Journal of Modern Applied Statistical Methods

Volume 4 | Issue 1 Article 23

5-1-2005

An Exploration of Using Data Mining in Educational Research

Yonghong Jade Xu
The University of Memphis

DOI: 10.22237/jmasm/1114906980

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 251-274

1538 – 9472/05/$95.00

An Exploration of Using Data Mining in Educational Research

Yonghong Jade Xu
Department of Counseling, Educational Psychology, and Research
The University of Memphis

Technology advances popularized large databases in education. Traditional statistics have limitations for analyzing large quantities of data. This article discusses data mining by analyzing a data set with three models: multiple regression, data mining, and a combination of the two. It is concluded that data mining is applicable in educational research.

Key words: Data mining, large scale data analysis, quantitative educational research, Bayesian network, prediction

Introduction

In the last decade, with the availability of high-speed computers and low-cost computer memory (RAM), electronic data acquisition and database technology have allowed data collection methods that are substantially different from the traditional approach (Wegman, 1995). As a result, large data sets and databases are becoming increasingly popular in every aspect of human endeavor including educational research. Different from the small, low-dimensional homogeneous data sets collected in traditional research activities, computer-based data collection results in data sets of large volume and high dimensionality (Hand, Mannila, & Smyth, 2001; Wegman, 1995).

Yonghong Jade Xu is an Assistant Professor at the Department of Counseling, Educational Psychology, and Research, College of Education, the University of Memphis. The author wishes to thank Professor Darrell L. Sabers, Head of the Department of Educational Psychology at the University of Arizona, and Dr. Patricia B. Jones, Principal Research Specialist at the Center for Computing Information Technology at the University of Arizona, for their invaluable input pertaining to this study. Correspondence concerning this article should be addressed to Yonghong Jade Xu, Email: yxu@memphis.edu

Many statisticians (e.g., Fayyad, 1997; Hand et al., 2001; Wegman, 1995) noticed some drawbacks of traditional statistical techniques when trying to extract valid and useful information from a large volume of data, especially those of a large number of variables. As Wegman (1995) argued, applying traditional statistical methods to massive data sets is most likely to fail because "homogeneity is almost surely gone; any parametric model will almost surely be rejected by any hypothesis testing procedure; fashionable techniques such as bootstrapping are computationally too complex to be seriously considered for many of these data sets; random subsampling and dimensional reduction techniques are very likely to hide the very substructure that may be pertinent to the correct analysis of the data" (p. 292). Moreover, because most of the large data sets are collected from convenient or opportunistic samples, selection bias puts in question any inferences from sample data to target population (Hand, 1999; Hand et al., 2001).

The statistical challenge has stimulated research aiming at methods that can effectively examine large data sets to extract valid information (e.g., Daszykowski, Walczak, & Massart, 2002). New analytical techniques have been proposed and explored. Among them, some statisticians (e.g., Elder & Pregibon, 1996; Friedman, 1997; Hand, 1998, 1999, 2001; Wegman, 1995) paid attention to a new data analysis tool called data mining and knowledge discovery in database. Data mining is a process

of nontrivial extraction of implicit, previously unknown, and potentially useful information from a large volume of data (Frawley & Piatetsky-Shapiro, 1991).

Although data mining has been used in business and scientific research for over a decade, a thorough literature review has found no educational study that used data mining as the method of analysis. To explore the usefulness of data mining in quantitative research, the current study provides a demonstration of the analysis of a large education-related data set with several different approaches, including traditional statistical methods, data mining, and a combination of these two. With different analysis techniques laid side-by-side working on the same data set, the virtue of the illustrated methods, models, outputs, conclusions, and unique characteristics is ready for assessment.

Research Background

According to its advocates, data mining has prevailed as an analysis tool for large data sets because it can efficiently and intelligently probe through an immense amount of material to discover valuable information and make meaningful predictions that are especially important for decision-making under uncertain conditions.

Data mining uses many statistical techniques, including regression, cluster analysis, multidimensional analysis, stochastic models, time series analysis, nonlinear estimation techniques, just to name a few (Michalski, Bratko, & Kubat, 1998).

However, data mining is not a simple rework of statistics; it implements statistical techniques through an automated machine learning system and acquires high-level concepts and/or problem-solving strategies through examples (input data) in a way analogous to human knowledge induction to attack problems that lack algorithmic solutions or have only ill-defined or informally stated solutions (Michalski et al., 1998).

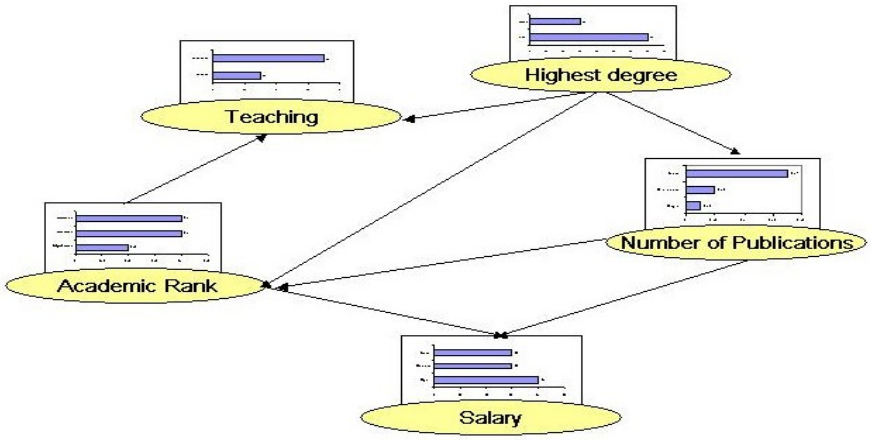
Data mining generates descriptions of rules as output using algorithms such as Bayesian probability, artificial neural networks, decision trees, and generic algorithms that do not assume any parametric form of the appropriate model. Automated analysis processes that reduce or eliminate the need for human interventions become critical when the volume of data goes beyond human ability of visualization and comprehension.

Due to its applied importance, data mining as an academic discipline continues to grow with input from statistics, machine learning, and database management (Fayyad, 1997; Zhou, 2003). One popular algorithm in recent research is the Bayesian Belief Network (BBN), which started from a set of probability rules discovered by Thomas Bayes in the 18th century. The tree-like network based upon Bayesian probability can be used as a prediction model (Friedman et al., 1997). To build such a model, various events (variables) have to be defined, along with the dependencies among them and the conditional probabilities (CP) involved in those dependencies.

Once the variables are ready and the topology is defined, they become the information used to calculate the probabilities of various possible paths being the actual path leading to an event or a particular value of a variable. Through an extensive iteration, a full joint probability distribution is to be constructed over the product state space (defined as the complete combinations of distinct values of all variables) of the model variables. The computational task is enormous because elicitation at a later stage in the sequence results in back-tracking and changing the information that has been elicited at an earlier point (Yu & Johnson, 2002). With the iterative feedback and calculation, a BBN is able to update the prediction probability, the so-called belief values, using probabilistic inference.

