Learning the Structure of Retention Data using Bayesian Networks

Amy McGovern, Christopher M. Utz, Susan E. Walden, Deborah A. Trytten
amcgovern@ou.edu, Christopher.M.Utz-1@ou.edu, Susan.E.Walden-1@ou.edu, dtrytten@ou.edu

Abstract – We introduce a novel approach to examining retention data by learning Bayesian Networks automatically from survey data administered to minority students in the College of Engineering at the University of Oklahoma. Bayesian networks provide a human readable model of correlations in large data sets, which enables researchers to improve their understanding of the data without preconceptions. We compare the results of our learned structures with human expectations and interpretation of the data as well as with cross-validation on the data. The average Area Under the Curve of the networks using cross-validation was 0.6. The domain experts believe the methodology of automatically learning such structures is promising and we are continuing to improve the structure learning process.

Index Terms—machine learning, Bayesian networks, retention, minority students.

INTRODUCTION

Inference is the task of answering a query given evidence. For example, “Will this student be retained in the College of Engineering given academic and survey data?” In this paper we demonstrate how we can use Bayesian networks to identify factors that are relevant to the retention of minority students.

This work is part of a larger project describing the differential strategies used by members of four minority groups (Native Americans (NA), Hispanic Americans (HA), Asian Americans (ASA), and African Americans (AFA)) to be successful as engineering undergraduates at the University of Oklahoma (DUE-0431642). Our research methodology uses a mixed method approach that blends both quantitative and qualitative elements. Unique features of our previous work include the focus on students who succeed at our institution, disaggregation of minority groups, and the inclusion of a minority group that is not under-represented (ASA).

Our primary data sources include one time and longitudinal interviews with minority engineering students, academic transcripts and a survey instrument. Most of our previous work has focused on interpreting the narratives in the qualitative interview data, using the quantitative data for triangulation [1-3]. In this work, we reverse this pattern and focus on discovering the knowledge within the quantitative survey data using machine learning.

Machine learning has been previously applied to engineering education data, including both assessment of student learning [4, 5] and retention [6, 7]. Researchers have applied statistical analysis to large survey- and academic record- based data sets [8-10]. These studies typically compare pre-college characteristics or experiences with the likelihood of choosing or of completing a degree in a science or engineering major. More recently institutions or consortia have cooperated on wide-spread data collection to allow researchers to focus surveys and academic record analysis specifically on questions surrounding retention in engineering [11-18]. Several of these studies have employed logistic regression models to identify factors predicting retention. These studies vary widely in the consideration of background factors (e.g. high school GPA) and college-life factors (e.g. grades in calculus or change in confidence in engineering skills).

One of the initial attempts to include affective measures in studying persistence in undergraduate engineering was the Pittsburgh Freshmen Engineering Attitude Survey (PFEAS)[11, 12]. The original PFEAS survey included measures of student attitudes toward engineering as a career and student perceptions of their high school preparation for engineering related coursework. This survey was evaluated and found to be both valid and reliable [11, 12] and later modified to include measures of confidence in performing ABET EAC outcomes [13]. Additional modifications and evaluation produced appropriate versions to use for sophomores, juniors, and graduating seniors with the addition of questions about engineering internship or research experiences and replacing high school preparedness questions with ones related to the previous year’s curriculum. Most of the research based on PFEAS used a pre-test and post-test design to look for statistically significant differences demonstrating changes in student attitudes or confidence [11-13, 19].

DATA

The data for this experiment is derived from a modification of the Pittsburgh Junior Engineering Learning and Curriculum Evaluation Instrument [19]. The modified instrument was administered as supplemental data for a qualitative study of persistence in engineering. Our data also included demographic data for the participants and
information excerpted from the participants’ academic transcripts.

Since our participants were recruited from any non-freshman academic classification and because we wanted a single instrument for all engineering majors, we made the following slight modifications to the Junior instrument.

- Removed questions specific to developers’ context,
- Mirrored engineering-related work questions with non-engineering-related work questions,
- Created a computer science specific survey.

This last change was similar to the changes in ABET outcomes between the Engineering Accreditation Commission [20] and the Computing Accreditation Commission [21]. The validity and reliability of the original survey has not transferred to our instrument.

