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

© COPYRIGHT UPM

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

© COPYRIGHT UPM

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-

© COPYRIGHT UPM

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.

© COPYRIGHT UPM

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

© COPYRIGHT UPM

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 diingini 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-

© COPYRIGHT UPM

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.

© COPYRIGHT UPM

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.

© COPYRIGHT UPM

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 Mahmod, 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:

© COPYRIGHT UPM

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:

© COPYRIGHT UPM

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:

© COPYRIGHT UPM

TABLE OF CONTENTS

© COPYRIGHT UPM

© COPYRIGHT UPM

© COPYRIGHT UPM