BBN combines a sound mathematical basis with the advantages of an intuitive visual representation. The final model of a BBN is expressed as a special type of diagram together with an associated set of probability tables (Heckerman, 1997), as shown in the example in Figure 1. The three major classes of elements are a set of uncertain variables presented as nodes, a

Figure 1. An example of a BBN model. This graph illustrates the three major classes of elements of a Bayesian network; all variables, edges, and CP tables are for demonstration only and do not reflect the data and results of the current study in any way.



set of directed edges (arcs) between variables showing the causal/relevance relationships between variables, and also, a CP table P(A| B_1, B_2,…, B_n) attached to each variable A with parents B_1, B_2, …, B_n. The CPs describe the strength of the beliefs given that the prior probabilities are true.

Because in learning a previously unknown BBN, the calculation of the probability of any branch requires all branches of the network to be calculated (Niedermayer, 1998), the practical difficulty of performing the propagation, even with the availability of high-speed computers, delayed the availability of software tools that could interpret the BBN and perform the complex computation until recently. Although the resulting ability to describe the network can be performed in linear time, given a relatively large number of variables and their product state space, the process of network discovery remains computationally impossible if an exhaustive search in the entire model space is required for finding the network of best prediction accuracy.

As a compromise, some algorithms and utility functions are adopted to direct random selection of variable subsets in the BBN modeling process and to guide the search for the optimal subset with an evaluation function tracking the prediction accuracy (measured by the classification error rate) of every attempted model (Friedman et al., 1997). That is, a stochastic variable subset selection is embedded into the BBN algorithms. The variable selection function conducts a search for the optimal subset using the BBN itself as a part of the evaluation function, the same algorithm that will be used to induce the final BBN prediction model.

Some special features of the BBN are considered beneficial to analyzing large data sets. For instance, to define a finite product state space for calculating the CPs and learning the network, all continuous variables have to be discretized into a number of intervals (bins). With such discretization, variable relationships are measured as associations that do not assume linearity and normality, which minimizes the negative impacts of outliers and other types of irregularities inherent in secondary data sources. Variable discretization also makes a BBN flexible in handling different types of variables and eliminates the sample size as a factor influencing the amount of computation.

With large databases available for research and policy making in education, this study is designed to assess whether the data mining approach can provide educational researchers with extra means and benefits in analyzing large-scale data sets.

Methodology

To examine the usefulness of data mining in educational research, the current study demonstrated the analysis of a large post-secondary faculty data set with three different approaches, including data mining, traditional statistical methods, and a combination of these two. Because data mining shares a few common concerns with traditional statistics, such as estimation of uncertainty, construction of models in defined problem scope, prediction, and so on (Glymour, Madigan, Pregibon, & Smyth, 1997), in order to narrow down the research problem, prediction functions were chosen as a focus of this article to see whether data mining could offer any unique outlook when processing large data sets.

To be specific, all three models were set to search for factors that were most efficient in predicting post-secondary faculty salary. On the statistical side, multiple linear regression was used because it is an established dynamic procedure of prediction; for data mining, prediction was performed with a BBN. Although the major concern of faculty compensation studies is the evaluation of variable importance in salary determination rather than prediction, the purpose of this study was to illustrate a new data analysis technique, rather than to advance the knowledge in the area of faculty compensation. Unless specified otherwise, α = .01 was used in all significance tests.

Data Set

In order to compare different data analysis approaches, the post-secondary faculty data set collected using the National Survey of Postsecondary Faculty 1999 (NSOPF:99) was chosen as a laboratory setting for demonstrating the statistical and data mining methods.

The NSOPF:99 was a survey conducted by the National Center for Education Statistics (NCES) in 1999. The initial sample included 960 degree granting postsecondary institutions and 27,044 full and part-time faculty employed at these institutions. Both the sample of institutions and the sample of faculty were stratified and systematic samples. Approximately 18,000 faculty and instructional staff questionnaires were completed at a weighted response rate of 83 percent. The response rate for the institution survey was 93 percent.

In this study, only faculty data were used which included 18,043 records and 439 original and derived measures. Information was available on faculty demographic backgrounds, workloads, responsibilities, salaries, benefits, and more. The data set was considered appropriate because it is an education-related survey data set, neither too large for traditional analysis approaches nor too small for data mining techniques.

To focus on the salary prediction of regular faculty in postsecondary institutions, only respondents who reported fulltime faculty status were included. Faculty assigned by religious order was excluded as well as those having affiliated or adjunct titles. Also, some respondents were removed from the data set to eliminate invalid salary measures. As a result, the total number of records available for analysis was 9,963. Two-thirds of the records were randomly selected as training data and used to build the prediction models; the remaining one-third were saved as testing data for purpose of cross-validation.

Variables in the data set were also manually screened so that only the most salient measures of professional characteristics were kept to quantify factors considered relevant in determining salary level according to the general guidelines of salary schema in postsecondary institutions and to the compensation literature in higher education. At the end, only 91 (including salary) were kept in the study out of the entire set of variables.

Among them, a few variables were derived from the original answers to the questionnaire in order to avoid redundant or overly specific information. However, multiple measures were kept on teaching, publication, and some other constructs because they quantified different aspects of the underlying constructs; the redundant information among them also offered a chance of testing the differentiation power of the variable selection procedures. Table 1 provides a list of all the 91 variables and their definitions.

Table 1. Name, Definition, and Measurement Scale of the 91 Variables from NSOPF:99.

Variable name	Variable definition	Scale
Q25	Years teaching in higher education institution	Interval
Q26	Positions outside higher education during career	Interval
Q29A1	Career creative works, juried media	Interval
Q29A2	Career creative works, non-juried media	Interval
Q29A3	Career reviews of books, creative works	Interval
Q29A4	Career books, textbooks, reports	Interval
Q29A5	Career exhibitions, performances	Interval
Q29B1	Recent sole creative works, juried media	Interval
Q29B2	Recent sole creative works, non-juried media	Interval
Q29B3	Recent sole reviews of books, works	Interval
Q29B4	Recent sole books, textbooks, reports	Interval
Q29B5	Recent sole presentations, performances	Interval
Q29C1	Recent joint creative works, juried media	Interval
Q29C2	Recent joint creative works, non-juried media	Interval
Q29C3	Recent joint reviews of books, creative works	Interval
Q29C4	Recent joint books, reports	Interval
Q29C5	Recent joint presentations, performances	Interval
Q2REC	Teaching credit or noncredit courses	Ordinal
Q30B	Hours/week unpaid activities at the institution	Interval
Q30C	Hours/week paid activities not at the institution	Interval
Q30D	Hours/week unpaid activities not at the institution	Interval

Table 1 Continued.