For the machine-learning test on these data, responses from 150 students across 9 engineering disciplines and computer science were used. Given the small number of participants in this pilot study and the resulting need to limit the number of variables, we grouped the questions on the survey into the following categories: (1) Preparedness, (2) Attitudes about engineering, (3) Confidence in ability to perform ABET outcomes, (4) Outcome Expectations of persistence and engineering career, (5) Engineering (or CS) related work experience, and (6) Non-engineering (or CS) related work experience. For the survey questions, a student’s Likert-type responses to the questions assigned to each category were averaged. Demographic and academic transcript variables included ethnicity, gender, major (divided into enrollment managed discipline or not), hours of transfer credit, high school type (rural or urban), number of hours of advanced standing credit, and grade point average in STEM classes. The final variable of retention was defined as positive if a student had graduated with an engineering or CS degree or was continuing in good standing along that path and negative otherwise. Thus each student was represented by 14 variables. Many of these variables are continuous and with so few students to train on, we discretized the variables by examining the mean of each variable and choosing breakpoints at the mean plus and minus one half of the standard deviation. Of our 150 students, 16 were not retained.

**BAYESIAN NETWORKS**

For any data set with more than a few variables, representing a full joint probability distribution in a tabular format can quickly become intractable. However, this information is necessary for inference. Bayesian networks [22, 23] make use of Bayes’ Rule and conditional independencies to compactly represent the full joint probability distribution. Bayesian networks are usually shown graphically, such as the example network built from our data shown in Figure 1. Bayesian networks with discrete variables store conditional probability tables (CPTs) at each node. These tables give the probability of observing a particular value for that variable given the values of its parent variables.
For example, in Figure 1, the CPT for outcome expectations depends on the observed value of high school type. This means that calculating the probability of a new but similar student having a high outcome expectation value depends directly on their observed values for high school type. Furthermore, the outcome expectation value directly impacts the probability calculations for GPA, engineering related work experience, and a hidden node. A more complicated example from this network is the enrollment management node, which is depends on the student’s confidence and gender. More broadly, we can calculate the probability of a new student staying in the College of Engineering from the network and its associated CPTs.

Survey instruments are unable to measure all relevant variables about a student. However, Bayesian networks can handle unobserved measures through hidden variables. Hidden variables are a common strategy used to improve the performance of the network and to simplify the CPTs [22-24].

The size of a CPT depends on the number of parent nodes for the variable in question as well as the number of possible values for that variable. Without the hidden nodes, the network had CPTs that had more rows than students in our training set. Such a network would be unable to generalize accurately to new students.

We learned the CPTs for the hidden nodes using Expectation Maximization (EM) [25]. Each student is initially assigned a value for every hidden node uniformly randomly between 0 and 1, respecting the laws of probability. EM then modifies these probabilities iteratively using the values of the observed variables. EM is a standard gradient ascent search technique, which converges to a local maximum. We chose six binary hidden nodes for our training although the network could choose to ignore some of them by not connecting them into the overall network.

We initially created several Bayesian networks by hand, drawing on existing literature in retention to construct plausible networks. However, the performance of the handcrafted networks was not as high as we expected. We hypothesized that an automated approach would enable us to find additional structure in the data and that the learned structures would outperform the handcrafted structures.

The networks were trained and evaluated using eight fold cross validation. The data set was partitioned into eight subsets, called folds, such that each fold contained an approximately equal number of students who were not retained. One fold is designated as the test set, and the remaining folds are combined into a training set. A structure is learned on the training set and then evaluated through inference on the test set. In the inference step, the probability of retention for each student in the test set is calculated according to the learned Bayesian network. This is repeated eight times, so that each individual subset becomes the test set once. We chose to have eight folds because it maximized the number of folds without isolating students that were not retained.

The following sum squared error was used to assign a numerical score to each network during learning.

\[ \sum_{\text{training set}} \text{(actual probability } - \text{ inferred probability)}^2 \] (1)

The actual probability of a student being retained was 1.0 if the student was retained in the College of Engineering or 0.0 if the student was not retained.

Structure learning for Bayesian networks is an unsolved problem, particularly because the search space is so large and the best networks for the training data tend to overfit to the data. In this study, we used a randomized beam search approach with a beam width of twenty [24]. The initial network had all 14 of the observed variables plus six hidden nodes representing characteristics that we could not directly measure. The initial network had zero edges, or links between a variable. At each step of the search, twenty copies of the network were made and a new random edge was added to each structure. To ensure the random edges were distributed somewhat uniformly, a heuristic was used to add new edges to the structure. Edges were added such that no new edge created a cycle in the structure, no node had an in-degree (number of parents) of more than five and no node had an out-degree (number of children) of more than five. In addition, ethnicity, gender, and high school type were not allowed to have a parent node and retained was disallowed from having child nodes. Each of the networks was evaluated using Equation 1 to determine the starting network for the next iteration of the search. Search continued until the network contained either a number of edges of at least 1.5 times the number of nodes or the score from one iteration of search to the next iteration changed by less than one percent. The beam search was repeated 28 times for each fold, each with a different starting seed. We chose the network with the lowest summed squared error on the training data as the final network for each fold.

