Accident Analysis and Prevention 97 (2016) 87–95

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Rule extraction from an optimized neural network for traffic crash frequency modeling

Qiang Zeng[a,b], Helai Huang[b,*], Xin Pei[c], S.C. Wong[d], Mingyun Gao[e]

[a] School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, Guangdong 510641, PR China
[b] Urban Transport Research Center, School of Traffic and Transportation Engineering, Central South University, Changsha, Hunan 410075, PR China
[c] Department of Automation, Tsinghua University, Beijing, PR China
[d] Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong
[e] Business School of Hunan University, Changsha, Hunan 410082, PR China

ARTICLE INFO

Article history:
Received 29 March 2015
Received in revised form 22 April 2016
Accepted 17 August 2016

Keywords:
Crash frequency
Neural network
Over-fitting
Structure optimization
Rule extraction

ABSTRACT

This study develops a neural network (NN) model to explore the nonlinear relationship between crash frequency and risk factors. To eliminate the possibility of over-fitting and to deal with the black-box characteristic, a network structure optimization algorithm and a rule extraction method are proposed. A case study compares the performance of the trained and modified NN models with that of the traditional negative binomial (NB) model for analyzing crash frequency on road segments in Hong Kong. The results indicate that the optimized NNs have somewhat better fitting and predictive performance than the NB models. Moreover, the smaller training/testing errors in the optimized NNs with pruned input and hidden nodes demonstrate the ability of the structure optimization algorithm to identify the insignificant factors and to improve the model generalization capacity. Furthermore, the rule-set extracted from the optimized NN model can reveal the effect of each explanatory variable on the crash frequency under different conditions, and implies the existence of nonlinear relationship between factors and crash frequency. With the structure optimization algorithm and rule extraction method, the modified NN model has great potential for modeling crash frequency, and may be considered as a good alternative for road safety analysis.

© 2016 Published by Elsevier Ltd.

1. Introduction

In recent decades, numerous models of crash frequency have been proposed to model the relationship between crash frequency at road segments or intersections and risk factors related to traffic and geometrical characteristics of the sites. Most studies of this sort have employed statistical count modeling techniques, since these models provide explicit forms for the random, discrete and non-negative nature of counting crash data and the effects of major contributing factors on crash occurrence. In addition to statistical models, some artificial intelligence (AI) models have been proposed (Chang, 2005; Li et al., 2008). As a common class of AI models, neural network (NN) models have been successfully used in many fields of transportation research, including highway safety analysis (Karlaftis and Vlahogianni, 2011).

In modeling crash frequency, NNs are able to approximate the potential nonlinear and complicated relationship between crash frequency and risk factors. Several studies have demonstrated the better model fitting and predictive performance of NNs over traditional negative binomial (NB) models, in which nonlinear safety effects of risk factors have been identified (Chang, 2005; Xie et al., 2007). The recently-developed random parameters (Anastasopoulos and Mannering, 2009) and Markov switching (Malyshkina et al., 2009) count models indicate that loosening the constraint of fixed parameters could significantly improve their performance on modeling crash frequency, which also partially reflects the existence of nonlinear relationship in crash modeling.

However, NNs have two primary drawbacks that limit their application to traffic safety research, including the so-called "black-box" characteristic and the possible over-fitting problem (Xie et al., 2007). The black-box characteristic has limited NNs' ability to explicitly illustrate the effects of explanatory variables on crash frequency. Even for studies using sensitivity analysis, the impacts on safety of each risk factor cannot be systematically or globally interpreted either. To overcome this problem, a more general approach

* Corresponding author.
E-mail addresses: 641459622@qq.com (Q. Zeng), huanghelai@csu.edu.cn, huanghelai@hotmail.com (H. Huang), peixin@mail.tsinghua.edu.cn (X. Pei), hhecwsc@hku.hk (S.C. Wong), 1198915787@qq.com (M. Gao).

http://dx.doi.org/10.1016/j.aap.2016.08.017
0001-4575/© 2016 Published by Elsevier Ltd.

is to extract the knowledge from the NNs. Using regression analysis, Setiono and Thong (2004) proposed a rule extraction method that generated a group of piecewise linear functions to approximate NNs. This method may be adopted in road safety analysis to clarify the relationship between network output(s) and input risk factors.

The possible over-fitting problem may be caused by the weak generalization ability of models, which was also observed in generalized linear models (Marzban and Witt, 2001). Sample size and model architecture are two factors that may have a profound effect on NNs' generalization performance (Haykin, 2009). Although Bayesian neural network (BNN) has been proposed to reduce the over-fitting phenomenon, it is not suitable for rule extraction (Xie et al., 2007). For a given sample size, optimizing the structure of NN models, that is, adjusting the number of units or neurons in each layer and the connections between different neurons, is a useful method for improving the model's generalization ability. Moreover, this method can identify and prune factors that have no significant effects on the crash frequency. In previous studies (Chang, 2005; Xie et al., 2007), only the number of hidden layer units was locally optimized by using cross-validation, which can neither guarantee the models' generalization performance nor verify the importance of the input variables. Recently, many advanced methods for NN model structure optimization have been proposed to achieve an optimized network that are able to not only represent more generalized relationship between crash frequency and risk factors, but also create a simpler set of extracted rules (Setiono and Thong, 2004).

Therefore, it would be interesting to research on the possibility of employing the emerging NN techniques in better modeling crash frequency. This study aims to develop a generalized NN model for crash frequency analysis, in which only the significant risk factors are retained with estimation of their effects on crash frequency.

2. Literature review

2.1. Statistical models of crash frequency

Statistical models have always been the most popular approach for modeling crash frequency. To handle the possible over-dispersion, multilevel heterogeneities, and spatiotemporal correlation among observations, models ranging from the negative binomial (NB) (Miaou, 1994), Poisson-lognormal (Miaou et al., 2005), and zero-inflation models (Shankar et al., 1997; Huang and Chin, 2010) to the Conway-Maxwell-Poisson (Lord et al., 2008), finite mixture/latent class (Park and Lord, 2009; Park et al., 2010), Markov switching (Malyshkina et al., 2009), random effects or random parameters (Shankar et al., 1998; Anastasopoulos and Mannering, 2009), multilevel (Huang and Abdel-Aty, 2010; Lee et al., 2015; Wang and Huang, 2016), and Bayesian spatial models (Dong et al., 2014, 2015, 2016; Huang et al., 2016; Xu et al., 2014; Xu and Huang, 2015), have been widely investigated. Most of these models are based on a generalized linear function framework and certain assumed distributions of crash data. If these assumptions are violated, the inferences about the effects of the related factors may become biased (Li et al., 2008). Lord and Mannering (2010) and Mannering and Bhat (2014) have presented more detailed descriptions and assessments of these models.

2.2. NN models of crash frequency

Unlike the statistical models, NN models are not limited by data assumptions and have been used to model the potential nonlinear relationship between crash frequency and related factors. Probably because of the aforementioned two limitations, only a few studies have focused on predicting crash frequency using NN models. Chang (2005) compared the use of NB and NN models for crash frequency analysis, and found that the NN model has better predictive performance. Xie et al. (2007) developed a BNN model for analyzing crash frequency and compared the BNN model, NB model, and NN model trained with a back-propagation (BP) algorithm (BPNN). The results showed that both the BNN and BPNN had higher prediction accuracies than the NB model.

