

**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

"Copyright in relation to this thesis Under Section 35 of the Copyright Act of 1968, the author of this thesis is the owner of any copyright subsisting in the work, even though it is unpublished.

Under section 31(I)(a)(i), copyright includes the exclusive right to 'reproduce the work in a material form'. Thus, copyright is infringed by a person who, not being the owner of the copyright, reproduces or authorises the reproduction of the work, or of more than a reasonable part of the work, in a material form, unless the reproduction is a 'fair dealing' with the work 'for the purpose of research or study' as further defined in Sections 40 and 41 of the Act.

This thesis must therefore be copied or used only under the normal conditions of scholarly fair dealing for the purposes of research, criticism or review, as outlined in the provisions of the Copyright Act 1968. In particular, no results or conclusions should be extracted from it, nor should it be copied or closely paraphrased in whole or in part without the written consent of the author. Proper written acknowledgement should be made for any assistance obtained from this thesis.

Copies of the thesis may be made by a library on behalf of another person provided the officer in charge of the library is satisfied that the copy is being made for the purposes of research or study."

## **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

Statement of Originality .....	ii
Form B.....	ii
Certificate of Authorship of Thesis .....	ii
Copyright .....	iii
Acknowledgement .....	iv
Abstract.....	v
List of Figures.....	ix
List of Tables .....	x
<b>Chapter 1.....</b>	<b>1</b>
<b>INTRODUCTION .....</b>	<b>1</b>
1.1. Background.....	1
1.2. Motivation.....	1
1.3. Objectives .....	2
1.4. Research Questions.....	2
1.5. Research Scope .....	3
1.6. Thesis Roadmap.....	3
<b>Chapter 2.....</b>	<b>4</b>
<b>DATA MINING AND MULTI-AGENTS SYSTEM: A REVIEW.....</b>	<b>4</b>
2.1. Introduction.....	4
2.2. Classification in Data Mining .....	4
2.3. Classification Algorithms .....	6
2.3.1. Decision Tree.....	7
2.3.2. SLIQ .....	8
2.3.3. SPRINT .....	9
2.3.4. CLOUDS .....	9
2.3.5. Meta Decision Tree .....	10
2.3.6. Ensemble Learning.....	10
2.3.6.1. Bagging.....	10
2.3.6.2. Boosting .....	12

2.4. Multi-agents System .....	13
2.3.7. MAS Motivation .....	14
2.3.8. Features and Capabilities .....	15
2.3.9. Multi-agents Toolkits.....	15
2.3.9.1. JADE.....	16
2.3.9.2. JACK .....	18
2.3.9.3. MASDK.....	18
2.4. Why Agent Mining? .....	18
2.4.1. Applications.....	20
2.5. Summary .....	21
<b>Chapter 3.....</b>	<b>23</b>
<b>THE RESEARCH PROBLEM AND THE PROPOSED DMMAS SOLUTION.....</b>	<b>23</b>
3.1. Introduction.....	23
3.2. The Problem.....	23
3.3. The Problem Characteristics .....	23
3.4. Analysis of the Problem.....	24
3.5. Algorithm.....	25
3.6. DMMAS Design .....	30
3.7. System Architecture.....	27
3.8. Summary.....	31
<b>Chapter 4.....</b>	<b>32</b>
<b>IMPLEMENTATION AND EXPERIMENTS.....</b>	<b>32</b>
4.1. Introduction.....	32
4.2. DMMAS Implementation .....	32
4.2.1. Platform Initialization.....	32
4.2.2. Data Source Configuration .....	36
4.2.3. Agents Training .....	39
4.2.4. Execution.....	41
4.2.5. Dataset Update.....	43
4.2.6. Classification in DMMAS .....	44

4.2.7.	Data Compression .....	45
4.2.8.	Agents Communication .....	45
4.3.	DMMAS Experiments .....	46
4.3.1.	Experiment Design .....	46
4.3.2.	Infrastructure .....	50
4.3.3.	Agent Container Setup .....	51
4.3.4.	Experiment Data .....	52
4.3.5.	Batch Mining .....	53
4.3.6.	DMMAS Algorithm .....	53
4.3.7.	Evaluation Method .....	54
4.4.	Summary .....	55
<b>Chapter 5</b>	.....	<b>56</b>
<b>RESULTS AND ANALYSIS</b>	.....	<b>56</b>
5.1.	Introduction.....	56
5.2.	Batch Mining Results.....	56
5.3.	DMMAS Results.....	61
5.4.	Comparison Analysis .....	65
5.5.	Summary .....	68
<b>Chapter 6</b>	.....	<b>69</b>
<b>CONCLUSION AND FUTURE WORK</b>	.....	<b>69</b>
	Research Limitations .....	71
	Future work.....	71
<b>BIBLIOGRAPHY</b>	.....	<b>73</b>
<b>APPENDIX A: PUBLICATION</b>	.....	<b>78</b>
<b>APPENDIX B: TYPE II DIABETES</b>	.....	<b>79</b>
<b>APPENDIX C: DATASETS</b>	.....	<b>81</b>
<b>APPENDIX D: UTILITIES FEATURES</b>	.....	<b>82</b>
<b>APPENDIX E: UML DIAGRAM</b>	.....	<b>82</b>

## List of Figures

Figure 2.1 Platform, Container and Agent Relationship (Bellifemine, et al., 2007).....	17
Figure 3.1 Out of memory thrown by Weka .....	25
Figure 3.2 System Architecture Overview .....	28
Figure 4.1 Container Relationship.....	33
Figure 4.2 Platform Setup.....	33
Figure 4.3 DMMAS Platform Ready .....	34
Figure 4.4 Non Main Container Joining.....	35
Figure 4.5 Platform is ready .....	36
Figure 4.6 Dataset Configuration .....	37
Figure 4.7 Dataset Metadata.....	38
Figure 4.8 Training.....	39
Figure 4.9 Dataset Allocation.....	41
Figure 4.10 Experiment Iteration Flowchart .....	48
Figure 4.11 DMMAS Mining Sub Process .....	49
Figure 5.1 Adult Dataset – Single Process Training Time and Testing Time.....	58
Figure 5.2 Adult Dataset – Batch Mining Accuracy .....	59
Figure 5.3 Adult Dataset – Batch Mining Overall Performance .....	59
Figure 5.4 DMMAS Training Time .....	63
Figure 5.5 Adult Dataset DMMAS Testing Time .....	63
Figure 5.6 DMMAS Accuracy .....	64



**List of Tables**

Table 2.1 Bagging Algorithm (Breiman 1996) .....	11
Table 4.1 Dataset Update Algorithm .....	43
Table 4.2: Experiment Repetition Dataset Parameters .....	47
Table 4.3 Infrastructure of computers used for the experiments .....	51
Table 5.1 Adult Dataset Single Process Experiment's Result .....	57
Table 5.2 DMMAS Results .....	59
Table 5.3 DMMAS Performance Gain .....	65