

This following is Chapter 2 in the New Direction for Institutional Research #113, published by Jossey-Bass. This chapter is reformatted for web by the author.

To cite this paper:

Luan, J. (2002) Chapter 2: Data Mining and Its Applications in Higher Education. Knowledge Management – Building a Competitive Advantage in Higher Education. Serban, A. & Luan, J (eds.) Jossey-Bass.

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 ©

Introduction

In the first chapter, readers have had an opportunity to review the definitions and components of knowledge management. The chapter has 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 not complete) 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 the 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, 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 the scholars in higher education, but data mining was not first applied to higher education. Suffice it to say that higher education is still a virgin territory for data mining. Just 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 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 could not even be 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 first known for having exotic names, the field has begun to include certain kind of descriptive statistics and visualization techniques into data mining (Westphal and Blaxton, 1998). Statsoft, an online 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 non-hypothesis based analytical techniques as data mining and adopt the concept developed by Berthold and Hand that all statistical techniques developed prior to the 60s are "classic" and "conformatory."

The definition from Gartner Group seems to be most comprehensive, as they define data mining as "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." The author refines the notion of data mining as the purpose of uncovering hidden trends and patterns and making accuracy based predictions through higher level of analytical sophistication. It is producing new observations from existing observations. Or, 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 Online 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 threshold, or examining multi-way frequency tables (e.g., "slice by slice" systematically reviewing combinations of levels of control variables)." It reserves the term of 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 (e.g., Logit)

Regression, Correspondence Analysis, Time Series Analysis, and Classification Trees.

Essential Concepts and Definitions

Data mining assumes the existence of spherical multi-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 are 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 6. Frequently cited/used tools such as C&RT (Classification and Regression Trees) and CHAID (Chi-squared Automatic Induction), 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
- Description

Table 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*	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

* 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 non-linear. Running neural networks may take a very long time due to backpropagating. 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 produce the outcome for a record (Garson, 1998). However, if the researcher terminates an active neural network at any given time while it is being training 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 is called "bagging."

This chapter focuses on Decision Trees, a Rule Induction technique and Back-propagation Neural Networks, a most frequently used Artificial Neural Networks, as both of them are the tools selected in the case study presented later in the chapter.

Decision Trees (CART and C5.0)

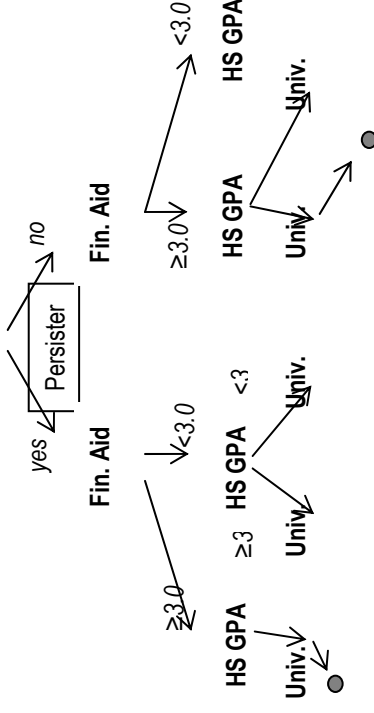
Decision trees, also called Rule Induction techniques, are easier to explain, since the notions of trees, leaves, and splits are generally understood. Inductive reasoning refers to estimation of a sample while the population is known. Decision trees use splits to conduct modeling and produce rule sets. 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...*)
 - 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 the best, while multiple splits are best taken by C5.0. If a tree has only two-way splits, it is considered binary tree, otherwise, a ternary tree. For most of its 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 1 is a graphical rendition of a decision tree using rules similar to the above.

Figure 1: Diagram of A Decision Tree



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 utilizes 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 (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 $\xi = n$. As uncertainty reduces, bits are reduced. Suppose the issue is a decision on yes vs. 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." On the other hand, 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 person'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 the Occam's Razor that states a simpler explanation is more likely to capture the essence of the problem (Eijkel, 1999). Only humans can intuitively make that decision. This reason alone is why most certainly humans cannot be replaced completely by any or a combination thereof all the data mining algorithms.

CHAID

CHAID (Chi-squared automatic induction) 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)