Variable name	Variable definition	Scale
Q31A1	Time actually spent teaching undergrads (percentage)	Ratio
Q31A2	Time actually spent teaching graduates (percentage)	Ratio
Q31A3	Time actually spent at research (percentage)	Ratio
Q31A4	Time actually spent on professional growth (percentage)	Ratio
Q31A5	Time actually spent at administration (percentage)	Ratio
Q31A6	Time actually spent on service activity (percentage)	Ratio
Q31A7	Time actually spent on consulting (percentage)	Ratio
Q32A1	Number of undergraduate committees served on	Interval
Q32A2	Number of graduate committees served on	Interval
Q32B1	Number of undergraduate committees chaired	Interval
Q32B2	Number of graduate committees chaired	Interval
Q33	Total classes taught	Interval
Q40	Total credit classes taught	Interval
Q50	Total contact hours/week with students	Interval
Q51	Total office hours/week	Interval
Q52	Any creative work/writing/research	Categorical
Q54_55RE	PI / Co-PI on grants or contracts	Ordinal
Q58	Total number of grants or contracts	Interval
Q59A	Total funds from all sources	Ratio
Q61SREC	Work support availability	Ordinal
Q64	Union status	Categorical

Table 1 Continued.

Variable name	Variable definition	Scale
Q76G	Consulting/freelance income	Ratio
Q7REC	Years on current job	Interval
Q80	Number of dependents	Interval
Q81	Gender	Categorical
Q85	Disability	Categorical
Q87	Marital status	Categorical
Q90	Citizenship status	Categorical
Q9REC	Years on achieved rank	Interval
X01_3	Principal activity	Categorical
X01_60	Overall quality of research index	Ordinal
X01_66	Job satisfaction: other aspects of job	Ordinal
X01_82	Age	Interval
X01_8REC	Academic rank	Ordinal
X01_91RE	Highest educational level of parents	Ordinal
DISCIPLINE	Principal field of teaching/researching	Categorical
X02_49	Individual instruction w/grad &1st professional students	Interval
X03_49	Number of students receiving individual instructions	Interval
X04_0	Carnegie classification of institution	Categorical
X04_41	Total classroom credit hours	Interval
X04_84	Ethnicity in single category	Categorical
X08_0D	Doctoral, 4-year, or 2-year institution	Ordinal

Table 1 Continued.

Variable name	Variable definition	Scale
X08_0P	Private or public institution	Categorical
X09_0RE	Degree of urbanization of location city	Ordinal
X09_76	Total income not from the institution	Ratio
X10_0	Ratio: FTE enrollment / FTE faculty	Ratio
X15_16	Years since highest degree	Interval
X21_0	Institution size: FTE graduate enrollment	Interval
X25_0	Institution size: Total FTE enrollment	Interval
X37_0	Bureau of Economic Analysis (BEA) regional codes	Categorical
X46_41	Undergraduate classroom credit hours	Interval
X47_41	Graduate and First professional classroom credit hours	Interval
SALARY	Basic academic year salary	Ratio

Note. All data were based on respondent' reported status during the 1998-99 academic year.

Analysis

Three different prediction models were constructed and compared through the analysis of NSOPF:99; each of them had a variable reduction procedure and a prediction model based on the selected measures. The first model, Model I, was a multiple regression model with variables selected through statistical data reduction techniques; Model II was a data mining BBN model with an embedded variable selection procedure. A combination model, Model III, was also a multiple regression model, but built on variables selected by the data mining BBN approach.

Model I. The first model started with variable reduction procedures that reduced the 90 NSOPF:99 variables (salary measure excluded) to a smaller group that can be efficiently manipulated by a multiple regression procedure, and resulted in an optimal regression model based on the selected variables. According to the compensation theory and characteristics of the current data set, basic salary of the academic year as the dependent variable was log-transformed to improve its linear relationship with candidate independent variables.

The variable reduction for Model I was completed in two phases. In the first phase, the dimensional structure of the variable space was examined with Exploratory Factor Analysis (EFA) and K-Means Cluster (KMC) analysis; based on the outcomes of the two techniques, variables were classified into a number of major dimensions. Because EFA measures variable relationships by linear correlation and KMC by Euclidian distance, only 82 variables on

dichotomous, ordinal, interval, or ratio scales were included. Two different techniques were used to scrutinize the underlying variable structure such that any potential bias associated with each of the individual approaches could be reduced.

In EFA, different factor extraction methods were tried and followed by both orthogonal and oblique rotations of the set of extracted factors. The variable grouping was determined based on the matrices of factor loadings: variables that had a minimum loading of .35 on the same factor were considered as belonging to the same group. In the KMC analysis, the number of output clusters usually needs to be specified. When the exact number of variable clusters is unknown, the results of other procedures (e.g., EFA) can provide helpful information for estimating a range of possible number of clusters. Then the KMC can be run several times, each time with a different number of clusters specified within the range. The multiple runs of the KMC can also help to reduce the chance of getting a local optimal solution. Because variables were separated into mutually exclusive clusters, the interpretation of cluster identity was based on variables that had short distance from the cluster seed (the centroid).

The results of the KMC analysis were compared with that of the EFA for similarities and differences. A final dimensional structure of the variable space was determined based on the consensus of the EFA and KMC outputs; each of the variable dimensions was labeled with a meaningful interpretation.

During the second phase, one variable was selected from each dimension. Because of the different clustering methods used, variables in the same dimension might not share linear relationships. Taking into consideration that the final model of the analysis was of linear prediction, a method of extracting variables that account for more salary variance was desirable. Thus, for each cluster, the log-transformed salary was regressed on the variables within that cluster, and only one variable was chosen that associated with the greatest partial R^2 change.

Variables that did not show any strong relationships with any of the major groups, along with multilevel nominal variables that could not be classified, were carried directly into the second stage of multiple regression modeling as candidate predictors and tested for their significance. Nominal variables were recoded into binary variables and possible interactions among the predictor variables were checked and included in the model if significant. Both forced entry and stepwise selection were used to search for the optimal model structure; if any of the variables was significant in one variable selection method, but nonsignificant in the other, a separate test on the variable was conducted in order to decide whether to include the variable in the final regression model. Finally, the proposed model was cross-checked with All Possible Subsets regression techniques including Max R and C_p evaluations to make sure the model was a good fit in terms of the model R^2, adjusted R^2, and the C_p value.

Model II. The second prediction model was a BBN-based data mining model. To build the BBN model, all 91 original variables were input into a piece of software called the Belief Network Powersoft ; variables on interval and ratio scales were binned into category-like intervals because the network-learning algorithms require discrete values for a clear definition of a finite product state space of the input variables. Rather than logarithmical transformation, salary was binned into 24 intervals for the following reasons: first, log-transformation was not necessary because BBN is a robust nonmetric algorithm independent of any monotonic variable transformation. And second, a finite number of output classes is required in a Bayesian network construction. During the modeling process, variable selection was performed internally to find the subset with the best prediction accuracy.

The BBN model learning was an automated process after reading in the input data. According to Chen and Greiner (1999), the authors of the software, two major tasks in the process are learning the graphical structure (variable relationships) and learning the parameters (CP tables). Learning the structure is the most computationally intensive task. The BBN software used in this study takes the network structure as a group of CP relationships (measured by statistical functions such as χ^2 statistic or mutual information test) connecting

the variables, and proceeds with the model construction by identifying the CPs that are stronger than a specified threshold value.

