A factor extraction method using deep learning technique on traffic accident risk

Fernando, Celso: Ehime University
Yoshii, Toshio: Ehime University
Tsubota, Takahiro: Ehime University
Shirayanagi, Hirotoshi: Ehime University

Keywords
Deep Learning, Neural network, Data mining, Maximal itemset, Accident risk

1. Introduction

To send the information about the likelihood to cause an accident to drivers is one of the effective ways to improve traffic safety. Because, the information, which notifies drivers that the likelihood (hereafter, 'accident risk') is high, can call the driver's attention to accident prevention. In addition, drivers who have got the information can choose the route with high safety on Spatio-temporal space. In order to provide the information, the accident risk should be estimated with high accuracy. It is well-known that the accident risk is affected by driving environments, which can be defined by various factors such as road alignment, road environment, weather condition, traffic flow state and etc.

In traffic accident analysis, the statistical method has long been a major analysis approach to understand the impact of these factors on accident risk. However, this approach fails when it deals with complex and highly nonlinear relationships [1]. For including the traffic states as a risk factor, regression analysis has been done by categorizing the traffic states defined by macroscopic observation data such as flow, density, and velocity [2]. In this study, continuous variables are converted into discrete variables, the impact of the factors are investigated by carrying out a multivariate analysis. However, the number of considering factors are limited in the analysis due to the characteristic of the accident risk. As traffic accidents are a rare event, the majority of the data indicates 'NO accident'. Therefore, a huge number of sample data should be collected for carrying out a multivariate analysis with many factors. Also, even if the sufficient number of sample data can be collected, it is difficult to carry out model estimation due to the huge number of samples, especially when interaction terms are tried to be analyzed. For those reasons, Neural Network (hereafter, 'NN') is applied to the analysis instead of statistical analysis. NN models are computational ones defined as a set of processing units, represented by artificial neurons. They can handle nonlinear relationships and interactions among factors, [3] but, they fail in the interpretability of the model.

While the multivariate models deal with each factor as an explanatory variable, the factors are not directly dealt with in NN model. Therefore, it is an important task in the analysis to evaluate the importance of each factor, which is used as part of the input data when using a NN model.

Feature selection has mainly been conducted by random forest. However, the performance of random forest has been questioned, and researchers recently suggest an alternative method based on data mining such as association rule mining [4][5]. The association rule mining has widely used in health science to identify cancerous genes and HIV-1 virus replication [6], [7]. The method has also been shown to be promising in variable selection for accident analysis [8]. The previous applications have focused on individual feature importance. In accident analysis, however, the effect of interaction among individual features should be considered to deal with the complexity and randomness of the accident occurrence.

This study establishes a factor extraction method on the NN model with a deep learning technique on traffic accident risk. By focusing on the maximal itemset, the way how to select the interaction terms, which have significant impacts on the accident risk, is developed. Then, the extraction method is verified by comparing the performances of the NN models, with and without the selected interaction terms as input.

2. Methodology

2.1 Neural network model for accident risk prediction

The neural network consists of four layers. The input layer, two hidden layers, and one output layer. The input data consist of six variables and the model outputs the probability of accident occurrence.

The activation function for the input layer and hidden layers is *relu* and the output layer is *sigmoid.* The loss function is *binary-crossentropy*.
Let assume that we are using neural networks to classify whether the driving environment condition is accident-prone or not. The input of the neural network is defined by the features from the driving environment condition data, and the interaction factors are defined by the maximal itemset extracted from the driving environment condition data.

We develop several neural network models where the inputs are the driving environmental attributes plus one of the maximal itemset. Each maximal itemset has its own model.

2.2 Proposed variable selection method

Association rule mining finds the sets of features (hereafter, 'set') that are most often found together in the database. A set {A, B} is said to be a rule, represented as A⇒B, if {A, B} occurs in the database in at least t times, t is the minimum support or threshold for rules definition [9]. The number of times that a set occur is called support. A set has two parts, the antecedent "A" and the consequent "B".

One thing that challenges the application of the association rule is a large number of rules generated during mining. Therefore, a compact representation of the rules which synopsizes the property of the sets is the most effective way for knowledge discovering [9]. In this study, we apply a maximal itemset.

