Minireview

Genomics & Informatics Vol. 1(2) 65-74, December 2003

Data Mining for High Dimensional Data in Drug Discovery and Development

Kwan R. Lee*, Daniel C. Park, Xiwu Lin and Sergio Eslava

GlaxoSmithKline, Research & Development, Data Exploration Sciences 1250 South Collegeville Road Collegeville, PA 19426, USA

Abstract

Data mining differs primarily from traditional data analysis on an important dimension, namely the scale of the data. That is the reason why not only statistical but also computer science principles are needed to extract information from large data sets. In this paper we briefly review data mining, its characteristics, typical data mining algorithms, and potential and ongoing applications of data mining at biopharmaceutical industries. The distinguishing characteristics of data mining lie in its understandability, scalability, its problem driven nature, and its analysis of retrospective or observational data in contrast to experimentally designed data. At a high level one can identify three types of problems for which data mining is useful: description, prediction and search. Brief review of data mining algorithms include decision trees and rules, nonlinear classification methods, memory-based methods, model-based clustering, and graphical dependency models. Application areas covered are discovery compound libraries, clinical trial and disease management data, genomics and proteomics, structural databases for candidate drug compounds, and other applications of pharmaceutical relevance.

Keywords: data mining; high dimensional data, genomics; proteomics, pharmacogenomics

Introduction

Data mining has generated significant interest recently, both in industry at large and in research laboratories and academia. What exactly is "data mining"? Unfortunately there is no precise answer. Loosely it can be described as the application of statistical and computer science principles to the problem of extracting information from large data sets. A key point is that neither statistics on its own, nor computer science on its own, are sufficient (for typical applications) as the sole basis for data mining. Rather, statistics provides the component of a data mining algorithm, which takes care of the details of how one fits parameters and models to data. There is a vast body of work in statistics on parameter and model estimation which can be (and often is) leveraged to good effect in data mining. The computer science component of a data mining algorithm handles the storage and access of information in an efficient manner and the heuristic (search) component of the parameter and model-fitting algorithm. Again, there is a vast store of techniques for optimization and search that can be leveraged.

A good example of the interplay between computer science and statistics is data mining using rule induction. The idea is to find rules of the form, "If $A=a$ and $B=b$ then $C=c$" with high probability from the data. Clearly one needs statistical methods to determine reliably (with some statistical confidence) which rules are worthwhile and which are noisy (it will be a function of how often the left-hand side of the rule occurs and how accurate it is). But a statistical quality measure for rules is almost useless on its own for this problem since there are so many possible rules to search over (if we are searching for rules relating K variables of interest). In fact the search space explodes in a combinatorial fashion with K, so it is critical that we have an efficient search method to prevent the enumeration of all possible patterns. Typical search methods used for this problem would be "branch and bound" where one can bound the quality of the solution in large parts of the search space (without actually searching there) and ignore low quality regions in this manner. The key point here is that it is the marriage of statistical and computer science techniques that allows this problem to be solved. There are many such rule induction algorithms available in the market and in the research literature. Those that rely only on statistical methods, or only on computer science methods, have been found to be inferior in performance to those algorithms that take advantage of methods from both fields.

In this paper we will review data mining, its characteristics, typical data mining algorithms, and

* Corresponding author:
E-mail kwan.lee@gsk.com, Tel +610-917-4041, Fax +610-917-4716

Accepted 29 November 2003

potential applications of data mining at biopharmaceutical industries. This paper is intended to be a brief overview of the field: for more in-depth discussions the reader is directed towards the collections of papers (Piatetsky-Shapiro G. *et al*., 1991 and Fayyad U.M. *et al*., 1996), edited Proceedings of the International Conferences on Knowledge Discovery and Data Mining (Fayyad U.M. *et al*., 1995 and Simoudis E. *et al*., 1996). Useful overviews are contained in the papers (Decker K.M. *et al*. ,1995, Mannila H., 1996, Glymour C. *et al*., 1996, Elder J. *et al*., 1996, and Fayyad U.M. *et al*., 1996). Chatfield C. presents the more traditional statistical viewpoint on data mining. Recently Friedman *et al*. gave a general review on data mining opportunities for statisticians working in biopharmaceutical industries which is complementary to our presentation here.

Finally, the reader should be aware of the excellent resources available on the World Wide Web in relation to data mining. The website at www.kdd.gte.com is an excellent resource for general information on data mining, technical reports, and pointers to many publicly available software systems. The website at www.kdnuggets.com is also another useful site on data mining.

Why do we need data mining?

