Computational Diagnostic Classification Model using Deep Feedforward Network based Semi-Supervised Learning

Kang Xue
Department of Educational Psychology
University of Georiga
kangxue@uga.edu

ABSTRACT

The purpose of cognitive diagnostic modeling is to classify students' latent attribute profiles from the designed assessments. When analyzing a particular assessment dataset, inappropriate diagnostic classification model and inaccurate Q-matrix impact the classification accuracy. In contrast to existing research which added new parameters or rebuilt statistic models, the classification rate of DINA and DINO models were not accurate enough in experimental comparisons. In this paper the misclassification results using DINA and DINO models were viewed as incomplete labels for examinees. A semi-supervised learning framework combining Co-Training module and Deep Feedforward Networks (DFN) module is proposed to achieve a robust classification result using these incompletely classification results. Both simulated study and real assessment data based study were conducted to compare the performance between the proposed method and 5 widely used DCMs. The experimental results showed that the proposed method obtained accurate and robust classification rates across different test conditions and was more robust to the Q-matrix misspecification.

1. INTRODUCTION

The purpose of cognitive diagnostic modeling (CDM) or diagnostic measurement is to provide students' skill/attributes mastery status (mastery or non-mastery) through their responses to items from carefully designed assessments. Because of the ability to provide educators diagnostic feedback from students assessment results, CDM have been the focus of much research in the last decade. Various types of diagnostic classification models (DCMs), such as the deterministic inputs, noisy and gate (DINA) [13], the reparametrized unified model/fusion model (RUM) [10], and the log-linear cognitive diagnosis model (LCDM) [12], are designed based on different cognitive theories or hypotheses about how attributes behave, or interact, to produce individual item.

When analyzing a particular assessment dataset, selecting inappropriate DCMs (model misspecification) impacts the classification accuracy and parameter estimation. For example, when the attributes measured by an assessment are non-compensatory, which indicates that non-mastery on one attribute cannot be compensated by mastery on another attribute, selecting a compensatory model will decrease the performance of classification and measurement. In the most recent research, a common method of selecting appropriate DCM is to apply various potential DCMs to the data and compare their performance using some statistic model evaluation criteria, such AIC and BIC. Whereas the conclusion of such comparison might differ when using different criteria. For example, when analyzing Michigan English Language Assessment Battery (MELAB) reading test dataset [15], the full generalized deterministic input noisy and gate model (G-DINA) [7], was found a better fit in terms of its Akaike information criterion (AIC), whereas the additive CDM (A-CDM) [7], one of the main effects models, was superior in terms of the Bayesian information criterion (BIC). Meanwhile, the restriction of models also affects the performance of applying DCMs. In one recent comparison among DCMs to analyze TIMSS 2007 fourth grade mathematic assessment, DINA [13] and DINO [21] models achieved worse fit than did the other more relaxed DCMs, such as G-DINA, LCDM and R-RUM because both DINA and DINO might be too restrictive to reflect actual students' knowledge status [23].

Although more general DCMs provide more accurate classification results through adding more parameters to the statistic model, the performance is sensitive to the assessment conditions and priori information such as attributes architecture and accuracy of Q-matrices. A Q-matrix indicates the relationship between items and attributes in an assessment. Q-matrices are often carefully designed by assessment experts, whereas some existing research and their experimental results have shown that Q-matrices constructed by content experts do not always reflect the relationship precisely and may require empirically-driven modifications [2; 22]. Furthermore, the Q-matrices for most large-scale assessment are not always known completely and must be estimated to establish the associations between items and attributes. The existing research showed that, generally, the DINA, attribute hierarchy model (AHM), and rule space model (RSM) were mostly used with math; the R-RUM and general models (e.g. G-DINA, LCDM) were mostly used with reading [20], however, selecting an appropriate DCM or design new statistic models to reduce misclassification for a particular assessment still took lots of research effects.

To find the best way to convert student's response pattern to a diagnostic classification, Artificial Neural Networks (ANNs) have been proposed as an attractive approach [5; 6], according to the increasing of data size and development of computational power. In the existing research, both ANNs for supervised learning (e.g. multi-layer perceptron, MLP) and unsupervised learning (e.g. self-organizing map, SOM) were used to classify students into different latent groups.

For supervised learning ANNs, the challenge is that true latent class labels for students are not available to train the parameters of ANNs. To overcome this problem, a procedure [5] was conducted to synthesize the fake ideal item responses using ideal attribute patterns and hypothesized DINA model (both slipping and guessing were equal to 0) and used such synthesized responses to train MLP. Their simulated study showed that the classification accuracy of MLP is not as good as DINA because the training process could not find the best optimization for ANN using such insufficient training data. For unsupervised learning ANNs, because the computation process of ANN is viewed as "black box", the outputs of ANNs may cause class switching and unexplainable results. The experimental results [5] showed that the performance in classification is not as good as MLP. In addition, all current ANN methods were only applied to simulated data based on DINA model and only focused on supervised learning and unsupervised learning.

