

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.

Signature of Student: Production Note:
Signature removed prior to publication.

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 pursue 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

1	Introduction	19
1.1	MOTIVATION	19
1.2	CONTRIBUTIONS	26
1.3	PUBLICATIONS	28
1.4	THESIS STRUCTURE	29
2	Literature Review	31
2.1	PRELIMINARY	31
2.2	GRPAH CLASSIFICATION	32
2.3	FREQUENT SUB-GRAPH MINING (FSM)	34
2.4	SUB-GRPAH FEATURE SELECTION	35
2.5	DATA STREAM MINING	36
2.6	REAL-TIME ANALYSIS	37
2.7	ROADMAP	38
3	Understanding the Roles of Sub-graph Features for Graph Classification: An Empirical Study Perspective	39
3.1	INTRODUCTION	39
3.2	PROBLEM FORMULATION	43
3.2.1	Graph and Sub-graph	43
3.2.2	Frequent Sub-graph Mining	44
3.2.3	Graph Classification	45
3.3	EXPERIMENTAL STUDY	45

3.3.1	Benchmark Data	46
3.3.2	Sub-graph Features	48
3.3.3	Experimental Settings	50
3.3.4	Results and Analysis	50
4	Graph Hashing and Factorization for Fast Graph Stream Classification	59
4.1	INTRODUCTION	59
4.2	PROBLEM DEFINITION	62
4.3	GRAPH FACTORIZATION	63
4.3.1	Factorization Model	63
4.3.2	Learning Algorithm	66
4.4	FAST GRAPH STREAM CLASSIFICATION	68
4.4.1	Overall Framework	68
4.4.2	Graph Clique Mining	70
4.4.3	Clique Set Matrix and Graph Factorization	72
	Discriminative Frequent Cliques	72
	Feature Mapping	74
4.4.4	Graph Stream Classification	75
4.5	EXPERIMENTS	75
4.5.1	Benchmark Data	75
4.5.2	Experimental Settings	78
4.5.3	Experimental Results	79
	Graph Steams Classification Accuracy	79
	Graph Steam Classification Efficiency	83
5	Super-graph based Classification	85
5.1	INTRODUCTION	85
5.2	PROBLEM DEFINITION	88
5.3	OVERALL FRAMEWORK	89
5.4	WEIGHTED RANDOM WALK KERNEL	89
5.4.1	Kernel on Single-attribute Graphs	91

5.4.2	Kernel on Super-Graphs	94
5.5	SUPER-GRAPH CLASSIFICATION	96
5.6	THEORETICAL STUDY	96
5.7	EXPERIMENTS AND ANALYSIS	98
5.7.1	Benchmark Data	98
5.7.2	Experimental Settings	100
5.7.3	Results and Analysis	101
6	Streaming Network Node Classification	105
6.1	INTRODUCTION	105
6.2	PROBLEM DEFINITION AND FRAMEWORK	109
6.3	THE PROPOSED METHOD	110
6.3.1	Streaming Network Feature Selection	113
	Feature Selection on a Static Network	113
	Feature Selection on Streaming Networks	117
6.3.2	Node Classification on Streaming Networks	121
6.4	EXPERIMENTS	123
6.4.1	Experimental Settings	123
6.4.2	Performance on Static Networks	125
6.4.3	Performance on Streaming Networks	128
6.4.4	Case Study	130
7	Conclusion	133
7.1	SUMMARY OF THIS THESIS	133
7.2	FUTURE WORK	135

List of Figures

2-1	Graph examples collected from different domains.	33
2-2	The overall roadmap of this thesis.	38
3-1	An example of sub-graph pattern representation. Left panel shows two graphs, G_1 and G_2 and right panel gives the two indicator vectors showing whether a sub-graph exists in the graphs.	40
3-2	The runtime of frequent sub-graph pattern mining with respect to the increasing number of edges of sub-graphs.	42
3-3	A conceptual view of graph <i>vs.</i> sub-graph. (b) is a sub-graph of (a).	44
3-4	An example of graph isomorphism.	45
3-5	Graph representation for a paper (ID17890) in DBLP. Node in red is the main paper. Nodes in black ellipse are citations. While nodes in black box are keywords.	48
3-6	Classification accuracy on five NCI chemical compound datasets with respect to different sizes of sub-graph features (using Support Vector Machines: Lib-SVM).	52
3-7	Classification accuracy on D&D protein dataset and DBLP citation dataset with respect to different sizes of sub-graph features (using Support Vector Machines: Lib-SVM).	53
3-8	Classification accuracy on one NCI chemical compound dataset, D&D protein dataset, and DBLP citation dataset with respect to different sizes of sub-graph features (using Nearest Neighbours: NN).	54
4-1	Coarse-grained <i>vs.</i> fine-grained representation.	60

