A Q-matrix Model Based on Binarized Neural Network

Jingjiang Li¹, Chong Jiang², Shiwei Ye¹,*

¹School of Electrical and Electronic Engineering University of Chinese Academy of Science, Beijing, China
²School of Computer Science and Technology University of Chinese Academy of Science, Beijing, China
*email: shiwye@ucas.ac.cn

Abstract. Q-matrix theory plays an important role in the field of cognitive diagnosis assessment. It is time-consuming and laborious for experts to define Q-matrix from a large number of data. In order to deal with this problem, this article comes up with a Q-matrix generation model based on binarized neural network. We combine the neural network with the Boolean operation relations of Q-matrix, A matrix and R matrix in item response theory, so that the model can mine Q-matrix more effectively.

1. Introduction

In recent years, many data mining methods have been applied to the field of education to mine students' knowledge state from their practice records, such as traditional classical test theory (CTT), item response theory (IRT) [1], and cognitive diagnosis assessment.

The classical test theory (CTT) and item response theory (IRT) only focus on the evaluation of students' level by using exam scores and they cannot reflect students' knowledge status. Cognitive diagnosis assessment [2,3] judges students' mastery of specific attributes by analyzing their responses to questions, which makes up for the shortcomings of CTT and IRT.

Let us suppose that there are a number of students and items, and a matrix \( R \in \{0,1\}^{m \times n} \) indicating the students' answer. \( R_{ij} \) express that student \( i \) respond to item \( j \) and \( R_{ij} = 1 \) indicates that the student \( i \) reply to the item \( j \) exactly. As shown in Table 1, 2 and 3, Q-matrix theory in cognitive diagnosis can well connect the knowledge status of students with the performance of answering questions. For example, to solve item 2+3-1 one must master the addition and subtraction concepts. In addition, it is necessary to determine the knowledge status of students, namely A matrix. For example, student 2 only mastered the two attributes of subtraction and division, so he cannot solve the item 2+3-1.

| Student | Item       |
|---------|------------|
|         | 2+3-1  | 4/2 | 4×2+3 |
| 1       | 1       | 0   | 0     |
| 2       | 0       | 1   | 1     |
| 3       | 1       | 0   | 1     |
### Tab. 2 Q-Matrix Example

| Item     | Knowledge concept |
|----------|------------------|
| +        | -                | ×    | ÷    |
| 2+3-1    | 1                | 1    | 0    | 0    |
| 4/2      | 0                | 0    | 0    | 1    |
| 4×2+3    | 1                | 0    | 1    | 0    |

### Tab. 3 A Matrix Example

| Student | Knowledge concept |
|---------|------------------|
| +       | -                | ×    | ÷    |
| 1       | 1                | 1    | 0    | 0    |
| 2       | 0                | 1    | 0    | 1    |
| 3       | 1                | 1    | 1    | 0    |

Based on the Q-matrix theory, many methods have been invented for cognitive diagnosis assessment, such as the DINA models [4], G-DINA models [5], NIDA model [6,7], RSM [8,9], DINO models [10-12], AHM [13], non-negative matrix factorization (NMF) [14], Boolean matrix completion (BMC) [15], Auto-Encoders (AEs) [16]. The Q-matrix records the corresponding relations between items and knowledge concepts[17], which was presented by Tatsuoka in 1983[18]. But all these methods, need to define Q-matrix manually.

Generally speaking, the main contributions of this paper has the following two aspects:

We propose a neural network-based model to automatically extract Q-matrix from students’ practice records. The model combines the Boolean operation relations of Q-matrix, R-matrix and A-matrix in cognitive diagnosis theory on the basis of binarized neural network.

The proposed binarized network model, can not only extract the Q-matrix, but also extract the knowledge state (A-matrix) from the hidden layer of neural network.

## 2. Related Work

### 2.1 Binarized Neural Network

The Binarized Neural Network (BNN) [19] is a neural network obtained by binarizing its weights and activation functions. Specifically, the weight $W$ and hidden layer activation values are changed from floating-point (31 bit) to $+1$ or $-1$ (1 bit), so as to occupy less storage space. At the same time, the binarized neural network uses bit operation to replace the multiplication and addition operation in the original neural network, which does not change the structure of the network, but binarize the parameters and activation function values of the network.