In contrast to existing methods which tried to find the best way for diagnostic classification by adding new parameters, rebuilding statistic models or simply using ANNs, the misclassification obtained by an inappropriate DCM is viewed as noisy or incomplete labels for examinees in this paper. According to this novel point of view, a semi-supervised learning is introduced to make use of such incomplete label obtained by inappropriate DCMs for training. In machine learning field, semi-supervised learning [24] falls between supervised learning (with completely labelled training data; e.g. regression, classification) and unsupervised learning (without any labelled training data; e.g. clustering, dimensional reduction). To handle the incomplete labels, Bootstrapping ("self-training") [9] built an initial classifier using the correctly labelled examples, and then iteratively classified unlabeled/mislabeled examples, updating the rules for the classifier using the expanded training data, and repeating these steps until convergence; Co-Training [18] uses a pair of classifiers with separate views of the data to iteratively learn and generate additional training labels. More recently, the techniques to solve the training using noisy labels using artificial neural networks have begun to receive attention, such as Restricted Boltzmann Machine (RBM) [14] and Generative Stochastic Networks [1]; also developed the deep neural network with robust loss function was also developed [16] to handle label-omission and registration error.

In this paper, to find a better way to convert response pattern to latent attribute profiles under different assessment conditions, a semi-supervised learning method combined a Co-Training module and a Deep Feedforward Network (DFN) module was developed to refine the classification accuracy using inappropriate DCMs. In the following sections, the structure of this framework is firstly described. In addition, the experimental results to compare the proposed method and 5 widely used DCMs under both simulated and real assessment-based experiments are illustrated. Lastly, the benefits and challenges of this methodology are summarized, and the future research is also outlined.

2. PROPOSED FRAMEWORK

The proposed framework in this paper consisted of two modules: Co-Training module and DFN module. The Co-Training module used a pair of DCMs with separate hypothesis of the attributes measured in an assessment to learn and generate the incomplete training labels. The DFN module was to conduct a semi-supervised learning procedure using response patterns to convert an observed response pattern to a latent attribute profile. Figure 1 shows the structure of the proposed framework. In this section, detailed description of these two modules is introduced.

2.1 Co-Training Module

Typical Co-Training algorithm is a semi-supervised learning requires two views of the data. It assumes that each observation can be represented using two different types of descriptions that with various and compensatory information about the instance. In this paper, we designed the Co-Training module according to the idea of typically Co-Training algorithm. To make the hypotheses of the attributes from two classifiers separated and compensatory, DINA and DINO models were selected because the attribute information contained by DINA and DINO models were compensatory.

DINA model is a non-compensatory or conjunctive DCM means that lack of one attribute cannot be compensated by the mastery of another attribute measured by an item. The DINA model classifies students into two groups for each item, those who have mastered all the attributes required by an item and those who have not. The jth item response probability of the ith student can be written as:

$$P(y_{ij} = 1|\xi_{ij}, s_j, g_j) = (1 - s_j)^{\xi_{ij}} g_j^{(1-\xi_{ij})} \quad (1)$$

where $\xi_{ij} = 1$ indicates the student has mastered all required attributes and $\xi_{ij} = 0$ refers to no-mastery status, s_j and g_j are the slipping parameter and guessing parameter of the jth item.

In contrast to DINA model, DINO model is a compensatory or disjunctive DCM which means that a non-mastery on one latent attribute can be compensated for by a mastery status on another attribute. The jth item response probability of the ith student can be written as:

$$P(y_{ij} = 1|\omega_{ij}, s_j, g_j) = (1 - s_j)^{\omega_{ij}} g_j^{(1-\omega_{ij})} \quad (2)$$

where the latent response $\omega_{ij} = 0$ indicates mastery of at least one measured attribute and $\omega_{ij} = 1$ indicates absence of all required attributes, like DINA, s_j and g_j are the slipping parameter and guessing parameter of the jth item.

In addition to the information compensation between DINA and DINO, the second reason of choosing these two DCMs as classifiers of Co-Training module was that both DINA and DINO are two earliest DCMs and have been widely used as baselines when introducing a new DCM. Due to the constraints of DINA and DINO, the classification processes can converge easily under different assessment data and the outputs are valid in contrast to relaxed DCMs. However, the classification rates of DINA and DINO models are not as good as the general DCMs in most research papers.

In the Co-Training module, once both DINA and DINO model were fitted according to responses and Q-matrix, the probability of ith student with response pattern X_i belongs to latent class c_1 under DINA and latent class c_2 under DINO models were denoted as $P(\{\tilde{Y}^1_{i,c_1}, \tilde{Y}^2_{i,c_2}\}|X_i)$, c_1 and c_2 could be either same or different.

