한국ITS학회논문지 J. Korea Inst. Intell. Transp. Syst.

돌발상황 처리시간 예측을 위한 영향요인 분석 및 SMOGN-DNN 모델 개발

Analysis of Incident Impact Factors and Development of SMOGN-DNN Model for Prediction of Incident Clearance Time

윤 규 리* · 배 상 훈**

* 주저자 : 부경대학교 공간정보시스템공학과 석사과정
** 교신저자 : 부경대학교 공간정보시스템공학과 교수

Gyu Ri Yun* · Sang Hoon Bae**

* Master's Student, Dept. of Spatial Information Eng., Pukyong National University
** Professor, Dept. of Spatial Information Eng., Pukyong National University

† Corresponding author : Sang Hoon Bae, sbae@pknu.ac.kr

Vol.20 No.4(2021)
August, 2021
pp.46~56

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2021.20.4.46

Received 22 June 2021
Revised 12 July 2021
Accepted 6 August 2021

요 약

돌발상황으로 인한 비반복정체로 발생하는 높은 교통비용과 혼잡을 효과적으로 해소하기 위해서 돌발상황 처리시간을 예측하는 것은 중요하다. 본 연구에서는 인공신경망을 활용한 예측모델 개발을 위해 국내 도로상황에 적합한 돌발상황 처리시간 영향요인을 분석하고, 이를 학습데이터로 생성하였다. 기존 연구에서 장시간 소요되는 돌발상황 처리시간에 대한 과소 예측 문제가 발생하여 이에 대한 해결방안으로 본 연구에서는 SMOGN기법을 적용한 오버샘플링 학습데이터를 생성하여 이를 모델에 적용하였다. 그 결과 SMOGN기법을 적용한 DNN모델이 MAE 18.3분으로 연구 과정에서 구축된 모델 중 가장 높은 정확도로 돌발상황 처리시간을 예측하여, 기존에 개발된 예측모델의 한계점을 보완할 수 있을 것으로 기대한다.

핵심어 : 돌발상황 영향 요인 분석, 돌발상황 처리시간 예측, 인공신경망 모델, 도시 고속도로, 오버샘플링

ABSTRACT

Predicting the incident clearance time is important for eliminating the high transportation costs and congestion from non-repetitive congestion caused by incidents. In this study, the factors influencing the clearance time suitable for domestic road conditions were analyzed, using a training dataset for predicting the incident clearance time using artificial neural networks. In a previous study, the under-prediction problem for high incident clearance time was used. In the present study, over-sampling training data applied using the SMOGN technique was obtained and applied to the model as a solution. As a result, the DNN model applying the SMOGN technique could compensate for the limitations of the previously developed prediction model by predicting the clearance time with the highest accuracy among the models developed in the research process with MAE = 18.3 minutes.

Key words : Analysis of Incident Impact Factors, Prediction of Incident Clearance Time, Artificial neural network, City Highway, Over-Sampling

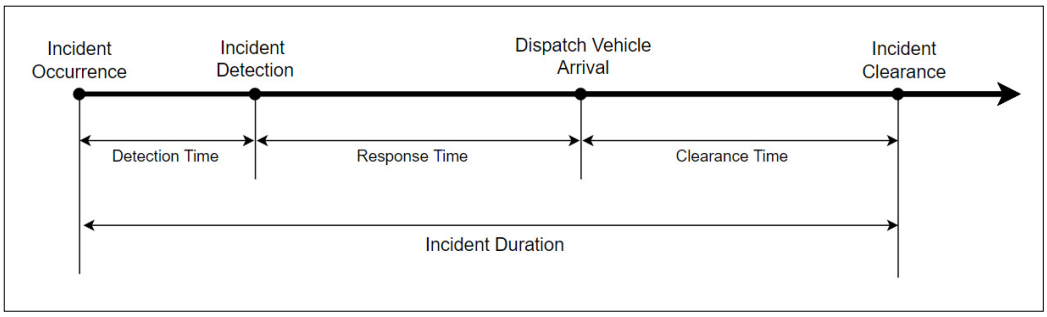
Ⅰ. 서 론

교통사고, 공사, 기상이변 등으로 인한 돌발상황은 도로 소통을 방해하여 차량 평균 속도를 급격하게 감소시키고, 밀도가 증가하면서 심각한 교통 혼잡을 일으킨다. 출퇴근 시간에 발생하는 반복적인 정체와는 달리 돌발상황은 예측 불가하게 발생하면서 비반복적인 정체를 일으킨다. 2020년도 기준으로 국내에서 발생한 교통사고는 총 209,654건이며, 최근 5년간 평균 20만 건 이상 꾸준하게 높은 통계치를 보이고 있다(KoROAD, 2021). 교통사고와 교통 혼잡으로 인해 발생한 국내 교통비용은 2017년 기준 약 290조로 조사되었으며, 전체 교통비용 중 약 35%를 차지한다(KOTI, 2019).

돌발상황이 발생할 경우, 도로가 차단되면서 도로 정체에 직접적인 영향을 끼침에 따라 돌발상황이 얼마나 지속될 것인가를 아는 것은 매우 중요하다. 이에 따라 2000년대 초기부터 관련 연구에서는 상관분석을 통해 유의한 관계에 있는 영향요인들을 파악하고, 이를 바탕으로 돌발상황 지속시간 예측 모형을 도출하였다. 과거의 연구에서는 회귀식의 통계적 모형을 주로 활용하였으며, 국내 데이터가 부족함에 따라 국외 데이터를 활용하거나 건수가 적은 국내 데이터를 활용하여 연구가 수행되었다. 최근에는 빅데이터 및 인공지능 기술의 발전에 따라 국내 데이터 수집 과정이 용이해져 대용량 국내 데이터 기반으로 의사결정모형과 KNN알고리즘을 이용한 연구가 수행되었으며, 이에 따라 과거의 연구보다 구체성과 정확성 측면에서 향상된 모형을 얻을 수 있었다.

