Data Mining Functions in Advanced Education

Nitin Trivedi
Research Scholar

Sachin Ahuja
Research Scholar

Jugal Kishor Gupta
Assistant Professor
Vidya College of Engineering
Meerut

Vaibhav Mittal
Assistant Professor
Vidya College of Engineering
Meerut

Ankit Jain
Assistant Professor
Vidya College of Engineering
Meerut

ABSTRACT
This paper is based out on the applications of data mining in advanced education. We have given three case studies for predicting the results on the basis of data mining functions. This paper will help and motivate advanced education institutions to look for improved way out.

General Terms
Data Mining, Advanced Education, Oaths.

Keywords
GPA, Models, Scholar.

1. INTRODUCTION
One of the major challenges today is predicting the paths of scholar and alumni. Institutions would want to make out, for example, which scholars will register in particular course programs, and which scholars will require support in order to graduate. Are some scholars more likely to transfer than others? What groups of alumni are most likely to offer oaths? In accumulation to this test, customary issues such as enrollment management and time-to-degree continue to motivate advance education institutions to look for improved way out.

One way to efficiently address these scholar and alumni tests is through the study and presentation of data, or data mining. Data mining makes possible organizations to use their present reporting capabilities to reveal and recognize unseen patterns in huge databases. These patterns are then built into data mining models and used to predict individual behavior with high accuracy. As a result of this insight, institutions are able to allocate resources and staff more effectively. Data mining may, for example, give an institution the information essential to get action before a scholar drops out, or to efficiently allocate resources with an accurate estimate of how many scholars will take a particular course [1].

This white paper deals with the abilities of data mining and its functions in advanced education. Three case studies demonstrate how data mining saves resources while maximizing effectiveness, and increases productivity without increasing cost. The paper starts with an overview of data mining abilities.

2. DATA MINING OVERVIEW
Data mining uses a combination of a clear knowledge base, sophisticated logical skills, and domain information to find out unseen trends and patterns. These trends and patterns form the basis of analytical models that allow analysts to make new annotations from accessible data.

Gartner Inc.'s definition of data mining is the 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." Data mining should be performed on very large or raw datasets using either supervised or unsupervised data mining algorithms [5]. Note that data mining cannot occur without direct interaction with unitary data.

2.1 Data Mining Models and Algorithms
Models house the steps, modules, and resources of the data mining process. Some data mining models include the entire process for a particular purpose, be it to cluster or predict. A model is, however, different from an algorithm. An algorithm is a specific, mathematically driven data mining function, such as a neural network, classification and regression tree (C&RT), or K-means [1].

Beyond those mentioned in this paper, there are the genetic, market basket analysis, Kohonen network, link analysis, time/sequence, and text mining algorithms, to name just a few. Most of the traditional statistics, such as logistic regression and principal component analysis, are also treated as data mining tools. In addition, university laboratories often produce new algorithms for specific business or scientific research purposes.

2.2 Data Mining in Advanced Education
Data mining is a controlling tool for academic interference. Through data mining, a university could, for example, predict with 80 percent accuracy which scholars will or will not graduate. The institution of advance education could use this information to focus academic support on those scholars most at danger. In order to know how and why data mining works, it's important to know a few essential concepts. First, data mining relies on four vital techniques: Classification, categorization, estimation, and visualization [2]. Classification recognizes associations and clusters, and divides subjects under learning. Categorization uses rule initiation algorithms to handle categorical results, such as "persist" or "dropout," and "transfer" or "stay." Estimation includes predictive functions or chances and deals with continuous result variables, such as GPA and salary level. Visualization uses

interactive graphs to demonstrate mathematically induced rules and scores, and is far classier than pie or bar charts. Visualization is used mainly to depict three-dimensional geographic locations of mathematical coordinates.
Advance education institutions can use classification, for example, for a comprehensive analysis of scholar characteristics, or use estimation to predict the likelihood of a variety of outcomes, such as transferability, persistence, retention, and course success [1].

