International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 4, Number 4 (2014), pp. 335-342

http://www. irphouse.com /ijict.htm

DRUG Discovery Using Data Mining

Charanpreet Kaur and Shweta Bhardwaj

Trinity Institute of Professional Studies, Dwarka, New Delhi.

Abstract

Drug Discovery today conceptualizes on the involvement of chemo informatics to overcome the shortcomings of the traditional Drug development process. Drug Discovery with the use of Chemo informatics and Data Mining generates large numbers of related chemical compounds. It speeds up the drug discovery in two ways: It can generate several million structurally related chemical molecules. By increasing the chemicals available for testing, the chances of finding a drug lead may be higher. If you know the biological effects of different molecules, you can combine them to make chemicals with particular designed biological effects.

Chemo informatics (Chemical Informatics) is the use of Computer and Information Technology, applied to a range of problems in the field of Chemistry.It transforms the data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug identification and optimization.

With Data Mining, throughout drug discovery, data is collected relating chemical structures to each other. The Data Mining Technique “Clustering Process” divides the databases of unknown drugs in clusters based on their similarity. It makes use of Lipinski Rule which defines those compounds as Drug like which have properties implicit to drug likeness such as log p, Molecular Weight, Number of Hydrogen bond acceptors and donors in a molecule etc. The clusters of unknown similar drugs are evaluated and compared with the clusters of some specific (e.g. HIV) drugs to discover from unknown drugs those drugs that are similar in properties with the known drugs.

This paper highlights the use of Chemo Informatics and Data Mining Techniques for exploiting the widespread database of unknown drugs to discover new Drug like compounds that contain functional groups and have physical as well as therapeutic properties consistent with the majority of known drugs.

Keywords: Data Warehouse, Data Mining, Drug Discovery, Chemo Informatics, Clustering Technique, Data Marts, Knowledge Discovery.

1. Introduction

Today, in the modern business world the financial crisis has increased the focus on Business Intelligence. Executives cannot afford to make decisions based on financial statements that compare last month's results to a budget created up to a year ago. They need information that helps them quickly answer the basic questions: What still works? What continues to sell? How can cash be conserved? What costs can be cut without causing long-term harm?

The giants that pioneered Business Intelligence (BI) made an important discovery that the path to true business intelligence passes through a Data Warehouse. "A data warehouse is a subject-oriented, integrated, time variant and non-volatile collection of data in support of management decision making process."

In today's world, an organization generates more information in a week than most people can read in a lifetime. It is humanly impossible to study, decipher, and interpret all the data to find useful information. A Data Warehouse pools all the data after proper transformation and cleansing into well organizes data structures.Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories. We are drowning in data, but starving for knowledge!

The solution for the data explosion problem is Data Mining.Data mining is the Knowledge Discovery in the databases that is the Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases.Data mining is the process of extracting hidden patterns from large amounts of data.

2. Relevance of Data Mining in Drug Discovery

There are different Data Mining Techniques that can be applied on the data warehouse to obtain knowledge or useful information. **In our research work, our basic aim was to study and analyze the various Data Mining Techniques and apply those algorithms in the real time. The 'Drug Discovery' application uses the Data Mining techniques and Chemo Informatics to find the unknown drugs that are similar to certain specific known drugs and study the relationship between them.**

The application named Drug Discovery is being developed keeping in mind the slow process of developing a new Drug. Drug Discovery today conceptualizes on the involvement of chemo informatics to overcome the shortcomings of the traditional Drug development process. The platform used for creating the application of Drug Discovery is primarily Java. Also the software tool named NetBeans is used which is an API and is used to develop the code, algorithms and the interface. Microsoft Access is used as the backend for storing the drugs.

The system would allow the user to find the clusters of unknown similar drugs and compare them with the clusters of some specific (e.g. HIV) drugs to discover from unknown drugs those drugs that are similar in properties with the known drugs. The access rights are not an issue because this software is created keeping in mind the freeware concept.

The summary of the major functions the software will perform are:

(i) Creating of DNS directly if the database is already attached or browse for data marts if not already uploaded.

(ii) After the DNS creation, the Data marts can be integrated into a Data Warehouse.

(iii)Cleaning can be done on the Data Warehouse to remove noisy and inconsistent data.

(iv)Clustering can be done to find clusters of drugs similar in chemical properties.

(v) Pattern evaluation can be done using Tanimoto Coefficient and structural similarity.

3. Implementation of Drug Discovery Application

3.1 Data Pre-processing

Data pre-processing component defines the pre-processing steps of data. Mainly we have three types of steps:

3.2 Data Selection

The selection process involves selecting a set of data marts that are to be used in discovery. The data used in the KDD process is selected based on an evaluation of its potential to yield knowledge. The selected Data sources are individually referred to as Data Marts. Results are highly dependent on the target dataset; therefore care should be taken in selecting the data.