하지만 최근 연구에서는 돌발상황을 등급 또는 유형으로 먼저 분류하여 모형을 도출하는 방식이 사용되었으며, 확률적으로 가능성이 가장 높은 예측 값을 도출하는 방식이 사용되면서 대형사고와 같이 빈도가 낮게 발생하는 긴 지속시간의 돌발상황에 대해서는 예측에 어려움이 있는 한계가 존재하였다. 돌발상황 지속시간이 길어질수록 교통 흐름에는 더 큰 영향을 미치기 때문에 기존 연구에서 이상치로 처리되었던 돌발상황들을 오히려 더 높은 정확도로 예측하는 것이 필요하다고 판단하였다. 이에 따라 본 연구에서는 돌발상황 지속시간 예측의 편의성과 정확성 향상을 위해 돌발상황에 대한 사전의 분류과정은 거치지 않고 모형을 구축하고자 하였으며, 특히 소수데이터로 분류되는 긴 지속시간에 대해 과소 예측되는 문제를 해결하는 것에 중점을 두었다.

돌발상황 지속시간은 <Fig. 1>과 같이 돌발상황 발생 시각부터 돌발상황을 검지하고, 이를 처리하기 위한 차량들이 출동한 후, 실제로 돌발상황 상황이 처리된 시각까지를 말한다. 현실에서는 돌발상황이 발생한 정확한 시각을 알기 어렵기 때문에 돌발상황이 검지된 시간부터 돌발상황을 처리 완료한 시간까지의 기간이 기록된다. 따라서 본 연구에서 얻어진 데이터에 따라 돌발상황 검지부터 처리완료까지의 돌발상황 처리시간을 활용하여 연구를 수행하였다.



<Fig. 1> Definition of Incident Duration

Ⅱ. 문헌고찰

초기의 돌발상황 지속시간 예측에 대한 국내 연구에서는 국내 데이터 획득의 어려움으로 미국 워싱턴의 돌발상황 데이터 395건을 사용하여 지속시간과 상관분석을 통해 주요 영향요인을 독립변수로 선정하고, SPSS 통계프로그램을 활용하여 다중회귀모형을 도출하여 R-Squared(결정계수) 0.882의 최적모형을 산출하였다(Han, 2001). 하지만 국내와 국외의 교통 시스템 및 도로 환경이 상이함에 따라 국외 데이터를 사용한 모형을 국내에 적용하는 것에는 한계가 존재하였으므로 이에 따라 국내 고속도로의 돌발상황 168건을 활용해 다중회귀모형을 도출한 연구가 수행되었다. MAE 7.7분의 오차로 우수한 모형을 도출하였으나 연구에 활용된 자료수가 매우 적어 세부적인 통계분석에 어려움과 실무에 적용할 경우의 정확성이 떨어질 가능성이 존재하였다(Shin and Kim, 2002).

하지만 이후 국내 교통시스템이 고도화되고, 인공지능이 발전하면서 총 960건의 서울도시고속도로 돌발 DB 자료를 기반으로 돌발상황 유형에 따라 분류된 Decision Tree모형이 개발되었으며(Kim, 2005), 국내 경부고속도로의 돌발상황 2,060건을 기반으로 사상자수에 따른 사고등급별 Decision Tree모형이 도출되었다(Ha, 2010). 사고 등급별로 지속시간을 분류한 또 다른 연구로 전국고속도로의 돌발상황 60,473건을 기반으로 세부요인별 가중치를 달리하여 가장 가까운 예측 값을 도출해내는 KNN알고리즘을 활용한 비모수모형이 제시되었다(Lee et al., 2015). 기존 연구에서는 폐쇄차로수, 돌발상황 유형, 돌발상황의 등급에 따라 지속시간에 큰 영향을 주는 것을 증명하였으며, 이를 중점적으로 지속시간을 분류하여 더 세부적이고, 정확한 모형을 도출하였다.

과거의 연구들에서 나타나는 공통적인 한계점으로 지속시간이 긴 대형사고에 대해서는 이력자료의 수가 부족함에 따라 도로에 큰 영향을 미치는 돌발상황임에도 불구하고 예측에 어려움이 존재하였다. 통계기반 모형의 예측 값은 확률적으로 정답일 가능성이 높은 값에 가까워지고, 이상치에서는 멀어지는 경향성을 가지기 때문에 오히려 어느 한 쪽으로 편향된 예측 값을 도출할 가능성이 높은 것으로 알려져 있다. 이에 따라 국외의 연구에서도 소수데이터가 고려되지 못한 편향된 예측 값을 도출하면서 과소 예측되는 문제가 발생하면서 향후 연구로 해결되어야할 중요한 과제임을 언급했다. 기존 연구의 유의미한 결과로는 ANN모델을 적용한 예측 모형이 평균적으로 가장 높은 정확도의 예측 값을 도출하는 것을 확인하였다(Ruimin et al., 2018). 또한, 기존 연구에서는 통계적 모형과 머신러닝 기법을 포함한 총 5가지 모델 중 ANN모델은 90분 이상의 긴 돌발상황 지속시간을 다른 모형에 비해 월등히 낮은 오차로 유일하게 예측 값을 도출하였다(Valenti et al., 2010).

