



**UNIVERSITI PUTRA MALAYSIA**

**CLUSTERING ENSEMBLE LEARNING METHOD  
BASED ON INCREMENTAL GENETIC ALGORITHMS**

**REZA GHAEMI**

**FSKTM 2012 8**

**CLUSTERING ENSEMBLE LEARNING METHOD BASED ON  
INCREMENTAL GENETIC ALGORITHMS**



**REZA GHAEMI**

**Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia,  
in Partial Fulfilment of the Requirement for the Degree of Doctor of Philosophy**

**August 2012**

*Dedicated to my Mother's soul who taught  
me how to Love, my Father who taught me  
to be Patient, my Brother who taught me  
Manhood, and my Sisters who taught me  
Affection*

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in partial fulfilment of the requirement for the degree of Doctor of Philosophy

**CLUSTERING ENSEMBLE LEARNING METHOD BASED ON  
INCREMENTAL GENETIC ALGORITHMS**

By

**REZA GHAEMI**

**August 2012**

**Chairman: Associate Professor Md Nasir Sulaiman, PhD**

**Faculty: Computer Science and Information Technology**

Over the past decade, the clustering ensemble has been emerged as a prominent method as far as the improving of clustering accuracy is concerned. Two major difficulties in clustering ensemble include diversity of clustering and consensus functions. Genetic algorithms are well known methods with high ability to resolve optimization problems including clustering. So far, limited genetic-based clustering ensemble algorithms have been developed. However, their clustering accuracy and convergence to group unlabeled samples are not still satisfied. Generally, associated common problems in traditional genetic algorithms include lose population diversity, clustering invalidity, and context insensitivity. In order to address the above-mentioned challenges, this study is devoted towards the development of a clusterer and a clustering ensemble learning method based on incremental genetic algorithms addressing group unlabeled samples. Firstly, an architecture for the clustering ensemble based on incremental genetic-based algorithms is proposed consisting of two phases: (i) to produce cluster partitions as initial populations, (ii) to combine cluster partitions and to generate final clustering solution by incremental genetic-

based clustering ensemble learning algorithm. In the first and second phases, a threshold fuzzy  $c$ -means clustering algorithm as a clusterer and a pattern ensemble learning method based on the incremental genetic-based algorithms are proposed respectively. In the first phase, the quality of cluster partitions belonging to initial populations is measured, in terms of diversity and clustering accuracy. In the second phase, the performance of incremental genetic-based clustering ensemble algorithms is measured, in terms of clustering accuracy and convergence.

A comprehensive experimental analysis is conducted by several experiments to evaluate the performance of the proposed clusterer and incremental genetic-based clustering ensemble algorithm which has been tested on the twelve benchmark datasets. In comparison to different clusterers, experimental results show that the proposed clusterer is able to produce cluster partitions with various diversity and desirable clustering accuracy. Moreover, experiments demonstrate that final clustering solution generated by the proposed incremental genetic-based clustering ensemble algorithm using the pattern ensemble learning method possess comparative or better clustering accuracy than clustering solutions generated by the incremental genetic-based clustering ensemble algorithms using other recombination operators. In addition, experiments prove that incremental genetic-based clustering ensemble algorithm speed up to converge into an optimal clustering solution, where pattern ensemble learning method and the cluster partitions produced by the threshold fuzzy  $c$ -means clustering algorithm are employed as recombination operator and initial population, respectively.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