2.2 Deep Feedforward Network Module

Since the classification results from DINA $\tilde{Y}^1_{i,c_1}$ and DINO $\tilde{Y}^2_{i,c_2}$ were incomplete if they were inappropriate models for

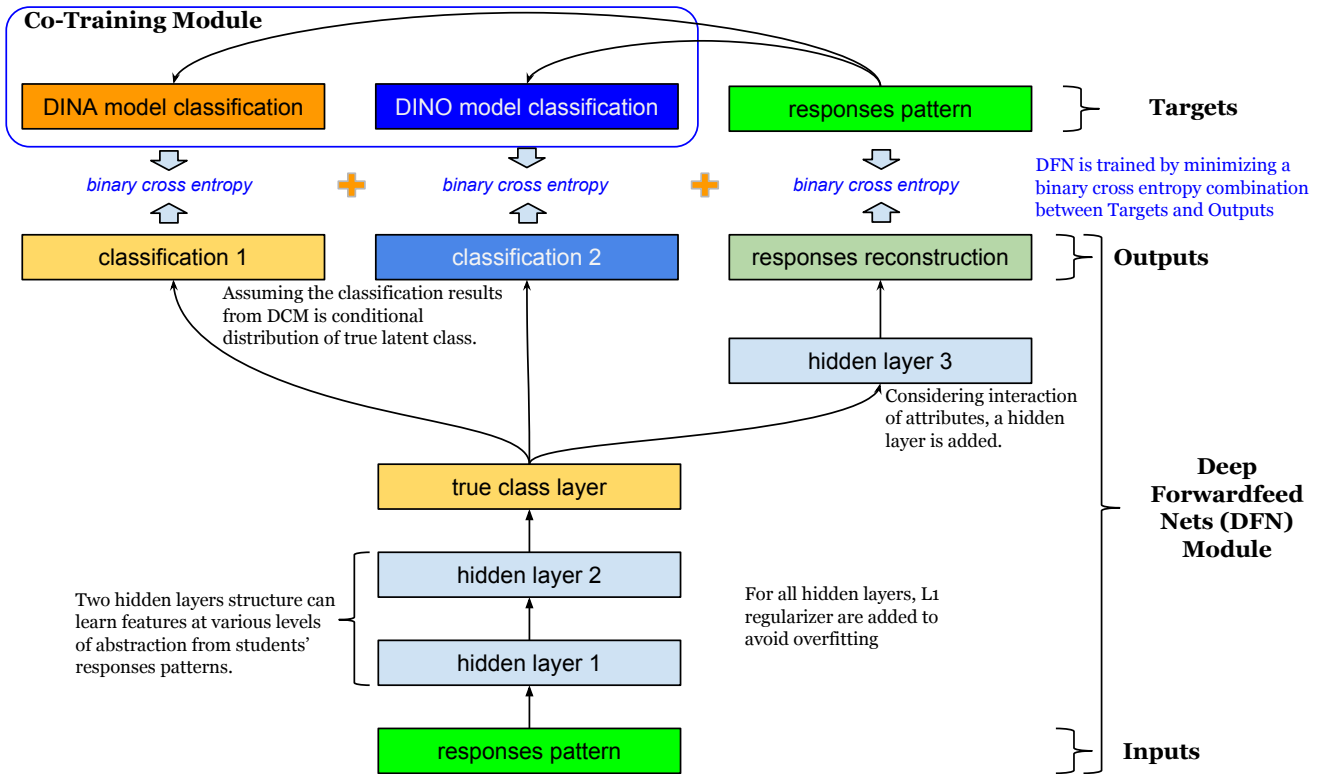

Figure 1: The diagram shows the structure of our semi-supervised learning process by combining Deep Feedforward Network module and Co-Training module. Although in our experiment, we will use DINA and DINO models, any combination of two DCMs is available for this network structure.

assessment data, these two classifications were assumed to be conditional distribution of the true latent class of the ith student $T_{i,c}$. Suppose $P(T_{i,c}|X_i)$ refers to the probability that ith student with response pattern X_i belongs to the true latent class $T_{i,c}$, $\{\tilde{Y}^1_{i,c_1}, \tilde{Y}^2_{i,c_2}\}$ and X_i are conditional independent when giving $T_{i,c}$, and the classification results of DINA and DINO can be represented as following:

$$P(\{\tilde{Y}^1_{i,c_1}, \tilde{Y}^2_{i,c_2}\}|X_i) = P(\{\tilde{Y}^1_{i,c_1}, \tilde{Y}^2_{i,c_2}\}|T_{i,c})P(T_{i,c}|X_i) \quad (3)$$

