2019-12

Network Classification with Incomplete Information

Yoshida, Ruriko

Monterey, California: Naval Postgraduate School

http://hdl.handle.net/10945/70018

MONTEREY, CALIFORNIA

NETWORK CLASSIFICATION WITH INCOMPLETE INFORMATION

Executive Summary Type: Final Report
Period of Performance: 03/15/2019–03/30/2020

Researchers:
Principal Investigator (PI): Dr. Ruriko Yoshida, Graduate School of Operations and Information Sciences (GSOIS), Operations Research (OR)
Additional Researcher(s): N/A
Student Participation: LT. Carolyne Vu and LT. Ross Spinelli, USN, OR

Prepared for:
Topic Sponsor Lead Organization: N1
Topic Sponsor Organization (if different): N/A
Topic Sponsor Name: Office of Naval Intelligence, LCDR, James DeWitt, Navy
Topic Sponsor Contact Information: james.p.dewitt@navy.mil

NPS NRP Executive Summary

Title: Network Classification with Incomplete Information

Report Date: 03/20/20 Project Number (IREF ID): NPS-19-N072-A

Naval Postgraduate School / School: Graduate School of Operational and Information Sciences

EXECUTIVE SUMMARY

Project Summary

With the growth of accessible data, particularly for incomplete networks, a demand for effective methods of analyzing networks has emerged. Even as means for data collection advance, incomplete information remains a reality for numerous reasons. For example, data can be obscured by excessive noise, and surveys for information typically contain some non-respondents. In other cases, simple inaccessibility restricts observation. Also, for illicit groups, we are confronted with attempts to conceal important elements or propagation of false information. In the real-world, it is difficult to determine when the observed network is both accurate and complete.

In this research, we consider two objectives: (1) a method for classification of incomplete networks (network classification) and (2) inference on how much missing information (network completion problem). In contrast to the current method of training models with only complete information, we examine the effects of training our classification model, and training network completion problem, with both complete and incomplete network information.

Our results strongly indicate the need to include incomplete network representations in training the classification model and network completion. Incorporating incomplete networks at various stages of completeness allow the machine to examine and learn the nuances of incomplete networks. By allowing the machine to study incomplete network structural features, it has an improved ability to recognize and classify other incomplete networks. We also confirm these simple, easily calculated network features are sufficient to classify an incomplete network and network completion.

Keywords: *machine learning, network classifications, network completion problem, random forests*

Background

An intelligence community's assessment of enemy organizations requires accurate classification of the observed network before the intelligence team can develop a strategy for combating the adversary. Problems are typically time-sensitive; however, gathering this complete and actionable intelligence is a challenging mission that could span years. An adversary's actions are secretive in nature, making it extremely difficult to collect a complete observation of the network. Crucial information is deliberately concealed. Intentionally dubious information might create problematic noise or false imputations. Thus, if an observed incomplete network can be classified as-is without delay, the network can be properly analyzed for a strategy to be devised and acted upon earlier.

With a method to accurately classify an incomplete network, techniques of imputation can be reserved for post-classification. This allows for the estimation to be tailored accordingly by network class in an effort to maintain the network's true structure. These techniques could provide the intelligence team with a reasonable evaluation of an enemy's prospective associations or activities.

It is commonly understood that graphs of a class will have similar characteristics in its topology. Under this assumption, unique network features should be leveraged to classify an unknown network. Li et al., (2012) proposed an alternative approach to kernel methods, and conducted a study of biological network classification based on attribute vectors generated from global topological and label features. They discovered that networks from similar classes have similar characteristics, and network characteristics carry distinctions leverageable in classification algorithms. This study found their feature-based classification models produced similar accuracy rates, with less computational requirements, than conventional kernel methods of measuring similarity between networks based on shared patterns.

Canning et al., (2018) investigated the use of network features for classification of real-world observed networks. Their research found that networks from differing classes do contain distinguishing structural features useful in network classification. Research prior to this study was mainly focused on classification of synthetic networks or different networks within one specific class. Their study included synthetically generated networks, and Canning et al., (2018) discovered "synthetic graphs are trivial to classify, as the classification model can predict with near-certainty the network model used to generate it." Their multiclass classification model using random forest (RF) was successful in classifying real-world networks using network features.

All of the aforementioned studies of feature-based classification presume complete network information in their methods. In contrast, we seek to examine an RF model that classifies a graph as it is observed—even while incomplete. Incomplete data is a reality of analyzing real-world networks as portions of the observed data may remain unknown for different reasons, such as: data obstruction by excessive noise, non-respondent survey answers, deliberate concealment, or inaccessibility for observation. The proper handling of incomplete data is a critical requirement for accurate classification, and an inapt approach could cause significant errors in classification results.

Our approach for handling incomplete data is the use of machine learning (ML) techniques, such as support vector machines and decision trees. However, when using any of these methods, we must be attentive to potential incidents of significant bias, added variance, or risks of generalizing estimated data. Thus, we seek to develop a method for classifying an incomplete network without estimations to complete the network. Once classified, the methods of predicting unknown data can be customized to consider that network class's known properties, not just its observed features.

