



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

工學博士學位論文

딥 러닝을 이용한 고속도로 교통사고
예측모델 개발

Development of Expressway Traffic Accident Prediction Model
Using Deep Learning

亞洲大學校 大學院

建設交通工學科

柳 鍾 得

딥 러닝을 이용한 고속도로 교통사고 예측모델 개발

Development of Expressway Traffic Accident Prediction Model
Using Deep Learning

指導教授 尹 一 守

이 論文을 工學博士 學位 論文으로 提出함.

2018年 2月

亞洲大學校 大學院

建設交通工學科

柳 鍾 得

柳 鍾 得 의 工學博士 學位 論文을 認准함.

審査委員長 吳 世 昌 (印)

審 查 委 員 李 相 洙 (印)

審 查 委 員 柳 政 勳 (印)

審 查 委 員 趙 漢 宣 (印)

審 查 委 員 尹 一 守 (印)

亞洲大學校 大學院

2017年 12月

요 약 문

기존에는 대부분의 교통사고 자료 분석이 전통적인 통계적 방법인 포아송 회귀모형 또는 음이항 회귀모형 등을 기반으로 시행되어져 왔다. 이러한 통계적 방법은 교통사고와 관련된 다양한 인적, 도로 기하구조적 그리고 환경적 요인들과 교통사고 간의 인과관계를 찾고, 교통사고 빈도를 예측하고 그리고 분석된 결과를 바탕으로 교통안전 등급을 산출하는 등 다양한 방식으로 활용되어져 왔다. 하지만, 최근 머신 러닝 및 딥 러닝과 같은 빅데이터 분석 기법을 활용한 새로운 접근 방법들이 주목을 받기 시작하였다. 이러한 머신 러닝 및 딥 러닝 기법은 이종(異種)의 대량 자료를 활용하여 교통사고와 관련된 요인들을 분석하는 데 장점을 보이고 있으며, 이미 교통 및 다른 분야에서는 활발하게 적용되어 우리들의 일상을 변화시키고 있다. 이에 본 연구의 목적은 고속도로 교통사고 자료를 이용하여 고속도로의 주요 분석 단위인 콘존의 교통사고 빈도수를 예측하기 위하여 전통적인 통계적 기법과 딥 러닝을 이용한 기법을 적용하고 각 기법들의 예측 성능을 비교하였다. 예측 성능 비교 결과, 딥 러닝 모형의 MOE들이 전통적인 통계 모형에 비해 다소 우수한 것으로 나타났다. 하지만 MAD 기준으로 차이가 0.27로 전통적인 통계적 기법 기반으로도 교통사고 건수를 충분히 예측이 가능하다고 판단되며, 특히 음이항 회귀모형이 포아송 회귀모형보다 우수한 것으로 나타났다. 또한 노출 계수(AADT, 콘존 길이)를 활용하는 모형이 더욱 우수한 것으로 판단된다. 하지만 딥 러닝을 이용할 경우 예측 신뢰도를 더욱 증가 시킬 수 있다. 또한, 딥 러닝의 경우에는 아직 교통사고 건수 예측에 활용한 사례가 적기 때문에 적절한 구조 등에 대한 추가 연구가 필요하다. 특히 본 연구에서 은닉층 및 노드 개수뿐만 아니라 Optimizer, 노드 구조 등에도 많은 영향을 받는 것으로 확인 되었다.

목 차

제1장 서론	1
제1절 연구의 배경 및 목적	1
1. 연구의 배경	1
2. 연구의 목적	2
제2절 연구의 범위	3
제3절 연구의 수행절차 및 방법	4
제2장 관련 이론 및 연구 고찰	6
제1절 관련 이론 고찰	6
1. 안전성능함수	6
2. 포아송 회귀모형	9
3. 음이항 회귀모형	11
4. 인공 신경망	13
5. K-means 클러스터링	18
제2절 기존 연구 고찰	20
1. 전통적인 통계기법을 이용한 안전성능함수 구축 사례	20
2. Neural Network 및 딥 러닝을 이용한 안전성능함수	26
제3장 딥러닝을 이용한 고속도로 교통사고 예측모형 개발을 위한 자료 수집 및 분석	28
제1절 자료 수집 및 분석 개요	28

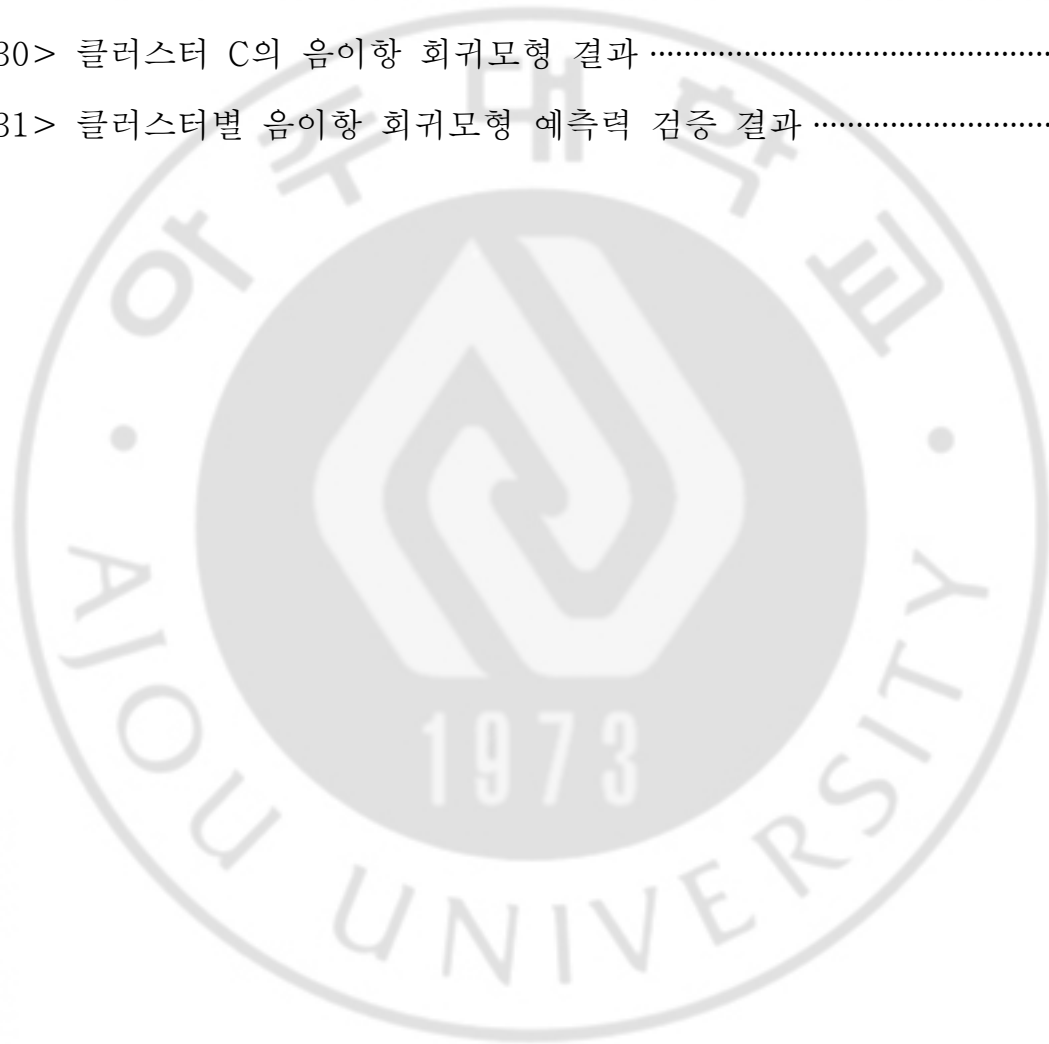
제2절 고속도로 교통사고 현황 분석	29
1. 고속도로 교통사고 추세 분석	29
2. 교통사고 등급별 발생 현황	31
3. 교통사고 위치별 발생 현황	32
제3절 변수 선정 및 자료 수집	34
제4절 수집 자료 가공	37
1. 자료 가공 단위 개요	37
2. 콘존 매칭을 통한 자료 가공	37
3. 공간 연산을 통한 자료 가공	38
제5절 분석 테이블 구축 및 기초 통계분석	39
1. 분석 테이블 구축	39
2. 기초 통계분석	40
제4장 고속도로 교통사고 예측모형 구축	41
제1절 고속도로 교통사고 예측모형 구축 방법론	41
1. 개요	41
2. 데이터 구분	42
3. 모형 검증 방법론	42
제2절 전통적인 통계 방법의 고속도로 교통사고 예측모형 구축	44
1. 구축 절차	44
2. 모형식 종류	45
3. 다중공선성 분석	46
4. 과분산 검정	47
5. 최종 안전성능함수 구축 및 검증	53

제3절 딥 러닝을 이용한 교통사고 예측모형 구축	59
1. 딥 러닝을 이용한 교통사고 예측모형 구축 배경	59
2. 딥 러닝 개요	61
3. 구축 절차	61
4. 딥 러닝을 이용한 교통사고 예측모형 구축	62
5. 최종 선정 모형 및 결과	71
6. 고속도로 교통사고 예측모형 비교	72
제4절 딥 러닝을 이용한 교통사고 건수 예측모형 구축 절차 및 활용 방안	74
1. 구축 절차	74
2. 활용 방안 및 기대효과	75
제5장 결론 및 향후 연구 과제	77
제1절 결론	77
제2절 연구의 한계 및 향후 연구과제	78
1. 연구의 한계	78
2. 향후 연구과제	79
참고문헌	80
부록 1. 클러스터링 기법을 이용한 고속도로 교통사고 예측모형 구축	85
1. 구축 배경 및 절차	85
2. 클러스터링 기법을 통한 자료 유형 구분	86
3. 안전성능함수 구축 및 검증	88
4. 시사점 도출	91

표 목 차

<표 1> 연구항목별 세부 연구내용	4
<표 2> 년도별 교통사고 발생 건수, 사망자 수 및 부상자 수	30
<표 3> 년도별 교통사고 등급별 사망사고 및 부상사고 현황	31
<표 4> 년도별 교통사고 위치별 사망자 현황	32
<표 5> 년도별 교통사고 위치별 부상자 현황	33
<표 6> 년도별 교통사고 위치별 치사율 현황	33
<표 7> 관련 연구 상의 독립변수 사용 내역	34
<표 8> 모형 구축을 위한 종속변수 및 독립변수 선정	36
<표 9> 수집 자료에 대한 기초 통계분석 결과	40
<표 10> 독립변수 다중공선성 분석 결과	46
<표 11> 포아송 회귀모형 A의 분석 결과	48
<표 12> 음이항 회귀모형 A의 분석 결과	49
<표 13> 포아송 회귀모형 B의 분석 결과	50
<표 14> 음이항 회귀모형 B의 분석 결과	51
<표 15> 음이항 회귀모형 A 도출 결과	53
<표 16> 음이항 회귀모형 B 도출 결과	54
<표 17> 음이항 회귀모형 검증 결과	55
<표 18> 딥 러닝 모형의 종류	63
<표 19> Gradient Descent 방법의 학습 및 테스트 Cost	66
<표 20> Adam 방법의 학습 및 테스트 Cost	66
<표 21> Adagrad 방법의 학습 및 테스트 Cost	67
<표 22> 각 Optimizer별 네트워크 구조	68
<표 23> 노드 구조 변경 및 최종 모형 선정	70

<표 24> 모형별 MOE 비교 결과	72
<표 25> 클러스터별 데이터 수	87
<표 26> 클러스터별 콘존길이 기초 통계	87
<표 27> 클러스터별 AADT 기초 통계	87
<표 28> 클러스터 A의 음이향 회귀모형 결과	88
<표 29> 클러스터 B의 음이향 회귀모형 결과	89
<표 30> 클러스터 C의 음이향 회귀모형 결과	90
<표 31> 클러스터별 음이향 회귀모형 예측력 검증 결과	90



그 림 목 차

<그림 1> 연구의 공간적 범위	3
<그림 2> 연구 수행절차	5
<그림 3> ANN 구조	14
<그림 4> 자료 수집 및 분석 절차	28
<그림 5> 연도별 교통사고 발생 건수, 사망자 수, 부상자 수	29
<그림 6> 노선별 교통사고 발생 건수	30
<그림 7> 교통사고 등급별 발생 현황	31
<그림 8> 교통사고 위치별 발생 현황	32
<그림 9> 콘존 마스터 예시	37
<그림 10> 콘존 매칭 코드	37
<그림 11> ArcGIS Spatial Join 결과	38
<그림 12> 분석 테이블 구축 결과	39
<그림 13> 데이터 구분 결과	42
<그림 14> 구축절차	44
<그림 15> 휴게소 개수와 교통사고 건수의 상관관계	56
<그림 16> 차로수와 교통사고 건수의 상관관계	56
<그림 17> 버스전용차로 유무와 교통사고 건수와 상관관계	57
<그림 18> 교량수와 교통사고 건수와 상관관계	57
<그림 19> 딥 러닝을 이용한 교통사고 예측모형 구축 절차	62
<그림 20> 학습 Cost 및 테스트 Cost 차이	68
<그림 21> 학습 Batch Size 및 Epoch별 R2	69
<그림 22> 테스트 Batch Size 및 Epoch별 R2	69
<그림 23> 최종 선정 모형	71
<그림 24> 모형 비교 및 검증	73
<그림 25> 통계기법 및 딥 러닝 기반 교통사고 건수 예측모형 구축 절차 ..	74
<그림 26> 클러스터링 기법을 이용한 안전성능함수 고도화 절차	85
<그림 27> 콘존길이 및 AADT 기준 클러스터링 결과 시각화	91

약 어 표

ANN(인공 신경망) : Artificial Neural Network

CDBN(합성곱 심층 신뢰 신경망) : Convolutional Deep Belief Network

CMF(사고보정계수) : Crash Modification Factor

CNN(합성곱 신경망) : Convolutional Neural Network

DBN(심층 신뢰 신경망) : Deep Belief Network

DNN(심층 신경망) : Deep Neural Network

HSM(도로안전매뉴얼) : Highway Safety Manual

LR(우도비 검정 통계량) : Likelihood Ratio

MAD(절대평균편차) : Mean Absolute Deviation

MLE(최대우도 추정법) : Maximum Likelihood Estimation

RBFNN(방사상 인공 신경망) : Radial Basis Function Neural Network

RBM(제한 볼츠만 머신) : Restricted Boltzmann Machine

RMSE(평균제곱오차의 제곱근) : Root Mean Square Error

SPF(안전성능함수) : Safety Performance Function

제1장 서론

제1절 연구의 배경 및 목적

1. 연구의 배경

2016년 우리나라에서는 총 220,917건의 교통사고가 발생하였고, 이로 인해 총 4,292명이 사망하고, 331,720명이 부상을 당하였다(경찰청, 2017). 이러한 수치는 2015년에 비해 교통사고 건수, 사망자수 그리고 부상자수가 각각 4.8%, 7.2%, 5.4%가 감소하였고 13년 만에 교통사고 사망자가 절반 이하로 줄어드는 등 괄목할만한 발전을 이루었다. 하지만 여전히 다른 국가에 비하여 높은 수준이다. 2016년 자동차 1만대 당 교통사고 사망자 수는 1.68명으로, OECD 회원국 평균인 1.1명¹⁾보다 여전히 높은 수치를 보이고 있다.

교통사고는 우리나라 국가기간교통망을 형성하고 있는 고속도로에서도 예외 없이 발생하고 있다. 2016년 말 기준 우리나라 주요 간선도로망인 고속도로는 28개 노선 3,989km에 이르고, 연간 154,033만대가 고속도로를 이용하였다(한국도로공사, 2017). 2016년 한 해 동안 고속도로에서는 총 2,195건의 교통사고가 발생하였으며, 그로 인해 총 239명이 사망하고, 1,424명이 부상을 당하였다(한국도로공사, 2017). 구체적으로 살펴보면 고속도로에서 발생한 교통사고 중 운전자 과실로 인한 교통사고 비율이 85.5%로 대부분을 차지하고 있다. 그 다음으로는 차량결함 교통사고가 7.9%, 마지막으로 노면잡물 등 기타 교통사고가 6.7%를 차지하고 있다.

한국도로공사는 고속도로 상의 교통안전 향상을 위해 다양한 사업 및 프로그램을 추진하고 있다. 정부가 목표로 하는 교통사고 감소 목표를 달성하기 위해서는 더욱 많은 노력이 요구되는 시점이다. 이러한 고속도로 안전성 향상을 위한 사업들을 효과적으로 수행하기 위해서는 교통사고 자료에 근거한 철저한 분

1) 최신 통계인 2014년 기준임

석의 선행되어야 할 필요가 있다. 최근 들어 교통사고 자료가 철저히 관리되며, 빅데이터 시대의 도래와 함께 교통사고와 관련된 요인을 설명할 수 있는 자료들의 종류와 양이 늘어나고 있는 시점에서 보다 최신 분석 기법을 이용하여 보다 엄격하고 정밀하게 교통사고 자료를 분석하고 시사점을 도출할 필요가 있다.

기존에는 대부분의 교통사고 자료 분석이 전통적인 통계적 방법인 포아송 회귀모형 또는 음이항 회귀모형 등을 기반으로 시행되어져 왔다. 이러한 통계적 방법은 교통사고와 관련된 다양한 인적, 도로 기하구조적 그리고 환경적 요인들과 교통사고 간의 인과관계를 찾고, 교통사고 빈도를 예측하고 그리고 분석된 결과를 바탕으로 교통안전 등급을 산출하는 등 다양한 방식으로 활용되어져 왔다. 하지만, 최근 머신 러닝 및 딥 러닝과 같은 빅데이터 분석 기법을 활용한 새로운 접근 방법들이 주목을 받기 시작하였다. 이러한 머신 러닝 및 딥 러닝 기법은 이종(異種)의 대량 자료를 활용하여 교통사고와 관련된 요인들을 분석하는 데 장점을 보이고 있으며, 이미 교통 및 다른 분야에서는 활발하게 적용되어 우리들의 일상을 변화시키고 있다.

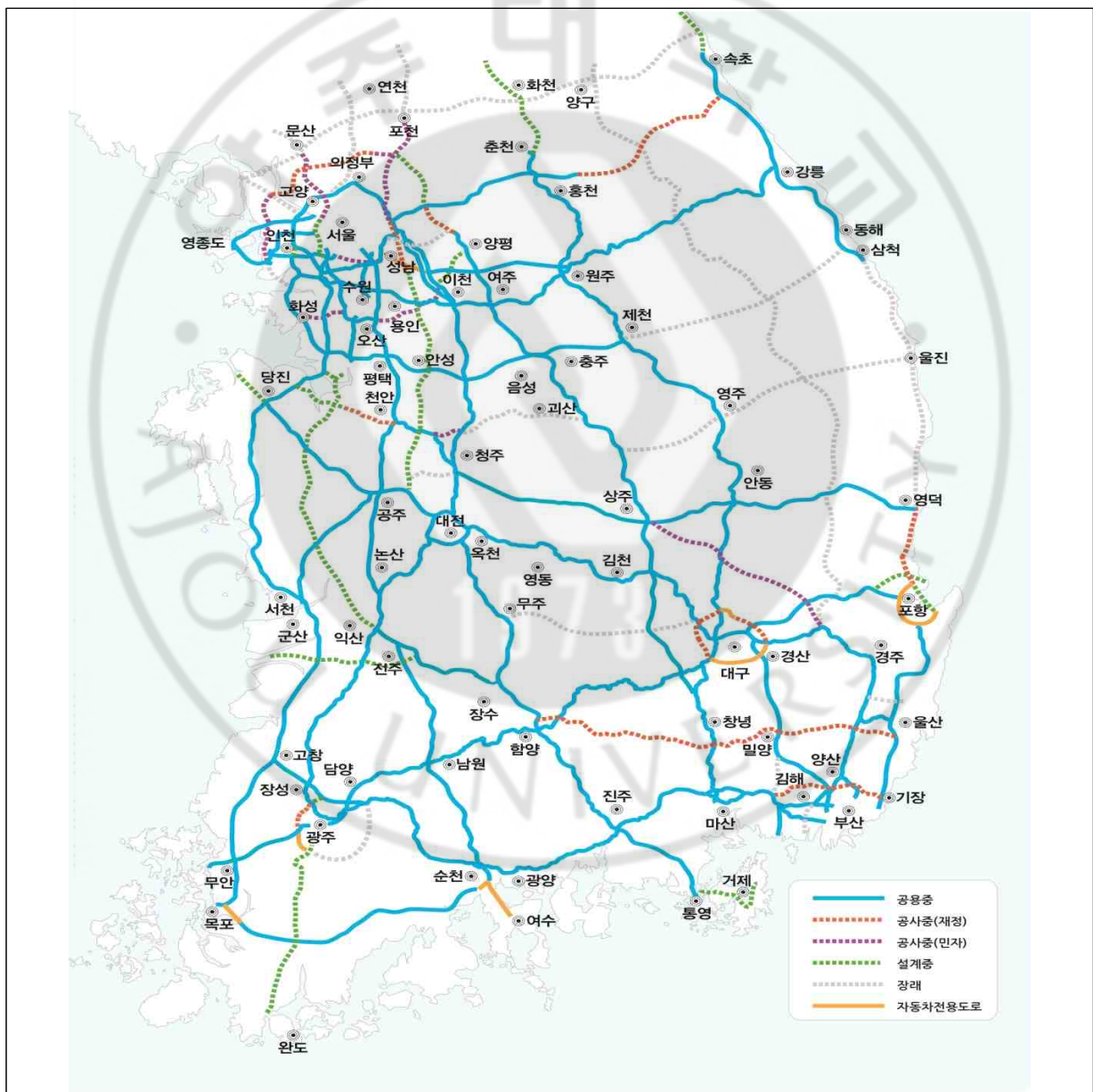
따라서 이러한 빅데이터 분석 기법을 활용하여 기존의 통계적 분석 방법으로 분석하기 곤란한 영역을 분석하고 또한 분석의 정밀도를 높이기 위한 노력이 필요한 시기이다.

2. 연구의 목적

본 연구의 목적은 고속도로 교통사고 자료를 이용하여 고속도로의 주요 분석 단위인 혼잡(Congestion Zone, Conzone)의 교통사고 빈도수를 예측하기 위하여 전통적인 통계적 기법, 머신 러닝을 이용한 기법 그리고 딥 러닝을 이용한 기법을 적용하고 각 기법들의 성능을 비교하고자 한다.

제2절 연구의 범위

본 연구의 공간적 범위는 자료 수집의 용이성 등을 고려하여 우리나라의 공용 중인 고속도로 중 한국도로공사가 관리하고 있는 고속도로 및 민자 고속도로로 한정한다. 또한 시간적 범위는 2013~2015년까지 3년간의 고속도로 교통사고 및 기하구조 자료를 활용하고자 한다.



<그림 1> 연구의 공간적 범위

제3절 연구의 수행절차 및 방법

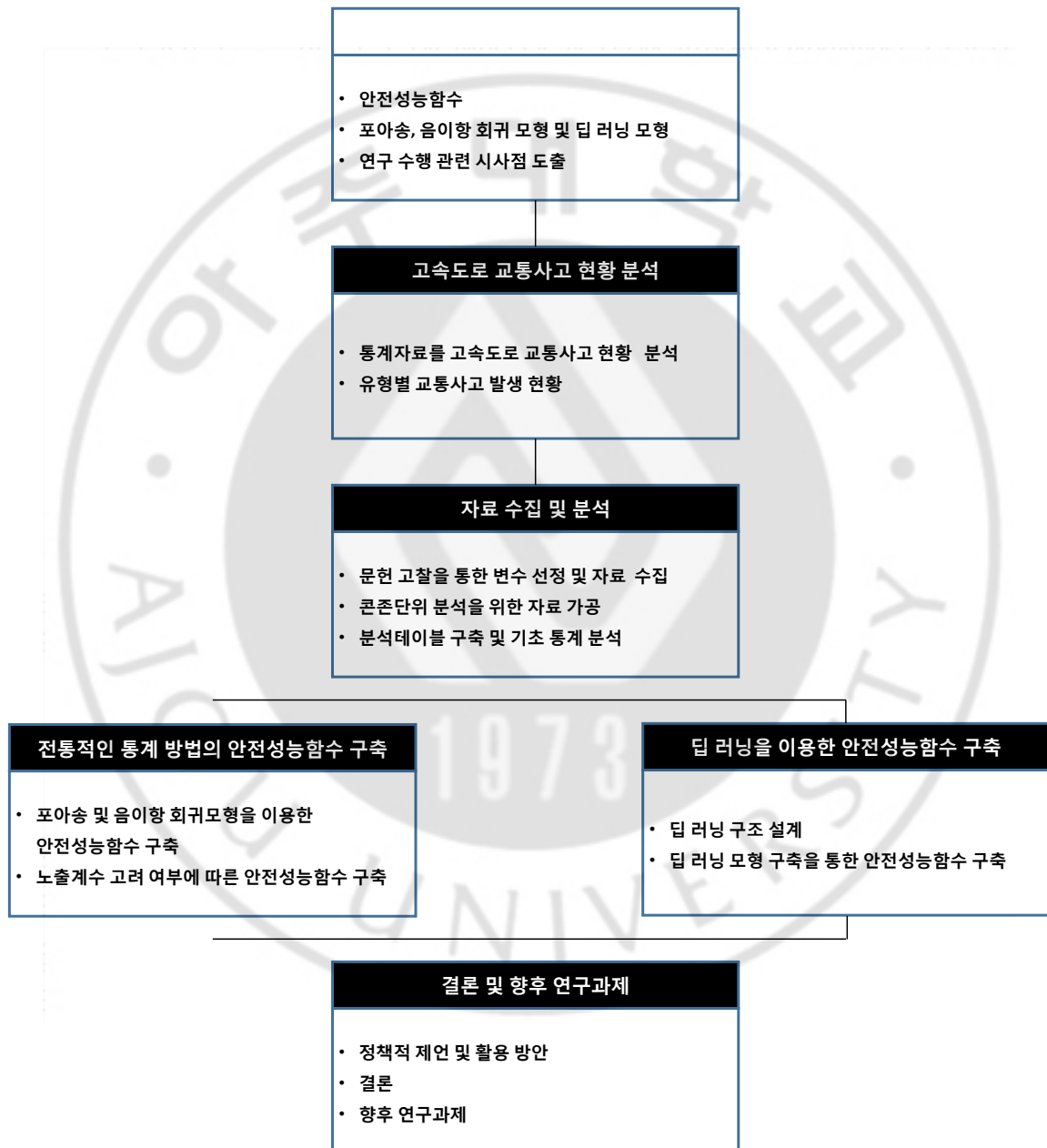
본 연구내용은 크게 이론적 배경 및 선행연구 고찰, 고속도로 교통사고 특성 분석, 고속도로 교통사고 관련 자료 수집, 통계적 기법을 이용한 고속도로 교통사고 예측, 그리고 딥 러닝을 이용한 교통사고 예측 고도화로 구성되어 있다. 이 중 본 연구의 핵심은 고속도로 콘존과 관련된 교통, 환경, 기하구조 등의 요인과 교통사고 간의 관계를 규명함으로써 교통사고 빈도를 예측할 수 있는 사고 예측모형을 개발하는 것이다.

이와 관련하여 연구항목별 세부 연구내용은 다음과 같다.

