A MULTI-OBJECTIVES GENETIC ALGORITHM CLUSTERING ENSEMBLES BASED APPROACH TO SUMMARIZE RELATIONAL DATA

GABRIEL JONG CHIYE

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH
2015

A MULTI-OBJECTIVES GENETIC ALGORITHM CLUSTERING ENSEMBLES BASED APPROACH TO SUMMARIZE RELATIONAL DATA

GABRIEL JONG CHIYE

SUBMITTED IN PARTIAL FULFILLMENT FOR THE REQUIREMENTS FOR THE BACHELOR OF COMPUTER SCIENCE (SOFTWARE ENGINEERING)

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH
2015

SUPERVISOR'S CONFIRMATION

NAME	:	**GABRIEL JONG CHIYE**
MATRIK NO.	:	**BK11110089**
TITLE	:	**A MULTI-OBJECTIVES GENETIC ALGORITHM CLUSTERING ENSEMBLES BASED APPROACH TO SUMMARIZE RELATIONAL DATA**
DEGREE	:	**DEGREE IN BACHELOR OF COMPUTER SCIENCE (SOFTWARE ENGINEERING)**
VIVA DATE	:	**29 JUNE 2015**

CERTIFIED BY

1. SUPERVISOR
Associate Professor Dr. Rayner Alfred Signature

_____

DECLARATION

I hereby declare that the material in this thesis is my own except for quotations, excerpts, equations, summaries and references. All information from these other sources has been duly acknowledged.

June 29, 2015 _____

Gabriel Jong Chiye

BK11110089

ACKNOWLEDGEMENT

First of all, I would like to express my very great appreciation to my main supervisor Associate Professor Dr. Rayner Alfred for provided me with valuable suggestions during the planning of this project, constructive guidance during the development of this project work. His willingness to give his time and share his knowledge so generously has been very much appreciated. I would also like to offer my special thanks to my examiner Dr. Lau Hui Keng and panel Dr. Mohd Hanafi Ahmad Hijazi for their encouragement, value recommendations and advise to keep my progress on schedule.

I would also like to extend my thanks to senior lecturer Dr. Chin Kim On for his help in offering me the research laboratory and necessary resources to run the experiments. My special thanks extended to my doctorate degree and master degree colleagues from the Center of Excellence of Semantic Agents for their professional guidance.

Finally, I wish to thank my parents and brother for their continuous love, support and encouragement throughout my study.

Gabriel Jong Chiye
June 2015

ABSTRACT

K-means algorithm is one of the well-known clustering algorithms that promise to converge to a local optimum in few iterative. However, traditional k-means algorithm is designed to cluster data of single target table. Due to the nature of data collected in real life applications, many data have been collected and stored in relational databases. Traditional clustering and classification learning algorithms cannot be applied directly in learning multi-relational databases. Several approaches have been designed and proposed to learn relational data which includes Inductive Logic Programming based approaches, Graph based approaches, Multi-View approaches and also Dynamic Aggregation of Relational Attributes approach. Dynamic Aggregation of Relational Attributes approach is very effective in learning relational data set. Dynamic Aggregation of Relational Attributes summarizes relational data by clustering records exist in non-target tables. However, the quality of summarization of data depends highly on the position of initial centroids selected. Thus, it may affect the overall classification task. Thus, this project proposes a Genetic Algorithm-based Clustering Ensembles in learning relational datasets by combining the results obtained from several k-means clustering runs with different values of number of clusters, in which the location of centroids are optimal for every sets of clusters. The effects of using different similarity measurements and applying different fitness functions for the genetic algorithm on the predictive accuracies of the classifiers are also studied. Based on the results obtained, it can be concluded that using the consensus result of several clustering results can increase the predictive accuracy of classification task. It can be concluded that the Euclidean distance has better performance on mutagenesis datasets and cosine similarity has better performance on hepatitis datasets when evaluated with Weka C4.5 classifier, but the other way round when Naïve Bayes classifier is used for evaluation.

