# Clustering for 2D Chemical Structures

A Study Submitted in Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy


at
The University of Sheffield

by
Chia-Wei Chu

Department of Information Studies
June 2010

# Acknowledgements

I would first like to thank my supervisors for their guidance, encouragement, patience and support throughout this study. I regard myself as extremely fortunate to have had the chance to learn from them; Prof. Peter Willett for his remarkable supervision and expertly managing this work, Dr. John Holliday for his invaluable support with all things technical.

Next, I would like to acknowledge the past and present members of Chemoinformatics Research Group: Prof. Val Gillet, Dr. Eleanor Gardiner, Kirstin Moffat, David Wood, Yogendra Patel, Kris Birchall, Hina Patel, Iain Mott, Richard Martin, Georgios Papadatos, Shereena Arif, Christoph Mueller, Aryati Bakri, Nurul Malim, Richard Sherhod, Daniel Butler, Sonny Gan and Jorge Valencia for their help, advice and special kinds of encouragement during my time in the Michael Lynch Laboratory.

Many thanks go to the people who have kept me sane in many ways, Chiung-Yao Kao deserves my gratitude for her faith in me, Wen-Chin Hsu and Yiping Hsu for taking good care of me when I was hospitalized, Kaichen Ti for sparking my interest in cooking, my flatmates Somsak Sriborisutsakul and Lucia Li for putting up with me.

Very special thanks to the staff at Royal Hallamshire Hospital and Northern General Hospital for their attentive care, especially the staff of Metabolic Bone Centre, who keep me in good condition for the past two years.

A massive thank you goes to my parents for their continued love and faith, also to my wife and son for their understanding.

# Abstract

The clustering of chemical structures is important and widely used in several areas of chemoinformatics. A little-discussed aspect of clustering is standardization, it ensures all descriptors in a chemical representation make a comparable contribution to the measurement of similarity. The initial study compares the effectiveness of seven different standardization procedures that have been suggested previously, the results were also compared with unstandardized datasets. It was found that no one standardization method offered consistently the best performance.

Comparative studies of clustering effectiveness are helpful in providing suitability and guidelines of different methods. In order to examine the suitability of different clustering methods for the application in chemoinformatics, especially those had not previously been applied to chemoinformatics, the second piece of study carries out an effectiveness comparison of nine clustering methods. However, the result revealed that it is unlikely that a single clustering method can provide consistently the best partition under all circumstances.

Consensus clustering is a technique to combine multiple input partitions of the same set of objects to achieve a single clustering that is expected to provide a more robust and more generally effective representation of the partitions that are submitted. The third piece of study reports the use of seven different consensus clustering methods which had not previously been used on sets of chemical compounds represented by 2D fingerprints. Their effectiveness was compared with some traditional clustering methods discussed in the second study. It was observed that no consistently best consensus clustering method was found.

# Contents of Thesis

# List of Figures

# List of Tables

# Chapter 1 : Introduction

Drug discovery is a time-consuming and costly process. To bring a new drug to market, it generally takes approximately 15 years and costs approximately 800 million US dollars (DiMasi et al., 2003), and this reveals the complex process of drug discovery. This process typically involves dealing with vast amount of information to find compounds with desired properties, using techniques such as high-throughput screening and virtual screening. In addition, more than 53 million unique chemical substances are known and the number is growing rapidly (CAS, 2010). The complex and enormous information can only be operated by computer techniques.

In fact, computer technology has been applied to the pharmaceutical industry, especially in drug discovery, for many years. The development of chemoinformatics was well reviewed by Willett (2008). These techniques eventually resulted in a new discipline, chemoinformatics, which was first introduced by Dr. Frank K. Brown in 1998:

"*The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization.*"

Chemoinformatics is simply the use of information techniques to deal with the chemical data explosion and to solve chemical problems; it speeds up the process and increases the efficiency of drug discovery (Oprea, 2005). Cluster analysis is one of these information techniques that find application in chemoinformatics; it is extensively used to find the representative subsets from high-throughput screening and combinatorial chemistry for chemical datasets (Downs and Barnard, 2002). The focus of this thesis is on the method to group

2D chemical structures.

Much previous research in chemical clustering is on methods, implementation and applications, whereas we consider the following three new aspects in this thesis:

1. role of standardization, which has been little studied in the literature of chemical clustering, as one component of chemical similarity measures
2. evaluation of clustering methods which have not previously been considered for chemoinformatics applications
3. consensus clustering methods, which have not been applied to chemoinformatics applications

Chapter 2 ('An Introduction to Chemical Information') first introduces common and machine-readable representations of molecular structures, which are the basis for similarity-based chemical computing. Similarity measures are then discussed along with their crucial component, similarity coefficients. With these, cluster analysis on chemical structures can be carried out. An overall discussion of clustering is described in Chapter 3 ('Clustering'). The traditional Ward's and K-Means methods are widely used in chemical applications, and also used in this thesis. In addition, some novel methods which are reported to be effective in other applications are employed to compare with the traditional ones.

In Chapter 4 ('Experimental and Evaluation Methods'), we describe the chemical datasets and their representations, clustering methods and evaluation methods, which have been applied to the experiments of the following three chapters.

The aim of Chapter 5 ('Effect of Standardization on Three Different Representations of Structural Similarity') is to discuss the effect of standardization procedures on chemical clustering of structural representations. The initial study employs two traditional clustering methods, i.e. Ward's and K-Means; the extensive study in the second part of the chapter uses another seven clustering methods to obtain more generalized results.

Chapter 6 ('Comparison of Chemical Clustering Methods Using Fingerprint-based Similarity Measures') seeks to find the most effective clustering method for the application of fingerprint-based similarity measures, traditional and novel clustering methods are mixed together to investigate their performance. The clustering results are evaluated using four different criteria. A good clustering method should be able to satisfy as many evaluation criteria as possible.

Consensus clustering offers a way to combine different clustering results with more confidence. Chapter 7 ('Comparison of Chemical Consensus Clustering Methods Using Fingerprint-based Similarity Measures') is an extended study of Chapter 6. The results from different clustering methods are integrated into a consensus result, and then compared with the performance of the traditional Ward's method and the single best clustering method in Chapter 6.

Finally, Chapter 8 ('Conclusion and Future Work') summarizes the results of this thesis and offers some suggested directions of how this work can be extended.

# Chapter 2 : An Introduction to Chemical Information

## 2.1 Chemical Databases

Chemical databases store vast amounts of chemical information such as compound names, chemical structure representations, or molecular data; they may contain millions of entries for the purpose of search and retrieval. Hence, they enable users to search the interesting data in databases and obtain the results within seconds (Leach and Gillet, 2007; Paris, 2003). They provide an efficient and convenient manner of storing enormous amounts of chemical information.

There are varied types of chemical database. However, it depends both on the properties of chemical information to be stored such as reaction or patent, 2D or 3D structure, etc., and on the methods of data storage, for example the tables in a relational database or the objects in an object oriented database (Attwood and Smith, 1999). All these well-organized chemical databases play an essential role as a communication tool for chemists, and have been used for assisting chemists.

Chemists usually need to know how chemical databases may be used to solve their problems, the functions that chemical databases provide, and the efficiency and accuracy of the information that can be retrieved (Paris, 2003). There is a huge number of databases with varied chemical information that can be accessed on the Internet and these Internet chemical databases usually provide chemists with a friendly and a simple interface which enables users to retrieve information, providing a convenient, global networking and high-performance operating environment (Tarkhov, 2003).

### 2.1.1 The Importance of Chemical Databases

Over 53 million chemical compounds (CAS, 2010) have been reported. Moreover, there are also over one million new compounds per year and more than 500,000 publications each year that are concerned with chemical information (Marshall, 2005; Willett, 2007a). It is hard to deal with such a vast and constantly increasing amount of chemical data by non-electronic methods. Moreover, the variety of chemical information such as literature, chemical properties and spectra, can only be encompassed by storing them in electronic format. Hence the useful chemical information can be obtained only by accessing chemical databases.

The storage and searching of chemical structures are probably the earliest applications of chemical databases and these are an essential component of what many now call chemoinformatics (Gasteiger, 2003). Thus, chemoinformatics should support the chemists with their essential problems, which they meet in their daily work, and offer a platform for the necessary communication between theoretical sciences and experimental chemistry (Gasteiger, 2003). In short, chemical databases play an important role in chemoinformatics.

Chemical structure databases contain the computer-readable structure representations of a huge number of chemical molecules. Chemoinformatics provides a variety of tools that can be used for data mining in these databases, so as to assist directly in the discovery of new molecules. It plays a major role in drug discovery (Marshall, 2005). With the increasing costs on drug discovery, it is expected that more applications will be made of such tools. Furthermore, the advent of more effective software will enable more accurate predictions of activity, and thus will enhance the cost-effectiveness of research (Leach and Gillet, 2007).

The application and development of chemical structures can not only be applied in a similarity search (see Section 2.3.3) from the original collection or any other databases but also in the usage of identifying other related compounds. In addition, the application of 2D structures or 3D models may construct a pharmacophore, and then be used in a 3D search for models which may adopt relevant molecular conformations using a conformationally flexible search (Paris, 2003).

### 2.1.2 Examples of Chemical Databases

There are a variety of chemical databases, and their categories can be generally classified into literature, factual (alphanumeric) and structural types (Engel, 2003a). However, a common manner of classification of chemical databases is based on the properties of chemical data, such as chemical structure databases, organic and inorganic databases, spectroscopic databases, chemical reaction databases, environmental information databases, patent databases, biochemistry, molecular biology databases etc.. In addition, different types of database can also be integrated into one single resource providing more information, such as Chemical Abstracts Service (CAS). Some well known chemical databases are discussed in the following paragraphs.

The primary service of Chemical Abstracts Service (CAS) databases is the Registry File, which currently contains more than 53 million (CAS, 2010) substance entries including organic compounds, peptides, and a wide variety of other chemical information (Fisanick and Shively, 2003). Another service from CAS is the CAplus file, it contains more than 32 million (CAS, 2010) patents and journal article references in chemistry related fields. Also, the CAS Reaction Search Service (CASREACT) is a chemical reaction database containing 25 million single- and multi-step reactions which were derived from 750,000 records of journals and patents (CAS, 2010).

The Beilstein database was transformed from the Beilstein Handbook of Organic Chemistry. It is the most complete and systematic collection of evaluated data on organic compounds, and contains information on reactions, substances, structures and properties. Similar to CAS databases, the Beilstein database is also a large collection of different types of chemical information (Wiggins, 2003). The Cambridge Structural Database (CSD) was created and managed by Cambridge Crystallographic Data Centre (CCDC); it is used to represent the crystal structures of small organic and organometallic compounds, and contains crystal structure information for more than 500,000 organic and organometallic structures (CCDC, 2010) analyzed using X-ray or neutron-diffraction techniques (Engel, 2003a). The Protein Data Bank (PDB) currently contains over 65,000 (PDB, 2010) experimentally determined, X-ray and Nuclear Magnetic Resonance (NMR) structures of proteins and protein-ligand complexes. Both CSD and PDB are continuously increasing in size (Engel, 2003a; Homeyer and Reitz, 2003).

Probably, the most important application of chemical structure databases is structure retrieval, for example exact 2D structure and substructure search, 2D and 3D similarity search, 3D volume-based searching and docking.

### 2.1.3 Summary

The central role played by 2D chemical database systems is reflected in the significant amount of effort that has been expended to implement and optimize methods for the storage, search and retrieval of chemical structures and molecular data (Leach and Gillet, 2007). Besides, chemical structures also play an important role in the organization, indexing and access to the continually growing chemical literatures and compounds. The application can apply not only in chemical structures searching but also in chemical patent searching and reaction databases (Paris, 2003). They will, hence, continue to play a critical role in chemoinformatics and will remain vital in the future research.

## 2.2 Representation of Molecules

Chemical structures are the easiest notation for chemists but not for computers. Hence, for the purpose of searching chemical structures, a machine-readable structure representation is needed; therefore, it is necessary for searching methods to develop some machine-readable structure representations of the way in which the atoms and bonds of a molecule are connected together (Willett, 2003a). This is necessary for chemists to search for all compounds in chemical databases containing a specific structure or a particular substructure (Barnard, 2003).

Although chemical structure diagrams are the most common and the most natural means of communication for chemists, such graphical images are not suitable for the purpose of chemical information retrieval (Engel, 2003; Paris, 2003). Such structural images are of only limited usefulness in chemoinformatics and computational chemistry, and structure diagrams have to be represented in machine-readable forms. With these representations of chemical structures, molecules and compounds can be stored in a database for retrieval and search. Although chemical entities can be named according to varied naming schemes e.g. International Union of Pure and Applied Chemistry (IUPAC) convention, names are not ideal for chemical information retrieval because of the lack of flexibility in the representation (Paris, 2003; Willett, 1987). Hence, such naming schemes usually need to be converted into another type of representation. Different types of chemical representation for a compound are discussed in the following sections.

### 2.2.1 Representation of 2D Molecular Structures

There are a variety of structure representations which have been discussed in the literatures; three common types of molecular representation are systematic nomenclatures, linear notations and connection table, but only the latter two representations are used extensively in modern chemoinformatics (Willett, 1987; Willett, 2003). Systematic nomenclature represents a chemical structure as a unique alphanumerical string, however the relationship between compound

names and chemical structures is many-to-one, because many different valid compound names may refer to the same chemical structure. Hence, it is not suitable for some manipulations in chemical information systems. With such disadvantage and its complicated naming, it has some limitations in the development of chemical structure representations (Engel, 2003).

#### 2.2.1.1 Line Notations

Linear notations represent a molecular structure in the form of a linear sequence of alphanumeric characters. They are simple and compact, and hence are especially suitable for manipulation, such as storing and transferring large numbers of molecules, in a chemical information system (Leach and Gillet, 2007). There are varied types of linear notations discussed in the literature but only some of them are widely accepted and especially important: the Wiswesser Line Notation (WLN), Simplified Molecular Input Line Entry Specification (SMILES) and Sybyl Line Notation (SLN) (Engel, 2003; Willett, 2003). These traditional line notations describe chemical structures by alphanumeric strings mainly based on atomic symbols and bond types. However, a new and increasingly-used line notation, called InChI (IUPAC International Chemical Identifier), was proposed by IUPAC (International Union of Pure and Applied Chemistry) and NIST (National Institute of Standards and Technology) (McNaught, 2006). It characterizes chemical structures also by the manner of alphanumeric strings, but contains more information than traditional line notations, such as the atoms and their bond connectivity, tautomeric information, isotopic information, stereochemical and electronic charge information. Figure 2-1 is an example of phenylalanine represented by above four popular line notations (Engel, 2003; IUPAC, 2010).

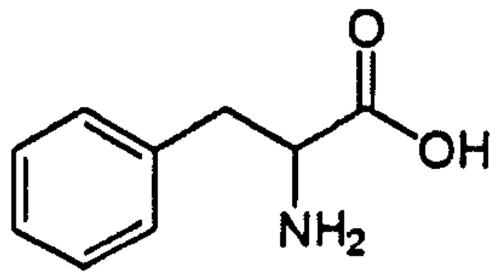
| Systematic Name: | Phenylalanine |
|---|---|
| Structure Diagram: |  |
| WLN: | VQYZ1R |
| SMILES: | NC(Cc1ccccc1)C(=O)O |
| SLN: | C[1]H:CH:CH:CH:CH:C(:@1)CH2CH(NH2)C(=O)OH |
| InChI | 1/C9H13NO.CH2O/c10-9(7-11)6-8-4-2-1-3-5-8;1-2/h1-5,9,11H,6-7,10H2;1H2 |

Figure 2-1 Example of various line notations of phenylalanine

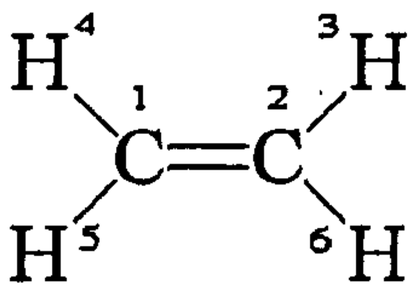
SMILES notation was used as the input chemical representation to convert into other file formats for the studies in Chapters 6 and 7. We hence discuss SMILES in the following paragraphs. It was proposed by Weininger (Weininger, 1988), and uses a few simple rules to build chemical structures by alphanumeric strings of characters based on atomic symbols; relative to WLN, that is also the reason why it is extensively accepted and widely used. With some significant rules of SMILES notation, atoms are represented by their atomic symbol, but hydrogen atoms are normally omitted, for SMILES is a hydrogen-suppressed notation (Engel, 2003).

There may be a variety ways to form the SMILES string for a given molecule, since, in SMILES notation, the string may be written by a different starting atom resulting in a different sequence. Hence, several SMILES strings may represent the same chemical structure. To get rid of the disadvantage of ambiguity, a method called the Morgan algorithm for generating a canonical sequence of the atoms has been widely used (Morgan, 1965). The other well-known technique called CANGEN algorithm has been devised to create a unique SMILES string for each molecule in the chemical databases (Weininger et al., 1989), and this unique SMILES string is usually termed Canonical

SMILES. It provides the simplest linear code; hence it is highly compact and easy to learn. Moreover, the fast data exchange format and unambiguity are also advantageous.

### 2.2.1.2 Connection Tables

Connection tables are the most significant format of chemical structure representation in a computer system and are also an alternative manner of representing molecular graphs (Engel, 2003). A connection table is a 2D matrix containing information about all the atoms and bonds in a 2D structure. In comparison with SMILES notation, a connection table provides the same information but in a different form; each row lists information about a particular atom such as the atom number, symbol, and number of atoms to which it is directly bonded and their bond types. A common example of connection tables is Tripos *mol* and *mol2* file format (Tripos, 2007). Figure 2-2 illustrates a simple example of connection table of ethylene (Engel, 2003). Each atom is numbered arbitrarily as an index forming an atom list; moreover each row in the bond list shows the indices of two atoms connected by a particular bond type (1 indicates single bond, 2 indicates double bond, analogically).




SMILES: C=C

Compound Name: Ethylene

| Atom List | |
|---|---|
| 1 | C |
| 2 | C |
| 3 | H |
| 4 | H |
| 5 | H |
| 6 | H |

| Bond List | | |
|---|---|---|
| 1st atom | 2nd atom | Bond Type |
| 1 | 2 | 2 |
| 1 | 4 | 1 |
| 1 | 5 | 1 |
| 2 | 3 | 1 |
| 2 | 6 | 1 |

Figure 2-2 Example of the connection table of ethylene

Connection tables are the most commonly used representation of chemical structures. However, many varied types of connection table have been discussed in the literature; thus, there are also translation programs to convert between the different forms. Most chemical software can exchange and store these as external connection tables. SMILES notation, and molecular fingerprints also can be generated from connection tables (Engel, 2003; Willett, 1987).

Connection tables are unambiguous because they offer a detailed and exact description of the topology of the compound that they represent but they are not unique. Thus, a specific molecule could be represented by different connection tables (Willett, 1987), because in a connection table the users can choose a different order to number the atoms. To find the unique identity by renumbering one of the connection tables in all possible types will be an important function. For instance, the Morgan algorithm (Morgan, 1965) is a widely used method to generate a unique order of the atoms. Since the connection tables involve a complete representation of the inter-connections between the atoms in a molecule, they can be considered as a labeled graph. Connection tables are particularly suitable for manipulation of such topological information, such as structure search, substructure search, and graphical structure input and output.

### 2.2.2 Representation of 3D Molecular Structures

There are more than 500,000 compounds whose 3D structures have been stored by the Cambridge Crystallographic Data Centre (CCDC, 2010), but such a number is really small when compared with the number of known compounds, which is over 53 million (CAS, 2010). Moreover, the experimental sources of 3D structures are not sufficient and there is an essential demand for computer-generated models. Some theoretical techniques such as quantum mechanics or molecular mechanics have good performance both on producing 3D molecular models and predicting a number of molecular attributes. These methods, nevertheless, still need at least some rational 3D geometry of the

molecule to be carried out.

There are two widely used methods for representing a 3D chemical structure. The major difference of these two methods is that they use different coordinate systems to characterize the spatial arrangement of the atoms of a molecule of interest. The first and common method is to store each atom in a molecule as their three space coordinates, x-, y- and z-coordinate values. It represents the 3D feature and conformation of a molecule. Such connectivity information or coordinate values can be collected either implicitly by approximating bonding distances between the atoms, or explicitly by a connection table. The other method uses internal coordinates, such as bond length, bond angles, and torsion angles to represent the 3D structure of a molecule. Such representations describe the spatial arrangement of the atoms relative to each other (Engel, 2003).

Automatic 3D structure generation, the transformation of a 2D connection table into a 3D molecular model, has become a standard technique commonly used in many fields of computational chemistry. Much research has focused on making these 3D structure generators as rapid as possible in order to apply them to large datasets of molecules (Sadowski, 2003). Since the useful representation of 3D structures can be transformed from 2D methods, it may be a better method to devise an efficient 2D method and then transform appropriately to its 3D usage.

### 2.2.3 Molecular Descriptors

Molecular descriptors are numerical values resulting from a procedure which transforms the structural information encoded within a symbolic representation of a molecule to describe properties of molecules (Leach and Gillet, 2007). With the use of molecular descriptors, it becomes possible to manipulate and analyze the chemical structural information very easily. Molecular descriptors, for example, may represent the physicochemical features of molecules that may be calculated by applying algorithmic techniques to the molecular structures. Many different molecular descriptors have been described and used

for a wide variety of purposes; they can be classified by the data type, such as Boolean, integer or real number, vector etc. of the molecular descriptor and the molecular representation of the compounds (Terfloth, 2003). The major difference of varied descriptors is the complexity of the information they encode and in the time required to calculate them (Leach and Gillet, 2007). However, the selection of the appropriate set of molecular descriptors is often the key to success.

Here, we concentrate on the three common types of descriptor that have been used in similarity search (as discussed in Section 2.3.3): whole-molecule descriptors, 2D descriptors and 3D descriptors (Willett & Gillet, 2007). The whole-molecule descriptors are the simplest, they describe a molecule by some simple properties such as molecular weight and *logP*, but a single descriptor is usually insufficient to find the similarity between a pair of molecules. Hence, it is normal to use several different types of descriptors together for similarity searching. Topological indices and fragment-based indices are two common types of 2D descriptor which can be generated from 2D molecular representations.

A topological index is a single number that encodes a molecular structure by its basic properties such as size and shape. With describing such simple properties, a combination of varied topological indices is usually used for similarity searching as in whole-molecule descriptors (Willett & Gillet, 2007); this is described in more detail in Section 2.4.1. Fragment-based descriptors characterize a molecule by the substructural features. Among varied types of 2D descriptor, 2D fingerprints are the most widely and commonly used descriptor for similarity searching, and were originally devised for substructure searching. They are considered one of the earliest similarity searching methods in the literature by Willett et al. (1998).

Fingerprint encoding is the process of transforming a chemical structure into a binary format, they capture the topological features of chemical compounds and convert them into a linear, binary string format which identifies the presence or absence of specific structural features in a chemical compound

(Eckert and Bajorath, 2006). There are a number of ways to generate fingerprints from chemical structures, however all these techniques generally have been categorized into two different types of 2D fingerprints: dictionary-based fingerprints, and hashed fingerprints (Flower, 1998; Leach and Gillet, 2007).

In dictionary-based fingerprints, a structural fragment dictionary is required, which contains typically from hundreds to thousands of structural fragments for 2D fingerprints and millions of structural fragments for 3D pharmacophore fingerprints (Xue et al., 2003); and such a dictionary will be used to determine whether each bit in the binary string is set or not. Each bit usually maps to a certain substructure fragment or structural feature in a predefined fragment dictionary. Hence, if a certain feature is present in a molecule, then the bit which corresponds to it will be set to '1'; otherwise it will be set to '0'. Thus, fingerprints transform the presence or absence of certain features within a molecule into a binary bit string. One limitation of dictionary-based fingerprints is that the optimum fragment dictionary is dataset dependent; another is that they are sparse, since most of the bits in the bit string are set to '0', sometimes a typical molecule has only a few fragments for the bit positions to represent.

On the other hand, hashed-based fingerprints do not need a predefined fragment dictionary, and are a very dense representation of the structural features in a molecule, typically capturing all possible connectivity pathways through a molecule up to a certain and defined path length. So, a molecular fingerprint is generated from a hash of all the unique connection paths, up to a certain maximum size which is predefined, into a fixed length bit string, and any fragment present in the molecule will be encoded in the fingerprint, (Willett & Gillet, 2007). Hashed fingerprints generate the bit patterns which are highly characterized, but several different fragments may set the same bit, that is the relationship between bit position and fragment is not one-to-one as in dictionary-based fingerprints. Therefore it becomes impossible to map from a bit position back to a unique fragment; that is, single bit positions no longer

correspond to specific structural features, and this leads to the possibility of ambiguity (Eckert and Bajorath, 2006).

Descriptors that can be generated from 3D molecular representations include basic fragment-based descriptors and also more complicated representations that describe molecular properties such as 3D shape and electrostatic fields (Willett & Gillet, 2007). 3D fingerprints were originally devised for substructure search as for 2D fingerprints; eventually they have been used for similarity searching. The 3D fingerprints describe the conformational features of molecules, such as interatomic distances and angles, by recording the absence or presence of specific 3D features. With making use of molecular descriptors, there are a wide variety of further applications of computational chemoinformatics, such as QSAR, data analysis, similarity searching and calculation, techniques for selecting diverse compound sets etc. (Leach and Gillet, 2007).

## 2.3 Some Common Searching Methods

When a new compound is added into the large chemical database, a structure search technique is required to ensure that the compound is really a new one, and it should not exist already. There are three major types of searching in chemical databases for structures: exact structure searching, substructure searching and similarity searching (Paris, 2003). Each of these types of search employs different methods because they are aiming to retrieve different types of information.

Generally speaking, all types of systems for retrieving information from a variety of databases will basically provide three different searching modes (Willett, 2003a): exact-match, partial-match, and best-match. These three modes are equivalent to structure searching, to substructure searching and to similarity searching respectively, in the chemical context.

### 2.3.1 Exact Structure Searching

Exact structure searching is the simplest chemical retrieval technique; it involves the retrieval of all entities in chemical databases that match exactly and completely a structure of interest. It involves simply identifying the presence or absence of a specific molecule in a database and will be efficient if a canonical notation has been devised (Willett, 2003a).

The canonical representation is significant for exact structure searching, and it must be unique otherwise that would be problematic. However, a hash function is usually associated with the canonical representation to accelerate structural retrieval such as finding items in a database (Leach and Gillet, 2007).

### 2.3.2 Substructure Searching

Substructure searching is probably the most widely used technique and it is the process of identifying parts of a given structure that are equivalent to a specified query substructure (Leach and Gillet, 2007); it identifies all the molecules in the database that contain a specified substructure. A two-stage mechanism is usually used in substructure searching. First, a screen search is executed to eliminate those substructures that cannot possibly match the query and to generate a subset of the database which might possibly match the query. Second, each molecule in the subset will pass through a detailed atom-by-atom graph matching search to decide whether a subgraph isomorphism does exist for the substructure of interest. Such atom-by-atom matching procedures are very time-consuming (Willett, 1987).

There are some restrictions of substructure searching. First, the users require sufficient knowledge in order to construct a meaningful substructure, and this knowledge is not always available. Second, the users have only limited control over the size of the searching results: that is, a generic query can result in a huge amount of hits, but a very specific query may retrieve only a very small number of hits (Leach and Gillet, 2007).

The substructure searching technique is usually the first step in the implementation of other important topological procedures for the analysis of chemical structures, such as identification of equivalent atoms, determination of maximal common structure, ring detection, calculation of topological indices, etc. (Kochev at. el., 2003).

### 2.3.3 Similarity Searching

Similarity searching provides a complementary, alternative technique to exact searching or substructure searching. It involves comparing the query with every compound in the database and retrieves objects that are similar to a query, sorted in order of their decreasing similarity (Kochev at. el., 2003).

There are several advantages of similarity searching when compared to substructure searching. First, one does not need to define a precise substructure query, since a single active compound is sufficient to undertake a search. Second, users are able to manage the size of the output because every compound in the database is given a numerical score, which is calculated by a similarity descriptor. So it can be used to generate a complete ranking. Alternatively, users can specify a particular value or level of similarity and retrieve just those compounds that exceed the threshold. Finally, similarity searching facilitates an iterative approach to searching chemical databases since the top-scoring compounds resulting from one search can be used as queries in subsequent similarity searches (Leach and Gillet, 2007).

## 2.4 Molecular Similarity Methods

Substructure searching is the major technique for retrieving information from chemical structure databases, however the focus on such retrieval techniques is increasingly transferring to similarity searching (Willett, 2003a). There are many similarity methods in the literature, and each single method has its application on certain query and biological activity. By evaluating the results from a single experiment, it is difficult to find a similarity method that is the best and also will be superior in other type of query and activity (Sheridan & Kearsley, 2002). Sheridan and Kearsley (2002) therefore, argued that the

combination of different similarity methods may be needed and the same method with several variations to get sufficient information to form a query as well.

The effectiveness of a similarity measure, in terms of its ability to retrieve bioactive molecules, is usually a crucial factor in similarity searching and some research has concentrated on the key components of a similarity measure that influence the effectiveness of similarity searching. There are usually three crucial components when computing the similarity between a pair of objects and each component can affect the effectiveness on similarity searching. The first component is the representation. An appropriate structural representation must be picked and be used to describe the molecules that are being compared. The second component is the weighting scheme. It is used to allocate different levels of significance to the varied components of representations, that is, important molecular features and less important ones can be distinguished. The final component is the similarity coefficients. They are used to determine the degree of resemblance between a pair of representations of chemical structures. Overall, the first component is the most important, since the representation can influence very strongly the manipulations that are possible and appropriate when calculating the similarity between a pair of molecules (Willett, 1987; Willett, 2003a).

### 2.4.1 Similarity Searching in 2D Databases

Similarity techniques for searching chemical databases were proposed initially in the mid-1980s (Willett et al., 1986), and their effectiveness usually causes users most concern and is usually a key factor on similarity searching. Some research has paid attention to the crucial components of a similarity measure that influence the effectiveness of similarity searching (Willett & Gillet, 2007). The similarity score is the basic component on similarity searching. For calculating the similarity value, there are three major types of representation which have been used to measure the degree of resemblance between two chemical structures of 2D databases. These are based on fragment substructures, on topological indices, and on maximum common subgraphs (Willett, 2003a).

Fragment substructures were originally devised for the representation of chemical structures but they are imprecise, as they do not encode how the fragments are linked together. Hence their usage then became common in the initial screening stage of 2D substructure searching and then they have eventually been applied to similarity searching. In similarity searching, fragment substructures are usually encoded as binary vectors or bit strings that are based on a pre-defined fragment dictionary or fingerprints. Similar to the nature of binary fingerprint encoding (discussed in Section 2.3.3), a bit is set to '1' indicating a certain feature or substructure is contained, and otherwise a bit is set to '0'. If the bit strings representing two molecules have a large number of fragment substructures in common, then these two molecules will have a high similarity (Willett, 2003a).

As molecular descriptors characterize properties of a molecule, topological indices describe more specific information on molecular structures. It is normally a single numeric value that can be generated from 2D representation of a molecule (Hall & Kier, 2001). A great number of varied topological indices have been devised in the literature. The general types of topological indices encode structures by their size, degree of branching such as electronic information based on the paradigm, and overall shape. For example, one of the most common indices is the molecular connectivity indices.

In brief, topological indices characterize the structures according to their topological properties such as size, amount of branching, amount of unsaturation and other complicated features. With the similarity calculation using topological indices, it usually needs to operate with many different indices, and then it uses a multivariate method, such as principal components analysis (PCA), to generate a smaller number of uncorrelated variables (indices) to encode all the molecules, i.e. using a smaller number of principal components to replace those indices with high correlation on some particular properties. All of these varied indices that can describe the molecular features have not only been widely used in 2D similarity searching but also increasingly in 3D. (Willett, 2003a)

Most similarity measures such as measures based on fragment substructures and topological indices are global similarity measures; they do not identify the resemblance of local areas but overall similarity between two molecules. Willett (2003a) concluded that some local similarity measures, graph-based approaches, such as maximum common subgraph (MCS) are not only an alternative but also an effective method for similarity-based virtual screening, and can carry out feature mapping between two molecules. The similarity calculation of local regions is operated by creating a mapping from the atoms of one molecule on to another. With structural diagram representations, graph matching techniques can easily be used with both 2D and 3D representations for identifying the MCS. The MCS techniques are devised to find the subgraph that is the largest set of atoms and bonds, including inter-atomic distances in the 3D case, in common or shared between two molecules. Furthermore, the number of atoms and bonds in the MCS can be used to calculate a Tanimoto-like coefficient that quantifies the degree of similarity between two molecules (Willett, 2003a; Willett & Gillet, 2007).

### 2.4.2 Similarity Coefficients

A similarity coefficient is used to quantify the degree of resemblance between pairs of objects; each object can be described by some number of attributes or descriptors (Holliday, 2002; Willett & Gillet, 2007). Similarity coefficients are used in a wide range of disciplines such as, biology, information retrieval, multivariate statistics, numeric taxonomy and marketing (Willett et al., 1998).

With the wide usage of similarity coefficients in different disciplines, there is a shortage of the canonical forms of coefficients. Hence, some similarity coefficients have been re-devised with different names, and many of them are closely related to each other. For example, some pairs of coefficients are different when they are used to manipulate continuous attributes but they become equivalent when they manipulate binary attributes (Willett et al., 1998). For example, on measuring similarity with binary variables, the Tanimoto similarity coefficient is equivalent to the Soergel distance, since the Soergel

distance is the complement of the Tanimoto coefficient (Leach and Gillet, 2007).

There are various types of similarity coefficient and the detail has been discussed in the reviews by Holliday et al. (2002; 2003), moreover three types are commonly discussed in the literature as follows: distance coefficients, association coefficients, and correlation coefficients (Holliday et al., 2002; Willett, 1987). The first two classifications, distance and association, are commonly used for similarity searching. Distance coefficients are a widely used type of similarity measure because their geometric representation is simple. Two well-known distance coefficients are the Euclidean distance and Hamming distance (Holliday et al., 2002; Willett, 1987). As for association coefficients, the Tanimoto coefficient is the most widely used similarity coefficient. It can be used for both continuous attributes and binary attributes. With continuous attributes such as topological indices, the value of the data may be real numbers over any range. While with binary attributes such as 2D fingerprints, the data are coded as 0 or 1 denoting respectively the absence or presence of specific substructure features. 2D fingerprints in combination with the Tanimoto coefficient provide a simple but effective way of quantifying the similarity relationships between pairs of molecules (Leach and Gillet, 2007).

For example, the similarity between two binary bit strings *A* and *B* (denoted by $S_{ab}$) can be computed by the commonly used Tanimoto coefficient which is represented as follows:

$$S_{AB} = \frac{c}{a+b-c}$$

where *a* is the number of bits set to "1" in bit string *A*, and
*b* is the number of bits set to "1" in bit string *B*, and
*c* indicates the number of bits set to "1" in both *A* and *B*

bit string A: 0 1 0 1 0 0 1 1 0 0 a=4
bit string B: 1 0 0 1 0 0 1 1 1 0 b=5, and c is 3

$$S_{AB} = \frac{3}{4+5-3} = 0.5$$

Figure 2-3 Example of calculating similarity based on Tanimoto coefficient

Different types of coefficients calculate similarity in various ways. For example some coefficients, such as the Tanimoto coefficient and the Dice coefficient, compute similarity directly. Others, such as the Hamming coefficient and the Euclidean coefficient, generate the distance or dissimilarity between pairs of molecules. Moreover, in the case of binary attributes, some coefficients such as Tanimoto generate a real number within the range from zero to one but others such as Euclidean provide a wider range from zero to infinity. Hence, a standardization procedure is required to convert the attribute value between similarity and distance coefficients. When the attribute values are limited to the range from zero to one, the measure used for different similarity and distance measures is simplified and standardized (Holliday et al., 2002; Leach and Gillet, 2007).

In addition to the normalization on attribute values for different coefficients mentioned above, the molecular size may also affect the calculation of similarity especially on the representation with binary fingerprints. For

example, the Tanimoto, Dice and Cosine coefficients directly compute the similarity according to the number of bits in common. On the other hand, the Hamming and Euclidean distances also calculate the similarity by the common absence of molecular features. Hence, the common presence or absence of molecular features will influence the similarity score (Leach and Gillet, 2007).

In some cases, the molecular size will directly influence the calculation of similarity measures by association coefficients such as Tanimoto coefficients (Holliday et al., 2003; Haranczyk and Holliday, 2008). They cause a bias of similarity calculation on different size of molecules. For example, in a similarity measure using fingerprints such as Tanimoto coefficients, the small molecules will usually have lower similarity score or larger distance value since they are likely to have fewer bits set in a fingerprint than large molecules. Conversely, when using the Hamming distance, small molecules tend to be more similar (Leach and Gillet, 2007). With such bias of coefficients on small molecules and larger molecules, it also requires some degree of size standardization to avoid such problem.

Even for a particular application of chemoinformatics, it should not be considered that a certain coefficient will always give better performance than others (Willett et al., 1998; Willett, 2003a), and some research has suggested that using mixed indices which combine two or more standard measures may have better performance on similarity searching (Leach and Gillet, 2007). Eventually, it might be true that there is still a need to find the most appropriate coefficient or combination of coefficients for any specific similarity searching application. Holliday et al. (2002) combined different coefficients for similarity searching using the application of data fusion. Different combinations of similarity coefficients were employed and the performance with the individual coefficients was compared; thus, the technique of data fusion has been shown to improve the performance of similarity searching.

### 2.4.3 3D Similarity

It is natural that there are differences between 2D molecular features and 3D, hence 3D similarity measures need different molecular properties such as conformational properties to be considered and are more complicated computational processes than 2D methods. The 2D similarity methods have been developed earlier than 3D methods and they are also the standard retrieval principles at present. Hence 2D methods have widely been developed as the fundamental principles for 3D methods. For instance, 3D substructure searching fingerprints can be used for similarity searching as well as 2D fingerprints.

There are some common 3D methods which have been discussed in literature, for example, the 3D equivalents of fragment and MCS methods, and the alignment methods based on molecular field information. However, some literature simply divides 3D similarity measures into two categories (Leach and Gillet, 2007; Willett & Gillet, 2007): alignment methods that are implemented by manipulating the molecules in 3D space and alignment-independent methods that do not need such geometric spatial information to be derived.

As mentioned above, 3D fingerprints were originally applied to 3D substructure searching and then to similarity searching like 2D fingerprints. But the major difference is that the molecular features, such as spatial characteristics of conformation that 3D fingerprints encoded are more complex than 2D fingerprints. The fingerprint can encode the presence or absence, or the frequency of occurrence of 3D molecular features. 3D molecular descriptors, such as inter-atomic distance, valence and torsion angles, and atom triplets, can be represented in a binary fingerprint similar to a 2D fingerprint and then be used by Tanimoto coefficients. Although, such manipulations of 3D fingerprints are simple, when the conformation flexibility has been involved, the calculations of all descriptors are quite time consuming (Leach and Gillet, 2007; Willett, 2003a).

The 3D graph-matching approaches can also be derived from 2D such as 2D MCS. The principle of 3D MCS is similar to 2D; it creates a mapping from the atoms of one molecule on to another and finds the largest set of atoms which match the distance between atoms. The similarity calculation is still time-consuming. As for the alignment methods, they take the degrees of freedom related to the conformational flexibility into account. They mainly arrange the alignment of two or more molecular structures, and the comparison between them is based on their shape and 3D confirmation (Willett, 2003a; Willett & Gillet, 2007).

The development of many varied 3D methods is currently at an early stage and there is still a need to find an efficient method on 3D similarity searching since most of their manipulations are time consuming or some factors such as conformational flexibility involving in the similarity calculation will be complex (Willett, 2003a).

## 2.5 Summary

There are many ways in which we can calculate the similarity between pairs of molecules, but the great majority of current similarity-searching systems employ simple 2D fragment-based measures. The applications of the similarity measures include chemical database clustering, reaction similarity searching, and the analysis of molecular diversity (Willett et al., 1998). One very important application of similarity measures is cluster analysis, it is discussed in the next chapter.

# Chapter 3 : Clustering

Cluster analysis, or clustering, in the most general sense of the term, is a process of partitioning which divides data into a number of groups, so data in one group are similar and data in different groups are not similar (Halkidi et al., 2001; Kantardzic, 2003; Milligan and Cooper, 1987). Clustering is a technique for exploratory data analysis and is used increasingly in preliminary analyses of large datasets of medium and high dimensionality as a method of selection, diversity analysis, and data reduction (Downs and Barnard, 2002). The literature is full of discussions surrounding the applications of cluster analysis, and that is also the evidence of its importance. With the increasing and continuing uses of cluster analysis in many research fields, a number of varied definitions have been proposed in the past several decades, however the favorite definition may be given according to the discipline involved and the aim of the researchers (Punj and Stewart, 1983). There are many synonyms of cluster analysis such as unsupervised learning, numerical taxonomy, typology, partition (Halkidi et al., 2001), automatic classification (Willett, 1985), unsupervised classification (Kantardzic, 2003), and unsupervised pattern recognition (Everitt, 2001).

Some reviews regard cluster analysis as a specific mode of classification (Dunham, 2003). Clearly, cluster analysis may differ in a number of ways from classification. For example, in contrast to classification, cluster analysis has no predefined classes and no examples to show the relations among samples, that is, there is no prior knowledge concerning the clusters, yet classification allocates a data item to a predefined set of categories. On the other hand, the results of clustering are dynamic. It follows from what has been said why cluster analysis is viewed as an unsupervised process (Halkidi et al., 2001).

## 3.1 The Key Components of Clustering

Cluster analysis may be of crucial importance in a wealth of applications in many disciplines such as business and science, and is one of the most useful tools for discovering patterns in the underlying data. Several studies (Everitt et al., 2001; Halkidi et al., 2001; Punj and Stewart, 1983) have proposed the fundamental functions of cluster analysis such as the following: prediction based on groups, hypothesis generation and testing, and data reduction and exploration.

A cluster analysis encompasses a sequence of processes. The sequence shows the important processes or decisions which have to be made in a cluster analysis. Sometimes, it may be necessary to adjust the processes in a sequence to fit a specific application in a certain research field. However, it is also important for the user to recognize that key decisions have been made. Although it may seem preferable when the user has no prior knowledge or even positive information to make a selection, it cannot be assumed that the original selection is optimal or even correct (Milligan, 1996).

The key processes in clustering can be summarized as follows (Everitt et al., 2001; Halkidi et al., 2001; Ketchen and Shook, 1996; Punj and Stewart, 1983):

### 3.1.1 Weighting Variables and Standardization

Choosing and weighting clustering variables for grouping objects are two of the most troublesome processes in the application of cluster analysis, and thus, perhaps the most important (Gnanadesikan et al., 1995; Ketchen and Shook, 1996). In addition, in many applications the variables that describe the objects to be clustered will not be measured in the same units or scales. Indeed they may often be variables of completely different types, and yet others having an interval scale. Thus, a simple standardization is needed.

A standardization process allows variables to contribute equally to the definition of clusters but may also eliminate the meaningful and important differences among elements (Ketchen and Shook, 1996). Whether to standardize clustering variables is an ambiguous issue. Some studies report standardization is needed to eliminate the potential effects of scale differences among variables. Others offer experimental evidence that standardization has no significant effects or generates limited improvement (Bath et al., 1993; Ketchen and Shook, 1996). Aldenderfer and Blashfield (1984) suggested that since standardizations may generate adverse effects, it should be carried out based on a case-dependent basis. Milligan and Cooper (1988) investigated a study of eight different standardization methods in the cluster analysis and reported that standardization techniques based on division by the range of observations were consistently superior to any other standardization approaches. Conversely, Gnanadesikan et al. (1995) highlighted the drawbacks of weighting based on the standard deviation or range of variables.

### 3.1.2 Selection of Similarity or Dissimilarity Measures

As discussed in Section 2.4.2, a similarity or dissimilarity measure is not only important for similarity searching but also critical to the application of cluster analysis. These measures reflect the degree of similarity or diversity between objects, a clustering hence can be carried out based on it. No single coefficient is applicable to all applications, and different similarity measures generate various clustering results. This reflects the importance of choosing an appropriate similarity measure for a particular application. A dissimilarity measure, such as distance, assumes larger values as two objects become less similar. Whereas a similarity measure, such as correlation, assumes larger values as two objects become more similar. The Tanimoto coefficient and Euclidean distance are two well-known and widely used measures for similarity and dissimilarity respectively.

Table 3-1 shows some commonly used similarity and distance coefficients in chemical application (Willett et al., 1998). In which, $S_{AB}$ denotes the similarity between $A$ and $B$, and $D_{AB}$ indicates the distance between $A$ and $B$. In addition, $i$ represents the attribute, and the $N$ is the number of attributes. As for binary variables (e.g. fingerprints), $a$ is the number of bits set to "1" in $A$, while $b$ is the number of bits set to "1" in $B$, and $c$ is the number of bits set to "1" in both $A$ and $B$.

| | Formula for continuous variables | Formula for dichotomous variables |
|---|---|---|
| Cosine Similarity | $S_{AB} = \frac{\sum_{i=1}^{N} x_{iA} x_{iB}}{\left( \sum_{i=1}^{N} (x_{iA})^2 \sum_{i=1}^{N} (x_{iB})^2 \right)^{1/2}}$ | $S_{AB} = \frac{c}{\sqrt{ab}}$ |
| Tanimoto Coefficient | $S_{AB} = \frac{\sum_{i=1}^{N} x_{iA} x_{iB}}{\sum_{i=1}^{N} (x_{iA})^2 + \sum_{i=1}^{N} (x_{iB})^2 - \sum_{i=1}^{N} x_{iA} x_{iB}}$ | $S_{AB} = \frac{c}{a+b-c}$ |
| Hamming Distance | $D_{AB} = \sum_{i=1}^{N} \lvert x_{iA} - x_{iB} \rvert$ | $D_{AB} = a+b-2c$ |
| Euclidean Distance | $D_{AB} = \left( \sum_{i=1}^{N} (x_{iA} - x_{iB})^2 \right)^{1/2}$ | $D_{AB} = \frac{a+b-2c}{a+b-c}$ |

Table 3-1 Some commonly used similarity and distance coefficients

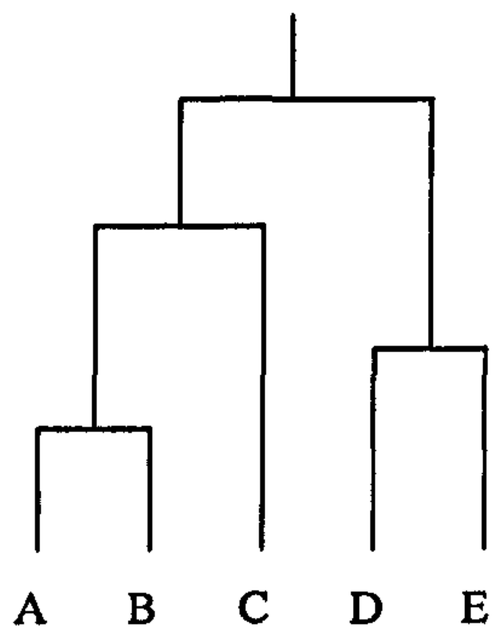
### 3.1.3 Selection of Clustering Methods

The selection of appropriate clustering methods is an important process for effective clustering (Punj and Stewart, 1983). An efficient good clustering method is definitely superior to an inefficient bad one; however researchers have to determine the choice between an efficient bad clustering method and an inefficient good one; besides, each clustering method has its suitability on certain areas, hence the decision of these considerations may depend on the demands of users.

Two types of clustering methods are common in the literature: hierarchical and non-hierarchical methods, which are discussed in Section 3.2. With distinct clustering approaches, each of them has its suitable application and limitation. For example, the Jarvis-Patrick method was reported to be suitable for chemical application rather than other fields. Some non-hierarchical methods usually require a prior setting before clustering, for example a user-defined number of clusters for K-Means method or a pre-determined $k$ nearest neighbours for Jarvis-Patrick method, whereas there is no such requirement for hierarchical methods. In addition, some methods are suitable for dealing with large datasets, such as CLARA. Some studies (Milligan, 1980; Punj and Stewart, 1983) proposed that the combination of hierarchical and non-hierarchical methods offers better performance; these use hierarchical methods to determine the number of clusters and the cluster centroids, and then use non-hierarchical methods based on these results. However, the shortcoming is the extra cost of time and effort.

### 3.1.4 Decision on the Number of Clusters

A prior assignment of the number of clusters is needed when the non-hierarchical methods are carried out, but not for hierarchical methods (Punj and Stewart, 1983). The hierarchical relationship in hierarchical clustering may be represented by a dendrogram, which represents the fusions or divisions made at each continuous stage of the analysis. The visual examination of a dendrogram is a commonly used and a basic technique to decide the number of clusters in dealing with hierarchical clustering (Ketchen and Shook, 1996; Leach and Gillet, 2007). Figure 3-1 illustrates an example of a dendrogram and the members of clusters in the hierarchical relationship.




Members of Clusters

[A, B, C, D, E]

[A, B, C]; [D, E]

[A, B]; [C]; [D, E]

[A, B]; [C]; [D]; [E]

[A]; [B]; [C]; [D]; [E]

Figure 3-1 Example of dendrogram and the members of clusters

Neither hierarchical nor non-hierarchical clustering methods directly address the issue of determining the number of groups within the data. Different techniques have been reported for determining the number of clusters on hierarchical and non-hierarchical clustering methods (Dubes, 1987; Fraley and Raftery, 1998; Milligan and Cooper, 1987) and their experimental results concluded some techniques are effective. Ketchen and Shook suggested (1996) that multiple techniques should be used to determine the number of clusters, rather than using a single approach, in order to get rid of the drawbacks of each other.

