Item-based Bayesian Student Models

Michel C. Desmarais, Michel Gagnon, Peyman Meshkinfam
École Polytechnique de Montréal

Abstract

Many intelligent educational systems require a component that represents and assesses the knowledge state and the skills of the student. We review how student models can be induced from data and how the skills assessment can be conducted. We show that by relying on graph models with observable nodes, learned student models can be built from small data sets with standard Bayesian Network techniques and Naïve Bayesian models. We also show how to feed a concept assessment model from a learned observable nodes model. Different experiments are reported to evaluate the ability of the models to predict item outcome and concept mastery.

Introduction

The advantages of using data mining and automated learning techniques in educational systems are compelling. When data is available, they can waive the efforts required by the domain and modeling expert for building student models. The elimination of this human effort brings a large number of benefits beyond the efficiency and economic issues. It also implies greater reliability by removing the subjectivity and variability induced by a human intervention in the modeling process. Although it does not necessarily imply improved accuracy, automated model learning does imply accrued model *predictability*, in the sense that confidence intervals can be derived from the data sample and the probability of making a wrong decision can be assessed. This factor is very important to avoid the loss of user confidence and even the rejection of the system that is often observed when the user gets frustrated by too many wrong decisions on the part of the system (Horvitz 1999). Having a user model that yields a measure of parameter confidence allows the system to refrain from taking initiatives that can be ill advised and frustrating for the user given their perceived utility.

Of course, these advantages vanish if the amount of data required to build the student model outweighs the benefits. The sensitivity of the model to the data set size is thus critical and cannot be ignored.

We review some work to build student models from small data samples and the means to assess the student skills with such models. We propose an approach based on *item to item* structures and assess the predictive performance of different models with this approach. We also show how to use this approach to assess concepts by reusing an existing Bayesian Network that models relations at the concept mastery level.

Representing student cognitive state

Not every student model lends itself to learning from data. A large number of approaches have been proposed to represent the student knowledge and skills, but only a subset is amenable to learning. We will focus on Bayesian graphical models of student proficiency. These models are among the most commonly used and allow a great level of flexibility (Mislevy *et al.* 1999), and they lend themselves to learning.

Graphical student models are generally organized as a hierarchy of concepts with observable nodes, namely test items, as leaves of this hierarchy. Figure 1 illustrates a hypothetical example of such network. The "non observable" nodes are concepts, skills, and misconceptions. They are considered hidden nodes in the sense that they cannot be directly observed. However, because hierarchical student models can contain a large number of hidden nodes, their structure generally cannot readily be learned from data without some human intervention (Vomlel 2004, for eg.).

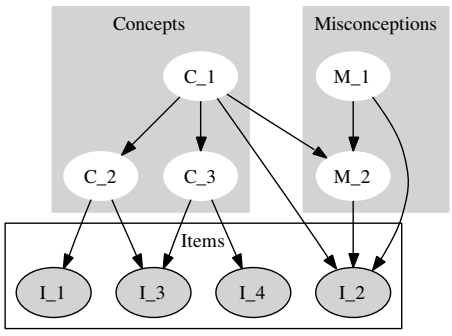

Figure 1: Example of a BN structure with items, concepts, and misconceptions.

One family of models departs from the hierarchical approach by building links among observable item nodes

themselves, bypassing concept links (Dowling & Hockemeyer 2001; Kambouri *et al.* 1994; Desmarais, Maluf, & Liu 1996). They emerge from the work of Falmagne *et al.* (1990) and Doignon & Falmagne (1999) on the theory of knowledge spaces. Our own work on Partial Order Knowledge Structures (POKS) (Desmarais, Maluf, & Liu 1996; Desmarais & Pu 2005) falls under this line of research as well. We refer to this type of student models as *item to item structures*.

Item to item structures are good candidates for learned student models because their nodes are observable, in contrast to concept nodes. We briefly review the theory behind item to item structures and report some results on learned student models with such an approach.

Knowledge Spaces

Item to item structures are based on a cognitive modeling theory named knowledge spaces (Doignon & Falmagne 1999). The theory of knowledge spaces asserts that knowledge items—observable elements that define a knowledge state such as question items—are mastered in a constrained order. In the knowledge space theory, a student's knowledge state is simply a subset of items that are mastered by an individual and the knowledge space determines which other states the person can move to. Viewed differently, the knowledge space defines the structure of prerequisites among knowledge items. For example, we learn to solve Figure 2's problems in an order that complies with the inverse of the arrow directions. It follows from this structure that if one masters knowledge item (c), it is likely she will also master item (d). Conversely, if she fails item (c), she will likely fail item (a). However, item (c) does not significantly inform us about item (b). This structure defines the following possible knowledge states (subsets of the set $\{a, b, c, d\}$):

$$\{\emptyset, \{d\}, \{c, d\}, \{b, d\}, \{b, c, d\}, \{a, b, c, d\}\}$$

Other knowledge states are deemed impossible (or *unlikely* in a probabilistic framework).

Formally, it can be shown that if the space of individual knowledge states is closed under *union*, then that knowledge space—the set of all possible knowledge states—can be represented by an AND/OR graph (Doignon & Falmagne 1999). In other words, if we combine two individuals' knowledge states, then that combined knowledge state is also plausible (i.e. part of the knowledge space). However, knowledge spaces are not closed under *intersection*, meaning that if we take the common knowledge items between two individuals' knowledge states, then we can obtain an invalid knowledge state. This phenomenon occurs when a knowledge item has two alternative prerequisites. For example, one individual might learn to add two fractions by first transforming them into a common denominator, whereas someone else might have learned to transform them into decimal form first, and then transform it back into a rational form. If each of them ignores the other individual's method, then the intersection of their knowledge states yields a state with the mastery of the fraction addition problem while none of the two alternative prerequisite knowledge items is mastered.

