Computing and Informatics, Vol. 41, 2022, 1240–1259, doi: 10.31577/cai_2022_5_1240

GENERAL DEEP MULTINOMIAL LOGIT MODEL

Peng Su

School of Information Science and Engineering
Qilu Normal University
Jinan 250013, China
e-mail: peng_su@139.com

Yuan Liu

School of Literature
Qilu Normal University
Jinan 250013, China

Lingyun Zhao

School of Information Science and Engineering
Qilu Normal University
Jinan 250013, China

Abstract. Multinomial logit model (MNL) is by far the most widely used discrete choice model that is widely used to explain or predict a choice from a set of two or more discrete alternatives. MNL operates within a framework of the random utility model (RUM) in which the utility of an alternative perceived by an individual consists of two components: systematic component and random component. The systematic component is usually defined as a linear function. However, practical decision processes involve complex considerations regarding various aspects of the alternatives and individual which cannot be adequately represented by simple linear models. To overcome the weakness of linear utility model and improve the performance of MNL, in this paper, we propose a general deep multinomial logit model (GDMNL) that takes advantage of both traditional MNL and deep learning. In this model, deep neural networks are applied to extend MNL by learning differ-

ent nonlinear utility functions of various alternatives. The empirical study in the domain of transit route choice analysis demonstrates the validity and superiority of the proposed model.

Keywords: Deep learning, discrete choice model, feedforward neural networks, multinomial logit model, nonlinear logit model

1 INTRODUCTION

A fundamental concern of economics is understanding human choice behavior. Models or hypotheses are construed on the nature of decision processes, and are evaluated in terms of observed behavior. Among the most important and well-known models are discrete choice models (DCMs), which are widely used to explain or predict a choice from a set of two or more discrete (i.e. distinct and separable; mutually exclusive) alternatives. For example, a DCM may be used to explain decision such as labor force participation, brand choice, whether to invest, traffic model choice, and predict a recession.

DCMs operate within a framework of rational choice; that is, it is assumed that when confronted with a discrete set of alternatives, people choose the alternative of maximal benefit or utility. In the light of the random utility model (RUM) [1], the utility of alternative j perceived by individual i, U_{ij}, consists of two components: systematic component, which we can observe, and random component, which we cannot observe. Thus, the model is expressed as:

$$U_{ij} = V_{ij} + \varepsilon_{ij}, i \in \{1, 2, \ldots, N\} \text{ and } j \in \{1, 2, \ldots, J\} \tag{1}$$

where V_{ij} is nonstochastic and reflects the "representitive" tastes of the population, ε_{ij} is stochastic and reflects the idiosyncracies of individual i in tastes for alternative j, N is the number of individuals, and J is the number of alternatives available to N individuals. Note that an alternative available to one individual does not necessarily apply to another.

V_{ij} is typically defined as a utility function v_{ij}, which relates all relevant observed factors to the utility level of the alternative. Part of the DCM is that we assume v_{ij} is linear in all relevant parameters. Thus, v_{ij} can be broken down in terms of two vectors:

$$v_{ij}(\boldsymbol{x}_{ij}) = \boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{x}_{ij} \tag{2}$$

where $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijk})^{\mathrm{T}}$ is a vector that describes alternative j and/or covariates describing either the individual (e.g., sex, income) or some aspect related to the decision context (e.g., whether the purchase is for personal use of the individual or a gift), and $\boldsymbol{\beta}_j$ is an alternative specific vector of equivalent dimension, made up of fixed regression coefficients; it factors in how much each of the attributes of $\boldsymbol{x}_{ij}$ relates to the overall utility level.

The utility model based on linear assumption of v_{ij} can be easily implemented and offer high interpretability about the contribution of different attributes of the alternatives and individual. However, many researchers have pointed out that practical decision processes involve complex considerations regarding various aspects of the alternatives and individual, e.g. threshold effects (utility changes only if a certain value of an attribute is exceeded) or saturation effects (utility does not change if a certain value of an attribute is exceeded), which cannot be adequately represented by simple linear models [2, 3, 4].

Artificial neural networks (ANNs) have strong ability to approximate arbitrarily complex nonlinear multivariate function and its derivatives with the desired level of precision given a sufficient number of hidden units with nonlinear activation functions [5, 6]. This inspired some researchers to use ANN or deep learning to construct nonlinear utility model to capture nonlinear impacts in choice decision making.

For example, to discover nonlinear effects on brands' utilities in a flexible way, reference [7] specified deterministic utility by means of a certain type of neural net. By combining such a neural net with the most widespread choice model, the multinomial logit, this work remains within the predominant utility-maximizing framework.

Take another example, reference [8] proposed a deep choice model which extends the classical conditional/multinomial logit model by learning a nonlinear utility function identically for each bidder within of the job posts via a pointwise convolutional neural network. The pointwise property ensures that the same nonlinear function is applied identically to each bidder, defining a fully connected neural network architecture that maps each individual bidder's attributes to a scalar utility value.

Though all these works show that the extended logit model is superior to the corresponding conventional model in terms of some evaluation metrics such as negative log-likelihood, they have two common shortcomings.

Depending on what the alternatives being described are, there may exist a set of attributes that are common across all, or at least a subset of the alternatives. In such cases, the analyst may choose to constrain the parameters in linear utility functions (Equation (2)) across two or more of the alternatives to be the same and estimate what are known as generic parameters. Linear utility functions may also be specified to contain what are termed alternative specific parameter estimates. An alternative-specific parameter is one which is allowed to differ across alternatives.

In the above-mentioned extended logit models using ANN or deep learning to construct nonlinear utility model, all the attributes are common across all the alternatives. Hence, the alternatives share the same feedforward neural network architecture. However, in many domains, the alternatives should correspond network architectures with different connection weights.

In addition, all the proposed models lack the regularization strategies that encode specific kinds of prior knowledge. In choice analysis domain, there exists crucial prior knowledge about suitable values of the model parameters, which can be used to design effective regularization strategies to improve the model performance.

To overcome these two shortcomings, in this paper, we propose a general deep multinomial logit model that takes advantage of both traditional MNL and deep learning. In this model, deep neural networks are applied to extend the MNL by learning different nonlinear utility functions of various alternatives. In addition, a regularization strategy of parameter tying is proposed, which expresses prior knowledge of choice modelling about suitable values of the model parameters. The experiment results demonstrate the validity and superiority of the proposed model in terms of negative log-likelihood.

