A Machine Learning Algorithm for Classification of Difficulty Levels of CATs Considering Education Institution Circumstances

Koki Araseki1[1], Haruki Ueno[2] and Hiroshi Komatsugawa[3]
[1]Graduate School of Science and Engineering, Chitose Institute of Science and Technology
758-65 Bibi, Chitose, Hokkaido, Japan
e-mail: m2230020@photon.chitose.ac.jp
[2]Faculty of Science and Technology, Chitose Institute of Science and Technology, Japan
758-65 Bibi, Chitose, Hokkaido, Japan
e-mail: h-ueno@photon.chitose.ac.jp
[3]Graduate School of Science and Engineering, Chitose Institute of Science and Technology
758-65 Bibi, Chitose, Hokkaido, Japan
e-mail: hiroshi@photon.chitose.ac.jp

In this paper, we constructed and evaluated a difficulty classification algorithm for a wide variety of exercises consisting of images and texts managed by each material provider, as well as for a group of test materials. In the previous study, classification based on test theory was performed to supplement missing values based on learning history. In the present study, we used machine learning for the classification considering learning history including missing values and the learning policy of the teacher. In addition, to take into account the context of the questions, we calculated the similarity of the questions based using BERT. As a result, we succeeded in performing difficulty classification with an accuracy of over 90%.
Key words: Computer-Adaptive-Testing, Test Theory, Machine Learning, Teacher Policy, BERT

1. INTRODUCTION

In recent years, the Giga School concept has been promoted in elementary and junior high school and one mobile device per person is provided. So, it is expected that Computer-based Testing (CBT) will be used to provide appropriate materials for individual learners. For example, “MEXCBT” managed by the Ministry of Education provides CBT contents in collaboration with educational institution and companies [1]. However, MEXCBT has no function of the Computer-adaptive Testing (CAT), which allows for adaptive questioning, and leads to supporting individualized and optimal learning. The CAT functions needs teaching materials structured at difficulty levels, which are in general provided as vendor-specific paid materials. Therefore, the CAT service is not provided in public for students, in particular, who need optimal learning support in daily educational activities. Furthermore, the logic of vendor-provided CAT is not open, and there remain issues for quality assurance of learning assessment in applying the system to public education. From the viewpoint of individualized and optimal learning support in each educational field, it is better for teachers to provide CAT materials flexibly and freely in accordance with their educational policy.

The purpose of this study is to construct an automatic difficulty classification algorithm for a wide variety of exercises in the educational field by utilizing a test theory and machine learning algorithms. In order to achieve the objectives, two goals are set as follows.

(1) Establishment of a classification model focusing on difficulty and its algorithm
(2) Evaluation of the classification model

2. POSITIONIG OF PREVIOUS STUDIES AND THIS STUDY

2.1 Base system

CIST-Solomon is an e-learning system developed and operated by the university of CIST since 1999.The system has a wide range of teaching materials, from elementary and secondary education to specialized fields of study at universities. For the empirical research, we use this system as the basic system with the CAT function and adaptive teaching materials.

2.2 Test theory

The recent(non-classical) test theory is a known as one of effective statistical models for learner's learning ability using learner's right/wrong data and consideration of difficulty degree at each question, in which Item Response Theory (IRT) and Latent Rank Theory (LRT) are representative ones. The IRT assumes that examinees' abilities follow a normal distribution and estimates examinees' abilities in the range from -3.0 to 3.0. On the other hand, LRT is a classification model based on SOM (Self-Organizing Map) and estimates learners' ability along with discrete numbers. This method realizes a relative rating for learners' assessment. In our research, we aim to consider assessment in each educational field, that means the relative rating is an appropriate approach. Then, we adopt the LRT in our research.

2.3 Previous study

Abe constructed classification algorithm of questions using Exametrika with LRT logic to update the difficulty based on the estimated values [2]. Fig.1 shows Abe's algorithm. First, the learning history of a learning unit is extracted from the CIST-Solomon database, and the provisional ability is estimated using Exametrika. Since there are missing values in the learning history, the probability of correct answers is assumed based on the provisional ability, and the missing values are completed after weighting. Next, the completed learning history was analyzed again using Exametrika, and the estimated value was used as the updated difficulty level for each question.

