Evaluation of FCV and FCM Clustering Algorithms in Cluster-Based Compound Selection

Sinarwati Mohamad Suhaili
School of Information Technology
International College of Advanced Technology Sarawak
93350 Kuching
Sarawak, Malaysia
sinarwati@ppks.edu.my

Mohamad Nazim Jambli
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
94300 Kota Samarahan
Sarawak, Malaysia
jmnazim@fit.unimas.my

Sharin Hazlin Huspi
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia
81310 Skudai
Johor, Malaysia
sharing@utm.my

Abstract—**In the last few years, a number of available screening compounds has been growing rapidly due to the recent developments of high-throughput screening in drug discovery. Chemical vendors provide millions of compounds for drug lead identification; however, these compounds are highly redundant. Clustering method that groups similar compounds into families, can be used to analyze such redundancy. One of most used clustering method is cluster-based compound selection, which involves subdividing a set of compounds into clusters and choosing one compound or a small number of compounds from each cluster. However, little research has been done on overlapping method fuzzy c-means (FCM) and fuzzy c-varieties (FCV) clustering algorithms in compound selection research. Therefore, these two clustering algorithms are implemented and the performance is analyzed based on the effectiveness of the clustering results in terms of mean intercluster molecular dissimilarity (MIMDS) where these results are compared with one another. The analysis shows that in terms of MIMDS, the FCV is better than FCM because it clearly shown the uniform results compare to FCM clustering algorithm.**

Keywords- Compound Selection; FCM; FCV; MIMDS.

I. INTRODUCTION

There have been a lot of investments in new technologies especially in chemoinformatics for early stage of drug discovery. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for intended purpose of making better decisions faster especially in the area of drug lead identification and organization [1]. However, so far, these are not resulting in more drug profits as discovery and marketing of a new drug cost a pharmaceutical company up to 650-800 million US dollars and take an average of 12 to 24 years. In addition, the research of novel compound in chemical industry is vastly complex and expensive when using traditional techniques and needs a long period of trials. For example, it could take a chemist 27 million weeks or 0.5 million years to synthesize 1,000 compounds per week [2]. Nevertheless, the drug discovery process is a very risky business because most of the recently found compounds do not result in a drug. In addition, the molecule that has the potential to become drugs may cause unexpected long-term side effects. The increasing numbers of molecules with different features in chemical databases, the time-consuming and expensive process increases the complexity for the chemist bring difficulty to chemist in analyze the large dataset as fast as possible has been the factor for pharmaceutical industries to find the best method in compound selection.

One of the potential ways to reduce the processing time and cost of drug discovery is using the compound selection method to select diverse sets of molecules in lead identification process [3]. This method can be used to screen, synthesize and analyze millions of compounds in order to find a possible useful compound. It involves subdividing a set of compounds into clusters and choosing one compound or a small number of compounds from each cluster. It is also used to groups the data into classes or clusters so that the objects within the cluster have high similarity in comparison to one another, but are very dissimilar to those data objects in other clusters [4]. Indirectly, by using this method, it has helped the researches in finding lead compounds faster and more effectively.

There are two types of clusters, namely overlapped and non-overlapped. The non-overlapped clustering method occurs when each compound is a member of only one cluster. In contrast to non-overlapped clustering, overlapped clustering method can allow some molecules or compound to become members of more than one cluster. The hierarchical and non-hierarchical clustering methods are the two major categories of non-overlapped clustering. Currently, nonoverlapping method is the clustering methods mostly used in chemical datasets [4]. This is because this clustering method is simpler, easier and widely used as compared to other overlapping methods in terms of development and analysis. This is proven by Willett in

978-1-61284-130-4/11/$26.00 ©2011 IEEE