<표 1> 연구항목별 세부 연구내용

항 목	세부 연구내용
관련 이론 및 선행 연구 고찰	<ul style="list-style-type: none"> • 고속도로 교통사고 관련 이론 고찰 • 고속도로 교통사고 예측모형 관련 국내·외 연구 분석 • 이론 및 기존 연구 고찰을 통한 연구 수행을 위한 시사점 도출
고속도로 교통사고 특성 분석	<ul style="list-style-type: none"> • 고속도로 전체 교통사고 특성 분석 • 연구의 수행 및 자료 수집 관련 주요 시사점 도출
모형 개발을 위한 자료 수집	<ul style="list-style-type: none"> • 고속도로 콘존별 교통사고 자료 수집 <ul style="list-style-type: none"> – 교통, 환경 및 기하구조 관련 자료 수집 및 데이터베이스 구축 • 수집된 자료의 전처리 및 기초 통계분석
교통사고예측 모형 개발	<ul style="list-style-type: none"> • 통계적 기법을 이용한 고속도로 교통사고예측모형 개발 <ul style="list-style-type: none"> – 이산회귀모형(포아송 회귀모형 및 음이항 회귀모형) • 딥 러닝 기법을 이용한 교통사고예측모형 개발 <ul style="list-style-type: none"> – 딥 러닝 모형 구조 및 파라미터 조정 • 개발된 모형의 성능평가를 위한 효과척도 선정 및 비교
활용방안 및 향후 연구과제	<ul style="list-style-type: none"> • 고속도로 교통사고예측모형 활용방안 도출 • 고속도로 교통사고예측모형 발전을 위한 향후 연구과제

- 연구의 배경 및 목적 설정
- 연구의 시간적, 공간적, 내용적 범위 설정



<그림 2> 연구 수행절차

제2장 관련 이론 및 연구 고찰

제1절 관련 이론 고찰

1. 안전성능함수

최근까지 가장 보편적으로 사용되고 있는 교통안전성 분석을 위한 모형은 미국 도로안전편람(Highway Safety Manual, HSM)²⁾에서 제시하고 있는 모형이라고 할 수 있다. HSM에서 제시하는 모형은 안전성능함수(Safety Performance Function, SPF)³⁾와 사고보정계수(Crash Modification Factor, CMF)⁴⁾ 및 지역보정계수(C)를 이용하여 특정 도로에서 발생할 것으로 예상되는 사고건수를 예측하는 것이다(최윤환, 2012). SPF는 일반적으로 이상적인 상태에서 해당 도로의 교통사고 발생빈도를 예측 또는 추정하는 데 활용하며, 일평균교통량(AADT)과 도로구간연장 등의 노출계수(exposure)와 곡선반경 등 설계요소 등을 독립변수들을 이용하여 산출한다(오영태 및 강동수, 2017). HSM 뿐만 아니라 국내·외 다양한 학술연구에서도 교통사고 건수를 예측하기 위하여 SPF를 개발하거나 고도화하는 연구가 활발하게 진행되고 있다.

SPF는 사용하는 독립변수의 종류에 따라 단순 안전성능함수(Simple SPF), 통합 안전성능함수(Inclusive SPF)로 구분할 수 있다. Simple SPF는 독립변수로서 AADT 및 구간길이만을 적용하는 반면, Inclusive SPF는 AADT와 구간길이 이외에도 기하구조, 교통시설, 도로의 운영 정보 등의 다양한 변수들이 독립변수로서 사용된다. HSM는 기본적으로 단순 고속도로 교통사고 예측모형을 사용하고 있지만, 다양한 연구들에서 SPF 정확도 향상을 위하여 연평균일교통량(AADT), 구간길이, 차로 폭, 조명시설 존재 여부, 도로의 기하구조 등 다

2) 미국의 도로안전편람(HSM : Highway Safety Manual)은 교통정책 결정과정을 지원하기 위해 1990년대 초부터 개발하여 2010년부터 운용되고 있는 기법이다.

3) SPF(기본사고예측건수)는 기하구조 등 다른 위험요소가 없다고 가정하고 구간길이와 교통량만을 고려(기본조건) 했을 때 기본적으로 발생이 예상되는 사고건수로 안전성능 함수식으로 계산한다.

4) CMF는 도로조건에 따른 사고 영향계수이며, 회귀분석을 통해 도출된다.

양한 독립변수들을 사용하는 통합 안전성능함수가 개발 및 사용되고 있다(한국도로공사, 2014).

가. 단순 안전성능함수

단순 안전성능함수(Simple SPF)는 독립변수로서 교통량 관련 정보만을 사용하며 교통량을 대변할 수 있는 일평균교통량(ADT) 또는 연평균 일교통량(AADT)가 보편적으로 사용된다. 식(2.1)이 Simple SPF의 대표적인 형태를 보여주고 있다(한국도로공사, 2014).

$$N_{SPF} = e^{\alpha} \times ADT^{\beta} \quad (2.1)$$

- N_{SPF} : 예측사고건수 또는 예측사고율 (건/km/년)
- ADT : 해당 구간의 일평균 교통량 (대/일)
- α, β : 회귀 계수

Simple SPF의 경우 교통량 관련 정보만을 독립변수로 이용하기 때문에 나머지 설명인자들은 CMF의 형태로 고려된다. Simple SPF와 CMF를 이용한 교통사고 건수를 예측하는 방법은 식(2.2)와 같다(AASHTO, 2010).

$$N_{predicted} = N_{spf_x} \times (CMF_{1x} \times CMF_{2x} \times \dots \times CMF_{yx}) \times C_x \quad (2.2)$$

- $N_{predicted}$: 지점 유형(x)에서 특정 연도에 대해 예측된 평균 사고건수
- N_{spf_x} : 지점 형태(x)에 대해 개발된 SPF 기본조건의 예측 평균 사고건수
- CMF_{yx} : 지역 형태(x)에 대한 사고수정계수
- C_x : 지역 형태(x)에 대해 지역 조건들을 조정하는 보정계수

Simple SPF의 경우 교통사고 건수를 예측하는데 있어 SPF와 CMF가 분리된 형태를 가짐으로써 기본조건에 따른 일반적인 사고의 경향을 표현하는 데

용이하다. 그리고, CMF를 통해 다양한 개선사업의 효과를 정량적으로 표현할 수 있다는 장점이 있다. 그러나 이러한 교통사고 건수 예측방법은 기하구조, 교통안전 시설물 등에 대한 효과를 개별적으로 측정하여 CMF를 만들어야 하고 시행 가능한 모든 개선사업의 CMF를 추정하여야 하는 단점이 있다(한국도로공사, 2014).

나. 통합 안전성능함수

통합 안전성능함수(Inclusive SPF)는 교통량 및 구간길 이와 같은 노출변수(exposure)만을 독립변수로 사용하는 Simple SPF을 발전시킨 개념으로서 도로의 기하구조 정보, 교통안전시설물 설치 정보, 도로운영 정보 등이 모형 식의 독립변수로서 포함된다. Inclusive SPF는 식(2.3)과 같이 표현될 수 있다(한국도로공사, 2013).

$$N_{SPF} = e^{\alpha} \times ADT^{\beta} \times V_1^{\beta_1} \times V_2^{\beta_2} \times V_3^{\beta_3} \dots \times V_n^{\beta_n} \quad (2.3)$$

- N_{SPF} : 예측사고건수 또는 예측사고율 (건/km/년)
- ADT : 해당 구간의 일평균 교통량(대/일)
- α, β_n : 회귀 계수
- V_1, V_n : 독립변수

Inclusive SPF는 노출변수와 다양한 독립변수들이 교통사고에 미치는 영향 정도를 회귀계수의 형태로서 표현할 수 있으며, 각 회귀계수 값들의 한계효과를 추정함으로써 교통사고에 미치는 영향 정도를 계량화시킬 수 있다. Inclusive SPF의 경우 다양한 변수들을 동시에 고려할 수 있기 때문에 교통사고 건수 예측체계의 구성이 상대적으로 용이하지만, 모형 내 반영되지 못하는 타 요인들에 대한 영향력을 고려하지 못한다. 이를 고려하기 위해서는 모형 전체를 재추정해야 한다는 한계를 갖고 있다(한국도로공사, 2014).

2. 포아송 회귀모형

포아송 회귀모형(Poisson Regression Model)에서 사용하는 분포인 포아송 분포(Poisson Distribution)는 교통분야에서 가장 빈번하게 사용되는 확률분포 중 하나이다. 포아송 분포는 종속변수의 값이 이산적(discrete)일 때 사용된다. 즉, 교통사고 발생건수는 연속확률분포가 아닌 이산확률분포를 따르며 발생확률이 적은 특성을 고려하여 포아송 분포를 가정한 모형을 구축하는 것이 일반적이다.

포아송 분포를 사용할 경우 일반적으로 오차항을 정규분포(Normal Distribution)로 가정하는 회귀모형을 가지고는 분석이 불가능하다. 왜냐하면 교통사고 건수는 무조건 '0' 이상의 이산 값(Discrete Value)을 가지게 되며, 발생확률이 일반적으로 작은 특징을 가지고 있다. 하지만 이러한 특성을 무시하고 정규분포를 가정한 일반선형회귀모형을 사용하면 정확한 분석 결과를 기대할 수 없다. 따라서 교통사고 건수 분석에는 이산형 종속변수와 연속형 또는 이산형 독립변수에 대한 분포인 포아송 분포를 기반으로 하는 포아송 회귀모형을 사용하는 것이 타당하다(Simon et al., 2010).

포아송 분포를 고속도로 교통사고에 적용하면, 고속도로 기하구조를 독립변수로 가지며 교통사고 건수를 종속변수로 가지는 확률변수 Y_i 의 확률함수로 표현할 수 있으며, 이는 식(2.4)과 같다. 또한 식(2.5)와 같이, 포아송 분포의 기댓값 μ_i 의 로그변환 형태는 독립변수(X_1, X_2, \dots, X_k)들의 선형결합으로 표현할 수 있다(정재풍, 2014).

$$P(Y_i = y_i; X_1, X_2, \dots, X_k) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \quad y = 0, 1, 2, \dots \quad (2.4)$$

$$\log(\mu_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k \quad (2.5)$$

식(2.5)를 μ_i 에 대하여 다시 표현하면 식(2.6)과 같이 독립변수들의 지수함수 형태로 표현할 수 있다(정재풍, 2014).

$$\begin{aligned} \mu_i &= \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k) \\ &= \exp\left(\sum_{k=0}^k \beta_k X_k\right) \end{aligned} \quad (2.6)$$

여기서,

β_k : 회귀식의 추정계수

X_k : 회귀식의 독립변수

포아송 회귀모형의 회귀계수 $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ 의 추정은 최대우도 추정법(Maximum Likelihood Estimation, MLE)을 기반으로 하는 수치해석적 방법을 이용하여 β 의 최우추정값을 찾는다(정재풍, 2014).

하지만, 포아송 분포를 사용하기 위해서는 교통사고 건수의 평균과 분산이 같다는 가정이 필요하다. 실제 교통사고 자료를 살펴보면 많은 경우에 분산이 기댓값보다 큰 과대산포(Over-Dispersion) 또는 과분산 문제가 발생한다. 분석하고자 하는 교통사고 자료가 과대산포 문제를 보일 경우 평균과 분산이 같다는 제약으로부터 자유로운 음이항 분포를 포아송 분포를 대신해서 사용할 수 있다(성낙문, 2002).

3. 음이항 회귀모형

포아송 회귀모형은 평균과 분산이 같다는 포아송 분포의 가정이 필요하다. 하지만 실제 교통사고 건수의 경우 이러한 가정을 만족하지 못해 과소산포(Under-Dispersion) 혹은 과대산포(Over-Dispersion)의 문제가 발생하는 경우가 발생한다. 이러한 경우 음이항 회귀모형(Negative Binomial Regression Model)은 이러한 포아송 분포의 한계를 오차항(ϵ_k) 추가를 통해 해결할 수 있다. 이때 기댓값은 아래 식(2.7)와 같이 표현할 수 있다(Simon et al., 2010; 정재풍, 2014).

$$\begin{aligned}\mu_i &= \exp(\beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \epsilon_k) \\ &= \exp\left(\sum_0^k \beta_k X_k + e_k\right)\end{aligned}\quad (2.7)$$

여기서,

β_k : 회귀식의 추정계수

X_k : 회귀식의 독립변수

ϵ_k : 오차항

식(2.7)를 이용하여 음이항 분포를 따르는 Y_i 의 기댓값과 분산의 관계는 식(2.8)과 같이 나타낼 수 있다.

$$\begin{aligned}Var(Y_i) &= \mu_i(1 + \alpha\mu_i) \\ &\quad \mu_i + \alpha\mu_i^2\end{aligned}\quad (2.8)$$

여기서,

$Var[Y_i]$: 종속변수(Y_i)의 분산

μ_i : 종속변수의 기댓값(평균)

α : 과분산계수

위의 식(2.8)를 통해 알 수 있듯이 과분산계수(Over-Dispersion Parameter) α 값이 0일 경우 포아송 분포와 동일한 형태이다. 일반적으로 α 값이 0.5 이하일 경우 포아송 분포를 사용하며 0.5 이상일 경우 음이항 분포를 사용하는 것이 적절한 것으로 알려져 있다(최윤환, 2012). 하지만 이러한 적용은 임의적으로 구분한 것이기 때문에 우도비 검정을 이용하여 과분산 여부를 평가하는 것이 더욱 타당하다(황경성 외, 2010). 과분산 여부를 검정하기 위한 우도비 검정 통계량(Likelihood Ratio, LR)은 아래 식(2.9)을 통해 구할 수 있다(이일현, 2014).

$$LR = -2(LL_{poisson} - LL_{negative}) \quad (2.9)$$

과분산 검정을 위한 귀무가설(H_0)과 대립가설(H_1)은 식(2.10)과 같다. 식(2.10)에서 LR 값은 자유도가 1인 χ^2 분포를 따르고 유의수준 임계값과 LR 값을 비교하여 과분산 여부를 평가할 수 있다. LR 값이 유의수준 임계값 보다 클 경우 기각역에 해당하여 과분산이 발생한 것으로 볼 수 있고 음이항 회귀모형을 사용하는 것이 바람직하다(황경성 외, 2010).

$$H_0 : \alpha = 0, \quad H_1 : \alpha > 0 \quad (2.10)$$

여기서,

α : 과분산계수

그리고 모형의 적합성을 비교할 때 사용하는 통계량으로서 AIC(Akaike Information Criterion)과 BIC(Bayesian Information Criterion)가 있다. 이 두 개의 통계량은 동일한 종속변수와 독립변수로 실시한 포아송 및 음이항 회귀모형 중에 더 좋은 모형을 평가한다. 여기서 AIC 및 BIC 값이 낮을수록 더 우수한 회귀모형임을 의미한다(이일현, 2014).

4. 인공 신경망

인공 신경망(Artificial Neural Network, ANN)은 인간의 뇌에 존재하는 수많은 신경세포(Neuron)가 복잡하게 연결되어 있는 형태를 모형화한 기법이다(오주택 외, 2014). 즉, 인간의 인식과정이나 신경생태를 수학적으로 모형화한 것이라고 할 수 있다. ANN은 어떤 문제에 대한 해결방안을 보여주거나 혹은 기존의 방법으로 해결하지 못했던 문제들을 해결 가능성을 제시하였다. 최근에는 ANN을 이용하여 패턴을 인식하거나 의사를 결정하는 등 많은 분야에서 이용되고 있다. ANN은 변수들의 연결강도(Weight)들을 통하여 결과 값을 결정하는데 이러한 구조를 수학적식(2.11)과 같이 표현할 수 있다(오주택 외, 2004).

$$n = x_1w_1 + x_2w_2 + \cdots + x_nw_n \quad (2.11)$$

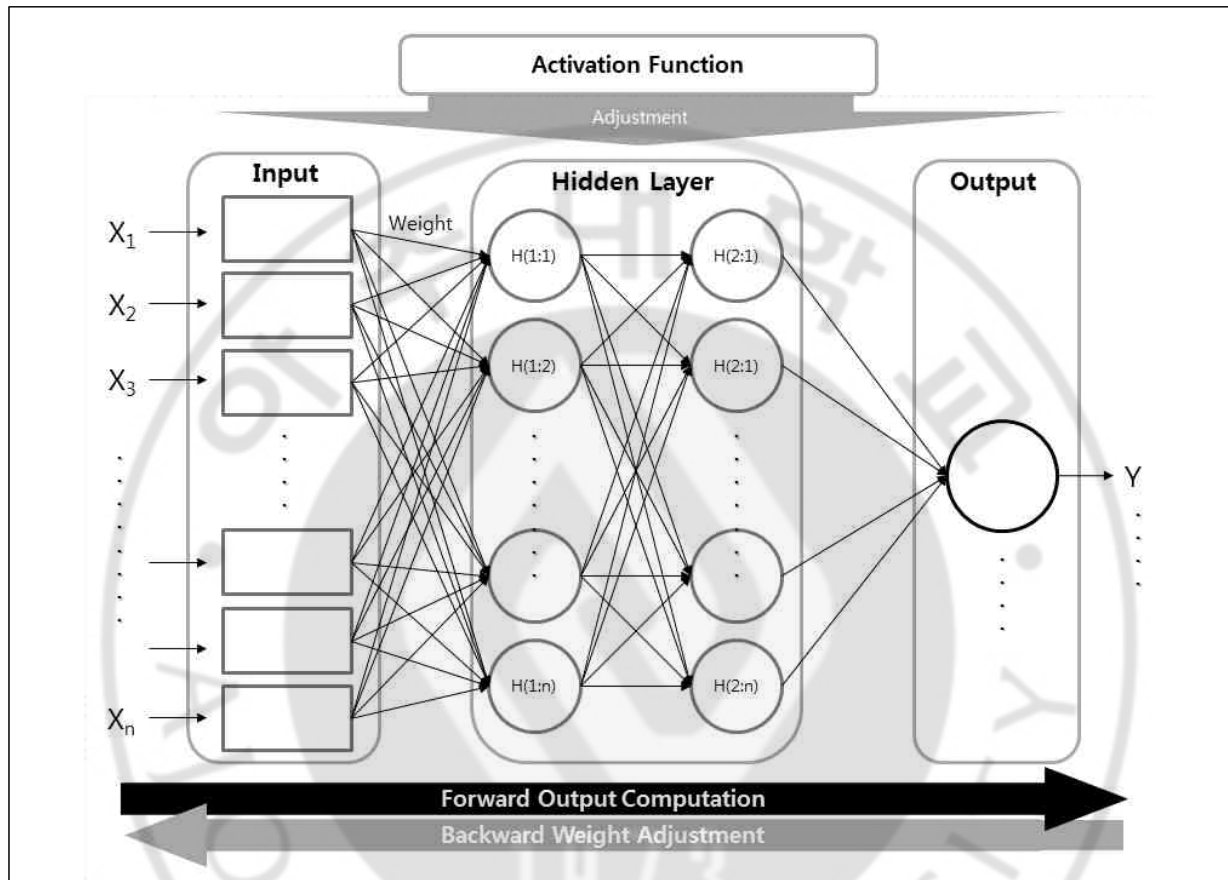
여기서,

x_n : 입력변수

w_n : 연결강도(weight)

ANN은 다른 모형들과 달리 학습(Learning)을 이용한다. 변수들 간의 관계를 계량화하는 연결강도는 학습이라는 과정을 통해서 조정된다. 일반적으로 사용되는 학습 방법으로는 지도학습(Supervised Learning)에 기반한 오류 역전파(backpropagation) 알고리즘을 사용한다. 오류 역전파 알고리즘은 일반화된 델타 규칙으로도 불린다. 다중 퍼셉트론을 학습시키기 위하여 사용되는 오류 역전파 알고리즘은 체계적으로 모형의 비선형성을 해결하여 출력되는 값과 목표로 하는 값의 차이를 최소화한다(오주택 외, 2014). ANN의 구조는 자료를 입력하는 입력층(Input Layer), 은닉층(Hidden Layer), 그리고 마지막 층인 출력층(Output Layer)으로 구성된다. 입력층은 예측값을 도출하기 위한 예측변수의 값들을 입력하는 역할을 하며 은닉층은 모든 입력노드부터 입력값을 받아 가중

합을 계산하고, 이 값을 전이함수에 적용하여 출력층에 전달하게 된다. 입력층에서 출력층으로 전달되는 과정 중 활성화 함수를 거쳐 결과를 도출하게된 것이다. 일반적인 오류역전파 인공 신경망의 구조는 다음 그림과 같다(Lee, 2009).



<그림 3> ANN 구조

ANN을 통한 예측값의 검증을 위한 방법은 다양하다. MPB(Mean Percentage Bias), MAD(Mean Absolute Deviation), R2, RMSE(Root Mean Square Error), %RMSE(Percent Root Mean Square Error) 등이 있는데 인공 신경망을 이용하여 예측한 값과 실제 값과의 비교를 통해 오차를 표현하는 방법인 MPB, MAD, RMSE, %RMSE 검증방법은 값이 '0'에 가까울수록 오차가 작다는 것을 의미하고 설명력을 나타내는 R2은 %가 높을수록 설명력이 높다고 할 수 있다(박병호 및 나희, 2012).

가. 딥 러닝 알고리즘

앞서 언급한 바와 같이 딥 러닝은 DNN을 학습하고 이를 이용하여 추론하는 연구이므로 DNN에서 사용하는 알고리즘을 주로 사용한다. 이 중에서 대표적인 것이 심층 신경망(Deep Neural Network, DNN), 합성곱 신경망(Convolution Neural Network, CNN), 제한 볼츠만 머신(Restricted Boltzmann Machine, RBM), 심층 신뢰 신경망(Deep Belief Network) 등이 있다.⁵⁾

(1) 심층 신경망

심층 신경망(Deep Neural Network, DNN)은 ANN 동일한 구조이나 입력층과 출력층 중간에 은닉층의 개수가 2개 이상이다. 은닉층이 늘어날수록 H/W 관련 한계가 있을 수 있고 학습을 시키기 위한 많은 데이터가 필요하다는 문제가 있다. 하지만, 이러한 한계들은 최근 빅데이터 분석 관련 H/W 및 S/W의 발달로 인하여 기술적으로 극복하고 있다(정완, 2017).

DNN은 기존 ANN과 마찬가지로 복잡한 비선형 관계들을 모형화 할 수 있다. 예를 들어 사물 식별을 위해 구축된 DNN 구조의 경우 각 객체를 이미지 기본 요소들의 계층적 구성으로 표현가능하다. 이때 추가 계층의 경우 점진적으로 모아진 하위 계층들의 특징들을 모을 수 있다(Christian Szegedy et al., 2013). DNN의 이러한 특징은 ANN에 비해 더 적은 수의 유닛들 만으로도 복잡한 데이터를 모델링 할 수 있게 해준다(Bengio, 2013).

이전의 DNN들은 보통 앞먹임 신경망(Feedforward Neural Network)으로 대부분 설계되어 왔다. 관련 프로그램 기술의 발달로 인해 최근의 연구들은 심층 학습 구조를 순환 신경망(Recurrent Neural Network, RNN)에 적용하는 것을 성공하였다(T. Mikolov et al., 2010).

DNN은 표준 오류 역전파 알고리즘을 통해 학습하게 되며 이때, 연결강도 또는 가중치(weight)들은 아래의 식과 같이 확률적 경사 하강법(Stochastic

5) 위키백과, 2017년 11월 접근

Gradient Descent)을 기반으로 갱신된다(Hinton et al., 2012).

$$\Delta w_{ij}(t+1) = \Delta w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}} \quad (2.12)$$

여기서 η 는 학습률(learning rate), C 는 비용함수(cost function)를 나타낸다. 비용함수의 선택은 사용되는 학습의 형태(지도학습, 비지도학습, 강화학습)와 활성화 함수(activation function)를 고려하여 결정하게 된다. 예를 들어, 다중 클래스 분류 문제(multi-class classification problem)에 지도학습을 수행할 때 일반적으로 활성화 함수와 비용함수는 각각 소프트맥스(softmax) 함수와 교차엔트로피 함수(cross entropy function)로 결정할 수 있다(G. E. Hinton et al., 2012).

(2) 합성곱 신경망⁶⁾

합성곱 신경망(Convolution Neural Network, CNN)은 전처리를 최소화하도록 설계된 다층 퍼셉트론(multilayer perceptrons)중의 한가지이다. CNN은 한 개 혹은 여러 개의 합성곱 계층과 그 위에 위치한 일반적인 ANN 계층들로 구성되어 있으며, 가중치 및 통합계층(pooling layer)들을 추가적으로 활용하게 된다. CNN은 2차원 이상의 구조를 가지는 입력 데이터를 충분히 활용할 수 있다. 또한 다른 딥 러닝 구조보다 영상, 음성 분야에서 좋은 성능을 보여주는 것으로 알려져 있다. CNN은 표준 오류 역전파 알고리즘을 통해 훈련될 수 있으며 CNN은 다른 앞먹임 신경망 기법들보다 쉽게 훈련이 가능하며 매개변수를 적게 사용한다는 장점이 있다.

최근 딥 러닝의 새로운 구조로서 합성곱 심층 신뢰 신경망(Convolutional Deep Belief Network, CDBN)이 개발되어 폭 넓게 사용되고 있다. 이는 기존 CNN과 구조적으로 유사하고 2차원 구조를 이용할 수 있으며 심층신뢰 신경망(Deep Belief Network, DBN)에서의 사전훈련(pre-training)에 의한 장점 또한 존재한다(Krizhevsky, 2010). CDBN은 다양한 영상 및 신호처리기법에 사용가능한 일반적인 구조를 제공하고 있다.

6) 위키백과, 2017년 11월 접근

(3) 제한 볼츠만 머신⁷⁾

볼츠만 머신(Boltzmann Machine)에서, 층간 연결을 없앤 형태가 제한 볼츠만 머신(Restricted Boltzmann Machine, RBM)이다. 층간 연결을 제거할 경우, 볼츠만 머신은 가시 유닛(visible unit)과 은닉 유닛(hidden unit)으로 이루어진 무방향 이분 그래프 형태로 구성된다. 이러한 RBM은 심층신뢰 신경망(Deep Belief Network, DBN)의 기본 골격이 된다.

(4) 심층신뢰 신경망⁸⁾

심층신뢰 신경망(Deep Belief Network, DBN)은 기계학습에서 사용되는 그래프 생성 모형(generative graphical model)의 한 종류이며, 다중계층으로 구성된 잠재변수(latent variable)를 이용한 계층 간 연결이 존재하지만 계층 내 유닛 간에는 연결이 없다는 것이 특징이다. 심층신뢰 신경망은 여러 개의 은닉 층으로 구성된 다층 신경망으로서 훈련용 데이터가 적을 때 유용한 구조를 가지고 비지도 방식으로 계층마다 학습을 진행한다. 심층신뢰 신경망은 생성모형이라는 특성 상 선행학습 시 사용가능하다. 또한 선행학습으로 알려진 사전학습을 이용하여 수행에 필요한 초기 가중치를 설정한 후 역전파 혹은 기타 판별 알고리즘을 통해 가중치의 세부적인 조정이 가능하다. 이와 같은 특성은 훈련용 데이터가 충분하지 않을 경우에 유용하며, 훈련용 데이터가 적을수록 가중치의 초기 값이 결과적인 모형에 끼치는 영향이 커진다. 선행학습을 통해 선택된 가중치 초기 값은 임의의 가중치 초기 값에 비해 최적의 가중치 값에 가까워지고 이는 곧 미조정 단계의 성능과 속도향상이 가능함을 뜻한다(이세진과 김동현, 2016).

7) 위키백과, 2017년 11월 접근

8) 위키백과, 2017년 11월 접근

5. K-means 클러스터링

클러스터링(clustering)은 주어진 개체들 중 유사한 개체들을 몇몇의 집단으로 군집화하여 각 집단의 성격을 파악하는데, 이때 각 개체의 유사성을 측정할 수 있는 방법이 필요하다. 이를 위해 개체를 속성과 속성값의 나열인 벡터의 형태로 변환하여 벡터간 유사도를 계산하는데, 개체들이 비슷한 속성값을 갖게 될수록 개체 간 유사성이 높아 하나의 클러스터를 이루며, 유사성이 낮은 개체들은 다른 클러스터에 속하도록 하는 것이 클러스터링 알고리즘의 기본 원리이다 (Ye, 2016).

물리적 혹은 추상적 객체들을 비슷한 객체들의 클래스로 그룹화하는 것을 군집화라 한다. 군집이란 동일한 군집 내의 객체들과는 비슷하고, 다른 군집의 객체들과는 상이한 데이터 객체들의 집합을 뜻한다. 데이터 객체들의 군집은 많은 응용에서 집합적으로 하나의 그룹으로 여겨진다(문지원, 2007).

K-Means 알고리즘이 클러스터링 알고리즘 중에서 가장 대표적으로 사용되는 알고리즘이라 할 수 있다. 이 알고리즘의 기본 개념은 데이터 집단들과 그 집단들이 속하는 클러스터 중심과의 평균 유클리디안(euclidean) 거리를 최소화하는 것이다. 데이터들의 집단인 클러스터의 중심은 그 클러스터에 속한 데이터들의 평균 혹은 중심(centroid) $\vec{\mu}$ 이며, 아래와 같이 정의할 수 있다(이원휘, 2010).

$$\vec{\mu}(w) = \frac{1}{|w|} \sum_{x \in w} \vec{x} \quad (2.13)$$

위의 식에서 w 는 클러스터에 속한 데이터 객체들의 집합을 의미하며, \vec{x} 는 클러스터에 속한 특정 데이터 객체를 말한다.

