CLUSTERING CHEMICAL DATA SET USING PARTICLE SWARM OPTIMIZATION BASED ALGORITHM

TRIYONO

UNIVERSITI TEKNOLOGI MALAYSIA

Dedicated to my beloved wife (Lita Rahmasari,S.Si), dad (Harjo Suwito), and my mom (Supiyati)

ACKNOWLEDGEMENTS

Alhamdulillah, it is with Allah S.W.T will that finally I get to finish this project. In completing this project, I was contacting many people and they have contributed toward my understanding. First of all, I wish to express my sincere appreciation to my supervisor, Associate Professor Dr. Naomie Binti Salim for her attention and generous help throughout the length of this study. I am very thankful for her kind guidance since the first day of this study. I am also very thankful to my co-supervisor, Associate Professor Dr. Siti Mariyam Binti Shamsuddin for her advices and motivation. Not forgotten to my evaluators, Dr. Siti Zaiton binti Mohd Hashim and Associate Professor Dr. Mohd Noor Md. Sap, I am also grateful for their helpful suggestions.

Thirdly, I would like to express my gratitude to Haza Nuzly bin Abdull Hamed for his sample PSO program, and special thanks to Nasser Omer Sahel Ba-Karait who had always shared his idea with me. Thank you to all my lecturers, staff, course mates and friends in the Faculty of Computer Science and Information System, Universiti Teknologi Malaysia for their help and support.

Nevertheless, thank you to my dear family who always support and encourage me. My family had encouraged me to do my best in this project. Besides, I would like to express my gratitude to my course mates and friends who had always shared their idea with me.

Finally, thank you to those who had ever helped me directly and indirectly which I cannot state out every one of them. Their helps, advices and encouragements were mostly appreciated. Thank you.

ABSTRACT

Clustering is the process of organizing similar objects into groups, with its main objective is to organize a collection of data items into some meaningful groups. Generally, clustering is the most suitable approach in dealing with huge amount dataset with higher resemblance such as chemical database. The chemical data sets contain a huge number of compounds and knowledge of the physiochemical properties. The biological activities of these compounds have a large significance in the process of designing and discovering new drugs. Many algorithms had been applied to cluster chemical data set such as Ward's algorithm. In this study, Particle Swarm Optimization (PSO) based clustering algorithm is exploited to optimize the results of other clustering algorithm such as K-means. Two chemical data sets were used and downloaded from MDDR (MDL Drug Database Report). The main difference between these two data sets is measured in terms of the similarities quantify of bioactivities between active compounds. The results are compared with Ward's algorithm in terms of proportion actives percentage in active clusters are. We found that PSO algorithm reveals better performance than Ward's algorithm on continuous data format; however for binary data format, Ward's algorithm outperforms arrogantly.

ABSTRAK

Kluster merupakan suatu proses bagi membolehkan objek yang serupa dikumpulkan ke dalam kumpulan yang sama. Tujuan utama kluster adalah untuk mencari kumpulan jenis data yang mempunyai makna dan cirri-ciri yang sama. Kepelbagaian jenis bahan kimia mengandungi satu jumlah sebatian yang sangat besar, pengetahuan physiochemical ciri-ciri dan aktiviti-aktiviti biologi sebatian-sebatian ini telah satu makna yang besar dalam proses mereka dan penemuan dadah baru. Banyak algoritma pernah digunakan ke atas kelompok set data kimia, Ward yang seumpama algoritma. Kajian ini dijalankan Particle Swarm Optimization (PSO) berpangkalan berkerumun algoritma dan juga memohon PSO untuk mengoptimumkan hasil-hasil lain berkerumun algoritma ibarat algoritma K-means. Dua set data kimia telah digunakan dan dimuat turun dari MDDR (MDL Drug Database Report). Perbezaan utama antara dua data ini set-set adalah langkah persamaan-persamaan bioactivities antara sebatian-sebatian aktif. Kajian ini juga memohon algoritma sebagai perbandingan Ward. Ukuran prestasi digunakan dalam kajian ini adalah peratusan aktif kadar dalam kelompok-kelompok aktif. Pihak kami mendapati algoritma PSO itu mempersembahkan lebih baik daripada algoritma Ward bentuk data yang berterusan, kecuali algoritma Ward mengalaahkan algoritma selainnya untuk mengelompokkan data perduaan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Clustering is the unsupervised classification of patterns. It deals with finding a structure in a collection of unlabeled data. Clustering is useful in several exploratory pattern-analysis, grouping, decision-making, and machine-learning situations, including data mining, document retrieval, image segmentation, and pattern classification. The clustering of chemical compounds is a widely used technique in the field of chemo informatics for the selection of compounds for screening, the analysis of substructure searching, and the prediction of molecular properties and biological activities from structural information.

The Particle Swarm Optimization (PSO) is a population-based optimization method, was introduced by Eberhart and Kennedy [Eberhart and Kennedy, 1995]. It was originally developed for optimization in a continuous space. It has been used to solve a range of optimization problems, including neural network training and function minimization. Recently, it was successful adapted to optimization in binary spaces, presenting good performance also when applied to discontinuous objective functions and used in the optimization of many nonlinear functions and in artificial neural networks training. Engelbrecht and Merwe also explored the applicability of PSO to cluster data vector, by modifying its basic algorithm [Engelbrecht and Merwe, 2003].

Chemical database is designed to store chemical information, such as structure diagrams. Traditional chemical structure diagrams have been used to support various tasks in chemical research and development. Large chemical databases are expected to handle the storage and searching of information on millions of molecules taking terabytes of physical memory. An important feature in a chemical database system is the ability to quantify the degree of structural similarity between pairs, or larger groups, of molecules.

