Seattle
Making Value for Society
122nd ASEE Annual Conference & Exposition
June 14 - 17, 2015
Seattle, WA

Paper ID #13379

Major Changes and Attrition: An Information Theoretic and Statistical Examination of Cohort Features Stratified on Major Switches

Dr. George D. Ricco, Purdue University, West Lafayette

George D. Ricco is the KEEN Program Coordinator at Gonzaga University in the School of Engineering and Applied Science. He completed his doctorate in engineering education from Purdue University's School of Engineering Education. Previously, he received a M.S. in earth and planetary sciences studying geospatial imaging and a M.S. in physics studying high-pressure, high-temperature FT-IR spectroscopy in heavy water, both from the University of California, Santa Cruz. He holds a B.S.E. in engineering physics with a concentration in electrical engineering from Case Western Reserve University. His academic interests include longitudinal analysis, visualization, semantics, team formation, gender issues, existential phenomenology, and lagomorph physiology. He lives in romantic Spokane with his leporidae partner, Rochelle Huffington Nibblesworth.

Mr. James F. Ryan III, Rensselaer Polytechnic Institute

Mr. James F. Ryan is a Ph. D. student at Rensselaer Polytechnic Institute in the Department of Mathematics. He attained a B.S. in both Mathematics and Mathematical Statistics from Purdue University and attained an M.S. in Mathematics with a focus on Mathematical Data Mining fro Tarleton State University. James' current research interests are in data analytics and mathematical techniques for data discovery and mining in myriad spaces. He has worked on case studies ranging from time series analysis of satellite data, risk analysis across shipping lanes and prescriptive analytics in the healthcare field.

Major Changes and Attrition: An Information Theoretic and Statistical Examination of Cohort Features Stratified on Major Switches

Introduction

Information Theory is a field derived from a seminal paper by Shannon[1] discussing the uncertainty extant in communication channels. We cover the details in the theory section but this paper focuses on a measure known as the mutual information. This measure, derived from Shannon's information entropy - a measure of uncertainty in a random variable - is the information gained with respect to one random variable given knowledge of another. In a sense, this measures the dependency between two random variables.

We consider this notion of mutual information as a way to measure the dependency between variables of interest in the Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) educational database among various cohorts of students. In particular we focus on how dependent the last term and last GPA are on the number of major switches done in an arbitrary pathway through a university. These measures can provide a framework to guide MIDFIELD drilldowns and data analytics —e.g. clustering using mutual information as a metric and Bayesian network analysis.

To the best of our knowledge information theory has not been fully explored in the education pathway space. We focus our literature review on theoretical papers that demonstrate the effectiveness of mutual information at supplementing analyses in other fields while drawing a bridge to ours. We then provide a theoretical foundation for our analysis before its performance and subsequent delineation.

Literature Review

One of the main utilizations of mutual information in data science is in the broad field of features selection. This involves selecting a group of the most relevant explanatory features from potentially thousands of candidate variables and the myriad combinations thereof; the aim is to provide a highly explanatory set of features without risking over-fitting the data. Work done by Fleuret[2] demonstrates this technique by using mutual information as a features selection criterion in a fast acting algorithm. Using that and a simple naïve Bayes model he is able to attain results bordering on state-of-the-art just by pruning the domain space. Given the sheer volume of potentially important variables that could describe major progression this technique demonstrates value in pruning features on any level of study.

A feature selection optimization paper by Brown and his colleagues[3] further solidifies the utility of such a measure. They demonstrate that many of the selection criterion that exist for feature selection are in fact derivative of an optimization problem; in particular, this is the maximization of the likelihood function which is related to maximizing the mutual information between sets of variables. Their paper mainly focused on dealing with this criterion for small scale data; for large scale data some of the variability issues do not exist.

