ADMIT - A Web-Based System to Facilitate Graduate Admission Process

Dmitriy Babichenko,[1] Marek J. Druzdzel[1,2]
[1]School of Information Sciences, University of Pittsburgh, Pittsburgh, PA, USA
[2]Faculty of Computer Science, Bialystok University of Technology, Bialystok, Poland

Abstract
In this paper we describe ADMIT, a software application developed to assist the graduate admissions process at the University of Pittsburgh School of Information Sciences (SIS). ADMIT uses a Bayesian network model built from historical admissions data and academic performance records to predict how likely each applicant is to succeed. The system rank-orders applicants based on the probability of their success in the Master of Science in Information Science (MSIS) program and presents results as an ordered list and as a histogram to the admission committee members. The system also enables users to see a graphical representation of the model (a causal graph) and observe how each input data point affects the system's suggestions.

Keywords: graduate admissions; Bayesian networks; machine learning;
doi: 10.9776/16240

Acknowledgements: While we take full responsibility for any errors and shortcomings of this paper, we thank anonymous reviewers for their insightful remarks that led to improvements of this paper. We thank Dr. Michael Lewis, Professor at the University of Pittsburgh School of Information Sciences and Ninette Kay (Admissions Coordinator) for providing information and insight about the SIS admission process. We also thank Olena Sherbinin (Data Architect) and Wesley Lipschultz (Director of Student Services) for making the data available to us. We also acknowledge the support the National Institute of Health under grant number U01HL101066-01. Models used in this project were built using GeNIe Modeler and SMILE engine, a Bayesian modeling environment developed at the Decision Systems Laboratory, University of Pittsburgh, and available free of charge for academic use at http://www.bayesfusion.com/.

Contact: dmb72@pitt.edu

1 Introduction

Graduate school admissions is a time-consuming process that puts a heavy burden on admissions committee faculty and staff. All SIS applicants must first complete an online application (?, ?), provide a written statement of goals, two letters of recommendation, academic transcripts from previous undergraduate and/or graduate degrees, and GRE scores. International applicants must also provide either TOEFL (?, ?) or IELTS (?, ?) scores. All applications and supplementary materials are then carefully reviewed by the admissions committee faculty who decide to either accept, provisionally accept, wait-list, or reject an applicant.

While there are numerous systems on the market that help higher education institutions manage the admissions process, surprisingly little work has been done to develop software that would go beyond streamlining admissions processes and organizing data, software that would actually help predict an applicant's success in his or her chosen field of studies.

A notable exception is GRADE, a statistical machine learning system developed by Waters and Miikkulainen (2013) to support the work of the PhD admissions committee at the University of Texas at Austin's Department of Computer Science (?, ?), which inspired our efforts.

In this paper we describe ADMIT, a graduate admissions support system designed to reduce the application review time and to help the admissions committee objectively identify students who are most likely to succeed in the MSIS program at the University of Pittsburgh's School of Information Sciences (SIS). ADMIT is based on a Bayesian network model built from historical admissions data and academic performance records.

The use of Bayesian networks in decision support systems is not a novel concept. Bayesian networks are widely used in medical diagnosis (?, ?, ?), business analytics (?, ?), and even forecasting the results of presidential elections (?, ?).

2 The Data

The data used to build and validate Bayesian models described in this paper consist of a set of de-identified admissions and academic records of students who applied to the MSIS program between 2009 and 2015. The initial data set received from the admissions office contained 1,482 applicants' records, 614 records for admitted and matriculated students, and 405 records for students who graduated.

"Honest brokers" approved by the University of Pittsburgh Institutional Review Board (IRB) de-identified the data by removing applicants' and students' names, addresses, dates of birth, and other information that could be used to determine their identity. The honest brokers also replaced students' and applicants' IDs with globally unique identifiers (GUIDs) to allow linking of applicants' and admitted students' records.

We divided the data into two sets. We used the data set containing 405 records for students that graduated from the MSIS program to build Bayesian network models used by our software. The second data set contains 200 records randomly selected from the pool of applicants who applied for the admission to the MSIS program for the fall of 2015. 25 applicants from the randomly selected group had been accepted by the SIS admissions committee. We used this second data set to validate our Bayesian networks, and to gauge the accuracy of the system's suggestions.

2.1 Variables In The Data Set

Age at Matriculation
We derived the age at which each student matriculated at SIS from that student's birth year and matriculation date.

