

DOCTORAL THESIS

A model-based approach for distributed data mining

Zhang, Xiaofeng

Date of Award:
2007

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

A Model-based Approach for Distributed Data Mining

ZHANG Xiaofeng

A thesis submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Principal Supervisor: Dr. William K. CHEUNG

Hong Kong Baptist University

December 2007

Abstract

Most data mining algorithms assume that data have been pooled together in a centralized repository so that analysis can be performed. Recently, there exist a number of cases where data are distributed and cannot be shared due to local constraints, such as privacy concerns or bandwidth limits. In this thesis, we focus on studying how a model-based approach can be applied to data mining in a distributed environment.

First, we demonstrate how a model-based approach can be applied to the web data clustering and visualization. In particular, we extend the latent class model (LCM) by modeling also the topological relationship of the latent classes and study how distributed learning of the LCM can be performed via merging local LCMs.

As a major contribution of this thesis, a distributed model-based data mining approach called *learning from abstraction* is proposed. At each source, it first computes local data abstraction using hierarchical clustering algorithms and then aggregates the local abstractions for global analysis. Gaussian mixture model is adopted as the representation of local data abstractions. Gaussian mixture model and generative topographic mapping are the global models we study for two applications — distributed data clustering and distributed manifold discovery respectively. An EM-like algorithm is derived for learning both global models solely based on the model parameters of the local abstractions. We tested the proposed approach using different scenarios regarding the size of the data sets and the distribution of the data over the different data sources. A number of synthetic and benchmark data sets are used to validate the proposed approach. Experimental results have shown that accurate global models can still be learned from properly abstracted data (privacy protected)

and the proposed approach is much more efficient (scalable) when compared with the model learned directly from the raw data. Also, its performance is found to be robust against heterogeneous data distributions among the local data sources.

While the proposed learning-from-abstraction approach is effective for distributed model-based data mining, how to obtain the right trade-off between the abstraction levels of the local data sources and the global model accuracy remains open. It is challenging because the local data sets could be inter-correlated to different extents. Therefore, the best abstraction strategy for a data source depends on how the other sources set their abstraction levels. We formulate this optimal abstraction task as a game and compute the Nash equilibrium as its solution. In addition, we investigate an iterative version of the game so that the Nash equilibrium can be computed by actively exploring the right level of details from the local sources in a need-to-know manner. In other words, based on the game theoretical approach, the local sources can self-organize to determine their own optimal granularity levels of abstraction so as to protect local data privacy at best and yet to acquire a good global model accuracy as far as possible.

Future research directions include (1) studying alternative data privacy measures, (2) extending the proposed approach to a peer-to-peer computing environment, (3) performing the theoretical study of the optimality of the proposed iterative game, (4) optimizing the local data abstraction, and (5) studying how the game theoretic based distributed data mining approach can be further enhanced for an untrusted and more dynamic environment.

Keywords: Model-based approach, clustering, manifold discovery, privacy preserving data mining, distributed data mining

Table of Contents

Declaration	i
Abstract	ii
Acknowledgements	iv
Table of Contents	v
List of Figures	x
List of Tables	xiii
Chapter 1 Introduction	1
1.1 Data Mining: Model-Based vs. Data Driven	1
1.2 Distributed Data Mining	2
1.3 Privacy Preserving Data Mining	3
1.4 Thesis Overview	5
1.4.1 Extending Latent Class Model for Web Mining	5
1.4.2 A Model-based Approach for Distributed Privacy Preserving Data Mining	6
1.4.3 A Game Theoretic Approach to Active Distributed Data Mining	6
1.5 Outline of the Thesis	7
Chapter 2 Model-based Web Content and Hyperlink Analysis	8
2.1 Introduction	8