The partition size for some clustering methods could be determined by a cut-off parameter or threshold, such as the CAST (Ben-Dor et al., 1999) and Yin-Chen (Yin and Chen, 1994) methods. However, in some cases, the partition size is sensitive to the threshold setting.

### 3.1.5 Validation and Interpretation of Results

Validation of clustering results is also one of the critical processes in cluster analysis because no clustering method assures offering superior performance even dealing with the datasets with no error or noise (Milligan, 1980). Interpretation of the clusters within the applied context requires the knowledge and expertise of the researcher's particular discipline (Halkidi et al., 2001).

### 3.1.6 Summary

Clustering methodology has been increasingly proposed and widely used in a variety of research fields such as, archaeology, astronomy, biology, computer science, electronics, engineering, information science, and medicine. Detailed review and general introductory texts on the topic of clustering were summarized by Milligan and Cooper (1987), Everitt et al. (2001), and Jain et al. (1999). In terms of its application in varied disciplines, there are some good reviews in a variety of areas such as marketing (Punj and Stewart, 1983), economics (Dunham, 2003), information retrieval (Willett, 2005), image segmentation, computer science, and data mining (Berkhin, 2002). In addition, as for chemical application, excellent review articles on the application to chemical data were summarized by Barnard and Downs (1992), Downs and Willett (1994), Willett (1987), and Downs and Barnard (2002). The importance of clustering in many disciplines is evident through its enormous literature and application in wide range of areas (Kantardzic, 2003).

## 3.2 Clustering Methods

It is important to distinguish a cluster analysis from a clustering method. A cluster analysis may refer to the overall sequence of processes that were discussed in section 3.1. Nevertheless, the clustering method represents a very important process in the cluster analysis.

Halkidi et al. (2001) proposed three criteria for the classification of clustering algorithms as follows: the type of data input to the algorithm; the clustering criterion defining the similarity between data points; and the theory and fundamental concepts on which clustering analysis techniques are based. For each clustering method, the type of variables used in the dataset can be generally classified into numeric data and categorical data.

Several clustering methods have been proposed in the reviews. However, with diverse algorithms on the basis of applied fields, the classification of clustering methods varies. Clustering methods can be generally classified into two popular categories, hierarchical and non-hierarchical clustering (Downs &

Barnard, 2002; Kantardzic, 2003; Willett, 1987).

### 3.2.1 Hierarchical Clustering

Hierarchical clustering methods create a cluster hierarchy. In other words, they organize data in a nested sequence of groups, which can be displayed in the form of a dendrogram or a tree-like structure (Kantardzic, 2003). Moreover, according to the methods that produce clusters, they can be further divided into agglomerative algorithms and divisive algorithms (Willett, 1987).

Agglomerative methods begin by considering each object as a single cluster, and gradually merge the objects into bigger clusters. The clustering procedure produced at each step results from the previous one by combining the two most similar clusters into a single cluster (Downs and Barnard, 2002; Halkidi et al., 2001). The most common agglomerative hierarchical methods are the *Sequential Agglomerative Hierarchical Non-overlapping* (SAHN) methods. A non-overlapping technique means that each object belongs to one cluster only. Some commonly used agglomerative methods can be found in the literature and they are varied in the measures of distance (or similarity) between clusters (Ketchen and Shook, 1996; Leach and Gillet, 2007). First, linkage methods group objects by different types of distance calculation such as: single linkage (nearest neighbour), calculating the minimum distance between objects; complete linkage (furthest neighbour), computing the maximum distance; and group average, measuring the average distance between all pairs of objects. Second, centroid methods cluster objects based on maximizing the distance between the centers of clusters. Finally, variance methods generate clusters by minimizing the increase of variance which is calculated by the error sum of squares. A well-known example is Ward's method.

On the other hand, divisive methods begin by treating all objects as a single cluster and gradually partition the objects into smaller clusters based on a single descriptor (Downs and Barnard, 2002; Halkidi et al., 2001). Because of the basis of a single descriptor, divisive methods are faster than the agglomerative methods. However, the chemical applications of divisive

methods offer poor performance in comparison with agglomerative methods (Rubin and Willett, 1983). Hence, little literature has been discussed on the use of hierarchical divisive methods to deal with chemical datasets.

There are several examples of well-known hierarchical clustering algorithms in the recent literature as follows: BIRCH (Zhang et al., 1996), CURE (Guha et al., 1998), and ROCK (Guha et al., 1999). However, in the application of chemoinformatics, Ward's clustering method has been widely used for analysis of chemical structure databases; it groups two clusters by the shortest Euclidean distance or variance between pairs of centroids (Ward, 1963). Another study of hierarchical clustering approach is that of El-Hamdouchi and Willett (1987) who employed Ward's hierarchic, single linkage, complete linkage, and group average clustering methods for document retrieval and found group average method has the best performance for document clustering.

### 3.2.2 Non-Hierarchical Clustering

Non-hierarchical clustering techniques, also known as partitioning clustering, split a dataset into a prior specified number of smaller datasets or clusters in some cases such as K-Means clustering. It begins by selecting an object as a cluster centre or "seed point", and then clusters all objects according to a certain threshold value or distance (Ketchen and Shook, 1996; Leach and Gillet, 2007). It is also a non-overlapping technique as hierarchical techniques, which means each object is assigned to one cluster only. Contrary to hierarchical clustering, non-hierarchical techniques split a dataset into groups that have no hierarchical relationship to each other. Therefore, the computational requirements for non-hierarchical clustering are generally less than for hierarchical techniques.

There are three major non-hierarchical methods as follows: relocation clustering, nearest-neighbour clustering, and single-pass clustering (Willett, 1987). Relocation methods begin with selecting (usually randomly) $k$ objects as "seed point", and then the rest are assigned to the closest seed generating a set of $k$ clusters. With the centroids re-calculated for each cluster, objects are

relocated to the closest new cluster centroid, and such process is usually repeated until no objects have been relocated. *K-Means* method is a commonly used relocation technique. Second, in nearest-neighbour methods, all pairwise similarities are measured to find the nearest neighbours of each object and ranked based on the similarities. A well-known example of chemical applications is the Jarvis-Patrick method (Jarvis and Patrick, 1973). Finally, in the single-pass methods, the first object is assigned to the first cluster, and the next object belongs to the first cluster or a new cluster depending whether their similarity is over a specified threshold value. Such methods cluster objects using only one pass over the dataset.

### 3.2.3 Summary

There is vast number of clustering algorithms available in the literature, and it may be difficult and confusing for users trying to choose a suitable algorithm for the problem. Thus, users undertaking a cluster analysis should take two important issues into account when they use clustering algorithms (Kantardzic, 2003).

First, it is essential for users who utilize a clustering algorithm to have a complete comprehension of the specific technique being used, as well as to know the details of the data grouping process. All of these will be the best criteria to choose an appropriate method. Moreover, the more information the user has relating to the data, the more likely the user would be able to succeed in a cluster analysis. Second, there is no single best clustering algorithm and no single method will be suitable for exploring the variety of structures present in all types of multidimensional datasets. Therefore, it is necessary for a user to try various algorithms on a given dataset to identify the most appropriate method for that application.

## 3.3 The Comparison of Clustering Methods

In recent years, there has been a dramatic increase in research in many fields concerned with clustering. Despite its frequent use, little is known about the applicability of available clustering methods, whether the method selected is suitable for user's problem at hand, or how clustering methods should be employed.

There have been various studies in the literature related to the comparison of clustering methods in varied disciplines such as marketing (Punj & Stewart, 1983), chemoinformatics (Raymond et al., 2003; Willett, 1987) and data mining (Berkhin, 2002). There also have been several extensive discussions of clustering validation; examples of comprehensive reviews are given by Willett (1985), the studies of Milligan (1996) and Halkidi et al. (2001) broadly cover clustering evaluation techniques, whilst discussions of some specific validation techniques can be found in the studies by Berkhin (2002), Halkidi et al. (2001) and Jain et al. (1999).

The evaluation of clustering results is always one of the most significant issues in cluster analysis, and is often done to find the clustering that best describes the underlying data (Halkidi et al., 2001). The researchers cannot assure that they have a set of useful and meaningful clusters even after careful analysis of a dataset and the selection of a final cluster method. Furthermore, to evaluate the quality of clustering results is always a significant issue of the procedure. On the other hand, the evaluation of clustering methods is also a critical issue in cluster analysis. Rand (1971) proposed several objective criteria which depend on a measure of similarity between two different clusterings of the same datasets, and the measure essentially considers how each pair of objects is assigned in each single cluster. In addition to evaluating clustering methods by their results, Murtagh (2000) evaluated clustering methods by their time and storage costs.

An empirical study by Milligan (1980) compared the performance of k-means methods and hierarchical methods and found that when using random seeds as the start points, K-means methods generated noticeably worse performance than hierarchical methods even under the condition of no error or noise. However if the optimal starting procedures, obtaining the starting seeds from hierarchical methods e.g. group average method, were carried out instead of random seeds selection, k-means methods offered similar or superior performance to the hierarchical methods.

Brown and Martin (1996) investigated clustering methods to compare their performance for compound selection by using varied fingerprints. Active or inactive data was available for the compounds in the datasets used, and then the evaluation was based on how well clustering separated active from inactive compounds. Although the Jarvis-Patrick technique was the fastest among all the methods, it offered the worst performance than any other. Overall, the Ward's method produced most consistent and the best performance.

## 3.4 Chemical Applications of Clustering

In discussions of chemical applications, clustering is one of the most important of the techniques that have been widely used in the literature. In recent years, clustering analysis is getting considerable attention not only in many disciplines such as business and computer sciences but also in Chemoinformatics; some common chemical applications of which are high-throughput screening, combinational chemistry, compound acquisition, and QSAR (Downs and Barnard, 2002).

The clustering of chemical structures may be the earliest and most important chemical application. The following serve as some examples: Adamson and Bush (1973) developed a method to classify automatically the chemical structures, comparing fragment bit-strings for similarity calculation by three different coefficients and the clustering results were reasonable from a qualitative viewpoint. Willett et al. (1986) summarized an empirical comparison of nonhierarchical clustering methods based on simulated property

prediction experiments, and clustered the outputs resulting from chemical substructure searches. The finding is that the Jarvis-Patrick method is effective in operation even with large datasets of many hundreds or thousands of chemical compounds. The study of Butina (1999) also found that using Jarvis-Patrick method with Daylight's fingerprints and the Tanimoto similarity index has a good performance in dealing with large datasets. Whilst, Reynolds et al. (1998) developed a simple clustering method to group structures based on 2D topology descriptors.

With the increasing needs of optimal clustering methods in chemical applications, a variety of novel methods are found in the literature; for example CAST (Ben-Dor et al., 1999), Raymond-Willett (Raymond and Willett, 2003), and Yin-Chen (Yin and Chen, 1994). Raymond et al. (2003) compared five clustering methods used for chemical structures by graph- and fingerprint-based similarity measures. Although the results based on graph similarities are different from fingerprint similarities, they cannot suggest that a certain method is consistently superior to the other; however, some novel clustering methods such as CAST and Yin-Chen generate superior performance to traditional clustering methods such as Ward's and Jarvis-Patrick over these tests, and may be useful alternatives for the clustering of chemical structure databases. Furthermore, they concluded that both graph- and fingerprint-based similarity measures can be used effectively for chemical clustering.

Hierarchical agglomerative techniques, for example Ward's method, are widely used for commercial purposes. The importance of current research is turning toward the quality of the clustering results. The achievements in chemical application of clustering are more hopeful than in other disciplines because the clustering methods in chemical application are able to deal with mixed or nonnumerical data and pay more attention on cluster size, shapes, and distribution (Downs and Barnard, 2002). For example, cluster-based and even dissimilarity-based algorithms, so far, are widely used to select compounds not only on the basis of chemical similarity or dissimilarity but also on the basis of other chemical characteristics such as cost, pharmacokinetic properties, and ease of synthesis (Willett, 2005).

Böcker et al. (2006) proposed a novel hierarchical clustering approach which is called NIPALSTREE to analyze large datasets in high-dimensional space. The clustering results of NIPALSTREE were compared with another hierarchical k-means clustering method; it was validated using ACE inhibitors in the COBRA dataset and shown to generate meaningful results.

As for the clustering applications on high-throughput screening in drug discovery, cluster analysis is a suitable tool for grouping similar compounds into classes. However, many available clustering methods focus on accurate classification of objects, and thus, they lead to a time-consuming process. It is not suitable to apply high-throughput screening on large scale compound libraries. Li (2006a) proposed a fast clustering method to group a very large scale dataset with millions of compounds in hours, and to analyze the redundant compounds of a very large high-throughput screening library. In addition, the use of clustering methods in high-throughput screening is discussed by Dunbar (1997).

## 3.5 Summary

Having introduced the main features of similarity and cluster analysis, the later three chapters (Chapter 5 to 7) describe the experiment work carried out in this thesis. One of the problems noted above (in Section 3.1) is the standardization of variables. This has been little studied in chemoinformatics, and hence Chapter 5 presents a detailed evaluation of standardization methods using both the similarity searching and cluster analysis to compare the various methods that have been suggested in the literature.

In addition, the applications of chemical clustering, especially on 2D structures, have room for improvement and extension, because there are limitations and drawbacks in the currently used clustering methods. It is worth employing some methods that are reported effective in other fields to the application of chemical clustering (as presented in Chapter 6).

# Chapter 4 : Experimental and Evaluation Methods

Due to the studies in the next three chapters containing some experiment details in common, all experimental contexts are hence presented in this chapter including the datasets and the chemical representations, clustering methods and the evaluation measures, which have been applied to the studies of the next three chapters.

## 4.1 Datasets

Two chemical databases are used in this thesis. The first is the MDL Drug Data Report (MDDR) containing 102,535 biologically relevant compounds with over 452 activity classes, produced formerly by MDL Information Systems and now by Symyx Technologies (Symyx, 2007). Each compound in the MDDR is classified into one or several activity classes corresponding to a certain therapeutic action. It is one of the largest databases of chemical structures with associated biological activities and the essential information about biological activity of the MDDR is mainly acquired from the patent literature, which is a popular example in the field of chemoinformatics. We randomly selected 10% from the entire MDDR database with SDF (Structure Data Format) format (for the experiments in Chapter 5) and SMILES format (for the experiments in Chapters 6 and 7) by SciTegic Pipeline Pilot software with default random seed 333 obtaining a total of 10,191 and 10,201 molecules respectively.

The other chemical database is the IDAlert containing 11,607 compounds across 834 activity classes classified by the pharmacological property, produced formerly by Current Drugs Ltd. and now by Thomson Reuters (Thomson Reuters, 2007). Similar to the MDDR, each compound in the IDAlert database is assigned to a certain activity class. This work used the

entire IDAlert database as the dataset for the studies in different chapters of this thesis.

Moreover, we chose eleven activity classes from the two databases, which have been reported previously by Hert et al. (2004), in a study of virtual screening methods on the MDDR database. The chosen eleven activity classes were employed as the indicators to evaluate the clustering results as shown in Tables 4-1 (for the MDDR) and 4-2 (for the IDAlert). Each row in the table contains an activity class, the number of molecules belonging to the class, and the indication (pairwise similarity and standard deviation) of the class's diversity. The diversity of an activity class is computed based on the pairwise Tanimoto similarities using the Pipeline Pilot ECFP_4 fingerprints (the manner of calculating Tanimoto similarity is discussed in Section 2.4.2 as Figure 2-3). However, some classes have different but similar names in these two databases, for example 5HT reuptake inhibitors and D2 antagonists in the MDDR are called 5HT uptake inhibitors and Dopamine D2 antagonists respectively in the IDAlert.

| Activity Class | Active Molecules | Average Pairwise Similarity | Pairwise Standard Deviation |
|---|---|---|---|
| 5HT3 antagonists | 89 | 0.34 | 0.11 |
| 5HT1A agonists | 94 | 0.33 | 0.10 |
| 5HT reuptake inhibitors | 38 | 0.35 | 0.14 |
| D2 antagonists | 40 | 0.35 | 0.09 |
| Renin inhibitors | 112 | 0.57 | 0.10 |
| Angiotensin II AT1 antagonists | 95 | 0.40 | 0.10 |
| Thrombin inhibitors | 108 | 0.42 | 0.13 |
| Substance P antagonists | 125 | 0.39 | 0.11 |
| HIV-1 protease inhibitors | 67 | 0.45 | 0.12 |
| Cyclooxygenase inhibitors | 54 | 0.27 | 0.09 |
| Protein Kinase C inhibitors | 48 | 0.31 | 0.13 |

Table 4-1 Eleven activity classes and their number of actives in the 10k MDDR dataset

| Activity Class | Active Molecules | Average Pairwise Similarity | Pairwise Standard Deviation |
|---|---|---|---|
| 5HT3 antagonists | 99 | 0.36 | 0.12 |
| 5HT1A agonists | 61 | 0.33 | 0.10 |
| 5HT uptake inhibitors [a] | 41 | 0.32 | 0.10 |
| Dopamine D2 antagonists [a] | 20 | 0.36 | 0.08 |
| Renin inhibitors | 123 | 0.48 | 0.14 |
| Angiotensin II AT1 antagonists | 12 | 0.43 | 0.08 |
| Thrombin inhibitors | 76 | 0.49 | 0.15 |
| Substance P antagonists | 66 | 0.41 | 0.13 |
| HIV-1 protease inhibitors | 32 | 0.42 | 0.13 |
| Cyclooxygenase inhibitors | 87 | 0.26 | 0.09 |
| Protein Kinase C inhibitors | 51 | 0.32 | 0.16 |

[a] MDDR activity classes 5HT reuptake inhibitors and D2 antagonists are called 5HT uptake inhibitors and Dopamine D2 antagonists respectively in the IDAlert dataset.

Table 4-2 *Eleven activity classes and their number of actives in the IDAlert dataset*

## 4.2 Chemical Representations

The two datasets were characterized by four different chemical representations. Molconn and Pipeline Pilot have similar data type, i.e. real number (numerical), of descriptors for structure description, but differ in the number of descriptors. Tripos molecular holograms and Pipeline Pilot ECFP_4 are fingerprint-based representations, but differ in the data type, integer and binary respectively, of their descriptors.

### 4.2.1 Molconn

Molconn structure descriptors are a set of varied types of topological indices of molecular structure. These indices (i.e. descriptors) show the molecular structure information which is useful. We used Tripos Sybyl software (Tripos, 2007) to calculate 523 Molconn descriptors from molecular structure (labeled Molconn-Z in this thesis) containing molecular connectivity (Chi) indices, electrotopological state (E-state) indices, shape (Kappa) indices, topological state and equivalence indices. These indices are suitable for QSAR (Quantitative Structure-Activity Relationships) and QSPR (Quantitative

Structure-Property Relationships) studies (Tripos, 2007), and are also ideal for statistical methods e.g. cluster analysis and regression. The Molconn-Z representation was employed in the experiments in Sections 5.3 to 5.7.

In addition, due to the license of Molconn-Z package in the Tripos Sybyl software being changed, we employed a new alternative of Molconn tool, called winMolconn software (HAC, 2010), which is available at http://www.molconn.com and denoted by win_Molconn in this thesis, for the extensive study of standardization methods in Chapter 5. It generates 668 descriptors from the connection table of chemical structures including three main categories of elementary structure information indices, molecular connectivity indices and electrotopological state (E-State) indices. The win_Molconn representation was used in the experiments in Sections 5.9 to 5.12.

The correlations between many Molconn descriptors, i.e. Molconn-Z and Win_Molconn, are highly correlated with each other (Shen et. al., 2003). Hence, the certain information of a set of highly correlated descriptors may usually be over-represented. In order to get rid of such problem, Principal Component Analysis (PCA) is commonly applied to transform a number of correlated variables, i.e. descriptors, into a small number of un-correlated variables which are usually called principal components. In other words, the number of descriptors, i.e. the dimensionality of a dataset, is hence reduced to generate a new set of small number of descriptors.

The process of Principal Component Analysis, in essence, usually involves the procedure of standardization (Leach & Gillet, 2007; Shen et. al., 2003), i.e. converting the source data to Z-score. However, one aim in the works of Chapter 5 is to compare the effectiveness of different standardization procedures on the chemical data with Molconn representations. To avoid the Molconn descriptors being re-standardized, the correlations between Molconn descriptors are ignored here, that is, all Molconn descriptors are kept in the datasets.

## 4.2.2 Pipeline Pilot

Similar to the data type of the Molconn descriptors, we used Scitegic Pipeline Pilot software (Accelrys, 2007) to generate twelve commonly used structural descriptors to form this chemical representation (labeled Pipeline Pilot in this thesis), such as AlogP, logD and PKa, molecular weight, Surface area and volume, and solubility (summarized in Table 4-3). The major difference between Pipeline Pilot and Molconn-Z representations is the number of descriptors they contained. The Pipeline Pilot representation was used in the study of Chapter 5.

| Descriptors | Descriptions |
|---|---|
| Minimized Energy | Gives the molecular energy after a fast minimization procedure |
| ALogP | The Ghose/Crippen group-contribution estimate for LogP, where P is the relative solubility of a compound octanol vs. water |
| ALogP_MR | The Ghose/Crippen estimate of molar refractivity, which contains information about molecular volume and polarizability of a compound |
| LogD | The ratio of the equilibrium concentrations of all species of a molecule in octanol to same species in the water phase at a given temperature. |
| Molecular weight | Molecular weight |
| Solubility | Molecular Solubility |
| Molecular_Volume | The 3D volume |
| Molecular_SurfaceArea | The total surface area and polar surface area for each molecule are calculated using a 2D approximation |
| Molecular_PolarSurfaceArea | |
| Molecular_SASA | The total solvent accessible surface area |
| Molecular_PolarSASA | The polar solvent accessible surface area |
| Molecular_SAVol | The solvent accessible volume |

Table 4-3 The summary of descriptors of Pipeline Pilot representation

### 4.2.3 Holograms

Molecular hologram representation is a technique of fingerprinting which consists of all varied molecular fragments within a molecule, and records the count of the frequency in which each unique fragment occurs rather than traditional 2D fingerprints that record only the status of absence or presence of a certain fragment (Tripos, 2007). We used Tripos Sybyl software to calculate molecular holograms (labeled Holograms in this thesis) containing 997 descriptors. Each descriptor represents a predefined molecular fragment, which is generated for all possible substructures between 4 and 7 atoms in size for all molecules, to record the number of times a unique fragment occurs in a given molecule. The Holograms representation was used in the study of Chapter 5.

### 4.2.4 ECFP_4 Fingerprints

Molecular fingerprints are one of the common chemical representations, and are widely used for similarity searching, virtual screening and clustering. Extended-connectivity fingerprints (ECFPs) are a commonly-used example of molecular fingerprints. They were designed to capture molecular features which correspond to molecular activity. We used SciTegic Pipeline Pilot software (Accelrys, 2007), which is available at http://www.accelrys.com to generate ECFP_4 circular fingerprints (labeled ECFP_4 in this thesis) with a fixed length of 1024 bits (descriptors).

The suffix number, i.e. 4, after the term ECFP indicates the diameter (in bonds) of the circular substructure. The data type, in essence, for ECFP_4 is binary. That is, each descriptor encodes simply the absence (zero) or presence (one) of a 2D structural fragment within a molecule. The main difference between Holograms and ECFP_4 is that the former records the counts for a certain fragment, whereas the latter records only the absence or presence of substructures. ECFP_4 is a type of Extended-Connectivity fingerprint (ECFPs), and such fingerprints encode circular substructures based on a hash function, a variation of the Morgan algorithm, which was initially proposed to solve the molecular isomorphism problem in order to generate a unique structural

description (Morgan, 1965; Leach and Gillet, 2007). ECFP_4 fingerprint was used in the studies in Chapters 6 and 7.

Table 4-4 summarizes the overall information for the above four chemical representations and the context they have been applied to.

| Chemical Representations | Data Types | Software Tools | Context |
|---|---|---|---|
| Molconn-Z<br>Win_Molconn | Real Number | Tripos SYBYL<br>winMolconn software | Sections 5.3 to 5.7<br>Sections 5.9 to 5.12 |
| Pipeline Pilot | Real Number | Scitegic Pipeline Pilot | Chapter 5 |
| Holograms | Integer | Tripos HQSAR | Chapter 5 |
| ECFP_4 | Binary | Scitegic Pipeline Pilot | Chapters 6 & 7 |

Table 4-4 Summary of four chemical representations

## 4.3 Clustering Methods

The clustering methods used in Chapters 5 and 6 are integrated and discussed in this section. Some methods, Yin-Chen and CAST, are coded, and the rest of the methods are carried out using the implementations in specific software packages. Due to the license of particular software package being changed, the Ward's method is carried out using different software packages in distinct experiments of this thesis but with the identical standard Ward's algorithm.

### 4.3.1 Yin-Chen

This clustering method involves a two-phase algorithm with fixed-radius selection (Yin and Chen; 1994). This approach examines the status of connectivity of pairwise objects: if the distance between them is less than a certain distance, i.e. two times the mean minimum distance (MMD), then they will be considered to be connected; otherwise, they will be considered to be noise and will be removed from the dataset. A graph theoretic procedure, in our study we chose Breadth First Search (BFS), is applied afterwards to find out

the connected components based on the status of adjacency. Each connected component is considered a cluster. In addition, the distance calculation in our study is based on the Tanimoto distance (Willett et al., 1998; Holliday et al., 2002; Li, 2006), and the number of clusters is determined by an adjustable parameter i.e. a cut-off threshold.

### 4.3.2 CAST

Cluster Affinity Search Technique (CAST) was proposed by Ben-Dor et al. (1999) for applications on clustering gene expression data. One feature of this method is using a cut-off parameter as a threshold to adjust the number of the clusters, therefore no predefined number of clusters is applicable to such method, and in some applications the number of clusters is usually unknown or hard to specify. The rationale of CAST, in short, is taking turns between moving the element with maximum similarity in the working cluster, and removing the element with minimum similarity from it until the working cluster is stable, i.e. a cluster has been generated; then a new cluster is started thereafter. In addition, the calculation of similarity is based on the Tanimoto coefficient (Leach and Gillet, 2007; Haranczyk and Holliday, 2008), and, similar to the Yin-Chen method, an adjustable parameter is needed to determine the number of clusters.

### 4.3.3 UPGMA

CLUTO is the abbreviation of CLUstering TOolkit and is a suitable software package for clustering with high dimensional datasets. It has been widely used in the application of document clustering (Steinbach et al., 2000; Zhao and Karypis, 2005), while in our study, we applied it to chemical clustering. Agglomerative clustering methods have been extensively used in a wide range of fields. Saad et al., (2006) compared the performance between agglomerative and partitional clusterings and found agglomerative method effective. In addition, the application of document clustering using CLUTO package also reported that the *agglo* method with UPGMA (Unweighted Pair Group Method using Arithmetic mean) criterion function and the *repeated bisection* method

had better performance (Steinbach et al., 2000). In our study, we employed a hierarchical agglomerative method, *agglo*, and two partitional-based methods, *direct* and *repeated bisection* (see next sections) for the application of chemical clustering.

The *agglo* method is the traditional hierarchical agglomerative method. Initially, it considers each object in a dataset as individual clusters and then keeps merging two clusters which are most similar until the desired number of clusters is found or certain criterion is reached. However, the critical process in such sort of methods is the scheme used to choose which two clusters to be merged next (Karypis, 2003). The default criterion function of *agglo* method in CLUTO is UPGMA, which is also known as average linkage. Two clusters with minimum distance are merged into one cluster, for which the distance is based on the average of pairwise distances in each cluster.

### 4.3.4 Direct

In terms of *direct* method, the desired *k* clusters are generated synchronously; it is similar to traditional K-means type of algorithms. The *direct* method is simply a two-step algorithm. The first step involves selecting randomly *k* objects from the dataset as the centroids and then assigning each of the rest of objects to its closest centroid. Hence the initial *k* clusters are obtained. The second step contains a number of iterations of refinements. The refinement is based on a best-one-element-move strategy (Zhao & Karypis, 2005). Each object is visited in a random order to see if any improvements in the value of a desired criterion function are found by moving one object to one of the rest of *k-1* clusters. If the improvements are found, then moves this object to the cluster which leads to the best improvement; if not, this object stays in its original cluster. The iteration of refinement stops on condition of no objects moved between clusters.

Both *i2* and *e1* criterion functions are used for each of the partitional-based methods i.e. *direct* and *repeated bisection*. The *i2* criterion is based on the within-cluster similarity; in this measure, each cluster is represented by its

centroid, and a cluster is generated by maximizing the similarity or minimizing the distance between a cluster centroid and each member in that cluster. The *e1* criterion, however, generates clusters by minimizing the similarity or maximizing the distance between the centroid of each cluster and the centroid of all clusters. For more detailed explanation of these criterion functions, the reader is referred to the study by Zhao and Karypis (2005). The equations for *i2* and *e1* are defined as follows (Karypis, 2003)

$$i2 = \text{maximize} \sum_{i=1}^{k} \sqrt{\sum_{a,b \in C_i} \text{similarity}(a,b)}$$

$$e1 = \text{minimize} \sum_{i=1}^{k} n_i \frac{\sum_{a \in C_i,\, b \in C} \text{similarity}(a,b)}{\sqrt{\sum_{a,b \in C_i} \text{similarity}(a,b)}}$$

*a* and *b* indicate two objects; *C* is the collection of all objects; $C_i$ represents the collection of objects in a certain cluster; *similarity(a,b)* indicates the similarity between object *a* and *b*.

### 4.3.5 Repeat Bisection

The *repeated bisection* method, which is a variation of K-Means but with hierarchical divisive method (Downs and Barnard, 2002; Willett, 2009) also named *Hierarchical K-means* (Böcker et al., 2005), it divides the dataset repeatedly into clusters. In a word, the dataset is initially split into two clusters using the original K-Means algorithm; and then one cluster is chosen and split. This process repeats until it reaches the desired number of clusters (Barnard and Downs, 1992). However, the critical process in repeated bisection is the measure employed to choose which cluster to be divided next, normally the largest cluster is selected for bisection (Steinbach et al., 2000; Saad et al., 2006). The criterion functions, *i2* and *e1*, used for this method were discussed in Section 4.3.4.

### 4.3.6 K-Means

K-Means algorithm, was first proposed by Stuart Lloyd in 1957 but was not published until 1982, whereas it was first used by MacQueen in 1967 (Jain, 2010), and is one of the best known partitional clustering methods. Basically, it is an iterative clustering algorithm in which objects are relocated among clusters until some convergence criterion is met. In this thesis, the traditional K-Means method was carried out using the implementation in the BCI (Barnard Chemical Information) software package, which is now Digital Chemistry Clustering Tools (Digital Chemistry, 2007), the main steps of this traditional K-Means method are

1. Choose *k* random objects as the centroids
2. Assign each object to its nearest centroid, i.e. cluster center
3. Compute the new cluster center as the centroid for each cluster
4. Repeat steps 2 and 3 until no object relocation is needed

The time complexity of K-Means is $O(tkn)$, where $t$ is the number of iterations, $k$ is the number of clusters, and the $n$ is the number of objects, i.e. size of dataset. Obviously, $k$ and $n$ can substantially influence its efficiency. It is time-consuming when dealing with large datasets, however it often generates good results. In addition, it is sensitive to the noise and outliers, since such data significantly influence the computing of cluster centers on relocating objects.

According to the algorithm of traditional K-Means listed above, it generates different results with each run, because the clustering results depend on the random selection of initial centroids. Moreover, it can obtain a local optimum, i.e. minimizing intra-cluster variance, but not assure the global optimum. Hence, extensive variations of the K-Means method are reported in the literature to obtain the overall optimum. Basically, they differ in the details of careful selecting the initial centroids, e.g. *Direct* method of CLUTO, or adjusting the partition, e.g. if the distance between two cluster centroids is less than a predefined threshold, then two clusters are merged (Dunham, 2003). Some methods also operate in a deterministic manner by removing the random

selection of centroids and the order-dependent processing of objects.

### 4.3.7 Ward's

Ward's method is a well-known hierarchical agglomerative clustering method, and is normally the method of choice in chemoinformatics especially in the application of chemical clustering of 2D structures (Barnard and Downs, 1997). Unlike many other clustering methods, Ward's method (Ward, 1963) considers clustering as an analysis of variance to evaluate the distance between clusters, instead of using distance or similarity metrics. The fusion criterion minimizes the increase of the error sum of squares computed based on Euclidean distance between two clusters in order to optimize the quality of the new cluster formed at each step (Everitt et al., 2001). Many hierarchical agglomerative techniques, e.g. complete, single or average linkage, obtain only the global optimum, i.e. minimum inter-cluster variance. However, the Ward's method obtains both local (intra-cluster) and global (inter-cluster) optimum by minimizing the increase of the intra-cluster error sum of squares.

### 4.3.8 Extended Ward's

This hierarchical clustering method was proposed by Szekely and Rizzo (2005); its rationale is based on joint between-within cluster distances. Similar to Ward's method, extended Ward's also minimizes the Euclidean distance between clusters. However, the distance for extended Ward's, named *e-distance*, is a measure of both the heterogeneity between clusters and homogeneity within clusters. In the proposed *e-distance* formula, with a power function $\alpha$ of the Euclidean distance will generate different clustering methods. For example, the objective function with $\alpha=1$ and $\alpha=2$ are equivalent to the extended Ward's and conventional Ward's method respectively. The formula was defined as

$$e^{\alpha}(A,B)=\frac{n_1 n_2}{n_1+n_2}\left(\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left\|a_i-b_j\right\|^{\alpha}-\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\left\|a_i-a_j\right\|^{\alpha}-\frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}\left\|b_i-b_j\right\|^{\alpha}\right)$$

where *A* and *B* represent two non-empty vector space (clusters)

$A = \{a_1, a_2, ..., a_{n1}\}$ and $B = \{b_1, b_2, ..., b_{n2}\}$

$n_1$ indicates the number of elements in cluster A, while $n_2$ represents the number of elements in cluster B. The α powers of Euclidean distance fall in the interval (0,2].

A summary of software tools for the above clustering methods and the context they are applied to is given in Table 4-5.

| Clustering Methods | Software Tools | Use in thesis |
|---|---|---|
| Ward's | BCI software | Sections 5.3 to 5.7 |
| | R software | Sections 5.9 to 5.12 and Chapter 6 |
| Extended Ward's | R software | Sections 5.9 to 5.12 Chapter 6 |
| K-Means | BCI software | Sections 5.3 to 5.7 |
| Yin-Chen | Coded by Perl Script | Chapter 6 |
| CAST | Coded by Perl Script | Chapter 6 |
| UPGMA | CLUTO | Sections 5.9 to 5.12 Chapters 6 and 7 |
| Direct | CLUTO | Sections 5.9 to 5.12 Chapters 6 and 7 |
| Repeated Bisection | CLUTO | Sections 5.9 to 5.12 and Chapter 6 |

Table 4-5 Summary of the software tools and use in thesis of all clustering methods

## 4.4 Evaluation of Clustering Results

Evaluation of clustering results is a critical process in cluster analysis, it not only assesses the clustering techniques but also provides the intensity of confidence for the clustering. Most clustering applications need an evaluation measure to assess the results from a certain method, such as the assignment of objects in clusters, the number of clusters, capturing the intra-cluster similarity and inter-cluster dissimilarity. There are extensive evaluation measures with different types in the literature; if a clustering method offers better performance than others over many evaluation measures, then that clustering method is the best for a certain type of application. Hence we chose five evaluation measures for our experiments. Shannon entropy and probability of correct prediction are two evaluation criteria used in the study of Böcker et al. (2006). Entropy based on cluster size is a measure which is similar to the conventional Shannon entropy to observe the distribution of partition sizes. F-measure is a measure widely used in document clustering for many years (Fung et al., 2003; Rosenberg and Hirschberg, 2007). Quality Clustering Index (QCI) is a new evaluation measure recently defined by Varin et al. (2008).

### 4.4.1 Shannon Entropy

Shannon Entropy (SE) is a technique to evaluate the distribution of active compounds from inactives for a given class across all clusters (Matter, 1997). Entropy-based approach assumes that the best possible classification is one in which all of the actives for a given particular activity class are located in the same cluster. Conversely, the worst possible classification is one in which they are distributed equally across the available clusters. The distribution of the actives was quantified using the Shannon Entropy (SE), which is defined (Godden and Bajorath, 2001; Batista et al., 2006) as

$$SE = -\sum_i p_i \log_2(p_i) \quad \text{and} \quad P_i = \frac{a}{A}$$

where $p_i$ is the fraction of the total number of active molecules that occur in the *i-th* cluster and where the summation is over all of the clusters

$a$ is the number of active molecules in a certain cluster, and

$A$ is the total number of molecules in a given activity class.

For example, if 4 of the 100 members of an activity class occur in some cluster A, then $p_i = 4/100 = 0.04$, yielding a contribution to $SE$ of 0.19. The performance measure is then the calculated entropy, with the results being averaged over all of the eleven activity classes. For this measure, small entropy values indicate good clustering results.

### 4.4.2 Probability of Correct Prediction

This evaluation criterion involves finding the fraction of clusters containing actives that are predicted to be active or inactive. The Shannon Entropy observes merely the distribution of actives and takes no account of actives' co-occurrence with inactives. Whereas, the evaluation using the probability of correct prediction takes account of both the actives and the inactives for a certain activity class. Let an *active cluster* be a cluster that contains at least one molecule from the chosen activity class. Define *P(active)* and *P(inactive)* for a particular cluster as

$$P(\text{active}) = \frac{a}{A} \quad \text{and} \quad P(\text{inactive}) = \frac{n-a}{N-A}$$

where $N$ is the total number of compounds in the dataset,

$n$ is the total number of molecules in the current active cluster,

$a$ is the number of active molecules in that cluster, and

$A$ is the total number of molecules exhibiting the chosen activity.

The two values *P(active)* and *P(inactive)* hence describe the proportion of the actives and the proportion of the inactives that are present in the chosen cluster. We would hope that *P(active)* would be greater than *P(inactive)* in the case of an active cluster, i.e., that there is a greater concentration of active molecules present (whereas the converse would imply the presence of some small number of "stray" actives in a cluster composed predominantly of inactives). We then use the number of times when this is in fact the case as a measure of the effectiveness of clustering: the more frequently this happens, the greater the degree of concentration of the actives in the active clusters. For example, assume that $a = 2$ and $n = 10$ for some cluster and that $N = 820$ and $A = 20$ for the dataset. Then the probabilities of activity and inactivity are

$$\mathrm{P(active)} = \frac{2}{20} = 0.1 \quad \text{and} \quad \mathrm{P(inactive)} = \frac{10-2}{820-20} = 0.01$$

with *P(active)* > *P(inactive)*, as would be predicted to be an active cluster. The performance measure is then the fraction of active clusters that are indeed predicted to be active for the chosen activity class, with the results being averaged over all of the eleven activity classes.

As the equations of *P(active)* and *P(inactive)* are listed above, the probability of a given cluster which is predicted to be active or inactive depends on two factors. The first factor is the size of dataset ($N$) and the other is the size of clusters ($n$). For example, suppose the size of the MDDR dataset is N=10,000 and the approximate size of clusters is n=20 to 10 (with the number of clusters 500 to 1000). Hence with the same conditions as in above example of $a$=2 and $n$=10 for some cluster and $A$=20 for the dataset; even if the number of active molecules ($a$) is small, the probability of *P(active)* tends to be much greater than *P(inactive)*.

$$\mathrm{P(active)} = \frac{2}{20} = 0.1 \quad \text{and} \quad \mathrm{P(inactive)} = \frac{10-2}{10000-20} \approx 0.0008$$

Obviously, *P(active)* is much higher than *P(inactive)*, that is, the chosen cluster is easily to be active. Hence, when dealing with very large dataset and small size of cluster, the clustering evaluation based on such measure may not be applicable.

Since this evaluation approach is strongly affected by the size of dataset and of cluster, we employed it only in the study described in Chapter 5 (Effect of Standardization on Three Different Representations of Structural Similarity), that is because the number of clusters in the experiment was set to be 25, 50 and 100. For other experiments, such as the extensive study in Chapter 5, and other studies in Chapters 6 and 7, the partitions contained 500, 600, 700, 800, 900 and 1000 clusters, in which case the partition size is much smaller. Consider the size of datasets (approximately 10,000) and the small partition size (20 to 10 in averages), and find that large size of dataset and small size of partitions will easily lead the clusters to be identified active. The evaluation using the probability of correct prediction is not applicable to above experiments but only to the experiments in Sections 5.3 to 5.7.

### 4.4.3 Entropy Based on Cluster Size

The rationale of entropy based on cluster size is similar to conventional entropy as discussed in Section 4.4.1. It evaluates the size distribution over all clusters. The only difference is the calculation of probability $p_i$ in the equation of Shannon entropy. The $p_i$ is defined as

$$p_i = \frac{n}{N}$$

where $n$ is the total number of molecules in a certain cluster, and

$N$ indicates the total number of molecules in the dataset.

This criterion hence considers only the sizes of the clusters, not the activity of the molecules in the clusters, and is hence biased towards a classification consisting of equal-sized clusters.

### 4.4.4 F-measure

F-measure (Rijsbergen, 1979) is the evaluation of external clustering quality which takes precision and recall into account, this evaluation measure is widely used in document clustering (Steinbach et al., 2000; Fung et al., 2003). For a certain cluster, the precision and recall can be computed based on a given activity class. Precision calculates the ratio of molecules in a cluster which belong to the given activity class to examine how this cluster is with respect to that activity class; while recall computes the ratio of molecules of the given activity class in a certain cluster to measure how complete this cluster is with respect to that activity class. Both can be defined as

$$\text{Precision} = \frac{a}{n}$$

$$\text{Recall} = \frac{a}{A}$$

where $a$ is the number of active molecules of a given class in a cluster,

$n$ is the total number of molecules in a cluster, and

$A$ is the total number of molecules exhibiting the chosen activity class.

The F-measure of a certain cluster and a given activity class can be defined as (Fung et al., 2003; Rosenberg and Hirschberg, 2007)

$$F = \frac{(2 * \text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$

In terms of the F value for entire clustering, it captures the maximum value for a chosen activity class over all clusters, i.e. finding the "best" cluster for a certain activity class. In conventional document clustering, the overall F value is computed using the weighted sum of such maximum values for all activity classes; and the sum is normally weighted by the ratio of size of a given class to size of the dataset. However, in our experiment, unlike its calculation in document clustering applications, the overall F score for entire clustering performance is the average of these maximum values for all activity classes

without a weighting scheme; because we chose only 11 activity classes which we are interested in from the datasets. The F score used in our study can be defined as

$$F_{overall} = \frac{1}{k}\sum_{ac=1}^{k} \max \{ F_{ac} \}$$

where $max\{F_{ac}\}$ is obtained by comparing the F values over all clusters for a certain activity class, and

$k$ is the number of activity classes, and

$ac$ indicated an activity class.

The value for F measure is between 0 and 1. In addition, larger F score value indicates better clustering results. This is an upper-bound criterion since it is based on identifying the best possible single cluster for a given activity class.

### 4.4.5 Quality Clustering Index

A previous study (Brown and Martin, 1996) used the ratio of active molecules in the active or inactive clusters as the clustering evaluation; however, this usually leads to bias when only a small number of active molecules exist in inactive clusters. In order to eliminate such bias on clustering evaluation, Varin et al. (2008) proposed a new index, Quality Clustering Index (QCI), to evaluate the separation between active and inactive molecules during the clustering process. They offered a new definition to verify a cluster as active or inactive by comparing the ratio of active molecules in a cluster and in the dataset. That is, if the ratio of active molecules is greater than the original ratio of total active molecules in the dataset, the cluster is considered to be active. The equation of QCI is defined as

$$QCI = \frac{a}{a+b+c+d} \times 100\%$$

where *a* is the number of active molecules in active clusters,

*b* represents the number of inactive molecules in active clusters,

*c* indicates the number of active molecules in inactive clusters, and

*d* is the number of active singletons.

## 4.5 Evaluation of Correlation

The evaluation of correlation is an important procedure in comparative studies where many experiments are carried out, and is a measure to compare the various methods under study. Here we are interested in the extent to which different conditions (e.g. different activity classes) rank a set of objects (e.g. different clustering methods), and the extent of the correlation between the different conditions. If a set of objects is always, or nearly always, ranked in the same order, then we can have some belief in the validity of that ordering.

Rank transformation procedures are a nonparametric approach that involves replacing the data values with their rankings, and this technique has been applied in clustering analysis, multiple regressions, and multiple comparisons (Conover and Iman, 1981). The Kendall rank coefficient is a technique to measure the degree of correlation between two rankings of *N* objects, as well as to assess the significance of the correlation. When there are more than two sets of rankings, Kendall's coefficient of concordance *W* can be used to measure the correspondence and its strength among them. For example, in our study in Chapter 5, we applied the Kendall *W* test to rank the performance of eight standardization methods, with each test comparing the three raters, Pipeline Pilot, Molconn-Z, and Holograms. In other study in Chapter 6, four evaluation criteria were considered as judges to rank the nine clustering methods in order to obtain a more quantitative view of the effectiveness of the clustering methods. The evaluation of correlation by Kendall's W test is applied to the

studies in Chapters 5 and 6.

The equation of calculation for the value of $W$ is listed as follows (Siegel and Castellan, 1988):

$$W = \frac{12\sum R_i^2 - 3k^2N(N+1)^2}{k^2N(N^2-1) - k\sum T_i}$$

where $k$ is the number of sets of ranking

$N$ is the number of objects being ranked

$R_i$ is the summation of the ranks

$T_i$ is the correction of tied observations

However, when the tied ranks are obtained, each observation is assigned the average of the ranking scores which would have been assigned when no ties occurred. In addition, the sample size $N$ will influence which approach for testing the significance of the Kendall coefficient of concordance should be used. When dealing with larger samples ($N > 7$), $W$ is approximately distributed as $\chi^2$ (chi square) with $N-1$ degrees of freedom and can be tested using

$$\chi^2 = k(N-1)W$$

In brief, the purpose of the Kendall $W$ test in our studies is to determine whether or not the agreement occurred on ranking different procedures. If a statistically significant level of correlation is obtained, then we have more confidence in the validity of that ranking.

## 4.6 Conclusions

In this chapter, we discussed the two datasets, MDDR and IDAlert, which are employed for the next three chapters. Four different chemical representations for these two datasets are also discussed, Molconn, Pipeline Pilot, and Holograms are used only in Chapter 5, while ECFP_4 fingerprint is used in Chapters 6 and 7.

Varied clustering methods discussed here are used in Chapters 5 and 6, and the clustering results in Chapter 6 are employed to generate a similarity matrix for the application of consensus clustering in Chapter 7. Five different evaluation criteria are discussed and verified their suitability for clustering evaluation. The evaluation using the probability of correct prediction is used only in the first experiment in Chapter 5 due to its applicability. While the evaluation criteria, Shannon Entropy, Entropy based on partition size, F-Measure, and QCI, are employed to evaluate clustering performance in the Chapters 6 and 7.

# Chapter 5 : Effect of Standardization on Three Different Representations of Structural Similarity

## 5.1 Introduction

The representation of chemical structures is one of the essential and crucial tasks in chemoinformatics (Engel, 2003). Once the structural descriptors of a certain chemical representation have been calculated, some critical chemoinformatics tasks, such as similarity search, as well as clustering of chemical structures or other applications, can be done. However, the standardization of descriptors or variables is a vital procedure when carrying out similarity searching or chemical clustering with different chemical representations, as well as when the descriptors have particularly varied characteristics. The aim of standardization is to adjust the magnitude or scale of the score of input variables to be equal.

Studies of standardization techniques in this chapter include two parts. The first part (Sections 5.3 to 5.7) involves the evaluation of standardization methods based on the results from similarity searching and clustering. Ward's and K-Means methods, which are commonly used in the application of Chemoinformatics, were employed for clustering in the first experiment. The second part (Sections 5.9 to 5.12) is an extension of the first study: it investigates the effect of the same standardization methods but with another seven clustering methods which were reported effective in the literature.

## 5.2 Standardization Methods

Milligan and Cooper (1988) discussed the use of standardization in cluster analysis, and evaluated the results of different standardization methods with artificial data. They concluded that, as far as standardization approaches are concerned, the standardizing transformations which involve the division by the range of variable have consistently better recovery of the underlying cluster structures; moreover, the most common *Z-Score* method proved to be less effective in some situations. Different standardization techniques reveal different performances in varied applications, however some reviews reported that the standardization procedures offer, at least, a limited advantage for those data needed to be grouped (Edelbrock, 1979; Milligan, 1980; Good et al., 2004). Moreover, Rogers et al. (1991) argued that poor standardizing of variables influences the performance of clustering procedures and algorithms. Instead of using traditional standardization procedures, Stoddard (1979) proposed a linear model for scaling measurements as the standardization procedure and concluded it is necessary to remove the variability of datasets but keep the differences in the size of the properties for generating the superior clustering results.

Bath et al. (1993) used eight different standardization techniques for the measures of intermolecular structural similarity and concluded that there was no significant difference in the effectiveness of various standardization methods when standardized fragment-based data was used for similarity searching on 2D chemical structures. Another study was carried out by Turner et al. (1995); both similarity coefficients and standardization methods were used on the calculation of field-based similarity search. The results showed that there is no significant difference among seven different standardization methods. Evaluations of different standardization methods with chemical properties are rarely seen. Dorans and Kulick (1986) summarized the utility of the standardization method to search unexpected differences in item performance over different subpopulations of educational test data. They concluded that the standardization approach is an effective technique for

comparing the item performance of groups of unequal properties; and the limitation is that relatively large sample sizes are required. Strike et al. (2001) developed quantitative models of software cost estimation with incomplete data and concluded the traditional *Z-Score* standardization method offered consistently the best performance. As for other non-chemical application of standardization, Doherty et al. (2004) investigated whether standardization will influence the clustering results generated by different norms such as the Minkowski or Euclidean norms, and their result showed that a significant improvement was obtained in the class accuracy recovery between standardized and un-standardized synthetic datasets. Account for the performance of different clustering methods, the Neural Gas clustering has the most remarkable improvement in the class recovery rate using standardization procedures in comparison with K-Means and nearest neighbour clustering methods.

