Statistics for Hospital Epidemiology

EDITED BY DAVID BIRNBAUM, PHD, MPH

Application of Data Mining Techniques to Healthcare Data

Mary K. Obenshain, MAT

ABSTRACT

A high-level introduction to data mining as it relates to surveillance of healthcare data is presented. Data mining is compared with traditional statistics, some advantages of automated data systems are identified, and some data mining strategies and algorithms are described. A concrete example illustrates steps involved in the data mining process, and three successful data mining applications in the healthcare arena are described (*Infect Control Hosp Epidemiol* 2004;25:690-695).

DATA MINING PERSPECTIVE

Data mining lies at the interface of statistics, database technology, pattern recognition, machine learning, data visualization, and expert systems. A database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated. Databases contain aggregations of data records or files, and a database manager provides users the capabilities of controlling read and write access, specifying report generation, and analyzing use. Databases and database managers are prevalent in large mainframe systems, but are also present in smaller systems and on personal computers. Databases usually include a query facility, and the database community has a tendency to view data mining methods as more complicated types of database queries. For example, standard query tools can answer questions such as, "How many surgeries resulted in hospital stays longer than 10 days?" Data mining is valuable for more complicated queries such as, "What are the important preoperative predictors of excessive length of stay?" Data mining techniques can be implemented retrospectively on massive data in an automated matter, whereas traditional statistical methods used in epidemiology require custom work by experts. Traditional methods generally require a certain number of predefined variables, whereas data mining can include new variables and accommodate a greater number of variables.

Traditional methods, such as statistical process control based on various underlying probability distribution functions, have been successfully implemented in hospital infection control.[1-8] Data mining techniques have been implemented separately, and some of these are described below. Direct comparison of traditional statistical methods with data mining would require competitive results on the same data. Application of either statistical or data mining techniques requires substantial human effort, and collaboration, rather than competition, needs to occur between the two fields. As more statisticians become involved in data mining, the two fields could contribute to each other more effectively by building on each other's strengths to create synergy than by having a "bake off" or taking an antagonistic approach.

Data mining encompasses a wide variety of analytical techniques and methods, and data mining tools reflect this diversity. Many database vendors are moving away from providing stand-alone data mining workbenches toward embedding the mining algorithms directly in the database. This process is known as "in place data mining" and it enables more efficient data management and processing. The driving factor for such systems is to derive value from large databases that were not originally designed to answer the kinds of questions that data mining can address. A data mining application suite (or bundle of products that have been "glued together" to provide capabilities for the entire data mining process) with embedded analytics (analytical tools or software products that are accessed via the user interface) and improved interoperability (the ability to work on other systems or products without special effort

Ms. Obenshain is an independent consultant in Chapel Hill, North Carolina.

Address reprint requests to Mary K. Obenshain, MAT, Data Quality Research Institute, UNC at Chapel Hill, CB#7226, 200 Timberhill Place, Suite 201, Chapel Hill, NC 27599-7226.

on the part of the user) can be substantially easier to use than a data mining product alone. A data mining application is especially useful when it focuses on a specific application area so that users can have a workspace and terminology that are appropriate for their role.

Data mining products such as SAS Enterprise Miner (SAS Institute, Inc., Cary, NC) are frequently included in data mining application suites for specific application areas such as fraud and abuse detection, customer relationship management, and financial management. For example, Morgan Stanley uses SAS Enterprise Miner as part of its customer relationship management system. Real world examples of customer relationship management and additional application areas are available at www.sas.com.

There are obvious advantages to an automated surveillance system, regardless of whether based on data mining or statistical methods such as statistical process control. These include (1) a reduction of time and effort on the part of the end user; (2) the ability to examine multiple areas simultaneously; (3) a decreased potential for human error; (4) data presented in the correct format; and (5) data that are accessible anytime and anywhere.

Automated surveillance systems raise citizen concern over privacy. But it is possible to detect events and monitor trends even after patient demographic and personally identifiable data are stripped out. Military medical researchers in the United States are using such a system that gathers data from military medical facilities worldwide as well as from other healthcare sources. The system detects outbreaks (eg, the Norwalk virus in San Diego in 2002) when individual healthcare practitioners may not be able to see the big picture, and it monitors progression of diseases (eg, West Nile virus and listeriosis).[9]

DATA MINING STRATEGIES

The goal of data mining is to learn from data, and there are two broad categories of data mining strategies: supervised and unsupervised learning.[10] The table presents data mining strategies by modeling objective, categorized by supervised and unsupervised distinctions. Modeling objectives are listed in the first column of the table, and are described below. Supervised and unsupervised strategies are listed in the second and third columns, respectively, and are also described below.

