2

Data mining is the process of discovering "hidden messages," patterns and knowledge within large amounts of data and of making predictions for outcomes or behaviors. This chapter discusses in detail the theoretical and practical aspects of data mining and provides a case study of its application to college transfer data.

Data Mining and Its Applications in Higher Education

Jing Luan

The first chapter has afforded readers an opportunity to review the definitions and components of knowledge management. The chapter also established that knowledge management is closely linked to technology. Explicit knowledge, which is a product of several major technologies, is the focus of this chapter. Specifically, this chapter addresses data mining.

One among a host of recent technology innovations, data mining is making changes to the entire makeup of our skills and comfort zones in information analysis. Not only does it introduce an array of new concepts, methods, and phrases, it also departs from the well-established, traditional, hypothesis-based statistical techniques. Data mining is a new type of exploratory and predictive data analysis whose purpose is to delineate systematic relations between variables when there are no (or incomplete) *a priori* expectations as to the nature of those relations.

Herman Hollerith's invention of punch cards in 1880 and of a counting machine for the 1890 census led to the development of modern data management and computing techniques. Thearling (1995) even chronicled the evolution of data as *data collection* in the 1960s, *data access* in the 1980s, *data navigation* in the 1990s, and *data mining* in the new century. Thearling (1995) and others foresaw the possibilities of data mining as a result of maturity of all three disciplines: massive data collection and storage, powerful multiprocessor computers, and data mining algorithms. According to Rubenking (2001), "data mining is a logical evolution in database technology. The earliest databases, which served as simple replacements for paper records, were data repositories that provided little more than the capability to summarize and report. With the development of query tools such as SQL

[Structured Query Language], database managers were able to query data more flexibly."

In summary, data mining is possible due to

- Storage and computing power
- Database technology
- Integrated and maturing data mining techniques
- Strong need for fast, vast, and production-driven outcome
- Learner relationship management

Learner relationship management, discussed in the opening chapter, acts as an agent for moving fast on data mining. Higher education is transitioning from the enrollment mode to recruitment mode (Roueche and Roueche, 2000). Higher education institutions find that they cannot continue to operate in the "receive and process" mode. Instead, they must actively seek prospective students. They must cater to students' needs instead of designing a program with the attitude of "take it or leave it." This transition alone will exert great pressure for finding ways to make recruitment more efficient and institutions more attuned to learners' needs. Last, but not least, is the notion of accountability to which higher education can better respond with more powerful tools.

What Is Data Mining?

Artificial intelligence and artificial neural networks, along with almost all data mining techniques, were the brainchild of scholars in higher education, but data mining was not first applied to higher education. Suffice it to say that higher education is still virgin territory for data mining. The amount of data produced in higher education alone calls for some serious data mining. With institutions adopting Enterprise Resource Planning applications, such as Peoplesoft, Datatel, or SAP, kilobytes of data are being created and stored every hour when school is in session. Built for handling extremely large datasets, data mining has enjoyed tremendous growth in the corporate world and several government agencies, such as the FBI. The benefits range from finding hidden patterns in the customer mix, outliers in fraud detection, and targeted product promotion, to name just a few.

Data mining is an evolving field with new concepts born monthly and current concepts struggling to retain their place. Many of the new and interdisciplinary concepts, such as the stochastic search methods (including genetic algorithms), market basket analysis, memory based reasoning, and Bayesian averages, were not even imagined less than a decade ago. Researchers from different branches of mathematics, statistics, marketing, or artificial intelligence will use different terminologies. Where a statistician sees dependent and independent variables, and an artificial intelligence researcher sees features and attributes, others see records and fields (Berry

and Linoff, 1997). The phrase "neural networks" is synonymous with data mining.

Although data mining is known for having exotic names, the field has begun to include certain kinds of descriptive statistics and visualization techniques into data mining (Westphal and Blaxton, 1998). Statsoft, an on-line statistical software provider, seemed to favor "exploratory data analysis." Berthold and Hand (1999) called their work "intelligent data analysis." This chapter will refer to all activities involving modeling and nonhypothesis-based analytical techniques as data mining and adopt the concept developed by Berthold and Hand that all statistical techniques developed prior to the 1960s are "classic" and "conformatory."

The definition of data mining from Gartner Group seems to be most comprehensive: "the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques" (Gartner Group, 2000). I refine the notion of data mining as the purpose of uncovering hidden trends and patterns and making accuracy based predictions through higher levels of analytical sophistication. It is producing new observations from existing observations. As explained by Rubenking (2001), "data mining is the process of automatically extracting useful information and relationships from immense quantities of data. In its purest form, data mining doesn't involve looking for specific information. Rather than starting from a question or a hypothesis, data mining simply finds patterns that are already present in the data."

Finally, in statistical language, Statsoft (2001) categorizes typical on-line analytical processing (OLAP) techniques as basic statistical exploratory methods or exploratory data analysis that include such techniques as "examining distributions of variables (e.g., to identify highly skewed or non-normal, such as bi-modal patterns), reviewing large correlation matrices for coefficients that meet certain thresholds, or examining multi-way frequency tables (e.g., 'slice by slice' systematically reviewing combinations of levels of control variables)." It reserves the term "multivariate exploratory techniques" for data mining. These techniques are designed specifically to identify patterns in multivariate (or univariate, such as sequences of measurements) data sets that include cluster analysis, factor analysis, discriminant function analysis, multidimensional scaling, log-linear analysis, canonical correlation, stepwise linear and nonlinear (for example, logit) regression, correspondence analysis, time series analysis, and classification trees.