College Ranking and Rank Types
Because many of the SIS applicants are international students, we had to obtain rankings for each university that granted each applicant's most recent degree from multiple sources. Rankings of US universities came from the U.S. News and World Report's website (?, ?). Since U.S. News and World Report offers separate rankings for national universities and liberal arts colleges, we created an additional column to differentiate national universities and liberal arts colleges and used this differentiation in conjunction with the obtained rankings. We obtained rankings for international universities from the global universities rankings section of the U.S. News and World Report website (?, ?), as well as from the topuniversities.com rankings of the top universities in the five BRICS countries (Brazil, Russia, India, China and South Africa) (?, ?). We used the same column that allowed us to differentiate U.S. national universities and liberal colleges to specify the country of origin for each ranked university. If a ranking for an institution was not available, we designated it as "unranked."

Applicants' GPA
This GPA value refers to the Grade Point Average from the most recent degree prior to the application to SIS. The GPA values for the applicants who graduated from U.S. universities are on a 4.0 scale, while GPA values for the applicants who graduated from universities outside of the U.S. are on a 100-point scale. We converted GPA values from both scales to z-scores. A z-score, or a standard score, identifies and describes the exact location of each raw score in a distribution. In other words, a z-score measures how many standard deviations below or above the population mean a raw score is (?, ?).

Major Category
We grouped all applicants' majors into the following broad categories: Business, Computer and Information Sciences, Education, Engineering, Library Information Sciences, Mathematics and Statistics, Physical and Biological Sciences, Social Sciences and Liberal Arts.

GRE Scores
GRE examinations changed their scoring scales on August 1, 2011. Since we reviewed applicants' records beginning with 2009, the data set contained GRE scores in two different scales. We converted each raw analytical writing, quantitative, and verbal score to z-scores for both scales.

Test Of English As Foreign Language (TOEFL) and International English Language Testing System (IELTS) Scores
SIS requires all international applicants to provide their GRE scores and either TOEFL (?, ?) or IELTS (?, ?) scores. For international applicants, we substituted verbal GRE section z-scores with corresponding TOEFL or IELTS z-scores. We believe that TOEFL exam is more reflective of foreign students' command of English than verbal GRE or GMAT.

Student's GPA at Graduation This GPA value refers to the Grade Point Average at the time of a student's graduation from SIS. All applicants' GPA values in the data set were on a 4.0 scale.

2.2 Discretization

The PC causal discovery algorithm can analyze data sets with either discrete data or continuous data, but not both, and the Tree Augmented Naive Bayes (TAN) algorithm can only work with discrete data (?, ?). To bring all the data to a common denominator, we chose to discretize all continuous variables accodring to the categories shown in Table 1. The thresholds are based on frequency distributions of each variable and on how the pre-admission data are used in admission decisions by the admission committee members.

Variable	Discrete Ranges
Age at Matriculation	[<25], [25 to 26], [26 to 28], [28 to 31], [>31]
College Ranking	[0 to 10], [10 to 20], [20 to 50], [< 50]
Applicant's GPA (z-score)	[<-2], [-2 to -1], [-1 to 0], [0 to 1], [1 to 2], [>2]
GRE Analytical Writing Score (z-score)	[<-2], [-2 to -1], [-1 to 0], [0 to 1], [1 to 2], [>2]
GRE Quantitative Score (z-score)	[<-2], [-2 to -1], [-1 to 0], [0 to 1], [1 to 2], [>2]
GRE Verbal Score (z-score)	[<-2], [-2 to -1], [-1 to 0], [0 to 1], [1 to 2], [>2]
GPA at Graduation	[<3.2], [3.2 to 3.7], [>3.7]

Table 1: Discrete categories for the pre-admissions data

3 Methodology

3.1 Bayesian Networks

We used GeNIe Modeler to create Bayesian networks described in this paper. Because we wanted to test the robustness of the system, as well as to create a network that would best predict values for dependent features, we used two different algorithms to learn from data and to create network structures.

3.1.1 The PC Algorithm

The PC algorithm (?, ?) is a member of the constraint-based search class of algorithms that use a chi-square test on each pair of discrete variables in a data set to determine conditional independence. Such algorithms produce a set of directed graphs representing the dependencies discovered in data. We used the PC algorithm to learn the causal structure from admissions records at the significance level $\alpha = 0.5$ (Figure 1).

It is important to note that since the PC algorithm attempts to discover the causal structure that has generated the data, and since causation has direction (e.g., a student's major may be a causal factor in success, not the other way around), it is helpful to specify the time precedence among the studied variables so that the PC algorithm can automatically discard the possibility of a causal relationship going backward in time. For example, since college rank precedes GRE verbal scores (Table 2), the algorithm would not create an arc pointing from GRE verbal score to college rank.

