Bayesian Networks in Educational Testing

Jiří Vomlel
vomlel@utia.cas.cz

Inst. of Information Theory and Automation, Academy of Sciences, Prague, Czech Republic

Department of Computer Science, Aalborg University, Denmark

Abstract

In this paper we discuss applications of Bayesian networks to educational testing. Namely, we deal with the diagnosis of person's skills. We show that when modeling dependence between skills we can get faster a better diagnosis. We also show that there exist test setups for which it is impossible to get a complete and correct diagnosis unless we model dependence between skills. We present results of experiments with basic operations with fractions. The experiments suggest that the test design can benefit from a Bayesian network modeling relations between skills not only in the case of an adaptive test but also when designing a fixed (non-adaptive) test.

1 Introduction

One task in educational testing is the diagnosis of the presence or the absence of person's skills. A possible scenario is that an institution provides a certain course and wishes to test the applicants for gaps in their prerequisites for attending the course. This task is more difficult than giving grades to students or the classification of examinees according to their overall level since the goal is not only to say whether an examinee is good or not but to pinpoint what are examinee's abilities and weak points.

Typically, a test designer specifies the tested skills $\mathcal{S} = \{S_1, \ldots, S_k\}$ and a bank of questions $\mathcal{X} = \{X_1, \ldots, X_m\}$. For every question the designer should specify the skills that are related to the question. Relations are often probabilistic, especially if a multiple choice test is used.

One approach is to construct a test that consists of a fixed sequence of questions covering all tested skills. We will call this type of test a *fixed test*. Another approach aims at constructing an optimal test for each examinee. After each response on a question the system selects next question based on the answers of the previous questions. Since this approach requires computers for the test administration it is often referred to as *computerized adaptive testing* (CAT), see Wainer et al. (2000) and Linden and Glas (2000). Tests that are automatically tailored to the level of the individual examinees will be referred as *adaptive tests*.

Almond and Mislevy (1999) proposed to use graphical models for CAT. Their model consists of one *student model* and several *evidence models*, one for each task or question. Let $\mathcal{S} = \{S_1, \ldots, S_k\}$ denote the set of examinees' skills, abilities, misconceptions, etc. For simplicity, we will call the elements of set $\mathcal{S}$ skills. The student model describes relations between skills. Every skill S_i is represented by a random variable having a finite set $\mathbb{S}_i$ of skill values. The knowledge about a student is expressed by use of a joint probability distribution $P(S_1, \ldots, S_k)$ defined on the variables of the student model. Let $\mathcal{X}$ denote the set of questions or tasks. For each question or task $X_j \in \mathcal{X}$ there is one evidence model describing the dependence of X_j on relevant skills from the student model.

In Section 2 we discuss criteria and methods for the test construction. We present a new algorithm exploiting Bayesian network model for the construction of a fixed test. In Section 3 we introduce student and evidence models for testing basic operations with fractions.

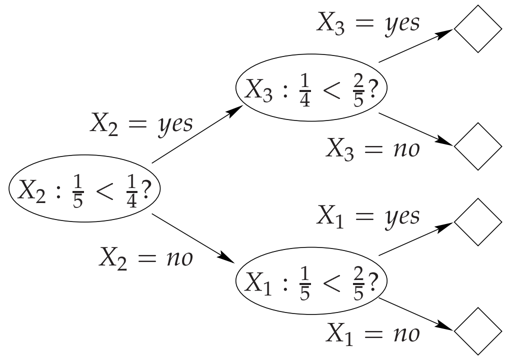

Figure 1: Example of an adaptive test

In Section 4 we describe the process of learning models and present results of experiments performed with tests built using the learned models.

2 Test construction

Every test $\mathbf{t}$ can be represented by a directed tree. See Figure 1 for a simple example. If the student's answer to the first question (X_2) is correct then the second question is X_3, which is more difficult, otherwise the student gets question X_1, which is easier. Every node corresponds to a question. Every edge $n \to m$ is labeled by by an answer to the question corresponding to the node n, $outcome(edge)$ is a function that provides the *edge* labels. The test terminates after a given number of questions was answered or if sufficient information about the tested examinee was achieved. The set of all terminal nodes of a test $\mathbf{t}$ (leaves of the corresponding tree) will be denoted by $\mathcal{L}(\mathbf{t})$. The root node will be denoted by ϑ. Let $path(n_1, n_k)$ be the sequence of edges constituting the path from node n_1 to node n_k in a test, then

$$\mathbf{e}_n = \bigcup_{edge \in path(\vartheta, n)} outcome(edge)$$

defines the evidence compiled through the performance of questions in $path(\vartheta, n)$. In every node n of test $\mathbf{t}$ we can compute probability $P(\mathbf{e}_n)$ of getting to this node.

