

Adaptive Test Design with a Naive Bayes Framework

Michel C. Desmarais, Alejandro Villarreal, and Michel Gagnon

Computer and Software Engineering Department

Polytechnique Montréal

{michel.desmarais, alejandro.villarreal, michel.gagnon}@polymtl.ca

Abstract. Bayesian graphical models are commonly used to build student models from data. A number of standard algorithms are available to train Bayesian models from student skills assessment data. These models can assess student knowledge and skills from a few observations. They are useful for Computer Adaptive Testing (CAT), for example, where the test items can be administered in order to maximize the information they will provide. In practice, such data often contains missing values and, under some circumstances, missing values far outnumber observed values. However, when collecting data from test results, one can often choose which values will be present or missing by a consequent test design. We study how to optimize the choice of test items for collecting the data that will be used for training a Bayesian CAT model, such as to maximize the predictive performance of the model. We explore the use of a simple heuristic for test item choice based on the level of uncertainty. The uncertainty of an item is derived from its initial probability of success and, thus, from its difficulty. The results show that this choice does affect model performance and that the heuristic can lead to better performance. Although the study's results are more exploratory than conclusive, they suggest interesting research avenues.

1 Introduction

Applications such as study guides [6] and adaptive tutoring [7, 10] must rely on a fine grained student model to tailor their interaction with the user [2]. Bayesian models data can be used for this purpose and they can be trained with skills assessment data to overcome a knowledge engineering effort [12, 13, 9, 4]. In many contexts, such as Computer Adaptive Testing (CAT), this data will contain missing values. We investigate the problem of choosing which items will be administered in each test with the aim of optimizing the choice of missing values distribution.

In CAT, this problem is often referred to as sampling design or test design/construction. See for example [1, 8, 11]. The goal of student modeling in CAT is to build a statistical model of an examinee's chances of success or failure to a test based on previous observations of correct or incorrect answers to a subset of test items. For a number of

practical reasons, the pool of test items often needs to be quite large, such as a few hundreds of items. However, for model training, it is impractical to administer a test of hundreds of questions to examinees in order to gather the necessary data, as such test could last tens of hours. We are thus forced to administer a subset of these test items to each examinee and, hence, we run into the problem of choosing how to distribute the test items in order to maximise the information we get for model building.

The work on test design is generally conducted within the Item Response Theory framework. We investigate this issue in the context of the Bayesian framework.

The next section describes the Bayesian model used for this study. Then, a few heuristics are proposed to design the missing values scheme that we hope will bring the most relevant information for model building given a fixed number of observations. Experimental studies of the performance of these heuristics are later reported and further studies are proposed.

2 The POKS Naive Bayes Framework

To investigate the issue of test design within the Bayesian framework, we use the POKS framework [4] (Partial Order Knowledge Structures). This model is fully automated and it was shown to perform well compared to other Bayesian student frameworks that involve a knowledge engineering effort, while providing much finer grained assessment than IRT [3].

This model is based on the Naive Bayes conditional independence assumption.

The POKS model aims to predict the outcome of each individual item based on the observed items. For each item, the model determines which set of evidence items, \mathbf{X}_e , are considered relevant to determine success to item X_c :

$$P(X_c|\mathbf{X}_e) = P(X_c|X_1, X_2, \dots, X_k)$$

where $\{X_1, X_2, \dots, X_k\}$ is the set of evidence items considered as relevant. Given the conditional independence assumption mentioned, that equation becomes:

$$P(X_c|\mathbf{X}_e) = \frac{P(X_c)}{P(X_1, X_2, \dots, X_k)} \prod_i^k P(X_i|X_c)$$

To determine which items are included in \mathbf{X}_e three statistical tests are applied for each possible pairs of nodes, (X_c, X_i) . Two of these tests aim to verify the strength of the conditional probabilities $P(X_e|X_c) \geq p_c$ and $P(\bar{X}_c|\bar{X}_e) \geq p_c$, and another one to verify the interaction between X_c and X_e . See [4] for details.

This approach has been shown to perform at least as well as the IRT framework for predicting global test mastery [5] and also provides a fine grained student model.