The main contributions of this paper are highlighted as follows:

1. Propose a general deep multinomial logit model (GDMNL), in which the alternatives can correspond different network architectures.
2. Design a regularization strategy of parameter tying using prior knowledge.
3. Propose a hybrid model that combines GDMNL and MNL to enhance interpretability while maintaining high performance.

The rest of the paper is organized as follows. Section 2 reviews and discusses related work. Section 3 presents the proposed GDMNL model. Section 4 presents our experimental study demonstrating the validity and superiority of the GDMNL model. Section 5 discusses the interpretability of the proposed model and proposes a hybrid model that integrates the proposed model with the traditional MNL to enhance the model interpretability. Finally, Section 6 summarizes our contributions and outlines future research directions.

2 RELATED WORK

2.1 Accommodation of Nonlinearity by Employing the Box-Cox Transformation

Box-Cox logit model [9] was proposed to accommodate nonlinearity in parameters by employing the Box-Cox (BC) transformation.

Generalized Box-Cox logit model was proposed in [10], which yields better results than Box-Cox logit model. It provides an estimable form of the universal logit and brings logit analysis back into the fold of classical demand analysis where additive separability of utility is not generally credible among close substitutes.

A Box-Cox mixed logit model was proposed in [11], which includes the estimation of the Box-Cox exponents in addition to the parameters of the random coefficients distribution. Probability of choose an alternative is an integral that will be calculated by simulation. The estimation of the model is carried out by maximizing the simulated log-likelihood of a sample of observed individual choices between alternatives.

These models can accommodate to some extent, but cannot accurately characterize the nonlinearity in parameters.

2.2 Representation of Nonlinearity in Choice Utilities

Various studies in discrete choice modelling directly represented nonlinearity in choice utilities based on domain knowledge.

Market shares were calculated in [12] by formulating a mixed logit model based on improved nonlinear utility functions taking different factors into consideration, such as seat grades, fares, running time, passenger income levels and so on.

In order to dynamically adjust the operation plan of overnight D-trains with skylights coordinately, transfer passenger demand was predicted by formulating a mixed logit model based on nonlinear random utility functions for different transport modes in [13].

In [14], advice was given about presenting interaction effects with a focus on different techniques that are well suited to a particular type of interaction effect, which depends on the measurement level of the independent variables that are included in the interaction (i.e., nominal or continuous).

The conventional MNL was extended to consider the interaction among the attributes, and also the decision maker's unique attitudinal character in [15]. The adjustable parameters in the proposed models help to represent a fine range of attitudinal effects. The proposed models hold potential in several applications such as in studying the decision making behavior of a large set of population, consumer behavior, or to predict the response of a population to a new norm or law.

The advanced DCMs such as nested logit, probit, and mixed logit were extended to consider the interaction in [16]. The adjustable parameters dedicated to represent the attributes interaction, and the attitudinal character facilitate to take into account the real world factors in determining the choice probabilities, which otherwise remain unconsidered in the conventional models.

An estimation of nonlinear logit panel data model with fixed effects was discussed in [17]. There are two main estimators for such models: "unconditional maximum likelihood" and "conditional maximum likelihood". Application study was designed to determine the most important factors affecting delayed completion of adjuvant chemotherapy among patients with breast cancer and adjuvant chemotherapy improvement outcomes of patients with breast cancer to determine the relationship between time to chemotherapy and outcome according to breast cancer.

Some studies proposed various domain-specific logit models with nonlinear utility functions based on prospect theory [18]. For instance, reference [19] incorporated, in a generalized nonlinear (in parameters) logit model, alternative functional forms for perceptual conditioning (known as probability weighting) and risk attitude in the utility function to account for travel time variability, and then derived an empirical estimate of the willingness to pay for trip time variability-embedded travel time savings as an alternative to separate estimates of time savings and trip time reliability. For another instance, reference [20] proposed cumulative prospect theory-based passenger behavioral logit models for dynamic pricing and transactive control of shared mobility on demand.

The above studies have two disadvantages. First, the proposed approaches are specific for application domain. Second, due to the complexity of decision-making process in the real world, it is difficult for the proposed approaches to fully represent the interaction effects, threshold effects, saturation effects, etc.

2.3 Generalized Approach to Capture Nonlinearity Using ANNs

Many generalized approaches to capture nonlinearity in DCMs using ANNs have been explored in the past several years [21, 22, 23].

Hybrid neural network was applied in [24] to the classification problems by integrating the variables selected by the statistical models and the outputs of statistical models with those of an ordinary network to create hybrid models that might be more accurate than either of the techniques when considered individually.

A logit-ANN ensemble was proposed in [25] for mode choice modeling. The proposed ensemble uses the technique of ANN to enhance/fine-tune the predictions of the logit model and thus increases its accuracy. The mode choice behavior of the travelers can be interpreted by using the logit model in the first phase of ensemble development. The use of ANN model in the second phase of ensemble is expected to improve the accuracy of the overall mode choice predictions as well as mode choice predictions for each mode choice.

A deep learning-based travel behavior choice model was presented in [26]. The proposed Residual Logit (ResLogit) model formulation seamlessly integrates an ANN architecture into MNL. It extends the systematic utility function to incorporate nonlinear cross effects using a series of residual layers and using skipped connections to handle model identifiability in estimating a large number of parameters.

The RBM (restricted Boltzmann machine) choice model was extended in [4] to a deep choice model to deal with the features of items, which are ignored in the RBM choice model. Deep learning was then used to extract latent features from images and plug those latent features as input to the proposed model.

These hybrid models were designed to contain nonlinear interaction effects, threshold effects, saturation effects, etc., using ANNs. However, all the model structures can only accommodate common parameters, not alternative specific ones. In addition, they all lack the regularization strategies that encode specific kinds of prior knowledge.

3 PROPOSED GDMNL MODEL

In this section, we will follow the problem specification presented in the introduction.

