AN ADAPTIVE E-ASSESSMENT TO ESTIMATE EXAMINEES’ ABILITY BASED ON NEURAL NETWORK APPROACH

M.A. Azmi Murad and A. Kazemi
Department of Software Engineering and Information Systems, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia 43400 UPM Serdang MALAYSIA
Email: masrah@upm.edu.my

ABSTRACT

The advancements in computer-based assessment provide the technological foundation for e-assessment in measuring students’ learning. The knowledge of a student (also known as an examinee) is measured through exams. A key purpose of using an exam is to determine the proficiency level of each examinee based on his/her responses to the administered test. The main problem of traditional test is that the asked questions did not match the actual ability of examinees and did not measure examinee’s proficiency accurately. Therefore, Computer Adaptive Testing (CAT) has been developed to address this issue. In CAT, each examinee has to answer the questions that are tailored to his/her ability level. It uses models of proficiency estimation such as Item Response Theory (IRT). IRT model relates the response of an examinee to a specific item to his/her ability level and characteristics of the item. However, in IRT model, the relationship between items characteristics and person’s skill are very complex and nonlinear. In this work, we proposed a neural network model to estimate examinees’ ability for small sample size and based on the experiments, we obtained a low mean square error (MSE) compared to IRT model.

Keywords: computer adaptive testing, e-assessment, examinee ability, item response theory, neural network, mean square error

INTRODUCTION

With the growth of computers and increase in speed and storage capabilities in technology, computer programs were designed to assess educational achievement, professional licensure and graduate student candidates with improved scoring, reliability and standardization compared to human examiners [1]. The modern tests such as Computer Adaptive Testing (CAT) were emerged to overcome the problems of traditional paper and pencil (P&P) tests.

The history of the first adaptive test is the old Binet test [4], which include features of a computer adaptive test. However, the Binet test was unsuitable for large test administered because it required one-on-one interaction between the examinee and test administrator. Adaptive testing is not developed until the extensive availability of desktop computer [2]. In CAT, items selected are tailored to the individual examinee’s ability level. As an examinee answers test items, the CAT appropriately choose the next item that will provide the most information about that examinee, based on the examinee’s current ability estimation [2]. Unlike the traditional testing, in CAT, different examinees are administered by different sets of items, and since the abilities of the examinees are on a common scale, the examinees can be compared. CAT has been utilized extensively for education, licensure programs and certification as well as within the military for selection purposes. Armed Services Vocational Aptitude Battery (ASVAB), Graduate Management Admission Test (GMAT-CAT), Microsoft© Certified Professional Exams, TOEFL, and GRE have applied CAT technology to assess examinee’s knowledge [5]. Besides educational field, CAT can also be utilized in other fields like marketing strategies [6], health outcome [7, 8] and medicine [3]. The advantages of using CAT are reduction in time, increase in accuracy of estimating an examinee’s knowledge level and minimizing the unwanted behaviours such as carelessness and guessing [9].

One of the most important components of adaptive testing is Item Response Theory (IRT) [10]. IRT is a statistical model with theoretical foundation that is widely used in modern educational testing technology and psychological testing. IRT model describes a relationship between responses of examinee to test items and his/her latent variable that is named trait, ability or proficiency. The latent variable is not directly measurable but can be estimated [11]. However, IRT needs a large size of sample data to achieve high precision in the ability estimation [13] and it is based on strict assumptions [9, 12]. With the
advantages of neural network and CAT, and its application in many areas, serve as an attractive tool to solve many complex nonlinear problems. Thus, to overcome the limitations of IRT in CAT, we proposed a neural network model to estimate examinees’ ability accurately by having the lowest estimation error especially for small sample size. The remaining of this paper is organized as follows: in the next section we present an overview of some related works. Section 2 describes the methodology of this work, section 3 discusses the results, and finally, section 4 concludes the work.

MATERIALS AND METHODS

In this section, we described the methodology implemented in this study. We employed Monte Carlo simulation and back-propagation algorithms in the experiment. Three common methods of estimation ability such as Maximum Likelihood Estimation (MLE), Expected A Posteriori (EAP), and Maximum A Posteriori (MAP) [16] are used as a comparison against Multi Layer Perceptron (MLP) neural network model. The test data generated will be submitted to MULTILOG program [15] to estimate the ability parameter and performed on the 3-Parameter Logistic (3-PL) dichotomous items (difficulty, discrimination, and guessing).