3 Heuristics for Effective Partial Data Sampling Scheme

As explained above, the context of adaptive testing is a typical case where we have the opportunity to decide upon a specific scheme of missing values for each item. We can decide which subset of questions we wish to administer to each examinee, leaving unanswered items as missing values. For a pool of hundreds of question items, which is a size often found in CAT, each examinee can be administered only 50 question items, leaving many more missing values than answered items.

Given that we have the choice of how many observations will be assigned to each item, we need to determine which item are most critical and, potentially, ought to be allotted more observations.

Without knowing in advance the topology of the POKS networks, or the topology of a Bayes Network over the items, we have to revert to simpler means of choosing critical items. Potential candidates are the highly uncertain items that have an initial probability of 0.5. They correspond to items of average difficulty. To the opposite spectrum, we can also favor choosing low uncertainty items that have a high or a low initial probability. They correspond to the most difficult and the easiest items. In this study, we do not discriminate between easy and difficult items, as, in terms of uncertainty, or entropy, they are equivalent. Of course, further investigations could explore such distinction.

Initial probabilities can be obtained from relatively small samples, in contrast to joint probabilities, so it is reasonable to assume that a pilot sample can provide these with sufficient reliability to guide further sampling.

We have conducted a simulation study of such a sampling scheme. The details of the experimental conditions and the results are described below.

4 Experimental Design

We define three sampling schemes to determine missing values in order to investigate their respective effects over the predictive accuracy of POKS Bayesian models :

1. *Uniform*: Uniform random samples of missing values.
2. *Most uncertain*: Higher sampling rate of missing values for uncertain items (favoring the choice of average difficulty items).
3. *Least uncertain*: Lower sampling rate of missing values for uncertain items (favoring the choice of difficult and easy items alike).

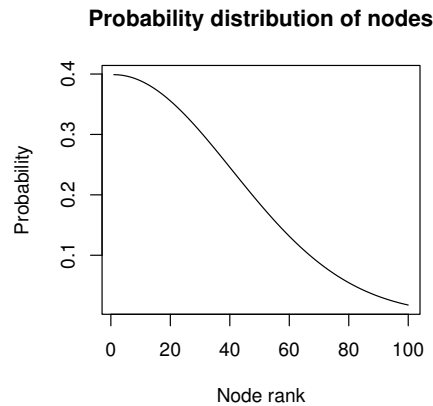


Figure 1: Sampling probability distribution of items used for the *most uncertain* and *least uncertain* sampling schemes.

Uncertain items are the items that, obviously, are closest to an initial probability of 0.5. For the *most uncertain* and *least uncertain* conditions, the probability of sampling is based on the $x = [0, 2.5]$ segment of a normal (Gaussian) distribution as reported in Figure 1. The probability of an item being sampled will therefore vary from 0.40 to 0.0175 as a function of its rank, from the most to the least uncertain item on that scale. Items are first ranked according to their uncertainty and they are attributed a probability of being sampled following this distribution. The distributions are the same for both conditions (2) and (3), but the ranking is reversed between the two of them. For the *uniform* condition (1), all items have equal probability of being sampled.

Ten samples are created according to the three sampling schemes above. They are used to validate the effect of the sampling scheme by performing CAT simulations and measuring the predictive power of the models based on different sampling schemes.

4.1 Simulation process

The experiment consists in simulating the question answering process with the real subjects. An item is chosen and the outcome of the answer, success or failure, is fed to the inference engine (POKS). An updated probability of success is computed given this new evidence. All items for which the probability is above 0.5 are considered mastered and all others are considered non-mastered. We then compare the results with the real answers to obtain a measure of how accurate the predictions are. The process is repeated from 0 item administered until all items are “observed”. Observed items are bound to their true value, such that after all items are administered, the score always converges to 100%.

The simulations replicate a context of computer adaptive testing (CAT) where the system chooses the question items in order to optimize skills assessment. The choice of item relies on a measure of the most informative question that gets administered to the examinee. This can be achieved in a number of ways and the results are often relatively close. We

use a heuristic that our exploratory results has shown to approach the performance of the information gain approach (see [5]), but which is computationally much faster. It consists in choosing the item i that has a high entropy and is highly connected to other nodes:

$$\max_i E(i) \frac{\log(links(i + 1) + E(0.5))}{E(0.5)}$$

where $E(i)$ is the entropy of item i and $links(i)$ is the number of incoming and outgoing links. The use of the maximal entropy of a item, $E(0.5)$, in the above equation is a normalizing factor that ensures the weights between the number of links and the entropy are similar.