2.3. NN structure optimization

Basically, the structure of NN models can be optimized by either constructing or pruning the network. In the constructing method, an NN starts with a small number of hidden layer neurons, and then hidden units are incrementally added during training until the training error cannot be reduced. The most common constructing algorithms include the growing cell structure (GCS) (Fritzke, 1994), constructive back-propagation (CBP) (Lehtokangas, 1999), and adaptively constructing methods (Ma and Khorasani, 2003). Although these constructing algorithms are computationally efficient, they cannot ensure that all of the added units in the hidden layers are properly trained.

For the pruning algorithms, an NN model is firstly created with sufficient hidden layer units. During or after network training, irrelative connections or redundant neurons in the network are removed. Popular pruning algorithms include the optimal brain surgeon (OBS) (Haykin, 2009), subset-based training and pruning (SBTP) (Xu and Ho, 2006), and independent component analysis (ICA) (Nielsen and Hansen, 2008). In contrast to the methods that delete one connection at a time, the NN pruning of the function approximation (N2PFA) algorithm proposed by Setiono and Leow (2000) removes one hidden/input node each time, which could significantly shorten the computational time.

2.4. Rule extraction of NN

A large number of rule extraction methods have been developed to make up the black-box characteristic of NNs (Elalfi et al., 2004; Hruschka and Ebecken, 2006). However, only a couple of methods have been devised to extract rules from NNs used for regression problems. Setiono et al. (2002) proposed extracting piecewise linear function rules from NNs. In that study, the hidden unit transfer function was approximated by either a three-piece or a five-piece linear function, which minimized the approximation errors. To generate a simpler and more accurate rule-set, Setiono and Thong (2004) developed a new three-piece linear function form that had comparable approximation performance with the NN model. This method can be modified to further improve its performance and be adopted in road safety analysis to reveal the relationship between safety performance and input factors.

3. Methodology

The NB model is one of the most widely used statistical models in crash frequency analysis. As in previous research, it is used as a benchmark in this study, and various techniques are used to compare its predictive performance with that of the NN models. In this section, the model architectures of the NB and NN models are specified. Then, the training, structure optimization, and rule extraction algorithms for the NN model are presented.

3.1. Model specification

3.1.1. NB model

The NB model, also known as the Poisson-gamma model, is a modification of the basic Poisson model that can address the com-

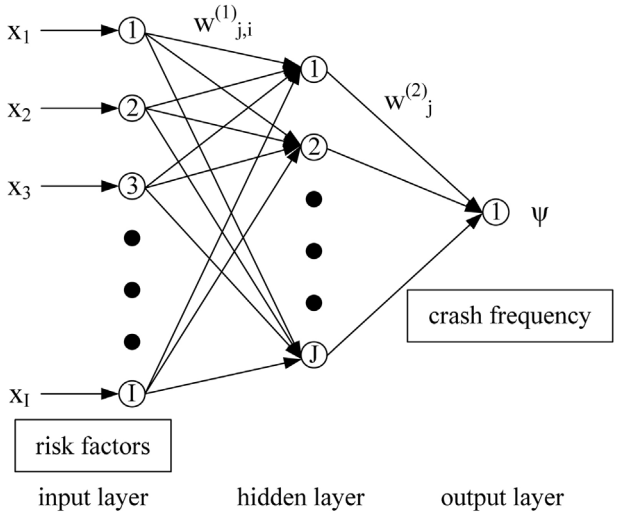

Fig. 1. Developed MLP structure.

mon over-dispersion of crash data. The crash count Y_{it} at site i during period t follows an NB distribution (Washington et al., 2011), that is,

$$P(Y_{it}=y_{it})=\frac{\Gamma(1/\alpha+y_{it})}{\Gamma(1/\alpha)+y_{it}!}\left(\frac{1/\alpha}{1/\alpha+\lambda_{it}}\right)^{1/\alpha}\left(\frac{\lambda_{it}}{1/\alpha+\lambda_{it}}\right)^{y_{it}},\ y_{it}=0,1,2,3,\cdots, \tag{1}$$

where $\Gamma(\cdot)$ is the Gamma function and α is the over-dispersion parameter. The mean of Y_{it}, λ_{it}, is assumed to have a generalized linear relationship with the risk factors,$\mathbf{X}_{it}$, such that

$$\ln\lambda_{it}=\ln e_{it}+\boldsymbol{X}'_{it}\boldsymbol{\beta}, \tag{2}$$

in which e_{it} is the crash exposure and $\boldsymbol{\beta}$ are the coefficients to be estimated.

3.1.2. NN Model

NNs are information processing mechanisms that are inspired by biological nervous systems (Haykin, 2009). Mathematically, an NN is a complex function, which is designed to learn from the collected data. Depending on their learning mechanisms, NNs can generally be divided into two groups, namely, supervised and unsupervised. Supervised NNs are only suited for learning from labeled samples (those with expected outputs, such as crash observations). Classical supervised NNs include multilayer perceptron (MLP) and radial basis function (RBF) network. On the contrary, unsupervised NNs are usually used for learning from unlabeled samples (those without expected outputs). Fuzzy adaptive resonance theory map (ART) network and self-organizing feature map (SOFM or SOM) network are typical examples of unsupervised NNs. The MLP, known as a universal approximator, is the most popular supervised NN for data mining; here it is used to model the underlying nonlinear relationship between crash frequency and related risk factors. Fig. 1 shows the structure of the developed MLP with fully connected neurons.

Consider a dataset containing N_1 continuous attributes and N_2 categorical attributes that may affect the crash frequency. As in many statistical modeling, each categorical attribute $A_n(n=1,2,\cdots,N_2)$ is transformed into m_n-1 binary attribute(s) $a_1^n,\cdots a_j^n,\cdots,a_{m_n-1}^n$, where m_n is the number of possible values for A_n. $a_j^n=1$ if A_n is equal to category j; and $a_j^n=0$ otherwise. Each of the transformed attributes together with the continuous ones is represented by a node $x_i(i=2,\cdots I)$ in the input layer. Besides, an input node, with $x_1=1$, is added. The weights of its connections with hidden neurons are the biases. The number of units I in the input layer is

$$I=N_1+\sum_{n=1}^{N_2}(m_n-1)+1. \tag{3}$$

Although two or more hidden layers are feasible, a single hidden layer is preferred in the MLP, as experimental evidence suggests that NNs with the latter give a similar performance to NNs with the former, and are less likely to be trapped at a local minimum during network training (Villiers and Barnard, 1993). To fit the training data well, the number of neurons in the hidden layer must be sufficiently large. If it is assumed to be J, then the connection weight between hidden node j, $j=1,\cdots,J$ and input node i, $i=1,\cdots,I$ is $w_{j,i}^{(1)}$. The hyperbolic function, $\tanh(\cdot)$, which is an odd sigmoid transfer function, is used for all of the hidden nodes.