<Table 1> Summary of Traffic incident duration prediction studies in South Korea

Paper	Data	The Number of Incidents	Method Category	Accuracy
Han, 2001	USA, Washington	395 incidents	Regression model	R-Squared : 0.882
Shin and Kim, 2002	Gyeongbu Expressway	168 incidents	Regression model	MAE : 7.7 minutes
Kim, 2005	Seoul Urban Highway	960 incidents	Decision Tree	Mean, IQR
Ha, 2010	Gyeongbu Expressway	2,060 incidents	Decision Tree	Original Value : 95 minutes Expected Value : 89.51±9.24 minutes
Lee et al., 2015	Korean Expressway	60,473 incidents	KNN Algorithm	MAE : 14 minutes

국내 연구에서 ANN모델이 적용된 경우가 없음에 따라 본 연구에서는 ANN모델을 적용해보는 것으로 유의미한 결과를 도출하고, 기존 연구와 달리 예측모형 구축에서 데이터 분류 과정을 거치지 않고도 더 넓은 범위의 지속시간에 대해 섬세한 예측 결과를 도출하는 것에 중점을 두었다. 또한, 자료부족에 따른 과소예측 문제를 보완하여 더 높은 정확도의 예측을 하기 위해서 데이터 불균형 문제에 보편적으로 활용되는 오버샘플링 기법을 적용하여 연구를 수행하였다.

Ⅲ. 연구 내용

본 연구에서는 모델 개발을 위한 기초 데이터로 돌발상황 처리시간과 돌발상황 특성을 포함한 데이터와 같은 시간적 범위의 속도 및 교통량 데이터를 수집하였다. 분석을 통해 예측모델의 학습에 필요한 변수들을 선정하고, 적합한 형태의 학습데이터를 생성하기 위해 전처리 과정을 수행하였다. 학습데이터 내의 높은 처리시간에 대한 과소 예측의 문제가 발생함에 따라 SMOGN기법을 적용한 데이터셋을 추가로 생성하였다. 이에 따라 ANN모델, DNN모델과 기본 학습데이터, SMOGN 데이터를 각각 적용하여 4가지 모델을 구축하고, 하이퍼파라미터 조절을 통해 최적모델과 최하모델을 선정하여, 각 모델 간 예측 정확도 평가를 수행하였다.

1. 연구 데이터

1) 수집 및 파악

본 연구의 데이터로 서울시설공단 교통정보처의 돌발상황 정보와 구간별 시간대별 평균 속도 및 교통량 정보가 수집되었다. 연구 데이터의 공간적 범위는 서울특별시 마포구에서부터 성동구까지 이어지는 도시 고속도로인 내부순환로이며 도로 왕복기준 총연장 약 42km에 10개의 IC로 이루어진 연속류 도로이다. 시간적 범위로는 획득된 두 데이터에서 일치하는 기간인 2014년 9월 4일부터 2019년 9월 23일까지의 돌발상황 정보를 연구데이터로 사용하였다.

본 연구에서 획득한 돌발상황 특성으로는 돌발상황 처리시간을 포함하여 돌발상황 시작일시, 돌발상황 종료일시, 돌발상황 발생 구간, 돌발상황 유형, 통제된 차선 정보 그리고 출동 차량 종류 및 대수이며, 돌발상황 발생 시의 소통상태 특성에 대한 평균 속도 및 교통량 정보를 수집하였다. 여러 종류의 데이터를 융합해서 사용하는 것이 적절한 예측 값을 도출할 가능성이 높음에 따라 해당 데이터에서 얻을 수 없었던 돌발상황 발생도로의 기하구조 특성, 날씨 등의 외부 데이터를 추가적으로 수집하였다. 기하구조 특성에 대해서는 위성지도를 통해 총 차선 수, 진입/진출로 여부, 일반도로/터널 여부, 진입로와 떨어진 구간 수에 대한 정보를 획득하였으며, 환경 특성에 대해서는 기상청의 공공데이터를 활용하여 연구데이터의 시간적 범위와 일치하는 강수량, 적설량 등의 날씨 데이터를 수집하였다.

2) 전처리 및 분석

돌발상황 정보에는 연속적인 숫자 값이 아닌 문자형 데이터와 범주형 데이터(Categorical Data)가 대부분이며 날짜와 같은 순서형 데이터(Ordinal Data)도 존재한다. 해당 값들을 컴퓨터가 이해할 수 있도록 의미 있는 숫자로 변환하는 과정은 필수적이다. 이에 따라 본 연구에서는 각 데이터 특성을 분석하여 의미 있는 값을 부여해주고, 컴퓨터가 이해할 수 있도록 인코딩(Encoding) 과정을 수행하였다.

먼저, 발생일시, 날씨와 같은 학습에 어려운 값들은 요일은 주말과 평일로 구분하고, 시간은 첨두시, 비첨

두시로 구분하였으며, 날씨는 맑음, 비, 눈으로 범주를 나누어주었다. 속도와 교통량의 경우에는 돌발상황이 발생한 도로의 정보와 상행도로의 정보를 칼럼으로 생성하였으며, 속도 정보를 활용하여 30km/h미만은 정체, 30km/h이상 50km/h미만은 지체, 50km/h이상은 원활로 구분한 소통상태 칼럼을 생성하였다. 이외에도 기하구조에 대한 변수로 진입로, 진출로, 터널인지에 대해 나타내주는 칼럼과 해당 도로가 진입로로부터 얼마나 떨어져있는지에 대한 칼럼을 생성하였다.

생성된 칼럼들의 값은 <Table 2>에 정리된 것과 같이 각 데이터 특성에 적합한 인코딩 방식을 적용하여 모든 데이터를 수치형으로 변환하는 과정을 수행하였다. 범주형 데이터 변환에 가장 보편적으로 사용되는 원-핫 인코딩(One-Hot Encoding)기법은 범주가 많아질수록 희소데이터(sparse data)가 되는 단점이 있어 도로유형 칼럼에만 이를 적용하였다. 나머지 칼럼에는 각 범주마다 순위를 매길 수 있을 때 자주 사용되는 레이블 인코딩(Label Encoding)방법을 적용해주었다.

