3
Data Mining and Clinical Decision Support Systems

J. Michael Hardin and David C. Chhieng

Introduction

Data mining is a process of pattern and relationship discovery within large sets of data. The context encompasses several fields, including pattern recognition, statistics, computer science, and database management. Thus the definition of data mining largely depends on the point of view of the writer giving the definitions. For example, from the perspective of pattern recognition, data mining is defined as the process of identifying valid, novel, and easily understood patterns within the data set.[1]

In still broader terms, the main goal of data mining is to convert data into meaningful information. More specifically, one major primary goal of data mining is to discover new patterns for the users. The discovery of new patterns can serve two purposes: description and prediction. The former focuses on finding patterns and presenting them to users in an interpretable and understandable form. Prediction involves identifying variables or fields in the database and using them to predict future values or behavior of some entities.

Data mining is well suited to provide decision support in the healthcare setting. Healthcare organizations face increasing pressures to improve the quality of care while reducing costs. Because of the large volume of data generated in healthcare settings, it is not surprising that healthcare organizations have been interested in data mining to enhance physician practices, disease management, and resource utilization.

Example 3.1

One early application of data mining to health care was done in the early 1990s by United HealthCare Corporation. United HealthCare Corporation was a managed-care company, and developed its first data mining system, Quality Screening and Management (QSM), to analyze treatment records from its members.[2] QSM examined 15 measures for studying patients with chronic illness and compared the care received by its members to that recommended by national standards and guidelines. Results of the analyses

were then used to identify appropriate quality management improvement strategies, and to monitor the effectiveness of such actions. Although not providing direct support for decision making at the point of care, these data could be used to improve the way clinical guidelines are used.

Data Mining and Statistical Pattern Recognition

Pattern recognition is a field within the area of data mining. It is the science that seeks analytical models with the ability to describe or classify data/measurements. The objective is to infer from a collection of data/measurements mechanisms to facilitate decision-making processes.[3,4] With time, pattern recognition methodologies have evolved into an interdisciplinary field that covers multiple areas, including statistics, engineering, computer science, and artificial intelligence. Because of cross-disciplinary interest and participation, it is not surprising that pattern recognition is comprised of a variety of approaches. One approach to pattern recognition is called statistical pattern recognition.

Statistical pattern recognition implies the use of a statistical approach to the modeling of measurements or data.[5] Briefly, each pattern is represented by a set of features or variables related to an object. The goal is to select features that enable the objects to be classified into one or more groups or classes.

Data Mining and Clinical Decision Support Systems

With the advent of computing power and medical technology, large data sets as well as diverse and elaborate methods for data classification have been developed and studied. As a result, data mining has attracted considerable attention during the past several decades, and has found its way into a large number of applications that have included both data mining and clinical decision support systems. Decision support systems refer to a class of computer-based systems that aids the process of decision making.[6] Table 3.1 lists some examples of decision support systems that utilize data mining tools in healthcare settings.

A typical decision support system consists of five components: the data management, the model management, the knowledge engine, the user interface, and the user(s).[7] One of the major differences between decision support systems employing data mining tools and those that employ rule-based expert systems rests in the knowledge engine. In the decision support systems that utilize rule-based expert systems, the inference engine must be supplied with the facts and the rules associated with them that, as described in Chapter 2, are often expressed in sets of "if–then" rules. In this sense, the decision support system requires a vast amount of a priori

TABLE 3.1. Examples of clinical decision support systems and data mining tools that utilize statistical pattern recognition.

System (reference)	Description
Medical imaging recognition and interpretation system	
Computer-aided diagnosis of melanoma[23]	Analysis of digitized images of skin lesions to diagnose melanoma
Computer-aided diagnosis of breast cancer[21]	Differentiation between benign and malignant breast nodules, based on multiple ultrasonographic features
Monitoring tumor response to chemotherapy[30]	Computer-assisted texture analysis of ultrasound images aids monitoring of tumor response to chemotherapy
Diagnosis of neuromuscular disorder[31]	Classification of electromyographic (EMG) signals, based on the shapes and firing rates of motor unit action potentials (MUAPs)
Discrimination of neoplastic and non-neoplastic brain lesions[27]	Predicting the presence of brain neoplasm with magnetic resonance spectroscopy
Gene and protein expression analysis	
Molecular profiling of breast cancer[25]	Identification of breast cancer subtypes distinguished by pervasive differences in their gene expression patterns
Screening for prostate cancer[32]	Early detection of prostate cancer based on serum protein patterns detected by surface enhanced laser description ionization time-of-flight mass spectometry (SELDI-TOF MS)
Educational system	
Mining biomedical literature[33]	Automated system to mine MEDLINE for references to genes and proteins and to assess the relevance of each reference assignment
Laboratory system	
ISPAHAN[34]	Classification of immature and mature white blood cells (neutrophils series) using morphometrical parameters
Histologic diagnosis of Alzheimer's disease[35]	Analysis of digital images of tissue sections to identify and quantify senile plagues for diagnosing and evaluating the severity of Alzheimer's disease
Diagnosis of inherited metabolic diseases in newborns[36]	Identification of novel patterns in high-dimensional metabolic data for the construction of classification system to aid the diagnosis of inherited metabolic diseases
Acute care system	
Identification of hospitals with potential quality problems[37]	Using logistic regression models to compare hospital profiles based on risk-adjusted death with 30 days of noncardiac surgery
Prediction of disposition for children with bronchiolitis[22]	Neural network system to predict the disposition in children presenting to the emergency room with bronchiolitis
Estimating the outcome of hospitalized cancer patients[28]	Predicting the risk of in-hospital mortality in cancer patients with nonterminal disease
Miscellaneous	
Flat foot functional evaluation[38]	Gait analysis to diagnosis "flat foot" and to monitor recovery after surgical treatment

knowledge on the part of the decision maker in order to provide the right answers to well formed questions. On the contrary, the decision support systems employing data mining tools do not require a priori knowledge on the part of the decision maker. Instead, the system is designed to find new and unsuspected patterns and relationships in a given set of data; the system then applies this newly discovered knowledge to a new set of data. This is most useful when a priori knowledge is limited or nonexistent.