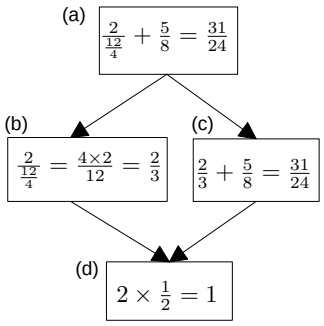

Figure 2: A simple knowledge space composed of 4 items ($\{a, b, c, d\}$) and with a partial order that constrains possible knowledge states to $\{\emptyset, \{d\}, \{b, d\}, \{c, d\}, \{b, c, d\}, \{d, b, c, a\}\}$.

It can be seen that the theory of knowledge spaces makes no attempt to structure knowledge in a hierarchy of concepts or any other structure containing latent variables (often called *latent traits*). The knowledge state of an individual is solely defined in terms of observable evidence of skills such as test question items. Of course, that does not preclude the possibility to re-structure knowledge items into higher level concepts and skills. In fact, this precisely is what a teacher does for developing a quiz or an exam, for example.

Item to Item Structures and Partial Orders

For our purpose, we make the assumption/approximation that knowledge spaces are closed under *union* **and** *intersection* and ignore the possibility of representing alternate prerequisite knowledge items. We refer to this variant as *partial order knowledge structures*, or POKS. Such structures can be represented by a DAG (Directed Acyclic Graph)[1], such as the one in Figure 2, because we further impose the assumption of closure under intersection. This assumption allows a considerable reduction the space of knowledge states. It greatly simplifies the algorithms for inducing a knowledge structure from data and reduces the amount of data cases required. Whether this assumption is warranted for knowledge assessment is a question we investigate empirically here.

Although a POKS network like the one in Figure 2 can be conveniently represented graphically by a DAG that resembles to a BN, the semantics of links is different. BN directed links usually represent causal relationships (although they can represent any kind of probabilistic relationship) and the structure explicitly represents conditional independence between variables. A Knowledge space directed link is similar to a logical implication relation, but it represents a prerequisite, or, to use Doignon and Falmagne terminology, a *surmise* relation. For example, if we have a surmise relation $A \succ B$, it implies that the mastery of B will precede the mastery of A, and thus if a student has a success for A,

[1] See Doignon & Falmagne (1999) for a formal proof and thorough analysis of the formal properties of knowledge spaces.

that student is likely to have a success for B.

The Induction of Item to Item Structures from Data

Item to item structures that are compliant with Partial Order Knowledge Structures (POKS) can be learned from data and we will describe two Bayesian frameworks for this purpose: a general Bayesian Network and a Naïve Bayes framework. We describe each of these two approaches and compare their respective predictive performance.

Bayesian Network Induction

In spite of the semantic differences between the links of a BN and those of an item to item structure like Figure 2's, the relations of both structures can be thought of as probabilistic implications between nodes. Both can represent evidence that influences the probabilities of neighboring nodes, in accordance to a Bayesian framework. It follows that any BN structural learning algorithm is a reasonable candidate for learning item to item structures. However, it must be emphasized that the semantics of a POKS structure is different from a BN and that, consequently the structure induced by a BN would have a different topology than the corresponding POKS, as we see later.

We conducted a study on learning item to item BN structures with the K2 (Cooper & Herskovits 1992) and PC algorithms (Spirtes, Glymour, & Scheines 2000). These algorithms are regularly used in the BN learning literature.

K2 The general principle of the K2 algorithm is to maximize the probability of a given topology given observed data. It uses a greedy search algorithm over the space of network topologies (Cooper & Herskovits 1992). The search is constrained by a given initial node ordering pattern to reduce the search space. For our experiments we use the topological order obtained from running the Maximum Weight Spanning Tree (MWST) algorithm by (Chow & Liu 1968) to derive a network topology, and by extracting a topological order from this structure. François & Leray (2003) has shown that the initial DAG obtained by the MWST is an effective replacement to a random ordering.

PC In contrast to searching the space of network topologies using a global Bayesian metric to score the topologies, the PC algorithm (Spirtes, Glymour, & Scheines 2000) falls into the *constraint-based structural learning* approach. It uses local conditional independence tests between a set of nodes to determine the network topology. Heuristic search consists in adding and deleting links according to the results of the independence tests and the search strategy. Murphy (2001) reports that the PC algorithm is in fact a faster but otherwise equivalent version of the IC algorithm from Pearl & Verma (1991).

In the experiment reported below, the BN parameters for both algorithms were initialized with Dirichlet uniform priors, which correspond to Beta priors in the case of binomial variables. We use the original Bayesian metric of Cooper & Herskovits (1992) to score the structures.

The PC algorithm must be given a value for the interaction test significance level. We use a value of 0.2.

We use Ken Murphy's BNT package for learning the BN structures of all the experiments conducted (`http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html`). Note that it was not possible to test the PC algorithm for the Unix test because of resource limitations with Matlab™.