클러스터 중심이 클러스터에 속한 데이터들을 얼마나 잘 표현했는가를 나타내는 척도인 RSS(Residual Sum of Squares)는 각 클러스터에 속하는 모든 데이터 객체에 대하여 각 데이터와 중심까지의 제곱거리의 합으로 산출되며 아래 식과 같다(Christopher et al., 2008; 권순재 외, 2017).

$$RSS_k = \sum_{x \in w_k} |\vec{x} - \vec{u}(w_k)|^2 \quad (2.14)$$

$$RSS = \sum_{k=1}^K RSS_k \quad (2.15)$$

K-Means Clustering 알고리즘은 입력 값을 k로 취하고 군집 내 유사성은 높고, 군집끼리 유사성이 낮게 되므로 n개 객체의 집합을 k개의 군집으로 분해하고, 유사성은 객체들의 평균값으로 측정한다고 한다. 즉, K-Means Clustering 알고리즘은 특정 성질의 데이터들이 유사성을 기초로 한 고정된 수인 k의 군집을 찾는 알고리즘을 의미한다. 이와 같은 작업을 통해 사용자가 설정한 임의의 임계치를 만족할 때 까지 반복적으로 군집분류를 진행한다(Arthur and Vassilvitskii, 2006).

K-Means는 초기 중심의 선정에 따라 성능이 크게 달라진다. 기존의 알고리즘은 초기 중심을 선정할 때 무작위로 설정되어 왔다. 하지만, 무작위로 선정된 초기 중심에서 진행되어온 클러스터링은 편차가 크므로, 선행 연구를 통해 초기 중심 설정의 문제점을 해결하고자 하였다(이신원, 2012).

제2절 기존 연구 고찰

1. 전통적인 통계기법을 이용한 안전성능함수 구축 사례

이수범 외(2003)는 도로 신설 및 개량 사업에 대한 타당성 조사 시 도로의 물리적인 특성이 충분히 반영되지 못하는 점을 해결하기 위하여 교통 특성과 도로의 물리적 특성을 고려한 교통사고 예측모형을 개발하였다. 이 연구에서는 도로 계획단계에서도 수집할 수 있는 자료들을 활용하여 고속도로, 도시지역의 일반국도, 지방도를 대상으로 도로 특성별 교통사고 예측모형을 구축하였다. 이 연구에서 사용한 변수는 중앙분리대의 유·무, 교통량, 교차로 수, 횡단신호등 수, 연결로 수, IC밀도 및 차로수이다. 다중회귀모형의 stepwise법을 이용하여 모형을 구축하였으며, 모든 변수들이 통계적으로 유의한 영향을 주는 것으로 분석되었다. 특히, 교차로 수, 횡단신호등이 사고와 연관성이 높은 것으로 나타났다. 개발된 모형을 이용하여 교통사고 빈도를 예측한 결과, 고속도로의 경우 교통량이 사고에 미치는 영향이 가장 큰 것으로 분석되었으며, 2차로 도로의 경우는 교차로 수 및 횡단신호등, 4차로·중앙분리대가 있는 경우는 교차로수가 교통사고에 영향을 미치는 주요요인으로 나타났다. 이 연구의 결과는 도로 신설 및 개량 사업 계획 시 여러 가지 대안에 대한 교통상의 안전성 평가를 위하여 활용될 수 있을 것으로 사료된다.

이근희 외(2015)는 서울, 수도권 그리고 부산광역시의 4지 신호교차로를 대상 음이향 회귀모형을 이용하여 교통사고 예측모형을 개발하였다. 개발된 모형을 이용하여 교통사고 빈도 및 특성을 분석한 결과, 기존의 음이향 회귀모형보다 확률적 음이향 회귀모형의 설명력이 높게 나타났다. 또한 총 52개의 변수 중 10개의 변수가(주도로의 차로수, 주도로의 좌회전 교통량, 주도로의 주행제약시설 수, 부도로의 우회전 교통량, 부도로의 교차로 시거, 교차로의 총현시, 부도로의 중앙분리대 유무, 부도로의 제한속도, 부도로의 교통섬 유무, 부도로의 속도제약시설 수)가 도시부 4지 신호교차로에서 교통사고에 영향을 미치는

유의한 변수인 것으로 확인되었다. 지금까지 대부분의 연구에서 사용되었던 회귀모형에서는 확인 할 수 없었던 확률적 변수를 도출하였다는 것이 이 연구의 학술적 기여도라고 할 수 있다. 상세히 설명하면, 교통사고에 유의한 영향을 미치는 것으로 확인된 10개의 유의한 변수 중 2개의 변수(부도로의 교차로 시거, 부도로의 차량 주행속도 제약 시설물 수)가 확률적 변수로 확인되었다. 또한 향후 연구 방향으로서 종속변수를 총 교통사고건수가 아닌 심각도별 사고건수를 적용한다면 각 변수가 단순 사고 발생이 아닌 사고 심각도에 미치는 변수를 파악할 수 있을 뿐 아니라 향후 교통사고 예방을 위한 개선 방안의 우선순위 등에 많은 도움을 줄 수 있을 것이라고 밝히고 있다.

강동운(2014)은 기존의 연구들에서 고려하지 못했던 안전성능함수 구축 시 독립변수들의 상관성 고려에 관한 연구를 수행하였다. 이 연구에서 안전성능함수 구축 시 독립변수들 간 상관성 문제를 해결하기 위하여 주성분 분석이 활용 가능하다고 제안하고 있다. 이를 증명하기 위하여 경부, 영동, 서해안 고속도로의 2008년부터 2012년까지 3년간의 자료를 사용하였고, 종속변수는 사고 건수, 독립변수로는 AADT, 구간길이, 평면선형, 종단선형, 차로수를 사용하였다. 교통사고 건수 예측 결과, AADT와 차로수의 상관관계가 높았고, 두 변수가 동시에 모형에 적용되었을 때 문제를 야기할 수 있음을 확인하였다. 또한 이를 해결하기 위해 주성분 분석을 적용한 후 고속도로 교통사고 예측모형을 구축하였다. CURE plot과 MAD와 RMSE를 통해 모형의 적합도와 예측력을 확인한 결과, 모형의 적합도가 준수하며 기존 모형에 비해 예측력도 비교적 높은 것으로 나타났다.

서임기 외(2015)는 경부고속도로, 호남고속도로, 영동고속도로, 서해안고속도로, 중부내륙고속도로, 중앙고속도로를 대상으로 교통사고의 특성을 분석하고, 교통사고 건수 예측을 위한 안전성능함수를 개발하였다. 교통량을 기반으로 개발된 안전성능함수는 분석대상 노선을 통합한 고속도로 교통사고 예측모형을 구축하였다. 또한 각 고속도로 노선별 교통사고 건수를 예측하기 위하여 각 노선별 교통사고 보정계수를 산출하였다.

Hochan Kwak et al.(2010)는 고속도로 교통사고에 영향을 미치는 기하구조 및 환경적 요인을 반영하여 고속도로 교통사고 건수 예측모형을 개발하였다. 이 연구는 예측모형의 구조를 결정하기 위해 적합도 검정방법(goodness of fit test)을 제시하고 있으며, 최종적으로 선정된 모형의 통계적 유의성을 검토하기 위해 Cumulative Scaled Residuals(CURE) plotting 방법론을 활용한 것이 특징이라 할 수 있다. 공간적 범위는 경부고속도로 양방향 전구간이며 3년간(2006년부터 2008년)의 기하구조 및 교통량 자료와 사고자료(사고발생위치, 사고시간, 사고유형, 사고원인 및 심각도)를 활용하였다. 교통사고의 경우 매우 확률이 적은 사건(rare and random event)로서 고속도로 구간 내에 발생한 사고의 대부분이 0~6건 사이에 분포하고 있는 것으로 확인되었다. 그 중 사고가 발생하지 않는 구간의 비중이 상당히 높은 것으로 나타났다. 이러한 '0'과잉 현상을 극복하기 위해 포아송 및 음이항 회귀모형 이외에 영과잉 포아송 및 영과잉 음이항 회귀모형을 함께 구축하였다. 음이항 회귀모형을 이용해 추정된 안전성능함수는 포아송 회귀모형을 통해 구축된 모형보다 통계적으로 우수한 것으로 확인되었다. 영과잉모형 사용 적합성을 확인할 수 있는 Vuong 검정을 통해 포아송 회귀모형과 영과잉 포아송 회귀모형을 비교했을 때 영과잉 포아송 회귀모형이 통계적으로 적합함을 확인하였다.

Jinyan Lu et al.(2013)은 교통사고 예측과 교통사고 위험이 높은 위치 검증을 위하여 safety analyst에서 제공하는 형태인 단순 안전성능함수와 교통 특성 및 기하구조 자료를 활용한 통합 안전성능함수를 비교하였다. 본 연구의 공간적 범위는 미국 플로리다 주의 도시부 4차로 고속도로이며 4년간(2007년부터 2010년)의 자료를 활용하였다. 기하구조 자료를 추출하기 위하여 플로리다 주 교통부의 Roadway Characteristic Inventory(RCI) DB를 사용하였다. 통합 안전성능함수의 독립변수로는 AADT, 차로폭, 중앙 간격, 오른쪽 바깥 길어깨 폭, 왼쪽 안 길어깨 폭, 오른쪽 바깥 길어깨 유형, 왼쪽 안 길어깨 유형, 제한 속도, 트럭 비율을 이용하였다. 단순 안전성능함수의 독립변수는 AADT만을 사용하였다. 이 연구는 HSM을 비롯한 학술연구에 사용되고 있는 음이항 회귀모

형을 이용하여 고속도로 교통사고 예측모형을 구축하였다. 이 연구에서는 통합 안전성능함수와 단순 안전성능함수의 교통사고 예측력 비교를 실시하며, 최종적으로 MAD와 Mean Squared Prediction Error(MSPE)를 이용하여 모형을 선정하였다. 선정 결과, 두 가지 모형 모두 교통사고 예측과 network screening에 있어 유사한 성능을 보이고 있는 것으로 확인되었다.

Salvatore Cafiso et al.(2013)는 고속도로 교통사고 예측모형을 구축하는데 있어 구간 길이를 짧은 구간으로 분할할 때 발생하는 문제점과 긴 구간으로 분할할 때 발생하는 문제점을 연구하여 최적의 분할 기준을 제시하였다. 이 연구의 공간적 범위는 이탈리아 지방부 고속도로이며, 시간적 범위는 8개년(2002년부터 2009년)이다. AADT, 사고위험도(사망사고 및 부상사고 합), 구간 길이, 곡률변화율, 경사변화율, 길가위험성을 독립변수로 사용하였다. 이 연구에서는 안전성능함수는 독립변수로 고려된 모든 변수들을 반영하는 모형, AADT와 곡률변화율(CCR)을 반영하는 모형, AADT만 반영하는 기본 모형 등 총 세 가지 모형을 개발하였다. 개발된 세 가지 모형에 대한 비교 결과, 최종적으로 Quasilikelihood under the Independence model Criterion(QIC)를 이용하여 적합도가 가장 높은 분할 형태를 결정하였다. 가장 좋은 분할 형태는 Quasilikelihood under the Independence model Criterion(QIC)가 가장 작은 segmentation 2인 2 curves와 2 tangents 기반 분할이었다. segmentation 4인 고정된 길이 기반 분할은 실제 적용에서 가장 유연하였다. 그 이유는 분할 길이가 SPF를 산정하는 요소들과 자료의 이용 가능성 및 질에 의해 결정되기 때문이었다. 가장 좋지 않은 분할 형태는 segmentation 5인 반영된 모든 변수들이 동질하다는 가정 기반 분할인 것으로 확인되었다.

Ducknyung Kim et al.(2013)은 고속도로 구간에 따른 특성을 반영한 고속도로 교통사고 건수 예측모형을 구축하였다. 이 연구의 공간적 범위로는 경부고속도로, 서해안고속도로, 영동고속도로, 호남고속도로이며, 시간적 범위는 총 4개년도(2007년부터 2010년)이다. 이 연구는 AADT, 사고 위험도(총 사고건수), 구간길이를 독립변수로 사용하였다. 분석 과정으로는 먼저 사고 위험도에 따라 AADT, 구간길이, 교통사고 건수를 바탕으로 구간 그룹화를 수행하였다.

따라서 각 사고 위험도에 따라 군집분석을 통해 3개의 그룹으로 분류하였다. 분류된 그룹을 토대로 각 그룹의 사고 분포에 대한 적합도 검정을 수행하였다. 검정 결과, 세 가지 모형 모두에서 음이항 분포를 따르는 것으로 확인되었다. 다음으로 The Bayesian information Criterion(BIC), Over-dispersion parameter, McFadden's pseudo R^2 와 같은 통계적 지표를 최종 안전성능함수를 선정하였다. 선정된 모형은 AADT와 구간길이가 노출계수로 사용된 모형이었다. 선택된 사고 분포와 모형의 형태를 고려하여 FI 사고와 total 사고에 대한 SPF를 각 그룹에서 추정하였다. 군집화된 그룹 사이 validity of differences를 설명하기 위해서 log-likelihood function value를 사용한 taste variation test를 실시하고, 추정된 계수들 사이의 차이를 보이기 위해 asymptotic t-test를 실시하였다. 2개의 test를 통해 FI 사고와 total 사고에 대한 안전성능함수 최종 모델을 산출하였다. 분석결과, Fatal-injury 사고와 total 사고에 관한 2가지 최종 모형을 제시하였고, 제시한 2가지 모델 모두 높은 통계적 유의성을 보였다.

Mohamadreza Banihashemi(2012)는 전체 데이터를 기준으로 보정계수를 산정하는 것을 참값(ideal calibration factor)으로 가정하고 각각의 data %에 따라 보정계수가 어떻게 달라지는 지에 대한 민감도 분석을 수행하였다. 최종적으로 본 연구의 결과를 통해 신뢰성이 수반된 보정계수의 산출방법, 데이터 규모를 제시하였다. 본 연구의 공간적 범위로는 워싱턴 주의 8가지 종류의 highway이고 시간적 범위로는 2006 ~ 2008년 총 3개년도이며 도로, 곡선, 경사, 차로, 사고 정보 자료를 이용하였다. 먼저 시뮬레이션을 통하여 민감도 분석을 실시하였다. percentage 비율 group별로 subset data를 10개 씩 구축하였으며, 각각에 대한 보정계수를 산출, 그 값의 평균과 표준편차를 도출하였다. 민감도 분석 결과 R2U(Rural Two lane - Undivided) 도로의 이상적인 보정계수 값은 1.472로 도출되었다. 유의수준 5%일 경우 전체 구간길이의 50% 이상을 대상으로 보정계수를 산출할 때 통계적으로 동일한 값이 산출된다. 유의수준 10%일 경우 전체 구간길이의 20% 이상을 대상으로 보정계수를 산출할

때 통계적으로 동일한 값이 산출되었다. 본 연구의 분석 결과 Rural two lane - undivided highway를 예로 들었을 때 유의수준 10%에서 전체 구간길이의 5~10%의 도로를 대상으로 보정계수를 산출하였을 경우, 보정계수 참값과 통계적으로 동일해 지는 확률이 65~80%로 도출되었다. 5~10%의 percentage는 전체 구간길이와 비교했을 때 약 240~480 mile의 수준이며, 사고건수와 비교했을 때 약 330~660 건/년의 수준으로 분석되었다. 만약 데이터가 전체 구간길이의 20% 이상 이용 가능하다면, 보정계수의 참값과 통계적으로 동일한 값을 산출 가능하다. 따라서 HSM에서 30~50 sites, 100 건/년으로 규정하고 있는 최소표본기준(minimum sample size)은 통계적으로 유의한 보정계수를 산출하는데 부족한 기준인 것으로 제시하고 있다.

Fillip Martinelli et al.(2009)는 different environmental, road characteristics, driver behaviour, crash reporting system을 갖는 타 지역에 대한 HSM accident predictive model의 전이성을 살펴보는 것을 주목적으로 설정하였다. 또한, 보정계수(calibration factors)를 산출하는 4가지 방법론을 비교검토하였다. 본 연구의 공간적 범위는 Italia Arezzo province의 rural two lane highways (1,300km)이며 시간적 범위는 2002 ~ 2004년 총 3개년도이다. geometric data, traffic data, accident data, driveways data를 이용하였다. 분석 과정으로는 먼저 accident predictive model의 형태를 정하였다. HSM에서 정의하는 바와 동일하게 전체 구간을 동질한 다수의 Homogeneous section (Segmentation)을 정의하였으며, 총 938km의 도로를 대상으로 8,379개의 Homogeneous section으로 구분하였다. Section의 구분기준은 1차로 교통량의 범위에 따라 5개의 Class로 구분하고, 기하구조 특성에 따라 차로 폭 기준 4개의 Class, 길어깨 폭 기준 4개의 Class로 구분하였으며, 총 80개의 Class로 동질구간을 분리하였다. 동질구간의 기본 구분기준에 따라 구간을 세분화 하였을 때 동질구간의 평균 길이는 112m이며, 8,379개의 Total section 중 8,106개의 section에서 사고건수가 0건으로 기록되는 Zero-accident 문제가 발생하는 것을 확인하였다.

2. Neural Network 및 딥 러닝을 이용한 안전성능함수

오주택 외(2014)는 교통사고예측모형 구축에 주로 사용되는 회귀모형, 인공 신경망, 구조방정식을 이용하여 교통사고 건수 예측모형을 각각 개발하였다. 구축된 모형의 예측력을 평균절대오차와 평균제곱예측오차를 이용하여 평가하였다. 이 연구에서는 2007년도 자료를 이용하여 전라북도와 공주시의 지방부 국도 인근 4지 신호교차로를 대상으로 도로주변 환경, 교통량, 도로시설을 포함하는 독립변수들과 종속변수로서 교통사고(단순물피사고를 포함한 경상, 중상, 사망의 전체 사고) 자료를 사용하였다. 모형 개발을 위한 90개소의 교차로 데이터와 모형의 검증을 위한 33개소의 교차로 데이터를 수집하였다. 이 연구에서는 신뢰수준 90%에서 7가지 요소(주도로 좌회전 전용 차로수, 부도로 차로수, 주도로 속도, 주도로 횡단보도 유무)가 4지 신호교차로 교통사고에 영향을 미치는 것으로 분석하였다. 또한 음이항 회귀 모형보다 포아송 회귀 모형이 적합한 것으로 분석하였다. 인공 신경망의 경우 모형 구축에 사용할 변수선정을 위해 비선형 회귀모형에서 유의성이 나타난 변수를 종합한 후 다중공선성 분석을 통해 최종 변수를 선정하였다. ANN 구축을 위해서는 관련 문헌에서 ANN 성능에 영향을 끼치는 것으로 제시된 학습률과 은닉층의 노드 수를 달리하는 총 8개의 시나리오 중에서 가장 우수한 MAD와 MAPE 값을 보이는 모형을 선정하였다. 구조방정식은 요인분석을 통해 4개의 요인으로 구분하였으며 구성된 요인을 통하여 모형을 구축하였다. 적합성 검증을 위해 NNFI(TLI), CFI, RMSEA를 기준으로 평가하였다. 구축된 모형을 MAD와 MAPE를 이용하여 평가한 결과, 인공 신경망 모형, 비선형 회귀모형, 구조방정식 모형 순으로 뛰어난 예측력을 보이는 것을 확인하였다.

Guangyuan Pan et al.(2017)은 다양한 지역의 다른 고속도로의 충돌 예상 빈도를 예측하는 데 사용할 수 있는 전 세계적인 도로안전 성능함수(SPF)을 개발하기 위해 머신 러닝 기법을 적용하는 방법에 대해 설명하였다. 가장 인기 있는 딥 러닝 모델 중 하나인 DBN은 crash modeling을 위한 기존 회귀 모델의 대안

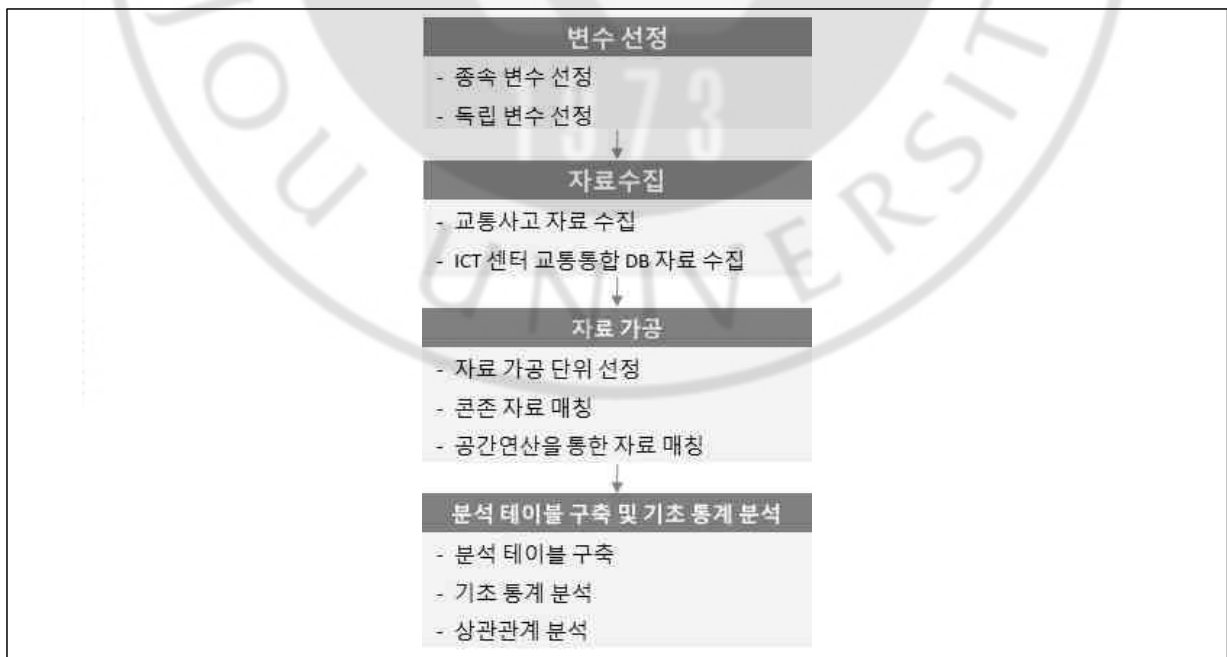
으로 도입되었다. 광범위한 경험적 연구는 위치(도심 vs. 농촌), 차선 수, 액세스 제어 및 지역으로 정의 된 6 가지 고속도로를 다루는 3 가지 실제 충돌 데이터 세트를 사용하였다. 본 연구는 기존 회귀 모델과 비교하여 네트워크 구조, 교육 방법, 데이터 크기 및 일반화 능력과 관련하여 DBN의 상대적인 성능에 관한 몇 가지 중요한 의문을 해결하는 여러 가지 분석을 포함하였다. 분석 결과는 DBN 모델이 다른 충돌 데이터 세트로 훈련 될 수 있고 예측 성능이 로컬로 보정 된 음이항 모형의 성능과 적어도 비교 될 수 있음을 보여주었다.

Helai Huang et al.(2016)는 고속도로 사고로 인해 사회에 엄청난 손실이 발생하면서 교통사고 발생에 영향을 미치는 위험 요인을 파악하는 것은 안전 관련 연구의 중요한 시사점이었다. 본 연구에서는 충돌 주파수와 관련 위험 요인 간의 비선형 관계를 근사화하기 위해 RBFNN(Radial Basis Function Neural Network) 모델을 개발하였다. 본 연구는 RBFNN 모델의 성능을 평가하기 위해 홍콩의 도로 구간에서의 사고 빈도 예측을 하고 기존 음이항 (NB) 방법과 Back-Propagation Neural Network (BPNN) 모델의 성능과 비교분석하였다. 분석 결과는 RBFNN이 NB 및 BPNN 모델보다 더 나은 fitting 및 예측 성능을 가지고 있음을 알 수 있었다. RBFNN이 최적화 된 후에는 근사성능이 향상되지만, 충돌 발생 빈도에 거의 영향을 주지 않는 몇 가지 변수를 발견하였다. 또한, 본 연구에서는 최적화된 RBFNN에서 나머지 입력 변수의 효과를 확인하기 위해 민감도 분석을 수행하였다. 분석 결과, 대부분의 위험 요인과 충돌 빈도가 서로 비선형적인 관계가 있음을 보여 주며 도로 안전 분석을 위해 수정된 RBFNN 모델의 사용을 지원하는 NB 모델보다 위험 요인의 영향에 대한 더 깊은 통찰력을 제공한다.

제3장 딥러닝을 이용한 고속도로 교통사고 예측모형 개발을 위한 자료 수집 및 분석

제1절 자료 수집 및 분석 개요

고속도로 안전성능함수 구축을 위해서는 종속변수와 독립변수를 선정하는 것이 필요하며, 교통사고건수에 영향을 줄 수 있는 독립변수를 선정하여 자료 수집하는 것이 필요하다. 본 연구에서는 기존 문헌고찰을 통해 교통사고에 영향력이 있는 변수를 선정하고, 이에 대한 자료를 수집하였다. 또한 수집된 자료를 고속도로 안전성능함수 구축 단위인 콘존 단위로 가공하였다. 여기서 사용된 안전성능함수 구축 단위인 콘존(Conzone)은 한국도로공사가 고속도로 구간을 IC, Jct, 그리고 TG 등 통행하는 차량수가 일정한 고속도로 구간으로 분류한 개념을 의미한다. 가공된 자료들을 이용하여 분석 테이블을 구축하고 기초 통계분석 및 상관관계 분석을 수행하였다.



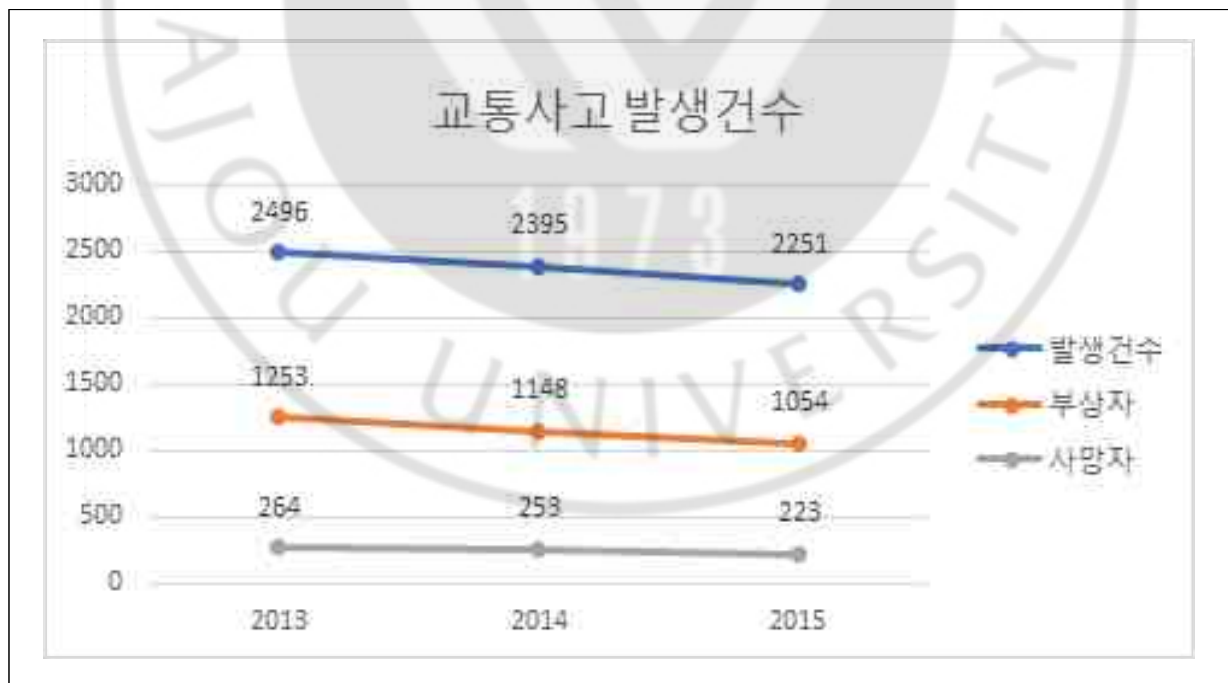
<그림 4> 자료 수집 및 분석 절차

제2절 고속도로 교통사고 현황 분석

교통사고는 인적, 물적, 환경적 요인이 복합적으로 작용하여 발생하는 것으로 교통사고를 통계적 비교분석을 중심으로 교통사고특성과 원인을 분석 검토하고자 한다. 따라서 본 절에서는 교통사고에 대한 종합적 분석을 위해 2013년부터 2015년까지 최근 3년간의 교통사고 자료를 살펴보고자 한다.

1. 고속도로 교통사고 추세 분석

본 연구의 시간적 범위인 3개년(2013~2015년) 동안의 고속도로 전체 교통사고는 2013년 2,496건, 2014년 2,395건, 2015년 2,251건이 발생하였다. 교통사고로 인한 부상자는 2013년 1,253명, 2014년 1,148명, 2015년 1,054명이 고, 사망자는 2013년 264명, 2014년 253명, 2015년 223명을 기록하였다.