There are mainly two methods to binarize neural network. One is the stochastic method as formula (1), the weight values and the activation function values are converted into a probability through a function, and then the value is $+1$ or $-1$ with a certain probability. The $\sigma$ is a “hard sigmoid” function in formula (2). The other method is the direct determination method, it is directly determined by the floating-point weight value or the sign of the activation function value. It can also be seen as binarizing the weight $W$ and the hidden layer activation value to 1 or -1 as shown in formula (3).

$$x^b = \begin{cases} 
+1 & \text{with probability } p = \sigma(x) \\
-1 & \text{with probability } 1 - p 
\end{cases}$$  
(1)
The stochastic method is difficult to achieve, so most experiments use deterministic methods. In the process of BNN forward propagation, we first binarize the weights by \( \sigma(x) = \text{clip}(\frac{x+1}{2}, 0, 1) = \max(0, \min(\frac{x+1}{2})) \) (2).

Then, the binarized parameters are used to calculate the real number intermediate vector, and then the real number hidden layer activation vector is obtained by batch normalization operation. If it is not the output layer, the vector is binarized.

In the process of BNN backward propagation, we need to use the derivative of a symbolic function, which is almost zero everywhere. Hence, we have to relax the symbolic function as shown in formula (4) and formula(5). The derivative is formula (6). Suppose the loss function is \( C = f(q) \) and the binarization function is \( q = \text{sign}(x) \), then the derivative of \( C \) to \( x \) is as in formula (7).

\[
H \tanh(x) = \text{clip}(x, 1, -1) = \max(-1, \min(1, x))
\] (4)

\[
H \tanh(x) = \begin{cases} 1 & x \geq 1 \\ x & -1 < x < 1 \\ -1 & x \leq 1 \end{cases}
\] (5)

\[
H' \tanh(x) = \begin{cases} 0 & |x| < 1 \\ 1 & \text{else} \end{cases}
\] (6)

\[
g(x) = \frac{\partial(C)}{\partial(x)} = \frac{\partial(C)}{\partial(q)} \frac{\partial(q)}{\partial(x)} = g(q)H' \tanh(x)
\] (7)

![Fig. 1 The symbolic function relaxation](image)

2.2. Matrix factorization
As shown in formula (8), matrix factorization aims to decompose matrix $R$ to two matrices $Q$ and $A$. In cognitive diagnose assessment, the Q-matrix reflects the connection between items and knowledge concepts while the A matrix indicates the knowledge state of students.

The objective function of matrix decomposition is show in formula (9).

$$
\overline{R} \approx QA
$$

$$
\arg \min_{Q,A} L(R,QA)
$$

(8)

(9)

3. Q-Matrix generation model of binarized neural network

It is an long-standing topic in educational data mining to mine students' knowledge states from their practice records. Matrix decomposition and neural network have been proved to be effective in solving this problem.

This part presents the proposed model to combine matrix decomposition and binarized neural network to mine students' practice records to generate Q-matrix.

3.1. Task Overview

Given the record of the students' practice, containing each student's answer to each item, it can be abstracted as $R$ matrix, $R = \{R_y\}_{M \times N}$, where $R_y = 1$ if student $i$ correctly grasped the item $j$ and $R_y = 0$ otherwise. The Q-matrix defined by experts is also given, $Q = \{Q_y\}_{N \times K}$, where $Q_y = 1$ if item $i$ relates to knowledge concept $k_j$ and $Q_y = 0$ otherwise. Our model aims to mine Q-matrix from datasets, and reconstruct R matrix based on CMD.