POKS Structural Induction

The second approach for inducing the relations among items is based on Desmarais, Maluf, & Liu (1996). We refer to it as the POKS induction algorithm. This approach to learning can be considered as a constraint-based structural learning approach since it uses conditional independence tests to determine the structure. It can also be considered a Naïve Bayes approach because it makes a local independence assumption that greatly simplifies the structural learning and evidence propagation algorithms.

Given the local independence assumption, the POKS induction algorithm relies on a pairwise analysis of item to item relationships. The analysis attempts to identify the order in which we master knowledge items in accordance to the theory of knowledge spaces (Doignon & Falmagne 1999). However, it imposes a stronger assumption than their original work, namely that the skill acquisition order can be modeled by a directed acyclic graph, or DAG.

The tests to establish a relation $A \rightarrow B$ consists in three conditions for which a statistical test is applied:

$$P(B|A) \geq p_c \quad (1)$$

$$P(\overline{A}|\overline{B}) \geq p_c \quad (2)$$

$$P(B|A) \neq P(B) \quad (3)$$

Conditions (1) and (2) respectively correspond to the confidence to predict that B is true given that A is observed true (*mastered*), and the confidence that A is false (*non mastered*) given that B is false. The third condition verifies that the conditional probabilities are different from the non conditional probabilities (i.e. there is an interaction between the probability distributions of A and B). The first two conditions are verified by a Binomial test with parameters:

- p_c the minimal conditional probability of equations (1) and (2),
- α the alpha error tolerance level.

The conditional independence test is verified by the Fisher exact test. The χ^2 test could also be used. See Desmarais, Maluf, & Liu (1996) or Desmarais & Pu (2005) for further details about the parameters.

For this study, p_c is set at 0.5. Condition (3) is the independence test verified through a χ^2 statistic with an alpha error set to $\alpha < 0.2$. The greater the values for α, the more relations will be retained in the POKS network.

Inference

Once we obtain an item to item structure, a probability of success over all items can be computed from partial evidence (a subset of observed items). We will evaluate the

validity of the two frameworks over their item outcome predictive ability. We do not attempt to assess the actual item to item structures themselves, because we have no means to determine their true structures. In fact, that issue belongs to the field of cognitive science and was already thoroughly investigated by Doignon and Falmagne (Doignon & Falmagne 1999, see) and a number of other researchers. Our interest lies in the predictive power of the models which is measured by their ability to perform accurate assessment.

Inference in BN For the BN structure, there exist a number of standard and well documented algorithms (refer to Neapolitan, 2004, for eg.). We use the junction-tree algorithm (Jensen 1996) which preforms an exact computation of posterior probabilities within a tree whose vertices are derived from a triangulated graph, which is itself derived from the DAG in the BN.

Inference in POKS For the POKS framework, computation of the nodes' probabilities are essentially based on standard bayesian posteriors under the local independence assumption. Given a relation $E \rightarrow H$, where E stands for an evidence node (eg. a knowledge item) and H stands for a hypothesis node (eg. a prerequisite of E), the posterior probability of H is computed from the odds likelihood version of Bayes' Theorem:

$$O(H \mid E) \quad = \quad O(H)\, \frac{P(E \mid H)}{P(E \mid \overline{H})} \tag{4}$$

where $O(H)$ is the prior odds ratio and $O(H|E)$ represents the odds of H given evidence of E, and assumes the usual odds definition $O(H|E) = \frac{P(H|E)}{1-P(H|E)}$.

In order to make inference from combined evidence sources, the knowledge structure inference process makes the local independence assumption. In the standard Bayesian network graphical notation, this assumption corresponds to the network in Figure 3. Given that assumption, the computation of a joint probability of evidence nodes, $E_1, E_2, \ldots, E_i$, and the hypothesis node, H, is a straightforward product of likelihood ratios. For example, assuming that we have n number of relations of the form $E_i \rightarrow H$, then it follows from this assumption that:

$$P(E_1, \ldots, E_n \mid H) = \prod_i^n P(E_i \mid H) \tag{5}$$

From equation (5), it follows that the probability update of H given $E_1, ..., E_n$ can be written in following posterior odds form:

$$O(H \mid E_1, E_2, \ldots, E_n) = O(H) \prod_i^n \frac{P(E_i \mid H)}{P(E_i \mid \overline{H})} \tag{6}$$

Local independence is a strong assumption that is a characteristic of the Naïve Bayes framework. It greatly simplifies the amount of data required to calibrate conditional probabilities. Although this assumption is very likely violated to a certain degree in many cases, it was shown to be relatively robust in many situations (Domingos & Pazzani 1997). The extent to which the violation affects the model performance is empirically explored in the following section.

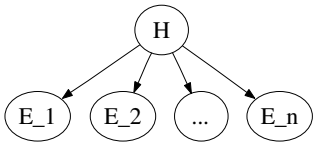

Figure 3: A Naïve Bayes network.

Predictive Performance of Item to Item Structures

The BN and POKS structural learning approaches of item to item structures are compared over their ability to predict item response outcome. We use data from real tests to conduct simulations and measure the performance of each approach for predicting the outcome over the full set of item answers from a subset of observed answers. This validation technique is identical to the ones used by Vomlel (2004) and Desmarais & Pu (2005).