However, in equation 3, the two conditional distributions $P(\{\tilde{Y}^1_{i,c_1}, \tilde{Y}^2_{i,c_2}\}|T_{i,c})$ and $P(T_{i,c}|X_i)$ are not easy to be directly represented using statistical representations before doing data analysis. Instead, we conducted a Deep Feedforward Network (DFN) module to approximate these two distributions. Deep feedforward networks [8], also called feedforward neural networks, or MLPs, are the quintessential deep learning models. The goal of a feedforward network is to approximate some functions { because the universal approximation theory [4] states that every continuous function that maps intervals of real numbers to some output interval of real numbers can be approximated arbitrarily closely by a deep feedforward networks with just one hidden layer. Using the DFN, these two conditional distributions $P(\{\tilde{Y}^1_{i,c_1}, \tilde{Y}^2_{i,c_2}\}|T_{i,c})$ and $P(T_{i,c}|X_i)$ can be approximated as following:

$$P(\{\tilde{Y}^1_{i,c_1}, \tilde{Y}^2_{i,c_2}\}|T_{i,c}) \approx P(\{\hat{Y}^1_{i,c_1}, \hat{Y}^2_{i,c_2}\}|\hat{T}_{i,c}) = \Psi(\hat{T}_{i,c}) \quad (4)$$

$$P(T_{i,c}|X_i) \approx P(\hat{T}_{i,c}|X_i) = \Phi(X_i) \quad (5)$$

As shown in Figure 1, $\Phi(\cdot)$ took a student's response pattern X_i as input (input layer) and used 2 hidden layers (hidden layer 1 & 2) to compute the nodes values of true label layer (true latent class of examinees). In the framework, the aim of adding multiple hidden layers was to learn more abstractive information from the students' responses. $\Psi(\cdot)$ took the estimated values of true label layer $\hat{T}_{i,c}$ as inputs and computed the classification 1 ($\hat{Y}^1_{i,c_1}$) and classification 2 ($\hat{Y}^2_{i,c_2}$) as the outputs. Thus, $P(\{\tilde{Y}^1_{i,c_1}, \tilde{Y}^2_{i,c_2}\}|X_i)$ were approximated by $\Psi \circ \Phi(X_i)$.

In the applications of DFN for supervised learning, when giving a completely labelled training dataset, the parameters used in DFN (e.g. weights and biases) can be estimated by minimizing the difference between outputs and targets using the back-propagation algorithm [8]. Whereas, the targets from Co-Training module $\{\tilde{Y}^1_{i,c_1}, \tilde{Y}^2_{i,c_2}\}$ are noisy if the DINA and DINO are inappropriate DCMs. Using such incomplete classification results, the assumption that targets are to be unambiguous and accurate cannot hold any longer. The DFN will exhibit poor performance because the parameter training depended critically on the accuracy of training samples [5]. To handle the impact from misclassification from inappropriate DCMs, as shown in Figure 1, a new reconstruction target was added to the DFN. Since in theoretical DCMs, a student's response pattern is conditional distribution of the attribute profile and item parameters $P(X_i|T_{i,c}, \theta)$ (θ are the set of item parameters), the reconstruction of a response pattern can be calculated using the values of true label layer. The reconstructed response

pattern of ith examinee was calculated as $\hat{X}_i = \Upsilon(\hat{T}_{i,c})$, where $\Upsilon(\cdot)$ was the calculation procedure through hidden layer 3. The training of our DFN is to minimize the following binary cross-entropy combination of two targets:

$$\{\omega, b\} = \arg\min[w_1 H(\hat{Y}^1_{i,c_1}, \tilde{Y}^1_{i,c_1}) + w_2 H(\hat{Y}^2_{i,c_2}, \tilde{Y}^2_{i,c_2}) + w_3 H(\hat{X}_i, X_i)] \quad (6)$$

where $\{\omega, b\}$ are the weights and biases set of DFN. Noting that $Y^1_{i,c_1}, Y^2_{i,c_2}, T_{i,c} \in \{0,1\}^C$, $\sum_{c_1} Y^1_{i,c_1} = \sum_{c_2} Y^2_{i,c_2} = \sum_c T_{i,c} = 1$, C refers to the number of latent classes. Considering computational efficiency, we chose Rectified Linear Unit function ($ReLU$) [17] as the activation function to compute the output of the nodes for all three hidden layers (hidden layer 1, 2 & 3) and added L_1 regularization term to avoid over-fitting for each hidden layer. The activation functions of true label layer and output layer were *SoftMax* function because each node on these two layers was binary variable. The estimates of parameters within the DFN, $\{\omega, b\}$, and the weights parameter $\{w_1, w_2, w_3\}$ of the binary cross-entropy combination in equation 6 were determined using cross validation.