Once the outcome of the answer to a question item is obtained, the probability of success to each other questions is then recalculated according to the POKS framework described above.

Simulations consist in ten-fold cross-evaluation runs. Each run consists of a different random sampling for test design (the choice of items according to the three schemes described in section 4) and a different random sampling of the examinee used for training and testing. We report the average results of the 10 simulations for each experimental condition.

4.2 Data sets

Four data sets are used for the simulations. They are based on real data from tests in four different domains :

Table 1 reports general statistics on these data sets as well as the sizes of the training and testing samples used for the simulations.

Table 1: Data sets

	Data Set	Set size		Number of items	Average success rate
		Training	Testing		
1	College math	375	56	60	61%
2	UNIX	30	18	34	53%
3	Arithmetic	100	49	20	61%
4	French	25	17	160	58%

The proportion of missing values inserted in the training set is half of the data. The testing data sets contain no missing values.

1. *College math*: a 60 question items test covering different topics in mathematics, from general high school math, to college level geometry, linear algebra, and calculus. The test was administered to 426 candidates newly admitted at an engineering school.

2. *UNIX*: a 34 question items test covering knowledge of the UNIX shell commands, from the basic “change directory” (`cd`) to advanced data manipulation commands with `awk`. The test was administered to 48 individuals with a wide variety of knowledge about UNIX.
3. *Arithmetic*: a 20 questions test on basic fraction arithmetic. The test was administered to 149 pupils in grade 10 to 12. More details can be found in [13].
4. *French*: a 160 general French grammar, reading, and comprehension test administered to 42 adults.

5 Results

The results of the simulation experiments are reported in Figures 2(a) to 2(d). The Y axis represents the proportion of correct predictions while the X axis reports the number of items administered. As mentioned, administered items are considered correctly classified, and thus, after all items are administered, the score reaches 100%. Given that the items are initialized to their unconditional probabilities, the prediction score generally starts above 70%, which indicates that more than two thirds of items are already correctly classified initially. A 90% confidence interval over the 10 simulations is reported around each data point.

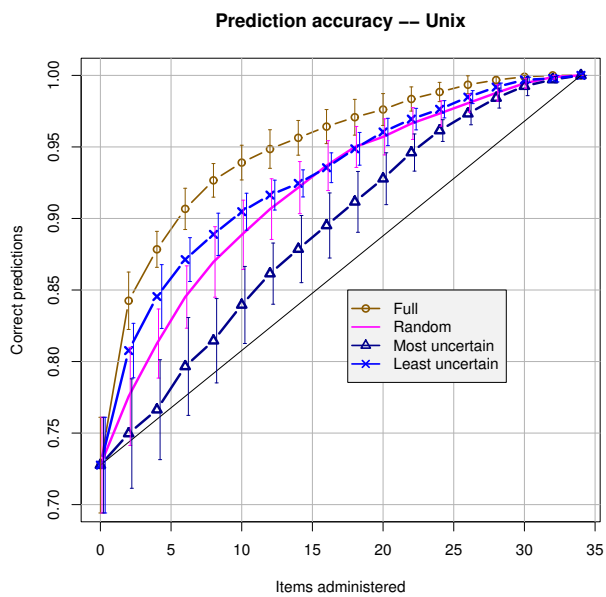
Each figure contains four curves. For comparison purpose, we report the *Full* condition which corresponds to the results for the full data set, without any missing values. The other three conditions are described in section 4.

The results show non significant differences for the *Arithmetic* and *College* math data sets. However, more significant differences are observed for the other two data sets (*UNIX* and *French*), and they follow a regular pattern: The *least uncertain* condition systematically outperforms the *most uncertain* condition, which, in turn, performs systematically worst than the *uniform condition*. As expected, the *full* data set is systematically better than, or equal to, the data sets that contains missing values.