The output of the BBN model was a network in which the nodes (variables) were connected by arcs (CP relationships between variables) and a table of CP entries (probability) for each arc. Only the subset of variables that was evaluated as having the best prediction accuracy stayed in the network. The prediction accuracy was measured by the percentage of correct classifications of all observations in the data set.

Model III. Finally, a combination model was created that synchronized data mining and statistical techniques: the variables selected by the data mining BBN model were put into a multiple regression procedure for an optimal prediction model. The final BBN model contained a subset of variables that was expected to have the best prediction accuracy. Once the BBN model was available, the variables in that model were put through a multiple regression procedure for another prediction model. If it results in a better model, it would be evident that BBN could be used together with traditional statistical techniques when appropriate. As in Model I, categorical variables were recoded and salary as the dependent variable was log-transformed. Multiple variable selection techniques were used including forced entry and stepwise selection.

Model Comparison

The algorithms, input variables, final models, outputs, and interpretations of the three prediction models were presented. The two multiple regression models were comparable because they shared some common evaluation criteria, including the model standard error of estimate, residuals, R^2, and adjusted R^2. The data mining BBN model offered a different form of output, and is less quantitatively comparable with the regression models because they had little in common.

Software

SAS and SPSS were used for the statistical analyses. The software for learning the BBN model is called Belief Network Powersoft, a shareware developed and provided by Chen and Greiner (1999) on the World Wide Web. The Belief Network Powersoft was the winner of the yearly competition of the Knowledge Discovery and Data mining (KDD) – KDDCup 2001 Data Mining Competition Task One, for having the best prediction accuracy among 114 submissions from all over the world.

Results

Model I

The result of the variable space simplification through EFA and KMC was that 70 of the 82 variables were clustered into 17 groups. Ten of the groups were distinct clusters that did not seem to overlap with each other: academic rank, administrative responsibility, beginning work status, education level, institution parameter, other employment, research, teaching, experience, and work environment index. Another seven groups were 1) teaching: undergraduate committee, 2) teaching: graduate, 3) teaching: individual instruction, 4) publications: books, 5) publications: reviews, 6) publication: performances and presentations, and 7) institutional parameters: miscellaneous. In general, the dimensional structure underlying the large number of variables provided a schema of clustering similar measures and therefore made it possible to simplify the data modeling by means of variable extraction.

Following the final grouping of variables, one variable was extracted from each of the clusters by regressing the log-transformed salary on variables within the same cluster and selecting the variable that contributed the greatest partial R^2 change in the dependent variable. The 17 extracted variables, along with the 20 variables that could not be clustered, are listed in Table 2 as the candidate independent variables for a multiple regression model.

After a thorough model building and evaluation process, a final regression model was selected having 16 predictor variables (47 degrees of freedom due to binary-coded nominal measures) from the pool of 37 candidates. The parameter estimates and model summary information are in Tables 3 and 5. The model R^2 is .5036 and adjusted R^2 .5001.

Table 2. Candidate Independent Variables of Model I.

Variable name	Variable Definition	*df*
	Variables from the clusters	
Q29A1	Career creative works, juried media	1
X15_16	Years since highest degree	1
Q31A1	Time actually spent teaching undergraduates (percentage)	1
Q31A2	Time actually spent at teaching graduates (percentage)	1
X02_49	Individual instruction w/grad &1st professional students	1
Q32B1	Number of undergraduate committees chaired	1
Q31A5	Time actually spent at administration (percentage)	1
Q16A1REC	Highest degree type	1
Q24A5REC	Rank at hire for 1st job in higher education	1
Q29A3	Career reviews of books, creative works	1
Q29A5	Career presentations, performances	1
X08_0D	Doctoral, 4-year, or 2-year institution	1
Q29A4	Career books, textbooks, reports	1
X10_0	Ratio: FTE enrollment / FTE faculty	1
Q76G	Consulting/freelance income	1
X01_66	Job satisfaction: other aspects of job	1
X01_8REC	Academic rank	1

Table 2 Continued.

Variable name	Variable definition	*df*
	Variables from the original set	
DISCIPLINE	Principal field of teaching/research	10
Q12A	Appointments: Acting	1
Q12E	Appointments: Clinical	1
Q12F	Appointments: Research	1
Q19	Current position as primary employment	1
Q26	Positions outside higher education during career	1
Q30B	Hours/week unpaid activities at the institution	1
Q31A4	Time actually spent on professional growth (percentage)	1
Q31A6	Time actually spent on service activity (percentage)	1
Q64	Union status	3
Q80	Number of dependents	1
Q81	Gender	1
Q85	Disability	1
Q87	Marital status	3
Q90	Citizenship status	3
X01_3	Principal activity	1
X01_91RE	Highest educational level of parents	1
X04_0	Carnegie classification of institution	14
X04_84	Ethnicity in single category	3
X37_0	Bureau of Economic Analysis (BEA) region code	8

Table 3. Parameter Estimates of Model I.

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$
Intercept	Intercept	10.0399	0.0485	207.10	<.0001
Q29A1	Career creative works, juried media	0.0019	0.0002	11.87	<.0001
X15_16	Years since highest degree	0.0077	0.0004	17.82	<.0001
Q31A1	Time actually spent teaching undergrads (%)	-0.0011	0.0002	-6.04	<.0001
Q31A5	Time actually spent at administration (%)	0.0017	0.0003	5.95	<.0001
Q16A1REC	Highest degree type	0.0841	0.0050	16.68	<.0001
Q29A3	Career reviews of books, creative works	0.0018	0.0004	4.22	<.0001
Q76G	Consulting/freelance income	0.0000037	0.0000	5.75	<.0001
X01_66	Other aspects of job	0.0519	0.0058	8.89	<.0001
X01_8REC	Academic rank	0.0510	0.0031	16.27	<.0001
Q31A4	Time actually spent on professional growth (%)	-0.0023	0.0006	-3.86	0.0001
Q31A6	Time actually spent on service activity (%)	0.0013	0.0003	3.80	0.0001
Q81	Gender	-0.0667	0.0084	-7.97	<.0001
<u>BEA region codes (Baseline: Far West)</u>					
BEA1	New England	-0.0608	0.0058	8.89	0.0021
BEA2	Mid East	0.0082	0.0031	16.27	0.5788
BEA3	Great Lakes	-0.0545	0.0006	-3.86	0.0001
BEA4	Plains	-0.0868	0.0003	3.80	<.0001

Table 3 Continued.

