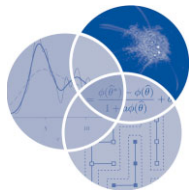
Advanced Review



Selection of multinomial logit models via association rules analysis

Pannapa Changpetch[1]* and Dennis K.J. Lin[2]

In this research, we propose a novel approach for a multinomial logit model selection procedure: specifically, we apply association rules analysis to identifying potential interactions for multinomial logit modeling. Interaction effects are very common in reality, but conventional multinomial logit model selection methods typically ignore them. This is especially true for higher-order interactions. Here, we develop a model selection framework to address this problem. Specifically, we focus on building an optimal multinomial logit model by (1) exploring the combinations of input variables that have a significant impact on response (via association rules analysis); (2) selecting potential (low-order and high-order) interactions; (3) converting these potential interactions into new dummy variables; and (4) performing variable selections among all the input variables and the newly created dummy variables (interactions). Our model selection procedure is the first approach to provide a global search for potential interactions and establish the optimal combination of main effects and interaction effects in the multinomial logit model. In our investigation, we consider both simulated and real-life datasets, thereby confirming the effectiveness and efficiency of this method.

How to cite this article:
WIREs Comput Stat 2013, 5:68–77. doi: 10.1002/wics.1242

Keywords: association rules analysis; interaction effect; model selection; multinomial logit model

INTRODUCTION

The multinomial logit model is one of the most important models for multicategorical responses. This model is used to make predictions about and explain relationships among variables in a wide variety of areas, including business, economics, education, health care, and geography. As it is an enhanced version of logistic regression, multinomial logistic regression shares the problem associated with logistic regression but with more complications involved.

Selecting a multinomial logit model when there are many main effects and interactions involved is difficult. Incorporating interactions into models becomes more challenging as the number of main effects increases, because the number of interactions between the effects grows at an accelerated rate. Because they tend to make the model selection process overly complicated, interactions among variables are usually omitted from the model selection process. In this study, however, we develop a model selection procedure that effectively selects significant interactions for the multinomial logit model and establishes the optimal combination of main effects and potential interactions.

We employ association rules analysis, a methodology that aids in selecting potential interactions among categorical variables from a large pool of possibilities. In market analysis, this method offers a way to understand relationships between products and consumers' purchasing behaviors. Using this method,

*Correspondence to: pchangpetch@bentley.edu

[1]Department of Mathematical Sciences, Bentley University, Waltham, MA, USA

[2]Department of Statistics, Pennsylvania State University, University Park, PA, USA

we are able to narrow the field of possible combinations by selecting interactions that are likely to contribute to the multinomial logit model. As association rules analysis works only with categorical variables, all of the predictors for our proposed method are required to be categorical variables. If the continuous predictors are of interest, we may discretize them into categorical variables before using our method.

Our model selection framework improves on classical model building's ability to consider potential interactions. Association rules analysis streamlines the selection of important rules, converts the selected rules into interaction variables, and determines the optimal multinomial logit model by implementing a subset selection method that considers all the main effects and potential interactions. The key advantages of the proposed framework include its ability to deal with a large number of interactions, to select potential interactions, and to provide alternative setups for interactions. In this method, interactions are incorporated into the multinomial logit model. Further, our model selection procedure has the distinguishing feature of allowing higher-order interactions to be included in the model.

We illustrate an effectiveness of our method by comparing the performances of our method with the classical method using best subset selection criteria. It is shown via both the simulated dataset and the real dataset that the proposed method provides a better explanation and a better fit for multinomial logit modeling than the classical method.

This paper is organized as follows. Next section provides a review of academic literature pertaining to multinomial logistic regression modeling and association rules analysis. Section *The Proposed Method* presents the proposed framework and method in detail. Section *Proposed Method with the Multinomial Logit Model* gives an example of how our framework works in practice using a simulated dataset. Section *Application: Alligator Food Choice Dataset* presents the application of the proposed method to a real dataset. The proposed method is shown to be effective and efficient in selecting the optimal multinomial logit model. Final section offers a discussion and concluding remarks.

MULTINOMIAL LOGISTIC REGRESSION MODEL AND ASSOCIATION RULES ANALYSIS

Used to make predictions and explain relationships among variables, the multinomial logit model has been widely used for many decades for multicategorical responses. On the basis of Ref 1, early work pertaining to the model includes studies by Gurland et al.[2] Cox,[3] Mantel,[4] and Theil.[5] Multinomial logit models have been applied in a range of areas including agriculture,[6,7] finance,[8,9] economics,[10] marketing,[11,12] nursing and heath care,[13–15] and education.[16–20]

In this study, we focus on the multicategory response without orders or nominal data. The multinomial logit model for this type of response is the baseline-category logit model, which is combined with a separate binary logit for each pair of response categories.[21] The baseline-category logit model with all the main effects is illustrated here:

$$\text{Let } \pi_j(x) = P(Y = j|x) \quad \text{at a fixed setting of } x \text{ and } \sum_j^{\pi_j}(x) = 1.$$

Logit models pair each response with a baseline category, often the last or the most common category. The model

$$\log \frac{\pi_j(x)}{\pi_j(x)} = \alpha_j + \beta_j' x, \quad j = 1, \ldots, J-1, \qquad (1)$$

simultaneously describes the effects of x on these $J-1$ logits.