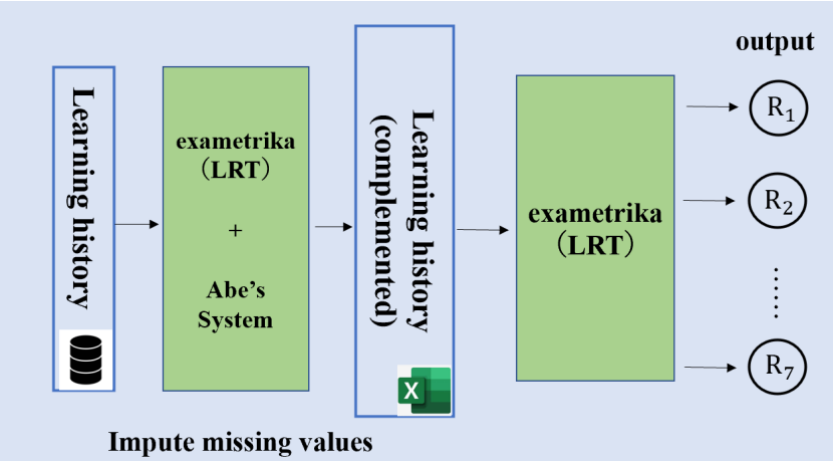

Fig.1 Abe's algorithm

2.4 Positioning of this study

In this study, we aim to construct an algorithm for a fully coupled neural network (NN) based on the learning history of CIST-Solomon, with teacher's policies (rubrics) added as input parameters, in order to solve the issues raised by the previous studies. Since LRT shows better performance in classification than IRT, we will build an algorithm based on Abe's previous work.

3. ALGORITHM CONSTRUCTED IN RESEARCH

3.1 Algorithm1

Fig.2 shows the first algorithm constructed in the present study that is a model using both the input data from LRT used in Abe's research and the teacher's policy. In this study, the difficulty level was updated on basis of assessment by the expert. Exametrika performs relative evaluation and is used for data completion and first order classification. On the other hand, the teachers' policy consists of absolute evaluation. So, this algorithm includes both relative and absolute evaluations.

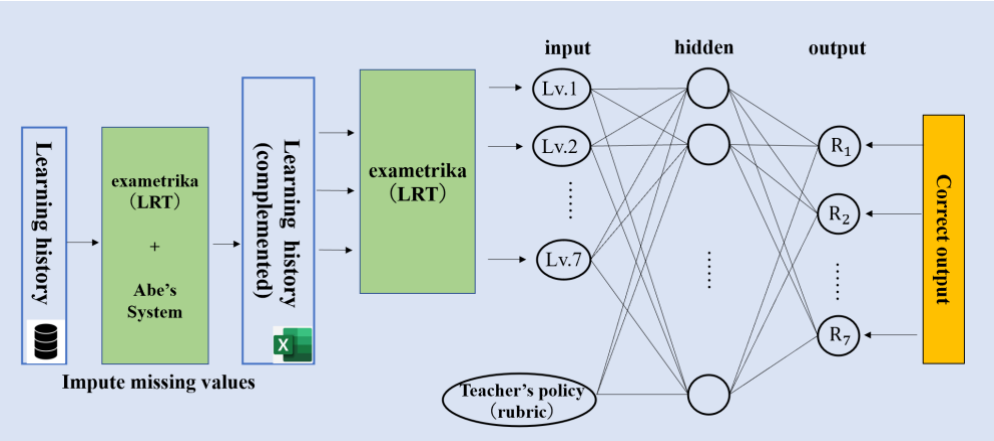

Fig.2 Algorithm 1

3.2 Algorithm2

Fig.3 shows the second algorithm that does not depend on LRT but instead utilizes statistical data obtained from the learning history including missing data and teacher's policy classified into 3 categories of rubric (i.e. 1st level of knowledge acquisition, 2nd level of knowledge utilization, and 3rd level of knowledge application). This is the simplest algorithm in this study. However, in this algorithm, absolute evaluation is used due to the rubric of teacher's policy but not introduced due to the omission of LRT.