We evaluated the final networks in several ways. First, we measured the average summed squared error as given in Equation 1. The average error for the eight-fold run was 2.2. Since this was the measure used to train the network, we expected it to be low on average. Although we could have used accuracy of prediction to evaluate the networks, this measure is problematic with these data and this network since predicting every student was retained would be 90% correct. A more precise measurement of an algorithm’s performance under unknown misclassification costs and unequal class distributions comes from the Area Under the receiver operator Curve (AUC) [26]. The AUC can range from zero to one with 0.5 being the score for a random algorithm and one being the score for a perfect algorithm. The average AUC for our eight-fold cross validation was 0.6. While this number is not as high it likely could be, it does indicate that the networks’ identified structure has predictive power. We believe the reason for the relatively low AUC is overfitting, in other words the network is learning extraneous structure from data. This assumption
was confirmed by examining the performance on the training set, where we achieved perfect AUCs of 1.0. We experimented with using six of the eight folds as training data, one as test, and one as a validation set to stop network growth when overfitting occurred but this decreased the size of the training set too much and increased overfitting. With more training data, this could be easily addressed.

Figure 2 shows an example composite network learned using this approach. The summed squared error for this network was 2.6 and the AUC for this network was 0.8. This network indicates that a student’s GPA, engineering related work experience, non engineering related work experience, and major (enrollment managed or not) directly relate to the probability of that student being retained in the College of Engineering. In the example network, only 5 of the 6 possible hidden nodes were connected to the network. The other networks often used similar numbers of the hidden nodes. It is likely that we need fewer nodes than we introduced and in current work, we are exploring the ability of the network to choose the appropriate number of hidden nodes.

Our final evaluation of the networks comes from domain experts evaluating and examining the learned networks. Although the experts examined the best network for each fold of the eight fold cross validation, we were more interested in which linkages were found repeatedly across the cross validation, demonstrating a robust correlation between the variables. Figure 2 shows the composite network for all edges that occurred three or more times across the eight folds of cross validation. The direction of the edges is ignored when creating the composite network. This means that if A is connected to B three times and B is connected to A two times, the composite network will show an undirected edge from A to B with the label of 5.

INTERPRETING THE RESULTS

At first glance, the composite network illustrated in Figure 2 may seem unsurprising. Proceeding counter clockwise around the retention node, non-engineering related work experience, ethnicity, engineering related work experience, GPA and outcome expectations all are directly related to retention in five or more of the eight folds. The average probabilities derived from the individual folds containing the ethnicity to retained link verified that the nature of the relationship was reasonable based on prior knowledge of our participants. Among our research participants, 83% of Native American and 84% of African American students graduated or are persisting while 93% of the Hispanic and Asian American students meet that criterion. The learned network shows the probability of retention as Hispanic (94%) ≈ Asian American (93%) >> African American (86%) > Native American (83%). Although their actual retention rates are similar, the unequal representation of NA and AA participants (NA=35, AA=44) in our data may be slightly magnifying the learned probability differences. Another study has shown that ethnicity influences retention [17].

One of the strongest relationships in the composite is between GPA and retained, appearing in six of the eight folds. High school GPA has previously been linked to enrollment or retention in STEM [17, 18]. Students who had a STEM GPA of less than 2.5 (out of 4.0) have a low probability of retention (70%). Students who declined to allow us to access to their academic transcript also had a low probability of being retained in engineering (76%). The decision to not allow us access to transcripts possibly meant that the student had a less than ideal academic record. An anomaly is that students with GPAs between 2.5 and 3.2 have a higher probability of being retained (99.6%) than students with GPAs higher than 3.2 (95%). We would expect this relationship to be reversed.

FIGURE 2
COMPOSITE LEARNED BAYESIAN NETWORK

38th ASEE/IEEE Frontiers in Education Conference
F3D-10
The relationship between engineering related work experience (defined in the survey as an internship, co-op, or research experience) and retention is striking. Students who described their engineering related work experience as not contributing meaningfully to their engineering education were less likely to be retained (79%) in engineering than students who had a moderate (97%) or high quality (92%) experience. Beyond that, students with a low quality experience were less likely to be retained than students who had no engineering related work experience (87%). This result points out the importance of vetting engineering related work experiences for students to avoid losing students unnecessarily.