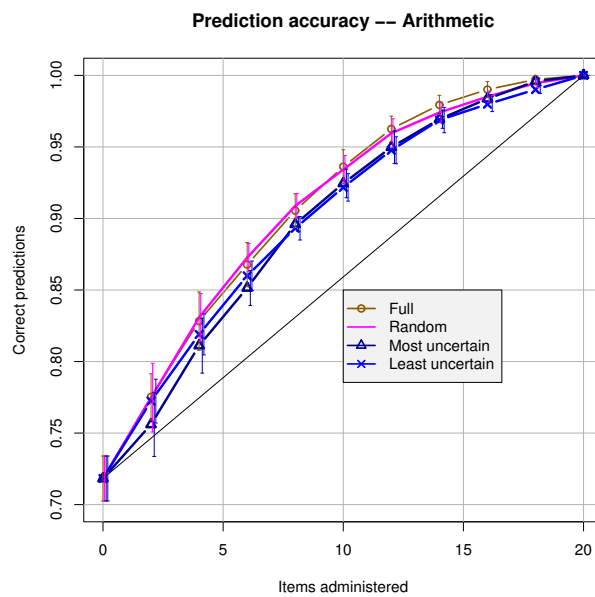
6 Discussion

These results suggest that higher sampling rates for the *least uncertain* items generally bring a higher predictive performance than for the *most uncertain* or the *uniform* choice, although this gain is not systematic.

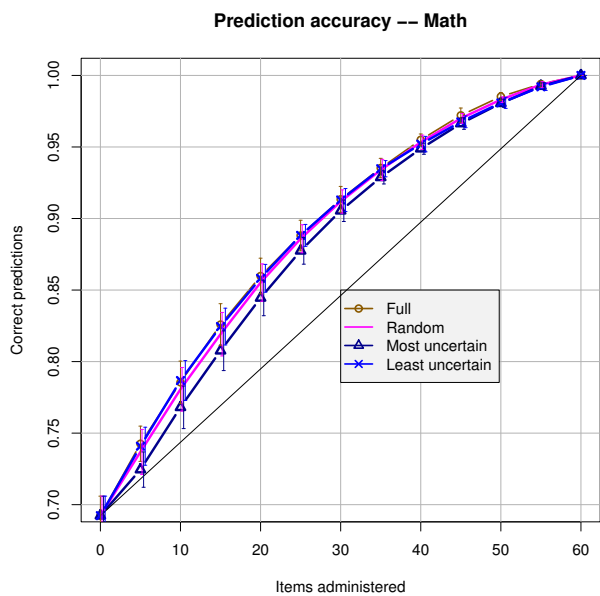
These results remain exploratory and a number of questions are left open. For one, how should we explain the patterns of differences found between the *least uncertain*, and the *most uncertain*? We initially hypothesized that the most uncertain items are the ones that would benefit the most from a higher sampling frequency. These items are generally the ones that bring the most information, and it seems reasonable to gather more data for them to correctly establish their relations to other items. However and contrary to these



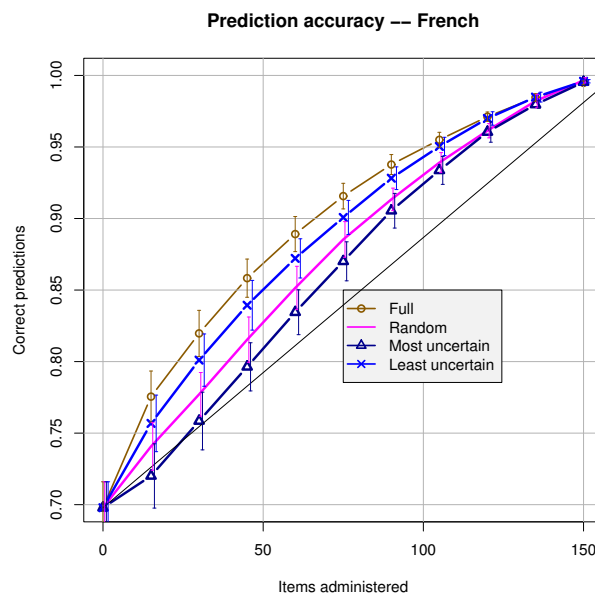
(a) UNIX



(b) Arithmetic



(c) College math



(d) French

Figure 2: Results of the four data test. 90% confidence intervals are displayed around each data point.

expectations, higher sampling of uncertain items yields the models with the poorest performance. One plausible explanation is that the estimation of the model's conditional probabilities is more subject to noise and to miscalibration for probabilities closer to 0 or 1 than for mid-range probabilities. As a consequence, a higher sampling rate for these items is required. This hypothesis also explains why we observe a large difference between *least uncertain* and *most uncertain* for small data sets (UNIX and French) than for the larger ones (College math and French): larger data sets are not as subject to sampling noise as smaller ones are.

