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.
Ph : +91-0120-3002219, +91-0-9811443110
Email : jranjan@imt.edu.

ABSTRACT

Almost two decades ago, the information flow in the pharmaceutical industry was relatively simple and the application of technology was limited. However, as we progress into a more integrated world where technology has become an integral part of the business processes, the process of transfer of information has become more complicated. Today increasingly technology is being used to help the pharmaceutical firms manage their inventories and to develop new product and services. The implications are such that by a simple process of merging the drug usage and cost of medicines (after completing the legal requirements) with the patient care records of doctors and hospitals helping firms to conduct nation wide trials for its new drugs. Other possible uses of information technology in the field of pharmaceuticals include pricing (two-tier pricing strategy) and exchange of information between vertically integrated drug companies for mutual benefit. Nevertheless, the challenge remains though data collection methods have improved data manipulation techniques are yet to keep pace with them.

Data mining fondly called patterns analysis on large sets of data uses tools like association, clustering, segmentation and classification for helping better manipulation of the data help the pharma firms compete on lower costs while improving the quality of drug discovery and delivery methods. A deep understanding of the knowledge hidden in the Pharma data is vital to a firm's competitive position and organizational decision-making. The paper explains the role of data mining in pharmaceutical industry.

The paper presents how Data Mining discovers and extracts useful patterns from this large data to find observable patterns. The paper demonstrates the ability of Data Mining in improving the quality of decision making process in pharma industry.

Keywords: Data Mining, drug discovery, pharma industry.

1. INTRODUCTION.

Data Mining 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 (Feelders, Daniels and Holsheimer, 2000). Traditional data analysis methods often involve manual work and interpretation of data that is slow, expensive and highly subjective (Fayyad, Piatsky Shapiro and Smyth, 1996). Data Mining, popularly called as knowledge discovery in large data, enables firms and organizations to make calculated decisions by assembling, accumulating, analyzing and accessing corporate data. It uses variety of tools like query and reporting tools, analytical processing tools, and Decision Support System (DSS) tools.

Recently, numerous experts have predicted that revenue growth for the pharmaceutical industry will slow from the healthy 12% rate to a much slower 5-6% rate. (Cosper Nate, 2003) describes this trend, which is becoming increasingly accepted, has numerous implications for the drug discovery technologies companies. Most significantly, slower revenue growth will necessitate decreased expenditures throughout the organization. Many drug discovery technology companies are attempting to address this challenge through developing solutions that will force new drugs to "fail faster and safer." Although this is a noble goal and if realized, would definitely create value for the industry, these solutions often over-promise and underestimate the obstacles that stand in the path to lower clinical failure rates. Marketing strategies centered on increasing revenues will be more convincing than those that address reducing expense. Demonstrating that technologies will enable pharmaceutical companies to better target and market to certain customer segments will increase adoption of that technology, and will open the door for projects aimed at reducing costs and increasing clinical trials throughput.

The importance of decision support in the delivery of managed healthcare can hardly be overemphasized (Hampshire and Rosborough, 1993). A variety of decision support capabilities will be necessary to increase the productivity of medical personnel, analyze care outcomes, and continually refine care delivery processes to remain profitable while holding the line on costs and maintaining quality of care (Dutta and Heda, 2000). Healthcare decision support is faced with the challenges of complex and diverse data and knowledge forms and tasks (Prins and Stegwee, 2000, Sheng, 2000), the lack of standardized terminology compared to basic sciences, the stringent performance and accuracy requirements and the prevalence of legacy systems (Sheng, 2000).

Data mining the life sciences researcher to mine data to understand safety and efficacy profiles within the patient population. By tackling the question of patient selection within the framework of demonstrating groups that are most responsive, Data mining is sure to penetrate the drug development marketplace. Data mining framework enables specialists to create customized nodes that can be shared throughout the organization, making the application attractive to skilled modelers in a pharmaceutical company's bioinformatics division.

The paper discusses how Data Mining discovers and extracts useful patterns from this large data to find observable patterns. The paper demonstrates the ability of Data Mining in improving the quality of decision making process in pharma industry.