Remark 1 Questions can be allowed to be repeated in a test if there are different versions of each question that have the same relation to the tested skills.

Design of a diagnostic test differs according to the goal of the test. Next we discuss different criteria that can be used when designing a test.

2.1 Maximum information

Naturally, every examiner would like to maximize information about the diagnosed examinee at the end of a test. In our context more information corresponds to a decrease in the uncertainty about the presence or the absence of the skills represented by the probability distribution over the skills. It means that we prefer probability distributions that have values close to zero or one. A way to formalize this preference is to aim at a probability distribution minimizing the Shannon entropy. Ben-Bassat (1978) used the Shannon entropy and variance to measure the impact of new information.

We will formalize the maximum information approach as a testing process aiming at distributions having minimal value of entropy at the end of the process.

Definition 1 Let $\mathbf{S}$ be a possibly multidimensional random variable with states $\mathbf{s}$ from a finite set $\mathbb{S}$ with probability distribution $P(\mathbf{S})$. Entropy $H(P(\mathbf{S}))$ of $P(\mathbf{S})$ is defined as

$$H(P(\mathbf{S})) = -\sum_{\mathbf{s} \in \mathbb{S}} P(\mathbf{S} = \mathbf{s}) \cdot \log P(\mathbf{S} = \mathbf{s}) \ .$$

Assume we are interested in all skills of the student model independently, i.e. in marginal probability distributions $P(S_i), i = 1, \ldots, k$.[1] In every node n of a test $\mathbf{t}$ we can compute the total entropy $H(\mathbf{e}_n)$.

Definition 2 Total entropy $H(\mathbf{e}_n)$ in a node n is

$$H(\mathbf{e}_n) = \sum_{S_i \in \mathcal{S}} H(P(S_i \mid \mathbf{e}_n))$$

We will simplify notation using function

$$EH(\mathbf{e}_n) = P(\mathbf{e}_n) \cdot H(\mathbf{e}_n) \ .$$

[1] Sometimes the goal of an examiner can be better formalized by use of the joint probability distribution $P(S_1, \ldots, S_k)$ or more dimensional marginals of the joint probability distribution.

We use the expected value of entropy after a test to define an optimal test.

Definition 3 Let $\mathcal{T}$ be the set of all considered tests. Test $\mathbf{t}^\star$ is optimal iff for all $\mathbf{t} \in \mathcal{T}$

$$\sum_{\ell^\star \in \mathcal{L}(\mathbf{t}^\star)} EH(\mathbf{e}_{\ell^\star}) \quad \leq \quad \sum_{\ell \in \mathcal{L}(\mathbf{t})} EH(\mathbf{e}_\ell) \enspace .$$

In practice it is often impossible to find an optimal test through the evaluation of all considered tests because of the combinatorial explosion. Instead a greedy heuristic is used to construct a myopically optimal test. A myopically optimal test is a test that consists of questions such that each question minimizes the expected value of entropy after the question is answered.

Definition 4 Let $ch(n)$ denote children of a node n in a test $\mathbf{t}$. A test $\mathbf{t}$ is myopically optimal iff in every nonterminal node n it holds that for all $X \in \mathcal{X}$

$$\sum_{m \in ch(n)} EH(\mathbf{e}_m) \quad \leq \quad \sum_{x \in \mathbb{X}} EH(\mathbf{e}_n \cup \{X = x\}) \enspace .$$

Other criteria

An alternative to the maximum information criteria is the minimization of *expected decision error*. This criteria is better suitable if, for example, the test is used to classify students into certain predefined groups. The criteria is then defined as minimization of probability of misclassification of students at the end of the test.

There can be several *psychometric constraints* on the order of items. For example, if there are related questions then they should be presented together. This restriction leads to the notion of a *testlet*, which is something like a group of related questions. Another restrictions can be that certain proportions of questions from different groups are required. Important constraints are constraints given by *security* issues. They should avoid frequent use of questions.

How should a test designer combine several criteria like maximum of information, reasonable ordering of questions, security issues, and other constraints? One solution that combines the security issue with a myopically optimal test is that instead of selecting a most informative question a probabilistic selection is used with probability of selecting question X being inversely proportional to

$$\sum_{x \in \mathbb{X}} EH(\mathbf{e}_n \cup \{X = x\}) \enspace .$$

Generally a test designer tries to minimize the number of constraints that are not fulfilled while maximizing the information at the end of a test. These questions are topics of an active research, see Stocking and Swanson (1993), Swanson and Stocking (1993), and Stocking and Lewis (1998). For a review of different test selection strategies see Madigan and Almond (1996). For the criteria often used for item selection in adaptive testing see Meijer and Nering (1999).