Moving into some applications, Tourassi and company[4] use mutual information criterion for aiding in disease diagnosis via medical image interpretation. Along a similar biological line, work done by Steuer and his team[5] demonstrate various approaches to the estimation of mutual information. In the latter work they look at both discrete and continuous approximations of mutual information, finding that the approximations may lead to discrepancies in the results. We will discuss how this affects our study later. The work of Liu[6] applies mutual information to the education space, albeit in the sense of developing tutoring programs. Using Bayesian networks as a framework and comparing various methods of relevance metrics, he finds that the mutual information measure performs best at the task of student classification in his testing simulation.

Though we have not exactly arrived at looking at larger patterns of student movements, a clear pattern of application has emerged. A few of the papers[3-5] in the field have shown that using mutual information as an underlying measure of variable relevance vastly improves classification models. Incorporating such a framework into education analytics would provide a solid baseline for any mathematical model being considered.

Theory

Let X, Y denote two random variables. A pertinent question regarding said variables is measuring their dependence upon each other. One way to describe this measure of dependence is by exploiting the concept of information entropy. As defined by Shannon[1] the entropy of a random variable can be expressed by the following quantity:

$$H(X) = -\sum_{i} p_i \log(p_i) \tag{1}$$

Under this definition we often utilize the base two logarithm in order to provide a *dimensionless* unit of measure—though in the case of communication channels this is often prescribed as bits. $p_{i,}$ then is the probability of event i given that i is located within the event space of our random variable. Intuitively, this is a measure of the uncertainty around specifying the random variable X.

We may also define a quantity known as the conditional entropy between two random variables. Formally this can be denoted as:

$$H(Y|X) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\big(p(y|x)\big) \tag{2}$$

This quantity, much like the regular delineation of entropy, has an intuitive understanding: it is the measure of uncertainty of specifying the random variable Y given knowledge about the random variable X.

A reasonable intuition for considering the effect that conditioning on knowledge has on a random variable would be to take the difference between the entropy and conditional entropies, i.e. $H(X) - H(X|Y)$. This would be a measure of the loss of uncertainty in X given our knowledge of Y. This is shown to be equivalent to the relative entropy between the joint and product distributions of X, Y.[7] Hence, we define:

$$I(X;Y) := \sum_{x \in X} \sum_{y \in Y} p(x,y) \log(\frac{p(x,y)}{p(x)p(y)}) = H(X) - H(X|Y) \qquad (3)$$

We refer to this quantity as the *mutual information* between the random variables X, Y; if this mutual information is zero it effectively means the random variables are independent since no amount of uncertainty in one is reduced by knowledge of the other. We can hence use this as a surrogate to measure the dependence between sets of random variables, analogous to Pearson's coefficient except generalizable over all probabilistic relationships.

Data and Design

As aforementioned, we employ the MIDFIELD database for our analysis. MIDFIELD The data contain 871,741 records of first time in college (FTIC) students at eleven partner institutions, covering over 13% of the United States' engineering students. We clean the data by removing institutions that operate on a trimester schedule and excise transfer students. We cap the number of major changes in a path at 8 and the last term in the system at 23 to remove outliers. This leaves us with $N = 499{,}188$ student records to analyze.

From this data we select three cohorts to analyze: the full set of students; the subset of students who were only engineering students through their time in university ($N_E = 123{,}101$); and the students whose final group designation through their time in university was engineering ($N_L = 103{,}148$).

For each cohort we consider three subsets of interest: the entire space; the space of those students who graduated; and the space of those students who failed to graduate. Let S denote the number of major switches in a single observation. For each cohort and each subset of interest we compute the mutual information using equation (3) between S and the following two variables: the final GPA of the student (denoted G); and the final term of the student (denoted T). This will allow us to compare, for each subset, how knowledge of major switching affects the uncertainty we have in assessing the final GPA and final term of a given student. As the database has been presented such that there are equal rates of graduated and failed to graduate students in each cohort the data are comparable.

All computations were performed in the R programming environment[8]. The package "entropy"[9] was loaded for convenience in calculating the mutual information as delineated by equation (3).