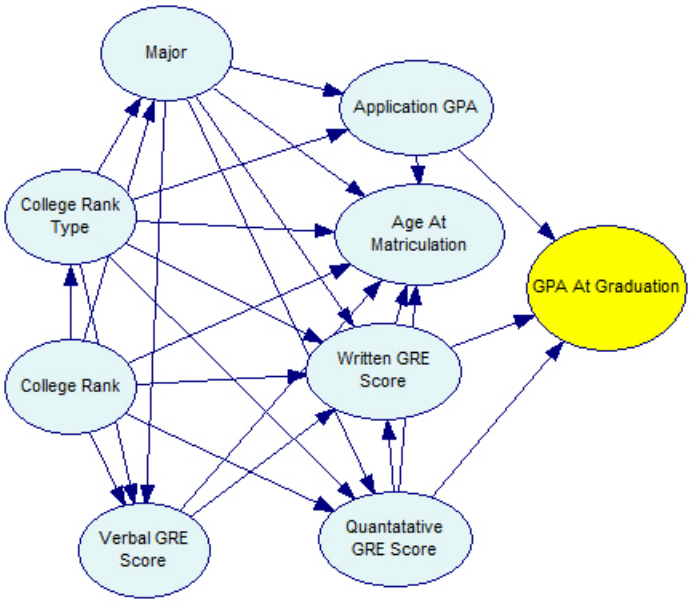


Figure 1: Causal structure generated by the PC algorithm from data at $\alpha = 0.5$

Temporal Precedence	Variable(s)
1	Gender College ranking Rank type
2	Major
3	GPA (prior to applying to SIS) GRE (analytical writing, verbal, quantitative)
4	Age at matriculation to SIS
5	GPA at graduation from SIS

Table 2: Temporal precedence for prior knowledge

3.1.2 Tree Augmented Naive Bayes Classifier (TAN)

Naive Bayes is a well-known Bayesian classifier that tends to outperform more sophisticated classifiers, especially in datasets where the variables are not strongly correlated (?, ?). Naive Bayes has several advantages over many other classifiers, including the PC algorithm. Because it assumes conditional independence among the features given the class in a data set, its structure is given a priori. Therefore, constructing a Bayesian network using the Naive Bayes classifier requires no structure learning procedure. Furthermore, because of the assumed independence, the classification process is very efficient (?, ?). However, in real life we cannot assume that variables in a data set are independent. The Tree Augmented Naive Bayes (TAN) algorithm adds a tree structure to a Naive Bayes model by connecting the most dependent attributes with directed arcs (?, ?). This results in a more accurate modeling of the join probability distributions and improves the predictive accuracy of the model.

We selected the students' GPA at the time of graduation from the MSIS program as the classifier variable for the TAN model, with all the other variables described in Section 2 of this paper as the features. The TAN algorithm yielded the structure shown in Figure 2.

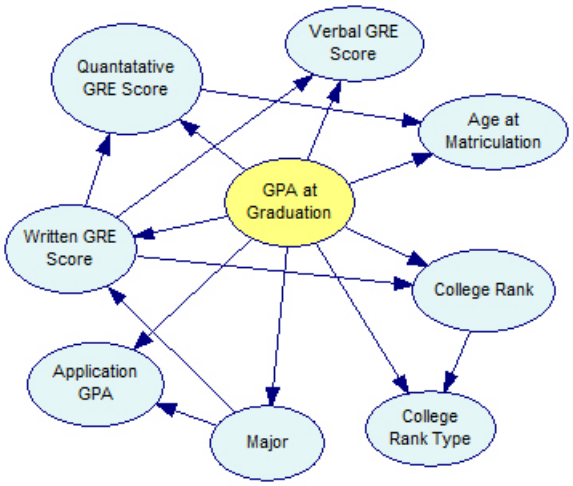


Figure 2: Tree Augmented Naive Bayes network generated from data by GeNIe

3.2 Implementation Details

ADMIT is a web-based database-driven application written in C#. The application includes SMILE API to learn and build Bayesian networks from data, as well as to use Bayesian networks designed in GeNIe to rank-order applicants and to make suggestions about which applicants to admit.

ADMIT can read data from multiple data sources, including SQL queries to Oracle, Microsoft SQL Server, and MySQL databases, as well as from comma-separated values (.csv) files.