Assuming the set $\{x_i, y\}$, let x_i be the antecedent and y be consequent. Let the antecedent x_i be $x_1 = \{a, b, c\}$, $sup(x_1 y) \geq t$. Let x_2 be a superset of x_1, meaning that $x_2 \supset x_1$ and $x_2 \neq x_1$, i.e. x_2={a,b,c,d}.

The set $\{x_1, y\}$ is said to be a maximal itemset if all of its supersets have the support lower than the minimum support t, i.e. $sup(x_2 y) < t$.

This study focuses on the maximal itemset extracted from the driving environment condition data (hereafter, 'antecedent' refer to 'driving environment feature interaction' and, 'consequent' refer to'accident occurrence').

Assuming that the maximal itemset is defined based on given minimum support, it is crucial to know how likely is the occurrence of the antecedent, it tells whether the set is a rare, common or random event. Likewise is the dependency between the antecedent and the consequent is an important indicator because it tells how strong is their relationship.

This study proposes the variable selection method considering the degree of uncertainty of the antecedent occurrence and the dependence between the antecedent with the consequent.

2.2.1 The degree of uncertainty of the antecedent

Given a maximal itemset defined by $\{x_i, y\}$.

First. Let the ratio of the support of the antecedent x_i to the whole database be called relative support of the antecedent $rsup(x_i)$, and the $rsup(x_i)$ lies between]0, 1]. If the $rsup(x_i)$ is closed to zero, the x_i is a rare event. The $rsup(x_i)$ can be considered as the probability of occurrence of the antecedent x_i in the database, and based on the information theory, the degree of the uncertainty of x_i is calculated by equation (1).The H_{x_i} takes convex-upward shape as shown in Fig 1; It is maximum when $rsup(x_i) = 0.5$, where the uncertainty of the antecedent x_i is maximum. The uncertainty is also known as entropy

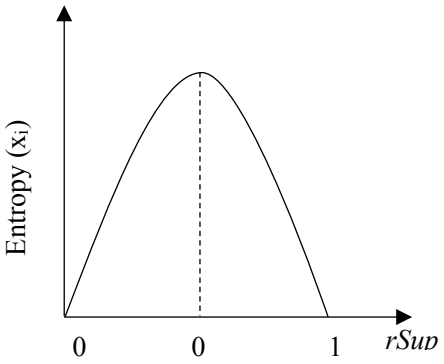

Fig 1. uncertainty of the antecedent

$$H_{x_i} = -rsup(x_i) \sum_{i=1}^{n} \log\left(rsup(x_i)\right) \quad (1)$$

2.2.2 The dependency between the antecedent and the consequent

Assuming that the rule $x_i \Rightarrow y$ consists of two components x_i and y, we can calculate the dependence relationship between this pair. The coefficient of correlation gives the strengths of their linear relationship.

$$|\rho_{x_i y}| = \frac{sup(x_i y) - sup(x_i) * sup(y)}{\sqrt{sup(x_i) * sup(y) * (1 - sup(x_i)) * (1 - sup(y))}} \quad (2)$$

2.2.3 Proposed measure of variable importance

In our proposed method we argue that the importance of the x_i is determined by the product of the coefficient of correlation of $\{x_i, y\}$ with the uncertainty of the occurrence of x_i.

$$I(x_i) = \left|\rho_{x_i y}\right| * H_{x_i} \quad (3)$$

2.3 Evaluation method for the variable selection method

One way to evaluate the performance of the models is by the receiver operating characteristic (ROC) curve. Assuming that the best model is the one that produces the highest ROC, we, therefore, argue that given a set of maximal itemsets, the importance of each set is determined by the respective ROC produce by the model where the set is one of the input.

3. Result

3.1 Data and study site

The data is from an expressway route in Osaka, Japan. The length of the route is 23.2 km. The study includes data from one year (April 2010 – March 2011).

The data includes weather conditions, vertical and horizontal alignment of each 100 meters, pavement material and daily traffic volume.