Noticeably absent from the composite are direct links from gender (1 fold), enrollment management (2 folds), confidence values (2 folds), or attitude values (1 fold) to retention. These links exist in some folds, as shown parenthetically above, but were not common enough to show up in the composite graph. Variability between folds is expected. If the algorithm produced identical results from every fold it might be an indication that the structure of the data was sufficiently straightforward that a simple search would have sufficed and that machine learning wasn’t necessary.

Another source of variability is the relatively small number of students used in the learning of the network. While 150 students is a large sample for gathering interview data, only 16 of them left the College of Engineering. With more students, the issue of multiple correct networks would likely be mitigated and the results would be even more consistent. Another aspect of the variation comes from the random search algorithm used in the learning process. The results here have convinced us that using Bayesian networks to help interpret a diverse data set has potential; therefore, we are continuing to explore other machine learning approaches, particularly ones designed for small data sets.

CONCLUSIONS AND FUTURE WORK

We demonstrated that the technique of automatically learning Bayesian networks can provide a human readable model of factors influencing retention within the College of Engineering for students who had achieved sophomore or higher status using survey data and academic and demographic information. With 150 student and fourteen observed variables with 46 possible values, human based analysis is already nearly intractable. Although the data set size is small for machine learning (and for statistical inference), we are able to produce a reasonable human readable model that can be used to improve the analysis of the data.

Because of the small size of the data set from a statistical point of view, overfitting is an issue. Overfitting can be detected by a large performance difference between the training and testing sets. Additionally there may be multiple correct networks to the same prediction task. While each of these networks is valid given the training data, domain experts often expect more consistency. In addition, while a given composite network may not include all of the dependencies that a domain expert might expect, they also have not included any dependencies that seemed irrational.

We are currently experimenting with new learning strategies. We are pursuing alternative approaches to automate network construction. We initially chose a randomized greedy approach to building the networks and we are planning to incorporate additional information including prior knowledge provided by a domain expert and statistical dependencies implied by the data. We are planning to allow the network to learn the appropriate number of hidden nodes rather than defining this network feature. This should both improve the overall results and the readability of the final networks. We are also pursuing collaborations to apply Bayesian networks to larger data sets. This will also address overfitting issues.

Although we are not including student quotes in this paper as we have in earlier works, our merger of qualitative and quantitative data and methodology aids in the interpretation of the Bayesian network. Understanding the full range of issues related to student retention benefits from both mathematical descriptions of common characteristics and textured examinations of individual narratives.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation’s Directorate of Undergraduate Education’s STEM Talent Expansion Program Grant No. DUE-0431642. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

The authors would like to thank the following collaborators who have contributed to this work: Teri J. Murphy, Teri Reed-Rhoads, Randa Shehab, Jeanette Davidson, Cindy E. Foor, Rosa Cintron, Paul Rocha, Francey Freeman, Kimberly Rutland, Mayra Olivares, Tony Lee, Claudia Morales, Lisa Schmidt, Tyler Combrink, Lauren Rieken, Anna Wong Lowe, Sedelta Osahwee, Johanna Rojas, Andres Guerrero, Wen-Yu Chao, Monica Flippin-Wynn, Nathaniel Manzo, Tiffany Davis-Blackwood, Tracy Revis, Jeff Trevillion, Van Ha, Ben Lopez, Quintin Hughes, Bach Do, Yi Zhao, Brittany Shanel Norwood, Ruth Moaning, Ginger Murray, William Stephen Anderson, Elaine Seymour, Karina Walters, Larry Schuman, David Bugg, James Borgford-Parnell, Mary Anderson-Rowland, Tony Lopez, Adrianna Kruger, and Sean Williams.

REFERENCES


[2] Foor, C. E., Walden, S. E., and Trytten, D. A., “‘I wish I belonged more to this whole engineering group’: Achieving Individual


AUTHOR INFORMATION
Amy McGovern, Assistant Professor, School of Computer Science, University of Oklahoma, ammcgovern@ou.edu

Christopher M. Utz, School of Computer Science, University of Oklahoma, Christopher.MUtz-1@ou.edu

Susan E. Walden, Director, Research Institute for STEM Education, K20 Center for Educational and Community Renewal, University of Oklahoma, Susan.E.Walden-l@ou.edu

Deborah A. Trytten, Acting Associate Director & Associate Professor, School of Computer Science, University of Oklahoma, dtrytten@ou.edu