There are different approaches to validate CAT systems and IRT models. Commonly, these approaches are classified into two primary types: live-test studies and simulations. We employed the later approach where it can be divided into simulation with real data (called post hoc simulation) and pseudo random data (Monte Carlo simulation). Live testing involves the administration of real tests to real examinees. There are real items, which are applied to a group of real examinees. It is an expensive and time-consuming because these kinds of tests are famous in large-scale exams and moreover, providing too many real examinees and real items can be a difficult task.

In post hoc simulation, the “item bank” used consists of the actual answers of examinees to a full-length test. These are real data, which are used to evaluate different component of CAT. A Monte Carlo simulation [14] defines a theoretical examinee population and item banks and generates their responses over items by using a pseudo random data generation model. The researcher can define examinee and item population parameters according to the research needs. Moreover, the scale of sample can be very large. We simulate item responses using the Monte Carlo simulation method [14]. Once response data were simulated, the data is calibrated using MULTILOG program [15]. Ability parameter estimation provided from this program is compared with the actual simulated data in order to evaluate their accuracy.

To estimate examinees’ ability using dichotomous model, parameters for 30 items are simulated using a normal (0, 1) for discrimination parameters, a normal (0, 1) for difficulty parameters, and a normal (0, 0.05) for guessing parameters. Ability parameter is generated with a normal (0, 1), which is used to simulate 1,000 examinees. Item response data (coded 0 for wrong and 1 for correct) is obtained based on these simulated parameters. This process is repeated to generate parameters 30 items and 3000 examinees in the dichotomous model. Hence, the dataset sizes and complexities are different and each data set is represented as below.

Data sample I: 250 examinees each with 30 dichotomous responses to items;
Data sample II: 500 examinees each with 30 dichotomous responses to items.

We employed Multi Layer Perceptron (MLP) neural network in order to achieve the objective of the study. This type of network is trained with back-propagation learning algorithms. MLP neural network is chosen because it provides a quality result without being very complex. It is comprised of an input layer, any number of hidden layers, and an output layer. The responsibility of the input layer is to receive the external data and propagate it towards the hidden layer. Choosing the number of hidden layers and hidden neuron is done empirically and there is no specific rule in selecting the numbers.

In this work, the number of input is equal to the number of items (30). Number of examinees (1000, 3000) is considered as the number of sample for training while output layer represents examinee’s ability. Hence, the structure of network is composed of thirty input layers, one hidden layer and one output layer (30-X-1). The number of hidden neurons (X) should be determined empirically in order to offer a neural network with the best performance. Figure 1 shows the block diagram of neural network based on CAT.
Most of MLP networks have three layers, especially for the problem of function approximation and regression area. More layers will make the network much more complex and the convergence much slower, although more layers might increase the accuracy and decrease the error. In this study, the experiment is started using one hidden layer and then adding another hidden layer. To build a reliable and high performance model, this work adds neurons in the middle layer to get more precise results. Thus, the MLP neural network employed is thirty input layers, one hidden layer, and one output layer. This work aims to obtain the best number of hidden neuron in terms of small MSE. First MLP network with 5 hidden neurons is created and the effect of the increasing number of hidden neurons is investigated on the performance of the network. In hidden layer, we attempt to use 5, 10, 15 and 20 neurons. Thus, the four architectures of the network are created, including 30-5-1, 30-10-1, 30-15-1 and 30-20-1.

In the next stage, we trained the data in MLP network using Back Propagation (BP) algorithms. Different optimization BP algorithms are experimented which include gradient descent with momentum, variable learning rate method, resilient BP algorithm, conjugate gradient algorithm, Quasi-Newton algorithm, and Levenberg-Marquardt algorithm. After constructing the network with appropriate numbers of hidden layers and hidden neurons and after training the network, the parameters are fixed. The training process stops as soon as one of the stop criteria is met. After training of network, it is important to use a test set to evaluate the network performance before deploying it to the real world. In this work, the testing data consists of the generated responses that weren’t used in the training phase. The data response matrix with 30 items (dichotomous) and 250 and 500 examinees are taken into account to test the network. The same data set is used to estimate the ability parameter by IRT model. The ability estimations are obtained using EAP, MAP, and MLE estimation method based on IRT model.