In the output layer, the only unit, ψ, represents the expected crash frequency. $w_j^{(2)}$ denotes the weight of the connection between the output node and hidden node j, $j=1,\cdots,J$. A linear function is employed as the transfer function for the output node. As a result,

$$\psi=\sum_{j=1}^{J}w_j^{(2)}\tanh\Big(\sum_{i=1}^{I}w_{j,i}^{(1)}x_i\Big). \tag{4}$$

3.2. Network training

In network training, the conjugate gradient algorithm which possesses better learning performance than the popular BP algorithm is adopted in this study (Haykin, 2009). For training samples $\{\mathbf{x}(m),o(m)|m=1,2,\cdots,M\}$ where $\mathbf{x}(m)$ and $o(m)$ are a vector of risk factors and the corresponding real crash frequency respectively and M is the number of samples, the conjugate gradient updates the connection weight vector, $\mathbf{w}$, as follows:

$$\begin{aligned}\boldsymbol{w}&=(w_1,\cdots,w_{(j-1)I+i},\cdots w_{JI},w_{JI+1},\cdots,w_{JI+j},\cdots w_{J(1+I)})\\&=(w_{1,1}^{(1)},\cdots,w_{j,i}^{(1)},\cdots w_{J,I}^{(1)},w_1^{(2)},\cdots,w_j^{(2)},\cdots w_J^{(2)})\end{aligned}. \tag{5}$$

1. Randomly select $w_{j,i}^{(1)}(j=2,\cdots,J;\ i=1,\cdots,I)$ and $w_j^{(2)}(j=1,\cdots,J)$ from two uniform distributions. The means of both distributions are equal to 0, and their variances are $1/J$ and 1, respectively. Set the initial iteration, $t=0$.

2. According to weight vector $\mathbf{w}(0)$, calculate the expected network outputs, $\psi(m)(m=1,2,\cdots,M)$, the derivative of outputs on all weights, $\frac{\partial\psi(m)}{\partial\mathbf{w}(0)}(m=1,2,\cdots,M)$, and the gradient vector, $\mathbf{g}(0)$:

$$\frac{\partial\psi(m)}{\partial w_h}=\begin{cases}\dfrac{\partial\psi(m)}{\partial w_j^{(2)}}=\tanh\Big(\sum_{i=1}^{I}w_{j,i}^{(1)}x_i(m)\Big), & if w_h=w_j^{(2)}\\ \dfrac{\partial\psi_k(m)}{\partial w_{j,i}^{(1)}}=w_j^2\tan h'\Big(\sum_{i=1}^{I}w_{j,i}^{(1)}x_i(m)\Big)x_i(m), & fw_h=w_{j,i}^{(1)}\end{cases}, \tag{6}$$

$$\mathbf{g}(t)=\frac{1}{N}\sum_{m=1}^{M}[o(m)-\psi(m)]\frac{\partial\psi(m)}{\partial\mathbf{w}(t)}. \tag{7}$$

3. Set $\mathbf{s}(0)=\mathbf{r}(0)=-\mathbf{g}(0)$

4. In iteration t, for the fixed $\mathbf{w}(t)$ and $\mathbf{s}(t)$, use the advance and retreat method to linearly search the optimal $\eta(t)$ by minimizing the cost function, $\xi_{av}(\mathbf{w}(t)+\eta\mathbf{s}(t))$:

$$\xi_{av}(\boldsymbol{w})=\frac{1}{2M}\sum_{m=1}^{M}[o(m)-\psi(m)]^2. \tag{8}$$

5. Check convergence criteria. If the Euclidean norm of $\mathbf{r}(t)$ decreases to a certain small portion, ε, of its initial value, $\|\mathbf{r}(0)\|$, or

the iteration number meets its maximum value, T, the algorithm is done:

$\|\mathbf{r}(n)\| \le \varepsilon\|\mathbf{r}(0)\|$, or $t = \mathrm{T}$.

6. Update the connection weight vector:

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \eta(t)\boldsymbol{s}(t). \tag{9}$$

7. Calculate the gradient vector $\mathbf{g}(t+1)$ by formulas (6)–(7) according to $\mathbf{w}(t+1)$. Set $\mathbf{r}(t+1) = -\mathbf{g}(t+1)$.

8. Calculate $\beta(t+1)$ by the Polak-Ribiere method:

$$\beta(t+1) = max\{\frac{\mathbf{r}'(t+1)(\boldsymbol{r}(t+1) - \boldsymbol{r}(t))}{\boldsymbol{r}'(t)\boldsymbol{r}(t)}, 0\}. \tag{10}$$

9. Update the direction vector:

$$\boldsymbol{s}(t+1) = \boldsymbol{r}(t+1) + \beta(t+1)\boldsymbol{s}(t). \tag{11}$$

10. Set $t = t+1$, and turn to step 4.

3.3. Structure optimization

Owing to Setiono and Leow (2000), the N2PFA algorithm, which has been successfully used to develop an optimized NN model for crash injury severity analysis (Zeng and Huang, 2014b), is proposed to improve the generalization capacity of the NN model and to identify insignificant explanatory variables. This method prunes the nodes that do not cause significant deterioration of the network's accuracy. The mean absolute deviations (MADs) of the training set **T** and testing set **X**, that is p and q, are used to evaluate the fitting and predictive performance during network optimization:

$$p = \frac{1}{M_1}\sum_{o(m)\in\mathbf{T}} |o(m) - \psi(m)|, \tag{12}$$

$$q = \frac{1}{M_2}\sum_{o(m)\in\mathbf{X}} |o(m) - \psi(m)|, \tag{13}$$

where M_1 and M_2 are the number of samples in the training and testing sets, respectively.

The following steps describe the detailed pruning process.

1. Train the network with a relatively large number of hidden nodes using the conjugate gradient algorithm.

2. Calculate the p and q of the trained NN, and set $p\_b = p, q\_b = q$, and $ermax = \max\{p\_b, q\_b\}$.

3. For each $i(i = 1, \cdots, I)$, set $w_{i,j}^{(1)} = 0(j = 1, \cdots, J)$ and calculate the fitting errors p_i.

4. Retrain the network with $w_{l,j}^{(1)} = 0(j = 1, \cdots, J)$, where $p_l = \min_i p_i$, and compute p and q for the retrained network.

5. If $p \le (1+\sigma)ermax$ and $q \le (1+\sigma)ermax$, then remove the input node l, set $p\_b = \min\{p, p\_b\}$, $q\_b = \min\{q, q\_b\}$, $ermax = \max\{p\_b, q\_b\}$, $I = I - 1$, and go back to step 3; otherwise, keep the previous weights of the network connections.

6. For each $j(j = 1, \cdots, J)$, set $w_j^{(2)} = 0$ and calculate the fitting errors p_j.

7. Retrain the network with $w_h^{(2)} = 0$, where $p_h = \min_j p_j$, and compute p and q of the retrained network.

8. If $p \le (1+\sigma)ermax$ and $q \le (1+\sigma)ermax$, then remove the hidden node h. Set $p\_b = \min\{p, p\_b\}$, $q\_b = \min\{q, q\_b\}$, $ermax = \max\{p\_b, q\_b\}$, and $J = J - 1$, and go back to step 6; otherwise, keep the previous weights of the network connections.

