Exploring Instance Correlation for Advanced Active Learning

Yifan Fu

A Thesis submitted for the degree of Doctor of Philosophy

Faculty of Engineering and Information Technology

University of Technology, Sydney 2013

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.

____

Acknowledgements

On having completed this thesis, I am especially thankful to my supervisor Prof. Xingquan Zhu and co-supervisor Prof. Chengqi Zhang, 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. 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 Bin Li, Shirui Pan, Guodong Long, Lianyang Ma, 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 for her generosity and encouragement. This thesis is definitely impossible to be completed without her 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 and a half years.

Table of Contents

Table of Contents iii

Abstract ix

1 Introduction 1

1.1 Motivation 1

1.2 Problem Statement and Solutions 6

1.3 Contributions 8

1.4 Publications 11

1.5 Thesis Structure 11

2 Preliminary and Notations 13

2.1 Definitions 13

2.2 Notations 17

3 Literature Review 20

3.1 Instance Correlation 20

3.2 Categorization for active learning methods 22

3.3 Active learning based on IID instance information 23

3.3.1 How to Select Unlabeled Instances for Labeling 24

3.3.2 How to Evaluate Selected Unlabeled Instances 24

3.4 Active Learning Based on Instance Correlations 25

3.4.1 How to Select Unlabeled Instances 25

3.4.1.1 Exploration on Feature Correlation 25

3.4.1.2 Exploration on Label Correlation 30

3.4.1.3 Exploration on Feature and Label Correlation . . . 33

3.4.1.4 Exploration on Structural Correlation 33

3.4.2 How to Evaluate Selected Unlabeled Instances 35

3.5 Algorithm Performance Comparison 36

3.5.1 Time Complexity Analysis 36

3.5.2 Label Complexity Analysis 39

3.5.3 Lessons Learned 40

3.5.3.1 Lessons from IID Based Active Learning 40

3.5.3.2 Lessons from Instance Correlation Based Active Learning 41

3.6 Emerging Applications: Challenges and Trends 43

3.6.1 Active Learning on Streaming Data Platform 43

3.6.2 Active Learning with Complex Data Presentations 44

3.6.3 Active Learning with Crowdsourcing Labelers 45
3.6.4 Active Learning for Domain Adaptations 47
3.7 Summary 48

4 Active Learning with Optimal Instance Subset Selection 49
4.1 Introduction 49
4.2 Problem Definition & System Overview 51
4.2.1 Problem Definition 51
4.2.2 System Overview 51
4.3 ALOSS: Optimal Instance Subset Selection for AL 53
4.3.1 Correlation Matrix Construction 53
4.3.2 Optimal Subset Selection 56
4.3.3 ALOSS: System Framework 58
4.3.4 Time Complexity Analysis 60
4.4 Experiments 61
4.4.1 Benchmark Methods 62
4.4.2 Experimental Results with Different Batch Sizes 64
4.4.3 Comparisons between Different Distance Measures 68
4.4.4 Results with Different Percent of Labeled Samples 68
4.4.5 Detailed Comparisons for All Methods 69

5 Pairwise Homogeneity Based Active Learning 71
5.1 Introduction 71
5.2 Problem Formulation 75
5.2.1 Problem Setting 75
5.2.2 Method Overview 76
5.3 The Proposed Method 78
5.3.1 Graph Ensemble Construction 78
5.3.2 Weight Adjustment in Min-cut Sets 80
5.3.3 Confidence based Data Selection 82
5.3.4 Weight Update in Selected Subset 83
5.3.5 Time Complexity Analysis 83
5.4 Experiments 85
5.4.1 Data Description and Experimental Settings 85
5.4.2 Baseline Methods 86
5.4.3 Parameter Setting for k-NN Graphs 88
5.4.4 Sensitivity w.r.t. Different Percentages of Labeling Noise . . 90
5.4.5 Comparison of Different Pair Selection Strategies 91
5.4.6 Detailed Comparison of All Methods 93

6 Active Learning with Unknown Class Space and Weak Labeling Knowledge 96
6.1 Introduction 96
6.2 Problem Formulation 100
6.3 CSAL: Cold-Start Active Learning 100
6.3.1 Class Discovery 101
6.3.1.1 Representative Instance Subset Selection 101
6.3.1.2 MST Based Class Discovery 103
6.3.2 Class Exploration and Instance Labeling 104

6.3.2.1 Consensus Ensemble Learning 105
6.3.2.2 Adaptive Query Strategy 110
6.3.3 Time Complexity Analysis 111
6.4 Experiments 114
6.4.1 Baseline Methods. 114
6.4.2 Accuracy Gains with Different Ensemble Methods 116
6.4.3 Class Discovery using Different Query Strategies 116
6.4.4 Class Exploration Comparisons 119
6.4.5 Experiment Results with Dynamic Data 119
6.4.6 Detailed Comparisons for All Methods 122

