J. Korean Soc. Transp.
Vol.32, No.4, pp.369-379, August 2014
http://dx.doi.org/10.7470/jkst.2014.32.4.369

Journal of Korean Society of Transportation
pISSN : 1229-1366
eISSN : 2234-4217

Article

유전자 프로그래밍을 이용한 고속도로 사고예측모형

곽호찬[1] · 김동규[1] · 고승영[2] · 이청원[2]*

[1] 서울대학교 건설환경종합연구소, [2] 서울대학교 건설환경공학부

A Crash Prediction Model for Expressways Using Genetic Programming

KWAK, Ho-Chan[1] · KIM, Dong-Kyu[1] · KHO, Seung-Young[2] · LEE, Chungwon[2]*

[1] Integrated Research Institute of Construction and Environmental Engineering, Seoul National University, Seoul 151-744, Korea

[2] Department of Civil and Environmental Engineering, Seoul National University, Seoul 151-744, Korea

Abstract

The Statistical regression model has been used to construct crash prediction models, despite its limitations in assuming data distribution and functional form. In response to the limitations associated with the statistical regression models, a few studies based on non-parametric methods such as neural networks have been proposed to develop crash prediction models. However, these models have a major limitation in that they work as black boxes, and therefore cannot be directly used to identify the relationships between crash frequency and crash factors. A genetic programming model can find a solution to a problem without any specified assumptions and remove the black box effect. Hence, this paper investigates the application of the genetic programming technique to develope the crash prediction model. The data collected from the Gyeongbu expressway during the past three years (2010-2012), were separated into straight and curve sections. The random forest technique was applied to select the important variables that affect crash occurrence. The genetic programming model was developed based on the variables that were selected by the random forest. To test the goodness of fit of the genetic programming model, the RMSE of each model was compared to that of the negative binomial regression model. The test results indicate that the goodness of fit of the genetic programming models is superior to that of the negative binomial models.

전통적인 사고예측모형은 통계적 회귀분석에 주로 의존하였으나, 이는 자료 분포 및 함수 형태에 대한 가정에 따른 한계를 가지고 있다. 이에 따라 일부 연구는 신경망 등의 비모수적 기법을 모형 구축에 활용하였으나, 이는 독립변수와 종속변수 간의 직접적인 관계 규명이 어렵다는 한계가 있다. 유전자 프로그래밍 기법은 모형 개발에 특별한 가정이 필요없고, 사고요인 규명이 가능하다는 장점이 있다. 따라서 본 연구에서는 고속도로의 사고예측에 유전자 프로그래밍 기법을 적용함으로써 이러한 한계를 극복하고자 하였다. 이를 위하여 경부고속도로에서 최근 3년간(2010-2012년) 구득된 자료를 활용하였으며, 보다 세밀한 사고 특성 규명을 위해 고속도로 구간을 직선 구간과 곡선 구간으로 구분하였다. 사고 발생에 중요한 영향을 미치는 변수를 선택하기 위하여 랜덤 포레스트 기법을 이용하였으며, 최종 선택된 변수들을 활용하여 사고예측을 위한 유전자 프로그래밍 모형을 구축하였다. 구축된 모형의 예측 성능을 평가하기 위해 음이항 회귀모형과 비교해본 결과, 유전자 프로그래밍 모형의 예측 성능이 더 우수한 것으로 나타났다.

Keywords

crash prediction, expressway, genetic programming, random forest, traffic safety
사고 예측, 고속도로, 유전자 프로그래밍, 랜덤 포레스트, 교통 안전

* : Corresponding Author
chungwon@snu.ac.kr, Phone: +82-2-880-7368, Fax: +82-2-873-2684

Received 25 December 2013, Accepted 21 May 2014

서론

사고예측모형은 교통사고에 영향을 미칠 수 있는 도로의 기하구조, 교통 특성, 그리고 기타 주변 환경 등을 독립변수로 하여 해당 도로 구간의 단위 시간당 예측되는 사고 빈도수를 추정하는 모형이다(Zhong et al., 2009). 이는 예산의 효율적 집행을 위해 도로안전 개선사업의 우선순위를 정하는 기준으로 사용될 수 있으며, 교통사고에 영향을 미치는 요인에 대한 규명을 통해 도로안전을 개선하기 위한 방향을 제시할 수 있기 때문에 교통안전 분야에서 중요한 이슈가 되고 있다. 특히, 고속도로의 경우 다른 일반도로에 비해 교통사고 발생으로 인한 피해 심각도가 상당히 높은 것으로 알려져 있기 때문에 고속도로에 대한 보다 현실성 높은 사고예측모형이 구축된다면 안전성 증진 효과를 높일 수 있을 것이다.

사고예측모형 개발 및 적용과 관련된 기존의 연구들은 대부분 교통사고 발생의 확률적 분포를 가정한 통계적 회귀모형에 초점을 맞추어 왔다. 이들 연구에서는 교통사고의 발생분포를 정규 분포, 포아송 분포 혹은 음이항 분포를 따르는 것으로 가정하고 모형을 추정하였으며, 모수 추정을 위해 최소자승법 또는 최우추정법을 사용하였다. 이렇듯 전통적인 통계적 회귀분석은 일반적으로 자료의 분포와 독립변수 및 종속변수 간의 관계를 설명할 수 있는 함수 형태에 대한 가정이 필요하다. 이러한 기본적인 가정에 위배되었을 경우, 부적절한 추정결과가 도출될 수 있으며, 해당 분석 결과는 큰 의미를 가지기 어렵다(Hauer, 2004).

통계적 회귀분석과 관련된 이와 같은 한계로 인해 몇몇 연구자들은 사고예측모형 개발에 신경망 등의 비모수적 기법을 사용하였다. 하지만 이러한 비모수적 기법의 가장 중요한 한계는 해당 모형들이 블랙박스(black box)로 작용한다는 점이다. 즉, 비모수적 기법을 통해 구축된 사고예측모형은 사고빈도수와 다양한 독립변수 간의 관계를 직접적으로 규명할 수 없다. 따라서 이러한 방법론으로는 사고 요인 규명을 통한 개선방안 도출이라는 사고예측모형의 기본적인 목적을 달성할 수 없다.

유전자 프로그래밍(genetic programming)은 진화론적 이론에 기반한 최적화 기법으로, 분류 및 회귀 문제를 풀이하기 위한 방법론으로 많이 사용되고 있다(Koza, 1994). 통계적 회귀모형 및 비모수적 모형과 비교하여 유전자 프로그래밍 모형은 크게 두 가지 장점을 가지고 있다. 첫째, 자료 분포 및 함수 형태에 대한 특별한 가정 없이 문제의 해법을 찾을 수 있다. 유전자 프로그래밍 모형의 해는 수학연산자로 표현 가능한 모든 선형 및 비선형 함수 형태를 모사할 수 있다. 둘째, 비모수적 기법과 달리 블랙박스 효과를 제거하고 수학적으로 해석가능한 모형을 도출한다. 이는 독립변수와 종속변수간의 관계를 정의할 수 있음을 의미하며, 이를 통해 사고 요인의 개선 등 공학 측면에서의 실제적인 적용이 가능하다. 이러한 장점에도 불구하고 고속도로의 사고예측을 위한 유전자 프로그래밍 모형의 적용성을 규명한 연구는 거의 이루어지지 않았다.