3.3 Data Integration

Data warehousing involves integration of the data. At this stage, multiple data sources are combined in a common source. The data warehouse is designed to centralize data. The data stored in a data warehouse should have the following characteristics: time dependent, non-volatile, subject oriented and integrated.

The integration process is composed of following steps:

- Let’s say there are n data marts
- Take a new data base.
- for each data marts (from 1 to n)
- for each compound in the data mart
- if compound name does not already exist in the database

- add compound to the database
- else
- go to next compound
- end if
- end inner for loop
- end outer for loop

3.4 Data Cleaning

Data cleaning is labour intensive and involves examining the data for completeness and integrity. A few examples of dirty data include: missing field information, names that are the same but spelled slightly different, duplicate data. Thus it is a phase in which noise data and irrelevant data are removed from the collection. The cleaning process is composed of following steps:

- For each structure in the database
- check the missing field in the attributes of the compound
- if missing field found
- take a default value from the user or assign null to that missing field
- end if
- end for loop

3.5 Data Mining

Data Mining component uses the***K-Means Clustering Technique***. K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.K-means clustering is an algorithm to classify or to group your objects based on attributes/features into K number of group. K is positive integer number.The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bar centres of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function is:-

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.

The algorithm is composed of the following steps:

- Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
- Assign each object to the group that has the closest centroidWhen all objects have been assigned, recalculate the positions of the K centroids.
- Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

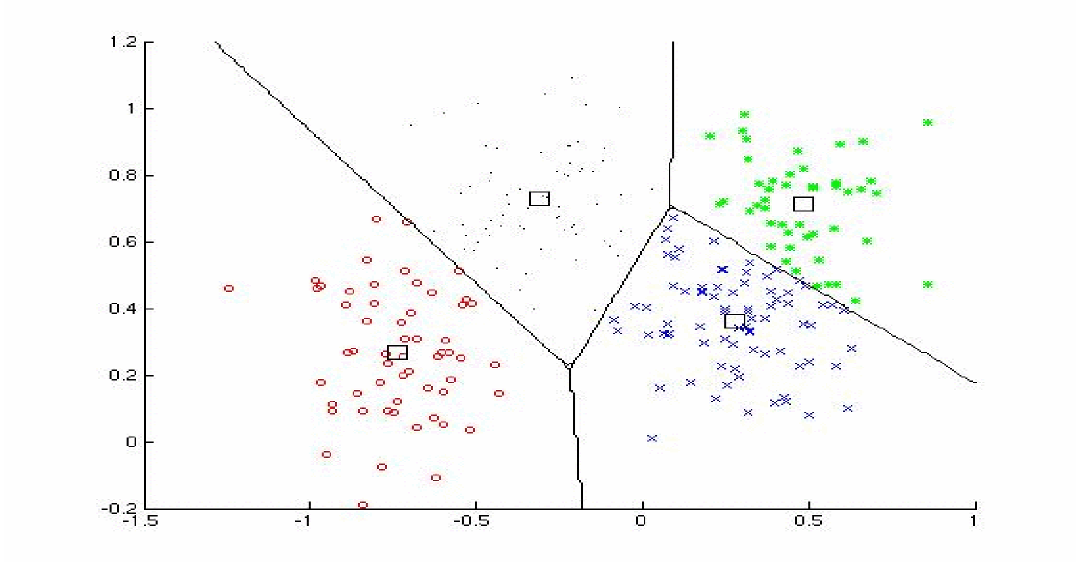

Fig. 1: K-means clustering.

3.6 Pattern Evaluation

In Pattern Evaluation, the clusters of unknown similar drugs are evaluated and compared with the clusters of some specific (e.g. HIV) drugs to discover from unknown drugs those drugs that are similar in properties with the known drugs. To represent them visually is an interesting new feature introduced by our application. The second step in pattern evaluation is the calculation of structural similarity between the drugs using the well-known Tanimoto Coefficient. It compares two drugs on the basis of structural keys that are used to identify features contained in a molecule. The structural fragments or features that are present in the given molecule are turned ON (set as 1) and the ones that are absent are kept OFF (set as 0). Thus, for each molecule one ends up having a string containing 1s and 0s (bit string). These bit strings are then evaluated to find the similarity.

3.7 Smile

Smiles stand for Simplified Molecular Input Line Entry System. It is a notation for entering and representing chemical compounds using short ASCII strings. However the term SMILES is also commonly used to refer to both a single SMILES string and a number of SMILES strings and the exact meaning is usually apparent from the context. The storage space needed for storing the structure of chemical compounds in system is very high, so we make use of smile notations to store the structures of chemical compounds in system. For each chemical structure we have well defined smile representation. From a smile notation we can make structures again by applying some rules and vice versa.

Examples showing conversion of structure to smiles notation:-

Table 1: Smiles Notation Representation of Compounds.

