Comparison of uncertainty of clustering algorithms in random variables networks

Aroslankin A.D. Kalyagin V.A..
National Research University Higher School of Economics, Nizhny Novgorod,
Laboratory of Algorithms and Technologies for Networks Analysis

Problem statement. Cluster analysis is a powerful tool in network science and it is well developed in many directions. However, the uncertainty analysis of clustering algorithms and the development of robust clustering algorithms are still not sufficiently investigated in the literature. Probabilistic approach to robust clustering based on the theory of labeled random point processes was proposed in (Dalton at al, 2018). In the present paper we propose to study uncertainty of clustering algorithms within the framework of the random variable network model (Kalyagin at al, 2020). This model can be considered as generalization of various network models: gene expression network, brain network, climate observation network, and stock market network.

Model. We study the problem of uncertainty of clustering algorithms in the framework of random variable network. Random variables network is a pair (X, γ), where $X = (X_1, \dots, X_N)$ is a random vector and $\gamma = \gamma(U, V)$ is a measure of dependence between pair of random variables U and V. Random variable network generates a network model, complete weighted graph with N nodes. The node i is associated with the random variable X_i, and the weight of edge (i,j) is given by $\gamma_{i,j} = \gamma(X_i, X_j)$. Clustering algorithms will be applied for this network model. Our research question is the following: given a sample of observations of the random vector X, and a clustering algorithm, evaluate uncertainty of identification of clusters by this algorithm in associated random variable network model.

Methodology. We use the concept of reference or true network and the concept of sample network (Kalyagin at al, 2020). Associated cluster structures will be called reference and sample cluster structures. Uncertainty of identification of the cluster structure will be measured by expected value of loss related with error of

identification. To measure this loss we compare reference and sample cluster structures with the use of popular RAND index (measure of the difference of two partitions of the set of nodes of graph).

Results. We compare uncertainty of clustering algorithms by numerical simulations. Distribution of the random vector X is supposed to be multivariate normal distribution. Following clustering algorithms are compared from uncertainty point of view:

- single linkage hierarchical algorithm,
- spectral clustering algorithm,
- spectral optimization of modularity
- Louvain algorithm
- hedonic game clustering algorithm
- asynchronous fluid communities algorithm

Different types of covariance matrix are considered. In particular we use an appropriate form of block stochastic models to investigate "phase transition" phenomena for above mentioned clustering algorithms. Finally, we discuss an impact of uncertainty of clustering on portfolio optimization in stock market network (Tola at al., 2008).

Acknowledgments. The work was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics.

References:

1. Dalton LA, Benalcázar ME, Dougherty ER (2018) Optimal clustering under uncertainty. PLoS ONE 13(10): e0204627
2. Kalyagin V. A., Koldanov A. P., Koldanov P.A., Pardalos P.M. (2020) Statistical analysis of graph structures in random variable networks, Springer Brief in Optimization, Springer.
3. Tola, V., Lillo, F., Gallegati, M., & Mantegna, R. N. (2008). Cluster analysis for portfolio optimization. Journal of Economic Dynamics and Control, 32(1), 235-258.