Simulation Methodology

The experiment consists of simulating the question answering process with the real subject data. The process starts with an initial estimate of each item probability based on the data sample. For a given subject, an item is chosen as the next *observed evidence* and the actual subject's answer is fed to the inference algorithm. An updated probability of success for each item is computed given this new evidence, and a new item is chosen next based on the updated probability. This cycle is repeated until all items are "observed". After each observed item, we compare the estimated probabilities with the real answers to obtain a measure of how accurate the predictions are. All items for which the estimated probability of success is above 0.5 are considered mastered, and all others are considered non-mastered. Observed items are bound to their true value, such that after all items are administered, the score always converges to 1.

The simulations replicate a context of computer adaptive testing (CAT) where the system chooses the question items in order to optimize skills assessment. This context is typical of study guide applications, where a quiz is administered prior to providing pedagogical assistance (Falmagne *et al.* 2006; Dösinger 2002). However, the choice of question may not entirely be driven by the need to optimize skills assessment, but also by an adaptive pedagogical strategy such as in Heller *et al.* (2006), for example.

For this experiment, the choice of the question to ask is determined by an entropy reduction optimization algorithm. The same algorithm is used for both the BN and POKS frameworks and is described in Vomlel (2004) and also in Desmarais & Pu (2005). Essentially, the choice of the next question to administer corresponds to the one that reduces the entropy of a set of network nodes. The algorithm will choose the item that is expected to reduce entropy the most. Items with very high or low probability of success are generally excluded because their expected entropy reduction value will be low.

Table 1: Data sets

Data set	nb. items	nb. cases			Average success rate
		Training	Test	Total	
Arithmetic	20	100	49	149	61%
Unix	33	30	17	47	53%

Data Sets

The data sets are taken from two tests administered to human subjects :

1. *Arithmetic test.* Vomlel (2004) gathered data from 149 pupils who completed a 20 question items test of basic fraction arithmetic. This data has the advantage of also containing independent concept assessment which we will return to when assessing the approaches' ability to predict concepts.
2. *UNIX shell command test.* The second data set is composed of 47 test results over a 33 question items test on knowledge of different Unix shell commands. The questions range from simple and essential commands (eg. *cd*, *ls*), to more complex data processing utilities (eg. *awk*, *sed*) and system administration tools (eg. *ps*, *chgrp*).

For each data set, a portion of the data is used for training and the remaining ones for testing. Table 1 provides the size of the training and testing sets along with the average success rate of each test.

For each data set, six training and test sets were randomly sampled from both corpus. All performance reports represent the average over all six sampled sets.

Learned Structures

Over all six randomly sampled sets, the POKS structural learning algorithm created structures that, for the arithmetic data set, contains between 181 and 218 relations, of which 117 to 126 are respectively symmetric, for an average between 9.1 to 10.9 links per node. For the Unix data set, the number of relations varies between 582 and 691, with the number of symmetric relations that varies between 348 and 297. The average relations per node varies between 17.6 to 20.9. The structure of the Unix data set is thus much more populated with an average link per node about twice that of the arithmetic test. These structures are too dense to be shown graphically here.

For the BN structural learning results, Figure 4 displays the first two structures (of the six sample runs) learned with the K2 algorithm with the arithmetic data set. It can be seen that the topology differs significantly between the two networks shown in this figure. In general, about only half of the relations are common between BN from two samples. However, and as mentioned, we do not focus on the actual topologies in this study but solely on the ability of the induced structures to perform accurate inferences.

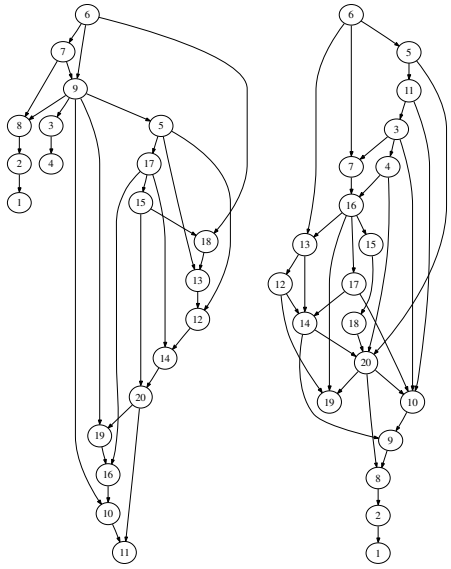

Figure 4: Two examples of BN structures learned with the K2 algorithm.

Results

Figure 5 reports the simulations results. It shows that, for both data sets, the POKS algorithm yields more accurate predictions of item outcome than the two BN algorithms. As a comparison, a random item selection strategy that does not make any inference would yield a straight line starting at the same first data point of the POKS curve, and ending at 1.

Although the difference is only a few percentage points, it is relatively significant, at least for the Unix data set. For example, after 20 items, the difference between the BN and POKS for the Unix data set is about 95% compared to 98% (see Figure 5). Although this represents a 3% difference in absolute values, it actually represents a 60% relative reduction in terms of the remaining error. In other words, the system would reduce the number of wrong decisions by a factor of over 2. In a context where, for example, we need strong confidence that a specific item is mastered and avoid making wrong decisions from an incorrectly assessed item, the difference in reliability can be quite meaningful.

The relative error reduction between both tests of the BN vs. POKS algorithms is significantly greater for the Unix than for the arithmetic test. This is potentially due to the fact that this test was meant to discriminate between a wide range of expertise, from basic file manipulation commands to sophisticated data manipulation scripts. In contrast, the arithmetic test is more focused on closely related skills and notions typical of early high school children. Moreover, the span of scores is wider for the Unix test than for the arithmetic one, with a respective variance of 29% compared to 25%. As a consequence, the ordering between items that is typical of knowledge structures is more likely to appear for the Unix than for the arithmetic test. Yet, another explanation is that POKS may be more robust to small sample

size than the BN algorithms (recall that 30 training cases are used for the Unix data set whereas 100 are used the arithmetic one). These issues remain to be investigated.