2.2 Fixed tests

In some situations a fixed test is more suitable than an adaptive test. One of the reasons can be that adaptive tests require computers for their application. Computers may not be available or persons taking the tests are expected to have problems with a computer based test. The design of a fixed test can also benefit from a Bayesian network model. The same criteria as in the case of adaptive tests can be used. The only difference is the set of considered tests $\mathcal{T}$. While adaptive tests corresponds to trees, every fixed test can be represented by a single sequence of questions.

In Table 1 we provide an algorithm for the construction of a fixed test. The algorithm performs a greedy search in the graph of the state space corresponding to all possible adaptive tests. It searches for a fixed test, therefore, at each stage it selects only one question.[2] The search is greedy, it means that at each stage a next question is selected using information available looking one step ahead. The selected question X_i maximizes probability of being used at the current stage or before in the myopically optimal adaptive test. This probability can be computed as the total sum of prob-

[2]Every stage in the state space graph corresponds to a position in the test.

Table 1: Construction of a fixed test

```
e_list := [∅];
test := [ ];
for i := 1 to |𝒳| do counts[i] := 0;
for position := 1 to test_length do
    new_e_list := [ ];
    for all e ∈ e_list do
        i := most_informative_X(e);
        counts[i] := counts[i] + P(e);
        for all x_i ∈ 𝕏_i do
            append(new_e_list, {e ∪ {X_i = x_i}});
    e_list := new_e_list;
    i* := arg max_i counts[i];
    append(test, X_{i*});
    counts[i*] := 0;
return(test);
```

abilities over all possible states s in the current and all previous stages where the question X_i would be selected, i.e. where

$$most\_informative\_X(\mathbf{e}_s) \quad = \quad i \ ,$$

where $\mathbf{e}_s$ denotes the evidence corresponding to the node of state s in the state space graph. The search continues only in the subspace corresponding to the selected question. We will use an example to better explain how the algorithm works.

Example 1 Figure 2 depicts the state space of all possible tests of the length two. Rectangular nodes correspond to states where a decision on a next question is made. Ovals correspond to states where examinee's answer to a given question is observed. Edges corresponds to answers. Assume question X_2 was selected as the first question. In Figure 2 the selection of the second question is shown. The evidence list $e\_list = \{\{X_2 = 0\}, \{X_2 = 1\}\}$. The most informative question for evidence $\{X_2 = 0\}$ is X_3:

$$most\_informative\_X(\{X_2 = 0\}) \quad = \quad 3$$

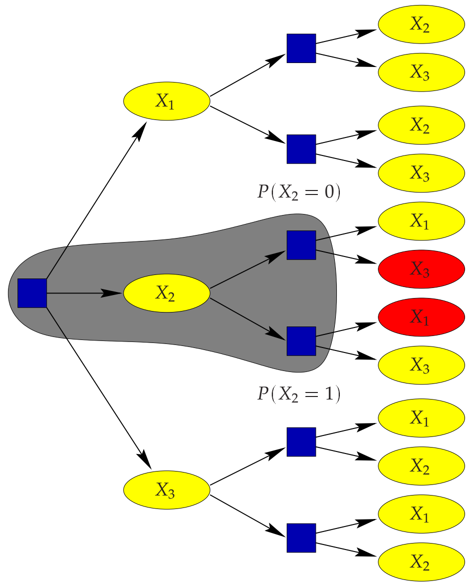

Figure 2: Selection of the second question in the algorithm from Table 1

therefore X_3 gets a count corresponding to $P(X_2 = 0)$, i.e. $counts[3] = P(X_2 = 0)$. Since

$$most\_informative\_X(\{X_2 = 1\}) \quad = \quad 1$$

X_1 gets a count corresponding to $P(X_2 = 1)$, i.e. $counts[1] = P(X_2 = 1)$. A question that has the maximal total count is selected as the second question. □

Modeling skill dependence

We will use Examples 2 and 3 to show that by modeling probabilistic or deterministic dependence between skills we can get a better diagnosis. Actually, for some test setups we can not get complete skill diagnosis without modeling skill dependence.[3]

Example 2 Assume a test that aims at diagnosing the absence (state 0) or the presence (state 1) of three skills S_1, S_2, and S_3 by use of a bank of three questions $X_{1,2}$, $X_{1,3}$, $X_{2,3}$ each with two states (state 0 denotes wrong answer, 1 correct). The questions are related to pairs of skills

[3]It may seem obvious that it is beneficial to encode relations between variables into a Bayesian network model. However, some standard techniques used in educational testing do not make use of it. For example, the standard Item Response Theory (Hambleton and Swaminathan, 1985) uses only one parameter θ to model the student.