The rest of the paper is organized as follows. Section 2 focuses on data mining and its techniques. Section 3 describes the relevance of data mining techniques in pharma industry. Section 4 briefly explains the difference between statistics and data mining. Section 5 concludes the paper.

2. DATA MINING TECHNIQUES.

Pharma industries rely on decision-oriented, systemic selection models that enable the decision maker to evaluate the payoff that is expected to result from the implementation of a proposed selection program. Such models go beyond an examination of the size of the validity coefficient and take a host of issues such as capital budgeting and strategic outcomes at the group and organizational levels. Many organizations generate mountains of data about their new drugs discovered and its performance reports, etc. This data is a strategic resource. Now, making use of most of these strategic resources will lead to improving the quality of pharma industries.

(Feelders, Daniels and Holsheimer, 2000) give six important steps in the Data Mining process as

1. Problem Definition.
2. Knowledge acquisition.
3. Data selection.
4. Data Preprocessing.
5. Analysis and Interpretation.
6. Reporting and Use.

(Berthold Michael and Hand David, 1999) identify the Data Mining process as

1. Definition of the objectives of the analysis.
2. Selection &Pretreatment of the data.
4. Explanatory analysis.
5. Specification of the statistical methods.
6. Analysis of the data.
7. Evaluation and comparison of methods.
8. Interpretation of the chosen model.

The techniques and methods in Data Mining need brief mention to have better understanding.

(A) ASSOCIATIONS, MINING FREQUENT PATTERNS.

These methods identify rules of affinities among the collections. (Hand, Mannila and Smyth, 2001) mention that patterns occur frequently during Data Mining process. The applications of association rules include market basket analysis, attached mailing in direct marketing, fraud detection, department store floor/shelf planning etc.

(B) CLASSIFICATION AND PREDICTION.

The classification and prediction models are two data analysis techniques that are used to describe data classes and predict future data classes. A credit card company whose customer credit history is known can classify its customer record as Good, Medium, or Poor. Similarly, the income levels of the customer can be classified as High, Low, and Medium. (Adriaans Peiter and Zantinge Dolf, 2005) explain that if we have records containing customer behavior and we want to classify the data or make prediction, we will find that the tasks of classification and prediction are very closely linked. The models of decision trees, neural networks based classifications schemes are very much useful in pharma industry. Classification works on discrete and unordered data, while prediction works on continuous data. Regression is often used as it is a statistical method used for numeric prediction. Primary emphasis should be made on the selection measurement accuracy and predicative efficiency of any new drug discovery. Simple or multiple regressions is the basic prediction model that enables a decision maker to forecast each criterion status based on predictor information. (Smith and Gupta, 2002) show through case studies how neural network technology is useful from different areas of business. We limited our discussion on algorithms and proof here.

(C) CLUSTERING.

It is a method by which similar records are grouped together. Clustering is usually used to mean segmentation. An organization can take the hierarchy of classes that group similar events. Using clustering, employees can be grouped based on income, age, occupation, housing etc. In business, clustering helps identify groups of similarities; characterize customer groups based on purchasing patterns, etc.

3. DATA MINING AND STATISTICS.

The ability to build a successful predictive model depends on past data. Data Mining is designed to learn from past success and failures and will be able to predict what will happen next (future prediction). One may think why use Data Mining in pharma industry organizations when statistical analysis is already been performed. The Data Mining tool checks the statistical significance of the predicted patterns and reports.

Data Mining will tell that it is likely that something unlikely (Berson and Smith, 2005) will happen. If Data Mining tool finds that 100 percent of the drugs of some particular large group have included for the performance analysis, but among them only 10 drugs have the characteristics of high performance ratings, then the tool can warn that it is very likely to be an idiosyncrasy of the data base rather than a usual predictive pattern

The difference between Data Mining and statistics is that Data Mining automates the statistical process requiring in several tools. Statistical inference is assumption driven in the sense that a hypothesis is formed and tested against data. Data Mining, in contrast is discovery driven. That is, the hypothesis is automatically extracted from the given data. The other reason is Data Mining techniques tend to be more robust for real-world messy data and also used less by expert users (Berson et al., 1999).

