Bayesian Network Approach to Computerized Adaptive Testing

Kyung Soo Kim[1] and Yong Suk Choi[2]*

[1]Department of Electronics and Computer Engineering
Hanyang University, Seoul, Korea
[2]Division of Computer Science and Engineering
Hanyang University, Seoul, Korea
kyungskim@hanyang.ac.kr, cys@hanyang.ac.kr

Abstract

For the personalized learning, a good testing method, which can effectively estimate a learner's proficiency, is required. In this paper, we propose a novel testing method, Bayesian network-based approach to Computerized Adaptive Testing (CAT). Our novel approach can estimate proficiency of the examinee effectively and efficiently because it reflects complicated relationships between all items and their categories, and can estimate detailed proficiency about each specific category. In experimental results, we show that our approach can improve accuracy and speed of estimating examinee's proficiency as compared with classical testing methods like paper-based test and conventional IRT-based CAT.

Keywords: *Computerized Adaptive Testing, Bayesian Network, EM algorithm*

1. Introduction

As amount of knowledge is increasing rapidly, an ability which can effectively understand and well-organize the knowledge, namely creativity, has been widely required. In order to support such creative human in learning, it is very important to develop effective testing methods which correctly estimate examinee's proficiency and then provide detailed learning information based on it. However, the classical testing methods such as paper-based testing are inadequate in modern testing environment because the classical estimation methods not only need time-consuming testing task but also are often troubled with inaccuracy problem due to guess or error effects and inconsistent responses. To overcome this weakness Computerized Adaptive Testing (CAT) was proposed. Conventional CAT can adaptively estimate proficiency of the examinee by adopting Item-Response Theory (IRT) and Item-Information Function (IIF) into estimation system [1]. However, it cannot adapt to a broad spectrum of proficiencies and especially, structurally inter-related proficiencies across a number of categories. Because conventional CAT cannot consider correlation between items in nature, accuracy of proficiency estimation is prone to deteriorate. Besides, even though CAT can improve testing speed and time as compared with paper-based test, it still requires quite many items to be selected in order to estimate proficiency of the examinee.

In this paper, we propose a novel testing method, which makes use of Bayesian network [2] to estimate not only examinee's overall proficiency but also the spectrum of proficiencies across a number of categories. Differently to existing testing methods, our novel testing method constructs Bayesian network topology, which represents implicative relations between items and their categories. Then, the Bayesian network is learned by exploiting machine learning algorithm to generate CPTs of all nodes automatically. The learned Bayesian network is used to estimate detailed proficiencies of the examinee.

* Corresponding author at Hanyang University, Korea.

This paper is organized as follows: In section 2, we describe our method and Bayesian network-based CAT system applying it. Then, experimental results are showed in section 3. Finally, in section 4, we summarize our work and future plans.

2. Bayesian network approach to Computerized Adaptive Testing

Because IRT-based CAT system does not consider correlation information between items, the system cannot correctly estimate proficiency of the examinee if the given item affects other items. In this section, we show how to utilize Bayesian network as an estimation model and then, how to estimate the examinee's proficiency using Bayesian network.

2.1. Constructing Bayesian Network Topology

Since Bayesian network can represent cause and effect relationships between random variables, all cause and effect relationships between items and categories also can be effectively modeled as a knowledge topology.

Many researches and methods for modeling knowledge structure using Bayesian network have been proposed. E.Millan [3] proposed a hierarchical knowledge structure using Bayesian network and illustrated how to apply it to the CAT system as an estimation model. Jiri [4] showed how to estimate proficiency of the examinee and influences between items using Bayesian network. Based on these existing researches, our system represents each category and item as each *node* in Bayesian network; and all parent-child relationships between items and categories nodes are represented as *arcs* (or arcs between category nodes).

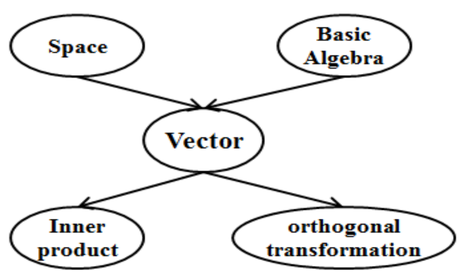


Figure 1. Bayesian network for "Linear Algebra" field

Figure 1shows a simple knowledge topology of a linear algebra field constructed by using Bayesian network. If the examinee does not know a concept of vector, he maybe cannot understand concepts of inner product and orthogonal transformation. But, if the examinee correctly understands concepts of inner product and orthogonal transformation, he maybe can correctly understand a concept of vector; and a probability that he correctly answer other items belonging to the categories will also be high.