2.3 Supervised and Unsupervised Modeling

Classification and estimation use either unsupervised or supervised modeling techniques. Unsupervised data mining is used for situations in which particular groupings or patterns are unknown. In scholar course databases, for example, little is known about which courses are usually taken as a group, or which course types are linked with which scholar types. Unsupervised data mining is often used first to study patterns and search for previously unseen patterns, in order to know, organize, symbolize, and code the objects of study before applying theories.
Supervised data mining, however, is used with records that have a known outcome. A graduation database, for example, contains records of scholars who completed their studies, as well as of those who dropped out. Supervised data mining is used to study the academic behavior of both groups, with the intention of linking behavior patterns to academic histories and other recorded information.
This so-called "machine learning" uses artificial intelligence to install rules and outline patterns that analysts can apply to fresh data. Once a model does well, the analyst can feed in another scholar group, such as new scholars, and the model applies the erudite information to the new group to predict the possibility of graduation. All of these steps are programmed to produce accurate estimations rapidly, saving time and resources compared to conformist behavior prediction methods.

3. Data Mining Applications in Advanced Education

Data mining is previously essential to the private sector. Many of the data mining practices used in the corporate world, however, are moveable to advance education [3]. **Figure 1**, below, shows the advance education equivalents of vital business questions answered by data mining.
The following three case studies illustrates key applications of data mining in advances education

Case Study One: Creating meaningful learning outcome typologies.

Challenge

"What do institutions know about their scholars?" If the answer is a recital of enrollment percentages or other basic counts, institutions do not know their scholars as well as they could. This case study demonstrates how suburban community colleges can establish learning outcome typologies for scholars using unsupervised data mining.
A typical suburban community college with an enrollment of 15,000 traditionally identifies its scholars as "transfer oriented," "vocational education directed," or "basic skill upgraders." These identifications, however, are based on scholars' initial declarations of educational goals at enrollment. While these are inclusive classifications, they don't help to illustrate the differences between each scholar type.

Private Sector Questions	**Advanced Education Equivalents**
Who are my most profitable customers?	Which scholars are talking the most credit hours?
Who are my repeat Web site visitors?	Which scholars are most likely to return for more classes?
Who are my loyal customers?	Who are the "persisters" at my university/college?
Who is likely to increase his/her purchases?	Which alumni are likely to make larger donations?
Which customers are likely to defect to competitors?	What types of courses will attract more scholars?

Fig 1: Data mining questions in the private sector and their advanced education equivalents

Solution
To establish appropriate typologies for the 15,000 scholars, researchers used both TwoStep and K-means, two powerful clustering algorithms. They first applied the algorithms to the general groupings identified above, with mixed results. The boundaries among clusters were unclear and dispersed, and even after repeated testing on holdout datasets, as well as the removal of suspected outliers (cases that do not appear to belong to any group), the results did not improve significantly.
It's possible that the scholars' initial declaration of goals did not dictate their academic behavior.
The researchers then used a replacement method that looked at educational outcomes in combination with lengths of study. Defining educational outcomes is easier said than done. Enough time must pass to conclude that a scholar has reached a certain milestone. Dropping out is also an outcome by itself. Further work was conducted to determine length of study, which required decisions on how to deal with "stopouts," scholars who left school and later returned. All of these situations test the data miner's domain knowledge. There are no absolutely right or wrong typologies. In essence, a typology is a good one if it serves a particular research objective.
After either removing the outliers or adding them to a particular cluster, the TwoStep algorithm produced the following clusters: "Transfers," "vocational scholars," "basic skills scholars," "scholars with mixed outcomes," and "dropouts." K-means validated these clusters. Introducing the length-of-study element gave new dimensions to each cluster. Some transfer scholars completed their studies quickly; some vocational scholars took longer; and other scholars appeared to simply take one or two courses at a time.
Results
Data mining, combined with scholar demographics and other information, enabled the college to improve its understanding of its scholar types. Certain older scholars, for example, tended to take their time, while younger scholars with more privileged socioeconomic backgrounds often took high credit courses and graduated quickly. One of the most interesting

steps in classification is naming the typologies. The college used the term "transfer speeders," for example, to describe scholars who quickly accumulated units, while those who took classes for a considerable length of time were "college historians." Other scholar clusters were "fence sitters," "skill upgraders," etc.

