Knowledge-Based Neural Networks and its Application in Discrete Choice Analysis

XIANYU JianChuan
Antai College of Economics & Management, Shanghai Jiao Tong University, China
jianchuanxy@sjtu.edu.cn
GAO LinJie
Antai College of Economics & Management, Shanghai Jiao Tong University, China

JUAN ZhiCai
Antai College of Economics & Management, Shanghai Jiao Tong University, China
zcjuan@sjtu.edu.cn
lj.gao@163.com

Abstract

Travel mode choice forecast has received wide attention in travel behavior analysis. Mode choice is a pattern recognition problem, where different human behavior patterns determine the choices among alternative travel modes. Based on the functional similarity between artificial neural networks (ANN) and decision tree, the method of knowledge-based neural networks (KBNN) combines the rule induction of decision tree (DT) and the accurate approximation of ANN. One appeal of KBNN is the use of pattern association and error correction to represent a problem. This contrasts considerably with the random utility maximization framework in discrete choice modeling. So a network built by this method and a nested logit (NL) model are specified, estimated and comparatively evaluated. The prediction results show that KBNN model demonstrates the highest performance. The analysis of actual investigation data shows that the proposed KBNN model has fast convergence and high precision, which is of great importance for travel mode choice prediction.

Key words: travel mode choice; ANN; DT; KBNN; NL

1. Introduction

Travel mode choice has received the most attention among travel behavior analysis, which has a direct impact on the efficiency of travel demand management and traffic control. Since multinomial logit model was developed in the 1970s [1], discrete choice models based on random utility maximization have become widely used for mode choice analysis [2-4]. And NL model is the most common tool for eliminating the property of independence of irrelevant alternatives. However, the pre-determined mode structure and the linear property of the utility function may not comprehensively represent the complex interrelations among explanatory variable and their effects on the dependent variables [5].

Recently, development in artificial intelligence and machine learning has provided new analysis tools for discrete choice modeling. These algorithms model choice behavior without making parametric functional form assumptions but treat discrete choice modeling as a pattern recognition problem, in which multiple complex patterns formed by the combination and interaction of explanatory variables determine the choice decisions among alternatives. Several studies of applying data mining techniques, especially DT [6-8] and ANN [9-11], for travel behavior analysis demonstrate considerable advantages over traditional NL model by providing insights into pattern characteristics extracted from sample data and from adaptive structures through computational process.

2. Artificial neural networks

ANNs are computational paradigms with initial inspirations rooted in biology and physics [12]. The computational framework for an ANN entails a large number of inter-connected processing elements, called neurons, which for the most part, implement similar and relatively simplistic computational tasks. The true computational power on the other hand derives from a combination of adaptable interconnections, called the weights, a layered topology, and nonlinearities associated with the neuron computations. However, recognized challenges of ANN based approach as classifiers or function approximators include the determination of the appropriate number of hidden layers and number of neurons in each hidden layer, the initialization of connection weights, the identification of an appropriate training algorithm among others and

978-0-7695-3322-3/08 $25.00 © 2008 IEEE
DOI 10.1109/NCM.2008.54

the difficulty in understanding the behavior of trained neural networks [13].

One way is to extract rules that can be provided to the users [14, 15]. The integration of domain knowledge into ANN provides efficient and reliable forms of neural networks learning. KBNN is such a hybrid intelligent system. It was first provided by Towell as a method to build neural networks according to extracted rules from experiences of expert system or from sample data [16, 17].

In this paper, a KBNN is developed for the prediction of commuter travel choice. First, useful attributes and concepts are extracted by DT from the data and linked in a way consistent with initial domain knowledge then the links are weighted properly. The capability and performance on work travel mode choice modeling of the developed KBNN are then compared with the traditional NL model. The data for estimating and evaluating the models are the diary datasets from the Changchun Travel Survey 2003.

3. KBNN method

DT model as a supervised learning algorithm uses recursive partitioning to form a tree structure with IF-THEN rules as splitting criteria, each of which is applied with an explanatory variable. Each branch on different levels of the tree represents a subgroup of observations with homogeneity of different degrees. Homogeneity increases from top to bottom where the bottom leaves contain the cases of the same group while the top branches offer the roughest split. Each branch from the top node to a bottom leaf node can be described as an IF-THEN rule sequence or rule set. There is a corresponding relation between multilayer feed-foreword neural networks and DT (Table 1), which provides favorable supports for rule extraction by DT model for neural networks learning process. A KNBB on an XOR problem (Figure 1) shows the procedure of building up the networks with rules extracted by a DT model.