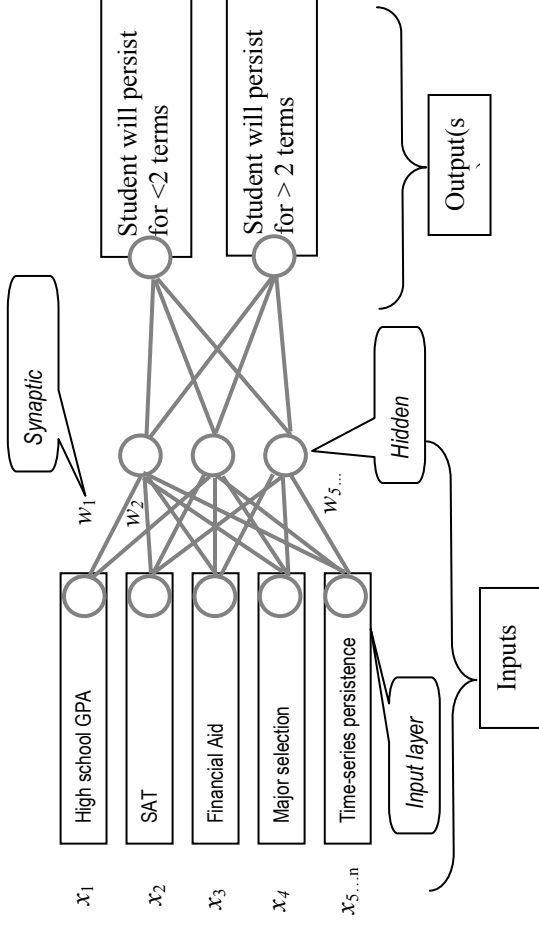
Our brain has 10^9 neurons interconnected in a complex way. There can be several thousand connections per neuron, which can potentially amount to 60 trillion synapses (Garson, 1998). The precise manner of how neurons, or the inner layer of these neural networks operate remains unknown (Brieman, 1994). The artificial intelligence developed in the 60s 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 was 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 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 backpropagation method (sometimes called error back propagation) that once again rekindled the interest in ANN. 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 presents an example of a Multi-layer Perceptron (MLP) of an ANN with two outputs, which is used in the case study presented later in this chapter.

Figure 2: Diagram of the Case Study Neural Net



Several authors have described mathematically the formulae for a backpropagating neural network (Hand, 1999). In a book helpful for institutional researchers, "Neural Networks, an Introductory Guide to Social Scientists," 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. It has 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 Online Analytical Processing

Delmater and Hancock (2001, p192) 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 to carry out pattern identification or rule setting is for using 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, developed by the author, provides a crosswalk comparison among the major concepts in data mining, statistics and online 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.

Table 2: Cross-walk of data mining models/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	Multi-dimensional cube
Spatial visualization	2-3 dimension charts	2-3 dimension charts
Euclidean Space	Structured equations, Linear and non-linear regression	Sequential files
Classification	Logistical regression	Multi-dimensional cube
Estimation	Regression equations, Chi-square, Structural equations	...
Segmentation	Cluster Analysis, Factor Analysis	Multi-dimensional cube
Prediction accuracy	Statistical significance	Temporal/trend reporting
Outliers detection	Standard deviation, error analysis	Aggregation
Supervised learning	Hypothesis, distributional assumptions, a-priori	...
Unsupervised learning	Descriptive statistics, Cluster Analysis	Temporal/trend reporting

DATA MINING	STATISTICS	DATA WAREHOUSING/OLAP
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

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 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. On the other hand, 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 technical 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 pre-conceived 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."

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 concern 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 by 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. Recently, visualization and database data mining are increasingly 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 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 3 depicts some of the core questions most often used in the business world and their analogs in higher education.

Table 3. Comparing data mining questions in education and the corporate world

Questions in the Business World	Counter-part 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?

Questions in the Business World	Counter-part Questions in Higher Education
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 the Learner Relationship Management. Marketing concerns the service area, enrollment, annual campaign, alumni, college image, and combined with institutional research, it expands into student feedback and satisfaction, course availability, faculty and staff hiring. A university service area now includes online course offerings, which brings the concept of mining the course data in 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 predict 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% 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 when they graduated from the 4-year institution. This data goldmine holds answers to many policy and research questions.

Data Mining Approach to Transfer Data

Both University of California Santa Cruz and San Jose State University provided data going back to 1992. California State University Monterey Bay, created recently in 1995, did not participate in this study because the recency of their data. The author 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 the author has 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% 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 as possible enrollment history and demographic information 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 2 - 4 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, which prevented 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, which 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 32,000. 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, Liberal Arts Courses Taken

Total Units Earned, and Grade Points by Course Type

Clementine, a software by SPSS Inc., enjoys the reputation for being the easiest to deploy models. The study chose Clementine as the data mining software (please refer to Chapter 6 for data mining tools). Experiences led the author to use Neural Networks (NN) and two rule induction algorithms, C5.0 and C&RT, to compare models and to compliment 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 non-transferred students, which was quite large initially.

NN model resulted in an accuracy of 76.5%. It contained 52 neurons, 7 hidden neurons, and 1 neuron (for a dichotomous output). The top 10 fields listed in the Relative Importance of

Inputs were:

Number of Liberal Arts classes taken (.315)

High School Origin (.189)

Race (.161)

Planned Work Hours (.159)

Initial Education Status (.145)

Grade Points (.085)

Number of Non-basic skills courses taken (.084)

Number of UCSC transferable courses taken (.081)

Gender (.079)

Number of degree applicable courses taken (.074)

The values in parenthesis would range from 0 to 1, but in practice they were rarely above the .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, 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 Non-Transfer Course <= 5
  and # of MATH > 0
  then -> Transferred (452, 0.877)
```

Rule #2 for Transferred:

```
if GENDER == F
and # of Non-Transfer 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 Pre-collegiate 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 Non-transferable 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)
      Non-transferable 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 4 contains the matrixes detailing these findings.