Essential Concepts and Definitions

Data mining assumes the existence of spherical multi-Euclidean dimensions (Delmater and Hancock, 2001). The *n*-dimensional Euclidean space, or Euclidean hyperspace, is called feature space, where any given coordinates

of ordered triples or ordered pairs are viewed as feature vectors. The understanding of Gaussian distribution, *z*-scores, and regression equations is very useful in data mining. One of the fundamental concepts operating within the data mining hyperspace is the cluster, which is formed of sets of feature vectors that are understood by examining their standard deviations. The tighter the vectors cluster, the better it is for classification purposes. In this case, the clusters are considered as good features, or gestalts.

Both rule induction and neural network data mining techniques fall under the category of machine learning (Hand, 1999), and they are based on various sophisticated and high-speed modeling techniques for predicting outcomes or uncovering hidden patterns. Tools for data mining are constantly emerging, and there are perhaps as many vendors of data mining software as data mining techniques. Some examples of data mining products and vendors are provided in Chapter Six. Frequently cited and used tools such as C&RT (classification and regression trees) and CHAID (chi-squared automatic induction), artificial neural networks (ANN), K-means, nearest neighbor, MBA (market basket analysis), MBR (memory based reasoning), automatic cluster detection, link analysis, decision trees, and genetic algorithms are most familiar to the data mining community.

Data mining is further divided into supervised and unsupervised knowledge discovery. Unsupervised knowledge discovery is to recognize relationships in the data and supervised knowledge discovery is to explain those relationships once they have been found (Berry and Linoff, 1997; Thearling, 1995; Westphal and Blaxton, 1998). Berry and Linoff (1997) described unsupervised knowledge discovery as a bottom-up approach that makes no prior assumptions; the data are allowed to speak for themselves.

To begin to better understand how data mining can be of use to institutional researchers is to examine the tasks performed and the tools used. Data mining tasks are categorized as follows: classification, estimation, segmentation, and description. Table 2.1 lists the tasks and the corresponding tools.

Table 2.1. Classification of Data Mining Tasks and Tools

Tasks	*Supervised*	*Unsupervised*
Classification	Memory based reasoning, genetic algorithm, C&RT, link analysis, C5.0, ANN	Kohonen nets
Estimation	ANN, C&RT	—
Segmentation[a]	Market basket analysis, memory based reasoning, link analysis, rule induction	Cluster detection, K-means, generalized rule induction, APRIORI
Description	Rule induction, market basket analysis	Spatial visualization

[a]For ease of understanding, the author includes tasks of affinity grouping, association, and clustering in Segmentation.

The main goal of a classification task is using data mining models to label output that is defined as a category of good-bad or yes-no. According to Berry and Linoff (2000), the estimation tasks refer to data mining models with outputs that are likelihood functions, or even more directly, sizes or length. Classification also functions for filling in missing values (data imputing). Segmentation includes tasks of affinity grouping and association, and clustering. Description has surpassed conventional visualization of a final outcome of data mining. Techniques or models used for descriptions are applicable to data modeling process in its entirety.

Cluster detection is inherently unsupervised data mining (Berry and Linoff, 1997) and decision trees are used for supervised data mining. C&RT and CHAID fall under this class. Genetic algorithms are similar to statistics, in that the form of the model needs to be known in advance. Genetic algorithms use the selection, crossover, and mutation operators to evolve successive generation of solutions. As the generations evolve, only the most predictive survive, until the functions converge on an optimal solution. When the inputs have many categorical variables, decision trees often work well. When the relationship between the inputs and the output is difficult to figure out, neural networks are frequently the technique of choice (Berry and Linoff, 1997).

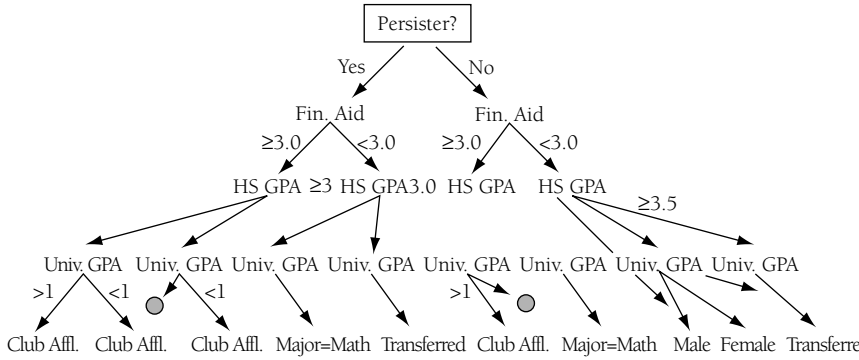
Neural networks work best when the nature of the data is nonlinear. Running neural networks may take a very long time due to back-propagating. Most neural networks rely on the process for the hidden layer to perform the summation and constantly adjust the weights until it reaches an optimal threshold, which then produces the outcome for a record (Garson, 1998). However, if the researcher terminates an active neural network at any given time while it is being trained on a test dataset, it will still provide useful results. Everything else being equal, fewer inputs will shorten the training time. Some in the data mining community advocate the use of decision trees before neural networks. Others prefer comparing both types of models simultaneously to see which one produces the most accurate model. The latter technique is called "bagging."

This chapter focuses on decision trees, also called rule induction technique, and back-propagation neural networks, a frequently used artificial neural network (ANN). These are also the tools selected in the case study presented later in the chapter.