(S_1, S_2), (S_1, S_3), and (S_2, S_3) by deterministic AND relations, which means that if $(s_i, s_j) = (1,1)$ then $P(X_{i,j} = 1 | S_i = s_i, S_j = s_j) = 1$ and 0 otherwise. Assume answers to all questions from the item bank are wrong, i.e. $X_{1,2} = 0$, $X_{1,3} = 0$, and $X_{2,3} = 0$.

Assume there is no student model describing dependence between skills, i.e. all skills are assumed to be independent, and probability distributions $P(S_j), j = 1, \ldots, k$ are uniform. The model is given in Figure 3. In this case we

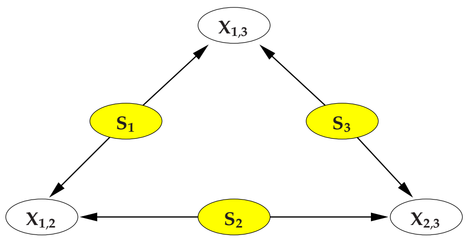

Figure 3: Overall model from Example 2.

can not decide which skills are actually present and which are missing. The probabilities for $j = 1, 2, 3$ are:

$$P(S_j = 0 \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) = 0.75.$$

We would need questions testing each skill independently, which may not be possible in practice. □

Example 3 If we extend Example 2 by a student model (see Figure 4) describing relations between skills we can make conclusions at least about some skills. Assume there is a determin-

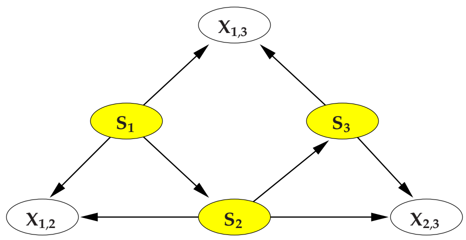

Figure 4: Overall model from Example 3.

istic hierarchy over the skills:

$$S_1 \Rightarrow S_2, \; S_2 \Rightarrow S_3 \; \text{ and}$$

$$\begin{aligned} P(S_1 = 0) &= 0.5 \\ P(S_2 = 0 \mid S_1 = 0) &= 0.5 \\ P(S_3 = 0 \mid S_2 = 0) &= 0.5 \;, \end{aligned}$$

Then we can infer that

$$\begin{aligned} P(S_1 = 0 \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) &= 1 \\ P(S_2 = 0 \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) &= 1 \\ P(S_3 = 0 \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) &= \frac{1}{2} \end{aligned}$$

Thus we can conclude that the student does not have skills S_1, S_2. The only uncertain skill is skill S_3.

Observe, that for $i = 1,2,3$ $P(S_i \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) = P(S_i \mid X_{2,3} = 0)$. Therefore, if we get the wrong answer to question $X_{2,3}$ then we get the same information as if we asked (and got wrong answers) all three questions. Thus, we can see that by using the student model we can get the same conclusions more efficiently. □

Often, the relations between skills are probabilistic. However, even with probabilistic relations encoded in the student model we can make more informed conclusions and in most cases we can get them also faster.

The length of the test (the number of questions) is related to the reliability of results. We will not discuss this issue in this paper. We only mention that there exist an extensive literature on sequential tests that are used for decision making - starting with Wald's sequential probability test (Wald, 1947).

3 Model for basic operations with fractions

3.1 Student model

The learning process that resulted in a student model consisted of several steps. First, a group of students from Aalborg University prepared paper tests that were given to students at Brønderslev High School. Four elementary skills, four operational skills, and abilities to apply operational skills to complex tasks were tested. See Table 2 for elementary and operational skills[4]. 149 students solved the test. The university students analyzed the tests and summarized the results, see Būtėnas et al. (2001). During this phase, seven types of misconception were discovered. See Table 3 for misconceptions observed in Brønderslev High School.

Table 2: Elementary and operational skills.

Label	Description	Example
CP	Comparison (common numerator or denominator)	$\frac{1}{2} > \frac{1}{3},\ \frac{2}{3} > \frac{1}{3}$
AD	Addition (comm. denom.)	$\frac{1}{7} + \frac{2}{7} = \frac{1+2}{7} = \frac{3}{7}$
SB	Subtract. (comm. denom.)	$\frac{2}{5} - \frac{1}{5} = \frac{2-1}{5} = \frac{1}{5}$
MT	Multiplication	$\frac{1}{2} \cdot \frac{3}{5} = \frac{3}{10}$
CD	Common denominator	$\left(\frac{1}{2}, \frac{2}{3}\right) = \left(\frac{3}{6}, \frac{4}{6}\right)$
CL	Canceling out	$\frac{4}{6} = \frac{2 \cdot 2}{2 \cdot 3} = \frac{2}{3}$
CIM	Conv. to mixed numbers	$\frac{7}{2} = \frac{3 \cdot 2+1}{2} = 3\frac{1}{2}$
CMI	Conv. to improp. fractions	$3\frac{1}{2} = \frac{3 \cdot 2+1}{2} = \frac{7}{2}$

