



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](#)

석 사 학 위 논 문

DSRC 데이터를 활용한
k-NN 알고리즘 기반
교통사고 위험예측 기법 연구

계 명 대 학 교 대 학 원

도 시 계 획 및 교 통 공 학 과

강 민 지

지도교수 박 신 형

2 0 1 8 년 2 월

DSRC 데이터를 활용한 k-NN 알고리즘 기반 교통사고 위험예측 기법 연구

강
민
지

2
0
1
8
년

2
월

DSRC 데이터를 활용한
k-NN 알고리즘 기반
교통사고 위험예측 기법 연구

지도교수 박 신 형

이 논문을 석사학위 논문으로 제출함

2 0 1 8 년 2 월

계 명 대 학 교 대 학 원

도 시 계 획 및 교 통 공 학 과

강 민 지

강민지의 석사학위 논문을 인준함

주 심 권 오 훈

부 심 박 신 형

부 심 박 용 진

계 명 대 학 교 대 학 원

2 0 1 8 년 2 월

목 차

제 1 장 서론	1
1.1 연구의 배경 및 목적	1
1.2 연구의 범위 및 방법	3
제 2 장 선행연구 고찰	5
2.1 선행연구 고찰	5
2.1.1 사고예측모형에 관한 연구	5
2.1.2 k-NN 알고리즘을 활용한 연구	7
2.2 연구방향 설정	10
제 3 장 자료구축	13
3.1 자료수집	13
3.1.1 교통소통 이력자료	13
3.1.2 교통사고 이력자료	14
3.1.3 링크 정보 자료	16
3.1.4 기상데이터	17
3.2 데이터 전처리 및 가공	19
3.2.1 교통소통 이력자료에 대한 전처리 및 가공	19
3.2.2 교통사고 이력자료에 대한 전처리	21
3.3 통합데이터블 구축	23
제 4 장 사고 위험 예측 모형 설계	25
4.1 분석 방법론	25
4.1.1 k-Nearest Neighbors Algorithm	25
4.1.2 Random Forest Model	26

4.2 변수선정	27
4.2.1 개요	27
4.2.2 차원의 축소	28
4.2.3 변수의 중요도 평가	31
4.3 사고 위험 예측 모형 개발	33
4.3.1 개요	34
4.3.2 불균형 데이터 처리(Imbalanced Data Processing)	34
4.3.3 데이터 정규화(Data Normalization)	36
4.3.4 분석 데이터	38
4.3.5 k-NN의 하이퍼파라미터(Hyper Parameter) 선정	38
제 5 장 모형 검증	46
5.1 검증 방법	46
5.1.1 Confusion Matrix	46
5.1.2 평가 매트릭(Evaluation Metric)	47
5.1.3 ROC Curve	49
5.1.4 종합	50
5.2 사례 연구	50
5.2.1 사례 연구 대상지 개요	50
5.2.2 모형의 적용 결과	53
5.2.3 결과 해석	57
제 6 장 결론 및 향후연구	62
6.1 결론 및 의의	62
6.2 향후연구	64
참 고 문 헌	67
영 문 초 록	71
국 문 초 록	74

표 목 차

<표 2-1> 모수 모형과 비모수 모형의 비교	10
<표 2-2> 본 연구의 약어 정리	12
<표 3-1> 교통소통 이력데이터의 수집 - 5분 단위 수집(예시)	14
<표 3-2> 사고 이력자료의 속성	14
<표 3-3> DSRC 링크 정보 데이터	16
<표 3-4> 기상데이터 자료의 수집(예시)	18
<표 3-5> 통합테이블(예시)	24
<표 4-1> 19개의 변수후보	27
<표 4-2> 최종 선정 변수	30
<표 4-3> RF Model을 이용한 각 변수의 중요도	32
<표 4-4> k값 설정 기준	40
<표 4-5> 거리 가중치 산정을 위한 Kernel Function	42
<표 4-6> 최종 선정 하이퍼파라미터	45
<표 5-1> Confusion Matrix	46
<표 5-2> Confusion Matrix의 속성	47
<표 5-3> 본 연구에서 정의한 Threshold	47
<표 5-4> 평가 메트릭	48
<표 5-5> 달구벌대로 DSRC 링크 - 사고 다발 순	52
<표 5-6> Weight 산출 값	53
<표 5-7> Distance 산출 값	54
<표 5-8> 사고 발생 확률 결과 값	54
<표 5-9> Threshold에 따른 각 평가 메트릭의 변화	56
<표 5-10> 실제 값과 예측 값 비교	57
<표 5-11> AUC에 따른 모형의 성능	58

그 립 목 차

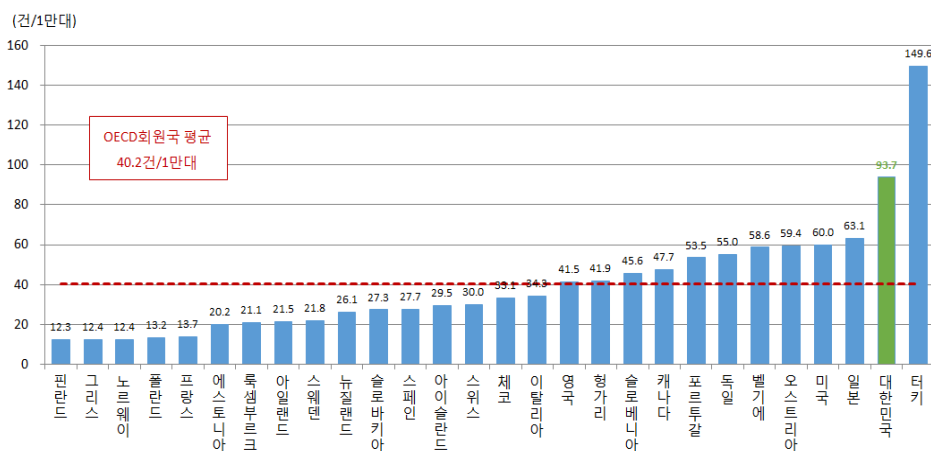
<그림 1-1> 2014년 자동차 1만대 당 교통사고 발생건수	1
<그림 1-2> 머신러닝 개념도	3
<그림 1-3> 연구의 수행과정	4
<그림 2-1> k-NN을 이용한 사고위험예측 개념도	11
<그림 3-1> DSRC 데이터 수집 과정	13
<그림 3-2> x, y 좌표 정보가 포함된 교통사고 이력데이터	15
<그림 3-3> 수치지도에 표출한 2014~2015년 대구광역시 교통사고	15
<그림 3-4> 174개의 DSRC 수집 장치 위치도	17
<그림 3-5> 교통소통 이력데이터의 표준편차 데이터 추가 생성	20
<그림 3-6> ArcGIS의 Proximity Analysis 도식	21
<그림 3-7> 사고 이력 자료의 DSRC 링크 정보 추가 생성	22
<그림 3-8> DSRC 링크 상에서 발생한 데이터 추출	22
<그림 3-9> 통합테이블의 도식화	23
<그림 3-10> 구축한 통합테이블 도식화	24
<그림 4-1> Random Forest 도식화	26
<그림 4-2> 차원의 저주(Curse of Dimensionality)	28
<그림 4-3> 변수 간 Scatter Plot Matrix	29
<그림 4-4> 변수 간 상관성(1)	30
<그림 4-5> 변수 간 상관성(2)	30
<그림 4-6> Variable Importance Plot	31
<그림 4-7> 알고리즘 개발 과정	33
<그림 4-8> 불균형 데이터 해결 방법(Under sampling, Over sampling)	35
<그림 4-9> SMOTE 도식	36

<그림 4-10> 개발 k-NN 알고리즘	39
<그림 4-11> Kernel Function의 종류별 분포	41
<그림 4-12> Kernel Function별 ROC Curve 비교	43
<그림 5-1> ROC curve(예시)	49
<그림 5-2> 링크 정보 Sheet	51
<그림 5-3> 검증 대상지 설정 기준	52
<그림 5-4> ROC curve (AUC:0.704)	58
<그림 5-5> 예측 결과의 확률밀도분포	59
<그림 5-6> Threshold에 따른 정확도	60

제 1 장 서론

1.1 연구의 배경 및 목적

우리나라는 1970년대부터 고도의 경제성장기를 거치며, 상당 기간 도로(인프라)의 확충과 교통소통관리 위주의 교통정책을 우선시 해왔기 때문에 상대적으로 교통안전에 대한 대책과 투자 노력이 미흡하였다. 그 결과, 지난 25년간 OECD(Organization for Economic Cooperation and Development, 경제협력개발기구) 국가 중 가장 빠르게 성장하며 세계 11위 경제 대국으로서의 위상은 높은 반면, 교통안전 분야에서는 OECD 국가 32개국 중 31위를 기록하며 여전히 최하위권의 교통안전 수준을 보이고 있다. 2014년 자동차 1만 대당 교통사고 발생 건수는 93.7건으로 평균 40.2건에 비해 약 2.3배 높으며, 자료가 파악된 OECD 28개국 중 27위를 기록하였다. 또한, 인구 10만 명당 교통사고 발생건수는 28개국 중 25위, 10억 주행 km 당 교통사고 발생건수는 16개국 중 16위를 차지하였다(도로교통공단, 2016).



<그림 1-1> 2014년 자동차 1만대 당 교통사고 발생건수

(자료 : 도로교통공단, 2016)

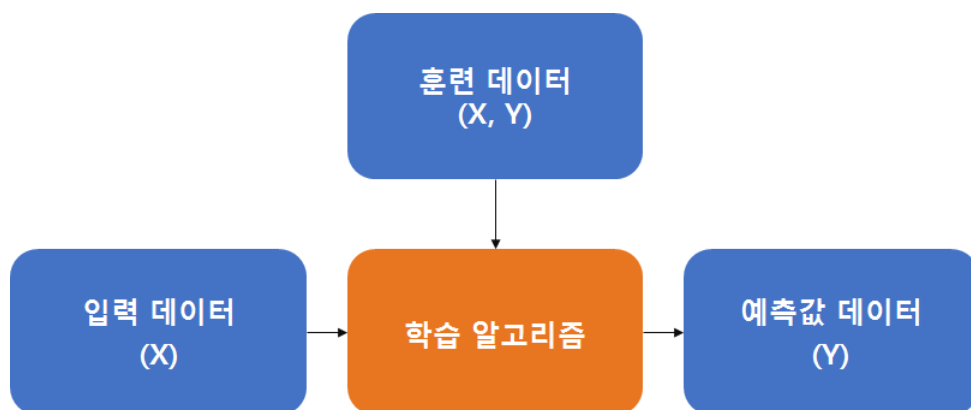
이러한 통계 결과는 우리나라가 교통안전 선진화를 위한 노력이 절실함을 보여준다. 현재 정부에서는 교통안전 선진화를 위한 노력 중 하나로 2016년 제8차 국가교통안전기본계획¹⁾을 통해 ‘2021년까지 도로부문 교통사고 사망자 수 2,700명까지 줄여나가겠다고 밝히며 교통사고 감소를 위한 정책적 기조를 유지하고 있다(국토교통부, 2016; 국토교통부, 2017). 우리나라는 교통사고 감소를 위한 대표적인 대책 중 하나로 교통사고 잦은 곳의 문제점을 개선하는 등 교통 환경 개선 사업을 활발히 실시하고 있다. 이는 사후적인 조치로서 중요하지만, 반드시 사고다발지점에만 사고 위험이 있는 것은 아니기 때문에 사고위험이 있는 도로를 찾아 위험정보를 알려주는 사전적인 조치(Proactive-response) 또한 중요하다.

최근 정보통신기술의 발전으로 대용량 데이터의 수집과 저장, 분석이 가능해지면서 빅 데이터의 활용이 주목받고 있다. 이에 따라 정부에서도 데이터의 개방과 활용을 적극적으로 장려하고 있는데, 교통 분야에서는 ITS(Intelligent Transport System, 지능형 교통 시스템)의 세부 분야인 ATMS(Advanced Traffic Management System, 첨단교통관리시스템)을 통해 실시간으로 교통정보를 수집 및 저장하고 있어 대용량의 교통정보를 활용한 연구가 가능해졌다. 이러한 데이터 기반을 통해 대용량의 데이터로부터 패턴이나 규칙을 찾아 유용한 정보를 추출하는 데이터 마이닝(Data Mining) 기법 또한 많은 관심을 받고 있다. 본 연구에서는 기 수집된 교통정보 빅데이터를 바탕으로 데이터 마이닝을 통해 교통사고 위험을 예측하는 것을 목적으로 한다.

