Estimating Prerequisite Structure From Noisy Data

EMMA BRUNSKILL, University of California, Berkeley

There has been a long-standing interest in how to estimate the prerequisite structure among a set of skills. This prerequisite structure is important for intelligent tutoring systems and the educational data mining community, because it has important implications for student modeling, and for automatically constructing teaching policies. Here we present preliminary work towards inferring skill prerequisite structure given a set of noisy observations of student knowledge. We compare models using likelihood calculations and provide experimental results on a set of Algebra skills.

1. INTRODUCTION

At least since the work of Robert Gagné and his colleagues [1974], there has been interest in estimating the prerequisite structure among items to be learned. In this paper we will use prerequisite structure to refer to relationships among skills that provide strict constraints on the order in which these skills can be acquired.

There are two primary reasons why inferring the prerequisite structure amongst skills is important for the intelligent tutoring systems and educational data mining community. First, such structure is relevant for student modeling. In this paper we follow the popular Bayesian approach (see e.g. Corbett and Anderson [1995] and Conati et al. [2002]) to modeling student learning. Though there are a number of differences between the specific models and approaches used, broadly Bayesian models of learning involve considering the student's knowledge as a hidden state that changes stochastically as pedagogical activities are provided. Information about this hidden state may also be gleaned from the student responses to the provided activities, such as correct or incorrect quiz answers. If there is prerequisite structure among the set of skills a student is trying to learn, that provides direct constraints on the dynamics of learning, since a student will not be able to learn a new skill if its prerequisite is not yet mastered. In addition, if the student correctly answers a question involving a particular skill j, that not only provides positive evidence that the student has mastered skill j, but also positive evidence that the student has mastered all of skill j's prerequisite skills.

Second, the skill prerequisite structure affects what constitutes a good sequence of teaching activities. Providing a student with exercises and activities on solving a linear equation will be superfluous if the student has not yet understood fractions, or other key prerequisite skills needed in order to master solving linear equations. In addition to affecting the quality of different conditional teaching strategies, recent work by Brunskill and Russell [2010] has demonstrated that prerequisite structure can be leveraged to significantly reduce the computational burden of calculating good teaching strategies in simulated teaching domains.

However, despite the importance of prerequisite structure, there are a number of challenges to estimating this structure within the context of a particular curriculum. In particular, in this paper we focus on the challenge of estimating this structure given access only to noisy observations of students' hidden knowledge. This is the most natural setting for a number of common scenarios, such as inferring the prerequisite structure among a set of skills on an exam. Equally importantly, we suggest that such data is likely to be more easily obtainable, and therefore a candidate for automatic mining and structure extraction, compared to more intensive data collection methods such as talk aloud methods, which may be able to get a better estimate of a student's true knowledge state.

As a preliminary step towards determining prerequisite structure from noisy observations of student data, in this paper we estimate the probability of the observations under alternate possible prerequisite structures. In the rest of this short paper we will briefly discuss related work, our method, and some preliminary results, before finishing with a discussion of future work.

2. RELATED WORK

Interest in inferring prerequisite structure amongst knowledge items has been present in the education community for decades. Gagné coined the term "learning hierarchy" to describe the prerequisite relationship among skills, and developed methods for extracting learning hierarchies from student data. Following this work, there were a number of studies in the education literature on estimating prerequisite structure in mathematics curriculums, including work by Miller and Phillips [1974], Uprichard and Phillips [1977], and Close and Murtagh [1986].

More recently there has been interest from the user modeling and educational data mining community. Desmarais, Maluf and Liu [1996] introduced a method for calculating partially ordered knowledge structures (POKS), and Pavlik, Cen, Wu and Koedinger [2008] have recently provided a hierarchical agglomerative clustering method for student tutoring data which computes and leverages POKS as an input to this clustering.

However, to our knowledge, none of these studies explicitly modeled the observation generation process that might distort measurements of student knowledge and affect the inferred prerequisite structure.

3. METHOD

We now discuss our preliminary data analysis towards inferring prerequisite structure (if any) among a set of N skills. The data is assumed to be a set of binary observations that indicate if a particular student got each skill correct or incorrect, collected over a set of students. For N skills there are an enormous number of potential prerequisite structures. Therefore, to start we focus on considering each pair of skills separately, and evaluate the evidence as to whether a prerequisite relation exists among the skill pair. In this initial analysis we assume all data comes from a single evaluation setting, such as a test, and focus on comparing the likelihood of the observed data under different precondition structures. We will consider alternative methods in the discussion section.