Table 3: Misconceptions

Label	Description	Occurrence
MAD	$\frac{a}{b} + \frac{c}{d} = \frac{a+c}{b+d}$	14.8%
MSB	$\frac{a}{b} - \frac{c}{d} = \frac{a-c}{b-d}$	9.4%
MMT1	$\frac{a}{b} \cdot \frac{c}{b} = \frac{a \cdot c}{b}$	14.1%
MMT2	$\frac{a}{b} \cdot \frac{c}{b} = \frac{a+c}{b \cdot b}$	8.1%
MMT3	$\frac{a}{b} \cdot \frac{c}{d} = \frac{a \cdot d}{b \cdot c}$	15.4%
MMT4	$\frac{a}{b} \cdot \frac{c}{d} = \frac{a \cdot c}{b+d}$	8.1%
MC	$a\frac{b}{c} = \frac{a \cdot b}{c}$	4.0%

An example of student model is given in Figure 5. Dark grey nodes correspond to misconceptions, light grey nodes to elementary skills, nodes with no fill correspond to operational skills, and shaded nodes to application skills. We discuss the model selection process in Section 4.

[4]For each operational skill there is a corresponding application skill labeled with the prefix "A" in the student model. Application skills are not listed in Table 2.

3.2 Evidence models

For each task or question an evidence model is created. An example of a task is

$$T_1 \quad \left(\frac{3}{4} \cdot \frac{5}{6}\right) - \frac{1}{8} = \frac{15}{24} - \frac{1}{8} = \frac{5}{8} - \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$$

First, assume that a student is able to solve certain tasks if and only if she has the necessary skills and does not have certain misconceptions (later we will relax this assumption). Thus, we can describe the task T_1 formally by a logical formula:

$$T_1 \quad \Leftrightarrow \quad MT \,\&\, CL \,\&\, ACL \,\&\, SB \;\&\, \neg MMT3 \,\&\, \neg MMT4 \,\&\, \neg MSB \ .$$

Of course, the assumption of deterministic relations between skills and the actual outcome of a task is unrealistic. A student can make a mistake even if she has all abilities needed to solve a given task. On the other hand, a correct answer does not necessarily mean that the student has all abilities since she may guess the right answer, e.g. in a test where she is to select one answer from a given set of answers. However, it turned out to be reasonable to understand a task variable T_i as the ability to solve the corresponding task and to use a non-deterministic model only for the description of the dependence between the skill T_i and the actual outcome of the corresponding task X_i. Thus we can model "guessing" using conditional probability $P(X_i \mid \neg T_i)$ and "mistakes" using $P(\neg X_i \mid T_i)$. See Figure 6.

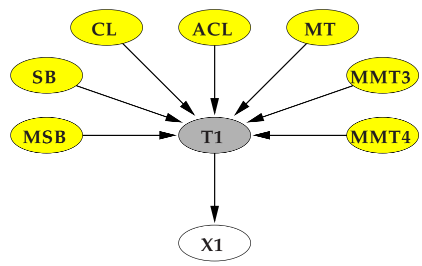

Figure 6: Evidence model of task T_1.

Conditional probability tables representing relations between variables in the student and evidence models can often be simplified, for example, by use of models of independence of causal influence. Examples

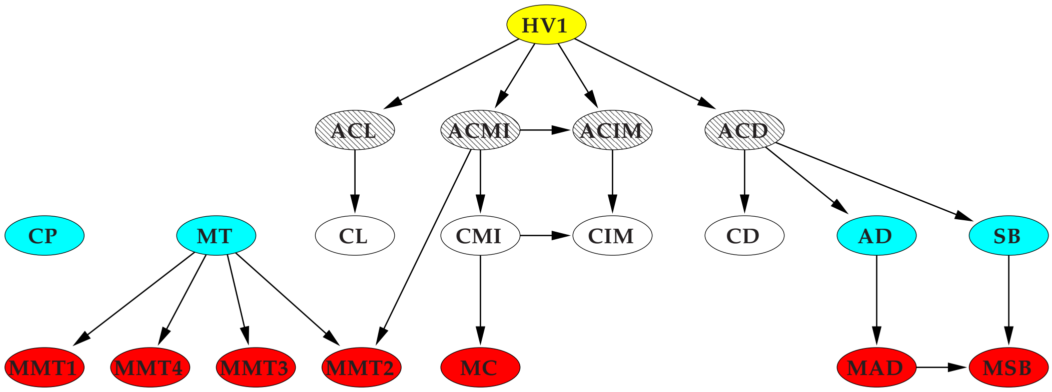

Figure 5: Student model describing relations between skills and misconceptions.

of such models are Noisy-OR or Noisy-AND. See Conati et al. (1997), Mislevy et al. (1999), and Almond et al. (2001) where design aspects of Bayesian network models used in educational testing and tutoring are discussed.