Numerous standardization techniques have been discussed in the literature, the detailed review are well described by Milligan and Cooper (1988). Seven standardization techniques were used in this study, based on those used previously in a study of fragment-based similarity applications (Bath et al., 1993). With all these standardization methods, we used the standard statistics notation, in which $X$ denotes the observation value of the variable, $\mu$ denotes the average of observations for the variable, $s$ denotes the standard deviation for the variable. *MAX(X)* and *MIN(X)* denote the maximum and minimum values of the variable respectively. All these standardization forms are as follows:

1. The most common and traditional standardization method is the Z-Score, which has been proposed by Sokal (1961) and Williams & Lambert (1966), and the transformation of variable will have a zero mean and a variance of 1.

$$S_1 = \frac{X - \mu}{s}$$

2. The second form of standardization is similar to Z-Score and has been proposed by Cormack (1971),

$$S_2 = \frac{X}{s}$$

which will result in a variance of 1 and a transformed mean of $\frac{\mu}{s}$.

3. Cain and Harrison (1958) suggest a similar transformation, which involves dividing the value of each variable by the maximum value.

$$S_3 = \frac{X}{MAX(X)}$$

4. Carmichael et al. (1968) proposed a standardization that involves the use of the variable's range as the divisor.

$$S_4 = \frac{X}{MAX(X) - MIN(X)}$$

5. The fifth standardization is similar to $S_4$ using the variable range as well, which has been proposed by Gower (1971).

$$S_5 = \frac{X - MIN(X)}{MAX(X) - MIN(X)}$$

6. Another standardization approach normalizes by the sum of the observations for a variable (Romesburg, 1984), which will result in a mean of 1/n.

$$S_6 = \frac{X}{\sum X}$$

7. Sneath and Sokal (1973) proposed a distinctive standardization method, which differs from the above six methods in using the rank of data instead of its value.

$$S_7 = Rank(X)$$

The advantage of such method is that it can reduce the influence of outliers in the sample. Hence, for all data, the mean will be (n+1)/2, and variance $(n+1)[\frac{2n+1}{6}-\frac{n+1}{4}]$. Moreover, when observations have same value, each observation will have the same rank. To get rid of such tied ranks, each observation will be assigned the average of ranks to adjust the ranking tied scores.

Finally, $S_0$ denotes the original data that is unstandardized. The eight standardization procedures above are summarized in Table 5-1.

| Standardization Methods | Description |
|---|---|
| $S_0$ | Un-standardized dataset |
| $S_1$ | Normalized by the standard deviation |
| $S_2$ | |
| $S_3$ | Normalized by the maximum |
| $S_4$ | Normalized by the variable's range |
| $S_5$ | |
| $S_6$ | Normalized by the sum of the variable |
| $S_7$ | Using rank of data instead of its value |

Table 5-1 Summary of standardization methods

## 5.3 Experimental Details

The MDDR and IDAlert datasets were represented by three different types of chemical representations, Pipeline Pilot, Holograms and Molconn-Z (both datasets and representations are discussed in detail in Chapter 4). However, for the Molconn-Z representation, some molecules failed to generate descriptors for both datasets and were removed. Hence, in order to obtain the equal size of datasets, we also removed those molecules in the datasets with Pipeline Pilot and molecular holograms representations. We then standardized each

representation with the eight different standardization methods defined in Section 5.2. For each dataset, with the combination of standardization methods and chemical representations, we hence had twenty-four test-datasets. With each test-dataset, the standard K-Means and Ward's methods are carried out using the implementations in BCI (Barnard Chemical Information) software, which is now Digital Chemistry Clustering Tools provided by Digital Chemistry (Digital Chemistry, 2007). These methods were carried out to generate partitions containing 25, 50 and 100 clusters.

## 5.4 Evaluation of Standardization Methods

The evaluation of standardization methods was carried out according to the clustering results and similarity searching results in this study, discussed in Sections 5.4.1 and 5.4.2 respectively.

### 5.4.1 Evaluation Based on Clustering Results

Two types of evaluation techniques were employed for analyzing the clustering results. One is the calculation of average probability that clusters could be active over the eleven activity classes with each standardization method (Matter, 1997). The other technique, Shannon Entropy, is to evaluate the distribution of active compounds from inactives for a given activity class across all clusters (Matter, 1997). Both evaluation techniques were discussed in detail in Chapter 4.

### 5.4.2 Evaluation Based on Similarity Searching Results

A similarity searching technique was carried out for analyzing the varied standardization methods. The majority of molecular attributes in this study are calculated physicochemical properties such as Pipeline Pilot and Molconn-Z representations; hence the distance coefficient, Euclidean distance listed as follows (Willett et al., 1998), was used for similarity searching.

Euclidean distance: $D_{AB} = \sqrt{\sum_{i=1}^{N}(x_{iA} - x_{iB})^2}$

where $D_{AB}$ is the distance between compound A and B.

$x_{iA}$ is the value of $i^{th}$ descriptor for compound A.

$x_{iB}$ is the value of $i^{th}$ descriptor for compound B.

$N$ is the number of descriptors.

A random set of 10 known active compounds for each of the 11 activity classes (listed in Tables 4-1 & 4-2) was selected as the reference compounds for similarity calculation, and then the distance was calculated for database compounds based on the Euclidean coefficient. The number of compounds within the same activity class was counted from the top-ranked 100 and 500 database compounds. These counts were employed to compute the recovery rate, i.e., the number of actives divided by the size of a given activity class. Eventually, for each activity class, the mean recovery rate was computed averaged over 10 independent reference compounds.

### 5.4.3 Evaluation of Correlation among Structural Representations

Kendall's $W$ test of concordance was used here to evaluate the significance of the correlation; this was discussed in detail in Chapter 4. As mentioned earlier, two datasets with three chemical representations, Pipeline Pilot, Molconn-Z and Holograms, were used in this experiment. Hence, we considered each single representation as a judge, i.e. k=3, ranking the eight different standardization methods, i.e. N=8, according to the results from clustering and similarity searching in the order of decreasing effectiveness.

## 5.5 Results and Discussions of Clustering Results

In this section, we first consider the performance of clustering methods (Section 5.5.1), then have a more detailed analysis on the effect of the representations (Section 5.5.2), the number of clusters (Section 5.5.3) and, most importantly, the standardization methods (Section 5.5.4).

### 5.5.1 Evaluation of Clustering Methods

The overall results that we obtained are detailed in Table 5-2 (for (a) MDDR and (b) IDAlert datasets). In each case, the results are averaged probability or Shannon Entropy over all eight different standardization methods. The probability shows the percentage of successful prediction, that is, the higher the probability the better quality of clustering results. On the contrary, the Shannon Entropy represents how split the actives are. Hence, the larger the Shannon Entropy, the worse quality of clustering results, since for a good clustering result, all actives of a certain class should be grouped together.

As for the overall results of the MDDR datasets (Table 5-2(a)), no significant benefit was found on using K-Means and Ward's methods in the evaluation using the probability of correct prediction, whereas the evaluation using Shannon Entropy, Ward's method has consistently better performance over different numbers of clusters than K-Means method. The actives in the Ward's clustering are more concentrated among clusters than K-Means. In addition, the Hologram representation with either clustering method always has the best values of Shannon Entropy.

The overall results for the IDAlert datasets (Table 5-2(b)) have a similar trend to the MDDR datasets. Again, no clustering method was found offering consistently better probability of correct prediction, that is, no significant difference between using K-Means and Ward's methods. Moreover, the Ward's method has consistently better values of Entropy across all numbers of clusters than the K-Means method, however no specific chemical representation was

found providing consistently better performance with either clustering method.

MDDR datasets

| Probability | K-Means | | | Ward's | | |
|---|---|---|---|---|---|---|
| # clusters | Pipeline Pilot | Molconn-Z | Holograms | Pipeline Pilot | Molconn-Z | Holograms |
| 100 | 0.84 | 0.83 | 0.70 | 0.77 | 0.75 | 0.68 |
| 50 | 0.64 | 0.64 | 0.60 | 0.61 | 0.67 | 0.62 |
| 25 | 0.55 | 0.54 | 0.54 | 0.56 | 0.58 | 0.55 |
| Entropy | K-Means | | | Ward's | | |
| # clusters | Pipeline Pilot | Molconn-Z | Holograms | Pipeline Pilot | Molconn-Z | Holograms |
| 100 | 4.27 | 4.00 | 3.02 | 3.94 | 3.54 | 2.63 |
| 50 | 3.54 | 3.40 | 2.58 | 3.27 | 2.83 | 2.08 |
| 25 | 2.77 | 2.59 | 2.14 | 2.42 | 2.17 | 1.64 |

(a)

IDAlert datasets

| Probability | K-Means | | | Ward's | | |
|---|---|---|---|---|---|---|
| # clusters | Pipeline Pilot | Molconn-Z | Holograms | Pipeline Pilot | Molconn-Z | Holograms |
| 100 | 0.77 | 0.71 | 0.72 | 0.71 | 0.69 | 0.67 |
| 50 | 0.61 | 0.59 | 0.60 | 0.57 | 0.61 | 0.60 |
| 25 | 0.46 | 0.51 | 0.55 | 0.47 | 0.54 | 0.54 |
| Entropy | K-Means | | | Ward's | | |
| # clusters | Pipeline Pilot | Molconn-Z | Holograms | Pipeline Pilot | Molconn-Z | Holograms |
| 100 | 4.62 | 4.71 | 4.30 | 4.29 | 4.05 | 4.18 |
| 50 | 4.02 | 3.93 | 3.77 | 3.69 | 3.11 | 3.57 |
| 25 | 3.24 | 2.97 | 3.17 | 2.99 | 2.37 | 2.78 |

(b)

Table 5-2 The overall clustering results of the (a) MDDR and (b) IDAlert datasets

### 5.5.2 Evaluation of Structural Representations

We first consider the inspection of the overall clustering results across varied standardization methods on the MDDR datasets Table 5-2(a). No structural representation offers the consistently best probability, whereas the Hologram representation provides consistently the best Entropy with either clustering method. The other manner of inspection is carried out to analyze the clustering results on individual structural representation on the MDDR datasets (Figures 5-1 and 5-2). We first inspect the evaluation using probability (Figures 5-1) on the dataset with no standardization procedure, i.e. $S_0$, the Molconn-Z representation has consistently the best performance with only K-Means clustering across all partition sizes, whereas the Holograms has consistently the worst. Moreover, using the same evaluation criterion on the datasets with different standardization procedures, no chemical representation was found offering consistently the best values of probability. Secondly, the evaluation using Shannon Entropy (Figures 5-2) on the datasets with all standardization procedures, i.e. $S_0$-$S_7$, no structural representation was found providing consistently the best values of Shannon Entropy. Hence, for the clustering results of MDDR datasets here, we can conclude that there is no obvious difference for selecting any one of the chemical representations.

Similar pattern of analysis is also carried out on the IDAlert datasets. According to the overall clustering results over all standardization procedures listed in Table 5-2(b), it shows that there is no significant difference on choosing any one of these three chemical representations. The other type of inspection of the clustering results on individual chemical representation is also carried out (Figures 5-3 and 5-4). Figure 5-3 shows the evaluation using probability of correct prediction on the IDAlert datasets with different clustering methods over three different numbers of clusters. We found that there is no chemical representation consistently providing the best or worst performance among all standardization methods. Moreover, with the evaluation using Shannon Entropy (Figure 5-4), the Molconn-Z representation with $S_0$ and $S_6$ standardization methods has significantly low Entropy values especially

clustering by Ward's method. In addition, when the datasets employing no standardization procedure, i.e. $S_0$, with Ward's clustering, Molconn-Z has consistently the best and Pipeline Pilot has consistently the worst performance.

Apparently, some outliers were found in the above evaluation of structural representations (Figures 5-1 to 5-4). For example, the MDDR datasets using $S_6$ standardization with Molconn-Z representation clustered by the Ward's method has extremely high values of probability especially on smaller partition sizes, such as 50 and 25 clusters. A similar outlier was also found in the evaluation using Shannon Entropy with the same standardization procedure, structural representation and clustering method on 25 clusters. The other example is the IDAlert datasets using $S_6$ standardization: the Molconn-Z representation with Ward's method has the smallest probability on 100 clusters partitioning and has extremely largest probability on 50 and 25 clusters partitionings. Moreover, this was also found in the evaluation using Shannon Entropy especially on 50 and 25 clusters partitionings.

To sum up, there is something in common in the cases of above abnormal outliers. These extreme values of evaluation always occurred with smaller partition sizes of the Ward's clustering on the Molconn-Z representation dataset using $S_6$ standardization. It is difficult to identify the cause of these extreme values coming from clustering method, standardization procedure, chemical representation or even partition size.

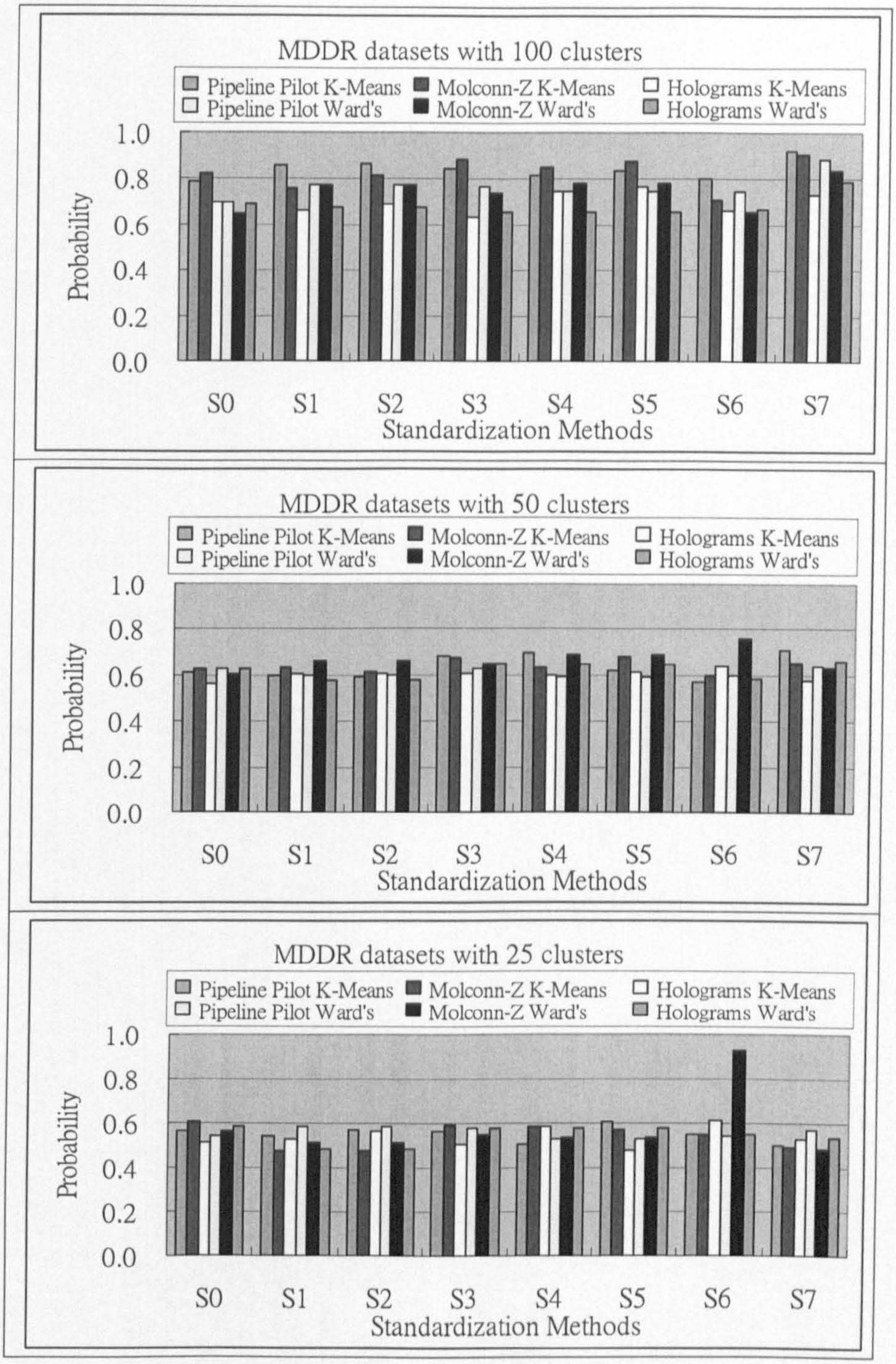



Figure 5-1 The evaluation using probability of correct prediction of the combination of clustering methods and representations on different standardization procedures of the MDDR datasets

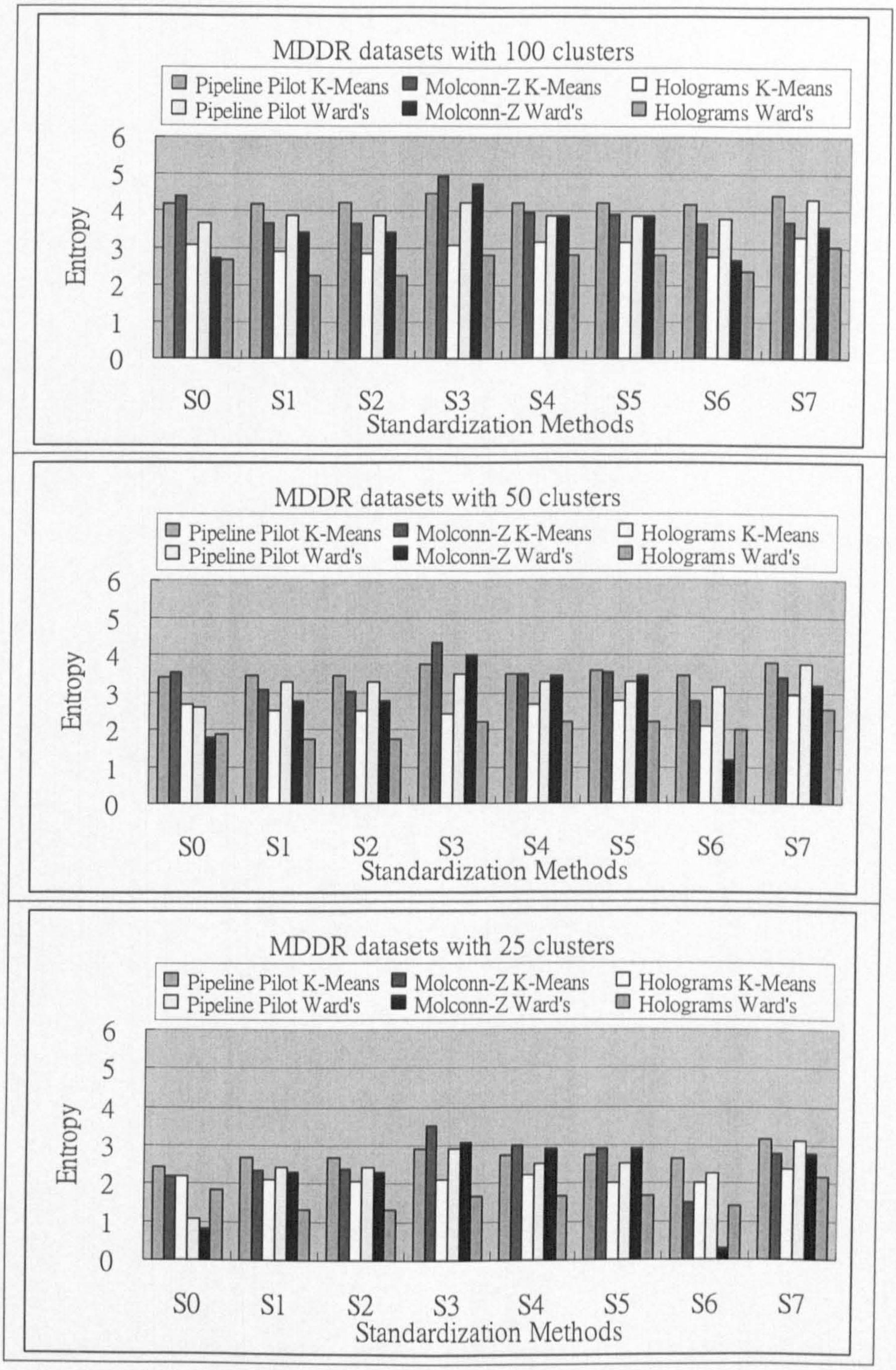



Figure 5-2 The evaluation using Entropy of the combination of clustering methods and representations on different standardization procedures of the MDDR datasets

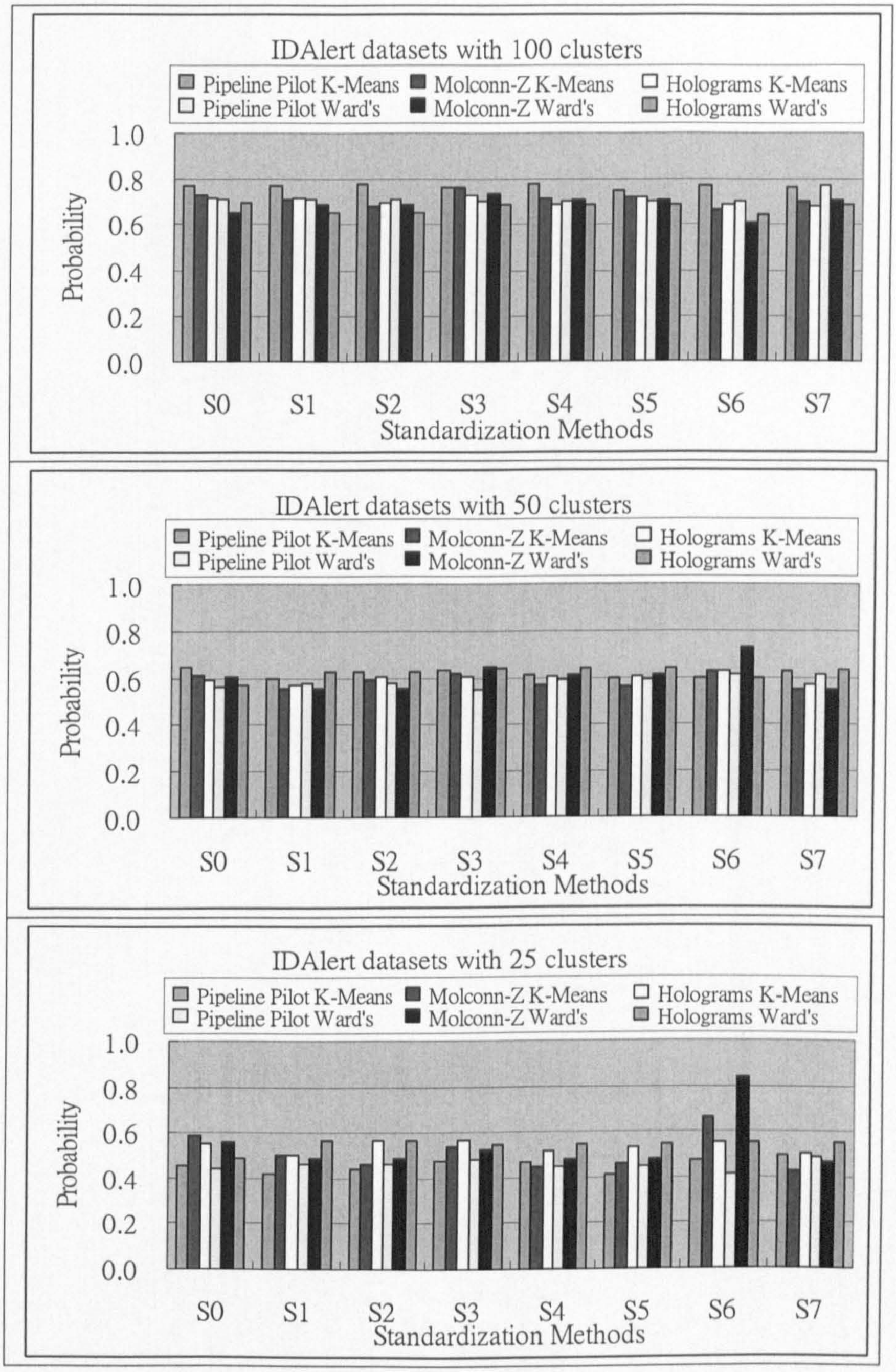



Figure 5-3 The evaluation using probability of correct prediction of the combination of clustering methods and representations on different standardization procedures of the IDAlert datasets

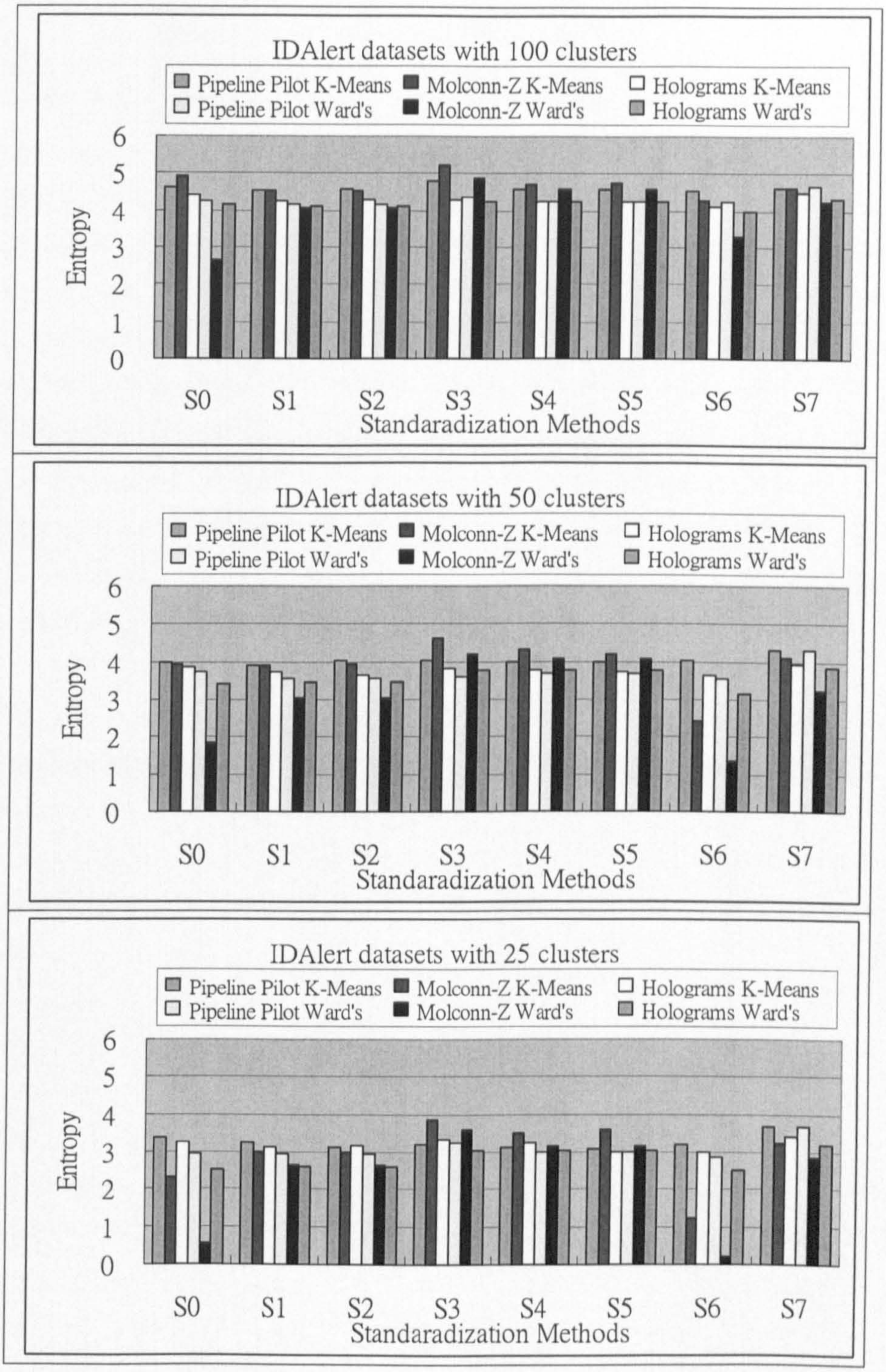



Figure 5-4 The evaluation using Entropy of the combination of clustering methods and representations on different standardization procedures of the IDAlert datasets

### 5.5.3 Evaluation of the Number of Clusters

In this section, the focus in on finding the optimal number of clusters to the clustering results. First, consider first the representation of Pipeline Pilot. Figure 5-5 illustrates the comparison of evaluation based on different numbers of clusters on the MDDR datasets. The consistent trend reveals that the larger number, e.g. 100, of clusters has larger probability and Entropy. As discussed in the Section 4.4.2, evaluation using probability of correct prediction, the smaller partition size is, the cluster is more likely to be active, i.e. high value of probability. Also, evaluations using Shannon Entropy, the smaller the partition size is, the more likely the actives are to be scattered, i.e. high value of Entropy. This is because, in addition to the applicability of a clustering algorithm to the dataset, these two evaluation criteria, in essence, naturally depend on the number of clusters.

The similar trend was also found in the MDDR datasets with the other two chemical representations, i.e. Molconn-Z and Holograms. Generally speaking, the overall trend is that the larger number of clusters, the larger probability and Entropy. However, some exception occurs on the Molconn-Z datasets using $S_6$ standardization with Ward's clustering on small numbers, e.g. 25 and 50, of clusters.

We also pick the IDAlert datasets with Pipeline Pilot representation (Figure 5-6) as an example to inspect the performance over different numbers of clusters. Similar trend was also found in the IDAlert datasets with the other two representations as in the MDDR datasets. To combine the analysis on the MDDR and IDAlert datasets together, it is concluded that the performance of varied cluster sizes strongly depends on the evaluation criteria. For the study here, the optimal number of clusters can be determined based on individual evaluation criterion, but it is hard to decide based on overall evaluation criteria.

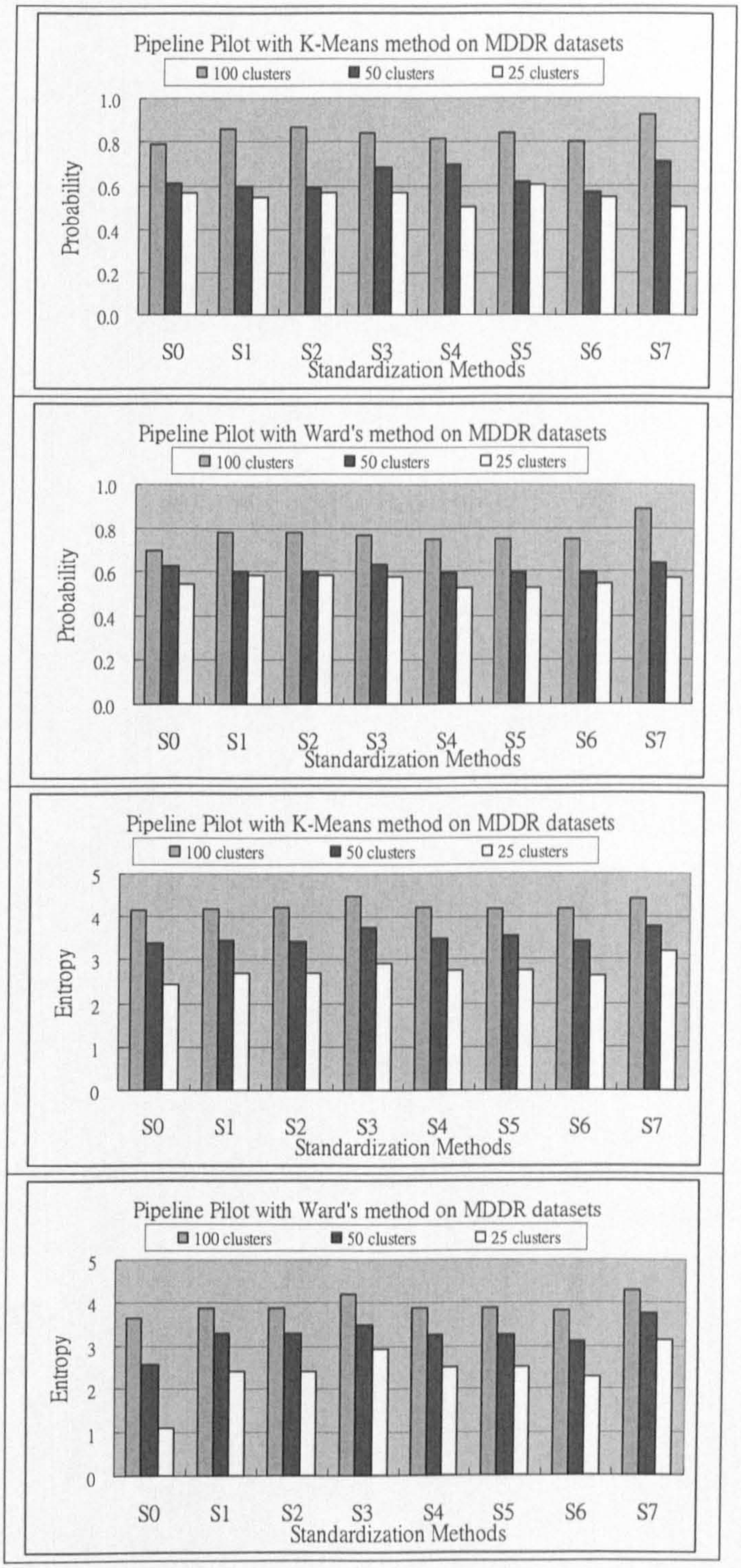



Figure 5-5 Comparison of the evaluation based on the number of clusters on the MDDR datasets

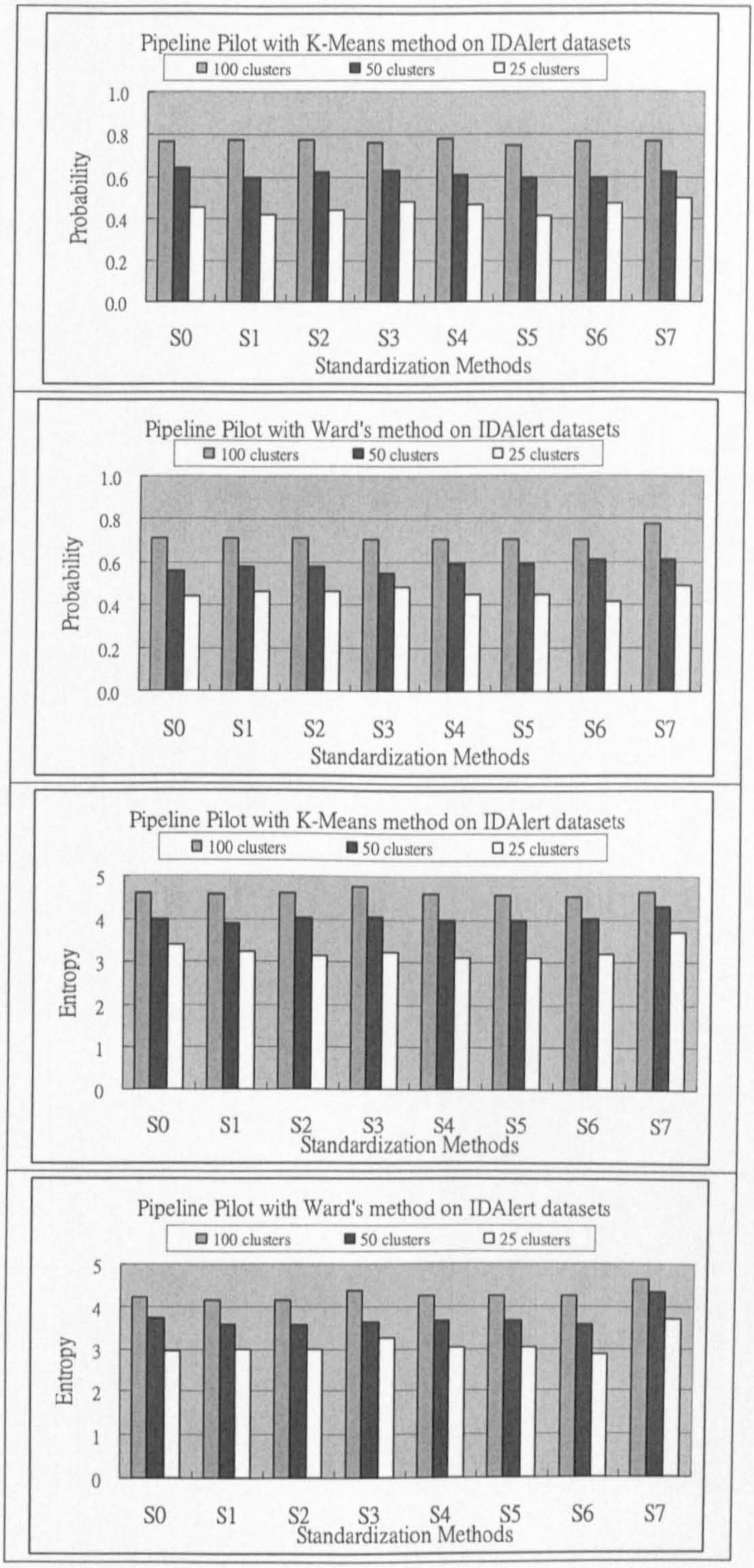



Figure 5-6 Comparison of the evaluation based on the number of clusters on the IDAlert datasets

### 5.5.4 Evaluation of Data Standardization Methods

In the context of this study, the most important part of the results is the effect of the various standardization methods. These results are shown in Figures 5-7 and 5-8 (MDDR datasets) and 5-9 and 5-10 (IDAlert datasets).

We first consider the results on the MDDR datasets based on individual evaluation criterion. Figure 5-7(a) shows the Pipeline Pilot representation evaluated by the probability of correct prediction. Apparently, $S_7$ procedure has consistently the best performance when the cluster sizes are 100 and 50. However, no standardization method was found providing consistently the best performance over all combinations of clustering methods and numbers of clusters. With the inspection of Figure 5-7(b), $S_7$ method has consistently the best performance only when the partition size is 100. As we discussed in Section 5.5.2, $S_6$ method has extremely good results when dealing with Molconn-Z datasets using Ward's clustering on small number of clusters. Hence, $S_6$ procedure has the best values of probability with Ward's clustering on partition size 50 and 25. As for the performance of Holograms listed in Figure 5-7(c), no single best standardization procedure was found consistently effective over all combinations of clustering methods and numbers of clusters.

The evaluation using Shannon Entropy on the MDDR datasets with different structural representations is listed in Figure 5-8. In terms of the Pipeline Pilot (Figure 5-8(a)), $S_0$ method offers consistently the best values of Shannon Entropy across all combinations of clustering methods and partition sizes, and this also indicates that no benefit can be obtained from using any one of these standardization methods. However, with the evaluation on Molconn-Z datasets (Figure 5-8(b)), the result was also consistent. $S_6$ standardization provides consistently the best performance over all combinations of clustering methods and cluster sizes. As for the Holograms datasets (Figure 5-8(c)), when it employs the K-Means clustering, $S_6$ procedure has consistently the best performance over all partition sizes. While clustering using Ward's method, $S_1$ and $S_2$ methods have the identical best Entropy over all partition sizes.

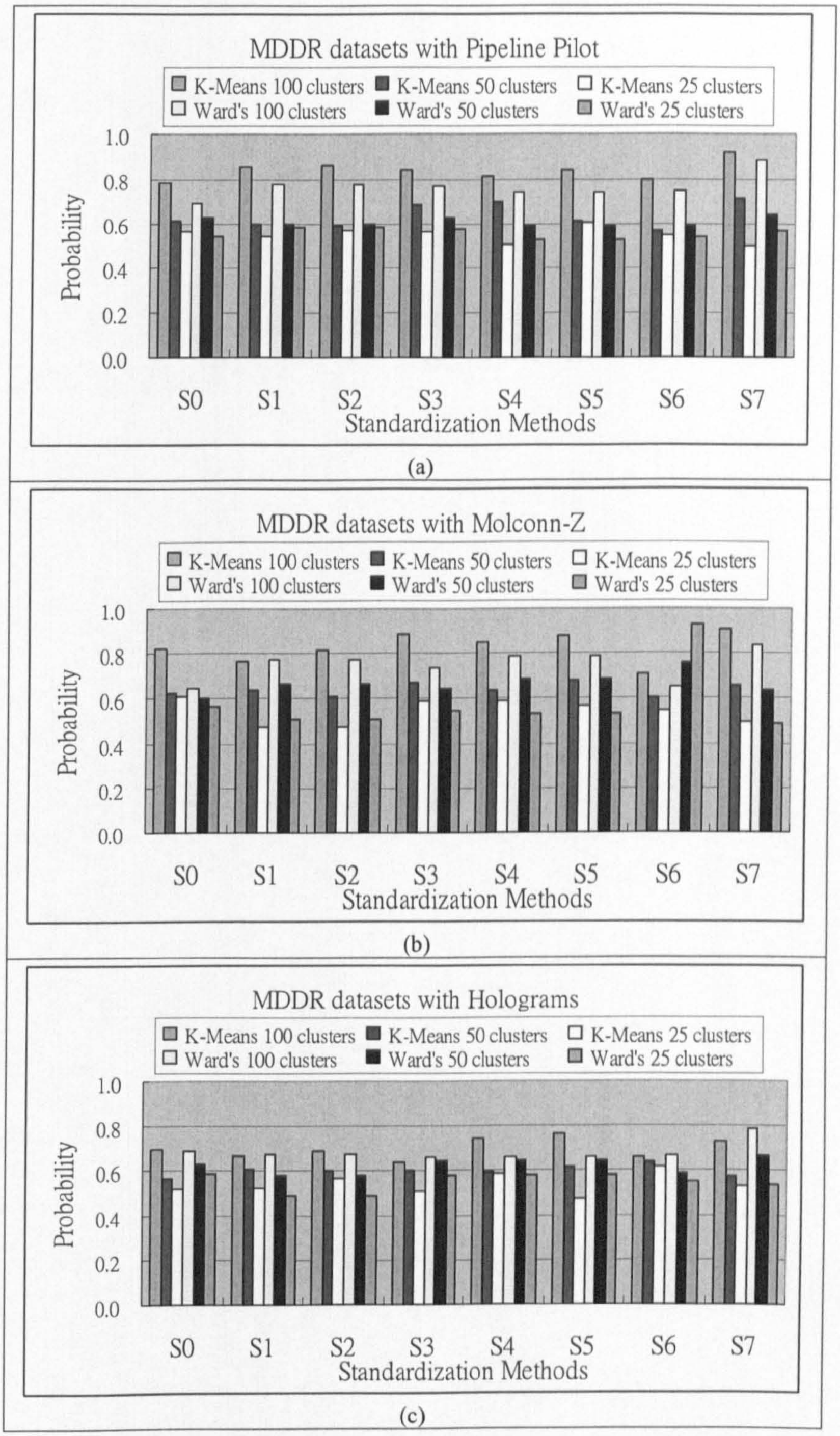



Figure 5-7 Comparison of standardization methods evaluating by probability of correct prediction on the MDDR datasets with (a) Pipeline Pilot, (b) Molconn-Z and (c) Holograms

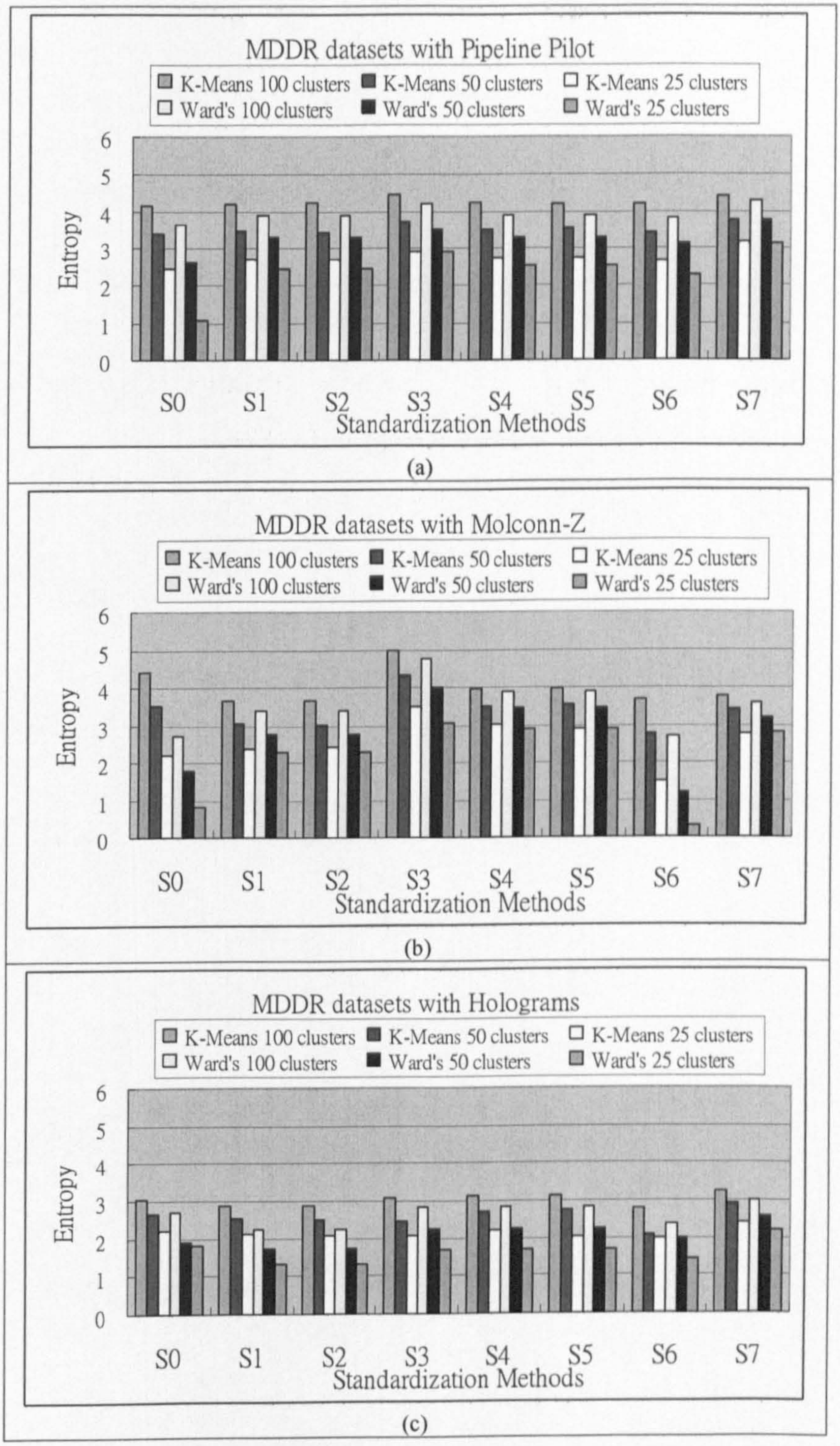



Figure 5-8 Comparison of standardization methods evaluating by Shannon Entropy on the MDDR datasets with (a) Pipeline Pilot, (b) Molconn-Z and (c) Holograms

Combining the above analysis in Figures 5-7 and 5-8, no single standardization procedure was found consistently effective over all combinations of evaluation criteria, clustering methods and partition sizes. Some results (Figures 5-8(a)) even show that the datasets without using any standardization methods would have a superior performance to standardized datasets. We hence summarized the single best standardization method for the combination of chemical representations, evaluation criteria, clustering methods and partition sizes in Table 5-3.

Evaluation using Probability

| | K-Means | | | Ward's | | |
|---|---|---|---|---|---|---|
| #clusters | 100 | 50 | 25 | 100 | 50 | 25 |
| Pipeline Pilot | $S_7$ | $S_7$ | $S_5$ | $S_7$ | $S_7$ | $S_1$,$S_2$ |
| Molconn-Z | $S_7$ | $S_5$ | $S_0$ | $S_7$ | $S_6$ | $S_6$ |
| Holograms | $S_5$ | $S_6$ | $S_6$ | $S_7$ | $S_7$ | $S_0$ |

Evaluation using Entropy

| | K-Means | | | Ward's | | |
|---|---|---|---|---|---|---|
| #clusters | 100 | 50 | 25 | 100 | 50 | 25 |
| Pipeline Pilot | $S_0$ | $S_0$ | $S_0$ | $S_0$ | $S_0$ | $S_0$ |
| Molconn-Z | $S_6$ | $S_6$ | $S_6$ | $S_6$ | $S_6$ | $S_6$ |
| Holograms | $S_6$ | $S_6$ | $S_6$ | $S_1$,$S_2$ | $S_1$,$S_2$ | $S_1$,$S_2$ |

Table 5-3 The best standardization method(s) evaluated by different criteria on the MDDR datasets

According to the overall results listed in Table 5-3, $S_6$ method tends to be more effective than any others on the MDDR datasets, and has the best performance 13 times out of 36. The study of Milligan and Cooper (1988) reported that those standardization approaches involving division by the range, such as $S_4$ and $S_5$, have better performance than other methods. But in our study, we did not find any obvious advantage from those standardization methods. In addition, the effectiveness of standardization method tends to depend on the types of chemical representation and evaluation criterion.