Interactions, especially those between categorical variables, can be addressed in different ways. In the classical method, interactions are omitted from the logit models.[21] In our study, we consider interactions and main effects simultaneously in the multinomial logit model. As the number of interactions increases at an accelerated rate with a higher number of main effects, the subset selection method is extremely inefficient when all the interactions are considered at the same time. Therefore, an efficient method for selecting potential interaction variables is required. We have developed a methodology for implementing association rules analysis for exactly this purpose.

Association rules analysis is a popular data mining technique that was introduced in the early 1990s.[22,23] The basic idea is to determine the rules important to a given dataset that helps in predicting the correct classes of the multinomial model. These rules must satisfy some constraints, for example, minimum support and minimum confidence. A large number of studies make use of association rules analysis in a wide variety of areas: biology,[24–26] business and marketing,[27–29] geography,[30–32] agriculture,[33,34] education,[35–37] photography,[23,38,39] economics,[40,41] and so on.

Association rules analysis is a methodology for exploring relationships among items in the form of rules. Each rule has two parts. The first part pertains to left-hand side item(s) or the condition(s), and the second part to the right-hand side item or the result. The rule is always represented as a statement: if *condition* then *result*.[42] Two measurements are attached to each rule. The first measurement, support (s), is computed by $s = \text{Prob}(condition \text{ and } result)$. The second measurement, confidence (c), is computed by $c = \text{Prob}(result \mid condition)$. Association rules analysis finds all the rules that satisfy both of these key thresholds: minimum support and minimum confidence.[22]

This set of rules can be used for other purposes, including classification. A technique called classification rule mining (CRM), which is a subset of association rules analysis, was developed to find a set of rules in a database that would form an accurate classifier.[23,43] This technique uses an item to represent a pair consisting of a main effect and its corresponding integer value. More specific than association rules analysis, CRM has only one target, which must be specified in advance. In general, the target of CRM is the response, which means the result of the rule (the right-hand side item) can only be the response and its class. Therefore, the condition (the left-hand side item) consists of the explanatory variable and its level. For example, assume that there are k binary factors, $X_1, X_2, \ldots, X_k$, and a binary response Y. Each variable has two levels, one denoted by 0 and the other by 1. Many rules can be generated by CRM. One such rule could be 'if $X_1 = 1$, then $Y = 1$' with $s = \text{P}(X_1 = 1 \text{ and } Y = 1)$ and $c = \text{P}(X_1 = 1 \text{ and } Y = 1)/\text{P}(X_1 = 1)$. Another rule could be 'if $X_1 = 1$, then $Y = 0$' with $s = \text{P}(X_1 = 1 \text{ and } Y = 0)$ and $c = \text{P}(X_1 = 1 \text{ and } Y = 0)/\text{P}(X_1 = 1)$.

In our study, we apply classification rule mining to screen out insignificant or irrelevant interactions and keep only those that are potentially significant to consider in building the multinomial logit model. This methodology is a major aspect of the selection of variables in our process. To our knowledge, there are no studies linking association rules analysis and classification rules mining to multinomial logit modeling. The method proposed by Changpetch and Lin[44] is limited to binary response variable, which cannot be straightforwardly extended to multilevel cases. New considerations need to be further developed.

Some studies have developed other techniques to select and screen variables for the multinomial logit model. Zahid and Tutz[45] use the likelihood-based boosting technique with one step of Fisher scoring in the variable selection, whereas Cherrie[46] developed the five-step technique that involves ANOVA and bootstrapping aggregation in variable screening. However, they do not consider the interactions among variables in the selection process. Lucadamo et al.[47] use principle component analysis to eliminate the problem of multicollinearity data, which can also reduce the number of variables in the multinomial logit model fitting.

Here, we focus on searching for potential interactions for the multinomial logit model. A study by Kim and Kim[48] has a similar interest in interactions: the authors developed a methodology that combines a decision tree with the multinomial logit model. In this two-stage method, the decision tree is used to select the influential interaction effects that act as the explanatory variables for the subsequent multinomial logit model fitting. However, the decision tree has the disadvantage of a hierarchical structure, whereas association rules analysis enables a global search such that more potential interactions can be located and thus considered than in the decision tree structure.

THE PROPOSED METHOD

The proposed framework for building a model to predict a multicategorical response from binary explanatory variables consists of four key steps. As shown in Figure 1, the four steps in our framework are:

Step 1: Generate the rules from association rules analysis.

Step 2: Select the rules based on confidence.

Step 3: Generate the variables for each rule from step 2.

Step 4: From the variables in step 3 and all the main effects, search for the optimal model.