We also note that the PC algorithm performs better than the K2 algorithm, apparently due to more accurate priors. However, the difference quickly vanishes after 2 items observed, after which the difference is insignificant.

Figure 6 reports the variability of the estimates across subjects, averaged over the six random samples. The actual variance of subject results for a single sample is actually wider, but for the purpose of comparing the approaches we average over the six runs. The plot represents the median and quartiles for the different algorithms. The middle line of a box indicates the median and the upper and lower regions of the box spans over roughly one quartiles around the median. The 'whiskers' represent the outer quartiles and outliers are shown as individual data points. The arithmetic result 'boxplot' of Figure 6 contains three series of boxes. The first corresponds to the POKS results and the next two are for the BN results of the PC and K2 algorithms. The Unix plot contains only the POKS and K2 algorithm results.

Discussion

The better performance of the POKS approach over a BN approach may appear surprising since both schemes rely on the Bayesian framework and the POKS approach makes stronger assumptions than the BN approach. However, this is not an exception as the Naïve Bayes framework, which shares the local independence assumption with POKS, was shown very effective in many context (Domingos & Pazzani 1997). The context of POKS may well be the case too.

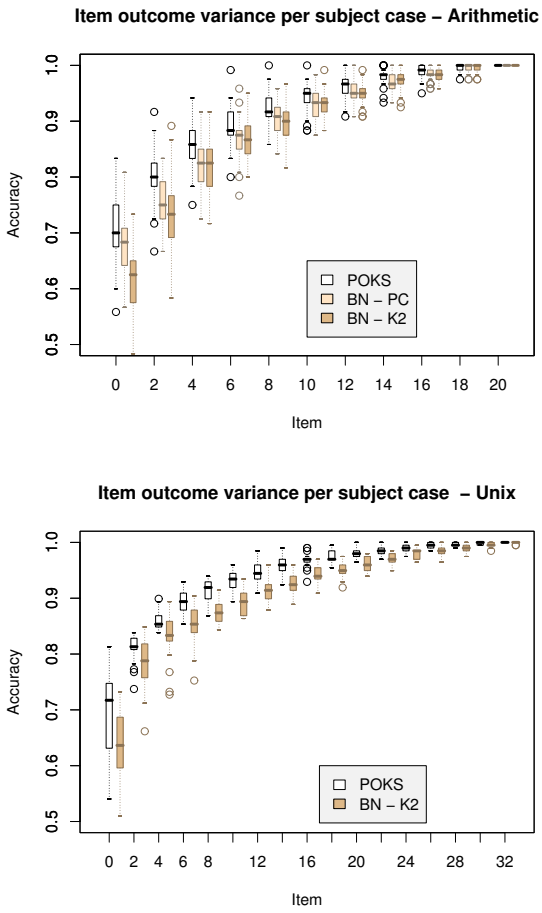

Figure 6: Box plots of the variance across subjects (49 for Arithmetic and 17 for Unix) and averaged over the six random samples. Central quartiles are within the box whereas the two extreme quartiles are represented by the "whiskers". The middle line in the box represents the median and outliers are represented by data points. Only the K2 results are shown for Unix.

From Observable Nodes to Concepts Assessment

So far, we showed how to learn a student model that contains solely observable nodes, namely items that can represent test questions or, more generally, any observable manifestations of a skill that a student could master, or fail to master. However, the systems that can make use of student models, be it adaptive hypertext, intelligent tutoring system, or study guides, need to work at the level of concepts, not at the item level. Linking observable items to hidden concept nodes is thus a problem that every student modeling approach based on BN has had to tackle.

This is generally done by defining a hierarchy, with items as leaf nodes and concepts and misconceptions as higher level nodes (see Figure 1). However, a number of issues remain on how to define and parametrize the BN structure. In particular, given that concept nodes are not directly observable, how can the conditional probabilities be derived without prior data? How can the topology be defined?

Let us briefly review some of the previous work on using BN to link items to concepts and introduce the approach tested in this study.

Previous Work with BN

VanLehn *et al.* (1998) investigated some variations on Pearl's noisy-And model (Pearl 1988) to link observable items to concepts, and found them effective with simulated student test data. Whereas this technique represents a means to introduce evidence from items into a BN in the absence of the required conditional probability tables, Millán *et al.* (2000) introduces a means to fill such tables in the absence of sufficient data. They used a combination of expert judgments and IRT's logistic function to parametrize the conditional probabilities between test items and concepts.