본 연구의 목적은 고속도로에서 발생하는 사고에 영향을 미치는 변수와 사고 빈도수와의 관계를 규명하고, 사고를 예측하는데 있어 유전자 프로그래밍 모형의 적용성을 평가하는 것이다. 유전자 프로그래밍의 경우 자체적으로 사고에 중요한 영향을 미치는 변수를 선택하는 것이 어렵기 때문에 본 연구에서는 랜덤 포레스트(random forest, RF) 기법을 활용하여 변수를 선택하였다. 이 방법은 변수의 중요도를 평가하는데 가장 효율적인 방법 중 하나로 알려져 있으며(Breiman, 2001), 이를 통해 도출된 변수들을 기반으로 유전자 프로그래밍 사고예측모형을 구축하였다. 유전자 프로그래밍 모형의 예측 성능을 평가하기 위하여 사고예측모형에 가장 많이 사용되는 음이항 회귀모형을 동일한 자료에 대해 구축하여 각 모형의 평균 제곱근 오차(root mean square error, RMSE)를 비교하였다.

Ⅱ장에서는 사고예측모형과 관련된 기존문헌을 고찰하고, 분석에 사용된 랜덤 포레스트와 유전자 프로그래밍에 대한 이론적 고찰을 수행하였다. Ⅲ장에서는 사고예측모형 구축을 위해 사용된 자료의 범주 및 내용을 기술하였으며, Ⅳ장에서는 유전자 프로그래밍 모형의 구축결과를 제시하고 예측 성능 평가 결과를 비교하였다. 마지막으로 Ⅴ장에서는 본 연구의 결론을 요약하고 향후 연구에 대해 기술하였다.

기존문헌 고찰 및 분석 방법론

1. 기존문헌 고찰

Abdel-Aty and Radwan(2000)은 대부분의 사고자료에서 과분산 현상이 나타나고 있으며, 이에 따라 사고예측 모형에서는 음이항 모형이 포아송 모형에 비해 우수한 결과를 도출한다는 결론을 도출하였다. Lord et

al.(2005)은 포아송 모형과 음이항 모형을 통해 사고에 대한 통계학적 접근이 가능하다고 결론지었다. 포아송 모형은 일부 제한적인 조건 하에서 사고를 잘 설명하는 반면, 음이항 모형은 대부분의 경우에 더 나은 결과를 도출한다고 언급하였다. Zhong et al.(2009)은 중국 고속도로에서 발생한 사고 자료에 대한 분포를 포아송, 음이항, ZIP(zero inflated Poisson), ZINB(zero inflated negative binomial) 네 가지로 가정하여 사고예측모형을 구축하였으며, 최적모형으로 음이항 모형을 선택하였다. Kononov et al.(2008)과 Kononov et al.(2011)은 신경망을 이용한 사고예측모형을 구축하였으며, Li et al.(2008)의 연구에서는 서포트벡터머신(support vector machine, SVM)을 이용하여 사고예측모형을 구축하였다. 이들 연구에서는 음이항 회귀모형과의 모형 적합도 비교를 통해 해당 모형의 적합도가 더 높음을 증명하였다.

사고예측모형과 관련된 연구는 국내에서도 활발히 이루어지고 있는데, 국내 연구의 대부분은 아직 통계학적 회귀모형에 의존하고 있다. Kang and Lee(2002)는 고속도로의 직선부와 곡선부에 대해 사고를 예측하는 선형회귀식을 개발하였으며, Kang et al.(2002)은 고속도로 곡선부에 대한 음이항 회귀모형을 개발하였다. Park(2007)은 고속도로 트럼펫 연결로 전체와 연결로 형식별로 사고를 예측하기 위한 음이항 회귀모형을 개발하였다. Han et al.(2008)은 사고예측을 위하여 Hauer(2004)가 제시한 방법론을 한국의 고속도로에 적용하였다.

이와 같이 국내외의 사고예측모형 개발과 관련된 연구는 대부분 통계적 회귀모형과 일부 비모수적 기법에 한정되어 있었다. 따라서 본 연구에서는 이들의 한계를 보완할 수 있는 유전자 프로그래밍 모형을 통한 사고예측 프로세스를 제시하고 모형의 적용성을 검토하였다.

2. 분석 방법론

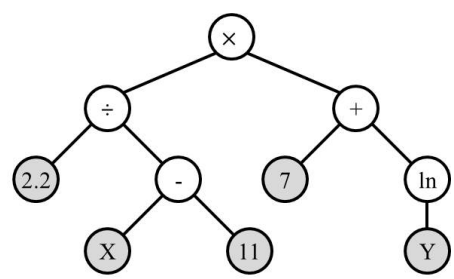
1) Random Forest (RF)

RF 기법은 트리 모형에 기반한 분석 방법론 중 하나로, 일반적인 트리 모형은 결과에 대한 불안정성이 단점으로 지적되고 있는데 반해, RF 기법은 임의로 생성되는 다수의 트리를 통해 평균적인 결과를 도출한다는 측면에서 일반적인 트리 모형에 비해 상당히 안정적인 결과를 도출할 수 있다는 장점이 있다. 또한 RF 모형의 학습 과정에서 각 트리는 학습 자료 중 랜덤하게 선택된 부트스트랩 샘플(bootstrap sample)에 기반하여 생성되기 때문에 과적합 문제없이 비교적 정확한 결과를 도출할 수 있다(Breiman, 2001).

RF 모형에서는 미리 지정된 수의 트리가 임의적으로 생성되고 각 트리에서 생성된 결과를 바탕으로 예측결과를 도출하는데, 회귀 문제의 경우 각 트리 결과에 대한 평균값이 최종 예측결과로 도출된다. 이러한 과정에서 RF 기법은 out-of-bag(OOB) 자료에 대한 예측 정확도 평가를 통해 각 설명변수의 중요도를 평가할 수 있다. 회귀 문제에 대한 RF 모형에서는 평균제곱오차(mean square error, MSE)에 기초한 지표가 변수 중요도를 평가하기 위해 주로 사용된다. 즉, 트리 내 각 노드의 분할(split) 단계에서 분할에 따른 MSE의 감소분이 분할하기 위해 사용되는 설명변수에 대해 계산된다. 그리고 각 설명변수에 대한 변수 중요도 지표는 모든 트리에서의 MSE 평균 감소분으로 계산된다. 이와 같은 과정을 통해 최종적으로 MSE 감소분이 큰 변수일수록 더 큰 변수 중요도 값을 가지게 된다.