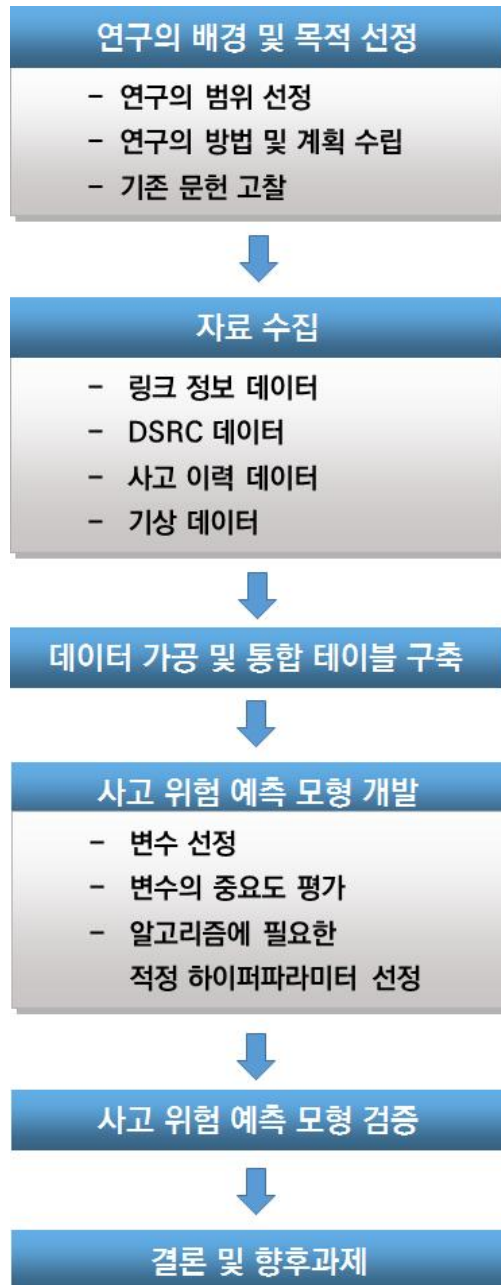
1) 5개년 단위로 도로·철도·항공·해양 분야를 포함하여 교통안전에 관한 중·장기 목표 및 교통안전 정책방향을 제시하는 국가 종합계획

1.2 연구의 범위 및 방법

본 연구는 대용량의 이력자료를 데이터 마이닝 한 사고위험예측 기법 개발을 목적으로 한다. 연구의 공간적 범위는 대구광역시 도시부 간선도로이며, 모형 검증은 대구광역시의 주요 간선도로 중 하나인 달구벌대로의 최상위 사고다발링크를 대상으로 하였다. 사고위험예측은 데이터 마이닝 기법의 머신 러닝(Machine Learning)을 활용하여 개발하였으며, 기존 문헌 고찰을 통해 여러 머신 러닝 기법 중 본 연구에서 사용할 알고리즘을 선정하였다. 머신러닝은 경험적 데이터를 기반으로 학습을 하고 학습된 내용을 통해 예측을 수행하는 알고리즘을 의미한다. 따라서 예측의 기반이 되는 학습 데이터의 구축이 필요한데, 본 연구에서는 2년간(2014~2015년) 대구광역시에서 발생한 사고데이터와 동일 기간의 ATMS로부터 수집한 교통소통정보, 기상청의 기상데이터를 토대로 이를 구축하였다. 사고 위험 예측 알고리즘을 개발하고 개발 알고리즘의 검증 및 절차를 설계하였으며, 검증 과정에서 파라미터를 최적화하여 보정함으로써 최적의 모형이 도출될 수 있도록 하였다. 본 연구의 전체적인 수행과정은 <그림 1-3>과 같다.



<그림 1-2> 머신러닝 개념도



<그림 1-3> 연구의 수행과정

제 2 장 선행연구 고찰

2.1 선행연구 고찰

2.1.1 사고예측모형에 관한 연구

한상진, 김근정, 오순미(2008)는 회귀분석을 이용한 사고예측모형은 다중 공선성으로 인해 중요한 변수가 누락될 가능성이 있고, 각 설명변수들과 사고율 사이의 인과관계를 구하기가 어려우며 변수별로 함수식을 다르게 하는 경우가 거의 없어 변수의 특성이 모형식에 반영하기 어렵다는 회귀분석의 한계점에 대해 정리하였다.

백승걸, 장현호, 강정규(2005)는 기존 사고 예측 관련 연구들은 주로 특정 지점에서의 도로기하 구조조건, 교통 및 환경조건들과 교통사고와의 관계를 설명하기 위한 모형의 개발에 초점을 두었으며, 그렇기 때문에 도로와 교통조건이 양호한 구간에서의 높은 교통사고율을 설명하기 어렵다는 한계를 가지고 있다고 언급하였다.

박준태, 이수범, 이동민(2011)은 국내의 사고예측과 관련된 연구는 대부분 상관분석 등을 통해 교통사고를 설명할 수 있는 변수들을 선정하여 선형회귀, 비선형회귀, 로지스틱 곡선, 음지수 분포 등의 회귀식과 회귀분석, 시계열분석, 수량화 분석 등을 이용하여 변수들과 교통사고와의 관계를 규명하였으나, 연구의 대부분이 사고 관련 자료의 불충분으로 인한 변수의 부족과 표본수의 부족, 사고건수 대신 상충수와의 비교 등으로 인하여 모형에 대한 정확성이 낮게 나타나고 있다고 언급하였다. 또한, 국외의 사고예측과 관련된 연구는 비선형 회귀분석을 이용한 연구가 주를 이루었으며, 사고에 영향을 미치는 변수를 고려함에 있어 보다 다양한 변수를 고려하고 있다고 하였다.

이재명, 김태호, 이용택, 원제무(2008)는 최근 사고예측 관련 연구에서는

도로구간의 개별특성 자료를 분석하는 미시적 분석을 도입하고 있으며 분석기법도 확률회귀분석모형(포아송, 음이항 회귀분석) 또는 변수의 비선형 특성을 설명하는 데에 우수한 인공신경망모형 등을 다양하게 이용하고 있다고 언급하였다. 그러나 기존 교통사고예측모형은 설명력을 높이기 위하여 교통사고와의 상관관계가 높은 변수로만 선형회귀모형을 개발하여 다양한 요인을 모형 내에 반영하지 못하거나, 교통사고의 비선형적 특성을 반영하기 위해 비선형 모형을 이용하더라도 교통사고관련 자료의 분산이 매우 커서 모형의 적용성에는 한계를 안고 있다는 문제점을 제시하였다.

Pirdavani et al.(2014)는 루프검지기의 데이터와 실시간 교통 데이터, 사고데이터를 통합하여 실시간 충돌위험예측 모델을 개발하였다. 예측모델은 이항 로지스틱 회귀분석을 사용하였으며, Confusion Matrix와 ROC Curve를 이용하여 모형을 검증하였다. 모델의 검증 결과, 사고가 발생하는 경우의 60% 이상을 예측할 수 있는 반면, 사고가 발생하지 않는 경우는 90% 이상을 예측하는 성능을 나타냈다.

Lin, Wang과 Sadek(2015)는 실시간 교통사고 위험 예측 모형을 위해 빈발패턴(Frequent Pattern)에 기반한 새로운 변수 선택 방법에 대해 연구하였다. 빈발패턴나무와 랜덤 포레스트(Random Forest)를 이용하여 8가지 사고발생에 위험을 주는 변수를 선정하였으며, k-NN 알고리즘과 베이지안 네트워크를 이용하여 사고예측을 통해 어떤 방법의 변수 선정이 예측률이 더 높았는지 확인하였다.

Lv, Tan과 Zhao(2009)는 k-NN 알고리즘을 활용하여 고속도로 교통사고 예측모형을 개발하였다. k-NN 알고리즘을 적용함으로써 교통사고의 유발가능성이 더 높은 교통 조건을 식별하고, 사고 유발 요인이 교통사고 발생에 미치는 영향을 고려하였다. ATMS 실시간 교통데이터와 사고이력자료를 통합하여 분석데이터로 활용하였으며, C-means 알고리즘과의 예측모형의 성능 비교 결과 k-NN 알고리즘이 더 우수함을 증명하였다. 또한, 이 연구는 k-NN 알고리즘을 활용하여 교통사고 예측모형을 적용한 최초의 연구라 할 수 있다.

Sun과 Sun(2016)은 랜덤 포레스트 모델을 사용하여 중요한 변수를 선정하

는데 사용하고, 서포트 벡터 머신(Support Vector Machine)과 k-means clustering을 결합한 하이브리드 모델로 사고예측모형을 개발하였다. 변수 선택 없이 서포트 벡터 머신을 활용한 경우와, 변수를 선택한 서포트 벡터 머신을 활용한 경우, 서포트 벡터 머신과 k-means clustering을 결합한 하이브리드 모델 총 3가지 경우를 비교하였다. 비교 결과, 하이브리드 모델과 변수를 선택하지 않았을 때 보다 랜덤 포레스트를 이용하여 변수를 선택한 경우의 예측 성능이 더 뛰어난 것을 확인하였다.

국내의 사고예측모형 관련 연구는 주로 모수적 기법을 이용한 연구가 다수였으며, 모수적 기법의 문제점을 설명한 연구들이 있었으나, 비모수적 기법을 활용한 연구는 많이 이루어지지 않았다. 국외의 사고예측모형 관련 연구에서는 실시간으로 얻을 수 있는 교통데이터를 활용한 연구가 많았다. 교통데이터, 사고 이력데이터 등 여러 가지 수집 데이터를 결합하여 사고위험을 예측하는 연구를 하였으며, 다수의 연구가 비모수 기법을 이용하였다.

2.1.2 k-NN 알고리즘을 활용한 연구

신강원, 심상우, 최기주, 김수희(2014)는 k-NN 알고리즘을 활용한 고속도로 통행시간 예측 연구를 수행하였다. 통행시간 예측과 관련해서 여러 연구가 진행되었으나 신경망은 학습과정이 매우 복잡하다는 단점이 있으며 회귀모형은 다수의 모형(기종 점, 경로)을 개발해야한다는 문제점이 있는 반면, k-NN 알고리즘의 경우 이런 문제를 해결할 수 있고 참조할 수 있는 데이터가 충분하다면 타 모형에 비해 정확도가 우수한 장점이 있다고 하였다. k-NN 기반의 예측 모형 입력 자료는 실시간 교통상황을 반영할 수 있는 TCS(Toll Collection System) 교통량과, DSRC(Dedicated Short Range Communications) 링크 통행시간을 활용하였고 유사성은 유클리디안 거리를 통해 산출하였으며, 최 근접 이웃은 5% 이내로 산정하였다.

최재익(2016)은 k-NN이 분류나 회귀에 사용되는 비모수방식이며, k-NN 알고리즘을 적용한 교통예보의 기본적인 개념은 기준일과의 유사도를 측

정하고 가장 유사도가 높은 k 개의 일자를 선택하여 유사도가 높은 k 개 일자의 이력자료에 가중치를 이용해 기준일의 예측을 수행하는 것이라 하였다. 예측모형에 활용한 데이터는 TCS, VDS(Vehicle Detector System), DSRC, 기상데이터이다

이승봉, 한동희, 이영인(2015)은 k -NN 알고리즘을 이용하여 교통사고 처리시간 예측 모형을 개발하였으며, k -NN 알고리즘이 현재 조건과 유사한 과거의 조건을 탐색하여 장래의 상태를 예측하는데 적용이 용이하다고 하였다. 기존 모형은 어떤 상황이 변화되면 기존의 수학적 모형의 경우 새로이 추가되는 입력 값과 변수를 수정해야하는 번거로움이 있지만 k -NN은 입출력 값과 모형의 재구성이 매우 용이한 탄력성이 있다고 하였다. k 값 설정은 예측오차를 최소화 할 수 있는 k 값을 이용하였고, 평균적으로 약 10개의 이웃을 추출할 경우에 모형의 정확도가 높은 것으로 나타났다.

김혜원과 이영인(2015)은 k -NN 알고리즘을 사용하여 고속도로 교통사고 발생 시 사고대응시간 예측모형을 개발하였다. 과거의 교통사고 이력자료를 바탕으로 NPR(Non Parametric Regression) 모형의 일종인 k -NN을 사용하였으며, NPR 모형은 설명변수를 고려하는 모수 회귀식과 달리 설명변수를 고려하지 않는 특징이 있어 각각의 독립변수에 대한 특별한 가정 없이 예측 할 수 있다고 하였다. 예측모형은 유클리디안 거리를 이용하여 데이터 간 거리를 산출하고, 실제 사고처리시간과 추정된 통행시간 예측오차를 최소화 할 수 있는 k 개를 직접 분석하여 설정하였다. 모형의 검증방법은 MAPE(Mean Absolute Percentage Error), MAE(Mean Absolute Error)를 활용하였다.