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$
BEA5	Southeast	-0.0921	0.0084	-7.97	<.0001
BEA6	Southwest	-0.0972	0.0198	-3.07	<.0001
BEA7	Rocky Mountain	-0.1056	0.0148	0.56	<.0001
BEA8	U.S. Service schools	0.1480	0.0142	-3.82	0.2879
Principal field of teaching/research (Baseline: legitimate skip)					
DSCPL1	Agriculture & home economics	-0.0279	0.0306	-0.91	0.3624
DSCPL2	Business	0.1103	0.0228	4.84	<.0001
DSCPL3	Education	-0.0643	0.0216	-2.98	0.0029
DSCPL4	Engineering	0.0695	0.0246	2.82	0.0048
DSCPL5	Fine arts	-0.0449	0.0241	-1.86	0.0627
DSCPL6	Health sciences	0.0933	0.0182	5.12	<.0001
DSCPL7	Humanities	-0.0641	0.0195	-3.29	0.001
DSCPL8	Natural sciences	-0.0276	0.0190	-1.45	0.148
DSCPL9	Social sciences	-0.0249	0.0202	-1.23	0.2173
DSCPL10	All other programs	0.0130	0.0194	0.67	0.502
Carnegie classification (Baseline: Private other Ph.D.)					
STRATA1	Public comprehensive	0.0053	0.0236	0.22	0.8221
STRATA2	Private comprehensive	-0.0377	0.0263	-1.43	0.1525
STRATA3	Public liberal arts	-0.0041	0.0341	-0.12	0.9039
STRATA4	Private liberal arts	-0.0917	0.0260	-3.52	0.0004

Table 3 Continued.

Variable	Label	Parameter estimate	Standard error	t value	$p > \|t\|$
STRATA5	Public medical	0.2630	0.0326	8.07	<.0001
STRATA6	Private Medical	0.2588	0.0444	5.82	<.0001
STRATA7	Private religious	-0.1557	0.0523	-2.98	0.0029
STRATA8	Public 2-year	0.0386	0.0247	1.56	0.1185
STRATA9	Private 2-year	-0.0061	0.0574	-0.11	0.9155
STRATA10	Public other	-0.0207	0.0563	-0.37	0.7127
STRATA11	Private other	-0.0879	0.0428	-2.06	0.0399
STRATA12	Public research	0.0792	0.0228	3.47	0.0005
STRATA13	Private research	0.1428	0.0259	5.51	<.0001
STRATA14	Public other Ph.D.	0.0005	0.0254	0.02	0.984
Primary activity (Baseline: others)					
PRIMACT1	Primary activity: teaching	-0.0541	0.0169	-3.21	0.0013
PRIMACT2	Primary activity: research	-0.0133	0.0199	-0.67	0.5039
PRIMACT3	Primary activity: administration	0.0469	0.0203	2.31	0.0211

Note. The dependent variable was log-transformed SALARY (LOGSAL).

Model II

To make the findings of the data mining BBN model comparable to the result of regression Model I, the second model started without any pre-specified knowledge such as the order of variables in some dependence relationships, forbidden relations, or known causal relations. To evaluate variable relationships and simplify model structure, the data mining software makes it possible for users to provide a threshold value that determines how strong a mutual relationship between two variables is considered meaningful; relationships below this threshold are omitted from subsequent network structure learning (Chen & Greiner, 1999).

In the current analysis, a number of BBN learning processes were completed, each with a different threshold value specified, in order to search for an optimal model structure. Because generalizability to new data sets is an

important property of any prediction models, the model parameters were cross-validated with the testing data set. The results suggested that the model of best prediction power was the one having six variables connected by 10 CP arcs as shown in Figure 2. The prediction accuracy, quantified as the percentage of correct classification of the cases, was 25.66% for training data and 11.57% for testing data.

Model III

The final prediction model produced by the Belief Network Powersoft had six predictor variables. However, one of six, number of years since achieved tenure (Q10AREC), was only connected to another predictor variable (i.e., years since the highest degree), a strong relationship substantiated by their Pearson correlation ($r = .64$). Q10AREC also had a strong correlation with academic rank ($r = .43$), another variable in the model. After a test confirmed that Q10AREC was not a suppressor variable, it was excluded from the combination model. Therefore, Model III started with only five independent variables. Among them, the Carnegie classification of institutions as the only categorical measure was recoded into binary variables. With log-transformed salary as the dependent variable, the process of building Model III was straightforward because all five variables were significant at $p < .0001$ with both forced entry and stepwise variable selections. The model has R^2 of .4214 and adjusted R^2 .4199 (summary information is presented in Tables 4 and 5).

Model Comparison

Model I and Model II are comparable in many ways. First, both models are result of data-driven procedures; second, theoretically, they both selected the predictors from the original pool of 90 variables; and third, they share the same group of major variables even though Model I had a much larger group. With the common ground they share, the differences between the two models provide good insight to the differences between traditional statistics and data mining BBN in make predictions with large-scale data sets.

The differences between Model I and Model III are informative about the effects of statistical and data mining approaches in simplifying the variable space and identifying the critical measures in making accurate prediction, given both models used multiple regression for the final prediction. Models II and III share the same group of predictor variables; their similarities and differences shed light on the model presentations and prediction accuracy of different approaches as well.

Variable Selection and Transformation

Model I started with all 90 variables in the pool, and identified 17 of the 70 variables that could be clustered with EFA and KMC procedures. Along with the ungrouped 20 variables, a total of 37 independent variables were available as initial candidates, and 16 of them stayed in the final model with an R^2 of .5036 (df = 47 and adjusted $R^2 = .5001$). With a clear goal of prediction, the modeling process was exploratory without theoretical considerations from variable reduction through model building. During this process, variable relationships were measured as linear correlations; consequently, the dependent variable was transformed to improve its linear relationships with the independent variables. Also, multilevel categorical measures were recoded into binary variables.

The data mining model, Model II, also started with all 90 variables. An automated random search was performed internally to select a subset of variables that provided the most accurate salary prediction. In contrast to regression models that explicitly or implicitly recode categorical data, data mining models usually keep the categorical variables unchanged, but bin continuous variables into intervals. The information loss associated with variable downgrade in binning is a threat to model accuracy, but it helps to relax model assumptions and as a result BBN requires no linear relationships among variables. The network structure discovery uses some statistical tests (e.g., χ^2 test of statistical independence) to compare how frequently different values of two variables are associated with how likely they happen to be together by random chance in order to build conditional probability statistics among variables (Chen, Greiner, Kelly, Bell, & Liu, 2001).

Figure 2. The BBN model of salary prediction. Some of the directional relationships may be counterintuitive (e.g., Q31A1 → X04_0) as a result of data-driven learning. The CP tables are not included to avoid complexity. The definitions of the seven variables are

a. SALARY: Basic salary of the academic year.
b. Q29A1: Career creative works, juried media
c. Q31A1: Percentage of time actually spent teaching undergrads
d. X15_16: Years since highest degree
e. X01_8REC: Academic rank
f. X04_0: Carnegie classification of institutions
g. Q10AREC: Years since achieved tenure

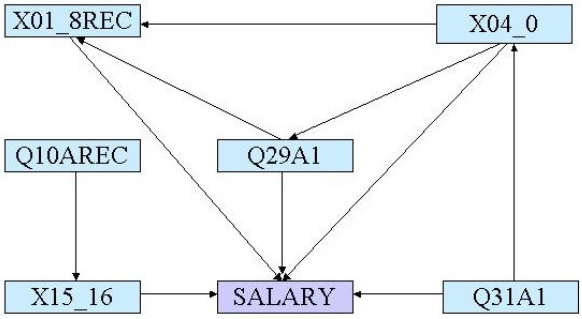


Table 4. Parameter Estimates of Model III.

