Crowdsourcing the Creation of Quality Multiple Choice Exams

Sarah K. K. Luger
Institute for Language, Cognition and Computation
The University of Edinburgh
Edinburgh, UK
s.k.k.luger@sms.ed.ac.uk

Abstract

Automated systems that can measure the difficulty and discrimination power for Multiple Choice Questions (MCQs) have value for both educators, who spend large amounts of time creating novel questions, and students, who spend a great deal of effort both practicing for and taking tests. The current approach for measuring question difficulty relies on models of how good pupils will perform and contrasts that with their lower-performing peers. This paper covers both a method for automatically judging the difficulty and discriminating power of MCQs and how best to build sufficient exams from these good questions. The MCQ data used in these experiments was voluntarily generated by students and allows a wider discussion of this method as a version of micro-task difficulty and worker quality measurement. Future work includes comparing the initial results to those used in broader crowdsourcing tasks to measure worker quality.

Introduction

Crowdsourcing presents an alternate method from academic, institutional, and research-oriented document annotations for gathering useful, human judgment data. Extensible data sets that rely on micro-task-built data transform the way human judgment data is incorporated into solving problems facing areas as disparate as educational testing, disaster remediation, and marketing surveys. Measuring the quality of the participating workers and the difficulty of the individual tasks can be complex. This paper presents an example in educational testing and discusses further analysis using machine learning algorithms.

The use of standardized comprehension or aptitude exams requires having access to sets of exam data, which include the questions and detailed, question-by-question results from thousands of students. Unfortunately, such ideal data is very difficult, if not impossible, to obtain. The use of crowdsourced, human-annotated, or "human-in-the-loop" data has emerged as an important resource for human judgments including answering exam questions. For example, Amazon's Mechanical Turk (Amazon 2013) and the crowdsourcing company Crowdflower (Biewald and Van Pelt 2013) both provide avenues to gather human judgments on a myriad of tasks (Callison-Burch 2009). More specifically, there are other question authoring and answering environments available, including Piazza (Pooja 2013) but I have chosen PeerWise for this work because it is open source software.

To measure the usefulness of exam questions, researchers have devised methods for judging both the difficulty of the question and the differentiation power of the answer options (Patz and Junker 1999) and (Beguin and Glas 2001). One such approach is Item Analysis Theory (Gronlund 1981).

Comprehension and aptitude tests seek to present questions that can be correctly answered by students who understand the subject matter and to confuse all other students with seemingly viable alternate answer options (distractors). A *good* or difficult distractor is one that catches or distracts more bad students than good students.

A high-scoring student is one who answers most questions correctly, but when their answers are incorrect, chooses the best distractors. A low-scoring student will choose any of the answer options seemingly at random. A difficult question is one whose answer options are all deemed viable to a high-scoring student. With a difficult question, the high-scoring cohort will behave like low-scoring students, with a near equal spread of multiple distractors being chosen.

The Methodologies

An exam is a set of students who have answered the same questions. The PeerWise data sets consist of students who have answered some questions, but not necessarily the

same questions from a set. Thus, the data contains an incomplete, or sparse exam. In the question sets there are:

Course:	1	2
Total number of students:	1055	887
Total number of questions:	148	132
Shared edges between the questions and the students:	28049	31314

There are two approaches; one is to find those questions that most students answered in common (clique-based approach). I need to include the same students who have answered the same questions because we are attempting to use Item Analysis that is dependent on full exam-based results. Another approach is to see which questions are the most difficult and choose the exam questions based on rank of difficulty (the weighting-based method).

A graph-based representation is used for gathering training data from existing, web-based resources that increases access to such data and better directs the development of good questions. Then, I use a complementary method based on weighting questions by difficulty for building an exam. Further, using Item Analysis Theory, (Gronlund 1981), I analyze these new virtual exams and measure both the item difficulty and the discriminating power of the questions. Please see (Luger 2011) for a deeper explanation of the clique- and weighting-based methodologies.

Then, I use a method that efficiently builds new exams that consist of only these discriminating questions and we demonstrate the effectiveness of this new set of questions by monitoring student performance group movement across exams of different sizes. The results suggested using the most correlated 15% of students and the most correlated 25% of questions for further analysis in the new exams. Thus, in these courses 26 and 20 questions remain, and cohort movement is 44% and 46%, respectively.

Machine Learning Approaches

The initial algebraic approaches to extract the largest student-question graph that is approximately connected show positive results. Nonetheless, a comparison to state-of-the-art machine learning-based methodologies would reveal the strengths and weaknesses of the different approaches. Implementing Item Analysis depends on using a connected graph to determine student skill and question difficulty but other research in measuring worker quality and task difficulty has effectively used supervised learning techniques. Machine learning algorithms can simultaneously predict worker quality and task difficulty without requiring many workers to answer many questions. Notable contributions to measurement include using maximum likelihood estimation (Raykar, et al. 2010) and (Whitehall, et al.).

Results and Future Work

I demonstrated two sets of algorithms that identified appropriate MCQs from PeerWise and showed how these questions could be analyzed to determine both their difficulty and discrimination. The matrix-based method was presented for data analysis and then built exams out of sets of questions that have been answered by students. Discovering the maximal clique would be ideal; in this case, I only needed to find a sufficiently large clique. The weighting-based performance cohorts (composed of the 26 and 20 questions to mirror the clique results) were far less stable than the cohorts created from the clique-based method. Thus, this approach is less viable than the clique-based method of building exams. The next steps in this research focus on comparing these results to those generated from supervised learning methods on the same data.

Acknowledgments

The author would like to thank Paul Denny of PeerWise, R. Alexander Milowski, Professor Bonnie Webber, Jeff Bowles, and Professor George F. Luger.

References

Amazon. 2013. Amazon's mechanical turk. http://www.mturk.com/.

Beguin, A. A., and Glas, C. 2001. Mcmc estimation and some model-fit analysis of multidimensional irt models. In *Psychometrika, Vol. 66, No. 4, pp. 541-562*.

Biewald, L., and Van Pelt, C. 2013. Crowdflower. http://www.crowdflower.com/.

Callison-Burch, C. 2009. Fast, cheap, and creative: evaluating translation quality using amazons mechanical turk. In *EMNLP 09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Denny, P. 2009. Peerwise. http://peerwise.cs.auckland.ac.nz/.

Gronlund, N. E. 1981. *Measurement and Evaluation in Teaching*. Macmillan, 4 edition.

Luger, S. 2011. A graph theory approach for generating multiple choice exams. In *2011 AAAI Fall Symposium on Question Generation*.

Patz, R. J., and Junker, B. W. 1999. Applications and extensions of mcmc in irt: Multiple item types, missing data, and rated responses. In *Journal of Educational and Behavioral Statistics, Vol. 24, No. 4 (Winter, 1999), pp. 342-366*.

Pooja, S. 2013. Piazza. http://www.piazza.com/.

Raykar, V.C.; Yu, S.; Zhao, L.H.; Valadez, G.H.; Florin, C.; Bogoni, L.; Moy, L. 2010. Learning from crowds. *Journal of Machine Learning Research* Vol. 11 pp. 1297-1322.

Whitehall, J.; Ruvolo, P.; Wu, T.; Bergsma, J., and Movellan, J. 2009. Whose vote should count more: optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems.*