SMILES	NAME	SMILES	Name
CC	ethane	[OH3+]	hydronium ion
O=C=O	carbon dioxide	[2H]O[2H]	deuterium oxide
C#N	hydrogen cyanide	[235U]	uranium-235
CCN(CC)CC	triethylamine	F/C=C/F	E-difluoroethene
CC(=O)O	acetic acid	F/C=C\F	Z-difluoroethene
C1CCCCC1	cyclohexane	N[C@@H](C)C(=O)O	L-alanine

3.8 Fingerprints

Fingerprints are a very abstract representation of certain structural features of a molecule. It is in form of Boolean string consists of 0's and 1's. For making a finger print we have prerequisite of smile notation of some substructure compounds that is compounds that are part of other chemical compounds and chemical compounds whose finger prints have to be found. Many substructure compounds together forms a single chemical compound. We match each substructure of chemical compound with other compounds and notice whether the substructure is the part of that structure, if it is a part of compound then fingerprint is 1 for that substructure otherwise 0 is the fingerprint for that substructure. Likewise all the structures are checked for each substructure and a binary string of 0's and 1's is formed. This string has the length equal to the total number of substructures and it is the finger print of that chemical compound.

The algorithm is composed of following steps:

- For each chemical compound in the database
- Fingerprint = 0
- For each substructure sub in the database
- If sub is subpart of the chemical compound
- Fingerprint + = 1
- Else
- Fingerprint + = 0

- End of inner for loop
- End of outer for loop

3.9 Tanimoto Coefficient

The term Tanimoto coefficient is used to find similarity between the two chemical compounds on the basis of finger prints formed of the compounds. Before discussing it we have to make one point clear that in the finger print of a compound 1 denotes that a substructure is present in that compound and 0 denotes the absence of substructure in the compound. Now the formula of Tanimoto coefficient for finding the similarity between the two chemical compounds A and B is

T= NAB / NA + NB - NAB

Where

NAB denotes the total number of bits that are 1 in the finger prints of both the compounds (A and B) at the same position.

NA denotes the total number of bits that are 1 in compound A.

NB denotes the total number of bits that are 1 in compound B.

The algorithm is composed of following steps:

- Given finger prints of both the compounds whose similarity is to be found and a threshold value of the Tanimoto coefficient above which the two compounds are found to be almost similar in structure and properties.
    - Calculate NAB
    - Calculate NA
    - Calculate NB
    - Calculate Tanimoto coefficient as T= NAB / NA + NB – NAB
    - If T>=threshold
    - Print “Compounds are similar”
    - Else
    - Print “Compounds are not similar”

3.10 Knowledge Representation

The final phase of knowledge representation presents set of rules that would be a picture of the worked done by us, i.e. all the knowledge discovered during clustering and pattern evaluation.

4. Conclusion

About 60% (117) drugs of 200 drugs form a cluster having the M.wt (75 to 407); clogp (-6.4 to 6.0); H.acc (1 to 10); H.don (0 to 5) and about 7% drugs out of 200 drugs form a cluster having the M.wt (110 to 407); clogp (-4.0 to 4.3); H.acc (2 to 8); H.don (0 to 3). There are many other small clusters.

- The Anti-HIV drug Abacavir has similarity with Mafenide, Nikethamide, Sulfaperine, Adenosime Monophosphate.
- The Anti-HIV drug Amprenavir has similarity with Pindolol and Ethanzamide.
- The Anti-HIV drug Deoxynojirmy has similarity with Trimethadione.
- The Anti-HIV drug Stavudine has similarity with Oxycodone and Mafenide.
- The Anti-HIV drug Zalcitabine has similarity with Prednisone and Metharbital.
- The Anti-HIV drug Tenofovir has similarity with Oxycodone, Nikethamide, Sulfaperine and Nicergoline.

References

Books

[1] Jiawei Han and MichelineKamber,“*Data Mining concepts and technology*”.
[2] Paul Raj Poonia, “*Fundamentals of Data Warehousing*”, John Wiley & Sons,2004

Reference Links

[3] http://bioinf.charite.de/superdrug/
[4] http://drugbank.ca/
[5] http://books.google.co.in/books?id=PoJs4C3MgNIC&pg=PA102&lpg=PA102&dq=MACCS+Keys&source=bl&ots=rrZ8rb-cTn&sig=RsaK3mp8Z15yrxZUIf_oheaGy24&hl=en&ei=mF-zSf7kApz47APgkIW6BQ&sa=X&oi=book_result&resnum=5&ct=result#PPA102,M1
[6] http://www.dalkescientific.com/writings/NBN/fingerprints.html
[7] www.qsarworld.com/files/Get_MACCS_Keys.pdf
[8] www.qsarworld.com/files/Tanimoto_Coefficient-1.pdf
[9] http://www.mesaac.com/Fingerprint.htm