Caveats

An issue that arises in computational studies using mutual information is in the estimation of the quantity itself. Since we limited the number of major switches to 8 through the data cleaning we may treat S as a categorical variable. In a similar light we may also treat T as a categorical variable. We may hence exactly compute the mutual information between the two variables using equation (3) directly. Tourassi[4] refers to this as the "histogram" method since the data may be directly binned and the random variable naturally split. An illustrative computation will be provided in the analysis section.

An issue arises when considering the continuous variable G. Steuer[5] notes that the histogram method is still often used in this setup for some selected split on the continuous variable in question but that this may encounter systematic errors in estimation, proposing the use of kernel density estimators to correct for this issue. Kraskov[10] also identifies this issue in mutual information estimation, noting that the optimal binning size for convergence is difficult. His paper suggests a correction using k-nearest neighbors to compute mutual information, noting its use in information theoretic computation before.

When considering these issues we note that, in the above papers, much of these considerations were made when the data were unbalanced and, in particular, small. As we have at least 50,000 observations in each of our sets of study we eschew utilizing a nearest neighbor or kernel density method and instead bin the last GPA until any further binning produces empty bins.

To further justify this while addressing another concern of Tourassi[4], we check the proportion of students in each cohort who switched majors a given number of times. If they are similar then the discrete approximation of the continuous grade variable is more likely to converge. Moreover we will demonstrate that our data cohort have similar proportions and are hence comparable. The results are provided in Table 1.

We note in the table that the data are somewhat balanced with respect to the number of major switches. Only engineers may have a different distribution as opposed to the other two cohorts. The zero proportions come from throwing out outlier values for major switches: these values caused errors in mutual information computation and were hence excised. As our data are large this will stabilize any discrete approximation of mutual information. We proceed to compute and compare the values noting there could be some discrepancies resulting from the only engineering cohort.

Major Switches	Full Cohort	Only Engineers	Ended as Engineers
0	5.69E-01	6.30E-01	6.88E-01
1	3.06E-01	3.08E-01	2.42E-01
2	9.44E-02	5.52E-02	5.66E-02
3	2.40E-02	6.53E-03	1.09E-02
4	5.40E-03	5.36E-04	1.82E-03
5	1.05E-03	5.69E-05	3.59E-04
6	1.74E-04	0.00E+00	6.79E-05
7	4.21E-05	0.00E+00	9.69E-06
8	1.20E-05	0.00E+00	0.00E+00

Table 1: Proportion of observations in a cohort broken up by major switches

Analysis

Let's consider a sample computation of the mutual information between S and T for the full cohort of students. We opt for our analysis to use the natural logarithm for computation; hence our information entropy units will be "nats". The choice of units is arbitrary as the mutual information is scale invariant.

Our first task is to create the bins. Since major changes and last term are both discrete the bin sizes are natural. We then construct a frequency of table giving us the number of observations that occur for a given major switch given it was their last term in the university system. Table 2 shows the discovered contingency table for this example.