Variable	Label	Parameter estimate	Standard error	*t* value	*p* > |t|
Intercept	Intercept	10.5410	0.0272	387.28	<.0001
Q29A1	Career creative works, juried media	0.0024	0.0002	15.34	<.0001
Q31A1	Time actually spent teaching undergrads (%)	-0.0030	0.0002	-20.06	<.0001
X01_8REC	Academic rank	0.0664	0.0032	21.01	<.0001
X15_16	Years since highest degree	0.0088	0.0004	19.97	<.0001

Table 4 Continued.

Carnegie classification (Baseline: Private other Ph.D.)					
STRATA1	Public comprehensive	-0.0385	0.0250	-1.54	0.1236
STRATA2	Private comprehensive	-0.0645	0.0281	-2.29	0.0218
STRATA3	Public liberal arts	-0.0315	0.0363	-0.87	0.3853
STRATA4	Private liberal arts	-0.1221	0.0276	-4.42	<.0001
STRATA5	Public medical	0.2933	0.0339	8.66	<.0001
STRATA6	Private Medical	0.2915	0.0471	6.20	<.0001
STRATA7	Private religious	-0.2095	0.0551	-3.80	0.0001
STRATA8	Public 2-year	-0.0403	0.0258	-1.56	0.1179
STRATA9	Private 2-year	-0.0371	0.0611	-0.61	0.544
STRATA10	Public other	-0.0245	0.0594	-0.41	0.6802
STRATA11	Private other	-0.0871	0.0456	-1.91	0.0563
STRATA12	Public research	0.0479	0.0242	1.98	0.0472
STRATA13	Private research	0.1543	0.0276	5.60	<.0001
STRATA14	Public other Ph.D.	-0.0496	0.0268	-1.85	0.0648

Variable	Label	Parameter estimate	Standard error	*t* value	$p > \|t\|$

Note. The dependent variable was log-transformed SALARY (LOGSAL).

Table 5. Summary Information of Multiple Regression Models I and III

Source	*df*	Sum of squares	Mean square	F	Pr > F
Model I: Multiple regression with statistical variable selection					
Model	47	621.4482	13.2223	142.46	<.0001
Error	6599	612.4897	0.0928		
Corrected total	6646	1233.9379			

Table 5 Continued.

Source	*df*	Sum of Squares	Mean square	F	Pr > F
Model III: Multiple regression with variables selected through BBN					
Model	18	520.2949	28.90527	268.4	<.0001
Error	6632	714.3279	0.10769		
Corrected total	6651	1234.6228			

Note:
1. For Model I, R^2 = .5036, adjusted R^2 = .5001, and the standard error of estimate is 0.305.
2. For Model II, R^2 = .4214, adjusted R^2 = .4199 and the standard error of estimate is 0.328

Given the measures of variable associations that do not assume any probabilistic forms of variable distributions, neither linearity nor normality was required in the analysis. Consequently, the non-metric algorithms used to build the BBN model binned the original SALARY measure as the predicted values.

Model Selection

In the multiple regression analysis, every unique combination of the independent variables theoretically makes a candidate prediction model, albeit the modeling techniques produce candidate models that are mostly in a nested structural schema. Model comparison is part of the analysis process; human intervention is necessary to select the final model that usually has a higher R^2 along with simple and stable structure. In contrast, the learning of an optimal BBN model is a result of search in a model space that consists of candidate models of substantially different structures. In the automated model discovery process, numerous candidate models were constructed, evaluated with criteria called score functions, and the one with best prediction accuracy is output as the optimal choice.

Model Presentation

As a result of different approaches to summarizing data and different algorithms of analyzing data, the outputs of the multiple regression and the BBN models are different. The final result of a multiple regression analysis is usually presented as a mathematical equation. For example, Model III can be written as:

Log (Salary) = 10.5410 + 0.0024 × Q29A1 - 0.0030 × Q31A1 + 0.0664 × X01_8REC + 0.0088 × X15_16 - 0.0385 × STRATA1 - 0.0645 × STRATA2 - 0.0315 × STRATA3 - 0.1221 × STRATA4 + 0.2933 × STRATA5 + 0.2915 × STRATA6 - 0.2095 × STRATA7 - 0.0403 × STRATA8 - 0.0371 × STRATA9 - 0.0245 × STRATA10 - 0.0871 × STRATA11 + 0.0479 × STRATA12 + 0.1543 × STRATA13 - 0.0496 × STRATA14 + error. (1)

If a respondent received the highest degree three years ago (X15_16 = 3), had three publications in juried media (Q29A1 = 3), spent 20% of work time teaching undergraduate classes (Q31A1 = 20) as an assistant professor (X01_8REC = 4) in a public research institution (STRATA12 = 1 and all other STRATA variables were 0), the predicted value of this individual's log-transformed salary should be 10.83 according to Equation 1 (about $50,418), with an estimated standard error indicating the level of uncertainty.

The result of the BBN model is presented in a quite different way. For the above case, the BBN model would make a prediction

of salary for such faculty with a salary conditional probability table as shown in Table 6. The predicted salary fell in a range between $48,325 and $50,035 because it has the highest probability (p= 15.9%) in the CP table for this particular combination of variable values. A CP table like this is available for every unique combination of variable values (i.e., an instance in the variable product state space).

Using the conditional mean as a point estimator in most statistical predictions implicitly expresses the prediction uncertainty with a standard error of estimate based on the assumption of normal distribution. In contrast, the BBN model makes predictions based on the distributional mode of the posterior probability of the predicted variable. The prediction based on the mode of a probabilistic distribution is a robust feature of BBN; the mode is not sensitive to outliers or skewed distribution as the arithmetic mean is. Moreover, the presentation of posterior probability as a random variable explicitly expresses the prediction uncertainty in terms of probability. Without the assumption of normality, the conditional probability of a predicted value is the outcome of binning continuous variables and treating all variables as on a nominal scale in the computation. However, one problem of the classification approach is that it is difficult to tell how far the predicted value missed the observed value when a case was misclassified.

Prediction Accuracy

In multiple regression, predication accuracy is usually quantified by residuals or studentized residuals. Also, the model R^2 is an index of how well the model fits the data. For example, Model III had a R^2 of .4214, which was considered an acceptable level of explained variance in regression given such a complex data set. The predication accuracy of the BBN model was the ratio of the number of correct classifications to the total number of predictions. In this study, the prediction accuracy of the BBN model was only 25.66% on the same training data.