Table 1 Correspondence between DT and NN

DT	ANN
Internal Nodes	Input Units
Branches	Hidden Units
Distinct Classes	Output Units
Connections between Nodes	Weighted Connections

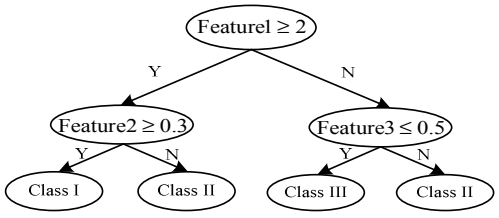

Figure 1 DT of the XOR problem

First, the number of input units is preset according to the rule set of the DT model, where each attribute responds to one unit. So here three input units are selected, corresponding to Feature 1 to 3.

Then the rules from the decision tree are translated from the original "IF A≥a ∧…∧B≤b THEN Class is I" to "(A-a)≥0 ∧…∧(B-b)≤0", in which (A-a) can be considered as a simple hyperplane. Mutually different hyperplanes are obtained from the translated decision rules and the unit number of the first hidden layer is set to the number of hyperplanes. In this XOR problem there are 3 units in the first hidden layer corresponding to the three hyperplanes, Feature1-2⩾0, Feature2-0.3 ⩾0 and Feature3-0.5⩽0.

Next, the unit number of the second hidden layer and the output layer are determined according to the branches of the decision tree and the number of objective pattern respectively. So for the XOR problem there are four units in the second hidden layer and three output units corresponding to the branches and number of classes.

Next the networks initialization is guided by two theorems proposed by Towell [17].

Theorem 1: Mapping conjunctive rules into a neural network using

$$\omega_p = -\omega_n = \omega > 0,\ \theta = -(2P-1)\omega/2$$

creates a network that accurately encodes rules given that the rules have a sufficiently small number of antecedents.

Theorem 2: Mapping disjunctive rules into a neural network using

$$\omega_p = \omega > 0,\ \theta = -\omega/2$$

creates a network that accurately encodes a disjunctive rule given that there are a sufficiently small number of rules. ω_p is the weight on links corresponding to positive dependencies, ω_n is the weight on links corresponding to negative dependencies, P is the number of positive antecedents to the rule, and ω is a constant to be determined by trial and error.

For units of the first hidden layer, the weights on links with the input layer are initialized based on the connection relation of extracted hyperplanes, with one represents a connection and zero represents none. The thresholds are set to the constants of the related hyperplanes. Units from the second hidden layer complete conjunction and each corresponds to one decision rule. They are initialized according to theorem 1. The output units finish disjunction and each corresponds to a recognized category. These units are initialized according to theorem 2. Finally important attributes not utilized by the decision tree are selected

guided by professional knowledge and added to the input layer and the hidden layers are also expanded. These newly added units and links are often randomly initialized [18]. Figure 2 shows the KNBB for the XOR problem

And after finishing the networks establishment the training process can be carried out based on various ANN training algorithms.

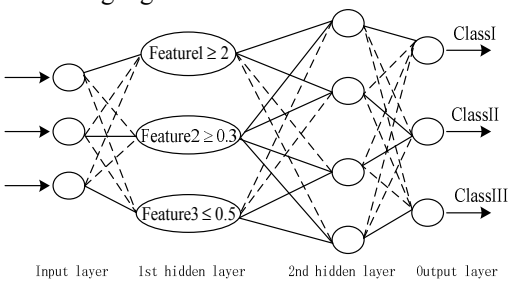

Figure 2 KBNN of the XOR problem

4. KBNN implementation

4.1. Data set and sample composition

The data used for implementing the models come from the travel diary data from the Changchun Travel Survey 2003. This study emphasizes the trip to work mode choice. The alternatives for work travel mode choice used in this study include walk, bicycle, motorcycle, transit and car. The database was divided randomly into two datasets with 60 percent used for the model estimation and another 40 percent used for the subsequent validation test. The actual mode split of training and test sets are shown in Table 2.