Many successful clinical decision support systems using rule-based expert systems have been developed for very specialized areas in health care.[8–14] One early example of a rule-based expert system is MYCIN, which used its rules to identify micro-organisms that caused bacteremia and meningitis.[14] However, such systems can be challenging to maintain due to the fact that they often contain several thousand rules or more. In addition, these "if–then" rule systems have difficulty dealing with uncertainty. Bayesian systems (see Chapter 2) are one way of addressing uncertainty. Statistical pattern recognition approaches are another.

Supervised Versus Unsupervised Learning

Data mining and predictive modeling can be understood as learning from data. In this context, data mining comes in two categories: supervised learning and unsupervised learning.

Supervised Learning

Supervised learning, also called directed data mining, assumes that the user knows ahead of time what the classes are and that there are examples of each class available. (Figure 3.1A) This knowledge is transferred to the system through a process called training. The data set used in this process is called the training sample. The training sample is composed of dependent or target variables, and independent variables or input. The system is adjusted based on the training sample and the error signal (the difference between the desired response and the actual response of the system). In other words, a supervised learning system can be viewed as an operation that attempts to reduce the discrepancy between the expected and observed values as the training process progresses. With enough examples in the training data, the discrepancy will be minimized and the pattern recognition will be more accurate.

The goal of this approach is to establish a relationship or predictive model between the dependent and independent variables. Predictive modeling falls into the category of supervised learning because one variable is designated at the target that will be explained as a function of other variables. Predictive models are often built to predict the future values or behavior of an object or entity. The nature of the target/dependent variable

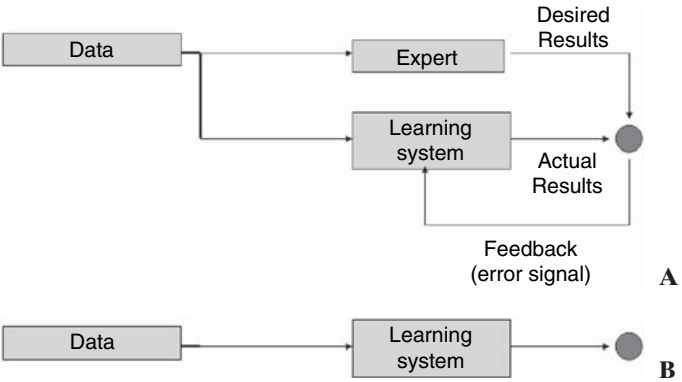

FIGURE 3.1. A, Supervised learning. B, Unsupervised learning.

determines the type of model: a model is called a classification model if the target variable is discrete; and a regression model if the target variable is continuous.

Example 3.2

Goldman et al. described the construction of a clinical decision support system to predict the presence of myocardial infraction in a cohort of 4,770 patients presenting with acute chest pain at two university hospitals and four community hospitals.[15] Based on the patient's symptoms and signs, the clinical decision support system had similar sensitivity (88.0% versus 87.8%) but a significantly higher specificity (74% versus 71%) in predicting the absence of myocardial infarction when compared to physicians' decisions if the patients were required to be admitted to the coronary care unit. If the decision to admit was based solely on the decision support system, the admission of patients without infarction to the coronary care unit would have been reduced by 11.5% without adversely affecting patient outcomes or quality of care.

A Priori Probability

In supervised learning, the frequency distribution, or a priori probability, of the classes of a certain training set (or a sample taken from the general population) may be quite different from that of the general population to which the classifier is intended to be applied. In other words, the training set/sample may not represent the general population. For example, a par-

ticular training set may consist of 50% of the subjects with disease and 50% without the disease. In this case, a priori probabilities of the two classes in the training set are 0.5 for each class. However, the actual a priori probability or the actual prevalence of disease may be very different (less than or greater than 0.5) from that of the training set. In some instances, the actual a priori probability of the general population may be unknown to the researchers. This may have a negative effect on the performance of the classifier when applied to a real world data set. Therefore, it is necessary to adjust the output of a classifier with respect to the new condition to ensure the optimal performance of the classifier.[16]

Unsupervised Learning

In unsupervised or undirected learning, the system is presented with a set of data but no information is available as to how to group the data into more meaningful classes (Figure 3.1B). Based on perceived similarities that the learning system detects within the data set, the system develops classes or clusters until a set of definable patterns begins to emerge. There are no target variables; all variables are treated the same way without the distinction between dependent and independent variables.

Example 3.3

Avanzolini et al. analyzed 13 commonly monitored physiological variables in a group of 200 patients in the six-hour period immediately following cardiac surgery in an attempt to identify patients who were at risk for developing postoperative complications.[17] Using an unsupervised learning (clustering) method, the investigators showed the existence of two well defined categories of patients: those with low risk of developing postoperative complications and those with high risk.