Data Mining can answer analytical questions such as: what are discovery of new molecules and issues over it? What factors or combinations are directly impacting the drugs? What are the best and outstanding drugs? Which drugs are likely to be retained? How to optimally allocate resources to ensure effectiveness and efficiency? etc. Since the major chunk of literary information is in the form of unstructured text, an intelligent text mining system could provide a platform for extracting and managing specific information at the entity level. For e.g. Information pertaining to genes, proteins, diseases, organisms, chemical substance etc can be analytically extracted for patterns . It would also aid in providing insights into inter-relationships such as protein-protein, Gene-gene, Protein-Chemical, Gene-Disease and Drug-Drug interactions. Text mining can be applied to biomedical literature, clinical documents and other medical literary

sources for data curation and database population in a semi-automated manner.

4. APPLICATIONS OF DATA MINING IN THE PHARMACEUTICAL INDUSTRY

Most healthcare institutions lack the appropriate information systems to produce reliable reports with respect to other information than purely financial and volume related statements (Prins & Stegwee, 2000). The management of pharma industry starts to recognize the relevance of the definition of drugs and products in relation to management information. In the turmoil between costs, care-results and patient satisfaction the right balance is needed and can be found in upcoming information and Communication technology.

The delivery of healthcare has always been information intensive, and there are signs that the industry is recognizing the increasing importance of information processing in the new managed care environment (Morrisey, 1995). Most automated systems are used as a tool for daily work: they are focused on ‘production’ (daily registration). All the data, which are used to keep the organization running, operational data, are in these automated systems. These systems are also called legacy systems. There is a growing need to do more with the data of an organization than to use them for administration only. A lot of information is hidden in the legacy systems. This information can easily be extracted. Most of the times this can not be done directly from the legacy systems, because these are not build to answer questions that are unpredictable. Research shows that (Zuckerman and Alan, 2006); Armoni, 2002; Rada, 2002) that successful decision systems enriched with analytical solutions are necessary for healthcare information systems.

Given the size of the databases being queried, there is likely to be a trade-off in accuracy of information and processing time. Sampling techniques and tests of significance may be satisfactory to identify some of the more common relationships; however, uncommon relationships may require substantial search time. The thoroughness of the search depends on the importance of the query (e.g., life threatening vs. "curious to know"), the indexing structures used, and the level of detail supplied in the query. Of course, the real data mining challenge comes when the user supplies only a minimal amount of information. For example: find possible serious side effects (not necessarily reported in the manufacturer's product literature) involving food and any type or brand of antacid.

A user-interface may be designed to accept all kinds of information from the user (e.g., weight, sex, age, foods consumed, reactions reported, dosage, length of usage). Then, based upon the information in the databases and the relevant data entered by the user, a list of warnings or known reactions (accompanied by probabilities) should be reported. Note that user profiles can contain large amounts of information, and efficient and effective data mining tools need to be developed to probe the databases for relevant information. Secondly, the patient's (anonymous) profile should be recorded along with any adverse reactions reported by the patient, so that future correlations can be reported. Over time, the databases will become much larger, and interaction data for existing medicines will become more complete.

The amount of existing pharmaceutical information (pharmacological properties, dosages, contraindications, warnings, etc.) is enormous; however, this fact reflects the number of medicines on the market, rather than an abundance of detailed information about each product.

One of the major problems with pharmaceutical data is actually a lack of information. For example, an food and drug administration department estimated that only about 1% of serious events are reported to the food and drug administration department. Fear of litigation may be a contributing factor; however, most health care providers simply don't have the time to fill out reports of possible adverse drug reactions. Furthermore, it is expensive and time-consuming for pharmaceutical companies to perform a thorough job of data collection, especially when most of the information is not required by law. Finally, one should note that the food and drug administration department does not require manufacturers to test new medicines for potential interactions.

There are in general three stages of drug development namely finding of new drugs, development tests and predicts drug behavior, clinical trials test the drug in humans and commercialization takes drug and sells it to likely consumers (doctors and patients).

DEVELOPMENT OF NEW DRUGS.