We then inspect the results on the IDAlert datasets also based on individual evaluation criterion. Figure 5-9 represents the evaluation using the probability of correct prediction with different types of structural representations. As for the Pipeline Pilot datasets Figure 5-9(a), the performance of standardization methods with a certain partition size is very close to each other. However, no single best procedure was found to keep offering the best probability values. Similar trend was also found in Figure 5-9(c), with the Holograms datasets, no standardization method has consistently the best results over all combinations of clustering methods and partition sizes. With the inspection of the evaluation on the Molconn-Z datasets (Figure 5-9(b)), $S_3$ standardization has consistently the best performance over the two clustering methods when dealing with the partitioning of 100 clusters. While $S_6$ method yields consistently the best values of probability over these two different partitioning approaches when the number of clusters is 50 or 25.

Figure 5-10 represents the evaluation using Shannon Entropy on the IDAlert datasets with different structural representations. In terms of the Pipeline Pilot (Figure 5-10(a)), $S_6$ standardization offers the best values of Shannon Entropy with Ward's clustering on partition size is 50 or 25. However, with the Molconn-Z datasets (Figure 5-10(b)), $S_6$ method has consistently the best performance on K-Means clustering over all numbers of clusters, and is not surprisingly having good results on Ward's clustering with small partition sizes of 25 and 50. A similar trend was also found in the Holograms datasets (Figure 5-10(c)), $S_6$ method, again, provides consistently the best Entropy on Ward's clustering over all partition sizes, and also has the leading performance on K-Means partitioning when the number of clusters is 100 or 50.

Combining the results of the MDDR and IDAlert datasets, two findings are worth noticing. First, the Ward's results of the standardization method pairs of ($S_1$,$S_2$) and ($S_4$,$S_5$) are identical, i.e. $S_1$=$S_2$ and $S_4$=$S_5$ (Figures 5-7 to 5-10) over all structural representations, evaluation criteria and partition sizes. One explanation is that there is a linear relation (see equations below) between the pairs of ($S_1$,$S_2$) and ($S_4$,$S_5$).

$$S_1 = \frac{X - \mu}{s} \qquad S_4 = \frac{X}{MAX(X) - MIN(X)}$$

$$S_2 = \frac{X}{s} \qquad S_5 = \frac{X - MIN(X)}{MAX(X) - MIN(X)}$$

Take ($S_1$,$S_2$) as an example, objects in the $S_1$ dataset are equivalent to the objects in $S_2$ dataset have a move of $\mu$ offset, in essence, on the linear relationship. Although their coordinates or positions in the vector space are different, their pairwise distances are the same. Hence, clustering algorithms based on the distance measuring, e.g. Ward's method, with the ($S_1$,$S_2$) or ($S_4$ ,$S_5$) standardized datasets will obtain equivalent dissimilarity (or similarity), and this will also naturally lead to the identical clustering result. However, above behaviour would not apply to K-Means clustering, since different sets of random seeds are picked in different runs from the dataset by the BCI software as the initial centroids of each cluster, and the final clustering result strongly depends on these initial random seeds.

Secondly, when Ward's clustering deals with Holograms datasets, $S_3$, $S_4$ and $S_5$ methods have the exactly identical values of probability (Figures 5-7(c) and 5-9(c)) and Entropy (Figures 5-8(c) and 5-10(c)). Such behaviour simply comes from the characteristic of Holograms fingerprints. As we mentioned in Section 4.2.3, each descriptor (or bit) in the molecular holograms representation records the number of times a unique fragment occurs in a given molecule. A Hologram molecule contains 997 descriptors, however most of the descriptors have the value of zero, i.e. absence of a certain fragment. In that case, considering the standardization procedures of $S_4$ and $S_5$, the minimum for those descriptors is zero.

$$S_3 = \frac{X}{MAX(X)}$$

Since the minimum is zero, i.e. $MIN(X) = 0$, the standardization procedures of $S_4$ and $S_5$ will hence be identical to $S_3$. In addition, this situation would not happen on the Monconn-Z and Pipeline Pilot representations, because their manner of descriptors calculating is different from Holograms.

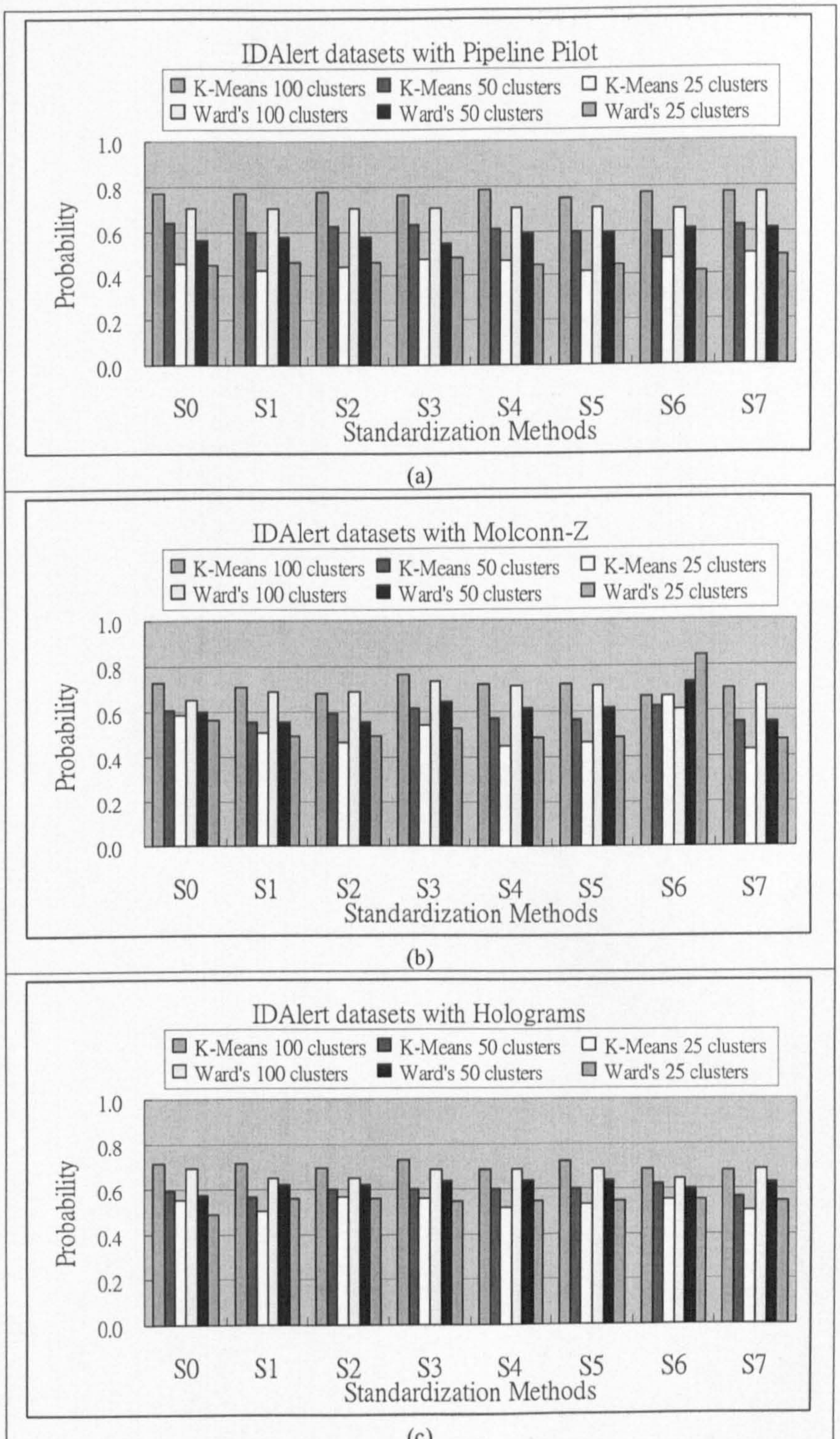



Figure 5-9 Comparison of standardization methods evaluating by probability of correct prediction on the IDAlert datasets with (a) Pipeline Pilot, (b) Molconn-Z and (c) Holograms

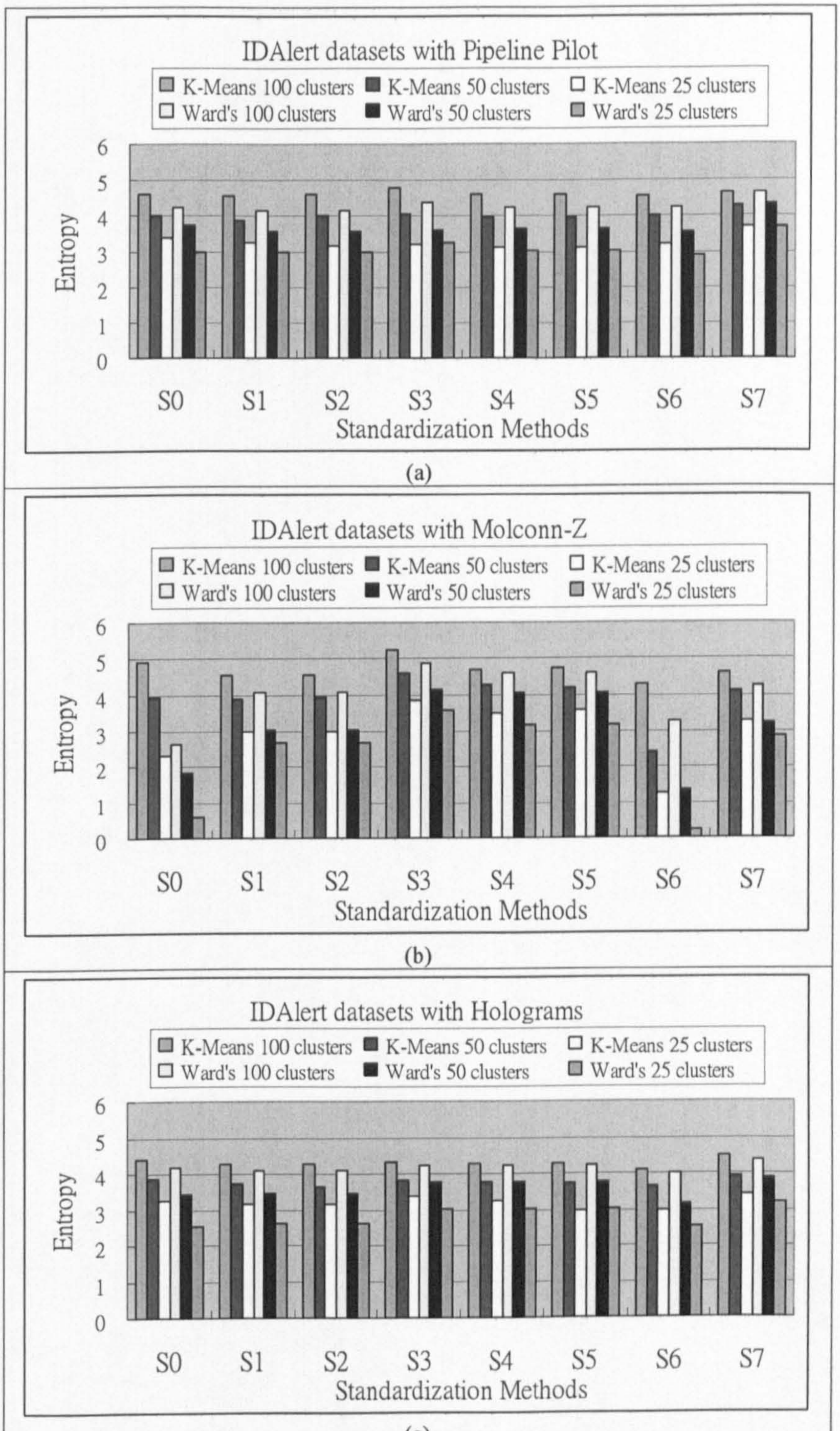



Figure 5-10 Comparison of standardization methods evaluating by Shannon Entropy on the IDAlert datasets with (a) Pipeline Pilot, (b) Molconn-Z and (c) Holograms

Above analysis of the IDAlert datasets were summarized in Figures 5-9 and 5-10: as with the MDDR datasets, no significant benefit was obtained on choosing any one of the standardization procedures for our study here. Table 5-4 summarized the single best standardization method for the combination of chemical representations, evaluation criteria, clustering methods and partition sizes on the IDAlert datasets.

Evaluation using Probability

| | K-Means | | | Ward's | | |
|---|---|---|---|---|---|---|
| #clusters | 100 | 50 | 25 | 100 | 50 | 25 |
| Pipeline Pilot | $S_4$ | $S_0$ | $S_7$ | $S_7$ | $S_6$ | $S_7$ |
| Molconn-Z | $S_3$ | $S_6$ | $S_6$ | $S_3$ | $S_6$ | $S_6$ |
| Holograms | $S_3$ | $S_6$ | $S_2$ | $S_0$ | $S_3S_4S_5$ | $S_1S_2$ |

Evaluation using Entropy

| | K-Means | | | Ward's | | |
|---|---|---|---|---|---|---|
| #clusters | 100 | 50 | 25 | 100 | 50 | 25 |
| Pipeline Pilot | $S_6$ | $S_1$ | $S_5$ | $S_1S_2$ | $S_6$ | $S_6$ |
| Molconn-Z | $S_6$ | $S_6$ | $S_6$ | $S_0$ | $S_6$ | $S_6$ |
| Holograms | $S_6$ | $S_6$ | $S_5$ | $S_6$ | $S_6$ | $S_6$ |

Table 5-4 The best standardization method(s) evaluating by different criteria on the IDAlert datasets

According to the overall results listed in Table 5-4, $S_6$ method tends to be more effective on the IDAlert datasets, and has the best performance 19 times out of 36. However, $S_4$ and $S_5$ were not found effective as reported in the study of Milligan and Cooper (1988). Moreover, $S_7$ tends to yield worse results on the values of Shannon Entropy especially with the Pipeline Pilot and Holograms representations. One possible cause is that the object function of clustering algorithm and standardization method applied on a dataset are two vital components for clustering. The aim of standardization is to adjust the magnitude or scale of the score of input variables to be equal. However, a proper standardization procedure can keep the magnitude of dissimilarity (or similarity) between objects after standardizing, and which is ideally to obtain good quality of clustering. Among these eight standardization methods discussed in Section 5-2, $S_7$ is one of the procedures, which loses more dissimilarity or similarity between objects after standardizing. Take $S_6$ and $S_7$ together as an example:

$$S_6 = \frac{X}{\sum X} \qquad S_7 = Rank(X)$$

with four objects A=1, B=100, C=9999 and D=10000. We standardized these four objects by above two procedures, and get the results as follows:

| | | | | |
|---|---|---|---|---|
| $S_6$: | A≒0.00005 | B≒0.005 | C≒0. 5 | D≒0. 5 |
| $S_7$: | A=1 | B=2 | C=3 | D=4 |

Obviously, according to the above example, the significant difference between these four objects remains after using $S_6$ procedure, while the difference becomes much less significant after using $S_7$ procedure.

In addition, with the wide data range of the Molconn-Z descriptors, the performances of $S_7$ were average. It is interesting that the traditional *Z-Score* standardization procedure, i.e. $S_1$, revealed only ordinary performances, since it was placed either in the superior or worse group. This finding is in line with previous study (Milligan and Copper, 1988). Finally, the performance of no standardized procedure $S_0$ was not as bad as expected, and no complementary relation of performances was found between $S_0$ and $S_1$.

As mentioned in the previous passage, in the case of the IDAlert datasets here, the effectiveness of standardization method remains depending on the types of chemical representation and evaluation criterion. In order to obtain a more quantitative view of the effectiveness on these standardization procedures, the Kendall's *W* test of statistical significance was carried out to evaluate the consistency of ranking judged by these three chemical representations in the next section (5.6).

Moreover, some standardization procedures were also found having a linear equivalent relationship as in the MDDR datasets. However, in the case of the IDAlert datasets here, the pairs of ($S_1$,$S_2$) and ($S_4$,$S_5$) also have linear equivalent relationship with Ward's clustering over all structural representations, and Ward's clustering with $S_3$, $S_4$ and $S_5$ standardizations also obtains identical results on the Holograms representation.

## 5.6 Results and Discussions of Correlation Tests

The Kendall coefficient of concordance *W* was used to measure the degree of association among three different chemical representations. Thus, the three types of representation are regarded as judges of the effectiveness of the eight types of standardization. If it can be shown that there is a statistically significant level of correlation between the rankings, it will be possible to provide an overall ranking of them (Siegel and Castellan, 1988). Hence, ranked the performance obtained from a certain evaluation criterion with a given partitioning method. The values of *W* and $X^2$ (chi square) of Kendall test based on the set of rankings are then calculated. For example, the values of probability based on K-Means clustering with partition size of 100 were obtained (see Table 5-5), and then ranked the performance in descending order based on individual representation, i.e. judge (see Table 5-6).

K-Means clustering with 100 clusters on the MDDR datasets

| | $S_0$ | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ |
|---|---|---|---|---|---|---|---|---|
| Pipeline Pilot | 0.78 | 0.85 | 0.86 | 0.84 | 0.81 | 0.84 | 0.80 | 0.92 |
| Molconn-Z | 0.82 | 0.76 | 0.81 | 0.88 | 0.85 | 0.88 | 0.71 | 0.91 |
| Holograms | 0.69 | 0.66 | 0.68 | 0.63 | 0.74 | 0.76 | 0.66 | 0.72 |

Table 5-5 Evaluation using probability of K-Means clustering with 100 clusters on the MDDR datasets

K-Means clustering with 100 clusters on the MDDR datasets

| | $S_0$ | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ |
|---|---|---|---|---|---|---|---|---|
| Pipeline Pilot | 8 | 3 | 2 | 4 | 6 | 5 | 7 | 1 |
| Molconn-Z | 5 | 7 | 6 | 2 | 4 | 3 | 8 | 1 |
| Holograms | 4 | 6 | 5 | 8 | 2 | 1 | 7 | 3 |

Table 5-6 Ranks obtained by the performance of K-Means clustering with 100 clusters on the MDDR dataset

With the comparison of ranks averaged by structural representations (as the example listed in Table 5-6), the values of $W$ and $X^2$ (chi square) of Kendall test were calculated and listed in Tables 5-7 and 5-8 (MDDR datasets) and Tables 5-9 and 5-10 (IDAlert datasets). Consider first the Kendall test on the MDDR datasets. In Table 5-7, the calculation of $W$ and $X^2$ were based on the probability of correct prediction of active clusters in the MDDR datasets. The lower value of $W$ indicates that the agreement among the three structural representations is also lower, i.e. less significant. It can be seen that the $W$ value for 100 clusters by K-Means and Ward's clustering is 0.49 and 0.51 respectively, obviously higher than partition sizes of 50 and 25. It can be concluded that the ranking by three structural representations of 100 clusters is more consistent than 50 and 25 clusters. Inspect the value of $X^2$ (chi square) of each single Kendall's test to see if it reaches the significant level of 95% ($\alpha$=0.05) and 99% ($\alpha$=0.01). The critical value of the chi square distribution at $\alpha$=0.05 level with 7 degrees of freedom is 14.07 and at $\alpha$=0.01 level is 18.48. There are 8 standardization methods to be evaluated, so the degree of freedom is 7 (Siegel and Castellan, 1988).

However, none of the three combinations listed in Table 5-7 with K-Means clustering methods are significant at $\alpha$=0.05 or $\alpha$=0.01 level. It is hence can be concluded that with K-Means clustering method across all partition sizes, there is no correlation between the rankings of the standardization methods by the three chemical representations. Moreover, the Kendall's tests of Ward's clustering also show the same results, no significant agreement between the different chemical representations at both $\alpha$=0.05 and $\alpha$=0.01 level.

| | Evaluation using Probability | | | |
|---|---|---|---|---|
| | K-Means | | Ward's | |
| #clusters | $W$ | $X^2$ | $W$ | $X^2$ |
| 100 | 0.49 | 10.33 | 0.51 | 10.69 |
| 50 | 0.33 | 6.94 | 0.14 | 2.97 |
| 25 | 0.17 | 3.64 | 0.22 | 4.58 |

Table 5-7 Kendall $W$ and $X^2$ values based on the evaluation using probability of active clusters correct prediction on the MDDR datasets

Table 5-8 shows the values of $W$ and $X^2$ based on Shannon Entropy in the MDDR datasets. Inspect first the K-Means results, and found that the $X^2$ value for 25 clusters by K-Means clustering is 14.00, a little greater than the values of 50 and 100 clusters. It almost reaches the significant level of 95% (14.07). In addition, a similar trend was also found with Ward's clustering. We obtained the same values of $W$ and $X^2$ with Ward's clustering, although they had similar ranks. However, the results are close to the significant level at $\alpha$=0.05 level. Combining all tests listed in Table 5-8, none of the six combinations are significant at $\alpha$=0.05 and $\alpha$=0.01 level.

| | Evaluation using Entropy | | | |
|---|---|---|---|---|
| | K-Means | | Ward's | |
| #clusters | $W$ | $X^2$ | $W$ | $X^2$ |
| 100 | 0.60 | 12.67 | 0.66 | 13.91 |
| 50 | 0.58 | 12.11 | 0.66 | 13.91 |
| 25 | 0.67 | 14.00 | 0.66 | 13.91 |

Table 5-8 Kendall $W$ and $X^2$ values based on the evaluation using Shannon Entropy on the MDDR datasets

Overall, considering all analysis in Tables 5-7 and 5-8, none of the twelve combinations are significant at either $\alpha$=0.05 or $\alpha$=0.01 level. It is hence can be concluded that, for chemical data of the sort considered here, there is no obvious performance benefit that is likely to be obtained from the use of any particular standardization method. The choice of standardization method is hence not a critical component of a procedure for chemical clustering.

The results of Kendall test for the IDAlert datasets are shown in Tables 5-9 and 5-10. Inspect first the calculation of $W$ and $X^2$ based on the probability of correct prediction of active clusters (Table 5-9). As in the MDDR datasets, the $W$ value for clusters numbers of 100 by K-Means clustering method is 0.35, obviously higher than when clusters numbers are 50 and 25. It indicates that with larger partition size, e.g. 100 clusters, there is more correspondent among the ranking by three structural representations than the clusters numbers of 50

and 25. However, none of the three combinations listed in Table 5-9 with K-Means clustering methods are significant at either α=0.05 or α=0.01 level, their chi square values ($X^2$) are still far from the critical value (14.07 at α=0.05 level). Similar results of Ward's clustering also showed that no combinations were found significant at either α=0.05 or α=0.01 level.

| | Evaluation using Probability | | | |
|---|---|---|---|---|
| | K-Means | | Ward's | |
| #clusters | *W* | $X^2$ | *W* | $X^2$ |
| 100 | 0.35 | 7.33 | 0.57 | 12.07 |
| 50 | 0.25 | 5.33 | 0.12 | 2.62 |
| 25 | 0.28 | 5.89 | 0.51 | 10.80 |

Table 5-9 Kendall *W* and $X^2$ values based on the probability of active clusters correct prediction on the IDAlert datasets

The calculation of *W* and $X^2$ based on Shannon Entropy is hence carried out (Table 5-10). A similar trend as in Tables 5-7 and 5-9 is also found here, clustering with larger partition size, e.g. 100, tends to have higher values of *W* and $X^2$. K-Means clustering with 100 clusters, and Ward's clustering with 100 and 50 clusters have reached the significant level at α=0.05. However, none of the combinations are significant at α=0.01 level. Six Kendall tests were carried out in Table 5-10, whereas only three out of six are significant at α=0.05, but none of them is significant at α=0.01 level. The result for the IDAlert datasets is insufficient to show the significant correlation among standardization methods. As the finding in the case of MDDR datasets, no consistence between these three structural representations in the case of IDAlert datasets, no benefit can be obtained from choosing any particular standardization method.

| | Evaluation using Entropy | | | |
|---|---|---|---|---|
| | K-Means | | Ward's | |
| #clusters | *W* | $X^2$ | *W* | $X^2$ |
| 100 | 0.74 | 15.56 | 0.83 | 17.37 |
| 50 | 0.66 | 13.89 | 0.88 | 18.41 |
| 25 | 0.59 | 12.44 | 0.63 | 13.22 |

Table 5-10 Kendall *W* and $X^2$ values based on the values of Shannon Entropy on the IDAlert datasets

Overall, taking the results of the MDDR and IDAlert datasets together, it is found that the Kendall *W* and $X^2$ values based on the evaluation using Shannon Entropy are higher than the probability of correct prediction. Moreover, among the overall 24 Kendall tests, only three of them reach the significant level of $\alpha$=0.05. It can be concluded that, for chemical data of the sort considered here, there is no obvious performance benefit that is likely to be obtained from the use of any particular standardization method. The choice of standardization method is hence not a critical component of a procedure for chemical clustering.

## 5.7 Results and Discussions of Similarity Searching

The evaluation of standardization in the previous two sections (5.5 and 5.6) is based on the clustering results. However, in this section, the evaluation is based on the recovery rates from similarity searching, which was discussed in detail in Section 5.4.2. The evaluation based on similarity searching results for the MDDR and IDAlert datasets is discussed in Sections 5.7.1 and 5.7.2 respectively.

### 5.7.1 Analysis of Similarity Searching Results of the MDDR Dataset

The similarity searching results that we obtained are detailed in Table 5-11. In each case, the results are averaged recovery rates over all eleven different activity classes used in the MDDR dataset.

| | MDDR Datasets | | | | | |
|---|---|---|---|---|---|---|
| Standardization | Pipeline Pilot | | MolconnZ | | Holograms | |
| Methods | Top 100 | Top 500 | Top 100 | Top 500 | Top 100 | Top 500 |
| $S_0$ | 5.53% | 18.76% | 4.23% | 16.03% | 14.44% | 22.95% |
| $S_1$ | 6.60% | 22.48% | 12.62% | 27.69% | 13.48% | 20.92% |
| $S_2$ | 6.60% | 22.48% | 12.62% | 27.69% | 13.48% | 20.92% |
| $S_3$ | 4.55% | 14.87% | 1.99% | 9.78% | 13.17% | 20.27% |
| $S_4$ | 6.65% | 22.04% | 10.98% | 25.25% | 13.17% | 20.27% |
| $S_5$ | 6.65% | 22.04% | 10.98% | 25.25% | 13.17% | 20.27% |
| $S_6$ | 6.07% | 19.33% | 8.44% | 21.61% | 13.16% | 20.36% |
| $S_7$ | 6.14% | 20.37% | 13.38% | 29.79% | 15.23% | 26.61% |
| Average | 6.10% | 20.30% | 9.41% | 22.89% | 13.66% | 21.57% |

Table 5-11 The recovery rates of 3 chemical representations of the MDDR datasets over 11 different activity classes

#### 5.7.1.1 Evaluation of Standardization Methods Based on Similarity Searching Results of the MDDR Dataset

According to the details listed in Table 5-11, $S_3$ offers noticeably worst performance with Molconn-Z (1.99% in top 100; 9.78% in top 500) and Pipeline Pilot (4.55% and 14.87% in top 100 and 500 hit list respectively), while $S_7$ provides the best performance with Molconn-Z and Holograms. However, no standardization method provides consistently superior or worst recovery rate over the three structural representations. The Holograms standardized datasets have similar recovery rates (between 13.16 and 13.48% in top 100; between 20.27 and 20.92% in top 500) except $S_7$. Comparing the no standardization procedure ($S_0$) with others ($S_1$ to $S_7$), the performance of unstandardized datasets ($S_0$) is better than some standardized datasets, such as

$S_3$ in Pipeline Pilot and Molconn-Z, and all others in Holograms.

One noticeable finding is worth discussing here. The pairs of ($S_1$,$S_2$) and ($S_4$,$S_5$) have identical recovery rates for each individual structural representation, while $S_3$, $S_4$ and $S_5$ standardizations generate exactly the same results with the representation of Holograms. As we discussed in the Section 5.5.4, the pairs of ($S_1$,$S_2$) and ($S_4$,$S_5$) also have a linear equivalent relationship with the distance-based Ward's clustering over all structural representations, and same clustering with $S_3$, $S_4$ and $S_5$ standardizations also obtains identical results on the Holograms representation. However, the similarity searching we carried out here is based on the Euclidean distance. Hence these pairs of standardization methods will obtain the identical recovery rates as shown in Table 5-11.

#### 5.7.1.2 Evaluation of Structural Representations Based on Similarity Searching Results of the MDDR Dataset

According to the Table 5-11, in the hit list of top-ranked 100 compounds, Holograms has superior overall average recovery rate, while Molconn-Z offers better overall average recovery rate in the aspect of 500 most similar database compounds list. However, there is no chemical representation that provides consistently better performance on both top 100 and 500 hit lists. Comparing the effect between no standardization ($S_0$) and standardization ($S_1$ to $S_7$) methods, $S_0$ with Holograms tends to offer superior performances than others except $S_7$. Hence, no significant difference of performance was obtained between standardized and unstandardized datasets with Holograms. On the other hand, as for the standardized datasets, Molconn-Z standardized datasets offer apparently better recovery rates than unstandardized dataset except $S_3$. To sum up, for the application of similarity searching, dataset with Holograms may offer better recovery rate than Molconn-Z and Pipeline Pilot.

#### 5.7.1.3 Measures of Correlation among Three Structural Representations of the MDDR Dataset

As discussed in previous passage (Section 5.6), the Kendall coefficient of concordance $W$ has been employed to measure the degree of association among three structural representations. The same analysis was also done in this section. The $W$ and $X^2$ values of Kendall test are shown in Table 5-12.

In Table 5-12, the calculation of W and $X^2$ were based on the recovery rates with different numbers of compounds in the hit list. The higher value of $W$ indicates that the agreement among these three chemical representations is also higher. According to the agreement test of the top 100 searching, it can be seen that the $W$ value is 0.55 and $X^2$ is 11.49. However the critical values of the chi square distribution at $\alpha$=0.01 and $\alpha$=0.05 significant level with 7 degree of freedom are 18.48 and 14.07. Obviously, none of these two tests has reached these two significant levels. Hence, we can conclude that there is no correlation among these three chemical representations by the datasets when the data is standardized, and there is no consistent ranking of the standardization methods.

| | MDDR datasets | |
|---|---|---|
| Hit rates | $W$ | $X^2$ |
| Top 100 | 0.55 | 11.49 |
| Top 500 | 0.62 | 13.11 |

Table 5-12 Kendall W and $X^2$ values based on the Recovery Rates of the MDDR datasets

### 5.7.2 Evaluation of Similarity Searching Results of the IDAlert Dataset

The recovery rates of similarity searching on the IDAlert dataset with different structural representations are listed in Table 5-13. In each case, the recovery rates are averaged over all eleven activity classes used in the IDAlert dataset. The evaluations of standardization methods (Section 5.7.2.1) and three structural representations (Section 5.7.2.2) were carried out based on the

recovery rates.

| Standardization Methods | IDAlert Datasets | | | | | |
|---|---|---|---|---|---|---|
| | Pipeline Pilot | | Molconn-Z | | Holograms | |
| | Top 100 | Top 500 | Top 100 | Top 500 | Top 100 | Top 500 |
| $S_0$ | 5.38% | 19.21% | 4.39% | 15.74% | 11.62% | 23.55% |
| $S_1$ | 6.64% | 21.05% | 11.13% | 26.16% | 9.14% | 18.30% |
| $S_2$ | 6.64% | 21.05% | 11.13% | 26.16% | 9.14% | 18.30% |
| $S_3$ | 4.44% | 14.63% | 2.78% | 9.66% | 9.05% | 17.95% |
| $S_4$ | 6.36% | 21.10% | 8.33% | 23.71% | 9.05% | 17.95% |
| $S_5$ | 6.36% | 21.10% | 8.33% | 23.77% | 9.05% | 17.95% |
| $S_6$ | 6.62% | 20.69% | 6.89% | 20.34% | 8.97% | 16.78% |
| $S_7$ | 6.86% | 21.39% | 13.94% | 30.86% | 11.70% | 23.93% |
| Average | 6.10% | 20.30% | 8.36% | 22.04% | 9.71% | 19.34% |

Table 5-13 The Recovery Rates of 3 Chemical Representations of the IDAlert datasets over 11 Different Activity Classes

#### 5.7.2.1 Evaluation of Standardization Methods Based on Similarity Searching Results of IDAlert dataset

With the visual inspection on Table 5-13, $S_3$ offers noticeably the worst performance with Molconn-Z (2.78% in top 100; 9.66% in top 500) and Pipeline Pilot (4.44% and 14.63% in top 100 and 500 hit list respectively), while $S_7$ provides consistently superior performance over three different chemical representations. We can conclude that for the IDAlert datasets we used here, $S_7$ method is the optimal choice on the standardization of dataset to obtain the better recovery rate.

In addition, the relationship of linear equivalence between pairs of standardization methods was also found in the case of IDAlert datasets. The pairs of ($S_1$,$S_2$) and ($S_4$,$S_5$) have identical recovery rates for each individual structural representation, while $S_3$, $S_4$ and $S_5$ standardizations generate exactly the same results with the representation of Holograms.

#### 5.7.2.2 Evaluation of Structural Representations Based on Similarity Searching Results of IDAlert dataset

According to the Table 5-13, in the hit list of top-ranked 100 compounds, Holograms has superior overall recovery rate (9.71%), while Molconn-Z offers the best overall recovery rate (22.04%) in the aspect of 500 most similar database compounds list. However, there is no chemical representation that provides consistently better performance on both top 100 and 500 hit lists. As for the unstandardized datasets ($S_0$), Holograms have remarkably better performances in both top 100 and 500 hit lists than standardized datasets ($S_1$ to $S_6$). On the contrary, Molconn-Z datasets using standardization procedures offer apparently better recovery rates than unstandardized datasets except $S_3$, whereas Holograms datasets using standardization methods provide worse performance than unstandardized datasets except $S_7$.

#### 5.7.2.3 Measures of Correlation among Three Structural Representations of IDAlert dataset

The Kendall test was employed to measure the concordance of three structural representations as listed in Table 5-11. Considering first the agreement test of top 100 searching, it can be seen that the $W$ value is 0.74 and $X^2$ is 15.64, and the critical values for chi square distribution at $\alpha$=0.01 and $\alpha$=0.05 significant level with 7 degree of freedom are 18.48 and 14.07 respectively. These two tests have reached the significant level at $\alpha$=0.05. Hence, we can conclude that there is correlation among these three chemical representations by the datasets when the data is standardized by a particular procedure. According to the analysis in Section 5.7.2.1, $S_7$ method is the best choice for the application of similarity searching.

| IDAlert datasets | | |
|---|---|---|
| Hit rates | $W$ | $X^2$ |
| Top 100 | 0.74 | 15.64 |
| Top 500 | 0.68 | 14.37 |

Table 5-14 Kendall $W$ and $X^2$ values based on the Recovery Rates of the IDAlert datasets

### 5.7.3 Summary

The first experiment in this chapter was carried out to evaluate the effect of standardization methods based on clustering results (Sections 5.5 and 5.6) and the results of similarity searching (Section 5.7). According to the analysis in those sections, there is no standardization method that provides consistently superior or worse performance in both the MDDR and IDAlert datasets at the $\alpha$=0.01 level of statistical significance, however we found statistically significant at the $\alpha$=0.05 level on the tests based on the results of similarity searching only on the IDAlert datasets. Hence, we conclude that there is no obvious performance benefit that is likely to be obtained from the use of any particular standardization method.

In terms of the comparison of structural representation, according to the analyses in the Sections 5.5.2 and 5.7, no chemical representation is found offering the consistently superior performance for the evaluation either based on the clustering results or based on the results of similarity searching. We hence conclude that there is no obvious difference for selecting any one of the chemical representations.

The first experiment in this chapter is largely focusing on evaluating the effect of standardization methods using clustering results. However, our findings show that the performance is affected by the clustering methods. For example, the pairs of ($S_1$,$S_2$) and ($S_4$,$S_5$) generate the same results when using Ward's method (distance-based method). We hence carried out the evaluation of standardization methods with more and diverse clustering methods in the extensive study in the next sections.

## 5.8 Extensive Study of the Effect of Standardization Methods

We have carried out an extensive study based on the previous sections (5.5 to 5.7) in this chapter. There are three differences between these two studies of this chapter. First, seven clustering methods were employed instead of two. Second, the partition size here contains 500, 600, 700, 800, 900 and 1000 clusters. Finally, the evaluation using the probability of correct prediction and Shannon Entropy have been replaced by F-Measure and QCI (discussed in Chapter 4) here. The main goal is again to find the effect of standardization procedures on the chemical clustering.

## 5.9 Experimental Details of the Extensive Study

### 5.9.1 Datasets

The same datasets, MDDR and IDAlert, were used in this experiment. The same three chemical representations were also employed, the only difference is that we used an alternative tool, winMolconn software, to generate win_Molconn representation for the experiment here. Again, during the process of calculating Molconn descriptors, some molecules fail to generate descriptors. We hence removed those molecules from datasets with all chemical representations to obtain equal size of datasets. These datasets eventually comprised 10,179 molecules from the MDDR dataset and 11,447 molecules from the IDAlert dataset. For each dataset, with the combination of standardization procedures and chemical representations, we hence obtained 24 test-datasets.

### 5.9.2 Clustering Methods

Seven clustering methods were used in this experiment. The Ward's and extended Ward's methods were carried out using the Energy package in the R software, and denoted by *WD* and *EW* respectively in the tables and figures of later context. The other five methods were carried out using the implementations in CLUTO (for CLUstering TOolkit) software package.

Direct and Repeated Bisection methods employed two criterion functions, *e1* and *i2*, for each method, and denoted by *D_e1*, *D_i2*, *RB_e1* and *RB_i2* in the tables and figures of the later paragraphs. The final method is the traditional agglomerative clustering with the criterion function of UPGMA (for Unweighted Pair Group Method using Arithmetic mean), also known as average linkage, and denoted by *UPGMA* in the tables and figures of the later context. All these clustering methods were discussed in detail in Chapter 4.

### 5.9.3 Standardization Procedures

Eight standardization methods, i.e. $Z_0$ to $Z_7$, were employed in this experiment, where $Z_0$ denotes the original, unstandardized dataset. All these standardization procedures are discussed in detail in Section 5.2. However for the same eight standardization methods used here, we use a different notation, $Z_0$ to $Z_7$ to distinguish different experiments in the same chapter.

### 5.9.4 Evaluation Criteria

The combinations of 2 datasets, 3 chemical representations, 8 standardization procedures, 7 clustering methods and with 6 partition sizes, hence generated 2016 clustering results, which were evaluated by F-Measure and QCI (Quality Clustering Index), which are discussed in detail in Chapter 4. As discussed in previous section (4.4.2), the evaluation using probability of correct prediction is not applicable to the case of small clusters here.

## 5.10 The Comparison between Standardization Procedures

The performance criteria of F-Measure and QCI were computed for each clustering result which was based on the combination of dataset, partition size, chemical representation and standardization procedure. Tables 5-15 (for the MDDR dataset) and 5-16 (for the IDAlert dataset) show the best single standardization procedure offering the best evaluation for the combination of dataset, partition size and chemical representation. For example, in Table 5-15(a), the best F-Measure value on Ward's 600-cluster clustering is generated

by the standardization procedures of Z12, which denotes $Z_1$ and $Z_2$ having the same best performance in the combination of the MDDR dataset, 600 clusters, and win_Molconn representation. The shaded grids in Tables 5-15 and 5-16 indicate the no standardization procedure, i.e. $Z_0$, has better result than standardization procedures.

With the visual inspection in Tables 5-15 and 5-16, for each table, no single standardization procedure provides consistently the best performance across all 252 possible combinations of clustering method, partition size, evaluation criterion and chemical representation. However, at least, the results suggest standardization procedures have better performance 235 times out of 252 (93%) on the MDDR datasets (Table 5-15) and 229 times out of 252 (91%) on the IDAlert datasets (Table 5-16) than non-standardization procedure, i.e. $Z_0$ (shown in shaded boldface in tables). This suggests that the use of standardization procedures is a critical component to improve the performance on chemical clustering.

Moreover, the visual inspection for the most effective standardization method was carried out, the results in Table 5-15 shows that $Z_7$ is the most consistently effective of the standardization procedures on the evaluation using QCI on Ward's and eWard's clusterings across three chemical representations on the MDDR datasets, and the same trend was also found on the IDAlert datasets (Table 5-16). This would suggest that $Z_7$ is the best choice of standardization method on Ward's and eWard's clusterings evaluated using QCI over all datasets and representations. In addition, focusing on the win_Molconn datasets in each table, $Z_7$ also tends to be more effective over all 84 possible combinations of clustering method, partition size and evaluation criterion, which is the best 54 times out of 84 on the MDDR dataset (Table 5-15(a)) and 72 times out of 84 on the IDAlert dataset (Table 5-16(a)).

MDDR win_Molconn Datasets

| | WD | | EW | | UPGMA | | D_e1 | | D_i2 | | RB_e1 | | RB_i2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # clusters | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q |
| 500 | Z7 | Z7 | Z7 | Z7 | Z1 | Z7 | Z7 | Z7 | Z7 | Z1 | Z2 | Z7 | Z2 | Z1 |
| 600 | Z12 | Z7 | Z7 | Z7 | Z1 | Z7 | Z7 | Z7 | Z7 | Z7 | Z2 | Z7 | Z2 | Z1 |
| 700 | Z12 | Z7 | Z7 | Z7 | Z1 | Z5 | Z7 | Z7 | Z7 | Z7 | Z2 | Z7 | Z2 | Z1 |
| 800 | Z12 | Z7 | Z7 | Z7 | Z7 | Z5 | Z7 | Z7 | Z7 | Z7 | Z2 | Z7 | Z2 | Z1 |
| 900 | Z12 | Z7 | Z7 | Z7 | Z7 | Z5 | Z7 | Z7 | Z2 | Z7 | Z2 | Z7 | Z1 | Z1 |
| 1000 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z2 | Z7 | Z1 | Z1 |

(a)

MDDR Pipeline Pilot Datasets

| | WD | | EW | | UPGMA | | D_e1 | | D_i2 | | RB_e1 | | RB_i2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # clusters | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q |
| 500 | Z12 | Z7 | Z6 | Z7 | Z1 | Z7 | Z7 | Z7 | Z1 | Z7 | Z7 | Z7 | Z17 | Z7 |
| 600 | Z12 | Z7 | Z12 | Z7 | Z1 | Z7 | Z2 | Z5 | Z1 | Z7 | Z7 | Z7 | Z1 | Z7 |
| 700 | Z12 | Z7 | Z6 | Z7 | Z1 | Z7 | Z4 | Z5 | Z1 | Z5 | Z7 | Z7 | Z1 | Z7 |
| 800 | Z12 | Z7 | Z6 | Z7 | Z1 | Z7 | Z2 | Z7 | Z4 | Z5 | Z7 | Z7 | Z7 | Z7 |
| 900 | Z12 | Z7 | Z45 | Z7 | Z1 | Z7 | Z6 | Z5 | Z6 | Z1 | Z7 | Z7 | Z7 | Z7 |
| 1000 | Z12 | Z7 | Z45 | Z7 | Z1 | Z7 | Z7 | Z7 | Z2 | Z1 | Z7 | Z7 | Z1 | Z7 |

(b)

MDDR Holograms Datasets

| | WD | | EW | | UPGMA | | D_e1 | | D_i2 | | RB_e1 | | RB_i2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # clusters | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q |
| 500 | **Z0** | Z7 | Z345 | Z7 | Z1 | Z1 | **Z0** | Z7 | **Z0** | Z1 | Z345 | Z345 | Z7 | Z1 |
| 600 | **Z0** | Z7 | Z1 | Z7 | Z1 | Z1 | Z6 | Z7 | **Z0** | Z6 | Z345 | Z345 | **Z0** | Z2 |
| 700 | **Z0** | Z7 | Z1 | Z7 | Z1 | Z1 | **Z0** | Z7 | **Z0** | Z1 | Z345 | Z2 | **Z0** | Z6 |
| 800 | **Z0** | Z7 | Z345 | Z7 | Z1 | Z1 | Z6 | Z7 | Z6 | Z2 | Z345 | Z2 | **Z0** | Z6 |
| 900 | **Z0** | Z7 | Z345 | Z7 | Z1 | Z1 | Z7 | Z7 | Z6 | Z7 | **Z0** | Z345 | **Z0** | Z6 |
| 1000 | **Z0** | Z7 | Z1 | Z7 | Z1 | Z1 | Z7 | Z7 | Z6 | Z7 | Z7 | Z2 | **Z0** | Z6 |

(c)

Table 5-15 The best standardization procedure(s) of 7 clustering methods over 6 different numbers of clusters using 2 types of evaluation on the MDDR datasets with (a) win_Molconn, (b) Pipeline Pilot, and (c) Holograms representations. (F represents F-Measure, and Q: QCI)

IDAlert win_Molconn Datasets

| | WD | | EW | | UPGMA | | D_e1 | | D_i2 | | RB_e1 | | RB_i2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # clusters | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q |
| 500 | Z7 | Z7 | Z7 | Z7 | Z7 | Z5 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 |
| 600 | Z7 | Z7 | Z7 | Z7 | Z7 | Z5 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 |
| 700 | Z7 | Z7 | Z7 | Z7 | Z7 | Z5 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z2 | Z7 |
| 800 | Z7 | Z7 | Z7 | Z7 | Z7 | Z5 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z2 | Z7 |
| 900 | Z7 | Z7 | Z4 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z7 | Z2 | Z7 |
| 1000 | Z7 | Z7 | Z4 | Z7 | Z7 | Z7 | Z4 | Z7 | Z5 | Z7 | Z7 | Z7 | Z2 | Z7 |

(a)

IDAlert Pipeline Pilot Datasets

| | WD | | EW | | UPGMA | | D_e1 | | D_i2 | | RB_e1 | | RB_i2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # clusters | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q |
| 500 | Z2 | Z7 | Z2 | Z7 | Z7 | Z7 | Z4 | Z7 | Z4 | Z7 | Z5 | Z7 | Z1 | Z7 |
| 600 | Z2 | Z7 | Z1 | Z7 | Z2 | Z7 | Z4 | Z7 | Z1 | Z7 | Z5 | Z7 | Z1 | Z7 |
| 700 | Z4 | Z7 | Z1 | Z7 | Z1 | Z7 | Z4 | Z7 | Z1 | Z7 | Z5 | Z7 | Z5 | Z7 |
| 800 | Z4 | Z7 | Z5 | Z7 | Z2 | Z7 | Z4 | Z7 | Z1 | Z5 | **Z0** | Z7 | Z5 | Z7 |
| 900 | Z6 | Z7 | Z5 | Z7 | Z4 | Z7 | **Z0** | Z7 | Z1 | Z7 | **Z0** | Z7 | Z5 | Z7 |
| 1000 | Z2 | Z7 | Z4 | Z7 | Z1 | Z7 | Z5 | Z7 | Z1 | Z7 | Z4 | Z7 | Z1 | Z7 |

(b)

IDAlert Holograms Datasets

| | WD | | EW | | UPGMA | | D_e1 | | D_i2 | | RB_e1 | | RB_i2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # clusters | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q |
| 500 | **Z0** | Z7 | **Z0** | Z7 | Z1 | Z1 | Z7 | Z2 | **Z0** | Z1 | Z2 | Z2 | **Z0** | Z2 |
| 600 | **Z0** | Z7 | **Z0** | Z7 | Z1 | Z1 | **Z0** | Z2 | Z1 | Z6 | Z2 | Z6 | Z1 | Z2 |
| 700 | **Z0** | Z7 | **Z0** | Z7 | Z1 | Z1 | **Z0** | Z6 | **Z0** | Z2 | Z2 | Z6 | **Z0** | Z2 |
| 800 | **Z0** | Z7 | Z1 | Z7 | Z1 | Z1 | Z7 | Z2 | Z7 | Z345 | Z2 | Z6 | Z7 | Z6 |
| 900 | **Z0** | Z7 | **Z0** | Z7 | **Z0** | Z1 | **Z0** | Z7 | **Z0** | Z7 | Z345 | Z7 | Z7 | Z2 |
| 1000 | **Z0** | Z7 | **Z0** | Z7 | Z7 | Z1 | Z345 | Z2 | Z2 | Z7 | Z345 | Z7 | Z7 | Z2 |

(c)

Table 5-16 The best standardization procedure(s) of 7 clustering methods over 6 different numbers of clusters using 2 types of evaluation on the IDAlert datasets with (a) win_Molconn, (b) Pipeline Pilot, and (c) Holograms representations. (F represents F-Measure, and Q: QCI)

The performance of standardization procedures seems to depend on the clustering method, chemical representation or dataset. For example, using specific standardization methods on the MDDR datasets with win_Molconn and Pipeline Pilot representations always improves performance significantly. However, using no standardization procedure, i.e. $Z_0$, on the datasets with Holograms sometimes has better results than standardization procedures (see shaded grids in Tables 5-15(c) and 5-16(c)). Hence, in order to obtain a more quantitative view of the effectiveness on the standardization methods, we employed Kendall's *W* test of statistical significance to evaluate the consistency of *k* different sets of ranked judgements of the same set of *N* different objects. Here, we have considered each of the representations i.e. Pipeline Pilot, win_Molconn and Holograms, as a judge ranking the different standardization procedures in order of decreasing effectiveness, i.e., $k$=3 and $N$=8.

We ranked the performance obtained from a certain clustering method with a predefined partition size based on a certain evaluation measure. For example, we obtained the F-Measure values based on Ward's clustering with 500 clusters on three different representations of the MDDR datasets (as shown in Table 5-17), then ranked the performance in descending order based on individual representation (Table 5-18). In addition, averaging ranks is used to deal with tied values if any. It simply averages the ranks of all tied observations if they are distinguishable. Finally, the *W* and chi-square ($\chi^2$) values of Kendall's *W* test can be computed based on the equations listed in Chapter 4.