Traditional methods of turning data into useful information rely heavily on manual analysis and interpretation. Let us consider two examples. In the first example, planetary geologists at NASA have been collecting images of the planets for decades from remote spacecraft. The images are painstakingly examined and catalogs are compiled which tabulate the location, size and characteristics of various geologic features of interest (such as craters, volcanoes, etc.). In the last few years the volume of available data has increased by more than 2 orders of magnitudes. For example, for the planet Venus, the Magellan spacecraft returned more than 30,000 one-megabyte images, far more than all previous planetary missions combined. Planetary geologists are swamped with data and are cataloging only certain fractions of the planet or generating low-resolution catalogs (ignoring the high-resolution data).

The second example involves the health-care industry. Specialists analyze current trends and changes in healthcare data on a quarterly basis. The specialists then provide a report detailing the analysis to the sponsoring healthcare organization and this report is used as the basis for future decision-making and planning for health-care management. The problem is that as more detailed data is collected routinely on each patient, more and more specialists are needed to analyze the data and a potentially important fraction of significant patterns are being missed.

There are many, many other fields where similar data analyzes are carried out and where only a fraction of the data can be examined. For these types of problems, manual probing of a data set can be slow, inefficient, expensive, and highly subjective. This type of manual data analysis is becoming completely impractical in many domains as data volumes grow exponentially. Who could be expected to digest millions of data points, each having potentially thousands of fields? Much of the current interest in data mining is very much problem-driven with users in a diverse set of business, science, medicine, engineering, and government communities demanding new ways to navigate and understand large data sets.

Characteristics of data mining

We have said earlier that it is difficult to precisely define data mining in a manner that makes it distinct from its "parent" disciplines of statistics and computer science. However, we can identify some specific characteristics of data mining which are relatively unique to the field:

Understandable models

The fundamental goal of data mining is to extract information (or "knowledge") from data. In this context, it is critical that the patterns and models produced by a data mining algorithm are understandable to the user. Thus, for example, data mining algorithms tend to focus on models and patterns that can be expressed easily (perhaps visually) to the human user, such as rules, trees, graphs, and clusters in data space. In contrast, models composed of non-linear equations (such as nonlinear regression or neural networks) are relatively difficult to interpret and tend not to be used in data mining systems. It is important here to note the importance of goals. If one's goal is to build a black box which is the most accurate model possible, then understandability of the black box may be irrelevant. However, as in data mining, one may be more interested in the qualitative structure of the model and would prefer an understandable model that is reasonably accurate over an incomprehensible one which is slightly more accurate.

Scalability to large problems

One of the key features of many (but not all) data mining applications is the sheer size of the data sets being used. Size can largely be measured in 2 dimensions (at least for "flat file" data sets): the number of variables (say K) and the number of data points or samples (say N). The number of variables K can be in the thousands, or even tens of thousands For example, a database on Alzheimer's patients at UC Irvine contains about one thousand variables measured per patient including variables such as age, employment status, questionnaire responses, cognitive test responses, and physical test results. Having to analyze data with 1000 variables is beyond the scope of many standard statistical techniques. Furthermore, this many variables make it largely impossible for a human to explore the data (e.g., visualizing it) in any meaningful manner. Thus, data mining algorithms typically are designed "upfront" to handle very large numbers of variables, and are often well-suited to the problem of finding low-dimensional structure in high-dimensional problems. The other aspect of size, the number of data points N, can also challenge conventional data analysis methods. It is not unusual to hear of data sets where N is in the billions for science applications for example. Clearly conventional algorithms and software simply can not handle data on this scale. Again, data mining focuses on specific solutions to large N problems: solutions such as sub-sampling, problem decomposition for parallel computing, etc., are being actively explored and applied.

Problem driven nature

Data mining is a field that is driven by practical problems, specifically the problem faced by many organizations of having data sets that are too large to explore manually. In this sense, it is an applications-oriented field, rather than a theory-driven field. (As an example, many of the most well-known researchers in the field work in applied research at industry labs rather than in academia). The impact of this is that data mining is more fragmented and uses fewer common formalisms (e.g., for describing algorithms) than other fields. Data mining borrows techniques liberally from statistics, database theory, artificial intelligence, operations research, etc., and focuses on the practical problem of how these techniques can be integrated and applied to particular problems. Thus, there is relatively little fundamental research going on in the field: research ideas are borrowed and "cannibalized" from other areas. Nonetheless there are some fundamental theoretical problems that are relatively unique to data mining: for example, controlling the interaction of search and statistical significance. Overall, however, the field is characterized by a focus on solving practical problems in specific domains.

Retrospective data analysis

Unlike statistics, data mining typically analyzes data in a retrospective fashion. Rather than designing an experiment, collecting data, and then testing a hypothesis (as in statistics), data mining is usually applied to problems where the data has been collected in the past or in some manner that is outside the control of data analyst. In this sense, data mining can not replace designed experiments/analysis: this is where traditional statistics must be applied. In fact, one can view data mining and statistics as being complementary to each other in this regard: patterns found during data mining may need to be confirmed by more traditional designed experiments (if possible) at a later stage.