3.1 Multinomial Logit Model

The probability that individual i chooses alternative j is given as the probability that outcome j will have the maximum utility:

$$P_{ij} = Pr(U_{ij} > U_{im}, \forall m \neq j) = Pr(V_{ij} + \varepsilon_{ij} > V_{im} + \varepsilon_{im}, \forall m \neq j). \tag{3}$$

The individual chooses the alternative which maximizes utility:

$$y_i = \underset{j}{\operatorname{argmax}}\, P_{ij}. \tag{4}$$

Multinomial logit model (MNL) [27] is by far the most widely used DCM. It is based on three hypothesis. The first hypothesis is the independence of the errors. The second one is that each ε follows a Gumbel distribution. The last one is that the errors are identically distributed. As the location parameter is not identified for any error term, this hypothesis is essentially a homoscedasticity hypothesis, which means that the scale parameter of the Gumbel distribution is the same for all the alternatives.

The probabilities for a MNL are considerably simpler than other popular DCM models, such as multinomial probit model (MPL), and can be computed in closed form. It has been shown in many sources, such as [28], that for a MNL:

$$P_{ij} = \frac{a_j \exp(V_{ij})}{\sum_{k=1}^{J} a_k \exp(V_{ik})}, \quad j \in \{1, 2, \ldots, J\} \tag{5}$$

where $a_k = 0$, if alternative k is available to individual i; else $a_k = 1$.

MNL is estimated by maximizing the following log likelihood function:

$$\sum_{i=1}^{N} \sum_{j=1}^{J} y_{ij} \log P_{ij} \tag{6}$$

where y_{ij} is the choice indicator of individual i of alternative j (one if alternative j is selected, else zero).

3.2 Multilayer Perceptrons

Multilayer perceptrons (MLPs) [29, 6, 30, 31], also often called deep feedforward networks, or neural networks, are the quintessential deep learning models. The goal of a feedforward network is to approximate some function f^*. These models are called feedforward because information flows through the function being evaluated from x, through the intermediate computations used to define f^*, and finally to the output y. There are no feedback connections in which outputs of the model are fed back into itself.

Input to the MLP is usually linear transformation (i.e. input × weight + bias), but most of the real world data are nonlinear. So, to make that input nonlinear, nonlinear activation functions are used to add nonlinearity into the network.

MLPs regained the attention of researchers due to the successes of deep learning, and became extreme importance to machine learning practitioners. They form the basis of many important commercial applications. For instance, the convolutional networks (CNNs) used for object recognition from photos are a specialized kind of feedforward network. In this study, we will construct the deep multinomial logit model based on MLPs.

3.3 Model Architecture Design

The architecture of the proposed GDMNL is shown in Figure 1.

The input layer consists of J vectors. j^{th} vector $\boldsymbol{x_j}$ denotes the attributes of alternative j and/or the attributes of individual i. Note that if j is not available to i, $\boldsymbol{x}_j = \vec{0}$.

The unit $H_{1m}^{(j)}$ of the first hidden layer is given by

$$H_{1m}^{(j)} = g(\boldsymbol{w}_m^{(j)(1)\mathrm{T}} \boldsymbol{x}_j + b_m^{(j)(1)}), \quad 1 \le j \le J, 1 \le m \le B_1 \tag{7}$$

where g is a nonlinear activation function and B_1 denotes the number of the units of the first hidden layer.

The unit $H_{lm}^{(j)}$ of the l^{th} $(2 \le l \le n)$ hidden layer is given by

$$H_{lm}^{(j)} = g(\boldsymbol{w}_m^{(j)(l)\mathrm{T}} \boldsymbol{H}_l^{(j)} + b_m^{(j)(l)}), \quad 1 \le j \le J, 1 \le m \le B_l \tag{8}$$

where B_l denotes the number of the units of the l^{th} hidden layer.

The $(n+1)^{\text{th}}$ hidden layer denotes the observed utilities of J alternatives perceived by individual i. The unit V_j of this layer is given by

$$V_j = g(\boldsymbol{w}^{(j)(n)\mathrm{T}} \boldsymbol{H}_{n+l}^{(j)} + b_j), \quad 1 \le j \le J. \tag{9}$$

According to Equations (7), (8) and (9), V_j can be regarded as a nonlinear utility function which relates $\boldsymbol{x}_{ij}$ to the utility of alternative j for individual i.

As in the multinomial logit model, a softmax-like function is applied to normalize the vector $(V_1, V_2, \ldots, V_J)^{\mathrm{T}}$ into a probability distribution $(P_1, P_2, \ldots, P_J)^{\mathrm{T}}$ as the output layer. Thus, the unit P_j is given by

$$P_j = \frac{a_j \exp(V_j)}{\sum_{k=1}^{J} a_k \exp(V_k)}, \quad 1 \le j \le J \tag{10}$$

where $a_k = 0$, if $\boldsymbol{x}_k = \vec{0}$; else $a_k = 1$.

Note that if g is a linear function $y = x$, the proposed model is reduced to the traditional MNL.