Another open question is whether these results apply to other domains and to other Bayesian frameworks for student modeling. For example, would the results be the same if we used a more general Bayesian Network approach that captures independence relations, such as in [13]?

A potentially interesting question to investigate is the hypothesis that the items that would be most informative are the ones that are central and highly connected in a Bayesian Network, that is, the nodes that are likely to influence the greatest number of nodes in the network. These nodes could benefit from a higher sampling rate. Moreover, given an initial topology of a Bayesian Network, we could guide the sampling beyond individual nodes, to pairs or to n-tuples of nodes that are deemed more critical.

However, the topology of a Bayesian network might not be reliably established with small sample sizes, in contrast to the heuristic that we used which is based on estimating the individual items non conditional probabilities: they require relatively small sample size. It is feasible to design the tests with an initial sample of a few tens of data records, and then collect a larger sample for estimating the conditional, joint probabilities. Whether this can be effectively done for a Bayesian Network remains open.

Further analysis and investigations are obviously required to bring some understanding to these results. Nevertheless, this investigation shows that we can influence the predictive performance of a Naive Bayes framework with partial data when we have the opportunity to select the missing values. It opens interesting questions and can prove valuable in some contexts of application.

References

- [1] Berger, M. P. F. A general approach to algorithmic design of fixed-form tests, adaptive tests, and testlets. *Applied Psychological Measurement* 18, 2 (1994), 141–153.
- [2] Bra, P. D., Brusilovsky, P., and Houben, G.-J. Adaptive hypermedia: from systems to framework. *ACM Comput. Surv.* 31, 4es (1999), 12.
- [3] Desmarais, M. C., Gagnon, M., and Meshkinfam, P. Item to item student models. In *AAAI'2006 Workshop 06, Data mining in education* (Sept. Boston MA 2006), p. 10.

- [4] Desmarais, M. C., Maluf, A., and Liu, J. User-expertise modeling with empirically derived probabilistic implication networks. *User Modeling and User-Adapted Interaction* 5, 3-4 (1996), 283–315.
- [5] Desmarais, M. C., and Pu, X. A bayesian inference adaptive testing framework and its comparison with Item Response Theory. *International Journal of Artificial Intelligence in Education* 15 (2005), 291–323.
- [6] Falmagne, J.-C., Cosyn, E., Doignon, J.-P., and Thiéry, N. The assessment of knowledge, in theory and in practice. In *ICFCA (2006)*, R. Missaoui and J. Schmid, Eds., vol. 3874 of *Lecture Notes in Computer Science*, Springer, pp. 61–79.
- [7] Heller, J., Steiner, C., Hockemeyer, C., and Albert, D. Competence-based knowledge structures for personalised learning. *International Journal on E-Learning* 5, 1 (2006), 75–88.
- [8] Henson, R., and Douglas, J. Test construction for cognitive diagnosis. *Applied Psychological Measurement* 29, 4 (2005), 262–277.
- [9] Millán, E., de-la Cruz, J.-L. P., and Suárez, E. Adaptive Bayesian networks for multilevel student modelling. In *ITS'00: Proceedings of the 5th International Conference on Intelligent Tutoring Systems (2000)*, Springer-Verlag, pp. 534–543.
- [10] Millán, E., Garcia-Herve, E., Rueda, A., and de-la Cruz, J. P. Adaptation and generation in a web-based tutor for linear programming. *Lecture Notes in Computer Science* 2722 (January 2003), 124–127.
- [11] Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement* 29, 2 (The National Assessment of Educational Progress (Summer, 1992) 1992), 133–161.
- [12] Pardos, Z. A., Feng, M., Heffernan, N. T., Lindquistheffernan, C., and Ruiz, C. Analyzing fine-grained skill models using bayesian and mixed effects methods. In *Workshop of Educational Data Mining (Los Angeles 2007)*.
- [13] Vomlel, J. Bayesian networks in educational testing. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* 12, Supplementary Issue 1 (2004), 83–100.