Classifiers for Supervised Learning

In supervised learning, classification refers to the mapping of data items into one of the predefined classes. In the development of data mining tools and clinical decision support systems that use statistical approaches like those described here, one of the critical tasks is to create a classification model, known as a classifier, which will predict the class of some entities or patterns based on the values of the input attributes. Choosing the right classifier is a critical step in the pattern recognition process. A variety of techniques have been used to obtain good classifiers. Some of the more widely used and well known techniques that are used in data mining include decision trees, logistic regression, neural networks, and nearest neighbor approach.

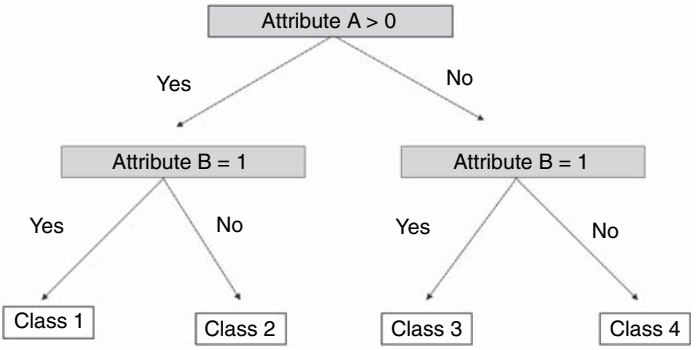

FIGURE 3.2. A simple decision tree with the tests on attributes A and B.

Decision Trees

The use of decision trees is perhaps the easiest to understand and the most widely used method that falls into the category of supervised learning. Figure 3.2 is the graphical representation of a simple decision tree using two attributes. A typical decision tree system adopts a top-down strategy in searching for a solution. It consists of nodes where predictor attributes are tested. At each node, the algorithm examines all attributes and all values of each attribute with respect to determining the attribute and a value of the attribute that will "best" separate the data into more homogeneous subgroups with respect to the target variable. In other words, each node is a classification question and the branches of the tree are partitions of the data set into different classes. This process repeats itself in a recursive, iterative manner until no further separation of the data is feasible or a single classification can be applied to each member of the derived subgroups. Therefore, the terminal nodes at the end of the branches of the decision tree represent the different classes.

Example 3.4

An example of a clinical decision support system using decision trees can be found in a study by Gerald et al.[18] The authors developed a decision tree that assisted health workers in predicting which contacts of tuberculosis patients were most likely to have positive tuberculin skin tests. The model was developed based on 292 consecutive cases and close to 3,000 contacts and subsequently tested prospectively on 366 new cases and 3,162 contacts. Testing showed that the decision tree model had a sensitivity of 94%, a specificity of 28%, and a false negative rate of 7%. The authors concluded that the use of decision trees would decrease the number of contacts investigated by 25% while maintaining a false negative rate that was close to

that of the presumed background rate of latent tuberculosis infection in the region.

Logistic Regression

Logistic regression is used to model data in which the target or dependent variable is binary, i.e., the dependent variable can take the value 1 with a probability of success p, or the value 0 with the probability of failure $1 - p$. The main objective is to develop a regression type model relating the binary variable to the independent variables. As such it is a form of supervised learning. It can also be used to examine the variation in the dependent variable that can be explained by the independent variables, to rank the independent variables based on their relative importance in predicting the target variable, and to determine the interaction effects among independent variables. Rather than predicting the values of the dependent variable, logistic regression estimates the probability that a dependent variable will have a given value. For example, instead of predicting whether a patient is suffering from a certain disease, logistic regression tries to estimate the probability of the patient having the disease. If the estimated probability is greater than 0.5, then there is a higher probability of the patient having the disease than not having the disease. The function relating the probabilities to the independent variables is not a linear function and is represented by the following equation:

$$p(y) = 1/\{1 + e^{(-a-bx)}\}$$

where $p(y)$ is the probability that y, the dependent variable, occurs based on x, the value of an attribute/independent variable, a is the constant, and b is the coefficient of the independent variable. Figure 3.3 shows a graphical representation of the logistic regression model which fits the relationship between the value of the independent variable, x and the probability of dependent variable, y occurring with a special S-shaped curve that is mathematically constrained to remain within the range of 0.0 to 1.0 on the Y axis.

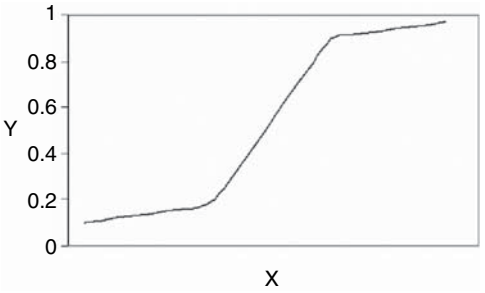

FIGURE 3.3. Logistic regression model.

Example 3.5

The following is an example that applies logistic regression to decision making. In the earliest stage of the epidemic of severe acute respiratory syndrome (SARS) when reliable rapid confirmatory tests were lacking, a group of researchers from Taiwan attempted to establish a scoring system to improve the diagnostic accuracy of SARS.[19] The scoring system was developed based on the clinical and laboratory findings of 175 suspected cases using a multivariate, stepwise logistic regression model. The authors then applied the scoring system to 232 patients and were able to achieve a sensitivity and specificity of 100% and 93%, respectively, in diagnosing SARS.