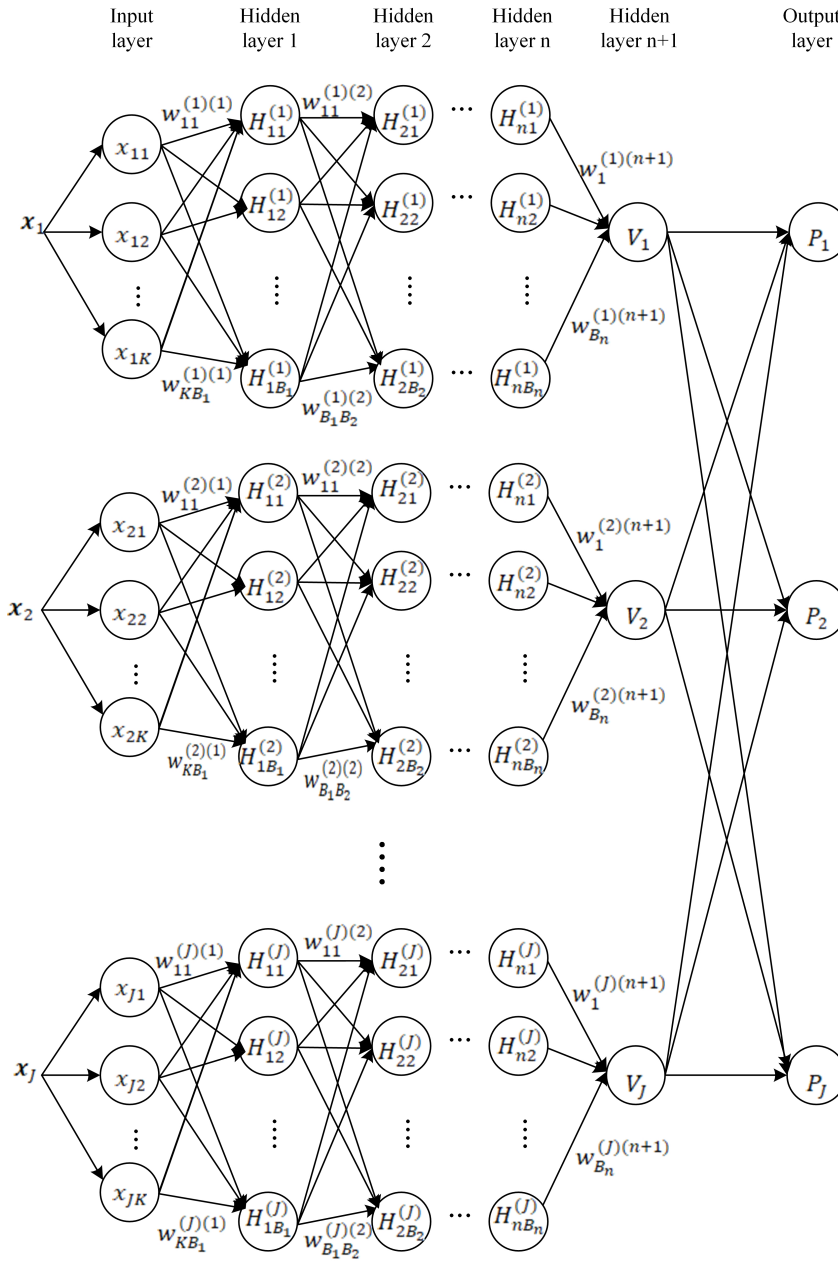

Figure 1. The architecture of the GDMNL model

3.4 Learning Algorithm

The proposed GDMNL is trained using maximum likelihood. This means that the cost function is simply the negative log-likelihood, equivalently described as the cross-entropy between the training data and the model output distribution $(P_1, P_2, \ldots, P_J)^{\mathrm{T}}$.

We use a famous stochastic gradient descent method, Adam optimization [32] as the learning algorithm, which is based on adaptive estimation of first-order and second-order moments. The Adam method is computationally efficient, has little memory requirement, invariant to diagonal rescaling of gradients, and it is well suited for problems that are large in terms of data/parameters.

3.5 Model Regularization

A central problem in deep learning is how to make an algorithm perform well not just on the training data, but also on new inputs. Many strategies to solve this problem used in deep learning are explicitly designed to reduce the test error, which are known collectively as regularization.

Most common regularization strategy may be putting extra constraints on a deep learning model, such as adding restrictions on the parameter values. If chosen carefully, these extra constraints can lead to improved performance. These constraints are usually designed to express a generic preference for a simpler model class in order to promote generalization. However, sometimes we may need other ways to express our prior knowledge about suitable values of the model parameters, such as some dependencies between them. A common type of dependency is that certain parameters should be close to one another, which is called parameter tying.

In linear utility functions, the generic parameters are constrained to be same and the alternative-specific parameters should not be far from each other across the alternatives. Similar to the constraint of the parameters in linear utility functions, in the proposed deep model, the corresponding parameters associated to different alternatives should be constrained to be close to one another. Hence, we use an L^2 penalty as the constraint which is given by

$$||\boldsymbol{\omega}^i - \boldsymbol{\omega}^j||_2^2 \leq d, \quad d \geq 0 \tag{11}$$

where $\boldsymbol{\omega}^{(i)}$, $\boldsymbol{\omega}^{(j)}$ denote the vectors formed by the parameters associated to $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively.

When $d = 0$, $\boldsymbol{\omega}^{(i)} = \boldsymbol{\omega}^{(j)}$, and this regularization strategy is referred to as parameter sharing. This means that all the attributes are common across all of the alternatives. A significant advantage of parameter sharing over parameter tying is that only a subset of the parameters (the unique set) need to be stored in memory.

4 EMPIRICAL STUDY

4.1 Application Domain and Dataset

Public transport has a significant role in urban transport systems. It is paramount to analyze the route choice behavior of transit passengers to capture the spatiotemporal distribution pattern of passenger flow in a transit network [33]. Also, maintaining service reliability would be improved by knowing flow assignment patterns in a transit network [34]. Passengers make their route choice decisions by considering several factors such as in-vehicle time, the number of transfers, crowdedness, etc. Therefore, the proposed deep multinomial logit model is promising for modeling the route choice behavior of transit passengers.

In this study, the main focus is on those journeys with a commuting purpose which start within a weekday morning peak (7–9 am) and night peak (4:30–6 pm). Using the smart card fare payment system of a city in China, all essential data such as date, time, boarding location, bus number, route number, and direction of each transaction is provided. Each commuter who uses smart card has a unique ID that is available in our data.

Based on the original data gathered in one year, we estimated the origin-destination (OD) pair for each card ID, and then generated the corresponding route choice set. The elements selected less than five times in a route choice set were removed as the corresponding journeys are very likely not for commuting purpose.

We extracted three datasets about Dezhou, Anqing and Jinan. These three datasets have 1 252, 2 730 and 6 685 records, respectively.

We quantified the relevant attributes that have potential influence on the route choice decision. A summary of all variables and their descriptions is given in Table 1.

Variable	Description
Time	Travel time in bus (minutes)
Distance	Distance traveled in bus (kilometers)
Tran-count	Number of the transfers
Tran-distance	Sum of all the transfer distances (meters)
Fare	Sum of the bus fares (Chinese yuan)
C-type	Card type (1: Common card; 2: Student card)

Table 1. Summary of variables

As a preprocessing step, we normalized the data to ensure commensurability in the attribute values.

4.2 Experimental Setup

We conduct an experiment with TensorFlow, the second generation of Google artificial intelligence learning system, and a desktop computer with Intel Core i5 CPU (2.90 GHz) with 16.0 GB memory.

We used the negative log-likelihood as the evaluation metric, and the 10-fold cross-validation as the technique to evaluate the models. In addition, we averaged the results of 5 evaluations for each model.

