Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
AAAI Technical Report CR-13-01

Two Methods for Measuring Question Difficulty and Discrimination in Incomplete Crowdsourced Data

Sarah K. K. Luger

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

Jeff Bowles

The Department of Computer Science
The University of New Mexico
Albuquerque, New Mexico
jbowles@cs.unm.edu

Abstract

Assistance in creating high-quality exams would be welcomed by educators who do not have direct access to the proprietary data and methods used by educational testing companies. The current approach for measuring question difficulty relies on models of how good pupils will perform and contrasts that with their lower-performing peers. Inverting this process and allowing educators to test their questions before students answer them will speed up question development and utility. We cover two methods for automatically judging the difficulty and discriminating power of MCQs and how best to build sufficient exams from good questions.

Introduction

We present two methodologies for creating quality exams which filter out the least discriminating questions in an exam. The first approach analyzes the best balance of students and questions based on creating a more dense matrix of those students and questions. The second approach initially analyzes the questions' difficulty to find the best new exam set that includes the most discriminating questions. Creating viable human performance data or new, virtual exams out of sets of questions allows downstream analysis of the difficulty of the questions that would not be otherwise available.

Crowdsourcing MCQs

We have procured data for sets of MCQs from University-level introductory biology classes using the PeerWise question creation system (Denny 2009). PeerWise is a free, web-based, question repository that allows classes to set up shared environments where students create questions that are subsequently used by their peers as a study aid. Instructors can review the questions or use some of the better questions for future exams. Because answering these questions is not compulsory, the resulting data can be sparse, with sets of questions that have been answered by some students, but not all of the questions have been answered by the same students. There are other similar datasets in the education domain (Pooja 2013) and many crowdsourcing tasks have multiple-choice elements (Munro, Erle, and Schnoebelen 2013).

Item Analysis

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 (Béguin and Glas 2001). One such approach is Item Analysis Theory (Gronlund 1981). Once a cohort (for this example, 100 students) has taken a test containing suitable questions, the exams are graded and ranked from highest score to lowest. The set of 100 students is split into three groups that represent the top-scoring, middle-scoring, and lowest-scoring students. These three groups are commonly split, lower 27%-middle 46%-upper 27%. The middle set of (46) exams is excluded because they confer no meaningful information. Item Difficulty is measured by the percentage of students who answered a question correctly. Item Discriminating Power is the difference between the number of high-scoring students versus the number of low-scoring students who chose the same answer option.

Our Clique-based Methodology

Our approach for representing the individual student question answering relationship is with a graph: an 'exam', where every student answers every question would be a complete bipartite graph (or biclique). We are seeking a good set that is similar to an exam.

To sort and build the covariance matrices we collect the data in triples of student ID, question ID, and answer choice. Then, the students are ordered by the number of questions they answered and the incidence matrix M is built with students listed as rows and the questions as columns. We compute $S = M \times M^{\mathrm{T}}$ and compute $Q = M^{\mathrm{T}} \times M$. We can find the most correlated students by computing the vector $\mathbf{s}$ by summing over the rows of S. Thus $\mathbf{s} = \sum_i S_{ij}$. We can then sort the rows and columns of S based on the ordering of $\mathbf{s}$ as S is symmetric. As above, we can find the most correlated questions by computing the vector $\mathbf{q} = \sum_i Q_{ij}$. We can then sort the rows and columns of Q based on the ordering of $\mathbf{q}$.

This sorting process provides a sound heuristic for selecting highly correlated students and questions. We then se-

Method:	Clique	Clique	W	W
Course:	1	2	1	2
Total no. S:	1055	887	886	807
Total no. Q:	148	132	148	132
% top correlated S:	0.15	0.15	N/A	N/A
% top correlated Q:	0.25	0.25	N/A	N/A
Omissions:	YES	YES	N/A	N/A
New no. S:	158	133	886	807
Initial exam size:	37	32	148	132
New exam size:	26	20	26	20
Cohort movement:				
Low to middle:	4	5	43	25
Low to high:	0	0	29	24
Middle to low:	4	5	125	121
Middle to high:	5	8	70	72
High to low:	0	0	22	27
High to middle:	5	8	105	106
Numerical total:	18	26	393	375
% total:	0.11	0.2	0.44	0.46

Figure 1: The weighted (W) method eliminated students who answered fewer than 4 questions.

lected the top 15% most correlated students and the top 15% most correlated questions from the dense group of students who have answered the same questions based on the adjacency methodology. This presents a realistic exam where there are a few omitted questions.

Our Weighting-based Methodology

A "weight" vector, $\mathbf{w}$ is created where each element of the vector is the weight for a question. The questions are weighted based on the number of times a question was answered correctly. Weights are normalized, or in the range $[0, 1]$. A question with weight 0 is a question that was never answered correctly by any student, and a weight of 1 is given to a question that was always answered correctly. Calculating components of the weight vector are:

$$w(x) = \frac{\sum_{i=0}^{n} c(i, x)}{n(x)}$$

Where x is the position in the vector $\mathbf{w}$, $n(x)$ is the number of answers to question x, $c(i, x)$ is the correctness of student i's answer to question x. Values for $c(i, x)$ are 1 if the answer is correct, 0 if wrong. Weights are in the range [0,1] where weights closer to 0 correspond to very difficult questions and weights closer to 1 correspond to very easy questions. The goal is to find the middle band of discriminating questions. In general, the questions are of moderate to easy difficulty. A few of the hard questions were answered correctly by about 1 in 5 students, but the majority of questions were answered correctly by more than 1 in 2 students.

Results

When questions with low and high weights were removed from the list to find exam sizes that were the same as the clique-based methodology, we found that 44% and 46% of the students were scored so significantly differently that they would be moved into different cohorts. As a comparison, in the clique-based method only 11% and 20% of the students moved into a different cohort. This indicates that performing analysis based on question weights is not an attractive method for finding the most discriminating questions. Question weighting was viewed as a simple alternate method for finding the most discriminating questions, but it appears that this analysis does not take into account enough contextual data to discover the most discriminating questions.

Conclusion and Future Research

We have addressed the difficult and resource-expensive task of making MCQ exams by considering crowdsourced, incomplete data. We demonstrated two sets of algorithms that identified appropriate MCQs and analyzed them to determine both their difficulty and discrimination. There are multiple methods for analyzing data in sparse matrices. We look forward to using additional machine learning approaches to dealing with sparse data and including supplementary data sets.

Finding algorithms that allow meaningful analysis with incomplete data sets have great benefit in crowdsourced data collection. Many problems that do not originally appear to be MCQs behave like them, such as recent efforts to crowdsource relief and recovery efforts after Hurricane Sandy (Munro, Erle, and Schnoebelen 2013). Ranking photos of the Sandy devastation could be viewed as a 3-distractor MCQ.

References

Béguin, 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*.

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

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

Munro, R.; Erle, S. E.; and Schnoebelen, T. 2013. 10th international conference on information systems for crisis response and management. baden baden, germany. In *Analysis After Action Report for the Crowdsourced Aerial Imagery Assessment Following Hurricane Sandy*.

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.