One challenge of applying ANNs in the field of CDM is that the estimation of the latent variables (e.g. attributes or latent classes) can vary, sometimes dramatically [3] when the ANN is trained multiple times using the same data. To handle this issue in our framework, a voting strategy was introduced by running the training procedures multiple times. The estimated probabilities of belonging to each latent class $P(\hat{T}_{i,c}|X_i)$ from multiple training were averaged to get the averaging probability $\bar{P}(\hat{T}_{i,c}|X_i)$. Then, the latent class c_i that ith student belonged to were determined as following:

$$c_i = \arg\max(\bar{P}(\hat{T}_{i,c}|X_i)) \quad (7)$$

3. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed framework and make comparison with the widely used DCMs, we conducted both simulation study and real assessment data study. In this section, the methods and results of these two studies will be discussed.

3.1 Simulated study

3.1.1 Method

In the previous research of DCMs which simulated data based on the framework of a specific DCM (e.g. DINA, RRUM and G-DINA), in this paper, the response data were simulated using a general $I \times C$ item by latent class matrix without the specific mathematic representation:

$$\Pi = \begin{bmatrix} \pi_{1,1} & \pi_{1,2} & \cdots & \pi_{1,C} \\ \pi_{2,1} & \pi_{2,2} & \cdots & \pi_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{I,1} & \pi_{I,2} & \cdots & \pi_{I,C} \end{bmatrix} \quad (8)$$

where the conditional probability that students in lth latent class answer ith item correctly $P(X_i = 1|\alpha_c) = \pi_{i,c}$. I indicated the number of items, C indicated the number of latent classes.

Under the framework of item by latent class matrix, we manipulated four assessment factors in the simulation, including the number of items (20 or 30), number of attributes (3 or 4), item discrimination (high or mixed), Q-matrix accuracy (100% or 90% accurate) and sample size (1000). The number of items (20 or 30) and the number of attributes were selected to reflect the current real assessment applications which often contained between 20 to 30 items and measured 3 or 4 attributes (e.g. MELAB data, DTMR data). For 3 attributes, we only generated 20 items and for 4 attributes, 20 and 30 items were generated respectively. Item discriminating power is another factor impact performance of DCMs. Students who have mastered the attributes measured by an item with high discrimination are expected to have a higher probability of responding correctly than the student who have not mastered the attributes. Two levels of item discrimination were examined in the simulation: high discrimination indicated the probability differences between two groups of students to respond to all items are over 0.3; mixed discrimination indicated that the differences for 50% items were over 0.3 and for the rest 50% items were less than 0.3. Two levels of Q-matrix accuracy were also simulated because the impact of Q-matrix accuracy is critical to link the DCMs to students' responses. 100% accuracy indicated that the Q-matrix were completely known; 90% accuracy indicated that 10% of elements in each Q-matrix were mis-specified.

In this simulated study, as comparison, 5 types of widely used DCMs were introduced as baselines to evaluate the diagnostic classification performance of the proposed framework. As the two classifiers used in the Co-Training module, DINA and DINO models were the selected. In addition, we chose two general models G-DINA, LCDM and a non-compensatory model RRUM as the other three base models. All 5 DCMs were conducted using "CDM" package in **R**, the proposed semi-supervised learning method was conducted using "tensorflow" library in **Python**.

3.1.2 Results

Table 1 showed the classification rates for tests with 3 attributes and 20 items using different classification methods, with respect to item discrimination and Q matrix accuracy. Table 2 and 3 showed the classification rate for tests measured 4 attributes and contained 20 and 30 items respectively, with respect to item discrimination and Q matrix accuracy. For each condition, as expected, the more relaxed DCMs (LCDM, G-DINA and RRUM) had a better classification performance at both individual attribute level and the class level (i.e. attribute pattern level) than DINA and DINO which hold a hard constraint. When the assessment condition was ideal (i.e. high item discrimination and 100% accurate Q-matrix), the two general models, LCDM and G-DINA, always achieved the best classification accuracy.

Simulation results indicated that using the proposed method (DFN), the classification rates were higher DINA and DINO, the two initial classifiers used in Co-Training module. Compared to DINA and DINO models, at the attribute level, the average improvements of classification using DFN was .0218 and .0140, and at the attribute pattern (class) level, the average improvements were .0589 and .0432. Compared to the general models LCDM and G-DINA, which often achieved the best performance in classification, the performance of DFN was also better than these two methods. The improvements at attribute level were .0056 and .0055 in contrast to LCDM and G-DINA models respectively. At attribute pat-

Table 1: Comparison of classification rates for 3 attributes using 20 items. The DFN indicates our proposed method; discriminating is the discriminative power of the test; Q-matrix is the accuracy of the Q-matrix; A1, A2, A3 and Class are the classification accuracy at three attribute level and pattern level respectively.