2) 유전자 프로그래밍

유전자 프로그래밍은 진화론적 알고리즘의 일종으로 어떤 문제의 정확한 해 또는 추정치를 나타내는 수학적 모형을 생성하기 위해 사용될 수 있다(Koza, 1992). 이는 대부분의 배후이론이 유전자 알고리즘(genetic algorithm)과 같아 유전자 알고리즘의 확장이라고 볼 수 있으며, 둘 사이의 주요한 차이는 개체를 표현하는 방법이다. 유전자 알고리즘 모형에서 개체는 고정된 길이의 이진 스트링(binary string)으로 코드화된 숫자이지만, 유전자 프로그래밍 모형에서의 개체는 수학적인 기호 및 설명변수로 이루어진 트리로 코드화된 수학적인 모형이다(Figure 1). 여기서 종착 노드는 설명변수 또



$$[2.2 \div (X - 11)] \times [7 + \ln(Y)]$$

Figure 1. Example of genetic programming model

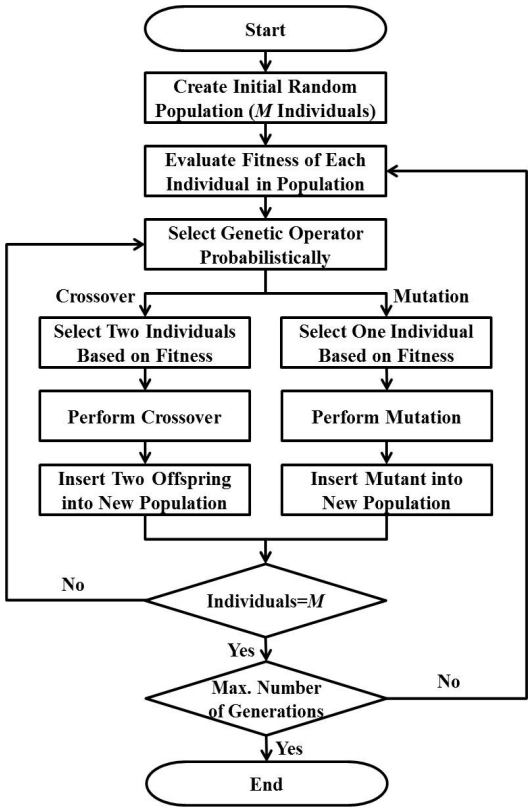


Figure 2. Flowchart of genetic programming

는 상수를 나타내며, 연결 노드들은 이들을 조합하기 위한 수학연산자를 의미한다.

유전자 프로그래밍은 유전자 알고리즘과 마찬가지로 유전 연산자를 통한 최적 적합체들의 생존이라는 진화론적 이론에 기초하여 수행된다. 즉, 각 세대에서 다수의 개체가 적합도에 기초하여 확률적으로 선택되고, 교배(crossover), 돌연변이(mutation) 등의 유전 연산자에 의해 새로운 개체군이 형성된다. 새로 생성된 개체군이 또 알고리즘의 다음 순서에 사용되고, 이러한 과정을 반복하다 종료조건을 만족할 경우 알고리즘이 종료된다(Figure 2).

유전자 프로그래밍에서 가장 중요한 구성요소는 바로 유전 연산자와 적합도 함수이다. 유전자 프로그래밍에서 주로 사용되는 유전 연산자는 교배, 재생성, 돌연변이 등이다. 유전자 알고리즘에서와 유사하게 교배 연산자는 두 개의 개체를 선택해서 각 개체의 일부분을 교환시키는 것이다. 재생성은 단순히 현재 세대의 구성원을 다음 세대로 복사하는 것이다. 돌연변이는 개체군의 다양성을 제공해주는 중요한 연산자로써 임의의 노드를 선택하여 임의적으로 변경시키는 것이다. 적합도 함수는 개체군에 대한 트리가 얼마나 잘 문제를 해결할 수 있는지를 결정하는 요소이다. 적합도 함수는 문제의 형태별로 상당히 다양하며, 일반적으로 모형에 의해 예측된 값과 실제 값 간의 오차에 기초하여 개발된다.

분석자료 구축

1. 분석자료 개요

본 연구에서는 사고예측모형 구축을 위하여 경부고속도로에 대해 수집된 자료를 활용하였다. 경부고속도로는 서울과 부산을 연결하는 416.0km의 연장을 가지는 한국에서 가장 긴 고속도로이며, 연간 가장 많은 사고가 발생하는 고속도로이기도 하다. 분석에 필요한 사고 자료 및 교통량 자료는 2010-2012년의 3년 동안 한국도로공사에 의해 집계된 자료를 활용하였다. 본 연구의 분석 범위는 고속도로 본선부를 대상으로 하고 있으며, 이에 따라 램프부와 휴게소 등 본선부 이외에서 발생한 사고를 제외한 5,191건의 사고를 대상으로 분석을 수행하였다. 보다 세밀한 사고 특성을 반영하기 위하여 도로 구간을 직선 구간과 곡선 구간으로 구분하였으며, 2010-2011년 자료는 모형의 구축을 위한 학습 자료로, 2012년 자료는 모형의 예측성능 비교를 위한 검증 자료로 활용하였다.

사고예측모형을 구축하기 위해서는 도로 구간을 적절한 분석 단위로 나눌 필요가 있는데, 이를 위하여 도로 구간을 일정한 단위 길이로 나누는 고정 길이 방법론(fixed length method)과 도로의 동질성에 기초하여 나누는 가변 길이 방법론(variable length method)이 사용되고 있다(Zhong et al., 2009). 본 연구에서는 상대적으로 통계적 적합도가 높고, 도로 구간의 기하구조 특성 및 교통 특성을 보다 정확히 반영할 수 있는 가변 길이 방법론을 사용하였다. 이에 따라 연평균일교통량(AADT) 및 도로의 기하구조, 즉 평면곡선반경과 종단구배가 동일한 구간을 하나의 분석단위로 설정하였으며, 경부고속도로 구간을 세분한 결과, 상행선의 경우 1,696개의 구간으로 나누어져 각 구간에 대한 평균 길이는 약 245m인 것으로 나타났으며, 하행선의 경우 1,694개의 구간으로 나누어져 각 구간에 대한 평균 길이는 약 246m인 것으로 나타났다.

Table 1. Candidate variables

Factor	Variable (Symbol)	Description
Exposure	EXPO (X1)	Exposure variable (10^6 veh-km)
Geometry	Curve (X2)	Radius of horizontal curve (1,000m)
	Slope (X3)	Grade of vertical curve (%)
	In (X4)	On-ramp (0 : no, 1 : yes)
	Out (X5)	Off-ramp (0 : no, 1 : yes)
Traffic	HV (X6)	Heavy vehicle ratio
Environment	E1 (X7)	Safety sign (0 : no, 1 : yes)
	E2 (X8)	Roadway surface roughness (0 : no, 1 : yes)
	E3 (X9)	Road side barrier (0 : no, 1 : yes)
	E4 (X10)	Speed camera (0 : no, 1 : yes)
	E5 (X11)	Antiskid sign (0 : no, 1 : yes)
	E6 (X12)	Tubular marker (0 : no, 1 : yes)
	E7 (X13)	Visual guidance facility (0 : no, 1 : yes)
	E8 (X14)	Lighting (0 : no, 1 : yes)
	E9 (X15)	Lighting (night) (0 : no, 1 : yes)
	E10 (X16)	Median barrier (0 : no, 1 : yes)
	E11 (X17)	Crash cushion (0 : no, 1 : yes)

2. 후보변수 설정

본 연구에서 사고예측모형 구축을 위해 사용되는 종속변수는 연간 발생한 사고 빈도수이다. 또한 이를 설명하기 위한 독립변수로 사고발생에 영향을 미치는 요인으로 알려진 노출도 변수, 기하구조 변수, 교통류 특성 변수, 그리고 주변 환경과 관련된 변수들을 본 연구에서 검토하였다.