Step 1: Association Rules Analysis

First, we use association rules analysis to create rules from datasets. Specifically, we perform CRM. For each rule, the condition (left-hand side items) represents the combination of explanatory variables and their levels, whereas the result (the right-hand side item) is the response and its class. To perform CRM, we use the CBA program developed by the Department of Information Systems and Computer Sciences at the National University of Singapore[23] (http://www.comp.nus.edu.sg/~dm2/). With this program, we are able to obtain all the active rules using the

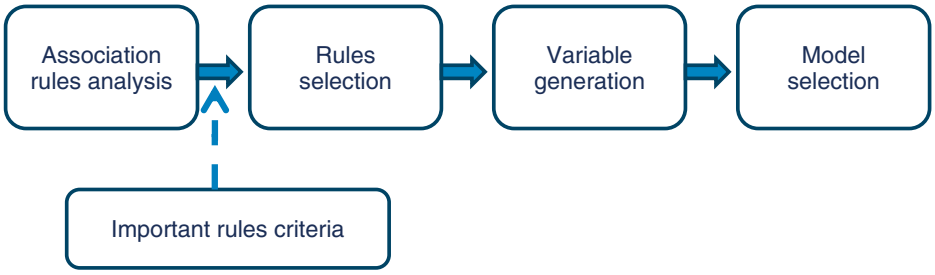


FIGURE 1 | Framework for the proposed model building.

given minimum support and minimum confidence. In general, we recommend the level of 5% for the minimum support based on our experience with both simulated datasets and real datasets. Note that the recommended minimum support in Changpetch and Lin[44] is higher since the number of classes of response for the logistic regression model is less than the number of classes of response for the multinomial logit model. The expected results from this step are the rules in the form 'if X_i's $= x_i$'s, then $Y = y$', where x_i in {0, 1} is the level of variable X_i, and y is the level of response Y in $\{1, \ldots, J\}$. With each rule, the respective support and confidence are attached. All active rules become inputs for the second step.

Step 2: Rule Selection

In this step, we select the rules to convert into interaction variables for the next step. The rules selection criterion used here is confidence. Therefore, rules with the highest confidence are selected from the active rules obtained in the first step. We call the rules selected at this stage as *potential rules*. Note that the number of rules selected at this stage is relatively small compared to the total number of possible interactions for the dataset. In this work, we set the number of potential rules at between 30 and 50 (the same number recommended by Changpetch and Lin[44]). The higher the number of variables, the higher the number of potential rules we select. All the potential rules are inputs for the third step.

Step 3: Variable Generation

In this step, we generate the variables for baseline-category logit model from the potential rules. To convert a rule into an interaction, we create interactions among the main effects on the left-hand side with the same settings that appear in the rule. Suppose that the selected rule has three predictors with the form 'if $X_i = x_i$, $X_j = x_j$, and $X_k = x_k$, then $Y = y$', where x_i is the level of variable X_i, x_j is the level of variable X_j, x_k is the level of variable X_k, and y is the level of response Y. We generate an interaction among X_i, X_j, and X_k by labeling this interaction as 1, if $X_i = x_i$, $X_j = x_j$, and $X_k = x_k$; and as 0 otherwise. This interaction is denoted by $X_i(x_i)X_j(x_j)X_k(x_k)$. For example, for the rule if $X_1 = 0$, $X_2 = 1$, and $X_3 = 1$ then $Y = 0$, we create an interaction between X_1, X_2, and X_3 denoted by $X_1(0)X_2(1)X_3(1)$. We have $X_1(0)X_2(1)X_3(1) = 1$ if $X_1 = 0$, $X_2 = 1$, and $X_3 = 1$, and 0 otherwise. Note that the level of Y does not play any role in generating the variables. The results from this step, that is, the binary variables generated from the potential rules, are inputs for the fourth step.

Step 4: Model Selection

In principle, any model selection criterion can be used. Here, the Akaike information criterion (AIC) is used for illustration.[49] The best subset selection method is performed by testing all the possible combinations of variables and selecting the one that gives the optimal AIC. In other words, the model selected is the one that gives the minimum AIC among all the models. The variables in this step consist of all the interactions generated in Step 3 plus all the main effects. Other model selection methodologies and information criteria can be applied here as well. Note that one association rule can only be converted into only one interaction.

PROPOSED METHOD WITH THE MULTINOMIAL LOGIT MODEL

For illustration and comparison, we modify the MONK's dataset (the first dataset from the MONK's problem) by giving more levels of response (from two classes to three classes) and adjusting the criteria for multiple classes as shown in details next. The MONK's problem is the basis for the first international comparison of learning algorithms. The original MONK's dataset was compiled by Sebastian Thrun of Carnegie Mellon University using propositional formulas based on over six factors.[50]