Major Switches x Last Term	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
0	0.057187	0.096106	0.045452	0.043827	2.69E-02	2.59E-02	2.31E-02	6.52E-02	5.01E-02	4.62E-02	2.74E-02	1.90E-02	1.30E-02	8.76E-03	6.07E-03	4.55E-03	2.96E-03	2.48E-03	2.06E-03	1.28E-03	7.81E-04	5.45E-04	3.29E-04
1	0	0.00997	0.010914	0.020119	1.43E-02	1.64E-02	1.50E-02	3.37E-02	3.84E-02	4.17E-02	3.07E-02	2.33E-02	1.61E-02	1.15E-02	7.56E-03	4.96E-03	3.42E-03	2.87E-03	2.05E-03	1.22E-03	8.71E-04	5.19E-04	3.47E-04
2	0	0	0.000777	0.003089	3.17E-03	4.40E-03	4.32E-03	1.00E-02	1.16E-02	1.41E-02	1.17E-02	9.30E-03	6.80E-03	4.59E-03	3.27E-03	2.34E-03	1.62E-03	1.15E-03	7.79E-04	5.15E-04	3.31E-04	2.68E-04	1.54E-04
3	0	0	0	0.000196	4.41E-04	7.69E-04	1.09E-03	1.98E-03	2.76E-03	3.57E-03	3.30E-03	2.90E-03	2.05E-03	1.57E-03	1.03E-03	7.69E-04	5.35E-04	4.01E-04	2.08E-04	2.00E-04	1.08E-04	9.82E-05	6.01E-05
4	0	0	0	0	5.21E-05	6.81E-05	1.70E-04	4.01E-04	5.87E-04	8.17E-04	7.93E-04	6.49E-04	5.87E-04	3.71E-04	2.56E-04	2.04E-04	1.38E-04	1.00E-04	7.41E-05	5.01E-05	3.00E-05	3.00E-05	2.40E-05
5	0	0	0	0	0.00E+00	8.01E-06	2.20E-05	6.81E-05	8.41E-05	1.18E-04	1.34E-04	1.54E-04	1.06E-04	1.06E-04	8.01E-05	3.61E-05	3.41E-05	3.21E-05	1.80E-05	1.60E-05	2.20E-05	4.01E-06	4.01E-06
6	0	0	0	0	0.00E+00	0.00E+00	0.00E+00	4.01E-06	8.01E-06	1.20E-05	2.40E-05	2.80E-05	2.80E-05	1.20E-05	1.40E-05	1.40E-05	1.00E-05	4.01E-06	2.00E-06	6.01E-06	4.01E-06	2.00E-06	2.00E-06
7	0	0	0	0	0.00E+00	0.00E+00	0.00E+00	2.00E-06	0.00E+00	2.00E-06	8.01E-06	4.01E-06	4.01E-06	4.01E-06	4.01E-06	2.00E-06	2.00E-06	2.00E-06	2.00E-06	2.00E-06	0.00E+00	2.00E-06	2.00E-06
8	0	0	0	0	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	4.01E-06	0.00E+00	2.00E-06	2.00E-06	0.00E+00	0.00E+00	0.00E+00	2.00E-06	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.00E-06

Table 2: Frequency Table for Major Switches vs Last Term in the Full Cohort of Study

Once we have these tables we may compute the mutual information. Using equation (3) we have a natural binning for the computation of the value $I(T;S)$, the relative change in uncertainty of last term given knowledge about major switches. We find for our example that $I(T;S) = 0.111103$, suggesting that knowledge of the student's history of switching majors reduces the uncertainty in determining the last term.

We proceed in this way to find mutual information values for the cohorts and subsets described in the Data and Design section. The results are presented in Table 3.

	MI	Full Cohort	Only Engineers	Last Group Engineering
Whole Space	I(T;S)	0.111103	0.083926	0.087116
	I(G;S)	0.031648	0.031923	0.031516
Did Not Graduate	I(T;S)	0.14089	0.113506	0.102177
	I(G;S)	0.031166	0.035844	0.035115
Graduated	I(T;S)	0.028349	0.03692	0.040165
	I(G;S)	0.018705	0.019646	0.020513

Table 3: Mutual Information Values for the Three Cohorts

We compute the mutual information between last term and major switches, $I(T;S)$, and the mutual information between last gpa and major switches, $I(G;S)$, and compare across our desired cohorts. The whole spaces indicates that the entire set was analyzed with the double lines separating this from the analyses done on the graduated and failed to graduate subsets.

Discussion

The values presented in Table 3 give us a measure of the decrease of uncertainty last term and last GPA with respect to the number of times a student switched majors. In each cohort we notice that, at most, our uncertainty with respect to the last GPA decreases at most by ~0.032 nats. Compared to last term this seems to suggest that the number of major switches and the last GPA are not dependent variables. Furthermore this reduces significantly if we only consider those students who graduated. In each cohort this value is ~0.02 nats. This suggests that other factors in student progression have more influence in determining final GPA, especially for those students who graduate.