Vomlel (2004) explored a number of learning approaches to define the structure of a BN of concepts and misconceptions in arithmetic. His study is noteworthy because he used independent concept mastery assessment data for the BN structural learning and conditional probability calibration. The best model was derived by an iterative process, where initial structures are first derived from the data and using the PC algorithm, and constraints on the structure is imposed

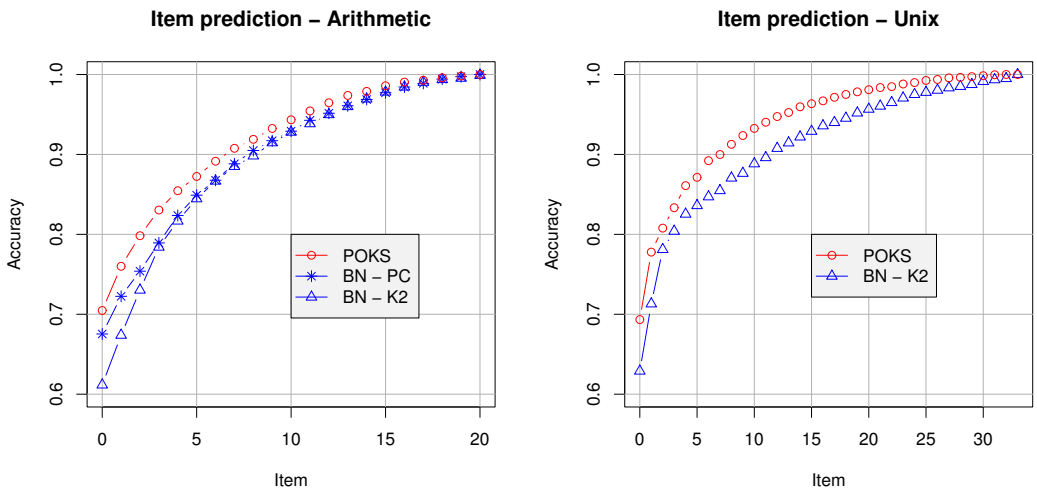

Figure 5: Item prediction performance. The graph on the left reports the accuracy of item outcome predictions for the Arithmetic test averaged over all 49 test cases. The graph on the right reports the Unix accuracy and it is averaged over 17 test cases. Each line represents averages over 6 simulation samples.

by domain experts to refine the structure. We return to this study since its results are used in the current study.

Rule Space, Weighted Means

Of course, not all approaches rely on a BN, and other techniques such as Tatsuoka's Rule Space or a simple weighted means are valid alternatives to assess concepts.

Tatsuoka (1983) introduced the concepts of Rule space and Q-matrices. Each "rule" (concept or skill) that is necessary to be successful at each test item is indicated in a matrix of rules by items. A probabilistic version of this framework actually corresponds to the approach of VanLehn *et al.* (1998) mentioned above. See also Barnes (2003) and Winters (2006) who investigated a number of probabilistic approaches of using Q-matrices and a number of other techniques for knowledge assessment.

An simple alternative to the Q-matrix is to decompose the mastery of a given concept as a weighted mean of items, much in the same manner as every teacher does when points are allotted to different test items in a exam. That approach has the advantage of being readily understood by teachers who frequently go through this process of determining which test items assess which concepts or topics.

Augmenting the Observed Evidence Set

Assuming we linked observable evidence to concepts that are linked within a BN, we could use the item to item model to augment the initial set of observed evidence and feed this augmented evidence set to the concept level model. For example, an item to item model could feed a BN with an augmented response vector that complements the information used by the BN at the concept level. To the extent that the item to item model provides an accurate assessment, we would expect that the assessment at the concept level would also be improved.

This scheme is further detailed and evaluated in the next section.

Evaluation of the Augmented Evidence Scheme

In accordance to the objective of assessing concept mastery from observable items, we investigate the effectiveness of combining item-to-item structures with a BN that contains concept and misconception nodes in accordance with the scheme outlined in the previous section. We used an existing hierarchical BN and combined it with a POKS built from the same data. The BN is taken from Vomlel (2004). Vomlel defined a list of 20 concepts and misconceptions and defined a BN structure over these nodes. The BN parameters are calibrated with the same data as the arithmetic test in table 1, and also with data from an independent assessment by experts of each of the 149 pupils mastery of these 20 concepts. In addition to calibration, this independent assessment of concepts also allows us to determine the concept predictive accuracy of the algorithms.

In this experiment, we use the POKS inference engine as a filter to augment the actual number of observations fed to the BN. Hence, the initial set of items observed mastered and non mastered is first given to the POKS module. All of the yet unobserved items, for which the chances of mastery are above (below) a given threshold, are considered mastered (non mastered). They are then given to the BN as additional evidence on top of the observed evidence.

More formally, assuming a set of observed responses S, POKS infers a set of additional responses, S'. The original set, S, is thereby augmented by the inferences from POKS, S' and the set of evidence fed to the BN represents the union of S and S'. This process is illustrated in Fig-

ure 8. It is repeated for every new observation, from 0 to all 20 items.

In order to determine that an item is considered inferred by POKS, a threshold is used, δ. Every item for which the probability of mastery of POKS is greater than $1 - \delta$ is considered mastered, whereas items with a probability smaller than δ is considered non mastered.

Results

We assume that, with the POKS augmented set of evidence, the BN estimated probability of concept and item mastery will be more accurate than the non augmented set. This is verified through a simulation using the same methodology as the item-to-item simulation. However, in this case, we can also report the accuracy of concept assessment based on Vomlel's independent concept assessment data and compare with his original results..

The simulation results of the combination algorithm of POKS and the BN are reported in figure 7. A threshold $\delta = 0.1$ is used for this experiment. This value provided the best results, although there were only small differences between thresholds values of 0.30 to 0.95.

The graph reports the prediction accuracy for concepts and items separately and for three conditions:

BN+POKS: *augmented* inferences.

BN: BN *non augmented* inferences.

POKS: replication of POKS item prediction performance (replication of the arithmetic results in Figure 5). based on global entropy (items and concepts).