김형주, 박신형, 장기태(2016)는 실시간 자료를 기반으로 k -NN을 이용한 단기교통상황을 예측하였다. 회귀모형은 교통상황이 변화되는 상황은 반영하지 못하는 단점 이 있으며, k -NN은 단순한 유사도 기반 자료매칭으로 모형이 간단하고 연 산시간이 짧아 단기 교통상황예측에 장점이 있다고 언급하였다. 짧은 연산 시간으로 급변하는 교통상태 변화에 적합한 유클리디안 거리를 적용하고 k 값은 시행착오 방법(Trial and Error)을 통하여 결정하였다.

김은미와 홍태호(2015)는 사례기반추론 기반(Case Based Reasoning, CBR)의 예측모형을 제시하였다. 사례기반추론은 적용이 쉽고 간단하며 모형의 갱신이 실시간으로 이루어질 수 있다고 하였다. 또한, 일반적으로 타 인공지능법에 비해 성과가 낮다고 알려져 있으나 입력변수의 중요도에 따라 가중치를 상이하게 적용할 경우에는 예측성과를 향상시킬 수 있다고 하였으며, 변수의 중요도를 부여한 예측모형(Weighted CBR)과 중요도를 부여하지 않은 일반 예측모형(Pure CBR)을 비교하였다. 사례기반추론 중 k-NN 알고리즘을 이용하였고 최적의 k를 찾기 위해 1부터 11까지 변화시켜가며 휴리스틱하게 최적의 이웃 수를 찾았다. 로짓 모형의 계수를 적용하여 입력변수의 중요도에 따라 가중치를 산출하였으며, 실증분석 결과 각 변수의 중요도에 기반하여 가중치를 적용한 예측모형이 동일한 가중치를 적용한 예측모형보다 높은 예측성과를 보여주었다.

k-NN 알고리즘과 관련한 연구를 통해 국내 교통 분야에서 k-NN 알고리즘 관련 연구는 주로 통행시간 예측이나 사고 대응시간, 돌발시간 예측하는 연구가 다수임을 알 수 있다. 알고리즘에서 거리 산정 방법은 주로 유클리디안을 사용하였으며, k값은 여러 개의 값을 적용하여 예측 오류가 가장 적은 값을 최적의 값으로 선정하였다. k-NN 알고리즘은 풍부한 데이터가 있을 경우 모수적 기법을 이용한 예측결과를 능가하며, 간단하고 모형의 갱신이 실시간으로 이루어질 수 있다는 점에서 대용량의 데이터를 활용하여 사고위험을 예측하고자하는 본 연구에 적합하다고 판단하였고 연구에 활용하고자 하였다.

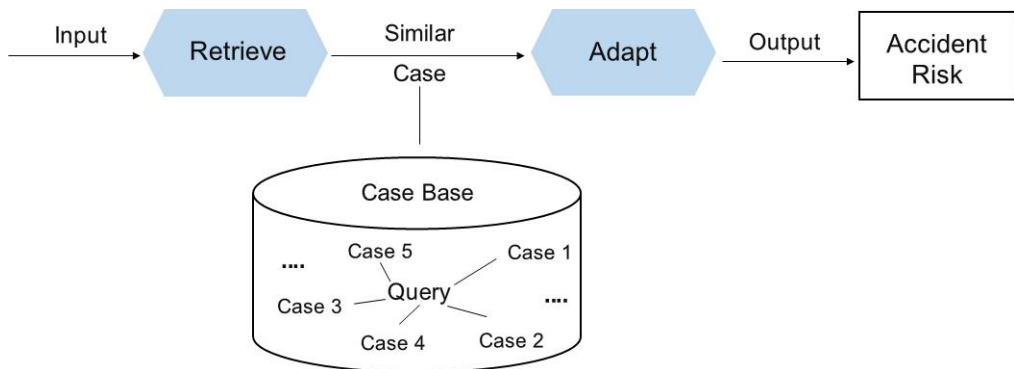
2.2 연구방향 설정

선행 문헌 고찰을 통해 국내 연구에서는 회귀모델, 로지스틱 회귀모델, 판별모델 등 모수적 기법을 사용한 사고예측모형 연구가 많이 진행되었다. 모수 모형(Parametric Model)은 구조적으로 설명변수를 정의하여 종속변수를 추정하는 일련의 수학적 과정이다. 반면, 비모수 모형(Non-parametric Model)은 구조와 변수가 데이터에 따라 결정되고 수학적 방정식에 의한 모델링이 아닌 과거이력자료에 내재된 경험적 지식을 기반으로 문제를 해결한다. 또한 모수 모형은 독립변수의 설명력이 장래에도 종속변수에 동일하게 영향을 미친다는 전제가 있으며 장래에 상황이 변할 경우 문제가 생길 수 있다(안병탁, 2015). 그리고 모수를 추정하는 과정에서 특정 분포를 따른다는 가정이 틀릴 경우 발생하는 오류의 가능성이 있으며, 종속변수와 설명변수 간의 관계를 정의하여 모델을 생성하기 때문에 최근 추세를 반영하기 위해서는 새로운 모델을 만들어야 할 필요가 있다. 반면, 비모수적 기법은 모수를 추정하지 않아 가정이 틀릴 경우 발생하는 오류의 가능성이 없으며, 독립변수와 종속변수 간의 직접적인 관계를 규명하지 않기 때문에 최근 자료의 추세를 잘 반영할 수 있다.

<표 2-1> 모수 모형과 비모수 모형의 비교

모수 모형 (Parametric Model)	비모수 모형 (Non-parametric Model)
- 모수 추정 시 특정 확률분포를 가정함	- 모수를 추정하지 않음
- 독립변수의 설명력이 장래에도 동일하게 종속변수에 영향을 미친다는 가정을 전제로 함	- 최근 자료의 추세를 잘 반영함
- 구조적으로 설명변수와 독립변수를 정의함	- 변수 간의 상관관계를 알 수 없음
- 회귀모델, 로지스틱회귀모델, 판별모델	- 의사결정나무, 랜덤포레스트, k-최근접 이웃

도로조건과 교통조건, 환경조건과 교통사고 간의 관계를 정의하여 모델을 생성하는 모수 모형은 도로와 교통조건이 양호한 구간에서의 높은 교통사고율을 설명하기 어렵다는 한계를 가지고 있다. 또한, 대용량 이력자료에 내재된 상태변화의 패턴을 이용하여 교통사고위험을 예측하고자하는 본 연구의 목적에 따라 본 연구에서는 비모수 모형이 적합하다고 판단하였다. 국외 사고예측모형에 관한 연구에서는 k-NN 알고리즘을 이용한 연구가 활발히 진행되었다. k-NN 알고리즘은 간단하며 연산시간이 짧고 대용량의 데이터에서 예측성능이 뛰어 나기 때문에 본 연구에서도 k-NN 알고리즘을 이용하여 교통사고위험 예측 기법을 개발하고자 하였다. k-NN 알고리즘을 적용한 교통사고위험 예측의 기본적인 개념은 입력조건(현재)과 유사한 과거의 사고가 발생한 사례를 찾아 그 사례를 바탕으로 위험이 있는지 예측하는 것이다. k-NN 알고리즘을 이용한 사고위험예측 개념도는 <그림 2-1>과 같다.



<그림 2-1> k-NN을 이용한 사고위험예측 개념도

본 연구에서 자주 활용되는 약어는 <표 2-2>와 같이 정리하였다.

<표 2-2> 본 연구의 약어 정리

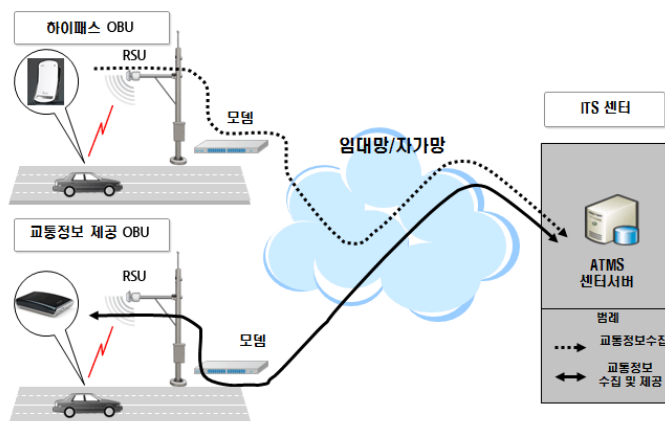
약어	내용
ITS	Intelligent Transport System - 지능형 교통시스템
ATMS	Advanced Traffic Management System - 첨단 교통 관리 시스템
DSRC	Dedicated Short Range Communications - 단거리 전용 통신
VMS	Variable Message Sign - 도로전광표지판
VDS	Vehicle Detector System - 차량검지시스템
TCS	Toll Collection System - 톨게이트 요금징수 시스템
OBU	On Board Unit - 차내단말기
RSE	Road Side Equipment - 노변기지국
NPR	Non Parametric Regression - 비모수 회귀
CBR	Case Based Reasoning - 사례기반추론
kNN	k-Nearest Neighbors - 최근접 이웃
RF	Random Forest - 랜덤 포레스트
MSE	Mean Square Error - 평균제곱 오차
ML	Machine Learning - 머신 러닝(기계 학습)
ED	Euclidean Distance - 유클리디안 거리
DT	Decision Tree - 의사결정나무
GIS	Geographic Information System - 지리 정보 체계
SMOTE	Synthetic Minority Over-sampling Technique

제 3 장 자료 구축

3.1 자료수집

3.1.1 교통소통 이력자료 - DSRC 데이터

첨단교통관리시스템(ATMS)은 도로교통정보를 자동으로 감지하여 실시간으로 도로이용자에게 도로전광판(Variable Message Signs, VMS) 등을 통해 제공하는 정보체계이다(국토교통부, 2011). 대구광역시 ATMS는 단거리 전용 통신(DSRC)을 이용하여 하이패스 단말기(On Board Unit, OBU)를 장착한 차량의 개별차량데이터를 수집하고 있다. <그림 3-1>과 같이 OBU가 도로 곳곳에 설치된 노변기지국(Road Side Equipment, RSE)을 통과하면서 단말기와의 통신을 통해 개별차량의 정보를 수집하며, 각 RSE를 통신한 시각과 RSE 간의 거리인 도로 구간 길이를 계산하여 개별 차량의 구간 데이터를 생성하고 있다. 수집되는 자료의 예시는 <표 3-1>과 같다.



<그림 3-1> DSRC 데이터 수집 과정

(자료 : 국토교통부, 2016)

<표 3-1> 교통소통 이력데이터의 수집 - 5분 단위 수집(예시)

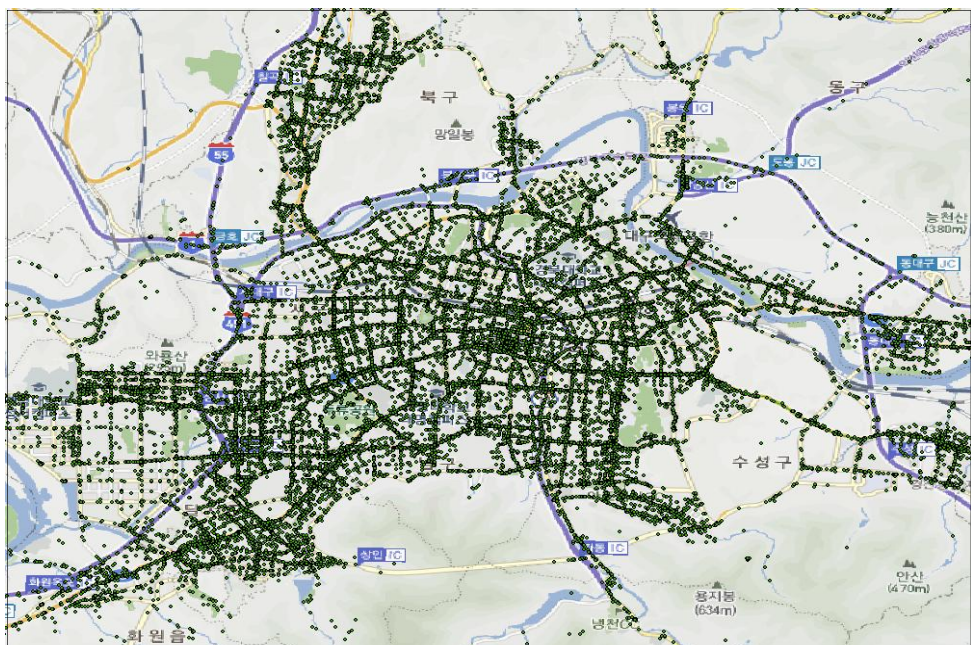
DateTime	링크 일련번호	요일	평균속도	평균 통과시간	검지대수	유효대수
2014011800	326	토요일	33.55	176.36	214	42
2014121515	51	월요일	57	118	153	45
2014061800	308	수요일	48.33	250.33	136	24
2015040106	594	수요일	55.7	84.9	72	9
2015012519	299	일요일	21.5	258.5	47	13
⋮	⋮		⋮	⋮	⋮	⋮

3.1.2 교통사고 이력자료

교통사고 이력자료는 도로교통공단에서 제공하는 2014~2015년까지 2년간 대구광역시에서 발생한 교통사고 자료를 활용하였다. 수집한 교통사고 자료는 <표 3-2>에 제시된 바와 같이 발생지, 발생일시, 요일, 1당 정보, 2당 정보, 사고유형, 사고 장소 등의 사고 속성을 포함하고 있다. 사고 이력 자료의 속성에는 사고 발생 위치의 주소 정보뿐만 아니라 각 사고가 발생한 경·위도의 좌표 정보를 포함하고 있다. 따라서 각 사고의 좌표 정보를 통해 ArcGIS를 이용한 교통소통 이력자료와의 매칭이 가능하다. <그림 3-3>은 사고 자료를 수치지도에 표출한 것이다. 수집 자료는 2014년, 2015년 각각 14,519건, 14,228건으로 총 28,747건의 사고가 발생하였다.