먼저, 도로 구간의 길이와 AADT는 사고 빈도수에 가장 큰 영향을 미치는 것으로 연구된 바 있다(Zhong et al., 2009). 따라서 식(1)을 통해 이들을 사고에의 노출도를 나타내는 변수(EXPO)로 설정하였다.

$$EXPO = \frac{AADT \times 365 \times L}{10^6} \quad (1)$$

여기서, L : 구간 길이(km)

Table 2. Descriptive statistics of continuous variables

Section	Variable	Avg.	S.D.	Max.	Min.
Straight	EXPO	4.23	5.37	45.72	0.12
	Slope	0.02	0.97	6.40	-5.02
	HV	0.38	0.08	0.54	0.19
Curve	EXPO	2.54	2.36	16.31	0.12
	Curve	2.10	3.04	16.00	0.01
	Slope	0.03	1.03	6.00	-5.00
	HV	0.41	0.07	0.54	0.19

다음으로 기하구조 변수로는 도로의 선형을 나타내는 평면선형의 곡선반경 변수(Curve)와 종단선형의 기울기 변수(Slope)를 설정하였으며, 진입램프(In)와 진출램프(Out)의 존재유무 또한 더미변수로 설정하였다. 버스 및 트럭 등의 중차량들은 일반 승용차에 비해 더 낮은 속도로 도로를 주행하기 때문에 이들로 인해 발생하는 교통류의 속도 차이는 사고 발생에 큰 영향을 주기 마련이다. 따라서 본 연구에서는 전체 교통량 중 버스 및 트럭 등의 중차량이 차지하는 비율을 교통류 특성 변수로 검토하였다. 주변 환경 관련 변수로는 교통안전표지, 노면요철포장, 노측방호울타리, 무인단속카메라, 미끄럼방지표지, 시선유도봉, 시선유도시설, 조명시설, 조명시설(야간), 중분대방호울타리, 충격흡수시설 등 도로 주변에 설치된 각종 시설물의 존재 유무를 더미변수로 설정하여 사고와의 관련성을 검토하였다. 이와 같이 본 연구에서는 Table 1에 나타난 것과 같이 총 17개의 후보 변수들을 검토하였으며, 이들 중 연속변수에 대한 기술통계량은 Table 2에 나와 있다.

분석 결과

1. 변수 선정 결과

본 연구에서는 직선 구간 및 곡선 구간에 대해 사고발생에 영향을 미치는 중요 변수가 RF 기법을 통해 선택된다. 본 연구에서는 MATLAB을 통해 코딩된 RF package를 사용하여 변수 중요도 분석을 수행하였다. RF 기법을 통한 변수 중요도 분석을 수행하기 위해서는 forest를 구성하는 트리의 개수와 각 노드에서 분할에 사용되는 독립변수의 수를 사전에 지정할 필요가 있다.

변수 중요도에 대한 안정적인 결과를 도출하기 위하여 트리의 개수를 변화시키면서 분석을 수행한 결과

Table 3. MSE error rates by number of variables

Number of variables	MSE error rates	
	Straight section	Curve section
3	1.320	0.505
6	1.224	0.501
12	1.258	0.529

400개의 트리부터 최소의 오차율이 일정하게 유지되는 패턴을 확인할 수 있었다. Liaw and Wiener(2002)에 따르면 트리 내 각 노드에서 분할을 위해 사용되는 변수의 수는 분석결과에 큰 영향을 미치지는 않는다고 언급하였다. 하지만 모형의 성능을 높이기 위하여 최적의 변수 수를 산정하는 방법론을 제시하였다. 알고리즘에서는 전체 변수의 1/3을 기본값으로 설정하고 있으며, 이에 기본값과 기본값의 1/2, 그리고 기본값의 2배를 대안으로 검토하여 최적의 오차율을 가지는 값을 선택할 것을 제시하였다. 본 연구에서 검토되는 전체 독립변수의 수는 17개이기 때문에 각 노드에서 사용되는 최적의 변수 수 산정을 위하여 6개, 3개, 12개의 대안을 분석해 보았으며, 직선 구간과 곡선 구간 모두에 대해 6개의 변수를 사용하였을 때 최소의 오차율을 가지는 것으로 분석되었다(Table 3). 따라서 본 연구에서는 트리의 수는 400개로, 각 노드에서 사용되는 변수의 수는 6개로 설정하여 RF 분석을 수행하였다.

이에 따라 직선 구간과 곡선 구간에 대해 RF 기법을 통한 변수 중요도 분석 결과는 Figure 3과 같다. 본 연구에서는 중요도가 가장 높은 EXPO 변수를 기준으로 표준화한 척도를 사용하였으며, 변수 중요도가 높을수록 해당 변수가 사고예측에 사용되었을 경우 모형의 정확도를 더 크게 향상시킬 수 있음을 의미한다. 이에 따라 직선 구간과 곡선 구간 모두 AADT와 구간 길이로부터 산출되는 노출도 변수와 전체 교통량 중 버스 및 트럭 등 중차량이 차지하는 비율이 사고 발생에 가장 중요한 영향을 미치는 것으로 분석되었다. 또한 도로의 기하구조와 관련된 변수들이 사고 발생에 중요한 영향을 미치는 것으로 분석되었는데, 직선 구간의 경우 종단 선형 및 진출입부의 존재유무가, 곡선 구간의 경우 평면선형, 종단 선형, 그리고 진출입부의 존재유무가 사고 발생에 중요한 영향을 미치는 것으로 나타났다. 주변 환경 변수는 이에 비해 상대적으로 사고 발생에 영향을 미치는 중요도가 작은 것으로 나타났으며, 이 중 무인단속카메라의 영향이 비교적 높은 것으로 분석되었다.

유전자 프로그래밍 모형 구축에 사용할 최종 독립변수를 선택하기 위하여 본 연구에서는 중요도가 높은 변수부터 순서대로 입력변수로 활용하여 RF 모형의 MSE 변화를 검토하였다. 직선 구간과 곡선 구간 각각에 대해 상위 1개 변수부터 상위 17개 변수까지를 순서대로 검토해본 결과 직선 구간은 상위 7개 변수를 선택했을 때 최소의 MSE 값이 도출되었으며, 곡선 구간은 상위 8개 변수를 선택했을 때 최소 MSE 값이 도출되었다. 따라서 본 연구에서는 직선 구간의 경우 EXPO, HV, Slope, In, Out, E4, E2 변수를, 곡선 구간의 경우 EXPO, HV, Curve, Slope, In, E4, E6, Out 변수를 유전자 프로그래밍 모형 구축에 사용하였다.