Example 3.6

In another study, the authors applied texture analysis to images of breast tissue generated by magnetic resonance imaging (MRI) for differentiating between benign and malignant lesions.[20] Using logistic regression analysis, a diagnostic accuracy of 0.8 +/− 0.07 was obtained with a model requiring only three parameters.

Neural Networks

The original development of the neural network programs was inspired by the way the brain recognizes patterns. A neural network is composed of a large number of processors known as neurons (after the brain cells that perform a similar function) that have a small amount of local memory and are connected unidirectionally (Figure 3.4). Each neuron can have more than one input and operates only on the inputs it receives via the connections. Like some of the data mining tools, neural networks can be supervised or unsupervised. In supervised neural networks, examples in the form of the training data are provided to the network one at a time. For each example, the network generates an output that is compared with the actual

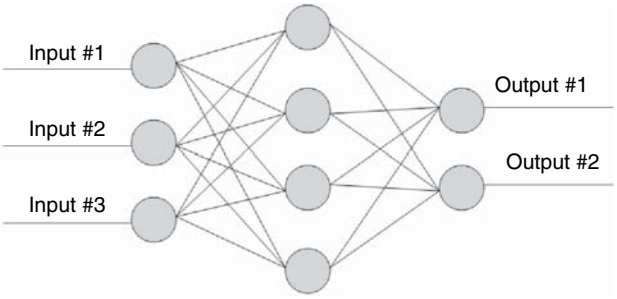

FIGURE 3.4. Neural network.

value as a form of feedback. Once the output of the neural network is the same as the actual value, no further training is required. If the output differs from the actual value, the network adjusts those parameters that contributed to the incorrect output. Once adjustment is made, another example is presented to the network and the whole process is repeated. The process terminates when all parameters are stabilized. The size and representativeness of the training data is obviously very important, since a neural network could work fine on the training set, but not generalize to a broader sample.

Example 3.7

One example of a neural network is the computer-aided diagnosis of solid breast nodules. In one study, ultrasonographic features were extracted from 300 benign and 284 malignant biopsy-confirmed breast nodules.[21] The neural network was trained with a randomly selected data set consisting of half of the breast nodule ultrasonographic images. Using the trained neural network, surgery could be avoided in over half of the patients with benign nodules with a sensitivity of 99%.

Example 3.8

In another example, a neural network was used to detect the disposition in children presenting to the emergency room with bronchiolitis (inflammation of small airways).[22] The neural network correctly predicted the disposition in 81% of test cases.

Nearest Neighbor Classifier

When a system uses the nearest neighbor (NN) classification, each attribute is assigned a dimension to form a multidimensional space. A training set of objects, whose classes are known, are analyzed for each attribute; each object is then plotted within the multidimensional space based on the values of all attributes. New objects, whose classes are yet to be determined, are then classified according to a simple rule; each new object is analyzed for the same set of attributes and is then plotted within the multidimensional space based on the value of each attribute. The new object is assigned to the same class of its closest neighbor based on appropriate metric/measurements. In other words, the NN rule assumes that observations which are the closest together (based on some form of measurement) belong to the same category (Figure 3.5). The NN rule is often used in situations where the user has no knowledge of the distribution of the categories.

One extension of this approach is the k-nearest neighbor approach (k-NN). Instead of comparing to a single nearest prototype, one can take into account k-neighboring points when classifying a data point, if the number of preclassified points is large. For each new pattern, the class is assigned by finding the most prominent class among the k-nearest data points in the

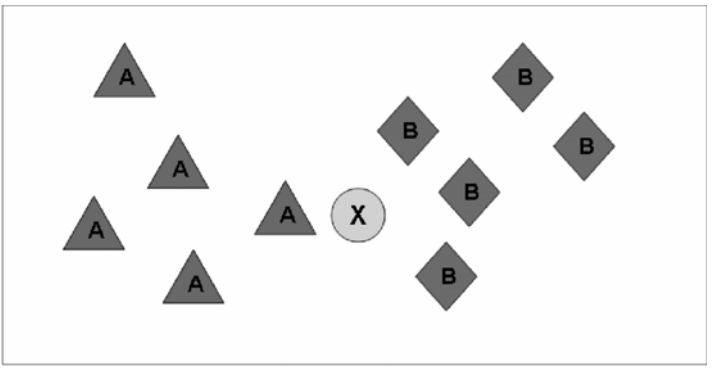

FIGURE 3.5. Nearest neighbor (NN) classifier. There are two classes: A (triangles) and B (diamonds). The circle represents the unknown sample, X. For the NN rule, the nearest neighbor of X comes from class A, so it would be labeled class A. Using the k-NN rule with k = 4, three of the nearest neighbors of sample X come from class B, so it would be labeled as B.

training set. (Figure 3.5) This approach works very well in cases where a class does not form a single coherent group but is a collection of more than one separate group.

Example 3.9

By applying the k-NN classifier, Burroni et al. developed a decision support system to assist clinicians with distinguishing early melanoma from benign skin lesions, based on the analysis of digitized images obtained by epiluminescence microscopy.[23] Digital images of 201 melanomas and 449 benign nevi were included in the study and were separated into two groups, a learning set and a test set. A k-NN pattern recognition classifier was constructed using all available image features and trained for a sensitivity of 98% with the learning set. Using an independent test set of images, a mean specificity of 79% was achieved with a sensitivity of 98%. The authors concluded that this approach might improve early diagnosis of melanoma and reduce unnecessary surgery.