7 Conclusions 126
7.1 Summary of This Thesis 126
7.2 Future Work 128

Bibliography 129

List of Figures

1.1 The general learning process of Active Learning [98] 3

3.1 Hierarchical structure of the categorization for active learning . . . 23

3.2 Relationships between the algorithms used in exploration on feature correlation and evaluation granularity 28

3.3 Active inference using Reflect and Correct Method 35

3.4 Query strategy comparison on representative algorithms from the instance correlation and time complexity perspectives 38

4.1 A toy example to demonstrate the tradeoff between uncertainty and diversity for sample selection in active learning 50

4.2 Accuracy comparisons with different batch sizes for active learning . 65

4.3 Head-to-Head comparisons between prediction distance and feature distance based instance disparity measures 66

4.4 Accuracy comparisons *w.r.t.* different portions of labeled samples . 66

4.5 Accuracy comparisons *w.r.t.* different portions of labeled samples, the batch size is fixed to 0.05 67

5.1 An example of using pairwise label homogeneity to ease a labeling task 73

5.2 Traditional active learning paradigm *vs.* the proposed PHAL paradigm 74

5.3 The proposed Pairwise Homogeneity based Active Learning (PHAL) framework. 76

5.4 The accuracy comparison between QHAL and PHAL 89

5.5 The sensitivity of different methods to labeling noise 91

5.6 Performance comparison of PHAL and the baseline methods with different pair selection strategies on 10 data sets. 95

6.1 A conceptual view demonstrating the difference between Hot-start *vs.* Cold-Start active learning 99

6.2 The overall framework of the proposed cold-start active learning . . 101

6.3 The class discovery process 102

6.4 The two steps in class exploration and instance labeling. 105

6.5 Illustration of the brief graph. 108

6.6 Accuracy comparisons with different ensemble methods 117

6.7 query number comparison with different query strategies 118

6.8 Iteration comparison with regard to Class Discovery 120

6.9 Iteration comparisons with regard to class exploration in a dynamic environment 123

6.10 Illustration of the brief graph. 124

List of Tables

2.1 Notations used in the paper 17

3.1 Different settings of active learning based on instance correlation . . 26

3.2 Summary of all reviewed instance-selection methods in terms of two dimensions: “how to select ”and “how to evaluate ”. 37

4.1 A toy example demonstrating classifier matrix construction 54

4.2 A simple description of the benchmark data 62

4.3 Detailed algorithm performance comparisons (using Decision Trees as the benchmark learner and labeling percentage is 15%) 70

4.4 Detailed algorithm performance comparisons (using Decision Trees as the benchmark learner and labeling percentage is 30%) 70

4.5 Detailed algorithm performance comparisons (using Decision Trees as the benchmark learner and labeling percentage is 50%) 70

5.1 Description of the benchmark data sets. 86

5.2 Detailed performance comparison 94

6.1 A simple description of the benchmark data 114

6.2 accuracy comparisons with different methods 122

6.3 Detailed algorithm performance comparisons (iteration number is 150)124

Abstract

Active learning (AL) aims to construct an accurate classifier with the minimum labeling cost by actively selecting a few number of most informative instances for labeling. AL traditionally relies on some instance-based utility measures to assess individual instances and label the ones with the maximum values for training. However, such approaches cannot produce good labeling subsets. Because instances exist some explicit / implicit relations between each other, instance-based utility measure evaluates instance informativeness independently without considering their interactions. Accordingly, this thesis explores instance correlation in AL and utilizes it to make AL's more accurate and applicable. To be specific, our objective is to explore instance correlation from different views and utilize them for three different tasks, including (1) reduce redundancy for optimal subset selection, (2) reduce labeling cost with a nonexpert labeler and (3) discover class spaces for dynamic data.

First of all, the thesis introduces existing works on active learning from an instance-correlation perspective. Then it summarizes their technical strengths/ weaknesses, followed by runtime and label complexity analysis, discussion about emerging active learning applications and instance-selection challenges therein.

Secondly, we propose three AL paradigms by integrating different instance correlations into three major issues of AL, respectively. 1) The first method is an optimal instance subset selection method (ALOSS), where an expert is employed to provide accurate class labels for the queried data. Due to instance-based utility measures assess individual instances and label the ones with the maximum values, this may result in the redundancy issue in the selected subset. To address this issue, ALOSS simultaneously considers the importance of individual instances and the disparity between instances for subset selection. 2) The second method introduces pairwise label homogeneity in AL setting, in which a non-expert labeler is only asked “whether a pair of instances belong to the same class ”. We explore label homogeneity information by using a non-expert labeler, aiming to further reducing the labeling cost of AL. 3) The last active learning method also utilizes pairwise label homogeneity for active class discovery and exploration in dynamic data, where some new classes may rapidly emerge and evolve, thereby making the labeler incapable of labeling the instances due to limited knowledge. Accordingly, we utilize pairwise label homogeneity information to uncover the hidden class spaces and find new classes timely. Empirical studies show that the proposed methods significantly outperform the state-of-the-art AL methods.