Supervised learning methods are deployed (or strategically put into service in an information technology context) when values of variables (inputs) are used to make predictions about another variable (target) with known values. Unsupervised learning methods can be used in similar situations, but are more frequently deployed on data for which a target with known values does not exist. An example of a supervised method would be a healthcare organization finding out through predictive modeling what attributes distinguish fraudulent claims. In supervised methods, the models and attributes are known and are applied to the data to predict and discover information. With unsupervised modeling, the attributes and models of fraud are not known, but the patterns and clusters of data uncovered by data mining can lead to new discoveries.

TABLE
MODELING OBJECTIVES AND DATA MINING TECHNIQUES

Modeling Objective	Supervised	Unsupervised
Prediction	Ordinary least squares regression Logistic regression Neural networks Decision trees Memory-based reasoning Support vector machines Multi-adaptive regression splines	Not feasible
Classification	Decision trees Neural networks Discriminant analysis Bagging and boosting ensembles Naïve Bayes classifiers	Clustering (eg, K means) Kohonen networks Self-organizing maps
Exploration	Decision trees	Principal components Clustering (eg, K means) Link analysis
Affinity		Associations Sequences Factor analysis

Prediction algorithms determine models or rules to predict continuous or discrete target values given input data. For example, a prediction problem could involve attempting to predict the amount of an insurance claim or a death rate given a set of inputs (pick one and then list the corresponding inputs).

Classification algorithms determine models to predict discrete values given input data. A classification problem might involve trying to determine whether a particular purchase represents anomalous behavior based on some indicators (eg, where the purchase was made, the amount of the purchase, or the type of purchase).

Exploration uncovers dimensionality in input data. Trying to identify groups of similar customers based on spending habits for a large, targeted mailing is an exploration problem.

Affinity analysis determines which events are likely to occur in conjunction with one another. Retailers use affinity analysis to analyze product purchase combinations in grocery stores. A potential medical example would be analysis of patients' signs and symptoms that occur together in a clinical trial.

The table lists both supervised and unsupervised

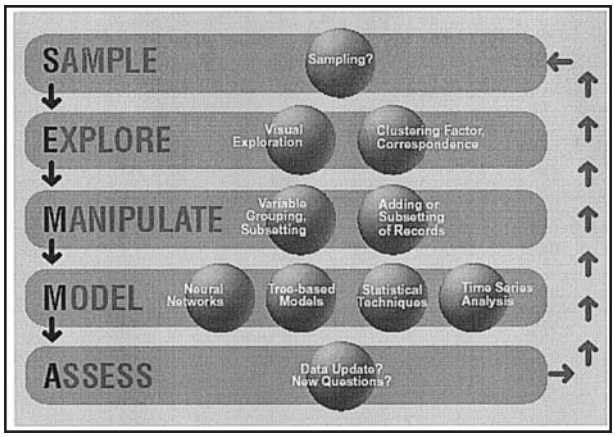

FIGURE. The SEMMA data mining process.

learning methods that are used in many industries (including retail, communications, financial, and insurance) for classification purposes. Specific examples of using data mining in these and other types of organizations are available at www.sas.com. In a particular business problem involving anomaly detection, the objective may be to establish a classification scheme for anomalies. Regression, decision trees, neural networks, and clustering can all address this problem. Decision trees build classification rules and other mechanisms for detecting anomalies. Clustering would indicate the types of groupings (based on a number of inputs) in a given population that are more at risk for exhibiting anomalies.

DATA MINING PROCESS

Data mining can be viewed as a process rather than a set of tools, and the acronym SEMMA (sample, explore, modify, model, and assess) refers to a methodology that clarifies this process. The SEMMA method divides data mining into five stages: (1) sample—to draw a statistically representative sample of data; (2) explore—to apply exploratory and statistical and visualization techniques; (3) modify (or manipulate)—to select and transform the most significant predictive variables; (4) model—to model the variables to predict outcomes; and (5) assess—to confirm a model's accuracy.

SEMMA is itself a cycle; the steps can be performed iteratively as needed. The process flow is depicted in the figure.