By utilizing topology of Bayesian network as a knowledge structure model for CAT, some key effectiveness is expected as follows: First, the examiner or examinee can dynamically confirm the spectrum of detailed proficiencies across a number of categories; and detailed characters of the examinee can be correctly estimated by representing the subject as detailed categories. Second, the entire accuracy of the proficiency estimation can be improved because our novel approach reflects complicated interaction relationships between all items and their categories.

2.2. Learning Bayesian network for constructing CAT and estimating proficiency

After constructing knowledge topology, all CPTs in Bayesian network should be made in order to estimate proficiency of the examinee using the Bayesian network. To compute all probability values in the CPTs, a training data set should be required. The training data set consists of item-responses (correct or incorrect) information of all students for the subject.

$$D = [U_1 \; U_2 \; ... \; U_M]^T \quad where \;\; \forall_{1 \le i \le M} U_i = [Q_{i1} \; Q_{i2} \; ... \; Q_{iN} \; C_{i1} \; C_{i2} \; .. \; C_{ik}] \qquad (1)$$

Equation 1 describes the training data set D as user-item response matrix. In equation 1, Q_{ij} represents whether the user i answer correctly the item j or not. C_{ik} shows proficiency of the user i. However, because proficiency (or ability) information for each category is not given by the training data set, the all CPTs of category nodes cannot be estimated by estimation method, such as MLE which uses a complete training data set. To use this incomplete data as the training data, some useful methods are been proposed such as Gibbs sampling [5], Sequential learning [6] and Expectation-Maximization (EM) algorithm [7,8]. Our CAT utilizes EM algorithm for estimating unobserved values of all CPTs in Bayesian network.

After all CPTs are estimated using training data and EM algorithm, a complete Bayesian network model for CAT is constructed. By exploiting the Bayesian network, we can effectively estimate not only total proficiency but also proficiency spectrum of the examinee.

Table 1 represents proficiency estimation algorithm for our CAT system. Whenever an item is given to the examinee, he response to the item; and probability values of all nodes are updated by message-propagation algorithm [9]. And then, the updated probability values of the items are used to select next item suitable for him. Testing process is continued until a converge condition is satisfied. Our CAT system uses item information of all items as the converge condition. In other words, if information of all items is sufficiently low, the testing is terminated because information gained by continuously testing the examinee using the remaining items is very few. After all testing process is completed; a total proficiency and detailed proficiency spectrum for each of categories are returned to the examinee or the examiner.

3. Experiment

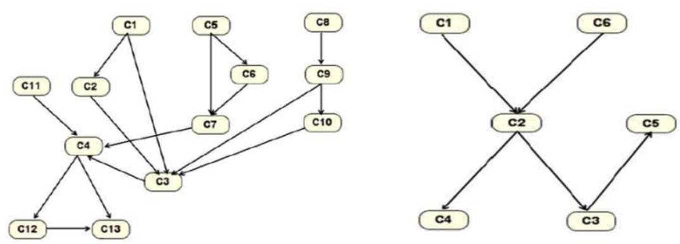
Table 1. Proficiency estimation algorithm

```
Algorithm estimateProficiency(Bayesian Network BN)
// Estimate proficiency of the examinee using Bayesian network constructed by topology and EM algorithm
Begin
   do {
            Q* ←selectItem(BN); // Select most suitable item having a probability that the examinee
                                   answers correctly the item closet to 0.5.
            if (the examinee answers correctly the item Q*) then P({Q* is true}) ← 1;
            else    P({Q* is true}) ← 0;
            for all nodes in BN do
                 recompute(BN); // Update probability values of all nodes in BN and re-compute answering
                                  probability of all items, P(Qn). This work is conducted by message-
propagation algorithm of Bayesian network
   } until (isConverged(BN) = true)   // This work is conducted by referencing probability values of all item nodes
   return detailed proficiencies in all categories and total proficiency of the examinee
End
```

Although Bayesian network-based CAT can effectively estimate detailed proficiency of each category, its testing way is very different from IRT-based CAT or paper-based test. Therefore, we should confirm accuracy and performance of our CAT system by comparing with other testing methods. In this section, we show some experiments and discuss about performance and accuracy of our CAT system.