TABLE 1 | Attributes for the MONK's Dataset

Attribute	Levels	Binary Variables
head_shape	Round, square, octagon	$X_1 = 1$ if head_shape is round and $X_1 = 0$, otherwise $X_2 = 1$ if head_shape is square and $X_2 = 0$, otherwise
body_shape	Round, square, octagon	$X_3 = 1$ if body_shape is round and $X_3 = 0$, otherwise $X_4 = 1$ if body_shape is square and $X_4 = 0$, otherwise
is_smiling	Yes, no	$X_5 = 1$ if smiling and $X_5 = 0$, otherwise
holding	Sword, balloon, flag	$X_6 = 1$ if holding a sword and $X_6 = 0$, otherwise $X_7 = 1$ if holding a balloon and $X_7 = 0$, otherwise
jacket_color	Red, yellow, green, blue	$X_8 = 1$ if jacket color is red and $X_8 = 0$, otherwise $X_9 = 1$ if jacket color is yellow and $X_9 = 0$, otherwise $X_{10} = 1$ if jacket color is green and $X_{10} = 0$, otherwise
has_tie	Yes, no	$X_{11} = 1$ if wearing a tie and $X_{11} = 0$, otherwise

The dataset used here was obtained from the UCI machine learning website (http://archive.ics.uci.edu/ml/datasets.html). It is constructed in such a way that there are interactions among variables.

The objective of this adaptation is to classify 432 robots into three classes (class 1, class 2, and class 3) based on six attributes. The details of all the attributes are shown in Table 1. The true model assigns each robot to one of the three classes, which are based on six attributes. However, given that the true model is unknown, many studies use their own methodologies to predict the class to which the robot belongs. In this study, we use the proposed method to find the multinomial logit model that best fits this dataset and to compare the results to those generated by the classical multinomial logistic regression techniques. Table 1 provides details about the original attributes and their levels. All are multilevel categorical variables. To construct the multinomial logistic regression, we convert the attributes into binary variables (i.e., $X_1, X_2, \ldots, X_{11}$), as shown in Table 1. According to the binary variables and the true model, the robot will belong to class 1 ($Y = 1$), class 2 ($Y = 2$), or class 3 ($Y = 3$), based on the conditions shown in Table 2. This is the underlying (true) model.

In this study, we use the proposed method to find the multinomial logit model that fits this dataset and then to compare the results to those generated by the classical multinomial logistic regression technique.

Implementing the Four-Step Method

Step 1: Use CBA to find the association rules. Note that we use a minimum support value of 5% to generate the active rules.

Step 2: Select the top 30 rules based on confidence criteria. Examples of the selected rules are:
Rule 1: If $X_8 = 1$, then $Y = 3$ with $s = 0.2500$ and $c = 1.000$;

TABLE 2 | Variable Class Results of the True Model for the Simulated Dataset

Class 3 ($Y = 3$)	Class 2 ($Y = 2$)	Class 1 ($Y = 1$)
$X_8 = 1$ (The robot wears a red jacket)	$X_1 = 1$ and $X_3 = 1$ (The robot has a round head shape and a round body shape)	$X_1 = 1$, $X_3 = 0$, and $X_8 = 0$ (The robot (1) has a round head shape, (2) does not have a round body shape, and (3) does not wear a red jacket)
	$X_2 = 1$ and $X_4 = 1$ (The robot has a square head shape and a square body shape)	$X_1 = 0$, $X_3 = 1$, and $X_8 = 0$ (The robot (1) does not have a round head shape, (2) has a round body shape, and (3) does not wear a red jacket)
	$X_1 = 0$, $X_2 = 0$, $X_3 = 0$, and $X_4 = 0$ (The robot has an octagonal head shape and an octagonal body shape)	$X_2 = 1$, $X_4 = 0$, and $X_8 = 0$ (The robot (1) has a square head shape, (2) does not have a square body shape, and (3) does not wear a red jacket)
		$X_2 = 0$, $X_4 = 1$, and $X_8 = 0$ (The robot (1) does not have a square head shape, (2) has a square body shape, and (3) does not wear a red jacket)

Rule 2: If $X_1 = 1$, $X_3 = 0$, and $X_8 = 0$, then $Y = 1$ with $s = 0.1667$ and $c = 1.000$.

Rule 3: If $X_1 = 0$, $X_3 = 1$, and $X_8 = 0$, then $Y = 1$ with $s = 0.1667$ and $c = 1.000$.

Rule 4: If $X_1 = 0$, $X_4 = 0$, and $X_8 = 0$, then $Y = 1$ with $s = 0.2500$ and $c = 0.750$.

Rule 5: If $X_2 = 0$, $X_3 = 0$, and $X_8 = 0$, then $Y = 1$ with $s = 0.2500$ and $c = 0.750$.

Step 3: Convert the 30 rules into 30 variables. For example, rule 2 is converted into the new variable called $X_1(1)X_3(0)X_8(0)$, where

$$X_1(1)X_3(0)X_8(0) = \begin{cases} 1 & \text{if } X_1 = 1, X_3 = 0, X_8 = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Step 4: Combine the 30 variables with the 11 main effects and use the subset selection method to search for the optimal baseline category logit model that yields the optimal AIC.