Typologies are important because they go beyond conventional scholar profiling to identify homogenous groups of scholars, thus increasing the accuracy of predictive modeling algorithms. Even if a data mining project ends with the discovery of appropriate typologies, the newly discovered patterns and relationships help educators and administrators better meet the needs of varied scholar groups.

Case Study Two: Academic planning and interventions-transfer prediction

Challenge

This case study showcases a solution to a vexing advance education problem: How to accurately predict academic outcomes in order to facilitate timely academic intervention. When institutions use data mining to predict which scholars are most at risk, institutions can prevent a scholar from failing before the scholar is even aware that he or she is at risk.

More than half of community college scholars identify transferring to four-year universities as their goal. Due to academic difficulties, however, many either take a long time to transfer or never transfer at all. While it has traditionally been difficult to discover which scholars transfer, the National Scholar Clearing House now allows community colleges and universities to match their data. This means that data miners and decision makers can link the academic behavior of community college scholars to their transfer outcomes.

Solution

Building an effective data mining model with this data involves a combination of typologies and domain knowledge. Transfer education domain knowledge emphasizes that the most effective means of increasing scholar transfers is to identify transfer directed scholars as early as possible. Grooming those who are most likely to transfer is far more meaningful than counting the number of scholars who have accumulated enough units to transfer.

Using the transfer outcome data, analysts built a dataset containing scholars who fell under the general transfer clusters of "speeders" and "laggards." The dataset was split into a test dataset and a validation dataset, using a proprietary randomization method. The outcome variable was transfer. Other variables, such as demographics, courses taken, units accumulated, and financial aid, were predictors to be analyzed without stepwise testing for significance. Data mining is very tolerant of variable interactions and non-linear relationships in data. Supervised data mining was the obvious and appropriate method; therefore, the analysts ran neural network and rule induction algorithms simultaneously in order to contrast and compare the prediction accuracy.

Results

Data mining enabled the college to accurately identify good transfer candidates. After extensive machine learning, the neural network algorithm, Neural Net, had a prediction accuracy of 72 percent, and the rule induction algorithms, C5.0 and C&RT, had a prediction accuracy of 80 percent. The models then ran against the test dataset and produced similar results, indicating their grasp of the patterns within the data.

Case Study Three: Predicting Alumni Pledges

Challenge

For a typical urban university of 25,000, the alumni population can be as much as ten times its enrollment. Most universities send mailings to alumni on a regular basis, even when alumni fail to respond. These mailings typically cost more than $100,000 a year. This case study shows how data mining helps universities focus on the alumni most likely to make pledges.

Solution

It's often difficult to determine whether mailings directly affect the volume and value of alumni pledges. Given the same type of mailing, one alumnus may contribute regularly while another may not. Adding to the confusion is the presence of outliers, such as alumni who unexpectedly contribute large sums. How do institutions identify and cultivate relationships with "outlier" alumni?

In **Figure 2**, on the next page, the chart shows the benefit of using data mining to determine alumni mail recipients versus simply mailing to all alumni. The curved line is the optimal return rate (alumni contributions) as predicted by data mining.