ABSTRAK

Bersama-samaan: Kekelompokan Berdasarkan Susunan

Algorithma k-means *adalah satu algoritma pengelompokan yang berjanji untuk menghasilkan satu optima tempatan dalam beberapa lelaran. Walau bagaimanapun, algorithma* k-means *yang tradisional telah direka untuk mengelompok data jadual sasaran tunggal. Oleh kerana aplikasi kehidupan sebanar, kita mengumpul dan menyimpan data di dalam pangkalan data berhubungan. Algoritma pembelajaran secara pengelompokan dan klasifikasi tidak bolek diguna secara langsung dengan pangkalan data berhubungan. Beberapa cara telah direka dan dicadangkan untuk belajar data berhubung, termasuk* Inductive Logic Programming*,* Graph*,* Multi-View*, dan* Dynamic Aggregation of Relational Attributes*.* Dynamic Aggregation of Relational Attributes *merupakan cara yang berkesan untuk belajar data berhubung.* Dynamic Aggregation of Relational Attributes *meringkaskan data berhubung dengan mengelompok rekod dalam jadual bukan sasaran. Walao bagaimanapun, kualiti ringkasan data sangat bergantung kepada kedudukan pusat awal yang dipilih. Dengan itu, ia boleh menjejaskan kerja keseluruhan klasifikasi. Oleh itu, kerja ini mencadang satu cara pengelompokan* ensemble *berdasarkan algoritma genetic bagi belajar data berhubung dengan menggabungkan keputusan yang didapat dari beberapa larian pengelompokan* k-mean *dengan bilang kelompokan yang berbeza, di mana lokasi pusat adalah optima untuk setiap pengelompokan. Kesan ketepatan ramalan oleh pengelas bagi mengguna pengukuran jarak yang berbeza dan mengguna fungsi kecergasan dalam algoritma genetik juga telah dikaji. Berdasarkan keputusan kajian yang telah diperolehi, ia boleh disimpulkan bahawa keputusan koncensus boleh meningkat ketepatan ramalan oleh tugas pengelasan. Ia juga boleh dikatakan bahawa jarak* Euclidean *boleh memperoleh prestasi yang lebih tinggi dalam data* mutagenesis*, persamaan kosinus memperoleh prestasi yang lebih tinggi dalam data* hepatitis *semasa dinilai dengan pengelas* Weka C4.5*, tetapi sebaliknya ketika dinilai dengan pengelas* Naïve Bayes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

BIBLIOGRAPHY

[1] R. Cattral, F. Oppacher, and K.J.L. Graham, "Techniques for evolutionary rule discovery in data mining," in *Evolutionary Computation, 2009. CEC '09. IEEE Congress*, Trondheim, May 2009, pp. 1737-1744.

[2] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, and Yong Ren, "Information Security in Big Data: Privacy and Data Mining," in *Access, IEEE*, 2014, pp. 1149-1176.

[3] Arno J. Knobbe, Marc de Haas, and Arno Siebes, "Principles of Data Mining and Knowledge Discovery," in *Principles of Data Mining and Knowledge Discovery.*: Springer Berlin Heidelberg, 2001, pp. 277-288.

[4] Jiawei Han, "Data Mining," in *Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship (2013).*, pp. 595-597.

[5] Dipankar Dutta, Paramartha Dutta, and Jaya Sil, "Data Clustering with Mixed Features by Multi Objective Generic Algorithm," in *12th International Conference on Hybrid Intelligent Systems*, Pune, 2012, pp. 336-341.

[6] Saso Dzeroski, "Relational Data Mining," in *Data Mining and Knowledge Discovery Handbook*, Oded Maimon and Lior Rokach, Eds.: Springer US, 2010, pp. 887-911.

[7] Ping Ling and Xiangsheng Rong, "Double-Phase Locality Sensitive Hashing of neighborhood development for multi-relational data," *Computational Intelligence (UKCI), 2013 13th UK Workshop*, pp. 206-213, September 2013.

[8] Urvashi Mistry and Amit R Thakkar, "Link-based classification for Multi-Relational database," *Recent Advances and Innovations in Engineering (ICRAIE), 2014*, pp. 1-6, May 2014.

[9] Wei Zhang, "Multi-Relational Data Mining Based on Higher-Order Inductive Logic," in *Intelligent Systems, 2009. GCIS '09. WRI Global Congress*, Xiamen, 2009, pp. 453-458.

[10] Jingfeng Guo, Lizhen Zheng, and Tieying Li, "An Efficient Graph-based Multi-relational Data Mining Algorithm," in *2007 International Conference on Computational Intelligence and Security*, Harbin, 2007, pp. 176-180.

[11] Hongyu Guo and Herna L. Viktor, "Mining Relational Databases with Multi-view Learning," in *Proceedings of the 4th international workshop on Multi-relational mining*, New York, 2005, pp. 15-24.