More precisely, we are interested in whether there exists a precondition relationship among each pair of skills s_i and s_j. We will consider two simple models: in the *precondition* model s_i must be mastered before s_j can be mastered, and in the *flat* model the two skills are independent. We will compare the two models by evaluating how well they model the observed student data, under the best fit of each model's parameters. The observed data will be denoted as $z_{ik} = (0, 1)$ to represent whether student k got skill i incorrect or correct: z_{jk} denotes the observation for skill j.

We will shortly describe the different parameters employed in the two models, but first recall that we presume that a student's mastery of a skill is not directly observed. Rather, following the knowledge tracing work of Corbett and Anderson [1995], we use a p_s slip parameter to model the probability of a student answering a question about a skill incorrectly given the student has mastered the skill, and a p_g guess parameter to model the probability of a student answering correctly even when the student has not mastered the skill in question.

In the flat model, there are two independent parameters, $p(s_i = 1)$ and $p(s_j = 1)$ which represent the probability of a student having mastered skill i and skill j respectively. These are unknown parameters. We also do not know whether or not each individual student k mastered or did not master skill i or skill j. Since our interest is in coming the likelihood of the observations z_{ik} and z_{jk} $forallk$, our aim is to compute the expected marginal likelihood of the observed data $z_{ik}z_{jk}$ for all students k, under the best possible fit of the hidden parameters $p(s_i = 1)$ and $p(s_j = 1)$, by summing over whether each student mastered or did not master each skill. We estimate this using Expectation Maximization. This involves first estimating of the probability of whether or not each student mastered or did not master skill i, given whether or not they got skill i correct. For example,

$$\begin{aligned} p(s_{ik}=1|z_{ik}=1) &= \frac{p(z_{ik}=1|s_{ik}=1)p(s_i=1)}{p(z_{ik}=1|s_{ik}=1)p(s_i=1)+p(z_{ik}=1|s_{ik}=0)(1-p(s_i=1)} \\ &= \frac{(1-p_s)p(s_i=1)}{(1-p_s)p(s_i=1)+p_g(1-p(s_i=1))}, \end{aligned} \quad (1)$$

and similar expressions exist for $p(s_{ik}=1|z_{ik}=0)$,$p(s_{ik}=0|z_{ik}=1)$ and $p(s_{ik}=0|z_{ik}=0)$. Note that these probabilities can be evaluated independently for skills i and j. These probabilities allow us to compute the expected value of the log likelihood. We then calculate the values of the two model parameters $p(s_i=1)$ and $p(s_j=1)$ that maximize this expected log likelihood and repeat.

In the precondition model we assume skill i must be mastered before skill j. Here only 3 possible student states are allowed:

—$s_i s_j=00$: both skills are not mastered
—$s_i s_j=10$: skill i is mastered but skill j is not mastered
—$s_i s_j=11$: both skills are mastered.

As the probability of these states must sum to 1, this model also only has two free parameters. We again wish to evaluate how well this model explains the observed data. Now the two skills are dependent so we must consider both when estimating a student's hidden skill mastery. For example, we can compute the probability that a student mastered both skills, given he got both skills correct, as

$$\begin{aligned} p(s_{ik}s_{jk}=11|z_{ik}z_{jk}=11) &= \frac{p(z_{ik}z_{jk}=11|s_is_j=11)p(s_is_j=11)}{p(z_{ik}z_{jk}=11|s_is_j=11)+p(z_{ik}z_{jk}=11|s_is_j=10)+p(z_{ik}z_{jk}=11|s_is_j=00)} \\ &= \frac{(1-p_s)^2p(s_is_j=11)}{(1-p_s)^2p(s_is_j=11)+(1-p_s)p_gp(s_is_j=10)+p_g^2p(s_is_j=00)}. \end{aligned} \quad (2)$$

We again estimated the model parameters (here, $p(s_is_j=00)$ and $p(s_is_j=10)$) using Expectation Maximization.

Our primary interest is to establish whether one model significantly better explains the observed data. We computed the expected value of the log likelihood of each model with respect to the conditional distribution of the hidden student mastery values given the estimated model parameters. As both models have the same number of parameters, we identified skill pairs i,j where the precondition model had a higher log likelihood than the flat model in one direction only (aka only for i,j or j,i but not both).