4 Experiments

4.1 Building models

We searched model structures using the PC-algorithm of Spirtes et al. (1993), implemented in Hugin (2002). At the beginning we run the PC-algorithm without any constraints on edges. It provided a first insight into the relations between skills and misconceptions. Using our ”expert knowledge” of the domain of fractions we explained some relations with the help of hidden variables and introduced certain constraints on edges.

An important parameter of the PC-algorithm is the significance level used in the independence tests. See Figure 7 from which one can see that significance level 5% appeared to be optimal also when learned models were compared using the Bayesian Information Criteria (BIC), a criteria derived by Schwarz (1978).

Applying different constraints on the resulting model we learned, using the PC-algorithm, nine different model structures, most of them containing hidden variables introduced by “domain experts”, see Table 4.

In most of the models certain edges were required to be present. We calibrated the

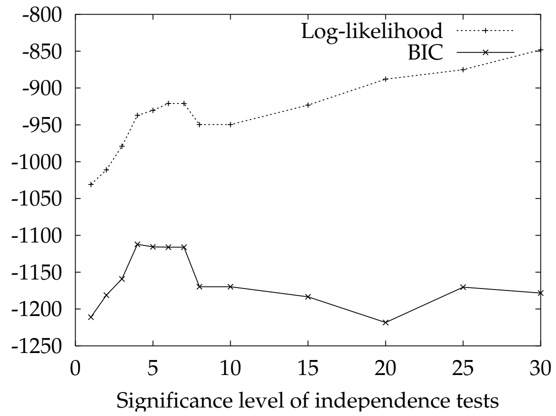

Figure 7: Comparison of models learned by the PC algorithm in Hugin with different significance levels.

overall models (composed from one student model and twenty evidence models) on a random subsample of 75 data vectors from the collected data consisting of 149 data vectors. Each data vector corresponds to the responses of one student. Parameter learning using the EM-algorithm was started from a non-uniform distribution to guide the EM-algorithm Lauritzen (1995) to reasonable parameters for hidden variables, the initial experience counts were very low (0.00001) to adjust the learned probability tables to the data.

Evidence models were designed by experts as described in Section 3.2, only the conditional probability distributions $P(X_i \mid T_i)$ were estimated from the collected data.

Table 4: Tested models.

(a)	model containing two hidden variables with several restrictions on presence or absence of edges,
(b)	model with two hidden variables and few restrictions,
(c)	model with one hidden variable and few restrictions, the model is given in Figure 5
(d)	model with few restrictions,
(e)	model with only obvious logical constraints as restrictions,
(f)	model with independent skills,
(g)	naive Bayes model with one hidden variable (with two states) being parent of all skills,
(h)	as above but the hidden variable has three states,
(i)	model without skills - all questions were children of one hidden variable with six states.

Table 5: Skill prediction quality.

model	4th	9th	14th	19th
(a)	84.8607	88.1737	89.2065	89.2883
(b)	86.0545	89.0925	89.9640	90.0198
(c)	86.5738	89.4621	90.1999	90.2671
(d)	86.4375	89.5329	90.1828	90.2500
(e)	85.3771	89.1015	89.8190	89.9293
(f)	84.3159	88.6483	89.4111	89.5164
(g)	84.6434	88.6397	89.3009	89.3886
(h)	85.0183	88.7521	89.3486	89.4263

4.2 Building tests from models

We used the results from the paper tests to test the models. We tested every model on remaining 74 data vectors that were not used for the calibration. Since each question was present only once in the test we could not allow questions to be repeated. Thus all experiments consisted of twenty questions only. We measured how well the models predicted the students' skills. For all nineteen skills we tested whether the most probable state (given observed answers) of each skill is equal to the true state observed in the data. The skill prediction quality is the percentage of skills that have the predicted state equal to the observed state. We repeated the whole procedure ten times on different subsamples of collected data. In Table 5 we compare models' skill prediction quality after 4th, 9th, 14th, and 19th questions in an adaptive test. Model (i) can not be tested on skill prediction since it does not contain skills in the model. The best model was model (c). Its structure was presented in Figure 5.

