# EFFECTIVE DATA MINING IN MOLECULAR DATABASES

**N. SreeRam**

Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, India, sriramnimmagadda@gmail.com

**Abstract.**

Data Mining is a process of extracting the “hidden" patterns from the databases and data warehouses. It is currently used in a wide range of applications such as marketing surveillance, fraud detection and scientific discovery. We can also implement these data mining techniques in comparison and analysis of molecular databases. This study lays down a basis for creating a user interface and search engine for molecular data base.

**Keywords:** hidden patterns, data mining techniques, molecular databases and search engine

## 1. Introduction

Data mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives and summarizing it into useful information. Data mining tools allow the users to analyse data from different dimensions or angles, categorize it, and summarize the relationships identified. Technically, Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. It is the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of Statistics, Machine Learning and Data Base Management Systems.[1][2][3].

The main techniques of data mining are as follows.

Association analysis: Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. Classification: Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. The classification analysis would generate a model that could be used to either accept or reject credit requests in the future.

Prediction: Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two majors' types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values.

Clustering: Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter class similarity).[9][17][20][21][26][35].

## 2. INTRODUCTION TO MOLECULAR DATA BASES

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

are expected to handle the storage and searching of information on millions of molecules taking terabytes of physical memory. An important feature in a chemical database system is the ability to quantify the degree of structural similarity between pairs, or larger groups, of molecules. Two principal factors that will affect the performance of similarity calculations are the representation used to characterize the molecules and the similarity coefficients used to compare them.[6][8][11][14][22][34][39].

### *A. REPRESENTATION OF MOLECULES*

A molecular structure may be represented as an ordered set of components with information concerning the relationship between those components. This information may be in the form of a list, i.e., the labeling of atoms and bonds (molecular codes), or in the form of the count of components of various types describing the mathematical properties of a structure one-dimensional constitutional information provided by the preceding formulas, the two-dimensional structural formula represents the arrangement of atoms using the topology of the molecule and the connectivity of the constituting atoms. The graph, a variant of the structural formula, omits the type of atom and nature of bonding. Alternative representations have been designed such as the Simplified Molecular Input Line Entry Specification (SMILES). Three-dimensional structures describe the structure of the molecule as a three dimensional entity with the atoms situated in specific positions in the space (x,y,z, coordinates), thus providing geometrical and spatial information

### *B. MOLECULAR SIMILARITY AND DISSIMILARITY*

The definition of similarity with respect to chemical molecules is more stringent than that in other fields. Basically, it consists of mapping “chemical space” (a representation of a molecule in structural or some property space) to one-dimensional space with entities of real numbers. Ideally similarity measures for molecules behave proportionally to all physical and biological properties of molecules in this representation. In other words, it groups together all molecules with very similar physical and biological properties in a confined area of chemical property space. Similar Property Principle says that Molecules having similar structures and properties should also exhibit similar activity. Consider two molecules A and B, a is the number of features (characteristics) present in A and absent in B, b is the number of features absent in A and present in B, c is the number of features common to both molecules, and d is the number of features absent from both molecules. Thus, c and d measure the present and the absent matches, respectively, i.e., similarity; while a and b measure the corresponding mismatches, i.e., dissimilarity. [6][8][11].

## 3. ASSOCIATION RULES IN CHEMICAL DATABASES

The prototypical data mining task is to find all frequently occurring patterns of a particular type. In its simplest form, known as association rule mining, the task is to find all frequent item sets, i.e. to list all combinations of items that are found together in a sufficient number of examples. A typical application of association rules is market basket analysis, where you identify all products which tend to be sold together - this information can then be used to influence product placement, etc. Directly translating association rule mining into a chemical or molecular context, with molecules as shopping baskets: the task would be to find all elements that occur frequently together in molecules. As this translation makes clear, the standard data mining task of association rule mining is not directly transferable to chemical databases. What is important in chemical databases is not the frequency of co-occurrence of individual atoms, but the frequency of occurrence of particular molecular substructures. There are two approaches to incorporating molecular sub structures. The standard one is to use attributes to represent structure. Attributes are descriptors which describe a property of a whole object. For example, typical attributes of a compound, the presence of a particular molecular substructure, the charge at a particular co-ordinate position, etc. It is a characteristic of the use of attributes that all the information about a particular example can be put into a single row of a table. The use of attributes is standard in statistics, neural networks, and machine learning. It is also standard in chemo informatics: traditional QSAR , CASE/MULTICASE , COMFA ,