**KAEDAH PEMBELAJARAN CLUSTERING ENSEMBLE BERASASKAN  
ALGORITMA GENETIK TOKOKAN**

Oleh

**REZA GHAEMI**

**August 2012**

**Pengerusi: Profesor Madya Md Nasir Sulaiman, PhD**

**Fakulti: Sains Komputer dan Teknologi Maklumat**

**ABSTRAK**

Sepanjang dekad yang lalu, 'clustering ensemble' telah muncul sebagai satu kaedah untuk menambahbaik ketepatan pengklusteran. Terdapat dua kesukaran utama dalam 'clustering ensemble' iaitu kepelbagaian pengklusteran dan fungsi konsensus. Algoritma genetik dikenali sebagai kaedah dengan keupayaan tinggi untuk menyelesaikan masalah pengoptimuman termasuk pengklusteran. Setakat ini, beberapa algoritma 'clustering ensemble' berasaskan-genetik telah dibangunkan yang ketepatan pengklusteran dan penumpuan untuk mengumpul sampel yang tidak berlabel masih tidak mencapai kepuasan dan juga sering mempunyai masalah yang sama dalam algoritma genetik tradisional seperti kehilangan kepelbagaian populasi, ketidaksahihan pengklusteran, dan ketidaksensitifan konteks. Sebagai tindakbalas kepada cabaran yang tersebut di atas, kajian ini tertumpu kepada membangunkan satu kaedah pembelajaran 'clustering ensemble' berasaskan algoritma genetik tokokan untuk mengumpul sampel yang tidak berlabel. Pada mulanya, satu senibina untuk satu algoritma 'clustering ensemble' berasaskan algoritma genetik tokokan

dicadangkan yang terdiri daripada dua fasa: untuk menghasikan partisi kluster sebagai populasi permulaan dalam fasa yang pertama, kemudian, untuk menggabungkan partisi-partisi kluster dan untuk menjana penyelesaian pengklusteran terakhir oleh algoritma 'clustering ensemble' berasaskan algoritma genetik peningkatan dalam fasa ke dua. Dalam fasa pertama dan kedua, satu ambang algoritma pengklusteran c-purata kabur sebagai satu pengkluster dan algoritma k-corak kaedah pembelajaran ensemble berasaskan genetik tokokan dicadangkan. Dalam fasa pertama, kualiti bagi partisi kluster yang dipunyai oleh populasi permulaan diukur dari segi kepelbagaian dan ketepatan pengklusteran. Dalam fasa kedua, prestasi bagi algoritma 'clustering ensemble' berasaskan-genetik tokokan diukur dari segi ketepatan pengklusteran dan penumpuan.

Analisis eksperimen yang komprehensif dikendali oleh beberapa eksperimen untuk menilai prestasi pengkluster yang dicadangkan dan algoritma 'clustering ensemble' berasaskan-genetik tokokan yang telah diuji ke atas dua belas set data penanda aras. Berbanding dengan pengkluster yang berbeza, keputusan eksperimen menunjukkan pengkluster yang dicadangkan mampu untuk menghasilkan partisi kluster dengan kepelbagaian dan ketepatan pengklusteran yang diinginkan berbanding dengan pengkluster yang lain. Selain itu, eksperimen menunjukkan bahawa penyelesaian pengkelasan terakhir yang dijana oleh algoritma 'clustering ensemble' berasaskan-genetik tokokan menggunakan kaedah pembelajaran ensemble k-corak mencapai ketepatan pengklusteran yang setanding atau lebih baik daripada penyelesaian yang dijana oleh algoritma 'clustering ensemble' berasaskan-genetik tokokan menggunakan penggabungan semula operator-operator lain. Disamping itu, eksperimen membuktikan bahawa algoritma 'clustering ensemble' berasaskan-

genetik tokokan yang dicadangkan adalah laju untuk menumpu ke dalam penyelesaian pengklusteran optimum, yang mana kaedah pembelajaran ensemble k-corak dan partisi-partisi kluster yang dihasilkan oleh satu ambang algoritma pengklusteran c-purata kabur yang digunakan sebagai operator penggabungan semula dan populasi permulaan.





## ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my supervisor Assoc. Prof. Dr. Md Nasir B Sulaiman for giving me an opportunity to start this study. Through the course of my study, I have had the great fortune to get to know and interact with him. His comments and suggestions for further development as well as his assistance during writing this thesis are invaluable to me. His specific background on data mining, interest, teaching and research style has provided for me an exceptional opportunity to learn more.

I would like to express my sincere thanks and appreciation to the supervisory committee members Professor Dr. Hamidah Ibrahim and Associate Professor Dr. Norwati Mustapha for their guidance, valuable suggestions and advice throughout this work in making this a success.