Table 2 Summary of mode splits

Travel Mode	Training Dataset	Test Dataset
Walk (M1)	884 (27.85%)	589 (27.85%)
Bicycle (M2)	809 (25.51%)	540 (25.51%)
Motorcycle (M3)	100(3.16%)	67 (3.16%)
Transit (M4)	1168 (36.79%)	778 (36.79%)
Car (M5)	212 (6.69%)	142 (6.69%)
Total	3173 (100%)	2116 (100%)

The socio-demographic characteristics of traveler and his family, attributes related to transportation systems and the trip to work are all important factors of mode choice. So a total of twelve variables were identified for travel mode choice modeling as defined in Table 3.

Table 3 Input variables

Variable	Definition
xb	0:Woman;1:Man
nl	Age of traveler in years, continuous
edu	Education level.1:Primary school and below;2:Secondary school;3:High school;4College and university; Graduate and above
zhy	Type of employment. 1:Unit responsible person;2: professional technical personnel;3:Clerk; 4:Business service;5:Agricultural and water conservancy labors;6:Production, transport equipment operators and related workers
rjzx	No. of bikes per household member, continuous
rjmt	No. of motorcycles per household member, continuous
rjxq	No. of cars per household member, continuouse
tq	No. of commuters in a household
ainc	Monthly income per household member in RMB
minc	Monthly income of the commuter in RMB
gxr	0:Without weekend; 1:With weekend
dis	Home to work travel distance in Km

Travel mode choice modeling predicts individual's or certain group's mode choice decisions and the induced demand for each mode or demand distribution across modes. Two types of prediction rates or match rates are used to evaluate and compare the mode choice modeling performance of different methods. One is individual match rate r_i , which is the ratio of the number of correctly predicted individual observations for one mode N_{pi} over the total number of the actual observations choosing this mode N_a ; the other is aggregate match rate r_a , which reflects the prediction accuracy on the mode aggregate level, defined as the ratio of the number of predicted observations for one mode N_{pa} over the number of the actual observations choosing this mode N_a . They are defined as follows,

$$r_i = N_{pi} / N_a$$

$$r_a = N_{pa} / N_a$$

4.2. Networks structure

The DT model (in Figure 3) applied for KBNN is based on the C4.5 algorithm. The C4.5 algorithm generates a DT in two phases, namely construction and pruning. In the construction of a DT model a training data group is divided at each stage subdivision according to an explanatory variable selected based on the splitting criterion. The division continues until all observations in a subgroup have the same mode choice at the bottom [6-8]. The twelve variables from Table 3 are used as input for the KBNN. Based on the procedures for networks setting up, ten mutually different hyperplanes are extracted, namely rjzx-0.5, rjmt-0, rjmt-1, rjxq-0, xb-1, ainc-1000, minc-2500, dis-1.5, dis-3, dis-7. The networks have 12 units on the second hidden layer and 5 units as output. It is initialized based on the method provided in the previous section.

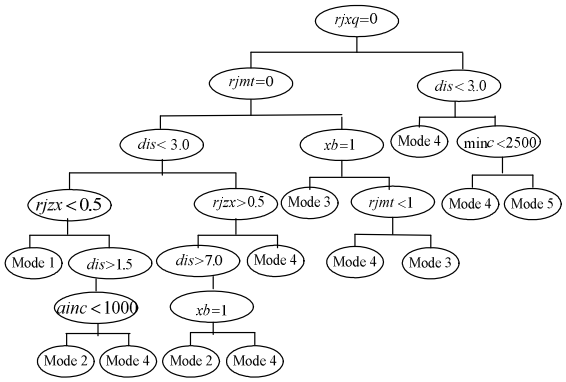

Figure 3 DT model for the KBNN

4.3. Performance comparison with NL model

A NL model was also estimated with the same dataset. The nested structure is shown in Figure 4. And this is the best fit version from a series of hierarchical trees. The model is estimated using maximum likelihood estimation and the results are summarized in Table 4. Walk mode is chosen as the baseline alternative.

