Real-Time Analytics for Complex Structure Data

Ting Guo

A Thesis submitted for the degree of Doctor of Philosophy

Faculty of Engineering and Information Technology University of Technology, Sydney 2015

Certificate of Authorship and Originality

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.

Production Note:
Signature removed prior to publication.

Signature of Student:

Date:

Acknowledgments

On having completed this thesis, I am especially thankful to my supervisor Prof. Chengqi Zhang and co-supervisor Prof. Xingquan Zhu, who had led me to an at one time unfamiliar area of academic research, and trusted me and given me as much as possible freedom to purse my own research interests. Prof. Zhu has taught me how to think and study independently and how to solve a difficult scientific problem in flexible but rigorous ways. He has sacrificed much of his precious time for developing my academic research skills. When I felt lost and terrified with my future, he always gave me the confidence and motivation to keep going and strive to get better. Prof. Zhang has also given me great help and support in life.

I am thankful to the group members I met in the University of Technology, Sydney, including Shirui Pan, Lianhua Chi, Jia Wu, and many others. I learned a lot from these smart people, and I was always inspired by the interesting and in-depth discussions with them. I enjoyed the wonderful atmosphere, being with them, of both academic research and daily life.

I am incredibly grateful to my mother and father for their generosity and encouragement. This thesis is definitely impossible to be completed without their constant support and understanding. I am also thankful to my friends who have companied me, though not always at my side, through the arduous journey of three years.

Contents

List of Figures

List of Tables

Abstract

The advancement of data acquisition and analysis technology has resulted in many real-world data being dynamic and containing rich content and structured information. More specifically, with the fast development of information technology, many current real-world data are always featured with dynamic changes, such as new instances, new nodes and edges, and modifications to the node content. Different from traditional data, which are represented as feature vectors, data with complex relationships are often represented as graphs to denote the content of the data entries and their structural relationships, where instances (nodes) are not only characterized by the content but are also subject to dependency relationships. Plus, real-time availability is one of outstanding features of today's data. Real-time analytics is dynamic analysis and reporting based on data entered into a system before the actual time of use. Real-time analytics emphasizes on deriving immediate knowledge from dynamic data sources, such as data streams, and knowledge discovery and pattern mining are facing complex, dynamic data sources. However, how to combine structure information and node content information for accurate and real-time data mining is still a big challenge. Accordingly, this thesis focuses on real-time analytics for complex structure data. We explore instance correlation in complex structure data and utilises it to make mining tasks more accurate and applicable. To be specific, our objective is to combine node correlation with node content and utilize them for three different tasks, including (1) graph stream classification, (2) super-graph classification and clustering, and (3) streaming network node classification.

Understanding the role of structured patterns for graph classification: the thesis introduces existing works on data mining from an complex structured perspective. Then we propose a graph factorization-based fine-grained representation model, where the main objective is to use linear combinations of a set of discriminative cliques to represent graphs for learning. The optimization-oriented factorization approach ensures minimum information loss for graph representation, and also avoids the expensive sub-graph isomorphism validation process. Based on this idea, we propose a novel framework for fast graph stream classification.

A new structure data classification algorithm: The second method introduces a new super-graph classification and clustering problem. Due to the inherent complex structure representation, all existing graph classification methods cannot be applied to super-graph classification. In the thesis, we propose a weighted random walk kernel which calculates the similarity between two super-graphs by assessing (a) the similarity between super-nodes of the super-graphs, and (b) the common walks of the super-graphs. Our key contribution is: (1) a new super-node and super-graph structure to enrich existing graph representation for real-world applications; (2) a weighted random walk kernel considering node and structure similarities between graphs; (3) a mixed-similarity considering structured content inside super-nodes and structural dependency between super-nodes; and (4) an effective kernel-based super-graph classification method with sound theoretical basis. Empirical studies show that the proposed methods significantly outperform the state-of-the-art methods.

Real-time analytics framework for dynamic complex structure data For streaming networks, the essential challenge is to properly capture the dynamic evolution of the node content and node interactions in order to support node classification. While streaming networks are dynamically evolving, for a short temporal period, a subset of salient features are essentially tied to the network content and structures, and therefore can be used to characterize the network for classification. To achieve this goal, we propose to carry out streaming network feature selection (SNF) from the network, and use selected features as gauge to classify unlabeled nodes. A Laplacian based quality criterion is proposed to guide the node classification, where the Laplacian matrix is generated based on node labels and network topology structures. Node classification is achieved by finding the class label that results in the minimal gauging value with respect to the selected features. By frequently updating the features selected from the network, node classification can quickly adapt to the changes in the network for maximal performance gain. Experiments and comparisons on real-world networks demonstrate that SNOC is able to capture dynamics in the network structures and node content, and outperforms baseline approaches with significant performance gain.