Methods	discriminating	Q-matrix	A1	A2	A3	Class
DINA	High	100%	0.949	0.864	0.957	0.778
DINO			0.953	0.871	0.952	0.784
LCDM			0.96	0.917	0.957	0.842
GDINA			0.96	0.917	0.957	0.842
RRUM			0.953	0.91	0.958	0.827
DFN			0.956	0.915	0.957	0.834
DINA		90%	0.944	0.824	0.957	0.741
DINO			0.946	0.852	0.944	0.757
LCDM			0.956	0.897	0.958	0.819
GDINA			0.956	0.897	0.958	0.819
RRUM			0.949	0.879	0.958	0.794
DFN			0.955	0.9	0.958	0.821
DINA	Mixed	100%	0.875	0.859	0.914	0.693
DINO			0.863	0.864	0.896	0.665
LCDM			0.879	0.884	0.913	0.712
GDINA			0.879	0.884	0.913	0.712
RRUM			0.873	0.9	0.917	0.724
DFN			0.883	0.884	0.915	0.720
DINA		90%	0.878	0.85	0.906	0.676
DINO			0.869	0.861	0.908	0.679
LCDM			0.878	0.85	0.918	0.685
GDINA			0.877	0.85	0.918	0.684
RRUM			0.877	0.85	0.915	0.685
DFN			0.874	0.888	0.908	0.704

tern (class) level, the improvements were .0130 and .0132.

The simulation results also indicated that when the Q-matrix became less accurate, the classification accuracy for each method dropped at both attribute level and attribute pattern (class) level when other test assessment factors were hold. When the Q-matrix accuracy decreased to 90% accurate, at the attribute level, the average reductions of classification accuracy were .0071, .0055, .0114, .0114, .0095 and **.0038** corresponding to DINA, DINO, LCDM, G-DINA, RRUM and **DFN** methods respectively. At the attribute pattern level, the average accuracy reductions were .0163, .0138, .0298, .0302, .0243 and **.0075** for DINA, DINO, LCDM, G-DINA, RRUM and **DFN** methods respectively. From this observation we could find that firstly, the relaxed models (LCDM, G-DINA and RRUM) were more sensitive to the accuracy of Q-matrix; secondly, the proposed method was more robust to the noise within the Q-matrix compared to the five DCMs.

In addition, high item discriminating was a positive impact on the classification accuracy of all 6 methods. When the discrimination of items decreased (from high to mixed), the classification rate dropped .0301, .0383, .0458, .0458, .0392 and **.0397** for DINA, DINO, LCDM, G-DINA, RRUM and **DFN** methods at the attribute level. The reductions were .0780, .1095, .1318, .1318, .1137 and **.1158** for DINA, DINO, LCDM, G-DINA, RRUM and **DFN** methods at the attribute pattern (class) level. The reason that DFN method dropped more than DINA, DINO and RRUM (only at the attribute level) was that when the items were high discriminating, the improvement of classification rate using DFN was more significant than using mixed discriminating items. Even though, the performance of DFN at both attribute level and attribute pattern level were the best among the six diagnostic classification methods.

3.2 Real data study

3.2.1 Data

In addition to the simulated study, we also tested our proposed method on the Elementary Probability Theory dataset which is available in the R package "pks" [11]. The dataset contains 12 items and 504 examinees. 4 different attributes are measured: 1) the classic probability of an event (pb); 2) the probability of the complement of an event (cp); 3) the probability of the union of two disjoint events (un); 4) the probability of two independent events (id). Since there is no ground truth for this real data, we replaced the classification results of RRUM by A-CDM and used the A-CDM classification rates as the base line because A-CDM obtained the lowest BIC when applying to the dataset [19].

3.2.2 Results

In Table 4, one of our initial classifiers in Co-Training module, DINO model, achieved much worse classification rate according to the A-CDM results. At the attribute level, the classification results were .869, .819, .804 and .946. The classification rate at attribute pattern (class) level was only .714. The DINA model's classification rates were .950, .984, .990 and .994 at attribute level and .928 at attribute pattern (class) level. By using the proposed method, the performance of DFN was better than both DINA and DINO models because the algorithm adjusted the weights of two targets in equation 6. The classification rate at attribute

Table 2: Comparison of classification rates for 4 attributes using 20 items. The DFN indicates our proposed method; discriminating is the discriminative power of the test; Q-matrix is the accuracy of the Q-matrix; A1, A2, A3, A4 and Class are the classification accuracy at four attribute level and pattern level respectively.