Decision Trees (CART and C5.0). Decision trees, also called rule induction techniques, are fairly easy to explain, since the notions of trees, leaves, and splits are generally understood. Inductive reasoning refers to estimation of a sample while the population is known. Decision trees use splits to conduct modeling and produce rule sets. For instance, a simple rule set might say

> If financial aid = "Yes", and high school GPA ≥ 3.0 and
> If university GPA ≥ 3.5, then persistence = *yes* (confidence = 0.87)
> (*sub-rules suppressed . . .*)

Figure 2.1. Diagram of a Decision Tree



If high school GPA < 3.0 and major = "math", and
If club affiliation < 1, then persistence = *no* (confidence = 0.90)
(*sub-rules suppressed* . . .)
. . .

Heuristic based decision trees, also called rule induction techniques, include classification and regression trees (C&RT) as well as C5.0. C&RT handles binary splits best, whereas multiple splits are best taken by C5.0. If a tree has only two-way splits, it is considered a binary tree, otherwise a ternary tree. For most of their applications, decision trees start the split from the root (root node) into leave nodes, but on occasion they reverse the course to move from the leaves back to the root. Figure 2.1 is a graphical rendition of a decision tree (binary).

The algorithms differ in the criterion used to drive the splitting. C5.0 relies on measures in the realm of the Information Theorem and C&RT uses the Gini coefficient (SPSS, 2000). Rule induction is fundamentally a task of reducing the uncertainty (entropy) by assigning data into partitions within the feature space based on information-theoretic approaches (van den Eijkel, 1999). The mathematical formula on discerning uncertainty is expressed as measurements in bits:

$$H(N) = \sum_{n=1}^{n} - P(n) \log_2 P(n)$$

where H(N) is the uncertainty defined as discrete information and $P(n)$ is the probability that $\epsilon = n$. As uncertainty reduces, bits are reduced. Suppose the issue is a decision on yes versus no, the conditional information $H(N|yes)$ is expressed as follows:

$$H(N|yes) = \sum_{n=1}^{n} - P(n|yes) \log_2 P(n|yes)$$

As with any artificial intelligence, algorithms tend to continue indefinitely once executed (Garson, 1998). As in developing rule sets, decision trees may split into fine leaf nodes that render themselves incapable of predicting, because no future records will be similar at such a fine level of splitting. This is considered to be underfitting. However, overfitting, a trade-off between bias and variance, is also a concern in using these techniques. The method to control the extent to which the tree splits is called pruning. Short trees tend to have higher bias. If the leaf node is not completely developed, or the tree is pruned too soon, then most anything will look alike to the model. A case befitting this scenario would be stopping the split at the node level of the first occurrence of gender. The model may determine that the rest of the relationships within or between the records will not be examined. In this case, a student may be predicted to be successful no matter what he or she does, so long as the student's gender is female or male. At some point, the development of leaf nodes needs to stop. One school of thought in handling the degree with which a tree is considered properly pruned is Occam's razor, which states a simpler explanation is more likely to capture the essence of the problem (van den Eijkel, 1999). Only humans can intuitively make that decision. This reason alone is why humans most certainly cannot be replaced completely by any or a combination of all the data mining algorithms.

Chi-Squared Automatic Induction (CHAID). CHAID was developed by J. A. Hartigan, who borrowed an earlier work by J. A. Morgan and J. N. Sonquest in automatic induction detection (AID). CHAID limits itself to categorical variables. Continuous variables need to be changed into ranges or classes. However, one benefit of CHAID is its ability to stop the split before overfitting occurs.

Artificial Neural Networks (ANN). Our brain has 10^9 neurons interconnected in a complex way. There can be several thousand connections per neuron, potentially amounting to 60 trillion synapses (Garson, 1998). The precise manner of how neurons, or the inner layer of these neural networks, operate remains unknown (Brieman, 1994). The artificial intelligence developed in the 1960s underwent tremendous modification to better mimic the inner functions of the brain. The current theory is a revisit to the Pavlovian theory, further advanced by renowned neurological system scientist Donald Hebb, who theorized that learning is a result of the strength of the synaptic connections, rather than the older concept that learning is a result of manipulations of symbols (Statsoft, 2001).

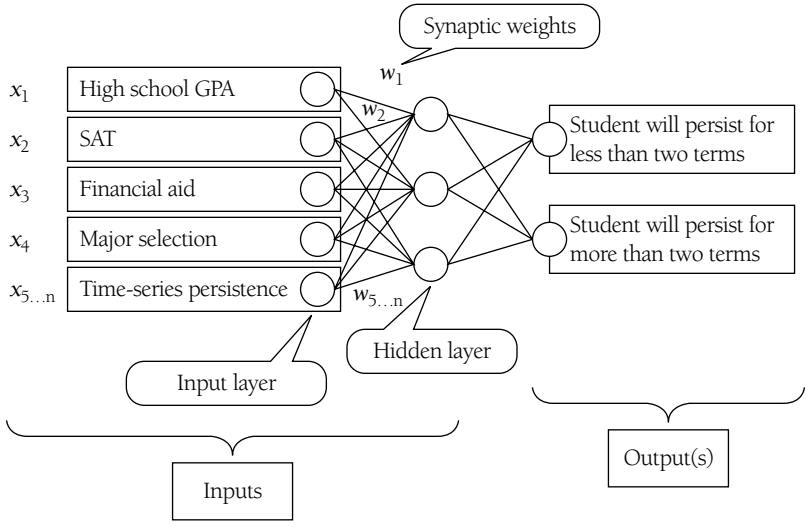
Developed as a mathematical rendition to explain the function of the nervous system by neurophysiologist Warren McCulloch and logician