The optimal baseline category logit model is the model that contains four variables as shown below:

$$\begin{aligned} \ln\left(\frac{p_1}{p_3}\right) &= 15.0488 - 29.3586X_8 - 14.7229X_1(1) \\ &\quad \times X_3(1) - 14.7229X_2(1)X_4(1) \\ &\quad -14.7229X_1(0)X_2(0)X_3(0)X_4(0), \\ \ln\left(\frac{p_2}{p_3}\right) &= 0.5357 - 27.9725X_8 + 13.9892X_1(1) \\ &\quad \times X_3(1) + 13.9892X_2(1)X_4(1) \\ &\quad + 13.9892X_1(0)X_2(0)X_3(0)X_4(0). \end{aligned}$$

According to the estimated coefficients from the four variables, the model can be explained using four conditions according to each variable:

1. If $X_8 = 1$, then $Y = 3$.
2. If $X_1(1)X_3(1) = 1$, meaning $X_1 = 1$ and $X_3 = 1$, then $Y = 2$.
3. If $X_2(1)X_4(1) = 1$, meaning $X_2 = 1$ and $X_4 = 1$, then $Y = 2$.
4. If $X_1(0)X_2(0)X_3(0)X_4(0) = 1$, meaning $X_1 = 0$, $X_2 = 0$, $X_3 = 0$, and $X_4 = 0$, then $Y = 2$.

These four conditions are exactly the same conditions as those shown in Table 2. Therefore, this model can capture the whole logic behind the true model.

We compare the performance of the proposed model with the optimal models (with optimal AICs) from the classical method which is to use the best subset selection considering all the main effects and all the two-way interactions. The optimal classical baseline category logit model is as shown below:

$$\begin{aligned} \ln\left(\frac{p_1}{p_3}\right) &= 4.0056 + 5.4198X_1 + 5.4198X_2 \\ &\quad + 5.4198X_3 + 5.4198X_4 - 23.3298X_8 \\ &\quad - 13.9539X_1X_3 - 13.9539X_2X_4, \\ \ln\left(\frac{p_2}{p_3}\right) &= 11.8191 - 11.1301X_1 - 11.1301X_2 \\ &\quad - 11.1301X_3 - 11.1301X_4 - 19.8371X_8 \\ &\quad + 20.8902X_1X_3 + 20.8902X_2X_4. \end{aligned}$$

According to this model, it captures the significance of the robot's head shape, the robot's body shape, and the color of the jacket (X_1, X_2, X_3, X_4, and X_8). However, to explain the model is much more complicated than to explain our proposed model. Especially, the relationship between head shape and body shape is not as simple as shown in the proposed model. Next, we consider the AIC among these two models. The AIC from our proposed model is 10, while the AIC for the model from the classical method is 16. Our proposed model is better than the model from classical method in terms of AIC as well.

The modified MONK's dataset adequately illustrates our proposed method and demonstrates its ability to capture the interactions between variables. We illustrate the implementation of the proposed method based on a real-life dataset in the next section.

APPLICATION: ALLIGATOR FOOD CHOICE DATASET

In this section, we use an alligator food choice dataset from Ref 21. As mentioned in Agresti's book, this data is courtesy of Clint Moore, from an unpublished study by M.F. Delaney and C.T. Moore. Here, we illustrate an application of our proposed method by selecting the model for this dataset using our framework.

The alligator food choice dataset is used to study factors that influence the primary food choice of alligators. There are five types: fish, invertebrate, reptile, bird, and other. There are three original categorical variables: lake, gender, and alligator size. We converted all the responses and attributes into binary variables, as listed in Table 3.

TABLE 3 | Responses and Attributes for the Alligator Food Choice Dataset

Attribute	Levels	Binary Variables
Primary food choice	Invertebrate, reptile, bird, other, and fish	$Y = 1$ if the class is invertebrate $Y = 2$ if the class is reptile $Y = 3$ if the class is bird $Y = 4$ if the class is other $Y = 5$ if the class is fish
Lake	Hancock, Oklawaha, Trafford, George	$X_1 = 1$ if Hancock and $X_1 = 0$, otherwise $X_2 = 1$ if Oklawaha and $X_2 = 0$, otherwise $X_3 = 1$ if Trafford and $X_3 = 0$, otherwise
Gender	Male and female	$X_4 = 1$ if male and $X_4 = 0$, otherwise
Size (m)	≤ 2.3 and >2.3	$X_5 = 1$ if ≤ 2.3 and $X_5 = 0$, otherwise

Note: We use fish as the baseline category (class 5).

We applied the proposed method to this dataset and obtained the following results:

Step 1: We used CBA to obtain the active rules. Note that we used a minimum support value of 5%.

Step 2: We selected the 30 rules with the highest confidence values from among all the active rules. Examples of the selected rules follow:

Rule 1: If $X_1 = 0$, $X_2 = 0$, $X_3 = 0$, and $X_5 = 0$, then $Y = 5$ with $s = 0.0776$ and $c = 0.7727$.