<Table 3>과 <Table 4>의 표와 같이 각 변수별 돌발상황 처리시간의 분포와 상관관계 분석에 따라 돌발상황 유형은 다른 변수들에 비해 돌발상황 처리시간과 유의한 관계를 갖는 것을 확인 할 수 있다. 또한, 타겟 인코딩은 단순 분류방식이 아닌 종속변수와의 관계를 고려해 확률로 계산된 값으로 인코딩하는 방식으로 데이터 차원 수는 증가시키지 않고, 신경망 모델이 더 쉽게 변수 간 관계를 학습할 수 있는 장점이 있어 이에 따라 돌발상황 유형 칼럼에는 더 유의한 가중치를 계산을 돕기 위해 타켓 인코딩(Target Encoding)방식을 적용하였다.

<Table 2> Columns for Predicting Incident Clearance Time

Feature	Column	Type	Unit	Encoding Method
Clearance Time	Clearance Time	Numeric	minute	-
Lane Closures	Closure_Lane	Categorical	1~4	Label Encoding
Vehicles reach the site	police, patrol, ambulance, fire_truck, tow_truck, cleaning_car, etc	Numeric	Number of vehicles	-
	total_cars	Numeric	Number of vehicles	-
Day of Week	day_cat	Categorical	1 = weekday 2 = weekend	Label Encoding
Time of Day	hour_cat	Categorical	1 = Non-Congestion 2 = Congestion	Label Encoding
Incident Type	incident_type_target	Numeric	Average of Incident Clearance Time of Incident Type	Target Encoding
roadway type	confluence, fractionation, tennel	Categorical	0, 1	One-Hot Encoding
Distance from Approach Road	access	Numeric	Number of Section	-
Traffic Flow Conditions	speed	Numeric	km/h	-
	speed_upstream	Numeric	km/h	-
	volume	Numeric	vehicles per hour	-
	volume_upstream	Numeric	vehicles per hour	-
	traffic_status	Categorical	1 = smooth, 2 = Delay 3 = Congestion	Label Encoding
Environmental conditions	weather	Categorical	1 = sunny, 2 = rain 3 = snow	Label Encoding

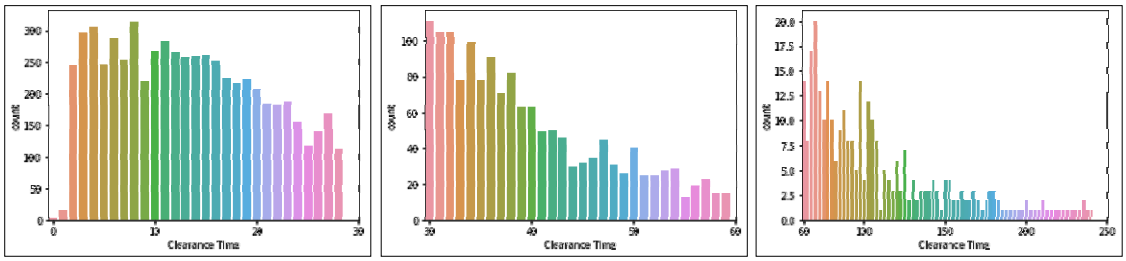
<Table 3> Summary of Incident Clearance Time by Features

Column	Features	count	sum	mean	std
Incident_Type	broken	3995	92389	23.126	18.333
	crash	3815	74214	19.453	15.387
	clash	133	4477	33.662	23.339
	falling	28	1838	65.643	55.465
	fire	16	772	48.25	35.311
	rollover	16	747	46.688	26.361
Lane Closures	Lane 3	4282	97545	22.780	18.465
	Lane 2	1315	25561	19.438	15.223
	Lane 1	2201	44131	20.050	15.950
	2 Lanes	205	7200	35.122	29.100
traffic_status	Congestion	1456	28599	19.642	15.013
	Delay	1969	40053	20.342	16.884
	smooth	4578	105785	23.107	18.949
day_cat	weekday	7141	156037	21.851	17.920
	weekend	862	18400	21.356	17.318
hour_cat	Non-Congestion	4360	99037	22.715	19.011
	Congestion	3643	75400	20.697	16.300
confluence	0	6194	130669	21.096	17.179
	1	1809	43768	24.195	19.818
fractionation	0	4113	85181	20.710	17.929
	1	3890	89256	22.945	17.708
tennel	0	4999	116515	23.308	18.318
	1	3004	57922	19.282	16.761
weather	sunny	7376	160646	21.780	17.660
	rain	335	7598	22.681	20.291
	snow	292	6193	21.209	19.737

<Table 4> Correlation of Independent Variables with Target Variable

Variables	total_cars	police	tow_truck	incident_type_target
Correlation	0.339	0.277	0.241	0.218
Variables	ambulance	cleaning_car	fire_truck	falling
Correlation	0.203	0.201	0.189	0.146
Variables	speed	clash	traffic_status	speed_upstream
Correlation	0.094	0.086	0.082	0.078
Variables	broken	confluence	fire	fractionation
Correlation	0.074	0.072	0.066	0.063
Variables	rollover	patrol	etc	incident_type
Correlation	0.062	0.055	0.048	0.018
Variables	weather	access	day_cat	day
Correlation	0	-0.005	-0.009	-0.013
Variables	Closure_Lane	hour	hour_cat	volume_upstream
Correlation	-0.019	-0.044	-0.056	-0.107
Variables	tennel	crash	volume	
Correlation	-0.109	-0.125	-0.136	

<Fig. 2>는 측 변수인 돌발상황 처리시간의 분포를 나타낸 그래프이다. 머신러닝을 활용하여 연속형 데이터를 예측할 경우, 정규분포를 따르는 데이터가 이상적이다. 하지만 특정 범위의 데이터가 부족함에 따라 한쪽으로 치우친 형태로 왜곡되어 있는 데이터의 경우 예측 정확도에 부정적인 영향을 끼칠 수 있다고 알려져 있다. <Fig. 2>와 같이 돌발상황 처리시간이 높아질수록 데이터 수가 급격히 희소해지는 형태를 보이며 정량적으로는 2.52의 높은 왜도 값이 측정됨에 따라 전반적인 데이터의 학습에 어려움이 있고, 예측 정확도가 낮아질 가능성이 있음을 확인하였다.