[12] Dan Roth and Wen-tau Yih, "Propositionalization of Relational Learning: An Information Extraction Case Study," in *Seventeenth International Joint Conference on Artificial Intelligence*, Seattle, 2001.

[13] Chung Seng Kheau, Rayner Alfred, and Lau Hui Keng, "Dimensionality Reduction in Data Summarization Approach to Learning Relational Data," in *Intelligent Information and Database Systems Lecture Notes in Computer Science*. Berlin, Germany: Springer, 2013, pp. 166-175.

[14] Rayner Alfred, "Summarizing Relational Data using Semi-Supervised Genetic Algorithm-based Clustering Techniques," *Journal of Computer*, vol. 6, no. 7, pp. 775-784, 2010.

[15] Rao M. Kotamarti, Douglas W. Raiford, Michael L. Raymer, and Margaret H. Dunham1, "A Data Mining Approach to Predicting Phylum for Microbial Organisms using Genome-Wide Sequence Data," in *Bioinformatics and BioEngineering, 2009. BIBE '09. Ninth IEEE International Conference*, Taichung, June 2009, pp. 161-167.

[16] W. Segretier, M. Clergue, M. Collard, and L. Izquierdo, "An evolutionary data mining approach on hydrological data with classifier juries," in *Evolutionary*

Computation (CEC), 2012 IEEE Congress, Brisbane, June 2012, pp. 1-8.

[17] Ming Xue and Changjun Zhu, "The Application of Data Mining In the Decision of Supermarket Extension and Businesses Expansion Based on Evolutionary Computation," in *Circuits, Communications and Systems, 2009. PACCS '09. Pacific-Asia Conference*, Chengdu, May 2009, pp. 778-780.

[18] Bin Lu and Fangyuan Ju, "An optimized genetic K-means clustering algorithm," in *International Conference on Computer Science and Information Processing*, Xi'an, 2012, pp. 1296-1299.

[19] Taoying Li and Yan Chen, "A Weight Entropy k-means Algorithm for Clustering Dataset with Mixed Numeric and Categorical Data," in *Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, Shandong, 2008, pp. 36-41.

[20] David Pettinger and Giuseppe Di Fatta, "Space Partitioning for Scalable K-Means," in *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference*, Washington, December 2010, pp. 319-324.

[21] Xiaohui Cui, Thomas E. Potok, and Paul Palathingal, "Document Clustering using Particle Swarm Optimization," in *Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE*, 2005, pp. 185-191.

[22] R.F. Abdel-Kader, "Genetically Improved PSO Algorithm for Efficient Data Clustering," in *Machine Learning and Computing (ICMLC), 2010 Second International Conference*, Bangalore, February 2010, pp. 71-75.

[23] Nisha M. N., Mohanavalli S., and Swathika R., "Improving the quality of Clustering using Cluster Ensembles," in *Proceedings of 2013 IEEE Conference on Information and Communication Technologies*, JeJu Island, 2013, pp. 88-92.

[24] Tan Pang-Ning, Michael Steinbach, and Vipin Kumar, *Introduction to Data Mining*. Boston, United States: Pearson, 2006.

[25] Soumen Chakrabarti et al., *Data Mining Know It All*. Burlington, United States: Morgan Kaufmann, 2009.

[26] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval: the concepts and technology behind search*. Harlow, England: Pearson Education Limited, 2011.

[27] Nileshkumar D. Bharwad and Mukesh M. Goswami, "Proposed efficient approach for classification for multi-relational data mining using Bayesian Belief Network," in *Green Computing Communication and Electrical Engineering (ICGCCEE), 2014 International Conference*, Coimbatore, 2014, pp. 1-4.

[28] Stephen Muggleton, "Inductive Logic Programming," *New Generation Computing*, vol. 8, no. 4, pp. 295-318, 1991.

[29] Lawrence B. Holder and Diane J. Cook, "Graph-Based Relational Mining: Current and Future Directions," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 1, pp. 90-93, July 2003.

[30] Rayner Alfred, "The Study of Dynamic Aggregation of Relational Attributes on Relational Data Mining," in *Third International Conference, ADMA*, Harbin, 2007, pp. 214-226.

[31] Florence Sia, Rayner Alfred, Leau Yu Beng, and Tan Soo Fun, "A Variable Length Feature Construction method for data summarization using DARA," in *Computing and Convergence Technology (ICCCT), 2012 7th International Conference*, Seoul, 2012, pp. 881-887.

