Consensus Methods for Network Analysis of Biomedical Data: Case Studies on Brain Functional Connectivity Network and Gene-Gene Association Networks

Reddy Rani Vangimalla and Jaya Sreevalsan-Nair, Senior Member, IEEE
Graphics-Visualization-Computing Laboratory, *and* E-Health Research Center,
International Institute of Information Technology Bangalore, Karnataka 560100, India.
reddyrani.vangimalla@iiitb.org | jnair@iiitb.ac.in

***Abstract*— Networks in biology are widely studied to understand biological patterns and significance. The studies are more significant for various disease diagnosis and treatments obtained from biomedical data. We are interested in exploring consensus methods to improve the accuracy of outcomes for data mining, e.g., community detection. In our doctoral study, we look at the novel applications of consensus methods in two different case studies. In the first study, we use consensus methods for community detection in resting-state brain functional connectivity network obtained from functional magnetic resonance imaging (fMRI) data. In the second problem, we use consensus methods in aggregating different regression models for constructing networks in the gene association study. Here, the outcome is a network between DNA methylation features and mRNA genes, derived from samples in cancer patients.**

I. Introduction

The use of networks for analysing biomedical datasets has been widely used for several decades. Networks generally either occur in biomedical datasets naturally or are derived based on the governing data model used for analysis. The pathway networks in gene association study [1] belong to naturally occurring networks, with appropriate semantics. The correlation networks in brain structural and functional connectivity studies [2], [3], as well as correlation and regression networks in gene association study [4]. In our doctoral work, we address some of the open problems in two case studies, namely, functional segregation of brain functional connectivity network at resting state [5], and identification of significant genes in DNA (deoxyribonucleic acid) methylation and mRNA (messenger ribonucleic acid) gene expression using integrated analysis. The former is derived from functional magnetic resonance (fMRI) for a specific cohort [6], and the latter is derived from gene expression profiles of subjects with specific cancer profile in The Cancer Genomic Atlas (TCGA) dataset[1].

II. Case Study 1: Resting-State Brain Functional Connectivity Network

The motivation for functional segregation, i.e., identifying communities in brain functional connectivity network is to understand how different regions of interest (ROIs) in the brain get co-activated for specific cognitive tasks or in resting-state. There have been several studies on the resting-state analysis of the brain, as the data collection is done in a relatively more controlled setting compared to the cognitive task-based studies. We focus on the functional segregation during resting-state here. The brain functional connectivity network is computed using the correlation between ROIs, where the fMRI signals for the region are aggregated over space and time. Functional connectivity is inferred from correlations between ROIs based on the blood-oxygenation level dependent (BOLD) signals in fMRI imaging.

However, there are several points of contention on proposed methods for analysing these networks, such as in the use of threshold for the understanding topology of significant sub-network [7], [8], global signal regression [9], eyes-open versus eyes-closed studies [10], etc. Here, we consider the case of finding functional segregation, in the form of node partitions, using the complete (correlation) network of brain functional connectivity, i.e., without applying a threshold for edge-filtering.

There has been active research in consensus clustering techniques across different domains. Our proposed method is a combination of identifying a co-association matrix using the consensus from an ensemble of clustering methods [11] as well as multi-resolution methods used in clustering in brain structural networks [12]. In our work, the challenge is to perform the node partitioning in a complete network, given we are not filtering edges using a pre-computed threshold. In the general sense of correlation network, we see the nodes or ROIs as random variables. Thus, we use exploratory factor analysis [13] using appropriate parameter settings on a correlation matrix, which is routinely done. We then identify the resolution in our proposed method to be the number of factors (n_F). We, thus, use a consensus method of identifying the degree of *co-association* of any two given nodes, which says how often two nodes will occur in the same module. We compute a co-association matrix for each resolution, and then aggregate them using an averaging operation. The co-association matrix is now treated as a sparsified connectivity matrix, which is then used for community detection as is done for edge-filtered functional connectivity network. We study the performance of different community detection methods, e.g., Louvain community detection [14], on the sparsified network, and analyse the biological significance

[1]https://www.cancer.gov/tcga

of the node partitions obtained using our consensus method.

III. Case Study 2: DNA Methylation-mRNA Association Network

The motivation behind an integrated study of DNA methylation and mRNA expression profiles is to identify the influence of genes in different stages implied in the dogma of molecular biology. An integrated study has been done using regression coefficients [4], where the multilayered regression networks have been constructed by aggregating similarity matrices based on regression coefficients. The multiple layers correspond to the gene-gene association at the DNA methylation process, and mRNA expression, respectively. Vangimalla et al. [4] have proposed the use of multiple regression models, e.g., LASSO, GLASSO, etc., and aggregating the outcomes of these models using similarity network fusion [15]. However, there is a debate on the use of regression coefficients analogous to a feature vector [16]. Hence, we propose new methods of generating networks using outcomes of multiple regression models.