<Fig. 2> Distribution of Incident Clearance Time by Level

3) 데이터 오버샘플링

본 연구에서는 이력정보가 부족함에 따라 불균형해진 데이터 분포의 문제를 해결하기 위해 SMOGN (Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise)기법을 적용하여 값을 변형시키지 않으면서 소수 데이터가 증폭된 오버샘플링 데이터를 사용하였다(Paula et al., 2017). 해당 기법은 Gaussian Noise의 기본 원리에 따라 기존 데이터를 훼손시키지 않으면서 과대 추정될 수 있는 범위의 데이터는 축소시키고, 과소 추정될 수 있는 소수 데이터 범위의 데이터는 증폭시켜주는 오버샘플링 기법이다. 이는 파이썬(Python) 라이브러리 형태로 제공되며, smoter함수의 advanced mode를 사용하여 샘플링해줄 소수 데이터의 범위와 얼마나 많은 샘플링을 해줄 것인지에 대한 세부인자 값들을 수동으로 설정해주었다. 세부인자 설정 과정에서 생성된 오버샘플링 데이터의 왜도 값을 비교하면서 여러 번의 시행착오를 거쳤으며, 이에 따라 0.346으로 가장 낮은 왜도 값을 가지는 데이터가 선정되었다.

2. 돌발상황 처리시간 예측 모델 개발

본 연구에서는 기본 데이터와 SMOGN기법을 적용한 오버샘플링 데이터로 2가지 데이터셋을 생성하여 정확도 향상에 대한 효과를 확인하기 위해 오버샘플링을 하지 않은 데이터로 구축된 모델과 오버샘플링 데이터를 적용한 모델 평가와 모델 간의 비교를 시도하였다. 인공신경망 모델인 ANN(Artificial Neural Network)과 DNN(Deep Neural Network)모델을 구축하였으며 모델 평가에는 MAE(Mean Absolute Error), RMSE(Root Mean Square Error), R-Squared 지수를 사용하였다. 모델의 여러 가지 하이퍼파라미터 조절을 통해 예측 정확도가 높은 모델을 선정하기 위한 최적화 과정을 수행하였다. 연구에 사용된 언어는 파이썬(Python) 3.8.3버전이며, 2.4.1버전 텐서플로우(Tensorflow) 프레임워크의 Keras API가 모델 구축에 사용되었다.

1) 입력데이터 구축

입력데이터 구축 과정에서 데이터의 불균형으로 인해 예측신뢰도가 떨어지는 것을 방지하기 위해서 <Table 5>와 같이 30분 미만의 낮음, 30분 이상 60분 미만의 중간, 60분 이상의 높음으로 데이터를 분류하고, 각 단계별로 일정한 비율의 데이터를 추출해 결합한 데이터를 학습, 검증, 테스트과정에 적용하였다. 이는 앞서 설명한 오버샘플링 기법을 적용하지 않은 데이터와 적용한 데이터 모두에 대해 같은 방식으로 샘플링되었다. 오버샘플링 기법은 전체 데이터 중 학습 및 검증 데이터에만 적용하였으며 테스트 데이터는 실제 데이터 그대로 기본 모델과 오버샘플링 모델에 동일하게 사용되었다.

2) 인공 신경망 모델 구축

본 연구에서는 기본 데이터를 적용한 Simple-ANN모델과 Simple-DNN모델을 구축하고, SMOGN기법을 통해 생성한 오버샘플링 데이터를 적용한 SMOGN-ANN모델과 SMOGN-DNN모델을 구축하였다. ANN모델의 경우, [32, 64, 128, 256, 512, 1028, 2056]의 노드 수, [500, 1000, 2000, 3000]의 학습 횟수(epoch), [0.01, 0.001, 0.0001]의 학습률에 대한 경우의 수를 거쳤으며, DNN모델의 경우에는 [[32,32], [64,64], [128, 128], [256, 256], [512, 512], [1028, 1028]]의 노드 수, [500, 1000, 2000]의 학습 횟수, [0.01, 0.001, 0.0001]의 학습률, [0.01, 0.1, 0.2]의 드롭아웃에 대한 경우의 수를 거쳐 최적 하이퍼파라미터를 선정하였다. 구축된 모델에는 Adam 최적화 함수와 ReLu(Rectified linear unit) 활성화 함수를 사용하였다.