In the above process, p_b and q_b represent respectively the minimal MADs of training and testing sets achieved so far. During the pruning process, generally, p_b increases while q_b decreases. In addition, σ is the margin by which the error is allowed to increase, when pruning a certain node.

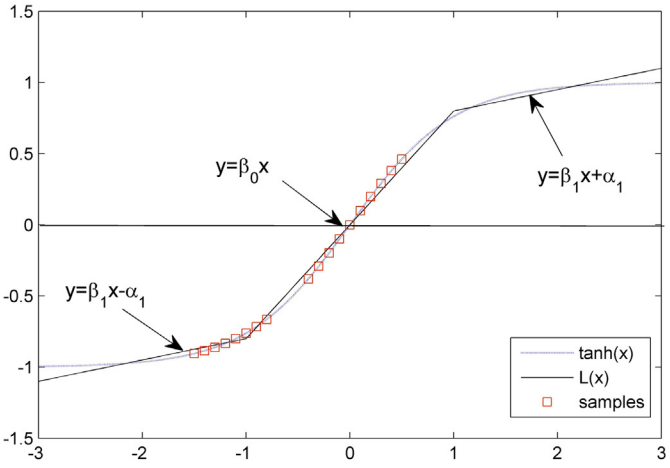

Fig. 2. Three-piece linear approximation of tanh(·).

3.4. Rule extraction

The rule extraction method developed by Setiono and Thong (2004) is modified to generate rules from the pruned NN for regression analysis. In the next subsections, a particle swarm optimization (PSO) algorithm based approach to approximating the transfer functions of hidden units is introduced as a critical step in the method, and then the rule extraction process is explicitly described.

3.4.1. Approximating transfer functions

The transfer functions of hidden nodes can be approximated by piecewise functions. Theoretically, more pieces fit the function, more accurate the rule-set may be, while more rules may be extracted. To balance the two aspects, a three-piece linear function suggested by Setiono and Thong (2004) is used to approximate the transfer function of each hidden node $j(j = 1, \cdots, J)$, tanh(·), as shown in Fig. 2. The slopes, β_{j0} and β_{j1}, and the cut-off point, ξ_{j0}, are three undetermined parameters which minimize the sum of the squared deviations, i.e.,

$$\min \sum_{m=1}^{M} (\tanh(\upsilon_j(m)) - L_j(\upsilon_j(m)))^2, \tag{14}$$

where

$$L_j(x) = \begin{cases} -\alpha_{j1} + \beta_{j1}x & if x < -\xi_{j0} \\ \beta_{j0}x & if -\xi_{j0} \le x \le \xi_{j0} \\ \alpha_{j1} + \beta_{j1}x & if x > \xi_{j0} \end{cases}, \tag{15}$$

$$\upsilon_j(m) = \sum_{i=1}^{I} w_{j,i}^{(1)} x_i(m), \tag{16}$$

$$\alpha_{j1} = (\beta_{j0} - \beta_{j1})\xi_{j0}. \tag{17}$$

3.4.2. Searching for the optimal parameters

To approximate the transfer function accurately, the PSO algorithm, an efficient global search method, is used to solve the aforementioned nonlinear optimization problem. The PSO algorithm is well-known for its exploration capacity, exploitation capacity and easy implementation (Poli et al., 2007). In the algorithm, each feasible solution $(\beta_{j0}, \beta_{j1}, \xi_{j0})$ is called a particle, **U**, and each particle flies around the three-dimensional search space with a velocity, **V**, which is updated iteratively according to the best solution of the particle achieved so far (particle best, **p***best*) and the

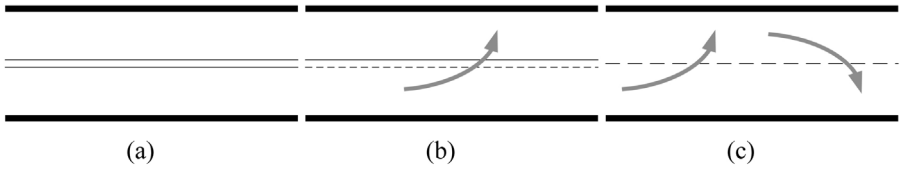

Fig. 3. Lane changing opportunity for different road section configurations.

Table 1
Descriptive statistics of the variables.

Variable	Description	Mean	S.D.	Min.	Max.
Response variable					
Crash	Crash count per segment per year	7.64	6.35	0	51
Exposure variables					
AADT	Average annual daily traffic (veh)	22077	19945	1164	101632
Length	Segment length (km)	1.47	1.55	0.15	9.07
Risk factors					
Lane	Number of lanes	2.41	1.18	1	7
Width	Average width of each lane	3.63	0.64	2.40	7.30
SL	Posted speed limit (km/h)	60.3	14.7	50	110
Merge	Number of merging ramps	0.84	1.00	0	4
Diverge	Number of diverging ramps	1.75	2.27	0	17
Inter	Number of intersections	1.90	2.37	0	16
Gradient	Average segment gradient (10^{-2})	0.04	2.74	−11	11
Curvature	Average segment curvature	21.9	17.5	0	85
LCO	Lane changing opportunity	2.43	1.61	0	7.85
Median	Presence of median barrier: yes = 1, no = 0	0.70	0.46	0	1
BS	Presence of bus stop: yes = 1, no = 0	0.64	0.48	0	1
Shoulder	Presence of hard shoulder: yes = 1, no = 0	0.13	0.34	0	1
Park	Presence of on-street parking: yes = 1, no = 0	0.51	0.49	0	1
Rainfall	Annual precipitation (mm)	2279	565	761	3215

best solution obtained by all particles in the swarm so far (global best, $\mathbf{g}best$):

$$\boldsymbol{V}_s^{r+1} = \mathbf{V}_s^r + c_1\lambda_1(\mathbf{pbest}_s^r - \boldsymbol{U}_s^r) + c_2\lambda_2(\mathbf{gbest}_s^r - \boldsymbol{U}_s^r), \quad (18)$$

$$\boldsymbol{U}_s^{r+1} = \mathbf{U}_s^r + \boldsymbol{V}_s^{r+1}, \quad (19)$$

$$\boldsymbol{U} = (\beta_{j0}, \beta_{j1}, \xi_{j0}); r = 1, 2, \cdots, R; s = 1, 2, \cdots, S.$$

where $\mathbf{U}_s^r$ is the sth particle at the rth iteration and $\mathbf{V}_s^r$ is its flying velocity to the rth iteration. c_1 and c_2 are two acceleration constants, while λ_1 and λ_2 are two uniform random numbers in [0, 1]. R is the maximum iteration number and S is the number of particles used for searching the optimal solution.

3.4.3. Generating regression rules

Once the transfer functions of hidden units have been approximated, the relationship between the network inputs and output can be formulated with piecewise linear functions. The detailed steps for extracting rules from the optimized NN are as follows:

1. For each hidden unit $j(j = 1, \cdots, J)$, generate a three-piece linear function $L_j(x)$ with the approach previously described.

2. According to the pair of cut-off points in $L_j(x)$, $-\xi_{j0}$ and ξ_{j0}, a certain input can be located in one of three sections of hidden node j. Then, J hidden nodes will result in $\underbrace{3 \times 3 \times \cdots \times 3}_{J}$ locations for inputs. Consequently, the whole input space can be separated into 3^J subspaces.