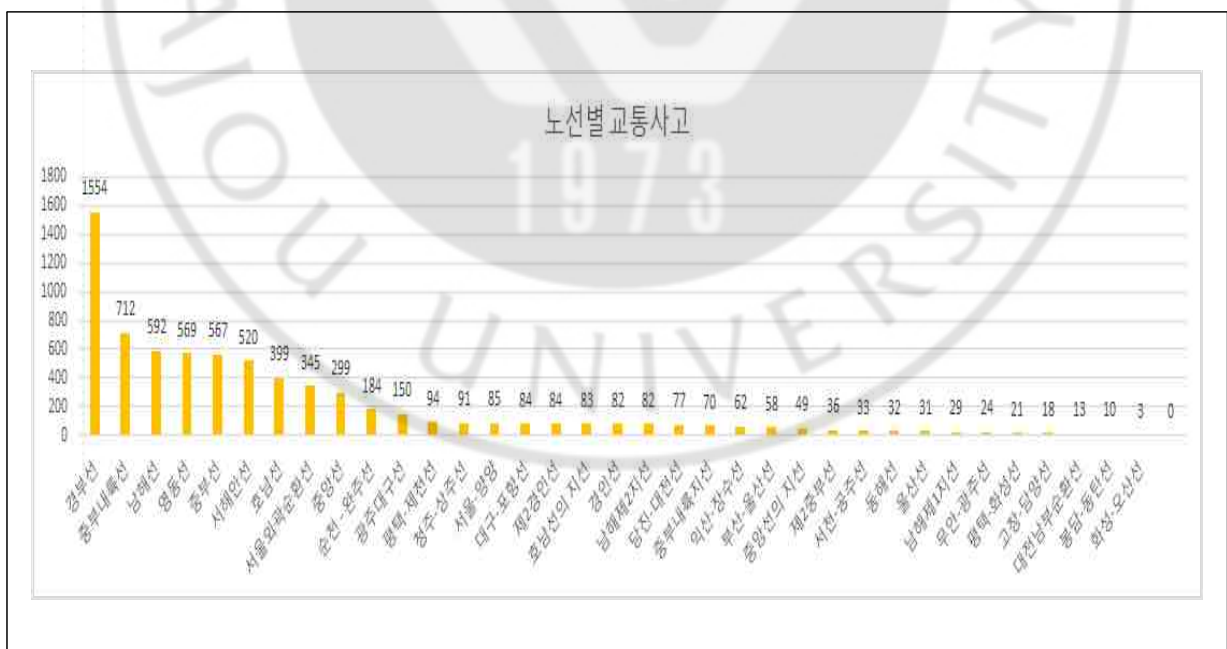


<그림 5> 연도별 교통사고 발생 건수, 사망자 수, 부상자 수

<표 2> 년도별 교통사고 발생 건수, 사망자 수 및 부상자 수

구분	발생건수 (건)		사망자(명)		부상자(명)	
		증감률(%)		증감률(%)		증감률(%)
2013	2,496	-4.0	264	-23.0	1,253	-22.6
2014	2,395	-4.0	253	-4.2	1,148	-8.4
2015	2,251	-6.0	223	-11.9	1,054	-8.2
평균	2,381	-5.0	247	-13.0	1,152	-13.0

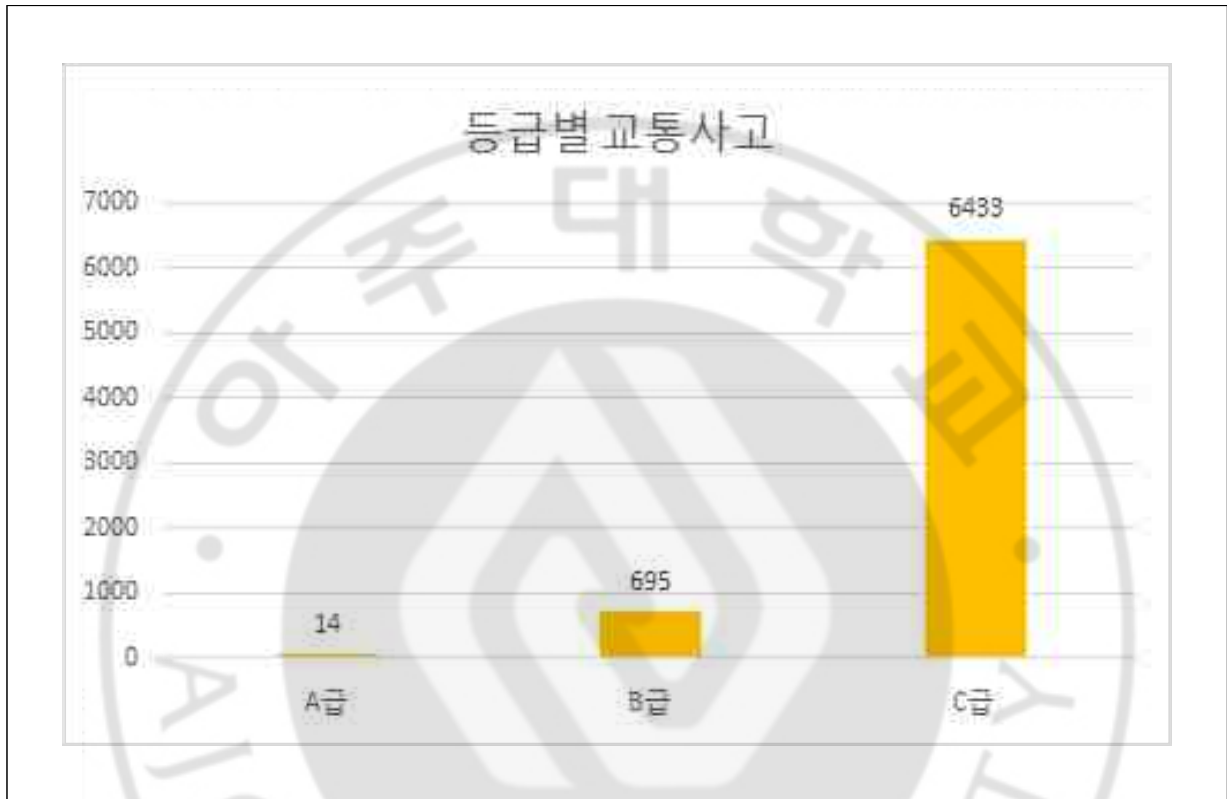
2013년부터 2015년까지 3년간 경부선은 1,554건의 교통사고가 발생하였고 중부내륙선, 중부선, 남해선, 영동선, 중부선, 서해안선, 호남선, 서울외관순환선 등 순으로 교통사고가 발생하였다.



<그림 6> 노선별 교통사고 발생 건수

2. 교통사고 등급별 발생 현황

2013년부터 2015년까지 총 3년간 A급은 14건, B급은 695건 C급은 6,433건 발생하였다.



<그림 7> 교통사고 등급별 발생 현황

<표 3> 년도별 교통사고 등급별 사망사고 및 부상사고 현황

구분	사고등급별 사망자 /부상자 현황					
	A급		B급		C급	
	사망자	부상자	사망자	부상자	사망자	부상자
2013	10	18	254	186	0	1,049
2014	19	15	234	186	0	947
2015	9	9	214	161	0	884
평균	13	14	234	178	0	960

3. 교통사고 위치별 발생 현황

최근 3년 고속도로 교통사고 자료를 이용하여 위치별로 분류하였다. 당연히 본선에서 교통사고가 발생하였고 램프, 터널, TG 순으로 발생하였다.



<그림 8> 교통사고 위치별 발생 현황

<표 4> 년도별 교통사고 위치별 사망자 현황

구분	사고위치별 사망자 현황										
	본선	램프	휴게소	정류장	TG	터널	교량	TG Hi	비상주차대	LCS차로	버스전용차로
2013	236	12	3	0	0	6	6	1	0	0	0
2014	221	10	3	0	6	5	6	1	1	0	0
2015	187	12	7	0	1	12	0	1	1	2	0
평균	215	11	4	0	2	8	4	1	1	1	0

<표 5> 년도별 교통사고 위치별 부상자 현황

구분	사고위치별 부상자 현황										
	본선	램프	휴게소	정류장	TG	터널	교량	TG Hi	비상주차대	LCS 차로	버스전용차로
2013	1,933	239	37	7	105	100	21	43	4	7	0
2014	1,804	262	41	1	88	110	23	44	7	15	0
2015	1,711	222	30	3	82	130	16	36	15	6	0
평균	215	11	4	0	2	8	4	1	1	1	0

사고 위치별로 분류한 자료의 교통사고 건수 대비 사망자수를 이용하여 치사율을 분석해보면 교량에서 사고 100건당 18.3명으로 가장 높았고, 휴게소 12.9명, 본선 11.8명, LCS 차로 11.1명 등을 기록하였다.

<표 6> 년도별 교통사고 위치별 치사율 현황

구분	사고위치별 치사율(사고 100건당 사망자수) (명)										
	본선	램프	휴게소	정류장	TG	터널	교량	TG Hi	비상주차대	LCS 차로	버스전용차로
2013	12.2	5.0	8.1	0	0	6.0	28.6	2.3	0	0	0
2014	12.2	3.8	7.3	0	6.8	4.5	26.3	2.3	14.3	0	0
2015	11.0	5.4	23.3	0	1.2	9.3	0	2.8	6.7	33.3	0
전체	11.8	4.7	12.9	0	2.7	6.6	18.3	2.5	7.0	11.1	0

제3절 변수 선정 및 자료 수집

본 연구에서는 앞서 수행된 고속도로 교통사고 현황 분석과 기존 문헌고찰을 통해 고속도로 안전성능함수 구축을 위한 변수를 선정하였다. 대부분의 관련 연구는 안전성능함수의 종속변수로 교통사고 건수를 사용하였다. 하지만, Salvatore Cafiso et al.(2013)에서는 사고 위험도를 종속변수 사용하였고, 윤일수 외(2012)에서는 교통사고 심각도(EPDO)를 이용하여 고속도로 교통사고 예측모형을 구축하였다. 본 연구에서는 다수의 연구에서 사용하고 있는 교통사고 건수를 안전성능함수의 종속변수로 설정하였다. 독립변수 선정에 위해 관련 연구 고찰 결과, 교통량, 중앙분리대의 유·무, 구간길이, 평면선형, 종단선형, 차로수 등이 사용 되었다. 다음 표는 관련 연구에서 사용된 독립변수를 정리하고 공통적인 변수를 본 연구의 독립변수로 선정하였다.

<표 7> 관련 연구 상의 독립변수 사용 내역

관련 연구	독립변수
이수범 외(2003)	교통량, 중앙분리대의 유·무, 교차로 수, 연결로 수, 횡단 신호등 수, IC 밀도 및 차로수
강동운 외(2014)	AADT, 구간길이, 평면선형, 종단선형, 차로수
이태헌 외(2015)	영업소 AADT, 하이패스 차로 비율, 중차량 혼입률
서임기 외(2015)	교통량, 구간길이
박효신 외(2007)	유출입여부, 곡선장, 교통량, 중차량비율, 곡률
박주환 외(2012)	연평균일교통량, 종단구배, 평면선형, 터널연장, 측방여유폭, 터널 높이, 설계속도, 교통사고 발생시 터널 진입전(진출후) 주야간상태
Jinyan Lu et al.(2013)	연평균일교통량
Bradford K. Brimley et al. (2012)	구간길이, 추월차로의 존재여부, AADT, 제한속도, 콤보유닛 트럭의 구성 비율 등
Yingfei TU (2012)	교통량, 속도, 중차량 비율, 차로수
Hochan Kwak et al.(2010)	기하구조 및 교통량 자료

문헌 고찰을 바탕으로 본 연구에서는 가장 많이 사용되고 있는 고속도로 교통사고 건수를 종속변수로 선정하였다. 또한 독립변수로는 교통량, 구간길이, 설계속도, 제한속도, 차로수 및 곡선반경 등 기하구조 자료와 그 외 시설물 및 구조물인 졸음쉼터, 터널, 교량 등 자료를 독립변수로 선정하였다.

앞서 선정된 변수에 대한 자료를 수집하기 위해 자료 수집단위를 선정하였다. 본 연구에서는 자료 수집 단위를 콘존 단위로 설정하고 자료를 수집하였다. 한국도로공사는 고속도로 구간을 콘존의 개념으로 분류하고 있다. 콘존은 IC, JC, 그리고 TG 등으로 구분된 구간(segment)로서 해당 구간을 통행하는 차량수에 변화가 없다(김상구 외, 2016). 종속변수인 교통사고 건수는 「고속도로 교통사고 속보자료」를 2013~2015년까지 3년간 C급 이상 사고에 대해 수집하였다⁹⁾. 「고속도로 교통사고 속보자료」에는 A~D 등급의 교통사고에 대해 사고위치, 사고발생지점, 사고원인, 주 사고원인 등 정보를 상세히 제공하고 있다.

독립변수에 사용될 자료를 수집하기 위해 한국도로공사에서 발간하는 2013~2015년 「고속도로 교통량」 자료 및 한국도로공사 ICT 센터가 보유한 교통통합 DB의 졸음쉼터 위치, 휴게소 위치, 교량 위치, 기하구조 자료 등을 수집하였다. 수집한 자료는 다음 표와 같다.

9) 한국도로공사의 자체적인 교통사고 분류 기법은 A, B, C 그리고 D 이다. 분류 기준은 교통사고의 피해정도이다. D급은 경미한 물피사고를 의미하며, C급은 부상 1명 이상, 도로시설물 피해액 30만원 이상, 관련 차량 3대 이상, 1개 차로 차단 사고를 말한다(최윤희, 2012).

<표 8> 모형 구축을 위한 종속변수 및 독립변수 선정

구분	보유 자료	자료 및 테이블명
종속변수	교통사고 건수	고속도로 교통사고 속보자료
독립변수	콘존길이	콘존 마스터 테이블
	줄음쉼터 개수	C_COMD_줄음쉼터
	휴게소개수	C_COMD_휴게소
	설계속도	C_COMD_콘존부가정보
	제한속도	C_COMD_콘존부가정보
	차로수	C_COMD_콘존부가정보
	갓길존재여부	C_COMD_콘존부가정보
	진입조절구간여부	C_COMD_콘존부가정보
	버스전용차로구간여부	C_COMD_콘존부가정보
	터널	T_BSFA_터널위치
	교량	T_BSFA_교량위치
	시설물	T_BSFA_시설물좌표
	AADT	고속도로 교통량 책자
	곡선반경길이	T_BSFA_도로기하정보
	종단경사값	T_BSFA_도로기하정보

제4절 수집 자료 가공

1. 자료 가공 단위 개요

수집된 자료들은 콘존 단위로 되어 있는 자료와 아닌 자료들이 존재하기 때문에, 콘존 단위로 가공이 필요하다. 수집한 자료들을 콘존 단위로 가공하기 위해서 콘존 마스터 테이블을 이용하였다. 콘존 마스터 테이블은 콘존 정보를 보유하고 있는 테이블로서, 콘존ID, 콘존길이, 기종점방향구분코드, 차로수, 노선번호 등 정보를 보유하고 있으며, 구성은 다음 그림과 같다.

콘존ID	콘존길이	기종점방향구분코드	시작노드ID	종료노드ID	차로수	노선번호	제한속도	노선구성순번	콘존명	버스전용차로유무	도로등급구분코드
0010CZE005	209	E	491	4	3	10	100	1	경부고속국도시점-구서IC	0	101
0010CZE010	1820	E	4	446	3	10	100	2	구서IC-영광IC	0	101
0010CZE011	1990	E	446	486	3	10	100	3	영광IC-부산TG	0	101
0010CZE020	1070	E	486	447	3	10	100	4	부산TG-노포	0	101
0010CZE030	7780	E	447	155	3	10	100	5	노포IC-왕산IC	0	101
0010CZE040	5080	E	155	453	4	10	100	6	왕산IC-왕산IC	0	101
0010CZE050	14110	E	453	652	3	10	100	7	왕산IC-통도시비	0	101
0010CZE055	1050	E	652	160	3	10	100	8	통도시비-서불산IC	0	101
0010CZE060	6430	E	160	203	3	10	100	9	통도시IC-서불산IC	0	101
0010CZE070	1630	E	203	204	3	10	100	10	서불산IC-영광IC	0	101

<그림 9> 콘존 마스터 예시

2. 콘존 매칭을 통한 자료 가공

콘존 부가정보 테이블은 콘존 ID를 기준으로 정리가 되어 있기 때문에, 콘존 ID를 Key값으로 하여 콘존 마스터와 매칭을 하였다. 하지만 교통사고 속보자료, 졸음쉼터자료, 휴게소 자료의 경우 콘존 ID가 없이, 노선명, 기종점방향, 이 정 자료가 구성되어 있어 이를 매칭 Key로 콘존 마스터와 매칭하였다.

```
# coding: utf-8
import pandas as pd
import numpy as np
data=pd.read_excel("D:/Code/Python_Code/CZ/data_7.xlsx", sheetname=0, header=0)
CZ=pd.read_excel("D:/Code/Python_Code/CZ/CONZONE_MASTER_NEW.xlsx", sheetname=0, header=0)
for i in range(0,len(data)):
    s=CZ.loc[CZ['ROUTE_NO']==data.ROUTE_NO[i]]
    b=s.loc[s['UPDOWN_DIV']==data.UPDOWN_DIV[i]]
    c=b.loc[b['FROM_MILEPOST'] <= data.FROM_MILEPOST[i]]
    d=c.loc[c['TO_MILEPOST'] > data.TO_MILEPOST[i]]
    if len(d)!=0:
        data.CONZONE_ID[i]=d.iloc[0]['CONZONE_ID']
    else:
        data.CONZONE_ID[i]=0
    print(i)
data.to_csv('D:/Code/Python_Code/CZ/result.csv', index=False, header=True)
```

<그림 10> 콘존 매칭 코드

3. 공간 연산을 통한 자료 가공

교량 위치, 시설물 위치, 터널 위치 자료의 경우 이점 정보를 가지지 않은 대신 GPS 좌표를 가지고 있다. 그러므로 콘존 ID가 포함된 전자지도 파일(shp)을 이용하여 공간연산을 수행하여 콘존 ID를 매칭 해야 한다. 본 연구에서는 ArcGIS 10.3의 Spatial Join 기능을 이용하여 공간연산을 하여 자료를 가공하였다.



<그림 11> ArcGIS Spatial Join 결과

제5절 분석 테이블 구축 및 기초 통계분석

1. 분석 테이블 구축

자료 가공을 통하여 모든 자료들은 콘존 ID를 기준으로 매칭 가능하도록 되었으며, 이를 콘존 마스터를 기준으로 매칭하여 분석테이블을 구축하였다. 분석테이블을 구축한 결과 종속변수와 독립변수를 포함하여 총 16개의 변수로 구성되었으며, 총 978개 데이터로 구축되었다. 분석테이블은 콘존 ID를 기준으로 콘존길이, AADT, 졸음쉼터개수 등 독립변수를 먼저 작성하고 종속변수인 사고건수를 마지막에 작성하였다. 다음 그림은 분석 테이블 구축 결과를 나타낸 그림이다.

CONZONE ID	CONZONE 길이(m)	AADT	졸음쉼터 개수	유계소 개수	설계속도	제한속도	차로수	차량종류 대부	신입조형 구간 여부	버스전용 차로여부	터널 개수	교량개수	시설물 수	곡선반경	중단경사	사고건수
0010C-ZE005	200	30961	0	0	100	100	3	0	0	0	0	0	0	0.00	1.41	0
0010C-ZE010	1830	30961	0	0	100	100	3	0	0	0	0	1	5	719.44	0.38	1
0010C-ZE011	1990	30961	0	0	100	100	3	0	0	0	0	0	0	386.84	-0.37	0
0010C-ZE020	1070	30961	0	0	100	100	3	0	0	0	0	2	5	155.00	1.07	1
0010C-ZE030	7780	36037	0	0	100	100	3	0	0	0	0	8	16	213.33	-0.29	13
0010C-ZE040	5050	40763	0	0	100	100	4	0	0	0	0	2	5	321.57	-0.10	4
0010C-ZE050	14130	32962	0	0	100	100	3	0	0	0	0	7	9	1631.52	0.47	15
0010C-ZE060	6430	30635	1	0	100	100	3	0	0	0	0	4	6	3015.63	-0.35	18
0010C-ZE070	1630	29223	0	0	100	100	3	0	0	0	0	5	3	346.25	-0.63	5
0010C-ZE080	20150	22271	0	1	100	100	3	0	0	0	0	6	3	687.70	-0.04	23
0010C-ZE090	13400	70893	0	1	100	100	3	0	0	0	0	6	3	699.13	0.23	8
0010C-ZE100	17730	21213	0	0	100	100	3	0	0	0	2	6	3	594.55	0.05	17
0010C-ZE110	15450	24247	0	1	100	100	3	0	0	0	0	13	20	1106.23	-0.30	17
0010C-ZE120	9280	41459	0	0	100	100	4	0	0	0	0	7	10	0.00	0.05	5
0010C-ZE130	4090	58076	0	0	100	100	4	0	0	0	0	4	6	97.56	0.30	5
0010C-ZE140	7630	73212	1	0	100	100	4	0	0	0	0	6	16	772.73	-0.50	14
0010C-ZE150	5480	74939	0	0	100	100	4	0	0	0	0	12	13	1282.61	0.02	7
0010C-ZE160	10740	60834	0	0	100	100	4	1	0	0	0	11	24	546.74	0.53	24
0010C-ZE170	3130	59967	0	0	100	100	4	0	0	0	0	2	2	342.86	0.47	9
0010C-ZE180	13930	52324	1	1	100	100	4	0	0	0	0	14	23	422.76	-0.15	20
0010C-ZE190	5070	43916	0	0	100	100	4	0	0	0	0	3	2	147.60	0.12	7
0010C-ZE200	32860	40172	1	0	100	100	4	0	0	0	0	7	7	420.00	0.32	19
0010C-ZE210	9130	19772	1	0	100	100	3	0	0	0	0	11	17	296.90	-0.23	6
0010C-ZE215	7930	38562	0	0	100	100	3	0	0	0	0	6	8	516.88	0.11	8
0010C-ZE220	12030	17346	1	1	100	100	3	0	0	0	0	11	24	484.85	0.98	10
0010C-ZE230	17050	17749	1	0	100	100	3	0	0	0	0	2	4	476.05	-0.36	11
0010C-ZE240	9880	37448	0	1	100	100	3	0	0	0	0	6	7	443.70	0.31	4
0010C-ZE250	16040	18134	0	1	100	100	3	0	0	0	12	10	28	1189.63	-0.23	11
0010C-ZE260	11400	18374	0	1	100	100	3	1	0	0	3	3	3	236.84	-0.05	15
0010C-ZE270	8670	25928	1	0	100	100	3	0	0	0	1	4	10	335.16	-0.26	3
0010C-ZE280	4080	36997	0	0	100	100	3	0	0	0	0	2	3	450.00	1.17	4
0010C-ZE290	5450	42796	0	0	100	100	3	0	0	0	0	1	2	375.56	0.44	10
0010C-ZE300	5030	38496	0	1	100	100	4	0	0	0	0	1	10	415.73	-0.47	5
0010C-ZE310	10070	56347	0	1	100	100	4	0	0	0	0	11	21	274.22	-0.04	9
0010C-ZE320	3530	46924	0	0	100	100	4	0	0	1	0	2	2	63.74	0.24	3
0010C-ZE330	3630	65723	0	0	100	100	3	0	0	1	0	2	1	956.44	0.24	3
0010C-ZE340	4630	35580	0	0	100	100	3	0	0	1	0	2	2	463.83	-0.63	1
0010C-ZE350	24800	41970	0	2	100	100	3	0	0	0	1	0	9	267.29	0.09	87
0010C-ZE360	3530	42508	0	0	100	100	3	0	0	1	0	1	8	245.14	0.27	18
0010C-ZE370	6700	73866	0	1	100	100	4	1	0	1	0	2	6	446.21	-0.19	37
0010C-ZE380	10760	82642	0	0	100	110	4	1	0	0	0	15	17	639.76	-0.16	32
0010C-ZE385	9210	81202	0	1	100	110	4	1	0	1	0	4	11	570.68	-0.14	45
0010C-ZE390	4830	82387	0	0	100	110	4	0	0	1	0	1	5	340.25	0.46	8
0010C-ZE400	13270	78523	0	1	100	110	4	1	0	1	0	5	4	117.42	-0.02	28
0010C-ZE410	3930	85296	1	0	100	110	4	1	0	1	0	0	0	36.49	0.32	11
0010C-ZE420	4390	87175	0	0	100	110	4	0	0	1	0	0	0	417.60	-0.04	10
0010C-ZE430	700	82135	0	0	100	110	5	0	0	1	0	1	0	550.29	0.94	2
0010C-ZE440	5030	103187	0	0	100	110	5	1	0	1	0	4	12	127.91	-0.09	6
0010C-ZE450	2630	107010	0	0	100	110	5	0	0	1	0	5	7	300.00	0.09	9
0010C-ZE460	8760	106240	0	1	100	110	5	0	0	1	0	10	16	0.00	0.07	20
0010C-ZE470	3220	106240	0	0	100	110	5	0	0	1	0	8	5	0.00	0.34	6
0010C-ZE480	1100	106240	0	0	100	110	5	0	0	1	0	15	7	207.00	0.76	6
0010C-ZE490	390	96044	0	0	100	110	5	0	0	1	0	0	0	666.67	-1.39	0
0010C-ZE500	7630	96044	0	0	100	110	5	1	0	1	0	4	10	101.97	0.15	22
0010C-ZE510	7620	87110	0	1	100	110	5	0	0	1	0	4	10	107.24	0.00	8

<그림 12> 분석 테이블 구축 결과

2. 기초 통계분석

콘존 단위로 가공된 16개 변수에 대하여, 평균, 표준편차, 최소값, 최대값의 기초 통계분석을 실시하였다.

<표 9> 수집 자료에 대한 기초 통계분석 결과

구분	평균	표준편차	최소	최대
사고건수	6.39	7.49	0.00	87.00
콘존길이(m)	7,379.37	5,408.78	110.00	30,790.00
교통량(AADT)	31,948.58	26,720.75	542.00	118,601.00
줄음선평터 개수	0.11	0.31	0.00	1.00
휴게소 개수	0.18	0.39	0.00	2.00
설계속도	103.27	8.81	80.00	120.00
제한속도	101.02	6.56	70.00	110.00
차로수	2.56	0.83	2.00	6.00
갓길차로 여부	0.03	0.18	0.00	1.00
진입조절 구간 여부	0.00	0.06	0.00	1.00
버스전용차로 여부	0.04	0.20	0.00	1.00
터널 개수	0.15	0.95	0.00	15.00
교량 개수	4.68	4.21	0.00	23.00
시설물 수	5.47	6.85	0.00	41.00
곡선반경	851.50	1,264.14	0.00	27,547.27
종단경사	0.01	0.54	-2.31	3.01

제4장 고속도로 교통사고 예측모형 구축

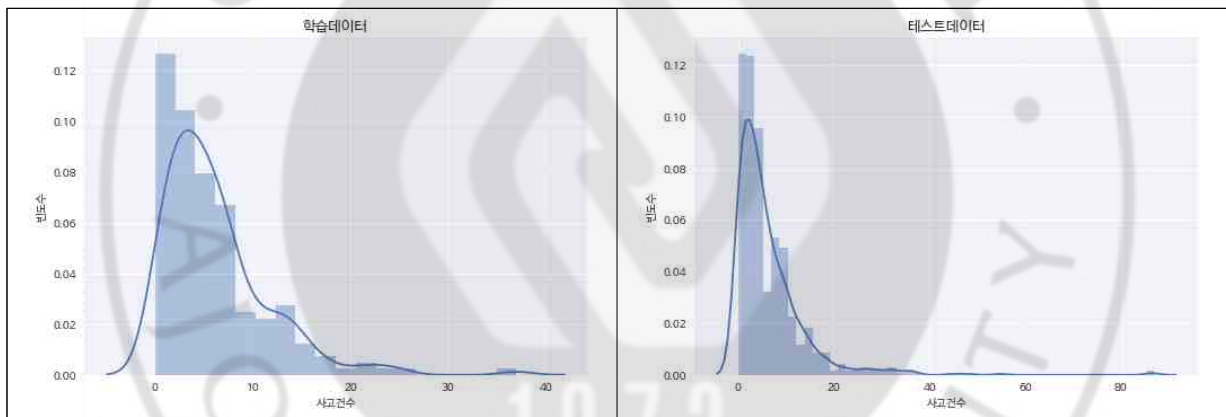
제1절 고속도로 교통사고 예측모형 구축 방법론

1. 개요

고속도로 교통사고 예측모형을 구축하기 위해서는 독립변수로 수집된 자료와 교통사고 건수의 관계를 모형화 할 수 있어야 한다. 본 연구에서는 교통사고 건수 추정에 사용되는 전통적인 통계 모형인 포아송 회귀모형과 음이항 회귀모형을 이용하여 고속도로 교통사고 예측모형을 구축하였다. 또한 독립변수의 특성에 따라 콘존을 유형화 하여 포아송 회귀모형과 음이항 회귀모형으로 고속도로 교통사고 예측모형을 구축한다. 마지막으로 최근 활발하게 사용되는 머신러닝 기법인 딥 러닝을 이용하여 고속도로 교통사고 예측모형을 구축하고 구축된 모형들에 의해 예측되는 사고 건수와 실제 사고건수를 비교하여 각 모형의 성능을 평가하고자 한다. 본 연구에서는 안전성능함수 구축 전 분석 테이블을 학습 데이터(학습 data)와 테스트 데이터(테스트 data)로 구분하였다. 학습 데이터는 전체 데이터의 80%를 사용하였으며, 테스트 데이터는 나머지 20%를 이용하였다. 본 연구에서는 안전성능함수 구축을 위해 통계분석, 머신러닝, 딥 러닝 구현이 우수한 프로그래밍 언어인 파이썬 ver 3.5를 이용하였다. 안전성능함수들을 구축하고 안전성능함수의 성능을 비교하기 위한 MOE로 MAD, SMAPE, RMSE를 사용하여 비교하였다.