[32] Neepa Shah and Sunita Mahajan, "Document Clustering: A Detailed Review," *International Journal of Applied Information Systems*, vol. 4, pp.

30-38, October 2012.

[33] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze, *An Introduction to Information Retrieval*, Online edition ed. Cambridge, England: Cambridge University Press, 2009.

[34] Chun-Ling Chen, Frank S.C. Tseng, and Tyne Liang, "An integration of WordNet and fuzzy association rule mining for multi-label document clustering," *Data & Knowledge Engineering*, vol. 69, no. 11, pp. 1208-1226, November 2010.

[35] Cheng-Fa Tsai, Han-Chang Wu, and Chun-Wei Tsai, "A new data clustering approach for data mining in large databases," in *Parallel Architectures, Algorithms and Networks, 2002. I-SPAN '02. Proceedings. International Symposium*, Makati City, 2002, pp. 278-283.

[36] Stuart P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129-137, March 1982.

[37] Hans-Hermann Bock, "Clustering Methods: A History of k-Means Algorithms," in *Selected Contributions in Data Analysis and Classification*. Berlin: Springer Berlin Heidelberg, 2007, pp. 161-172.

[38] Xin Jin and Jiawei Han, "K-Means Clustering," in *Encyclopedia of Machine Learning*, Claude Sammut and Geoffrey I. Webb, Eds. United States: Springer US, 2010, pp. 563-564.

[39] Erendira Rendon, Alejandra Arizmendi Itzel Abundez, and Elvia M. Quiroz, "Internal versus External cluster validation indexes," *International Journal of Computers and Communications*, vol. 5, no. 1, pp. 27-32, 2011.

[40] M. Bilal, S. Masud, and S. Athar, "FPGA Design for Statistics-Inspired Approximate Sum-of-Squared-Error Computation in Multimedia Applications," *Circuits and Systems II: Express Briefs, IEEE Transactions*,

vol. 59, no. 8, pp. 506-510, July 2012.

[41] Thomas Weise and R. Chiong, "Evolutionary Data Mining Approaches for Rule-based and Tree-based Classifiers," *Cognitive Informatics (ICCI), 2010 9th IEEE International Conference*, pp. 696-703, July 2010.

[42] Melanie Mitchell, *An Introduction To Genetic Algorithms*. London, England: MIT Press, 1999.

[43] Noraini Mohd Razali and John Geraghty, "Genetic Algorithm Performance with Different Selection Strategies in Solving TSP," in *Proceedings of the World Congress on Engineering 2011*, vol. II, London, 2011.

[44] Abdul Wahid, Xiaoying Gao, and Andreae Peter, "Multi-View Clustering of Web Documents using Multi-Objective Genetic Algorithm," in *Evolutionary Computation (CEC), 2014 IEEE Congress*, Beijing, 2014, pp. 2625 - 2632.

[45] Xiaoyu Wen, Xinyu Li, Liang Gao, Liang Wan, and Wenwen Wang, "Multi-objective genetic algorithm for integrated process planning and scheduling with fuzzy processing time," in *Advanced Computational Intelligence (ICACI), 2013 Sixth International Conference*, Hangzhou, 2013, pp. 293-298.

[46] Abdullah Konak, David W. Coit, and Alice E. Smith, "Multi-objective optimization using genetic algorithms: A tutorial," *Reliability Engineering & System Safety*, vol. 91, no. 9, pp. 992-1007, September 2006.

[47] Fatimah Sham Ismail, Rubiyah Yusof, and S Muhd Masduqi Waqiyuddin, "Multi-objective optimization problems: Method and application," in *Modeling, Simulation and Applied Optimization (ICMSAO), 2011 4th International Conference*, Kuala Lumpur, 2011, pp. 1-6.

[48] Nafissa Zeghichi, Mekki Assas, and Leila Hayet Mouss, "Genetic Algorithm with Pareto Fronts for Multi-Criteria Optimization Case Study Milling

Parameters Optimization," in *Software, Knowledge Information, Industrial Management and Applications (SKIMA), 2011 5th International Conference*, Benevento, 2011, pp. 1-5.

[49] K. Atashkari, N. NarimanZadeh, A. R. Ghavimi, M. J. Mahmoodabadi, and F. Aghaienezhad, "Multi-objective optimization of power and heating system based on artificial bee colony," in *Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium*, Istanbul, 2011, pp. 64-68.