4. EXPERIMENT & RESULTS

As part of a larger study, we collected test data from 113 ninth grade students over a set of 23 linear inequality skills. Test questions were designed to include particular skills so there existed a specific mapping between each test item and skill. Each instance of each test skill was graded as correct or incorrect. As our primary interest was in estimating prerequisite structure, and not on estimating the domain parameters, we fixed $p_s=0.3$ and $p_g=0.2$ for all skills, and both model structures, therefore removing the issue of observation modeling. [1].

Figure 1 shows the resulting prerequisite structure induced for this particular domain. Several skills did not appear in the precondition structure. This structure should be interpreted with caution, as this analysis

[1] Those particular parameter values were chosen based on expert consultation.

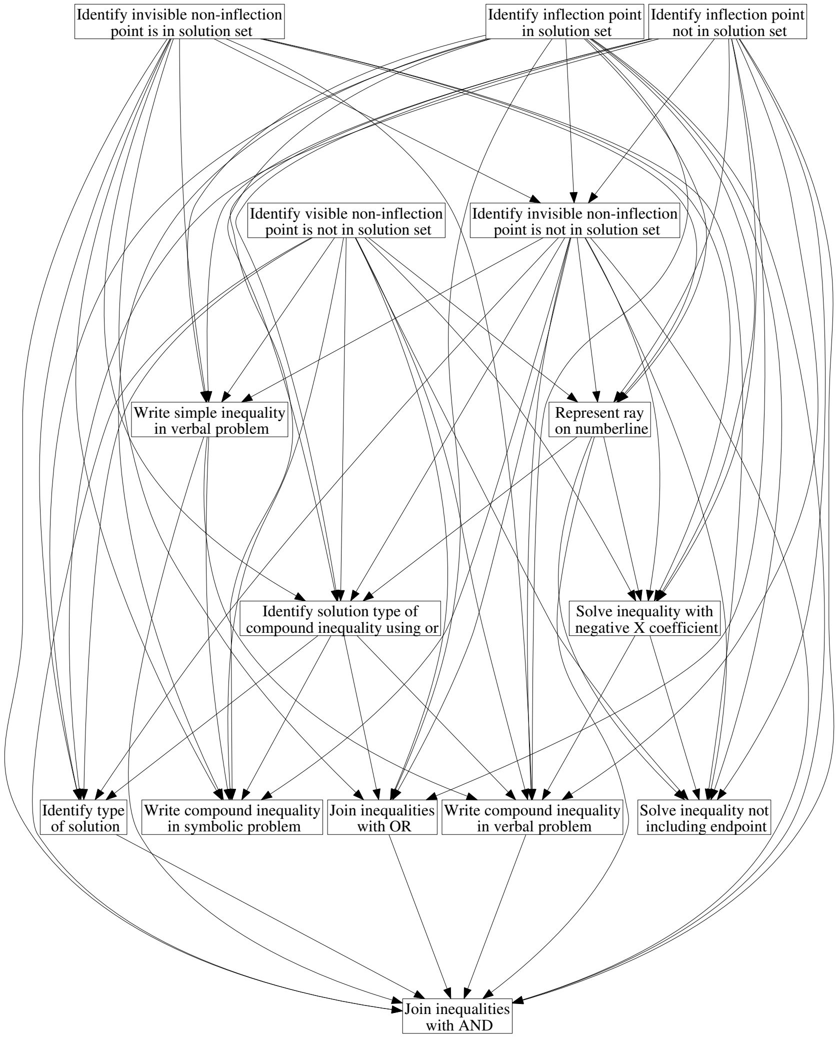
Identify invisible non-inflection point is in solution set
Identify inflection point in solution set
Identify inflection point not in solution set
Identify visible non-inflection point is not in solution set
Identify invisible non-inflection point is not in solution set
Write simple inequality in verbal problem
Represent ray on numberline
Identify solution type of compound inequality using or
Solve inequality with negative X coefficient
Identify type of solution
Write compound inequality in symbolic problem
Join inequalities with OR
Write compound inequality in verbal problem
Solve inequality not including endpoint
Join inequalities with AND

Fig. 1. Precondition structure inferred for linear inequalities curriculum. An arrow from skill i to skill j implies that i is a precondition of j.

is only preliminary and we will discuss limitations and further extensions shortly. However, some of the inferred preconditions were intuitively plausible, such as the skill of being able to write a simple inequality in a verbal problem being a precondition to the skill of writing a compound inequality in a verbal problem.