2. 데이터 구분

전통적인 통계 방법의 고속도로 교통사고 예측모형을 구축하기에 앞서 구축된 분석테이블을 모형을 구축하는 학습데이터와 결과를 검증하는 테스트 데이터로 구분하였다. 본 연구에서는 학습데이터는 전체 데이터의 80%, 테스트 데이터는 20%로 선정하였다. 학습데이터와 테스트 데이터를 구분하는데 랜덤샘플링을 통해 분리하였으며, 파이썬 라이브러리인 Scikit-Learn의 명령어를 이용하여 구분하였다. 구분한 결과 학습데이터는 781건이고, 테스트데이터는 196건이었다. 또한 학습데이터와 테스트 데이터의 종속변수인 사고건수는 유사한 분포를 가지는 것으로 나타났다.



(a) 학습데이터

(b) 테스트 데이터

<그림 13> 데이터 구분 결과

3. 모형 검증 방법론

분리한 테스트 데이터를 이용하여 구축된 모형의 예측력을 검증하는 것이 필요하다. 모형의 예측력을 검증하는데 절대평균편차(Mean Absolute Deviation, MAD), 대칭절대퍼센트 오차평균(Symmetric Mean Absolute Percentage Error, SMAPE), 평균제곱오차의 제곱근(Root Mean Square Error, RMSE)를 사용하여 모형간의 예측력을 비교한다.

가. 절대평균편차(MAD)

실제 사고건수와 모형을 통해 예측된 사고건수를 비교하기 위한 방법으로 실제 사고건수와 예측된 사고건수의 차에 절댓값을 구하여 산술평균한 것을 의미한다. 산출 방식은 다음 수식과 같다. MAD값이 작을수록 모형의 설명력이 높다고 판단할 수 있다.

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| \quad (4.1)$$

나. 대칭 절대 퍼센트 오차 평균(SMAPE)

통계적 기법으로 예측한 값에 대하여 정확도를 측정하는 방법인 MAPE의 단점을 보완한 방법으로 백분율 또는 상대 오류를 기반으로 하는 정확도 측정방법이다.

$$SMAPE = \frac{\sum_{t=1}^n |F_t - A_t|}{\sum_{t=1}^n (A_t + F_t)} \quad (4.2)$$

다. 평균제곱오차의 제곱근(RMSE)

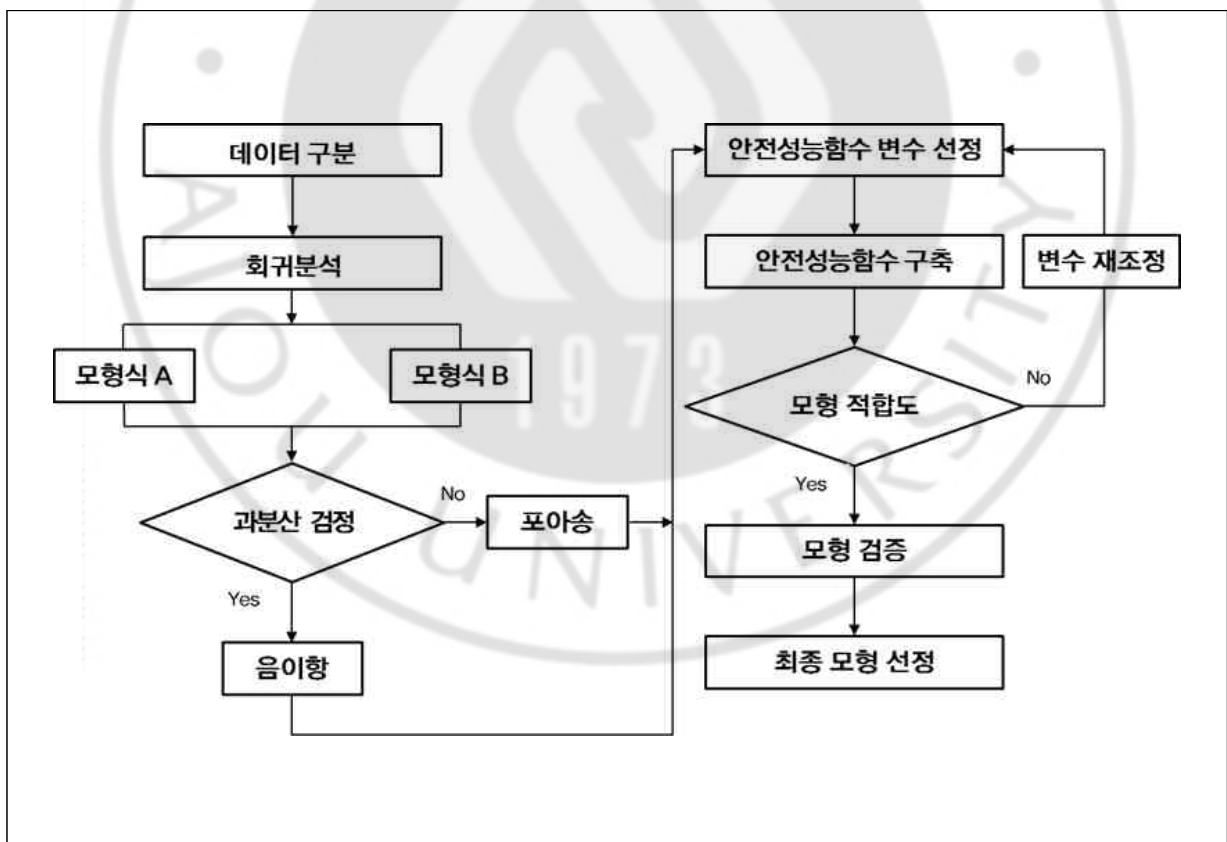
고속도로 교통사고 예측모형을 통해 예측된 값과 실제 사고 건수의 차이를 다루는 측도로서 정밀도를 표현하는데 적합하다. 각각의 차이는 잔차라고 하며, 평균제곱근 잔차들을 하나의 측도로 종합할 때 사용된다.

$$RMSE = \sqrt{MSE(\theta_1, \theta_2)} = \sqrt{E((\theta_1 - \theta_2)^2)} = \sqrt{\frac{\sum_{i=1}^n (\chi_{1,i} - \chi_{2,i})^2}{n}} \quad (4.3)$$

제2절 전통적인 통계 방법의 고속도로 교통사고 예측모형 구축

1. 구축 절차

전통적인 통계 방법의 고속도로 교통사고 예측모형을 구축하기 위해 앞서 구분된 데이터 중 학습데이터를 이용하여 회귀분석을 모형식 A와 모형식 B에 대해 수행하였다. 이를 바탕으로 과분산 검증을 수행하여 과분산이 아니면 포아송 회귀모형을 과분산이면 음이항 회귀모형을 적용한다. 선택된 회귀모형에 적용할 안전성능함수 변수를 선정하여 고속도로 교통사고 예측모형을 구축하고 모형 적합도를 확인하여 적합하지 않은 경우 변수를 재조정하고 적합도를 통과하면 모형 검증 단계를 거쳐 최종 모형을 선택하게 된다.



<그림 14> 구축절차

2. 모형식 종류

고속도로 교통사고 예측모형을 구축하기 위해 전통적인 통계 모형 방법인 포아송 회귀모형과 음이항 회귀모형을 이용하였다. 문헌 고찰을 통해 고속도로 교통사고 예측모형을 구축하는데 있어 모형식의 종류는 다양한 형태로 구축하고 있었다. 일반적인 방법으로는 노출계수에 대한 고려 없이 모든 독립변수들이 Exponential 함수의 지수 형태로 반영되고 있는 모형식 A 형태이다.

$$Y_A = e^{(\alpha + \beta_1 AADT + \beta_2 Conzonelength + \beta_3, \dots)} \quad (4.4)$$

Y_A : 교통사고 예측 건수

$AADT$: 해당 구간의 연평균일평균 교통량(대/일)

$Conzonelength$: 콘존길이

α, β : 회귀 계수

또한, $AADT$ 와 구간길이 같은 노출계수를 로그-변환하여 사용하고 있는 B 형태이다.

$$Y_B = AADT^{\beta_1} \times Conzonelength^{\beta_2} \times e^{(\alpha + \beta_3, \dots)} \quad (4.5)$$

Y_B : 교통사고 예측 건수

$AADT$: 해당 구간의 연평균일평균 교통량(대/일)

$Conzonelength$: 콘존길이

α, β : 회귀 계수

3. 다중공선성 분석

모형 구축 전 독립변수 간에 서로 영향을 미치는지 확인하기 위해 다중공선성 분석을 수행하였다. 다중공선성 분석은 분산팽창지수(Variance inflation factor, VIF)를 이용하여 체크하며, 통계적으로 VIF가 10 이하이면 독립변수 간에 다중공선성이 없어 변수를 제거하지 하고 다중회귀분석 수행이 가능하다. 분석 결과, 구축된 데이터의 모든 변수에는 VIF가 10이하로 다중공선성이 없는 것으로 파악되었다. 이로서 모든 독립변수들은 다른 독립변수 간에 상관 정도가 낮아 고속도로 교통사고 예측모형 구축 시 부정적 영향이 없을 것으로 판단된다.

<표 10> 독립변수 다중공선성 분석 결과

독립변수	VIF
콘존길이	2.06
AADT	3.18
줄음쉼터 개수	1.13
휴게소 개수	1.27
설계속도	3.31
제한속도	3.39
차로수	2.92
갓길차로 여부	1.14
진출입제한 여부	1.05
버스 전용차로 여부	1.48
터널 개수	1.10
교량 개수	1.57
시설물 개수	1.50
곡선반경 길이	1.03
종단경사 길이	1.00

4. 과분산 검정

가. 개요

포아송 회귀모형은 종속변수의 분산이 일정한 경우 수행하고, 분산이 일정하지 않고 과분산이면 음이항 회귀 모형을 수행한다(이일현, 2014). 구축된 데이터의 과분산을 검정하여 적합한 회귀모형 선정을 위해서 포아송 회귀 모형과 음이항 회귀모형을 각각 구축한 후 우도비 검정통계량(Likelihood Ratio, LR)을 이용하여 과분산을 검정하여야 한다(이일현, 2014). 본 연구에서는 포아송 회귀모형과 음이항 회귀모형을 구축하여 모형식 A와 모형식 B에 대하여 각각 구축 후 비교하여 모형식 A와 B에 적합한 회귀모형을 선택하였다.

나. 모형식 A

<표 11>은 포아송 회귀모형에 노출계수를 고려하지 않은 모형식 A를 적용하여 추정한 고속도로 교통사고 예측모형을 구축한 결과이다. 분석결과 유의수준 0.01에서 콘존길이, AADT, 졸음쉼터개수, 휴게소 개수, 차로수, 갓길존재 여부, 버스전용차로 여부, 시설물 개수, 곡선반경길이가 유의한 것으로 나타났다. 유의수준 0.05에서는 설계속도가 유의한 것으로 나타났으며, 유의수준 0.1에서는 제한속도가 유의한 것으로 나타났다. 계수의 부호를 살펴본 결과, 차로수가 적을수록 사고발생빈도가 늘어났으며 진입조절구간이 운영되는 곳에서 사고발생빈도가 늘어났으며, 터널 개수가 적을수록, 종단 경사 값이 (-)로 내리막 경사가 클수록 사고발생빈도가 늘어나는 것으로 분석되었다.

<표 11> 포아송 회귀모형 A의 분석 결과

모형	포아송 회귀모형 A			
관측수	781			
Log-Likelihood	-2482.7			
AIC	4997.4			
BIC	-2530.1			
변수	계수	SE	p-value	유의수준
상수	-0.306	0.244	0.209	
콘존길이	0.000	0.000	0.000	***
AADT	0.000	0.000	0.000	***
줄음쉼터 개수	0.278	0.041	0.000	***
휴게소개수	0.265	0.033	0.000	***
설계속도	0.006	0.003	0.029	**
제한속도	0.006	0.004	0.094	*
차로수	-0.250	0.033	0.000	***
갓길존재여부	0.368	0.057	0.000	***
진입조절구간여부	-0.333	0.215	0.121	
버스전용차로여부	0.504	0.063	0.000	***
터널 개수	-0.007	0.012	0.568	
교량 개수	0.000	0.004	0.916	
시설물 개수	0.010	0.002	0.000	***
곡선반경길이	0.000	0.000	0.006	***
종단경사값	-0.010	0.013	0.438	

주 : *유의수준 0.1, **유의수준 0.05, ***유의수준 0.01

<표 12>은 음이항 회귀 모형에 모형식 A를 적용하여 고속도로 교통사고 예측모형을 구축한 결과로 유의수준 0.01에서 콘존길이, AADT, 제한속도, 시설물 수가 유의한 것으로 나타났다. 유의수준 0.05에서는 차로수, 진입조절구간여부가 유의한 것으로 나타났으며, 유의수준 0.1에서는 줄음쉼터 개수가 유의한 것으로 나타났다. 계수의 부호를 살펴본 결과, 차로수가 적을수록 사고발생빈도가 늘어났으며 진입조절구간이 운영되는 곳에서 사고발생빈도가 늘어났으며, 종단경사 값이 (-)로 내리막 경사가 클수록 사고발생빈도가 늘어나는 것으로 분석되었다.

<표 12> 음이항 회귀모형 A의 분석 결과

모형	음이항 회귀 모형 A			
관측수	781			
Log-Likelihood	-2122.50			
AIC	4276.90			
BIC	-4601.37			
변수	계수	S.E	p-value	유의수준
상수	-0.554	0.513	0.280	
콘존길이	0.000	0.000	0.000	***
AADT	0.000	0.000	0.007	***
줄음쉼터 개수	0.266	0.098	0.051	*
휴게소개수	0.159	0.081	0.452	
설계속도	0.005	0.006	0.377	
제한속도	0.007	0.008	0.000	***
차로수	-0.255	0.061	0.017	**
갓길존재여부	0.389	0.162	0.461	
진입조절구간여부	-0.305	0.414	0.017	**
버스전용차로여부	0.396	0.165	0.809	
터널	0.007	0.029	0.180	
교량	0.012	0.009	0.174	
시설물 수	0.007	0.005	0.000	***
곡선반경길이	0.000	0.000	0.203	
종단경사값	-0.001	0.017	0.937	

주 : *유의수준 0.1, **유의수준 0.05, ***유의수준 0.01

과분산 검정을 실시한 결과, LR은 720.4($p < 0.001$)으로 유의하게 나타나 구축된 데이터는 과분산이 있는 것으로 판정되었다. 따라서 포아송 회귀분석보다는 음이항 회귀분석이 적합한 것으로 나타났다. 또한 모형의 비교 평가에서도 AIC와 BIC 값이 음이항 회귀 모형이 낮게 나타나 모형식 A에서는 음이항 회귀 모형으로 사고 건수를 예측하는 모형이 바람직하다.

다. 모형식 B

<표 13>은 포아송 회귀모형에 모형식 B를 적용하여 추정한 고속도로 교통사고 예측모형을 구축한 결과로 유의수준 0.01에서 콘존길이, AADT, 휴게소 개수, 버스 전용차로 여부가 유의한 것으로 나타났다. 유의수준 0.05에서는 차로수, 갓길 존재여부, 곡선반경 길이가 유의한 것으로 나타났으며, 유의수준 0.1에서는 졸음쉼터 개수가 유의한 것으로 나타났다. 계수의 부호를 살펴본 결과, 차로수가 적을수록 사고발생빈도가 늘어났으며, 제한속도가 낮을수록, 종단경사값이 (-)로 내리막 경사가 클수록 사고발생빈도가 늘어나는 것으로 분석되었다.

<표 13> 포아송 회귀모형 B의 분석 결과

모형	포아송 회귀 모형 B			
관측수	781			
Log-Likelihood	-2156.2			
AIC	4344.3			
BIC	-3183.2			
변수	계수	S.E	p-value	유의수준
상수	-15.357	0.534	0.000	***
콘존길이	1.018	0.032	0.000	***
AADT	0.795	0.032	0.000	***
졸음쉼터 개수	0.074	0.042	0.077	*
휴게소개수	0.148	0.032	0.000	***
설계속도	0.002	0.003	0.547	
제한속도	-0.003	0.004	0.451	
차로수	-0.247	0.028	0.000	**
갓길존재여부	0.128	0.058	0.026	**
진입조절구간여부	0.285	0.213	0.181	
버스전용차로여부	0.514	0.061	0.000	***
터널	0.008	0.013	0.542	
교량	0.001	0.004	0.723	
시설물	0.002	0.002	0.430	
곡선반경길이	0.000	0.000	0.025	**
종단경사값	-0.007	0.012	0.572	

주 : *유의수준 0.1, **유의수준 0.05, ***유의수준 0.01

<표 14>은 음이항 회귀모형에 모형식 B를 적용하여 추정한 고속도로 교통 사고 예측모형을 구축한 결과로 유의수준 0.01에서 콘존길이, AADT, 차로수, 버스전용차로 여부가 유의한 것으로 나타났다. 유의수준 0.1에서는 휴게소 개수, 교량 개수가 유의한 것으로 나타났다. 계수의 부호를 살펴본 결과, 차로수가 적을수록 사고발생빈도가 늘어났으며, 제한속도가 낮을수록, 종단경사 값이 (-)로 내리막 경사가 클수록 사고발생빈도가 늘어나는 것으로 분석되었다.

<표 14> 음이항 회귀모형 B의 분석 결과

모형	음이항 회귀 모형 B			
관측수	781			
Log-Likelihood	-2075.5			
AIC	4182.9			
BIC	-4695.3			
변수	계수	S.E	p-value	유의수준
상수	-13.232	0.874	0.000	***
콘존길이	0.870	0.051	0.000	***
AADT	0.801	0.057	0.000	***
줄음쉼터 개수	0.095	0.099	0.339	
휴게소개수	0.150	0.080	0.061	*
설계속도	0.003	0.006	0.609	
제한속도	-0.005	0.009	0.577	
차로수	-0.271	0.054	0.000	***
갓길존재여부	0.173	0.164	0.290	
진입조절구간 여부	0.173	0.411	0.674	
버스전용차로 여부	0.451	0.167	0.007	***
터널 개수	0.026	0.029	0.366	
교량 개수	0.015	0.009	0.088	*
시설물 개수	0.003	0.005	0.629	
곡선반경길이	0.000	0.000	0.193	
종단경사값	-0.003	0.017	0.863	

주 : *유의수준 0.1, **유의수준 0.05, ***유의수준 0.01

포아송 회귀 모형과 음이항 회귀 모형에 모형식 B를 적용하여 과분산 검정을 실시한 결과, LR은 161.4($p < 0.001$)으로 유의하게 나타나 데이터가 과분산으로 판정되었다. 따라서 포아송 회귀분석보다는 음이항 회귀분석이 적합한 것으로 나타났다. 또한 모형의 비교 평가에 사용되는 AIC와 BIC 값이 음이항 회귀분석이 낮게 나타나 모형식 B에서도 음이항 회귀분석으로 사고건수를 예측하는 모형을 선택하였다.



5. 최종 안전성능함수 구축 및 검증

가. 최종 안전성능함수 구축

최종 안전성능함수는 음이항 회귀모형을 이용하여 구축하였으며, 모형식 B에서 노출계수로 사용한 AADT와 콘존 길이를 기본 변수로 나머지 변수의 가능한 모든 조합을 산출하였다. 모형식 A의 경우 콘존길이, AADT, 줄음쉼터 개수, 휴게소 개수, 설계속도, 차로수, 버스전용차로 여부, 시설물 수가 최종 변수로 선정되었다. 선정된 변수를 이용하여 안전성능함수로 구축한 결과는 다음과 같다.

<표 15> 음이항 회귀모형 A 도출 결과

모형	음이항 회귀 모형 A			
관측수	781			
Log-Likelihood	-2125.6			
AIC	4269.2			
BIC	-4641.7			
변수	계수	S.E	p-value	유의수준
상수	-0.266	0.397	0.503	
콘존길이	0.000	0.000	0.000	***
AADT	0.000	0.000	0.000	***
줄음쉼터 개수	0.261	0.099	0.008	***
휴게소 개수	0.167	0.082	0.042	**
설계속도	0.009	0.004	0.014	**
차로수	-0.249	0.062	0.000	***
버스전용차로 여부	0.503	0.156	0.001	***
시설물 수	0.010	0.005	0.041	**

주 : *유의수준 0.1, **유의수준 0.05, ***유의수준 0.01

모형식 B의 경우, 콘존길이, AADT, 휴게소 개수, 차로수, 버스전용차로여부, 교량의 수가 최종변수로 선정되었다. 선정된 변수를 이용하여 안전성능함수로 구축한 결과는 다음과 같다.

<표 16> 음이항 회귀모형 B 도출 결과

모형	음이항 회귀 모형 B			
관측수	781			
Log-Likelihood	-2077.2			
AIC	4168.4			
BIC	-4751.8			
변수	계수	S.E	p-value	유의수준
상수	-13.704	0.745	0.000	***
콘존길이	0.894	0.048	0.000	***
AADT	0.812	0.055	0.000	***
휴게소 수	0.144	0.080	0.070	*
차로수	-0.269	0.053	0.000	***
버스전용차로 여부	0.460	0.156	0.003	***
교량 수	0.017	0.008	0.034	**

주 : *유의수준 0.1, **유의수준 0.05, ***유의수준 0.01

두 모형 중 나은 모형을 선택하기 위해 Akaike 정보의 기준으로 모형의 적합치를 설명하는 모형의 상대적 품질 측도를 측정하는 AIC와 Bayesian의 정보 기준으로 최적의 모형을 탐색하는 기준인 BIC를 이용하여 두 모형을 비교하였다. AIC와 BIC는 모형 들 간의 상대적 비교를 하는 척도로 낮을수록 좋다. AIC와 BIC를 비교한 결과 음이항 회귀 모형 B가 음이항 회귀 모형 A에 비해 적합한 것으로 나타났다.

나. 최종 모형 검증

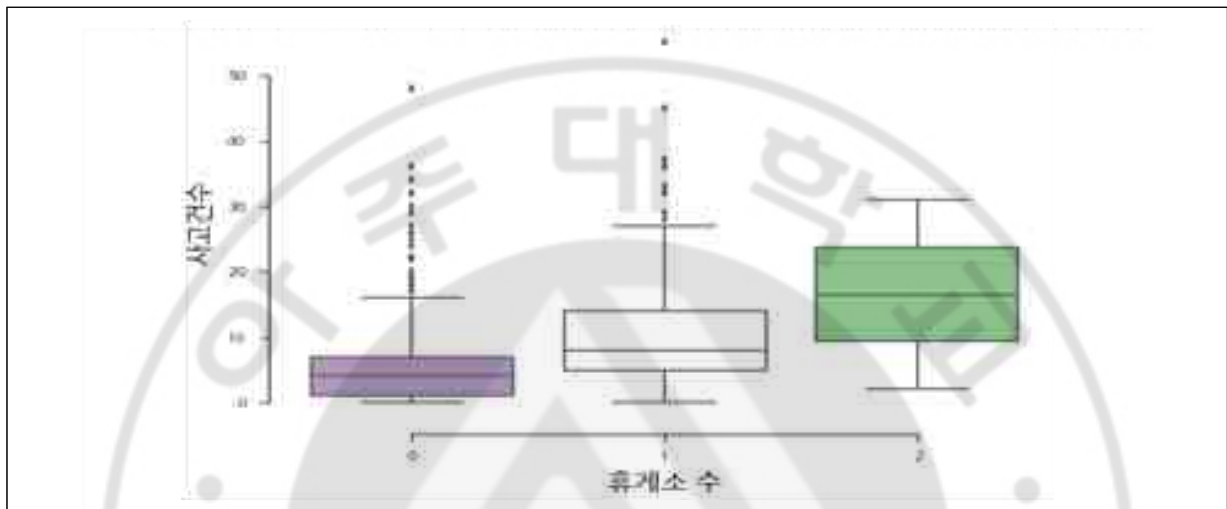
본 연구에서는 구축된 모형들을 검증하기 위해 테스트 데이터를 이용하여 예측력을 비교하였다. 모형의 예측력을 비교하는 수단으로 앞서 선정된 MOE인 MAD, SMAPE, RMSE를 이용하여 음이항 회귀 모형 A와 B를 비교하였다. 다음은 음이항 회귀 모형 A와 B를 선정된 MOE를 이용하여 모형을 검증하였다.

<표 17> 음이항 회귀모형 검증 결과

구분	MAD	RMSE	SMAPE
모형식 A	3.61	6.52	0.27
모형식 B	2.79	3.67	0.22
차이 (A-B)	0.82	2.85	0.05

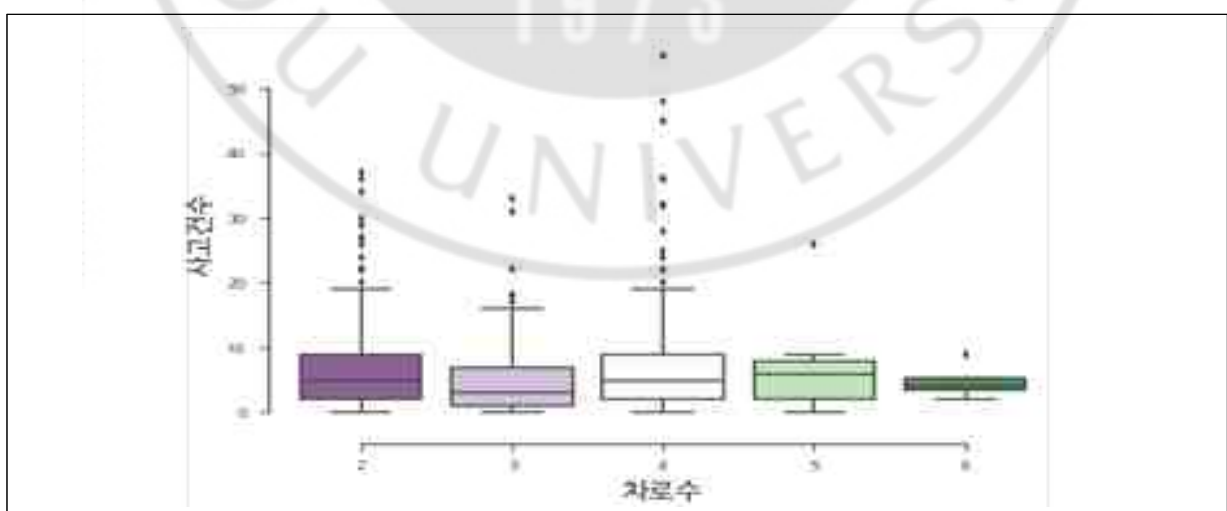
모형식 A와 B를 MAD 측면에서 비교 결과 모형식 A가 3.61, 모형식 B가 2.79로 모형식 A가 모형식 B에 비해 큰 오차를 갖는 것으로 나타났다. 또한 모형식 A의 RMSE는 6.52로 모형식 B의 3.67에 비해 큰 것으로 나타났다. SMAPE 또한 모형식 A가 0.05 더 크게 나타나 최종적으로 모형식 B가 적합한 안전성능함수 모형식으로 나타났다. 또한 음이항 회귀 모형식 B의 독립변수와 교통사고 건수와 상관관계를 분석하여 도출된 계수의 부호들을 검증하였다.