The application consists of a set of data pre-processing tools written in Python with Natural Language Processing Toolkit (NLTK) (?, ?), a Bayesian network selection screen, two screens for selecting data sources (comma delimited file or SQL query), and an applicants' analysis and ranking screen (Figure 3). While this design requires administrators to go through an extra step of linking models and data sets, the system has the flexibility to work with any model created in GeNIe, regardless of the algorithm that was used to create the model.

Admission officers can use the screen shown in Figure 3 to select an applicants' data set for the system to analyze. Once the user selects a data set, the system uses SMILE API to update values for each node in the network linked to the selected data set with the values from each record in the selected data set. The system recalculates the probabilities associated with the feature that we are trying to predict (in this case, the applicant's GPA at graduation from SIS). The system then rank-orders applicants in descending order based on the probability of their GPA at the time of graduation from SIS being ≥ 3.7.

The results of the analysis also include a graph that shows the probability distribution over the probability of the event "GPA ≥ 3.7"(Figure 4). Admissions officers can use this screen to specify two threshold values – a threshold for the lowest acceptable probability and a threshold for the class size. The system can use these values to color-code the results to clearly show which applicants to admit.

4 Results

In order to assess how well our models will generalize to future applicants' data, we validated the models using 10-fold cross-validation (?, ?).

Students' success (GPA ≥ 3.7) prediction accuracy for the model generated using the PC algorithm is 33% and 51% for the model generated using the TAN algorithm. Figures 5a and 5b show the ROC curves for the two models.

ID	gpaZscore	majorCategory	matriculationAge	quantatativeGRE	verbalGRE	writtenGRE	Score
663	-0.37	unknown	26	0.23	-0.31	-0.57	96.63
948	1.37	science	22	1.31	0.61	-0.57	92.61
972	1.7	cs	24	1.31	0.22	0.75	90.81
1871	1.1	cs	23	-0.74	0.74	-0.57	90.24
1713	1.5	cs	24	-0.01	-0.31	-0.57	89.42
1830	0.07	unknown	27	-0.13	1.53	2.07	82
869	1.77	cs	24	0.95	-0.57	-0.57	76.95
1912	1.54	cs	22	0.23	-0.84	-0.57	76.95
486	0.47	cs	23	1.07	0.61	-0.57	76.83
968	0.9	cs	23	1.31	0.61	-0.57	76.83
729	0.7	cs	23	1.07	-0.18	-0.57	75.46

Figure 3: Rank-ordered applicant list. Applicants highlighted in green are above admissions committee-specified acceptance threshold. Note that this figure does not display all variables used in the model, and that the values shown are fictitious.

5 Discussion and Future Work

One of ADMIT's weaknesses is that it still requires some manual data pre-processing. The applicants' data are imported into ADMIT from a web-based university admissions system called ApplyYourself (?, ?). While most of the data are structured, applicants complete fields such as university names and undergraduate majors by typing in responses instead of selecting them from a predefined list. Because of errors in data entry and variations in spelling (often caused by translation into English), we wrote a Python script that utilizes NLTK to identify and remove the aforementioned inconsistencies before importing the data into a MySQL database. The pre-processing script corrects approximately 83% of spelling errors and typos, but the remaining 17% have to be corrected manually.

Another possible issue may arise from the discrete categories that we chose for the PC algorithm. Spirtes et al. (?, ?, p. 271-272) show that discretization can possibly alter the results of the algorithm's decisions about conditional independence which may affect the structure learned by the PC algorithm.

The most obvious reason for the models' relatively weak performance is that we do not have performance data on rejected students. Furthermore, because the training data set consists of data from applicants who had already been admitted to SIS, there is little variance in GRE scores, undergraduate GPAs, and students' GPAs at graduation from SIS.

In order to improve the accuracy of ADMIT's recommendations, we are planning automating scoring of applicants' written goal statements and recommendation letters (?, ?) and augmenting the models with these additional variables.

There exists other classification algorithms (?, ?). However, there is no consensus as to which algorithms are consistently more accurate then others — it all depends on the problem and on the data. Our choice of Bayesian Networks was influenced by their solid grounding in statistics and their intuitive structural representation. To verify whether the accuracy we achieved was in line with other methods, we also built a support vector machine (SVM) and a Random Forest (RF) model using Weka (?, ?). Both algorithms belong to the very best classification algorithms. 10-fold validation showed predictive accuracy of 51.11% and 57.46% for SVM and RF models respectively, both in the same range as the TAN model.