Findings and Conclusions

In this research, we consider a method for classification of incomplete networks. We examine the effects of training the classification model with complete and incomplete information. Observed network data and their network features are classified into technological, social, information, and biological categories using supervised learning methods. This comparative analysis contributes to a better understanding of network characteristics for classification. Then we consider to create a robust method for rebuilding a graph network with missing information. We propose a method for classifying the percent of information missing in networks based on feature characteristics, which will determine how much information needs to be rebuilt. In this project two students, LT Carolyne Vu and LT Ross Spinelli participated. LT Vu worked on the classification of network from an incomplete network, and LT Spinelli worked on the network completion.

First, we consider a method for classification of incomplete networks, and classify real-world networks into technological, social, information, and biological categories by their structural features using supervised learning techniques. In contrast to the current method of training models with only complete information, we examine the effects of training our classification model with both complete and incomplete network information. This technique enables our model to learn how to recognize and classify other incomplete networks. The full results are reported in LT Vu's thesis (Vu, 2019).

The representation of incomplete networks at various stages of completeness allows the machine to examine the nuances of incomplete networks. By allowing the machine to study incomplete networks, its ability to recognize and classify other incomplete networks improves drastically. Our method requires minimal computational effort and can accomplish an efficient classification. The results strongly confirm the effectiveness of training a classification model with incomplete network information.

In LT Vu's thesis, we found that if we train machines with not only complete networks but also incomplete networks the accuracy rates increased dramatically even with real life networks. Especially that our method of training with both complete and incomplete information achieves improves classification rates at all stages of network incompleteness. The foundation established in LT Vu's work allows for an enhanced understanding of incomplete networks (Vu, 2019). Opportunities for follow-on research extend to incorporation of this classification model into practical implementation and exploration of other ML techniques. In this project, we are taking the next step. Based on LT Vu's work, we can classify what kind of the model we should consider for fitting the observed network. However, we need to know how much information the observed network is missing.

For the second part of this project, the network completion problem, the full results are reported in LT Spinelli's thesis (2020). Essentially, LT Spinelli worked on developing a novel ML model to accurately infer how much information the observed network is missing. Here, we used simulated data sets generated by the following models: (1) Erdos-Renyi (ER), random graph, nodes and probability of edge connection; (2) Barabasi-Albert (BA), Scale-free network (Follows power law), preferential node attachment; (3) Random Geometric (RG), patial network, uses radius to determine edges, and; (4) Small World (SW), similar to Erdos-Renyi, but this model is used for analyzing social networks.

Then we set up the two scenarios: (1) heterogeneous graphs: the heterogeneous graphs contain graph calculated characteristics for varying parameters of the specific graph type. For each dataset (RG, SW, BA, and ER) we change two inputs, the amount of missing information and the parameter for the specific graph; and (2) homogeneous graphs: the homogeneous graph data sets contain calculated graph characteristics for a single given set of parameters. The only changing difference between each of the graphs is the percent of data missing as the ML methods are able to perform better with less changing parameters. These graphs are simpler, and will not be required to distinguish between percent missing and a graph with varying parameters.

We ran experiments with two scenarios: (1) missing nodes and edges associate with these nodes and (2) missing only edges of a network. For both scenarios, we can accurately infer how much information is missing with more than 98% accuracy from simulation studies if we use RF and Adaboost as classifiers.

Presently, the application of this classification method is being explored for a Department of Defense Unmanned Autonomous Vehicle (UAV) network control project in a joint effort between the Operations Analysis, Mechanical and Aerospace Engineering, and Computer Science Departments at the Naval Postgraduate School.

Recommendations for Further Research

This research establishes a foundation for the continued study of incomplete networks; efforts to incorporate this classification model in real applications is necessary for testing the model's practical implementation. Future research efforts should also allow for an enhanced understanding of incomplete networks, and how to classify them, including methods of accurately classifying sub-portions of a too-large network. Also, the networks we examine are limited to static observed networks, therefore, future efforts should include dynamic networks. Additionally, while our research confirms the advantages of a standard supervised learning method in classifying incomplete networks, a deep learning approach ought to be considered to harness its capability and flexibility to process larger amounts of raw data through its incremental layered learning (LeCun et al., 2015).

References

Canning, J.P, Ingram, E.E., Nowak-Wolff, S., M. Ortiz, A., Ahmed, N.K., Rossi, R.A., Schmitt, K.R.B., & Soundarajan, S. (2018). *Predicting graph categories from structural properties*. http://arxiv.org/abs/1805.02682.

LeCun, Y., Bengio, Y., & Hinton, G. (2015.) Deep learning. *Nature*, *521*(7553), 436–444,

Li, G., Semerci, M., Yener, B., & Zaki, M. J. (2012.) Effective graph classification based ontopological and label attributes. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *5*(4), 265–283.

Spinelli, R. (2020). *Inference on missing information in a social network*. [Master's thesis, submitted for publication]. https://calhoun.nps.edu

Vu, C. (2019.) *A method for classification of incomplete networks: training the model with complete and incomplete information.* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://calhoun.nps.edu/handle/10945/62313

Acronyms

Barabasi-Albert	BA
Erdos-Renyi	ER
Machine Learning	ML
Random Geometric	RG
Random Forest	RF
Small World	SW
unmanned autonomous vehicle	UAV