Often the final goal of data mining is to generalize findings for new data. The generation of predicted values for a new data set that may not contain a target is called "scoring." A complete data mining application is required to score a new or updated database by automatically applying the model calculated from the training data to the new or updated database. Results from automated scoring can then be surfaced to the appropriate individuals to determine whether any action or further investigation is needed. As an example, a data mining solution at Noel–Levitz helps public and private universities develop and implement sophisticated recruitment strategies faster than they could before implementation of the solution. Tim Thein, senior vice president at Noel–Levitz, said, "Predictive modeling for college enrollment has become more competitive in the last few years. By investing resources in the web-based scoring application, we can maintain an advantage in what has become a much more dynamic marketplace." The Noel–Levitz story and examples of other data mining solutions are available at www.sas.com.

Example of the Data Mining Process

The following drug discovery example illustrates how data mining (using Enterprise Miner software from SAS Institute, Inc.) might be implemented to derive knowledge from great volumes of data generated by high throughput screening systems. Details are not included because the purpose of the example is to illustrate how basic elements of the data mining process can be implemented to obtain useful results.

Twenty years ago, pharmaceutical scientists could screen approximately 30 drug candidates per week. Today, with combinatorial chemistry and high throughput screening, the same scientists can screen thousands of compounds per day in a controlled biological assay to determine potentially effective treatments. The most effective compounds, called "hits," are classified and retested. The best "leads" may result in a chemical synthesis program to optimize the chemical structure regarding the biological activity of interest.

A scientist in drug discovery described using data mining to uncover potentially useful relationships between chemical structure and biological activity that can be exploited to find new therapeutic candidates.[11] Steps in the data mining process implemented are based on the SEMMA paradigm described earlier and depicted in the figure.

Sample. Databases subject to data mining are typically in the terabyte or gigabyte range and continue to grow. Mining a representative sample instead of the whole volume drastically reduces the processing time required to get crucial information.

Estimating population parameters given a representative sample is well established in most research areas in statistics. Sampling strategies include simple random samples and stratification of samples to reflect subgroups of interest correctly.

A sampling strategy extracts a reliable, statistically representative sample of the full detail data. If general patterns appear in the data as a whole, these will be traceable in a representative sample. If a subgroup is so tiny that it is not represented in a sample and yet so important that it could influence results, it can be discovered with data mining techniques.

Sampling avoids problems in model validation. If all data are used for setting up the model, then there are no data left to check how well the model would adapt to new data. Sampling also avoids the problem of producing

results of substandard quality in terms of reliability that could occur when faster search algorithms on more data produce more findings but with lower confidence in any of them.

One of the crucial questions in data mining is how well its results can be generalized to other data sets not currently examined or available. The most common way to deal with this problem is to partition data before modeling into parts that are used to model and others that are set aside.

In our drug discovery example the goal is to relate chemical structure to biological activity. Just as a variable named AGE might represent age of individuals in years, in our example a variable named SCORE represents the biological activity of compounds and is measured on a range of 0 to 3. SCORE is the target or independent variable. Two additional activity variables were defined to facilitate additional modeling where activity is characterized as either Active or Inactive, and they were named ACT and ACT2.

BCUT numbers are continuous variables that describe the structure of compounds such as their surface area, bonding patterns, and charges. As an example, two-dimensional BCUTs are based on the two-dimensional representation of the chemical structure and the corresponding adjacency matrix. In our example there are 16 two-dimensional BCUT variables (named BCT2D1 through BCT2D16) along with 47 other BCUT variables so that altogether there are 63 BCUT variables in the input data set. The BCUT approach[12] attempts to describe molecules regarding the way they might interact with a bioreceptor.

As part of the sampling process, partitioned data sets are created as follows: training—used for model fitting (55% of the data); validation—used for assessment and to prevent overfitting (35% of the data); and test—used to obtain an honest assessment of how well a model generalizes (10% of the data). The stratified partitioning method is chosen so that the proportion of compounds with the various scores is the same across the three data sets.

Explore. There are graphical and analytical means of exploration in data mining, and visualization is one of the most versatile techniques. As a simple example of visualization, a histogram is a graphic presentation of a frequency distribution that can reveal variables that are heavily skewed on visual inspection. Another popular visualization method is to geocode or pinpoint data to a geographic location in a building or on the earth, and visually examine data points superimposed on a map of the building or geographic region. The geocode method had been useful in diverse applications, including investigation of the space shuttle Columbia accident and analysis of dead bird clusters as an early warning system for West Nile virus activity.[13]

If data are not suitable for visual exploration (eg, complex n-dimensional data), there is an alternate route that involves summarization using established statistical or advanced data mining methods. Exploring the data may reveal findings that suggest a sequence of questions including testing specific models.