Methods	discriminating	Q-matrix	A1	A2	A3	A4	Class
DINA	High	100%	0.908	0.924	0.79	0.893	0.591
DINO			0.909	0.928	0.858	0.899	0.653
LCDM			0.918	0.929	0.858	0.919	0.67
GDINA			0.918	0.929	0.858	0.919	0.67
RRUM			0.923	0.921	0.853	0.917	0.664
DFN			0.919	0.925	0.858	0.922	0.67
DINA		90%	0.909	0.922	0.74	0.886	0.56
DINO			0.903	0.924	0.852	0.879	0.621
LCDM			0.904	0.922	0.824	0.887	0.616
GDINA			0.904	0.922	0.824	0.887	0.616
RRUM			0.905	0.922	0.8	0.884	0.599
DFN			0.912	0.923	0.862	0.89	0.648
DINA	Mixed	100%	0.854	0.836	0.824	0.851	0.503
DINO			0.863	0.817	0.855	0.816	0.484
LCDM			0.867	0.823	0.855	0.84	0.509
GDINA			0.867	0.824	0.855	0.84	0.51
RRUM			0.878	0.831	0.855	0.837	0.522
DFN			0.864	0.842	0.857	0.859	0.531
DINA		90%	0.856	0.826	0.744	0.854	0.448
DINO			0.854	0.817	0.855	0.851	0.503
LCDM			0.865	0.817	0.776	0.844	0.469
GDINA			0.865	0.817	0.776	0.844	0.469
RRUM			0.864	0.821	0.855	0.84	0.509
DFN			0.852	0.871	0.855	0.852	0.542

level were .964, .988, .988 and .996. The attribute pattern (class) classification rate was .952. The performance was close to the general model LCDM and G-DINA. The results showed that even one initial classifier cannot achieve a good performance, the proposed method still has ability to obtain a more accurate classification rate than both two initial classifiers.

4. CONCLUSION AND DISCUSSION

The propose of this paper is to design a new semi-supervised learning method used to interpret student performance on diagnostic measurement assessment and to evaluate the performances of the proposed method using both simulation study and real assessment data. In the proposed framework, we viewed the classification results of inappropriate DCMs as incomplete labels and introduced a method by combining a Co-Training module and deep feedforward networks together. In Co-Training module, we used two basic DCMs, DINA and DINO models, as the initial classifiers. In the DFN module, by using two types of targets, the outputs of Co-Training module and true response patterns, was to find the correct classification results.

In the simulated study, we compared the proposed method with other five DCMs. Beside the two initial classifiers, DINA and DINO, 3 widely used reflaxed DCMs, LCDM, G-DINA and RRUM were also introduced. By varying the factors (item discrimination, Q-matrix accuracy, number of attributes and items) which impact the performance of DCMs, the comparison results indicated that the proposed method achieved better classification rates than the five DCMs across all assessment conditions at both attribute level and attribute pattern (class) level. In addition, the proposed method was robust to the Q-matrix mis-specification because the classification rate dropped less than the other five DCMs when the Q-matrix accuracy decreased to 90% accuracy. Although the classification rates of the proposed method dropped more than DINA and DINO when the item discriminating power reduced, the proposed method was more robust to the item discriminating reduction than the general DCMs. In the real assessment data-based study, the results indicated that even one of the classifiers achieved bad classification rate, the performance of our proposed method was better than both initial classifiers and also achieved as classification rates compared to general DCMs.

One concern of this study is that the current analysis only focused on the classification rate of the proposed method. In the future study, the classification results could be used to analyze item parameters to evaluate item discriminating power among students' mastery level for specific attributes or determine the relationship between items and attributes to explore the attribute structures. The classification results could also be useful to explore new type of DCMs when giving new assessment dataset.

In summary, the proposed method provided a novel point of view to classify students' attributes using a computational psychometric method by combining semi-supervised learning and theoretic DCMs. Both simulated study and real assessment data-based study showed the advantage of using this new strategy. Considering some limitations in this paper, a future study is needed to test the performance of this method on assessment data exploration.

Table 3: Comparison of classification rates for 4 attributes using 30 items. The DFN indicates our proposed method; discriminating is the discriminative power of the test; Q-matrix is the accuracy of the Q-matrix; A1, A2, A3, A4 and Class are the classification accuracy at four attribute level and pattern level respectively.