3. For each non-empty subspace, the rule consequence is $\tilde{y} = \sum_{j=1}^{J} w_j^{(2)} \cdot L_j(\upsilon_j)$, where $\upsilon_j = \sum_{i=1}^{I} w_{i,j}^{(1)} \cdot x_i$, and the rule condition is $C_1 \& C_2 \& \cdots C_J$, where C_j is either $\upsilon_j < -\xi_{j0}$, $-\xi_{j0} \le \upsilon_j \le \xi_{j0}$ or $\upsilon_j > \xi_{j0}$.

4. Data preparation and preliminary analysis

A crash dataset obtained from the Traffic Information System (TIS) maintained by the Transport Department of Hong Kong is used to demonstrate the proposed NN models and to compare them with the NB model. This dataset contains 211 road segments evenly and widely distributed across Hong Kong. Geographical information system (GIS) techniques are used to map crashes to these segments, and the annual crash numbers of each site during 2002–2006 are obtained. The road geometric and traffic information is also included in the dataset. Table 1 illustrates the definitions and descriptive statistics of the variables used in the model development.

The lane changing opportunity (LCO) variable refers to the length-weighted average number of eligible lane changing in a subsegment with identical lane markings. For double continuous lines, no lane changing is allowed (shown in Fig. 3(a)), thus $LCO = 0$. For double lines with one continuous line and one broken line, lane changing is only allowed from the side of the broken line to the side of the continuous line (shown in Fig. 3(b)), thus $LCO = 1$. For a single broken line, lane changing between both adjacent lanes is allowed (shown in Fig. 3(c)), thus $LCO = 2$. Pei et al. (2012) provided a more detailed description of LCO.

According to Table 1, the mean and variance of the crash frequencies are 7.64 and 40.32, respectively, which indicates the possible over-dispersion in the crash data. NB model is developed in which crash exposure is formulated by the product of a power of the average annual daily traffic (AADT) and a power of segment length, or $e = (AADT)^{\alpha_1}(Length)^{\alpha_2}$, where α_1 and α_2 are two parameters that can be estimated (Zeng and Huang, 2014a).

Correlation tests and multi-collinearity diagnoses for the risk factors are conducted. Table 2 shows the results of Pearson correlation tests. According to the results, we can find that ln(AADT) and Lane, ln(AADT) and Park, Lane and LCO, SL and Shoulder, SL

Table 2
Pearson correlation coefficients between explanatory variables.

	Lane	SL	Merge	Diverge	Inter	Median	BS	Gradient	Shoulder	Curvature	LCO	Width	Park	Rainfall	ln(AADT)	ln(Length)
Lane	1.00	0.35	0.24	−0.18	−0.31	0.57	−0.20	−0.01	0.20	−0.08	0.65	−0.14	−0.38	−0.02	0.71	−0.15
SL	0.35	1.00	0.24	−0.13	−0.41	0.46	−0.46	−0.02	0.69	0.17	0.28	0.09	−0.67	−0.13	0.57	0.50
Merge	0.24	0.24	1.00	0.19	−0.19	0.09	−0.03	−0.02	0.19	0.24	0.13	−0.01	−0.28	0.02	0.36	0.23
Diverge	−0.18	−0.13	0.19	1.00	0.56	−0.24	0.26	0.10	−0.16	0.22	−0.19	−0.06	0.09	0.01	−0.20	0.22
Inter	−0.31	−0.41	−0.19	0.56	1.00	−0.33	0.44	0.03	−0.31	−0.07	−0.17	0.00	0.47	0.04	−0.47	−0.06
Median	0.57	0.46	0.09	−0.24	−0.33	1.00	−0.39	−0.03	0.25	−0.25	0.51	0.16	−0.54	−0.05	0.59	−0.04
BS	−0.20	−0.46	−0.03	0.26	0.44	−0.39	1.00	0.05	−0.35	0.02	−0.13	−0.16	0.46	0.07	−0.36	−0.12
Gradient	−0.01	−0.02	−0.02	0.10	0.03	−0.03	0.05	1.00	−0.02	0.02	0.02	0.00	0.02	0.00	−0.04	0.03
Shoulder	0.20	0.69	0.19	−0.16	−0.31	0.25	−0.35	−0.02	1.00	0.21	0.32	0.19	−0.39	−0.17	0.31	0.38
Curvature	−0.08	0.17	0.24	0.22	−0.07	−0.25	0.02	0.02	0.21	1.00	−0.17	−0.06	−0.08	−0.02	0.00	0.40
LCO	0.65	0.28	0.13	−0.19	−0.17	0.51	−0.13	0.02	0.32	−0.17	1.00	0.07	−0.24	−0.03	0.48	−0.24
Width	−0.14	0.09	−0.01	−0.06	0.00	0.16	−0.16	0.00	0.19	−0.06	0.07	1.00	0.03	−0.03	−0.05	−0.08
Park	−0.38	−0.67	−0.28	0.09	0.47	−0.54	0.46	0.02	−0.39	−0.08	−0.24	0.03	1.00	0.09	−0.65	−0.34
Rainfall	−0.02	−0.13	0.02	0.01	0.04	−0.05	0.07	0.00	−0.17	−0.02	−0.03	−0.03	0.09	1.00	0.01	−0.10
ln(AADT)	0.71	0.57	0.36	−0.20	−0.47	0.59	−0.36	−0.04	0.31	0.00	0.48	−0.05	−0.65	0.01	1.00	0.09
ln(Length)	−0.15	0.50	0.23	0.22	−0.06	−0.04	−0.12	0.03	0.38	0.40	−0.24	−0.08	−0.34	−0.10	0.09	1.00

and Park are significantly correlated with correlation coefficients greater than 0.6. To reduce the model complexity, Lane, Park, and Shoulder are excluded from the models. The results of the diagnoses indicate that there is no significant collinearity in the remaining factors. In this study, the modified dataset is used for the development of both NB and NN models. Therefore, the exclusion would have little impact on the results of model comparison, although it possibly brings about omitted variable biases.

5. Model implementation and result analysis

5.1. Model implementation

The NB model is estimated with the Stata software. The training, structure optimization, and rule extraction algorithms of the NN model are programmed in MATLAB. All of the variables are normalized for improving the efficiency of network training. To compare the performance of the models fully, a 5-fold cross validation is conducted, where the dataset is randomly divided into five parts with equal number of observations/patterns. Each time, the subdataset of any four parts is input for training the models while the rest is used for testing the predictive performance. The number of input nodes $I = 14$, and initially set $J = 10$. In the network training, $\varepsilon = 0.001$ and $T = 50$. We assume that $\sigma = 0.05$ in the N2PFA algorithm, while $R = 300$ and $S = 700$ in the PSO algorithm.

5.2. Model comparison

The results of the model comparison are summarized in Table 3. With regard to the five folds of model comparison, in terms of the MAD criteria, all of the trained and optimized NNs have a bit smaller errors for the training and testing datasets than the NB models. It demonstrates that NNs of crash frequency offer slightly better approximation performance than NB models, which is consistent to the findings of the previous studies (Chang, 2005; Xie et al., 2007).