Table 4. Matrixes of Model Performance for Test and

Validation Sets

Neural Networks on Test Set		On Validation Set	
NoTran	Tran	NoTran	Tran
67.9	32.1	78.7	21.3
20.8	79.2	22.5	77.5

C5.0 on Test Set		On Validation Set	
NoTran	Tran	NoTran	Tran
72.1	28.0	70.0	30.0
12.5	87.5	8.0	92.0

C&RT on Test Set		On Validation Set	
NoTran	Tran	NoTran	Tran
81.3	18.7	82.8	17.2
18.4	81.6	17.9	82.1

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 non-transferred. Overall, C&RT appeared to be the best model to use.

C5.0 initially produced a perfect estimation with close to 100% 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%) 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.

There are three additional strategies researchers may use when conducting data mining. One is that the results can be verified 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 non-significant 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 classical statistics that we are familiar with. The third strategy is highly recommended by the author. 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 5 or 6 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 data mining algorithm can significantly increase its accuracy. The author 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

It is a great challenge to synthesize the vast amount of research and ideas and to condense them into one chapter for the aim of introducing data mining to the institutional research audience in higher education. 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 enhances output and at the same time reduces cost. The One-percent Doctrine (Luan, 2000) states that a one-percentage point change means one unit of gain and one unit in savings. For example, a one-percentage point increase in enrollment change may mean \$500,000 for a typical college of 20,000 students and it achieves such 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 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 the potential student and prepare the right counseling for him or her. The potential of data mining in education cannot be underestimated.

Addendum

Steps for Data Mining Preparation

(Based on Cross Industry Standard Process for Data Mining)

Step One – Investigate the possibility of overlaying data mining algorithms directly on a data warehouse. This may require extra efforts and diplomatic skills with the IT department, but it pays off in the long run. This 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 End-to-End data mining solution, also called Knowledge Discovery in Database (KDD). (Total time usage: 5% - 15%).

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 online analytical processing) serves this purpose well. Except for APPORRI, 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% - 75%).

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 they may play out in the analysis. Missing data should not be treated in the same manner in every situation. In certain cases, missing data is 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 in discovering previously unknown patterns. In fraud detection, it is these outliers that will flag the system to avoid loss. (Total time usage: 10% - 20%).

Step Four – Mine your data! (Total time usage: 10% - 20%).

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 Berthold, M. & Hand, D. (eds.), *Intelligent Data Analysis*. Milan: Springer, 1999.
- Breiman L. “Comment.” *Statistical Science*, 9(1), 1994, pp. 38-42
- Crowley, B. “Knowledge Management for the Information Professional.” In Srikantaiah and Koenig (eds.), *Tacit Knowledge and Quality Assurance: Bridging the Theory-Practice Divide*. Medford: Informational Today Inc., 2000.
- Davenport T. and Prusak L. *Working Knowledge. How organisations manage what they know*. Boston: Harvard Business School Press, 1998.
- Delmater R., and Handcock 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, CA, 2000.
- Garson, G. D. *Neural Networks. An introductory guide for Social Scientists*. London: SAGE, 1998.
- Hand, D. “Introduction.” In Berthold, M. and Hand, D. (eds.), *Intelligent Data Analysis*. Milan: Springer, 1999.
- Jacob, C. “Stochastic Search Method.” In Berthold, M. and Hand, D. (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, WA, 2000.
- Luan, J. “Learner Relationship Management: Knowledge Management and Data Mining as Applied in Higher Education.” Paper presented at the Annual Conference of the Community College League of California, Riverside, CA, 2001.
- Mena, J. “Data-mining FAQs.” *DM Review*, January 1998.

- Rouche, J. E., Rouche, S. S. *High Stakes, High Performance: Making Remediation Works*. Community College Press, 1999.
- Rubinking, N. "Hidden Messages." *PC Magazine*, May 22, 2001.
- Silipo, R. "Neural Networks." In Berthold, M. and Hand, D. (eds.), *Intelligent Data Analysis*. Milan: Springer, 1999.
- SPSS. *SPSS Clementine 6.0 User's Guide*. Chicago: SPSS, 2001.
- Statsoft. (2001). <http://www.statsoft.com/textbook/glosfra.html>
- Therling, K. "An Overview of Data Mining at Dun and Bradstreet." DIG White Paper, 1995.
- van den Eijkel, G. C. "Rule Induction." In Berthold, M. and Hand, D. (eds.), *Intelligent Data Analysis*. Milan: Springer, 1999.
- Watkins, D. "Neural Network Master Class." Presentation at CLUG 2000.
- Westphal, C. and Blaxton, T. *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*. New York: Wiley Computer Publishing, 1998.
- Jing Luan, Ph.D., is Chief Planning, Research & Knowledge Systems Officer at Cabrillo College in Aptos, California.