In this context, data mining has much in common with exploratory data analysis techniques in statistics. Rather than approaching a data set with a predefined set of hypotheses, one "lets the data speak" and considers a large number of possible models and methods for representing structure in the data. Clearly this requires some discipline in order to prevent fitting noise in the data. Such overfitting was termed "data-dredging" or "data fishing" in statistics in the 1960's. The more modern view is that as long as certain precautions are taken (such as cross-validated sampling methods to test for generality), it is possible to reliably find structure in data without having strong a priori hypotheses on what expects to find. This of course is the "holy grail" of data mining: finding unexpected and valuable "nuggets" of knowledge.

Local patterns and global models

Another feature of data mining algorithms (as implemented today) is that they often produce lists of patterns of the form "if X increases Z also increases" as the output of the algorithm. This is in contrast to a fully specified model for the data, which in this case might be $Z = aX + bY$ where X, Y, and Z are variables and a and b are parameters of the model. The advantage of searching for local patterns is that they can be more robust than full modeling. In the above example the model must assume a linear form while the pattern is much more non-committal. Thus, pattern-finding can be very useful for suggesting structure in low-level data without having to make any strong model assumptions. However, the downside is that the process of finding patterns is not as well-founded from a statistical viewpoint, and furthermore, the sets of patterns produced by many data mining algorithms can be redundant, dependent, and very difficult to understand. Nonetheless, pattern-based data mining is a useful addition to the conventional repertoire of exploratory data analysis techniques.

Non-standard data types

Measurements on variables can be real-valued (continuous), integer-valued (discrete), categorical (names), and so forth. In addition one can also have richer data types such as text, sequences and time series, audio, images, video, etc. Traditional methods (such as much of statistics) often deal with only one fixed data type such as real-valued data. In practice, however, there is an increasing number of practical applications where the data is naturally represented across several different data types. For example, in medicine, a patient may have real-valued variables describing blood pressure, categorical variables such as ethnicity and sex, free text containing comments written by specialists during a consultation, recordings of biomedical monitoring over time, and diagnostic images. Furthermore, any of these data types may be annotated with physician diagnoses (perhaps multiple diagnoses over time). How can one handle, integrate, and model data from vastly different sources? As yet, there are relatively few techniques for handling such non-standard heterogeneous data sources. However, these data sets are the type of data to which data mining is being applied and for which more conventional analysis methods do not exist.

What can one do with data mining?

At a high level one can identify three types of problems that data mining is useful for, namely: description, prediction, and "retrieval by content." This 3-way separation is not perfect (there is a fair degree of overlapping between these types of problems) but nonetheless the breakdown helps to clarify the main application areas to which data mining is applied.

Description means that one is interested in the extraction and an understanding of one's data, e.g., the structure of how the variables in one's database relate to each other. Which variables are directly dependent on which other variables? Which variables are relatively independent? Which variables can be grouped together? Is there an understandable model that can effectively simulate or generate the data we have? These types of questions are typically answered by a variety of techniques in exploratory data analysis known as clustering, dependency analysis, density estimation and so forth. The key point here is that the information extracted by the model is intended to be presented and interpreted by a human user. This is probably the main application of data mining today since it directly addresses the problem of finding interpretable information in large data sets.

Prediction is the process of building a model which can be used for generating future predictions for some variable of interest whose value is unknown, e.g., predicting the value of the stock market tomorrow given today's economic indicators, or classifying a patient as having a particular disease or not based on a set of diagnostic test results. This is the traditional domain of statistical methods, e.g., regression, discrimination, and so forth. The emphasis here is usually on the performance of the model (how well does it predict?) rather than on understanding the data or the model. Thus, predictive modeling could be viewed as somewhat ancillary to much of data mining since a predictive model need not provide the user with any insight into the data-generating process (but it can still be very useful).

In practice the interplay between predictive and descriptive modeling is quite close. In most predictive modeling applications, there is usually a desire on the part of the user to understand how the model works and why it works at all. Conversely, in descriptive modeling there is also the element of predictive accuracy in the sense that one can evaluate a descriptive model in terms of how well it would describe unseen data. For large-scale practical problems with many variables, descriptive modeling may be used to gain insight into the structure of the problem and to guide the user in the model selection process before a predictive model is applied.

The final primary category of data mining applications is "retrieval by content." This is best described by an example. A neurologist examines Magnetic Resonance Images (MRIs) looking for reductions in hippocampal volume that may indicate the onset of Alzheimer's disease. He/she may have access to MRIs from thousands of patients, a large database of MRIs that is impossible to search or annotate manually. The neurologist finds an interestingly looking visual pattern in the hippocampal region of one particular patient. What he/she would like to do is then to explore the database for other similar patterns: "find me the 10 patterns which look most like this" or "find me the 10 patients which have patterns like this and similar cognitive test scores". For such image data (and other "non-standard" data types) it is very difficult to translate the human-level notion of similarity into algorithmic constraints. The relevant questions in such problems are how can one define notions of pattern similarity and how can one efficiently and effectively organize the search to find patterns of interest. Data mining systems for this problem can be constructed either by building models for the domain or by applying appropriate distance metrics to the raw data.