<표 3-2> 사고 이력자료의 속성

구분	속성
사고 이력자료	발생지, 발생일시, 요일, 1당 정보, 2당 정보 사고 유형, 사망자수, 중상자수, 경상자수 부상신고자수, 범규위반, 주야 구분, 기상상태 도로형태, x좌표, y좌표, 사고 장소

[illegible]

3.1.3 링크 정보 자료

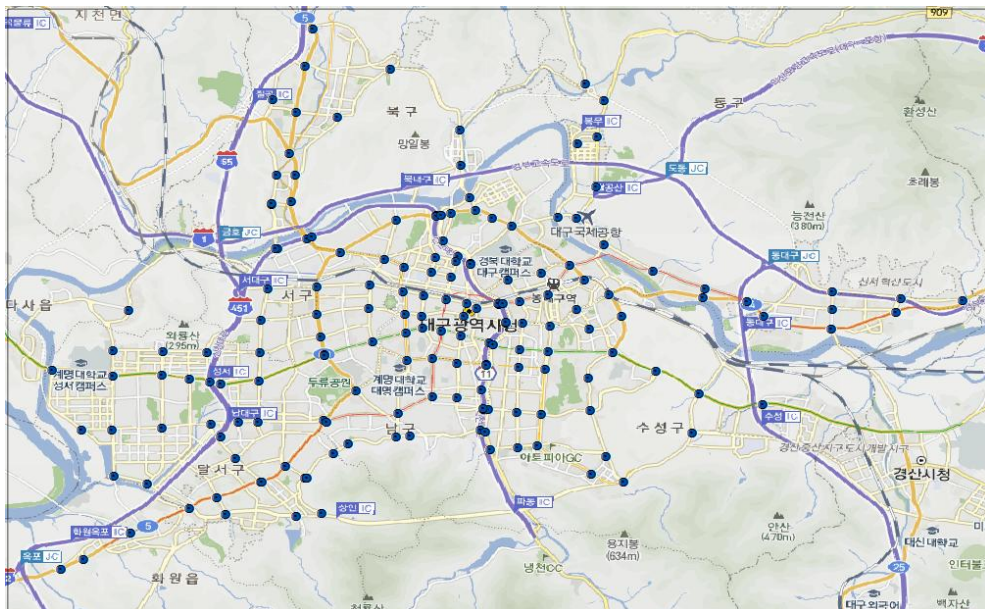
본 연구에서 활용하는 교통소통 이력자료는 RSE와 RSE 사이의 구간을 기준으로 수집된 교통 소통 정보 자료이다. RSE는 교차로의 교차점인 주요 노드²⁾에 설치되어 있으며, 노드와 다른 노드 사이를 잇는 링크는 DSRC 데이터를 수집하고 있는 단위에 해당된다. 본 연구에서는 대구광역시 ATM S의 DSRC 데이터 수집 단위인 링크를 편의상 ‘DSRC 링크’로 명명한다. <표 3-3>은 DSRC 링크에 대한 정보 자료로, 각 DSRC 링크의 노드정보와 구간정보(교차로명), 구간길이 정보를 포함하고 있다.

<표 3-3> DSRC 링크 정보 데이터

DSRC 링크 일련번호	시작 Node	끝 Node	교차로명	거리(m)
1	1	2	팔거교교차로-칠곡우체국사거리	1,361
2	2	1	칠곡우체국사거리-팔거교교차로	1,369
3	2	3	칠곡우체국사거리-칠곡네거리	1,599
4	3	2	칠곡네거리-칠곡우체국사거리	1,606
5	3	4	칠곡네거리-대전삼거리	1,390
6	4	3	대전삼거리-칠곡네거리	1,427
9	5	6	매천주유소-팔달교	1,385
10	6	5	팔달교-매천주유소	1,382
⋮	⋮	⋮	⋮	⋮
3,222	75	34	성서IC-남대구IC삼거리	1,547
3,285	69	7	침산교남단(남)-팔달교(입구)	4,104
3,286	7	69	팔달교(입구)-침산교남단(남)	4,048
4,285	69	15	침산교남단(북)-팔달교(입구)	4,104

2) 차량이 도로를 주행함에 있어 속도의 변화가 발생하는 곳을 표현한 것(교차로, 교량의 시종점, 도로의 시종점 등)

대구광역시의 DSRC 링크는 총 654개이며, DSRC 수집 장치는 <그림 3-4>와 같이 주요 간선도로의 노드 지점에 총 174개의 수집 장치가 설치되어 있다. 본 연구에서는 링크 정보 자료를 통해 교통소통 이력자료의 위치를 파악하고, 교통사고 이력자료와 매칭하여 교통소통 이력자료와 교통사고 이력자료의 두 데이터를 통합하는 매개체로서 활용하고자 하였다.



<그림 3-4> 174개의 DSRC 수집 장치 위치도

3.1.4 기상데이터

본 연구에서는 사고 위험 예측을 위한 데이터로서 기상데이터를 활용하였다. 기상데이터의 사용 근거를 마련하기 위해 기상데이터와 사고와의 관계를 나타낸 문헌들을 고찰하였으며, 그 내용은 다음과 같다.

Edwards(1999)는 기상 이력데이터와 도로 교통사고의 심각성의 관계를 분석한 결과, 날씨가 좋을 때에 비해 비가 왔을 때의 사고 심각도는 감소한 반면, 안개가 짙을 때는 지리적 변동을 나타냈다고 하였다. 이경준, 정임국,

노윤환, 윤상경, 조영석(2015)은 2013년도 교통사고 발생 자료와 지역별 상세 기상 관측 자료인 AWS(Automatic Weather System) 기상자료(시간당 강수량, 강수유무, 기온, 풍속)와 시간대, 요일 정보를 활용하여 기상과 교통사고와의 관계를 분석하였다. 로지스틱 회귀모형과 의사결정나무 모형을 통해 분석한 결과, 교통사고가 발생한 경우는 교통사고가 발생하지 않은 경우에 비해 기온이 더 높고, 비가 온다는 것을 확인함으로써 기상 요인 중 강수 유·무와 기온은 교통사고 발생에 영향을 미치는 요인이라 하였다. 기상청(2001)에서는 1년 동안 시간당 발생한 사고건수의 비교를 통해 기온이 높을수록, 습도가 낮을수록, 강수량이 높을수록 교통사고가 많이 발생했다고 하였다. 최새로나, 이기영, 오철, 김동균(2012)은 교통사고와 날씨와의 관계를 연구한 기존 문헌들을 통해 기상요소가 교통사고에 영향이 있다고 판단하였다.

이러한 문헌들을 통해 기상이 사고 발생에 영향을 끼칠 것이라는 판단 하에, 기상청의 국가 기후 데이터 센터로부터 대구광역시의 기상 이력데이터를 수집하였다. 수집 범위는 다른 수집 데이터와 동일한 기간인 2014~2015년까지 2년간의 데이터를 활용하였다. 기상 데이터는 1시간 단위로 수집되어 있지만 데이터의 지리적 구분은 세분화 되어있지 않고 지역 단위로 제공되고 있다. 이 점을 고려하여 동일한 시각에서 대구광역시 내 전 지역의 날씨는 동일하다는 가정 하에 분석을 진행하였다. 기상데이터의 구성은 <표 3-4>와 같다.

<표 3-4> 기상데이터 자료의 수집(예시)

일시	기온 (℃)	이슬점 온도 (℃)	시간 강수량 (mm)	풍향 (deg)	풍속 (m/s)	해면 기압 (hPa)	증기압 (hPa)	습도 (%)	일사 (MJ/m ²)
2014-02-26 15	9.4	5.1	0.5	140	1.5	1024.9	8.8	75	0.25
2014-02-26 16	8.6	5.8	-	290	0.8	1024.3	9.2	83	0.21
2014-02-26 17	8.5	6.2	-	0	0.3	1024	9.5	86	0.16
2014-02-26 18	8.6	6.5	3	0	0.2	1023.8	9.7	87	0.08
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

주) 자료 : 기상청 국가기후데이터센터

3.2 데이터 전처리 및 가공

본 연구는 머신 러닝을 통한 교통사고 위험 예측을 목적으로 하고 있으며, 머신 러닝은 학습 알고리즘이 훈련데이터를 기반으로 새로운 값을 예측하는 것을 의미한다. 본 연구에서는 과거 교통사고와 관련된 정보들을 훈련데이터로 적용한 학습 알고리즘을 활용하여 교통사고 위험을 예측하고자 한다. 따라서 훈련데이터를 구축하기 위해 사고 발생 당시의 상황을 파악할 수 있는 데이터들을 하나의 Dataset으로 통합하였고, 교통소통 이력자료와 교통사고 이력자료, 기상 데이터 등 3가지 데이터를 하나의 통합 테이블로 구축하기 위한 전처리 작업은 다음과 같이 진행하였다.

3.2.1 교통소통 이력자료에 대한 전처리 및 가공

교통소통 이력자료는 5분 단위로 교통 정보가 수집된 데이터를 활용하였다. 해당 데이터는 DSRC 링크별로 수집되고 있으며, DSRC 링크를 통행한 차량들의 통행시간과 속도 등을 평균하여 구축된 데이터이다. 수집한 데이터들은 각기 다른 기관에서 수집한 데이터이므로 수집 형식과 기준이 다른 이종(異種)데이터³⁾이다. 따라서 하나의 통합 테이블로 구축하기 위해서는 데이터의 공통 필드를 맞추는 필요가 있으며, 본 연구에서는 동일한 일시(연, 월, 일, 시)를 기준으로 데이터를 통합하였다. 사고 이력데이터는 사고 발생 일시가 1시간 단위까지만 기록되어 있기 때문에 교통소통 이력자료 또한 사고 이력데이터와 매칭 시키기 위해 1시간 단위로 변환하였다. 1시간 단위의 변환은 5분 단위 데이터를 평균하여 산정하였고, 이 과정에서 결측 데이터는 제외를 하고 평균으로 산출되도록 하였다. 데이터 가공은 범용 DBMS(Database Management System)인 MS SQL을 이용하였으며, 2014~2015년까지 1시간 단위로 구축한 교통소통 이력자료는 결측 데이터를 제

3) 다른 분야에서 발생하여 서로 다른 데이터들을 명명(네이버 지식백과) (Heterogeneous Data)

외하고 총 827,271개의 행으로 이루어진다.

또한, Oh, Oh, Ritchie와 Chang(2001)의 연구에 따르면 t-test를 통해 총 6개 변수 후보인 속도의 5분 평균값과 5분 속도편차, 교통량의 5분 평균과 5분 속도편차, 점유율의 5분 평균값과 5분 속도편차 중에서 사고가 발생했을 때와 사고가 발생하지 않았을 때의 차이가 가장 큰 지표는 속도의 5분 표준편차인 것으로 나타났다. 표준편차가 사고 유무에 큰 차이를 나타낸 변수라는 기존 연구 결과에 따라 본 연구에서도 5분 평균속도 데이터와 평균 통행시간 데이터를 활용하여 속도의 표준편차와 통행시간의 표준편차를 산출하였으며, 산출된 값은 마찬가지로 1시간 단위로 변환한 구축데이터에 추가하였다.

<그림 3-5>는 전처리 및 가공을 완료한 교통소통 이력자료의 일부이다.