3.1. Experimental Configuration

To evaluate overall performance of our CAT system, we used real test results of 160 high school students about two subjects, Data processing and Multimedia, as training and evaluating data set. Each subject was built as a hierarchical knowledge structure using Bayesian network. Each subject consists of thirteen or six categories; and these categories were mutually connected using direct arc. All items were linked to the category node to which the item belonging. Figure 2 show Bayesian network topologies for two subjects: Data processing (Figure 2-(a)) and Multimedia (Figure 2-(b)).



(a) Case A (b) Case B

Figure 2. Bayesian networks for "Data Processing" (Case A) and "Multimedia" (Case B)

In the experiment, we converted scores, which the 160 students get on paper-based test of the two subject, to ranking scores and compared how two ranking scores, where a ranking score is estimated by our CAT and other is estimated by paper-based test, are similar whenever the examinee response to an item. Similarity between two ranking scores was computed using correlation coefficient. Equation 2 shows a correlation coefficient where x means a ranking score estimated by CAT model and y means a ranking score estimated by paper-based test.

$$r = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \bullet \sigma_y} \qquad (2)$$

In (2), $\bar{x}$ and $\bar{y}$ means average value of x and y; and σ_x and σ_y mean standard deviation of each score. If r>0, two correlation is very high; however if r<0, two correlation is very low. In other words, the more the correlation is close to 1, the more score estimated by CAT is equal to a score estimated by paper-based test. Thus, this measure describes how correctly the CAT systems can estimate proficiency of the examinee by comparing with paper-based test. Where, we used Bayesian network-based CAT, IRT-based CAT and Naïve Bayes model as comparison model for estimating performance and accuracy of each CAT model.

Meanwhile, because CAT systems uses values of various types as testing result depending on each CAT systems, a converting function, which converts proficiency values of all items to a total score, is required. Equation 3 shows a converting function used in our experiment. The converting function converts proficiency values of all item nodes to a total score by multiplying probability value of each item by item point defined on the paper-based test.

$$Score = \sum_j score_j \times P(j) \quad (3)$$

where j means an item number and $score_j$ is point of the item j on the paper-based test. P(j) is a probability that the examinee responses to the item j correctly. Our CAT system used probability values of all items in Bayesian network as P(j) of equation 3 because the probability value of item node in Bayesian network describe a probability that the examinee answers correctly the item j.

However, in IRT-based CAT, the proficiency values of the items cannot be directly used as P(j) of equation 3 because this model uses Maximum Likelihood Estimation (MLE) method to estimate proficiency; and if the examinee answers all items correctly (or all incorrectly), the estimation of proficiency is infinite.

In order to overcome this weakness, we utilized Rasch model [10] as P(j) of equation 3instead of proficiency values of the examinee. Equation 4 shows 1-Parameter Logistic (1-PL) Rasch model which describes a probability that the examinee with proficiency θ answers correctly the item j with difficulty b. By applying this model to P(j) of equation 3, we could effectively standardize the examinee's proficiency estimated by IRT-based CAT.

$$P_j(\theta) = \frac{1}{1+e^{-(\theta-b)}} \quad (4)$$

We also utilized the number of items required to converge on correlation coefficient 0.9 between paper-based test and the CAT systems as a performance criterion for measuring performance of the CAT systems. And then, by using the number of the items, we evaluated how similarity between two scores is rapidly increased. This criterion demonstrates how many items are required to estimate the proficiency of examinee on equal accuracy with paper-based test. Moreover, in order to evaluate how the structural complexity of Bayesian network affects the estimation performance, we involved a Naïve Bayesian model, that all item nodes are linked to a category node, in our experiment.