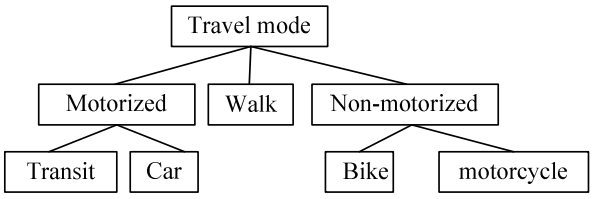

Figure 4 NL model structure

Table 4 Estimation results of NL training model

Alternative specific constant	Estimation	t-statistic
M2	1.63	7.00
M3	1.05	4.16
M4	1.23	2.38
M5	1.45	3.79
Non-motorized	Inclusive value=0.65, t-statistic=4.22	
Factors for M2		
dis	-0.83	-9.816
ainc	-0.0002	-1.742
nl	0.021	1.93
below edu4	0.38	4.15
Factors for M3		
rjzx+rjzd+rjxq	1.26	9.75
nl	0.0381	2.74
Motorized	Inclusive value=0.72, t-statistic=3.74	
Factors for M4		
zhy4	-0.63	-4.41
below edu4	-0.30	-5.39
gxr	-0.19	-4.26
rjzx+rjzd+rjxq	-0.26	-2.08
Factors for M5		
zhy1	0.75	3.21
edu4	-0.84	-2.18
rjzx+rjzd+rjxq	1.15	2.37

Number of cases=3173 $LL(0) = -5106.74$

$LL(c) = -3104.25$, $LL(\hat{\beta}) = -2093.76$

$-2[LL(0) - LL(\hat{\beta})] = 6025.96$

$-2[LL(c) - LL(\hat{\beta})] = 2020.98$ $\rho^2 = 0.59$

The comparison of estimation results between KBNN and NL model are demonstrated in Table 5. It is clear that KBNN model outperforms the NL model on both the training and test dataset; NL has worse transferability in that the match rates decrease considerably from the training dataset to the test dataset; the KBNN model is estimated using the standard back propagation algorithm while the NL model is estimated using maximum likelihood estimation method, so in the MNL model the magnitude and sign of the coefficients (Table 4) indicated the importance and impact of the corresponding variables on mode choice and the value of the t-statistic indicate the confidence level. While the KBNN model shows week interpretability, although the importance of explanatory variables can be obtained from sensitivity analysis.

Table 5 Results of KBNN and NL

Method	Training Accuracy (3173)	Test Accuracy (2116)
KBNN	84.5%	86.8%
NL	80.2%	76.5%

For the KBNN model, the estimation results of training dataset and test dataset are shown in Table 6 and Table 7, in which the actual number of observations of each mode are shown in the first column, while the predicted mode choice are shown in the following five columns. From the estimation results it is shown that KBNN offer higher individual prediction accuracy for motorcycle and car in both training and test datasets. But the model can't predict bike and transit mode very accurately in either the training or the test dataset. Patterns from these two modes are often mismatched mutually. This is very likely due to the fact that walk and bike travel modes share many common attributes such as physical strength depletion, travel distance and time consumption characters and travel cost, which makes the classification very difficult. The same trend also exists in the identification between bike and transit, which also shows similarity in travel time and travel distance consumption attributes.

Table 6 Estimation results of KBNN on training dataset

	Predicted Mode Choice					r_i (%)	r_a(%)
	M1	M2	M3	M4	M5		

	824	812	108	1135	189		
M1 884	757	118	0	0	0	85.6	93.2
M2 809	50	654	2	122	0	80.8	100.4
M3 100	14	40	97	24	0	97.0	108.0
M4 1168	3	0	9	989	4	84.7	97.2
M5 212	0	0	0	0	185	87.3	89.2

Table 7 Estimation results of KBNN on test dataset

	Predicted Mode Choice					r_i(%)	r_a(%)
	M1 574	M2 540	M3 70	M4 793	M5 146		
M1 589	528	37	0	0	0	89.7	97.3
M2 540	32	461	0	82	0	85.4	100.0
M3 67	0	0	66	57	0	98.5	104.0
M4 778	14	42	4	654	18	84.1	101.9
M5 142	0	0	0	0	128	90.2	102.8

5. Discussion and conclusions

With the capacity of learning driven by pattern recognition and effort correction mechanisms, the KBNN method based on the similarity between neural networks and decision tree, which combines the rule extraction and the accurate approximation of these two algorithms was used to construct commuter mode choice model for Shanghai, China. Its capacity for generalization and estimation performance was compared with the classical discrete choice model, namely the nested logit model. The research results demonstrated that the KBNN model has fast convergence and high precision, which is of great importance for travel mode choice prediction. The model offers considerable advantages on predictive power and comparatively appeal in transferability from training dataset to test dataset over NL model. The proposed KNBB model performs with its structure flexibility to adapt to the training data and has the capability to produce explainable if-then rules through hyperplanes in the first hidden layer.

It is found in this research that a behavior rich teaching data is very important to construct and train a KBNN model. And one important property of neural networks models is their capability of processing missing data and noise. So further research should be carried out to find if the KBNN model is able to work well with noise and missing data.