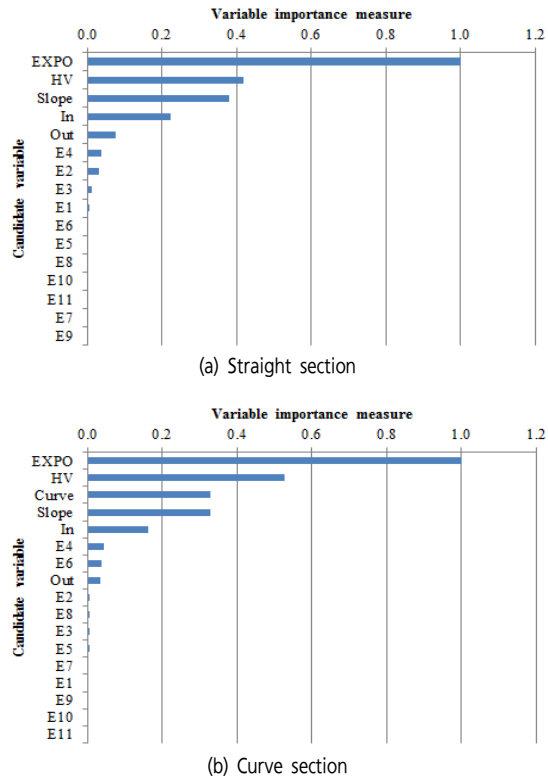


Figure 3. Normalized variable importance measure based on mean decrease in MSE

2. 유전자 프로그래밍 모형

본 연구에서는 고속도로의 직선 구간과 곡선 구간에 대해 앞서 선택된 중요 변수들을 기반으로 사고예측을 위한 유전자 프로그래밍 모형을 구축하였다. 유전자 프로그래밍 모형은 MATLAB으로 코딩된 GPLAB toolbox v3.0을 사용하여 구축되었으며, 유전자 프로그래밍 알고

Table 4. Summary of parameters

Parameter	Selected value
Number of generations	50
Number of individuals	1,000
Depth limited to	30
Initial maximum depth	6
Probability of crossover	automatic adaptation procedure
Probability of mutation	automatic adaptation procedure
Probability of reproduction	0
Selection	Lexictour
Function set	+, -, ×, ÷, √, ln
Terminal set	Selected variables

리즘 구현을 위해 본 연구에서 사용된 파라미터는 Table 4와 같다.

먼저 한 세대에 존재하는 개체군의 크기는 다양한 개체들을 생성시키기 위하여 충분히 큰 1,000개로 설정하였다. 또한 알고리즘의 종료조건으로 사용되는 세대수 파라미터는 50으로 설정하였다. 적합도 함수는 개선되지 않으면서 모형의 크기만 커지는 "bloat" 현상을 방지하기 위해 트리의 깊이를 30으로 한정시켰다.

본 연구에서는 부모 세대 개체의 선택 방법론으로 lexictour 방법을 사용하였다. 이 방법은 개체군 중에서 임의의 개체를 선택하여 이 중 가장 좋은 적합도를 가지는 개체를 최종적으로 선택하는 방법으로, 동일한 적합도를 가지는 개체에 대해서는 노드수가 적은 개체를 선택하여 bloat 현상을 최소화한다(Silva, 2007). 모형 구축을 위한 유전 연산자는 새로운 개체의 출현에 가장 중요한 역할을 수행하는 교배와 돌연변이만을 사용하였으며, +, -, ×, ÷, 루트, 자연로그 등 6개의 표준적인 수학연산자가 사용되었다.

유전 연산자의 선택확률은 유전자 프로그래밍의 구현 과정에서 자동적으로 연산된다. 만약 해당 연산자가 모형의 적합도를 향상시킨다면 해당 연산자의 선택 확률값은 증가할 것이고, 반대의 경우라면 해당 연산자의 선택 확률값은 감소할 것이다. 알고리즘의 연산에 사용되는 변수는

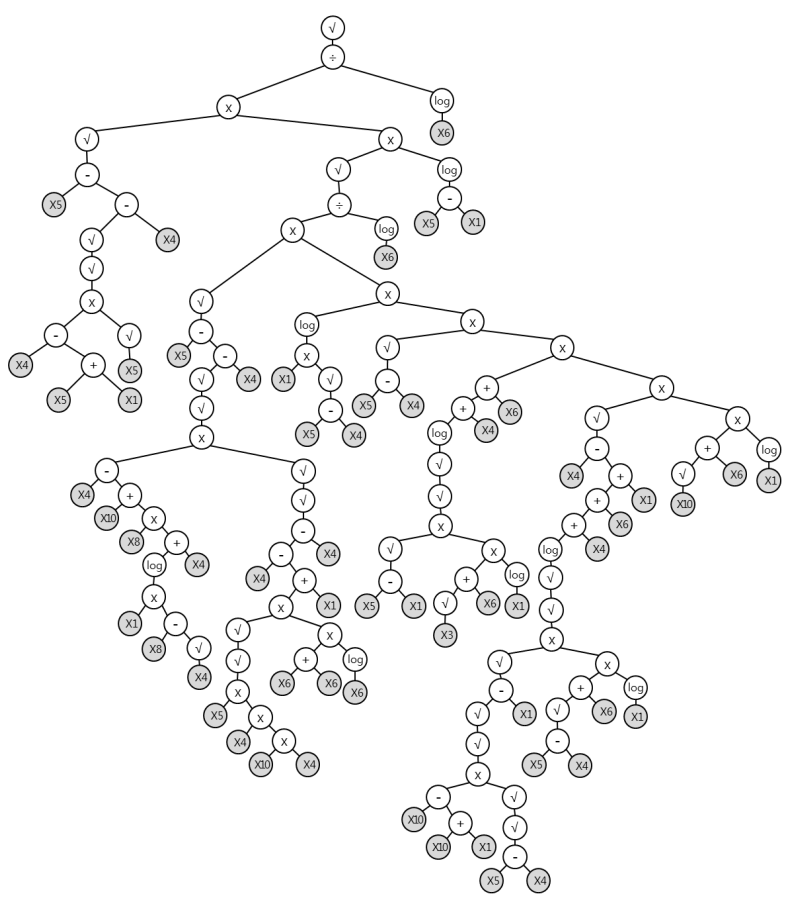


Figure 4. Genetic programming model for the straight section

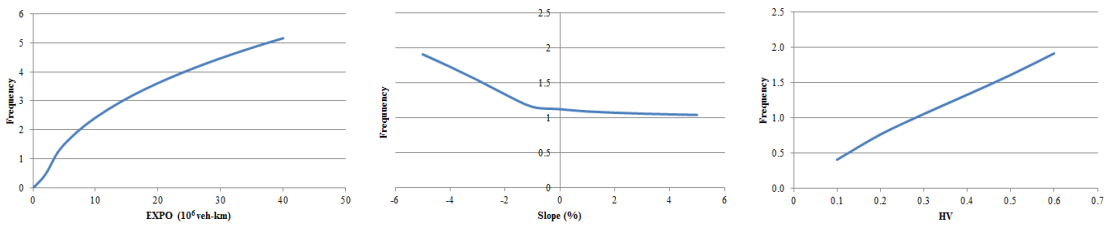


Figure 5. Relationship between crash frequency and continuous variables on straight section

앞서 RF 분석을 통해 선택된 변수들이 사용되며, 적합도 함수로는 회귀 문제에 효과적인 RMSE가 사용된다.