<Table 5> Shape of Dataset

dataset	clearance time level	raw dataset	dataset applied SMOGN
train dataset	low	5493	2362
	median	1301	1874
	high	316	2027
validation dataset	low	610	267
	median	145	210
	high	28	225
test dataset	low	62	62
	median	76	76
	high	32	32

<Table 6> Comparison of Worst&Best Models with Hyperparameter

Model		Dense	Epoch	Learning Rate	Dropout	MAE (min)	RMSE	R-Squared
Worst	Simple-ANN	[512]	3000	0.01	-	36.75	46.39	-2.36
	Simple-DNN	[128, 128]	2000	0.01	0.01	26.25	35.11	-0.93
	SMOGN-ANN	[128]	2000	0.01	-	31.33	41.08	-1.64
	SMOGN-DNN	[512, 512]	500	0.01	0.01	23.86	31.83	-0.58
Best	Simple-ANN	[32]	1000	0.0001	-	18.79	25.57	-0.02
	Simple-DNN	[64, 64]	500	0.001	0.1	17.97	24.53	0.06
	SMOGN-ANN	[128]	500	0.0001	-	18.72	24.29	0.08
	SMOGN-DNN	[64, 64]	500	0.0001	0.01	18.3	23.52	0.14

Ⅳ. 연구 결과

각 모델의 예측 정확도 분석을 위해 <Table 6>과 같이 가장 높은 오차 값을 가진 최하모델과 가장 낮은 오차 값을 가진 최고모델을 비교하였다. 모델의 정량적인 평가 지표로는 평균절대오차(Mean Absolute Error; MAE), 평균 제곱근 오차(Root Mean Square Error; RMSE), 결정계수(Coefficient of Determination; R-Squared)가 사용되었다. MAE가 보편적으로 사용되는 지표이지만 해당 오차 값은 관측 값을 기준으로 상대적인 값이 산정되기 때문에 관측 값이 낮을수록 오차 값은 더 크게 나오는 한계점이 존재한다. 이에 따라 MAE 지표만으로는 명확한 검증에 어려움이 존재하기 때문에 3가지 지표를 활용하여 모델 비교 및 평가를 수행하였다.

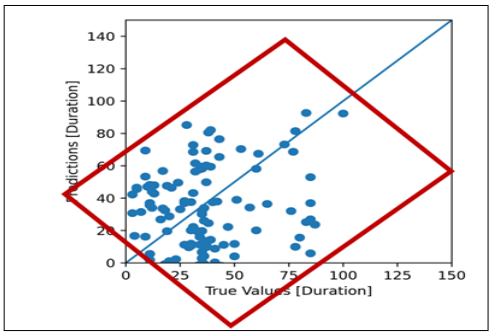
먼저 최하모델에서 오버샘플링 데이터도 적용하지 않고, Dense Layer가 1개인 Simple-ANN모델의 경우, 36.75분의 MAE가 계산되었으며, 오버샘플링 데이터를 적용하고, Dense Layer가 1개 더 추가된 SMOGN-DNN 모델의 경우에는 23.86분의 MAE가 도출되었다. MAE뿐만 아니라 RMSE와 R-Squared 값은 ANN 모델보다 DNN의 정확도가 월등히 높게 나왔으며, SMOGN기법을 적용한 오버샘플링 데이터로 학습된 모델의 경우 오버샘플링 데이터를 사용하지 않은 모델보다 훨씬 더 낮은 오차 값이 도출되었다. 또한, 최하모델과 최고모델의 하이퍼파라미터를 비교한 결과로 Dense layer의 노드 수, 학습 횟수, 학습률에 대한 입력 값이 높은 경우에 오히려 오차 값이 더 낮게 도출되는 경향이 확인되었다.

최고모델 간의 평가 지표에 대해서는 R-Squared는 최하모델의 분석과 비슷하게 모델이 업그레이드 될수록 값이 점점 높아지는 것이 확인되나 MAE와 RMSE 값의 경우에는 거의 차이가 없는 것을 볼 수 있다. 이에 대해 더 신뢰성 있는 모델 검증을 위해 <Fig. 3>~<Fig. 10>와 같이 최하모델과 최고모델에 대한 실제 데이터와 예측 데이터를 비교한 그래프를 통해 정성적 평가를 시도하였다.

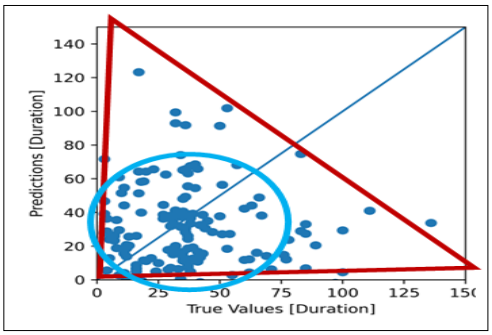
먼저 Simple-ANN모델의 최하모델에 대한 예측 그래프를 분석했을 때, 오버피팅되거나 어느 한쪽으로 편향되지 않으면서 부분적으로 매우 낮은 오차의 예측을 한 것으로 보인다. 하지만 매우 넓은 오차범위를 가지면서 모든 모델 중에서 가장 높은 MAE값인 36.75분이 도출되었다. SMOGN-ANN의 최하모델에서는 Simple-ANN모델에 비해 낮은 처리시간에 대해 예측 정확도가 증가한 것으로 확인되지만 높은 처리시간에 대해서는 매우 큰 오차로 예측 정확도가 매우 낮은 것을 볼 수 있다. Simple-DNN과 SMOGN-DNN모델 또한 이전 모델들과 시각적으로 보기에 크게 다르지 않은 결과가 나왔으나 낮은 처리시간에 대한 예측오차범위가

조금씩 좁아지는 것으로 보아 <Table 6>의 MAE값이 줄어드는 것과 일치하는 결과로 ANN모델보다 DNN모델을 적용할수록 정확도가 높아지지만 높은 처리시간에 대해서는 오버샘플링 데이터를 적용하였을 때의 효과를 확인하기엔 어려움이 있다.

