a multi-resolution community detection algorithm to the similarity matrix
- Detect the best partition in the screening of resolution scales and propose the classification.

3. Results on the IRIS data set

The unsupervised cqlassification of the IRIS data set is a major challenge in artificial intelligence and statistical theory. The three types of flowers, Setosa,

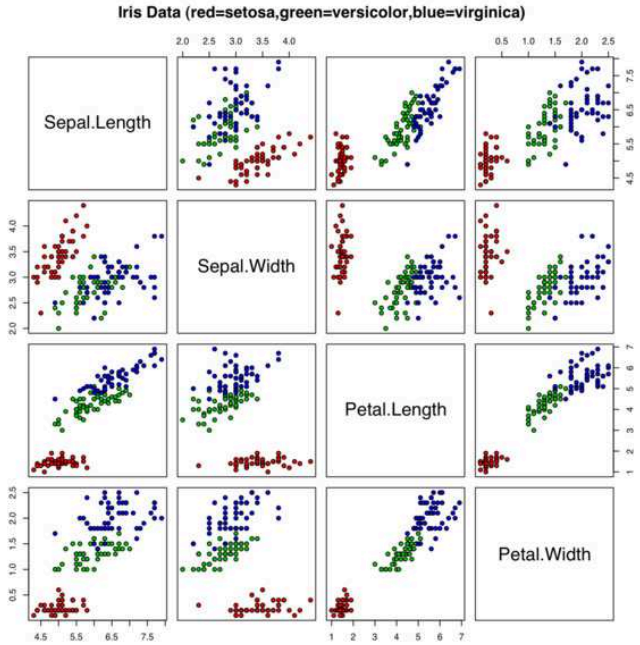

Figure 1: Feature vectors for the IRIS data set. The feature selection process raises features petal width and length as the most relevant variables. From Wikipedia Commons.

Versicolor and Virginica form a linear separable problem (Setosa and the other two), and a non-linear separable problem (Versicolor and Virginica). Plots for the cross-variables and type of flowers are represented in Fig. 1. A standard feature selection algorithm Minimum-redundancy-maximum-relevance based on mutual information [5] gives us two variables from the four variables set, petal length and petal width. Working with these two variables, we propose to build up a similarity matrix as the distance in the two dimensional space mentioned with respect to the center of mass of the data set in this space. For any pair of flowers i and j, we define the distance $d_{ij} = \bar{d} - \|x^i - x^j\|)$, where $\bar{d}$ stands for the average distance of the set, and $\|\cdot\|$ is the euclidean distance between the feature vectors of each flower. The resulting similarity matrix is interpreted as a weighted network whose communities will, in principle, reproduce the right clustering of the data. Using a multi-resolution scheme we find that the most relevant structure found in the data corresponds to a partition in three clusters with a 100% success detecting Setosa, and an approximately 86% success disentangling Versicolor and Virginica, see Fig.2.

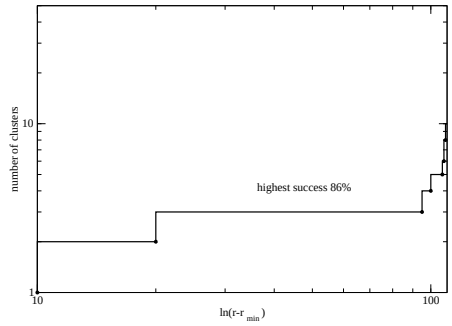

Figure 2: Number of clusters as a function of the resolution parameter. The highest success is achieved for three clusters.

The method proposed can be extended to other classification problems and could be understood as a new data clustering algorithm.

References

[1] A.K. Jain, M.N. Murty and P.J. Flynn, *ACM Computing Surveys* **31**, 3 (1999).

[2] S. Gomez, P. Jensen and A. Arenas, *Physical Review E* **80**, 016114 (2009).

[3] R.A. Fisher, *Annals of Eugenics*, **7**, 179 (1936).

[4] A. Arenas, A. Fernández and S. Gomez, *New Journal of Physics*, **10**, 053039 (2008).

[5] H. Peng, F. Long, and C. Ding, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226-1238 (2005).