Here, we remodel the multivariate multiple linear regression between DNA methylation features, as independent variables, and mRNA genes, as dependent variables, as multi-layered networks. In our network model, we construct correlation network of independent variables, partial correlation network of dependent variables, and a bipartite graph of cross-links between the two layers using regression coefficients. We then analyse each of the layers and the cross-links separately, first, and then, together. The idea here is to modularise the dependencies within the sets of variables and across the sets. Thus, we analyse each network *component* in isolation as well as integrated. Some of the analytical tasks performed on the layers are community detection, ranking of genes based on their local neighborhood, and identification of common genes across layers.

We use consensus of different regression models in arriving at the cross-links as well as the partial correlation network of the dependent variable. We determine appropriate aggregation operation for obtaining the consensus. For example, we have experimented the use of similarity network fusion [15] for aggregating the partial correlation matrices. We use this analysis to study cancer genes obtained from subjects for specific cancer profiles, e.g., lung, brain, breast, etc., of the TCGA datasets. Our preliminary results of this analysis have given significant genes as well as gene-pairs within each genomic dataset as well as across different genomic datasets. We have used gene ontological tools, such as DAVID, STRING, to validate the identified genes in specific cancer profile.

IV. Conclusions

The goal of our doctoral study is to determine novel methods of constructing and analysing networks from biomedical datasets, using a combination of methods in network science, matrix analysis and statistics. In this paper, we have demonstrated two different case studies where we have re-modeled an existing network as well as constructed new complex (multilayered) networks using our proposed methodology. In our work, we have exclusively focused on non-overlapping communities, however, the consensus results can be generated using overlapping communities. Our future work is in further network analysis as well as appropriate biological validation using standard tools and findings in the literature.

Acknowledgments

This doctoral study has been supported by the Visvesvaraya PhD Scheme for Electronics and IT, initiated by the Ministry of Electronics and Information Technology, Government of India.

References

[1] Pau Creixell, Jüri Reimand, Syed Haider, Guanming Wu, Tatsuhiro Shibata, Miguel Vazquez, Ville Mustonen, Abel Gonzalez-Perez, John Pearson, Chris Sander, et al., "Pathway and network analysis of cancer genomes," *Nature methods*, vol. 12, no. 7, pp. 615, 2015.

[2] Olaf Sporns, "Structure and function of complex brain networks," *Dialogues in clinical neuroscience*, vol. 15, no. 3, pp. 247, 2013.

[3] Ed Bullmore and Olaf Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature reviews neuroscience*, vol. 10, no. 3, pp. 186, 2009.

[4] Reddy Rani Vangimalla, Hyun-hwan Jeong, and Kyung-Ah Sohn, "Integrative regression network for genomic association study," *BMC medical genomics*, vol. 9, no. 1, pp. 31, 2016.

[5] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol, "Exploring the brain network: a review on resting-state fmri functional connectivity," *European neuropsychopharmacology*, vol. 20, no. 8, pp. 519–534, 2010.

[6] Mingrui Xia, Jinhui Wang, and Yong He, "Brainnet viewer: a network visualization tool for human brain connectomics," *PloS one*, vol. 8, no. 7, pp. e68910, 2013.

[7] Nicolas Langer, Andreas Pedroni, and Lutz Jäncke, "The problem of thresholding in small-world network analysis," *PloS one*, vol. 8, no. 1, pp. e53199, 2013.

[8] Fabrizio de Vico Fallani, Jonas Richiardi, Mario Chavez, and Sophie Achard, "Graph analysis of functional brain networks: practical issues in translational neuroscience," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1653, pp. 20130521, 2014.

[9] Kevin Murphy, Rasmus M Birn, Daniel A Handwerker, Tyler B Jones, and Peter A Bandettini, "The impact of global signal regression on resting state correlations: are anti-correlated networks introduced?," *Neuroimage*, vol. 44, no. 3, pp. 893–905, 2009.

[10] Jie Wei, Tong Chen, Chuandong Li, Guangyuan Liu, Jiang Qiu, and Dongtao Wei, "Eyes-open and eyes-closed resting states with opposite brain activity in sensorimotor and occipital regions: multidimensional evidences from machine learning perspective," *Frontiers in human neuroscience*, vol. 12, pp. 422, 2018.

[11] Ana LN Fred and Anil K Jain, "Combining multiple clusterings using evidence accumulation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 6, pp. 835–850, 2005.

[12] Lucas GS Jeub, Olaf Sporns, and Santo Fortunato, "Multiresolution consensus clustering in networks," *Scientific reports*, vol. 8, no. 1, pp. 3259, 2018.

[13] Anna B Costello and Jason Osborne, "Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis," *Practical assessment, research, and evaluation*, vol. 10, no. 1, pp. 7, 2005.

[14] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, pp. P10008, 2008.

[15] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature methods*, vol. 11, no. 3, pp. 333, 2014.

[16] Kirsten Bulteel, Francis Tuerlinckx, Annette Brose, and Eva Ceulemans, "Using raw var regression coefficients to build networks can be misleading," *Multivariate behavioral research*, vol. 51, no. 2-3, pp. 330–344, 2016.