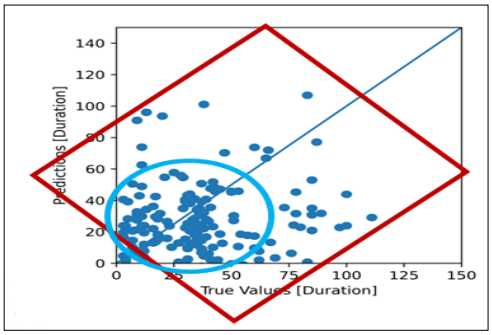
최고모델 간 비교 및 분석으로 Simple-ANN모델의 경우에는 정량적으로 더 낮은 오차를 가지면서 정성적으로도 낮은 처리시간에 대해 매우 높은 정확도를 가지는 것으로 보인다. 하지만 해당 모델에서 거의 모든 값들을 낮은 처리시간의 값으로 예측하는 것을 보아 오버피팅이 된 것으로 확인되며, 그럼에도 불구하고 정량적으로 오차가 낮은 이유는 대체로 낮은 값들에 의해 평가 지표 계산이 이루어진 까닭으로 평균 오차의 값도 함께 낮아진 것으로 유추할 수 있다. Simple-DNN모델의 경우에는 앞의 모델보다 더 높은 기울기로 예측 점들이 분포되었지만 여전히 낮은 값에 대해 오버피팅된 경향을 보이고 있어 앞선 모델과 정량적으로 오차 값에 큰 차이가 나타나지 않았다. 또한, SMOGN-ANN모델은 약간 높은 기울기로 예측이 이루어지면서 낮은 값에 대한 오버피팅 경향도 약해졌으나 매우 넓은 오차 범위를 가지고 있어 앞선 두 모델보다 더 낮은 정확도를 나타내고 있다. 이와 달리. SMOGN-DNN 모델의 경우에는 앞선 모든 모델들에 비해 낮은 처리시간에 오버피팅되지 않아 높은 처리시간에 대해서 대부분의 예측점이 회귀선과 가까이 분포하고, 이와 더불어 낮은 처리시간에 대해서도 넓지 않은 오차 범위를 가지는 것으로 나타난다. <Table 6>에 정리된 것처럼 최고모델의 평가 지표 중에서 각 모델의 MAE값은 매우 비슷한 수준으로 계산되었지만 RMSE값은 가장 낮고, R-Squared값이 가장 높다는 점과 더불어 <Fig. 10>의 그래프를 분석해본 결과, SMOGN-DNN모델이 다른 모델에 비해 훨씬 우수한 예측결과를 나타냈다고 판단할 수 있다.



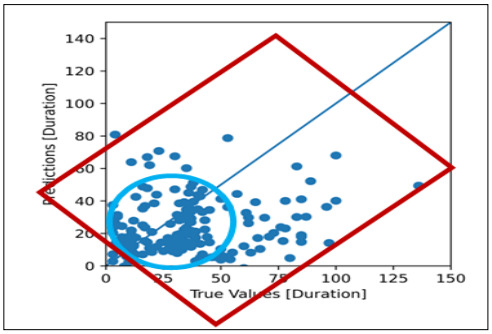
<Fig. 3> Prediction of Incident Clearance Time with Worst Simple-ANN Model



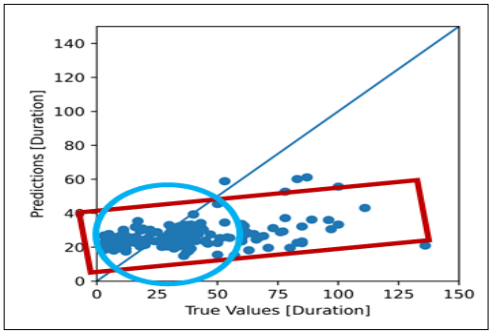
<Fig. 4> Prediction of Incident Clearance Time with Worst SMOGN-ANN Model



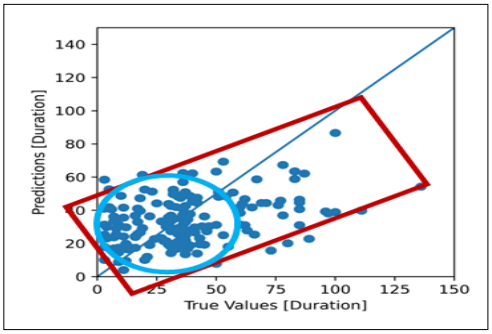
<Fig. 5> Prediction of Incident Clearance Time with Worst Simple-DNN Model



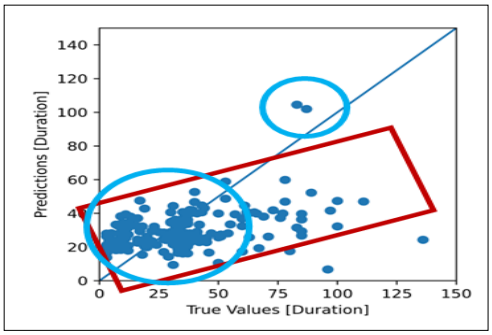
<Fig. 6> Prediction of Incident Clearance Time with Worst SMOGN-DNN Model



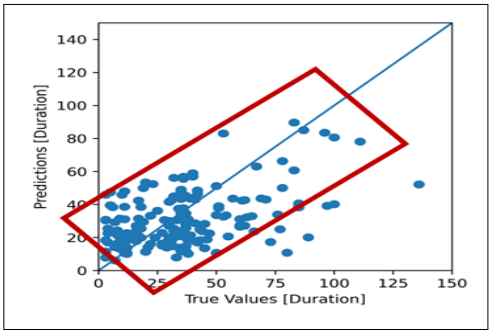
<Fig. 7> Prediction of Incident Clearance Time with Best Simple-ANN Model



