Data Mining and Warehousing in Pharma Industry

Andrew Kusiak
The University of Iowa, USA

Shital C. Shah
The University of Iowa, USA

INTRODUCTION

Most processes in pharmaceutical industry are data driven. Company's ability to capture the data and making use of it will grow in significance and may become the main factor differentiating the industry. Basic concepts of data mining, data warehousing, and data modeling are introduced. These new data-driven concepts lead to a paradigm shift in pharmaceutical industry.

BACKGROUND

The volume of data in pharmaceutical industry has been growing at an unprecedented rate. For example, a microarray (equivalent of one test) may provide thousands of data points. These huge datasets offer challenges and opportunities. The primary challenges are data storage, management, and knowledge discovery. The pharmaceutical industry is one of the most data-driven industries and getting value out of the data is key to its success. Thus processing the data with novel tools and methods for extracting useful knowledge for speedy drug discovery and optimal matching of the drugs with the patients may well become the main factor differentiating the pharmaceutical companies.

The main sources of pharmaceutical data are patients, genetic tests, regulatory sources, pharmaceutical literature, and so on. Data and information collected from different sources, agencies, and clinics, has been traditionally used for narrow reporting (Berndt et al., 2001). Massive clinical trials may lead to errors such as protocol violations, data integrity, data formats, transfer errors, and so on. Traditionally the analysis in pharmaceutical industry [from prediction of drug stability (King et al., 1984) to drug discovery techniques (Tye, 2004)] is performed using the population based statistical techniques. To minimize possible errors and increase confidence in predictions there is a need for a comprehensive methodology for data collection, storage, and analysis. Automation of the pharmaceutical data over its lifecycle seems to be an obvious alternative.

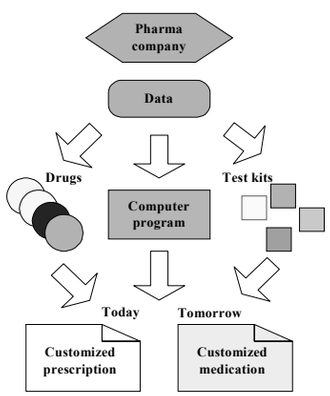
The vision for pharmaceutical industry is captured in Figure 1. It has become apparent that data will play a central role in drug discovery and delivery to an individual patient. We are about to see pharmaceutical companies delivering drugs, developing test kits (including genetic tests), and computer programs to deliver the best drug to the patient. The current data mining (DM) tools are capable of issuing customized prescriptions, providing most effective drugs, and dosages with minimal adverse effects. It has become practical to think of designing, producing, and delivering drugs intended for an individual patient (or a small population of patients). Here, analogy of the one-of-a-kind production of today versus mass production is worthy attention. Many industries are attempting to produce customized products. The pharmaceutical industry may soon become a new addition to this collection of industries.

MAIN THRUST

Paradigm Shift

Predicting implies knowing in advance, which in a business environment translates into competitive advantage.

Figure 1. The future of pharmaceutical industry



Copyright © 2006, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.

A future event can be predicted in two major ways: population-based and individual-based. The population-based prediction says, for example, a drug A has been effective in treating 80% of patients in the population P as symbolically illustrated in Figure 2. Of course, any patient would like to belong to the 80% rather than the 20% category before the drug A is administered. Statistics and other tools have been widely used in support of the population-paradigm, among others in medicine and pharmaceutical industry.

The individual-based approach supported by numerous DM algorithms emphasizes an individual patient rather than the population (Kusiak et al., 2005). One of many decision-making scenarios is illustrated in Figure 3, where the original population P of patients has been partitioned into two segments 1 and 2. The decisions for each patient in Segment 1 are made with high confidence; say 99%, while the decisions for Segment 2 are predicted with lower confidence.