Walter Pitts, their concept of an artificial neural net composed of binary-valued neurons opened a brand new chapter in data analysis. Their mathematical model, as a step function, mimics the way nerve cells process information either for excitatory synapse or inhibitory synapse or for enhancing and reducing the transmitted signal (Silipo, 1999). Although biological neurons are inherently analog in nature, the artificial neuron by McCulloch and Pitts can perform Boolean operations (Not, Or, and And) with the proper adjustment of weights and threshold at which a perceptron would produce an output (called neuron firing). All contemporary neural networks bear the imprints of the McCulloch-Pitts model. The field of neural networks almost suffered irreparable failure in the late 1960s due to the discovery of its inability to model the Boolean operation of exclusive-OR (XOR) by two researchers at MIT. It was not until the 1980s, when John Hopfield invented the back-propagation method (sometimes called error back propagation), that interest in ANN was rekindled. The networks feed back through the network errors discovered in prediction, modifying the weights by a small amount each time, until all example records have been processed, perhaps many times, while discarding unneeded inputs (Watkins, 2000). With the advent of computing technology, ANN flourishes today.

Figure 2.2 presents an example of a multilayer perceptron (MLP) of an ANN with two outputs, which is used in the case study presented later in this chapter.

Several authors have described mathematically the formulae for back-propagating neural networks (Hand, 1999). In a book helpful for institutional researchers, *Neural Networks: an Introductory Guide for Social Scientists,*

Figure 2.2. Diagram of the Case Study Neural Net



Garson (1998) treated the inner workings of an MLP as a weighted summation function and a sigmoid transfer function. He explained the process of dendrites (inputs) passing information to reach a threshold for axons (output) to signal the connected neural nets using the mathematical formulae

$$o_j = f\left(\sum_{i=1}^{n} o_i w_{ji}\right), \text{ where } f'(x) = -(1 + e^{-x})^{-2} e^{-x}(-1)$$

where o is the outcome, x_i is the input vector, w_i is the weight, which is set at random upon first feed.

Kohonen Neural Networks. There are many alternative neural networks. One of the most well known is the Kohonen neural network. Developed by Finnish researcher Tuevo Kohonen, Kohonen neural networks primarily act as an unsupervised knowledge discovery technique. Garson (1998) stated that Kohonen nets are estimators of the probability density function for the input vector. Some data miners call it self-organizing maps (SOM), which means intuitively that the outcome is a result of allowing the algorithms to analyze the variables until certain patterns emerge (Berry and Linoff, 2000; Silipo, 1999). In formulaic terms,

$$d_j = \sqrt{\left[\sum_{i=1}^{n} (x_{ji} - w_{ji})^2\right]}$$

where the Euclidean distance d of neuron j is the sum of the squared distances from x (inputs) and the assigned weights of x.

Kohonen nets are useful for discerning patterns and groups within a feature space. Researchers may use Kohonen nets to learn about the data before building other models. They have great value in understanding who takes what clusters of courses or what groups of students tend to have similar course-taking patterns.

Statistics, Data Mining, and On-Line Analytical Processing

Delmater and Hancock (2001, p. 192) wrote: "The science underlying predictive modeling is a mixture of mathematics, computer science, and domain expertise." Their point is very well taken and is the focus of this section. Data mining is a knowledge discovery process to reveal patterns and relationships in large and complex data sets (De Veaux, 2000). Moreover, data mining can be used to predict an outcome for a given entity. The ultimate reason for carrying out pattern identification or rule setting is to use the knowledge gained from this exercise to influence the policy makers.

Most of the processes involved in data mining are explainable by mathematics, statistics in particular, and are to a certain extent familiar to researchers who are comfortable with explication statistics. Even in the so-called "data fishing expedition" of conducting unsupervised data mining, the algorithms are still based on logics and formulae.

Table 2.2, which I developed, provides a crosswalk comparison among the major concepts in data mining, statistics, and on-line analytical processing (OLAP). A crosswalk like this provides a guide for understanding data mining terminologies and concepts. It is not intended, however, to be all-inclusive, as researchers can spend a lifetime collecting and categorizing the ever-growing data mining models (Garson, 1998) and terminologies.

In the early days of computing when classical statistics were the only tools of choice, reducing data size was crucial (Berry and Linoff, 1997). The power delivered by data warehousing to data mining software has challenged traditional statistical methodologies (Mena, 1998). Rather than approaching a problem in a limited source domain that typically is a sample of data identified by the guidance of *a priori* hypotheses, researchers can now overlay data mining algorithms on the entire population. This entire

Table 2.2. Crosswalk of Data Mining Models and Algorithms to Statistics and Data Warehouse Based OLAP

Data Mining	*Statistics*	*Data Warehousing, OLAP*
Artificial neural networks	Regression equations, chi-square, structural equations	—
Rule induction	Principle components, discriminant function, factor analysis, logistic R	—
Kohonen networks	Cluster analysis, probability density function	Multidimensional cube
Spatial visualization	Two–Three dimension charts	Two–Three dimension charts
Euclidean space	Structured equations, linear and non-linear regression	Sequential files
Classification	Logistical regression	Multidimensional cube
Estimation	Regression equations, chi-square, structural equations	—
Segmentation	Cluster analysis, factor analysis	Multidimensional cube
Prediction accuracy	Statistical significance	Temporal, trend reporting
Outliers detection	Standard deviation, error analysis	Aggregation
Supervised learning	Hypothesis, distributional assumptions, APRIORI	—
Unsupervised learning	Descriptive statistics, cluster analysis	Temporal, trend reporting
Population, universe	Samples	Fact tables and dimension tables
Feature vectors	Histogram, correlation	Cross-tabs
Feature extraction	Flat files	Extract, transform, load (ETL)
Machine learning, artificial intelligence	Mathematics	Structured query language (SQL)
Attributes, features	Variables, values	Fields, records
Outputs or scoring	Independents	Fields

population can be terabytes in size, and in the very near future data mining modeling can happen to live data, called knowledge discovery in databases.