Table 6: Question prediction quality.

model	4th	9th	14th	19th
(a)	83.2703	90.5000	94.8851	99.1486
(b)	83.2331	90.0845	94.6655	99.1520
(c)	83.1757	89.8761	94.6126	99.1622
(d)	82.9662	89.8007	94.5321	99.1571
(e)	82.4122	89.7527	94.5081	99.1622
(f)	82.0991	89.7173	94.5394	99.1813
(g)	82.5145	89.8224	94.6052	99.1931
(h)	82.7264	89.8834	94.6250	99.1968
(i)	82.5173	89.7230	94.5345	99.1622

However, one can see there is not a significant difference between quality of models. We conjecture that it is because little cases (75) were available for learning, so that the more complex models could not be calibrated well. This conjecture can be supported by an observation: when we calibrated models using all 149 data records the difference between models became more significant. To see the importance of the calibration we also tested models that were not well calibrated by giving our original model experience count 50. We observed significantly worse predictions, e.g. after 4th answer it was only 72.6038 for the best model - model (c). In Table 6 we present comparisons of question prediction quality[5]. The differences between models are even smaller there.

We compared test design methods using the best model - model (c). The tested methods were: (1) the myopically optimal adaptive test, (2) the average fixed test constructed using the method described in Section 2, (3) fixed test where questions are taken in the reverse (descending) order as they were ordered in the paper tests given to the students, and (4) test where questions are in the ascending same order, i.e. as they were ordered in the paper tests given to the students.

In Figure 8 we can see how the quality of

[5] In order to provide a stable criteria for comparing the predictions of answers the presented number is the average computed from the predictions of all questions. It means that also the questions that were answered are included as if they could be asked again. Therefore the plots converge to 100% when all questions are answered.

skill prediction for different selection methods evolves with more questions being answered. The adaptive test provides best results. The average fixed test provides nearly as good results as the adaptive test. The two fixed tests are substantially worse, especially the one where questions are taken in the same order as they were ordered in the paper tests given to the students. Note that in the adaptive test we needed only six questions to reach prediction quality for which the fixed (with questions in the ascending order) required sixteen questions.

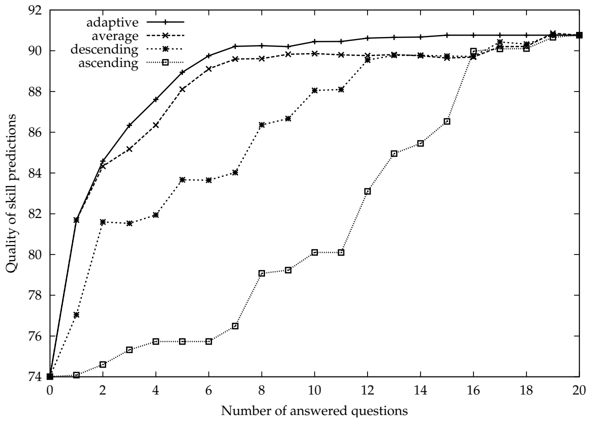

Figure 8: Quality of skill prediction

The criteria used to select next question in the adaptive test was the total expected entropy on skills. In Figure 9 we present how the actual total entropy on skills evolved. Observe that the quality of skill prediction and the total expected entropy are inversely proportional - the lower the entropy the better predictions.

To complete the comparisons we also present a plot showing how the quality of prediction of answers evolved for different selection methods in Figure 10. Our primary goal was not to make good predictions of answers, but it is natural that with better knowledge about students skills the model provides better predictions of answers.

5 Conclusions

We provided empirical evidence showing that educational testing can benefit from the application of Bayesian networks. We showed