There are a variety of other data mining applications which do not naturally fall under the above categories, including "change detection" (detecting whether changes in data have occurred over time) and detecting unusual patterns (such as detecting fraud).

An outline of some popular data mining algorithms

Naturally there are some data mining algorithms which have been found useful across a broad variety of problems. In this section we provide a very brief overview of some of these algorithms.

Decision Trees and Rules

Decision trees and rules have a simple representational form, making the inferred model relatively easy to comprehend by the user. However, the restriction to a particular tree or rule representation can significantly restrict the functional form (and thus the approximation power) of the model. If one enlarges the model space to allow more general expressions (such as multivariate hyperplanes at arbitrary angles), then the model is more powerful for prediction but may be much more difficult to comprehend. There are a large number of decision trees and rule induction algorithms described in the machine learning and applied statistics literature (Quinlan J.R., 1993, Breiman L. *et al.*, 1984, and Michie D. *et al.*, 1994). To a large extent they are all based on likelihood-based model evaluation methods with varying degrees of sophistication in terms of penalizing model complexity. Greedy search methods, which involve growing and pruning rule and tree structures, are typically employed to explore the super-exponential space of possible models. Trees and rules are primarily used for predictive modeling, both for classification and regression, although they can also be applied to descriptive pattern generation. A popular application of rule learning techniques in data mining is that of "association rules" (Agrawal R. *et al.*, 1995 and Mannila H. *et al.*, 1996) which look for patterns of the form "if A and B occur then C also occurs", etc. Most tree and rule learning algorithms are predictive rather than descriptive: see Smyth P. *et al.* for a description of how descriptive rules can be learned from data.

Nonlinear Regression and Classification Methods

These methods consist of a family of techniques for prediction that fit linear and non-linear combinations of basis functions (sigmoids, splines, and polynomials) to combinations of the input variables. Examples include feedforward neural networks, adaptive spline methods, projection pursuit regression, and so forth. Consider neural networks, for example. In terms of model evaluation, while networks of the appropriate size can universally approximate any smooth function to any desired degree of accuracy, relatively little is known about the representation properties of fixed size networks estimated from finite data sets. In terms of model evaluation, the standard squared error and cross entropy loss functions used to train neural networks can be viewed as log-likelihood functions for regression and classification respectively. The backpropagation technique corresponds to a parameter search method that performs gradient descent in parameter space to find a local maximum of the likelihood function starting from random initial conditions. Given the approximation power of the underlying non-linear model, nonlinear regression methods often provide excellent predictors from data for both classification and regression functions: conversely, however, these models can be very difficult to interpret and, consequently, have found limited application in data mining applications.

Memory-based Methods

The representation is simple: use representative examples ("memory") from the database to approximate a model, i.e., predictions on new examples are derived from the properties of "similar" examples in the model whose prediction is known. Techniques include nearest-neighbor classification and regression algorithms and case-based reasoning systems. A potential disadvantage of example-based methods (compared with tree-based methods for example) is that a well-defined distance metric for evaluating the distance between data points is required. Model evaluation is usually based on cross-validation estimates of a prediction error: "parameters" of the model to be estimated can include the number of neighbors to use for prediction and the distance metric itself. Like non-linear regression methods, example-based methods are often asymptotically quite powerful in terms of approximation properties, but conversely can be difficult to interpret since the model is implicit in the data and not explicitly formulated. Related techniques include kernel density estimation for descriptive modeling of joint probability densities.

Probabilistic (Model-Based) Clustering

This is a descriptive modeling technique where one wishes to group one's data in some manner into "natural" groups or clusters. Because it is difficult to formally specify what a natural grouping is, non-probabilistic clustering algorithms are often quite ad hoc and difficult to compare. Probabilistic clustering on the other hand assumes that the data are being generated by a probabilistic model, proceeds to find the parameters of this model, and identifies the component densities in the model as clusters. The most commonly used representation is that of linear mixtures: the data is assumed to have been generated by a linear combination of M component densities, often chosen to be Gaussian. The fit function used is maximum likelihood, and the search is a technique from statistics known as the Expectation-Maximization procedure which is an effective method for finding a local likelihood maximum in parameter space with hidden data (here the hidden data are the "labels" telling us which data point belongs to which cluster). Determining M, the number of clusters, is quite difficult: recent work has shown that Bayesian and cross-validation techniques are useful in this regard (Smyth P. *et al.*, 1996).