5. DISCUSSION AND FUTURE WORK

We have just described preliminary work towards inferring the prerequisite structure from noisy observations of student knowledge. In this paper we evaluated the likelihood of the observed student performance on a set of test items given different model structures. While test data is abundant, and may be possible to obtain in settings where temporal data on student learning is not available, test data has the significant limitation that it may distort the inferred precondition structure. Ultimately we are primarily interested in the dynamics of instruction. Precondition structure could help improve automated instruction by informing whether it will be possible to successfully teach skill j if skill i has never been taught. To infer such instructional precondition structure will require analyzing temporal data of student learning. Extracting precondition structure solely from static test data can be misleading, since it both may mask a precondition between skill i and skill j (if both skills have been successfully taught, student performance on both will be very good) as well as create false preconditions (if skill i has been taught extensively, but skill j has only been recently taught, it may appear that skill i is a prerequisite for skill j, even if the two skills are independent or could have been taught in reverse order). Going forward we intend to develop approaches to computing precondition structure based on temporal data.

There are a number of other interesting issues we would like to consider. In our described approach, we assumed the observation parameters p_s and p_g were provided, and were tied to be the same across all skills. These parameters could be learned and allowed to vary amongst skill, or even problem type: for example, the probability of guessing on a multiple choice question is different than a write in answer. Another interesting issue is whether prior information about each student could be used to influence the parameters specifying the probability of a student having mastered or not mastered each skill. There exist intelligent tutoring systems that maintain a Bayesian estimate over the student knowledge state that could be used to provide such a prior. However, one limitation of leveraging such estimates would be that the Bayesian models employed by such systems make their own assumptions about the prerequisite structure, and so it might not be possible to estimate all the parameters desired. It would also be useful to compute full structures, rather than pairwise models. We are also interested in comparing to other previous approaches to inferring structure. In doing so we would need to identify an appropriate metric for comparison, which could include evaluating how well the models explain the data, or perhaps how well the models predict student performance.

Ultimately it may be impossible to infer a single correct structure. In the longer term, it is interesting to think about how one could compute teaching policies that could take as input a distribution of possible precondition structures, and select policies that were robust to uncertainty over this set of structures.

ACKNOWLEDGMENTS

This research has been supported in part by NSF grant SBE-0836012 to the Pittsburgh Science of Learning Center, and a NSF Mathematical Sciences Postdoctoral Fellowship. We thank Gail Kusbit and Carnegie Learning for help with data collection, and Stuart Russell's research group for useful discussions.

REFERENCES

Brunskill, E. and Russell, S. 2010. RAPID: A reachable anytime planner for imprecisely-sensed domains. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Close, J. and Murtagh, F. 1986. An analysis of the relationships among computation-related skills using a hierarchical-clustering technique. *Journal for Research in Mathematics Education 17,* 2, 112–129.

Conati, C., Gertner, A., and VanLehn, K. 2002. Using Bayesian networks to manage uncertainty in student modeling. *Journal of User Modeling and User-Adapted Interaction 12*, 371–417.

Corbett, A. and Anderson, J. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction 4*, 253–278.

Desmarais, M., Maluf, A., and Liu, J. 1996. User-expertise modeling with empirically derived probabilistic implication networks. *User Modeling and User-Adapted Interaction 5*, 283–315.

Gagné, R. and Briggs, L. 1974. *Principles of Instructional Design.* Holt, Rinehard, and Winston.

Miller, P. and Phillips, E. R. 1974. Developed of a learning hierarchy for the computational skills of fractional number subtraction. In *American Educational Research Association Meeting.*

Pavlik Jr., P., Cen, H., Wu, L., and Koedinger, K. 2008. Using item-type performance covariance to improve the skill model of an existing tutor. In *Proceedings of the 1st International Conference on Educational Data Mining (EDM).*

Uprichard, A. E. and Phillips, E. 1977. An intraconcept analysis of rational number addition: A validation study. *Journal for Research in Mathematics Education 8,* 1, 7–16.