Tracking Performance and Forming Study Groups for Prep Courses Using Probabilistic Graphical Models

(Extended Abstract)

Yoram Bachrach
yobach@microsoft.com
Microsoft Research

Yoad Lewenberg
yoadlew@cs.huji.ac.il
Hebrew Univ. of Jerusalem

Jeffrey S. Rosenschein
jeff@cs.huji.ac.il
Hebrew Univ. of Jerusalem

Yair Zick
yairzick@cs.cmu.edu
Carnegie Mellon Univ.

ABSTRACT

Efficient tracking of class performance across topics is an important aspect of classroom teaching; this is especially true for psychometric general intelligence exams, which test a varied range of abilities. We develop a framework that uncovers a hidden thematic structure underlying student responses to a large pool of questions, using a probabilistic graphical model.

General Terms

Experimentation, Theory

Keywords

Probabilistic Graphical Models, Bayesian PCA, Education

1. INTRODUCTION

When teaching a class, students often exhibit varied abilities across topics; this is especially evident when a curriculum is diverse. Preparatory courses for psychometric entrance exams such as the SAT or GMAT tutor students on various topics including mathematics, writing skills, and logical puzzles. Naturally, student proficiency across topics is not uniform: a student may be a math whiz, but have poor writing skills. Prep courses try to ensure that students perform well on all topics. This requires the teachers to shift focus based on class performance: if students are doing poorly in logical puzzles, additional practice on this sub-topic is required.

Our Contribution: We develop a framework for *course personalization* that uncovers a hidden thematic structure underlying students' responses to a large pool of questions. Our framework identifies abstract skills required to correctly answer the different questions in the pool.

Further, our model provides educators a way to keep track of students' overall course performance and offer personalized assistance on the course and individual student level. Using a form of Bayesian Principal Component Analysis, akin to other Bayesian matrix factorization methods [4], we recover latent abilities that explain student performance. We show that these latent abilities offer a *meaningful interpretation* of the data, as they match the actual sub-topic divisions in the case of the verbal and English sub-topics.

A key application of our framework is a method for dividing students into *study groups*, consisting of students with similar difficulties in the different skills tested in the exam. Our system automatically identifies the different skills tested in the exam, and the students' proficiency levels in these areas. We use this information to partition the students into student sets with homogeneous proficiencies in topics. We show that our latent ability-based clustering method has better performance than a baseline clustering method based directly on the observed responses. Our method for clustering students results in study groups that are more homogeneous, the student in every group have roughly the same skill level in each topic, and share the same difficulties.

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.*

2. SLAB MODEL

We are interested in identifying an underlying reason for students' performance; to do so, we introduce a model of Student Latent ABilities (SLAB, for short). For a fixed dimension K, the model assumes there are K latent abilities tested by a topic (K is assumed to be much smaller than the number of students and questions). Therefore every student s and question q are represented by a K-dimensional latent vector $\vec{x}_s \in \mathbb{R}^K$ and $\vec{y}_q \in \mathbb{R}^K$, respectively, where $(\vec{x}_s)_i$ is student s's proficiency in ability i and $(\vec{y}_q)_i$ is the level that q requires in ability i.

For every student s and question q we denote by $X_{s,q}$ the random variable that represents the performance of s on question q; $C_{s,q}$ is the Boolean random variable that equals "True" if the student s gave the correct answer to the question q, and "False" otherwise. We assume $X_{s,q}$ is normally distributed with mean $\vec{x}_s^T \vec{y}_q + b$ and standard deviation β, where b is the student and question bias. We assume that $\Pr[C_{s,q} = \text{True}] = \Pr[X_{s,q} > 0]$. This model resembles other Bayesian models for matrix factorization (e.g., [1, 4]).

SLAB's input is a set of student responses to questions (correct\incorrect). Based on the observed data, and using the Infer.NET [2] framework for graphical models and the Expectation Propagation algorithm [3], the model infers the values of the vector of each student and question.

	st_1	st_2	st_3
st_1	**59 (20)**	3 (36)	33 (39)
st_2	4 (14)	**76 (47)**	41 (60)
st_3	1 (2)	18 (14)	**29 (32)**

(a) English

	st_1	st_2	st_3	st_4
st_1	**48 (34)**	1 (7)	8 (18)	11 (8)
st_2	0 (24)	**20(0)**	13 (7)	0 (2)
st_3	26 32	10 6	**51 (35)**	1 (15)
st_4	15 23	0 2	2 23	**47 (16)**

(b) Verbal

	st_1	st_2	st_3	st_4
st_1	**15 (21)**	7 (5)	29 (4)	39 (24)
st_2	14 18	**10 (7)**	32 (41)	27 (17)
st_3	7 (22)	10 (4)	**47 (48)**	28 (18)
st_4	6 (13)	13 (4)	12 (34)	**44 (24)**