Several explanations are available for this relatively low prediction accuracy of Model II compared to that of Model III. First, information was lost when continuous variables were binned: five of the six predictors were on an interval or ratio scale. Second, the final class identity of an individual case was algorithmically determined to be the salary bin that had the highest probability, which might not be substantially strong when the predictor variable was divided into many narrow bins (as in the above example p = .16). Third, when the bin widths are relatively narrow, misclassification may increase due to weakened differences among the levels of a variable. Finally, scoring functions used for model evaluation in the Bayesian network learning could be another factor. According to Friedman et al. (1997), when the structure of the network is not constrained with any prior knowledge as in the current case, nonspecialized scoring functions may result in a poor classifier function when there are many attributes.

Dimensional Simplification

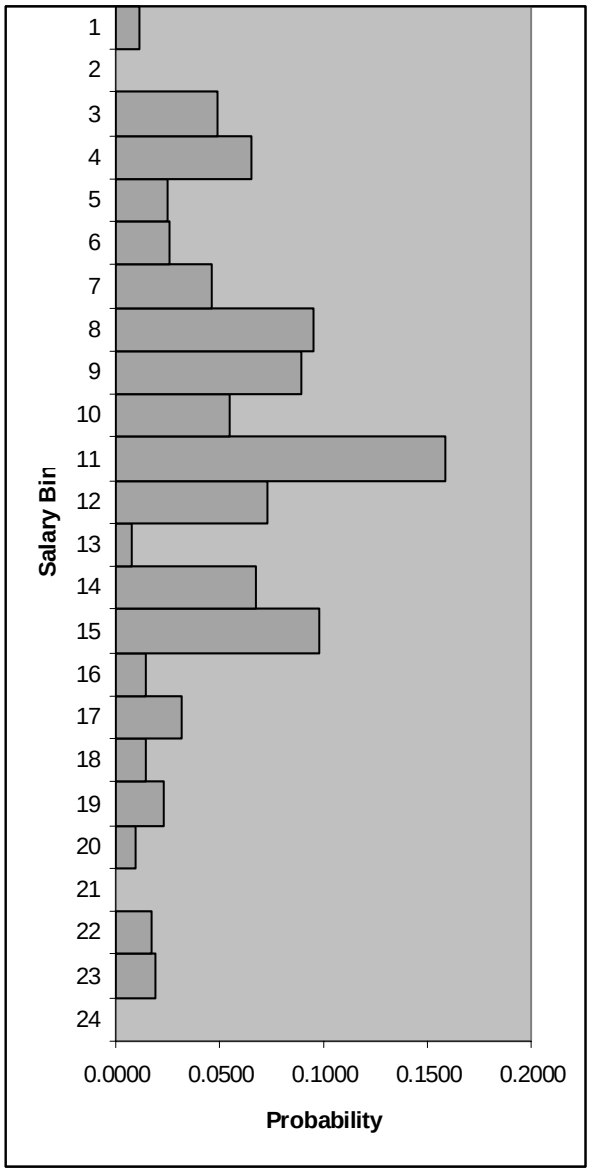
One important similarity between Models I and III is the final predictor variables. Model III had only five variables selected by the BBN model, and they were among the top six variables in the stepwise selection of Model I. Both models captured variables that shared strong covariance with the predicted variable. The overlap of the predictor variables is an indication that they both can serve the purpose of dimensional simplification.

In comparison to the automated process of variable selection and dimensional simplification in the BBN algorithms, the statistical approach was relatively laborious. However, the automation in BBN learning blinded researchers from having a detailed picture of variable relationships. In the statistical variable reduction, the clustering structure of variables was clear, and so were the variables that were similar or dissimilar to each other. Therefore, the high automation is only desirable when the underlying variable relationships are not of concern, or when the number of variables is extremely large.

The BBN data mining Model II identified five predictor variables that were subsequently used in Model III for prediction, all five independent variables were significant at $p < 0.0001$, and resulted in a final model with an

Table 6. An Example of the BBN Conditional Probability Tables.

Bin #	Salary range	Probability
1	Salary < 29600	0.0114
2	29600 < Salary < 32615	0.0012
3	32615 < Salary < 35015	0.0487
4	35015 < Salary < 37455	0.0655
5	37455 < Salary < 39025	0.0254
6	39025 < Salary < 40015	0.0263
7	40015 < Salary < 42010	0.0460
8	42010 < Salary < 44150	0.0950
9	44150 < Salary < 46025	0.0894
10	46025 < Salary < 48325	0.0552
11	48325 < Salary < 50035	0.1590
12	50035 < Salary < 53040	0.0728
13	53040 < Salary < 55080	0.0081
14	55080 < Salary < 58525	0.0672
15	58525 < Salary < 60010	0.0985
16	60010 < Salary < 64040	0.0140
17	64040 < Salary < 68010	0.0321
18	68010 < Salary < 72050	0.0142
19	72050 < Salary < 78250	0.0228
20	78250 < Salary < 85030	0.0098
21	85030 < Salary < 97320	0.0005
22	97320 < Salary < 116600	0.0170
23	116600 < Salary < 175090	0.0190
24	175090 < Salary	0.0005



Note. Salary was binned into 24 intervals. For this particular case, the product state is that the highest degree was obtained three years ago (X15_16 = 3), had three publications in juried media (Q29A1 = 3), spent 20% of the time teaching undergraduate classes (Q31A1 = .2) as an untenured (Q10AREC = 0) assistant professor (X01_8REC = 5) in a public research institution (STRATA = 12 and all other binary variables were 0).

R^2 = .4214 (df = 18 and adjusted R^2 = .4199). Although Model I has a greater R^2 than Model III, it also has more model degrees of freedom (47 vs. 18). Given an R^2 about .0822 higher than that of Model III at the expense of 29 more variables, each additional variable in Model I only increased the model R^2 by .0028 on average.

One of the negative effects associated with large numbers of independent variables in a multiple regression model is the threat of multicollinearity caused by possible strong correlations among the predictors. Model R^2 never decreases when the number of predictor variables increases, but if the variables bring along multicollinearity, estimated model parameters can have large standard errors, leading to an unreliable model. For the two regression models, Model I has 31 out of 47 variable with a VIF > 1.5 (66%). Model II has 10 out of 18 variables with a VIF > 1.5 (55%), and most of high VIF values are associated with the binary variables recoded from categorical variables.

Because the ordinary least square (OLS) method in prediction analysis produces a regression equation that is optimized for the training data, model generalizability should be considered as another important index of good prediction models. Model generalizability was measured by cross validating the proposed models with the holdout testing data set. Model I and III were applied to the 3,311 records to obtain their predicted values, and the R^2s of the testing data set were found to be .5055, and .4489, respectively, as compared with .5036 and .4214 in the original data set.

Large Data Volume

Multiple regression models have some problems when applied to massive data sets. First, many graphical procedures, including scatter plots for checking variable relationships become problematic when the large number of observations turns the plots into indiscernible black clouds. Second, with a large number of observations the statistical significance tests are oversensitive to minor differences. For example, a few variables with extremely small partial R^2s had significant p values in the stepwise selection of Model I. One particular case was the union status, which had a partial R^2 = .0009, given a sample size of 6,652, the variable was still added at a significant p = 0.0073.