It is quite possible that Segment 2 patients would seek an alternative drug or a treatment. There are different ways of using DM algorithms. They cover the range between the population and individual-based paradigms. The exiting DM algorithms can be grouped into the following basic ten classes (Kusiak, 2001):

A. Classical statistical methods (e.g., linear, quadratic, and logistic discriminant analyses)
B. Modern statistical techniques (e.g., projection pursuit classification, density estimation, *k*-nearest neighbor, Bayes algorithm)

Figure 2. Population-based paradigm

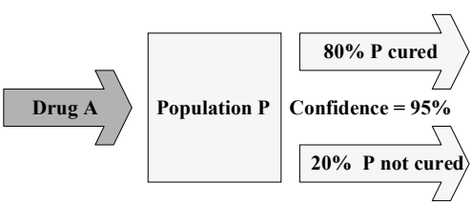
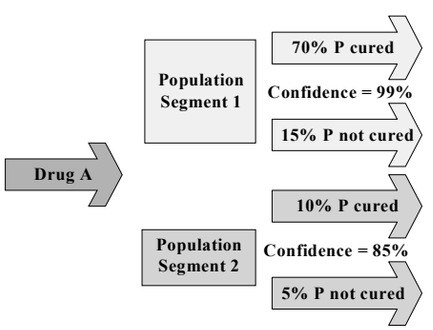


Figure 3. Individual-based paradigm using data mining tools



C. Neural network (Mitchell, 1997)
D. Support vector machines
E. Decision tree algorithms [C4.5 (Quinlan, 1992)]
F. Decision rule algorithms [Rough set algorithms (Pawlak, 1991)]
G. Association rule algorithms
H. Learning classifier systems
I. Inductive learning algorithms
J. Text learning algorithms

Each class containing numerous algorithms, for example, there are more than 100 implementations of the decision tree algorithm (class E).

Data Warehouse Design

A warehouse has to be designed to meet users' requirements. DM, online analytical processing (OLAP), and reporting are the top items on the list of requirements. Systems design methodologies, and tools can be used to facilitate the requirements capture. Examples of methodologies for analysis of data warehouse (DW) requirements include AND/OR graphs and the house of quality (Kusiak, 2000).

The architecture of a typical DW embedded in a pharmaceutical environment is shown in Figure 4. The pharmaceutical data is extracted from numerous sources and preprocessed to minimize inconsistencies. Also, data transformation will capture intricate solution spaces to improve knowledge discovery. The cleaned and transformed data is loaded and refreshed directly into a DW or data marts. A data mart might be a precursor to the full-fledged DW or function as a specialized DW. A special purpose (exploratory) data mart or a DW might be created for exploratory data analysis and research. The warehouse and data marts serve various applications that justify the development and maintenance cost in this data storage technology. The range of services that could be developed off a DW could be expanded beyond OLAP and DM into almost all pharmaceutical business areas, including interactions with federal agencies and other businesses.

Data Flow Analysis

A task that may parallel the capture of requirements for a DW involves analysis of data flow. A warehouse is to integrate various streams of data that have to be identified. The information analysts and users need to feel comfortable with the data flow methodology selected for capturing the data flow logic. At minimum this data flow modeling exercise should increase efficiency of the data handling and management. An example to a methodology that can be used to model data flow is the

4 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/data-mining-warehousing-pharma-industry/10600

Related Content

Query Optimisation for Data Mining in Peer-to-Peer Sensor Networks

Mark Roantree, Alan F. Smeaton, Noel E. O'Connor, Vincent Andrieu, Nicolas Legeayand Fabrice Camous (2010). *Intelligent Techniques for Warehousing and Mining Sensor Network Data (pp. 234-256).*

www.irma-international.org/chapter/query-optimisation-data-mining-peer/39548

Biological Data Mining

George Tzanis, Christos Berberidisand Ioannis Vlahavas (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 1696-1705).*

www.irma-international.org/chapter/biological-data-mining/7725

Preference-Based Frequent Pattern Mining

Moonjung Cho, Jian Pei, Haixun Wangand Wei Wang (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 1280-1299).*

www.irma-international.org/chapter/preference-based-frequent-pattern-mining/7699

Intelligence Density

David Sundaramand Victor Portougal (2005). *Encyclopedia of Data Warehousing and Mining (pp. 630-633).*

www.irma-international.org/chapter/intelligence-density/10673

Classification Methods

Aijun An (2005). *Encyclopedia of Data Warehousing and Mining (pp. 144-149).*

www.irma-international.org/chapter/classification-methods/10582