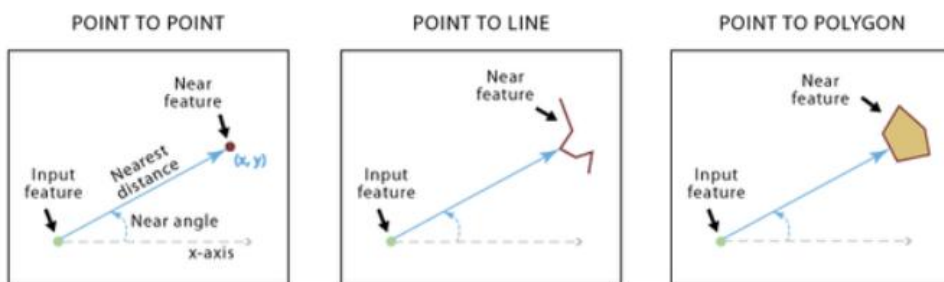
결과		메시지				
	DSRC_LINK_SN	DateTime	AVGSpeed	AVGPasstime	SDSpeed	SDPasstime
4	222	2014102715	27,36	213,91	5,63	57,74
5	2220	2014102715	49,89	361,56	2,42	19,1
6	65	2014100710	45,17	65,5	3,59	5,49
7	1066	2014101608	47,42	59,5	1,98	2,43
8	543	2014102409	33,5	59,92	2,58	4,94
9	199	2014100919	5,42	841,08	0,67	91,4
10	729	2014100300	31,2	325,4	7,16	57,24
11	82	2014103109	37,33	92,92	1,56	3,85
12	1046	2014101906	68,5	85,5	0,71	0,71
13	411	2014102714	17,83	166	1,27	12,2
14	386	2014101105	31,29	171,29	2,56	14,23

<그림 3-5> 교통소통 이력데이터의 표준편차 데이터 추가 생성

3.2.2 교통사고 이력자료에 대한 전처리

교통사고 이력자료에 필요한 전처리 작업은 두 가지로 나뉜다. 첫 번째는 각 사고의 발생이 어느 DSRC 링크에서 발생하였는지를 파악하기 위한 전처리 작업과 두 번째는 교통소통정보를 수집하는 DSRC 링크 상에서 발생한 사고 자료를 추출하는 전처리 작업이다.

먼저, 좌표 정보가 포함된 사고 데이터와 링크 데이터를 shp파일로 생성하여 지도상에 표출하고, ArcGIS의 공간분석을 통해 각 사고 건별로 해당하는 DSRC 링크를 매칭 하였다. 사용한 공간분석은 Proximity Analysis이며 <그림 3-6>과 같이 입력 Feature Class와 Near Feature class의 Feature들 중 가장 가까운 점과의 거리와 점의 좌표, 각도와 가장 가까운 Feature가 포함된 Feature Class를 속성테이블에 입력하는 분석이다. DSRC 링크는 Line, 사고 데이터는 Point로 이루어져있으므로 <그림 3-6>의 Point to Line에 해당되며, 최단 거리 계산을 통해 사고 데이터에 가장 가까운 사고 발생 링크 정보가 추가로 생성된다. 생성된 결과는 <그림 3-7>과 같다.



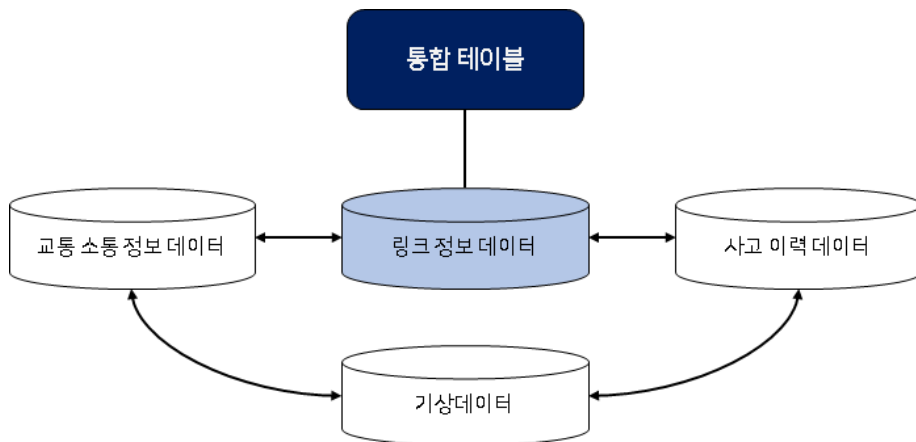
<그림 3-6 > ArcGIS의 Proximity Analysis 도식

(자료 : 한국에스리, 2012)

교통소통 이력데이터는 DSRC 링크를 통행하는 차량의 정보를 수집하기 때문에 사고 이력 데이터 또한 교통소통 정보를 수집하는 링크 상에서 발생한 데이터를 추출하였다. <그림 3-8>는 DSRC 링크 상에서 발생한 사고

3.3 통합테이블 구축

앞서 전처리 및 가공을 한 교통소통 이력자료와 교통사고 이력자료의 두 데이터는 <그림 3-9>에서 통합테이블을 도식화한 바와 같이 링크 정보 데이터를 매개체로 통합하였다. 기상데이터는 3.1.4 기상데이터에서 기술하였듯이 링크 단위의 구분 없이 동일한 데이터를 매칭하였으며, 시간 단위로만 구분하였다.



<그림 3-9> 통합테이블의 도식화

<그림 3-10>은 구축한 통합테이블을 도식화 한 것이며, 2년간의 데이터는 654개의 링크가 1시간 단위로 구성되어 전체 약 11백만 행으로 이루어졌다. <표 3-5>는 통합테이블의 구조를 설명하기 위해 간략한 예시를 보여주는 표이며, 2014~2015년까지 1시간 단위로 링크별 소통 정보와 사고 정보, 기상정보를 하나의 Dataset으로 구축하였다. 사고가 발생하지 않은 시간일 경우, 사고 데이터는 NA값을 부여받게 되고, 동일한 링크에서 동일한 시간에 사고가 2건 이상 발생할 경우에는 각각의 사고 데이터가 따로 생성하여 1시간 단위로 구성된 데이터에서 동일한 시간에 대한 데이터가 사고 발생 수만큼 중복이 되도록 하였다. 예시를 들자면, <표 3-11>의 통합테이블에서 같은 DSRC 링크의 2014년 1월 12일 21시에 사고가 두 건이 발생했

제 4 장 사고 위험 예측 모형 설계

4.1 적용 알고리즘

4.1.1 k-Nearest Neighbors Algorithm

최근접 이웃(k-NN) 알고리즘은 분류나 회귀에 사용되는 비모수방식의 머신러닝 기법이다. 학습데이터를 바탕으로 새로운 값을 입력할 때 입력 값과 가장 유사한 k개의 데이터 사용하여 새로운 데이터를 예측한다. 유사성의 척도로서 거리를 많이 사용하며 주로 유클리디안 거리를 이용하여 거리를 계산한다.

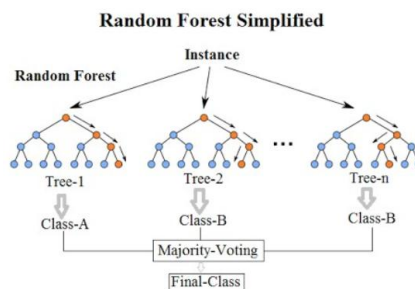
제 2 장 선행 연구를 통해 k-NN 알고리즘은 단순하고 효율적이며, 훈련 데이터의 양이 많을수록 효과적이기 때문에 대용량의 데이터를 활용하는 본 연구에서 적합한 알고리즘이라 판단하였다. k-NN 알고리즘은 사례기반 추론(CBR) 중 하나로 Instance-based Reasoning에 해당한다. 사례기반 추론은 과거에 적용되었던 사례와 그 결과들을 참조하여 새로운 사례에 대한 결과 값을 예측하는 것으로, 새로운 사례와 가장 비슷한 과거의 사례를 일부 추출하여 추출한 사례의 특정지식을 통해 문제를 해결하는 방식이다(김은미, 홍태호, 2015). 또한, 사례기반 추론은 새로운 사례가 데이터베이스에 추가되더라도 특별한 학습과정을 거치지 않고 모형의 갱신이 즉각적으로 이루어질 수 있다는 장점이 있다(Shin & Han, 1999; 김은미, 홍태호, 2015에서 재인용).

본 연구에서는 이력 데이터들을 활용하여 충돌 상태의 정보와 정상적인 상태의 정보를 각각 매칭한 학습데이터를 k-NN 알고리즘이 학습하고, 새로운 교통정보와 날씨 정보를 입력하였을 때 학습한 내용들을 바탕으로 사고 위험 예측 값이 산출될 수 있도록 하였다.

4.1.2 Random Forest Model

제 2 장 선행 연구의 김은미와 홍태호(2015)의 연구를 통해 사례기반추론 기법(CBR)에서 목표변수와 관련성이 적은 속성들이 관련성이 높은 속성들과 같은 중요도로 사용된다면 예측성가에 부정적인 영향을 줄 수 있기 때문에 관련성이 적은 속성에는 낮은 가중치를 적용하고 관련성이 많은 속성에는 높은 가중치를 적용하여 모형의 예측성가를 향상시킬 수 있다는 것을 알 수 있었다. Lin 등(2015)과 Sun과 Sun(2016)은 사고예측모형 개발 시 랜덤 포레스트를 이용하여 변수의 중요도를 평가하였다. 본 연구에서도 마찬가지로 사례기반추론인 k-NN 알고리즘의 예측 성능을 높이기 위해 랜덤 포레스트를 이용하여 각 변수마다 중요도를 평가하고 변수별 가중을 부여하고자 한다.

랜덤 포레스트(RF model)는 앙상블 학습 방법의 일종으로 <그림 4-1>과 같이 훈련 과정에서 구성한 다수의 결정 트리로부터 분류 또는 평균 예측치를 출력하는 머신러닝 기법이다. 주어진 데이터에 대해 복원 샘플링을 하여 다수의 샘플 데이터를 생성하고, 각 샘플 데이터를 모델링 한 후 결합하여 최종의 예측 모형을 산출한다. 랜덤 포레스트의 특징은 예측력이 높으며 변수의 중요정보를 제공하고, 다수의 모형 결합을 통해 과대적합(Overfitting)되는 것을 방지함으로써 의사결정나무(Decision Tree, DT)의 단점을 보완한 기법이라 할 수 있다.



<그림 4-1> Random Forest 도식화
(자료 : Linkedin, 2016)

4.2 변수선정

4.2.1 개요

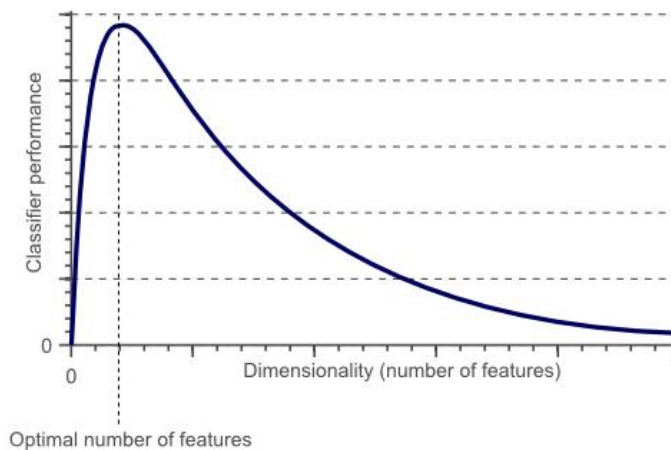
3.3 통합데이터블 구축에서 구축한 Dataset을 이용하여 사고 위험을 예측하기 위한 변수후보는 <표 4-1>과 같다. 기하구조 특성의 경우, 각 도로의 차로수와 차로폭, 종단구배 등 도로 기하구조 정보를 구득하기 어렵기 때문에 이를 대체할 수 있는 DSRC 링크 일련번호를 변수로 사용하였다. 같은 링크 내 도로 기하구조는 동일하며, 사고 위험 예측 시 링크별로 사고 위험을 예측하므로 각 DSRC 링크 일련번호는 해당 구간의 도로 기하구조에 관한 정보를 포함하는 것으로 판단하고 이와 같은 가정 하에 연구를 진행하였다. 교통 특성을 대표하는 변수 중 교통량 또한 데이터 구득이 어려워 차량 검지대수를 교통량의 대리지표로서 활용하였다. 차량의 검지대수는 링크를 통행하는 차량 중 RSE에 검지된 차량의 수를 의미하며, 검지된 차량의 교통정보를 통해 DSRC 교통정보를 생성하게 된다. 총 19개의 변수후보 중 변수 간 상관성이 높은 변수를 제거하고, 제거된 변수들 중에서 사고와의 영향을 각각 분석하여 위험 예측에 사용할 최종 변수를 선정하고 변수의 중요도를 결정하였다.

