DOCTORAL THESIS

A model-based approach for distributed data mining

Zhang, Xiaofeng

Date of Award:
2007

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