Probabilistic graphical dependency models

Graphical models consist of a graph (with a node for each variable) where the links in the graph show the dependency relations that exist in a joint probability distribution over the variables. There are several different types of graphical models, depending on the type of graph used (directed, undirected, mixed) and the form of the independence assumptions (arbitrary, Markov, etc.). One of the most widely used examples is a (so-called) Bayesian network (also known as belief networks): see Heckerman D. *et al*. for an overview of how such networks can be learned from data. Another well known class of graphical models is hidden Markov models: see Smyth P. *et al*. for a review of hidden Markov models within a graphical modeling framework.

One of the primary advantages of graphical dependency models is their understandability: a graph, where nodes represent variables, and links represent dependencies, can be a very clear and insightful way to visualize the structure of a model. Graphical models are typically used with categorical or discrete-valued variables, but extensions to special cases, such as Gaussian densities, for real-valued variables are also possible. Within the artificial intelligence community these models were initially developed within the framework of probabilistic expert systems: the structure of the model and the parameters (the conditional probabilities attached to the links of the graph) were elicited from experts. More recently there has been significant work in both the AI and statistical communities on methods whereby both the structure and parameters of graphical models can be learned from databases directly. Model search can consist of greedy hill-climbing methods over various graph structures: prior knowledge, such as a partial ordering of the variables based on causal relations, can be quite useful in terms of helping the model search phase. Although still primarily at the research phase, graphical model learning algorithms look quite promising for descriptive data mining tasks.

Given the broad spectrum of data mining methods and algorithms, this brief overview is inevitably limited in scope: there are many data mining techniques, particularly specialized methods for particular types of data and domains, which were not mentioned specifically in the discussion. Although different algorithms and applications may appear quite different on the surface, it is not uncommon to find that they share many common components.

Data mining applications in biopharmaceutical industries

Discovery in compound library databases

Scientists in GSK drug discovery are working to build libraries large numbers of drug compounds of 100,000 or more each. Each library contains structural and biological activity information (for various target compounds) of individual compounds. Important questions to ask are of the form "what are the characteristics of the compounds which have desirable activity for particular targets." This type of information is invaluable for drug design. Traditionally so-called quantitative structure activity (QSAR) studies have been carried out on such data but only on a small scale (a few hundred compounds or less). Nonetheless, on data sets of this scale, classical multivariate regression/classification techniques have been applied to this data with success and will still be valuable to a large extent since such a study can start from the designed experiment. However the large size of the libraries (typically with 100 or more fields involved) are clearly beyond the scope of traditional statistical techniques. Data mining techniques can play a useful role in QSAR studies. For example, rule and tree-based methods may be able to identify low-dimensional sets of variables that are useful for activity prediction. Such information from exploratory data mining of QSAR data could be fed into the next stage of the drug design perhaps through designed experiment. An interesting application of several data mining algorithms to high-throughput screening data can be found in Engels M.F.M *et al*..

Clinical trial data/disease management/ outcomes research data

Large amounts of data are accumulated through various stages of clinical trials of drug developments. In addition, there are vast quantities of doctors' prescription records available internally and externally. Similar data can be obtained through outcomes research. Data mining could be used with such data to answer questions such as "what are the characteristics of the group of people for whom a certain drug was effective?" or "can we find a small subgroup of people who have adverse reactions to certain drug and characterize them?". Such information could be very valuable for the next stage of drug development or for target marketing development. However there are many more clinical and epidemiological questions which can be asked against such patients data collected and aggregated. A recent paper (Olaleye D. *et al*., 2001) discusses many practical issues in the clinical data mining and Lee K.R. *et al*. is about a specific application of data mining to merged clinical trial data on type II diabetes.

Genomics and proteomics database

In collaboration with Human Genome Sciences (HGS), GSK has accumulated a large number of human genome sequences in terms of ESTs (Expressed Sequence Tags). Exploratory data analytic techniques to make better use of this important database will be very valuable to GSK. Statistical methods based on hidden Markov models (HMMs) (see Fayyad U.M. *et al.*, 1996) are increasingly finding applications to this type of sequence data with significant success. However, the HMM is a model with very restrictive independence assumptions: there is considerable room for exploring techniques beyond the HMM, including grammar models, local pattern dependencies, and so forth. However the most exciting new technology lies in functional genomics, where the gene expressions of thousands of genes are measured simultaneously in a single sample for cells. The DNA microarrays, also known as "gene chips" is a new promising technology to find genes specifically responsible for certain diseases.

Many techniques for analyzing microarray data are proposed with contributions from prominent statisticians and data scientists. Some of the recent work can be found in Hastie T. et al. and Tibshirani R. et al.. Thus far, much of the work in DNA microarray data analysis has centered around cluster analysis to find genetic subtyping. Further work in this direction may be of good reason: a recent paper (Burr T. *et al.*, 2001) claims that model-based clustering (see e.g., Banfield J. *et al.*, 1993) has great potential for choosing the number of subtypes in genetic data.