Evaluation of Classifiers

ROC Graphs

In statistical pattern recognition, the goal is to map entities to classes. Therefore, the ultimate question is: which classifiers are more accurate in performing this classification task? Suppose one wanted to identify which classifiers would be best to determine whether a patient has cancer or not,

based on the results of certain laboratory tests. Given a classifier and an instance, there are four possible outcomes. If the patient has cancer and is diagnosed with cancer, based on the classifier, it is considered a true positive; if the patient is declared healthy by the classifier, but really has cancer, it is considered a false negative. If the patient has no cancer and is declared healthy, it is considered a true negative; if he is diagnosed as having cancer when he is really healthy, it is considered a false positive.

We can plot the true positive rate on the Y axis and the false positive rate on the X axis; a receiver operating characteristic (ROC) graph results (Figure 3.6). The true positive rate (also known as sensitivity) is obtained by dividing the number of true positives by the sum of true positives and false negatives. The false positive rate is obtained by dividing the number of false positives divided by the sum of true negatives and false positives; the false positive rate can also be expressed as "1 minus specificity," where specificity is equal to true negatives divided by the sum of true negatives and false positives. The ROC graph is a two-dimensional graph that depicts the trade-offs between benefits (detecting cancer correctly, or true positive) and costs (false alarm or false positive). Each classifier generates a pair of true positive and false positive rates, which corresponds to a point on the ROC graph. The point (0, 1) represents perfect classification, i.e., 100% true positive rate and 0% false positive rate. One classifier is considered superior to another if it has a higher true positive rate and a lower false positive rate, corresponding to a more "northwest" location relative to the other on the ROC graph. In general, the false alarm rates go up as one attempts to increase the true positive rate. Classifiers with points on the southwest corner of an ROC graph are more "conservative" since they make positive predictions only with strong evidence; therefore there is a low true positive rate, but also few false positive errors. On the other hand, classifiers on the

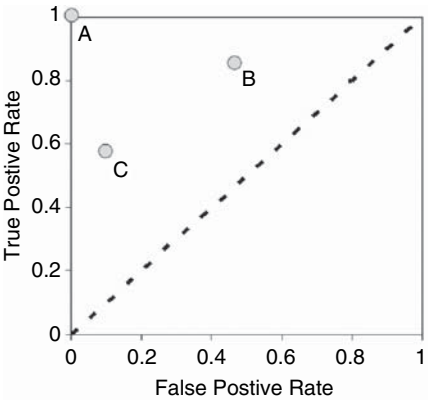

FIGURE 3.6. Receiver operating characteristic (ROC) curve. Point A represents perfect performance. The performance of C is more conservative than B.

northeast corner of an ROC graph are more "liberal" since they make positive prediction with weak evidence; therefore they have high true positive rates, but also high false positive rates.

Some classifiers, such as neural networks, yield a numeric value which can be in the form of a numeric score or probability that represents the likelihood an object belongs to a certain class. These classifiers can be converted into discrete, binary (yes versus no) classifiers by setting a threshold, i.e., if the output score is above the threshold, the classifier produces a "Yes, else a No". By choosing a different threshold, a different point in the ROC graph is produced. As a result, varying the thresholds will produce a curve in the ROC graph for a particular classifier. Given an ROC curve, one can select the threshold corresponding to a particular point on the ROC that produces the desired binary classifier with the best true positive rate (correctly diagnosed cancer) within the constraints of an acceptable false positive rate (false alarm). This is chosen based on the relative costs of the two types of errors: missing a diagnosis of cancer (type I error) versus creating a false alarm (type II error).

The area under the ROC curve (AUC) provides a single statistic (the C-Statistic) for comparing classifiers. It measures the accuracy of the classifiers. Consider the situation in which a classifier attempts to separate patients into two groups; those with disease and those without. One can randomly pick a patient from the disease group and one from the non-disease group and apply the classifier on both. The area under the curve represents the percentage of randomly drawn pairs where the classifier correctly classifies the two patients in the random pair. The value of AUC ranges from 0.5 to 1. A classifier with an AUC of 0.5 would be a poor classifier, roughly equivalent to flipping a coin to decide the class membership. A classifier with an AUC close to 1 results in better classification of entities to classes. For example, in Example 3.6, the resulting trained neural network model yielded a normalized area under the ROC curve of 0.95.

Computing the AUC is complex and beyond the scope of this chapter. Briefly, there are two commonly used methods. One method is based on the construction of trapezoids under the curve as an approximation of the area. The other method employs a maximum likelihood estimator to fit a smooth curve to the data points. Both methods are available as computer programs and give an estimate of area and standard error that can be used to compare different classifiers.

Kolmogorov-Smirnov Test

While the AUC provides a way of distinguishing groups overall, there are other statistical tests used to provide a more refined comparison of groups or subgroups. The Kolmogorov-Smirnov test, or KS test, is used to determine whether the distributions of two samples differ from each other or

whether the distribution of a sample differs from that of the general population. The KS test provides what is called the D-statistic for comparison of classifiers.[24]

Unsupervised Learning

Cluster Analysis

Unsupervised classification refers to situations where the goal is to classify a diverse collection of unlabeled data into different groups based on different features in a data set. Unsupervised classification, also known as cluster analysis or clustering, is a general term to describe methodologies that are designed to find natural groupings or clusters based on measured or perceived similarities among the items in the clusters using a multidimensional data set (Figure 3.7). There is no need to identify the groupings desired or the features that should be used to classify the data set. In addition, clustering offers a generalized description of each cluster, resulting in better understanding of the data set's characteristics and providing a starting point for exploring further relationships.