3.2 Results

From the association rule, we extract five maximal itemsets x_i Table 1.

x_1 = {Straight, Heavy traffic},
x_2 = {Straight, DPSA},
x_3 = {Sunny, Heavy traffic},
x_4 = {Sunny, DPSA} and,
x_5 = {Straight, Sunny}

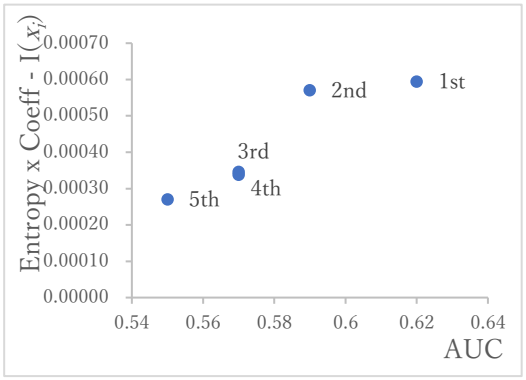

Fig. 2. Importance of the feacture

From fig. 2, we observe that the set which produces the highest ROC-AUC is rank as the most important and the one which produces the lowest ROC-AUC is ranked as the less important.
The fig. 2 shows that there is a linear relationship between our proposed method and the result of the performance of NN model.

Table 1. Comparison of the models

Models	x_i	*sup*(x_i, y)	*sup*(x_i)	AUC	$\rho_{x_1 y}$	*Hx*$_i$	*I*(x_i)	Rank
Basic Model				0.58				
Basic Model	Straight, Heavy traffic	36	425,000	0.59	0.0019	0.29156	0.00057	2nd
	Straight, DPSA	36	492,365	0.55	0.0009	0.29955	0.00027	5th
	Sunny, Heavy traffic	36	419,568	0.62	0.0020	0.29060	0.00059	1st
	Sunny, DPSA	37	492,694	0.57	0.0011	0.29957	0.00034	4rd
	Straight, Sunny	59	1,047,477	0.57	-0.0069	0.04938	0.00035	3th

4. Conclusion and future work

This study establishes a factor extraction method on NN model with deep learning technique on traffic accident risk. By focusing on the maximal itemset, the way how to select the interaction terms, which have significant impacts on the accident risk, is developed. The importance of the set is calculated based on entropy and the coefficient of correlation between the antecedent and the consequent of the association rule. Then, the extraction method is verified by comparing the performances of the NN models, with and without the selected interaction terms as input. As a result, model performance is improved when the selected interaction terms are added to the input data.

References

[1] M. G. Karlaftis and E. I. Vlahogianni, 'Statistical methods versus neural networks in transportation research: Differences, similarities and some insights', *Transp. Res. Part C Emerg. Technol.*, vol. 19, no. 3, pp. 387–399, 2011.

[2] Hyodo, S. and Yoshii, T., 'Analysis of the impact of the traffic states on traffic accident risk', Proceedings of 22st World Congress on Intelligent Transportation Systems (Scientific Paper), ITS-2863, Bordeaux, 2015.

[3] I. N. da Silva, D. H. Spatti, R. A. Flauzino, L. H. B. Liboni, and S. F. dos R. Alves, *Artificial Neural Networks*. Switzerland: Springer, 2017.

[4] Y. Liu, 'Random forest algorithm in big data environment', *Comput. Model. New Technol.*, vol. 18, pp. 147–151, 2014.

[5] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, 'A comparison of random forest variable selection methods for classification prediction modeling', *Expert Syst. Appl.*, vol. 134, pp. 93–101, 2019.

[6] L. Lin, Q. Wang, and A. W. Sadek, 'A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction', *Transp. Res. Part C Emerg. Technol.*, vol. 55, pp. 444–459, 2015.

[7] Y. Koçak, T. Özyer, and R. Alhajj, 'Utilizing maximal frequent itemsets and social network analysis for HIV data analysis', *J. Cheminform.*, vol. 8, no. 1, pp. 1–15, 2016.

[8] K. R. Seeja, 'Feature selection based on closed frequent itemset mining: A case study on SAGE data classification', *Neurocomputing*, vol. 151, no. P3, pp. 1027–1032, 2015.

[9] C. C. Aggarwal and J. Han, *Frequent Pattern Mining*. New York: Springer International Publishing, 2014.