1.2 Problem Background

The development process of new drugs is a lengthy and costly procedure. The historical method of drug discovery is by trial-and-error testing of chemical substances on animals, and matching the apparent effects to treatments. The new method of drug design begins with knowledge of specific chemical responses in the body or target organism, and tailoring combinations of these to fit a treatment profile.

The process needs clustering process in order to choose compounds from each cluster representative of the structural content of the original compound database, classify substitute properties that are present in a dataset and summarize the classes of compounds that exist in a given dataset. The clustering process also can be used to view range of structural classes that contains a user-defined sub-structure, Analyze structure-activity relationship, and also predict unknown properties of compounds from other compounds in the same cluster.

There are challenges caused by large chemical space describing potential new drugs without side-effects, to find drug-like compounds from a database of thousands and millions of compounds. According to the *similar property principle*, structurally similar molecules will exhibit similar physiochemical and biological properties [Fink, November 1996].

Recently, several chemical databases that contain thousands or millions of chemical compound data have been developed. Based on that database, several grouping or clustering techniques developed to accelerate drug design processes.

The thousands or millions of chemical compound grouped based on their attributes also called descriptors. However, clustering is a difficult problem combinatorially [Jain, 1999].

1.3 Problem Statement

Based on the background given in previous section, looking for new technique of clustering of chemical compound data is very importance. The compound chemical data need to be clustered (grouped) into many cluster because some need especially in food and drug design. There are many methods and techniques which we are going to use could help us in best way to do that job. This study tries to applying Particle Swarm Optimization (PSO) to cluster chemical compound data. This study also observes about the performance PSO algorithm to clustering continuous and binary data. The study expect that PSO algorithm perform better than other algorithm because synergizing more than process into best result, also PSO had been known as good algorithm in term to search optimal solution through the search space.

1.4 Objectives

The objectives of this study are:

- To cluster chemical compound data using Particle Swarm Optimization (PSO) for both continuous and binary representation of chemical data.
- To utilize PSO in optimizing the clustering results produced by other clustering algorithm on chemical data.
- To analyze performance of PSO algorithm by comparing with Ward’s algorithm in clustering different representation chemical compound data; continuous and binary.

REFERENCES

Downs, G.M., Willett, P. and Fisanick, W. (1994). Similarity searching and clustering of chemical structure databases using molecular property data. *Journal of Chemical Information and Computer Science.* 34:1094-1102.

Everitt, B.S. (1993). *Cluster Analysis, 3rd Ed.* Halsted Press, New York.

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999). *Data Clustering: A Review*, ACM Comp. Surveys, Vol.31, No. 3, September 1999.

Jolife. (1986). *Principal component analysis*. New York: Springer-Verlag.

Jones, Tim. (2003). *AI Application Programming.* Charles River Media.

Kennedy, J., and Eberhart, R. (1995). Particle swarm optimization. *Neural Networks, 1995. Proceedings*., IEEE International Conference on Volume 4, 27 Nov.- 1 Dec. 1995 Page(s):1942 - 1948 vol.4

Kennedy, J. and Eberhart, R.C. (1997). A discrete binary version of the particle swarm algorithm**,** Systems, *Man, and Cybernetics, 1997. 'Computational Cybernetics and Simulation*. IEEE International Conference on, Volume 5, 12-15 Oct. 1997 Page(s):4104 - 4108 vol.5

Kohonen, Teuvo. (2000) *Self Organizing Maps – 3rd ed.* Springer-Verlag Berlin Heidelberg New York.

Laurene, Fausett V. (1994). *Fundamental of Neural Network.* Prentice Hall.

Mao, J. and Jain, A. K. (1996). A self-organizing network for hyperellipsoidal clustering (HEC). *IEEE Trans. Neural Netw. 7*, 16–29.

Mojena, R. (1977). *Hierarchical grouping methods and stopping rules: an evaluation.* Computer Journal. 20:359-363.

Mojtaba Ahmadieh Khanesar, Mohammad Teshnehlab and Mahdi Aliyari Shoorehdeli. (2007). A novel binary particle swarm optimization**.** *Control & Automation, 2007. MED '07. Mediterranean Conference on*, 27-29 June 2007 Page(s):1 – 6

P. Berkhin. (2002). *Survey of Clustering Data Mining Techniques*, Accrue Software.

Pamela, Fink K., Herren, Tandy L. (1996). Modeling disease processes for drug development: bridging the gap between quantitative and heuristic models, November 1996.

Shah, Jehan Zeb., Salim, Naomie bt. (2006) A Fuzzy Kohonen SOM Implementation and Clustering of Bio-active Compound Structures for Drug Discovery. *Computational Intelligence and Bioinformatics and Computational Biology. CIBCB '06. 2006 IEEE Symposium on* , vol., no., pp.1-6, 28-29 Sept. 2006.

Shi, Y. and Eberhart, R. (1998). A modified particle swarm optimizer**,** *Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on*, 4-9 May 1998 Page(s):69 – 73

Shi, Y.and Eberhart, R.C. (1999). Empirical study of particle swarm optimization**,** Evolutionary Computation. *CEC 99, Proceedings of the 1999 Congress on*, Volume 3, 6-9 July 1999.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxanomy.* Freeman, San Francisco.

Suganthan. P.N. (1999) Particle swarm optimiser with neighborhood operator. *Evolutionary Computation. CEC 99. Proceedings of the 1999 Congress on,* Volume 3, 6-9 July 1999

Van der Merwe, D.W.and Engelbrecht, A.P. (2003). Data clustering using particle swarm optimization. *Evolutionary Computation. CEC '03, The 2003 Congress on*, Volume 1, 8-12 Dec. 2003 Page(s):215 - 220 Vol.1.