Ward's clustering with 500 clusters on the MDDR datasets

| | $Z_0$ | $Z_1$ | $Z_2$ | $Z_3$ | $Z_4$ | $Z_5$ | $Z_6$ | $Z_7$ |
|---|---|---|---|---|---|---|---|---|
| win_Molconn | 0.0876 | 0.2207 | 0.2207 | 0.0779 | 0.2206 | 0.2206 | 0.1811 | 0.2289 |
| Pipeline Pilot | 0.1098 | 0.1383 | 0.1383 | 0.0893 | 0.1368 | 0.1368 | 0.1304 | 0.1214 |
| Holograms | 0.2950 | 0.2353 | 0.2654 | 0.2634 | 0.2634 | 0.2634 | 0.2926 | 0.2309 |

Table 5-17 Evaluation using F-Measure of Ward's clustering with 500 clusters on the MDDR datasets

Ward's clustering 500 clusters on the MDDR datasets

| | $Z_0$ | $Z_1$ | $Z_2$ | $Z_3$ | $Z_4$ | $Z_5$ | $Z_6$ | $Z_7$ |
|---|---|---|---|---|---|---|---|---|
| win_Molconn | 7 | 2.5 | 2.5 | 8 | 4.5 | 4.5 | 6 | 1 |
| Pipeline Pilot | 7 | 1.5 | 1.5 | 8 | 3.5 | 3.5 | 5 | 6 |
| Holograms | 1 | 7 | 3 | 5 | 5 | 5 | 2 | 8 |

Table 5-18 Ranks obtained by the performance of Ward's clustering with 500 clusters on the MDDR dataset

Tables 5-19 (for MDDR) and 5-20 (for IDAlert) present the results of a Kendall's *W* analysis, showing the chi-square ($\chi^2$) values based on the combination of partition size, clustering method and performance criterion. The critical value of the chi-square ($\chi^2$) distribution at $\alpha$=0.05 level with 7 degrees of freedom is 14.07 and at $\alpha$=0.01 level is 18.48. The shaded grids in these two Tables indicate statistical significance was found at $\alpha$=0.05 level. The inspection based on individual clustering methods shows that using UPGMA method on the MDDR datasets has 12 combinations out of 14 (86%) found significant at $\alpha$=0.05 level. The results here would suggest there is obvious ranking of the eight standardization procedures. However, in order to obtain an overall and confident view of statistical significance, inspection of more combinations of clustering method and partition size is needed. We hence carried out an overall inspection based on datasets.

Inspection of Table 5-19 (for MDDR) shows that only 20 combinations out 84 (24%) were found significant at $\alpha$=0.05 level, and no combination reached the significant level of $\alpha$=0.01. A similar trend was also found in Table 5-20 (for IDAlert), only 14 combinations out of 84 (17%) were found significant at $\alpha$=0.05 level, and only one combination, D_e1 500 clusters evaluating by QCI, reached the significant level of $\alpha$=0.01. Taking the inspections from the MDDR and IDAlert together, there is no correlation between the rankings of the standardization methods by the three chemical representations. We can

conclude that, for the chemical datasets considered here, there is no obvious performance benefit that is likely to be obtained from the use of any particular standardization procedure. The choice of standardization method is hence not a critical component of a procedure for chemical clustering.

| # clusters | 500 | | 600 | | 700 | | 800 | | 900 | | 1000 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Evaluations | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q |
| Ward's | 6.31 | 12.28 | 7.69 | 11.93 | 7.69 | 12.74 | 5.05 | 10.56 | 8.03 | 11.93 | 4.02 | 14.80 |
| e-Ward's | 6.89 | 12.28 | 14.11 | 12.99 | 9.07 | 12.74 | 7.11 | 13.89 | 8.69 | 12.05 | 8.72 | 11.48 |
| UPGMA | 15.35 | 15.02 | 15.24 | 14.34 | 14.68 | 14.11 | 14.23 | 14.11 | 13.66 | 14.11 | 12.87 | 15.81 |
| Direct_e1 | 8.93 | 17.73 | 13.77 | 15.92 | 12.19 | 15.69 | 12.76 | 16.37 | 13.55 | 16.03 | 16.94 | 16.03 |
| Direct_i2 | 7.00 | 15.69 | 10.05 | 12.98 | 8.47 | 15.24 | 7.79 | 11.40 | 9.48 | 12.19 | 11.52 | 17.27 |
| RB_e1 | 11.97 | 7.00 | 8.58 | 7.11 | 9.60 | 8.47 | 7.34 | 10.06 | 9.82 | 8.92 | 10.84 | 10.50 |
| RB_i2 | 8.70 | 10.50 | 4.85 | 9.49 | 8.58 | 10.05 | 5.19 | 9.15 | 3.61 | 9.15 | 2.71 | 9.15 |

Table 5-19 The chi-square ($\chi^2$) values of the Kendall's test based on the ranking by F-Measure and QCI evaluations of clusterings over varied numbers of clusters on the MDDR datasets (F represents F-Measure, and Q: QCI)

| # clusters | 500 | | 600 | | 700 | | 800 | | 900 | | 1000 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Evaluations | F | Q | F | Q | F | Q | F | Q | F | Q | F | Q |
| Ward's | 8.76 | 9.33 | 7.29 | 9.06 | 6.66 | 10.76 | 8.12 | 10.93 | 2.30 | 10.86 | 4.03 | 12.18 |
| e-Ward's | 5.20 | 11.89 | 5.12 | 8.99 | 3.81 | 9.22 | 6.09 | 8.99 | 9.30 | 9.90 | 5.51 | 10.81 |
| UPGMA | 14.79 | 12.65 | 13.21 | 11.97 | 12.53 | 14.45 | 13.35 | 14.11 | 11.18 | 14.06 | 11.97 | 14.56 |
| Direct_e1 | 13.83 | 18.74 | 10.84 | 15.24 | 12.87 | 12.31 | 13.21 | 14.56 | 13.10 | 17.16 | 16.15 | 14.45 |
| Direct_i2 | 10.05 | 12.98 | 8.13 | 9.78 | 7.79 | 7.06 | 8.24 | 12.87 | 7.11 | 13.40 | 7.74 | 16.82 |
| RB_e1 | 10.39 | 9.37 | 7.90 | 8.35 | 11.63 | 11.74 | 10.27 | 11.74 | 10.27 | 13.10 | 9.60 | 13.77 |
| RB_i2 | 8.58 | 6.40 | 10.05 | 10.18 | 3.05 | 12.83 | 5.19 | 15.92 | 4.29 | 14.06 | 2.94 | 13.69 |

Table 5-20 The chi-square ($\chi^2$) values of the Kendall's test based on the ranking by F-Measure and QCI evaluations of clusterings over varied numbers of clusters on the IDAlert datasets (F represents F-Measure, and Q: QCI)

## 5.11 The Comparison between Clustering Methods

We carried out the comparison between clustering methods based on the datasets with no standardization procedure, i.e. $Z_0$ (Figures 5-11(a)) and the single best standardization procedure (as listed in Tables 5-15 and 5-16) on the individual clustering method (Figures 5-11(b)). First, we would like to know if any benefit can be obtained from the use of any particular standardization method by simply comparing Figures 5-11(a) and (b). Secondly, we would also like to know the performance of each clustering method with its single best standardization procedure in order to find the optimal, if any, combinations of clustering method and standardization procedure.

First consider the MDDR dataset with no standardization procedure, i.e. $Z_0$, no clustering method offers consistently the best performance across all numbers of clusters with three different representations (see Figure 5-11(a)). Hence, we inspect each criterion performance with different representation individually. As the evaluation using F-Measure shown in Figure 5-11(a), with the Holograms representation, Ward's method tends to have better values of F-Measure. While with win_Molconn, the agglomerative UPGMA method has consistently best performance. No clustering method was found offering consistently better results with Pipeline Pilot. Again, inspection on the single best standardization procedures (Figure 5-11(b)), the UPGMA method with Pipeline Pilot representation is consistently superior to all of the other approaches, and this clustering method also tends to have better F-Measure with win_Molconn, whilst no single clustering method can yield consistently better performance with the Holograms representation.

We compared Figures 5-11(a) with 5-11(b) to find the difference of the datasets with or without standardization procedure. Overall, the performance of the datasets with win_Molconn and Pipeline Pilot representations has improved significantly by using the single best standardization procedures. However, the performance of Holograms datasets has limited improvement on only some clustering methods, e.g. UPGMA.

A similar inspection on the MDDR dataset but with the evaluation using QCI, the Direct methods with criterion function of *e1* or *i2* have consistently the best performance over three representations with no standardization procedure (Figure 5-12(a)), and the results of Direct_e1 and Direct_i2 are close to each other. In addition, contrary to the criterion performance of F-Measure, the UPGMA method yields consistently the worst QCI over all representations. While with the single best standardization procedures (Figure 5-12(b)), Direct methods have consistently better results with the Holograms representation, and the Direct_e1 and Direct_i2 methods have similar values of QCI. Similar behaviour was found on the Pipeline Pilot representation, Ward's and e-Ward's methods have close and consistently better QCI values. With the win_Molconn representation, Ward's method yields consistently the best performance. In addition, the UPGMA method is still consistently inferior to all of other clustering methods over all representations on the MDDR datasets with the standardization procedures.

Again, we compared Figures 5-12(a) with 5-12(b) to find the difference of the datasets with or without standardization procedure. Significant improvement was found on the datasets with win_Molconn and Pipeline Pilot especially on Ward's, extended Ward's and Direct method. Moreover, limited improvement was obtained on the datasets with Holograms on only some clustering methods, e.g. UPGMA method.

Overall, the performance of the datasets with win_Molconn and Pipeline Pilot representations has improved significantly by using the single best standardization procedures. However, the performance of Holograms datasets has limited improvement on only some clustering methods, e.g. UPGMA. The overall comparison between clustering methods based on two criteria performance and three representations on the MDDR datasets is summarized in Table 5-21.

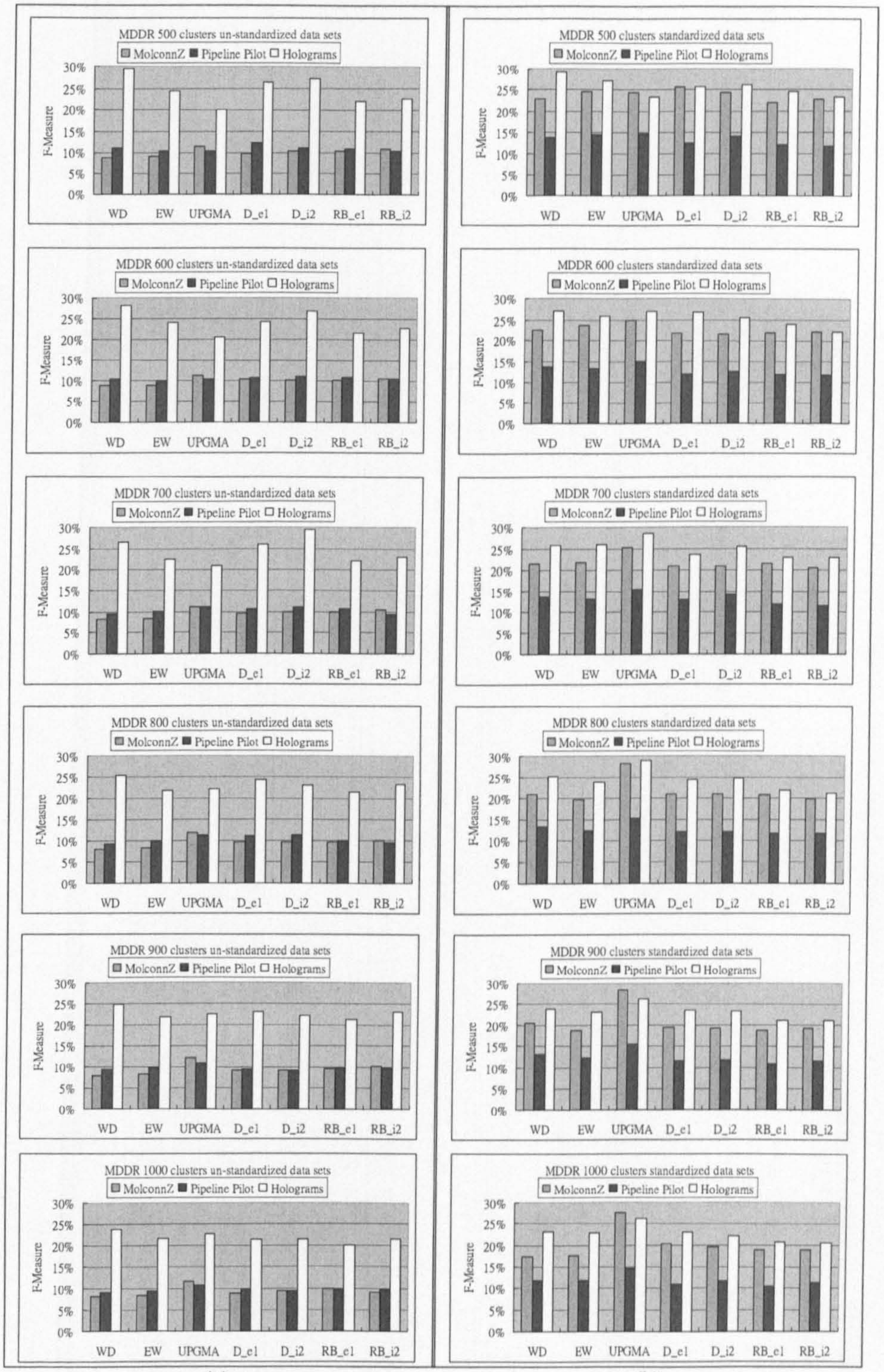



Figure 5-11 The evaluation using F-Measure of 7 clustering methods over 6 different numbers of clusters of (a) no standardization and (b) the single best standardization procedures standardization procedures on 3 chemical representations of the MDDR datasets

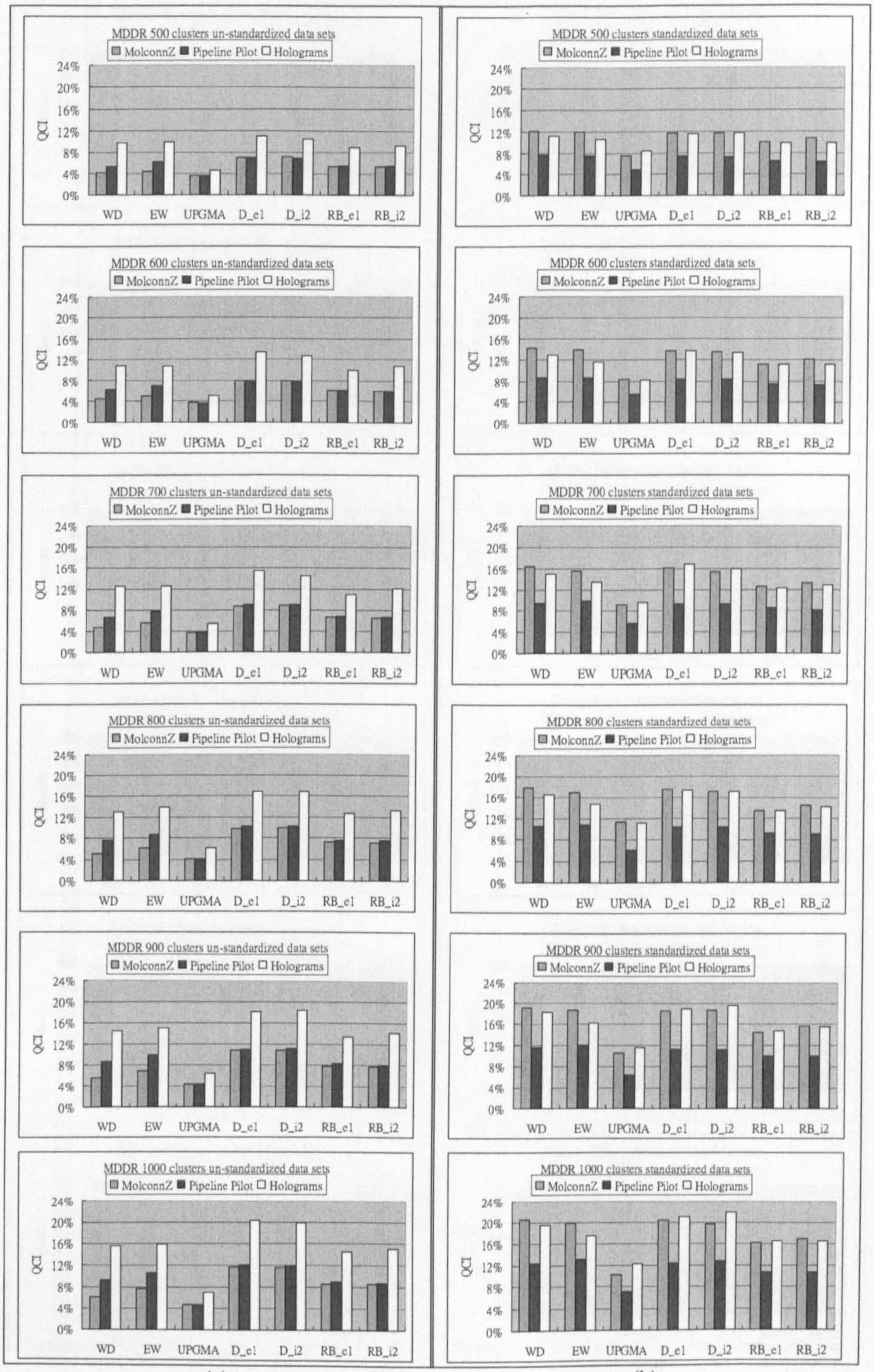



Figure 5-12 The evaluation using QCI of 7 clustering methods over 6 different numbers of clusters of (a) no standardization and (b) the single best standardization procedures on 3 chemical representations of the MDDR datasets

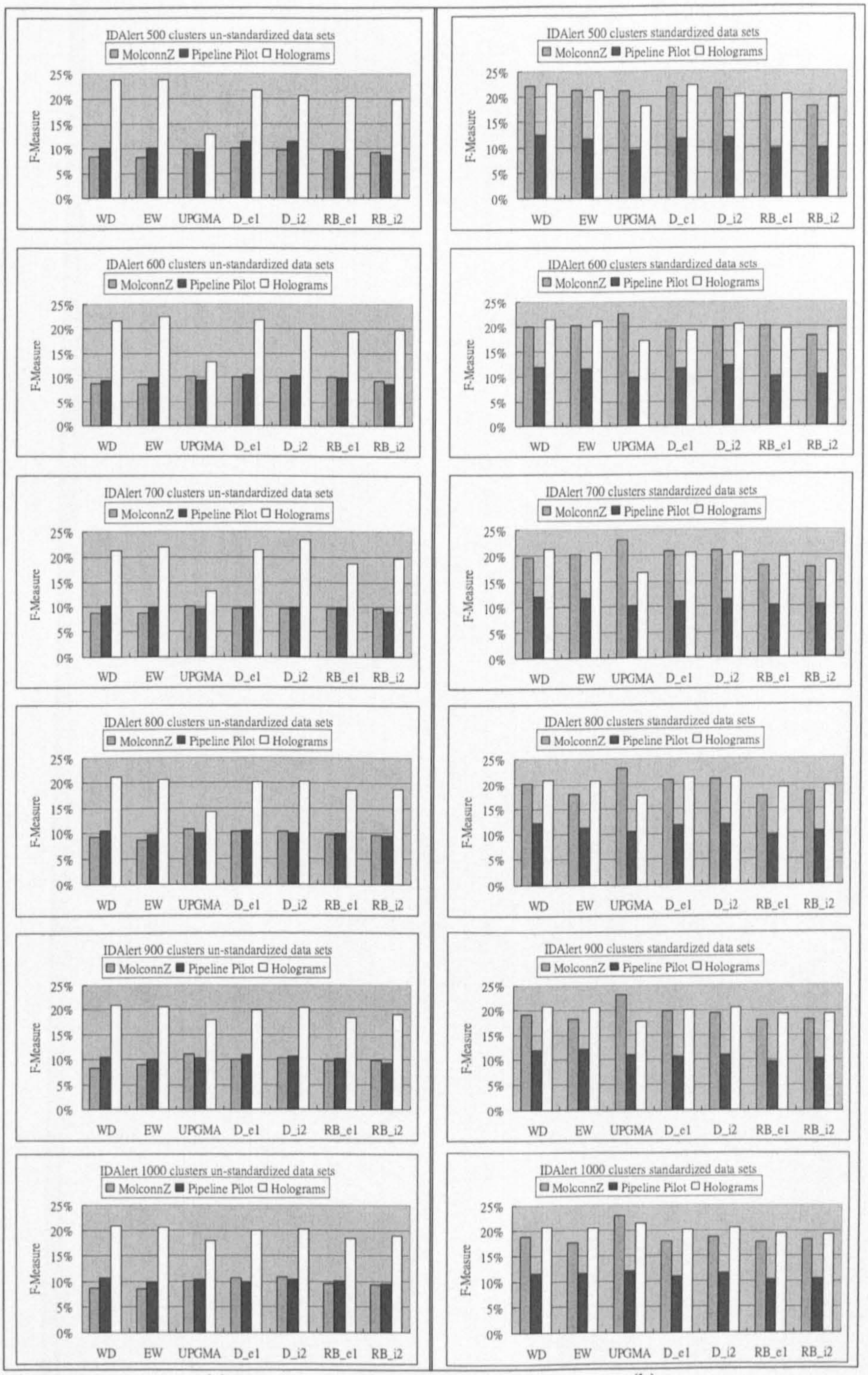



Figure 5-13 The evaluation using F-Measure of 7 clustering methods over 6 different numbers of clusters of(a) no standardization and (b) the single best standardization procedures on 3 chemical representations of the IDAlert datasets

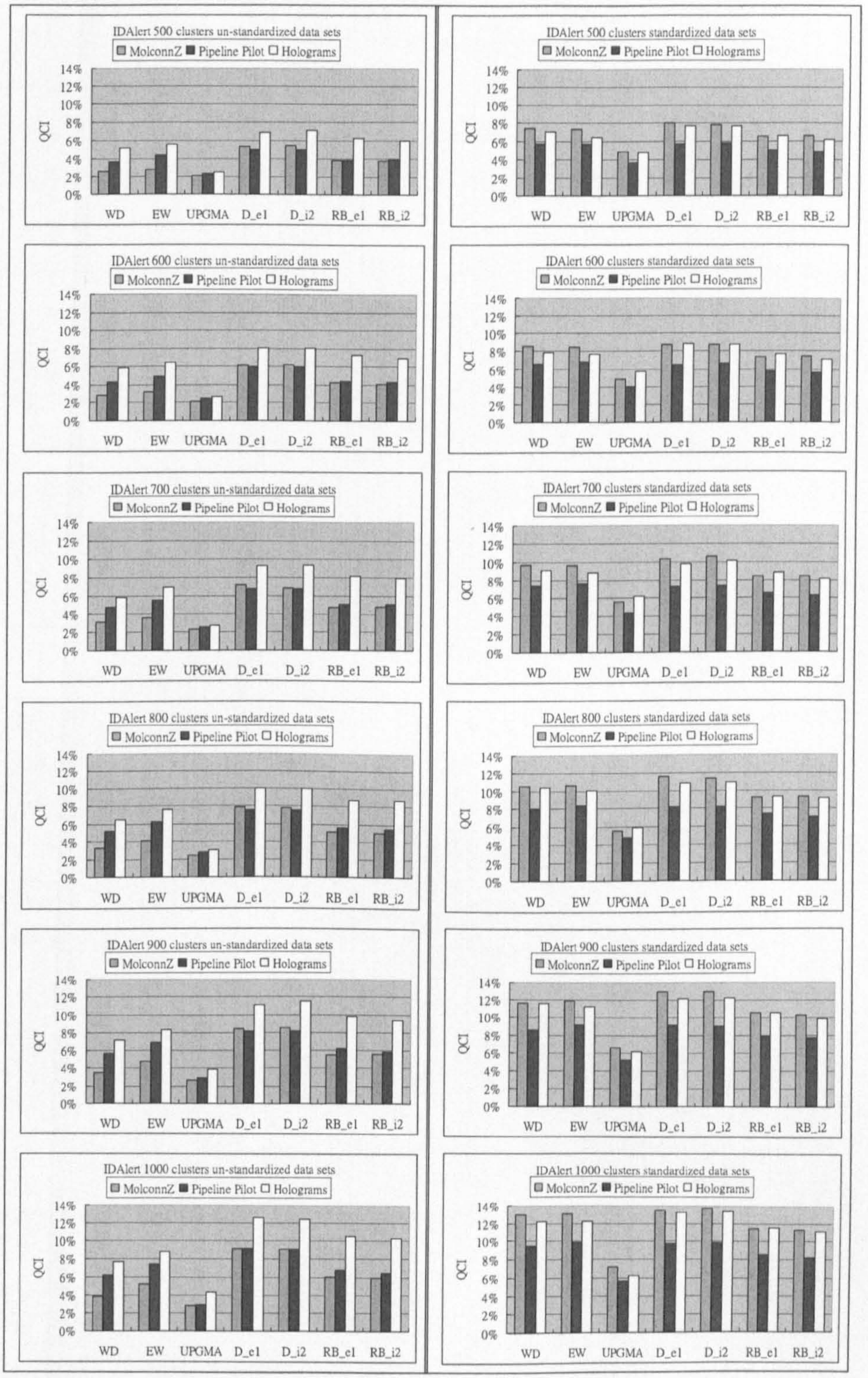



Figure 5-14 The evaluation using QCI of 7 clustering methods over 6 different numbers of clusters of (a) no standardization and (b) the single best standardization procedures on 3 chemical representations of the IDAlert datasets

With the IDAlert datasets, no overall best clustering method was found on the performance of F-Measure with no standardization or standardization procedures across all chemical representations (Figure 5-13). As for the performance of QCI (Figure 5-14), the Direct methods tend to have consistently better results. The results of Direct_e1 and Direct_i2 methods are very close, while the UPGMA method has the worst overall performance, as on the MDDR datasets. The overall comparison is summarized in Table 5-22.

As for the improvement on the use of standardization procedure, the comparison between Figure 5-13(a) and Figure 5-13(b), and Figure 5-14(a) and Figure 5-14(b) also shows the similar trend, in which noticeable improvement was found the datasets with win_Molconn and Pipeline Pilot, limited benefit was found with Holograms, when employing the single best standardization procedures.

With the summary in Tables 5-21 and 5-22, for those clustering methods having consistently better performance, we can find its corresponding single best standardization procedure in Tables 5-15 and 5-16. For example, the UPGMA method has the best overall performance on the MDDR datasets with Pipeline Pilot (see Table 5-21), hence we can find its corresponding single best standardization procedure is $Z_1$ for the evaluation using F-Measure, and is $Z_7$ for the evaluation using QCI.

| | MDDR datasets | | | |
|---|---|---|---|---|
| | F-Measure | | QCI | |
| | No standardization procedure ($Z_0$) | Standardization procedures ($Z_1$- $Z_7$) | No standardization procedure ($Z_0$) | Standardization procedures ($Z_1$- $Z_7$) |
| win_Molconn | UPGMA the best overall | UPGMA tends to have better results | Direct methods tend to be consistently better<br><br>UPGMA the worst overall | Ward's method the best overall |
| Pipeline Pilot | No overall best method | UPGMA the best overall | | Ward's and e-Ward's methods tend to have consistently better |
| Holograms | Ward's tends to have better results | No overall best method | | Direct methods tend to have consistently better |

Table 5-21 Summary of effectiveness of clustering methods on the MDDR datasets

| | IDAlert datasets | | | |
|---|---|---|---|---|
| | F-Measure | | QCI | |
| | No standardization procedure ($Z_0$) | Standardization procedures ($Z_1$- $Z_7$) | No standardization procedure ($Z_0$) | Standardization procedures ($Z_1$- $Z_7$) |
| win_Molconn | No overall best method | | Direct methods tend to be consistently better<br>UPGMA the worst overall | |
| Pipeline Pilot | | | | |
| Holograms | | | | |

Table 5-22 Summary of effectiveness of clustering methods on the IDAlert datasets

## 5.12 The Comparison between Chemical Representations

We first consider the comparison of the effect of three different types of chemical representations on the MDDR datasets. The Figures 5-11 and 5-12 show the F-Measure and QCI values respectively, which are obtained by (a) no standardization procedure, i.e. $Z_0$, and (b) the single best standardization procedure from the MDDR datasets with three different representations over six numbers of clusters. Obviously, the unstandardized dataset with the Holograms representation consistently offers the best performance of F-Measure and QCI values when compared to the other two representations, while the performance of win_Molconn and Pipeline Pilot representations is worse and similar to each other. Similarly, the performance of Holograms representation with the unstandardized IDAlert dataset (Figures 5-13 and 5-14) is also consistently yielding the best results.

In terms of the performance of the single best standardization procedures, the improvement of the Holograms representation was limited, especially in comparison with win_Molconn. For example, the F-Measure results of Holograms obtained by Direct_i2 clustering method with the standardized MDDR datasets are even worse on 500, 600, and 700 clusters (Figure 5-11(b)). By comparison, the performance of the win_Molconn representation with the single best standardization procedure improved significantly, and is similar to Holograms. Also, the performance of the Pipeline Pilot representation

improved but not as much as win_Molconn. That is, the performance of win_Molconn and Pipeline Pilot can be improved by choosing a proper standardization procedure. Similar trends are also found with the standardized IDAlert datasets, as shown in Figures 5-13(b) and 5-14(b). However there are no standardized datasets with a certain chemical representation that could offer consistently better performance.

The overall performance of three chemical representations with the unstandardized and standardized datasets discussed above, suggests (summarized in Table 5-23) that the un-standardized dataset with Holograms is the most effective chemical representation that we have tested here. As for the standardized dataset, although no consistent benefit can be obtained from choosing a certain chemical representation, for the datasets using either win_Molconn or Pipeline Pilot are suggested to employ a proper standardization procedure, if any, to improve its performance according to our finding in this study.

| MDDR and IDAlert datasets | | | |
|---|---|---|---|
| | Performance of using non-standardization procedure | Performance of using standardization procedure | Improvement made by standardization procedure |
| Pipeline Pilot | Worse performance | Has the worst performance | Improved |
| win_Molconn | Worse performance | Have significantly better performance than Pipeline Pilot, and similar results | Significantly improved |
| Holograms | Has consistently best results | | Limited improvement on only with certain clustering methods |

Table 5-23 Summary of effectiveness of three chemical representations

## 5.13 Conclusions

Account for the magnitude of the variability obtained in different BCI K-Means runs, in the first experiment of Chapter 5, the implementation of the BCI K-Means clustering was carried out in a default mode, using different random seeds in different runs. Hence, in order to investigate the variability caused by the different random seeds in different runs of BCI K-Means clustering, one dataset is picked, e.g. the MDDR with pipeline pilot representation and S1 standardization procedure in the following case, and run 20 times of K-Means clustering on the same dataset. The clustering results were evaluated by probability of correction prediction and Shannon Entropy, and are listed as follows:

| Run | Probability of correction prediction | Shannon Entropy |
|---|---|---|
| 1 | 0.6378 | 3.5089 |
| 2 | 0.6228 | 3.5335 |
| 3 | 0.6163 | 3.4995 |
| 4 | 0.5804 | 3.4803 |
| 5 | 0.6481 | 3.4644 |
| 6 | 0.6213 | 3.4966 |
| 7 | 0.6193 | 3.4507 |
| 8 | 0.6313 | 3.4601 |
| 9 | 0.6289 | 3.5182 |
| 10 | 0.6379 | 3.4862 |
| 11 | 0.5948 | 3.5088 |
| 12 | 0.6620 | 3.4319 |
| 13 | 0.6175 | 3.4749 |
| 14 | 0.6507 | 3.5164 |
| 15 | 0.6061 | 3.4958 |
| 16 | 0.6341 | 3.5054 |
| 17 | 0.6099 | 3.4439 |
| 18 | 0.6168 | 3.4962 |
| 19 | 0.6395 | 3.4610 |
| 20 | 0.6247 | 3.4482 |
| Average | 0.6249 | 3.4805 |
| Standard Deviation | 0.0191 | 0.0262 |

Table 5-24 The evaluation of 20 runs of K-Means clustering using probability of correct prediction and Shannon Entropy on the $S_1$ Pipeline Pilot MDDR dataset.

The averages of 20 runs of evaluation using probability and Shannon Entropy are 0.6249 and 3.4805 respectively, which are close to the results, 0.6 and 3.4665 respectively, listed in Figures 5-1 (probability) and 5-2 (Entropy). In addition, the standard deviations for the two evaluation measures are 0.0191 and 0.0262, which means the magnitude of the variability caused by the different random seeds in different runs of BCI K-Means clustering is not significantly large.

In addition, the variation of K-Means method, CLUTO *Direct* method, was employed in the extended work of Chapter 5; the *Direct* method was implemented in a default mode which the clustering result is the one has the best performance over 10 runs. However, the focus of chapter 5 is mainly on the effectiveness of different standardization procedures rather than the effectiveness of different clustering methods. Moreover, the effect of standardization was also carried out by means of similarity searching. All these findings lead to the same conclusion that no standardization procedure was found offering consistently best performance over the two datasets.

Combining the analysis and discussion from two experiments in this chapter, no standardization procedure was found offering consistently the best performance over two datasets, i.e. no overall best method. However, the use of standardization methods tends to provide significant improvement especially in the Molconn and Pipeline Pilot datasets, and limited improvement in the Hologram datasets.

The evaluation of clustering methods in this chapter shows that no single clustering approach has the best overall performance over the combination of representations and datasets. The result also shows that the clustering performance depends on the many factors, such as the use of standardization procedures, the feature of evaluation criterion, and the data type of dataset. For example, in the extensive experiment, $Z_7$ standardization procedure has overall the best performance on the Ward's and e-Ward's clusterings; $Z_1$ procedure with UPGMA clustering yields the overall best F-Measure only on the MDDR datasets; while the non-standardization procedure ($Z_0$) with Ward's clustering

offers the best F-Measure on the Holograms datasets only, and with UPGMA clustering provides the worst QCI on all representations.

Hence, the use of standardization procedures does not bring any consistent benefit in terms of clustering behaviour. In the next chapter, we investigate the applicability of nine clustering methods on the same datasets but characterized by binary fingerprints.

# Chapter 6 : Comparison of Chemical Clustering Methods Using Fingerprint-based Similarity Measures

## 6.1 Introduction

Cluster analysis is a process to identify groups of similar objects; the objects in the same cluster are similar, while the objects in different groups are dissimilar. Some introduction content and methodologies were well reviewed by Milligan and Cooper (1987), Jain et al. (1999), Berkhin (2002), and Xu and Wunsch (2005). However, it has also been extensively discussed in many disciplines, such as document clustering (El-Hamdouchi and Willett, 1986; Willett, 1988; Zhao and Karypis, 2002). In addition, there is also considerable interest in chemoinformatics including high-throughput screening, combinatorial chemistry, compound acquisition, and QSAR. Moreover the applications of chemical clustering are well reviewed previously by Willett (1985, 1987) and by Downs and Barnard (2002).

The early works of chemical clustering were largely by Willett and co-workers (Downs and Willett, 1994), their studies showed the Jarvis-Patrick clustering method offered better performances for different types of chemical datasets (Willett et al., 1986; Willett, 1987). Moreover, the later work by Brown and Martin (1996) reported Ward's method had better performance than other hierarchical methods for the biologically active and inactive molecules separation. In a more recent study, Holliday et al. (2004) verified the ability of fuzzy K-means method on small datasets by highlighting the multicluster membership and finding outlier objects in comparison with Ward's and original K-means methods.

A wide variety of clustering methods have been proposed in the literature, they are classified into hierarchical, partitional, and density-based clustering methods. Choosing the right clustering method is always a critical issue of clustering analysis. Some comparative studies on chemical datasets can be found in the literature. Rubin and Willett (1983) compared four hierarchical divisive methods on eleven small datasets represented by substructural fragments. Their results showed that no single method offered consistently better performance, and most of the clustering methods are not suitable for dealing with thousands of objects at that time. Downs et al. (1994) applied agglomerative hierarchical, divisive hierarchical and non-hierarchical clustering methods on physicochemical properties of large datasets, and found that hierarchical methods had better performances than the Jarvis-Patrick non-hierarchical method. Raymond et al. (2003) compared five published clustering methods using graph- and fingerprint-based similarity measures, and their study reported that two methods, CAST and Yin-Chen methods, which have been applied previously in clustering on gene expression patterns were found effective for the clustering of 2D chemical structures.

No single method is applicable for all types of data, and not all methods are equally applicable to all problems. Different clustering methods will generate different types of clustering results; that is, clustering different types of data with one single clustering method will have varied performances. Most clustering methods are dependent on the features of the dataset, e.g. data types, or are sensitive to parameter setting; that is, some algorithms will be more suitable for certain types of data than others (Gionis et al., 2007).

The Jarvis-Patrick and Ward's methods are the clustering procedures of choice in most chemoinformatics applications and software packages. However, cluster analysis is a strong focus in data mining research and this has resulted in the recent development of many new clustering methods that can be applied to large databases. In this chapter, we consider the utility of some of these new methods for the use in chemoinformatics. The main focus of this study is to investigate the suitability of different clustering which were reported effective in other applications to the chemical clustering on 2D structures, and also to

compare their performance with some commonly used methods in chemoinformatics.

## 6.2 Clustering Methods

Nine clustering methods were evaluated in this study. The first two clustering methods, Yin-Chen and CAST, were coded by *Perl* script in this study; the next two methods, Ward's and extended Ward's, used the implementations in the *Energy* package of the *R* statistical system (available at http://www.r-project.org/); while the other three clustering procedures, *agglomerative hierarchical*, *Direct* and *Repeated Bisection,* were carried out using the implementations in the CLUTO (for CLUstering TOolkit) software package (available at http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview). In addition, we used the default criterion function, UPGMA (Unweighted Pair Group Method using Arithmetic mean), for the agglomerative hierarchical method. However, the *Direct* and *Repeated Bisection* methods employed two criterion functions, *e1* and *i2*, hence with the combinations of clustering method and criterion function, we obtain four different methods.

In addition to the Ward's and UPGMA methods are commonly seen in chemical clustering, the other seven methods are new or little discussed to the application of clustering for 2D structure. However, all these clustering methods and their criterion functions were discussed in detail in Chapter 4. The software tools and denotations of these nine clustering methods are summarized in Table 6-1.

| Clustering Methods | Software Tool | Code in Tables and Figures |
|---|---|---|
| Yin-Chen | Coded using Perl Script | YC |
| CAST | Coded using Perl Script | CAST |
| Extended Ward's | R software | EW |
| Ward's | R software | WD |
| *agglo* method with *UPGMA* criterion function | CLUTO package | UPGMA |
| *Direct* method with *e1* criterion function | CLUTO package | DR-e1 |
| *Direct* method with *i2* criterion function | CLUTO package | DR-i2 |
| *repeated bisection* method with *e1* criterion function | CLUTO package | RB-e1 |
| *repeated bisection* method with *i2* criterion function | CLUTO package | RB-i2 |

Table 6-1 Summary of the software tools and denotations of the nine clustering methods

## 6.3 Experimental Details

The MDDR and IDAlert datasets were characterized by ECFP_4 fingerprints using SciTegic Pipeline Pilot software (both datasets and representation are discussed in detail in Chapter 4) , and were coded as being active or inactive in eleven activity classes that had been studied previously by Hert et al. (2004).

These two datasets were then clustered by above nine clustering procedures to generate partitions containing 500, 600, 700, 800, 900 and 1000 clusters. However, the number of clusters in some clustering methods, such as Yin-Chen and CAST, is determined by an adjustable parameter or a cut-off threshold and is sensitive to the setting of the threshold, we attempted to generate the number of clusters which are as close to above numbers as possible (as shown in Table 6-2).

| | Yin-Chen | | CAST | |
|---|---|---|---|---|
| # clusters | MDDR | IDAlert | MDDR | IDAlert |
| 500 | 513 | 499 | 505 | 501 |
| 600 | 599 | 600 | 594 | 598 |
| 700 | 704 | 698 | 695 | 701 |
| 800 | 801 | 798 | 799 | 803 |
| 900 | 903 | 904 | 892 | 899 |
| 1000 | 1002 | 999 | 994 | 1001 |

Table 6-2 The numbers of clusters determined by the adjustable parameter for the Yin-Chen and CAST clustering methods

## 6.4 Evaluation of Clustering Performance

Evaluation is one of the critical components in cluster analysis. Most clustering applications need the evaluation measures to assess the clustering results from a certain criterion such as, capturing the intra-cluster similarity and inter-cluster dissimilarity. There are extensive evaluation measures of different types in the literature, if a clustering method offers better performance than others over many evaluation measures; we can claim that clustering method should be the best for a certain type of application. Hence we employed four criteria in this study and each of them was discussed in detail in Chapter 4. The Shannon entropy is a criterion to observe the distribution of actives of a given class, while the Entropy based on cluster size is a measure similar to the conventional Shannon entropy to investigate the distribution of cluster sizes.

The F-Measure is a measure widely used in document clustering in information retrieval (Fung et al., 2003; Rosenberg and Hirschberg, 2007), and the Quality Clustering Index (QCI) is a new evaluation measure defined by Varin et al. (2008). Both the F-Measure and QCI are the measures based on the extent of how the compounds with the same bioactivities can be grouped together, especially the eleven active classes mentioned in Chapter 4. However, the evaluation using the probability of correct prediction is not applicable to the clustering validation here due to the smaller partition size as we discussed in Section 4.4.2.

## 6.5 Results & Analysis

The evaluation and analysis of clustering results are carried out in three aspects. The first is the evaluation of clustering methods over all criteria; the Sections 6.5.1 and 6.5.2 are for the evaluation of the MDDR and IDAlert datasets respectively. The next aspect of evaluation of clustering methods is based on individual criterion (Section 6.5.3). The final aspect (Section 6.5.4) focuses on the comparison of some particular clustering methods, such as the Ward's and extended Ward's methods, and the Direct and Repeated Bisection methods in the CLUTO tool kit.

### 6.5.1 The Evaluation of Clustering Results of the MDDR Dataset

Table 6-3 displays the clustering performance of 1000 clusters on the MDDR dataset over all evaluation criteria. The values in each row represent the clustering result evaluated by the four criteria for a specific clustering method. For example, the Yin-Chen clustering result evaluated by the two types of Shannon Entropy, F-Measure and QCI are 5.83, 9.89, 7.83%, and 11.07% respectively.

| | MDDR dataset | | 1000 clusters | |
|---|---|---|---|---|
| | Entropy | Entropy based on size | F-Measure | QCI |
| YC[a] | 5.83 | 9.89 | 7.83% | 11.07% |
| CAST[a] | 3.87 | 8.79 | 22.26% | 10.08% |
| EW | 4.26 | 9.68 | 21.86% | 23.40% |
| WD | 4.13 | 9.74 | 24.23% | 25.59% |
| UPGMA | 2.93 | 8.84 | 29.62% | 18.29% |
| DR-e1 | 4.18 | 9.84 | 24.21% | 28.81% |
| DR-i2 | 4.16 | 9.81 | 23.83% | 28.83% |
| RB-e1 | 4.57 | 9.81 | 21.72% | 20.40% |
| RB-i2 | 4.68 | 9.78 | 20.61% | 18.43% |

[a] The numbers of clusters for the Yin-Chen and CAST methods generated by their adjustable parameters are 1002 and 994 respectively.

Table 6-3 The evaluation of different clustering methods (1000 clusters) for the MDDR dataset based on the four different evaluation criteria

The visual inspection of results in Table 6-3 suggests that UPGMA method has significantly the best performance on the evaluation criteria of Shannon Entropy and F-Measure, and better result on the Shannon Entropy based on cluster size, but ordinary performance on the QCI. In addition, the Direct method has the noticeably the best performance on the QCI criterion. Contrary, the Yin-Chen method is consistently inferior. As for the comparison between the Ward's and extended Ward's, the Ward's method has consistently better performance than the extended Ward's. In terms of the two partitional clustering methods in the CLUTO toolkit, the Direct method offers consistently better performances than the Repeated Bisection on the evaluation criteria of Shannon Entropy, F-Measure and QCI. However, they have similar results on the Entropy based on cluster size. As for the effect of two different criterion functions, the performance of the use of *e1* and *i2* on either the Direct or Repeated Bisection methods is similar. The use of *e1* and *i2* reveal a high degree of variability offering inconsistently superior or inferior performance to each other, i.e. no significant difference between the use of *e1* and *i2*.

The evaluations for the rest of numbers of clusters on the MDDR dataset are similar to the case of 1000 clusters. With visual inspection, no clustering method is found offering consistently the best performance over all evaluation criteria.

In order to obtain a more quantitative view of the effectiveness of the clustering methods, we employed Kendall's $W$ test of statistical significance to evaluate the consistency of $k$ different sets of judgements of the same set of $N$ different objects. Here, we have considered each of the four evaluation criteria as a judge ranking the nine different clustering methods in order of decreasing effectiveness, i.e. $k$=4 and $N$=9, as shown in Table 6-4 (MDDR 1000 clusters). The equation of Kendall's $W$ test has been given in Chapter 4, together with the use of $\chi^2$ (chi-square) test to assess the significance of the calculated $W$ value. Since the sample size $N$ is greater than 7, it will be considered a large sample. The Kendall's $W$ values for large sample can be computed using these four sets of rankings, and the significance test can also be carried out using $\chi^2$

distribution with *N-1* degrees of freedom.

| | MDDR dataset | | 1000 clusters | |
|---|---|---|---|---|
| | Entropy | Entropy based on size | F-Measure | QCI |
| YC | 9 | 9 | 9 | 8 |
| CAST | 2 | 1 | 5 | 9 |
| EW | 6 | 3 | 6 | 4 |
| WD | 3 | 4 | 2 | 3 |
| UPGMA | 1 | 2 | 1 | 7 |
| DR-e1 | 5 | 8 | 3 | 2 |
| DR-i2 | 4 | 6.5 | 4 | 1 |
| RB-e1 | 7 | 6.5 | 7 | 5 |
| RB-i2 | 8 | 5 | 8 | 6 |

Table 6-4 The performance of clustering methods ranked by the four criteria functions for the MDDR dataset (1000 clusters)

The critical values of the chi-square ($\chi^2$) distribution at the $\alpha$=0.01 level of statistical significance is 20.09 for eight degrees of freedom, and at the $\alpha$=0.05 level of statistical significance is 15.51.

Table 6-5(a) (for the MDDR dataset) reveals the results of a Kendall's *W* analysis, showing the *W* and $\chi^2$ values for the combination of number of clusters and clustering performance generated by four evaluation criteria. For example, the computed values of *W* and $\chi^2$ for the MDDR dataset with 1000 clusters are 0.50 and 16.15 respectively (as shown in Table 6-5(a)). Five in six tests have reached the significant level of $\alpha$=0.05, while none of these values have reached the significant level of $\alpha$=0.01. Therefore, there would hence appear to be no strongly significant measure of agreement between the clustering methods and the evaluation criteria. Hence, it is not possible to recommend any particular clustering method as being of general applicability.

| # Clusters | (a) MDDR dataset | | (b) IDAlert dataset | |
|---|---|---|---|---|
| | $W$ | $\chi^2$ | $W$ | $\chi^2$ |
| 500 | 0.53 | 16.93 | 0.37 | 11.91 |
| 600 | 0.49 | 15.80 | 0.33 | 10.67 |
| 700 | 0.48 | 15.48 | 0.35 | 11.11 |
| 800 | 0.53 | 17.04 | 0.45 | 14.53 |
| 900 | 0.56 | 17.82 | 0.41 | 13.24 |
| 1000 | 0.50 | 16.15 | 0.26 | 8.22 |

Table 6-5 Kendall $W$ and $\chi^2$ values based on the four different evaluation measures for the (a) MDDR and (b) IDAlert datasets

## 6.5.2 The Evaluation of Clustering Results of the IDAlert Dataset

The overall evaluations of clustering performance for the IDAlert dataset are shown in Table 6-6 (1000 clusters). In this case, for a certain clustering method, the values are generated by the same evaluation criteria mentioned previously. For example, the computed values of two types of Shannon Entropy, F-Measure, and QCI of 1000-cluster CAST clustering for IDAlert 10K dataset are 4.25, 8.90, 17.74%, and 5.62% respectively (as shown in Table 6-6).

| | IDAlert dataset | | 1000 clusters | |
|---|---|---|---|---|
| | Entropy | Entropy based on size | F-Measure | QCI |
| YC | 4.84 | 9.64 | 14.18% | 9.29% |
| CAST | 4.25 | 8.90 | 17.74% | 5.62% |
| EW | 4.34 | 9.15 | 19.17% | 8.93% |
| WD | 4.29 | 9.10 | 20.60% | 10.38% |
| UPGMA | 3.17 | 8.61 | 22.92% | 8.72% |
| DR-e1 | 4.26 | 9.82 | 22.75% | 15.71% |
| DR-i2 | 4.29 | 9.81 | 21.42% | 15.29% |
| RB-e1 | 4.57 | 9.78 | 18.68% | 13.22% |
| RB-i2 | 4.53 | 9.77 | 20.65% | 12.24% |

[a] The numbers of clusters for the Yin-Chen and CAST methods generated by the adjustable parameters are 999 and 1001 respectively.