Gain Chart

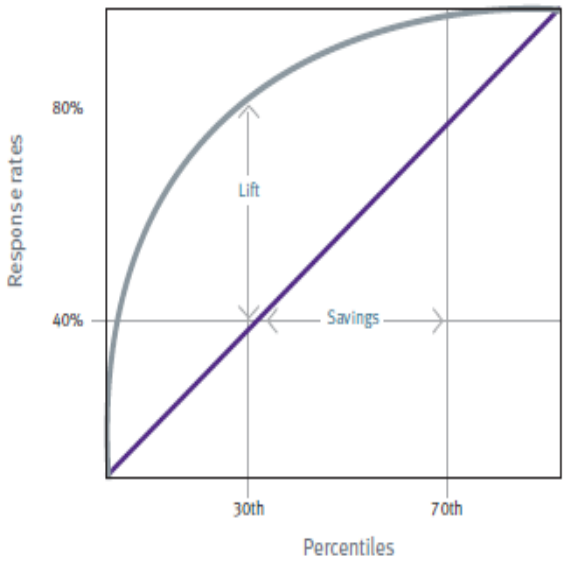

Fig 2: Gain Chart for hypothetical data mining of alumni pledges

Results

Using data mining, the college discovered a way to make its mailing more effective and increase alumni pledges, while reducing the mailing costs. This is the best described using concept called "lift." If 30 percent of alumni respond to a pledge request, the college should concentrate on those 30 percent. If data mining can quickly identify potential donors by a ratio of two to four (correctly predicting two out of four who will donate), then the university can achieve results by mailing only to the 50 percent of the alumni population, thus saving considerable time and money.

4. CONCLUSION

Data mining is a powerful analytical tool that enables educational institutions to better allocate resources and staff, proactively manage scholar outcomes, and improve the effectiveness of alumni development. With the ability to uncover hidden patterns in large databases, community colleges and universities can build models that predict—with a high degree of accuracy—the behavior of population clusters. By acting on these predictive models, educational institutions can effectively address issues ranging from transfers and retention, to marketing and alumni relations.

5. REFERENCES

[1] Jing Luan, Knowledge discovery laboratories/SPSS

[2] Berry, C. M. & Sackett, P. R. (2009). Individual differences in course choice result in underestimation of college admissions system validity.

[3] The g factor: the science of mental ability, Arthur R. Jensen; Praeger Publishers, 1998.

[4] Noftle, Erik E.; Robins, Richard W., Personality Predictors of Academic Outcomes: Big Five Correlates of GPA and SAT Scores, Journal of Personality and Social Psychology. Vol 93(1), July 2007, 116-130.

[5] Agrawal, R., Imielinski, T.; Swami A. (1993), "Mining Associations between Sets of Items in Massive Databases", *Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data*, Washington D.C., May 1993.

[6] Aeberhard, S., Coomans D., and de Vel, O. (1992) "Comparison of Classifiers in High Dimensional Settings", Tech. Rep. no. 92-02, (1992), Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland.

[7] Agresti, A. (2002), *Categorical data analysis.* 2nd edition, New York: Wiley, 2002.

[8] Albertelli, G., Minaei-Bigdoli, B., Punch, W.F., Kortemeyer, G., and Kashy, E., (2002) "Concept Feedback In Computer-Assisted Assignments", *Proceedings of the (IEEE/ASEE) Frontiers in Education conference*, 2002.

[9] Albertelli, G. Sakharuk, A., Kortemeyer, G., Kashy, E., (2003) "Individualized examinations for large on-campus courses", *Proceedings of the (IEEE/ASEE) Frontiers in Education Conference 2003, vol 33.*

[10] Azevedo, R, Bernard, R, M, "*A Meta-analysis of the Effects of Feedback in Computer-based Instruction*", J. Educational Computing Research 13, 111-127. (1995).

[11] Baker, J. E. (1987). Reducing bias and inefficiency in the selection algorithm, *Proceeding ICGA 2*, pp. 14-21, Lawrence Erlbuam Associates, Publishers, 1987.

[12] Cestnik, B. Konenenko, I. Bratko, I. (1987). *ASSISTANT 86: A Knowledge Elicitation Tool for Sophisticated Users*, in Bratko, I. and Navrac, N. (eds), Progress in Machine Learning, Sigma Press, UK.

[13] Dong, G., Li, J., (1998) "Interestingness of Discovered Association Rules in terms of Neighborhood-Based Unexpectedness", *Proceedings of Pacific Asia Conference on Knowledge Discovery in Databases (PAKDD)*, pp. 72-86 Melbourne, 1998.

[14] Ester, M., Kriegel, H.-P., Xu. X. (1995) "A Database Interface for Clustering in Large Spatial Databases", *Proceedings of the Knowledge Discovery and Data Mining Conference*, pages 94-99, Montreal, Canada, 1995.