My deepest appreciation to my father Mr. Rahim Ghaemi who has been supportive and patiently waiting for me to complete my study. Finally, I owe my sincere thanks to my brother and sisters, for their encouragement and affirmation, which made it possible for me to achieve this work.

For the others who have directly or indirectly helped me in the completion of my work, I thank you all.

I certify that a Thesis Examination Committee has met on 3 August 2012 to conduct the final examination of Reza Ghaemi on his thesis entitled "Clustering Ensemble Method Based on Incremental Genetic Algorithms" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the Doctor of Philosophy degree.

Members of the Examination Committee were as follows:

**Rahmita Wirza, PhD**

Associate Professor  
Faculty of Computer Science and Information Technology  
Universiti Putra Malaysia  
(Chairman)

**Ramlan Mahmud, PhD**

Full Professor  
Faculty of Computer Science and Information Technology  
Universiti Putra Malaysia  
(Member)

**Abu Bakar b Md. Sultan, PhD**

Associate Professor  
Faculty of Computer Science and Information Technology  
Universiti Putra Malaysia  
(Member)

**Anca L. Ralescu, PhD**

Full Professor  
School of Computing Sciences and Informatics  
University of Cincinnati - USA  
(External Examiner)

---

**SEOW HENG FONG, PhD**

Professor and Deputy Dean  
School of Graduate Studies  
Universiti Putra Malaysia

Date:

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of philosophy. The members of the Supervisory Committee were as follows:

**Md. Nasir Sulaiman, PhD**

Associate Professor  
Faculty of Computer Science and Information Technology  
Universiti Putra Malaysia  
(Chairman)

**Hamidah Ibrahim, PhD**

Professor  
Faculty of Computer Science and Information Technology  
Universiti Putra Malaysia  
(Member)

**Norwati Mustapha, PhD**

Associate Professor  
Faculty of Computer Science and Information Technology  
Universiti Putra Malaysia  
(Member)

---

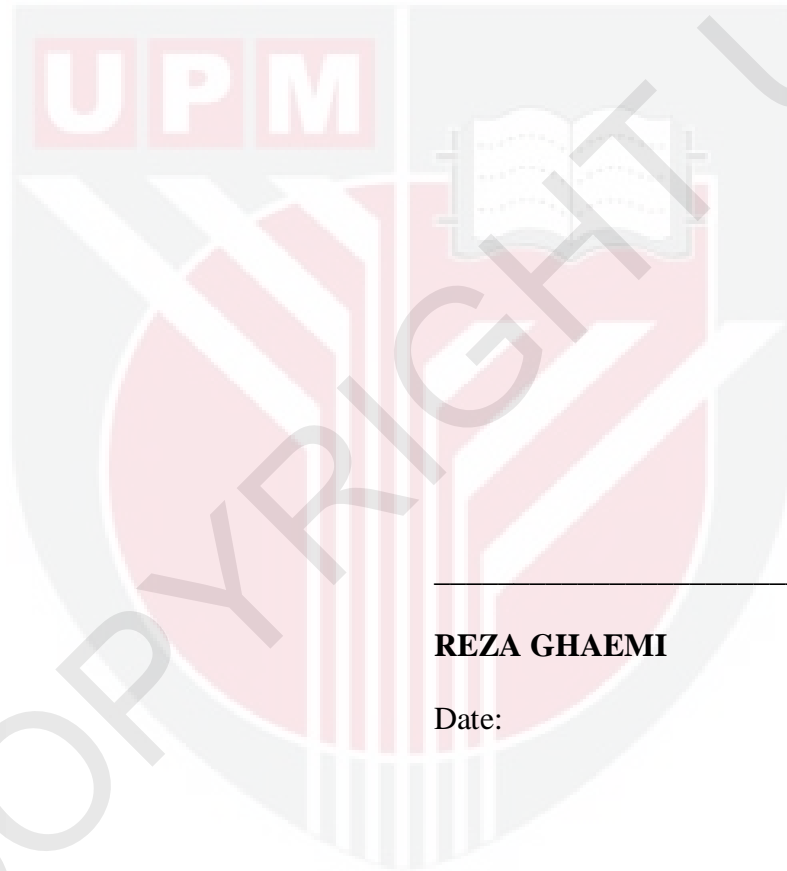
**BUJANG BIN KIM HUAT, PhD**

Professor and Dean  
School of Graduate Studies  
Universiti Putra Malaysia

Date:

## DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Universiti Putra Malaysia or any other institution.



---

**REZA GHAEMI**

Date:

## TABLE OF CONTENTS

|   | <b>Page</b> |
|---|-------------|
| <b>DEDICATION</b>                                     | ii          |
| <b>ABSTRACT</b>                                       | iii         |
| <b>ABSTRAK</b>  | v           |
| <b>ACKNOWLEDGEMENTS</b>                               | viii        |
| <b>APPROVAL</b>                                       | viii        |
| <b>DECLARATION</b>                                    | xi          |
| <b>LIST OF TABLES</b>                                 | xv          |
| <b>LIST OF FIGURES</b>                                | xv          |
| <b>LIST OF ABBREVIATIONS AND NOTATIONS</b>            | xvii        |
| <br>  |             |
| <b>CHAPTER</b>  |             |
| <br>  |             |
| <b>1 INTRODUCTION</b>                                 | <b>1</b>    |
| 1.1 Motivation  | 1           |
| 1.2 Problem Statement                                 | 3           |
| 1.3 Research Objectives                               | 4           |
| 1.4 Research Scope                                    | 5           |
| 1.5 Research Contributions                            | 6           |
| 1.6 Organization of Thesis                            | 8           |
| <br>  |             |
| <b>2 BACKGROUND</b>                                   | <b>11</b>   |
| 2.1 Introduction                                      | 11          |
| 2.2 Cluster Analysis                                  | 11          |
| 2.2.1 Data Mining and Its Main Tasks                  | 12          |
| 2.2.2 Clustering Problem                              | 14          |
| 2.2.3 A Categorization of Major Clustering Algorithms | 16          |
| 2.2.4 Clustering Measurement                          | 17          |
| 2.3 Clustering Ensemble Problem                       | 18          |
| 2.3.1 Formulation of Clustering Ensemble Problem      | 20          |
| 2.4 Challenges in Clustering Ensemble                 | 22          |
| 2.5 Genetic Algorithm as an Evolutionary Approach     | 23          |
| 2.5.1 Evolutionary Approaches                         | 23          |
| 2.5.2 Genetic Algorithms                              | 24          |
| 2.5.3 A Categorization of Genetic Algorithms          | 26          |
| 2.5.4 Generational GAs versus Incremental GAs         | 27          |
| 2.6 Summary   | 31          |
| <br>  |             |
| <b>3 LITERATURE REVIEW</b>                            | <b>32</b>   |
| 3.1 Introduction                                      | 32          |
| 3.2 Consensus Function Problem                        | 32          |
| 3.2.1 Hypergraph Partitioning Method                  | 33          |
| 3.2.2 Voting Approach                                 | 35          |
| 3.2.3 Information Theoric Method                      | 36          |
| 3.2.4 Co-Association Based Function Method            | 37          |