The mini-batch size of stochastic gradient descent was set to 20. The learning rate, exponential decay rate for the 1^{st} moment estimates, exponential decay rate for the 2^{nd} moment estimates, and epsilon hat of the Adam method were set to 0.001, 0.9, 0.999, and 1e−7, respectively.

We used the rectified linear unit (ReLU) as the activation function, as a model that uses it is easier to train and often achieves better performance. The number of the units of the first hidden layer was set to 32. The number of the units of the l^{th} $(2 \leq l \leq n)$ hidden layer was set to $2^{(6-l)}$.

We used Glorot initialization as weight initialization scheme, which initialized the weights in the model by drawing them from a distribution with zero mean and a specific variance. In addition, we initialized the biases with zeros.

4.3 Parameter Sharing

In this domain, all the attributes are common across all of the alternatives which share the same linear utility functions. Therefore, we designed a parameter sharing strategy, named PS-1, as model regularization strategy. PS-1 is described as follows:

$$\boldsymbol{w}_{pq}^{(i)(s)} = \boldsymbol{w}_{pq}^{(j)(s)}, \boldsymbol{w}_{r}^{(i)(n+1)} = \boldsymbol{w}_{r}^{(j)(n+1)}, \quad 1 \leq i, j \leq J, 1 \leq s \leq n. \tag{12}$$

To verify the superiority of PS-1, we proposed another parameter sharing strategy, named PS-2, which leads to higher model capacity and probability of overfitting. Assuming $\boldsymbol{x}_u$, $\boldsymbol{x}_v$ correspond m^{th} alternatives of any two individuals respectively, PS-2 is described as follows:

$$\boldsymbol{w}_{pq}^{(u)(s)} = \boldsymbol{w}_{pq}^{(v)(s)}, \boldsymbol{w}_{r}^{(u)(n+1)} = \boldsymbol{w}_{r}^{(v)(n+1)}, \quad 1 \leq u, v \leq J, 1 \leq s \leq n. \tag{13}$$

4.4 Experimental Results

Table 2 shows the experimental results on the three datasets. The table entries present the negative log-likelihood of the traditional MNL, the DDCM model proposed in [8], the ResLogit model proposed in [26], and the GDMNL models with different numbers of hidden layers and parameter sharing strategies. For DDCM and ResLogit, we selected their best models on each dataset for comparison. The best model for each dataset is highlighted.

In the following, we abbreviate GDMNL models with PS-i $(i = 1, 2)$, and PS-i and j $(j = 1, 2, 3, 4)$ hidden layers as PS-i and PS-i-j, respectively.

From Table 2 we can see that on all three datasets, both PS-1 and PS-2 outperform the traditional MNL and PS-1 outperform DDCM and ResLogit. This strongly suggests the validity and superiority of the proposed GDMNL model. On Dezhou dataset, PS-1-1 gets the best performance among all the models, while on the other

Dataset	MNL	DDCM	ResLogit	GDMNL				
				Number of Hidden Layers				Parameter Sharing Strategy
				1	2	3	4	
Dezhou	0.80288	0.78014	0.77381	**0.77232**	0.77236	0.77299	0.77312	**PS-1**
				0.78256	0.78491	0.78336	0.78261	**PS-2**
Anqing	0.79353	0.76119	0.75657	0.75266	**0.75187**	0.75349	0.75513	**PS-1**
				0.76319	0.76252	0.76408	0.76627	**PS-2**
Jinan	0.79455	0.74886	0.74302	0.74271	**0.73944**	0.74006	0.74108	**PS-1**
				0.75193	0.75011	0.74836	0.74996	**PS-2**

Table 2. The negative log-likelihoods of different models

two datasets, PS-1-2 performs best. The reason for the worse performance of PS-1-3 and PS-1-4 may be that higher model complexity leads to more severe overfitting.

Figure 2 contrasts the running times of MNL, 1 layer GDMNL and 2 layer GDMNL as dataset size increases from 1 000 to 6 000.

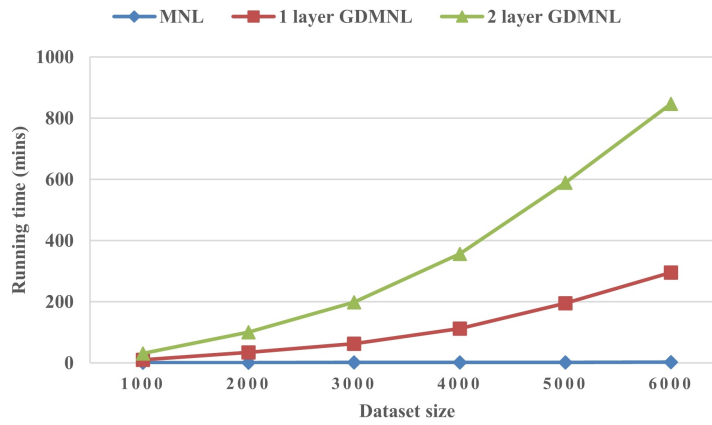

Figure 2. Running times with dataset size

5 DISCUSSION

5.1 The Interpretability of the GDMNL Model

Like many popular machine learning models such as neural networks, the proposed GDMNL models are black-box, and thus are not readily interpretable to the decision maker (DM). Recent researches have proposed approaches to addressing the interpretability of machine learning models [35], such as the gradient-based methods [36, 37], the sensitivity analysis [38, 39], and the mimic models [40, 41]. Among

these methods, a particular important stream of research focuses on learning explicit functions to characterize the nonlinear contributions of variables during the training process. A typical example is the generalized additive model (GAM), which uses a link function to build a connection between the mean of the prediction and a function of the variables [42]. It is good at both dealing with and presenting the nonlinear and non-monotonic relationship between the variables and the prediction [43]. Therefore, GAM is usually more accurate than linear additive models but difficult to be interpreted if the underlying function shapes are extremely complicated, and thus cannot be easily understood by the DM.

MNL and machine learning can provide natural benefits to each other. On one hand, machine learning techniques are capable of handling high-dimensional and nonlinear data because of the high complexity of the model. They can help MNL approaches relax conventional assumptions and improve model performance. On the other hand, the preference disaggregation approaches of MNL use a global value function (usually in an additive form with a predefined shape) to reveal the rationale of DM's judgment. Thus, MNL approaches provide convincing evidence to assist comprehending the decision-making. They can help enhance the interpretability of "black-box" machine learning models.