After pruning the network structure with the N2PFA algorithm, the model training performance is generally expected to be degraded to some extent but the model prediction should be improved as discussed in the Section 3.3. But in the results as shown in Table 3, it is surprisingly found that both the training and prediction errors of the NN models in three (No. 2, 4, 5) folds of models are reduced by the proposed model structure optimization algorithm. As generally known, like other training algorithms, the proposed conjugate gradient algorithm may have reached a local optimum (Haykin, 2009). Therefore, a presumable cause for the reduced model-training errors may be that pruning nodes and retraining network could help to escape from local minima and to search for better solutions. As a result, we may argue that the model generalization performance associated with the proposed algorithm is improved as reflected by the reduced model training and prediction errors.

Moreover, certain numbers of input and hidden nodes are removed from the trained NNs in all the five runs of models, indicating that the original models have redundant nodes and that the factors corresponding to those removed input nodes may have no significant effects on crash frequency.

Unsurprisingly, the rule-sets approximate the relationship between the crash frequency and the risk factors and the optimized NNs in terms of comparable training and testing MADs, as the generated three-piece linear functions could be considered substitutions for the transfer functions in the optimized NN. However, the linear function in each rule can illustrate the factors' effects, whereas the NN cannot.

It is also noticeable that the five optimized NNs end up with slight distinctions in MAD values and the final sets of input and hid-

Table 3
Model comparison.

	Method	Training MAD	Testing MAD	Number of input nodes	Number of hidden nodes
1	NB	3.560	3.912	–	–
	Trained NN	3.488	3.682	14	10
	Optimized NN	3.514	3.672	9	4
	Rule-set	3.524	3.631	–	–
2	NB	3.698	3.432	–	–
	Trained NN	3.590	3.412	14	10
	Optimized NN	3.458	3.142	9	5
	Rule-set	3.437	3.147	–	–
3	NB	3.575	3.769	–	–
	Trained NN	3.386	3.733	14	10
	Optimized NN	3.397	3.641	6	5
	Rule-set	3.392	3.652	–	–
4	NB	3.718	3.386	–	–
	Trained NN	3.463	3.300	14	10
	Optimized NN	3.402	3.121	8	3
	Rule-set	3.448	3.167	–	–
5	NB	3.504	4.009	–	–
	Trained NN	3.422	3.738	14	10
	Optimized NN	3.366	3.611	10	2
	Rule-set	3.346	3.647	–	–
Average	NB	3.611	3.702	–	–
	Trained NN	3.470	3.573	–	–
	Optimized NN	3.427	3.437	–	–
	Rule-set	3.429	3.449	–	–

den nodes. This instability is presumably attributable to the small sample size problem (Xie et al., 2007), given the important impact of sample size on the model generalization performance (Haykin, 2009).

5.3. Interpretation of the explanatory variables

This section presents a discussion for the identified exploratory factors for justifying the model validity. As an example, the rule-set generated on the first fold of model comparison is specified and analyzed in details. Tables 4 and 5 summarize the specific conditions and consequences of the extracted rules respectively. For comparison purpose, the estimation results of the significant parameters in the NB model are shown in Table 6, where the mean and 95% confidence interval of the over-dispersion parameter are 0.219 and [0.188, 0.255] respectively, suggesting that the employed crash data is over-dispersed.

The rule conditions in Table 4 may be difficult to be directly understood. Therefore, instead, we employ the characteristics of the road segments involved at certain particular rules to illustrate the effects of the risk factors. The analysis mainly focuses on the rule consequences in Table 5. Even so, according to the conditions, we can accurately determine which rule each observation in the analysis should be assigned to. Comparing the results in Table 5 with those in Table 6, we can find that the coefficients of the factors (including the constant) in the optimized NN are significant at 95% confidence level in the NB model.

Regarding the main effects of the risk factors identified, most of the risk factors have consistent signs as shown in Table 5, which also conform to the signs in NB model results shown in Table 6. In particular, for the factors AADT, speed limit (SL), bus stop (BS) and Rainfall, their signs of coefficients are identical in all rules. As for the other factors, it is interesting to find that they have positive coefficients in several rules while negative coefficients in the others. Moreover, it is observed that the coefficient values estimated are also distinct for several specific rules. This implies that those risk factors probably have variable safety effects at different road conditions. It could be an important evidence for the nonlinear relationship between crash frequency and the risk factors, which could not be identified and modeled with the traditional generalized linear regression models such as NB model.

According to the results based on the extracted rules, more crashes tend to occur on road segments with heavier daily traffic as observed by the coefficient estimations associated with all the ten rules for AADT (AASHTO, 2010; Zeng and Huang, 2014a). Nevertheless, the proposed NN model presents specific values for varied safety effects under different conditions. For example, increasing one unit AADT is expected to increase only 0.224 (based on the normalized data) crashes under Condition 8 but 1.295 (5.8 times of the former) crashes under Condition 5.

As expected, more crashes occur on longer roadways at most rules, because road segment length is often interpreted as a crash exposure variable in highway safety analysis. However, the length is found negatively related to the crash frequency under Condition 5, in which all crashes occurred on Tsing Long Highway. This is a high standard highway that is well designed and maintained, and is the longest road segment in the dataset. However, its average annual crash number is only 5.63, far smaller than the mean of the whole population (7.64). A possible reason for the negative effect of Length may be that some unobserved factors associated with this highway greatly promote the safety situation.

It is interesting to find that the crash frequency decreases with higher speed limits, which may contradict engineering intuitions and many existing studies (Aguero-Valverde and Jovanis, 2008). However, some previous researchers have argued that roadway segments designed for higher speeds are usually well- planned, constructed, and managed, features that promote road safety (Milton and Mannering, 1998).

Results show the presence of median under most conditions has positive effects in crash prevention. Donnell and Mason (2006) also found that median barriers could effectively prevent the occurrence of cross-median crashes. However, the estimated coefficients are positive at Rules 5, 9 and 10. For those observations at these rules, 92% of the road segments have median barriers, of which most are inner-city highways with heavy daily traffic (mean = 33,437 vehicle), long length (mean = 1.93 km) and many merging ramps (mean = 1.68). These factors may hinder safe driving and bring about more median-related collisions.

Table 4
Rule conditions.