Clustering techniques are very useful in data mining because of the speed, reliability, and consistency with which they can organize a large amount of data into distinct groupings. Despite the availability of a vast collection of clustering algorithms in the literature, they are based on two popular approaches: hierarchical clustering and nonhierarchical clustering. The former, which is the most frequently used technique, organizes data in a nested sequence of groups that can be displayed in a tree-like structure, or dendrogram.

There are several problems that are associated with clustering. One problem is that data can be grouped into clusters with different shapes and sizes. Another problem is the resolution or granularity, i.e., fine versus

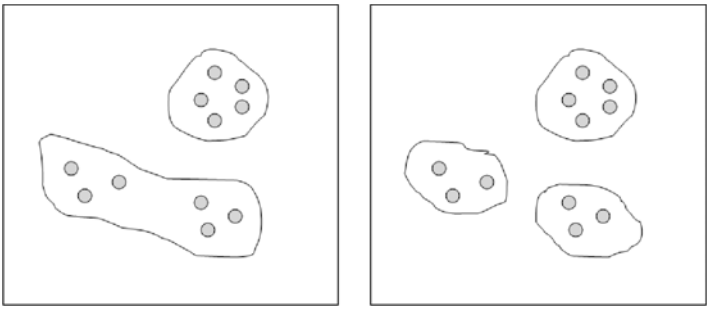

FIGURE 3.7. Cluster analysis. Two clusters of data (left); three clusters (right) using the same set of data.

coarse, with which the data are viewed. This problem is most obvious when one tries to delineate a region containing a high density of patterns compared to the background. Therefore, some authors define a cluster as one that consists of a relatively high density of points separated from other clusters by a relatively low density of points, whereas some define clusters containing samples that share more similarities to each other than to samples of different clusters. As a result, the selection of an appropriate measure of similarity to define clusters is a major challenge in cluster analysis.

Gene Expression Data Analysis

One of the applications of cluster analysis in medicine is the analysis of gene expression. With the completion of the human genome project, which identified more than 30,000 gene sequences, researchers are now able to examine the expression of several thousand genes from blood, body fluids, and tissue samples at the same time, in an attempt to identify gene subsets that are associated with various disease statistics. Since information is obtained from hundreds and thousands of gene sequences, an astronomical body of data is generated. Common research questions often fall under the following categories: class discovery, class prediction, and gene identification. Class prediction refers to the classification of samples based on certain behaviors or properties such as response to therapy, whereas gene identification involves the discovery of genes that are differentially expressed among different disease groups.

Class discovery refers to the discovery of previously unknown categories or subtypes based on some similarity measure calculated from the gene expression data. Cluster analysis is often the method of choice in accomplishing this task, because samples are clustered into groups based on the similarity of their gene expressions without utilizing any knowledge of any predefined classification schemes such as known histological tumor classification.

Example 3.10

In the future, it is likely that genomic data will be incorporated into clinical decision support systems to refine both diagnosis and therapy. The following is an example that used clustering to explore breast cancer classification using genomic data. In this study, Perou et al. evaluated the pattern of gene expression of 8,102 human genes in 65 breast cancers obtained from 42 patients.[25] Using hierarchical cluster analysis, the authors were able to classify 65 breast cancer samples into three distinct subtypes. One subtype was cancers that overexpressed the oncogene *erbB-2*. The remaining two subtypes were unknown prior to this study; they were estrogen receptor-positive luminal-like cancers and basaloid cancers. Subsequent survival analyses on a group of patients with locally advanced breast cancer

showed significantly different outcomes for the patients belonging to different subtypes; patients with basaloid cancers had a poor survival rate.[26] In the same study by Perou et al, the samples contained 20 primary tumors that were biopsied twice, before and after the completion of chemotherapy. Using clustering, the authors demonstrated that gene expression patterns were similar among samples from the same patients taken at different time points but not between samples taken from different patients.

Other Techniques

The goal of any tool that is used for pattern recognition is to arrive at an optimal solution within a given set of complex constraints. The development of sophisticated computer-based computation techniques has enabled analysts to attain better solutions than previous techniques. As improved techniques are developed to handle increasingly complex problems, there is a corresponding need for more innovative methods for arriving at optimal solutions. Genetic algorithms and biologic computing are two examples of innovative techniques that have gained increasing acceptance and application in the field of pattern recognition and data mining.