음이향 회귀모형 B의 독립변수들과 교통사고 건수와 상관계 분석을 위해 박스 플랏을 각각 그려 결과를 비교하였다. <그림 15>는 휴게소 개수와 교통사고 건수의 상관관계를 비교한 결과 콘존에 있는 휴게소 수가 증가할수록 사고건수가 증가하는 추세를 보였다. 이는 휴게소 진입 및 진출시 위빙이 발생하여 사고가 발생하는 것으로 판단된다.



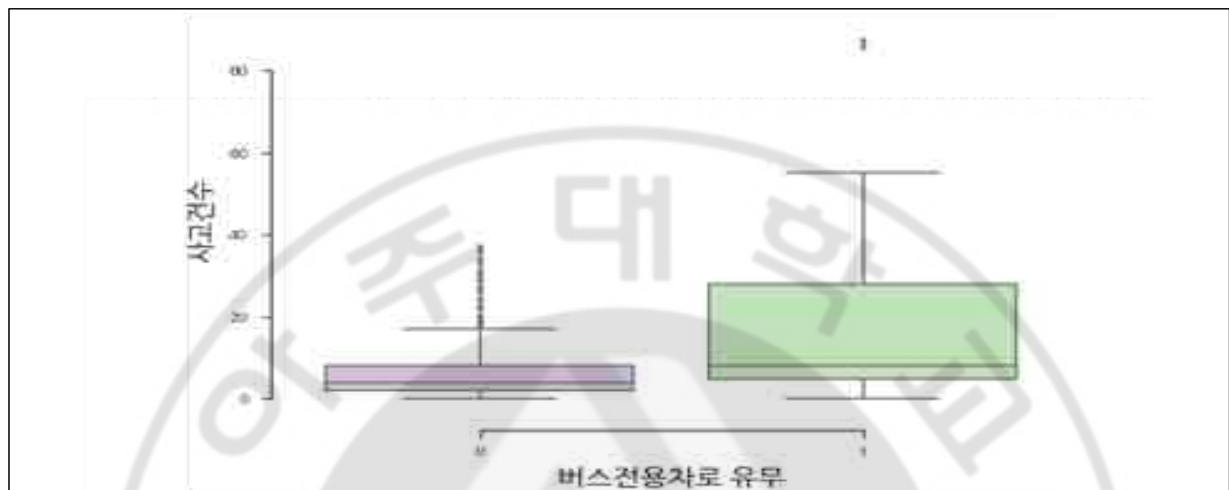
<그림 15> 휴게소 개수와 교통사고 건수의 상관관계

차로수의 경우 차로수가 증가할수록 사고건수가 감소하는 경향을 보였다. 이는 차로수가 적은 공간에서는 사고 발생 건수가 높은 것으로 알 수 있다.



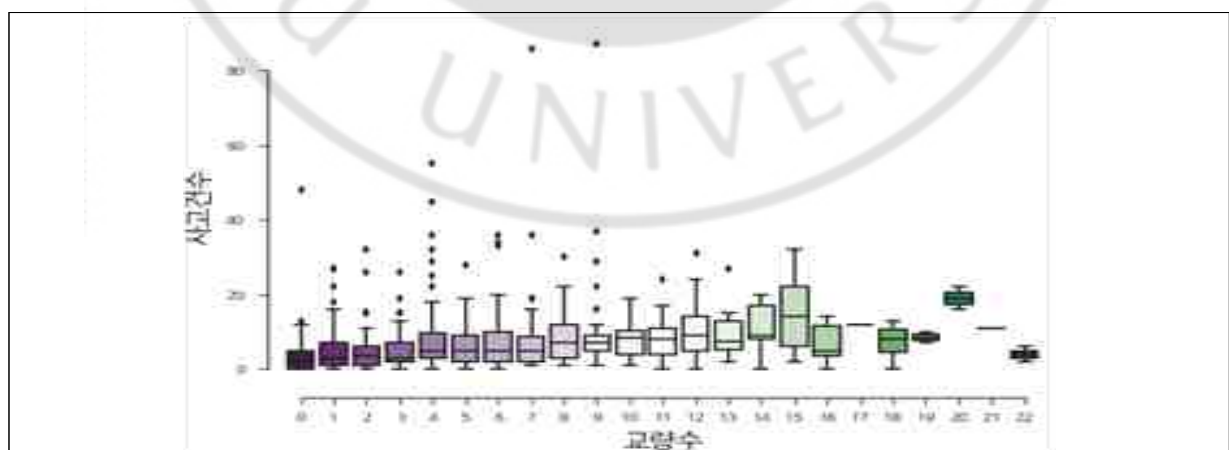
<그림 16> 차로수와 교통사고 건수의 상관관계

또한 버스전용차로가 있는 경우 사고건수가 증가하였다. 이는 고속도로의 버스전용차로는 중앙버스전용차로 운영되고 있는데 버스전용차로로 진입 및 진출 시 엇갈림이 발생하기 때문인 것으로 판단된다.



<그림 17> 버스전용차로 유무와 교통사고 건수와 상관관계

콘존에 교량 수가 많은 경우 사고가 증가하는 것으로 분석되었다. 이는 국내 교량구간에서는 차선 변경이 금지되어 있고, 이로 인한 교통사고가 발생하는 것으로 조사되었다.



<그림 18> 교량수와 교통사고 건수와 상관관계

특히, 버스전용차로 유무의 경우 적은 샘플 수에도 불구하고 유의한 변수로 선정되었다. 따라서 버스전용차로 유무에 대한 보다 상세한 분석이 필요할 것으로 판단된다. 또한 버스전용차로 여부가 다른 독립변수에 영향을 미치는 지에 대한 추가 분석을 실시하였다.

우선 버스전용차로가 존재하는 콘존의 수는 42개로서 전체 데이터 수에서 4.3%를 차지하고 있다. 앞서 살펴본 바와 같이, 다중공선성 분석 결과 VIF가 모두 10 이하로 산출되었기 때문에 버스전용차로 유무가 다른 변수와 다중공선성을 보이지는 않는 것으로 판단된다.

따라서 버스전용차로 유무에 해당하는 샘플 수가 차지하는 비율이 적음에도 불구하고, <그림 29>에서 보인 바와 같이 버스전용차로가 있고 없음에 대하여 교통사고 발생 빈도가 확연하게 차이가 나며, 다중공선성이 없기 때문에 다른 변수에도 영향을 미치지 않는 것으로 판단된다.

하지만, 버스전용차로 유무를 독립변수에서 제거할 경우 모형의 적합도 등에 미치는 영향을 추가로 분석한 결과 두 모형간의 큰 차이는 보이지 않았다. 버스전용차로 유무를 포함하지 않고 모형을 구축한 경우 AIC와 BIC가 각각 4,171.7과 -4,753.2로서 버스전용차로를 포함한 모형의 AIC와 BIC 값인 4,168.44와 -4,751.8보다 다소 높게 나와 모형 적합도도 근소하지만 조금 저하되는 것으로 나타났다. 참고로, 두 모형에서 모든 독립변수들이 유의수준 90%에서 통계적으로 유의한 것으로 나타났다.

또한 예측력 비교에서도 버스전용차로 유무를 포함하지 않고 모형을 구축한 경우 MAD, RMSE 그리고 SMAPE가 각각 2.80, 3.67, 0.22로서 버스전용차로를 포함한 모형의 MAD, RMSE 그리고 SMAPE 값인 2.79, 3.67, 2.22보다 모형의 예측력이 같거나 다소 낮은 것으로 나타났다.

결론적으로 버스전용차로가 존재하는 콘존의 수가 상대적으로 적지만 이를 독립변수에서 제거할만한 근거를 찾지 못하였다.

제3절 딥 러닝을 이용한 교통사고 예측모형 구축

1. 딥 러닝을 이용한 교통사고 예측모형 구축 배경

대부분의 교통사고 자료 분석이나 교통사고 건수 예측 등이 전통적인 통계적 방법인 포아송 회귀모형 또는 음이항 회귀모형 등을 기반으로 시행되어져 왔다. 이러한 전통적인 자료에 기반한 통계적 방법은 교통사고와 관련된 다양한 인적, 도로 기하구조적 그리고 환경적 요인들과 교통사고 간의 인과관계를 찾고, 교통사고 빈도를 예측하고 그리고 분석된 결과를 바탕으로 교통안전 등급을 산출하는 등 다양한 방식으로 활용되어져 왔다.

하지만, ICT기술의 발전으로 인하여 교통분야에서 과거에 경험해보지 못한 다양한 자료들이 대량으로 생산되고 있다. 또한 최근 컴퓨터 H/W 및 S/W 발달로 인하여 이러한 이종 및 대량의 정보를 빠른 시간 내에 분석하는 것들이 가능해지게 되었다. 이러한 자료 확보 및 분석 여건의 발전을 기반으로 최근에는 새로운 도전들이 시도되고 있다. 대표적인 예가 머신 러닝 및 딥 러닝과 같은 빅데이터 분석 기술이라 할 수 있다.

이러한 배경 하에 머신 러닝 및 딥 러닝과 같은 빅데이터 분석 기법을 활용한 새로운 접근 방법들이 주목을 받기 시작하였다. 이러한 머신 러닝 및 딥 러닝 기법은 이종(異種)의 대량 자료를 활용하여 교통사고와 관련된 요인들을 분석하는 데 장점을 보이고 있으며, 이미 교통 및 다른 분야에서는 활발하게 적용되어 우리들의 일상을 변화시키고 있다. 하지만, 이러한 장점에도 불구하고 교통분야, 특히 교통안전분야에는 그 활용도가 낮은 것이 사실이다.

머신 러닝과 딥 러닝의 큰 장점으로서는 첫째, 이종 및 대량의 자료를 이용하여 기존에 전통적 통계적 모형으로 찾지 못한 교통사고와 다른 요인들과의 관계를 추정할 수 있다는 것이다. 둘째, 빅데이터 환경에서 그 발전 가능성이 더욱 높아지는 것이라 할 수 있다. 특히, 향후 자율주행자동차 및 C-ITS 등에서 생산되는 센서 정보, 기상 정보 등을 결합할 경우 교통사고 예측의 정확도 및 신속성을 더욱 높일 수 있을 것으로 기대되고 있다.

전통적으로 딥 러닝의 경우에는 인식, 분류 및 군집 등에 많이 활용되어져 왔다. 하지만, 최근에는 이러한 활용 분야가 예측과 관련된 분야로 확대되고 있다. 예를 들면, 김호용 외(2016)은 「교통사고 지점 예측을 위한 딥 러닝 모델」에서 기상자료와 FFNN 딥 러닝 모형을 이용하여 교통사고 지점을 예측한 바 있다.

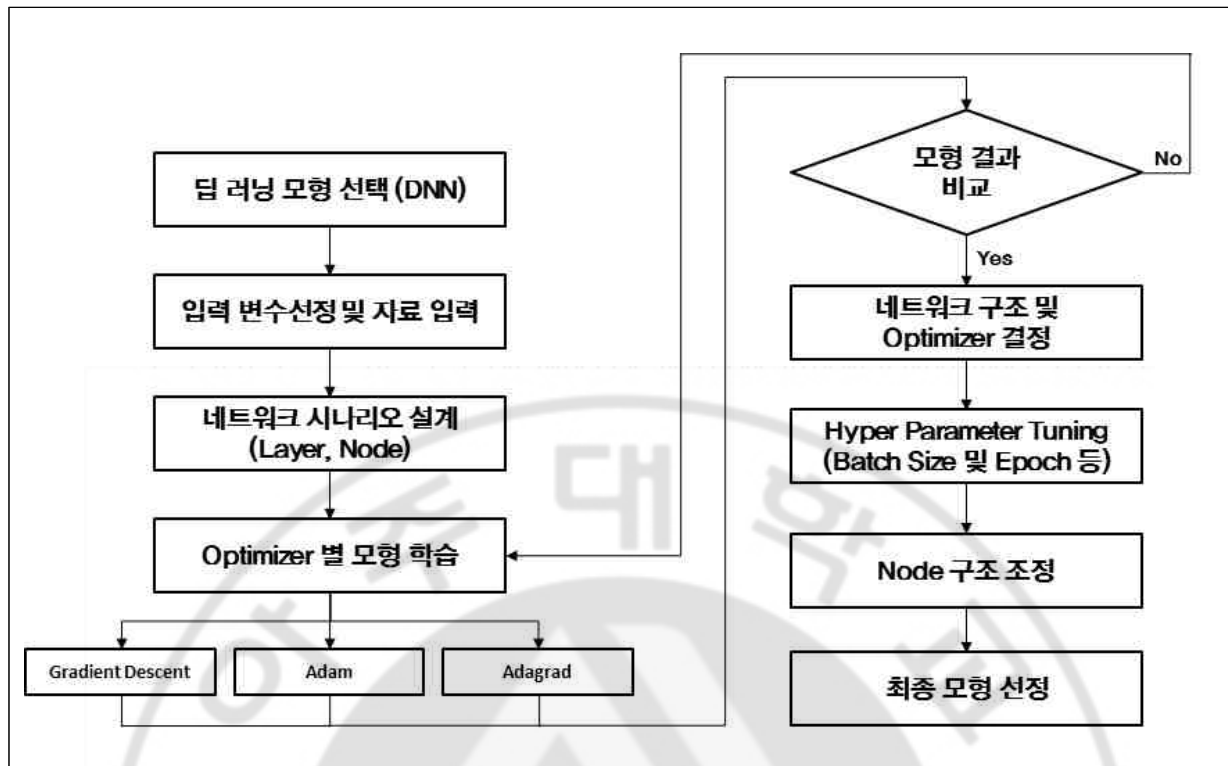
아직은 딥 러닝이 regression 분야에서 활발하게 이용되고 고도화되지는 않았으나, 기존에도 인공신경망이 regression 분야에서 많이 활용되어져 왔고, 텐서플로우를 중심으로 딥 러닝을 이용하여 regression을 수행할 수 있는 오픈소스들이 많이 제공되고 있다. 특히 딥 러닝은 데이터의 복잡성으로 인해 모델링이 어려운 경우 모델링을 가능하게 하는 장점이 있다. 이러한 개발 환경을 바탕으로 본 연구에서는 선제적으로 딥 러닝을 교통사고 예측에 활용함으로써 다른 후속 연구에서 활용될 수 있는 기술적 시사점을 제시하고자 한다.

2. 딥 러닝 개요

기존 안전성능함수에 비해 예측력이 우수한 교통사고 예측모형을 구축하기 위해 인식 분야와 예측 분야에서 다양하게 사용되고 있는 딥 러닝 기법을 이용하여 교통사고 예측모형을 구축하였다. 딥 러닝은 기존 인공신경망과 기본 구조가 같으나 은닉층을 2개 이상으로 구조를 다양하게 구성할 수 있는 장점을 가지고 있는 기법으로 다양한 데이터에 적용이 가능하다. 본 연구에서는 다양한 장점을 지닌 딥 러닝을 이용하여 교통사고 예측모형을 구축하고 딥 러닝을 이용한 교통사고 예측모형 구축 절차에 대해 제시하고자 한다. 또한, 딥 러닝을 이용한 교통사고 예측모형의 예측력과 전통적인 통계 기법을 이용한 교통사고 예측모형인 안전성능함수와 예측력을 비교하였다. 또한, 딥 러닝을 이용하여 교통사고 예측모형을 구축한 결과와 절차를 바탕으로 체계화된 방법론을 제시하고자 한다.

3. 구축 절차

본 연구에서는 딥 러닝 모형 중 인공신경망과 가장 유사한 형태인 Deep Neural Network(DNN)을 이용하였다. 우선 입력 변수를 선정하고 입력 자료를 결정하였다. 그 후 DNN의 네트워크 구조를 결정하기 위하여 Layer와 Node 시나리오를 설계하였다. 그 후 Optimizer 별로 모형을 구축하고 그 결과를 비교하여 네트워크 구조 및 Optimizer를 결정하였다. 결정된 네트워크 구조를 Batch Size 및 Epoch 등 Hyper Parameter Tuning을 통해 적합한 모형을 구축하고 Layer의 Node 구조를 조정하여 최종 모형을 선정하게 된다.



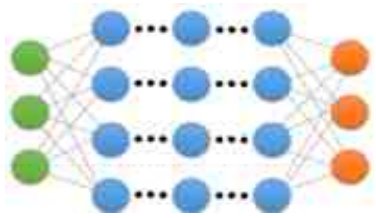
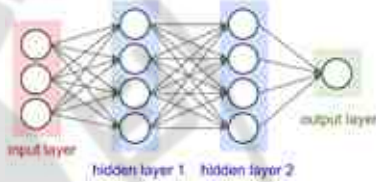
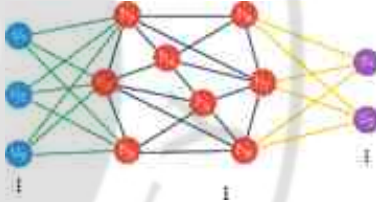
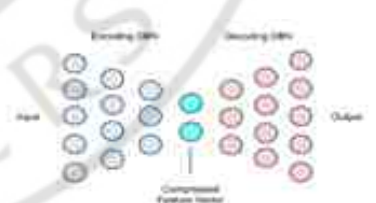
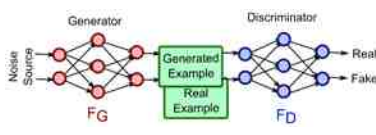
<그림 19> 딥 러닝을 이용한 교통사고 예측모형 구축 절차

4. 딥 러닝을 이용한 교통사고 예측모형 구축

가. 딥 러닝 모형 선택

딥 러닝(deep learning)은 머신러닝의 기법으로, 입력된 데이터에 대한 유형 분류나 회귀를 수행하는 방법이다. 일반적으로 딥 러닝은 여러 층으로 구성된 신경망을 말하며, 다양한 구조를 가진다. 딥 러닝은 구조 및 처리 방식에 따라 Deep Neural Network(DNN), Convolutional Neural Network(CNN) 및 Recurrent Neural Network(RNN) 등 다양한 모형을 가지고 있다. 본 연구에서는 다양한 모델링이 가능한 장점을 가진 Deep Neural Network를 이용하여 고속도로 교통사고 예측모형을 구축하였다. 다음 표는 딥 러닝 모형의 종류 및 구조를 나타내고 있다.

<표 18> 딥 러닝 모형의 종류

딥 러닝 모형	특징	구조
Deep Neural Network (DNN)	DNN은 일반적인 인공신경망과 마찬가지로 다양한 모델링이 가능한 장점이 있지만 과적합에 취약한 단점을 가짐 ex) 다양한 분야에서 예측 및 분류	
Convolutional Neural Network (CNN)	다른 딥 러닝 구조들에 비해 영상, 음성 분야에서 성능이 뛰어나다는 장점을 가지지만 많은 연산이 필요하다는 단점이 있음 ex) 2차원 이미지 인식, 음성 인식	
Recurrent Neural Network (RNN)	신경망 내부의 메모리를 활용할 수 있다는 장점이 있지만 그레디언트 소실 문제가 있어 패턴 학습을 못할 수 있는 단점이 있음 ex) 필기체 인식	
Auto Encoder (AE)	직관적이고 구현하기 쉽고 파라미터가 더 적어서 튜닝 하기가 쉽지만 입력 변수의 분포와 잠재 변수의 분포 사이에 상관관계가 없다는 단점이 있음 ex) 비정상 거래 검출, 추천 알고리즘	
Generative Adversarial Network (GAN)	결과가 좋고 빠르게 출력된다는 장점이 있지만 훈련이 불안정하다는 단점을 가지고 있음 ex) 이미지 복원, 사진과 사진의 전환	

나. 입력 변수 설정 및 자료 입력

전통적인 통계 방법에서 예측 성능이 우수하게 나온 음이항 회귀모형 B의 독립 변수를 딥 러닝을 이용한 교통사고 예측모형의 독립변수로 사용하였다. 본 연구에서 수집한 총 데이터는 977개로 많은 독립변수를 사용하여 예측할 경우 차원이 커질수록 학습이 어려워지고 더 많은 데이터를 필요로 하는 문제이다.

이를 해결하기 위해 음이항 회귀모형 B의 독립변수를 사용하여 딥 러닝 모형을 구축하였다. 사용된 입력 변수는 AADT, 콘존길이, 휴게소 개수, 차로수, 버스 전용차로 여부, 교량개수이다.

다. 네트워크 구조 시나리오 설계

딥 러닝 모형 구축을 위해서는 모형의 히든 레이어 수와 노드 수를 결정하는 것이 필요하다. 본 연구에서는 히든 레이어의 수를 3개와 5개로 구성하였으며, 노드의 수는 15개, 25개, 50개, 75개, 100개로 구성하였다. 즉 히든 레이어 수와 노드의 조합을 이용하여 민감도 분석(sensitivity analysis)을 수행하였으며 최종 네트워크를 구축할 수 있는 기반을 마련하였다.

라. Optimizer별 모형 학습

앞서 설계된 네트워크 구조 시나리오별로 최종적인 구조를 결정하기 위해 Optimizer별로 시나리오를 학습과 테스트를 진행하였다. 이때 Optimizer는 Gradient Descent, Adam, Adagrad를 사용하였는데 Optimizer별로 과적합 발생 등 차이가 발생하고, 학습 데이터와 테스트 데이터 모두 적합한 네트워크 구조를 찾기 위해서는 다양한 Optimizer를 이용하여 확인할 필요가 있다.

Gradient Descent 방법은 1차 근사값 발견용 최적화 알고리즘으로, 함수를 미분을 통해 기울기를 반복적으로 구하여 기울기가 낮은 쪽으로 이동시켜 나가면서 0에 가까워질 때까지 반복하는 시켜 최적해를 찾는다.

Adam 방법은 2015년에 제안된 새로운 방법으로 최근 가장 좋은 성능을 보이는 방법이다. 확률적 목적함수 기반의 1차 기울기 알고리즘이다. 특히 모멘트의 개념을 추가한 방법이다. 이 방법은 구현하기 쉽고, 계산상 효율적이다 (Kingma, 2015).

Adagrad 방법은 학습률을 조정하면서 학습을 진행하는 최적화 기법이다. 수들을 update할 때 각각의 변수마다 step size를 다르게 설정해서 이동하는 방식이다. 이 알고리즘의 기본적인 아이디어는 ‘지금까지 많이 변화하지 않은 변수들은 step size를 크게 하고, 지금까지 많이 변화했던 변수들은 step size를 작게 하자’ 라는 것이다.

최적 네트워크 구조 시나리오를 선정하기 위해서 비용함수(cost function)로는 MAD(Mean Absolute Deviation)을 사용하여 진행하였으며, 네트워크 구조 시나리오 별로 5회씩 학습 및 테스트를 진행하였다. 이에 따라, 학습 Cost와 테스트 Cost가 유사한 시나리오를 선정하였다.

(1) Gradient Descent 방법

Gradient Descent 방법으로 학습 및 테스트를 진행하여 학습과 테스트 Cost를 확인하고, 차이를 확인하였다. 확인 결과 히든 레이어별로 노드 수가 15인 경우 가장 우수한 성능을 보이는 것으로 나타났으며, 히든 레이어가 5개일 때 보다 3개일 경우 더욱 차이가 0.06으로 더 낮은 것으로 나타났다. 특히 히든 레이어 수 5인 경우에 노드 수가 100인 경우 학습 Cost가 낮았는데 이는 학습 데이터에 과적합(overfitting)이 되어 테스트 data에 맞지 않는 Cost를 보이는 것으로 나타났다. 즉, 일반적인 모형이 아닌 학습데이터에만 적합한 모형이 생성된 것이다.

<표 19> Gradient Descent 방법의 학습 및 테스트 Cost

Hidden Layer		3					5				
Node		15	25	50	75	100	15	25	50	75	100
1	학습	2.75	2.72	2.53	2.39	2.37	2.89	2.49	2.08	1.84	1.73
	테스트	2.81	2.86	2.88	2.90	2.88	2.80	2.97	3.27	2.94	3.10
2	학습	2.77	2.66	2.61	2.44	2.43	2.64	2.49	2.30	2.17	1.80
	테스트	2.82	2.81	2.80	2.89	2.92	2.94	2.97	3.12	3.08	3.03
3	학습	2.76	2.68	2.46	2.46	2.31	2.72	2.27	2.28	1.91	2.07
	테스트	2.88	2.85	2.92	2.95	2.89	2.87	2.96	3.14	3.16	3.26
4	학습	2.83	2.70	2.54	2.43	2.39	2.88	2.43	2.11	2.15	2.09
	테스트	2.82	2.89	2.86	2.93	2.91	2.80	3.05	3.04	3.20	3.16
5	학습	2.73	2.70	2.55	2.44	2.32	2.73	2.30	2.12	1.97	1.82
	테스트	2.83	2.86	2.85	2.87	2.90	2.85	3.01	3.06	3.11	3.12
평균	학습	2.77	2.69	2.54	2.43	2.36	2.77	2.40	2.18	2.01	1.90
	테스트	2.83	2.85	2.86	2.91	2.90	2.85	2.99	3.12	3.10	3.13
차이		0.06	0.16	0.32	0.47	0.54	0.08	0.60	0.95	1.09	1.23

(2) Adam 방법

Adam 방법으로 학습시킨 결과 히든 레이어와 노드 수가 증가할수록 학습데이터에 과적합하는 현상이 발생하는 것을 확인하였으며, 학습데이터와 테스트 데이터의 차이가 적은 히든 레이어 3과 노드 수 15를 최종적으로 선정하였다.

<표 20> Adam 방법의 학습 및 테스트 Cost

Hidden Layer		3					5				
Node		15	25	50	75	100	15	25	50	75	100
1	학습	2.06	1.68	0.96	0.81	0.58	2.14	1.50	0.61	0.39	0.19
	테스트	3.21	3.74	4.30	4.10	4.23	3.19	3.34	3.75	4.01	3.63
2	학습	2.06	1.68	1.21	0.83	0.57	2.04	1.65	0.75	0.24	0.21
	테스트	3.38	3.70	4.22	4.20	3.88	3.50	3.63	4.32	3.87	4.04
3	학습	1.95	1.62	1.10	0.80	0.60	1.84	1.70	0.78	0.48	0.17
	테스트	3.27	3.82	4.28	3.74	4.48	3.52	3.29	3.71	3.70	3.67
4	학습	2.21	1.75	0.99	0.72	0.55	1.88	1.50	0.73	0.38	0.19
	테스트	3.21	3.63	4.19	3.83	4.12	3.22	3.53	4.16	3.88	3.93
5	학습	2.10	1.81	1.16	0.80	0.61	2.15	1.51	0.54	0.37	0.17
	테스트	3.17	3.34	4.86	4.55	3.81	3.11	3.57	3.56	3.78	3.88
평균	학습	2.07	1.71	1.08	0.79	0.58	2.01	1.57	0.68	0.37	0.19
	테스트	3.25	3.65	4.37	4.09	4.10	3.31	3.47	3.90	3.85	3.83
차이		1.17	1.94	3.29	3.29	3.53	1.30	1.90	3.22	3.48	3.64

(3) Adagrad 방법

Adagrad 방법으로 학습 시킨 결과 히든 레이어 3과 노드 수 25의 학습 데이터와 테스트 데이터의 차이가 0.01로 가장 차이가 적게 나타나 일반적인 딥러닝 모형이 구성되었다고 볼 수 있다.

<표 21> Adagrad 방법의 학습 및 테스트 Cost

Hidden Layer		3					5				
Node		15	25	50	75	100	15	25	50	75	100
1	학습	3.36	2.87	2.61	2.61	2.44	2.81	2.66	2.31	2.12	1.90
	테스트	2.89	2.94	2.86	2.83	2.85	2.81	3.00	2.91	2.98	3.06
2	학습	3.47	2.81	2.63	2.56	2.52	2.86	2.63	2.42	2.22	1.86
	테스트	2.29	2.87	2.83	2.83	2.80	2.86	2.85	2.93	3.01	3.10
3	학습	3.05	2.84	2.62	2.06	2.46	2.90	2.71	2.34	2.19	1.95
	테스트	2.81	2.87	2.93	2.84	2.89	2.80	2.85	3.00	2.92	3.03
4	학습	3.31	2.91	2.69	2.55	2.48	2.85	2.65	2.44	2.15	1.81
	테스트	2.86	2.83	2.94	2.78	2.88	2.83	2.82	2.87	2.96	2.95
5	학습	3.53	2.93	2.68	2.52	2.49	2.83	2.68	2.34	2.19	1.88
	테스트	2.99	2.79	2.84	2.87	2.82	2.87	2.80	2.90	2.97	3.13
평균	학습	3.35	2.87	2.65	2.46	2.48	2.85	2.66	2.37	2.17	1.88
	테스트	2.77	2.86	2.88	2.83	2.85	2.83	2.87	2.92	2.97	3.05
차이		-0.58	-0.01	0.23	0.37	0.37	-0.02	0.20	0.55	0.79	1.17

마. 네트워크 구조 및 Optimizer 결정

각 Optimizer별로 평균 학습과 테스트 Cost 차이를 바탕으로 각각 최종시나리오를 선정하였다. 선정결과 모든 Optimizer에서 hidden layer는 3으로 선정되었고, Node의 경우 Optimizer별로 다르게 나타났다. 특히 Adagrad 방법의 경우 노드 수가 25로 나타났으며, 다른 방법들은 15로 나타났다.