1) 직선 구간 사고예측모형

고속도로 직선 구간에 대해 도출된 유전자 프로그래밍 모형은 Figure 4와 같다. 앞서 설명한 바와 같이 Figure 4는 선택된 6개의 수학연산자와 사고빈도수에 주요한 영향을 미치는 설명변수들이 조합된 수식을 나타내는 트리로 표현된다. 노출도 변수와 종단선형의 기울기, 중차량 비율, 진출입부 존재 유무, 무인단속카메라 및 노면요철포장 변수가 직선 구간 모형에서 사고 발생에 주요한 영향을 미치는 설명변수로 사용되었다. 유전자 프로그래밍 모형은 사고 빈도수와 설명변수들간의 복잡한 관계를 표현한다. 따라서 본 연구에서는 각 설명변수의 변화에 따라 사고 빈도수가 어떻게 변화하는지를 살펴보기 위하여 해당 변수 외 다른 변수들은 샘플 평균값으로 고정하고 해당 변수를 현실적인 범위 내에서 변화시키면서 사고 빈도수에 미치는 영향을 분석하였다.

직선 구간 모형에서 연속변수인 노출도, 종단선형 기울기, 중차량 비율 변수에 대한 분석 결과는 Figure 5와 같다. 노출도가 증가할수록 사고 빈도수 역시 증가하는 것으로 나타났으며, 종단선형의 기울기는 내리막 경사가 급해질수록 사고 빈도수가 증가하지만, 오르막 경사의 경우 사고 빈도수에 큰 영향을 미치지 않는 것으로 분석되었다. 또한 중차량 비율이 증가할수록 사고 빈도수가 증가하는 것으로 분석되었다.

이진변수인 진출입부 존재 유무, 무인단속카메라 및 노면요철포장 변수에 대한 분석 결과는 Table 5에 나와 있다. 연속변수와 마찬가지로 해당변수 외 다른 변수들은 샘플 평균값으로 고정하고 해당 이진변수만을 0과 1로 변화시키면서 해당 시설물의 존재유무에 따른 사고빈도수 및 변화량을 분석하였다. 분석 결과 진출입부의 존재는 사고 빈도수를 증가시키는 요인으로 나타났으며, 무인단속카메라와 노면요철포장 시설은 사고를 감소시키는 요인으로 분석되었다.

Table 5. Relationship between crash frequency and binary variables on straight section

variable	0	1	variation
In	0.225	0.981	+0.756
Out	0.510	0.611	+0.101
E4	0.720	0.472	-0.248
E2	0.613	0.552	-0.060

유전자 프로그래밍 모형의 적합도를 비교하기 위하여 본 연구에서는 동일한 자료를 활용하여 음이항 회귀모형을 구축하였으며, 직선구간에 대해 구축된 회귀모형은 식(2)와 같다. 유의수준 0.05 하에서 유전자 프로그래밍 모형과 동일한 변수들이 유의성을 가지는 것으로 나타났으며, 각 설명변수 별로 사고빈도수의 증가와 감소에 미치는 영향은 유전자 프로그래밍 모형의 결과와 유사한 패턴을 가지는 것으로 분석되었다.

$$ACC_s = X1 \cdot \exp\begin{pmatrix} -3.481 - 0.035X3 + 1.301X4 + \\ 1.036X5 + 3.840X6 - 0.278X8 - \\ 0.471X10 \end{pmatrix} \quad (2)$$

두 모형간 비교를 위한 비교 지표로는 모형에 의해 도출된 사고 빈도수와 실제 사고 빈도수와의 차이를 나타내는 RMSE를 사용하였다. 본 연구에서 구축된 직선 구간의 유전자 프로그래밍 모형의 RMSE는 0.573으로 나타났으며, 음이항 회귀모형의 RMSE는 0.688로 나타나 유전자 프로그래밍 모형의 적합도가 더 우수한 것으로 분석되었다.

이처럼 본 연구에서는 유전자 프로그래밍 모형에 의해 도출되는 함수식을 기반으로 사고 빈도수와 설명변수 사이의 관계에 대한 도식화를 통해 사고 요인에 대한 규명이 가능하고, 음이항 회귀모형에 비해 상대적으로 우수한 적합도를 가지는 사고예측모형을 구축할 수 있었다.

2) 곡선 구간 사고예측모형

고속도로 곡선 구간에 대해 구축된 유전자 프로그래밍 모형이 Figure 6에 나와 있다. 곡선 구간의 경우 앞서 설정한 6개의 수학연산자와 노출도, 평면선형의 곡선