<표 4-1> 19개의 변수후보

구 분		변 수
Class	-	사고유무
Feature	기하구조 특성	링크 일련번호
	교통 특성	차량 검지대수, 평균 통행시간, 평균 속도, 평균 통행시간 표준편차, 평균 속도 표준편차
	기상 특성	기온, 이슬점온도, 시간 강수량, 풍향, 풍속, 증기압, 습도, 전운량, 일사, 일조
	시간 특성	요일

4.2.2 차원의 축소

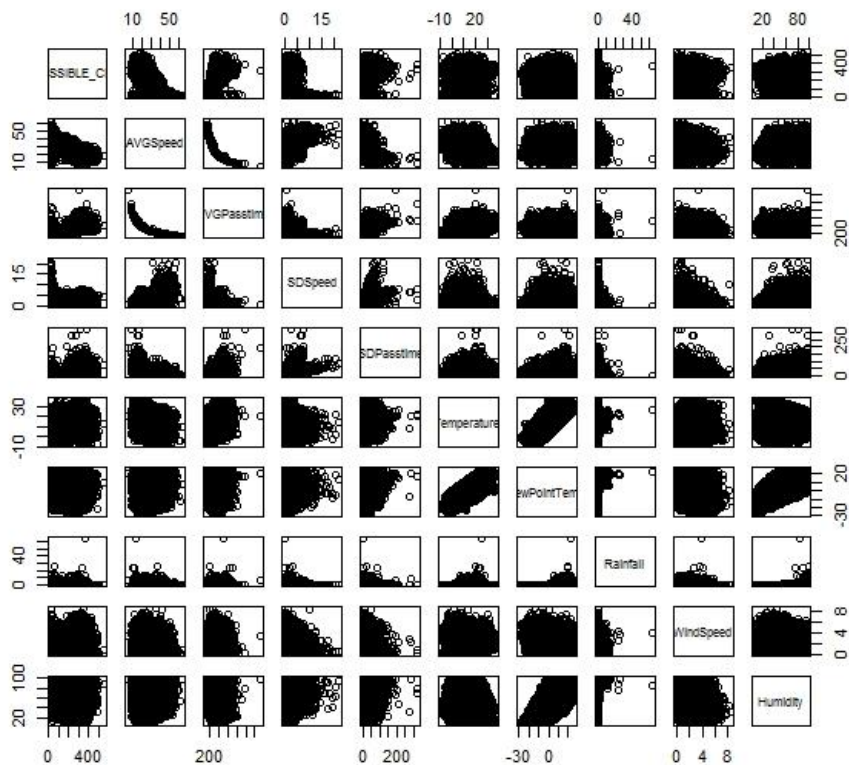
상관성이 높은 변수가 중복될 경우에는 불필요한 차원이 늘어나기 때문에 차원의 저주(Curse of Dimensionality)가 발생할 수 있다. 차원의 저주는 차원의 증가로 인해 부피가 커져 부피 안에 있는 데이터의 밀도가 낮아짐으로써 신뢰도가 낮아지고 학습시간이 오래 걸리며 정확도가 크게 감소하는 것을 의미한다. 여기서 차원은 변수의 수를 나타내고, 차원이 늘어날수록 필요한 데이터의 양은 기하급수적으로 늘어나게 된다. <그림 4-2>는 차원이 증가할 때 적정 개수까지는 성능이 증가하지만 그 이후부터는 차원이 증가할수록 성능이 나빠지는 것을 나타낸 그래프이다. 변수가 증가함에 따라 필요한 데이터양은 커지는 반면 데이터의 양은 고정되어 있기 때문에 모델의 성능은 떨어진다고 볼 수 있다. 이러한 차원의 저주 문제를 방지하기 위해서는 꼭 필요한 최소한의 변수를 사용하는 것이 중요하다. 본 연구에서는 최소한의 변수를 선택하기 위해 변수 간 상관성을 파악하였고, 상관성이 높은 경우에는 두 변수 중 중요도가 더 높은 변수를 선택하여 사용하였다.



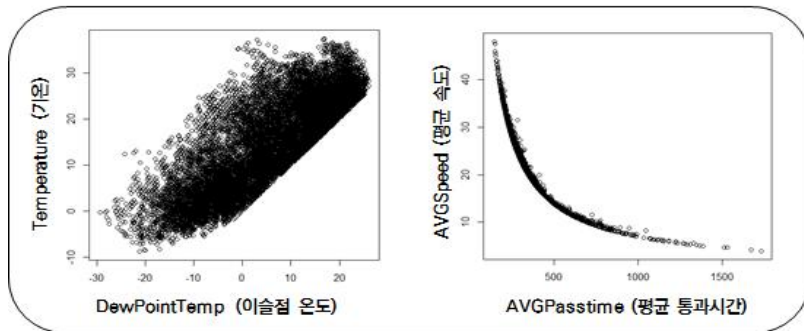
<그림 4-2> 차원의 저주 (Curse of Dimensionality)

(자료 : Data Science Central, 2014)

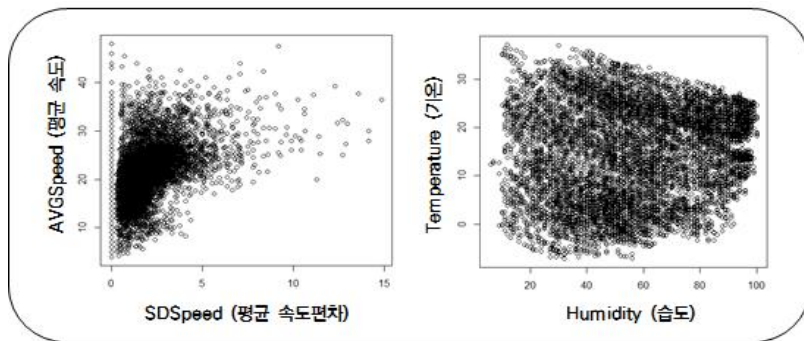
산점도 행렬(Scatter Plot Matrix)은 다변량 데이터에서 변수 쌍 간의 산점도를 그린 그래프이며, 산점도 행렬을 통해 각 변수 간의 상관성을 파악하였다. 명목형 변수를 제외한 변수의 산점도 행렬은 <그림 4-3>와 같다. <그림 4-4>와 <그림 4-5>는 각각 변수 간의 상관성이 보이고 있는 그래프와 변수 간 뚜렷한 관계가 나타나지 않는 그래프이다. 변수 간의 상관성이 나타난 변수들은 변수를 제거하기 전, 전체 변수에 대해 변수의 중요도 평가를 실시하여 중요도가 더 높은 변수는 남겨두고 나머지 변수는 제거하는 방식을 취하였다. 이 과정에서 이슬점 온도와 평균 통행속도는 분석 변수에서 제외되었으며, 최종 선정 변수는 <표 4-2>와 같다.



<그림 4-3> 변수 간 Scatter Plot Matrix



<그림 4-4> 변수 간 상관성(1)



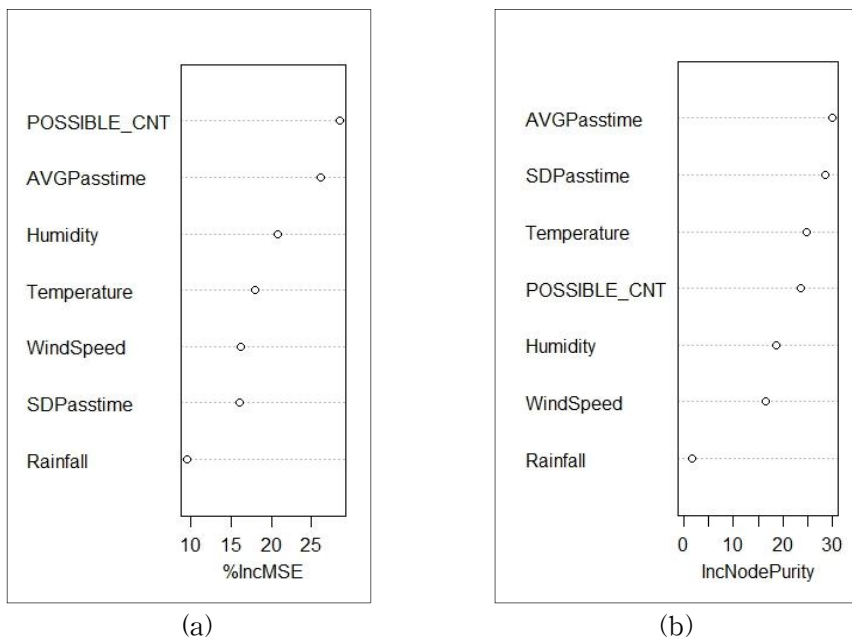
<그림 4-5> 변수 간 상관성(2)

<표 4-2> 최종 선정 변수

NO.	변수명	변수설명	데이터 타입
1	DSRC_LINK_SN	링크 일련번호	Nominal variables
2	DAY	요일	Nominal variables
3	POSSIBLE_CNT	검지대수	Continuous variables
4	AVG Passtime	평균통행시간	Continuous variables
5	SD Passtime	통행시간 표준편차	Continuous variables
6	Temperature	기온	Interval variables
7	Rainfall	강수량	Interval variables
8	Wind Speed	풍속	Interval variables
9	Humidity	습도	Interval variables

4.2.3 변수의 중요도 평가

본 연구에서는 사고 위험 예측 시, 각 변수가 사고 위험에 영향을 주는 정도에 따라 변수별로 가중치를 주고자 하였다. 일반적으로 변수의 가중치를 구할 때에는 카이제곱 검정(Chi-squared test), 피어슨 상관계수(Pearson's correlation), 스피어만 상관계수(Spearman's correlation), 랜덤포레스트(Random Forest) 등을 이용하는 방식이 있다. 각 방식마다 적합한 데이터의 형태가 정해져 있다. 카이제곱 검정은 이산형 Feature와 이산형 Class, 피어슨 상관계수와 스피어만 상관계수는 연속형 Feature와 연속형 Class가 가능하며, 랜덤포레스트는 이산형 Class와 이산형과 연속형 Feature 모두 가능하다. 본 연구에서 사용하는 데이터는 이산형 Class와 연속형 Feature 이므로 랜덤포레스트의 방식이 적합하였다.



<그림 4-6> Variable Importance Plot : (a) 각 변수의 MSE(Mean Square Error; (b) 각 변수의 Node Purity

따라서, 변수별로 가중치를 산정하기 위해 랜덤포레스트를 이용하여 중요도 평가를 실시하였다. 랜덤포레스트는 의사결정나무의 단점을 개선하기 위한 알고리즘으로 다수의 의사결정 나무를 결합하여 하나의 모델을 생성하며, 모델링에 사용할 변수를 선정하거나 변수의 중요도를 평가할 때 널리 사용되고 있다. 변수의 중요도는 변수가 정확도(Accuracy)와 노드 불순도(Node Impurity) 개선에 얼마만큼 기여하는지를 측정함으로써 산출되며, <그림 4-6>과 같이 각 변수의 MSE(Mean Square Error)와 Node Purity를 통해 결과 값이 산출된다.

<표 4-3> RF Model을 이용한 각 변수의 중요도

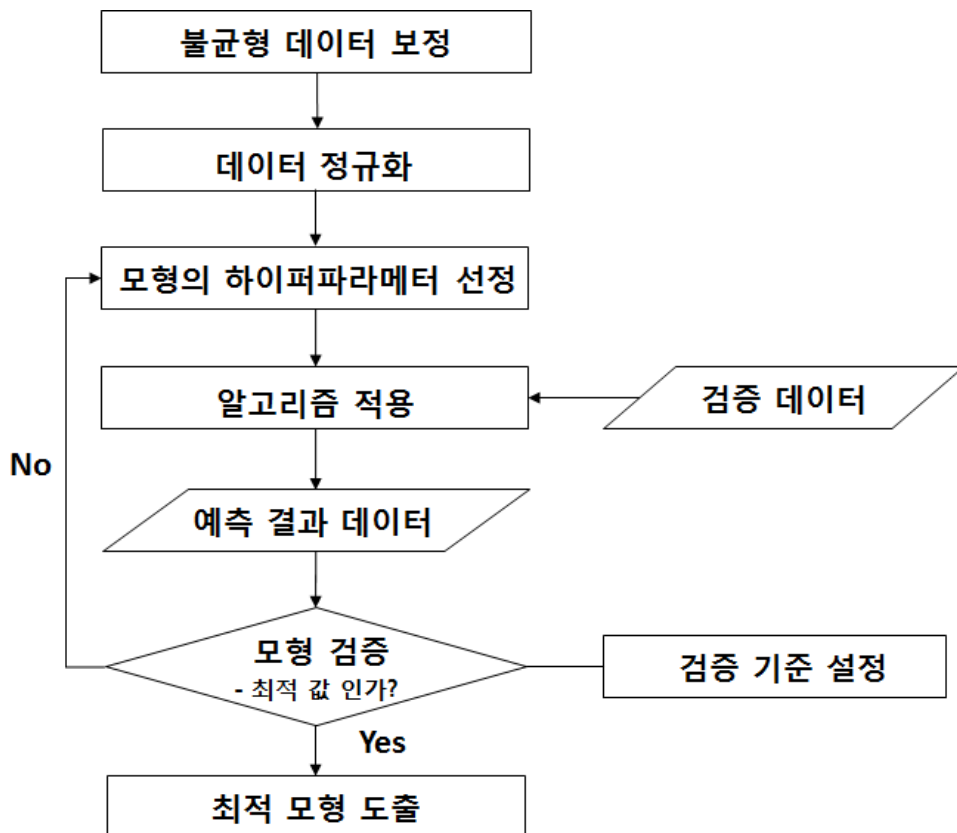
NO.	변수명	변수설명	변수 중요도
1	DSRC_LINK_SN	링크 일련번호	-
2	DAY	요일	-
3	POSSIBLE_CNT	검지대수	36.26089
4	AVGPasstime	평균통행시간	36.42249
5	SDPasstime	통행시간 표준편차	24.22104
6	Temperature	기온	22.54096
7	Rainfall	강수량	13.12835
8	WindSpeed	풍속	26.1088
9	Humidity	습도	28.19421

각 변수의 중요도는 <표 4-3>에서와 같이 AVGPasstime(평균통행시간), POSSIBLE_CNT(검지대수), Humidity(습도), WindSpeed(풍속), SDPasstime(통행시간 표준편차), Temperature(기온), Rainfall(강수량) 순으로 중요도가 나타났으며, 이후 각 변수의 중요도를 이용하여 위험 예측 알고리즘에서 변수의 가중치로 적용하고자 한다.