<표 22> 각 Optimizer별 네트워크 구조

Optimizer	Hidden Layer	Node
Gradient Descent	3	15
Adam	3	15
Adagrad	3	25

또한 선정된 3개의 네트워크 구조의 학습 Cost와 테스트 Cost 차이를 비교하였다. 그 결과 Adagrad의 차이가 0.01로 가장 작아 최종 Optimizer로 선정되었다. 학습 Cost와 테스트 Cost의 차이가 가장 적은 네트워크 구조를 선택하는 이유로는 특정 데이터에 과적합 되지 않은 일반적이며 보편적인 모형을 나타내기 때문이다. 최종 네트워크 구조는 히든 레이어 3개와 노드 25개로 선정되었다.

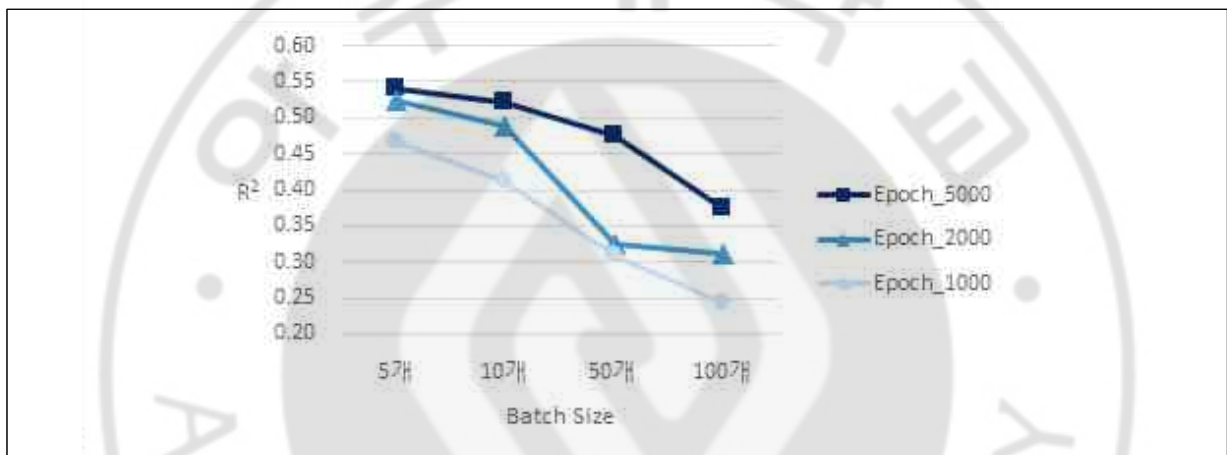


<그림 20> 학습 Cost 및 테스트 Cost 차이

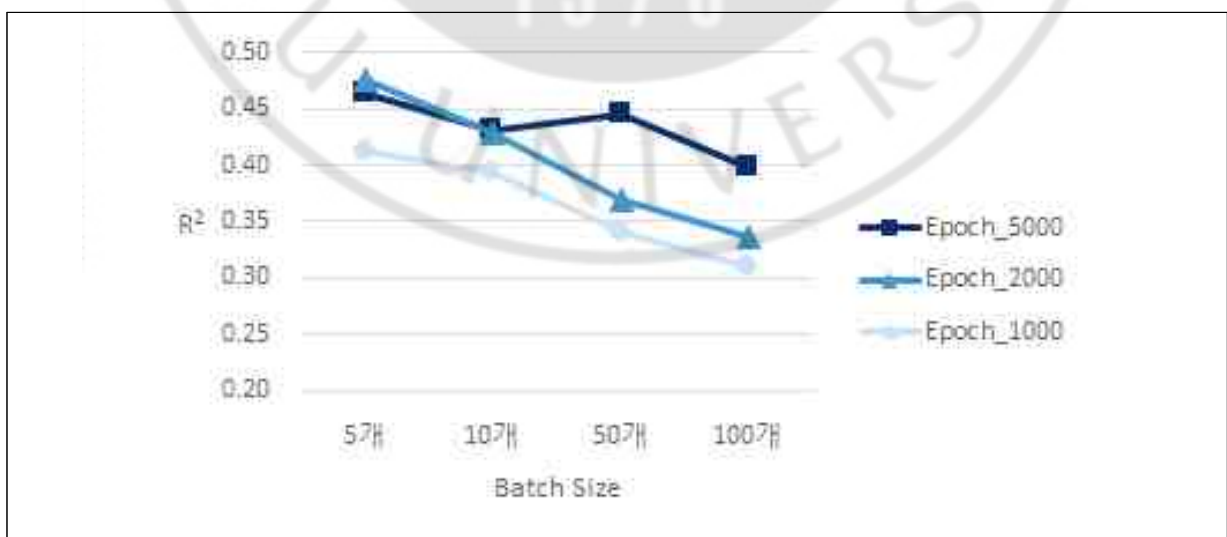
바. Hyper Parameter Tuning

앞서 선정된 네트워크 구조 및 Optimizer를 바탕으로 Batch Size 및 Epoch 수를 결정하는 것이 필요하다. 특히 Batch Size는 학습을 좀 더 깊게 하기 위해 필요하며 Batch Size 별로 모형의 수렴여부가 다르기 때문에 Epoch 또한 변경하며 학습을 진행하였다. Batch Size와 Epoch을 결정하기 위한 MOE로는

실제 사고건수와 예측 사고건수 간의 R^2 를 사용하였다. 학습 및 테스트별로 Batch Size와 Epoch 수를 이용하여 비교 하였다. 이때 Batch Size는 5개, 10개, 50개, 100개를 사용하였다. 학습 데이터의 수가 781개로 작아, Batch Size를 1개로 하여 학습을 할 경우 시간이 오래 걸리는 단점과 과적합되는 문제가 발생하여 Batch Size 1개는 제외하였다. Epoch는 1,000회, 2,000회, 5,000회를 이용하였다. 최종적으로 Batch Size는 5개, Epoch는 5,000회가 선정되었으며, 각각의 R^2 는 다음 그래프와 같다.



<그림 21> 학습 Batch Size 및 Epoch별 R^2



<그림 22> 테스트 Batch Size 및 Epoch별 R^2

사. 노드 구조 변경

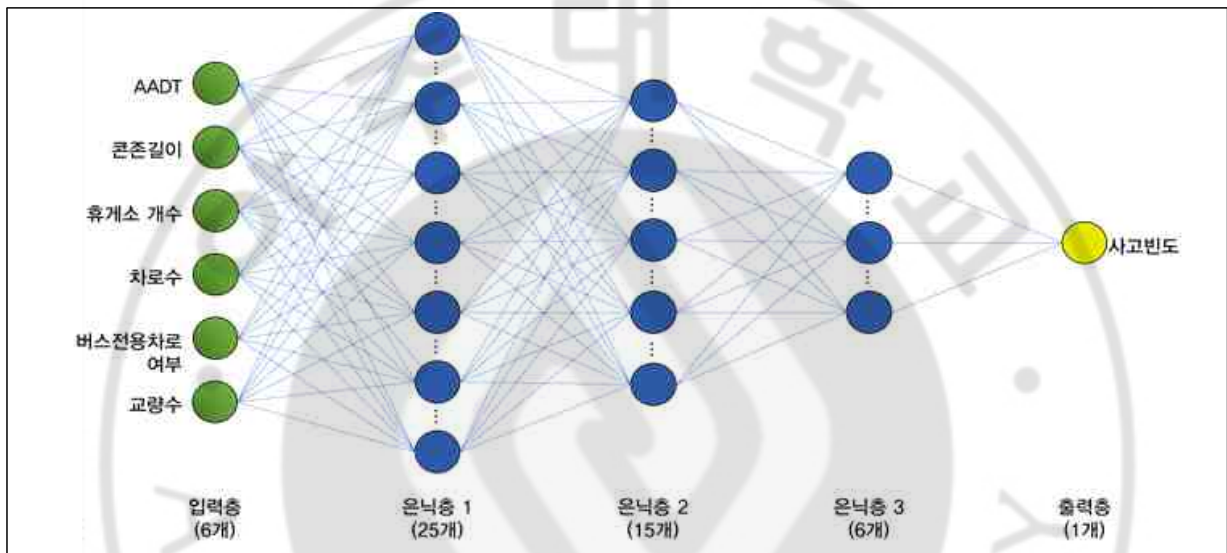
결정된 네트워크 구조 시나리오, Optimizer, Batch Size와 Epoch를 바탕으로 딥 러닝의 장점인 다양한 구조를 실험하였다. 우선 히든 레이어 별 노드 시나리오를 작성하여 구성하였다. 네트워크 구조로 선정된 히든 레이어의 수는 고정하고 히든 레이어들의 노드 수를 다양한 조합으로 변경하여 시나리오를 작성하였다. 첫 번째 시나리오는 구조가 히든 레이어 별로 노드 수가 같은 경우이다. 두 번째 시나리오는 2번째 히든 레이어의 노드 수를 2배로 늘리는 경우이다. 세 번째 시나리오는 2번째 히든 레이어의 노드 수를 1/2배하는 경우이다. 네 번째 시나리오는 1번째 히든 레이어의 노드 수를 2배로 늘리고, 3번째 히든 레이어의 노드 수를 1/2배하는 경우이다. 다섯 번째 시나리오는 1번째 히든 레이어는 1/2배하고, 3번째 히든 레이어의 노드 수를 2배로 하는 경우이다. 앞서 구성된 시나리오를 바탕으로 학습 및 테스트를 진행하였으며, 본 연구에서는 히든 레이어의 노드 수가 줄어들수록 성능이 좋아짐을 확인하였다. 이를 바탕으로 노드 수는 앞서 선정된 25개와 독립변수의 수인 6개 사이에서 반복 시도를 통해 본 연구에서 구축된 데이터에 적합한 최종 시나리오를 선정하게 되었다. 다음 표는 노드 구조를 변경하며, 최종 모형을 선정하는 과정이며, 25-15-6의 구조가 최종 모형으로 선정되었다.

<표 23> 노드 구조 변경 및 최종 모형 선정

Hidden Layer 별 Node시나리오	학습 Cost	테스트 Cost	차이	선정 여부
25-25-25	3.07	2.78	0.29	
25-50-25	3.00	2.85	0.15	
25-12-25	3.15	2.82	0.33	
50-25-12	3.00	2.84	0.16	
12-25-50	3.27	2.86	0.41	
25-15-6	2.59	2.52	0.07	선정

5. 최종 선정 모형 및 결과

노드 구조 변경을 통해 최종 모형으로 25-25-25 모형을 변형 시킨 25-15-6 모형이 선정되었다. 학습 Cost는 2.59이며, 테스트 Cost는 2.52로 딥 러닝 모형 중 성능이 좋은 모형이다. 본 연구에서 최종적으로 선정된 딥 러닝 모형의 구조는 다음 그림과 같다.



<그림 23> 최종 선정 모형

본 연구에서는 전통적인 통계 기법과 딥 러닝 모형을 이용하여 고속도로 교통사고 예측모형을 구축하고 예측력을 비교하였다. 예측력 비교 결과 딥 러닝 모형의 MOE들이 전통적인 통계 모형에 비해 다소 우수한 것으로 나타났다. 하지만 딥 러닝을 이용할 경우 예측 신뢰도를 더욱 증가 시킬 수 있는 것으로 확인되었다.

또한, 딥 러닝의 경우에는 아직 교통사고 건수 예측에 활용한 사례가 적기 때문에 적절한 구조 등에 대한 추가 연구가 부족하였다. 본 연구를 통해서 딥 러닝을 이용할 경우 기존에 중요하다고 언급된 은닉층 및 노드 개수뿐만 아니라 Optimizer, 노드 구조 등에도 많은 영향을 받는 것으로 확인 되었다.

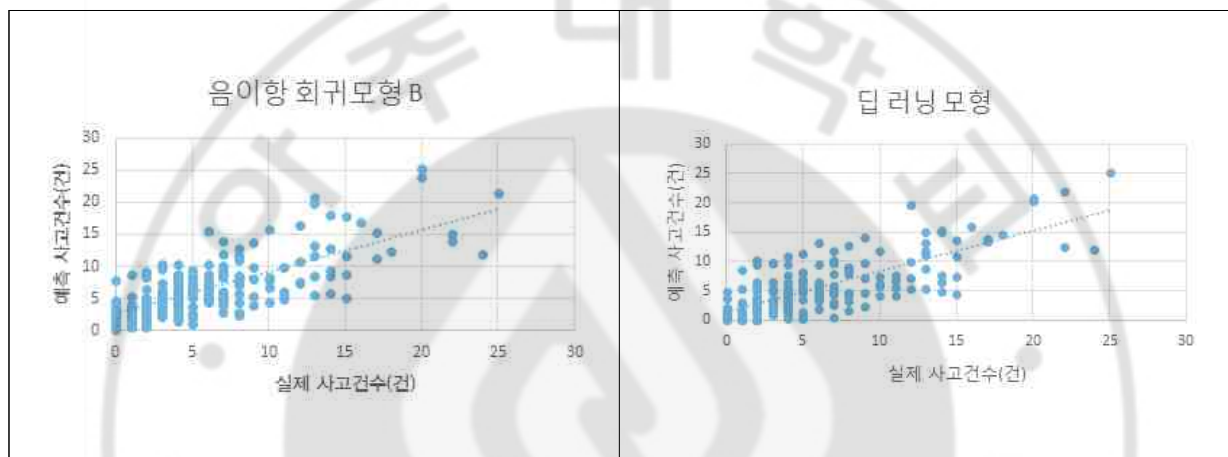
6. 고속도로 교통사고 예측모형 비교

구축된 고속도로 교통사고 예측모형들을 비교하기 위해, 전통적인 통계 모형에서 선정된 음이항 회귀 모형식 A, 음이항 회귀 모형식 B와 딥 러닝을 통해 구축된 모형을 비교하였다. 비교 결과, 딥 러닝 모형이 MOE인 MAD, RMSE, SMAPE 모두에서 가장 우수하게 나타났다. MAD의 경우 2.52로 전통적인 통계 모형에서 가장 우수했던 음이항 회귀 모형식 B의 MAD인 2.79보다 0.27 낮게 나타났다. 또한 딥 러닝 모형의 RMSE의 경우 3.43으로 음이항 회귀 모형식 B의 3.67에 비해 낮게 나타났으며, SMAPE도 0.01 낮게 나타났다. 전체적으로 딥 러닝 모형의 MOE가 전통적인 통계기법에 비해 좋은 성능으로 나타났으나, 차이는 미미한 것으로 나타났다. 이는 데이터 수의 한계로 판단된다. 두 모형의 MOE를 비교한 표는 다음과 같다.

<표 24> 모형별 MOE 비교 결과

구분	MAD	RMSE	SMAPE
① 음이항 모형식 A	3.61	6.52	0.27
② 음이항 모형식 B	2.79	3.67	0.22
③ 딥 러닝 모형	2.52	3.43	0.21
차이 (①-③)	1.09	3.09	0.06
차이 (②-③)	0.27	0.24	0.01

또한, 모형의 적합도를 확인하기 위해 음이항 회귀 모형 B와 딥 러닝 모형의 테스트 데이터를 이용하여 실제 사고건수와 예측 사고건수를 비교하는 산점도를 그려 모형을 적합도를 검증하였다. 검증 결과, 딥 러닝 모형의 경우 음이항 회귀 모형 B에 비해 실제 사고건수가 적게 나는 구간에서는 예측을 잘하는 것으로 판단된다. 하지만, 두 모형 간에는 두드러진 차이가 나타나지는 않는 것으로 판단된다.

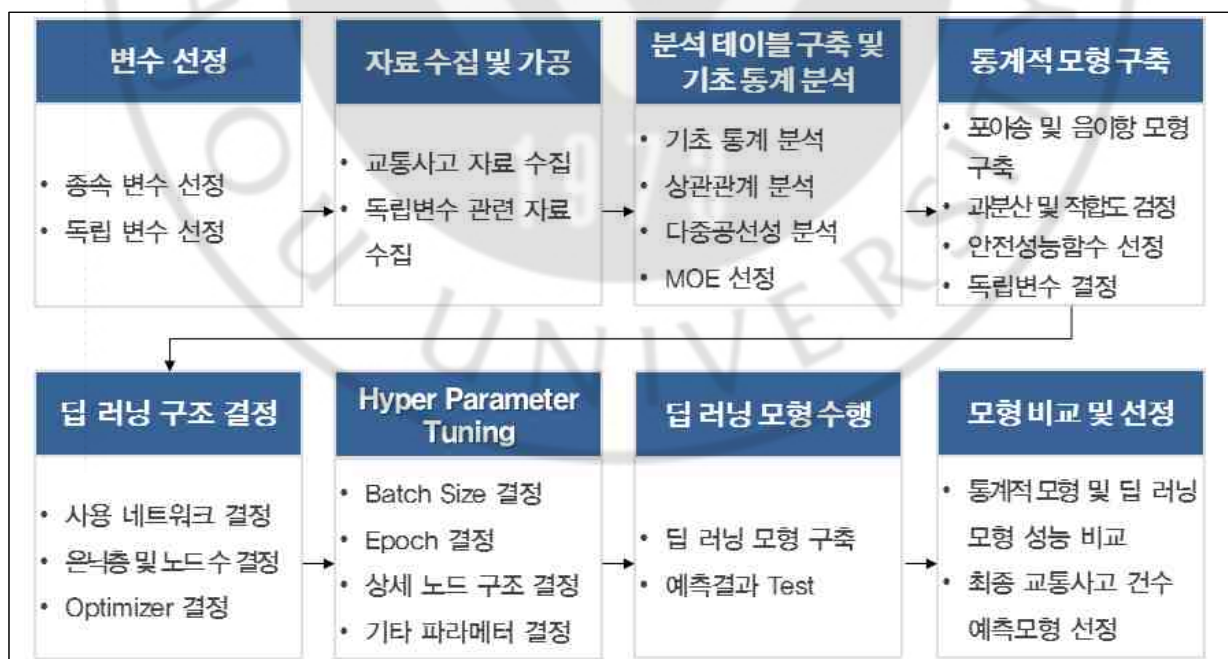


<그림 24> 모형 비교 및 검증

제4절 딥 러닝을 이용한 교통사고 건수 예측모형 구축 절차 및 활용 방안

1. 구축 절차

본 연구에서는 전통적인 통계 기법 및 딥 러닝 기반의 교통사고 건수를 예측하는 모형을 구축하는 절차를 제시하고자 한다. 전통적인 통계 기법과 딥 러닝 기법이 가진 장점이 상이하게 다르기 때문에, 이를 통합한 방법론을 제시하고자 한다. 우선 변수선정, 자료 수집 및 가공을 통해 전통적인 모형과 딥 러닝을 모형 구축을 위한 준비를 한다. 그 후 통계적 기법을 이용하여 데이터를 점검하고, 전통적인 통계적 모형을 구축하여 독립변수들의 영향력과 통계적 유의성을 파악하고, 최종적으로 구축된 전통적인 통계적 모형의 독립변수를 이용하여 딥 러닝 구조를 결정한다. Hyper Parameter를 Tuning 하여 최종 딥 러닝 모형을 구축한 후 모형 비교를 통해 선정하게 된다. 다음 그림은 전통적인 통계 기법 및 딥 러닝 기반 교통사고 건수 예측모형 구축 절차이다.



<그림 25> 통계기법 및 딥 러닝 기반 교통사고 건수 예측모형 구축 절차

2. 활용 방안 및 기대효과

본 연구에서는 최종적으로 전통적인 통계 기법 및 딥 러닝 기반 교통사고 건수 예측모형의 구축 방안을 제시하였다. 이에 제시된 고속도로 교통사고 건수를 실무적으로 활용할 수 있는 방안에 대하여 제안하고자 한다.

첫째, 교통사고 건수 예측모형의 경우 기본적으로 콘존 단위의 고속도로 구간에 대하여 교통사고 건수를 예측할 수 있다. 이렇게 모형에 의해서 교통사고 건수를 예측할 경우 교통사고 자료에서 흔히 발생하는 평균으로의 회귀(regression to the mean) 현상을 완화시키는 데 도움을 줄 수 있다. 이렇게 모형에서 나온 교통사고 건수의 경우 미국 도로안전편람 등에서 제시하고 있는 경험적 베이즈(empirical bayes) 방법론을 적용하여 교통사고 건수를 예측하고 이를 통하여 교통안전성 개선을 위해 투자가 필요한 콘존 등을 선정하는 데 활용될 수 있다.

둘째, 본 연구에서 개발된 교통사고 예측모형을 활용할 경우 교통사고를 기반으로 하는 콘존 단위의 교통안전등급을 산출하는 데 활용할 수 있다. 실제 교통사고 건수의 경우 앞서 언급한 바와 같이 평균으로의 회귀 현상이 발생할 수 있다. 즉, 올해 교통사고가 높은 지점으로 선정된 콘존들이 내년에는 평균적인 교통사고 건수를 보일 수도 있다. 따라서 교통사고 건수가 높은 콘존들을 기반으로 개선지점을 선정하는 것 보다는 모형에서 예측된 교통사고 건수와 실제 교통사고 건수의 차이를 기반으로 교통안전등급을 산출하는 것이 더욱 효과적일 수 있다. 즉, 해당 콘존과 유사한 특성을 보이는 콘존들에 비해 해당 콘존이 높은 교통사고 건수를 보인다고 하면, 해당 콘존이 위험한 구간이라고 판단할 수 있다.

마지막으로 본 연구에서 개발된 교통사고 건수 예측모형을 활용할 경우 해당 콘존의 기하구조, 교통시설 또는 교통여건을 개선할 경우 기대되는 교통안전성 향상 효과를 추정하는 데 활용할 수 있다. 이를 위해서는 지속적으로 교통사고 예측모형을 고도화시켜나가는 과정이 필요할 것으로 판단된다.

본 연구를 통해서 구축된 딥 러닝과 전통적인 통계적 기법을 접목한 고속도로 교통사고 빈도 예측모형은 고속도로의 교통사고 예측 기법을 고도화시켜 예측 정확도를 향상시킬 수 있을 것으로 기대된다. 또한 본 연구에서 개발한 딥 러닝을 기반으로 하는 교통사고 빈도 예측모형은 고속도로뿐만 아니라 다양한 종류의 도로와 교차로 등에 적용될 수 있으며, 이를 통하여 우리나라 전반의 교통안전성 향상에 기여할 수 있을 것으로 전망된다.

또한 본 연구는 다른 분야에선 활발히 활용 중이나 아직까지 교통사고 예측 분야에 그 활용성이 낮은 딥 러닝의 가능성을 인지하고 제4차 산업혁명 시대의 핵심기술 중에 하나인 딥 러닝을 교통사고 예측분야에 선제적으로 활용함으로써 앞으로 딥 러닝을 활용하고자 하는 연구자들에게 유의하고 활용 가능한 시사점을 제공할 수 있을 것으로 기대된다.

제5장 결론 및 향후 연구 과제

제1절 결론

기존에는 대부분의 교통사고 자료 분석과 교통사고 건수 예측 등이 전통적인 통계적 방법인 포아송 회귀모형 또는 음이항 회귀모형 등을 이용하여 수행되어져 왔다. 이러한 통계적 방법은 교통사고와 관련된 다양한 인적, 도로 기하구조적 그리고 환경적 요인들과 교통사고 간의 인과관계를 찾고, 교통사고 빈도를 예측하고 그리고 분석된 결과를 바탕으로 교통안전 등급을 산출하는 등 다양한 방식으로 활용되어져 왔다.

하지만, 최근 머신 러닝 및 딥 러닝과 같은 빅데이터 분석 기법을 활용한 새로운 접근 방법들이 주목을 받기 시작하였다. 이러한 머신 러닝 및 딥 러닝 기법은 이종(異種)의 대량 자료를 활용하여 교통사고와 관련된 요인들을 분석하는 데 장점을 보이고 있으며, 이미 교통 및 다른 분야에서는 활발하게 적용되어 우리들의 일상을 변화시키고 있다.

이에 본 연구의 목적은 고속도로 교통사고 자료를 이용하여 고속도로의 주요 분석 단위인 콘존의 교통사고 빈도수를 예측하기 위하여 전통적인 통계적 기법과 딥 러닝을 이용한 기법을 적용하고 각 기법들의 예측 성능을 비교하였다.

예측 성능 비교 결과, 딥 러닝 모형의 MOE들이 전통적인 통계 모형에 비해 다소 우수한 것으로 나타났다. 하지만 MAD 기준으로 차이가 0.27로 전통적인 통계적 기법 기반으로도 교통사고 건수를 충분히 예측이 가능하다고 판단되며, 특히 음이항 회귀모형이 포아송 회귀모형보다 우수한 것으로 나타났다. 또한 노출 계수(AADT, 콘존 길이)를 활용하는 모형이 더욱 우수한 것으로 판단된다. 하지만 딥 러닝을 이용할 경우 예측 신뢰도를 더욱 증가 시킬 수 있다. 또한, 딥 러닝의 경우에는 아직 교통사고 건수 예측에 활용한 사례가 적기 때문에 적절한 구조 등에 대한 추가 연구가 필요하다. 특히 본 연구에서 은닉층 및 노드 개수뿐만 아니라 Optimizer, 노드 구조 등에도 많은 영향을 받는 것으로 확인 되었다.

제2절 연구의 한계 및 향후 연구과제

1. 연구의 한계

본 연구는 다른 도로 유형에 비하여 상대적으로 교통사고 건수 예측에 필요한 기본 자료가 잘 구축되어 있는 고속도로를 대상으로 교통사고 건수 예측모형 구축을 실시하였다. 하지만, 상대적으로 관련 자료가 정확하고 많은 고속도로임에도 불구하고 교통사고 예측에 활용될 수 있는 상세한 자료가 부족하여 아주 다양한 변수들을 모형에서 고려하지 못한 한계가 존재한다.

또한 본 연구는 자료 수집의 용이성을 이유로 상대적으로 자료가 잘 정리되어 있는 콘존을 기준으로 교통사고 건수 예측모형을 구축하였다. 하지만, 콘존의 경우 상대적으로 구간길이가 길고 또한 구간길이가 다양하다. 이렇게 콘존을 대상으로 모형을 구축하는 경우, 콘존 단위로 가공하여 진행했기에 딥 러닝을 수행하기에 충분한 데이터 수를 확보하지 못하는 한계가 존재한다.

그리고 클러스터링 기법과 고속도로 교통사고 예측모형을 연계하기 위한 노력을 하였으나, 분석 결과는 부정적으로 나타났다. 하지만, 본 연구에서 다양한 머신 러닝 기법을 이용하여 고속도로 자료 및 고속도로 교통사고 자료를 유형화시키지 못하였다는 한계가 존재한다. 따라서 이에 대한 보다 체계적이고 심도있는 분석이 부족하였다.

마지막으로 딥 러닝의 경우에도 교통사고 건수 예측에 활용된 기존 연구가 부족하여 교통사고 건수 예측에 적절한 모형과 모형 구조를 선정하는 데 부족함이 존재한다. 본 연구에서 사용된 DNN 모형 이외에도 다양한 모형들이 딥 러닝에서 사용되고 있으며, 모형의 구조 또한 훨씬 다양하게 활용되고 있다. 본 연구가 딥 러닝을 이용한 교통사고 건수 예측 관련 연구가 부족한 상황에서 실행되었기는 하지만 향후 이러한 부분에 대한 연구가 더욱 진행될 필요가 있다고 판단된다.

2. 향후 연구과제

앞서 언급한 바와 같이 고속도로 교통사고 건수 예측과 관련하여 본 연구는 많은 한계점을 보이고 있다. 하지만, 본 연구에서 수행한 다양한 시도들은 충분히 가치가 있으며 앞으로 진행될 많은 연구들에게 유효한 시사점을 제시할 것으로 기대된다. 본 연구에서 수행한 내용과 본 연구의 한계점을 바탕으로 다음과 같은 향후 연구과제를 제안하고자 한다.