Methods	discriminating	Q-matrix	A1	A2	A3	A4	Class
DINA	High	100%	0.937	0.938	0.814	0.892	0.641
DINO			0.942	0.941	0.854	0.902	0.681
LCDM			0.947	0.949	0.873	0.925	0.732
GDINA			0.947	0.949	0.873	0.925	0.732
RRUM			0.948	0.945	0.872	0.917	0.719
DFN			0.949	0.944	0.872	0.916	0.722
DINA		90%	0.934	0.94	0.853	0.853	0.64
DINO			0.935	0.924	0.855	0.874	0.644
LCDM			0.948	0.946	0.858	0.92	0.708
GDINA			0.948	0.946	0.859	0.92	0.709
RRUM			0.945	0.945	0.869	0.915	0.713
DFN			0.952	0.948	0.873	0.916	0.723
DINA	Mixed	100%	0.903	0.876	0.8	0.882	0.56
DINO			0.911	0.884	0.858	0.858	0.586
LCDM			0.912	0.886	0.857	0.88	0.616
GDINA			0.912	0.886	0.858	0.88	0.617
RRUM			0.9	0.884	0.858	0.871	0.592
DFN			0.91	0.889	0.862	0.881	0.616
DINA		90%	0.908	0.887	0.847	0.876	0.603
DINO			0.906	0.883	0.852	0.836	0.566
LCDM			0.908	0.891	0.863	0.868	0.605
GDINA			0.908	0.891	0.863	0.868	0.605
RRUM			0.905	0.891	0.864	0.861	0.602
DFN			0.909	0.885	0.859	0.871	0.61

Table 4: Comparison of classification accuracy among 5 base models and the proposed method using the dataset in [11]; A-CDM was used as the true value because it achieved lowest BIC in [19].

Methods	pb	cp	un	id	Class
DINA	0.950	0.984	0.990	0.994	0.928
DINO	0.869	0.819	0.804	0.946	0.714
LCDM	0.986	0.976	0.988	0.998	0.958
GDINA	0.980	0.996	0.996	0.998	0.974
A-CDM	1.000	1.000	1.000	1.000	1.000
DFN	0.968	0.988	0.988	0.996	0.952

5. REFERENCES

[1] Y. Bengio, E. Laufer, G. Alain, and J. Yosinski. Deep generative stochastic networks trainable by backprop. In *International Conference on Machine Learning*, pages 226–234, 2014.

[2] L. Bradshaw, A. Izsák, J. Templin, and E. Jacobson. Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational measurement: Issues and practice*, 33(1):2–14, 2014.

[3] D. C. Briggs and R. Circi. Challenges to the use of artificial neural networks for diagnostic classifications with student test data. *International Journal of Testing*, 17(4):302–321, 2017.

[4] B. C. Csáji. Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24:48, 2001.

[5] Y. Cui, M. Gierl, and Q. Guo. Statistical classification for cognitive diagnostic assessment: An artificial neural network approach. *Educational Psychology*, 36(6):1065–1082, 2016.

[6] Y. Cui, Q. Guo, and M. Cutumisu. A neural network approach to estimate student skill mastery in cognitive diagnostic assessments. 2017.

[7] J. De La Torre. The generalized dina model framework. *Psychometrika*, 76(2):179–199, 2011.

[8] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[9] G. R. Haffari and A. Sarkar. Analysis of semi-supervised learning with the yarowsky algorithm. *arXiv preprint arXiv:1206.5240*, 2012.

[10] S. M. Hartz. *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. PhD thesis, ProQuest Information & Learning, 2002.

[11] J. Heller and F. Wickelmaier. Minimum discrepancy estimation in probabilistic knowledge structures. *Electronic Notes in Discrete Mathematics*, 42:49–56, 2013.

[12] R. A. Henson, J. L. Templin, and J. T. Willse. Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2):191, 2009.

[13] B. W. Junker and K. Sijtsma. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3):258–272, 2001.

[14] H. Larochelle and Y. Bengio. Classification using discriminative restricted boltzmann machines. In *Proceedings of the 25th international conference on Machine learning*, pages 536–543. ACM, 2008.

[15] H. Li, C. V. Hunter, and P.-W. Lei. The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, 33(3):391–409, 2016.

[16] V. Mnih and G. E. Hinton. Learning to label aerial images from noisy data. In *Proceedings of the 29th International conference on machine learning (ICML-12)*, pages 567–574, 2012.

[17] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[18] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Cikm*, volume 5, page 3, 2000.

[19] M. Philipp, C. Strobl, J. de la Torre, and A. Zeileis. On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics*, 43(1):88–115, 2018.

[20] J. Sessoms and R. A. Henson. Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, 16(1):1–17, 2018.

[21] J. L. Templin and R. A. Henson. Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, 11(3):287, 2006.

[22] H. Tjoe and J. de la Torre. On recognizing proportionality: Does the ability to solve missing value proportional problems presuppose the conception of proportional reasoning? *The Journal of Mathematical Behavior*, 33:1–7, 2014.

[23] K. Yamaguchi and K. Okada. Comparison among cognitive diagnostic models for the timss 2007 fourth grade mathematics assessment. *PloS one*, 13(2):e0188691, 2018.

[24] X. Zhu. Semi-supervised learning literature survey, department of computer sciences, university of wisconsin at madison, madison. Technical report, WI, Technical Report 1530. http://pages. cs. wisc. edu/˜ jerryzhu/pub . . . , 2006.