Cairo University

Faculty of Graduate Studies for Statistical Research
Department of Information Systems and Technology

Agent based system for distributed data mining

A thesis submitted in partial fulfillment of the requirements
For The M.SC Degree in Information System and Technology

Submitted by
Doaa Mostafa Hussein Ahmed

Supervised by

Prof. Dr. Hesham Ahmed Hefny
Vice-Dean for Graduate Studies
Faculty of Graduate Studies for Statistical Research
Cairo University, Egypt

Dr. Ammar Mohammed Ammar
Assistant professor at computer Science Department
Faculty of Graduate Studies for Statistical Research
Cairo University, Egypt

Dr. Maryam Mohey Eldin Hazman
Senior Researcher
Agriculture Research Center, Egypt

2021

ABSTRACT

Data mining technology has appeared to discover patterns and trends from large quantities of data. Distributed Data Mining (DDM) is emerged from the need of mining over decentralized data sources. When using a batch approach for distributed data is complex and expensive. It requires techniques to improve performance and reduce complexity in a proper way. Multi-Agent System (MAS) is one of such techniques which handle the complexity and the distribution of data by efficient way.

This research advances the understanding of a multi-agent approach to data mining of large datasets. An agent mining architecture called ADDM (Agent based Distributed Mining) is developed for the purpose of building accurate and transparent cluster and improving the efficiency of mining a large dataset by using MAS. In the proposed architecture, several agents are distributed over local servers. The novelty of our work is to select the best clustering result on each local server side and use MAS system to manage resources to impose the processing power of distributed infrastructures and reduce time and cost.

The ADDM approach provides innovative distributed data mining model, with great research and commercial prospect for distributing mining across multiple agents and different data sources. This thesis work to value greatly the idea that combining data mining and multi-agent approaches in large scale data mining applications.

TABLE OF CONTENTS

Chapter 1 Introduction 1

1.1 Problem definition 3
1.2 Research objective 3
1.3 Thesis organization 4

Chapter 2 The Distributed data mining problem: An overview 5

2.1 Introduction 5
2.2 Data mining 5
2.2.1 Knowledge discovery in databases (KDD) 7
2.2.3 Data mining techniques 8
2.3 Distributed Data Mining 9
2.3.1. Processes Steps in distributed data mining 10
2.3.1 Distributed Data Mining Techniques 11
2.4 Strategies of DDM 13
2.4.1 DDM based on Multi Agent System 13
2.4.2 DDM based on Meta learning 14
2.4.3 DDM based on grid 15
2.5 The challenges of distributed data mining 16
2.6 Advantages of using distributed data mining 17
2.7 Summary 18

Chapter 3 Distributed Clustering: Distributed k-means Algorithm 19

3.1 Introduction 19
3.2 Clustering 19
3.3 Centroid-based Clustering 21
3.4 Distributed Cluster Algorithm 22
3.5 The procecess of Distributed Cluster 23
3.6 K-means Algorithm 24
3.7 Distributed K-means Clustering Algorithm 26
3.8 Clustering Quality Metrics 27
4.8.1 External Validation Index 28
4.8.2 InternalValidation Index 29
3.9 Summary 32

Chapter 4 Distributed Data Mining and Multi Agent System 33

4.1 Introduction 33
4.2 Multi-agent System 34
4.2.1 Features of Agent Environment 35
4.2.2 MAS motivation 37

4.3 The structure of Agent based distributed data mining (ADDM) ... 37
4.4 Applications of ADDM ... 39
4.5 Comments and Limitations of Research Work on ADDM ... 40
4.6 Previous Work ... 40
4.7 Summary ... 45

Chapter 5 Proposed Model of Agent based DDM and Distributed K-means Algorithm....46

5.1 Introduction ... 46
5.2 The Proposed ADDM Architecture ... 46
5.3 The proposed algorithm of agent based distributed k-means clustering. ... 48
5.4 The Proposed combined clustering quality metrics (CCQM).. ... 50
5.5 Summary. ... 50