Table 6-6 The evaluation of different clustering methods (1000 clusters) for the IDAlert dataset based on the four different evaluation criteria

Focusing on Table 6-6 with the visual inspection, the result is similar to the MDDR 1000 clusters (as shown in Table 6-3), the agglomerative method, UPGMA, has the consistently and noticeably the best performance on two types of Shannon Entropy and F-Measure, but ordinary result on the QCI. The Direct method has the significantly best performance on the QCI, and the Ward's method has consistently better performance than the extended Ward's. As for the comparison between the Direct and Repeated Bisection methods, the Direct method has consistently better performance on the Shannon Entropy, F-Measure and QCI evaluations than Repeated Bisection, but has similar results on the evaluation using Entropy based on cluster size. However, the same clustering method with two different criterion functions, *e1* and *i2*, generated similar and closer values of evaluation criteria. In addition, either *e1* or *i2* offers inconsistently better or worse performance to each other. In addition, none of these clustering methods is the most consistently ineffective in this case. Again, for getting a more quantitative view of the effectiveness of the clustering methods, a Kendall's *W* test of statistical significance was employed to evaluate the consistency.

Table 6-5(b) (for the IDAlert dataset) reveals the results of a Kendall's *W* analysis, showing the *W* and $\chi^2$ values for the combination of number of clusters and clustering performance generated by four evaluation criteria. The computed values of *W* and $\chi^2$ for the IDAlert dataset with 1000 clusters are 0.26 and 8.22 respectively (as shown in Table 6-5(b)).

As we mentioned previously, the critical value for chi-square ($\chi^2$) distribution at $\alpha$=0.05 level of statistical significance is 15.51 for eight degrees of freedom. Similarly, an identical lack of consistency is also found in the IDAlert dataset, and it will hence be reported that none of these values in Table 6-5(b) are significant at either $\alpha$=0.05 level or $\alpha$=0.01 level. Therefore, there would hence appear to be no strongly significant measure of agreement between the clustering methods and the evaluation criteria. Hence, there is no obvious "best" clustering method recommended to be generally applicable.

In the next section, a different type of evaluation of correlation based on individual evaluation criterion was carried out.

### 6.5.3 The Evaluation of Clustering Methods Based on Individual Criterion

There is no consistency between clustering methods and evaluation criteria (see Table 6-5), we hence carried out a distinct condition of Kendall's $W$ test of statistical significance to find if there is any consistency between clustering methods and activity classes for a particular evaluation criterion. Taking the MDDR dataset as an example, we generated the Shannon Entropy (Table 6-7) and QCI (Table 6-9) values over eleven activity classes. In the case of the MDDR 500 clusters (Table 6-7), the Shannon Entropy values of the Yin-Chen and CAST clusterings based on AC1 (i.e. activity-class-1, which is the 5HT3 antagonists in Table 4-1) are 6.24 and 3.95 respectively. The eleven activity classes are denoted by AC1 to AC11 in Tables 6-7 to 6-10. The other case of the MDDR 800 clusters (Table 6-9), the QCI values of the Yin-Chen and CAST methods based on AC2 (5HT1A agonists) are 8.05% and 6.76% respectively.

These eleven activity classes were then considered as the judges ranking the nine clustering methods instead of the four evaluation criteria in order of decreasing effectiveness, i.e. $k$=11 and $N$=9. In Table 6-8, we rank these Shannon Entropy values based on each single activity class in ascending order (small Entropy value indicates good clustering). Hence, the rankings for the Yin-Chen and CAST methods based on AC1 are 9 and 3 respectively. Similarly, in Table 6-10, the QCI values are ranked based on individual activity class, the rankings for the Yin-Chen and CAST methods based on AC2 are 8 and 9 respectively.

| MDDR dataset 500 clusters | Evaluation using Shannon Entropy | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AC1 | AC2 | AC3 | AC4 | AC5 | AC6 | AC7 | AC8 | AC9 | AC10 | AC11 |
| YC | 6.24 | 6.25 | 5.07 | 5.03 | 6.32 | 6.18 | 6.34 | 6.67 | 5.83 | 5.53 | 5.42 |
| CAST | 3.95 | 3.46 | 3.59 | 2.94 | 0.15 | 0.57 | 2.84 | 3.70 | 1.62 | 4.38 | 4.12 |
| EW | 4.22 | 4.59 | 3.89 | 4.21 | 2.17 | 3.27 | 3.81 | 4.21 | 3.22 | 4.53 | 4.18 |
| WD | 4.16 | 4.75 | 3.84 | 4.05 | 2.13 | 3.31 | 3.86 | 4.21 | 3.25 | 4.48 | 4.02 |
| UPGMA | 3.32 | 3.31 | 3.61 | 2.74 | 0.07 | 0.53 | 2.74 | 3.32 | 1.92 | 4.21 | 3.73 |
| DR-e1 | 4.04 | 4.47 | 3.89 | 4.38 | 1.78 | 3.01 | 3.58 | 4.18 | 3.36 | 4.40 | 4.23 |
| DR-i2 | 4.23 | 4.45 | 3.86 | 4.23 | 2.24 | 2.55 | 3.64 | 4.22 | 3.51 | 4.65 | 4.33 |
| RB-e1 | 4.77 | 4.99 | 4.04 | 4.51 | 2.35 | 3.94 | 4.38 | 5.13 | 3.61 | 4.68 | 5.06 |
| RB-i2 | 4.83 | 5.07 | 4.32 | 4.78 | 2.24 | 3.73 | 4.46 | 5.50 | 3.91 | 4.96 | 4.79 |

Table 6-7 The Shannon Entropy values of clustering methods for each activity class of the MDDR dataset

| MDDR dataset 500 clusters | Ranked by the Shannon Entropy | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AC1 | AC2 | AC3 | AC4 | AC5 | AC6 | AC7 | AC8 | AC9 | AC10 | AC11 |
| YC | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| CAST | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 3 |
| EW | 5 | 5 | 5 | 4 | 5 | 5 | 5 | 4 | 3 | 5 | 4 |
| WD | 4 | 6 | 3 | 3 | 4 | 6 | 6 | 5 | 4 | 4 | 2 |
| UPGMA | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 |
| DR-e1 | 3 | 4 | 6 | 6 | 3 | 4 | 3 | 3 | 5 | 3 | 5 |
| DR-i2 | 6 | 3 | 4 | 5 | 6 | 3 | 4 | 6 | 6 | 6 | 6 |
| RB-e1 | 7 | 7 | 7 | 7 | 8 | 8 | 7 | 7 | 7 | 7 | 8 |
| RB-i2 | 8 | 8 | 8 | 8 | 7 | 7 | 8 | 8 | 8 | 8 | 7 |

Table 6-8 The performance of clustering methods for each activity class ranked by the Shannon Entropy values for the MDDR dataset

| MDDR dataset 800 clusters | Evaluation using QCI | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AC1 | AC2 | AC3 | AC4 | AC5 | AC6 | AC7 | AC8 | AC9 | AC10 | AC11 |
| YC | 7.85% | 8 05% | 8.43% | 8 00% | 10.42% | 8.82% | 9.13% | 9.62% | 7.69% | 8.45% | 8.07% |
| CAST | 7.73% | 6.76% | 3.51% | 3.31% | 13.91% | 25.47% | 7.45% | 9.50% | 6.67% | 6.55% | 6.77% |
| EW | 13.93% | 12.40% | 8 84% | 9.20% | 68.71% | 32.09% | 30.77% | 20.87% | 15.88% | 8 50% | 9.80% |
| WD | 17.98% | 14.87% | 10.73% | 9.83% | 69.14% | 30.45% | 30.51% | 21.78% | 18 06% | 10.51% | 12.03% |
| UPGMA | 18.00% | 12.20% | 6.08% | 6.68% | 14 93% | 38.78% | 9.38% | 8.78% | 5.48% | 10.47% | 10.29% |
| DR-e1 | 21.76% | 18.29% | 14 02% | 11.43% | 58.03% | 35.85% | 36.49% | 32.38% | 19.88% | 14.03% | 14.55% |
| DR-i2 | 17.91% | 17.47% | 11.55% | 10.15% | 77.24% | 38.15% | 27.84% | 28.09% | 18.77% | 13.14% | 14.95% |
| RB-e1 | 16.39% | 12.62% | 10.67% | 9.01% | 37.21% | 23.57% | 21.01% | 17.83% | 14.35% | 13.40% | 9.06% |
| RB-i2 | 13.76% | 11.41% | 8.90% | 7.75% | 39.72% | 19.51% | 14.10% | 13.57% | 15.73% | 9.59% | 9.86% |

Table 6-9 The QCI values of clustering methods for each activity class of the MDDR dataset

| MDDR dataset 800 clusters | Ranked by the QCI | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AC1 | AC2 | AC3 | AC4 | AC5 | AC6 | AC7 | AC8 | AC9 | AC10 | AC11 |
| YC | 8 | 8 | 7 | 6 | 9 | 9 | 8 | 7 | 7 | 8 | 8 |
| CAST | 9 | 9 | 9 | 9 | 8 | 6 | 9 | 8 | 8 | 9 | 9 |
| EW | 6 | 5 | 6 | 4 | 3 | 4 | 2 | 4 | 4 | 7 | 6 |
| WD | 3 | 3 | 3 | 3 | 2 | 5 | 3 | 3 | 3 | 4 | 3 |
| UPGMA | 2 | 6 | 8 | 8 | 7 | 1 | 7 | 9 | 9 | 5 | 4 |
| DR-e1 | 1 | 1 | 1 | 1 | 4 | 3 | 1 | 1 | 1 | 1 | 2 |
| DR-i2 | 4 | 2 | 2 | 2 | 1 | 2 | 4 | 2 | 2 | 3 | 1 |
| RB-e1 | 5 | 4 | 4 | 5 | 6 | 7 | 5 | 5 | 6 | 2 | 7 |
| RB-i2 | 7 | 7 | 5 | 7 | 5 | 8 | 6 | 6 | 5 | 6 | 5 |

Table 6-10 The performance of clustering methods for each activity class ranked by the QCI values for the MDDR dataset

In addition, all evaluation criteria can be judged by these eleven activity classes except the evaluation using Entropy based on cluster size, because the value of Entropy based on cluster size is simply based on the number of objects in a cluster instead of the activity classes. Hence, only three evaluation criteria, Shannon Entropy, F-Measure and QCI, are discussed in this section.

With these rankings judged by the eleven activity classes, the extent of the correlation among eleven sets of rankings for nine clustering methods can be checked by the Kendall's *W* test. The overall results were obtained and shown in Tables 6-11 (for the MDDR dataset) and 6-12 (for the IDAlert dataset).

| | MDDR dataset | | | | | |
|---|---|---|---|---|---|---|
| | Shannon Entropy | | F-Measure | | QCI | |
| # Clusters | *W* | *χ2* | *W* | *χ2* | *W* | *χ2* |
| 500 | 0.90 | 79.63 | 0.41 | 36.25 | 0.73 | 64.62 |
| 600 | 0.83 | 73.23 | 0.46 | 41.24 | 0.74 | 65.55 |
| 700 | 0.79 | 69.56 | 0.43 | 37.94 | 0.75 | 66.13 |
| 800 | 0.87 | 76.67 | 0.36 | 32.17 | 0.73 | 65.09 |
| 900 | 0.80 | 71.14 | 0.44 | 39.08 | 0.78 | 68.99 |
| 1000 | 0.74 | 65.30 | 0.39 | 34.86 | 0.76 | 67.48 |

Table 6-11 Kendall *W* and $\chi^2$ values based on 11 different activity classes for each evaluation measure on the MDDR dataset

Table 6-11 shows the Kendall *W* and $\chi^2$ values based on eleven different activity classes for each evaluation criterion on the MDDR dataset. For example, in the case of the MDDR dataset with the evaluation using Shannon Entropy, the values of *W* and chi-square ($\chi^2$) for the 500 clusters are 0.90 and 79.63.

The critical value of the chi-square ($\chi^2$) distribution mainly depends on the sample size i.e. the number of clustering methods in our study. However, even with the change of the number of judges, the critical values of the chi-square

($\chi^2$) distribution remain the same 20.09 and 15.51 (as in Section 6.5.1) for the $\alpha$=0.01 and $\alpha$=0.05 levels respectively for the degrees of freedom is eight.

In terms of evaluation using Shannon Entropy, all $\chi^2$ values of different numbers of clusters are significantly larger than the critical values at $\alpha$=0.01 and $\alpha$=0.05 levels, i.e. all tests are significant. Tables 6-13 and 6-14 summarize the top three best performances of clustering methods evaluated by individual criterion over varied partition sizes for the MDDR and IDAlert datasets respectively. Moreover, the best method for the combination of partition size and evaluation criterion is represented in bold font. We hence had a visual inspection on the Entropy values for each single Kendall's *W* test on the MDDR dataset (Table 6-13), and found the UPGMA method has consistently and significantly the best performance across all partition sizes. In addition, the CAST method is consistently in the leading group (i.e. top three best performances). Hence, we can conclude that the UPGMA method has obvious performance benefit, and the active molecules of clustering results are more concentrated in certain clusters.

All the tests in the evaluation using F-Measure of the MDDR dataset are also found statistical significance (Table 6-11). According to the visual inspection in Table 6-13, there is not a single clustering method offering consistently the best results over all partition sizes. The UPGMA and Ward's methods tend to be more effective than the others on the evaluation using F-Measure.

As for the evaluation using QCI for the MDDR dataset, all tests show statistical significance (Table 6-11). The visual inspection in Table 6-13 found that the Direct-e1 method has consistently and significantly the best performance over all numbers of clusters, and the Direct-i2 and Ward's methods consistently remain in the leading groups. Hence, we can conclude that the Direct-e1 method has obvious performance benefit.

| | IDAlert dataset | | | | | |
|---|---|---|---|---|---|---|
| | Shannon Entropy | | F-Measure | | QCI | |
| # Clusters | *W* | *χ2* | *W* | *χ2* | *W* | *χ2* |
| 500 | 0.80 | 71.07 | 0.59 | 52.52 | 0.45 | 40.15 |
| 600 | 0.79 | 69.87 | 0.59 | 52.32 | 0.45 | 40.02 |
| 700 | 0.79 | 69.92 | 0.65 | 57.77 | 0.46 | 41.21 |
| 800 | 0.79 | 69.86 | 0.67 | 58.98 | 0.43 | 38.23 |
| 900 | 0.79 | 69.91 | 0.60 | 53.24 | 0.40 | 35.52 |
| 1000 | 0.80 | 71.08 | 0.58 | 51.22 | 0.38 | 33.64 |

Table 6-12 Kendall *W* and $\chi^2$ values based on 11 different activity classes for each evaluation measure on the IDAlert dataset

Table 6-12 shows the results of the Kendall *W* test for the IDAlert dataset. A distinct condition of Kendall's *W* test of statistical significance was also carried out for this dataset. In terms of the evaluation using Shannon Entropy, all these six tests are found statistical significance. A visual inspection of Entropy values for each single Kendall's *W* test, we found the leading groups of CAST, Ward's and e-Ward's clustering methods consistently offer the better performance over all numbers of clusters (Table 6-14), of which the CAST method has the consistently best performance.

In terms of the evaluation using F-Measure, all the tests are found statistical significance (Table 6-12). However, a similar result to the Shannon Entropy, the individual inspection of F-Measure values for each single Kendall's *W* test shows that the leading group of Ward's, e-Ward's and CAST clustering methods has consistently better performance, of which the Ward's method consistently offers the best results.

As for the evaluation using QCI, all these tests are found statistical significance (Table 6-12). The individual inspection of QCI values in Table 6-14 shows that the Ward's method provided the consistently best results over all partition sizes. Furthermore, the e-Ward's and Yin-Chen clustering methods also consistently remain in the leading groups. Hence, we can conclude that the Ward's method

has obvious performance benefit.

Combining all the analysis in this section (6.5.3) can be summarized by following findings: first, according to the top three best performances of clustering methods evaluated by each criterion function across all partition sizes on two datasets (Tables 6-13 and 6-14), the leading group of clustering methods in the IDAlert dataset is more consistent than in the MDDR dataset. In addition, it is easy to identify the best method or the method tending to be the best, across the combinations of evaluation criterion and partition size. However, no single clustering method was found to be consistently effective over the combinations of evaluation criterion, partition size and dataset.

Second, the inspection of the leading groups in both datasets shows that the CAST method tends to have better Shannon Entropy, and the Ward's method tends to have better F-Measure and QCI values over two datasets. However, some methods have the best results only on a specific dataset. For example, the UPGMA and Direct-e1 methods have the consistently best Shannon Entropy and QCI respectively on the MDDR dataset only; the e-Ward's method tends to have better performance over all evaluation criteria on the IDAlert dataset only.

| | MDDR dataset | | |
|---|---|---|---|
| # Clusters | Shannon Entropy | F-Measure | QCI |
| 500 | **UPGMA**, CAST, DR-e1 | **WD**, DR-e1, EW | **DR-e1**, DR-i2, WD |
| 600 | **UPGMA**, CAST, DR-i2 | **DR-i2**, WD, DR-e1 | **DR-e1**, WD, DR-i2 |
| 700 | **UPGMA**, CAST, DR-e1 | **DR-e1**, WD, UPGMA | **DR-e1**, DR-i2, WD |
| 800 | **UPGMA**, CAST, DR-e1 | **UPGMA**, WD, DR-e1 | **DR-e1**, DR-i2, WD |
| 900 | **UPGMA**, CAST, DR-e1 | **UPGMA**, DR-i2, WD | **DR-e1**, DR-i2, WD |
| 1000 | **UPGMA**, CAST, DR-i2 | **UPGMA**, WD, DR-e1 | **DR-e1**, DR-i2, WD |

Table 6-13 The top three best performances of clustering methods evaluated by each criterion function for varied numbers of clusters of the MDDR dataset

| | IDAlert dataset | | |
|---|---|---|---|
| # Clusters | Shannon Entropy | F-Measure | QCI |
| 500 | **CAST**, WD, EW | **WD**, EW, CAST | **WD**, EW, YC |
| 600 | **CAST**, WD, EW | **WD**, EW, CAST | **WD**, EW, YC |
| 700 | **CAST**, WD, EW | **WD**, EW, CAST | **WD**, EW, YC |
| 800 | **CAST**, WD, EW | **WD**, EW, CAST | **WD**, EW, YC |
| 900 | **CAST**, WD, EW | **WD**, EW, CAST | **WD**, EW, YC |
| 1000 | **CAST**, WD, EW | **WD**, EW, CAST | **WD**, YC, EW |

Table.6-14 The top three best performances of clustering methods evaluated by each criterion function for varied numbers of clusters of the IDAlert dataset

### 6.5.4 The analysis of Comparative Clustering Methods for the MDDR and IDAlert datasets

In this section, we carried out conventional comparison of clustering methods, which can been found in the literature e.g. the comparison between hierarchical and partitional methods, Ward's and e-Ward's, and two divisive clustering methods in CLUTO tool kit.

We first consider the conventional comparison of hierarchical and partitional clusterings, taking Tables 6-13 and 6-14 together into account, the hierarchical clustering of UPGMA and CAST methods have better performance on the evaluation using Shannon Entropy than the partitional clustering of Direct and Repeated Bisection methods over the two datasets. As for the evaluation using F-Measure and QCI, hierarchical clustering of Ward's method is superior to partitional clustering methods only on the IDAlert dataset. Hence there is no consistent performance benefit on choosing either hierarchical or partitional clustering methods.

As for the comparison between Ward's method and its variation of e-Ward's method, the Ward's clustering method has consistently better performance than

e-Ward's across all types of evaluation on the two datasets (Tables 6-13 and 6-14). Hence, on the choice of Ward's and e-Ward's methods for the chemical data of the sort considered in this study, we suggest using Ward's method to generate performance benefit rather than e-Ward's method.

Two partitional clustering methods in CLUTO tool kit were also evaluated. The Direct method is consistently superior to the Repeated Bisection method over all types of evaluation on the two datasets (Tables 6-13 and 6-14). Although Repeated Bisection method has been reported effective in the application of document clustering (Steinbach et al., 2000), for chemical data of the sort applied in this study, we suggest that the use of the Direct method could bring performance benefit rather than the Repeated Bisection.

Figures 6-1 to 6-4 show the overall clustering performance evaluated by varied evaluation criteria over two datasets. Overall, the MDDR dataset has better performance on the evaluation using F-Measure and QCI (Figures 6-3 and 6-4). That is because the evaluation using F-Measure and QCI is based on the number of actives for a certain activity class. For some activity classes e.g. Angiotensin II AT1 antagonists, Substance P antagonists, HIV-1 protease inhibitors, and 5HT1A agonists, the number of actives for these classes in the MDDR dataset is much more (from 83 to 33 actives) than in the IDAlert dataset (see Table 4.1). Hence, under the condition of same number of clusters, the performance of F-Measure and QCI on the MDDR dataset is easily better than the IDAlert dataset.

The clustering performances evaluated by Shannon Entropy in Figure 6-1 are very similar except the Yin-Chen and CAST methods. That is because the number of clusters for those two clustering methods is determined by an adjustable parameter which sometimes may be sensitive and fail to generate exactly partition sizes (see Table 6-2). Generally speaking, in each Yin-Chen clustering result, the MDDR dataset tends to have slightly more number of clusters than the IDAlert; whereas, in each CAST clustering result, the MDDR dataset tends to have slightly less partition sizes than the IDAlert dataset. In addition, the Shannon Entropy, in essence, is basically dependent on the

number of clusters. The more number of clusters tends to lead worse Shannon Entropy. Hence, the more number of clusters with Yin-Chen method on the MDDR dataset is inferior (high Entropy value) to the IDAlert dataset; on the contrary, the less number of clusters with CAST method of MDDR dataset is superior (low Entropy value) to the IDAlert dataset.

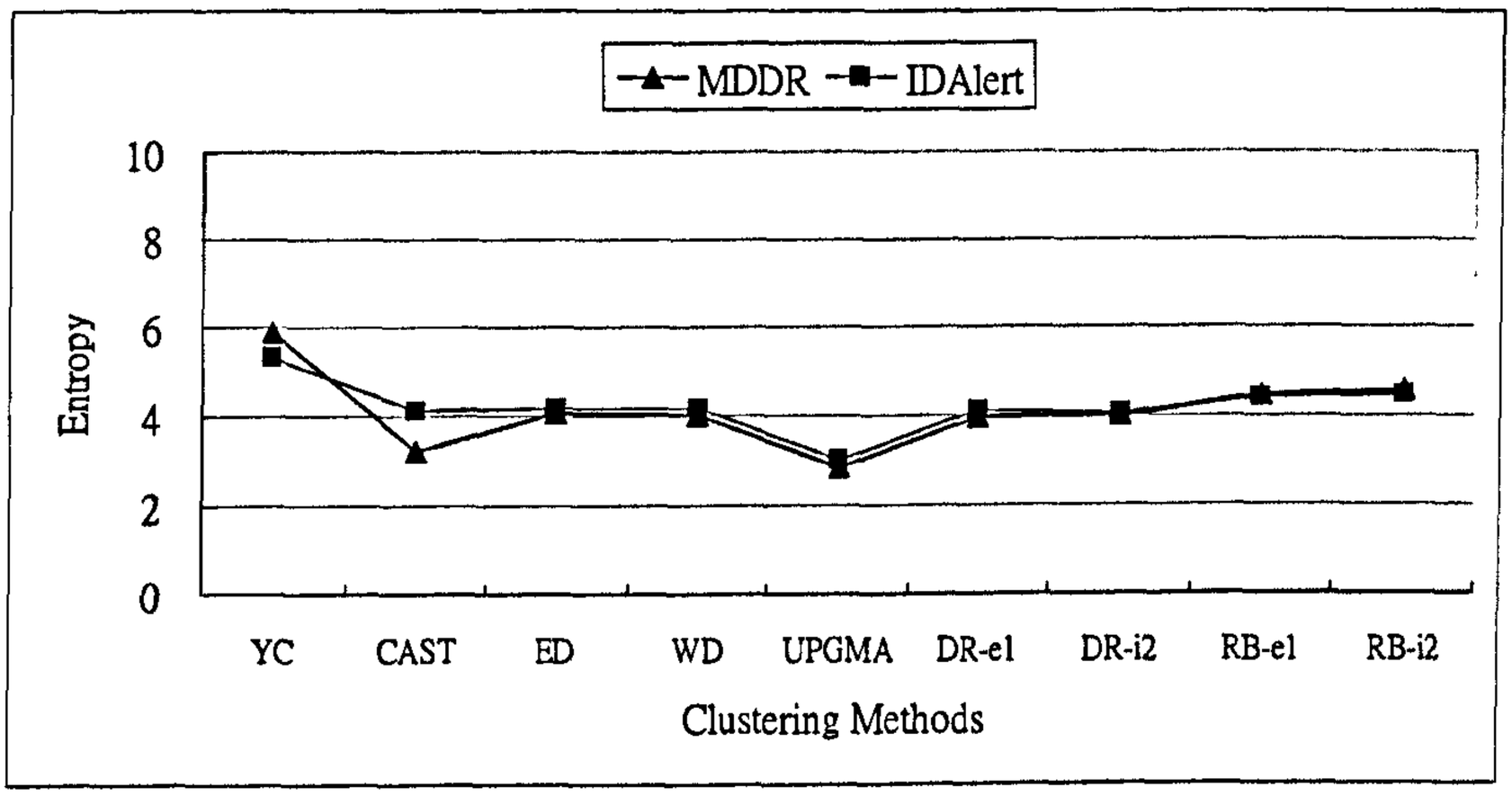



Figure 6-1 The overall performance evaluated by the Shannon Entropy over two datasets

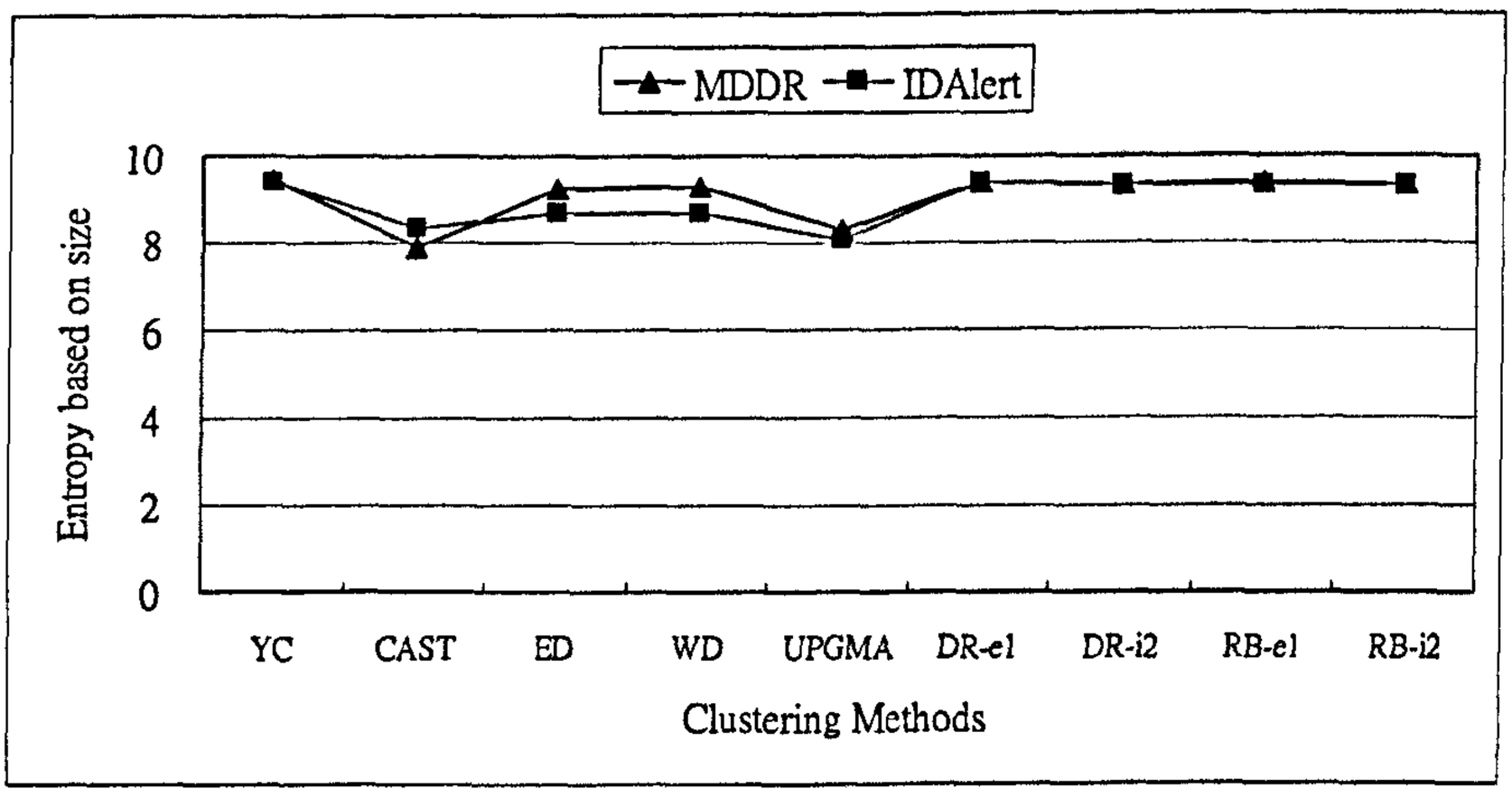



Figure 6-2 The overall performance evaluated by the Shannon Entropy based on cluster size over two datasets

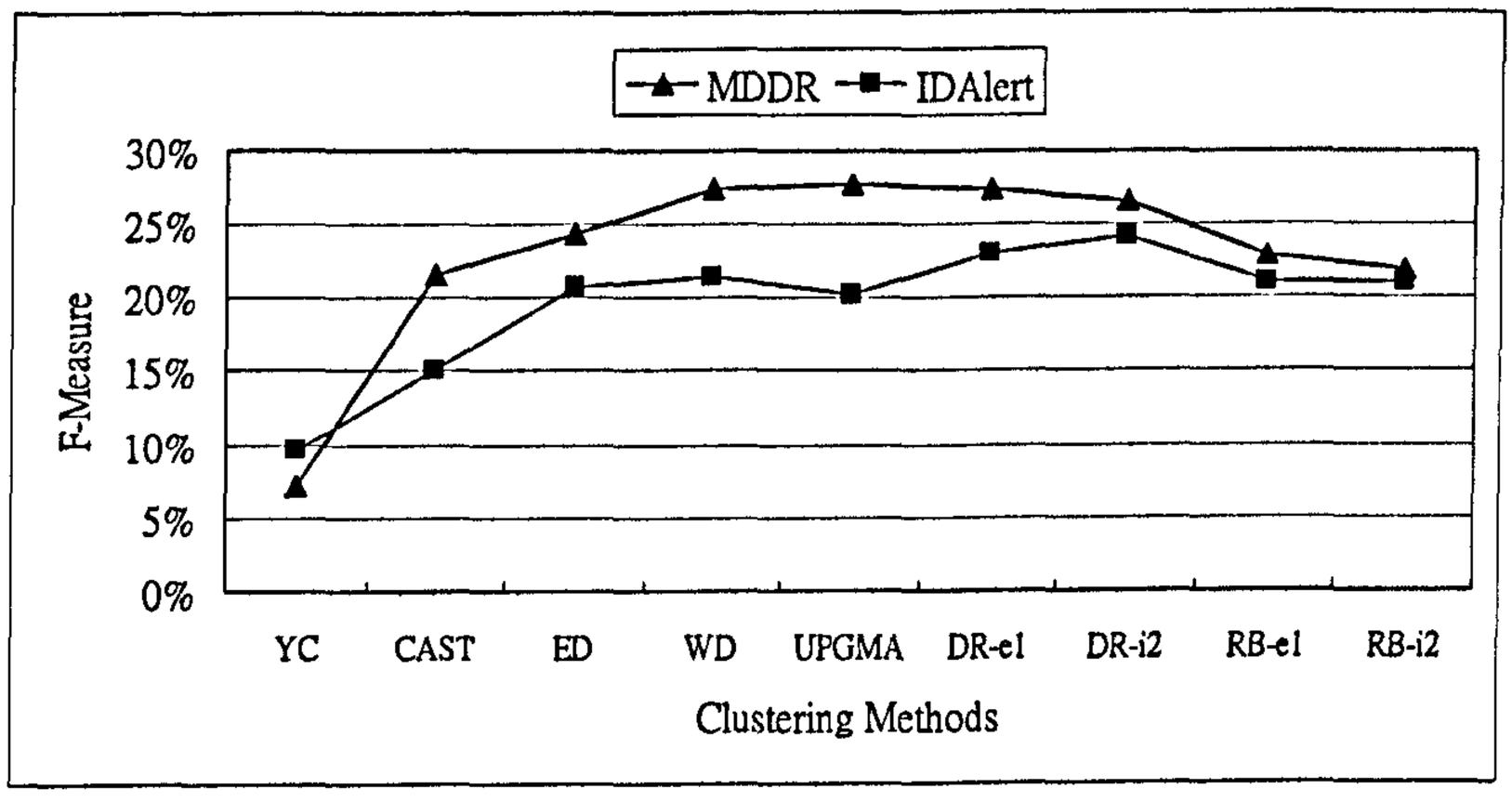



Figure 6-3 The overall performance evaluated by the F-Measure over two datasets

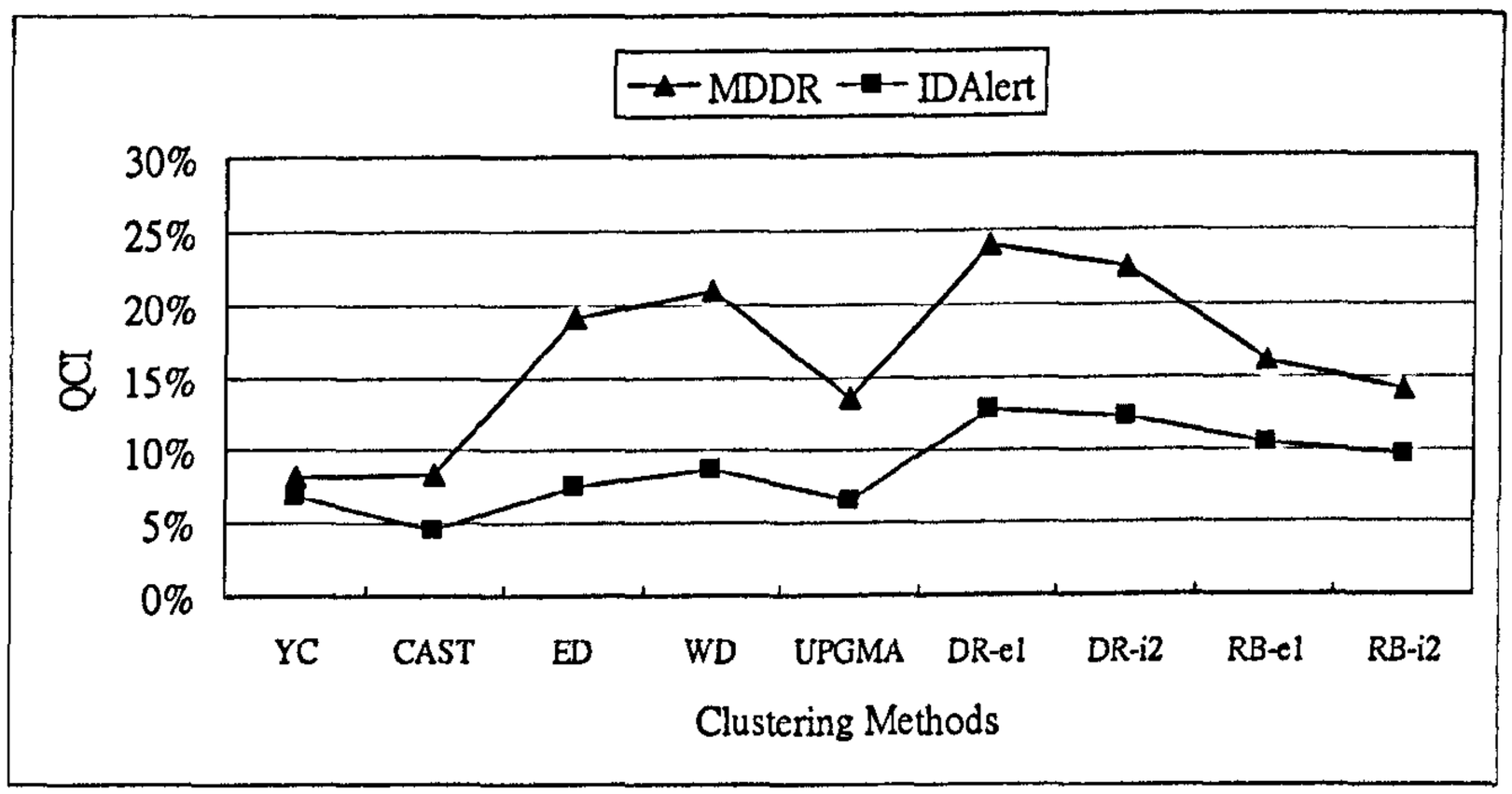



Figure 6-4 The overall performance evaluated by the QCI over two datasets

## 6.6 Conclusions

There are many different clustering methods published and used in a wide range of application domains, some are reported effective in certain applications or fields. Moreover, the type of data is also a critical factor of clustering quality. In this study, we experimentally evaluated nine clustering methods by means of four evaluation criteria to obtain clustering solutions from chemical datasets characterized by ECFP_4 fingerprint. A clustering method offering consistently better performance over more evaluation criteria indicates a superior partitioning. However, it is difficult for a single clustering method to provide consistently better performance over multiple evaluation criteria, especially over different types of criterion. One clustering method could have superior performance on a specific evaluation criterion but have ordinary result on another criterion as in our findings.

Two findings are worth discussing: first, according to the conclusions (Section 5.13) in Chapter 5, the non-standardization procedure ($Z_0$) with Ward's clustering offers the best F-Measure only on the datasets with Holograms fingerprints. However, in this chapter, the same datasets were unstandardized and characterized by the ECFP_4 fingerprints. The results (Tables 6-13 and 6-14) shows that the Ward's clustering offers the consistently best F-Measure on the IDAlert dataset, and tends to have better F-Measure (i.e. remains in the leading group) on the MDDR dataset, over all partition sizes. This finding suggests that the Ward's method tends to have better F-Measure on the chemical data with binary (e.g. ECFP_4) or non-binary (e.g. Holograms) fingerprints representation.

Second, according to the conclusions (Section 5.13) in Chapter 5, the non-standardization procedure ($Z_0$) with UPGMA method yields the worst QCI, and with Direct method tends to have better QCI, on the datasets with all representations (including, of course, the Holograms fingerprints). In this chapter (see Figure 6-4), the UPGMA method provides the consistently worst QCI, and the Direct method generates the consistently best QCI, in comparison

with the other six clustering methods, which are also used in the extensive study of Chapter 5 (i.e. the nine clustering methods in this chapter except the Yin-Chen and CAST methods). The finding here is in line with the conclusions in Chapter 5, and suggests that the Direct method tends to have better QCI for the chemical data of binary (e.g. ECFP_4) or non-binary (e.g. Holograms) fingerprints representation used in this thesis.

Our results suggest that, for chemical data of the sort considered here, no consistent performance benefit that is likely to be obtained from the use of any particular clustering method using the chosen evaluation methods. Since no single clustering method is universal to all applications, the study of consensus clustering is hence carried out in the next chapter to integrate the clustering results from different methods and with the aim of generating a representative consensus result which is reported robust and reliable.

# Chapter 7 : Comparison of Chemical Consensus Clustering Methods Using Fingerprint-based Similarity Measures

## 7.1 Introduction

An inherent feature of clustering is that distinct methods or even a single method on the same dataset will generate different clustering results. In addition, most clustering methods offer simply an approximation to the optimal result, and find only a single result based on some specific clustering criterion. Hence, instead of determining one specific clustering method, some typical issues have been discussed such as selecting the best result, verification of the best result, and fusion of all results to get a consensus clustering of a dataset.

Data fusion is the technique that combines the information from different results or data sources aiming to obtain the efficient and accurate output rather than using a single source. There is a growing interest in the literature e.g. chemoinformatics, because several studies found that data fusion improved the results significantly in virtual screening experiments (Holliday et al., 2002; Salim et al., 2003; Whittle et al., 2003; Willett, 2006). A similar technique to data fusion, consensus clustering is the process of combining the different clustering results in order to yield a result with robustness and confidence.

Consensus clustering, also known as clustering ensemble, clustering combination, median partition, clustering of clustering, and clustering aggregation (Gionis et al., 2007), is a technique that integrates the results of multiple runs from either different clustering methods or different initializations e.g. parameter or random values of a specific clustering approach, into a single representative consensus (Topchy et al., 2004 ). It can not only

enhance the robustness but also usually offers better clustering results than using a single clustering method, also it is less sensitive to the dataset variations, noise, and outliers (Nguyen & Caruana, 2007). For example, K-Means method is usually sensitive to the initialization, however by integrating the multiple runs of K-Means clustering, the consensus result will be more reliable. Also, when there is no prior knowledge for the number of clusters, it will be difficult to determine. However, the consensus clustering over multiple runs can be more confident in determining the number of clusters.

In theory, the aim of consensus clustering is to find a median point (clustering) among the clustering space, which minimizes disagreement between the input clusterings. Consensus clustering in essence is NP-complete as has been proven in the literature (Filkov & Skiena, 2004a), and a variety of approximations have been applied to such a problem. A simplified example of consensus clustering is shown as Figure 7-1, where Clustering 1 to 4 can be different runs of a single clustering method or varied clustering methods.

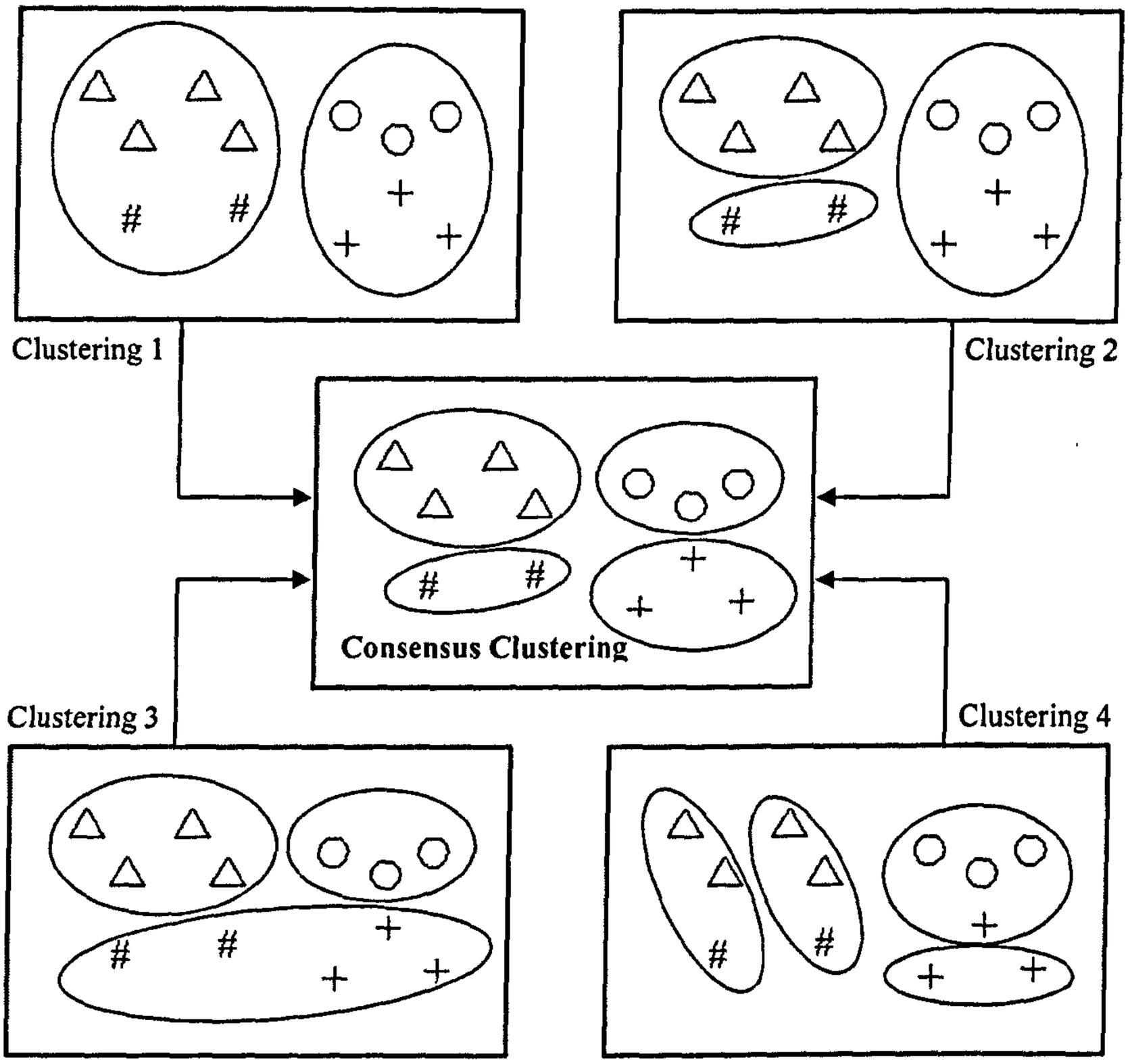



Figure 7-1 Example of consensus clustering

Few studies of consensus clustering employ weighting schemes. One reason is that most applications of consensus clustering involve multiple runs of a single clustering method; all results are treated equally. However, the consensus results may consist of very different clusterings, and these input clusterings could be significantly different or correlative. Hence treating all input clusterings equally may not be effective on the consensus result. Some studies (Gullo et al., 2009; Li & Ding, 2008; Domeniconi & Al-Razgan, 2009; Wang et al., 2009) addressed the importance of weighting schemes and showed that the performance of consensus clustering can be improved by using proper weighting schemes, and can even give results as good as the individual best clustering method (Al-Razgan & Domeniconi, 2006).

Previous studies of consensus clustering were mostly the applications of multiple runs of a single clustering method, and treated the various clustering results equally (Fred and Jain, 2002; Nguyen and Caruana, 2007). However, our study is based on different clustering methods, and considers the clustering results from distinct methods differently by employing a simple weighting scheme.

In addition, the application of consensus clustering has been shown to be effective in various fields, such as data mining, pattern recognition, and machine learning (Gionis et al., 2007); and is largely based on categorical data and heterogeneous data. For example, in dealing with the attribute of categorical data in the manner of consensus clustering, not only can each single attribute be treated as an input clustering, but also each single class can be considered as a cluster of its attribute, rather than finding a similarity or distance function which is believed difficult to determine (Goder & Filkov, 2008). Application in chemical information has mainly been reported in the field of bioinformatics e.g. gene expression data (Filkov & Skiena, 2004a; Monti et al., 2003), but there has been no application on the field of chemoinformatics. Our study is in this field using chemical fingerprints to represent molecules.

## 7.2 Related Work

A number of algorithms have been proposed to solve the consensus clustering problem, and classification of these algorithms may be different (Strehl & Ghosh, 2002), some well know classes are Clustering-based Similarity Algorithms, HyperGraph Partitioning Algorithms, and Meta-Clustering Algorithms. However, we only briefly review some commonly used consensus clustering algorithms.

Cluster-based Similarity Algorithms are mainly based on a similarity matrix in which each entry records pairwise relationship of the number of times objects *i* and *j* have been clustered together to the input clusterings; the details of constructing the similarity matrix will be discussed in a later section. It is the

simplest and most commonly used class of consensus algorithm. Most existing clustering algorithms perform clustering based on such a similarity matrix. Bertolacci and Wirth (2007) apply consensus clustering to the categorical datasets, Mushroom and 20 Newsgroups, from the UCI repository; and found the CCLP-pivot algorithm, a variation of the CC-pivot algorithm, proposed by Ailon et al. (2008) had better performance on the consensus clustering problem, while the Furthest Linkage Algorithm had the worst result. Although, the CCLP-pivot algorithm has been reported effective, it is a time consuming algorithm with $O(n^8)$, and is thus not suitable to deal with large datasets (Bertolacci & Wirth, 2007). Thus, we just simply employed the CC-pivot algorithm in our experiment, which has time complexity $O(kmn)$, where $k$ is the number of clusters and $m$ represents the number of input clusterings for consensus. Moreover, the Average Linkage Algorithm was found also offered as good performance as the CCLP-pivot algorithm in their study. Nguyen and Caruana (2007) proposed three iterative algorithms to carry out the consensus clustering, which can be considered as the variations of K-Means method, and compared with eleven commonly used algorithms in consensus clustering. The result showed that the performance was as good as, and often better than, others.

Meta-Clustering Algorithms (Caruana et al., 2002) offer many clusterings for users to select which are considered to be good, rather than just generating a single optimal clustering. Zeng et al. (2002) compared the meta-clustering algorithm with those algorithms that have been successfully applied on bioinformatics e.g. K-Means, average linkage, and self-organized-maps (SOM), on both artificial and real (categorical) datasets. Their result showed that the meta-clustering algorithm with the proposed distance measure is effective.

In addition to the above two types of consensus algorithm, some other algorithms have been used in the literature. Fred and Jain (2002) proposed a single linkage technique, Minimum-Spanning-Tree (MST) based algorithm, to combine the results from multiple runs of K-Means method on both synthetic and real datasets, and found it effective. Our work here uses the same clustering technique but with cluster-based similarity approach.

## 7.3 Experimental

Our consensus clustering experiments used the MDDR and IDAlert datasets with the molecules represented by ECFP_4 fingerprints discussed previously in Chapter 4.

### 7.3.1 Measuring Consensus

There are varied methods to measure the similarity of a set of clusterings, such as Rand (Rand, 1971), Fowlkes-Mallows (Fowlkes and Mallows, 1983) and Jaccard Indices (Ben-Hur et al., 2002). The application of consensus clustering in our study is based on the similarity or distance between the input clusterings, in order to measure the consensus between input clusterings, with $N$ objects of a dataset, we defined a ($N$ $X$ $N$) symmetric matrix to record the pairwise similarity relationship; each entry in the matrix represents the proportion of clustering runs or number of input clusterings in which two objects are clustered together. It simply counts the pairs of co-clustered objects in the set of clusterings (Filkov & Skiena, 2004). That is, the entry ($i$ , $j$) in the similarity matrix indicates the number of times objects $i$ and $j$ are assigned to the same cluster divided by the total number of clustering runs. The consensus similarity matrix in essence is similar to the well known Rand Index. Most of the commonly used consensus clustering algorithms as described in the later section can be carried out with the consensus similarity matrix. Figure 7-2 illustrates how the measure of consensus similarity is computed with an example.