6. Acknowledgements

This research is funded by the national high technology research and development program of China (2007AA11Z203) and National Natural Science Foundation of China (50578094). The authors appreciate the Shanghai City Comprehensive Transportation Planning Institute for providing the data used in this study. The contents of the paper reflect the views of the authors who are responsible for the facts and accuracy of the information presented herein. The contents do not necessarily reflect the official views of the Antai Collge of Economics and Management, Shanghai Jiao Tong University.

References

[1] M.C. Fadden, "Conditional Logit Analysis of Qualitative Choice Behavior", in *Frontiers in Econometrics,* Academic Press, New York, 1973.

[2] L. Cao, "A Model for Travel Mode Switching", New Jersey Institute of Technology: New Jersey, 1998.

[3] H. Sascha, V. N. Rob, H. Serge et al, "Home-Activity Approach to Multi-Modal Travel Choice Modeling", *the 85th Transportation Research Board Annual Meeting*, Washington, D. C., 2006.

[4] J. M. Eric, J. R. Matthew, and A. C. Juan, "A Tour-Based Model of Travel Mode Choice", *the 10th International Conference on Travel Behavior Research*, Lucerne, 2003.

[5] D. A. Hensher, and T. T. Ton, "A Comparison of the Predictive Potential of Artificial Neural Networks and Nested Logit Models for Commuter Mode Choice", *Transportation Research Part E*, Elsevier, 2000, pp. 155-172.

[6] G. Wets, K. Vanhoof, T. Arentze, and H. Timmermans, "Identifying Decision Structures Underlying Activity Patterns: An Exploration of Data Mining Algorithms", *Transportation Research Record*, Transportation Research Board, Washington, D. C., 2000, pp. 1-9.

[7] J.-C. Thill, and A. Wheeler, "Tree Induction of Spatial Choice Behavior", *Transportation Research Record*, Transportation Research Board, Washington, D. C., 2000, pp. 250-258.

[8] T. Yamamoto, R. Kitamura, and J. Fujii, "Driver's Route Choice Behavior: Analysis by Data Mining Algorithms", *Transportation Research Board*, Transportation Research Board, Washington, D. C., 2002, pp. 59-66.

[9] M. Abolfazl, and J. M. Eric, "Nested Logit Models and Artificial Neural Networks for Predicting Household Automobile Choices: Comparison of Performance", *Transportation Research Record*, Transportation Research Board, Washington, D. C., 2002, pp. 92-1000.

[10] D. A. Hensher, and T. T. Ton, "A Comparison of the Predictive Potential of Artificial Neural Networks and Nested Logit Models for Commuter Mode Choice", *Transportation Research Part E*, Elsevier, 2000, pp. 152-172.

[11] T. Arentze, H. Timmermans, "Parametric Action Decision Trees: Incorporating Continuous Attribute Variable

into Rule-Based Models of Discrete Choice", *Transportation Research Part B*, Elsevier, 2007, pp. 772-783.
[12] P. J. Werbos, The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting, Wiley, New York, 1994.
[13] P. Lauret, E. Fock, and R. N. Randrianarivony, "Bayasian Neural Network Approach to Short Time Load Forecasting", *Energy Conversion and Management*, Elsevier, 2008, pp. 1156-1166.
[14] C.-P. Lim, J.-H. Leong, and M.-M. Kuan, "A Hybrid Neural System for Pattern Classification Tasks with Missing Features", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, pp. 648-653
[15] K. Polat, and S. Gunes, "A Novel Hybrid Intelligent Method Based on C4.5 Decision Tree Classifier and One-Against-All Approach for Multi-Class Classification Problems", *Expert Systems with Applications*, Elsevier, 2008.
[16] G. Serpen, D. K. Tekkedil, and M. Orra, "A Knowledge-Based Artificial Neural Network Classifier for Pulmonary Embolism Diagnosis", *Computers in Biology and Medicine*, Elsevier, 2008, pp. 204-220.
[17] G. G. Towell, and J. W. Shavlik, "Knowledge-Based Artificial Neural Networks", *Artificial Intelligence*, Elsevier, 1994, pp. 119-165.
[18] Y. Yao, Y. V. Venkatesh, and C. C. Ko, "A Knowledge-Based Neural Network for Fusing Edge Maps of Multi-Sensor Images", *Information Fusion*, 2001, pp. 121-133.