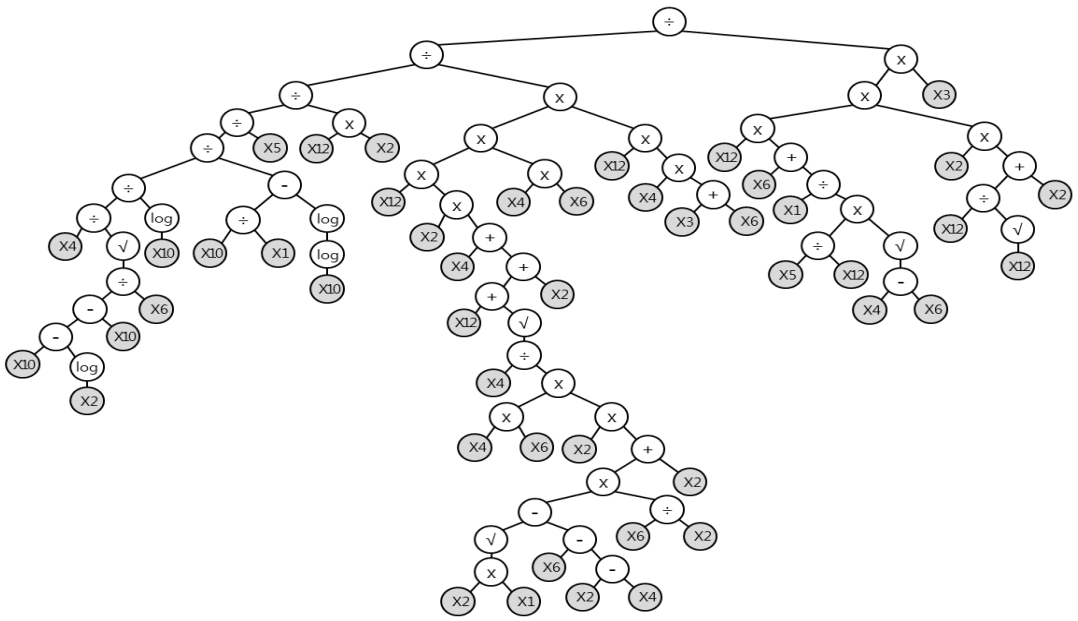


Figure 6. Genetic programming model for the curve section

반경, 종단선형의 기울기, 중차량 비율, 진출입부 존재 유무, 무인단속카메라, 시선유도봉 등 8개의 설명변수의 조합으로 표현되는 트리 모형이 사고빈도수를 예측하기 위해 구축되었다. 직선 구간과 마찬가지로 곡선 구간에 대해 구축된 유전자 프로그래밍 모형을 통해 설명변수와 사고 빈도수와의 관계를 도출하였다.

곡선 구간 모형에서 사용된 연속변수는 노출도, 평면선형의 곡선반경, 종단선형의 기울기, 중차량 비율 등 네 개의 변수이며, 이에 대한 분석결과는 Figure 7과 같다. 직선 구간과 마찬가지로 노출도가 증가할수록 사고 빈도수 역시 증가하는 것으로 나타났으며, 평면선형의 곡선반경이 증가할수록 사고빈도수가 감소하는 것으로 나타났다. 종

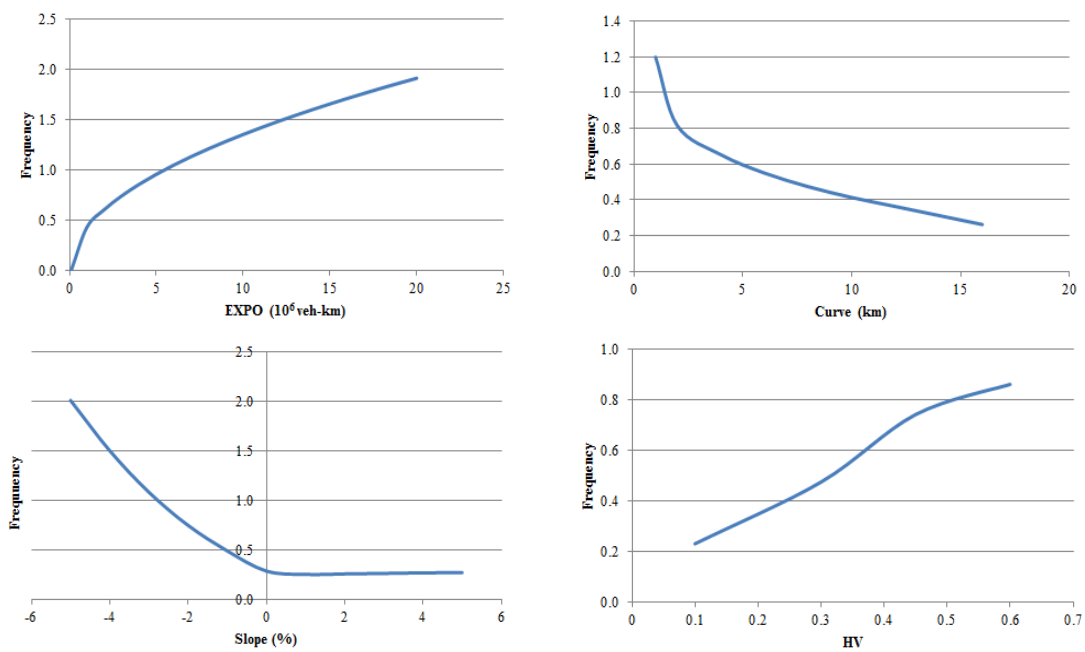


Figure 7. Relationship between crash frequency and continuous variables on curve section

Table 6. Relationship between crash frequency and binary variables on curve section

variable	0	1	variation
In	0.175	0.941	+0.766
Out	0.353	0.807	+0.454
E4	0.363	0.332	-0.032
E6	0.353	0.339	-0.014

단선형의 기울기는 내리막 경사의 기울기가 클수록 사고빈도수가 증가하는 것으로 나타났으며, 오르막 경사의 경우 사고빈도수에 큰 영향을 미치지 않는 것으로 분석되었다. 직선 구간과 마찬가지로 곡선 구간에서도 중차량 비율이 증가할수록 사고 빈도수가 증가하는 것으로 분석되었다.

곡선 구간 모형에서 사용된 이진변수인 진출입부 존재 유무, 무인단속카메라 및 시선유도봉 변수에 대한 분석 결과는 Table 6에 나와 있다. 앞서 설명한 바와 같이 이진변수의 존재 유무에 따라 0과 1로 변화시키면서 사고빈도수 및 변화량을 분석하였으며, 분석 결과 진출입부의 존재는 직선구간과 마찬가지로 사고 빈도수를 증가시키는 요인으로 나타났으며, 무인단속카메라와 시선유도봉 시설은 사고를 감소시키는 요인으로 분석되었다.

유전자 프로그래밍 모형의 적합도를 비교하기 위하여 곡선 구간에 대한 음이항 회귀모형을 구축하였으며, 이는 식(3)과 같다. 유의수준 0.05 하에서 시선유도봉을 제외한 대부분의 변수들이 유의성을 가지는 것으로 나타났다. 시선유도봉의 경우, 회귀모형 추정 결과 비상식적인 부호를 가지는 동시에 통계적으로도 유의하지 않은 것으로 분석되었다. 다른 설명변수의 경우, 사고빈도수의 증가와 감소에 미치는 영향은 유전자 프로그래밍 모형의 결과와 유사한 패턴을 가지는 것으로 분석되었다.

$$ACC_c = X1 \cdot \exp\begin{pmatrix} -4.051 - 0.012X2 - 0.135X3 + \\ 1.231X4 + 1.095X5 + 4.991X6 - \\ 0.182X10 \end{pmatrix} \quad (3)$$

그 결과 본 연구에서 구축된 곡선 구간의 유전자 프로그래밍 모형의 RMSE는 0.326으로 나타났으며, 음이항 회귀모형의 RMSE는 0.445로 나타나 유전자 프로그래밍 모형의 적합도가 더 우수한 것으로 분석되었다.

결론 및 향후 연구

본 연구에서는 고속도로의 사고 예측을 위한 유전자 프로그래밍 모형의 적용성을 조사하였다. 한국의 경부고속도로에 대해 3년(2010-2012년) 동안 구득된 사고빈도수, 노출도, 기하구조, 교통류 특성, 주변 환경 관련 변수를 기반으로 분석을 수행하였다. 고속도로의 직선 구간과 곡선 구간에 대해 사고 발생에 영향을 미치는 중요한 변수를 선택하기 위하여 RF 기법을 통한 변수 중요도 분석을 수행하였다. RF에 의해 선택된 변수들에 기초하여 직선 및 곡선 구간에 대해 사고예측을 위한 유전자 프로그래밍 모형을 구축하였다.