|          |   |            |
|----------|---|------------|
| 3.2.5    | Mixture Model   | 41         |
| 3.3      | Diversity of Clustering   | 42         |
| 3.3.1    | Using Different Clustering Algorithms   | 44         |
| 3.3.2    | Changing Initialization and Parameters  | 46         |
| 3.3.3    | Using Different Features  | 46         |
| 3.4      | Comparison Between Clustering Ensemble Methods  | 47         |
| 3.5      | Evidence Accumulation Clustering Method   | 51         |
| 3.6      | GA-based Clustering Ensemble  | 53         |
| 3.6.1    | Representation Scheme   | 54         |
| 3.6.2    | Population Initialization   | 56         |
| 3.6.3    | Fitness Function  | 57         |
| 3.6.4    | Selection Mechanism   | 58         |
| 3.6.5    | Recombination Operator  | 60         |
| 3.6.6    | Mutation Operator   | 66         |
| 3.6.7    | Reinsertion Operation   | 67         |
| 3.6.8    | Termination Condition   | 68         |
| 3.7      | GA-based Clustering Ensemble Algorithms   | 69         |
| 3.8      | Evaluation and Validation of Clustering Ensemble  | 72         |
| 3.8.1    | Clustering Validation   | 72         |
| 3.8.2    | Evaluation Diversity of Clustering  | 75         |
| 3.8.3    | Convergence Evaluation  | 77         |
| 3.9      | Summary   | 78         |
| <b>4</b> | <b>METHODOLOGY</b>  | <b>79</b>  |
| 4.1      | Introduction  | 79         |
| 4.2      | An Overview of The Problem  | 79         |
| 4.3      | Research Steps  | 81         |
| 4.4      | Experimental Design   | 88         |
| 4.4.1    | Benchmark Datasets  | 88         |
| 4.4.2    | Experimental Remarks  | 90         |
| 4.5      | Summary   | 93         |
| <b>5</b> | <b>THRESHOLD FUZZY C-MEANS CLUSTERING ALGORITHM AND<br/>PATTERN ENSEMBLE LEARNING METHOD BASED ON<br/>INCREMENTAL GA-BASED ALGORITHMS</b> | <b>94</b>  |
| 5.1      | Introduction  | 94         |
| 5.2      | The Clustering Ensemble Architecture for the Incremental GA-Based<br>Clustering Ensemble Algorithm  | 95         |
| 5.3      | Generation process  | 98         |
| 5.3.1    | Producing Cluster Partitions by Clusterers  | 99         |
| 5.3.2    | Threshold Fuzzy C-Means Clustering Algorithm  | 102        |
| 5.4      | Combination process   | 112        |
| 5.4.1    | Incremental GA-based Clustering Ensemble Algorithm using Pattern<br>Ensemble Learning Method  | 114        |
| 5.5      | Summary   | 139        |
| <b>6</b> | <b>RESULTS AND DISCUSSIONS</b>  | <b>140</b> |
| 6.1      | Introduction  | 140        |
| 6.2      | Experimental Environment  | 140        |
| 6.3      | Experimental Results of Generation Process  | 141        |

|          |  |            |
|----------|--|------------|
| 6.3.1    | Diversity Evaluation of Cluster Partitions Produced by Clusterers  | 142        |
| 6.3.2    | Clustering Accuracy Evaluation obtained by Cluster Partitions Produced by Clusterers   | 149        |
| 6.4      | Experimental Results of Combination Process  | 158        |
| 6.4.1    | Genetic Parameters Setting   | 160        |
| 6.4.2    | Comparable Clustering Ensemble Algorithms  | 161        |
| 6.4.3    | Clustering Accuracy Evaluation obtained by the Incremental GA-based Clustering Ensemble Algorithms   | 166        |
| 6.4.4    | Relationship between Clustering Accuracies Resulted by the Incremental GA-based Clustering Ensemble Algorithms and the Initial Populations | 175        |
| 6.4.5    | Comparison of the Clustering Accuracy obtained by IGCPCL with the First Group of the Comparable Clustering Ensemble Algorithms             | 180        |
| 6.4.6    | Convergence Evaluation   | 181        |
| 6.5      | Summary  | 192        |
| <b>7</b> | <b>CONCLUSIONS AND FUTURE WORK</b>   | <b>195</b> |
| 7.1      | Conclusions  | 195        |
| 7.2      | Future Work  | 199        |
|          | <b>REFERENCES</b>  | <b>200</b> |
|          | <b>APPENDIX A</b>  | <b>215</b> |
|          | <b>APPENDIX B</b>  | <b>218</b> |
|          | <b>APPENDIX C</b>  | <b>220</b> |
|          | <b>APPENDIX D</b>  | <b>222</b> |
|          | <b>BIODATA OF STUDENT</b>  | <b>226</b> |
|          | <b>LIST OF PUBLICATIONS</b>  | <b>227</b> |