Chapter 6 System Design and Implementation ... 51

6.1 Introduction ... 51
6.2 Experimental Results ... 51
6.2.1 Experiment of Tamilnadu Dataset ... 51
6.2.2 Experiment of Yeast Dataset ... 58
6.2.3 Jade Platform ... 64
6.3 Summary. ... 66

Chapter 7 Conclusion and future work ... 67

7.1 Thesis contribution ... 67
7.2 Future work ... 68

References ... 69

Appendices ... 78

Appendix A: ... 78
Appendix B: ... 84

LIST OF FIGURES

Figure 2.1 Architecture of data mining system…………………………………… 6

Figure 2.2 Knowledge discovery in databases……………………………....... 7

Figure 2.3 A general Distributed Data Mining framework..................................... 10

Figure 2.4 CAKE Architecture……………………………………………....... 14

Figure 2.5 JAM Architecture……………………………................................... 15

Figure 4.1 Agent Environment ………………………………….............……. 35

Figure 4.2 Structure of Agent based distributed data mining system……………... 37

Figure 5.1 The Proposed ADDM Architecture……………………………………. 47

Figure 6.1 Elbow Graph for Centralized Tam dataset1……………………………. 52

Figure 6.2 Elbow Graph for Part1Tam dataset1…………………………………… 53

Figure 6.3 Elbow Graph for Part2Tam dataset1…………………………………… 54

Figure 6.4 Elbow Graph for Part3Tam dataset1.. …………………………………. 54

Figure 6.5 Determining number of no of k for Master site………………………... 55

Figure 6.6 CCQM Score for distributed & centralized k-means clustering of Tam… 58

Figure 6.7 Elbow Graph for Centralized Tam dataset2……………………………. 59

Figure 6.8 Elbow Graph for Part1Yeast dataset2………………………………….. 60

Figure 6.9 Elbow Graph for Part2Yeast dataset2………………………………….. 60

Figure 6.10 Elbow Graph for Part3Yeast dataset2………………………………….. 61

Figure 6.11 Determining number of no of k for Master site………………………... 62

Figure 6.12 CCQM Score for distributed and centralized k-means clustering of Yeast… 64

Figure 6.13 Agent Platform………………………………………………………. 65

Figure 6.14 Agent Interaction ………………………………………………………… 66

Figure A.1 Calculate number of clusters by RStudio for Tamilnadu Dataset1……. 80

Figure A.2 Run k-means algorithm by weka API at local site for Dataset1………. 81

Figure A.3 Run Global K-means Algorithm for Dataset1…………………………. 82

Figure B.1 Calculate number of clusters by RStudio for Yeast Dataset2………….. 85

Figure B.2 Run k-means algorithm by weka API for Yeast at global site…………. 86

Figure B.3 Rsults k-means algorithm by weka API for Yeast at global site………. 87

Figure B.4 Jade agent Code……………………………………………………... 88

Figure B.5 Jade Agent Running…………………………………………………. 89

Figure B.6 Agent Interface……………………………………………………… 90

LIST OF TABLES

Table 2.1 DDM Strategies…………………………………………………………… 16

Table 4.1 Related Work ……………………………………………………………… 43

Table 6.1 The centralized k-means clustering result of Tamilnadu dataset1……….. 55

Table 6.2 The three distributed partitions of Tamilnadu dataset1……........................ 56

Table 6.3 Cluster results at global site over 16 agents' communication cycles……... 57

Table 6.4 The centralized k-means clustering result of Yeast dataset2………………. 59

Table 6.5 The three distributed partitions of Yeast dataset2……………….............. 62

Table 6.6 Cluster results at global site over 7 agents' communication cycles ……… 63

Table A.1 Samples of Tamilnadu records………………………………………. 79

Table1 B.1 Samples of Yeast records …………………………………………… 84

LIST OF ALGORITHMS

Algorithm 3.1 K-means clustering algorithm…………………………………………… 26

Algorithm 3.2 Distributed K-means Algorithm……………………………………....... 27

Algorithm 5.1 Proposed algorithm of agent based distributed k-means clustering…. 49