Rule 2: If $X_1 = 1$, $X_4 = 0$, and $X_5 = 1$, then $Y = 5$ with $s = 0.0731$ and $c = 0.6154$.

Rule 3: If $X_1 = 0$, $X_3 = 0$, $X_4 = 0$, and $X_5 = 1$, then $Y = 1$ with $s = 0.0822$ and $c = 0.6207$.

Rule 4: If $X_2 = 0$, $X_3 = 0$, $X_4 = 1$, and $X_5 = 0$, then $Y = 5$ with $s = 0.0594$ and $c = 0.6842$.

Rule 5: If $X_2 = 1$ and $X_5 = 1$, then $Y = 1$ with $s = 0.0502$ and $c = 0.5500$.

Step 3: We converted the 30 selected rules into variables. For example, rule 1 was converted into the new variable called $X_1(0)X_2(0)X_3(0)X_5(0)$, where

$$X_1(0)X_2(0)X_3(0)X_5(0) = \begin{cases} 1 & \text{if } X_1 = 0, X_2 = 0, X_3 = 0, X_5 = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Step 4: We combined the 30 variables with the main effects and used the subset selection method to search for the optimal baseline category logit model.

The baseline category logit model from the proposed method with five variables is as shown below:

$$\begin{aligned}
\ln\left(\frac{p_1}{p_5}\right) &= -0.9741 + 13.5729X_1(0)X_2(0)X_4(0) \\
&\quad \times X_5(0) - 15.2166X_2(0)X_3(0)X_5(0) \\
&\quad + 0.5383X_1(0)X_3(0)X_5(0) - 0.9731X_2(0)X_3(0) \\
&\quad \times X_4(1) + 1.7619X_1(0)X_5(1), \\
\ln\left(\frac{p_2}{p_5}\right) &= -1.0717 - 15.3623X_1(0)X_2(0)X_4(0) \\
&\quad \times X_5(0) - 0.0559X_2(0)X_3(0)X_5(0) \\
&\quad + 0.3483X_1(0)X_3(0)X_5(0) - 14.0207X_2(0)X_3(0) \\
&\quad \times X_4(1) - 0.1939X_1(0)X_5(1), \\
\ln\left(\frac{p_3}{p_5}\right) &= -1.6908 - 14.5898X_1(0)X_2(0)X_4(0) \\
&\quad \times X_5(0) + 1.1470X_2(0)X_3(0)X_5(0) \\
&\quad - 0.9893X_1(0)X_3(0)X_5(0) - 0.5903X_2(0)X_3(0) \\
&\quad \times X_4(1) - 0.2422X_1(0)X_5(1), \\
\ln\left(\frac{p_4}{p_5}\right) &= -0.8987 - 0.2288X_1(0)X_2(0)X_4(0) \\
&\quad \times X_5(0) + 0.9088X_2(0)X_3(0)X_5(0) \\
&\quad - 1.8607X_1(0)X_3(0)X_5(0) - 0.2183X_2(0)X_3(0) \\
&\quad \times X_4(1) + 0.1319X_1(0)X_5(1),
\end{aligned}$$

The probability of being in each class can be estimated from the model. Furthermore, according to the estimated coefficients from the five variables, the model can be explained using five conditions according to each variable.

1. If *Lake* is neither Hancock nor Oklawaha, *Gender* is female and *Size* >2.3 m, then the probability that the primary food choice is fish is always higher than the probability that the primary food choice is reptile, bird, or other.
2. If *Lake* is neither Oklawaha nor Trafford and *Size* >2.3 m, then the probability that the primary food choice is fish is always higher than the probability that the primary food choice is invertebrate or reptile.
3. If *Lake* is neither Hancock nor Traffort, *Gender* is female and *Size* >2.3 m, then the probability that the primary food choice is fish is always higher than the probability that the primary food choice is bird or other.

4. If *Lake* is neither Oklawaha or Traffort and *Gender* is male, then the probability that the primary food choice is fish is always higher than the probability that the primary food choice is reptile, bird, or other.
5. If *Lake* is not Hancock and *Size* $\leq$ 2.3 m, then the probability that the primary food choice is fish is always higher than the probability that the primary food choice is reptile or bird.

Please note that the AIC from the proposed model is 534.6545, while the AIC from the five-variable model from the classical method is 545.5595. Therefore, our proposed model is better than the model from classical method in terms of AIC.

CONCLUSIONS

Interaction effects are very common in reality, but they are typically ignored in the conventional methods. This is especially true for higher-order interactions. In this research, we develop a model selection procedure for a multinomial logit model that is capable of selecting potential interactions from a large number of candidates and including those selections in the variable selection process. Typically neglected in multinomial logit model building, significant higher-order interactions can be selected and incorporated into the model. Our study confirms that the methodology proposed herein is effective in selecting interactions that improve model fit and facilitate our understanding of datasets. The classical method in the multinomial logit model does not work well in general, as indicated in the simulation (refer to section *Proposed Method with the Multinomial Logit Model*).