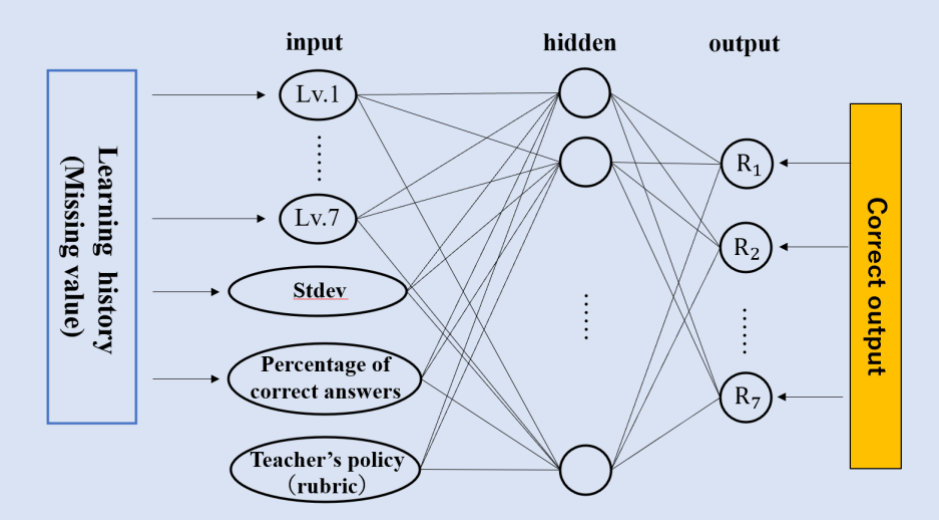

Fig.3 Algorithm 2

3.3 Algorithm 3

Fig.4 shows the third algorithm that is a revised version of Algorithm 2 with use of the sentence vectors of question sentences generated by BERT. The use of sentence vectors enables us to take into account the context information of the question. Furthermore, we realized to convert the CBT materials including image data to text ones, using OCR and input for BERT. We assume that adding the sentence vectors of questions to input labels of the neural network enables us to consider both the similarity of question sentences and the difficulty of each question that is the same way the actual teachers in the educational fields do.

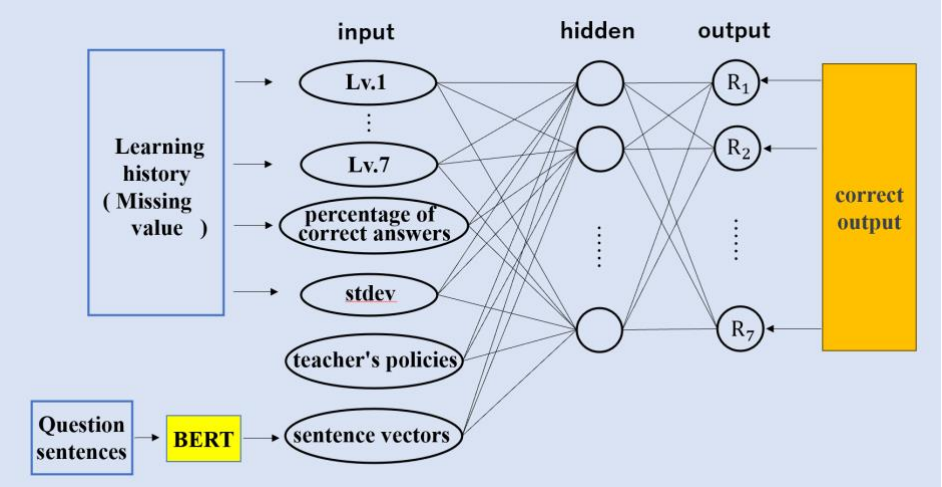

Fig.4 Algorithm 3

4. EVALUATION

For the evaluation of the constructed algorithms, we selected the units of recursive processing and sort algorithm. Table I shows the result of each algorithm. First, we investigated the effect of LRT in comparison with the cases of Abe's algorithm, Algorithm 1, and Algorithm 2. The result indicates that Algorithm 2 had the highest accuracy and worked most effectively. Since the algorithm using LRT supplements dummy values into missing learning data, the prediction accuracy may drop down.

Second, we compared Algorithm 2 and 3, and found that Algorithm 3 had higher accuracy and less variation in the number of problems at each level. Algorithm 3 takes in teacher's policy described into three categories determined on basis of difficulty of CBT problems and the interpretation on basis of the context obtained automatically from the question texts and image data of each CBT material. The judgment on basis of the policy and the interpretation is usually done by the teachers in the actual education field. We think that the similarity with teachers' assessment behavior may improve the accuracy of Algorithm3.

Table I Evaluation of each algorithm

	Recursive processing		Sort algorithm	
Algorithm	Accuracy(%)	Stdev	Accuracy(%)	Stdev
Abe's algorithm	65.8	7.12	62.4	14.1
Algorithm 1	79.8	6.39	54.8	22.6
Algorithm 2	84.8	7.68	78.5	10.2
Algorithm 3	**98.7**	**1.6**	**96.8**	**5.92**

5. CONCLUSION

The purpose of this study was to construct difficulty classification algorithms for exercises added by providers of educational materials. Based on Abe's previous research, three difficulty classification algorithms were constructed, using the NN algorithm, which includes input information of the learning history, teacher policy, and contextual information of the question text. The algorithm, including the teacher's policy and contextual judgments for CBT questions, achieved prediction accuracy close to the classification set by the teacher. In the future, It is necessary to confirm that the classification is consistent across diverse educational environments, since the classification set by the teacher is based on the judgment of a single expert.

REFERENCES

[1] Ministry of Education, Culture, Sports, Science and Technology, Japan. "About the Ministry of Education, Culture, Sports, Science and Technology CBT System (MEXCBT) ." Ministry of Education, Culture, Sports, Science and Technology, December 2023, *https://www.mext.go.jp/a_menu/shotou/zyouhou/mext_00001.html* . Accessed 15 January. 2024.

[2] Abe, Kodai. "Proposal of Practical Classification Method of Learning Quizzes.", 2021.