4-2	An example of graph factorization.	64
4-3	The framework of <i>FGSC</i> for graph stream classification.	69
4-4	An example of clique mining in a compressed graph.	71
4-5	An example of “in-memory” Clique-class table Γ	74
4-6	Graph representation for a paper (ID17890) in DBLP.	77
4-7	Accuracy w.r.t different chunk sizes on DBLP Stream . The number of features in each chunk is 142. The batch sizes vary as: (a) 1000; (b) 800; (c) 600.	80
4-8	Accuracy w.r.t different number of features on DBLP Stream with each chunk containing 1000 graphs. The number of features selected in each chunk is: (a) 307; (b) 142; (c) 62.	80
4-9	Accuracy w.r.t different classification methods on DBLP Stream with each chunk containing 1000 graphs, and the number of features in each chunk is 142. The classification methods selected here are: (a) NN; (b) SMO; (c) NaiveBayes.	80
4-10	Accuracy w.r.t different chunk sizes on IBM Stream . The number of features in each chunk is 75. The batch sizes vary from (a) 500; (b) 400; to (c) 300.	81
4-11	Accuracy w.r.t different number of features on IBM Stream with each chunk containing 400 graphs. The number of features selected in each chunk is: (a) 148; (b) 75; (c) 43.	81
4-12	Accuracy w.r.t different classification methods on IBM Stream with each chunk containing 400 graphs, and the number of features in each chunk is 75. The classification methods selected include: (a) NN; (b) SMO; (c) NaiveBayes.	81
4-13	System accumulated runtime-based by using NN classifier, where $ D = 1000$, $ m = 142$ (for DBLP) and $ D = 400$, $ m = 75$ (for IBM) respectively. (a) Results on DBLP stream; (b) Results on IBM stream.	83
5-1	(A): a single-attribute graph; (B): an attributed graph; and (C): a super-graph.	86

5-2	A conceptual view of a protein interaction network using super-graph representation.	87
5-3	$WRWK$ on the super-graphs (G, G') and the single-attribute graphs (g_1, g_2, g_3)	90
5-4	An example of using super-graph representation for scientific publications.	99
5-5	Super-graph and comparison graph representations.	100
5-6	Classification accuracy on DBLP and Beer Review datasets <i>w.r.t.</i> different classification methods (NB, DT, SVM, and NN).	102
5-7	Classification accuracy on Beer Review dataset <i>w.r.t.</i> different datasets and classification methods (NB, DT, SVM, and NN).	103
5-8	The performance <i>w.r.t.</i> different edge-cutting thresholds on DBLP and Beer Review datasets by using $WRWK$ method.	104
6-1	An example of streaming networks, where each color bar denotes a feature.	106
6-2	An example of using feature selection to capture changes in a streaming network (keywords inside each node denote node content).	108
6-3	The framework of the proposed streaming network node classification (SNOC) method.	111
6-4	An example of using feature selection to capture structure similarity.	115
6-5	The accuracies on three real-world static networks <i>w.r.t.</i> different numbers of selected features (from 50 to 300).	125
6-6	The accuracy on three networks <i>w.r.t.</i> (a) different maximal lengths of path l (from 1 to 5), (b) different values of weight parameter ξ (from 0 to 1), and (c) different percentages of labeled nodes.	126
6-7	The accuracy on streaming networks: (a) accuracy on DBLP citation network from 1991 to 2010, (b) accuracy on PubMed Diabetes network for 15 time points, and (c) accuracy on extended DBLP citation network from 1991 to 2010.	128
6-8	The cumulative runtime on DBLP and PubMed Diabetes networks corresponding to Fig. 6-5.	130

6-9 Case study on DBLP citation network. 131

List of Tables

3.1	The advantages and disadvantages comparisons between vector representation vs. graph representation	46
3.2	NCI datasets used in experiments	47
3.3	DBLP dataset used in experiments	47
3.4	Number of sub-graphs with respect to different sizes (<i>i.e.</i> number of edges)	49
4.1	DBLP dataset used in experiments.	76
6.1	Accuracy Results on Static Network.	126

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 SNOG is able to capture dynamics in the network structures and node content, and outperforms baseline approaches with significant performance gain.