In this case, the typical steps taken by researchers to make statistical assumptions about the population are not necessary. However, understanding the database in which data reside and the data characteristics (structured and unstructured) are essential to successful data mining. Throwing all variables in a database for data mining is not conducive to machine learning. For instance, a researcher may want to identify patterns of persistence. In the dataset entered, student social security number and the corresponding college assigned student IDs and student names served no other purpose than confusing the algorithms and hogging memory. They need not be selected. However, addresses may reveal important information about a student's inclination to relocate. Sometimes the use of factor analysis or principal components to root out the auxiliary features may be desirable. The use of several techniques to cross-validate particular extrapolatives, including classical statistical techniques, is a recommended approach, called "bagging."

Data mining works best in exploratory analysis scenarios that have no preconceived assumptions (Westphal and Blaxton, 1998). *A prior* hypothesis may guide classical statistical approach but cloud the judgment of a data miner. Data mining, neural networks in particular, is most useful for prediction and scoring but not for casual statistical analysis (Garson, 1998). If the traditional methods can be viewed as top down, data mining is truly bottom up. Research questions do not begin with "what *if*," instead, they begin with "what *is*" (Luan and Willette, 2001).

As expressed earlier in the chapter, basic and classical statistical knowledge is highly useful to a data miner in discerning minute significance in cluster boundaries—an important data mining task. Neural networks are in essence regression models adapted to conduct estimation. In this sense, what was a concern to a researcher, statistical significance, is now a question of how it translates into accuracy of the prediction or classification. Data mining has set researchers free by taking the chore of making distributional assumptions about data out of their hands and giving them the power of applying machine learning models to new data. This is how data mining transforms a researcher armed with statistical skills into a data miner who drives the engine of pattern recognition and behavior prediction.

Current Trends in Data Mining

Evolving out of traditional statistics, data mining started as an independent set of tools. More and more, visualization and database data mining are adopted. Conventional visualization techniques are aimed at the executives who are information consumers. Spatial visualization provides visual plots depicting members of the population in their feature space. It is not aggregation based computation, but faithful (powerful) rendition of the geometric relationships, be it orientation, density, or clustering (Delmater and

Hancock, 2001). Also, knowledge discovery in database attempts to seamlessly integrate data mining with databases, so as to eliminate the extra work of producing additional datasets. Knowledge discovery in database maintains data consistency and, most crucially, makes real-time scoring possible. Both these trends are here to stay.

Fuzzy Logic. Another data mining algorithm being developed is fuzzy logic, which can be applied to both rule induction techniques and neural networks. Silipo (1999) argued that the opaque nature of all neural network operations would be diminished via the implementation of fuzzy logic due to its relatively transparent decisional algorithms. Berthold (1999) applied fuzzy logic to imprecise data, most commonly found in social science where crisp measurements do not exist. Even though regular neural networks have redundancy computation built in, which alleviates some of the damage done by data degradation (Silipo, 1999), the use of fuzzy logic is deemed a good alternative.

Genetic Algorithm. Genetic algorithm (GA) is an optimization algorithm developed by John Holland at the University of Michigan. It is based on the two basic rules that govern the vast organic world, selection and variation. Genetic algorithm uses selection, crossover, and mutation parameters in evolutionary computation in reaching the solution (Jacob, 1999; Berry and Linoff, 1997). Genetic algorithms fall under the class of stochastic search methods.

Applications of Data Mining

Data mining has been recently discovered by academia but was first put to full use by the Fortune 500, who have since benefited tremendously. Data mining was behind numerous successful market campaigns and quality assurance. Table 2.3 depicts some of the core questions most often used in the business world and their analogs in higher education.

Table 2.3. Comparison of Data Mining Questions in Education and the Corporate World

Questions in the Business World	*Counterpart Questions in Higher Education*
Who are my most profitable customers?	Who are the students taking most credit hours?
Who are my repeat website visitors?	Who are the ones likely to return for more classes?
Who are my loyal customers?	Who are the persisters at our university, college?
What clients are likely to defect to my rivals?	What type of courses can we offer to attract more students?

Data mining was first implemented for marketing outside higher education. It certainly has parallel implications and value in higher education. As discussed earlier, marketing is part of learner relationship management. Marketing concerns the service area, enrollment, annual campaign, alumni, and college image. Combined with institutional research, it expands into student feedback and satisfaction, course availability, and faculty and staff hiring. A university service area now includes on-line course offerings, thus bringing the concept of mining course data to a new dimension. Data mining is quickly becoming a mission critical component for the decision-making and knowledge management processes.

