Data Mining of Undergraduate Course Evaluations

Sohail Javaad Syed, Yuheng Helen Jiang, Lukasz Golab
University of Waterloo, Canada
{sjavaad,y29jiang,lgolab}@uwaterloo.ca

1. INTRODUCTION

In this paper, we take a new look at an old problem of analyzing course evaluation data. We present an information-theoretic study to characterize courses whose ratings have high *entropy*, i.e., those which some classmates rate highly and some poorly. Our data set comes from the Engineering faculty of a large Canadian university, and, to the best of our knowledge, is an order of magnitude larger that those analyzed in previous work (see, e.g., [1, 2, 3]). After removing evaluations with fewer than 15 responses, we have 257,612 student evaluations of 5,740 undergraduate courses taught by 2,112 distinct instructors from 2003 till 2012.

Table 1 lists the 17 questions on our evaluation forms; we will refer to them by their abbreviations (e.g., Q1). Q1 through Q9 refer to teaching attributes and Q11 through Q16 refer to course attributes. Q10 and Q17 are the overall appraisals. Each question has five possible answers from A (best) to E (worst), where an A is assigned 100, B is 75, C is 50, D is 25 and E is zero. For each question, we have the frequencies of each possible answer and an average. We also have the course level, semester, and an anonymized instructor ID. Additionally, we obtained the following attributes by scraping online course calendars: class size, course type (compulsory or elective), time of lecture (we define morning classes as those which start before 10:00, day classes as those which start between 10:00 and 17:00, and evening classes as those which start after 17:00), and the number of lectures per week (one three-hour lecture, two 90-minute lectures or three one-hour lectures). Finally, we derived the following attributes for each course offering: teaching experience of the instructor (total number of times he or she taught in the past), attendance (the number of evaluations received divided by course enrolment–i.e., we assume that attendance on the evaluation day is a good indicator of average attendance throughout the course), and *specific* teaching experience (the number of times this instructor has taught this particular course).

Table 1: Questions on course evaluation form

Q1	Instructor's organization and clarity
Q2	Instructor's response to questions
Q3	Instructor's oral presentation
Q4	Instructor's visual presentation
Q5	Instructor's availability and approachability outside of class
Q6	Instructor's level of explanation
Q7	Instructor's encouragement to think independently
Q8	Instructor's attitude towards teaching
Q9	Professor-class relationship
Q10	**Overall appraisal of teaching quality**
Q11	Difficulty of concepts covered
Q12	Workload required to complete this course
Q13	Usefulness of textbooks
Q14	Contribution of assignments to understanding of concepts
Q15	How well tests reflect the course material
Q16	Value of tutorials
Q17	**Overall appraisal of the course**

Table 2: Average entropy of each question

QID	Avg	QID	Avg	QID	Avg	QID	Avg
Q1	1.39	Q2	1.49	Q3	1.18	Q4	1.5
Q5	1.43	Q6	1.29	Q7	1.63	Q8	1.25
Q9	1.3	**Q10**	**1.47**	Q11	1.57	Q12	1.5
Q13	1.95	Q14	1.72	Q15	1.66	Q16	1.92
Q17	**1.63**						

2. RESULTS

For each course evaluation, we compute the entropy of each of the 17 questions as follows. Let p_A, p_B, p_C, p_D and p_E be the relative fractions of the students who chose options A, B, C, D and E, respectively. Then the entropy is

$$-p_A \log_2 p_A - p_B \log_2 p_B - p_C \log_2 p_C - p_D \log_2 p_D - p_E \log_2 p_E.$$

Higher entropy means that there is more variability in the responses among the students in a given class.

We start by calculating the average entropy of each question, shown in Table 2. According to the t-test, Q17 has a statistically significantly higher average entropy than Q10, meaning that *classmates tend to agree more on teaching quality than overall course quality*. Of the teaching-related questions, quality of oral presentation (Q3) has the lowest entropy, which makes sense: good or bad speakers are uniformly perceived as such. Encouragement to think independently (Q7) has the highest entropy, which also makes sense since different things may make different students think. Of the course-related questions, usefulness of textbooks (Q13) and usefulness of tutorials (Q16) have the highest entropy. This is likely due to the different learning styles of different students: some learn on their own and/or from lectures, while others need a good textbook or effective tutorials. Workload (Q12) has the lowest entropy: a heavy course is perceived as heavy by the majority of students.

Table 3: Regression results

	Q10 RMSE	Q17 RMSE
Related survey attributes	0.15	0.24
All survey attributes	0.15	0.19
All survey attributes + other attributes	0.15	0.19

2.1 Predicting the Entropy of Q10 and Q17

We now turn our attention to predicting the entropy of Q10 and Q17 using linear regression. We compute the Root Mean Square Error (RMSE) of three models: First, we predict the entropy of Q10 and Q17 using only the entropy of the teaching or course-related survey attributes, respectively ("Related survey attributes"). Next, we use the entropy of all survey attributes ("All survey attributes"), followed by adding the values of other attributes we collected such as class size, instructor experience, etc. Results are shown in Table 3.