2.2	Web Page and Structure Analysis using Latent Class Model	10
2.2.1	Latent Class Model	11
2.2.2	Distributed Learning of LCM	12
2.2.3	Local Learning	13
2.2.4	Model Merging	13
2.2.5	Model Exchange Scheme with Additional Privacy Constraints	15
2.2.6	Experiments	15
2.2.7	Web Pages Preprocessing	16
2.2.8	Experimental Results	16
2.3	Web Manifold Visualization using Generative Topographic Mapping .	17
2.3.1	Extending GTM for Modeling Web Contents and HyperLinks	18
2.3.2	Experiments	21
2.4	Distributed Learning of LCM with Multiple Model Exchange	25
2.4.1	Communication Overhead and Computational Complexity . .	26
2.4.2	Experiment Setups for Different Model Exchange Schemes . .	28
2.4.3	Performance Comparison	29
2.5	Summary	32

Chapter 3 A Model-based Approach for Distributed Privacy Preserving Data Mining 33

3.1	Introduction	33
3.1.1	Model-Based Data Abstraction	34
3.1.2	Mining Abstracted Data	35
3.2	Learning from Abstraction	35
3.2.1	Generic Problem Formulation	36
3.2.2	Local Data Abstraction	37
3.2.3	An EM Algorithm for Mining Abstracted Data	40
3.3	Learning Global Models from Abstractions	41
3.3.1	Learning GMM for Clustering	41
3.3.2	Learning GMM from Data	41

3.3.3	Learning GMM from Abstraction	42
3.3.4	GMM Initialization Based on Abstraction	43
3.3.5	Learning GTM for Manifold Discovery	43
3.3.6	Learning GTM from Data	44
3.3.7	Learning GTM from Abstraction	45
3.3.8	GTM Initialization Based on Local Abstraction	46
3.4	Computational Gain by Learning from Abstraction	47
3.5	GMM: Experiments on Data Clustering	48
3.5.1	Accuracy and Speedup	48
3.5.2	Performance on Local Sources with Biased Distributions	52
3.5.3	Performance on Large Data Sets	52
3.5.4	Abstractions with k -item Per-grouping	55
3.6	GTM: Experiments on Manifold Discovery	55
3.6.1	Accuracy and Speedup	57
3.6.2	Performance on Local Sources with Heterogeneous Abstraction Quality	59
3.6.3	Performance on Local Sources with Biased Distributions	63
3.7	A Graph-based Local Data Abstraction for Manifold Discovery	65
3.7.1	Introduction	65
3.7.2	Minimum Cut Clustering	66
3.7.3	Incorporation of Graph-Based Abstractions	68
3.7.4	Experiments	70
3.8	Summary	73

Chapter 4 A Game Theoretic Approach to Active Distributed Data

	Mining	76
4.1	Introduction	77
4.1.1	Data Mining in Game Theory	77
4.1.2	Unsupervised Active Learning	78
4.2	Active LFA - A Game Theoretic Approach	79

4.2.1	Game Theory in A Nut Shell	79
4.2.2	Analogy	80
4.2.3	Formulation	80
4.2.4	An Approximated Utility Function	83
4.2.5	Nash Equilibrium as Optimal Abstraction	84
4.2.6	Reaching the Nash Equilibrium Based on Need-to-know Information	85
4.3	Experiments	87
4.3.1	Results on 2-Source Problems	87
4.3.2	Results on N -Source Problems	90
4.4	Summary	97
Chapter 5 Conclusions and Future Works		99
5.1	Summary of Thesis	99
5.1.1	Extending Latent Class Model for Web Mining	99
5.1.2	A Model-based Approach for Distributed Data Mining with Privacy Preservation	100
5.1.3	A Game Theoretic Approach to Active Distributed Data Mining	100
5.2	Contributions	101
5.3	Future Work	102
5.3.1	Extensions to LFA	102
5.3.2	Optimizing Graph-based Abstraction	103
5.3.3	Active Data Mining	104
Bibliography		105
Appendices		117
A	Detailed Derivation of The Modified EM Algorithm	117
A.1	GMM as The Global Model	117
A.2	GTM as The Global Model	119