We are also proposing a multi-year study where SIS will compare ADMIT's recommendations with

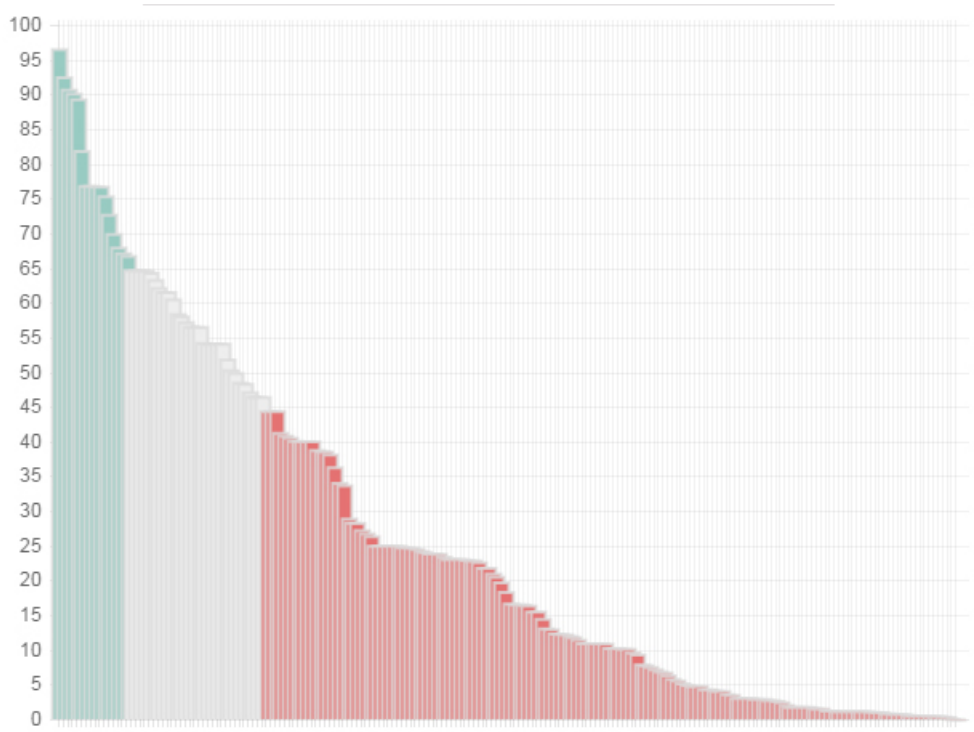
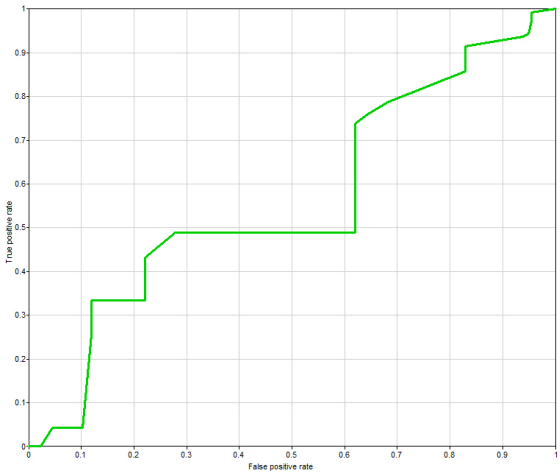


Figure 4: Graph representing the distribution of probabilities that applicants' GPA at graduation will be ≥ 3.7.

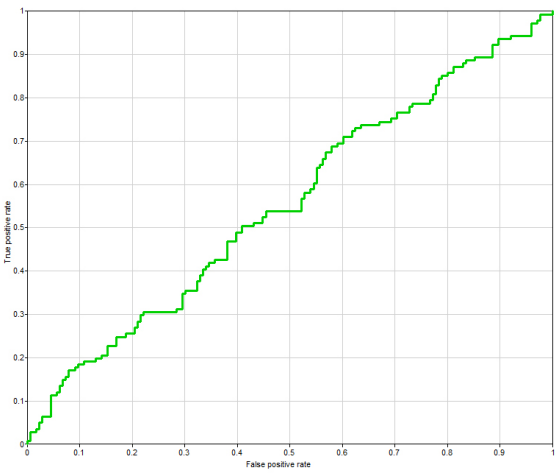
the recommendations of the admissions committee and observe how academic performance of the applicants initially rejected by ADMIT but accepted by the committee compares with that of the applicants recommended by ADMIT.

Table of Tables

Table of Figures



(a) ROC curve for model generated by the PC algorithm; predicting GPA at graduation ≥ 3.7.



(b) ROC curve for model generated by the TAN algorithm; predicting GPA at graduation ≥ 3.7.