CLUSTER ANALYSIS FOR NETWORKS USING A FUZZY APPROACH

Ilaria Bombelli [1][2], Ichcha Manipur [3] and Maria Brigida Ferraro [2]

[1] Italian National Institute of Statistics

[2] Department of Statistical Sciences, Sapienza University of Rome, (e-mail: ilaria.bombelli@uniroma1.it, mariabrigida.ferraro@uniroma1.it)

[3] Institute for High-Performance Computing and Networking, National Research Council, (e-mail: ichcha.manipur@icar.cnr.it)

ABSTRACT: As the network representation is widely used to describe problems in an increasing number of disciplines, novel methodologies are needed to handle such complexity. In particular, cluster analysis is an interesting and challenging task in the network framework. In this work, we focus on how to represent networks for fuzzy clustering and how to apply standard fuzzy algorithms for clustering multiple networks on synthetic data.

KEYWORDS: Ensembles of Networks, Fuzzy Clustering, Networks Clustering, Whole-graph Embedding.

1 Introduction

Networks represent a powerful model for problems in different scientific and technological fields, such as neuroscience, molecular biology, biomedicine, sociology, social network analysis and political science. The increasing number of network applications leads research on clustering analysis develop rapidly.

In a network framework, a well-known approach to the clustering problem is the detection of clusters of nodes (or *communities*). A new approach to the clustering problem is to consider a single network as the unit of interest and to detect clusters of networks.

What is proposed here is to apply fuzzy cluster analysis techniques to identify clusters of networks by choosing an adequate representation. The novelty here lies in the usage of a fuzzy approach: indeed, related works use only a hard approach to clustering, meaning that each network can belong to one cluster only. However, networks may have characteristics in common to more than one cluster, and therefore in such situations, a more flexible approach is more adequate. In this sense, the fuzzy approach guarantees major flexibility than the hard approach, by allowing each network to belong to all clusters according to different membership degrees.

2 Network representation

To cluster networks, we need to find an adequate representation. In the early proposals on this topic, networks have been represented using some topological characteristics, but very different networks might be represented by the same values of the chosen features, making the data analysis difficult. Moreover, the well-known *adjacency matrix* representation does not account for differences in specific parts of the network and therefore ignores its topological characteristics.

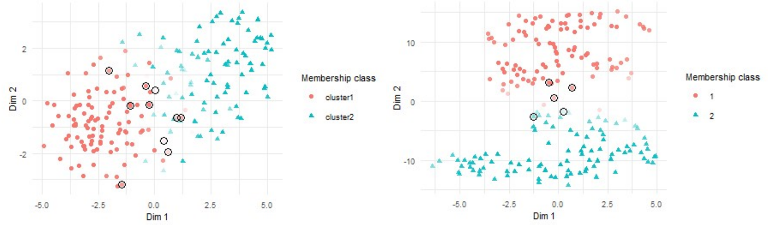
To overcome these limits, we study two types of network representations: a probabilistic representation of graphs (either Node Distance Distribution or Transition Matrices, see Granata *et al.*, 2020 for details) and a whole-graph embedding representation (Joint Embeddings by Wang *et al.*, 2021). By using the probabilistic representation, the Jensen-Shannon (JS) Divergence is then used to compute pairwise distances between networks and finally to obtain a distance matrix; instead, the embedding techniques provide a vector space representation of the networks to identify a space that is optimal with respect to some characteristics; the output is therefore a units by variables matrix, where units are networks and variables are networks' features.

3 Algorithms for fuzzy clustering

Once we have chosen how to adequately represent the networks, it is possible to apply fuzzy clustering algorithms. We use Non-Euclidean Fuzzy Relational Clustering, introduced by Davé & Sen, 2002, when the networks are represented by a matrix of distances; instead, we applied the Fuzzy k-Means (Bezdek, 1981), when they are in form of a feature matrix.

4 Simulation

We empirically analyze our proposal on synthetic dataset. In detail, the simulated networks are generated using the Multiple Random Eigen Graphs (MREG) model, defined in Wang *et al.*, 2021. Particularly, an MREG dataset with 200 graphs having 100 nodes each was generated using the MREG model. The



(a) NEFRC applied on $\mathcal{M}^N$

(b) [FkM applied on JE$_{d=2}$

Figure 1: t-SNE representation of clustering results of NEFRC, FkM (MREG networks). Misclassified units are circled in black. The intensity of the colors is given by the membership degree of each network to the corresponding assigned cluster.

graphs belong to 2 classes, with 100 graphs in each class. The clustering task consists of grouping networks with a similar distribution of edges.

Here, for the sake of brevity, we show two applications of fuzzy clustering algorithms (NEFRC and FkM) to two networks' representations: $\mathcal{M}^N$, i.e. the distance matrix obtained by applying JS divergence on Node Distance Distribution representation of networks; JE, i.e. the feature matrix resulting from Joint Embedding technique. Table 1 shows the algorithm's performance using the clustering validity indices. In detail, high ARI and AMI indices values show that most of the networks are correctly assigned to their original clusters.

Table 1: Main results of the application of NEFRC the Distance Matrix $\mathcal{M}^N$ and FkM to Feature Matrix JE (MREG networks)

	NEFRC		FkM	
	$\mathcal{M}^N$		JE	
	ARI	AMI	ARI	AMI
Median	**0.81**	0.72	**0.9**	0.83
IQR	0	0	0	0
SD	0.01	0.02	0.01	0.01

The graphical representation allows us to explore the results more in-depth in Figure 1. Figure 1 shows that the two clusters are well separated; misclassified networks are highlighted by the circled points. The fuzzy membership degrees allows us to deeply study the misclassified units. By applying NEFRC to distance matrices, we notice that, on average, approximately 40% of misclassified networks are in the middle of the two cluster prototypes, having membership degrees close to 0.5 and being represented by blurry colors in Figure 1 (a). Regarding the application of F*k*M to JE, we notice that 20% of misclassified units are represented by very blurry colors in Figure 1 (b) and are softly assigned to both the clusters. Therefore, membership degrees allow us to consider the uncertainty of an assignment of a unit to a cluster and then possibly add information on clustering interpretation: this represents one of the main advantages of a fuzzy approach.

5 Final Remarks

This study explores clustering analysis when the statistical units are networks. To this extent, we focus on different methodologies that can provide a suitable representation of the sample of the networks for subsequent data analysis. We applied fuzzy clustering algorithms on such representations, using standard metrics to evaluate their performance on synthetic datasets. Our analysis provides valuable hints for cluster analysis in a network framework.

References

BEZDEK, JAMES C. 1981. *Pattern recognition with fuzzy objective function algorithm*. Plenum Press, New York.

DAVÉ, RAJESH N, & SEN, SUMIT. 2002. Robust fuzzy clustering of relational data. *IEEE Transactions on Fuzzy Systems*, **10**(6), 713–727.

GRANATA, ILARIA, GUARRACINO, MARIO ROSARIO, MADDALENA, LUCIA, & MANIPUR, ICHCHA. 2020. Network Distances for Weighted Digraphs. *Pages 389–408 of: International Conference on Mathematical Optimization Theory and Operations Research*. Springer.

WANG, SHANGSI, ARROYO, JESÚS, VOGELSTEIN, JOSHUA T, & PRIEBE, CAREY E. 2021. Joint embedding of graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 1324–1336.