In the light of the above analysis, we propose a hybrid model that combines MNL and GDMNL to achieve good performance while capturing the explicit relationships between individual attributes and the prediction. The MNL uses marginal value functions to approximate the explicit relationship between the outcome and individual attributes whereas the GDMNL is used to capture the implicit high-order correlations between attributes in the model.

The architecture of the hybrid GDMNL is shown in Figure 3, which has two components. The one in dashed box is the linear component. Note that constant vector $\boldsymbol{\beta}^{(j)}$ corresponds to $\boldsymbol{\beta}^{(j)}$ in Equation (2), which is estimated using the traditional approaches of MNL. The other one outside dashed box is the nonlinear component.

The learning algorithm and model regularization of the hybrid GDMNL can be the same as that of GDMNL. The linear and nonlinear components of the model are jointly trained. Note that this joint training process is different from ensemble learning [44], in which multiple classifiers are trained individually and their predictions are simply combined after every model is optimized separately. For example, an ensemble learning approach could have a linear logistic regression model and an MLP model to make predictions for the same dataset separately, and then integrate the prediction results of the two models. The joint training process indicates that the linear and nonlinear components are connected.

5.2 Flexible Inclusion of Attributes in the Nonlinear Component

In practice, human decision-making usually focuses on a small number of well-chosen attributes/criteria [45, 46]. However, there could exist other minor attributes that do not directly contribute to the prediction but could affect the prediction through non-

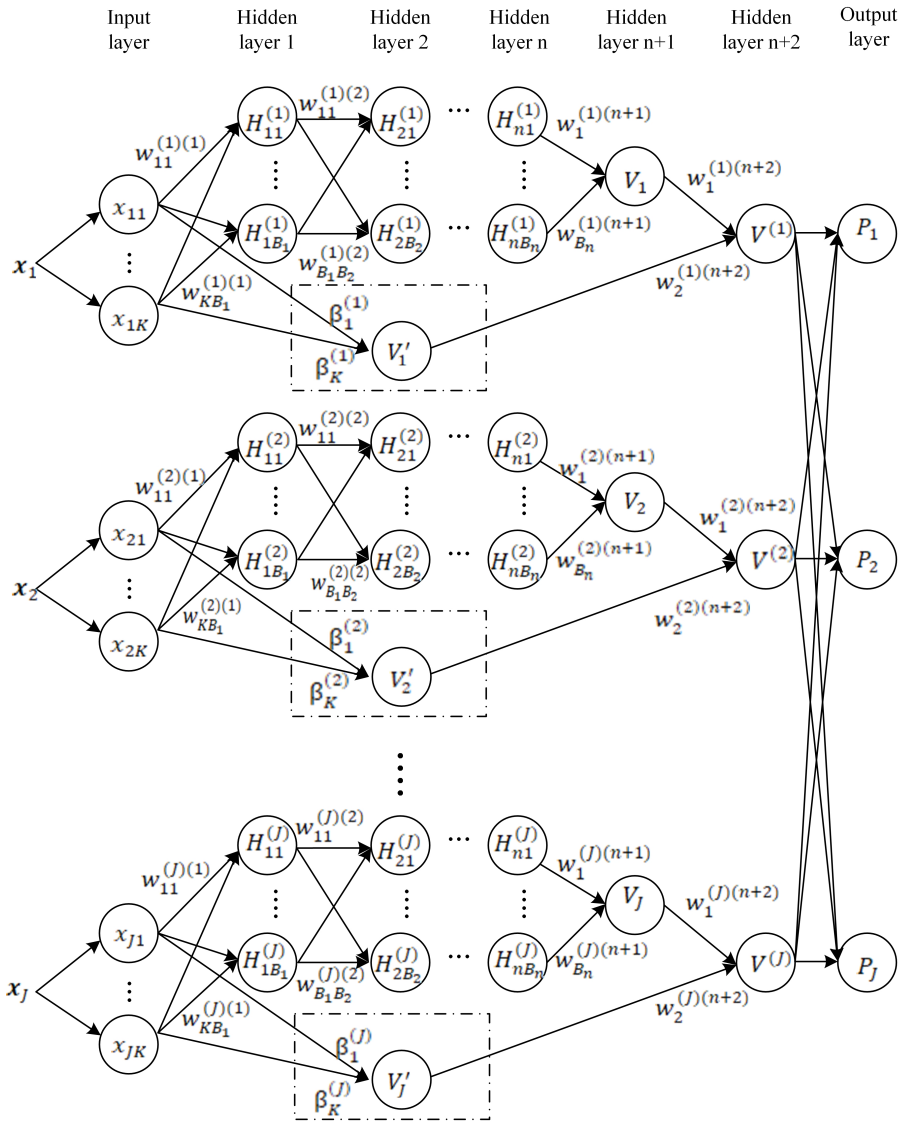

Figure 3. The architecture of the hybrid GDMNL model

traceable complex interactions with other attributes (for example, the interaction between the nonlinear transformation of an attribute and the nonlinear transformation of five other attributes). These minor attributes can be incorporated by the nonlinear component of the hybrid GDMNL model to improve model performance.

6 CONCLUSIONS AND FUTURE WORK

In this paper, a general deep multinomial logit model is proposed to adequately represent the complex considerations regarding various aspects of the alternatives and individual. Compared to the traditional MNL and two latest similar models, the proposed model gets better performance in terms of negative log-likelihood.

Our work opens up avenues for further research. While we have conducted a preliminary experiment using a transit route choice dataset, more experiments with datasets drawn from various domains can be conducted to validate the generalizability of our findings. In addition, the deep neural networks can be applied to extend the other sophisticated DCMs, such as multinomial probit model, nested logit model, etc., to get better performance.

REFERENCES

[1] Thurstone, L. L.: A Law of Comparative Judgement. Psychological Review, Vol. 34, 1927, No. 4, pp. 273–286, doi: 10.1037/0033-295X.101.2.266.

[2] Mottini, A.—Acuna-Agost, R.: Deep Choice Model Using Pointer Networks for Airline Itinerary Prediction. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17), 2017, pp. 1575–1583, doi: 10.1145/3097983.3098005.