<Fig. 8> Prediction of Incident Clearance Time with Best SMGON-ANN Model



<Fig. 9> Prediction of Incident Clearance Time with Best Simple-DNN Model



<Fig. 10> Prediction of Incident Clearance Time with Best SMOGN-DNN Model

V. 결 론

본 연구는 돌발상황 처리시간 예측 모델 개발을 위해 돌발상황 처리시간에 대한 영향 요인을 분석하여 총 14개의 요인을 통해 모델의 입력 데이터셋을 구축하였다. 또한, 기존 연구의 한계점인 장시간 소요되는 처리시간에 대해 과소 예측되는 문제에 대한 해결방안으로써 과대 추정의 가능성이 있는 데이터에 대해서는 축소시키고, 과소 추정의 가능성이 있는 데이터에 대해서는 증폭시켜주는 SMOGN기법을 적용하여 학습 및 검증 데이터에 오버샘플링된 데이터를 활용하여 모델을 구축하였다. 이에 따라 원본 데이터를 적용한 Simple-ANN모델, Simple-DNN모델과 오버샘플링 데이터로 구축된 SMOGN-ANN모델, SMOGN-DNN모델로 총 4가지 모델을 구축하였으며 각 모델에 대해 하이퍼파라미터를 조절하여 각 모델의 최하 및 최고 모델을 선정하였다. 기존 모델과 오버샘플링 데이터를 적용한 모델간의 비교 및 평가를 통해 본 연구에서 제시한 방법에 대한 효과를 검증하였다. 해당 예측 모델의 정량적 검증을 위해서는 MAE, RMSE, R-Squared의 3가지 평가 지표를 활용하였으며, 정성적인 평가로는 실제 값과 예측 값을 비교한 그래프를 통해 예측 정확도를 분석하고, 모델의 우수성을 검증하였다.

본 연구에서 도출된 유의미한 결과로 돌발상황 처리시간 예측에는 ANN모델보다 DNN모델이 더 낮은 평균 오차로 예측함에 따라 DNN모델이 우수한 모델임을 확인하였으며, 최적 모델 구축 과정에서 신경망 모델의 하이퍼파라미터인 Dense의 노드수, epoch수, 학습률(learning rate)은 값이 낮아질수록 예측 정확도는 높아

지는 경향성이 확인되었다. 또한, SMOGN기법이 적용된 오버샘플링 데이터로 학습된 모델의 경우, 기본 모델보다 정량적으로 월등히 더 높은 정확도로 예측되었다. 특히, 데이터가 밀집되어 있는 낮은 지속시간에 예측 값이 편향되지 않고, 높은 값에 대해서도 실제 값과 가까이 예측되는 것이 확인되었다. 본 연구의 결과에 따라 오버샘플링 기법을 적용하는 것은 기존 연구의 한계점이었던 과소 예측 문제를 해결하고, 좀 더 섬세한 지속시간 예측을 가능하게 하여 운전자들에게 최적 경로 및 통행시간 산정에 더 적절한 정보를 제공해줄 수 있을 것이라고 기대한다. 또한, 신경망기법을 활용한 예측모형은 기존 연구와 달리 다양한 돌발상황에 대해 추가적인 분류과정을 거치지 않고 학습된 모델로 편의성도 높을 것이라 판단된다.

하지만 국내 교통시스템 특성상 도로별로 관리기관이 상이함에 따라 여러 도로의 데이터 획득에 어려움이 있어 해당 연구에서는 국내 도시고속도로 중 내부순환로의 돌발상황 데이터만을 활용하였다. 이러한 한계점과 예측 모델의 신뢰성 향상을 위해 향후 다른 도로의 돌발상황 정보를 활용한 추가 연구가 필요하다. 또한 해당 모델이 긴 지속시간에 대해 높은 정확도로 예측함에도 불구하고, MAE기준 약 18분의 평균오차를 가지면서 아직 실무에 적용되기엔 여전히 큰 오차 범위라고 생각되어 더욱 정확한 예측을 할 수 있는 모델 개발에 추가적인 연구가 필요하며, 오버샘플링 데이터를 활용하여 신경망 모델이 아닌 다른 모델에 적용하였을 경우에도 정확도 향상에 효과가 있을지에 대해 이를 검증하기 위한 향후 연구가 필요하다.

REFERENCES

Ha O. K.(2010), “The Prediction Models for Clearance Times for the unexpected Incidences According to Traffic Accident Classifications in Highway,” *The Journal of the Korea Institute of Intelligent Transportation Systems*, vol. 9, no. 1, pp.101-110.

Han W. G.(2001), “A Model For Estimating Incident Duration,” *Korean Society of Civil Engineers*, vol. 21, no. 3-D, May.

Kim J. W.(2005), *Development of a model for estimating incident duration: Focusing on seoul urban expressway traffic management systems*, University of Seoul.

KoROAD(2021), *Traffic Accident Analysis System*, http://taas.koroad.or.kr

Lee K. Y. et al.(2012), “A Study on the Influencing Factors for Incident Duration Time by Expressway Accident,” *Korean Society of Road Engineers*, vol. 14, no. 1, pp.85-94.

Lee S. B., Han D. H. and Lee Y. I.(2015), “Development of Freeway Traffic Incident Clearance Time Prediction Model by Accident Level,” *J. Korean Soc. Transp.*, vol. 33, no. 5, pp.497-507.

Paula B., Luís T. and Rita P. R.(2017), “SMOGN: A Pre-processing Approach for Imbalanced Regression,” *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, PMLR, vol. 74, pp.36-50.

Ruimin L., Francisco C. P. and Moshe E. B. A.(2018), “Overview of traffic incident duration analysis and prediction,” *European Transport Research Review*, vol. 10, p.22.

Shin C. H. and Kim J. H.(2002), “Development of Freeway Incident Duration Prediction Models,” *Journal of Korean Society of Transportation*, vol. 20, no. 3, pp.17-30.

The Korea Transport Institute(2019), *Transportation cost calculation status and improvement plan*.

Valenti G., Lelli M. and Cucina D.(2010), “A Comparative study of models for the incident duration prediction,” *Eur Transp Res Rev.*, vol. 2, no. 2, pp.103-111.