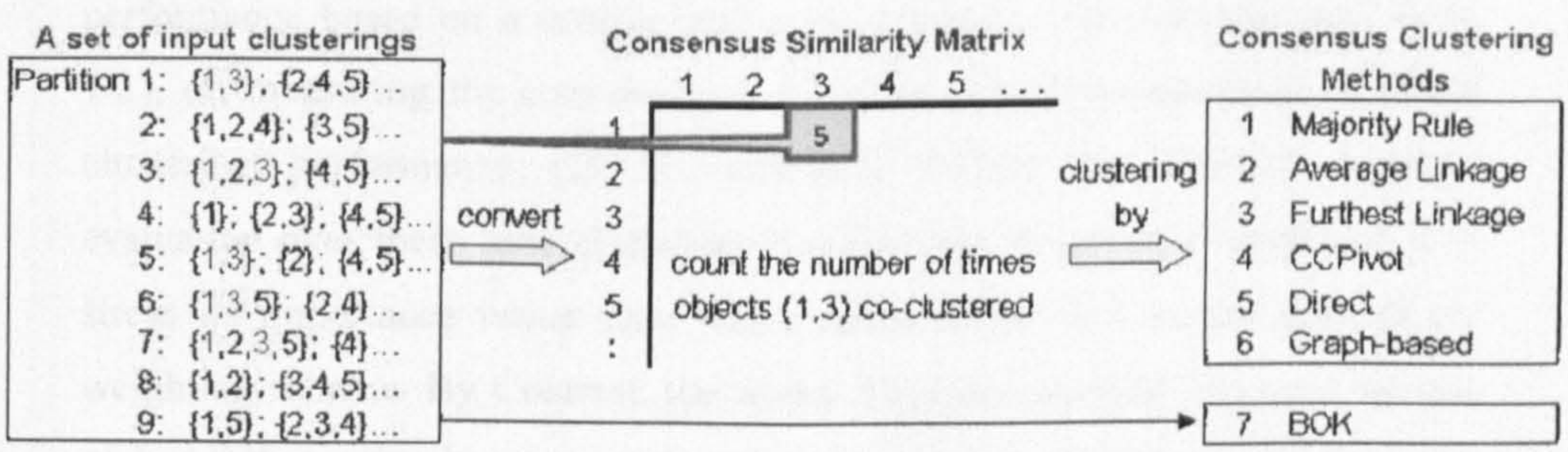



Figure 7-2 Example of consensus similarity measuring

### 7.3.2 Weighting Scheme

A common shortcoming for most consensus clustering methods is that the importance of all input clustering results is considered equally. Different clustering methods have different clustering performance. Weighting different clustering results for consensus clustering is expected to have better performance rather than simply averaging all results. However, only a few studies (Al-Razgan and Domeniconi, 2006; Li and Ding, 2008) in the literature have discussed the weighting schemes for consensus clustering. Gullo et al. (2009) proposed three types of diversity-based weighting schemes for consensus clustering, Single Weighting, Group Weighting and Dendrogram Weighting. These weighting schemes are designed to be independent of any specific consensus clustering method, moreover the correlations among input clusterings are also considered. Although there was no consistent benefit found from their weighting schemes over all datasets, the majority of results have been improved over the unweighting consensus clustering for certain datasets. In addition, for these studies, the classification of each object in the datasets is known, i.e. clustering on categorical data, and most datasets are from UCI (University of California in Irvine) data repository. Hence, the evaluation of clustering result is simply verifying if each object has been assigned to the right class, while in our study we used four different types of evaluation which have been discussed in Chapter 4.

In our study, the weighting scheme employed for measuring consensus is simply to apply larger weights to the clustering method with better performance based on a certain evaluation criterion. For example (see Table 7-1), on measuring the consensus of a pair of objects co-clustered, the best clustering performance (CLUTO-UPGMA) based on Shannon Entropy evaluation over these nine clustering methods will be given a weight of 9 to stress its importance rather than just given a count of 1 in the case of no weighting scheme. By Contrast, the worst, Yin-Chen method obtains a weight of 1. All these nine input clusterings in this experiment are basically from the previous study carried out in Chapter 6 and ranked by varied evaluation criteria.

However, in our example, a rank of 1 indicates the best performance, and a rank of 9 denotes the worst.

500 clusters of clustering on the MDDR dataset evaluated by Shannon Entropy

| Clustering Methods | Ranks | Weights |
|---|---|---|
| Yin-Chen | 9 | 1 |
| CAST | 2 | 8 |
| e-Ward's | 5 | 5 |
| Ward's | 4 | 6 |
| UPGMA | 1 | 9 |
| Direct-e1 | 3 | 7 |
| Direct-i2 | 6 | 4 |
| Repeated Bisection-e1 | 7 | 3 |
| Repeated Bisection-i2 | 8 | 2 |

Table 7-1 Example of the weights on measuring consensus

The consensus procedure could employ the above weighting scheme to generate a weighted consensus similarity matrix. Again, all the consensus algorithms except BOK (discussed in next section) can be carried out based on such a weighted consensus similarity matrix. This is because the essence of BOK algorithm is simply based on the calculation of average consensus similarity (or Rand distance) rather than by taking the ranking of different clustering methods into account.

### 7.3.3 Algorithms

Seven consensus clustering methods were employed in this study. CC-Pivot and BOK methods (Goder & Filkov, 2008) were coded using *Perl* script based on their algorithms, while the other five methods, Majority Rule, Average Linkage, Furthest Linkage, Direct and Graph-based, were carried out using the implementations in the CLUTO software package (CLUTO, 2003). In addition, the graph-based clustering method in CLUTO employs hypergraph partitioning algorithms as well as the efficient multilevel graph partitioning algorithms derived from METIS and hMETIS (Karypis, 2003) which are the commonly used packages in the application of graph-based consensus clustering (Nguyen

& Caruana, 2007; Strehl & Ghosh, 2002).

#### 7.3.3.1 Majority Rule

The Majority Rule is also called the Quota Rule (Goder & Filkov, 2008). In nature, it is a bottom-up agglomerative procedure in which every single object is considered as a cluster in the beginning, then for every pair of objects for which their consensus similarity is greater than a predefined threshold, these are merged into the same cluster; if objects are in the different clusters, then the clusters are merged. Those remaining objects which have not been assigned to any cluster will be considered as singletons (Fred, 2001). This technique is equivalent to the single linkage clustering (Fred & Jain, 2002). The threshold in this experiment is determined by the desired number of clusters. That is, the threshold is adjusted to generate the closest number of clusters to 500, 600, 700, 800, 900 and 1000.

#### 7.3.3.2 Average Linkage

The Average Linkage is a standard bottom-up agglomerative method which is also known as group average or Unweighted Pair-Group Method using Arithmetic averages (UPGMA). It begins with every object being assigned to a cluster; then, two clusters are merged with the minimum mean distance or maximum mean consensus similarity. However, the calculation of the mean distance or consensus similarity is based on the pairwise relationship between two clusters. It takes account of all possible pairs of objects between two clusters, not only the minimum or maximum distance (Everitt et al., 2001). Such iterative procedure of finding the maximum consensus similarity can be terminated when the maximum consensus similarity is smaller than a predefined threshold or when reaching the desired number of clusters. In our study we chose the latter as the terminating criterion in order to compare the performance of our previous study.

#### 7.3.3.3 Furthest Linkage

This is also known as complete linkage or farthest neighbour, which is the opposite of single linkage. The distance between two clusters is based on the

maximum of all possible pairwise distances i.e. farthest pair of objects, one object from each cluster. In each step, two clusters are merged with the smallest maximum pairwise distance, that is, the largest minimum pairwise consensus similarity in our study. However, there are variations of furthest linkage discussed in the literature (Bertolacci & Wirth, 2007; Gionis et al., 2007; Nguyen & Caruana, 2007), few are reported effective. Hence, in our study, we employed the traditional agglomerative furthest linkage clustering method provided in the CLUTO package to deal with the consensus problems.

#### 7.3.3.4 CC-Pivot

The CC-Pivot usually picks an object *p* randomly as a pivot, and then assigns every object having a consensus similarity with *p* greater than a predefined threshold to one cluster. It then iteratively chooses a new pivot object from the un-clustered objects. The procedure is executed repeatedly until all objects have been clustered. Again, the threshold is determined by the number of clusters as for the Majority Rule method. In addition, there are alternative ways to pick the pivot object in the literature (Zuylen, 2005), such as picking the pivot object with the smallest, average, or maximum consensus similarity or other similarity functions. However, in our study, we employed the most common random pivot object selection method.

#### 7.3.3.5 Direct

Nguyen and Caruana (2007) proposed a variation of the K-Means method, which is called Iterative Pairwise Consensus in their study, based on a consensus similarity matrix to solve the consensus clustering problem. The Iterative Pairwise Consensus (IPC) method offered better performance under some evaluation criteria rather than being consistently superior to others. However, we found another variation of the K-Means method useful in our previous study of Chapter 6, which is called the *direct* method in the CLUTO toolkit package developed by Karypis (2003) and is discussed in detail in Chapter 4. Hence we also employed the *direct* method based on the consensus similarity matrix to cope with the consensus problem in this study.

#### 7.3.3.6 Graph based

The graph-based consensus clustering method basically constructs a sparse graph to represent the similarity relations between the different objects (Karypis et al., 1999). In the essence of graph theory, the objects in the consensus similarity matrix correspond to the vertices or nodes, while the consensus similarities correspond to the edges. In the literature (Nguyen & Caruana, 2007; Strehl & Ghosh, 2002), METIS and hMETIS are the commonly used software packages in the application of graph-based consensus clustering such as HyperGraph partitioning clustering (HGPA) and Cluster-based Similarity Partitioning Algorithm (CSPA). In our study, we employed CLUTO's graph partitioning based clustering algorithm, *graph* method, since it integrates and exploits the advantage from the previous graph and hypergraph partitioning algorithms of METIS and hMETIS software packages. In *graph* method, each object (vertex) is connected to its most similar other objects using a nearest-neighbor algorithm to form a graph. The graph is then split into desired number of clusters using a min-cut graph partitioning algorithm (Karypis, 2003).

#### 7.3.3.7 BOK

This algorithm may be the simplest one and is also known as The Best Clustering Algorithm (Gionis et al., 2007; Bertolacci & Wirth, 2007). It arbitrarily picks one clustering from the input clusterings as the consensus in turn, and then calculates its Rand distance (Rand, 1979) between the consensus and the rest of input clusterings. The consensus which has the minimum average distance will be considered as the best of clustering (BOK) (Filkov & Skiena, 2004a; Goder & Filkov, 2008).

Above seven consensus clustering methods and their abbreviations used in the following sections are summarized and shown in Table 7-2.

| Consensus Clustering Methods | Software Tools | Code in Tables and Figures |
|---|---|---|
| Majority Rule | *Agglomerative* method with criterion function of *single linkage* in CLUTO | MR |
| Average Linkage | *Agglomerative* method with criterion function of *upgma* in CLUTO | AL |
| Furthest Linkage | *Agglomerative* method with criterion function of *furthest linkage* in CLUTO | FL |
| CC-Pivot | Coded by Perl script | CCP |
| Direct | *Direct* method in CLUTO | DR |
| Graph based | *Graph* method in CLUTO | GB |
| BOK | Coded by Perl script | BOK |

Table 7-2 Summary of consensus clustering methods

### 7.3.4 Determining the Number of Clusters

The number of clusters is normally determined by the consensus algorithm (Gionis et al., 2007; Bertolacci & Wirth, 2007). In our study, in order to compare the performance of consensus clustering with our previous study, we set the same number of clusters, which is 500, 600, 700, 800, 900, and 1000 clusters for each consensus clustering run for both datasets. For the Majority Rule and the CC-Pivot methods, the number of clusters is determined by the predefined threshold. Within these two consensus clustering methods, the CC-Pivot method is extremely sensitive to the number of clusters with regard to the initial threshold setting and the random pivot objects selecting. To work such problem out, we chose the closest number of clusters from over 30 clustering runs by adjusting the threshold.

## 7.4 Results and Analysis

Evaluating the clustering results is a critical issue of consensus clustering. There are extensive evaluation measures in the literature. If a clustering method offers better performance than others over many evaluation measures, then we can claim confidently such clustering method should be the best for a certain type of application. The four evaluation measures used in our previous study are also employed in this experiment as discussed in Chapter 4, entropy, entropy based on cluster size, F-Measure and QCI.

### 7.4.1 Evaluation of the MDDR Dataset

#### 7.4.1.1 Evaluation using the F-Measure on the MDDR Dataset

Our previous study carried out the comparison of performances between nine different clustering methods. However in this section, the performances of seven consensus clustering methods are compared with the single best clustering based on a certain evaluation from our previous study in Chapter 6; for example if evaluating the MDDR 500-cluster results using the F-Measure, single best clustering would be method that gave the best F-Measure results with the 500 MDDR clusters. In addition to the single best clustering, the Ward's method is the clustering procedure of choice in most Chemoinformatics applications and software packages. Hence, the performance of Ward's method in our previous study is also included in the comparison with consensus clustering.

The consensus clustering result for unweighted consensus similarity using the MDDR dataset was evaluated by the F-Measure and its evaluation is shown in Figure 7-3. The performances evaluated by the F-Measure of seven consensus clustering methods are significantly split into two groups, the Majority Rule method gives the consistently worst results over all numbers of clusters, and the rest of consensus clustering methods are in the leading group. The reason for the Majority Rule method offering the significantly worst results is because of its worse clustering generating many active singletons (Table 7-3) and few

big clusters, and this will lead to worse evaluation on the F-Measure.

| Consensus Clustering Methods | # clusters | Unweighted datasets | | Weighted datasets | |
|---|---|---|---|---|---|
| | | MDDR | IDAlert | MDDR | IDAlert |
| Majority Rule | 500 | 31 | 27 | 26 | 22 |
| | 600 | 25 | 31 | 28 | 22 |
| | 700 | 29 | 31 | 34 | 36 |
| | 800 | 44 | 45 | 44 | 39 |
| | 900 | 44 | 50 | 47 | 45 |
| | 1000 | 55 | 60 | 56 | 53 |

Table 7-3 The number of active singletons in the consensus clustering results of the Majority Rule method for unweighted and weighted datasets.

In the leading group, Average Linkage, K-Means based, and Furthest Linkage have the best performance with different numbers of clusters. However, no single consensus clustering method yields the consistently best results over all numbers of clusters. In comparison with the single best clustering (SBC) method from our previous study, the performance of single best consensus clustering methods provides superior results over 800, 900, and 1000 clusters; while over 500, 600, and 700 clusters, the single best conventional clustering method has better performance than the single best consensus clustering methods. There is thus no consistent benefit gained in comparison with the single best conventional clustering. However, in comparison with the most commonly used Ward's (WD) clustering, 5 in 6 single best consensus clusterings have shown better performance.

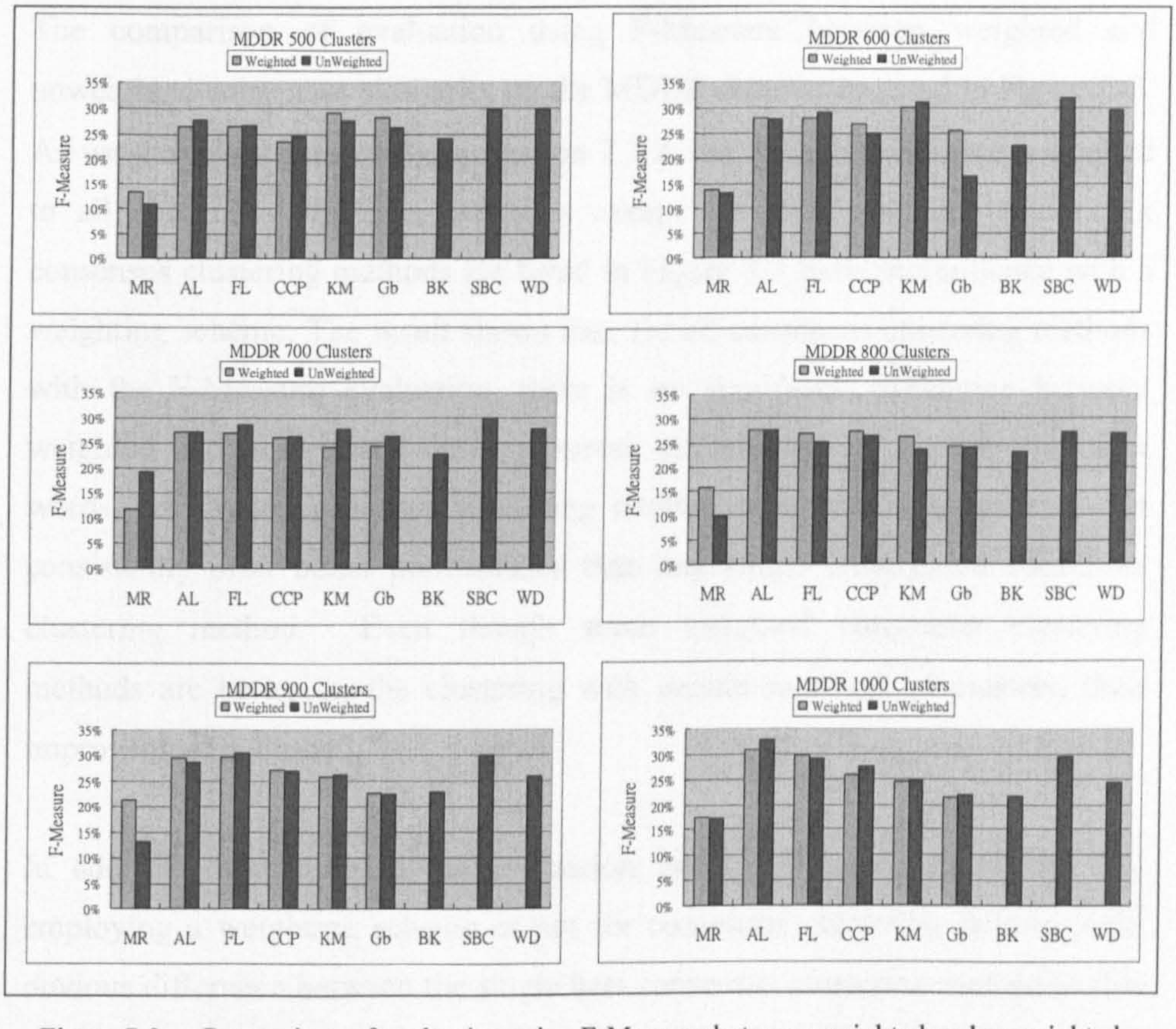



Figure 7-3 Comparison of evaluation using F-Measure between weighted and unweighted MDDR datasets

The performance evaluated by the F-Measure on weighted consensus similarity using the MDDR dataset is shown in Figure 7-3. Similar to the result of the unweighted MDDR dataset, the Majority Rule method provides consistently worst performances across all numbers of clusters. K-Means based, Average Linkage, and Furthest Linkage methods in the leading group still yield the best results with different numbers of clusters. Again, in comparison with the single best clustering method (SBC), consensus clustering methods tend to have better performance than the single best clustering method when there are large numbers of clusters e.g. 800 and 1000 clusters. Similarly, in comparison with the most commonly used Ward's (WD) clustering, 4 in 6 single best consensus clusterings have shown better performance.

The comparison of evaluation using F-Measure between weighted and unweighted consensus similarity on the MDDR datasets is listed in Figure 7-3. As we described previously in Section 7.3.2, the weighting scheme is applied to all consensus clustering methods except the BOK method. Hence, six consensus clustering methods are listed in Figure 7-3 have performance with a weighting scheme. The result shows that, for all consensus clustering methods with the F-Measure evaluation, there is no significant difference between weighted and unweighted datasets across all numbers of clusters. In other words, no single consensus clustering method with weighting scheme can consistently offer better performance than any single unweighted consensus clustering method. Even though some weighted consensus clustering methods are better on the clustering with certain numbers of clusters, their improvement is limited.

In addition, according to the evaluation using F-Measure in Figure 7-3, employing a weighting scheme or not for consensus clustering fails to yield obvious difference between the single best consensus clustering method in this study and the single best clustering method in our previous study. No notable benefit is gained from consensus clustering under the F-Measure evaluation. However, several cases where the best consensus method is better than the standard Ward's method.

### 7.4.1.2 Evaluation using the QCI on the MDDR dataset

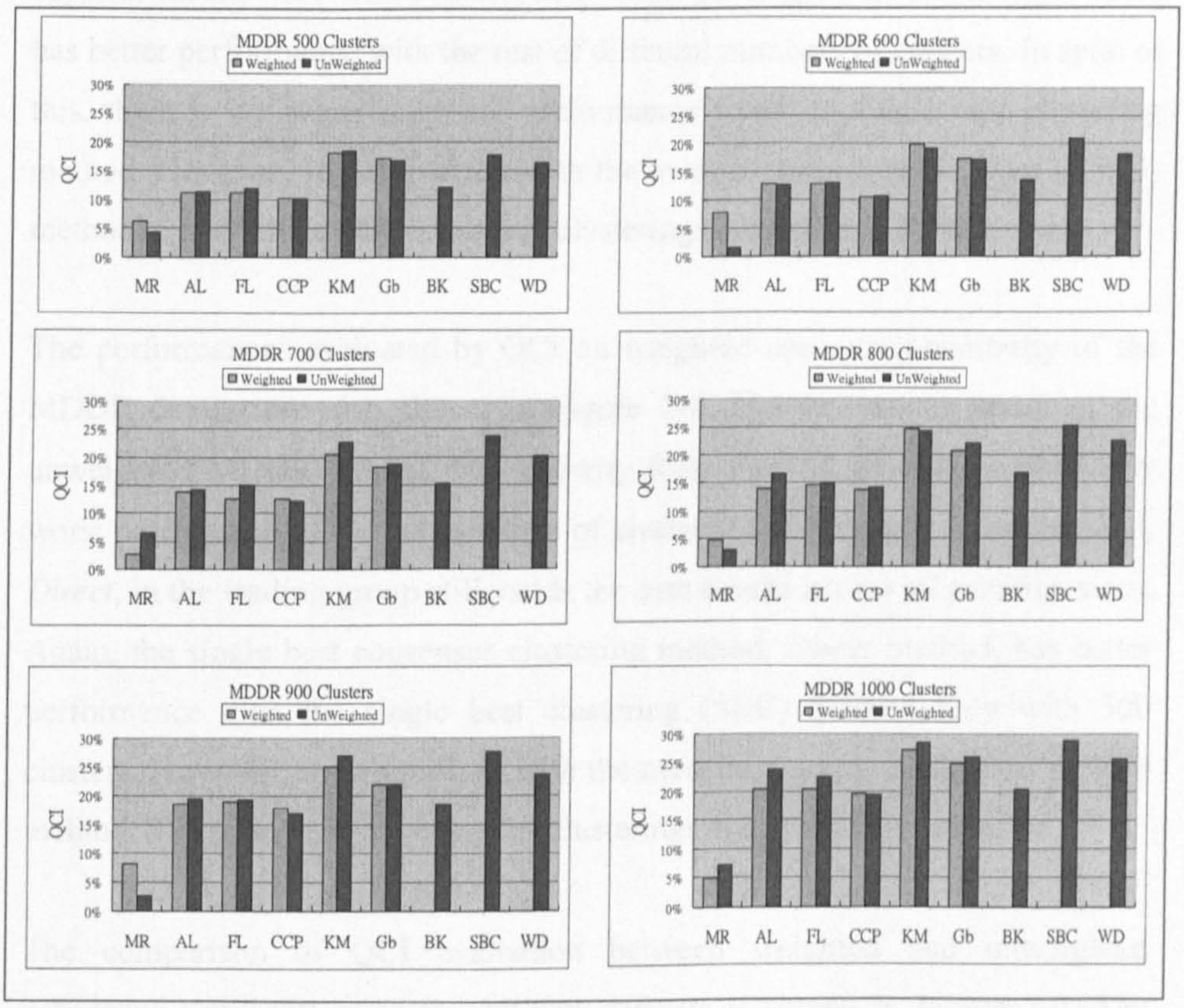



Figure 7-4 Comparison of evaluation using QCI between weighted and unweighted MDDR datasets

Figure 7-4 presents the results of QCI evaluation of consensus clustering on unweighted consensus similarity using the MDDR dataset, and includes the performance of the single best conventional clustering method. The performances evaluated by QCI of seven consensus clustering methods show that K-Means based and Graph based methods are consistently in the leading group and keep offering better QCI values. Similarly, the Majority Rule method consistently has worst results over all numbers of clusters due to its worse clustering containing many singletons and few big clusters. According to the equation of QCI discussed in Chapter 4, as the number of active singletons increased, the value of QCI decreased. In the leading group, K-Means based

method has the noticeably and consistently best performance with all numbers of clusters. In comparison with the single best clustering (SBC) method from our previous study, the performance of consensus clustering methods provide superior results only with 500 clusters, while the single best clustering method has better performance with the rest of different numbers of clusters. In spite of this, there is no overall superior performance found to single best clustering method. However, in comparison with the most commonly used Ward's (WD) method, 5 in 6 single best consensus clusterings are found to be superior.

The performances evaluated by QCI on weighted consensus similarity of the MDDR dataset are also shown in Figure 7-4. Identical to the result of the unweighted MDDR dataset, the Majority Rule method provides consistently worst performances over all numbers of clusters. The K-Means based method, *Direct*, in the leading group still yields the best results across all partition sizes. Again, the single best consensus clustering method, *Direct* method, has better performance than the single best clustering (SBC) method only with 500 clusters. However, in comparison with the most commonly used Ward's (WD) method, 5 in 6 single best consensus clusterings are found to be superior.

The comparison of QCI evaluation between weighted and unweighted consensus similarity using the MDDR datasets is shown in Figure 7-4. The result shows that for all consensus clustering methods with QCI evaluation, there is no significant difference between weighted and unweighted datasets over all numbers of clusters, that is, no single consensus clustering method with weighting scheme can consistently offer better performance. Even though some weighted consensus clustering method is better on the clustering with certain number of clusters, its improvement is limited.

In addition, according to the QCI evaluation in Figures 7-4, notwithstanding the consensus clustering method is with or without weighting scheme, there is no obvious difference between the single best consensus clustering method in this study and the single best clustering method in our previous study. No significant benefit is gained from consensus clustering with the QCI evaluation.

#### 7.4.1.3 Evaluation using Entropy and Entropy based on cluster size on the MDDR dataset

##### 7.4.1.3.1 Evaluation using Entropy on the MDDR dataset

Shannon Entropy was employed to evaluate the distribution of active compounds over all clusters, the smaller Entropy value the better the performance. Figure 7-5 shows the Entropy evaluation of six consensus clustering methods with weighted scheme, seven consensus clustering methods with unweighted scheme, and also the single best clustering and the Ward's method from our previous study.

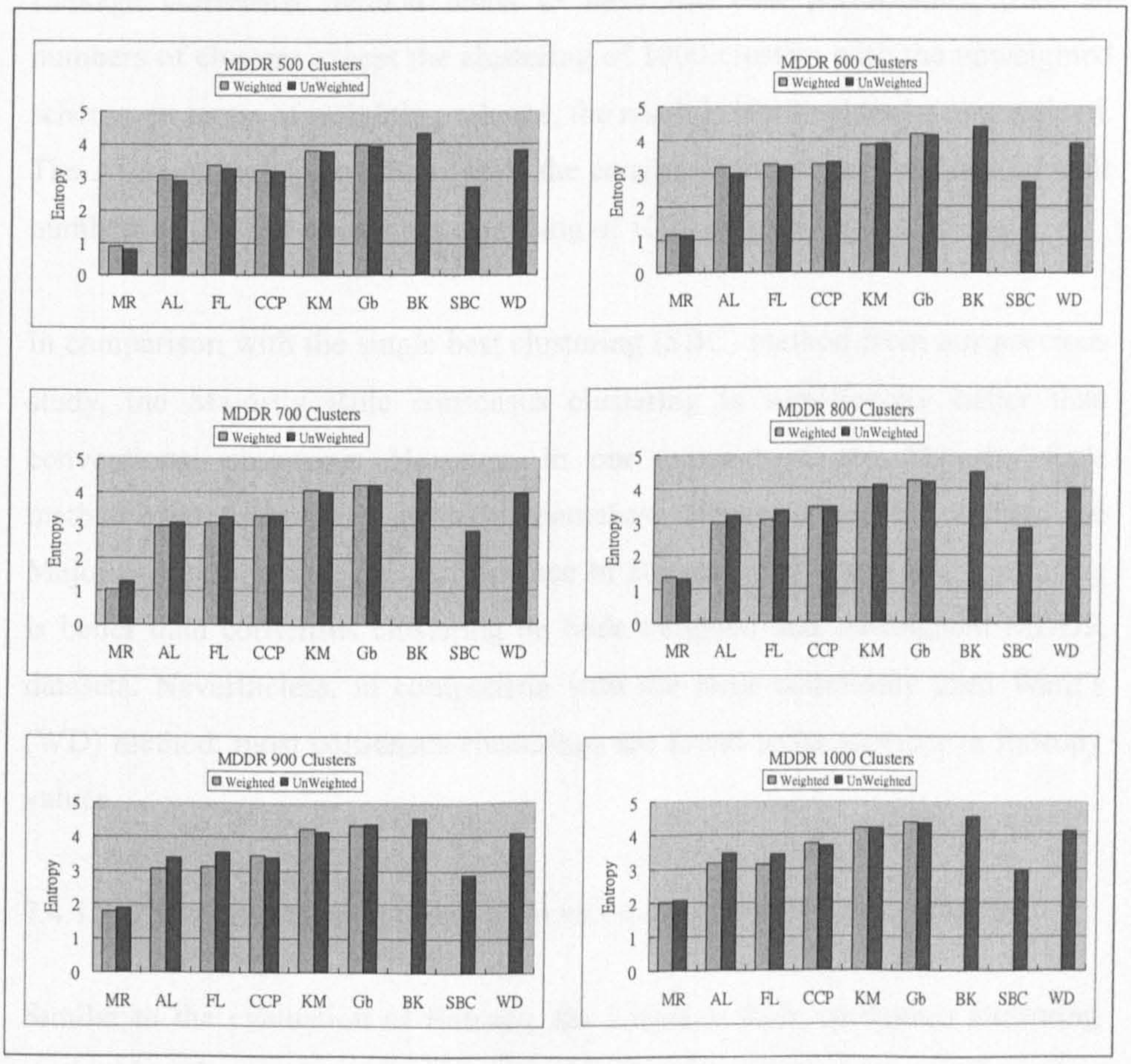



Figure 7-5 Comparison of evaluation using Shannon Entropy between weighted and unweighted MDDR datasets

Obviously, the Majority Rule consensus clustering method has the consistently and noticeably best (smallest) Entropy values than any others on either weighted or unweighted using the MDDR dataset. However, with visual inspection on the clustering results of the Majority Rule, there are many singletons, many small clusters, and few big clusters. Most of the active compounds have been assigned to these few big clusters, and this leads to the smaller Entropy value, since Shannon Entropy, in essence, focuses on the distribution of active compounds without taking the number of singletons into account.

However, evaluating without the abnormal Majority Rule method, the Average Linkage consensus method tends to have the best performance over all numbers of clusters except the clustering of 1000 clusters with the unweighted scheme. In terms of weighting scheme, the result is identical to the unweighted. The Average Linkage method yields the consistently best performance over all numbers of clusters except the clustering of 1000 clusters.

In comparison with the single best clustering (SBC) method from our previous study, the Majority Rule consensus clustering is significantly better than conventional clustering. However, in our experiment, the Majority Rule method offered abnormal clustering somehow. Hence, comparing without the Majority Rule method, the performance of conventional single best clustering is better than consensus clustering on both weighted and unweighted MDDR datasets. Nevertheless, in comparison with the most commonly used Ward's (WD) method, most consensus clusterings are found to be superior in Entropy values.

#### 7.4.1.3.2 Evaluation using Entropy Based on Cluster Size on the MDDR Dataset

Similar to the evaluation of Entropy, the Majority Rule consensus clustering method has the consistently and noticeably best (smallest) Entropy values than any others on either the weighted or unweighted MDDR dataset. However, with visual inspection of the clustering results of the Majority Rule, there are

many singletons and small clusters. Since Entropy based on cluster size measures the distribution of cluster size, multiples clusters with similar size will definitely give a better result.

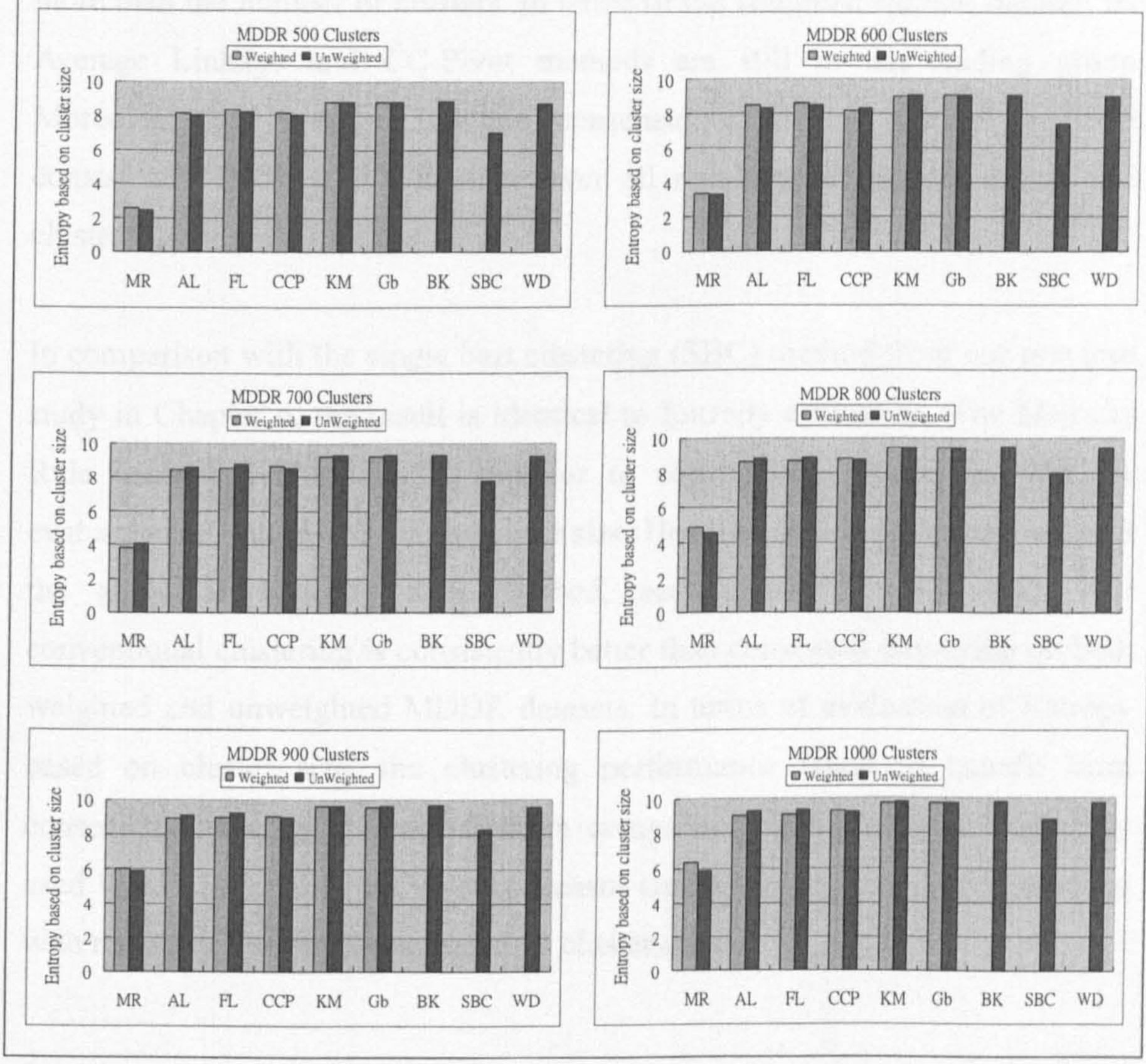



Figure 7-6 Comparison of evaluation using Entropy based on cluster size between weighted and unweighted MDDR datasets

After discarding the abnormal Majority Rule method, the leading group of CC-Pivot and Average Linkage methods have better results. Nevertheless, the CC-Pivot consensus method provides the consistently best performance over all numbers of clusters with the unweighted MDDR dataset. As we described previously, the CC-Pivot method is extremely sensitive to the number of clusters with its initial setting. For example, in our experiment, doing CC-Pivot consensus clustering with unweighted consensus similarity matrix generated by

500 clusters of different clusterings will eventually produce 541 clusters. In theory, the additional 41 clusters will more or less decrease the value of Entropy based on cluster size. However, as shown in Figure 7-6, as the number of clusters increase, the value of Entropy based on cluster size does not always decrease. The cluster size actually depends on the clustering algorithm itself more than the number of clusters. In terms of the weighted MDDR dataset, the Average Linkage and CC-Pivot methods are still in the leading group. Moreover the Average Linkage consensus clustering method provides consistently the best performance over all numbers of clusters except 600 clusters.

In comparison with the single best clustering (SBC) method from our previous study in Chapter 6, the result is identical to Entropy evaluation. The Majority Rule consensus clustering is superior to conventional clustering with the evaluation of Entropy based on cluster size. However, when comparing without the abnormal Majority Rule method, performance of the single best conventional clustering is consistently better than consensus clustering on both weighted and unweighted MDDR datasets. In terms of evaluation of Entropy based on cluster size, the clustering performance failed to benefit from consensus clustering. Nevertheless, in comparison with the most commonly used Ward's (WD) method, most consensus clusterings are found to be superior with regard to Entropy value based on cluster size.

#### 7.4.1.3.3 Summary

Table 7-4 summarizes the performance evaluated by four different criteria, and the comparison with the single best clustering and Ward's methods from our previous study in Chapter 6.

| Evaluations | MDDR dataset | | |
|---|---|---|---|
| | Performance of consensus clusterings | Comparison with the single best clustering | Comparison with Ward's method |
| F-Measure | No consistently better method | Except 500, 600 and 700 clusters of clusterings, the single best consensus clusterings is better | 5 in 6 the best consensus clustering method is better |
| QCI | K-Means based method is the best | The consensus clustering is better only with 500 clusters of clustering | 6 in 6 the best consensus clustering method is better |
| Shannon Entropy | 5 in 6 clusterings, AL method is the best (discard MR) | Consensus clustering methods are worse | 6 in 6 the best consensus clustering method is better |
| Entropy based on cluster size | No consistently better method (discard MR) | | |

Table 7-4 Summary of the performance of consensus clusterings and the comparison with previous study using the MDDR dataset

## 7.4.2 Evaluation of the IDAlert Dataset

### 7.4.2.1 Evaluation using F-Measure on the IDAlert dataset

The consensus clustering results of the unweighted IDAlert dataset were evaluated by the F-Measure as shown in Figure 7-7, and it also includes the single best conventional clustering method from our previous study in order to compare their performance. Obviously, the Majority Rule consensus method has the consistently worst results over all numbers of clusters. The reason is that it yields worse clustering with many singletons (Table 7-3) and a few big clusters as we discussed in Section 7.4.1.1. In addition, the Average Linkage method has the best F-Measure values with all numbers of clusters except 500 clusters; the K-Means based method also has better performance with most of the numbers of clusters.

However no single consensus clustering method is found to be effective providing consistently best results over all numbers of clusters. In comparison with the single best clustering method from our previous study, the performance of single best consensus clustering method, Average Linkage

method, provides superior results over all numbers of clusters except 500 and 700 clusters to the single best conventional clustering method. Nevertheless, the single best and some consensus clusterings consistently offer better performance than the Ward's method from our previous study.

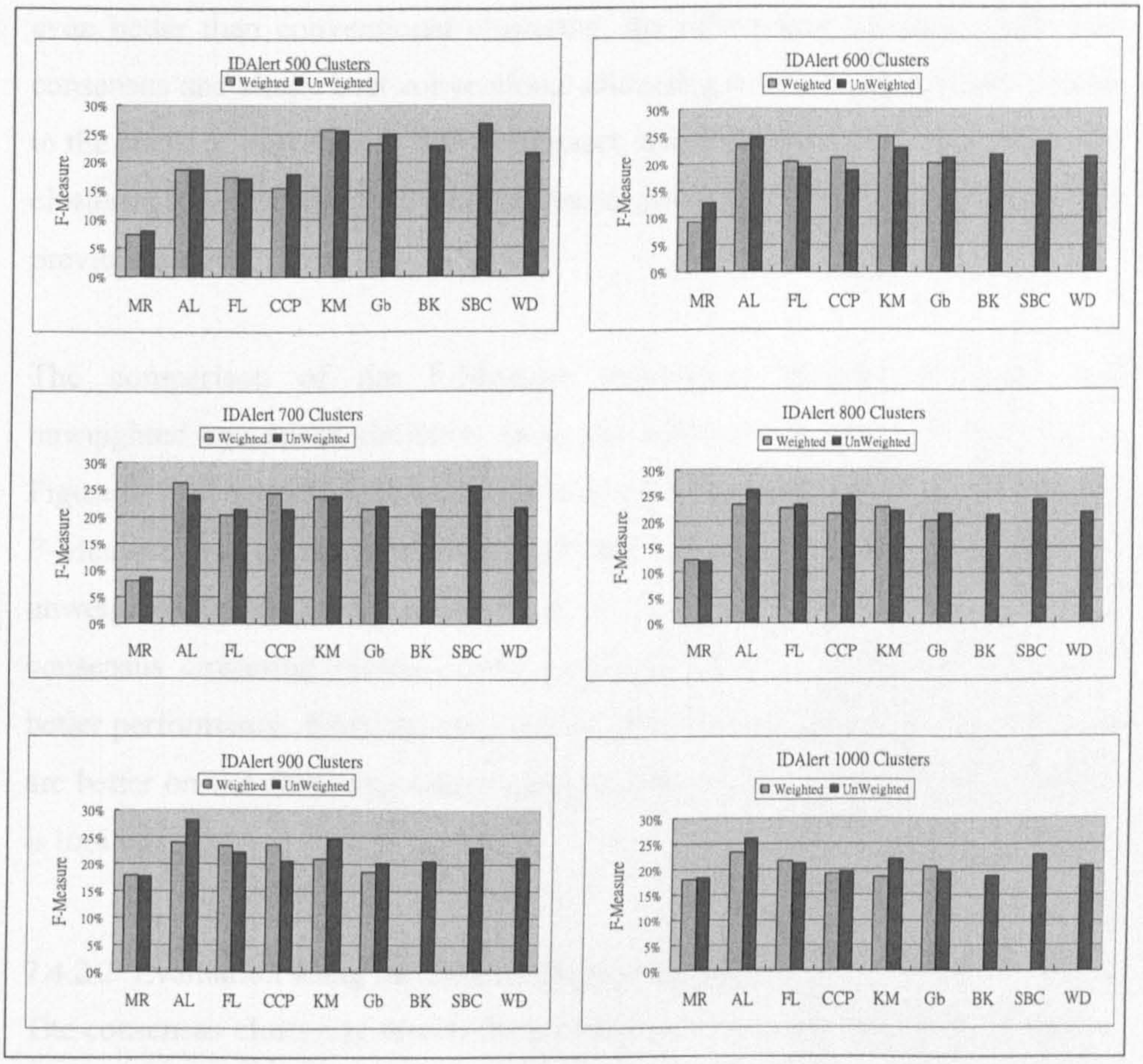



Figure 7-7 Comparison of evaluation using F-Measure between weighted and unweighted IDAlert datasets

The performances evaluated by the F-Measure on weighted consensus similarity using the IDAlert dataset are also listed in Figure 7-7. Similar to the result with the unweighted IDAlert dataset, the Majority Rule consensus clustering method provides consistently worst performances over all numbers of clusters due to its poor clustering with a large number of singletons. Similar to the unweighted IDAlert dataset, in addition to the Average Linkage method,

no others consistently remain in the leading group. The Average Linkage still offers the best results in most of the numbers of clusters. Again, the single best consensus clustering method, the Average Linkage, produces better performance than the single best clustering method over all numbers of clusters except 500 and 800 clusters. Although, with the F-Measure evaluation, no significant evidence shows consensus clustering improved the performance or even better than conventional clustering, the differences between single best consensus and single best conventional clustering methods are limited. Similar to the result of unweighted IDAlert dataset, the single best and some consensus clusterings have consistently better results than the Ward's method from our previous study.

The comparison of the F-Measure evaluation between weighted and unweighted consensus similarity using the IDAlert datasets is also shown in Figure 7-7. The result shows that for all consensus clustering methods with the F-Measure evaluation, there is no significant difference between weighted and unweighted datasets over all numbers of clusters. In other words, no single consensus clustering method with weighting scheme can consistently yield better performance. Even though some weighted consensus clustering methods are better on the clustering with certain number of clusters, their improvement is limited.

#### 7.4.2.2 Evaluation using the QCI on the IDAlert dataset

The consensus clustering results for unweighted consensus similarity using the IDAlert dataset were evaluated by QCI and the evaluation is shown in Figure 7.8. The performance evaluated by QCI of seven consensus clustering methods shows that K-Means based and Graph based methods consistently offer better and similar QCI values, and this is similar to the results on the MDDR dataset. The only difference is, in the leading group, the Graph based method yields the consistently best performance with all numbers of clusters except 1000 clusters. Again, the Majority Rule method consistently has the worst results over all numbers of clusters due to its poor clustering. The single best clustering method from our previous study in Chapter 6 provides superior results with all

numbers of clusters. However the single best consensus clustering method, i.e. Graph based method, has closer results to it. Nevertheless, the single best and some consensus clusterings consistently offer better performance than the Ward's method from our previous study.

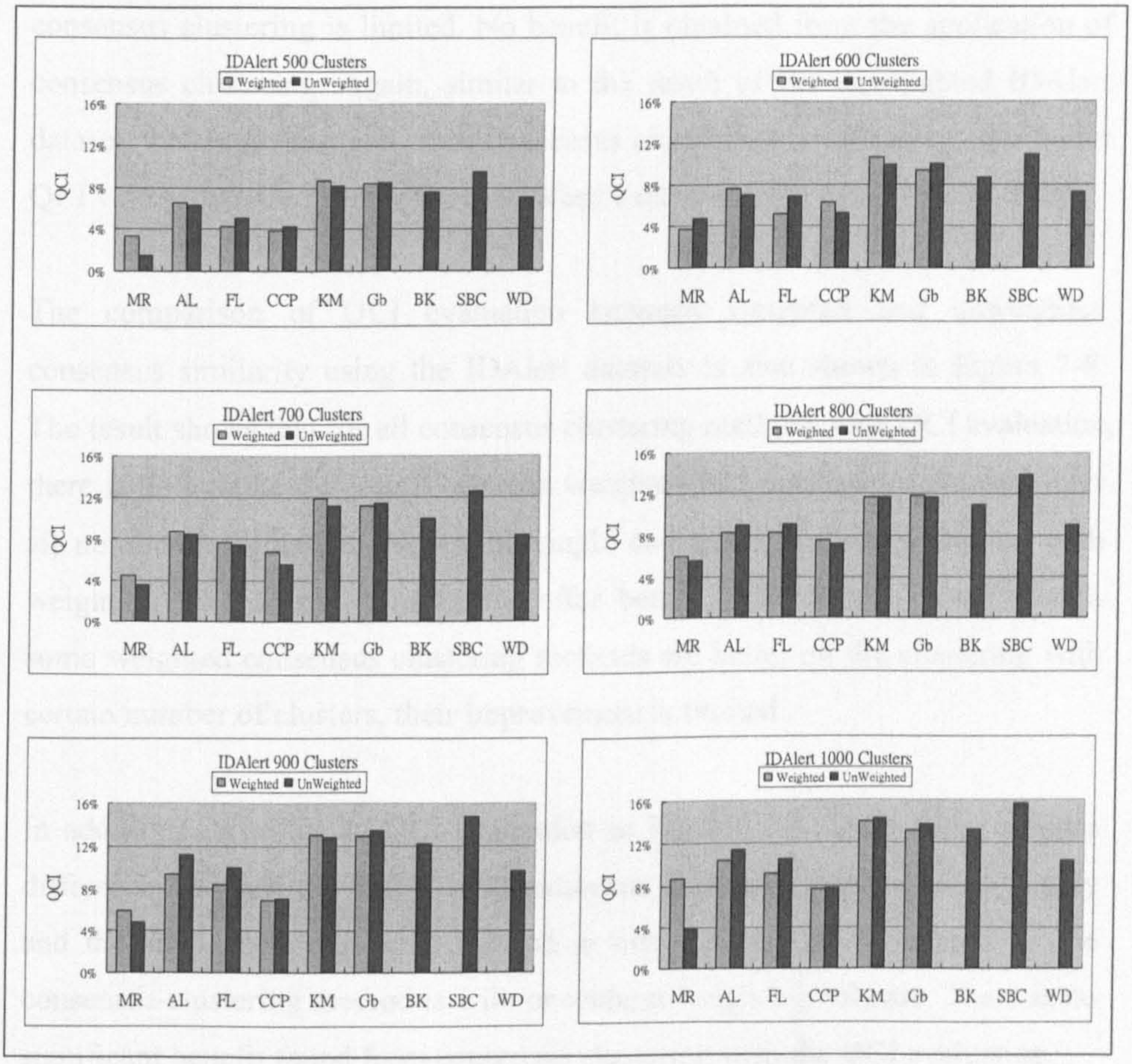



Figure 7-8 Comparison of evaluation using QCI between weighted and unweighted IDAlert datasets

Figure 7-8 also presents the performances evaluated by QCI on weighted consensus similarity using the IDAlert dataset. Similar to the result of the unweighted IDAlert dataset, the K-Means based and Graph based methods remain in the leading group offering better results. However, with the weighting scheme, the K-Means method yields the best performance instead of the Graph based method over all numbers of clusters except 800 clusters.