3.2. Binarized Neural Network model

Given $Q = [Q_y \in \{0,1\}^{N \times K}]$ as the relation matrix of items and attributes, row vector $q_j = [Q_{j1}, Q_{j2}, ..., Q_{jk}]$ is the attribute vector of the item $j$; Students' knowledge state matrix $A, A = [A_y \in \{0,1\}^{N \times K}]$, row vector $\alpha_j = [A_{j1}, A_{j2}, ..., A_{jk}]$ means the knowledge state of the student $i$; $R \in \{0,1\}^{m \times n}$ indicates the students’ answer. As shown in formula(10), $R_y = 1$ indicates that the student $i$ reply to the item $j$ exactly, while $R_y = 0$ indicates that the student $i$ reply to the item $j$ mistakenly.

$$
R_y = \begin{cases} 
1 & \text{if } \alpha_i \geq \alpha_j \\
0 & \text{otherwise}
\end{cases}
$$

(10)

$R$ is a matrix operation in formula (11), $\circ$ means Boolean multiplication. It is defined as formula (12).

$$
R = \overline{A} \circ Q
$$

$$
B = U \circ V = [b_{ij}], b_{ij} = V_k(u_{ik} \land v_{kj})
$$

(11)

(12)

Here, $0 \lor 0 = 0, 0 \lor 1 = 1 \lor 0 = 1 \lor 1 = 1, 0 \land 0 = 0 \land 1 = 1 \land 0 = 0, 1 \land 1 = 1$. Based on Boolean multiplication, we can define $R_y$ as formula (13). Based on this relationship, our network structure concept is shown in the Figure 2.

$$
R_y = \bigwedge_k I(A_{ik} \circ Q_{jk})
$$

(13)
Based on binarized neural network, the weights and the output of the neuron are constrained to +1 or -1. We need to convert the value of our input R matrix from \{1,0\} to \{+1,-1\}. As shown in formula (14), $\vee$ takes the maximum and $\wedge$ takes the minimum of input data and the sign function is the same in formula (3).

\[
\begin{align*}
\wedge(x_1,\ldots,x_K) &= \min(x_1,\ldots,x_K) = \text{sign}(\sum_{k=1}^{K} x_k - K + 1) \\
\vee(x_1,\ldots,x_K) &= \max(x_1,\ldots,x_K) = \text{sign}(\sum_{k=1}^{K} x_k + K - 1)
\end{align*}
\]

(14)

After the above formulation, can be obtained as formula (15). The vector form is written in formula (16). Where * means the row vector of Q-matrix, means the row vector of A-matrix, and $e$ means a unit vector whose components are all one.

\[
R_{ij} = \text{sign}(\frac{1}{2} \sum_{k=1}^{K} (1 + A_{ik} + A_{ik}Q_{jk} - Q_{jk}) - K + 1)
\]

(15)

\[
R_{ij} = \text{sign}(A_{ir}(Q_{jr} + e_{k+1}^T) + e_{k+1}^TQ_{jr}^T + (2 - k))
\]

(16)

The model of the simplest three-layer neural network is shown in the Figure 3.
The input to the network is the row vector of the R matrix (student's practice record), the input layer to the hidden layer is a fully connected neural network, the hidden layer to the output layer is also a fully connected neural network. We set the number of neurons in the input layer equal to the number of items, the number of neurons in the hidden layer equal to the number of attributes, and the number of neurons in the output layer equal to the number of items.

The activation functions in the network model are all sign functions, and the Q-matrix can be extracted from the weight of the middle layer. The weight value from the hidden layer to the output layer is the transpose of the Q-matrix mined from the data. And the activation value of neurons in the last hidden layer is A matrix.

The objective function of model optimization:

$$\theta^* = \arg \min_{\theta} \beta L_\beta(x, \hat{Q}) + \lambda \| W_1 \|_2$$  \hspace{1cm} (17)

In formula (17), means Cross entropy loss function, $\hat{Q}$ means the real Q-matrix defined by experts and $\beta$ means weigh parameters. We input row vectors of the R matrix to train the model. By continually narrowing down the value of the optimization function, we extract the mined Q-matrix between the output layer and the hidden layer.

Generally speaking, deep neural network has stronger expression ability and ability to extract features. On the basis of Figure 3, we can build a deep binary neural network model to achieve better fitting effect, as shown in Figure 4. In a way, our model is very similar to the idea of an automatic encoder. The output form of our model is consistent with the input form.

4. Experiments

4.1. Datasets
Simulated Data: This dataset is a dataset artificially simulated by IDR method. The data set contains 1000 students and 50 items and each topic has 8 attributes or more. After the data set is expanded, we extract the training data according to different proportion and use the rest as the test data.
FrcSub: This dataset was used for fractional subtraction tests and first appeared in Tatsuoka [8], which include 536 students, 20 items, and 8 implicit attribute.