In this article, we compare the performances of the proposed method with the classical method which employs the best subset selection criteria. As shown via both the simulated dataset and the real dataset, the results show the effectiveness of the proposed method with the multinomial logit modeling: specifically, the proposed method can provide a better explanation and a better fit for multinomial logit modeling than the classical method does. Note that it is not possible to select any of the five variables shown in the baseline category logit model using any conventional approach.

There is some arbitrariness in the proposed method, for example, the minimum support in step 1 and the number of selected rules in step 2. Additionally, determining the cut-off points has always been an issue for association rules analysis. The minimum thresholds always depend on the individual practitioner's determination. We propose a model-building procedure based on the commonly used threshold, which also works well with our dataset. The simulations also confirm that the recommended threshold works well with general datasets with the involvement of interactions. The goodness of fit could be measured by R^2 or AIC (for example). If necessary, we could employ cross validation (the so-called 'testing data') to verify the goodness of the model.

From our empirical observations, the proposed method performs well when the sample size n is reasonable large (say, 100 or more). This is independent with the number of factors (p). Recent literature, notably in genetic study, has a very large number of factors. We believe that the proposed method shall work well even when $p > n$, as long as n is reasonably large. This is, however, only based on our empirical observation, further investigation is needed for a solid conclusion.

We apply association rules with the multinomial logit modeling with the central purpose of capturing the potential low-order and high-order interactions. Among the data mining techniques, the decision tree is capable of finding interactions among variables. However, the decision tree has the disadvantage of a hierarchical structure. On the other hand, association rules analysis has an important advantage over the decision tree: association rules analysis enables a global search that allows more potential interactions to be considered than is possible with a decision tree structure.

REFERENCES

1. Cramer J. *The Origins and Development of the Logit Model*. Cambridge: Cambridge University Press; 2003.
2. Gurland J, Lee I, Dahm PA. Polychotomous quantal response in biological assay. *Biometrics* 1960, 16:382–398.
3. Cox DR. Some procedures connected with the logistic qualitative response curve. In: David F, ed. *Research Papers in Statistics: Festschrift for J. Neyman*. London: John Wiley & Sons; 1996, 55–71.
4. Mantel N. Models for complex contingency tables and polychotomous dosage response curves. *Biometrics* 1966, 22:83–95.
5. Theil H. A multinomial extension of the linear logit model. *Int Econ Rev* 1969, 10:251–259.