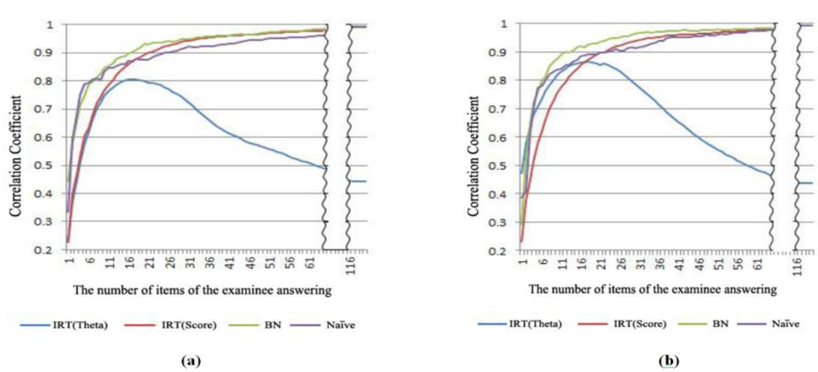


Figure 3.The experimental results of case A (a) and case B (b)

3.2. Experimental Results

The overall experimental results are showed in Figure 3. Figure3-(a) shows an experimental result about “Data processing (Case A);” and Figure3-(b) shows an experimental result about “Multimedia (Case B).”In Figure 3, we confirmed that all two experimental cases, A and B, had been converged to 0.97~0.98. Thus, we could confirm that Bayesian network-based CAT can estimate proficiency of the examinee on equal terms with IRT-based CAT.

Meanwhile, in case A, the number of items required to estimate the proficiency is 21 when IRT-based CAT is used; and the number of the items is 17 when Bayesian network-based CAT is used. In case B, the number of the items is 23 when IRT-based CAT is used; and the number of item is 14 when Bayesian network-based CAT is used. Thus, we could confirm that when Bayesian network-based CAT is used, the number of testing items is considerably reduced; and the time required for estimating proficiency also can be efficiently reduced.

In two experimental cases, we confirmed that in case A, the number of testing items required to converge to correlation coefficient 0.9 is more than the number of the items in case B. We can consider structural complexity of the network as cause of this situation. In case B, because the number of category nodes, namely the number of parameters which should be estimated by EM algorithm is less than the number of category nodes in case A, the overall accuracy of the estimated parameter values in case B will be higher than accuracy of the case A. Due to this reason, the total converging speed in case B is faster than speed in case A. Thus, we can know that if the number of categories is great many, the entire accuracy and performance of the proficiency estimation can be decreased. In other words, deciding the appropriate number of category nodes is very important; and we confirmed that the appropriate number of categories is 5~9 through many experiments.

Our CAT model can be variously applied according to the topology structure of Bayesian network. If a goal of the test is to estimate an entire proficiency of the examinee, Naïve Bayesian model-based CAT can be effectively used because a parent (category) node represents an entire proficiency of the examinee. In contrast, if estimating detailed proficiency spectrum about each of categories is goal, Bayesian network-based CAT can be utilized. In aspect of accuracy, since Bayesian network-based CAT compares with Naïve Bayesian model-based CAT, the two models can be appropriately applied according to goal of the test.

Additionally, we could discover some types of the examinees by analyzing proficiency spectrum which represents estimated proficiencies about each of categories. Table 2 shows detailed proficiency spectrum about each of categories and total score. In Table 2, we can know that although the total scores are mutually equivalence, their detailed proficiencies about each of categories can be reciprocally different. By utilizing this information, the examiner such as teacher or educational engineer can specifically check strength or weakness of the examinee and provide educational adaptive service such as personalized learning guide or reviewing service.

Table 2. Specific proficiency of each category (Case B)

C1	C2	C3	C4	C5	C6	Score
1.000	1.000	1.000	1.000	1.000	1.000	400
0.000	0.999	0.999	1.000	0.978	1.000	279.1
0.764	0.000	0.904	0.909	1.000	0.002	279.1
1.000	1.000	0.993	1.000	0.000	0.997	272.4
0.023	0.000	0.999	0.986	1.000	0.001	272.4
0.000	0.000	0.000	0.547	0.000	0.989	228.9
0.000	0.000	0.140	0.000	0.119	0.998	228.9
0.051	0.000	0.000	0.000	0.000	0.004	152.1
0.001	0.000	0.000	0.000	0.058	0.000	142.1

4. Conclusion

In this paper, we proposed a Bayesian network approach to CAT which can improve testing performance and accuracy by utilizing merits of probabilistic graphical model. To estimate the examinee’s proficiency, IRT-based CAT system has been widely used; however, in modern educational environment requiring to estimate multi-level proficiencies, the existing IRT-based CAT system is inadequate because it cannot effectively estimate the spectrum of proficiencies across a number of categories.