Genetic Algorithms

The fundamental concept of genetic algorithms has its roots in Darwin's evolutionary theories of natural selection and adaptation. According to Darwin, organisms that come up with successful solutions to best support them and protect themselves from harm survive, whereas those organisms that fail to adapt to their environment become extinct. Based on the same idea of "survival of the fittest," a genetic algorithm initially tries to solve a given problem with random solutions. These solutions are often referred to as the genomes, or a collection of genes. The gene represents the smallest unit of information for the construction of possible solutions. The next step is to evaluate or quantify the fitness of all the available genomes or solutions based on a fitness function. The latter returns a value of goodness or fitness so that a particular genome or solution may be ranked against all other genomes or solutions. Those solutions with better fit are ranked higher among others and are allowed to "breed." Once the initial evaluation is completed, the genetic algorithms examine new solutions by letting all the current solutions "evolve" through mutual exchange of "genetic materials" among solutions to improve the genomes and/or mutation (i.e., randomly changing the genetic materials) to "create" new solutions. The new solutions are then evaluated using the same fitness functions to determine which solutions are good and which are not and need to be eliminated. Thus the process repeats itself until an "optimal" solution is attained.

There are many benefits of genetic algorithms. One major advantage is that a genetic algorithm almost always guarantees finding some reasonable solution to problems, particularly those that we have no idea how to solve. Further, the final solution is often superior to the initial collection of possible solutions. Another benefit is that genetic algorithms tend to arrive at a solution much faster than other optimization techniques. Also, the strength of the genetic algorithm does not depend upon complex algorithms but rather on relatively simple concepts. Despite the power of genetic algorithms, however, some parameters, such as the size of the solution population, the rate of mutation and crossover, and the selection methods and criteria, can significantly affect their performance. For example, if the solution population size is too small, the genetic algorithm may have exhausted all the available solutions before the process can identify an optimal solution. If the rate of genetic mutation is too high, the process may be changing too fast for the selection to ever bring about convergence, resulting in the failure of generating an optimal solution.

Example 3.11

Genetic algorithms have been used to construct clinical decision support systems. In a study by Zellner et al., the authors evaluated the performance of a logistic regression model in diagnosing brain tumors with magnetic resonance spectroscopy using the genetic algorithms approach.[27] The genetic algorithm approach was superior to the conventional approach in 14 out of 18 trials, and the genetic algorithm had fewer false negatives and false positives. In addition, the authors also pointed out that the genetic algorithm approach was less costly.

Example 3.12

Genetic algorithms have also been used as a data mining technique in healthcare operations. One study investigated whether genetic algorithms could be used to predict the risk of in-hospital mortality of cancer patients.[28] A total of 201 cancer patients, over a two-year period of time, was retrospectively evaluated. Compared to other methods, such as multivariate logistic regression, neural networks, and recursive partitioning analysis, genetic algorithms selected the least number of explanatory variables with a comparable proportion of the cases explained (79%). The authors concluded that genetic algorithms reliably predicted in-hospital mortality of cancer patients and were as efficient as the other data mining techniques examined.

Biological Computing

Biological computing is another new discipline that has found its way into data mining applications. It cuts across two well established fields: computer

science and biology. While the genetic algorithm approach uses the analogy of natural selection to develop computer algorithms, the idea of biological computing actually involves the use of living organisms or their components, e.g., DNA strands, to perform computing operations. The benefits include the ability to hold enormous amounts of information, the capability of massive parallel processing, self-assembly, self-healing, self-adaptation, and energy efficiency. As of now, a biological computer can only perform rudimentary functions and it has no practical applications, but its potential continues to emerge. For example, some scientists have been working on the development of tiny DNA computers that circulate in a person's body to monitor his/her well-being and release the right drugs to repair damaged tissue or fight off infections and cancers.[29]

Conclusions

Data mining refers to the process of pattern and relationship discovery within large data sets. It holds promise in many areas of health care and medical research, with applications ranging from medical diagnosis to quality assurance. The power of data mining lies in its ability to allow users to consider data from a variety of perspectives in order to discover apparent or hidden patterns. There are two main divisions of classification: supervised learning or training, and unsupervised learning. Supervised training requires training samples to be labeled with a known category or outcome to be applied to the classifier. There are many classifiers available and their performance can be assessed using an ROC curve. Unsupervised learning, also known as clustering, refers to methodologies that are designed to find natural groupings or clusters without the benefit of a training set. The goal is to discover hidden or new relationships within the data set. One application of clustering is the analysis of gene expression data. Genetic algorithms and biological computing are two newer disciplines that have found their way into data mining applications and clinical decision support systems.

References

1. Fayyad UM, Piatetsky-Shapiro G, Smyth P. Knowledge discovery and data mining: towards a unifying framework. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 82–88. Portland, Oregon, August 1996. AAAI Press. Available from: http://ww-aig.jpl.nasa.gov.kdd96. Accessed July 17, 2006.
2. Leatherman S, Peterson E, Heinen L, Quam L. Quality screening and management using claims data in a managed care setting. QRB Qual Rev Bull 1991; 17:349–359.
3. Duda RO, Hart PE, Stork DG. Pattern classification and scene analysis, 2nd ed. New York: John Wiley and Sons; 2000.