The table 4 shows the difference between the Simulated data and FrcSub dataset.

We add slipping parameters and guessing parameters to simulate real data when using IDR method to generate artificial data. The composition of the data set is shown in Table 4.

**Table 4 Dataset Summary**

| Dataset    | Simulated Data-1 | Simulated Data-2 | Simulated Data-3 | FrcSub |
|------------|------------------|------------------|------------------|--------|
| Students   | 1000             | 1000             | 1000             | 536    |
| Items      | 50               | 50               | 50               | 20     |
| Knowledge  |                  |                  |                  |        |
| Concept    | 8                | 16               | 24               | 8      |

4.2. Evaluation Criteria

Let's start with two existing models, and then compare our model with them.

DINA: DINA model is one of the most commonly used and important models in cognitive diagnosis and is often used as a benchmark for cognitive diagnosis.

NMF: Nonnegative matrix decomposition is an effective multiple data decomposition technique, which has no negative value in all matrix properties. In the field of personalized learning, the method can also be used to predict students' performance.

Two evaluation indexes were adopted, including accuracy (Acc) and root-mean-square Error (RMSE). The accuracy is higher and the root-mean-square error is smaller, the model effect will be better.

\[
ACC = \frac{TP + TN}{TP + FP + TN + FN}
\]

\[
RMSE = \sqrt{\frac{1}{|T|} \sum_{ij} (R_{ij} - \hat{R}_{ij})^2}
\]

As shown in Formula (19), \( R_{ij} \) means the real response of student I on item J, and \( \hat{R}_{ij} \) means the predicted response of student I on item J, \(|T|\) is the total number of ratings in the test set.

In the process of model training, we use the five-layer model. In order to reduce the error of the result, we will train the model five times to obtain five groups of training parameters. By extracting five groups of Q-matrix values, the final predicted value of Q-matrix (\( Q_{pre} \)) was obtained by voting. For the test set, we only need to extract the A matrix in the last hidden layer of the model. By combining the Q-matrix (\( Q_{pre} \)) obtained from training with the A matrix extracted from the test set, we can reconstruct the R matrix \( \hat{R} \) of students.

In order to evaluate the performance of this model, we built the training sets based on different knowledge concepts. We randomly chose the training set from the simulation data set and use the remaining data as the test set.

Table 5 and Table 6 compare the RMSE of DINA, NMF algorithm and Q-matrix generation model based on binarized neural network. And in this paper, a method of generating Q-matrix and reconstructing R matrix on four datasets is proposed.

**Table 5 The rmse of Q-Matrix generation**

| Dataset       | Algorithms |
|---------------|------------|
|               | DINA       |
|               | NMF        |
|               | BNN        |
Simulated Data-1 & 0.2792 & 0.2504 & 0.1922 \\
Simulated Data-2 & 0.3288 & 0.3680 & 0.1820 \\
Simulated Data-3 & 0.3123 & 0.3465 & 0.1725 \\
FrcSub & 0.3275 & 0.2692 & 0.1823 \\
\hline
\textbf{Tab. 6} The rmse of RMatrix generation

| Dataset          | Algorithms |       |       |       |
|------------------|------------|-------|-------|-------|
|                  | DINA       | NMF   | BNN   |       |
| Simulated Data-1 | 0.3935     | 0.3635| 0.2563|       |
| Simulated Data-2 | 0.4189     | 0.4520| 0.2273|       |
| Simulated Data-3 | 0.5510     | 0.3786| 0.2021|       |
| FrcSub           | 0.3359     | 0.3953| 0.2060|       |