[3] Nam, D.—Kim, H.—Cho, J.—Jayakrishnan, R.: A Model Based on Deep Learning for Predicting Travel Mode Choice. Proceedings of the 96th Annual Meeting Transportation Research Board, 2017, pp. 8–12.

[4] Otsuka, M.—Osogami, T.: A Deep Choice Model. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30, 2016, No. 1, pp. 850–856, doi: 10.1609/aaai.v30i1.10059.

[5] Cybenko, G.: Approximation by Superpositions of a Sigmoidal Function. Mathematics of Control, Signals and Systems, Vol. 2, 1989, No. 4, pp. 303–314, doi: 10.1007/BF02551274.

[6] Hornik, K.—Stinchcombe, M.—White, H.: Multilayer Feedforward Networks Are Universal Approximators. Neural Networks, Vol. 2, 1989, No. 5, pp. 359–366, doi: 10.1016/0893-6080(89)90020-8.

[7] Hruschka, H.—Fettes, W.—Probst, M.: Analyzing Purchase Data by a Neural Net Extension of the Multinomial Logit Model. In: Dorffner, G., Bischof, H., Hornik, K. (Eds.): Artificial Neural Networks – ICANN 2001. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 2130, 2001, pp. 790–795, doi: 10.1007/3-540-44668-0_110.

[8] Ma, Y.—Zhang, Z.—Ihler, A.: A Deep Choice Model for Hiring Outcome Prediction in Online Labor Markets. International Journal of Computers Communications and Control, Vol. 15, 2020, No. 2, Art. No. 3760, doi: 10.15837/ijccc.2020.2.3760.

[9] Gaudry, M. J. I.—Wills, M. J.: Estimating the Functional Form of Travel Demand Models. Transportation Research, Vol. 12, 1978, No. 4, pp. 257–289, doi: 10.1016/0041-1647(78)90068-0.

[10] Gaudry, M.—Le Leyzour, A.: Improving a Fragile Linear Logit Model Specified for High Speed Rail Demand Analysis in the Quebec-Windsor Corridor of Canada. Technical Report No. 9413, Centre Interuniversitaire de Recherche en Économie Quantitative, CIREQ, 1994, https://ideas.repec.org/p/mtl/montec/9413.html.

[11] Orro, A.—Novales, M.—Benitez, F. G.: Box-Cox Mixed Logit Model for Travel Behavior Analysis. AIP Conference Proceedings, Vol. 1281, 2010, No. 1, pp. 679–682, doi: 10.1063/1.3498568.

[12] Han, B.—Ren, S.—Bao, J.: Mixed Logit Model Based on Improved Nonlinear Utility Functions: A Market Shares Solution Method of Different Railway Traffic Modes. Sustainability, Vol. 12, 2020, No. 4, Art. No. 1406, doi: 10.3390/su12041406.

[13] Han, B.—Ren, S.: Mixed Logit Model Based on Nonlinear Random Utility Functions: A Transfer Passenger Demand Prediction Method on Overnight D-Trains. Soft Computing, Vol. 26, 2022, No. 7, pp. 3411–3434, doi: 10.1007/s00500-021-06621-4.

[14] Mize, T. D.: Best Practices for Estimating, Interpreting, and Presenting Nonlinear Interaction Effects. Sociological Science, Vol. 6, 2019, pp. 81–117, doi: 10.15195/v6.a4.

[15] Aggarwal, M.: Logit Choice Models for Interactive Attributes. Information Sciences, Vol. 507, 2020, pp. 298–312, doi: 10.1016/j.ins.2019.08.013.

[16] Aggarwal, M.: Probit and Nested Logit Models Based on Fuzzy Measure. Iranian Journal of Fuzzy Systems, Vol. 17, 2020, No. 2, pp. 169–181, doi: 10.22111/IJFS.2020.5227.

[17] El-Masry, A. M.—Youssef, A. H.—Abonazel, M. R.: Using Logit Panel Data Modeling to Study Important Factors Affecting Delayed Completion of Adjuvant Chemotherapy for Breast Cancer Patients. Communications in Mathematical Biology and Neuroscience, Vol. 2021, 2021, Art. No. 48, doi: 10.28919/cmbn/5410.

[18] Kahneman, D.—Tversky, A.: Prospect Theory: An Analysis of Decision Under Risk. Econometrica, Vol. 47, 1979, No. 2, pp. 263–292, doi: 10.2307/1914185.

[19] Hensher, D. A.—Greene, W. H.—Li, Z.: Embedding Risk Attitude and Decision Weights in Non-Linear Logit to Accommodate Time Variability in the Value of Expected Travel Time Savings. Transportation Research Part B: Methodological, Vol. 45, 2011, No. 7, pp. 954–972, doi: 10.1016/j.trb.2011.05.023.

[20] Jagadeesan Nair, V.: Estimation of Cumulative Prospect Theory-Based Passenger Behavioral Models for Dynamic Pricing & Transactive Control of Shared Mobility on Demand. Ph.D. Thesis. Massachusetts Institute of Technology, 2021.

[21] Borysov, S. S.—Rich, J.—Pereira, F. C.: How to Generate Micro-Agents? A Deep Generative Modeling Approach to Population Synthesis. Transportation Research Part C: Emerging Technologies, Vol. 106, 2019, pp. 73–97, doi: 10.1016/j.trc.2019.07.006.

[22] Wong, M.—Farooq, B.: A Bi-Partite Generative Model Framework for Analyzing and Simulating Large Scale Multiple Discrete-Continuous Travel Behaviour Data. Transportation Research Part C: Emerging Technologies, Vol. 110, 2020, pp. 247–268, doi: 10.1016/j.trc.2019.11.022.

[23] Badu-Marfo, G.—Farooq, B.—Patterson, Z.: Composite Travel Generative Adversarial Networks for Tabular and Sequential Population Synthesis. IEEE Transactions on Intelligent Transportation Systems, Vol. 23, 2022, No. 10, pp. 17976–17985, doi: 10.1109/TITS.2022.3168232.

[24] Yim, J.—Mitchell, H.: A Comparison of Corporate Distress Prediction Models in Brazil: Hybrid Neural Networks, Logit Models and Discriminant Analysis. Nova Economia, Vol. 15, 2005, No. 1, pp. 73–93.

[25] Gazder, U.—Ratrout, N. T.: A New Logit-Artificial Neural Network Ensemble for Mode Choice Modeling: A Case Study for Border Transport. Journal of Advanced Transportation, Vol. 49, 2015, No. 8, pp. 855–866, doi: 10.1002/atr.1306.