This research need to use data mining tools and techniques. This can be achieved by clustering the molecules into groups according to the chemical properties of the molecules via cluster analysis (Cooman, 2005). This way every time a new molecule is discovered it can be grouped with other chemically similar molecules. This would help the researchers in finding out with therapeutic group the new molecule would belong to. Mining can help us to measure the chemical activity of the molecule on specific disease say tuberculosis and find out which part of the molecule is causing the action. This way we can combine a vast number of molecules forming a super molecule with only the specific part of the molecule which is responsible for the action and inhibiting the other parts. This would greatly reduce the adverse effects associated with drug actions.

Scientists run experiments to determine activity of potential drugs. They use high speed screening to test tens, hundreds, or thousands of drugs very quickly. The general goal is to find activity on relevant genes or to find drug compounds that have desirable characteristics (whatever those may be). The Data mining techniques that are used in developing of new drugs are clustering, classification and neural networks. The basic objective is to determine compounds with similar activity. The reason is for similar activity compounds behave similarly. This is possible only when we have known compound and looking for something better. When we don't have known compounds but have desired activity and want to find compound that exhibits this activity, then data mining rescues this.

DEVELOPMENT TESTS AND PREDICTS DRUG BEHAVIOR

There many issues which affect the success of a drug which has been marketed which can impact the future development of the drug. Firstly adverse reactions to the drugs are reported spontaneously and not in any organized manner. Secondly we can only compare the adverse reactions with the drugs of our own company and not with other drugs from competing firms. And thirdly we only have information on the patient taking the drug not the adverse reaction that the patient is suffering from. All this can be solved with creation of a data warehouse for drug reactions and running business intelligence tools on them a basic classification tool can solve much of the problems faced here. We could find out the adverse reactions associated with a specific drug and still go a step further to show if any specific condition aggravates the adverse reaction for eg age, sex, and obesity (Novartis Business Intelligence report, 2004). This could help the medical practitioner to describe the side effects to the patients being prescribed these drugs.

Pharma companies think that drugs might have some yielded benefits. The drug undergoes testing in animals and human tissue to observe effect and determines how much drug to consume for desired effect or how dangerous is the drug. The Data mining techniques can be here used is classification and neural networks. The goal here is to predict if treatment will aid patients. Because if drug will not aid patients, what purpose does drug serve. Predicting the drug behavior is essential when we have data supporting use of drug and also have training data that shows effects of drug (positive or negative). The test should be able to predict which patients will benefit and which treatment help sickle cell anemia patients. The information like gender, body weight, disease state, etc will play crucial role. This crucial data should be fed into neural network and predict whether patient will benefit from drug. Only one of two classifications *yes/no* will be available on training data. Network is trained for the *yes* classifications and a snapshot is taken of the neural network. Then network is trained for the *no* classifications and another snapshot is taken. The output is *yes* or *no*, depending on whether the inputs are more similar to the *yes* or the *no* training data.

CLINICAL TRIALS TEST THE DRUG IN HUMANS

Company tests drugs in actual patients on larger scale. The company has to keep track of data about patient progress. The Government wants to protect health of citizens, many rules govern clinical trials. In developed countries food and drug administration oversees trials. The Data mining techniques used here can be neural networks. Here data is collected by pharmaceutical company but undergoes statistical analysis to determine success of trial. Data is generally reported to food and drug administration department and inspected closely. Too many negative reactions might indicate drug is too dangerous. An adverse event might be medicine causing drowsiness. As a matter of fact, Data mining is performed by food and drug

administration, not as much by pharmaceutical companies. The goal is to detect when too many adverse events occur or detect link between drug and adverse event. Too many adverse events linked to a drug might indicate drug is too dangerous or health of patient is at risk. Adverse events are reported to food and drug administration when link is suspected. One can feed the information on drug causing too many adverse events pertaining to drugs into a neural network and let network lead us to what is meant by 'too many'.

5. OUTCOME RESEARCH

The effectiveness of the drug is often measured by how soon the drug deals with the medical condition (Joe and Teresa, 1996). A simple association technique could help us measure the outcomes that would greatly enhance the patient's quality of life say for e.g. faster restoration of the body's normal functioning. This could be a benefit much sought after by the patient and could help the firm better position the drug vis-à-vis the competition.