Data mining models usually respond to large samples positively due to their inductive learning nature. Data mining algorithms rarely use significance tests, but rely on the abundant information in large samples to improve the accuracy of the rules (descriptions of data structure) summarized from the data. In addition, more data are needed to validate the models and to avoid optimistic bias (overfit).

Conclusion

In the field of education, large data sets recorded in the format of computer databases range from student information in a school district to national surveys of some defined population. Although data are sometimes collected without predefined research concerns, they become valuable resources of information for collective knowledge that can inform educational policy and practice. The critical step is how to effectively and objectively turn the data into useful information and valid knowledge. Educational researchers have not been able to take full advantage of those large data sets, partly because data sets of very large volume have presented practical problems related to statistical and analytical techniques.

The objective of this article is to explore the potentials of using data mining techniques in studying large data sets or databases in educational research. Data analysis methods that can effectively handle a large number of variables is one of the major concerns in this study of 91 variables (one was salary, the predicted variable).

The major findings are as follows. The multiple regression models were cumbersome with a large number of independent variables. Although the loss of degrees of freedom was not a concern given a large sample size, a thorough examination of variable interactions became unrealistic. The data mining model BBN needed much less human intervention in its automated learning and selection process. With the BBN algorithm inductively studying and summarizing variable relationships without probabilistic assumptions, the defense against normality and

linearity was dismissed, and significance tests were rarely necessary. However, the BBN model had some drawbacks as well. First, the BBN model, as most data mining models, is adaptive to categorical variables. Continuous measures had to be binned to be appropriately handled. The downgrade of measurement scale definitely cost information accuracy.

It also became clear in the process of this study that the ability to identify the most important variable from a group of highly correlated measures is an important criterion for evaluating applied data analysis methods when handling a large number of variables because redundant measures on the same constructs are common in large data sets and databases. The findings of this study indicate that BBN is capable to perform such a task because Model II identified five variables from groups of measures on teaching, publication, experience, academic seniority, and institution parameter, the same five as those selected by the data reduction techniques in building Model I for the reason that the five variables accounted for more variance of the predicted variables than their alternatives.

In general, data mining has some unique features that can help to explore and analyze enormous amount of data. Combining statistical and machine learning techniques in automated computer algorithms, data mining can be used to explore very large volumes of data with robustness against poor data quality such as nonnormality, outliers, and missing data. The inductive nature of data mining techniques is very practical to overcome limitations of traditional statistics when dealing with large sample sizes. The random selection of subset variables in making accurate predictions simplifies the problem associated with large number of variables. Nevertheless, the applicability of this new technique in educational and behavioral science has to be tailored for the specific needs of individual researchers and the goal of their studies.

By introducing data mining, a tool that has been widely used in business management and scientific research, this study demonstrated an alternative approach to analyzing educational databases. A clear-cut answer is difficult regarding the differences and advantages of the individual approaches. However, looking at a problem from different viewpoints itself is the essence of the study, and hopefully it can provide critical information for researchers to make their own assessment about how well these different models work to provide insight into the structure of and to extract valuable information from large volumes of data. Using confirmatory analysis to follow up the findings generated by data mining, educational researchers can virtually turn their large collection of data into a reservoir of knowledge to serve public interests.

References

Chen, J., & Greiner, R. (1999). Comparing Bayesian network classifiers. *Proceedings of the Fifteenth Conference on Uncertainty In Artificial Intelligence (UAI), Sweden*, 101-108.

Chen, J., Greiner, R., Kelly, J., Bell, D., & Liu, W. (2001). Learning Bayesian networks from data: An information-theory based approach. *Artificially Intelligence, 137*(1-2), 43-100.

Daszykowski, M., Walczak, B., & Massart, D. L. (2002). Representative subset selection. *Analytical Chimica Acta, 468*(1), 91-103.

Elder, J. F. & Pregibon, D. (1996). A statistical perspective on knowledge discovery in databases. In U. M. Fayyad, G. Piatetsky-Shapiro, R. Smyth, & R. Uthurusamy (Eds.) *Advances in knowledge discovery and data mining* (pp.83-113). Menlo Park, California: AAAI Press.

Fayyad, U. M. (1997, August). *Data mining and knowledge discovery in databases: Implications for scientific databases.* Papered presented at 9th International Conference on Scientific and Statistical Database Management (SSDBM'97), Olympia, WA**.**

Frawley, W. J., Piatetsky-Shapiro, G., & Matheu, C. J. (1991). Knowledge discovery in database: An overview. In G. Piatetsky-Shapiro & W. J. Frawley (Eds.) *Knowledge Discovery in Databases* (pp. 1-27). MIT: AAAI Press.

Friedman, J. H. (1997). Data mining and statistics: What's the connection? In D. W. Scott (Ed.), *Computing Science and Statistics*: *Vol. 29(1). Mining and Modeling Massive Data Sets in Sciences, and Business with a Subtheme in Environmental Statistics* (pp. 3-9). (Available from the Interface Foundation of North America, Inc., Fairfax Station, VA 22039-7460).

Glymour, C., Madigan, D., Pregibon, D. & Smyth, P. (1997). Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery, 1*, 11-28.

Hand, D. J. (1998). Data mining: Statistics and more? *The American Statistician, 52*, 112-118.

Hand, D. J. (1999). Statistics and data mining: Intersecting disciplines. *SIGKDD Exploration, 1*, 16-19.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Cambridge, MA: MIT Press.

Heckerman, D. (1997). Bayesian networks for data mining. *Data Mining and Knowledge Discover, 1*, 79-119.

Michalski, R. S., Bratko, I., & Kubat, M. (1998). *Machine learning and data mining: Methods and applications.* Chichester: John Wiley & Sons.

National Center of Education Statistics. (2002). National Survey of Postsecondary Faculty 1999 (NCES Publication No. 2002151), [Restricted-use data file, CD-ROM]. Washington, DC: Author.

Niedermayer, D. (1998). *An introduction to Bayesian networks and their contemporary applications.* Retrieved on September 24, 2003 from http://www.niedermayer.ca/papers/bayesian/

Thearling, K. (2003). *An Introduction to Data Mining: Discovering hidden value in your data warehouse.* Retrieved on July 6, 2003 from http://www.thearling.com/text/dmwhite/ dmwhite.htm.

Wegman, E., J. (1995). Huge data sets and the frontiers of computational feasibility. *Journal of Computational and Graphical Statistics, 4*(4), 281-295.

Yu, Y., & Johnson, B. W. (2002). *Bayesian belief network and its applications* (Tech. Rep. UVA-CSCS-BBN-001). Charlottesville, VA: University of Virginia, Center for Safety-Critical Systems.

Zhou, Z. (2003). Three perspectives of data mining. *Artificial Intelligence*, *143*(1), 139-146.