4. Fukunaga K. Introduction to statistical pattern recognition, 2nd ed. New York: Academic Press; 1990.
5. Schalkoff RJ. Pattern recognition: statistical, structural and neural approaches. New York: John Wiley and Sons; 1991.
6. Finlay PN. Introducing decision support systems. Cambridge, MA: Blackwell Publishers; 1994.
7. Marakas GM. Decision support systems, 2nd ed. Princeton, NJ: Prentice Hall; 2002.
8. Ambrosiadou BV, Goulis DG, Pappas C. Clinical evaluation of the DIABETES expert system for decision support by multiple regimen insulin dose adjustment. Comp Methods Programs Biomed 1996;49:105–115.
9. Marchevsky AM, Coons G. Expert systems as an aid for the pathologist's role of clinical consultant: CANCER-STAGE. Mod Pathol 1993;6:265–269.
10. Nguyen AN, Hartwell EA, Milam JD. A rule-based expert system for laboratory diagnosis of hemoglobin disorders. Arch Pathol Lab Med 1996;120:817–827.
11. Papaloukas C, Fotiadis DI, Likas A, Stroumbis CS, Michalis LK. Use of a novel rule-based expert system in the detection of changes in the ST segment and the T wave in long duration ECGs. J Electrocardiol 2002;35:27–34.
12. Riss PA, Koelbl H, Reinthaller A, Deutinger J. Development and application of simple expert systems in obstetrics and gynecology. J Perinat Med 1988;16: 283–287.
13. Sailors RM, East TD. A model-based simulator for testing rule-based decision support systems for mechanical ventilation of ARDS patients. Proc Ann Symp Comp Appl Med Care 1994:1007.
14. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput Biomed Res 1975; 8:303–320.
15. Goldman L, Cook EF, Brand DA, et al. A computer protocol to predict myocardial infarction in emergency department patients with chest pain. N Engl J Med 1988;318:797–803.
16. Scott AJ, Wild CJ. Fitting logistic models under case-control or choice based sampling. J Roy Stat Soc B 1986;48:170–182.
17. Avanzolini G, Barbini P, Gnudi G. Unsupervised learning and discriminant analysis applied to identification of high risk postoperative cardiac patients. Int J Biomed Comp 1990;25:207–221.
18. Gerald LB, Tang S, Bruce F, et al. A decision tree for tuberculosis contact investigation [see comment]. Am J Respir Crit Care Med 2002;166:1122–1127.
19. Wang TL, Jang TN, Huang CH, et al. Establishing a clinical decision rule of severe acute respiratory syndrome at the emergency department. Ann Emerg Med 2004;43:17–22.
20. Gibbs P, Turnbull LW. Textural analysis of contrast-enhanced MR images of the breast. Magn Reson Med 2003;50:92–98.
21. Joo S, Yang YS, Moon WK, Kim HC. Computer-aided diagnosis of solid breast nodules: use of an artificial neural network based on multiple sonographic features. IEEE Transact Med Imaging 2004;23:1292–1300.
22. Walsh P, Cunningham P, Rothenberg SJ, O'Doherty S, Hoey H, Healy R. An artificial neural network ensemble to predict disposition and length of stay in children presenting with bronchiolitis. Eur J Emerg Med 2004;11:259–564.

23. Burroni M, Corona R, Dell'Eva G, et al. Melanoma computer-aided diagnosis: reliability and feasibility study. Clin Cancer Res 2004;10:1881–1886.
24. Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical recipes in FORTRAN example book: the art of scientific computing. 2nd Ed New York: Cambridge University Press; 1992.
25. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000;406:747–752.
26. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001;98:10869–10874.
27. Zellner BB, Rand SD, Prost R, Krouwer H, Chetty VK. A cost-minimizing diagnostic methodology for discrimination between neoplastic and non-neoplastic brain lesions: utilizing a genetic algorithm. Acad Radiol 2004;11:169–177.
28. Bozcuk H, Bilge U, Koyuncu E, Gulkesen H. An application of a genetic algorithm in conjunction with other data mining methods for estimating outcome after hospitalization in cancer patients. Med Sci Monit 2004;10: CR246–CR251.
29. Benenson Y, Gil B, Ben-Dor U, Adar R, Shapiro E. An autonomous molecular computer for logical control of gene expression [see comment]. Nature 2004; 429:423–429.
30. Huber S, Medl M, Vesely M, Czembirek H, Zuna I, Delorme S. Ultrasonographic tissue characterization in monitoring tumor response to neoadjuvant chemotherapy in locally advanced breast cancer (work in progress). J Ultrasound Med 2000;19:677–686.
31. Christodoulou CI, Pattichis CS. Unsupervided pattern recognition for the classification of EMG signals. IEEE Trans Biomed Eng 1999;46:169–178.
32. Banez LL, Prasanna P, Sun L, et al. Diagnostic potential of serum proteomic patterns in prostate cancer. J Urol 2003;170(2 Pt 1):442–426.
33. Leonard JE, Colombe JB, Levy JL. Finding relevant references to genes and proteins in Medline using a Bayesian approach. Bioinformatics 2002;18:1515–1522.
34. Bins M, van Montfort LH, Timmers T, Landeweerd GH, Gelsema ES, Halie MR. Classification of immature and mature cells of the neutrophil series using morphometrical parameters. Cytometry 1983;3:435–438.
35. Hibbard LS, McKeel DW Jr. Automated identification and quantitative morphometry of the senile plaques of Alzheimer's disease. Anal Quant Cytol Histol 1997;19:123–138.
36. Baumgartner C, Bohm C, Baumgartner D, et al. Supervised machine learning techniques for the classification of metabolic disorders in newborns. Bioinformatics 2004;20:2985–2996.
37. Gordon HS, Johnson ML, Wray NP, et al. Mortality after noncardiac surgery: prediction from administrative versus clinical data. Med Care 2005;43:159–167.
38. Bertani A, Cappello A, Benedetti MG, Simoncini L, Catani F. Flat foot functional evaluation using pattern recognition of ground reaction data. Clin Biomech 1999;14:484–493.