Rule	Condition[a]
1	$v_1 < -0.217$ & $v_2 > 0.877$ & $v_3 < -0.411$ & $-0.703 \le v_4 \le 0.703$
2	$v_1 < -0.217$ & $-0.877 \le v_2 \le 0.877$ & $-0.411 \le v_3 \le 0.411$ & $-0.703 \le v_4 \le 0.703$
3	$v_1 < -0.217$ & $v_2 > 0.877$ & $-0.411 \le v_3 \le 0.411$ & $-0.703 \le v_4 \le 0.703$
4	$v_1 < -0.217$ & $-0.877 \le v_2 \le 0.877$ & $v_3 > 0.411$ & $-0.703 \le v_4 \le 0.703$
5	$-0.217 \le v_1 \le 0.217$ & $-0.877 \le v_2 \le 0.877$ & $v_3 > 0.411$ & $-0.703 \le v_4 \le 0.703$
6	$v_1 > 0.217$ & $-0.877 \le v_2 \le 0.877$ & $v_3 > 0.411$ & $-0.703 \le v_4 \le 0.703$
7	$v_1 < -0.217$ & $v_2 > 0.877$ & $v_3 > 0.411$ & $-0.703 \le v_4 \le 0.703$
8	$v_1 < -0.217$ & $v_2 > 0.877$ & $v_3 < -0.411$ & $v_4 > 0.703$
9	$v_1 < -0.217$ & $-0.877 \le v_2 \le 0.877$ & $-0.411 \le v_3 \le 0.411$ & $v_4 > 0.703$
10	$v_1 < -0.217$ & $v_2 > 0.877$ & $-0.411 \le v_3 \le 0.411$ & $v_4 > 0.703$

[a] $v_1 = -0.525 - 1.111AADT + 1.045Length + 0.302SL - 0.144Median - 0.518BS - 0.476Gradient + 0.285LCO - 0.333Rainfall$,
$v_2 = 0.783 + 0.383AADT - 1.11Length - 0.535SL + 0.149Median - 0.014BS + 0.197Gradient + 0.44LCO - 0.49Rainfall$,
$v_3 = -0.533 + 0.226AADT + 1.381Length - 0.004SL + 0.045Median + 0.07BS - 0.4Gradient + 0.234LCO - 0.001Rainfall$,
$v_4 = 0.75 - 0.018AADT - 0.644Length + 0.253SL + 0.295Median + 0.213BS - 0.191Gradient - 0.094LCO + 0.04Rainfall$.

Table 5
Rule consequences.

Rule	Coefficient of the variable in the consequence (linear function)								
	Constant	AADT	Length	SL	Median	BS	Gradient	LCO	Rainfall
1	−0.025	0.226	0.504	−0.145	−0.050	0.014	−0.020	0.111	0.055
2	−0.056	0.316	0.494	−0.231	−0.020	0.025	−0.038	0.211	0.134
3	−0.040	0.254	0.674	−0.145	−0.045	0.023	−0.069	0.140	0.054
4	0.061	0.288	0.325	−0.231	−0.026	0.017	0.012	0.182	0.134
5	0.340	1.295	−0.622	−0.505	0.104	0.486	0.443	−0.076	0.435
6	−0.332	0.288	0.325	−0.231	−0.026	0.017	0.012	0.182	0.134
7	0.076	0.226	0.504	−0.145	−0.050	0.014	−0.020	0.111	0.055
8	−0.018	0.224	0.404	−0.105	−0.004	0.047	−0.050	0.097	0.061
9	−0.048	0.313	0.394	−0.192	0.025	0.058	−0.067	0.197	0.140
10	−0.033	0.251	0.574	−0.106	0.001	0.056	−0.099	0.126	0.061

The presence of bus stop (BS) is found to have positive effects on the crash frequency. This may be attributed to frequent pedestrian activity around bus stops and increased interaction between buses and other vehicles when entering or leaving bus bays. In fact, 93% of the pedestrian-involved crashes in the observed road segments occurred on roadways with bus stops (Pei et al., 2012).

The negative coefficients of Gradient under most conditions indicate that more crashes are expected to occur on the involved road segments with steeper downgrade slopes, which is generally consistent with engineering experience. Besides, Gradient is found to have positive effects on the crash frequency at Rules 4–6. The covered roadways consist of the three longest ones—Tsing Long Highway (9.07 km), Shek O Road (7.75 km) and Tolo Highway (5.60 km). Driving on the downgrade directions of these long highways, drivers may be more careful, thus reducing the crash risk.

The variable Lane changing opportunity (LCO) has significant effect to increase the crash frequency under most conditions. It means that more lane changing opportunity would lead to higher crash risk. Lane changing maneuver often increases vehicle interaction, such as overtaking, thereby raising the incidence of traffic conflict (Pei et al., 2012). Similar to the variable Length, LCO has a negative effect on crash frequency under Condition 5 only. On Tsing Long Highway, lane changing maneuver is less frequent than on busy inner-city roadways, which may reduce the vehicle speed variance. This may possibly explain why LCO negatively affects the crash frequency on this highway.

Generally, rainfall obstructs the visibility and makes the road surfaces slippery, thereby reducing skidding resistance, which raises the probability of crash occurrence. That is why Rainfall has positive model coefficients under all conditions, which indicates rainfall condition may always lead to more crashes (Pei et al., 2012).

Table 6
Estimation of significant parameters in NB.

Variables	Mean	S.D.	Confidence interval	
			2.5%	97.5%
Constant	−5.899	0.390	−6.664	−5.134
ln(AADT)	0.493	0.033	0.428	0.558
ln(Length)	0.573	0.030	0.516	0.631
SL	−0.025	0.002	−0.030	−0.020
Diverge	0.034	0.011	0.014	0.055
Median	−0.184	0.064	−0.309	−0.059
BS	0.373	0.050	0.274	0.471
Gradient	−1.755	0.717	−3.159	−0.350
LCO	0.110	0.016	0.079	0.140
Rainfall	0.00008	0.00004	0.00002	0.00015
α	0.219	0.017	0.188	0.255

6. Conclusions and future research

This study develops an NN for modeling possible nonlinear relationship between crash frequency and risk factors. To improve the generalization capacity and to deal with the black-box characteristic of the NN, a structure optimization N2PFA algorithm and a modified rule extraction algorithm are proposed. A crash dataset obtained from the TIS maintained by the Transport Department of Hong Kong is used to demonstrate the proposed methods and to compare them with the results of an NB model.

The results show that both the trained and optimized NNs outperform the NB models in fitting and predictive performance to somewhat extent. In the optimized NNs, certain numbers of

input and hidden nodes are dropped off, and better approximation performance is achieved, demonstrating the ability of the N2PFA algorithm to identify insignificant factors and to improve the model generalization capacity. The optimized NN generates ten rules in which the coefficients of the explanatory variables are different, which confirms that they are nonlinearly related to the crash frequency. The signs of these coefficients have identical directions under most conditions, and are consistent with those in the NB model. Moreover, most of the results for the explanatory variables are reasonable and conform to traffic engineering experience or the findings of previous studies, which further validates the proposed method.

Compared with NB and NN models as employed in previous traffic safety studies, the improved NN techniques not only achieve better fits when modeling crash frequency, but also illustrate the effects of the risk factors. As NN is a universal approximator, these methods may also be useful in other aspects of highway safety analysis, such as jointly modeling crash frequency and injury severity, identifying sites with promise and evaluating countermeasure effectiveness. Comparing the proposed approaches with Empirical Bayes or other state-of-the-art methods in the application of hotspot identification would be an interesting research topic. Finally, it is worth noting that the training and testing MADs ($\approx$3.5) are about 50% of the average crash count (7.64) per segment per year, which may be attributed to the heterogeneity among the observations. Moreover, a number of recently proposed statistical models (e.g. Bayesian hierarchical and random parameter models) may also outperform the fixed NB model. This study presents an additional vision for improving crash modeling techniques from the perspective of improved AI model. Further research efforts could also be made to compare the proposed NN model with the emerging advanced statistical models based on more field datasets.