[50] Zhi-Hua Zhou, "Ensemble Learning," in *Encyclopedia of Biometrics*, Stan Z. Li and Anil Jain, Eds. United States: Springer, 2009, pp. 270-273.

[51] Shaohua Wan and Hua Yang, "Comparison among Methods of Ensemble Learning," in *Biometrics and Security Technologies (ISBAST), 2013 International Symposium*, Chengdu, 2013, pp. 286-290.

[52] S. Kulkarni and V. Kelkar, "Classification of multispectral satellite images using ensemble techniques of bagging, boosting and adaboost," in *Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference*, Mumbai, April 2014, pp. 253-258.

[53] Xiang Hui and Yang Sheng Gang, "Using clustering-based bagging ensemble for credit scoring," in *Business Management and Electronic Information (BMEI), 2011 International Conference*, vol. 3, Guangzhou, May 2011, pp. 369-371.

[54] S. Bernard, S. Adam, and L. Heutte, "Using Random Forests for Handwritten Digit Recognition," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference*, vol. 2, Parana, September 2007, pp. 1043-1047.

[55] Jian Xu, Jianhua Chen, and Bin Li, "Random Forest for Relational Classification with Application to Terrorist Profiling," in *Granular Computing,*

2009, GRC '09. IEEE International Conference, Nanchang, August 2009, pp. 630-633.

[56] Peter Buhlmann,.: Springer Handbooks of Computational Statistics 2012, 2011, pp. 985-1022.

[57] Youqin Pan and Zaiyong Tang, "Ensemble methods in bank direct marketing," in *Service Systems and Service Management (ICSSSM), 2014 11th International Conference*, Beijing, June 2014, pp. 1-5.

[58] Yijun Chen and Man Leung Wong, "An Ant Colony Optimization Approach for Stacking Ensemble," in *2010 Second World Congress on Nature and Biologically Inspired Computing*, Fukuoka, 2010, pp. 146-151.

[59] R. Polikar, "Ensemble based systems in decision making," *Circuits and Systems Magazine, IEEE*, vol. 6, no. 3, pp. 21-45, September 2006.

[60] Gao Wei and Cheng Mingshu, "A new dynamic credit scoring model based on clustering ensemble," in *Computer Science and Network Technology (ICCSNT), 2013 3rd International Conference*, Dalian, 2013, pp. 421-425.

[61] Mao-ting Gao and Bing-jing Wang, "Text clustering ensemble based on genetic algorithms," in *Systems and Informatics (ICSAI), 2012 International Conference*, Yantai, 2012, pp. 2329-2332.

[62] Rodrigo A. Coelho, Fabricio R. N. Guimaraes, and Ahmed A. A. Esmin, "Applying Swarm Ensemble Clustering Technique for Fault Prediction Using Software Metrics," in *Machine Learning and Applications (ICMLA), 2014 13th International Conference*, Detroit, 2014, pp. 356-361.

[63] R Ghaemi, M. Sulaiman, N. Mustapha, and H. Ibrahim, "Improving of Initial Clusters Fitness in Genetic Guided-Clustering Ensembles," in *Information Technology: New Generations (ITNG), 2010 Seventh International*

Conference, Las Vegas, 2010, pp. 227-232.

[64] Armindokht Hashempour Sadeghian and Hossein Nezamabadi-pour, "Document clustering using gravitational ensemble clustering," in *Artificial Intelligence and Signal Processing (AISP), 2015 International Symposium*, Mashhad, 2015, pp. 240-245.

[65] Zhenya Zhang, Hongmei Cheng, Shuguang Zhang, Wanli Chen, and Qiansheng Fang, "Clustering aggregation based on genetic algorithm for documents clustering," in *Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence). IEEE Congress*, Hong Kong, 2008, pp. 3156-3161.

[66] Rayner Alfred and Dimitar Kazakov, "Pattern-Based Transformation Approach to Relational Domain Learning Using Dynamic Aggregation for Relational Attributes," in *The 2006 International Conference on Data Mining* , Las Vegas, 2006, pp. 118-124.

[67] Chung Seng Kheau, Rayner Alfred, and Lau Hui Keng, "Dimensionality Reduction in Data Summarization Approach to Learning Relational Data," in *Intelligent Information and Database Systems Lecture Notes in Computer Science*. Berlin, Germany: Springer, 2013, pp. 166-175.