DISTRIBUTED DATA MINING: A MULTIAGENT APPROACH

by

Cuong Trung Tong

BIT, GradDip (InfSc)

A thesis submitted in fulfilment of the requirements for the degree of Master of Information Science.

Faculty of Information Sciences and Engineering
University of Canberra

June 2011

Copyright

Acknowledgement

I feel so grateful to numerous people who have helped me with their guidance, time, support, and encouragement through out this research.

First and foremost, I would like to express my sincere gratitude to my supervisory panel Professor Dharmendra Sharma and Dr Fariba Shadabi for the professional and personal guidance that go far beyond their responsibilities. It is their patient guidance, gentle encouragement and advices that led me through this journey that seemed impossible at times.

I would like to thank my research colleagues and the staff at Faculty of Information Science and Engineering, University of Canberra, especially Associate Professor Dat Tran, Dr Kim Le , Professor John Campbell and Dr Wan Li Ma for their time and supports during my time at the University.

Finally, I would like to express my deepest gratitude to my parents, for their loves, sacrifices and encouragements.

Abstract

Data mining on large datasets using a batch approach is time consuming and expensive. Training a large dataset can be time-consuming and in some cases may not be practical or even possible. In addition, batch learning introduces a single point of failure – this means that the training process may crash at any one point during the job and the whole process would need to be restarted.

This research advances the understanding of a multi-agent approach to data mining of large datasets. An agent mining model called DMMAS (Distributed Mining Multi-Agent System) is developed for the purpose of building accurate and transparent classifiers and improving the efficiency of mining a large dataset.

In our case study utilising the DMMAS model, the Pima Indian Diabetes dataset and US Census Adult dataset were used. They are well-known benchmark data from the UCI (University of California, Irvine) machine learning repository. This study found that the processing speed is improved as the result of the multi-agent mining approach, although there can be a corresponding marginal loss of accuracy. This loss of accuracy gap tends to close over time as more data becomes available.

The DMMAS approach provides a new, innovative data mining model, with great research and commercial potential for distributing mining across several agents and possibly different data sources. This research also reinforces the idea that combining multiagent and data mining approaches is a logical extension for large scale data mining applications.

Table of Contents

List of Figures

List of Tables