Exploring Data Mining in Higher Education: A Case Study

Using data mining to monitor and predict community college students' transfer to four-year institutions provides significant benefits for decision makers, counselors, and students. For years, institutional researchers have not been able to clearly pinpoint the type of students who transfer and their course taking patterns. Analyses of the outcomes of transferred students in upper divisions can influence the curriculum design back at the community colleges. Data mining helps predict the transferability of each currently enrolled student. A model developed in this case study is aimed at providing a profile of the transferred students and predicting which student currently enrolled in a community college will transfer so that the college can personalize and time their interactions and interventions with these students, who may need certain assistance and support. This embodies the principles of learner relationship management.

A data exchange consortium, led by the planning and research office of Cabrillo College, including Cabrillo College, University of California Santa Cruz, San Jose State University, and California State University Monterey Bay, established in 1998 a longitudinal data warehouse of transferred students, including all their course information. Taken together, the records cover 75 percent of the total annual transfers from Cabrillo College. The transfer data warehouse is then combined with the existing data warehouse of the planning and research office at Cabrillo College to provide unitary records for every student from the moment they enrolled at Cabrillo College to the day they graduated from the four-year institution. This data gold mine holds answers to many policy and research questions.

Data Mining Approach to Transfer Data. Both University of California Santa Cruz (UCSC) and San Jose State University provided data going back to 1992. California State University Monterey Bay, created in 1995, did not participate in this study because of the recency of their data. I spent a significant amount of time staging the data that came from three disparate sources. Cross industry standard process for data mining (CRISP-DM)

lent guidance for this endeavor, and I have also listed steps in data preparation in the appendix of this chapter. It is a major rule in the data mining community that a data mining project cannot be successful if the investigator is not a domain expert who is very tuned to the granular data. The investigator must also have adequate skills in feature extraction, where more than 65 percent of the time can be spent on getting the features and attributes correctly presented and primed for mining purposes. The current trend is for researchers to educate themselves to master these contrasting sets of skills in order to adapt to the changing world of knowledge management.

With the outcome of transfer of students being clearly known, this was a supervised data mining. Owing to the need for predicting transfers and, as a consequence, planning for contacting these transfer-directed students, the dataset includes as much enrollment history and demographic information as possible for every student who had ever attended Cabrillo College, transferred or not. This constitutes a considerably deep and wide feature extraction. The evolution of this project is chronicled as follows.

Transfer tracking is a matter of latency. Research showed that it typically takes two to four years for the majority of the cohort to transfer. Therefore, the first task was to decide which time series in the database to use. The first year when data were available in the planning and research office data warehouse was 1992, which meant the corresponding first year of the two universities' data should be 1994. Since the last year of data that were congruent to each other from these universities was 1998, the cohort therefore should be former Cabrillo College students who enrolled between summer 1992 and spring 1997. The second task was to edit every field so that indexes could function properly. Many hidden problems, such as different coding for social security numbers and terms, were uncovered at this point, thus preventing future problems downstream. Also, data fields sent from the same university were not the same each year, when a new person was wearing the database administrator hat. The third task was to tackle the so-called "deep and wide" enrollment history data due to the nature of the original data source, governed by a transactional data structure, which meant that the enrollment data were highly normalized with enrollment records repeating for as many rows as needed for each student. Although data mining algorithms would run directly using this setup, it could only produce completely erroneous conclusions. Each subsequent enrollment record of a student needed to become a field by itself, a requirement that brought on the issue of dealing with potentially many dozens of fields for just the courses taken without yet introducing the grades for the courses and the term in which the student took each of the courses. The final dataset was a result of collapsing courses based on their type (transfer, remedial, vocational) at the expense of term and individual grades.

The total number of students in the dataset was thirty-two thousand. A proprietary data split algorithm divided the set into a test set and a validation set. Data mining was applied to the test set, until such time when the

models were considered optimal. The validation set, first time seen by all the models, was brought in for actual scoring.

The following is a partial list of the groups of features (fields) selected for this case study:

- Demographics: age, gender, ethnicity, high school, zip codes, planned employment hours, education status at initial enrollment
- Financial aid
- Transfer status (doubled as the reference variable)
- Total transfer, vocational, basic skills, science, and liberal arts courses taken
- Total units earned and grade points by course type

Clementine, a software by SPSS Inc., enjoys a reputation for being the easiest model to deploy. The study chose Clementine as the data mining software (please refer to Chapter Six for data mining tools). Experience led me to use neural networks (NN) and two rule induction algorithms, C5.0 and C&RT, to compare models and to complement the scoring. As already mentioned, some data mining experts call this "bagging." The type node in Clementine coded the fields into appropriate types, and the balance node reduced the imbalance between transferred and nontransferred students, which was quite large initially.

The NN model resulted in an accuracy of 76.5 percent. It contained fifty-two neurons, seven hidden neurons, and one neuron (for a dichotomous output). The top ten fields listed in the relative importance of inputs were

Number of liberal arts classes taken (0.315)
High school origin (0.189)
Race (0.161)
Planned work hours (0.159)
Initial education status (0.145)
Grade points (0.085)
Number of nonbasic skills courses taken (0.084)
Number of UCSC transferable courses taken (0.081)
Gender (0.079)
Number of degree applicable courses taken (0.074)

The values in parenthesis would range from zero to one, but in practice they were rarely above the 0.35 threshold. A couple of points worth noting here are that the investigator should pay attention to every field, even the one listed at the bottom, as data mining is both a task to identify the averages and rule of thumbs and a task to use outliers for a number of reasons, such as fraud detection. As neural networks results were a bit cryptic, it was necessary to use a rule induction model to list the rules uncovered. The following resulted from C5.0:

Rules for Transferred
Rule #1 for Transferred:
if units > 12
and # of nontransfer course ≤ 5
and # of math > 0
then transferred → (452, 0.877)