Acknowledgements

This research was jointly supported by the Natural Science Foundation of China (No. 71371192, 71301083), the Research Funds of Tsinghua University (No. 20151080412), the University Research Committee of the University of Hong Kong (201109176069), and the Hong Kong Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU 717512).

References

AASHTO, 2010. Highway Safety Manual, 1st edition.
Aguero-Valverde, J., Jovanis, P.P., 2008. Analysis of road crash frequency with spatial models. Transp. Res. Rec. 2061, 55–63.
Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accid. Anal. Prev. 41 (1), 153–159.
Chang, L., 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Saf. Sci. 43 (8), 541–557.
Dong, N., Huang, H., Xu, P., Ding, Z., Wang, D., 2014. Evaluating spatial proximity structures in TAZ-level crash prediction models. Transp. Res. Rec. 2432, 46–52.
Dong, N., Huang, H., Zheng, L., 2015. Support vector machine in crash prediction at the level of traffic analysis zones: assessing the spatial proximity effects. Accid. Anal. Prev. 82, 192–198.
Dong, N., Huang, H., Lee, J., Gao, M., Abdel-Aty, M., 2016. Macroscopic hotspots identification: a Bayesian spatio-temporal interaction approach. Accid. Anal. Prev. 92, 256–264.
Donnell, E.T., Mason Jr., J.M., 2006. Predicting the frequency of median barrier crashes on Pennsylvania interstate highways. Accid. Anal. Prev. 38 (3), 590–599.
Elalfi, A.E., Haque, R., Elalami, M.E., 2004. Extracting rules from trained neural network using GA for managing E-business. Appl. Soft Comput. 4 (1), 65–77.
Fritzke, B., 1994. Growing cell structures—a self-organizing network for unsupervised and supervised learning. Neural Netw. 7 (9), 1441–1460.
Haykin, S.S., 2009. Neural Networks and Learning Machines, 3rd edition. Prentice Hall, New York.
Hruschka, E.R., Ebecken, N.F., 2006. Extracting rules from multilayer perceptrons in classification problems: a clustering-based approach. Neurocomputing 70 (1), 384–397.
Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accid. Anal. Prev. 42 (6), 1556–1565.
Huang, H., Chin, H.C., 2010. Modeling road traffic crashes with zero-inflation and site-specific random effects. Stat. Methods Appl. 19 (3), 445–462.
Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., Abdel-Aty, M., 2016. Macro and micro models for zonal crash prediction with application in hot zones identification. J. Transp. Geogr. 54, 248–256.
Karlaftis, M.G., Vlahogianni, E.I., 2011. Statistical methods versus neural networks in transportation research: differences, similarities and some insights. Transp. Res. Part C: Emerg. Technol. 19 (3), 387–399.
Lee, J., Abdel-Aty, M., Choi, K., Huang, H., 2015. Multi-level hot zone identification for pedestrian safety. Accid. Anal. Prev. 76, 64–73.
Lehtokangas, M., 1999. Modelling with constructive backpropagation. Neural Netw. 12 (4), 707–716.
Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accid. Anal. Prev. 40 (4), 1611–1618.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A: Policy Pract. 44 (5), 291–305.
Lord, D., Guikema, S., Geedipally, S.R., 2008. Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. Accid. Anal. Prev. 40 (3), 1123–1134.
Ma, L., Khorasani, K., 2003. A new strategy for adaptively constructing multilayer feedforward neural networks. Neurocomputing 51, 361–385.
Malyshkina, N., Mannering, F., Tarko, A., 2009. Markov switching negative binomial models: an application to vehicle accident frequencies. Accid. Anal. Prev. 41 (2), 217–226.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22.
Marzban, C., Witt, A., 2001. A Bayesian neural network for severe-hail size prediction. Weather Forecast 16 (5), 600–610.
Miaou, S.-P., Bligh, R.P., Lord, D., 2005. Developing median barrier installation guidelines: a benefit/cost analysis using Texas data. Transp. Res. Rec. 1904, 3–19.
Miaou, S.-P., 1994. The relationship between truck accidents and geometric design of road sections: poisson versus negative binomial regressions. Accid. Anal. Prev. 26 (4), 471–482.
Milton, J., Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation 25 (4), 395–413.
Nielsen, A.B., Hansen, L.K., 2008. Structure learning by pruning in independent component analysis. Neurocomputing 71 (10), 2281–2290.
Park, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accid. Anal. Prev. 41 (4), 683–691.
Park, B.J., Lord, D., Hart, J.D., 2010. Bias properties of Bayesian statistics in finite mixture of negative regression models for crash data analysis. Accid. Anal. Prev. 42 (2), 741–749.
Pei, X., Wong, S.C., Sze, N.N., 2012. The roles of exposure and speed in road safety analysis. Accid. Anal. Prev. 48, 464–471.
Poli, R., Kennedy, J., Blackwell, T., 2007. Particle swarm optimization. Swarm Intell. 1 (1), 33–57.
Setiono, R., Leow, W.K., 2000. Pruned neural networks for regression. In: PRICAI 2000 Topics in Artificial Intelligence. Springer, Berlin Heidelberg, pp. 500–509.
Setiono, R., Thong, J.Y., 2004. An approach to generate rules from neural networks for regression problems. Eur. J. Oper. Res. 155 (1), 239–250.
Setiono, R., Leow, W.K., Zurada, J.M., 2002. Extraction of rules from artificial neural networks for nonlinear regression. IEEE Trans. Neural Netw. 13 (3), 564–577.
Shankar, V., Milton, J., Mannering, F.L., 1997. Modeling accident frequency as zero-altered probability processes: an empirical inquiry. Accid. Anal. Prev. 29 (6), 829–837.
Shankar, V.N., Albin, R.B., Milton, J.C., Mannering, F.L., 1998. Evaluating median cross-over likelihoods with clustered accident counts: an empirical inquiry using random effects negative binomial model. Transp. Res. Rec. 1635, 44–48.
Villiers, J., Barnard, E., 1993. Backpropagation neural nets with one and two hidden layers. IEEE Trans. Neural Netw. 4 (1), 136–141.
Wang, J., Huang, H., 2016. Road network safety evaluation using Bayesian hierarchical joint model. Accid. Anal. Prev. 90, 152–158.
Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2011. Statistical and Econometric Methods for Transportation Data Analysis, 2nd edition. CRC Press, New York.
Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian neural networks: an empirical analysis. Accid. Anal. Prev. 39 (5), 922–933.
Xu, J., Ho, D.W., 2006. A new training and pruning algorithm based on node dependence and Jacobian rank deficiency. Neurocomputing 70 (1), 544–558.
Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: random parameter versus geographically weighting. Accid. Anal. Prev. 75, 16–25.
Xu, P., Huang, H., Dong, N., Abdel-Aty, M., 2014. Sensitivity analysis in the context of regional safety modeling: identifying and assessing the MAUP effects. Accid. Anal. Prev. 70, 110–120.
Zeng, Q., Huang, H., 2014a. Bayesian spatial joint modeling of traffic crashes on an urban road network. Accid. Anal. Prev. 67, 105–112.
Zeng, Q., Huang, H., 2014b. A stable and optimized neural network model for crash injury severity prediction. Accid. Anal. Prev. 73, 351–358.