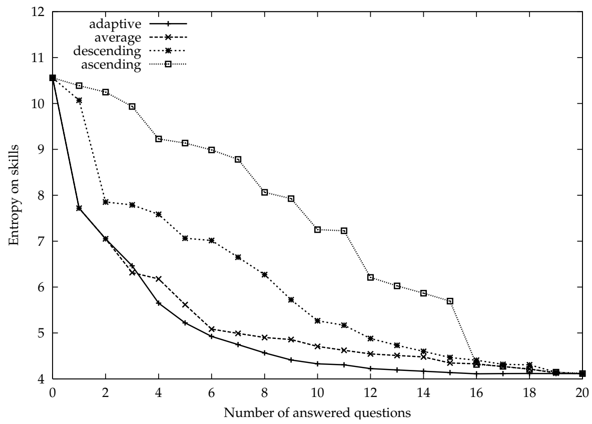

Figure 9: Total entropy of probability of skills

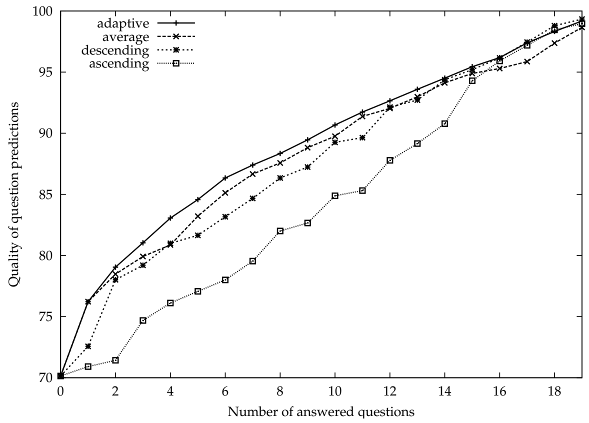

Figure 10: Quality of answers prediction

that adaptive tests may substantially reduce the number of questions that are necessary to be asked. We proposed a new method for the design of a fixed test, which provided good results on tested data. It may be regarded as a good cheap alternative to the computerized adaptive tests when they are not suitable. There are theoretical problems related to the application of Bayesian networks to educational testing that were not studied in this paper. One of them is efficient inference - a problem addressed in Vomlel (2002).

Acknowledgments

This paper is mainly based on results achieved when I was with the Department of Computer Science at Aalborg University, Denmark. I am grateful to Finn V. Jensen and the Decision Support Systems group at Aalborg University for the inspiring environment, Russell G. Almond and Robert J. Mislevy from Educational Testing Service (ETS) for interesting discussions during my visit at ETS, Frank Jensen and Anders L. Madsen for their assistance with Hugin, and Kirsten Bangsø Jensen for organizing the tests in Brønderselev High School. I was supported by the Grant Agency of the Czech Republic through grant number GA ČR 201/01/1482.

References

Russell G. Almond and Robert J. Mislevy. 1999. Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, 23(3):223–237.

Russell. G. Almond, Lou Dibello, Frank Jenkins, Deniz Senturk, Robert J. Mislevy, Linda S. Steinberg, and Duanli Yan. 2001. Models for conditional probability tables in educational assessment. In *Proc. of the 2001 Conference on AI and Statistics*. Society for AI and Statistics.

Moshe Ben-Bassat. 1978. Myopic policies in sequential classification. *IEEE Transactions on Computers*, 27(2):170–174.

Linas Būtėnas, Agnė Brilingaitė, Alminas Čivilis, Xuepeng Yin, and Nora Zokaitė. 2001. Computerized adaptive test based on Bayesian network for basic operations with fractions. Student project report, Aalborg University. `http://www.cs.auc.dk/library`.

Cristina Conati, Abigail S. Gertner, Kurt VanLehn, and Marek J. Druzdzel. 1997. On-line student modeling for coached problem solving using Bayesian networks. In Anthony Jameson, Cecile Paris, and Carlo Tasso, editors, *Proc. of the Sixth Int. Conf. on User Modeling (UM97), Chia Laguna, Sardinia, Italy*. Springer Verlag.

Ronald K. Hambleton and Hariharan Swaminathan. 1985. *Item response theory: Principles and applications*. Kluwer-Nijhoff, Boston.

Hugin. 2002. Hugin Explorer, ver. 6.0. Computer software. `http://www.hugin.com`.

Steffen L. Lauritzen. 1995. The EM-algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 1:191–201.

Wim J. Van Der Linden and Cees A. W. Glas, editors. 2000. *Computerized Adaptive Testing: Theory and Practice*. Kluwer Academic Publishers.

David Madigan and Russell G. Almond. 1996. On test selection strategies for belief networks. In D. D. Fisher and H. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics*, volume V. Springer Verlag.

Rob R. Meijer and Michael L. Nering. 1999. Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, 23(3):187–194. Special Issue: Computerized Adaptive Testing.

Robert J. Mislevy, Russell G. Almond, Duanli Yan, and Linda Steinberg. 1999. Bayes nets in educational assessment: Where do the numbers come from? In Kathryn B. Laskey and Henri Prade, editors, *Proc. of the Fifteenth Conf. on Uncertainty in AI*, San Francisco. Morgan Kaufmann Publishers, Inc.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 7(2):461–464.

Peter Spirtes, Clark Glymour, and Richard Scheines. 1993. *Causation, Prediction, and Search*. Number 81 in Lecture Notes in Statistics. Springer Verlag.

Martha L. Stocking and Charles Lewis. 1998. Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 23:57–75.

Martha L. Stocking and Len Swanson. 1993. A method for severly constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17:277–292.

Len Swanson and Martha L. Stocking. 1993. A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, 17:151–166.

Jiří Vomlel. 2002. Exploiting functional dependence in Bayesian network inference. In *Proc. of the* 18^{th} *Conf. on Uncertainty in AI*.

Howard Wainer, David Thissen, and Robert J. Mislevy. 2000. *Computerized adaptive testing : a primer*. Mahwah, N.J. : Lawrence Erlbaum Associates, second edition.

Abraham Wald. 1947. *Sequential Analysis*. Wiley, New York.