To overcome weakness of IRT-based CAT and improve accuracy and efficiency of testing, we proposed a Bayesian network-based CAT system. Through the several experiments, we confirmed that our novel CAT can effectively reduce the number of items which should be tested to estimate proficiency correctly as compared with other CAT models. Meanwhile, in order to evaluate accuracy of the proficiency estimation, we conducted some additional experiments, which compares ranking scores computed by our method to the scores gained by paper-based test, and confirmed that our CAT can equally estimate proficiency of the examinee as compared with the result of paper-based test. Specially, we confirmed that by using our novel approach, the examiner or examinee can get useful information, namely the spectrum of detailed proficiencies across all categories. This information may be utilized to provide personalized learning services, such as learning guide or reviewing system.

In the near future, we will study various methods which can improve the converge speed required to estimate proficiency and evaluate the accuracy of the proficiency estimation by comparing with various CAT models. Moreover, to provide intelligent services, a method that can automatically construct a Bayesian network topology from the training data set will also be researched. Maybe to accomplish these researches, some additional researches, such as heuristic or optimization algorithm, ought to be preceded.

Acknowledgements

This work was supported by the Technology Innovation Program (S1072514, Active Studying Management Program based on Adaptive Guide) funded by the Small & Medium Business Administration (SMBA, Korea).

References

[1] H. Wainer, “Computerized Adaptive Testing: A Primer”, Erlbaum Associates, New Jersey **(2001)**.
[2] Q.Yang, X. Wang, Z. Huang and S.Zheng, “Research of Student Model Based on Bayesian Network”, Proceedings of the Information Technologies and Applications in Education, ISITAE '07. First IEEE International Symposium, **(2007)**, pp. 514-519.
[3] E. Millán, M. Trella, J.L. Pérez-de-la-Cruz and R. Conejo, “Using Bayesian Networks in Computerized Adaptive Tests”, Computers and Education in the 21st Century, Part II, **(2002)**, pp. 217-228.

[4] JIŘÍ VOMLEL, “Bayesian Networks in Educational Testing”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol.12, Supplementary Issue 1, **(2004)**, pp. 83-100.
[5] S. Geman and D. Geman, “Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 6, **(1984)**, pp. 721-741.
[6] R. G. Cowell, A.P. Dawid and P. Sebastiani, “A comparison of sequential learning methods for incomplete data, Bayesian Statistics, vol. 5, **(1996)**, pp. 533-542.
[7] A. Dempster, D. Laird and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, Journal of the Royal Statistical Society, Series B, vol. 39, **(1977)** , pp. 1-38.
[8] S. Z. Zhang, Z. N. Zhang, N. H. Yang, J. Y. Zhang and X. K. Wang, “An improved EM algorithm for Bayesian networks parameter learning, Machine Learning and Cybernetics”, vol. 3, **(2004)**, pp. 1503-1508.
[9] EthemAlpaydin, “Introduction to machine learning”, MIT Press, Cambridge **(2010)**.
[10] F. M. Load, “Small N justifies the Rasch model”, In D. J. Weiss (Ed.), New horizons in testing, Academic Press, New York **(1983)**.

Authors

Kyung Soo Kim

Kyung Soo Kim received the B.S. degree in Computer education from Mokpo National University, Mokpo, Korea, in 2011. He is currently a M.S. student at the Division of Computer Science and Engineering, Hanyang University, Seoul, Korea. His current research interests include AI, machine learning and information retrieval.

Yong Suk Choi

Yong Suk Choi is a professor of Division of Computer Science and Engineering at Hanyang University in Seoul, Korea. He joined Hanyang University in 2000 after receiving Ph.D. in computer science from Seoul National University. His research interests include information retrieval, AI, machine learning, and software agents. He has done a number of researches in knowledge-based systems for fault diagnosis and in intelligent agents for text retrieval, filtering, and mining on the Internet.