In our drug discovery example, variable distributions were examined graphically and then it was decided not to transform or filter the variables examined. Although many data mining analysis methods are fairly robust regarding violations in the underlying assumptions of normality, homogeneity of variance, and additivity, it is still a good idea to check variable distributions to determine whether modifications are needed.

Modify. Data quality is an important requirement for data mining. Data sources can have records with missing values for one or more variables, and outliers could obstruct some of the patterns. A wide range of methods is available to deal with missing values and outliers. As an example, instead of ignoring records with missing data, the missing values can be replaced with a mean, a median, or a user-specified method.

Often it is desirable to segment records by placing objects into groups or clusters suggested by the data. Objects in each cluster tend to be similar to each other and those in different clusters tend to be dissimilar.

Data transformations are useful to improve model fitting. Of particular interest are binning transformations that allow collapsing an interval variable into a grouping variable.

No modifications were performed in the current drug discovery example.

Model. When patterns are found in the data, the next question is, "What causes those patterns?" Numerous modeling techniques are available, including traditional statistics. Some market analysts may include visualization as a modeling technique, but the SEMMA method finds an earlier place for this in the exploration phase.

Not all methods are suitable for all data. If only one or two methods are consistently advocated, due care should be taken to select the method according to the problem rather than bend the problem to fit a particular method.

Modeling methods explored in the current drug discovery example are stepwise regression, neural networks, and decision tree. Results from the regression model indicated a specific description of chemical structure is the most important variable in the model. The neural network model captured more of the "best" compounds than did the other models. The decision tree showed the two most important BCUT variables and their order of importance. This information is helpful for a chemist to know when searching for new therapeutic compounds because each BCUT variable corresponds to a particular chemical feature known to be important in binding to a receptor.

Assess. A common means of assessing a model is to apply it to a portion of data set aside during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model. Similarly, the model can be tested with known data.

The goal of data mining is to provide the best and most suitable explanation and to generalize this explanation for new but similar data (compounds in our drug dis-

covery example). SEMMA provides a common framework with a "yardstick" to compare models and predications from analytical methods. With the use of expected and actual values, cross-model comparisons and assessments are made independently of all other factors such as sample size or type of modeling tool used.

Assessment of models in the current drug discovery example revealed that the neural network model captured more of the best compounds more quickly than did the other models and potentially could be the best model to use for scoring new data.[11] Next, the test data set was scored using a scoring code generated by all three models that were run. Although the neural network model seemed to be the best model in the earlier assessment, examination of scores revealed that the neural network did better than the logistic regression but not as well as the decision tree in identifying active compounds. Thus, in the process flow diagram, the output model was changed from neural networks to decision tree so that the tree-based model will be used for subsequent predictions.

A goal in our drug discovery example was to generate a list of compounds most likely to have biological activity. Examination of the distribution of scores for the most potent inhibitor revealed a cut-off value to use for defining the predicted best compounds. A report summarizing results and the modeling process was generated in HyperText Markup Language (HTML).

Output from the data mining process can be customized to meet specific needs. Each stage of the process generates output, and graphical and tabular displays can be created to surface to the appropriate individual, in the appropriate format, at the appropriate time. In the current drug discovery example, final output might be a list of chemical structures that are most likely to have the desired biological activity or a list of potential potent inhibitors worthy of further investigation.

APPLICATIONS FOR MINING HEALTHCARE DATA

Business and marketing organizations may be ahead of healthcare in applying data mining to derive knowledge from data. This is quickly changing. Successful mining applications have been implemented in the healthcare arena, three of which are described below.

Hospital Infection Control

Nosocomial infections affect 2 million patients each year in the United States, and the number of drug-resistant infections has reached unprecedented levels.[14] Early recognition of outbreaks and emerging resistance requires proactive surveillance. Computer-assisted surveillance research has focused on identifying high-risk patients, expert systems, and possible cases and detecting deviations in the occurrence of predefined events.

A surveillance system that uses data mining techniques to identify new and interesting patterns in infection control data has been implemented at the University of Alabama.[15] The system uses association rules on culture and patient care data obtained from the laboratory information management systems and generates monthly patterns that are reviewed by an expert in infection control. Developers of the system conclude enhancing infection control with the data mining system is more sensitive than traditional infection control surveillance, and significantly more specific.