첫째, 교통사고 건수 예측을 위하여 딥 러닝을 활용함에 있어, 입력변수의 종류 및 량 증대를 통해 모형을 고도화 하는 것이 필요하다. 현재는 공간적 범위로 고속도로 콘존을 대상으로 모형을 구축하였으나, 훨씬 자세한 자료 및 수량이 많은 VDS존으로 확대하거나 아니면 다른 도로유형(일반국도, 지방도, 시도 등)으로 확대하는 것도 고려해 볼 필요가 있다.

둘째, 입력 자료를 확대하는 것이다. 입력변수도 세부 기하구조, 교통량 및 속도 시계열 자료 등으로 확대할 수 있다. 현재는 단면적인 자료를 입력하고 있으나, 교통량 또는 속도 시계열 자료들을 활용한다면 해당 도로의 사용 패턴에 대한 보다 적극적인 고려가 가능할 것으로 판단된다.

셋째, 교통사고 건수 예측을 위한 딥 러닝 심층 분석이 필요할 것으로 판단된다. 교통사고 건수 예측에 적합한 딥 러닝 네트워크 구조, 그리고 각종 파라미터 값을 심층 적으로 분석하여 예측력을 향상 시킬 필요가 있다.

마지막으로 향후 딥 러닝 활용 등을 염두에 두고 교통사고 자료 및 기하구조 자료 등을 DB로 구축하여 관리하는 것이 필요할 것으로 판단된다.

참고문헌

- 경찰청(2017), 2017 교통사고 통계(2016년 통계)
- 강동운(2014), 상관성이 있는 변수를 고려한 주성분 분석 기반의 안전성능함수, 서울대학교 대학원 건설환경공학부 석사학위 논문
- 권순재, 김성현, 탁은식, 정현희(2017), K-Menas Clustering 알고리즘과 헤도닉 모형을 활용한 서울시 연립 다세대 군집분류 방법에 관한 연구, 한국지능정보 시스템학회, Vol. 23, No. 3, pp. 95-118
- 김상구, 윤일수, 박재범, 박인기, 천승훈, 김경현, 안현경(2016), VDS 자료 기반 고속도로 교통혼잡비용 산정 방법론 연구, 한국도로학회논문집, Vol. 18, No. 1, pp. 99-107
- 김윤진(2017), 딥 러닝(Deep Learning)을 활용한 이미지 빅데이터(Big Data) 분석 연구, 중앙대학교 대학원 박사학위 논문
- 김호용, 김정재, 박상민, 전채남, 온병원(2016), 교통사고 지점 예측을 위한 딥 러닝 모델, Smart Connected World 2017
- 문지원(2007), K-Means 군집분석을 이용한 U-도시 유형분류에 관한 연구, 성균관대학교 석사학위 논문
- 박병호, 나희(2012), 로터리 사고발생 위치별 사고모형 개발, 도로학회논문집, Vol. 14, No. 4, pp. 83-91
- 박주환, 김상구(2012), 다중선형 회귀분석을 이용한 고속도로 터널구간의 교통사고 예측모형 개발, 한국ITS학회, Vol. 11, No. 6, pp. 145-154
- 박효신, 손봉수, 김형진(2007), 고속도로 인터체인지 연결로에서의 교통사고 예측모형 개발, 대한교통학회지, Vol. 25, No. 3, pp. 123-135
- 서임기, 강동운, 박제진, 박신행(2015), 고속도로 선형 동질구간 기반의 안전성능함수 개발, 대한토목학회논문집, Vol. 35, No. 2, pp. 397-405

- 성낙문(2002), 고속도로 인터체인지에서 교통사고 예측모델 개발, 대한토목학술논문지, Vol. 22, No. 4, pp. 617-625
- 오영태, 강동수(2017), 첨단교통안전공학
- 오주택, 윤일수, 황정원, 한음(2014), 비선형 회귀분석, 인공 신경망, 구조방정식을 이용한 지방부 4지 신호교차로 교통사고 예측모형 성능 비교 연구, 대한교통학회지, Vol. 32, No. 3, pp. 266-279
- 윤일수, 박성호, 윤정은, 최진형, 한음(2012), 유입·유출특성을 고려한 고속도로 연결로의 교통사고 심각도 예측모형, 한국도로학회, Vol. 14, No. 5, pp. 101-111
- 이근희, 노정현(2015), 확률모수를 이용한 교통사고예측모형 개발-수도권 및 부산광역시 4지 교차로를 대상으로, 한국ITS학회논문지, Vol. 14, No. 6, pp. 91-99
- 이세진, 김동현(2016), 도시 구조물 분류를 위한 3차원 점 군의 구형 특징 표현과 심층 신뢰 신경망 기반의 환경 형상 학습, 로봇학회논문지 로봇공학회 논문지, 제11권 제3호(통권 제41호), pp.115~126.
- 이수범, 김정현, 김태희(2003), 도로 및 교통특성에 따른 계획 단계의 도시부 도로 교통사고 예측모형개발, 대한교통학회지, Vol. 21, No. 4, pp. 133-144
- 이신원(2012), K-Means 클러스터링에서 초기 중심 선정 방법 비교, 인터넷정보학회, Vol. 13, No. 6, pp. 1-8
- 이일현(2014), EASYFLOW 회귀분석, 한나래
- 이원휘(2010), K-Means 알고리즘을 이용한 대용량 문서 클러스터링에서 개선된 초기 중심 선정 방법의 제안, 전북대학교 박사학위 논문
- 이태현, 곽호찬, 김동규, 고승영(2015), 고속도로 영업소 구간 안전성능함수 개발, 대한교통학회지, Vol. 33, No. 1, pp. 81-89
- 정완(2017), 딥 러닝을 활용한 자율주행 자동차의 사회적 딜레마 해결 방안, 단국대학교 대학원 석사학위 논문
- 정재풍(2014), 교통사고건수에 대한 포아송 회귀와 음이항 회귀모형 적합, 고려대학교 석사학위 논문

- 최윤환(2012), 고속도로 연결로 구간의 사고예측계수(AMF) 개발 및 활용방안 연구, 아주대학교 박사학위 논문
- 최희열, 민운홍(2015), 지능형 정보 시스템; 딥 러닝 소개 및 주요 이슈, 한국정보처리학회, Vol. 22, No. 1, pp. 7-15
- 한국도로공사(2017), 2016년도 고속도로 교통사고 통계
- 한국도로공사(2014), 고속도로 안전성능합수 사고수정계수 개발사례집
- 황경성, 최재성, 김상엽, 허태영, 조원범, 김용석(2010), 차량속도를 이용한 도로 구간분할에 따른 고속도로 사고빈도 모형 개발 연구, 대한교통학회지, Vol. 28, No. 2, pp. 151-159
- Li Ye(2016), Overlapping K-means 알고리즘을 이용한 키워드 기반 저널 중복 클러스터링 연구, 연세대학교 석사학위 논문
- AASHTO(2010), Highway Safety Manual
- Arthur, D. and S. Vassilvitskii(2006), How Slow is the K-means Method. Proceedings of the Twenty-Second Annual Symposium on Computational Geometry, ACM, p144-153.
- Bengio, Y. A. Courville, and P. Vincent(2013). Representation Learning: A Review and New Perspectives. IEEE Trans. PAMI, special issue Learning Deep Architectures, Vol. 35, No, 8, pp. 1798-1828
- Bradford K. Brimley, Mitsuru Saito, Grant G. Schultz(2012), Calibration of the Highway Safety Manual Safety Performance Function and Development of New Models for Rural Two-Lane Two-Way Highways, Transportation Research Board Annual Meeting
- Christian Szegedy, Christian, Alexander Toshev and Dumitru Erhan(2013), Deep neural networks for object detection, Advances in Neural Information Processing Systems
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze(2008), Introduction to Information Retrieval, Cambridge University Press,

pp.331–338

Diederik P. Kingma, Jimmy Lei Ba(2015), ADAM: A Method For Stochastic Optimization, International Conference for Learning Representations, Vol.9, pp. 1–15

Ducknyung Kim, Dong–Kyu Kim, Chungwon Lee(2013), Safety Performance Functions Reflecting Categorical Impact of Exposure Variables for Freeways, Transportation Research Board Annual Meeting

Fillip Martinelli, Francesca La Torre, Paolo Vadi(2009), Calibration of the Highway Safety Manual Accident Prediction Model for Italian Secondary road network, Transportation Research Board Annual Meeting

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel–rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury(2012), Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups, IEEE Signal Processing Magazine, Vol. 29, pp. 82–97

Guangyuan Pan, Liping Fu, Lalita Thakali(2017), Development of a global road safety performance function using deep neural networks, International Journal of Transportation Science and Technology, Vol.6, No.3, pp. 159–173

Helai Huang, Qiang Zeng, Xin Pei, S.C. Wong, Pengpeng Xu(2016), Predicting crash frequency using an optimised radial basis function neural network model, Transportmetrica A: Transport Science, Vol.12, No 4, pp.330–345

Hochan Kwak, Dong–kyu Kim, Shin Hyoung Park, Seung–Young Kho(2010), Development of a Safety Performance Functions for Korean Expressways, 12th World Conference on Transport Research (WCTR) – Lisbon, Portugal, pp. 1–12

Jinyan Lu, Kirolos Haleem, Priyanka Alluri, Albert Gan(2013), Full versus

Simple Safety Performance Functions: A Comparison Based on Urban Four-Lane Freeway Interchange Influence Areas in Florida, Transportation Research Board Annual Meeting

Krizhevsky, Alex(2010), Convolutional Deep Belief Networks on CIFAR-10, Unpublished manuscript (downloaded from cs.toronto.edu), Vol. 1, pp. 1-9

Lee J. K. (2009), A Study on Prediction of the Early-Age Strength of Concrete using Artificial Neural Network Theory, Joongbu University graduate school, A master Dissertation.

Mohamadreza Banihashemi(2012), Highway Safety Manual, Calibration Dataset Sensitivity Analysis, Transportation Research Board Annual Meeting

Salvatore Cafiso, Carmelo D'Agostino Bhagwant Persaud(2013), Investigating the influence of segmentation in estimating safety performance functions for roadway sections, Transportation Research Board Annual Meeting

Simon, W. P., Matthew, G. K., Fred L, M, (2010), Statistical and Econometric Method For Transportation Data Analysis, CRC Press

Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan "Honza" Cernocky, Sanjeev Khudanpur(2010), Recurrent neural network based language model, Interspeech, pp. 1045-1048

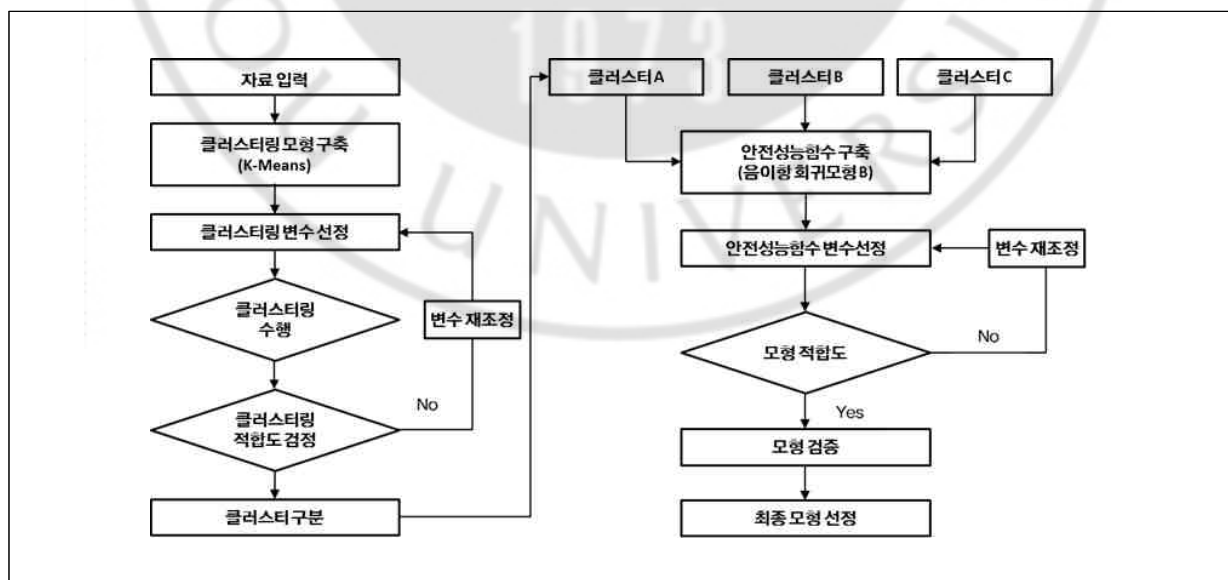
Yingfei TU, Jiangman ZHANG, Chao YANG, Xiaohing CHEN(2012), Crash Frequency Analysis for Urban Expressways by Considering Segment Type, Transportation Research Board Annual Meeting

부록 1. 클러스터링 기법을 이용한 고속도로 교통사고 예측모형 구축

1. 구축 배경 및 절차

본 연구에서는 고속도로 콘존을 유형화 하여 고속도로 교통사고 예측모형을 고도화하려는 시도를 수행하였다. 콘존은 유형화가 되어 있지 않지만, 각 콘존의 특성 자료(예, 교통량 수준 등)를 비슷한 유형으로 묶은 후 고속도로 교통사고 예측모형을 구축하면 예측력을 더욱 높일 수 있을 것이라는 기본 아이디어를 바탕으로 연구를 진행하였다. 콘존을 유형화하기 위해 독립변수들을 이용하여 클러스터링 기법인 K-Means 기법을 이용하여 유형화 하였으며, 각 유형화된 콘존 자료를 학습데이터와 테스트 데이터로 구분하여 안전성능함수들을 구축하고 예측 결과를 평가 하였다.

본 연구에서 수행된 K-Means 기법을 이용하여 콘존을 유형화하고 이를 바탕으로 고속도로 교통사고 예측모형을 도출하는 과정은 다음 그림과 같다.



<그림 26> 클러스터링 기법을 이용한 안전성능함수 고도화 절차

2. 클러스터링 기법을 통한 자료 유형 구분

본 연구에서는 클러스터링 기법을 통해 콘존을 유형화하기 위해 비지도 학습(unsupervised learning) 알고리즘인 K-means 알고리즘을 이용하여 데이터를 군집화 하였다. K-means 알고리즘은 클러스터의 개수인 K값을 기준으로 클러스터를 구분한다. 주요 장점은 계산이 빠르고 간편하지만, 적절한 클러스터 개수 선정이 필요한 단점이 있다. 본 연구에서는 클러스터 개수인 K 값을 3으로 하여 클러스터링을 수행하였다. 클러스터링 수행을 위해서는 모든 변수에 대해 전체 자료의 분표를 평균 0, 분산 1이 되도록 스케일링 하는 것이 필요하다. 본 연구에서는 Scikit-Learn의 전처리 기능을 이용하여 데이터들을 스케일링 하였다. 특히 Robust Scaler를 이용하여 outlier의 영향을 최소화 할 수 있도록 하였다.

연구의 총 데이터는 977개로 클러스터 수인 K 값을 증가시키면 클러스터별 데이터 수가 적어져 모형 구축에 어려움이 있을 것으로 판단되어 클러스터 수를 3으로 선정하여 분석하였다. 또한 클러스터링의 수행 결과를 파악하기 위해 Calinski-Harabasz Index와 Silhouette Coefficient를 사용하여 결과를 비교하는데 사용하였다. Calinski-Harabasz Index는 클러스터내 분산과 클러스터간 분산사이의 비율로 정의되며, 클러스터가 잘 정의되면 Index 가 높다. Silhouette Coefficient는 같은 클러스터내의 요소들 사이의 평균거리와 다른 클러스터 내에서의 요소들 사이에서의 평균거리를 이용한 방법으로 -1에서 1사이의 값을 갖는데 1에 가까울수록 군집분석이 잘 수행되었다고 볼 수 있다. 클러스터링에 사용된 변수는 콘존길이, AADT를 이용하여 클러스터링을 수행하였다. Calinski-Harabasz Index는 2,390.24, Silhouette Coefficient는 0.67의 결과가 나타났다.

<표 25> 클러스터별 데이터 수

K 값 (클러스터 수)	Calinski-Harabasz Index	Silhouette Coefficient	클러스터별 데이터 수	
3	2,390.24	0.67	클러스터 A	294
			클러스터 B	201
			클러스터 C	482

클러스터별 데이터 특성을 파악하기 위해 대표적인 변수의 콘존길이와 AADT에 대하여 기초 통계분석을 수행하였다.

<표 26> 클러스터별 콘존길이 기초 통계

구분	콘존길이(m)			
	평균	표준편차	최소	최대
클러스터 A	14,075.5	4,212.6	8,860.0	30,790.0
클러스터 B	3,795.7	2,564.6	240.0	13,270.0
클러스터 C	4,789.4	2,514.3	110.0	9,800.0

<표 27> 클러스터별 AADT 기초 통계

구분	AADT(대/일)			
	평균	표준편차	최소	최대
클러스터 A	15,096.1	8,978.0	1,625.0	53,520.0
클러스터 B	77,823.2	16,828.9	50,857.0	118,601.0
클러스터 C	23,097.6	12,263.9	542.0	49,050.0

3. 안전성능함수 구축 및 검증

클러스터 별 고속도로 교통사고 예측모형을 구축하기 위하여 클러스터된 자료를 각각 학습데이터 80%와 테스트 데이터 20%로 구분하였다. 학습데이터를 이용하여 고속도로 교통사고 예측모형을 구축하고 테스트 데이터를 이용하여 예측력을 검증하였다. 우선 클러스터별 안전성능함수 구축 결과는 다음과 같다.

가. 클러스터 A 안전성능함수

클러스터 A에 대하여 음이항 회귀모형을 수행한 결과 콘존길이와 AADT가 유의한 변수로 도출되었다. 그리고, 모형 적합도를 설명하는 AIC 및 BIC 값이 모두 통계적 기법을 이용한 기존 음이항 회귀모형 B에 비하여 떨어지는 것으로 분석되었다.

<표 28> 클러스터 A의 음이항 회귀모형 결과

모형	음이항 회귀모형 클러스터 A			
관측수	235			
Log-Likelihood	-749.88			
LL-Null:	-784.61			
Pseudo-R2	0.04			
AIC	1505.76			
BIC	-1188.16			
변수	계수	S.E	p-value	유의수준
상수	-15.135	1.466	0.000	***
콘존길이	1.039	0.137	0.000	***
AADT	0.779	0.065	0.000	***

나. 클러스터 B 안전성능함수

클러스터 B에 대하여 음이항 회귀모형을 수행한 결과 콘존길이와 AADT가 유의한 변수로 도출되었다. 그리고, 클러스터 A의 경우와 마찬가지로 모형 적합도를 설명하는 AIC 및 BIC 값이 모두 통계적 기법을 이용한 기존 음이항 회귀모형 B에 비하여 떨어지는 것으로 분석되었다.

<표 29> 클러스터 B의 음이항 회귀모형 결과

모형	음이항 회귀모형 클러스터 B			
관측수	160			
Log-Likelihood	-442.79			
LL-Null:	-485.44			
Pseudo-R2	0.09			
AIC	891.57			
BIC	-708.03			
변수	계수	S.E	p-value	유의수준
상수	-14.861	3.363	0.000	***
콘존길이	1.017	0.087	0.000	***
AADT	0.746	0.283	0.008	***

다. 클러스터 C 안전성능함수

클러스터 C에 대하여 음이항 회귀모형을 수행한 결과 콘존길이와 AADT가 유의한 변수로 도출되었다. 그리고, 클러스터 A 및 B의 경우와 마찬가지로 모형 적합도를 설명하는 AIC 및 BIC 값이 모두 통계적 기법을 이용한 기존 음이항 회귀모형 B에 비하여 떨어지는 것으로 분석되었다.

<표 30> 클러스터 C의 음이항 회귀모형 결과

모형	음이항 회귀모형 클러스터 C			
관측수	385			
Log-Likelihood	-898.45			
LL-Null:	-972.75			
Pseudo-R2	0.08			
AIC	1802.90			
BIC	-2037.96			
변수	계수	S.E	p-value	유의수준
상수	-14.178	1.103	0.000	***
콘존길이	0.774	0.071	0.000	***
AADT	0.908	0.085	0.000	***

라. 예측력 검증

구축된 클러스터별 안전성능함수의 예측력을 비교하기 클러스터별 테스트 데이터를 이용하여 MAD, RMSE, SMAPE로 검증하였다. 검증 결과는 다음과 같다.

<표 31> 클러스터별 음이항 회귀모형 예측력 검증 결과

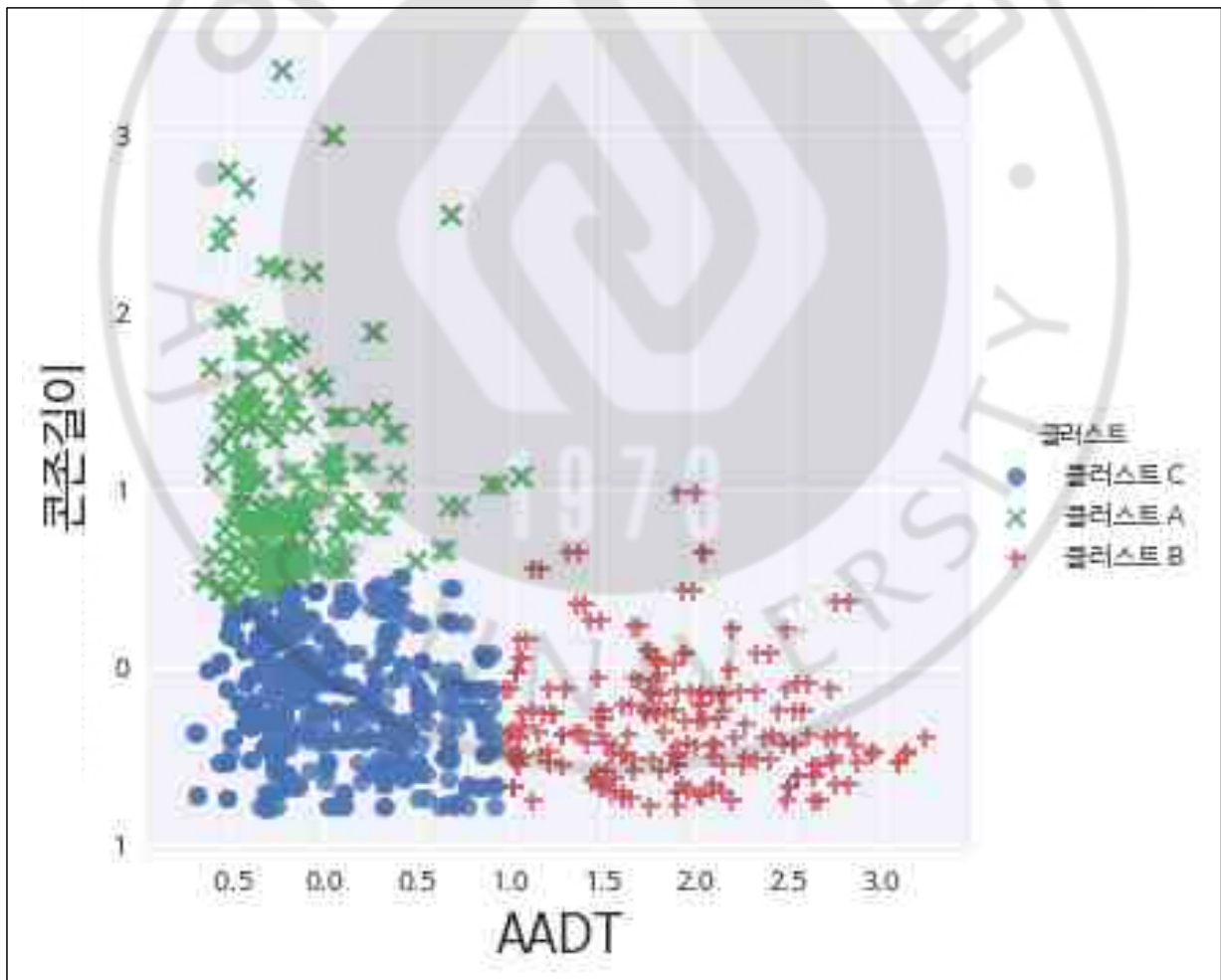
구분	클러스터 A	클러스터 B	클러스터 C	평균
MAD	3.96	2.51	2.39	2.89
RMSE	5.42	3.28	3.59	4.07
SMAPE	0.21	0.22	0.27	0.24
데이터수	294.00	201.00	482.00	977.00

검증 결과, 클러스터 A를 제외하고 클러스터 B, 클러스터 C는 MAD가 2.51, 2.39로 낮게 나타났으며, 클러스터링의 효과를 파악하기 위해 평균은 데이터 수 가중 평균을 이용하였다. 데이터수 가중 평균 결과, MAD는 2.89, RMSE는 4.07, SMAPE는 0.24로 나타났다.

4. 시사점 도출

본 연구에서 K-Means 기법을 안전성능함수 개발에 적용하고자 한 목적은 앞서서도 밝힌 바와 같이 콘존의 특성 자료(예, 교통량 수준 등)를 비슷한 유형으로 묶은 후 고속도로 교통사고 예측모형을 구축하면 모형의 예측력을 더욱 높일 수 있을 것이라는 기본 아이디어를 기본으로 한다.

하지만, 실질적으로 K-Means 기법을 적용하여 콘존을 세 가지 클러스터로 구분하여 고속도로 교통사고 예측모형을 각각 도출한 결과 모형 적합도 및 모형 예측력 모두에서 부정적인 결과를 보였다.



<그림 27> 콘존길이 및 AADT 기준 클러스터링 결과 시각화

이러한 결과가 나타난 원인에 대하여 분석한 결과, 클러스터링을 통하여 전체 데이터를 유형화한 경우 아래 그림에서 보이는 바와 같이 고속도로 교통사고 예측모형을 도출하는 범위를 축소시킴으로써 오히려 모형의 적합도 등을 떨어뜨리는 것으로 확인되었다. 클러스터링을 수행하게 되면, 종속변수의 교통사고 건수의 데이터가 존재하는 범위는 변화가 없으면서, 독립변수의 경우에는 데이터가 존재하는 범위는 줄어들게 된다. 즉, 클러스터 C의 경우 콘존 길이는 대략 0~1km 범위로 줄고, AADT는 0~50,000대/일 수준으로 줄어들게 된다. 결론적으로 종속변수의 범위는 그대로 유지되지만, 독립변수들의 범위가 줄어들어 적합한 모형을 찾는 데 더욱 어려움을 주게 되는 것으로 확인되었다.

따라서 본 연구에서 가정하였던 독립변수들의 특성을 기반으로 자료를 비슷한 유형으로 묶은 후 고속도로 교통사고 예측모형을 구축하려는 시도는 적절하지 않은 것으로 확인되었다. 또한 향후연구에서 이러한 문제점을 해결하기 위해서는 클러스터링 대신에 분류(classification)을 적용하는 등 다양한 시도를 해볼 필요는 있을 것으로 판단된다.

Abstract

In the past, most traffic accident data analysis has been carried out based on the traditional statistical methods, including Poisson regression model or negative binomial regression model. These statistical methods can be used to find important relationships between various human, road geometrical and environmental factors related to traffic accidents, to predict the frequency of traffic accidents, and to calculate the traffic safety grade based on the analyzed results. However, recent approaches using big data analysis techniques such as machine learning and deep learning have begun to attract attention. Such machine learning and deep learning techniques have shown advantages in analyzing factors related to traffic accidents by utilizing other kinds of mass data. It has been actively applied in traffic and other fields and is changing our daily life. The purpose of this study is to predict the frequency of traffic accidents in Congestion Zone(a.k.a, Conzone) using highway traffic accident data. To do this, we applied the traditional statistical technique and the method using the deep learning, and compared the prediction performance of each technique. As a result of the prediction performance comparison, it was found that the MOEs of the deep learning model are somewhat superior to those of the traditional statistical model. However, the difference is 0.27 on the MAD basis, and it was found that the number of traffic accidents can be sufficiently predicted even on the basis of traditional statistical techniques. In particular, the negative binomial regression model is superior to the Poisson regression model. Also, it is considered that the model using the exposure (i.e., AADT, Conzone length) is more excellent.

However, using deep learning can increase predictive reliability even further. Also, in case of deep learning, there are few cases that are used to predict the number of traffic accidents. In particular, it is confirmed that not only the number of hidden layers and nodes, but also the optimizer and node structure are affected.