6. CONCLUSION.

Imagine that all molecules being created in pharma industry are grouped using clustering analysis into *groups* via the characteristic chemical properties of the molecule. Once that these are grouped, one could find the elements most influencing the probability of belonging to group one and not to group two.

Normally, for proving the efficiency of a drug, the rules f are described. For example a drug for hay fevers, where one knows how to measure (relief) to compare the drugs. With data mining techniques, we could try to find alternative measures of relief, and promote the drug in another way: on top of curing the disease in a standard way, with our drug you get some extras compared to the competitor.

Until now, most of the statistics are used in the R&D department. It is also observed that Data mining techniques are seldom used in a pharmaceutical environment. This paper described that these techniques can be easily and successfully used. The paper presented on how Data mining discovers and extracts useful patterns from this large data to find observable patterns. The paper demonstrates the ability of Data Mining in improving the quality of decision making process in pharma industry.

REFERENCES

[1]. 'Novartis Business Intelligence report' Cognos press 2004 , www.cognos.com

[2]. Adriaans Peiter, Zantinge Dolf. (2005) *Data Mining*, Pearson Education, pp.69-71.

[3]. Armoni,A. (2002) 'Effective Healthcare information systems', IRM Press.

[4]. Berson A and Smith, S.J (2005) *Data Warehousing, Data Mining & OLAP*, Tata McGraw-Hill Edition, pp.5-6.

[5]. Berson A., Smith, S. and Thearling, K. (1999) *Building Data Mining Applications for CRM*. McGraw-Hill Professional.

[6]. Berthold Michael and Hand David, J. (1999) *Intelligent Data Analysis: An Introduction*. Springer, pp.3-10.

[7]. Cooman De Frankey. (2005) 'Data mining in a Pharmaceutical Environment', Belgium press.

[8]. Cosper Nate. (2003) 'Insightful Strategies For Increasing Revenues In The Pharmaceuticals Industry: Data Mining for Successful Drugs', retrieved 21 august from http://www.frost.com/prod/servlet/market-insight-top.pag?docid=5701902.

[9]. Dutta, A. and Heda, S. (2000) 'Information systems architecture to support managed care business process', *Decision Support Systems*, 30, pp 217-225.

[10]. Fayyad, U.M., Piatsky Shapiro, G. and Smyth, P. (1996) *From Data Mining to Knowledge Discovery in Data Base*, AI Magazine, pp.37-54.

[11]. Feelders, A., Daniels, H. and Holsheimer, M. (2000) 'Methodological and Practical Aspects of Data Mining', *Information and Management*, pp.271-281.

[12]. Hampshire, D. A. and Rosborough, B. J. (1993) 'The evolution of decision support in a manged care organization', *Topics in healthcare financing*, 20, 2, pp26-37.

[13]. Hand David, Mannila Heikki and Smyth Padhaic. (2001) *Principles of Data Mining*, The MIT Press, pp.9-10.

[14]. Joe and Teresa Graedon. (1996) 'People's guide to deadly drug interactions', New York St Martin's press.

[15]. Morrisey, J. (1995) 'Managed care steers info systems', *Modern Healthcare*, Vol-25, 8.

[16]. Prins, S and Stegwee, R. A. (2000) 'Zorgproducten en geintegreerde infromatiesystemen', *(in Dutch) Handboek sturen met zorgproducten*, F3100-3, december.

[17]. Rada,R. (2002) 'Information systems for Healthcare enterprises', Hypermedia Solutions Limited.

[18]. Roy Levy (1999) 'Pharmaceutical Industry :A discussion of legislative and anti trust issues in an environment of change', Federal trade commission report.

[19]. Sheng, O. R. Liu (2000) 'Decision support for healthcare in a new information age', *Decision Support Systems*, 30, pp101-103.

[20]. Smith Kate and Gupta Jatinder. (2002) *Neural Networks in Business: Techniques and Applications*, IGI Publishing.

[21]. Zuckerman and Alan, M. (2006) 'Healthcare Strategic Planning', *Prentice Hall of India*.