Similarly, the Majority Rule method provides consistently poor performances over all numbers of clusters because of its worse clustering. Again, no single best consensus clustering methods have consistently better performance than the single best clustering method for all numbers of clusters. The difference of performance between the single best conventional clustering and single best consensus clustering is limited. No benefit is obtained from the application of consensus clustering. Again, similar to the result of the unweighted IDAlert dataset, the single best and some consensus clusterings consistently offer better QCI values than the commonly used Ward's method from our previous study.

The comparison of QCI evaluation between weighted and unweighted consensus similarity using the IDAlert datasets is also shown in Figure 7-8. The result shows that for all consensus clustering methods with QCI evaluation, there is no notable difference between weighted and unweighted datasets over all numbers of clusters, that is, no single consensus clustering method with weighting scheme can consistently offer better performance. Even though some weighted consensus clustering methods are better on the clustering with certain number of clusters, their improvement is limited.

In addition, according to QCI evaluation in Figures 7-8, there is no obvious difference between the single best consensus clustering method in this study and the single best clustering method in our previous study in spite of the consensus clustering method is with or without weighting scheme. There is no significant benefit found from consensus clustering with the QCI evaluation.

#### 7.4.2.3 Evaluation using Entropy and Entropy based on cluster size on the IDAlert dataset

##### 7.4.2.3.1 Evaluation using Entropy on the IDAlert dataset

Shannon Entropy was employed to evaluate the distribution of active compounds over all clusters, the smaller Entropy value the better performance. Figure 7-9 shows the Entropy evaluation of seven consensus clustering methods with the unweighted scheme, and six methods with the weighted scheme. Figure 7-9 also represents the performance of the single best conventional clustering method and the performance of Ward's method.

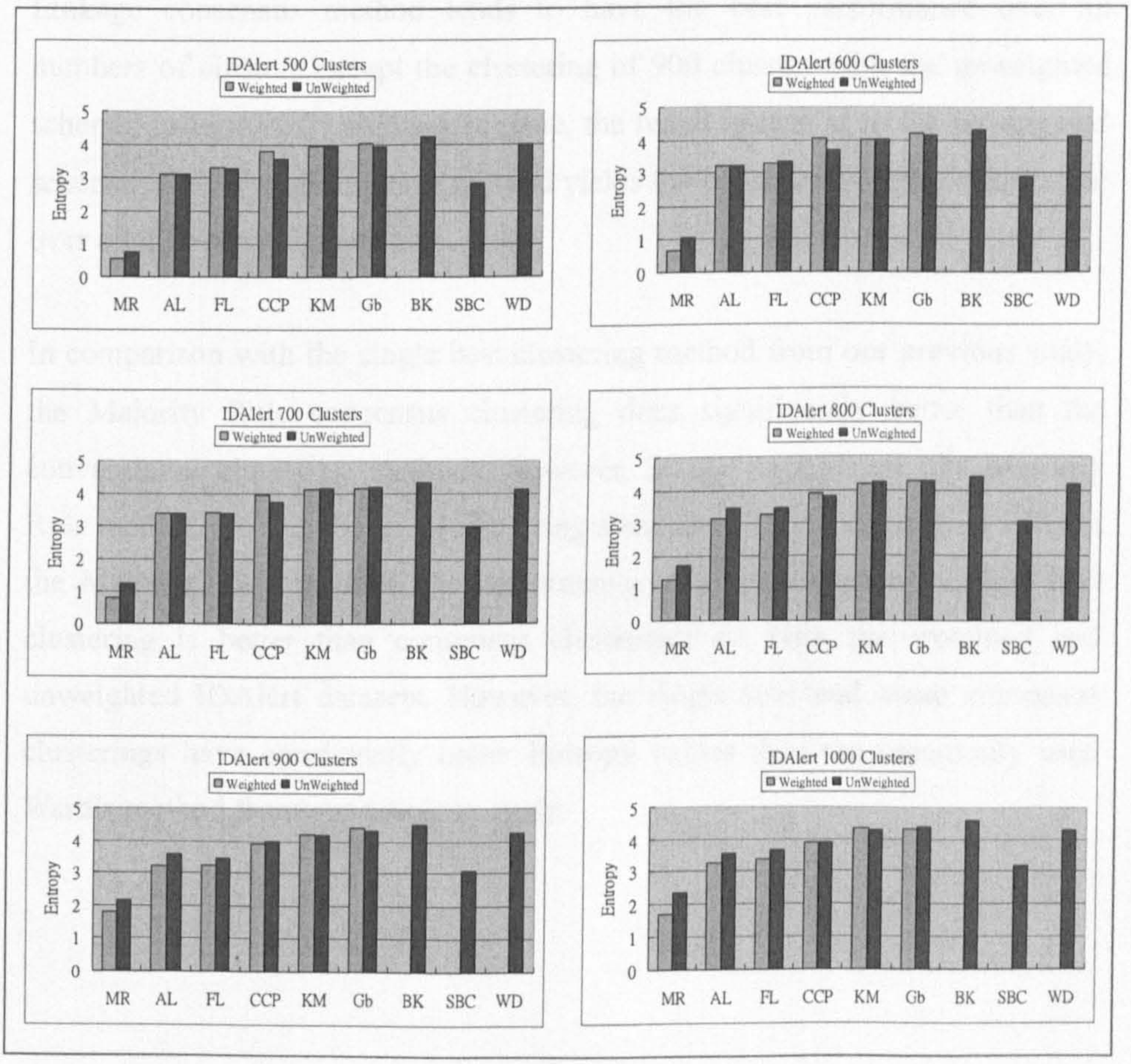



Figure 7-9 Comparison of evaluation using Shannon Entropy between weighted and unweighted IDAlert datasets

Similar to the result of the MDDR dataset, the Majority Rule consensus clustering method has the consistently and noticeably best (smallest) Entropy values for either the weighted or unweighted IDAlert dataset. However, with visual inspection on the clustering results of the Majority Rule, there are many singletons, many small clusters, and few big clusters. The active compounds are largely assigned to these few big clusters, and this leads to the smaller Entropy value. This is because Shannon Entropy, in essence, focuses on the distribution of active compounds without taking the number of singletons into account.

However, evaluating without the abnormal Majority Rule method, the Average Linkage consensus method tends to have the best performance over all numbers of clusters except the clustering of 900 clusters with the unweighted scheme. In terms of weighting scheme, the result is similar to the unweighted scheme. The Average Linkage method yields the consistently best performance over all numbers of clusters.

In comparison with the single best clustering method from our previous study, the Majority Rule consensus clustering does significantly better than the conventional clustering methods; however, in our experiment, the Majority Rule method offered abnormal clustering somehow. Hence, comparing without the Majority Rule method, the performance of the conventional single best clustering is better than consensus clusterings on both the weighted and unweighted IDAlert datasets. However, the single best and some consensus clusterings have consistently better Entropy values than the commonly used Ward's method from our previous study.

#### 7.4.2.3.2 Evaluation using Entropy Based on Cluster Size on the IDAlert Dataset

The evaluation of Entropy based on cluster size for the IDAlert dataset is shown in Figure 7-10. Similar to the evaluation of Entropy, the Majority Rule consensus clustering method has the consistently and noticeably best (smallest) Entropy values than any others on either the weighted or unweighted IDAlert dataset. However, with visual inspection of the clustering results of the Majority Rule, there are large numbers of singletons and small clusters, and this leads to the consistency of cluster size in the form of multiple small clusters, since Entropy based on cluster size is an evaluation to measure the distribution of cluster size, clusters with similar small size will definitely yield better entropy values.

Under the situation of discarding the abnormal Majority Rule method, the leading group of the CC-Pivot, Furthest Linkage and Average Linkage methods have better results in the unweighted IDAlert dataset. Identical to the result of MDDR, the CC-Pivot consensus method provides the consistently best performance over all numbers of clusters. The CC-Pivot method is extremely sensitive to the number of clusters with its initial setting as we described previously. In terms of the weighted IDAlert dataset, the Average Linkage, Furthest Linkage and CC-Pivot methods are still in the leading group. But no single consensus clustering method provides the consistently best performance over all numbers of clusters.

In comparison with the single best clustering method from our previous study, the result is identical to Entropy evaluation. The Majority Rule consensus clustering is superior to conventional clustering with the evaluation of Entropy based on cluster size; however, comparing without the abnormal Majority Rule method, performance of the single best conventional clustering is consistently better than consensus clustering on both weighted and unweighted IDAlert datasets. Nevertheless the single best and some consensus clusterings have consistently better Entropy values based on cluster size than the commonly used Ward's method from our previous study. In terms of evaluation of Entropy

based on cluster size, the clustering performance failed to benefit from consensus clustering.

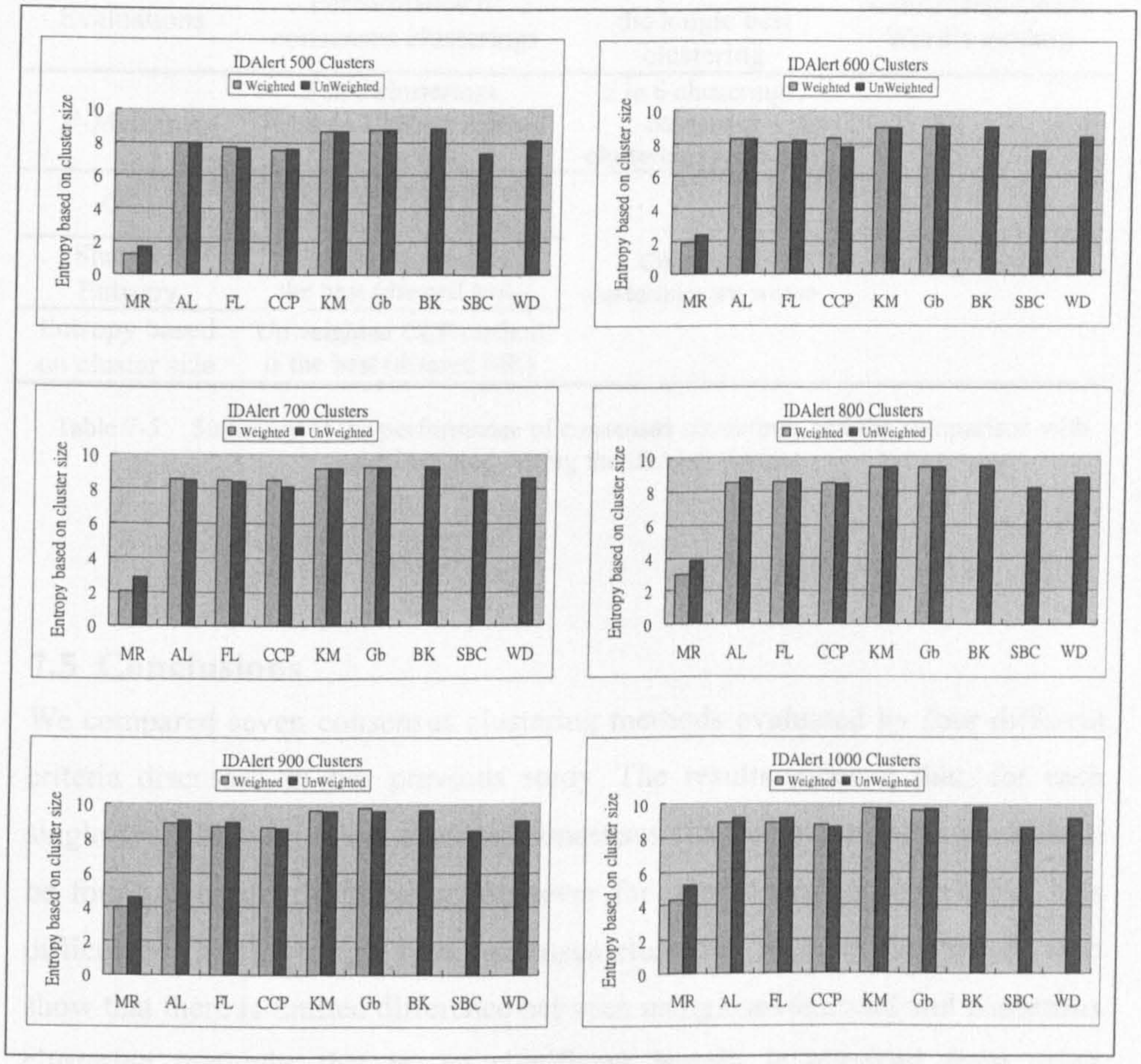



Figure 7-10 Comparison of evaluation using Entropy based on cluster size between weighted and unweighted IDAlert datasets

### 7.4.2.3.3 Summary

Table 7-5 summarizes the performance evaluated by four different criteria, and the comparison with the single best clustering and Ward's method from our previous study in Chapter 6.

| Evaluations | IDAlert dataset | | |
|---|---|---|---|
| | Performance of consensus clusterings | Comparison with the single best clustering | Comparison with Ward's method |
| F-Measure | 5 in 6 clusterings, Average Linkage method is the best | 5 in 6 clusterings, consensus clusterings are better | 6 in 6 the best consensus clustering method is better |
| QCI | 5 in 6 clusterings, Graph based method is the best | | |
| Shannon Entropy | Weighted AL method is the best (discard MR) | Consensus clusterings are worse | |
| Entropy based on cluster size | Unweighted CCP method is the best (discard MR) | | |

Table 7-5 Summary of the performance of consensus clusterings and the comparison with previous study using the IDAlert dataset

## 7.5 Conclusions

We compared seven consensus clustering methods evaluated by four different criteria discussed in our previous study. The results indicate that, for each single evaluation criterion, a certain consensus clustering method is possible to be found consistently effective; however for overall evaluation criteria, it is difficult to find the single best consensus clustering method. Our results also show that there is limited difference between using conventional and consensus clustering methods, that is, no significant benefit is obtained from using consensus clustering in our study. In terms of weighted scheme applied to the consensus similarity matrix, the improvement is also limited. It is suggested that other weighting schemes might be more successful. The results in our study still show that consensus clustering methods are dataset dependent as reported in the literature; no single best clustering method can be applied to all applications and all fields.

# Chapter 8 : Conclusions and Future Work

## 8.1 Conclusions

The work described in this thesis has discussed the application of clustering on 2D chemical structures.

The initial study of this thesis shows the effect of standardization procedures on chemical clustering and similarity searching. No standardization method was found that provides consistently superior or worse performance in both the MDDR and IDAlert datasets at the $\alpha$=0.01 level of statistical significance, moreover we found statistically significant at the $\alpha$=0.05 level on the tests based on the results of similarity searching only on the IDAlert dataset. We hence conclude that there is no obvious performance benefit that is likely to be obtained from the use of any particular standardization method. In a later extensive study, we employed more diverse clustering methods, but the performance of standardization methods is similar to the previous study. Overall, standardization procedures can improve the clustering performance more and less, but no method was found to be consistently effective.

Next, the comparison of nine clustering methods showed that, for the ECFP_4 chemical representation considered in this work, no consistent performance benefit is likely to be obtained from the use of any particular clustering method using the chosen evaluation methods. One possible reason to explain the inconsistent performance is the diverse evaluation criteria, for example *CLUTO-Direct* method has consistently better F-Measure and QCI results but worse Entropy based on cluster size. That is, the clustering results of CLUTO-Direct did not yield clusters with equal size but obtained good F-Measure and QCI values. Can we conclude a clustering with equal size of clusters is a good partition, or a clustering without equal size of clusters a worse partition? To sum up, the result reveals that it is difficult for a clustering

method to satisfy all our evaluation criteria.

Finally, the performance of seven consensus clustering methods evaluated by four different criteria shows that evaluating using only one single criterion, a certain consensus clustering method is possibly to be found consistently effective; whereas evaluating using overall evaluation criteria, it is difficult to find the single best consensus clustering method. Our results also show that there is limited difference between using conventional and consensus clustering methods. That is, no significant benefit is obtained from using consensus clustering in our study. In terms of weighted scheme applied to the consensus similarity matrix, the improvement is also limited. The results in our study still show that consensus clustering methods are dataset dependent as reported in the literature. Although no single best clustering method can be applied to all applications and all fields, consensus clustering still offers more confidence to the result.

Quantitative evaluation of clustering methods is not simple. The applicability of different evaluation measures is varied in essence. The use of several different evaluation measures in this thesis is expected to get some consistency to make results believable. Shannon Entropy based on cluster size takes only cluster size into account and ignores the number of actives, whereas Shannon Entropy considers only the distribution of actives and neglects the cluster sizes. However, both evaluation measures are not suitable to the clustering outcome containing one very large cluster and many small clusters. Such an abnormal clustering usually leads an extremely low Entropy value. Conversely, they are suitable to the clustering methods, e.g. Repeated Bisection method, which tend to generate similar cluster sizes. The combination of Shannon Entropy and Entropy by cluster size might not be used to fit above abnormal clustering, because Entropy, in essence, is an index to measure the distribution of a variable (e.g. actives or cluster sizes). If a clustering generates only one extremely large cluster, it is naturally not the case of distribution. Hence, a new or the other index to detect such abnormal situation may be needed

Moreover, F-Measure considers only the maximum F-value in a given active class rather than the average of F-value; for the case of abnormal clustering discussed above, it may not reflect the true quality of clustering. The evaluation using probability of correct prediction takes both actives and cluster sizes into account; however it is not applicable to the large datasets. Finally, the QCI takes both actives and inactives into account. In addition, for some cases of abnormal clustering, the number of singletons is also considered in the calculation of QCI. Hence, considering five evaluation measures used in this thesis, it is suggested that QCI is the evaluation measure of choice for the application of clustering on chemical structures.

Account for the upper and lower bounds of the evaluation measures. The boundary of evaluation using probability of correct prediction is [0, 1]. The essence of Entropy-based measures is to evaluate the distribution of a given variable, e.g. actives or cluster sizes. Hence, the Entropy value naturally depends on the partition size, i.e. scope of distribution, as well as the number of actives (for evaluation using Shannon Entropy) or dataset size (for evaluation using Entropy based on cluster size). In the worst case, a given variable is equally distributed over all clusters; the worst possible Entropy value could be varied, since it depends on above two factors, therefore, no upper bound for these two evaluation measures. Conversely, in the theoretically best case of Shannon Entropy, all actives of a certain class stay in one single cluster, the best possible Entropy value (lower bound) will be zero. Similarly, as for Entropy based on cluster size, clustering outcome containing one extremely large cluster and many singletons will leads the theoretically best (lowest) Entropy value, approaching zero, however it is actually an abnormal clustering result in practice.

According to the equations discussed in Sections 4.4.4 and 4.4.5 for the evaluation using F-Measure and QCI, both measures consider the number of actives and dataset size, that is, both evaluation measures depend on number of actives and dataset size. For example, large dataset size and small number of actives tend to generate low value of F-Measure or QCI; by contrast, if the number of actives is close to the dataset size, the value of F-Measure or QCI

will possibly be high. The boundary of evaluation using F-Measure and QCI is (0, 1].

In this thesis, it was initially expected that it would be possible to find a standardization method, clustering method or consensus clustering method offering consistent benefit in the applications of chemical clustering. However, the results show that this is not the case, as reported in the literature. No clustering technique is universal to all applications.

## 8.2 Future Work

This thesis involved the application of standardization procedures to chemical clustering, which is little studied in chemoinformatics, and the application of consensus clustering, which is discussed for the first time in chemoinformatics. Thus, there is obviously a lot of space for improvement and extension.

First, for the manners of measuring consensus similarity in our work, we just simply count the pairs of co-clustered objects in the set of clusterings. However, varied techniques (Saporta and Youness, 2002) to compute consensus similarity by comparing partitions were reported in the literature. Different techniques to calculate consensus similarity will result in different types of similarity matrix, and this will also, of course, lead to varied performance of consensus clustering methods.

Along with the measuring consensus similarity, the weighting scheme is also a component worth discussing in consensus clustering. In our study, we simply weight the consensus similarity based on the performance of a given clustering method from prior result. As we described in Section 8.1, many weighting schemes for consensus clustering were reported effective in the literature (Li and Ding, 2008; Domeniconi and Al-Razgan, 2009). Two types of weighting schemes might be worth applying to the field of chemoinformatics as follows:

Li & Ding (2008) proposed a weighted consensus clustering which is based on a non-negative matrix factorization (NMF) framework. A NMF is a matrix which can usually be factorized into two non-negative matrices (factors) (Berry et. al., 2006). In addition, each input clustering in the weighted consensus clustering is treated unequally with a weight which is automatically determined by a weighted aggregate connectivity matrix which records the co-clustered relationship of pairwise objects. They also showed their NMF framework is an instance of spares Principal Component Analysis, therefore their weighting scheme is able to deal with the case when some input clusterings are highly correlated, their weights will be small.

The second is that Gullo et. al. (2009) proposed diversity-based weighting schemes, as mentioned in Section 7.3.2. The main difference between above NMF-based and the diversity-based is that the consensus clustering problem has to be formulated into NMF framework, while diversity-based weighting schemes consider only general properties of consensus clustering and is based on different implementations of diversity functions e.g. Normalized Mutual Information (NMI) and F-Measure in their study. Moreover, the diversity-based weighting schemes can be applied to any Instance-based, Cluster-based and Hybrid consensus clustering method.

Our works in Chapters 6 and 7 deal with the chemical datasets represented by ECFP_4 fingerprints. Basically, clustering is dataset dependent as reported in the literature. Hence, clustering on the datasets represented by similar fingerprints (e.g. molecular holograms) or different chemical representations (Molconn-Z) may result in different results.

Finally, the evaluation of clustering is another critical component. Different evaluation criteria evaluate different features of a clustering. It is difficult to find a clustering method can fit all types of evaluation criterion. Evaluation using similar types of criterion may be more likely to result in consistent evaluation of clustering performance.

# References


Accelrys. (2007). SciTegic Pipeline Pilot Software version 6.1: Accelrys.

Adamson, G. W., & Bush, J. A. (1973). "A Method for the Automatic Classification of Chemical Structures". *Information Storage and Retrieval,* **9**, 561-568.

Ailon, N., et al. (2008). "Aggregating Inconsistent Information: Ranking and Clustering". *Journal of ACM,* **55**(5), 1-27.

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster Analysis*. Newbury Park, CA: Sage University Press.

Al-Razgan, M., & Domeniconi, C. (2006). "Weighted Clustering Ensembles". In: Proceedings of *The Sixth SIAM International Conference on Data Mining*, 20-22 April 2006, Bethesda, USA, 258-269.

Attwood, T. K., & Smith, D. J. P. (1999). *Introduction to Bioinformatics*. London: Addison Wesley Longman.

Barnard, J. M. (2003). "Representation of Molecular Structure-Overview". In: Gasteiger, J. (ed.), *Handbook of Chemoinformatics: From Data to Knowledge*. Weinheim: Wiley-VCH.

Barnard, J. M., & Downs, G. M. (1992). "Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures". *Journal of Chemical Information and Computer Sciences,* **32**(6), 644-649.

Barnard, J. M., & Downs, G. M. (1997). "Chemical Fragment Generation and Clustering Softwares". *Journal of Chemical Information and Computer Sciences,* **37**(1), 141-142.

Bath, P. A., et al. (1993). "Effect of Standardization on Fragment-Based Measures of Structural Similarity". *Journal of Chemometrics,* **7**(6), 543-550.

Batista, J., et al. (2006). "Assessment of Molecular Similarity from the Analysis of Randomly Generated Structural Fragment Populations". *Journal of Chemical Information and Computer Sciences,* **46**, 1937-1944.

Ben-Dor, A., et al. (1999). "Clustering Gene Expression Patterns". *Journal of Computational Biology,* **6**, 281-297.

Ben-Hur, A., et al. (2002). "A Stability Based Method for Discovering Structure in Clustered Data". In: Altman, B., et al. (eds.), *Pacific Symposium on Biocomputing*, pp. 6-17. Sydney: World Scientific Publishing Company.

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

Bertolacci, M., & Wirth, A. (2007). "Are Approximation Algorithms for Consensus Clustering Worthwhile". In: Proceedings of *The Seventh SIAM International Conference on Data Mining*, 26-28 April 2007, Minneapolis, USA, 437-442.

Böcker, A., et al. (2006). "NIPALSTREE: A New Hierarchical Clustering Approach for Large Compound Libraries and Its Application to Virtual Screening". *Journal of Chemical Information and Modeling*, **46**(6), 2220-2229.

Böcker, A., et al. (2005). "A Hierarchical Clustering Approach for Large Compound Libraries". *Journal of Chemical Information and Modeling*, **45**, 807-815.

Brown, F. K. (1998). "Chemoinformatics: What Is It and How Does It Impact Drug Discovery". *Annual Reports in Medicinal Chemistry*, **33**, 375-384.

Brown, R. D., & Martin, Y. C. (1996). "Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection". *Journal of Chemical Information and Computer Sciences*, **36**(3), 572-584.

Butina, D. (1999). "Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets". *Journal of Chemical Information and Computer Sciences*, **39**(4), 747-750.

Cain, A. J., & Harrison, G. A. (1958). "An Analysis of the Taxonomist's Judgment of Affinity". In: Proceedings of *Proceedings of the Zoological Society of London* 1958, London, UK, 85-98.

Carmichael, R. M. (1968). "Finding Natural Clusters". *Systematic Zoology*, **17**, 144-150.

Caruana, R., et al. (2006). "Meta Clustering". In: Proceedings of *Proceedings of the International Conference on Data Mining (ICDM)*, 26-29 June 2006, Hong Kong, China, 107-118.

CAS. (2010). *Chemical Abstracts Service*. http://www.cas.org/ [Accessed 13 March 2010].

CCDC. (2010). *Cambridge Crystallographic Data Centre.* http://www.ccdc.cam.ac.uk/ [Accessed 19 Jan. 2010].

CLUTO. (2003). Clustering Toolkit version 2.1.1: Karypis Lab.

Conover, W. J., & Iman, R. L. (1981). "Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics". *The American Statistician,* **35**(3), 124-129.

Cormack, R. M. (1971). "A Review of Classification". *Journal of the Royal Statistical Society. Series A (General),* **134**(3), 321-367.

Digital Chemistry. (2007). *Digital Chemistry Clustering Tools.* http://www.digitalchemistry.co.uk [Accessed 6 June 2007].

DiMasi, J. A., et al. (2003). "The Price of Innovation: New Estimates of Drug Development Costs". *Journal of Health Economics,* **22**, 151-185.

Doherty, K. A. J., et al. (2004). "Non-Euclidean Norms and Data Normalisation". In: Proceedings of *Proceedings of the 12th Euro. Symposium on Artificial Neural Networks*, 28-30 April 2004, Bruges, Belgium, 181-186.

Domeniconi, C., & Al-Razgan, M. (2009). "Weighted Cluster Ensembles: Methods and Analysis". *ACM Transactions on Knowledge Discovery from Data (TKDD),* **2**(4), 1-42.

Dorans, N. J., & Kulick, E. (1986). "Demostrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test". *Journal of Educational Measurement,* **23**(4), 355-368.

Downs, G. M., & Barnard, J. M. (2002). "Clustering Methods and Their Uses in Computational Chemistry". *Reviews in Computational Chemistry,* **18**, 1-40.

Downs, G. M., & Willett, P. (1994). "Clustering of Chemical Structure Databases for Compound Selection". In: Waterbeemd, H. V. D. (ed.), *Advanced Computer-Assisted Techniques in Drug Discovery.* Weinheim: VCH Publishers.

Downs, G. M., et al. (1994). "Similarity Searching and Clustering of Chemical-structure Databases Using Molecular Property Data". *Journal of Chemical Information and Computer Sciences,* **34**(5), 1094-1102.

Dubes, R. C. (1987). "How Many Clusters Are Best? An Experiment". *Pattern Recognition,* **20**(6), 645-663.

Dunbar, J. B. (1997). "Cluster-Based Selection". In: *Computational Methods for the Analysis of Molecular Diversity*, pp. 51-63. Dordrecht, The Netherlands: Kluwer/ESCOM.

Dunham, M. H. (2003). *Data Mining: Introductory and Advanced Topics*. Upper Saddle River, NJ: Prentice Hall.

Eckert, H., & Bajorath, J. (2006). "Design and Evaluation of a Novel Class-Directed 2D Fingerprint to Search for Structurally Diverse Active Compounds". *Journal of Chemical Information and Modeling,* **46,** 2515-2526.

Edelbrock, C. (1979). "Comparing the Accuracy of Hierarchical Clustering Algorithms: The Problem of Classifying Everybody". *Multivariate Behavioral Research,* **14,** 367-384.

El-Hamdouchi, A., & Willett, P. (1986). "Hierarchic Document Clustering Using Ward's Method". In: Proceedings of *The 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 8-10 September 1986, Pisa, Italy, 149-156.

El-Hamdouchi, A., & Willett, P. (1989). "Comparison of Hierarchic Agglomerative Clustering Methods for Document Retrieval". *The Computer Journal,* **32**(3), 220-227.

Engel, T. (2003a). "Databases and Data Sources in Chemistry". In: Gasteiger, J. & Engel, T. (eds.), *Chemoinformatics*, pp. 227-288. Weinheim: Wiley-VCH.

Engel, T. (2003). "Representation of Chemical Compounds". In: Gasteiger, J. & Engel, T. (eds.), *Chemoinformatics*, pp. 15-168. Weinheim: Wiley-VCH.

Everitt, B. S., et al. (2001). *Cluster Analysis*. London: Arnold.

Filkov, V., & Skiena, S. (2004). "Heterogeneous Data Integration with the Consensus Clustering Formalism". In: Proceedings of *Data Integration in the Life Sciences*, 25-26 March 2004, Leipzig, Germany, 110-123.

Filkov, V., & Skiena, S. (2004a). "Integrating Microarray Data by Consensus Clustering". *International Journal on Artificial Intelligence Tools,* **13**(4), 863-880.

Fisanick, W., & Shively, E. R. (2003). "The CAS Information System: Applying Scientific Knowledge and Technology for Better Information". In: Gasteiger, J. (ed.), *Handbook of Chemoinformatics: From Data to Knowledge*, pp. 556-606. Weinheim: Wiley-VCH.

Flower, D. R. (1998). "On the Properties of Bit String-Based Measures of Chemical Similarity". *Journal of Chemical Information and Computer Sciences,* **38**, 379-386.

Fowlkes, E. B., & Mallows, C. L. (1983). "A Method for Comparing Two Hierarchical Clusterings". *Journal of American Statistical Association,* **78**, 553-569.

Fraley, C., & Raftery, A. E. (1998). "How Many Clusters? Which Clustering Methods? Answer Via Model-Based Cluster Analysis". *The Computer Journal,* **41**(8), 1-11.

Fred, A. (2001). "Finding Consistent Clusters in Data Partitions". In: Proceedings of *The Second International Workshop on Multiple Classifier Systems (MCS)*, 2-4 July 2001, London, UK, 309-318.

Fred, A., & Jain, A. (2002). "Data Clustering Using Evidence Accumulation". In: Proceedings of *The Sixteenth International Conference on Pattern Recognition (ICPR)*, 11-15 August 2002, Quebec, Canada, 276-280.

Fung, B. C. M., et al. (2003). "Hierarchical Document Clustering Using Frequent Itemsets". In: Barbara, D. & Kamath, C. (eds.), Proceedings of *The Third SIAM International Conference on Data Mining*, 1-3 May 2003, San Francisco, USA, 59-70.

Gasteiger, J. (2003). "Introduction of Chemoinformatics". In: Gasteiger, J. & Engel, T. (eds.), *Chemoinformatics*, pp. 9-12. Weinheim: Wiley-VCH.

Ginn, C. M. R., et al. (1997). "Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion". *Journal of Chemical Information and Computer Sciences,* **37**(1), 23-37.

Gionis, A., et al. (2007). "Clustering Aggregation". *ACM Transactions on Knowledge Discovery from Data (TKDD),* **1**(1), 1-30.

Gnanadesikan, R., et al. (1995). "Weighting and Selection of Variables for Cluster Analysis". *Journal of Classification,* **12**, 113-136.

Godden, J. W., & Bajorath, J. (2001). "Differential Shannon Entropy as a Sensitive Measure of Differences in Database Variability of Molecular Descriptors". *Journal of Chemical Information and Computer Sciences,* **41**(4), 1060-1066.

Goder, A., & Filkov, V. (2008). "Consensus Clustering Algorithms: Comparison and Refinement". In: Proceedings of *The Tenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, 19 January 2008, San Francisco, USA, 109-117.

Good, A. C., et al. (2004). "Descriptors You Can Count On? Normalized and Filtered Pharmacophore Descriptors for Virtual Screening". *Journal of Computer-Aided Molecular Design,* **18**, 523-527.

Gower, J. C. (1971). "A General Coefficient of Similarity and Some of Its Properties". *Biometrics,* **27**(4), 857-871.

Guha, S., et al. (1998). "CURE: An Efficient Clustering Algorithm for Large Databases". In: Proceedings of *The ACM SIGMOD Conference on Management of Data*, 2-4 June 1998, Seattle, USA, 73-84.

Guha, S., et al. (1999). "ROCK: A Robust Clustering Algorithm for Categorical Attributes". In: Proceedings of *The IEEE International Conference on Data Engineering*, 23-26 March 1999, Sydney, Austrialia, 512-521.

Gullo, F., et al. (2009). "Diversity-based Weighting Schemes for Clustering Ensembles". In: Proceedings of *The Ninth SIAM International Conference on Data Mining*, 30 April - 2 May 2009, Nevada, USA, 437-448.

HAC. (2010). winMolconn Software: Hall Associates Consulting.

Halkidi, M., et al. (2001). "On Clustering Validation Techniques". *Journal of Intelligent Information Systems,* **17**(2), 107-145.

Hall, L. H., & Kier, L. B. (2001). "Issues in Representation of Molecular Structure : The Development of Molecular Connectivity". *Journal of Molecular Graphics and Modelling,* **20**(1), 4-18.

Haranczyk, M., & Holliday, J. (2008). "Comparison of Similarity Coefficients for Clustering and Compound Selection". *Journal of Chemical Information and Modeling,* **48**(3), 498-508.

Hert, J., et al. (2004). "Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures". *Journal of Chemical Information and Computer Sciences,* **44**, 1177-1185.

Holliday, J. D., et al. (2002). "Grouping of Coefficients for the Calculation of Inter-molecular Similarity and Dissimilarity Using 2D Fragment Bit-strings". *Combinatorial Chemistry and High Throughput Screening,* **5**, 155-166.

Holliday, J. D., et al. (2004). "Clustering Files of Chemical Structures Using the Fuzzy K-means Clustering Method". *Journal of Chemical Information and Computer Sciences,* **44**(3), 894-902.

Holliday, J. D., et al. (2003). "Analysis and Display of the Size Dependence of Chemical Similarity Coefficients". *Journal of Chemical Information and Computer Sciences,* **43**, 819-828.

Homeyer, A., & Reitz, M. (2003). "Databases in Biochemistry and Molecular Biology". In: Gasteiger, J. (ed.), *Handbook of Chemoinformatics: From Data to Knowledge*, pp. 756-789. Weinheim: Wiley-VCH.

IUPAC. (2010). InChI software version 1.01: International Union of Pure and Applied Chemistry.

Jain, A. (2010). "Data Clustering: 50 Years Beyond K-Means". *Pattern Recognition Letters,* **31**, 651-666.

Jain, A. K., et al. (1999). "Data Clustering: A Review". *ACM Computing Surveys,* **31**(3), 264-323.

Jarvis, R. A., & Patrick, E. A. (1973). "Clustering Using a Similarity Measure Based on Shared Near Neighbours". *IEEE Transactions in Computers,* **C-22**(11), 1025-1034.

Kantardzic, M. (2003). *Data Mining: Concepts, Models, Methods, and Algorithms*. Piscataway, NJ: IEEE Press.

Karypis, G. (2003). *CLUTO: A Clustering Toolkit, Release 2.1.1*: University of Minnesota.

Karypis, G., et al. (1999). "Multilevel Hypergraph Partitioning: Applications in VLSI Domain". *IEEE Transactions on Very Large Scale Intergration Systems,* 7(1), 69-79.

Ketchen, D. J., & Shook, C. L. (1996). "The Application of Cluster Analysis in Strategic Management Research : An Analysis and Critique". *Strategic Management Journal,* **17**, 441-458.

Khalifa, A. A., et al. (2009). "Comparison of Nonbinary Similarity Coefficients for Similarity Searching, Clustering and Compound Selection". *Journal of Chemical Information and Modeling,* **49**(5), 1193-1201.

Kochev, N. (2003). "Searching Chemical Structures". In: Gasteiger, J. & Engel, T. (eds.), *Chemoinformatics*, pp. 291-318. Weinheim: Wiely-VCH.

Leach, A. R., & Gillet, V. J. (2007). *An Introduction to Chemoinformatics*. Dordrecht: Kluwer

Li, T. (2006). "A Unified View on Clustering Binary Data". *Machine Learning,* **62**(3), 199-215.

Li, T., & Ding, C. (2008). "Weighted Consensus Clustering". In: Proceedings of *The Eighth SIAM International Conference on Data Mining*, 24-26 April 2008, Atlanta, USA, 798-809.

Li, W. (2006a). "A Fast Clustering Algorithm for Analyzing High Similar Compounds of Very Large Libraries". *Journal of Chemical Information and Modeling,* **46**, 1919-1923.

Marshall, G. R. (2005). "Introduction to Chemoinfromatics in Drug Discovery - A Personal Review". In: Oprea, T. I. (ed.), *Chemoinformatics in Drug Discovery*, pp. 1-18. Weinheim, Germany: Wiley-VCH.

Matter, H. (1997). "Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors". *Journal of Medicinal Chemistry,* **40**(8), 1219-1229.

Milligan, G. W. (1980). "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms". *Psychometrika,* **45**, 325-342.

Milligan, G. W. (1996). "Clustering Validation: Results and Implications for Applied Analyses". In: Arabie, P., et al. (eds.), *Clustering and Classification*, pp. 341-375. River Edge, NJ: World Scientific Publishing.

Milligan, G. W., & Cooper, M. C. (1987). "Methodology Review: Clustering Methods". *Applied Psychological Measurement,* **11**(4), 329-354.

Milligan, G. W., & Cooper, M. C. (1988). "A Study of Standardization of Variables in Cluster Analysis". *Journal of Classification,* **5**(2), 181-204.

Monti, S., et al. (2003). "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data". *Machine Learning,* **52**(1), 91-118.

Morgan, H. L. (1965). "The Generation of a Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Service". *Journal of Chemical Documentation,* **5**, 107-113.

Murtagh, F. (2000). "Clustering in Massive Data Sets". In: Proceedings of *The Chemical Data Analysis in the Large: The Challenge of the Automation Age*, 22-26 May 2000, Bozen, Italy, 28-51.

Nguyen, N., & Caruana, R. (2007). "Consensus Clusterings". In: Proceedings of *The Seventh IEEE International Conference on Data Mining*, 28-31 October 2007, Omaha, USA, 607-612.

Oprea, T. I. (2005). "Chemoinfromatics in Drug Discovery". In: Oprea, T. I. (ed.), *Chemoinformatics in Drug Discovery*, pp. 25-37. Weinheim, Germany: Wiley-VCH.

Paris, C. G. (2003). "Databases of Chemical Structures". In: Gasteiger, J. (ed.), *Handbook of Cheminformatics: From Data to Knowledge*, pp. 523-552. Weinheim: Wiley-VCH.

PDB. (2010). *Protein Data Bank.* http://www.pdb.org/ [Accessed 19 Jan. 2010].

Punj, G., & Stewart, D. W. (1983). "Cluster Analysis in Marketing Research: Review and Suggestions for Application". *Journal of Marketing Research,* **20**(2), 134-148.

Rand, W. M. (1971). "Objective Criteria for the Evaluation of Clustering Methods". *Journal of the American Statistical Association,* **66**(336), 846-850.

Raymond, J. W., et al. (2003). "Comparison of Chemical Clustering Methods Using Graph- and Fingerprint-based Similarity Measures". *Journal of Molecular Graphics and Modelling,* **21**(5), 421-433.

Raymond, J. W., & Willett, P. (2003). "A Line Graph Algorithm for Clustering Chemical Structures Based on Common Substructural Core". *MATCH,* **48**, 197-207.

Reynolds, C. H., et al. (1998). "Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds". *Journal of Chemical Information and Computer Sciences,* **38**(2), 305-312.

Rijsbergen, C. J. V. (1979). *Information Retrieval*: Butterworths.

Rogers, D. F., et al. (1991). "Aggregation and Disaggregation Techniques and Methodology in Optimization". *Operations Research,* **39**(4), 553-582.

Romesburg, H. C. (1984). *Cluster Analysis for Researchers*. North Carolina: LULU.

Rosenberg, A., & Hirschberg, J. (2007). "V-measure: A Conditional Entropy-based External Cluster Evaluation Measure". In: Proceedings of *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 28-30 June 2007, Prague, Czech Republic, 410-420.

Rubin, V., & Willett, P. (1983). "A Comparison of Some Hierarchical Monothetic Divisive Clustering Algorithms for Structure-Property Correlation". *Analytica Chemica Acta,* **151**, 161-166.

Saad, F. H., et al. (2006). "A Comparison of Two Document Clustering Approaches for Clustering Medical Documents". In: Crone, S. F., et al. (eds.), Proceedings of *International Conference on Data Mining*, 11 December 2006, Las Vegas, USA, 425-431.

Sadowski, J. (2003). "3D Structure Generation". In: Gasteiger, J. (ed.), *Handbook of Chemoinformatics: From Data to Knowledge*. Weinheim: Wiely-VCH.

Salim, N., et al. (2003). "Combination of Fingerprint-based Similarity Coefficients Using Data Fusion". *Journal of Chemical Information and Computer Sciences*, **43**(2), 435-442.

Saporta, G., & Youness, G. (2002). "Comparing Two Partitions: Some Proposals and Experiments". In: Hardle, W. & Bernd, R. (eds.), Proceedings of *Proceedings in Computational Statistics*, 24-28 August 2002, Berlin, Germany, 243-248.

Sheridan, R. P., & Kearsley, S. K. (2002). "Why Do We Need So Many Chemical Similarity Search Methods? " *Drug Discovery Today*, 7(17), 903-911.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. Singapore: McGraw-Hill.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical Taxonomy*. San Francisco, CA: Springer.

Sokal, R. R. (1961). "Distance as a Measure of Taxonomic Similarity". *Systematic Zoology*, **10**(2), 70-79.

Steinbach, M., et al. (2000). "A Comparison of Document Clustering Techniques". In: Proceedings of *The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 20-23 August 2000, Boston, USA, 35-54.

Stoddard, A. M. (1979). "Standardization of Measures Prior to Cluster Analysis". *Biometrics*, **35**(4), 765-773.

Strehl, A., & Ghosh, J. (2002). "Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions". *The Journal of Machine Learning Research*, **3**, 583-617.

Strike, K., et al. (2001). "Software Cost Estimation with Incomplete Data". *IEEE Transactions on Software Engineering*, **27**(10), 890-908.

Symyx Technologies. (2007). *MDL Drug Data Report*. http://www.symyx.com.

Szekely, G. J., & Rizzo, M. L. (2005). "Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method". *Journal of Classification,* **22**(2), 151-183.

Tarkhov, A. (2003). "Chemistry on the Internet". In: Gasteiger, J. (ed.), *Handbook of Chemoinformatics: From Data to Knowledge*, pp. 794-840. Weinheim: Wiley-VCH.

Terfloth, L. (2003). "Calculation of Structure Descriptors". In: Gasteiger, J. & Engel, T. (eds.), *Chemoinformatics*, pp. 401-437. Weinheim: Wiely-VCH.

Thomson Reuters. (2007). *IDAlert database*. http://thomsonreuters.com/.

Topchy, A., et al. (2004). "Adaptive Clustering Ensembles". In: Proceedings of *The Seventeenth International Conference on Pattern Recognition*, 23-26 August 2004, Cambridge, UK, 272-275.

Tripos. (2007). Sybyl Software version 7.2: Tripos.

Turner, D. B., et al. (1995). "Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of Similarity Coefficients and Standardisation Methods for Field-based Similarity Searching". *SAR and QSAR in Environmental Research,* **3**, 101–130.

Varin, T., et al. (2008). "3D Pharmacophore, Hierarchical Methods, and 5-HT4 Receptor Binding Data". *Journal of Enzyme Inhibition and Medicinal Chemistry,* **23**(5), 593-603.

Wang, F., et al. (2009). "Generalized Cluster Aggregation". In: Proceedings of *The Twenty-First International Joint Conference on Artificial Intelligence*, 11-17 July 2009, Pasadena, USA, 1279-1284.

Ward Jr, J. H. (1963). "Hierarchical Grouping to Optimize an Objective Function". *Journal of the American Statistical Association,* **58**(301), 236-244.

Weininger, D. (1988). "SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules". *Journal of Chemical Information and Computer Sciences,* **28**(1), 31-36.

Weininger, D., et al. (1989). "SMILES. 2. Algorithm for Generation of Unique SMILES Notation". *Journal of Chemical Information and Computer Sciences,* **29**(2), 97-101.

Wiggins, G. D. (2003). "Overview of Databases / Data Sources". In: Gasteiger, J. (ed.), *Handbook of Chemoinformatics: From Data to Knowledge*, pp. 496-522. Weinheim: Wiley-VCH.

Willett, P. (1985). "Clustering Tendency in Chemical Classifications". *Journal of Chemical Information and Computer Sciences,* **25**, 78-80.

Willett, P. (1987). *Similarity and Clustering in Chemical Information Systems*. Letchworth: Research Studies Press.

Willett, P. (1988). "Recent Trends in Hierarchic Document Clustering : A Critical Review". *Information Processing and Management,* **24**(5), 577-597.

Willett, P. (2003). "Similarity-Based Approaches to Virtual Screening". *Biochemical Society Transactions,* **31**, 603-606.

Willett, P. (2003a). "Similarity Searching in Chemical Structure Databases". In: Gasteiger, J. (ed.), *Handbook of Chemoinformatics: From Data to Knowledge*. Weinheim: Wiley-VCH.

Willett, P. (2005). "Chemoinformatics Techniques for Data Mining in Files of Two-dimensional and Three-dimensional Chemical Molecules". In: Petitjean, M. (ed.), Proceedings of *The Third Conference on the Foundations of Information Science*, 4-7 July 2005, Paris, France, 1-15.

Willett, P. (2006). "Similarity-based Virtual Screening Using 2D Fingerprints". *Drug Discovery Today,* **11**(23-24), 1046-1053.

Willett, P. (2008). "From Chemical Documentation to Chemoinformatics: Fifty Years of Chemical Information Science". *Journal of Information Science,* **34**(4), 477-499.

Willett, P. (2009). "Similarity Methods in Chemoinformatics". *Annual Review of Information Science and Technology,* **43**, 3-71.

Willett, P., et al. (1998). "Chemical Similarity Searching". *Journal of Chemical Information and Computer Sciences,* **38**, 983-996.

Willett, P., & Gillet, V. J. (2007). "Compound Selection Using Measures of Similarity and Dissimilarity". In: Talor, J. B. & Triggle, D. J. (eds.), *Comprehensive Medicinal Chemistry II.* Maryland, USA: Elsevier.

Willett, P., et al. (1986). "Implementation of Nonhierarchic Cluster Analysis Methods in Chemical Information Systems : Selection of Compounds for Biological Testing and Clustering of Substructure Search Output". *Journal of Chemical Information and Computer Sciences,* **26**(3), 109-118.

Williams, W. T., & Lambert, J. M. (1966). "Multivariate Methods in Plant Ecology: V. Similarity Analyses and Information-Analysis". *The Journal of Ecology,* **54**(2), 427-445.

Xu, R., & Donald Wunsch, II. (2005). "Survey of Clustering Algorithms". *IEEE Transactions on Neural Networks,* **16**(3), 645-678.

Xue, L., et al. (2003). "Design and Evaluation of a Molecular Fingerprint Involving the Transformation of Property Descriptor Values into a Binary Classificaion Scheme". *Journal of Chemical Information and Computer Sciences,* **43**, 1151-1157.

Yin, P. Y., & Chen, L. H. (1994). "A New Non-iterative Approach for Clustering". *Pattern Recognition Letters,* **15**(2), 125-133.

Zeng, Y., et al. (2002). "An Adaptive Meta-Clustering Approach: Combining the Information from Different Clustering Results". In: Proceedings of *The IEEE Computer Society Bioinformatics Conference*, 14-16 August 2002, Palo Alto, USA, 276-287.

Zhang, T., et al. (1996). "BIRCH: An Efficient Data Clustering Method for Very Large Databases". In: Proceedings of *The 1996 ACM SIGMOD International Conference on Management of Data*, 4-6 June 1996, Montreal, Canada, 103-114.

Zhao, Y., & Karypis, G. (2002). *Criterion Functions for Document Clustering Experiments and Analysis*: University of Minnesota.

Zhao, Y., & Karypis, G. (2005). "Hierarchical Clustering Algorithms for Document Datasets". *Data Mining and Knowledge Discovery,* **10**(2), 141-168.

Zuylen, v. A. (2005). *Deterministic Approximation Algorithms for Ranking and Clustering Problems*: Technical Report 1431, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY.