# Student Models Construction by Using Information Criteria

Maomi Ueno
*Nagaoka University of Technology*
*ueno@kjs.nagaokaut.ac.jp*

## Abstract

*This paper proposes a construction method of Student models for Intelligent Tutoring Systems(ITSs) by using information criteria. This proposal provides a method to automatically construct the optimum Student model from data. The main problem when the traditional information criteria are employed to construct a model is that large amount of data, which are difficult to obtain in actual school situations, need to be obtained. This paper proposes a new criterion for using a smaller amount of data by utilizing a teacher's expert knowledge. Concretely, 1) the general predictive distribution is derived, and 2) the determination method of the hyper parameters by using a teacher's expert knowledge is proposed. Finally, some Monte Carlo experiments comparing some information criteria (ABIC, BIC, MDL, and the exact predictive distribution) are performed. The results show that the proposed method provides the best performance.*

## 1. Introduction

Over the last few years, a method of reasoning using probabilities[1],[2] variously called Bayesian networks, belief networks, causal networks, and so on , has become popular within the Intelligent Tutoring System community. For example, in [3], the knowledge states diagnosis system is based on belief networks and decision theory. In [4],[5], [6],[7] ,[8]besides the diagnosis system, updating is concerned with the expected changes in student knowledge due to tutoring.

However, these methods subjectively constructed the student model structure without using students' response data. If the students' response data is available, then it helps our decision making for the student model construction. This paper proposes a construction method of Student models for Intelligent Tutoring Systems (ITSs) by using information criteria. The main problem when the traditional information criteria are employed to construct a model is that large amount of data, which are difficult to obtain in actual school situations, need to be obtained. This paper proposes a new criterion for using a smaller amount of data by utilizing a teacher's expert knowledge. Concretely, 1) the general predictive distribution is derived, and 2) the determination method of the hyper parameters by using a teacher's expert knowledge is proposed. Finally, some Monte Carlo experiments comparing some information criteria (ABIC, BIC, MDL, and the exact predictive distribution) are performed. The results show that the proposed method provides the best performance.

## 2. Representation of the student model by the Belief networks

In this paper, the student model is defined by belief networks. Let $X = \{X_1, X_2, \Lambda, X_N\}$ be a set of $N$ variables which represent students' knowledge states; each can take $r_i$ states in the set $\{1, \Lambda, r_i\}$. $x_i = k$ is written when it is observed that variable $x_i$ takes $k$. $p(x = k \mid y = j, \xi)$ is used to denote the probability of a person with background knowledge $\xi$ for the observation $x = k$, given the observation $y = j$. A student model is represented as a pair of knowledge structure $S$ and a set of conditional probability parameters $\Theta$, $(S, \Theta)$. An example of knowledge structure $S$ in the domain "the solution of the linear equation" in a junior high school is shown in Figure.1. Here, $A \longrightarrow B$ indicates that we have to acquire knowledge A in order to acquire knowledge B. The nodes which depends on the target node are called "parent nodes" of the target node. In addition, a set of problems which correspond to the nodes has to be prepared. If a student provides a correct

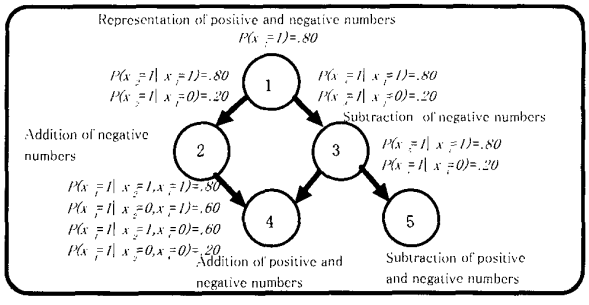


Figure.1 An example of Student model represented by belief networks



0-7695-1013-2/01 $10.00

answer to $i$-th problem, $x_i = 1$ (he has *i-th* knowledge) ,otherwise, $x_i = 0$ . A joint probability distribution over the network is shown by

$$P(X_1, X_2, \Lambda, X_N \mid S) = \prod_{i=1}^{N} p(x_i \mid \Pi_i, S) \quad , \quad (1)$$

where $\Pi_i \subseteq \{x_1, x_2, \Lambda, x_{q_i}\}$ is a set of parents nodes that renders $x_i$ and $\{x_1, x_2, \Lambda, x_{q_i}\}$ conditionally independent. In particular, $S$ is a directed acyclic graph such that (1) each variable corresponds to a node in S, and (2) the parents of the node corresponding to $x_i$ are the nodes corresponding to the variables in $\Pi_i$.

The next section will propose a method to automatically construct the student model structure from the students' response data.

## 3. Estimation of conditional parameters and posterior distribution

This paper proposes a method to automatically construct the student model from the data. It is necessary to define the model in (1) as a statistical model in which parameters are estimated. Now, consider a database $\mathbf{X} = \{x_{sik}\}, (s = 1, \Lambda, n, i = 1, \Lambda, N, k = 0, \Lambda, r_i - 1)$, where $x_{sik} = 1$ when the $s$-th student takes $k$-th knowledge states about the $i$-th node, otherwise $x_{sik} = 0$. Let $\theta_{ijk}$ be a conditional probability parameter of $x_i = k$ when $\Pi_i = j$, then the following likelihood function can be obtained.

$$p(\mathbf{X} \mid \Theta_S, S) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{n_{ijk}}{\sum_{\bar{k}=0}^{1} n_{ijk}} \prod_{k=0}^{1} \theta_{ijk}^{n_{ijk}} \quad , \quad (2)$$

where

$\Theta_S = \{\theta_{ijk}\}, (i = 1, \Lambda, N, j = 1, \Lambda, q_i, k = 0,1)$ , and it is assumed that the prior distribution has a Dirichret distribution ,which is a conjecture distribution of (2). That is,

$$p(\Theta_S \mid S) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\sum_{k=0}^{1} n'_{ijk})}{\prod_{\bar{k}=0}^{1} \Gamma(n'_{ijk})} \prod_{k=0}^{1} \theta_{ijk}^{n'_{ijk}-1} \quad , \quad (3)$$

where $\Gamma()$ indicates Gamma function. Then, the following posterior distribution is obtained,

$$p(\mathbf{X}, \Theta_S \mid S) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\sum_{k=0}^{1} n'_{ijk} + n_{ijk} - 1)}{\prod_{\bar{k}=0}^{1} \Gamma(n'_{ijk} + n_{ijk} - 1)} \prod_{k=0}^{1} \theta_{ijk}^{n'_{ijk}+n_{ijk}-1} \quad (4)$$

Therefore, the following maximum a posterior estimator can be derived ;

$$\hat{\theta}_{ijk} = \frac{n'_{ijk} + n_{ijk}}{n'_{ij} + n_{ij}} \quad , \quad (5)$$

where $n'_{ij} = \sum_{k=0}^{1} n'_{ijk}$ , $n_{ij} = \sum_{k=0}^{1} n_{ijk}$ .

Thus, the conditional probability parameters $\hat{\theta}_{ijk}$ and the posterior distribution $p(\mathbf{X}, \Theta_S \mid S)$ can be estimated from (4) and (5).

## 4. Student models construction by using information criteria

Let us consider a structure with just three nodes, then there are eight possible structures as shown in Figure 2. The problem is how to find the optimum structure of the student model. In this case, it is well known that information criteria are useful. Since Akaike's criterion[10], various criteria have been proposed. (For example, ABIC[10], BIC[11], MDL[12], and so on). The student model construction problem in this paper employs Bayesian approach, it is considered that employing Bayesian information criteria, ABIC, BIC, MDL, and so on, is valid.

ABIC is given by

$$\text{ABIC(model)} = -2\log p(\mathbf{X}, \Theta_S \mid S) + 2K. \quad (6)$$

BIC and MDL have the same formulation by

$$\text{BIC(MDL) (model )} = -2\log p(\mathbf{X}, \Theta_S \mid S) + 2K\log n. \quad (7)$$

Finding the structure is completed by minimizing these criteria. However, these criteria are generally derived by approximating the posterior distribution or predictive distribution, then, it is considered to expect better results by deriving directly a criterion from the belief networks model. For this motivation, Cooper [9] derived the

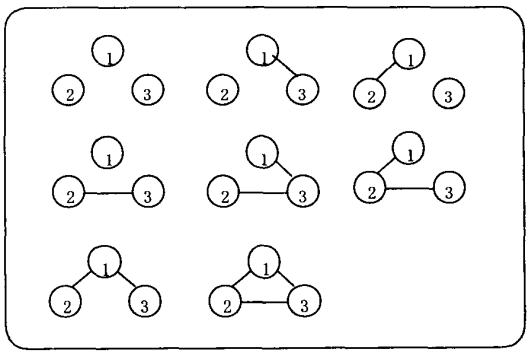


Figure 2. Possible structures of the student models with three nodes

predictive distribution of the belief networks when the prior distribution has a unique distribution. This criterion is given by

$$p(\mathbf{X}, S \mid \Theta_S) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(n_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} n_{ijk}! \quad (8)$$