직선 구간의 경우 노출도 및 중차량 비율, 종단선형의 기울기, 진출입부의 존재 유무, 무인단속카메라 및 노면요철포장 변수가 사고 빈도수에 영향을 미치는 것으로 나타났으며, 곡선 구간의 경우 노출도, 평면 선형의 곡선반경, 종단선형의 기울기, 중차량 비율, 진출입부의 존재 유무, 무인단속카메라, 시선유도봉 변수가 사고 빈도수에 중요한 영향을 미치는 것으로 분석되었다. 이들 변수와 사고 빈도수와의 관계를 도출해본 결과, 노출도가 증가할수록, 내리막 기울기의 절대값이 증가할수록, 곡선반경이 감소할수록, 중차량 비율이 높아질수록 사고 빈도수가 증가하는 것으로 나타났으며, 진출입부의 존재는 사고 빈도수를 증가시키고, 무인단속카메라와 노면요철포장, 그리고 시선유도봉 시설 등은 사고 빈도수를 감소시키는 요인으로 분석되었다.

본 연구 결과는 사고 요인 규명과 사고 빈도수의 정확한 예측을 통한 도로 안전도 평가, 그리고 도로안전사업에 대한 투자 우선순위 결정을 위한 의사결정에 사용될 수 있을 것이다. 도로의 계획 및 설계 측면에서는 기하구조 및 시설물 정보를 활용한 도로의 안전도 평가가 가능하며, 도로의 운영 측면에서 사고 발생 위험이 높은 지점의 사고 감소를 위하여 무인단속카메라, 노면요철포장 및 시선유도봉 시설 설치, 그리고 중차량 분리 운영 등의 대안을 제시할 수 있을 것이다.

본 연구에서 구축된 유전자 프로그래밍 모형의 적합도를 비교하기 위하여 동일한 자료를 활용하여 음이항 회귀모형을 구축하였다. 유전자 프로그래밍 모형과 음이항 회귀모형의 RMSE를 비교해본 결과 직선 구간과 곡선 구간 모두에 대해 유전자 프로그래밍 모형의 적합도가 높은 것으로 분석되었다. 즉, 본 연구에서 제시한 유전자 프로그래밍 기법을 사고예측모형 구축에 도입함으로써 설명변수와 사고 빈도수와의 관계를 규명하는 동시에 모형의 적합도를 향상시킬 수 있었다.

하지만 본 연구에서 제안된 방법이 실제적으로 사용되기 위해서는 몇 가지 추가적인 연구가 필요하다. 우선, 경부고속도로 이외에 다른 고속도로에서 수집된 자료를

이용하여 유전자 프로그래밍 모형의 공간적 전이성을 검토해볼 필요가 있다. 또한 본 연구에서는 유전자 프로그래밍 모형의 적합도를 비교하기 위하여 전통적인 회귀분석 모형을 비교대상으로 사용하였다. 하지만 본 연구에서 제시된 방법론의 우수성을 강조하기 위하여 회귀모형에 비해 모형의 적합도가 상대적으로 더 높은 것으로 알려진 신경망이나 SVM 같은 비모수적 기법과의 추가적인 성능 비교가 이루어질 필요가 있다. 그리고 통계적 회귀모형은 동일한 분석 자료에 대해 일관된 모형 추정 결과를 도출하지만, 유전자 프로그래밍 모형과 같은 학습기반의 분석법의 경우 모형 구축 결과에 차이가 있을 수 있으며, 이에 따라 최적 결과 도출을 위한 파라미터 산정에 대한 추가적인 연구가 필요하다. 마지막으로 사고발생에 영향을 미치는 추가적인 변수 및 보다 자세한 사고특성 자료 구득을 통해 교통사고에 영향을 미치는 추가적인 요인 규명 및 보다 세밀한 모형 구축이 필요할 것으로 판단된다. 이러한 추가 연구가 이루어진다면 본 연구에서 제시된 방법론을 통해 보다 안전한 고속도로의 설계 및 운영이 가능해질 것이다.

REFERENCES

Abdel-Aty M. A., Radwan A. E. (2000), Modeling Traffic Accident Occurrence and Involvement, Accid. Anal. Prev., 32(5), 633-642.

Breiman L. (2001), Random Forests, Mach. Learn., 45(1), 5-32.

Han S., Kim K., Oh S. (2008), What Goes Problematic in the Existing Accident Prediction Models and How to Make It Better. J. Korean Soc. Road Eng., 10(1), 19-29.

Hauer E. (2004), Statistical Road Safety Modeling, TRR, 1987, TRB, 81-87.

Kang J. G., Lee S. H. (2002), Traffic Accident Prediction Model by Freeway Geometric Types, J. Korean Soc. Transp., 20(4), Korean Society of Transportation, 163-175.

Kang M. W., Doh T. W., Son B. S. (2002), Fitting Distribution of Accident Frequency of Freeway Horizontal Curve Sections & Development of Negative Binomial Regression Models, J. Korean Soc. Transp., 20(7), Korean Society of Transportation, 197-204.

Kononov J., Bailey B., Allery B. K. (2008), Relationships Between Safety and Both Congestion and Number of Lanes on Urban Freeways, TRR, 2083, TRB, 26-39.

Kononov J., Lyon C., Allery B. K. (2011), Relation of Flow, Speed, and Density of Urban Freeways to Functional Form of a Safety Performance Function, TRR, 2236, TRB, 11-19.

Koza J. R. (1992), Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press (Cambridge, MA, USA), 73.

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.

Liaw A., Wiener M. (2002), Classification and Regression by randomForest, R news, 2(3), 18-22.

Lord D., Washington S. P., Ivan J. N. (2005), Poisson, Poisson-gamma and Zero-inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory, Accid. Anal. Prev., 37(1), 35-46.

Park H. S., Son B. S., Kim H. J. (2007), Development of Accident Prediction Models for Freeway Interchange Ramps, J. Korean Soc. Transp., 25(3), Korean Society of Transportation, 123-135.

Silva S. (2007), GPLAB: A Genetic Programming Toolbox for MATLAB, Mathworks (Natick, MA, USA), 10.

Zhong L., Sun X., Yulong H., Zhong X., Chen Y. (2009), Safety Performance Function for Freeway in China, 88th Annual Meeting of the TRB, Washington D.C.

♧ 주 작 성 자 : 곽호찬
♧ 교 신 저 자 : 이청원
♧ 논문투고일 : 2013. 12. 25
♧ 논문심사일 : 2014. 2. 17 (1차)
2014. 5. 8 (2차)
2014. 5. 21 (3차)
♧ 심사판정일 : 2014. 5. 21
♧ 반론접수기한 : 2014. 12. 31
♧ 3인 익명 심사필
♧ 1인 abstract 교정필