Recursive Partitioning etc. are all based on attributes. The alternative approach, and the one which we favor, is to use a relational language to describe chemicals. This approach is known as Inductive Logic Programming (ILP). Here chemical structure are naturally relational and they can only be approximated using attributes. ILP has shown its value in many conventional structure-function problems where it has found solutions not accessible to standard statistical, neural network, or genetic algorithms. ILP based drug design methods have been successfully extended from standard QSAR problems, and to pharmacophore discovery. Here we describe the ILP data mining algorithm Warmr. Warmr is a general purpose ILP data mining algorithm that finds frequent relational patterns in databases. It has been applied to a number of different application areas, e.g. telecommunications. The efficiency of Warmr scales linearly with database size and it has been applied to datasets containing many millions of data points. This answers some of the efficiency problems of ILP. The frequent relational patterns found by

Warmr can be used in predictive theories and contribute to scientific insight. Warmr is to find all frequent patterns in a database of chemicals. Warmr is a general-purpose Inductive Logic Programming (ILP) data mining algorithm. It uses datalog to represent both data and patterns. Datalog is a logic programming language (with no function symbols) specifically designed to implement deductive databases (databases that can incorporate rules as well as facts).

Warmr can discover knowledge in structured data, where patterns reflect the one-to-many and many-to-many relationships of several tables. This is not possible with standard data mining programs. Background knowledge is represented in a uniform manner and has an essential role in the discovery of frequent patterns, unlike in most data mining settings. Warmr used the efficient level wise method known from the Apriori algorithm This allows it to be used on very large databases. The Warmr level wise search algorithm is based on a breadth-first search of the pattern space. This space is ordered by the generality of patterns. The level wise method searches this space one level at a time, starting from the most general patterns. The method iterates between candidate generation and candidate evaluation phases: in candidate generation, the lattice structure is used for pruning non-frequent patterns from the next level; in the candidate evaluation phase, frequencies of with respect to the database. Pruning is based on the monotonicity of specificity with respect to frequency - if a pattern is not frequent then none of its specializations can be frequent. So, while generating candidates for the next level, all the patterns that are specializations of infrequent patterns can be pruned. The levelwise approach has two crucial useful properties. First, the database is scanned at most k+1 times, where k is the maximum level (size) of a frequent pattern; all candidates of a level are tested in single database pass. This is an important factor when mining large databases. Second, the time complexity is linear with the number of examples – assuming matching patterns against the data is fast. We have previously shown how Warmr can be tuned to simulate Apriori and some other well-known algorithms for frequent pattern discovery. Warmr is, in principle, capable of discovering arbitrary frequent data log queries from a given database. [4][5][7][10].

## 4. CLASSIFICATION IN CHEMICAL DATABASES

Molecules comparison is usually performed using either similarity coefficients or machine learning approaches.

Kernel Methods: They attempt to predict the output of a continuous output variable given continuous input variables. In drug-design, usually only the distinction between active and non-active entities is to be made. Then binary kernel methods are used, which can predict the output variable based on binary input vectors.

Artificial Neural Networks (ANNs) have been used to distinguish drug-like and nondrug-like molecules using a sub structural analysis [used electrostatic and steric properties at grid points for feeding a genetic artificial neural network in order to develop a QSAR model.

Support Vector Machines (SVMs) attempt to learn the maximum separating boundary compared to Neural Networks which do not optimize the decision boundary if the prediction performance does not change. Compared to C5.0 decision trees, multi-layer perceptions and other neural networks. SVMs need less training time and achieve slightly better prediction performance.

The following is based on the 2-dimensional structure of molecules. Using the classification given above, it belongs to the group of sub shape based molecular representations, combined with a machine learning approach in the form of a Naïve Bayesian Classifier for classification of structures We use atom environments as a molecular representation. Atom environments are similar to Signature Molecular Descriptors [Faulon 2003a; Faulon 2003b; Visco Jr. 2002; Faulon 1994]. They are translationally and rotationally invariant. Furthermore, they do not depend on a particular conformation as they are calculated from the connectivity table. This makes generating atom environments less difficult compared to alignment-dependent approaches. Another benefit with atom environments is that they are easily interpretable, as they resemble the chemical concept of functional groups.

## 3. CLUSTERING IN CHEMICAL DATABASES

Clustering methods can produce overlapping clusters or non-overlapping clusters. Overlapping clusters occur when each object can exist in more than one cluster, while in non-overlapping clusters; each object belongs to only one cluster.

The clustering process for chemical structures is outlined by Brown and Martin as follows:

(1) Select a set of attributes on which to base the comparison of the structures. These may be structural features and/or physicochemical properties.