In addition to cluster analysis, other techniques, including evolutionary computing as well as survival analysis, are increasingly being used to analyze bioinformatics data. Genetic programming, which falls under the umbrella of evolutionary computing, has been used to automatically generate predictors for some of the critical properties of drug-like chemicals (Moore J.S. *et al.*, 2002). And recently, Beer D. *et al.* used survival analysis techniques to determine whether gene-expression profiles were associated with variability in survival times. They went on to demonstrate a gene-expression risk profile that can distinguish stage I lung adenocarcinomas and differentiate prognoses. The techniques of microarray data analysis are also applicable to other bioinformatics data, including that of Single Nucleotide Polymorphisms (SNP) and protein arrays.

Recently, a challenge at the CAMDA03 (Critical Assessment of Microarray Data Analysis) conference was to propose schemes to effectively integrate information from different microarray platform data sets. This problem is both extremely important and relevant to ongoing work with microarrays, as researchers will often find that the number of genes in common between different sets may not be sufficient to produce conclusive results. One possible solution is to combine data from several studies using Q-Normalization (Lin X. *et al.*, 2003), thereby producing larger samples and providing more statistical power in analyzing data. A minor drawback is that due to differences in the design of probe sets for different microarray chips, information may be lost when using combined data. However, this method led to an excellent selection of important genes associated with diseases, which were highly consistent with prior biological findings. Finding new and effective way of merging data remains a continuing challenge today.

The other component of functional genomics is proteomics. The hype and speculation surrounding gene sequencing is now being switched to gene expression studies and proteomics. Proteomics (PROTEin + genOMICS) represents the effort to identify, quantify, and determine the structure and function of proteins using techniques such as 2D-PAGE (two-dimensional polyacrylamide gel electrophoresis) and Mass Spectrometry. 2D-PAGE is a procedure by which individual proteins of a given sample are separated by isoelectric charge in one dimension and molecular weight in the other dimension. Mass spectrometry is a powerful analytical technique that is used to help identify unknown compounds, to quantify known compounds, and to elucidate the structure and chemical properties of molecules. The technique uses dispersion or filtering to sort ions according to their mass-to-charge ratios or a related property. One significant advantage of mass spectrometry is that detection of compounds can be accomplished with very minute quantities.

Data mining needs in proteomics data are similar to that of microarray data, but the reproducibility issues and the need for proper transformation of data before analysis is even more important. Some of the statistical analysis of proteomics data appeared in Vohradsky J. *et al.*. Recently Lee K.R. *et al.* published statistical analysis of mass spectrometric proteomics data using latent variable projection method.

Pharmacogenomics and data mining

For every medication, response falls into three main groups: some patients respond well with minimal or no adverse effects, some respond but have unacceptable adverse effect, and some do not respond at all. The goal of pharmacogenomics is to predict, using specific information from patients' genomes, those that will respond well to treatment. Many research based pharmaceutical companies are collecting patients' genetic information together with other common biomarkers collected in clinical trials. The influx of genetics related data is so vast that pharmaceutical companies cannot effectively disentangle the complex relationship between genes, environmental factors and drug efficacy. Pharmacogenomics impacts clinical trials in two ways. As stated above, it allows researchers to be able to select patients that will favorably respond to the drug being tested. It can also be used to determine which genetic variations are related to adverse effects. This is complicated by the fact that the adverse effects may be unrelated to the known drug target. For example the genetic variation led to the idiosyncratic liver toxicity of certain diabetes drug is not obvious. Pharcogenomics will not only affect the clinical trial but also the prescription of drugs, and the development of genomics based diagnostics.

Many large biotechs an have pharmacogenomics effort and the number of start-up companies who are dedicated to this field is also large. Data mining needs of pharmacogenomics are more complicated than mere genomics or proteomics data since these data would be merged with traditional pheno-type data. Decision tree-like tools are commonly used for initial exploratory type analysis and a company like Golden Helix (http://www.goldenhelix.com) commercialized a version of decision tree for analysis of pharmacogenomics data. However, any prediction modeling tools can be potentially useful for analysis of such data.

Text mining

The technique of text mining is a departure from most other analysis techniques known today, namely in that it attempts to apply data mining techniques to a non-structured data format, i.e. text. More formally, text mining constitutes the search for local and global patterns in natural language text to extract information for clearly defined purposes. The main obstacle for such a method, as the reader can infer, involves attaining a fundamental understanding of natural language text, which remains a tremendous challenge. However, text mining recognizes that a solution to the aforementioned problem is not immediately attainable and instead focuses on extracting small amounts of information from text with high reliability.

In terms of successful biopharmaceutical applications, text mining was utilized by Ai C.S. *et al.* to automatically extract and organize chemical reaction information from a text database of the American Chemical Society by using logical ("grammar-like") representations.