It should be noted that the optimum structure is obtained by maximizing this criterion. The next interest is which criterion is best for the student model construction problem. Here, some Monte Carlo studies are demonstrated. For simplicity, Figure 1 is considered as the true model. The random data is generated from the Figure 1, and the sample sizes are 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000. 1000 realization for each sample size were generated. For each realization, the relative performances of the criteria mentioned before can be compared The results, the number of times when the criteria selects the true structure among 1000 iterations for each sample size are shown in Table 1

Table 1. The results of the Monte Carlo experiments

| Sample sizes | ABIC | BIC(MDL) | Cooper |
|---|---|---|---|
| 50 | 205 | 75 | 18 |
| 100 | 423 | 347 | 222 |
| 200 | 486 | 641 | 487 |
| 300 | 494 | 786 | 634 |
| 400 | 497 | 908 | 752 |
| 500 | 498 | 951 | 782 |
| 600 | 487 | 972 | 835 |
| 700 | 480 | 983 | 870 |
| 800 | 458 | 988 | 872 |
| 900 | 462 | 988 | 881 |
| 1000 | 424 | 989 | 897 |

From the table, BIC or MDL shows the best performances for large sample sizes, and ABIC shows the best performance for small sample sizes. In an actual school situation, it is difficult to gather a large sample of data, in this sense, it is considered that ABIC provides the best performance. However, for a small sample (100), ABIC selects the true structure with a probability of 0.42 at most. Then, this paper proposes more effective method to construct the student model in the next section.

## 5. Student models construction by using teacher's expert knowledge

### 5.1. The general predictive distribution

All information criteria mentioned in section 4 assumes that the prior distribution, which reflects prior knowledge about the student model, has a uniform distribution. However, in education, most teachers have prior knowledge about the student model. The main idea of this paper is to develop an efficient information criterion by integrating a teacher's expert knowledge into the prior distribution. To realize this idea, an exact general predictive distribution with various prior distributions should be derived. From the assumptions in this paper, the general predictive distribution can be derived as follow: That is, from (4), the predictive distribution is given by

$$p(\mathbf{X} \mid S) = \prod_{i=1}^{N} \frac{\Gamma(n'_{ijk})}{[\Gamma(n'_{ijk})]} \frac{\prod_{j}^{q_i} \Gamma(n'_{ijk} + n_{ijk})}{\Gamma(n'_{ijk} + n'_{ijk})}. \quad (9)$$

It should be noted that the general predictive distribution has the hyper parameter $n'_{ijk}$ . In fact, this hyper parameter acts the most important role. The predictive distribution converges to various information criteria by changing the value of the hyper parameter. When $n'_{ijk} = 1$ (the prior distribution has the unique distribution shown in Figure 3), the predictive distribution converges to Cooper's criterion, although it is natural from the definition. When $n'_{ijk} = 1/2$ (the prior distribution has the U distribution shown in Figure 3), the predictive distribution converges to BIC, or MDL. Moreover, when $n'_{ijk} = -\log\theta_{ijk} + 1/2$ (the prior distribution has the convex distribution shown in Figure 3), the predictive distribution converges to AIC. It should be noted that BIC, or MDL and Cooper's criterion assume stronger penalties for the complexity of the model, which is the number of parameters, than one of AIC. Then, AIC has a tendency to select a structure with more arcs than the true

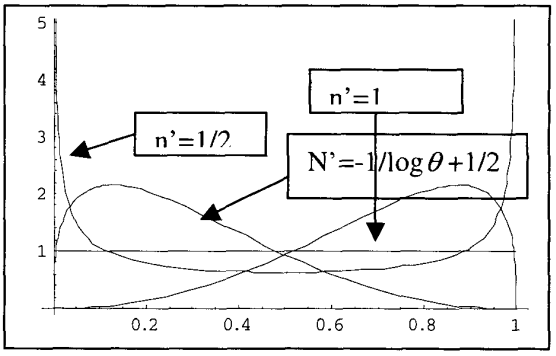


Figure 3. Prior distributions for various hyper parameters

structure, and BIC or MDL and Cooper's criterion have a tendency to select a structure with less arcs than the true structure.