Rule #2 for Transferred:
if gender = F
and # of nontransfer course ≤ 5
and # of math > 0
then → transferred (278, 0.871)

Rule #3 for Transferred:
if age >19.9
and age ≤ 24
and grade points ≥ 5
and # of UCSC transferable courses > 0
and # of precollegiate basic skills course ≤ 0
and # of vocational course ≤ 5
and # of math ≤ 0
then → transferred (29, 0.806)

Rules for Not Transferred:
Rule #1 for Not Transferred:
if race = Hispanic
and # of SJSU transferable course ≤ 21
and # of nontransferable course > 6
and # of math ≤ 3
then → not transferred (24, 0.962)

Rule #2 for Not Transferred:
if # of UCSC transferable course ≤ 7
then → not transferred (403, 0.736)

The first value in the parenthesis was the number of cases supporting this rule and the second value the confidence. The case study then used the C&RT node to generate a decision tree with the following tree branches:

Units < 21.5 [mode: not transferred] (369)
Units < 5.5 (156, 0.955) → not transferred
Units ≥ 5.5 [mode: not transferred] (213)
NTRCRS < 2.5 [mode: not transferred] (165)
Nontransferable courses ≥ 2.5 (48, 0.938) → not transferred
UNITS ≥ 21.5 [mode: transferred] (974)

MATH ≤ 0.5 [mode: transferred] (197)
UCCRS < 13.5 (83, 0.554) → not transferred
UCSC transferable courses ≥ 13.5 (114, 0.754) → transferred
Math ≥ 0.5 (777, 0.874) → transferred

Model Analysis. Clementine provides an efficient way to compare the classification for the test set and the scoring for the validation set. Table 2.4 contains the matrixes detailing these findings.

As indicated by these matrixes, the neural networks model produced decent and somewhat balanced accuracy but not as good when compared to the C&RT model. C5.0 provided the highest accuracy for predicting students who had transferred, but it was far less accurate in predicting nontransferred. Overall, C&RT appeared to be the best model to use.

C5.0 initially produced a perfect estimation with close to 100 percent accuracy. This was a signal of the model memorizing the rules, not necessarily learning the rules. Adjustment in the number of records allowed for each split and quickly eliminated this problem. During this process, the dataset had to be rebuilt twice due to informational redundancy (correlation) concerns.

Data mining is an iterative process and identifying patterns is even more so. It is highly possible that with enough time devoted to preparing the data and adjusting the model, a higher accuracy rate (<90 percent) is possible. Ideally, the research department will be able to overlay data mining on college data warehouse and use the above model to score new students on a yearly basis. This is a true end-to-end data mining solution. The counseling department can use the list containing students scored to be "transferring inclined" for targeted mailing and personalized assistance.

Table 2.4. Matrixes of Model Performance for Test and Validation Sets

Neural Networks on Test Set			On Validation Set		
	NoTran	Tran		NoTran	Tran
NoTran	67.9	32.1	NoTran	78.7	21.3
Tran	20.8	79.2	Tran	22.5	77.5
C5.0 on Test Set			**On Validation Set**		
	NoTran	Tran		NoTran	Tran
NoTran	72.1	28.0	NoTran	70.0	30.0
Tran	12.5	87.5	Tran	8.0	92.0
C&RT on Test Set			**On Validation Set**		
	NoTran	Tran		Notran	Tran
NoTran	81.3	18.7	NoTran	82.8	17.2
Tran	18.4	81.6	Tran	17.9	82.1

There are three additional strategies researchers may use when conducting data mining. The first is verifying the results by classical statistics for which Clementine has provided nodes such as linear regression and logistic regression. Applying the logistic regression node to the test set resulted in an identified group of most significant features, but they are ordered differently either due to their level of significance or the internal functions of the model. Nonetheless, it covered the spectrum very well. The second strategy is to use factor analysis and principle component analysis to weed out nonsignificant variables or variables that are highly correlated with each other. However, it is worthwhile to point out that data mining is very tolerant of correlated variables, compared to the classical statistics with which we are familiar. The third strategy is one that I highly recommend. The researcher should consider clustering and segmentation analysis using TwoStep, K-means, or Kohonen even though the target field(s) is known. For example, K-means can reveal that students sharing similar characteristics may form five or six giant clusters in the data. This gives the researcher additional insights into the population and may prompt the researcher to divide the population into cluster datasets with which the data mining algorithm can significantly increase its accuracy. I applied this strategy when mining for persistence and found that by concentrating on the students who were clustered by their educational goals and the type of courses taken, the model produced far better results.

Conclusion

Synthesizing the vast amount of research and ideas and condensing them into one chapter with the aim of introducing data mining to the institutional research audience in higher education is a great challenge. By using well-defined algorithms from the disciplines of machine learning and artificial intelligence to discern rules, associations, and likelihood of events, data mining has profound application significance. If it were not for the fast, vast, and real-time pattern identification and event prediction for enhanced business purposes, there would not have been such an exponential growth in dissertations, models, and the considerable amount of investment in data mining in the corporate world.