Other published data mining applications of pharmaceutical relevance

There have been relatively few published reports of data mining techniques being applied to drug and chemical compound databases. This does not mean that such applications do not exist: it is more likely that there are in fact such applications but that the companies involved would rather retain a competitive advantage by not publishing their methods. A few researchers have published work on applications that may be of relevance to GlaxoSmithKline and their work is briefly described below. This selection is intended to be illustrative rather than exhaustive.

Cook D.J. *et al.* describe an interesting algorithm (called SUBDUE) which uses graph-matching techniques, coupled with information-theoretic fitness criteria, to search for natural substructure in structured data. As an example of their approach, they applied their method to a chemical compound database where the individual atoms are mapped to labeled nodes in the graph, and mapping bonds between the atoms onto labeled edges in the graph. The SUBDUE algorithm was able to automatically find commonly used substructures such as isoprene and benzene rings from the database. While this was only a "toy" example (a small-scale experiment) and the patterns discovered were already well-known to chemists, the work nonetheless illustrates the potential for using such techniques for iterative, semi-automated pattern discovery in large scale chemical compound databases.

Data mining of rules was used in crystallography by Hennessy D. *et al.* to discover potentially significant new empirical relationships in crystal growth. The data mining was applied to the Biological Macromolecular Crystallization Database to discover relationships between experimental parameter settings and crystal growth.

Bahler D. *et al.* applied decision tree and rule learning algorithms to a database describing chemical compounds and their carcinogenicity. The learned rules confirmed expert heuristic knowledge and out-performed all previous computer-based prediction algorithms, while matching human performance.

Jain A.N. *et al.* describe a computer algorithm called COMPASS which uses an explicit representation of molecular shape along with neural network models, for accurate drug activity prediction in drug design. Again, the use of learning or mining methods to leverage previously unseen patterns in data was a critical component in the success of this work.

There has been a significant amount of work on learning models from protein and DNA sequences. The hidden Markov model methods pioneered by Haussler and his colleagues can be viewed as data mining applied to such data (see Fayyad U.M. *et al.* , 1996 for a data mining perspective on this work). Less well known, is the work of Muggleton S. *et al.* which describes the use of learning techniques to discover logical rules describing relations among proteins. This work characterizes the data mining approach in general in that the algorithm finds local patterns in rule form (rather than a full model). Also of interest is a rather novel technique for incorporating background knowledge into the algorithm using logical formulae to describe known protein relations.

Conclusions

Data mining is an applications-oriented field (rather than theory-oriented) which leverages well-known techniques from applied statistics and computer science to generate particular solutions to the problem of extracting useful information automatically from data. There are a number of data mining techniques which are unique to the field, and increasingly the field is developing its own identity via annual conferences and so forth. Data mining focuses on identifying understandable models from data rather than building models that predict well but which have little interpretative power.

From the business viewpoint, data mining is not a silver bullet to the problem of dealing with large data sets. The application of data mining algorithms still requires close attention to problem formulation, problem representation, matching of algorithm and problem, and interpretation of results. Nonetheless, data mining has already produced a variety of useful and novel methods for exploring large data sets. It is safe to predict continued progress in the field, driven both by research advances based on computer science and statistics, and practical advances for specific applications.

References

Agrawal, R., *et al.* (1995). Fast discovery of association rules. In *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*. (AAAI Press), 3-8.

Ai, C.S., Blower, P.E., and Ledwith, R.H. (1991). Extracting reaction information from chemical databases. In Piatetsky-Shapiro, G. and W. J. Frawley ,eds. *Knowledge Discovery in Databases*, (Cambridge, MA: AAAI/MIT Press), 367-381.

Bahler, D. and Bristol, D.W. (1993). The induction of rules for predicting chemical carcinogenesis in rodents. *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology* (Menlo Park, CA: AAAI Press), 29-37.

Banfield, J. and Raftery, A. (1993). Model-based Gaussian and non-Gaussian Clustering. *Biometrics*. 49, 803-821.

Beer, D. *et al.* (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*. 8, 816-824

Breiman, L., Friedman, J., Olshen, R., and Stone, C.J. CART: Classification and Regression Trees.(Belmont, CA: Wadsworth Press).

Burr, T., Gattiker, J.R., and LaBerge, G.S. (2001). Genetic Subtyping using Cluster Analysis. *SIGKDD Explorations*. 3, 33-42.

Chatfield, C. (1995). Model uncertainty, data mining, and statistical inference. *J. R. Statist. Soc.* (A).158, 419-466.

Cook, D.J. and Holder, L. (1994). Substructure discovery using minimum description length and background knowledge. *Journal of Artificial Intelligence Research*. 1, 231-255.

Decker, K.M. and Foccardi, S. (1995). Technology overview: a report on data mining. Technical Report CSCS TR-95-02. (Swiss Scientific Computing Center, Manno, Switwerland)

Elder, J. and Pregibon, D. (1996). A statistical perspective on KDD, Advances in Knowledge Discovery and Data Mining. U. Fayyad, *et al* eds. (Cambridge, MA: AAAI/MIT Press), 83-114.

Engels, M.F.M., Knapen, K., and Tollenaere, J.P. (2001). Approaches for Mining High-throughput Screening Data Sets. Paper presented on the 13th European Symposium on Quantitative Structure-Activity Relationships, Dusseldorf, Germany.

Fayyad, U.M., Piatetsky-Shapiro,G., Smyth, P., and Uthurasamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. (Cambridge, MA: AAAI/MIT Press)

Fayyad, U.M., Haussler, D., and Stolorz, P. (1996). KDD for science data analysis: issues and examples. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. E. Simoudis and J. Han eds. (Menlo Park, CA: AAAI Press), 50-56.

Friedman, H.P. and Goldberg, J.D. (2000). Knowledge Discovery from Databases and Data Mining: New Paradigms for Statistics and Data Analysis? Biopharmaceutical Report.8(2), Biopharmaceutical Section, American Statistical Association

Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. (1996). *Data mining and statistics Communications of the ACM*. 39, 35-41.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staut, L., Botstein, D., and Brown, P. (2000). Identifying distinct set of genes with similar expression patterns via gene shaving. *Genome Biology*. 1, 1-21.

Heckerman, D. (1996). Bayesian networks for knowledge discovery. In *Advanced in Knowledge Discovery and Data Mining*, U. Fayyad et al. eds. (AAAI/MIT Press), 273-305.

Hennessy, D. *et al.* (1995). Induction of rules for biological macromolecule cystallization. *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*. (Menlo Park, CA: AAAI Press), 179-187.

Jain, A. N., *et al.* (1994). Compass: a shape-based machine learning too for drug design. *Journal of Computer-Aided Molecular Design*. 8, 635-652.

Lee, K.R., Lin, X., Park, D.C., Eslava S. (2003). Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method. *Proteomics*. 3, 1680-1686.

Lee, K.R., Lydick, E., Park, D.C., Lin, X. (2001). Exploratory Data Analysis of Irregular Patterns of Longitudinal Laboratory Data from Clinical Trials - A case study of liver function test. *Proceedings of 10th World Congress on Medical Informatics*, London, UK. 873.

Lin, X., Park, D.C., Eslava, S., Lee, K.R., Lam, L.H., and Zhu L.A. (2003). Making Sense of Human Lung Carcinomas Gene

Expression Data: Integration and Analysis of Two Affymetrix Platform Experiments. *Proceedings of Critical Assessment of Microarray Data Analysis (CAMDA03)*, Durham, NC, USA, 23-27.

Mannila, H. (1996). Data mining: machine learning, statistics, and databases. *Proceedings of the 1996 International Conference on Machine Learning*, (San Mateo, CA: Morgan Kaufmann Publishers), also available on the Web at http://www.cs.helsinki.fi/~mannila.

Mannila, H. and Toivonen, H. (1996). Discovering generalized episodes using minimal occurences. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (AAAI Press), 146-151.

Moore, J.S., Parker J.S., Olsen, N.S., and Aune, T.M. (2002). Symbolic discriminant analysis of microarray data in automimmune disease. *Genetic Epidemiology*. 23, 57-69.

Muggleton, S., King, R., and Sternberg, M. (1992). Protein secondary structure prediction using logic. *Protein Engineering*. 5, 647-657.

Michie, D., Spiegelhalter, D.J., and Taylor, C.C. (1994). *Machine Learning, Neural and Statistical Classification*. (New York: Ellis Horwood).

Olaleye, D. and Tardiff, B.E. (2001). Practical Issues in and Applications of Clinical Data Mining. *Drug Information Journal*. 35, 791-808.

Piatetsky-Shapiro, G. and Frawley, W.J. (1991). *Knowledge Discovery in Databases*. (Cambridge, MA: AAAI/MIT Press).

Quinlan, J.R. (1993). C4.5: Programs for Machine Learning, San Mateo. (CA: Morgan Kaufmann).

Smyth, P. and Goodman, R.M. (1992). An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering*. 4, 301-316.

Smyth, P. (1996). Clustering using Monte Carlo cross-validation. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. (AAAI Press) 126-133.

Smyth, P., Heckerman, D., and Jordan, M.I. (1997). Probabilistic independence networks for hidden Markov probability models. *Neural Computation*. 9, 227-269.

Tibshirani, R., Hastie, T., Botstein, D., and Brown, P. (2001). Supervised harvesting of expression trees. *Genome Biology* 2, 1-12.

Vohradsky, J. and Thompson, C.J. (1997). Identification of procaryotic developmental stages by statistical analyzes of two-dimensional gel patterns. *Electrophoresis* 18, 1418-1428.