(2) Characterize every structure in the dataset in terms of the attributes selected in step one.

(3) Calculate a coefficient of similarity, dissimilarity, or distance between every pair of structures in the dataset, based on their attributes.

(4) Use a clustering method to group together similar structures based on the coefficients calculated in step 3.

(5) Analyze the resultant clusters or classification hierarchy to determine which of the possible sets of clusters should be chosen.

Non-overlapping methods are more widely used for compound datasets, and the clustering experiments here only involve non-overlapping methods. The two main non- overlapping clustering approaches are hierarchical methods and non-hierarchical methods. Various clustering algorithms are as follows.

1.Agglomerative

2.Divisive

Agglomerative methods are.

1.a. Single linkage method

1.b. Complete linkage method

1.c. Wards method

1.d. Group weighted average method

1.e. Weighted average method

Divisive methods are as follows

2.a. Bisecting K-means algorithm

2.b. Minimum diameter algorithm

Nonhierarchical methods are

1.a.K. means algorithm

1.b. Partition around medoids

1.c. CLARANS algorithm

1.d. Mixture model algorithm

1.e. Density based algorithm

1.f. Jarvis –Patrick algorithm

1.g. Single pass algorithm

## 4. Conclusions

The main applications of data mining in molecular databases are rational molecular design that is implemented in drug discovery. The efficient design of chemical structures used in Computer-Aided Molecular Design (CAMD) and Computer Aided-Drug Design (CADD). The concept of similarity in molecular DBs has been proven to be very successful in the pharmaceutical industry in combinatorial chemistry, enormous libraries of millions of compounds are analysed by High Throughput Screening (HTS) methods. Here we have discussed various data mining techniques and algorithms implemented for molecular databases. Implementation of information technology in chemical databases is so useful in various applications like drug discovery, pharmaceuticals industry and combinatorial chemistry. The research in this wing is still going on in order to find efficient chemical structures design in drug discovery and pharmaceutical industry,

## References

[1] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[2] http://www.exinfm.com/pdffiles/intro_dm.pdf

[3] CMPUT690 Principles of Knowledge Discovery in Databases Ó Osmar R. Zaïane, 1999.

[4] Warmr: A Data Mining Tool for Chemical DataRoss D. King1, Ashwin Srinivasan2, Luc Dehaspe3

[5] Finding frequent substructures in chemical compoundsLuc Dehaspe, Hannu Toivonen, Ross Donald King

[6] SMARTS Approach to Chemical Data Mining and Physicochemical Property Prediction by Adam C. Lee.

[7] Technical note Computational chemistry, data mining, high-throughput synthesis and screening—informatics and integration in drug discoveryCharles J. Manly Discovery Technologies, Neurogen Corporation, Branford, CT, USA

[8] New Paradigms in Computational Chemistry for Drug Discovery report by Dr Robert Brown and Oleg Fastovsky

[9] Principles of Knowledge Discovery in Databases Osmar R. Zaïane, 1999.

[10] Association Rules: Problems, solutions and new applications María N. Moreno, Saddys Segrera and Vivian F. López Universidad de Salamanca, Plaza Merced S/N, 37008, Salamanca.

[11] APPLICATIONS OF DATA MINING TECHNIQUES IN PHARMACEUTICAL INDUSTRY Jayanthi Ranjan Information Management and Technology Area Institute of Management Technology Raj Nagar, Ghaziabad. Uttar Pradesh , India

[12] Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries: The Diverse Property-Derived (DPD) Approach. J. Chem. Inf. Comput. Sci. 1997, 37, 559-614.

[13] Downs, G. M.; Willett, P. Clustering of Chemical-Structure Databases for Compound Selection. In AdVanced Computer-Assisted Techniques in Drug DiscoVery; van de Waterbeemd, H., Ed.; VCH: New York 1994; pp 111-130

[14] Development of Compound Clustering, Techniques Using Hybrid Soft-Computing Algorithms, PROJECT NUMBER: 04 - 02 - 06 - 0093 EA001, VOTE NUMBER: 74252 PROJECT LEADER: Assoc. Prof. Dr. Naomie Salim Faculty of Computer Science & Information Systems, Universiti Teknologi Malaysiam/ enterprise/ chemicals/ chempol/ whitepaper/ reach.htm

[15] A Similarity Based Approach for Chemical Category Classification Ana Gallegos Saliner, Grace Patlewicz, Andrew p. worth european commission directorate general clustering chemical data set using particle swarm optimization based algorithm triyono faculty of computer science and information system, universiti teknologi malaysia