(c) Quantitative

Table 1: Confusion matrices of subtopic (st) labels for all topics, based on the latent-abilities vectors (response vectors)

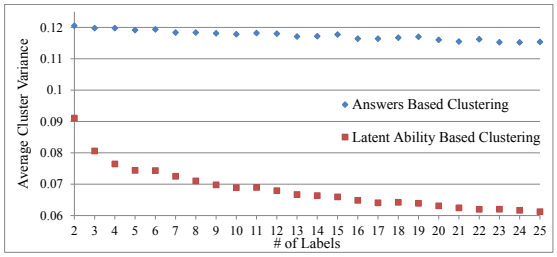

Figure 1: The average cluster variance for answers and latent abilities clustering

3. RESULTS

The Data: Our dataset consists of student responses to mock Israeli Psychometric Entrance Tests (PET), provided by a private company that administers PET prep courses. Questions are labeled according to topic (English, Verbal and Quantitative) and subtopic, as well as the correct answers to the questions. The database had 345 Verbal questions, 340 Quantitative questions, and 352 English questions. A total of 5874 students took the exam.

Eliciting Latent Problem-Solving Abilities: Using the SLAB model, we are able to use the latent ability vectors to recover the sub-topic divisions with relatively high accuracy. To demonstrate this, we do the following: each question q is associated with a vector of latent abilities $\vec{y}_q$. We cluster the questions to a number of clusters equal to the number of subtopics using k-means; now, each question has a label. To assess how well these labels match the actual divisions of the questions into sub-topics, we produce a *confusion matrix* for each subtopic. A confusion matrix C is a $k \times k$ matrix, where the C_{ij} are the number of times that a question of type j was labeled as a question of type i. A confusion matrix with a high trace is one which indicates that the prediction was accurate. Using k means clustering puts us at a disadvantage; performing k-means clustering may output a labeling that will perfectly match the original labels *up to a permutation*. In other words, it may assign the label 3 to all questions from topic 1, the label 2 to all questions from topic 3 and the label 1 to all questions from topic 2. To overcome this difficulty, once a latent-ability based labeling is produced, we permute the labels until we find a permutation that offers the highest trace for the confusion matrix. To benchmark our clustering method, we compare it to response-based clustering. The resulting confusion matrices are presented in Tables 1a, 1b and 1c. The tables show that clustering by latent abilities recovers questions well in the English and Verbal topics, but does not do well for the quantitative section. Nevertheless, in every topic latent-abilities clustering outperformed response-based clustering.

Focus Groups via Latent Abilities: The latent abilities elicited by our model can be effectively used in order to group students into focus groups. Grouping similar students is important: one would want to have students facing similar difficulties working together, in order to make effective use of time. In our dataset, we know the type of each question — the questions in each section have a distinct, known type — thus we can cluster students based on their competence in each sub-topic. However, this is not always the case; questions do not always belong to a specific type, which would pose a challenge for educators who wish to offer personalized tutoring to their students.

Our approach is simple: for each student s, we infer a vector of latent abilities, $\vec{x}_s$; next, we cluster the student body based on the student ability vectors. To assess the effectiveness of our latent ability based clustering, we measure the mean-squared variance in student subtopic proficiency within each cluster. If our methodology works, then all students from the same cluster should have similar capabilities in terms of their abilities to correctly answer questions from various sub-topics. We benchmark our latent ability clustering method by comparing it to response-based clustering. Our results are summarized in Figure 1. This is an encouraging finding; it further establishes that different latent abilities are needed to answer questions in each sub-topic; furthermore, ability-based clustering presents educators with an effective method of grouping students when there is no clear sub-topic division.

4. CONCLUSIONS AND FUTURE WORK

We have utilized probabilistic graphical models to elicit course metrics. Our framework can be extended to settings where there are no preset correct answers; an immediate use is in guiding experts towards tasks they do better, by assigning them problems at which they show higher proficiency. By eliciting latent abilities required by the tasks and those of the workers, we can better match tasks to experts. Our framework can also be used to make predictions about future student success.

Acknowledgements

This research has been partly funded by Microsoft Research through its PhD Scholarship Program, and by Israel Science Foundation grant #1227/12.

REFERENCES

[1] D. Agarwal and B.-C. Chen. flda: matrix factorization through latent dirichlet allocation. In *Proceedings of WSDA 2010*, pages 91–100. ACM, 2010.

[2] T. Minka, J. Winn, J. Guiver, S. Webster, Y. Zaykov, B. Yangel, A. Spengler, and J. Bronskill. Infer.NET 2.6, 2014. Microsoft Research Cambridge.

[3] T. P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of UAI 2001*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

[4] D. H. Stern, R. Herbrich, and T. Graepel. Matchbox: large scale online bayesian recommendations. In *Proceedings of WWW 2009*, pages 111–120. ACM, 2009.