[26] Wong, M.—Farooq, B.: ResLogit: A Residual Neural Network Logit Model for Data-Driven Choice Modelling. Transportation Research Part C: Emerging Technologies, Vol. 126, 2021, Art. No. 103050, doi: 10.1016/j.trc.2021.103050.

[27] McFadden, D.: The Measurement of Urban Travel Demand. Journal of Public Economics, Vol. 3, 1974, No. 4, pp. 303–328, doi: 10.1016/0047-2727(74)90003-6.

[28] Train, K. E.: Discrete Choice Methods with Simulation. 2nd Edition. Cambridge University Press, 2009, doi: 10.1017/CBO9780511805271.

[29] Rumelhart, D. E.—Hinton, G. E.—Williams, R. J.: Learning Representations by Back-Propagating Errors. Nature, Vol. 323, 1986, No. 6088, pp. 533–536, doi: 10.1038/323533a0.

[30] Lippmann, R. P.: Pattern Classification Using Neural Networks. IEEE Communications Magazine, Vol. 27, 1989, No. 11, pp. 47–50, doi: 10.1109/35.41401.

[31] Siegelmann, H. T.: Neural Networks and Analog Computation: Beyond the Turing Limit. Birkhäuser Boston, MA, 1999, doi: 10.1007/978-1-4612-0707-8.

[32] Kingma, D. P.—Ba, J.: Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference for Learning Representations (ICLR 2015), 2015, doi: 10.48550/arXiv.1412.6980.

[33] Tang, J.—Wang, Y.—Hao, W.—Liu, F.—Huang, H.—Wang, Y.: A Mixed Path Size Logit-Based Taxi Customer-Search Model Considering Spatio-Temporal Factors in Route Choice. IEEE Transactions on Intelligent Transportation Systems, Vol. 21, 2020, No. 4, pp. 1347–1358, doi: 10.1109/TITS.2019.2905579.

[34] Desaulniers, G.—Hickman, M. D.: Chapter 2 Public Transit. Handbooks in Operations Research and Management Science, Vol. 14, 2007, pp. 69–127, doi: 10.1016/S0927-0507(06)14002-5.

[35] Murdoch, W. J.—Singh, C.—Kumbier, K.—Abbasi-Asl, R.—Yu, B.: Definitions, Methods, and Applications in Interpretable Machine Learning. Proceedings of the National Academy of Sciences (PNAS), Vol. 116, 2019, No. 44, pp. 22071–22080, doi: 10.1073/pnas.1900654116.

[36] Sundararajan, M.—Taly, A.—Yan, Q.: Axiomatic Attribution for Deep Networks. In: Precup, D., Teh, Y. W. (Eds.): Proceedings of the 34th International Con-

ference on Machine Learning. Proceedings of Machine Learning Research (PMLR), Vol. 70, 2017, pp. 3319–3328, doi: 10.48550/arXiv.1703.01365.

[37] Shrikumar, A.—Greenside, P.—Kundaje, A.: Learning Important Features Through Propagating Activation Differences. In: Precup, D., Teh, Y. W. (Eds.): Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research (PMLR), Vol. 70, 2017, pp. 3145–3153, doi: 10.48550/arXiv.1704.02685.

[38] Ribeiro, M. T.—Singh, S.—Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[39] Lundberg, S. M.—Lee, S. I.: A Unified Approach to Interpreting Model Predictions. In: Guyon, I., Von Luxburg, U., S., B., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates, Inc., 2017, pp. 4765–4774, doi: 10.48550/arXiv.1705.07874.

[40] Baesens, B.—Setiono, R.—Mues, C.—Vanthienen, J.: Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation. Management Science, Vol. 49, 2003, No. 3, pp. 312–329, doi: 10.1287/mnsc.49.3.312.12739.

[41] Letham, B.—Rudin, C.—McCormick, T. H.—Madigan, D.: Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. The Annals of Applied Statistics, Vol. 9, 2015, No. 3, pp. 1350–1371, doi: 10.1214/15-AOAS848.

[42] Hastie, T.—Tibshirani, R.: Generalized Additive Models. Statistical Science, Vol. 1, 1986, No. 3, pp. 297–318.

[43] Caruana, R.—Lou, Y.—Gehrke, J.—Koch, P.—Sturm, M.—Elhadad, N.: Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15), 2015, pp. 1721–1730, doi: 10.1145/2783258.2788613.

[44] Cheng, H. T.—Koc, L.—Harmsen, J.—Shaked, T.—Chandra, T. et al.: Wide & Deep Learning for Recommender Systems. Proceedings of the 1st Workshop on Deep Learning for Recommender Systems (DLRS 2016), 2016, pp. 7–10, doi: 10.1145/2988450.2988454.

[45] Keeney, R. L.—Raiffa, H.: Decisions with Multiple Objectives: Preferences and Value Tradeoffs. Cambridge University Press, 1993, doi: 10.1017/CBO9781139174084.

[46] Zhang, J.: Research on the Mechanism of Time-Varying Evolution of Internet Public Opinion and Countermeasures. Social Sciences Press, 2020 (in Chinese).

Peng Su received his B.Eng. degree in computer application technology from the Shandong University of Technology, Jinan, China in 1997, the M.Eng. degree in computer science from the Shandong University, Jinan, China in 2004, and his Ph.D. degree in computer application technology from the Chinese Academy of Sciences, Beijing, China in 2011. He is Professor in the School of Information Science and Engineering at the Qilu Normal University. His research interests mainly focus on data mining and business intelligence. He is Member of IFAC Technical Committees on Economic and Business Systems (TC9.1). He regularly serves in program committees at international conferences such as IEEE ISI.

Yuan Liu graduated from the Jinan Vocational College majoring in Chinese language and literature education. She is Librarian in the School of Literature at the Qilu Normal University. Her research interests mainly focus on cultural computing.

Lingyun Zhao received her B.Eng. degree in computer Science and technology from the Taishan University, Taian, China in 2017, and the M.Eng. degree in computer technology from the Harbin University of Technology, Harbin, China in 2020. She is Lecturer in the School of Information Science and Engineering at the Qilu Normal University. Her research interests mainly focus on natural language processing and machine learning.