Two simulation studies are considered to estimate ability of examinees. The design of the first study includes the dichotomous IRT model (3-PL), test lengths (n=30), and two sample sizes (N=250 and 500). The two samples sizes represent small and moderate samples. The accuracy of ability estimation for each condition is analysed using mean square error (MSE) between the true and estimated parameter. MSE calculates average square error between the true and estimated parameter.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i)^2$$

where,

$x_i$ and $y_i$ are the true and estimated parameter.

**RESULTS AND DISCUSSION**

In this experiment, different hidden neurons (5, 10, 15, and 20) are experimented to achieve the minimum error between the estimated ability and true ability. The true ability parameter was generated by Monte Carlo simulation. The mean square error (MSE) is used to evaluate the performance of the network. The main goal is to achieve minimum error after training the network.

As stated before, there is no specific rule to determine the exact number of hidden layers and neurons for the network’s topology. To find out how many hidden layers and neurons can possibly provide better performance in terms of least error of estimation; we design a network with one hidden layer in which there are five hidden neurons. Therefore, four topologies of neural networks are created in this study; 30-5-1, 30-10-1, 30-15-1 and 30-20-1. Next, the network is trained with the above-mentioned four topologies. The response data matrix,
which includes responses of 1000 examinees to 30 dichotomous, is submitted to the network. The sample data is subdivided into training (500), validation (250), and test sets (250) to prevent from over fitting. The back-propagation algorithms are utilized to train the network since the purpose is to find which algorithm obtains the smallest error between the actual and estimated ability value. In this section, 24 experiments are performed (4 topologies with 6 algorithms) for each sample size to find the optimization network. After training, the test data (unseen data) are fed into the network to investigate the minimum error. Table 1 displays the MSE of the network in different topologies and back-propagation algorithms. Topology (30-10-1) via the gradient descent with momentum algorithm and the 30-10-1 topology using variable learning rate BP algorithm produce the smallest error (0.1975, 0.1918) compared to other topologies.

<table>
<thead>
<tr>
<th>Topology</th>
<th>Gradient Descent with momentum</th>
<th>Variable Learning Rate</th>
<th>Resilient BP</th>
<th>Scaled Conjugate Gradient</th>
<th>Quasi-Newton</th>
<th>Levenberg-Marquardt</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-5-1</td>
<td>0.1986</td>
<td>0.1991</td>
<td>0.1901</td>
<td>0.1986</td>
<td>0.1941</td>
<td>0.1981</td>
</tr>
<tr>
<td>30-10-1</td>
<td>0.1975</td>
<td>0.1918</td>
<td>0.206</td>
<td>0.1975</td>
<td>0.203</td>
<td>0.2016</td>
</tr>
<tr>
<td>30-15-1</td>
<td>0.3601</td>
<td>0.2116</td>
<td>0.2127</td>
<td>0.3601</td>
<td>0.1986</td>
<td>0.2044</td>
</tr>
<tr>
<td>30-20-1</td>
<td>0.4943</td>
<td>0.2098</td>
<td>0.2229</td>
<td>0.4943</td>
<td>0.2112</td>
<td>0.2387</td>
</tr>
</tbody>
</table>

The 30-5-1 topology of Resilient BP method, the 30-10-1 topology of Scaled Conjugate Gradient, the 30-5-1 topology of Quasi Newton, and the 30-5-1 topology of Levenberg-Marquardt generate the minimum error (0.1901, 0.1975, 0.1941 and 0.1981) respectively. The results show that topology 30-5-1 using Resilient BP algorithm obtained the smallest error (0.1901) compared to the other topologies (see table 1).

The next table shows the estimation ability level of 500 examinees. Since this is a moderate sample data, therefore, a large data set is needed for training. The matrix that includes responses of 3000 examinees to 30 dichotomous items is submitted into the network. The sample data is subdivided into training (1500), validation (1000), and test sets (500) to prevent over fitting. Four topologies of neural networks are designed again i.e. 30-5-1, 30-10-1, 30-15-1, and 30-20-1. Different BP algorithms are used to train the network since the goal is to find an algorithm that obtains the smallest error between the actual and estimated ability value.

