Time-variant Graph Learning and Classification

Haishuai Wang

Faculty of Engineering and Information Technology

University of Technology, Sydney

A thesis submitted for the degree of

Doctor of Philosophy

30 October 2016

This thesis is dedicated to my loving parents

CERTIFICATE OF ORIGINAL AUTHORSHIP

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Student: Haishuai Wang
Date: 30.10.2016

Acknowledgements

I benefited and learned a lot from my supervisors, my colleagues, and my friends during the PhD study at University of Technology, Sydney, Australia. I wish to take this opportunity to thank all of them.

Firstly, I would like to express my sincere gratitude to my advisors, Dr. Ling Chen, Dr. Peng Zhang and Prof. Xingquan Zhu, for the continuous support of my Ph.D study and related research, for their patience, motivation, and immense knowledge. Dr. Ling Chen has been supportive and has given me the freedom to pursue various projects without objection. She has also provided insightful discussions about the research. I am deeply indebted to Dr. Peng Zhang for his fundamental role in my doctoral work. Dr. Zhang provided me with every form of guidance, assistance, and expertise that I needed during my Ph.D study. In addition to our academic collaboration, I greatly value the close personal rapport that Dr. Zhang and I have forged over the years. I am also very grateful to Professor Xingquan Zhu for his scientific advice and knowledge and many insightful discussions and suggestions. Their friendship has also been important to me as they have often given me invaluable advice in a personal sense.

I would like to thank Professor Yixin Chen who has provided me with a great opportunity to visit Washington University in St. Louis and has given me a postdoctoral position. I am extremely grateful for his guidance and all the excellent discussions that I have had with him. His deep insights have helped me at various stages of my research. I also give thanks to Professor Chengqi Zhang, Professor Ivor W Tsang, Professor Huan Liu, Professor Xindong Wu for their very helpful comments and suggestions which were aimed at improving my research skills.

I would also like to take this opportunity to thank all my friends in the Quantum Computation & Intelligent Systems Centre at UTS for all the great times that we have shared, in particular, Jia Wu, Shirui Pan, Shaoli Huang, Bo Han, Qin Zhang, Sujuan Hou, Lianhua Chi, Chunyang Liu, Guodong Long, Bozhong Liu, Anjin Liu, Yu Bai, and Tongliang Liu. They are the ones who have given me support during both joyful and stressful times, and to whom I will always be thankful. I am also grateful to Jemima Moore for proof reading my submission drafts.

Finally, I am deeply thankful to my parents and sisters for their endless love, encouragement, support, and various sacrifices. Without them, this thesis would never have been written. I dedicate this thesis to them.

Abstract

Graph classification is an important tool for analyzing data with structure dependency. In traditional graph classification, graphs are assumed to be independent where each graph represents an object. In a dynamic world, it is very often the case that the underlying object continuously evolves over time. The change of node content and/or network structure, with respect to the temporal order, presents a new time-variant graph representation, where an object corresponds to a set of time-variant graphs (TVG). A time-variant graph can be used to characterize the changing nature of the structured object, including the node attribute and graph topological changing over time. Therefore, the evolution of time-variant graphs could be either network structure or node content over time. In this dissertation, we formulate a new time-variant graph learning and classification (TVGLC) task.

To learn and classify time-variant graphs, the vital steps are feature extraction, modeling and algorithm design. However, for time-variant graph classification, frequent subgraph features are very difficult to obtain. Because one has to consider the graph structure space and the temporal correlations to find subgraph candidates for validation, the search space for finding frequent subgraph features is infinite and unlikely to obtain stable structures. Secondly, graph structures that imply subgraph features may irregularly change over time. Thus, to extract effective and efficient features is a great challenge for TVGLC. In addition, carrying out applicable models and algorithms to cater for the extracted features for TVGLC is also a challenge.