The entropy of teaching quality ratings (Q10) is explained by the entropy of the teaching-related survey questions (Q1-Q9); adding other attributes to the model does not improve the RMSE. The entropy of response to questions (Q2) and organization and clarity (Q1) have the largest regression coefficients of 0.28 and 0.27, respectively. Thus, classmates disagree on the overall teaching quality largely because they disagree on the organization and clarity of the instructor or his or her effectiveness in responding to questions.

The entropy of the overall course appraisal (Q17) can be explained by the entropy of all the survey questions, both teaching-related and course-related (using only the course-related questions has a higher RMSE, showing that teaching quality significantly influences the overall course appraisal). The entropy of usefulness of assignments (Q14) has the largest regression coefficient of 0.35, whereas the entropy of usefulness of tutorials (Q16) has the smallest coefficient of 0.01. This suggests that if classmates disagree on the overall course appraisal, they do so because some enjoy working on the assignments but others do not. On the other hand, disagreement in the rating of tutorials does not lead to disagreement in the overall rating of the course. One possible explanation is that students who do not find tutorials useful may choose not attend them, but if they like other aspects of the course, they will still rate it highly.

As for the other attributes, class size is positively correlated with the entropy of Q10, and teaching experience is slightly negatively correlated with the entropy of Q10 and Q17. Interestingly, optional courses have higher entropy of teaching quality, but lower entropy of course quality. We hypothesize that students who take an optional course are interested in the material and may rate the course uniformly well regardless of how it turns out; at the same time, some of these students may rate the instructor more highly than they normally would have, just because they liked the topic of the course, while others may rate the instructor normally. In terms of the time of lecture, evening classes have higher entropy of their appraisals. We hypothesize that some students who attend evening classes may sit in the back and do their homework instead of paying attention, and may give lower ratings; however, students who make an effort to wake up early and attend morning classes tend to pay attention and provide more consistent feedback. Finally, in terms of the course level, the entropy of the overall course appraisal is lower in first year, and then it increases significantly in the second and third years, and drops in the fourth year. The increase from first year might be because as students take more courses, they develop a better idea of what they like and do not like in a course, and as a result they express stronger opinions. The fourth-year drop is likely due to the fact that many fourth-year courses are optional, which we found to have lower course appraisal entropy.

2.2 Detailed Analysis

Our entropy analysis does not fully capture the polarity of opinions expressed by different students in the same class. For example, a course appraisal with 50 percent A's and 50 percent B's (and no other ratings) has the same entropy as an appraisal with 50 percent A's and 50 percent E's (and no other ratings). Clearly, the latter is more "controversial" as some students love it and others hate it. Motivated by this observation, we now further investigate how the responses to Q10 and Q17 are distributed over the five possible options. In general, we found that highly-rated courses have low entropy (mostly A's and perhaps a few B's) but poorly-rated courses have high entropy, meaning that they may have a non-zero number of all five possible responses. This suggests that good courses and instructors are rated highly by the majority of students, but mediocre ones may be rated highly or poorly, depending on the student.

We informally define a teaching or course appraisal (Q10 or Q17) with *no gaps* as one that has at least one of every possible option (A through E). Intuitively, courses with no gaps elicit the most variable opinions, ranging from best (A) to worst (E). Upon further investigation, we found that many courses rated between 50 and 60 contain no gaps, meaning that the average appraisal is C, but there is also at least one A, B, D and E. More surprisingly, even some courses rated as poorly as 20 have no gaps (some students liked them), as do some courses rated as highly as 80 (some students hated them)! One possible explanation for the former is that some students in bad courses may not take the evaluations seriously and they will simply choose the first answer for every question—which happens to be A—so they can complete the survey as soon as possible and leave. If true, this means that the real average appraisal of such courses is even lower than reported. For the latter, we hypothesize that even highly-rated courses may have a handful of unhappy students for various reasons.

Finally, there are no courses whose appraisals only contain A's and E's, and no other ratings in between. However, there are 13 courses whose teaching appraisals only have A's, B's and E's, and no C's and D's. The teaching quality scores of these 13 courses range from 76 to 96. Thus, these are courses that obtained mostly A and B ratings, with only a few E's. Digging deeper, we noticed that the lowest-rated questions for these courses are encouraging to think independently (Q7) and how well test reflect the course material (Q15); both of these contained many D's and E's. We hypothesize that these courses had good instructors but poorly-designed tests (or perhaps unfairly-graded tests that did not reward independent thinking); most students rated the instructor highly despite the problems with tests, but a few may have found these problems so serious that they felt the instructor deserved to be rated poorly.

3. REFERENCES

[1] B. Badur and S. Mardikyan. Analyzing teaching performance of instructors using data mining techniques. *Informatics in Education*, 10(2):245–257, 2011.

[2] K. A. Feldman. The superior college teacher from the students' view. *Research in Higher Education*, 5(3):243–288, 1976.

[3] H. W. Marsh. The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. *Applied Psychological Measurement*, 6(1):47–59, 1982.