### 5.2. Integration of teacher's knowledge

By using these properties, utilizing the teacher's prior knowledge about student model into the predictive distribution (9) can be considered. The procedure is as follows: 1) A teacher constructs a student model structure by using his or her expert knowledge, 2)based on this

structure, let the value of the hyper parameter for the arcs which is considered to exist be $n'_{ijk} = -\log\theta_{ijk} + 1/2$, and let the value of the hyper parameter for the arcs which is not considered to exist be $n'_{ijk} = 1/2$. Now, consider three possible structures, as prior knowledge structures, A, B, C concerning the structure in Figure 1 as follows: A is the structure with full arcs, B is the true structure, and C is the structure with no arc. Let consider three cases of which each structure is considered as a prior knowledge about the student model. The same Monte Carlo studies as section 4 are demonstrated in Table 2. If a teacher knows the true structure, the criterion acts more exactly than the traditional criteria for small sample sizes. Moreover, if a teacher has a wrong knowledge as structures A and B, the proposed criterion acts the same as the traditional criteria.

Table 2. The results of the Monte Carlo experiments by using prior knowledge

| Sample sizes | Structure A | Structure B | Structure C |
|---|---|---|---|
| 50 | 205 | 982 | 75 |
| 100 | 423 | 972 | 347 |
| 200 | 486 | 1000 | 641 |
| 300 | 494 | 1000 | 786 |
| 400 | 497 | 1000 | 908 |
| 500 | 498 | 1000 | 951 |
| 600 | 487 | 1000 | 972 |
| 700 | 480 | 1000 | 983 |
| 800 | 458 | 1000 | 988 |
| 900 | 462 | 1000 | 988 |
| 1000 | 424 | 1000 | 989 |

## 7. Application

By using the data from 294 junior high school students data, the student model in a domain of a simple equation is estimated from (7). Expert knowledge is employed as prior knowledge, although it is omitted for want of space. It is known that the obtained structure is reasonable considering the meanings of the nodes.

## 8. Conclusions

This paper proposed a new information criterion for the Student model construction problem by using a teacher's expert knowledge. The Monte Carlo experiments showed the efficiency of the proposed model.

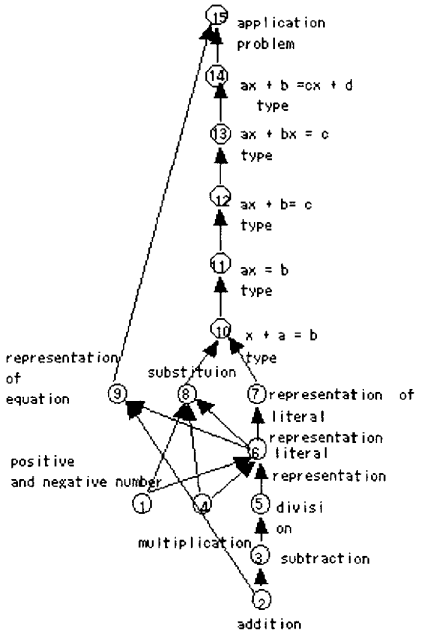


Figure 4. An example of estimated student model


## References

[1] J. Pearl, "*Probabilistic Reasoning in Intelligent System*", Morgan Kaufman Publishers, California, 1988.

[2] R.E. Neapolitan, "*Probabilistic Reasoning in Expert System*", John Willey & Sons, INC, New York, 1990.

[3] M. Ueno, "*A test theory using probabilistic network*", Electronics and Communications in Japan, 78(5), John Wiley & Sons, Inc. ,1996.

[4] J. Reya, "Two-Phase updating of student models based on daynamic belif networks", *Proceedings of 4th Internationa Conference of Aritificial Inttelegence*, 1998, pp, 274-283.

[5] A.H. Hashimoto *et al.*,"A probabilistic approach for student modeling task",*Proceeding of the 13th Annual* Conference of JSAI, Tokyo, 1998. pp.68-69

[6] M. Ueno, "Environments of learning by using Internet, Meta knowledge navigation system", *Proceedings of ICCE 99*, IOS press, Tokyo, 1999, pp.748-751

[7]J. Martin & K.VanLehn," Student assessment using Bayesian nets. International" Journal of Human-Computer Studies, Vol. **42**.,. Academic Press.1995, pp. 575-591

[8]K.VanLehn, & J. Martin, "Evaluation on an assessment system based on Bayesian student modeling.", International Journal of Artificial Intelligence and Education, Vol.**8**.2, 1998

[9] G.F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data ", *Machine Learning* , **9**, 1990, pp.54-62.

[10] H. Akaike, "A Bayesian extension of the minimum AIC procedure of an autoregressive model fitting", *Biometrika*, **66**, 237-242.

[11]G. Schwarz, "Estimating the dimension of a model ", Annals of Statistics **6**, 451-454

[12] J. Rissanen, "A universal prior for integrals and estimation by minimum description length", Annals of Statistics, **11**, 416-431