4.3 사고 위험 예측 모형 개발

4.3.1 개요

본 연구에서는 불균형 데이터 보정, k-NN 알고리즘의 하이퍼파라미터 (Hyper Parameter) 선정 등 여러 과정들을 거쳐 보정함으로써 최적의 모형을 도출하고자 하였으며, 전체적인 알고리즘 개발 과정은 <그림 4-7>과 같다. 알고리즘은 오픈소스 통계분석 프로그램인 R을 이용하여 구현하였다.



<그림 4-7> 알고리즘 개발 과정

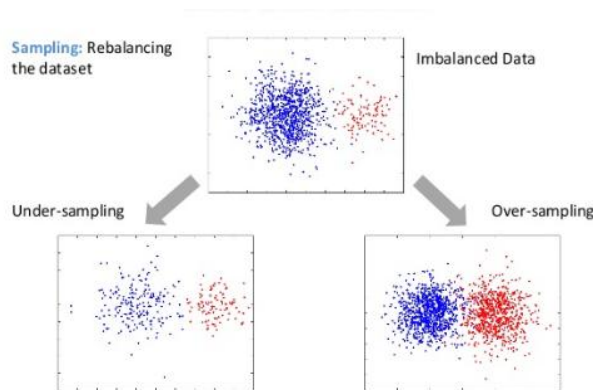
4.3.2 불균형 데이터 처리(Imbalanced data processing)

각 데이터를 클래스들이 비교적 균등한 개수의 레코드들을 포함하고 있을 때, 이 데이터의 집합을 균형 데이터 집합이라 하고, 어떤 특정 클래스가 다른 클래스들보다 현저히 많은 레코드들을 포함하고 있을 때 이를 불균형 데이터 집합 혹은 비대칭(Skewed) 데이터 집합이라고 한다(이은정, 2010). 불균형 데이터는 이진(Binary)분류로 된 데이터에서 하나의 범주에 속하는 데이터의 수가 다른 범주에 속하는 데이터의 수와 현저히 차이가 나타나는 데이터이며, 이는 머신 러닝 알고리즘의 성능을 저하시키는 요인으로 작용한다(강필성, 이형주, 조성준, 2004; 김은미, 홍태호, 2015에서 재인용).

본 연구에서 사용한 데이터는 2년간(2014~2015년)의 자료를 1시간 단위로 구축한 데이터로, 사고가 발생했을 때의 데이터가 사고가 나지 않았을 때의 데이터보다 현저하게 적기 때문에 불균형 데이터에 해당한다. 클래스가 불균형한 데이터를 머신러닝의 훈련데이터로서 그대로 사용할 경우, 데이터의 수가 월등히 많은 비사고 데이터에 대해서는 충분히 학습을 할 수 있지만 데이터의 수가 적은 사고데이터에 대한 학습은 충분히 하지 못할 가능성이 높아 예측 결과가 비율이 높은 비사고 데이터 쪽으로 치우쳐질 수 있다.

이러한 불균형 데이터를 해결하는 방법은 데이터 수준의 접근 방법과 알고리즘 수준의 접근방법으로 구분할 수 있다. 데이터 수준의 접근 방법은 <그림 4-8>과 같이 소수의 클래스를 중복 샘플링하는 오버샘플링(Over Sampling)과 다수의 클래스를 적게 샘플링하는 언더샘플링(Under Sampling) 등을 통해 훈련데이터를 조절하는 방식이다. 알고리즘 수준의 접근방법은 원래의 데이터를 그대로 유지하면서 오류율을 계산할 때 비용의 개념을 도입하여 오분류한 것에 페널티를 부과하는 방법이다. 오버샘플링은 이상치가 선택되었을 경우 계속적인 확산의 우려가 존재하며 데이터를 복제하는 것이기 때문에 훈련 데이터에 대한 정확도는 높을 수 있지만

검증데이터에 대한 정확도는 낮을 수 있다. 언더샘플링은 데이터의 잡음(Noise)을 제거하여 예측성과를 향상시키는 장점이 있으나 데이터에 대한 정보손실에 대한 우려가 있다(Liu, An, & Huang, 2006: 김은미, 홍태호, 2015에서 재인용). 오분류 비용을 통한 불균형의 문제 개선 방식은 오분류의 비용을 알고 있다는 가정 하에 사용하는 방식이며, 본 연구에서는 사고 데이터를 오분류 했을 때의 비용을 산정하기가 어렵기 때문에 해당 방식은 제외 하였다.



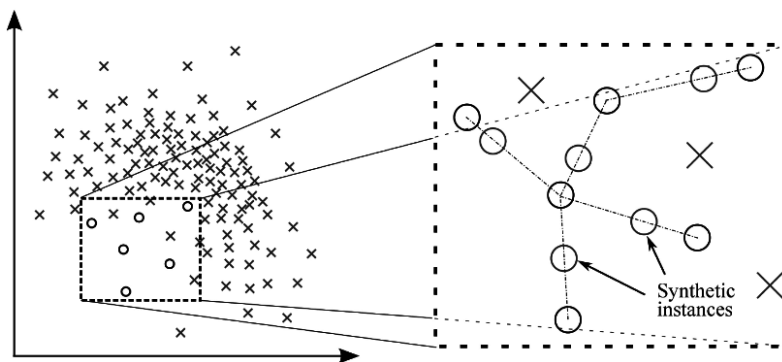
<그림 4-8> 불균형 데이터 해결 방법(Under sampling, Over sampling)
(자료 : Data Science Central, 2017)

여러 방법들의 장·단점을 검토하는 과정을 거쳐 본 연구에서는 SMOTE(Synthetic Minority Over-Sampling Technique) 기법을 이용하여 데이터 불균형 문제를 해결하고자 하였다. SMOTE는 기존 샘플들을 적절하게 조합하여 새로운 샘플을 만드는 기법으로 오버샘플링과 언더샘플링을 합성한 방법이다. SMOTE의 오버샘플링은 김한용과 이우주(2017)의 연구에서 제시한 <식 4.1>에 따라 데이터가 생성되며, 기존 오버샘플링과는 달리 적은 분류의 데이터 k개의 최근접 이웃을 고려하여 약간씩 이동시킨 점들을 추가하는 방식이다.

$$x_{new} = x_i + (\hat{x}_i - x_i) * \delta \quad \text{<식 4.1>}$$

여기서, δ : 0과 1사이의 값에 균일하게 분포하는 랜덤 변수

그렇기 때문에 기존 오버샘플링의 과적합(Overfitting) 문제를 어느 정도 개선할 수 있는 것으로 알려져 있어 오버샘플링과 언더샘플링을 보완하는 방법이라 할 수 있다. SMOTE의 오버샘플링 관련 그림은 <그림 4-9>와 같다.



<그림 4-9> SMOTE 도식
(자료 : Data Science Central, 2017)

4.3.3 데이터 정규화(Data Normalization)

데이터의 변수마다 범위와 단위가 다르기 때문에 같은 기준으로 맞출 수 있도록 재조정하는 과정이 필요하다. 특히, k-NN 알고리즘은 거리계산을 기반으로 k개의 데이터를 추출하고, 추출된 k개의 값은 결과 도출에 영향을 미치게 된다. 그렇기 때문에 각 변수의 단위나 범위가 다를 경우에는 단위 값이 매우 큰 하나의 속성이 거리계산에 미치는 영향은 클 수밖에 없다. 따라서 알고리즘을 적용하기 전, 각 데이터를 범위를 맞추는 정규화과정이 필요하다.

정규화 하는 방법은 모든 변수의 값을 0에서 1사이로 조정하는 최소-최대 정규화(Min-Max Normalization)방법과 평균과 표준편차를 이용하여 변수를 정규화 하는 정규분포의 표준화 방법(Z-Score Normalization)이 있다. Min-Max Normalization의 경우, <식 4.2>과 같이 데이터의 최대값과 최소값을 이용하고, Z-Score Normalization은 <식 4.3>와 같이 데이터의 평균

과 분산을 이용하여 정규화 값을 산출한다. Min-Max Normalization은 최대값과 최소값을 이용하기 때문에 Z-Score Normalization에 비해 데이터의 이상치(Outlier)가 정규화한 값에 영향을 많이 줄 수 있다. 주로 모수적 기법을 이용한 사고예측모형은 사고발생과 변수와의 관계를 계수 값으로 나타내기 때문에 사고와의 관계를 왜곡시킬 수 있는 이상치를 최대한 배제한다.

하지만 본 연구에서 사용하는 k-NN 알고리즘은 비모수 모형으로 사고발생과 각 변수의 관계를 명확히 규정하여 계수로 나타내지 않는다. 그렇기 때문에 이상치에 해당하는 데이터를 배제할 필요가 없으며, 과거의 모든 데이터들이 예측모형의 훈련데이터로서 적용된다. 따라서 정규화 과정에서도 이상치의 구분 없이 모든 데이터들의 값이 반영될 수 있도록 Min-Max Normalization을 이용하여 데이터 정규화를 진행하였다.

$$X_{new} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad <식\ 4.2>$$

여기서, X : 입력 값
 X_{new} : 입력 값의 정규화한 값

$$X_{new} = \frac{X - \mu}{\sigma} \quad <식\ 4.3>$$

여기서, X : 입력 값
 X_{new} : 입력 값의 정규화한 값
 μ : 전체 입력 값의 평균
 σ : 전체 입력 값의 표준 편차

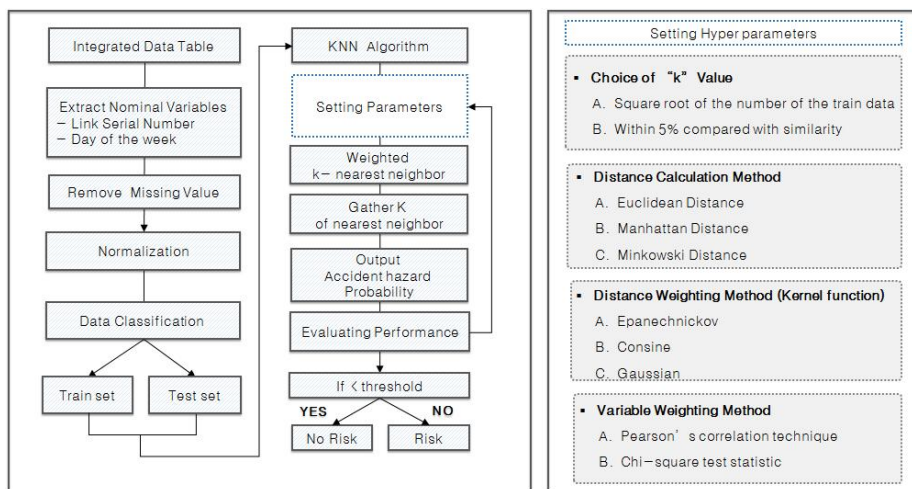
4.3.4 분석데이터

분석데이터의 구성은 훈련데이터(Training Data)와 검증데이터(Test Data)로 나누었다. 머신러닝의 지도학습은 Training Data를 통해 학습한 속성을 기반으로 예측하고, Test Data를 이용하여 개발 모형이 학습한 내용을 토대로 예측한 결과가 어느 정도의 좋은 성과를 갖는지 평가한다. 일반적으로 훈련데이터와 검증데이터의 설정은 전체 Dataset의 비율로 분류하는 방식과 기간으로 분류하는 방식이 있다. 비율로 분류하는 방식은 전체 Dataset의 80% 또는 70%를 훈련데이터, 20% 또는 30%를 검증데이터로 설정하는 방식이고, 기간으로 분류하는 방식은 특정 시점을 기준으로 전·후를 나누어 각각 훈련데이터와 검증데이터로 분류하는 방식이다. 그러나 본 연구에서는 링크 단위로 나누어 사고 위험 예측을 한다는 특성에 따라 일반적인 방식과 달리 분석데이터의 구성을 설정하였다. 링크별로 데이터를 추출하여 사고 위험을 예측하기 때문에 추출한 데이터를 다시 비율 또는 기간 등으로 나누는 일반적인 방식을 택할 경우, 학습을 할 수 있는 훈련데이터의 수가 줄어들게 된다. 따라서 훈련데이터는 추출한 데이터 전체를 훈련데이터로 설정하였다. 검증데이터 또한 사고데이터와 비사고 데이터의 수가 불균등하게 구성되어 있을 경우 한 쪽으로 치우쳐진 예측 성능이 전체 예측 성능으로 평가되는 결과를 초래할 수 있다. 따라서 검증 데이터는 균등한 예측 성능을 파악하기 위해 각 링크에서 발생한 사고건수와 동일한 비사고 데이터를 추출하여 사고 데이터와 비사고 데이터를 1:1로 구성하였다.