The item prediction results reveal that the highest performance is achieved when the BN inferences are augmented by the observations from the POKS inferences. The performance is significantly better than for the "non augmented" BN condition, but only marginally better than for POKS alone. These results suggest that item to item structures can provide additional, complementary inference to the BN when it comes to predicting item outcome.

In contrast to the item prediction results, the concept prediction accuracy results reveal that all four conditions are relatively similar. Surprisingly, the improvement seen for the item outcome prediction does not transfer to the concept prediction. A possible explanation is that by optimizing item selection on item entropy reduction for the POKS+BN condition, the gain from the augmented inferences is offset by targeting item entropy over the global, combined concept and item entropy. Another possibility is simply that the independent concept assessment is not reliable enough to reveal gains at the concept level. The span of accuracy for concepts only ranges between 75% and 90% and the maximum is almost reached after only 5 question items. There may be no room for improvement with such data. Moreover, many concepts have only 2 or 3 items from which to validate their mastery, which may be too little for a reliable assessment. In fact, we currently have no reason to expect that improvements at predicting item outcome would not be reflected in concept mastery assessment.

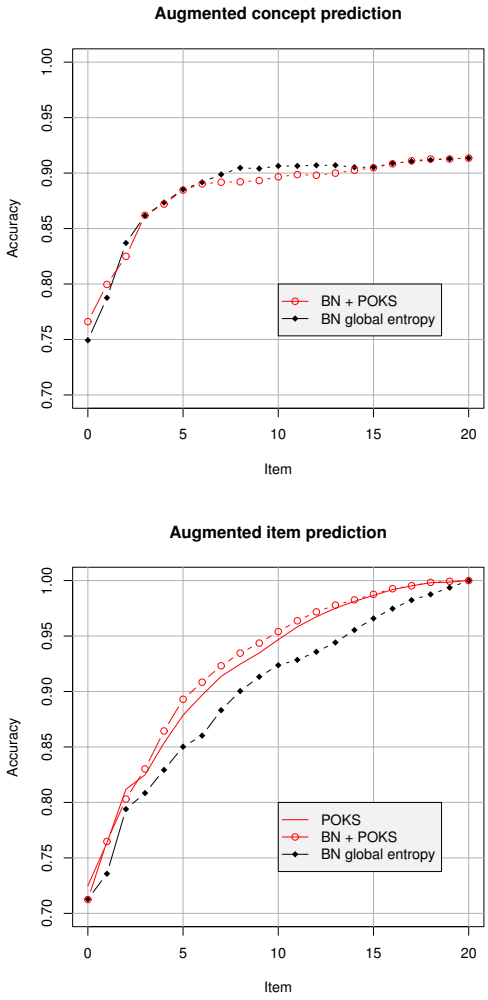

Figure 7: Results of the simulation where the POKS inferences are used to augment the observed set of items. A threshold value of $\delta = 0.1$ is used. Refer to the text for a description of each curve.

Discussion

Learned item to item student models have the potential to provide accurate, fine-grained skills assessment without the drawbacks of requiring significant human effort and expertise. This effort could be limited to the familiar task that every teacher goes through during the elaboration of an exam: linking and weighting items with respect to a list or a hierarchy of concepts.

This study shows that item to item structures can be constructed from data and yield effective predictive item outcome models. Two approaches were investigated, namely standard BN induction techniques and the POKS framework, which stands closer to the Naïve Bayes family of models.

Two simulations show that the POKS framework yields better predictive results for item outcome prediction than

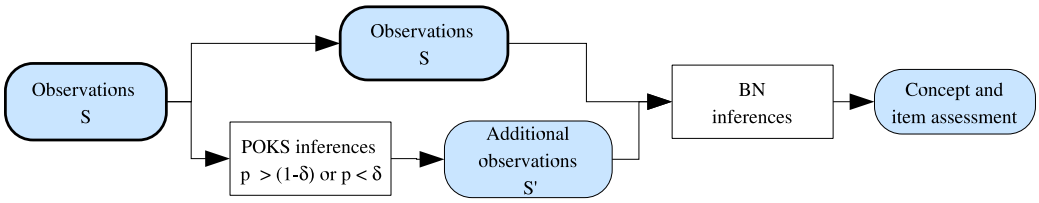

Figure 8: Combination algorithm of POKS with BN.

does the general BN framework. Although the stronger performance of a Naïve Bayes framework is by no means uncommon, we conjecture that, for this study, it is consequent with the constrained nature of knowledge spaces closed under union and intersection. The assumptions made by the POKS framework may be appropriate for the knowledge spaces and, consequently, allow the framework to be applied effectively with small data sets as the ones we experimented with. Moreover, the simplicity of the POKS framework is reflected in the computational cost: In our simulations, the POKS algorithms are faster than the BN algorithms by a factor of two orders of magnitude, both for model construction and inference. This is attributed in part to the fact that POKS does not use iterative algorithms but relies on a closed form solution.

We further investigated how to link item to item structures to an existing BN model, which offer high modeling flexibility at the concept level and enjoy great recognition in the student modeling community, and can lead to the re-use of student models. The specific approach studied is to augment the set of evidence from observed items using the item to item inference scheme.

The results show that we can, indeed, improve item outcome prediction with the augmented inference scheme. However, we could not demonstrate improvements at the concept assessment level from the simulation conducted. This could be a limitation of the data set used, or it could be a side effect of the item selection strategy that we can either gear towards item or concept entropy reduction. Nevertheless, given that item outcome is determined by the student skills and concept mastery, the improvement obtained at the item outcome level should eventually lead to a better skills assessment.