As we have discovered, insights from data sets and variable lists, previously seen as unwieldy and chaotic, can be obtained with data mining and developed into the foundations for program planning or to resolve operational issues. The power of data mining lies in the fact that it simultaneously enhances output and reduces cost. The One-Percent doctrine (Luan, 2000) states that a 1 percent change means one unit of gain and one unit in savings. For example, a 1 percent increase in enrollment may mean $500,000 for a typical college of twenty thousand students, and it is achieved with no additional cost. Data mining conducted for alumni

donations may correctly pinpoint the right donors and the right target amount. This saves campaign costs and increases the campaign's effectiveness. The ability to provide intervention to individual students who are seen as likely to drop out or to transfer also holds value beyond cost and savings. Data mining conducted to predict the likelihood of an applicant's enrollment following their initial application may allow the college to send the right kind of materials to potential students and prepare the right counseling for them. The potential of data mining in education cannot be underestimated.

Appendix: Steps for Data Mining Preparation, Based on Cross Industry Standard Process for Data Mining (CRISP-DM)

- *Step One.* Investigate the possibility of overlaying data mining algorithms directly on a data warehouse. Doing so may require extra effort and diplomatic skills with the information technology department, but it pays off in the long run. It avoids possible errors in field names, unexpected changes in data types, and extra effort to refresh multiple data domains. The scoring can also be directly performed to live database. This is called an end-to-end data mining solution, also called knowledge discovery in database (KDD). (Total time usage: 5 to 15 percent.)
- *Step Two.* Select a solid querying tool to build data mining files. These files closely resemble multidimensional cubes. As a matter of fact, MOLAP (multidimensional on-line analytical processing) serves this purpose well. Except for APRIORI, which can use transactional data files directly (alas!), all other algorithms need "tabular" files, which are relational database files queried to produce a file with unique records with multiple fields. A number of querying tools are available for this purpose. SQL skills are highly desirable. This step can be most time consuming. (Total time usage: 30 to 75 percent.)
- *Step Three.* Data visualization and validation. This means examining frequency counts as well as generating scatter plots, histograms, and other graphics, including clustering models. A graph is the best indication for a correlation estimate. This step gives the researcher the first impression of what each of the data fields contains and how it may play out in the analysis. Missing data should not be treated in the same manner in every situation. In certain cases, missing data are extremely diagnostic. In data mining, the outliers may be just what we are looking for, simply because they deviate from the norm. Therefore, they may hold truth for discovering previously unknown patterns. In fraud detection, it is these outliers that will flag the system to avoid loss. (Total time usage: 10 to 20 percent.)
- *Step Four.* Mine your data! (Total time usage: 10 to 20 percent.)

References

Berry, M., and Linoff, G. *Data Mining Technique: For Marketing, Sales, and Customer Support.* New York: Wiley Computer Publishing, 1997.

Berry, M., and Linoff, G. *Master Data Mining: The Art and Science of Customer Relationship Management.* New York: Wiley Computer Publishing, 2000.

Berthold, M. "Fuzzy Logic." In M. Berthold and D. Hand (eds.), *Intelligent Data Analysis.* Milan: Springer, 1999.

Brieman, L. "Comment." *Statistical Science,* 1994, *9*(1), 38–42.

Delmater, R., and Hancock, M. *Data Mining Explained: A Manager's Guide to Customer-Centric Business Intelligence.* Boston: Digital Press, 2001.

De Veaux, R. "Data Mining: What's New, What's Not." Presentation at a Data Mining Workshop, Long Beach, Calif., 2000.

Garson, G. D. *Neural Networks: An Introductory Guide for Social Scientists.* London: Sage, 1998.

Gartner Group. "The GartnerGroup CRM Glossary." [http://www.gartnerweb.com /public/static/hotc/hc00086148.html].

Hand, D. "Introduction." In M. Berthold and D. Hand (eds.), *Intelligent Data Analysis.* Milan: Springer, 1999.

Jacob, C. "Stochastic Search Method." In M. Berthold and D. Hand (eds.), *Intelligent Data Analysis.* Milan: Springer, 1999.

Luan, J. "An Exploratory Approach to Data Mining in Higher Education: A Primer and a Case Study." Paper presented at the AIR Forum, Seattle, Wash., 2000.

Luan, J., and Willette, T. "Data Mining and Knowledge Management." Paper presented at the Research and Planning Group Conference, Lake Arrowhead, Calif., 2001.

Mena, J. "Data-Mining FAQs." *DM Review,* January 1998. [http://www.dmreview.com /master.cfm?NAVID=198&EdiD=792].

Roueche, J. E., and Roueche, S. S. *High Stakes, High Performance: Making Remediation Work.* Washington, D.C.: Community College Press, 1999.

Rubenking, N. "Hidden Messages." *PC Magazine,* May 22, 2001, *20*(10), 86–88.

Silipo, R. "Neural Networks." In M. Berthold and D. Hand (eds.), *Intelligent Data Analysis.* Milan: Springer, 1999.

SPSS. *SPSS Clementine 6.0 User's Guide.* Chicago: SPSS, 2001.

Statsoft. [http://www.statsoft.com/textbook/glosfra.html], 2001.

Thearling, K. "An Overview of Data Mining at Dun and Bradstreet." *DIG White Paper,* 1995. [http://www3.shore.net/~kht/text/wp9501/wp9501.htm].

van den Eijkel, G. C. "Rule Induction." In M. Berthold and D. Hand (eds.), *Intelligent Data Analysis.* Milan: Springer, 1999.

Watkins, D. "Neural Network Master Class." Presented at Clementine User Group (CLUG), July 2000, Reading, United Kingdom.

Westphal, C., and Blaxton, T. *Data Mining Solutions: Methods and Tools for Solving Real-World Problems.* New York: Wiley Computer Publishing, 1998.

JING LUAN is chief planning and research officer at Cabrillo College in Aptos, California.