Ranking Hospitals

Organizations rank hospitals and healthcare plans based on information reported by healthcare providers. There is an assumption of uniform reporting, but research shows room for improvement in uniformity. Data mining techniques have been implemented to examine reporting practices. With the use of International Classification of Diseases, 9th revision, codes (risk factors) and by reconstructing patient profiles, cluster and association analyses can show how risk factors are reported.[16]

Standardized reporting is important because hospitals that underreport risk factors will have lower predications for patient mortality. Even if their success rates are equal to those of other hospitals, their ranking will be lower because they reported a greater difference between predicted and actual mortality.[16] Standardized reporting would also be important for meaningful comparisons across hospitals.

Identifying High-Risk Patients

American Healthways provides diabetes disease management services to hospitals and health plans designed to enhance the quality and lower the cost of treatment of individuals with diabetes. To augment the company's ability to prospectively identify high-risk patients, American Healthways uses predictive modeling technology.[17] Extensive patient information is combined and explored to predict the likelihood of short-term health problems and intervene proactively for better short-term and long-term results.

A robust data mining and model-building solution identifies patients who are trending toward a high-risk condition. This information gives nurse care coordinators a head start in identifying high-risk patients so that steps can be taken to improve the patients' quality of healthcare and to prevent health problems in the future.

CONCLUSION

Automated surveillance systems offer obvious advantages over manual ones. When analytical technologies are embedded in automated hospital infection surveillance systems, it is not clear whether data mining outperforms traditional statistical methods.

Further exploration of data mining for research related to infection control and hospital epidemiology seems in order, especially where the data volume exceeds capabilities of traditional statistical techniques. Data miners and statisticians should collaborate so that the two fields can contribute to each other. The challenge is for each to widen its focus to attain harmonious and productive collab-

oration to develop best practices for automated surveillance systems.

REFERENCES

1. Birnbaum D. Analysis of hospital infection surveillance data. *Infect Control* 1984;5:332-338.
2. Sellick JA Jr. The use of statistical process control charts in hospital epidemiology. *Infect Control Hosp Epidemiol* 1993;14:649-656.
3. Finison LJ, Spencer M, Finison KS. Total quality measurement in health care: using individuals charts in infection control. *ASQC Quality Congress Transactions* 1993;47:349-359.
4. Ngo L, Tager IB, Hadley D. Application of exponential smoothing for nosocomial infection surveillance. *Am J Epidemiol* 1996;143:637-647.
5. Benneyan JC. Statistical quality control methods in infection control and hospital epidemiology (parts I and II). *Infect Control Hosp Epidemiol* 1998;19:194-214, 265-283.
6. Kaminsky FC, Benneyan JC, Davis RD, et al. Statistical control charts based on a geometric distribution. *Journal of Quality Technology* 1992; 24:63-69.
7. Gustafson TL. Practical risk-adjusted quality control charts for infection control. *Am J Infect Control* 2000;28:406-414.
8. Benneyan JC. Number-between g-type statistical quality control charts for monitoring adverse events. *Health Care Manag Sci* 2001;4:305-318.
9. Johnston G. System adds to biodefense readiness. *Bio-IT World*. November 1, 2002. Available at www.bio-itworld.com/news/110102_report1436.html. Accessed July 21, 2004.
10. Matkovsky IP, Nauta KR. Overview of data mining techniques. Presented at the Federal Database Colloquium and Exposition; September 9-11, 1998; San Diego, CA.
11. Lajiness MS. Using Enterprise Miner to explore and exploit drug discovery data. Proceedings from the 25th Annual SAS User Group International; April 9-12, 2000; Indianapolis, IN.
12. Perleman RS, Smith KM. Novel software tools for chemical diversity. *Perspectives in Drug Discovery & Design* 1998;9:339-353.
13. Mostashari F, Kulldorff M, Hartman J, Miller J, Kulasekera V. Dead bird clusters as an early warning system for West Nile virus activity. *Emerg Infect Dis* 2003;9:641-646.
14. Gaynes R, Richards C, Edwards J, et al. Feeding back surveillance data to prevent hospital-acquired infections. *Emerg Infect Dis* 2001;7:295-298.
15. Brosette SE, Spragre AP, Jones WT, Moser SA. A data mining system for infection control surveillance. *Methods Inf Med* 2000;39:303-310.
16. Cerrito P. Using text analysis to examine ICD-9 codes to determine uniformity in the reporting of MedPAR data. Presented at the Annual Symposium of the American Medical Informatics Association; November 9-13, 2002; San Antonio, TX.
17. Ridinger M. American Healthways uses SAS to improve patient care. *DM Review* 2002;12:139.