In addition, it would seem that for each cohort, the graduated subset has drastically lower values of mutual information than for any other subset. This is especially interesting when noticing that, across each cohort, the mutual information between the whole space and the subset that failed to graduate are approximately equal. This suggests that studies that utilize major switching should not that there appear to be differences in the between-variable relationships for those students that graduate and those that fail to graduate.

Specifically, we see that major changes reduce the uncertainty with regards to last term more for students who fail to graduate than for students who graduate. This reduction in uncertainty is less for the only engineering and left in engineering cohorts across the whole space. In the graduated subset, there seems to be more of a dependency between last term and major changes compared to the full cohort. More interestingly, before the split the mutual information between last term and major switches for the engineering cohorts were ~0.085 nats but jumped above 0.1 nats when only considering those that did not graduate.

Conclusions

We computed the mutual information between the last term/last GPA and the number of major switches for students in the MIDFIELD database across subsets of cohorts. We found that last GPA and number of major switches do not appear to share many dependencies in all cases while there does appear to be some relationship between last term and the number of major switches. This trend changes when splitting the cohorts between graduated and failed to graduate: it remains the same for those that fail to graduate but last term and major switches appear to be more independent for those students that graduate.

We note that further study on MIDFIELD using concepts of mutual information theory will require more robust computation for continuous variables that address the approximation issues mentioned by Steuer[5]. Furthermore, a framework for testing the significance of these values should be established for statistical completeness.

Aside from comparison, statistical robustness provides a useful vehicle for model selection in the MIDFIELD and other data spaces. Steuer notes[5] that the mutual information can also be used as a sort of sanity check for Pearson's correlation coefficient with respect to linear relationships. We may hence use this measure with Pearson's coefficient to assess non-linearity between random variables in the aforementioned data. This would codify which variables are linearly related and which require nonlinear models when looking at major progression.

Our results demonstrate that there may exist subpopulations in MIDFIELD of students who graduate and fail to graduate. These subpopulations seem to exist among all our cohorts. Our computation of mutual information demonstrates that the dependencies between the same variables change over this cohort. This in turn shows the value of mutual information: before drilling down into the data we have identified, over a uniform set of variables, the presence of subpopulations with different statistical behavior. Future work can explore this phenomenon through descriptive analytics while using the mutual information itself as a metric for its analyses.

References

1. Shannon, C., *A Mathematical Theory of Communication* The Bell Systems Technical Journal, 1948. 27.
2. Fleuret, L., *Fast Binary Feature Selection with Conditional Mutual Information.* Jurnal of Machine Learning Research, 2004. 5: p. 1531-1555.
3. Brown, G., et al., *Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection.* Journal of Machine Learning Research, 2012. 13: p. 27-66.
4. Tourassi, G., et al., *Application of the Mutual Information Criterion for Feature Selection in Computer-Aided Diagnosis.* Journal of Medical Physics 2011. 28(12): p. 2394-2402.
5. Steuer, R.S., J, et al., *The mutual information: Detecting and evaluating dependencies beween variables.* Bioinformatics 2002. 18 (Suppl. 2): p. 231-240.
6. Liu, C., *Using Mutual Information for Adaptive Item Comparison and Student Assessment.* Educational Technology and Society, 2005. 8(4): p. 100-119.
7. Cover, T. and J. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing, ed. D. Schilling. 1991, New York City: John Wiley and Sons, Inc.
8. R Core Team. *R: Language and Environment for Statistical Computing*. Vienna, Austria, 2004. http://www.R-project.com.
9. Hausser, J. and Strimmer, K. *entropy: Estimation of Entropy, Mutual Information and Related Quantities*. R package version 1.2.1, http://CRAN.R-project.org/package=entropy.
10. Kraskov, A., Stögbauer, H. and Grassberger, P. *Estimating Mutual Information*. Physical Review E, 2004. 69: 066138.