A number of issues remain open over the current study, one of which is how general are the findings. We already see a different pattern of results between the simulations over the two data sets. We suggested that the Unix test showed greater relative error reduction because it was designed and tested over a very large span of expertise and is therefore highly consistent with the knowledge space framework. It is quite plausible that some domain of knowledge, or some types of tests may not conform to the underlying assumptions of POKS and knowledge spaces and therefore the framework would not perform as well. Similarly, the BN structural learning algorithms can display wide differences depending on the nature of the data set and the sample size (François & Leray 2003, see, for eg.,). As a consequence, the effectiveness of item to item approaches may vary and more investigations are required to address this issue.

Many of the qualities that we expect from a learned a student modeling framework are found item to item. The experiments we conducted showed their effectiveness for performing knowledge assessment with models learned from very small data sets (as few as 30 data cases for the Unix experiment with POKS). Yet, they display all the advantages of graphical probabilistic learned models, namely the automation of model building and the absence of human intervention, which avoids the human expertise bottleneck and subjectivity, and offers the possibility of estimating the reliability of the diagnostic. We also showed how they can be effectively used in combination with existing BN models to yield accurate concept assessment and within a perspective of reusing user models.

References

Barnes, T. M. 2003. *The q-matrix method of fault-tolerant teaching in knowledge assessment and data mining*. Ph.D. Dissertation, North Carolina State University.

Chow, C., and Liu, C. 1968. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Information Theory* 14(11):462–467.

Cooper, G. F., and Herskovits, E. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9:309–347.

Desmarais, M. C., and Pu, X. 2005. A bayesian inference adaptive testing framework and its comparison with item response theory. *International Journal of Artificial Intelligence in Education* 15:291–323.

Desmarais, M. C.; Maluf, A.; and Liu, J. 1996. User-expertise modeling with empirically derived probabilistic implication networks. *User Modeling and User-Adapted Interaction* 5(3-4):283–315.

Doignon, J.-P., and Falmagne, J.-C. 1999. *Knowledge Spaces*. Berlin: Springer-Verlag.

Domingos, P., and Pazzani, M. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29:103–130.

Dösinger, G. 2002. Adaptive competence testing in eLearning. *European Journal of Open and Distance Learning*.

Dowling, C. E., and Hockemeyer, C. 2001. Automata

for the assessment of knowledge. *IEEE Transactions on Knowledge and Data Engineering*.

Falmagne, J.-C.; Koppen, M.; Villano, M.; Doignon, J.-P.; and Johannesen, L. 1990. Introduction to knowledge spaces: How to build test and search them. *Psychological Review* 97:201–224.

Falmagne, J.-C.; Cosyn, E.; Doignon, J.-P.; and Thiéry, N. 2006. The assessment of knowledge, in theory and in practice. In Missaoui, R., and Schmid, J., eds., *ICFCA*, volume 3874 of *Lecture Notes in Computer Science*, 61–79. Springer.

François, O., and Leray, P. 2003. Etude comparative d'algorithmes d'apprentissage de structure dans les réseaux bayésiens. In *RJCIA03*, 167–180.

Heller, J.; Steiner, C.; Hockemeyer, C.; and Albert, D. 2006. Competence–based knowledge structures for personalised learning. *International Journal on E–Learning* 5(1):75–88.

Horvitz, E. 1999. Principles of mixed-initiative user interfaces. In *CHI*, 159–166.

Jensen, F. V. 1996. *An introduction to Bayesian Networks*. London, England: UCL Press.

Kambouri, M.; Koppen, M.; Villano, M.; and Falmagne, J.-C. 1994. Knowledge assessment: tapping human expertise by the query routine. *International Journal of Human-Computer Studies* 40(1):119–151.

Millán, E.; Trella, M.; Pérez-de-la-Cruz, J.-L.; and Conejo, R. 2000. Using Bayesian networks in computerized adaptive tests. In Ortega, M., and Bravo, J., eds., *Computers and Education in the 21st Century*. Kluwer. 217–228.

Mislevy, R. J.; Almond, R. G.; Yan, D.; and Steinberg, L. S. 1999. Bayes nets in educational assessment: Where the numbers come from. In Laskey, K. B., and Prade, H., eds., *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI-99)*, 437–446. S.F., Cal.: Morgan Kaufmann Publishers.

Murphy, K. P. 2001. The Bayes net toolbox for MATLAB. Technical report, University of California at Berkeley; Berkeley , CA.

Neapolitan, R. E. 2004. *Learning Bayesian Networks*. New Jersey: Prentice Hall.

Pearl, J., and Verma, T. 1991. A theory of inferred causation. In *KR*, 441–452.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. Cambridge, Massachusetts: The MIT Press, 2 edition.

Tatsuoka, K. 1983. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement* 20:345–354.

VanLehn, K.; Niu, Z.; Siler, S.; and Gertner, A. S. 1998. Student modeling from conventional test data: A Bayesian approach without priors. In *ITS'98: Proceedings of the 4th International Conference on Intelligent Tutoring Systems*, 434–443. London, UK: Springer-Verlag.

Vomlel, J. 2004. Bayesian networks in educational testing. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* 12(Supplementary Issue 1):83–100.

Winters, T. 2006. *Educational Data Mining: Collection and Analysis of Score Matrices for Outcomes-Based Assessment*. Ph.D. Dissertation, University of California Riverside.