Simulated Data-1 & 0.3935 & 0.3635 & 0.2563 \\
Simulated Data-2 & 0.4189 & 0.4520 & 0.2273 \\
Simulated Data-3 & 0.5510 & 0.3786 & 0.2021 \\
FrcSub           & 0.3359 & 0.3953 & 0.2060 \\
\hline
\textbf{Tab. 7} The ACC of Q-Matrix Generation

| Dataset          | Algorithms |       |       |       |
|------------------|------------|-------|-------|-------|
|                  | DINA       | NMF   | BNN   |       |
| Simulated Data-1 | 0.8562     | 0.8721| 0.8864|       |
| Simulated Data-2 | 0.8723     | 0.8706| 0.9042|       |
| Simulated Data-3 | 0.8652     | 0.8951| 0.9116|       |
| FrcSub           | 0.8663     | 0.8832| 0.9021|       |

\begin{figure}
\centering
\includegraphics[width=\columnwidth]{fig4.png}
\caption{The ACC of Q-matrix generation}
\end{figure}

On the whole, the performance of our model is better than that of the traditional cognitive diagnostic model. In particular, when there are fewer knowledge points involved in the item, our cognitive diagnostic model is not very prominent compared with the traditional cognitive diagnostic model. However, with the increase of knowledge points, the advantage of our model is gradually obvious.

Table 7 compare the ACC of DINA, NMF algorithm and Q-matrix generation model which based on binarized neural network proposed in this article to generate Q-matrix respectively on four datasets. On the whole, the performance of our model is superior to the DINA model and the NMF algorithm.

5. Conclusion
This paper has proposed a Q matrix generation model by combining matrix decomposition and binarized neural network in the field of cognitive diagnosis. Our model can automatically extract the Q-matrix from the student response matrix. The experiment results indicates that our model is superior to the other two models. In future experiment, we will improve the performance of the model and apply it to more data sets.

References
[1] WOOLF B P. 2006 Estimating student proficiency using an item response theory model Lecture Notes in Computer Science 4053(5) pp 473-480
[2] GIERL M 2007 Cognitive diagnostic assessment for education: Theory and applications Journal of Qingdao Technical College 45(4) pp 407-411
[3] ZHENG C 2007 Detect and track latent factors with online nonnegative matrix factorization International Joint Conference on Ijcai
[4] B. W. Junker and K. Sijtsma 2001 “Cognitive assessment models with few assumptions, and connections with nonparametric item response theory,” Applied Psychological Measurement vol. 25, no. 3, pp 258-272
[5] Delatorre J 2011 The Generalized DINA Model Framework Psychometrika 76(2) pp 179-199
[6] J. Dina model and parameter estimation: A didactic Journal of Educational & Behavioral Statistics 34(1) pp 115-130
[7] DING S L 2011 Research on mcmc parameter estimation and properties of high order dina model
[8] TATSUOKA K. Linkage between teacher quality, student achievement, and cognitive skills: A rule-space model Studies in Educational Evaluation 30(3) pp 0-223
[9] QI G 2016 The rule space and attribute hierarchy methods John Wiley & Sons, Inc.
[10] JUNKER K 2001 Brian/Sijtsma.Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. Applied Psychological Measurement 25(3) pp 258-272
[11] MARIS E. 1999 Estimating multiple classification latent class models Psychometrika 64(2) pp 187-212
[12] HENSON R A. Measurement of psychological disorders using cognitive diagnosis models Psychol Methods 11(3) pp 287-305
[13] HUNKA S 2002 The attribute hierarchy model for cognitive assessment
[14] D. D. Lee and H. S. Seung 2001 “Algorithms for non-negative matrix factorization,” in Advances in neural information processing systems pp 556-562
[15] H. Wang, G. Su, Y. Sun, S. Ye, P. Gao, and Y. Sun 2015 “A latent space perturbation algorithm for boolean matrix completion based on weighted frobenius norm,” in Behavioral, Economic and Socio-cultural Computing (BESC) International Conference on. IEEE pp 43-46
[16] Y. Bengio et al 2009 “Learning deep architectures for ai,” Foundations and trends® in Machine Learning, vol. 2 no. 1 pp 1-127
[17] M. Birenbaum, A. E. Kelly, and K. K. Tatsuoka 1992 “Diagnosing knowledge states in algebra using the rule space model,” ETS Research Report Series, vol. 1992, no. 2
[18] K. K. Tatsuoka 1983 “Rule space: An approach for dealing with misconceptions based on item response theory,” Journal of educational measurement vol. 20, no. 4, pp 345-354
[19] Courbariaux M, Hubara I, Soudry D, et al 2016 Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1