<table>
<thead>
<tr>
<th>Topology</th>
<th>Gradient Descent with momentum</th>
<th>Variable Learning Rate</th>
<th>Resilient BP</th>
<th>Scaled Conjugate Gradient</th>
<th>Quasi-Newton</th>
<th>Levenberg-Marquardt</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-5-1</td>
<td>0.1428</td>
<td>0.1440</td>
<td>0.1473</td>
<td>0.1436</td>
<td>0.1472</td>
<td>0.1462</td>
</tr>
<tr>
<td>30-10-1</td>
<td>0.1858</td>
<td>0.1447</td>
<td>0.1459</td>
<td>0.1439</td>
<td>0.1445</td>
<td>0.1425</td>
</tr>
<tr>
<td>30-15-1</td>
<td>0.1939</td>
<td>0.1526</td>
<td>0.1446</td>
<td>0.1459</td>
<td>0.1414</td>
<td>0.1506</td>
</tr>
<tr>
<td>30-20-1</td>
<td>0.2018</td>
<td>0.1455</td>
<td>0.157</td>
<td>0.1498</td>
<td>0.1441</td>
<td>0.1715</td>
</tr>
</tbody>
</table>

Table 2 above displays MSE of the network in different topologies and BP algorithms. The topology (30-5-1) via the Gradient Descent with momentum algorithm and the 30-10-1 topology using variable learning rate BP algorithm produce smaller error (0.1428, 0.1447) compared to the other topologies. The 30-15-1 topology via Resilient BP algorithm, the 30-5-1 topology with Scaled Conjugate Gradient, 30-15-1 topology using Quasi-Newton, and 30-10-1 topology with Levenberg-Marquardt generate the minimum error of (0.1446, 0.1436, 0.1414 and 0.1425) respectively. The results show that the topology 30-15-1 using Quasi-Newton BP algorithm obtained the smallest error (0.1414) compared to the other topologies and algorithms.

We performed the same experiment to estimate the ability parameters using MLE, MAP, and EAP based on IRT model. The responses of 250 and 500 examinees are imported into MULTILOG program. The two tables below show the MSE for both data samples.
Table 3 and 4 display the mean square error (MSE) of estimating examinees’ ability using IRT model. The results show that EAP produces the smallest error for data sample size of 250 examinees while MAP produces the smallest error for data sample size of 500 examinees.

### Table 3: MSE of ability estimation using IRT model for data sample I

<table>
<thead>
<tr>
<th>Estimation method</th>
<th>Mean Square Error (MSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>0.5596</td>
</tr>
<tr>
<td>MAP</td>
<td>0.4956</td>
</tr>
<tr>
<td>EAP</td>
<td>0.4373</td>
</tr>
</tbody>
</table>

### Table 4: MSE of ability estimation using IRT model for data sample II

<table>
<thead>
<tr>
<th>Estimation method</th>
<th>Mean Square Error (MSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>0.4339</td>
</tr>
<tr>
<td>MAP</td>
<td>0.3573</td>
</tr>
<tr>
<td>EAP</td>
<td>0.3862</td>
</tr>
</tbody>
</table>

MLE, MAP, and EAP methods are used to estimate the ability parameter. The item parameters (difficulty, discrimination and guessing) are known because they were generated by Monte Carlo simulation. These parameters are submitted to the MULTLOG program to approximate examinee's ability.

![Figure 2: Comparison of estimation error using neural network and IRT model for small and moderate sample size](image)

The figure above depicts the comparison of estimation error using MLP neural network model and IRT model. It was observed that MLP neural network model obtains lower estimation error for small and moderate sample size in comparison with MLE, MAP, and EAP methods of IRT model.

**CONCLUSIONS**

In this work, we have proposed an intelligent method to measure the examinees’ ability in Computer Adaptive Testing (CAT) using neural network. CAT technique uses a conventional approach i.e. Item Response Theory to estimate examinees’ ability. Due to the issues and limitations of IRT model, the MLP neural network is introduced to improve the ability estimation of CAT. The experimental result showed that the MLP neural network model obtained a smaller error for small sample sizes (N=250, N=500) against MLE, MAP, and EAP methods of IRT model. Hence, the MLP neural network has proven to be a suitable model to estimate examinees’ ability for small sample size.
REFERENCES