Considering the above challenges, this research aims to extract efficient features and design new algorithms to enable the learning of the

time-variant graph. Because time variant graphs may involve changes in the network structures and changes in the node content, which complicate the algorithm designs and solutions, our research employs a divide and conquer principle to first solve a simplified case where (1) network topology is fixed whereas the node content continuously evolves (i.e., networked time series classification). After that, we advance to the setting to (2) evolving network structure and propose solutions to TVGLC with incremental subgraph features. To enhance the subgraph feature exploration for time variant graph classification, we propose (3) graph-shapelet features for TVGLC. Last, but not the least, we study (4) an application of online diffusion provenance detection.

Temporal Feature Selection on Networked Time Series: As the time-variant graph can be graph node content and/or graph structure evolution, we first study a simple case where the structure is fixed but the node content continuously evolves. The problem forms time series data when the node content changes over time, and we combine time series data with a static graph to form a new problem called networked time series. We formulate the problem of learning discriminative features (*i.e.,* segments) from networked time series data considering the linked information among time series (*e.g.,* social users are taken as social sensors that continuously generate social signals (tweets) represented as time series). The discriminative segments are often referred to as *shapelets* of time series. Extracting shapelets for time series classification has been widely studied. However, existing works on shapelet selection assumes that time series are independent and identically distributed (i.i.d.). This assumption restricts their applications to social networked time series analysis. This thesis proposes a new Network Regularized Least Squares (NetRLS) feature selection model, which combines typical time series data and user network graph data for analysis.

Incremental Subgraph based TVGLC: To learn and classify the

time-variant graph with network structure evolve, the key challenges are to extract features and build models. To date, subgraphs are often used as features for graph learning. In reality, the dimension of the subgraphs has a crucial dependency on the threshold setting of the frequency support parameter, and the number may become extremely large. As a result, subgraphs may be incrementally discovered to form a feature stream and require the underlying graph classifier to effectively discover representative subgraph features from the subgraph feature stream. Moreover, we propose a *primal-dual incremental subgraph feature selection* algorithm (*ISF*) based on a max-margin graph classifier. The ISF algorithm constructs a sequence of solutions that are both primal and dual feasible. Each primal-dual pair shrinks the dual gap and renders a better solution for the optimal subgraph feature set. To avoid the bias of the ISF algorithm on short-pattern subgraph features, we present a new *incremental subgraph join feature selection* algorithm (*ISJF*) by forcing graph classifiers to join short-pattern subgraphs and generate long-pattern subgraph features.

Graph-shapelet based TVGLC: As graph structure continuously evolves over time, the search space for finding frequent subgraph features is infinite and unlikely to obtain stable structures. To tackle this challenge, we formulate a new time-variant graph classification task, and propose a new graph feature, *graph-shapelets*, for learning and classifying time-variant graphs. Graph-shapelet is compact and discriminative *graph transformation subsequences*. A graph-shapelet can be regarded as a graphical extension of *shapelets* – a class of discriminative features designed for vectorial temporal data classification. In order to discover graph-shapelets, we propose to convert a time-variant graph sequence as time-series data, and use shapelets discovered from the time-series data to find *graph transformation subsequences* as graph-shapelets. By converting each graph-shapelet as a unique tokenized graph transformation sequence, we can use the editing distance to calculate the distance between two graph-shapelets for time-variant graph classification.

Application of Online Diffusion Provenance Detection: In social network analysis, the information propagation graph (i.e., cascade) is a kind of time-variant graph because the information diffusion forms a graph at a certain time and the graph evolves over time. An important application of information diffusion networks (i.e., time-variant graph) is provenances detection. Existing work on network diffusion provenance identification focuses on offline learning where data collected from network detectors are static and a snapshot of the network is available before learning. However, an offline learning model does not meet the needs of early warning, real-time awareness and real-time response to malicious information spreading in networks. In this part, we study a new problem of online discovering diffusion provenances in large networks. To this end, we propose an online regression model for real-time diffusion provenance identification. Specifically, we first use offline collected network cascades to infer the edge transmission weights, and then use an online l_1 nonconvex regression model as the identification model. The proposed methods are empirically evaluated on both synthetic and real-world networks.

Experiments on synthetic and real-world data validate and demonstrate the effectiveness of the proposed methods for time-variant graph learning and classification.

Contents

List of Figures

LIST OF FIGURES

List of Tables