6. Zhou B, Kockelman KM. Neighborhood impacts on land use change: a multinomial logit model of spatial relationships. *Ann Region Sci* 2008, 42:321–340.
7. Caudill SB, Groothuis PA, Whitehead JC. The development and estimation of a latent choice multinomial logit model with application to contingent valuation. *Am J Agric Econ* 2011, 93:983–992.
8. Lawrence EC, Arshadi N. A multinomial logit analysis of problem loan resolution choices in banking. *J Money Credit Bank* 1995, 27:202–216.
9. Rahji M, Fakayode SB. A multinomial logit analysis of agricultural credit rationing by commercial banks in Nigeria. *Int Res J Finance Econ* 2009, 24:90–100.
10. Pryanishnikov I, Zigova K. Multinomial logit models for the Austrian labor market. *Aust J Stat* 2003, 32:267–282.
11. Gonul F, Srinivasan K. Modeling multiple sources of heterogeneity in multinomial logit models: methodological and managerial issues. *Market Sci* 1993, 12:213–229.
12. Basuroy S, Nguyen D. Multinomial logit market share models: equilibrium characteristics and strategic implications. *Manage Sci* 1998, 44: 1396–1408.
13. Yip WC, Wang H, Liu Y. Determinants of patient choice of medical provider: a case study in rural China. *Health Policy Plan* 1998, 13:311–322.
14. Kwak C, Clayton-Matthews A. Multinomial logistic regression. *Nursing Res* 2002, 51:404–410.
15. Mala A, Ravichandran B, Raghavan S, Rajmohan HR. Multinomial logistic regression model to assess the levels in trans, trans-muconic acid and inferential-risk age group among benzene-exposed group. *Ind J Occup Environ Med* 2010, 14:39–41.
16. Park KH, Kerr PM. Determinants of academic performance: a multinomial logit approach. *J Econ Ed* 1990, 21:101–111.
17. Peng C-YJ, Nichols RN. Using multinomial logistic models to predict adolescent behavior risk. *J Mod Appl Stat Meth* 2003, 2.
18. Stratton LS, O'Toole DM, Wetzel JN. A multinomial logit model of college stopout and dropout behavior. Discussion paper: Forschunginstitut zur Zukunft der Arbeit Institute for the Study of Labor, 2005.
19. Grilli L, Rampichini C. A multilevel multinomial logit model for the analysis of graduates' skills. *Stat Meth Appl* 2007, 26: 381–393.
20. Torre J. Multidimensional scoring of abilities: the ordered polytomous response case. *Appl Psychol Measure* 2008, 32: 355–370.
21. Agresti A. *Categorical Data Analysis*. 2nd ed. New Jersey: John Wiley & Sons; 2002.
22. Agrawal R, Srikant S. Fast algorithms for mining association rules. *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1994, 487–499.
23. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (*KDD'98*), New York, 1998.
24. Appice A, Ceci M, Lanza A, Lisi FA, Malerba D. Discovery of spatial association rules in georeferenced census data: a relational mining approach. *Intell Data Anal* 2003, 7:541–566.
25. Becquet C, Blachon S, Jeudy B, Boulicaut JF, Gandrillon O. Strong association rule mining for large gene expression data analysis: a case study on human SAGE data. *Genom Biol* 2002, 3.
26. Calders T, Rigotti C, Boulicaut JF. A survey on condensed representations for frequent sets. *Constraint Based Mining*. Springer-Verlag, LNAI; 2006 64–80.
27. Changchien W, Lu T. Mining association rules procedure to support on-line recommendation by customers and products fragmentation. *Expert Syst Appl* 2001, 20:325–335.
28. Ding Q, Ding Q, Perrizo W. Association rule mining on remotely sensed images using p-trees. *Proceedings of Pacific Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD'02*), 2002.
29. Dong G, Zhang X, Wong L, Li J. CAEP: Classification by aggregating emerging patterns. *Proceedings of the Second International Conference on Discovery Science* (*DS'99*), Japan, 1999.
30. Dopfer K, Potts J. Evolutionary realism: a new ontology for economics. *J Econ Meth* 2004, 11:195–212.
31. Etchells TA, Lisboa PJG. Orthogonal Search-based Rule Extraction (OSRE) method for trained neural networks: a practical and efficient approach. *IEEE Trans Neural Netw* 2006, 17:374–384.
32. Garcia J, Romero C, Ventura S, Calders T. Drawbacks and solutions of applying association rules mining in learning management systems. *Proceedings of the International Workshop on Applying Data Mining in e-learning (ADML'07)*, Crete, Greece, 2007, 13–22.
33. Garcia P, Amandi A, Schiaffino S, Campo M. Evaluating Bayesian networks' precision for detecting students' learning styles. *Comp Ed J* 2007, 49:794–808.
34. Guerreiro J, Trigueiros D. A unified approach to the extraction of rules from artificial neural networks and support vector machines. *Adv Data Mining Appl* 2010, 2:34–42.
35. Karaçali B, Krim H. Fast minimization of structural risk by nearest neighbor rule. *IEEE Trans Neural Netw* 2003, 14:127–137.
36. Kuo RJ, Lin SY, Shih CW. Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan. *Expert Syst Appl* 2007, 33:794–808.
37. Lee AJT, Hong RW, Ko WM, Tsao WK, Lin HH. Mining spatial association rules in image databases. *Inform Sci* 2007, 177:1593–1608.

38. Li J. On optimal rule discovery. *IEEE Trans Knowledge Data Eng* 2006, 18:460–471.
39. Liu KH, Weng MF, Tseng CY, Chuang YY, Chen MS. Association and temporal rule mining for post-processing of semantic concept detection in video. *IEEE TMM* 2008, 10:240–251.
40. Marghny MH, El-Semman IE. Extracting logical classification rules with gene expression programming: Microarray case study. *Proceedings of the International Conference on Artificial Intelligence and Machine Learning (AIML'05)*, Cairo, Egypt, 2005.
41. Matsumoto K. An experimental agricultural data mining system. *Proceedings of the First International Conference on Discovery Science* (*DS'98*), Japan, 1998, 439–440.
42. Berry MJA, Linoff G. *Data Mining Techniques: For Marketing, Sales, and Customer Support*. New York: John Wiley & Sons; 1997.
43. Quinlan JR. *C4.5: Programs for Machine Learning*. CA: Morgan Kaufmann; 1992.
44. Changpetch P, Lin DKJ. Model selection for logistic regression via association rules analysis. *J Comp Stat Simul* 2012. doi:10.1080/00949655.2012.662231.
45. Zahid FM, Tutz G. Multinomial logit models with implicit variable selection. Technical Report No. 89. Institute of Statistics, Ludwig-Maximilians-University Munich, Germany, 2010.
46. Cherrie JA. Variable screening for multinomial logistic regression on very large data sets as applied to direct response modeling, *SAS Global Forum*, Orlando, FL, 2007.
47. Camminatiello I, Lucadamo A. Estimating multinomial logit model with multicollinearity data. *Asian J Math Stat* 2010, 3:93–101.
48. Kim JH, Kim M. Two-stage multinomial logit model. *Expert Syst Appl* 2011, 38:6439–6446.
49. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Cont* 1974, 19:716–723.
50. Thrun SB, Bala J, Bloedorn E, Bratko I, Cestnik B, Cheng J, De Jong K, Dzeroski S, Fahlman SE, Fisher D, et al. The MONK's problems: a performance comparison of different learning algorithms.Technical Report CS-CMU-91-197, Carnegie Mellon University, Pittsburgh, PA, 1991.