4.3.5 k-NN의 하이퍼파라미터(Hyper Parameter) 선정

<그림 4-10>은 개발 k-NN 알고리즘과 알고리즘 개발 과정에서 필요한 하이퍼파라미터(초매개변수)의 선정과정을 나타낸 것이다. 하이퍼파라미터는 파라미터와 달리 데이터로부터 추정할 수 없는 값이며, 알고리즘을 적용

하기 위해 분석자가 설정해야 하는 매개변수를 뜻한다. k-NN 알고리즘의 하이퍼파라미터는 k값 설정, 거리 산정 방식 선정, 거리 가중치 산정 방식 선정, 변수 가중치 산정 방식 선정이 있다. 이 중 k값 설정과 가중치 산정 방식 선정은 4.3.1의 <그림 4-7>과 같이 경험적인 방법을 통해 모형 검증 값을 기준으로 최적의 하이퍼파라미터를 선정하도록 하였다.



<그림 4-10> 개발 k-NN 알고리즘

가. k값 설정

k-NN 알고리즘은 k개의 유사한 데이터를 통해 결과 값을 산출하므로 k값 설정은 매우 중요하다. k값이 너무 작으면 이상치에 민감하며, 평균값과 차이가 큰 데이터를 선택하게 된다. 반대로 k값이 너무 크면 둔감하게 반응하고 변별력이 떨어진다. 최적의 k값을 찾는 가장 좋은 방법은 for문을 이용하여 분석데이터 수만큼 하나씩 분석하여 결과 값이 좋은 k를 찾는 것이다. 이 방식은 데이터의 수가 적은 경우에는 가능한 방법이지만 본 연구에서 사용하는 훈련데이터의 수는 기본 10,000개 정도로 크기 때문에 현실적으로 적용이 불가능하였다.

본 연구에서는 링크마다 훈련데이터의 수가 각기 다르기 때문에 고정적

인 k값을 설정하기보다 훈련데이터의 비율에 맞춘 k값을 설정하는 것이 더 적합할 것이라 판단하였다. k-NN 알고리즘을 사용한 기존 연구들에서 적용한 방법과 종합하여 최종적으로 설정한 k값의 기준은 <표 4-4>와 같다. 3가지 기준 중에서 경험적인 방법을 통해 가장 최적의 k값을 설정하였다.

<표 4-4> k값 설정 기준

구 분	내 용
1	훈련데이터의 제공근
2	훈련데이터의 5%
3	훈련데이터의 1%

나. 거리 산정 방법

k개의 유사한 데이터를 추출하기 위해 유사성의 척도로서 거리를 사용한다. 거리를 산정하는 방법에는 민코우스키 거리(Minkowski distance), 맨하탄 거리(Manhattan distance), 유클리디안 거리(Euclidean distance), 마할라노비스 거리(Mahalanobis distance), 해밍 거리(Hamming distance) 등 여러 가지 산정식이 있다. 대표적인 산정식은 <식 4.4>, <식 4.5>, <식 4.6>와 같다. <식 4.4>는 민코우스키 거리 산정식이며, 맨하탄 거리와 유클리디안 거리를 일반화시킨 식이다. 산정식의 파라미터에 따라 맨하탄 거리와 유클리디안 거리 식으로 사용할 수 있으며, m=1일 경우 <식 4.5>의 맨하탄 거리가 된다. 맨하탄 거리는 격자에서 절대 거리를 계산한 값으로 주로 격자 도로의 이동거리를 측정할 때 사용된다. m=2일 경우 <식 4.6>의 유클리디안 거리 산정식이 되며, 두 점 사이의 거리를 계산할 때 일반적으로 사용하는 방법이다. 본 연구에서는 연속형 데이터의 두 벡터 $a = [a_1, \dots, a_j, \dots, a_J]^T$ 와 $b = [b_1, \dots, b_j, \dots, b_J]^T$ 간의 거리를 산정하는 것이므로 맨하탄 거리보다 유클리디안 거리를 이용하여 유사성을 측정하는 것이 더 적합하다고 판단된다. 알고리즘에는 기본 민코우스키 거리를 적용하였으며, 파라미터는 m=2를 사용하여 유클리디안 거리 산정식이 적용되도록 하였다.

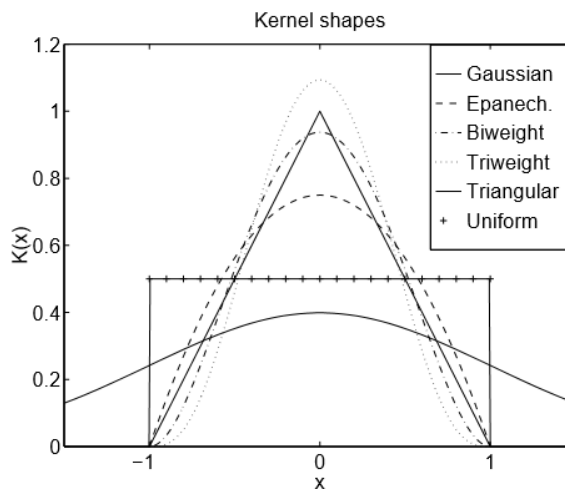
$$L_m(A, B) = \sqrt[m]{\sum_{i=1}^n (|a_i - b_i|)^m} = (\sum_{i=1}^n (|a_i - b_i|)^m)^{1/m} \quad \text{<식 4.4>}$$

$$L_1(A, B) = dist(A, B) = \sum_{i=1}^n |a_i - b_i| \quad \text{<식 4.5>}$$

$$L_2(A, B) = dist(A, B) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2} \quad \text{<식 4.6>}$$

다. 거리 가중치 산정 방법

k개의 데이터가 결과 값에 동일하게 영향을 주는 것이 아닌 가까운 거리일수록 중요하게 적용될 수 있도록 거리에 따른 가중치를 부여하였다. 가중치를 부여하는 방식은 Kernel Function을 이용하였다. 커널함수는 일반적으로 <표 4-5>와 같은 종류가 있다. <그림 4-11>은 각 커널 함수를 분포로 나타낸 것이다. 함수의 종류에 따라 가중치 부여방식이 조금씩 다르며, <그림 4-11>에서와 같이 Uniform 함수를 제외하고 대부분의 커널 함수는 거리가 가까울수록 가중치는 1에 가까워지고 거리가 멀수록 가중치가 줄어드는 형태이다.

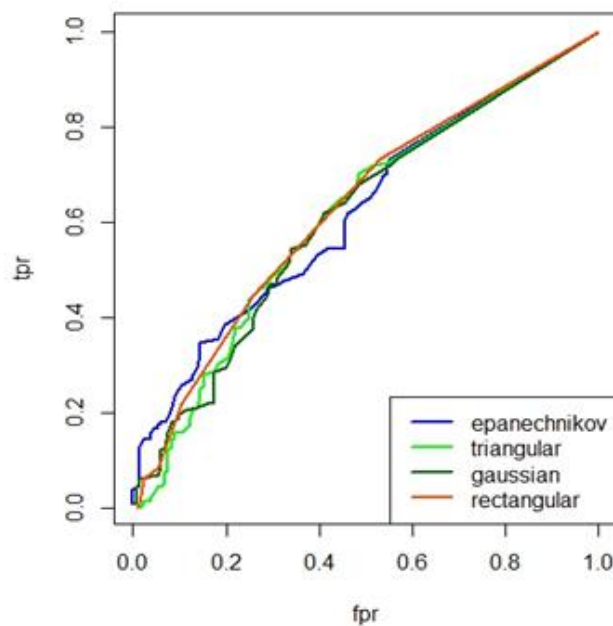


<그림 4-11> Kernel Function의 종류별 분포
(자료 : Goutte & Larsen, 2000)

<표 4-5> 거리 가중치 산정을 위한 Kernel Function

종 류	산정식
Epanechnikov	$K(u) = \frac{3}{4}(1-u^2)$ <p>Support: $u \leq 1$</p>
Cosine	$K(u) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}u\right)$ <p>Support: $u \leq 1$</p>
Quartic	$K(u) = \frac{15}{16}(1-u^2)^2$ <p>Support: $u \leq 1$</p>
Triweight	$K(u) = \frac{35}{32}(1-u^2)^3$ <p>Support: $u \leq 1$</p>
Triangular	$K(u) = (1 - u)$ <p>Support: $u \leq 1$</p>
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$
Uniform ("rectangular window")	$K(u) = \frac{1}{2}$ <p>Support: $u \leq 1$</p>

본 연구에서는 k 값의 설정과 마찬가지로 결과 값의 비교를 통해 최적의 커널함수를 선정하였다. 결과 값은 제 5장 모형 검증에서 사용할 ROC Curve의 아래 면적인 AUC(Area Under the Curve) 값을 기준으로 하였다. 여러 가지 함수들 중 대표적인 Epanechnikov와 Triangular, Gaussian, Rectangular 함수를 대상으로 분석하였다. 각 Kernel Function의 ROC Curve(Receiver Operating Characteristic Curve)를 비교한 그래프는 <그림 4-12>와 같으며, AUC 값은 Epanechnikov 함수는 0.6507978, Gaussian 함수는 0.634154, Triangular 함수 0.6022575, Rectangular 함수는 0.6022575로나와 Epanechnikov 함수가 미미한 차이로 가장 좋은 결과를 보였다. 여기서, Rectangular 함수는 가중치를 부여하지 않은 표준상태의 함수이다. 따라서 본 결과를 통해 가중치를 부여하지 않았을 때보다 유사성에 따른 가중치를 부여하였을 때의 결과가 더 좋다는 것을 확인할 수 있었다.



<그림 4-12> Kernel Function별 ROC Curve 비교

라. 변수 가중치 산정 방법

4.2.3 변수의 중요도 평가에서 나온 결과 값을 적용하여 변수의 중요도에 따라 가중치를 부여하도록 한다. 각 변수의 가중치는 <식 4.7>에서와 같이 전체 변수의 중요도 합에서 각 변수의 중요도의 비율로 정의하였다.

$$Weight = \frac{|x_i|}{\sum_{i=1}^n |x_i|} * 100 \quad \text{<식 4.7>}$$

여기서, x_i : 각 변수의 중요도
 $Weight$: x_i 의 가중치

변수의 가중치는 거리를 계산하는 과정에서 <식 4.8>과 같이 적용된다. 개발 알고리즘은 유클리디안 거리를 이용하여 거리를 산정하므로 변수 간의 거리차이를 구할 때 해당 산정식에서 각 변수마다 가중치를 곱할 수 있도록 하였다.

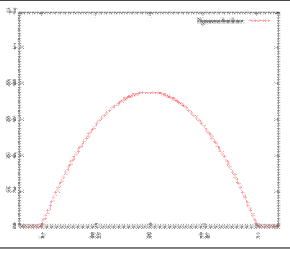
$$D_{weight} = \sqrt{w_1^2(x_1 - y_1)^2 + w_2^2(x_2 - y_2)^2 + \dots + w_n^2(x_n - y_n)^2} \quad \text{<식 4.8>}$$

여기서, D_{weight} : 각 변수의 가중치가 반영된 거리
 w_n : 각 변수의 가중치
 x_n : 첫 번째 데이터의 입력변수
 y_n : 두 번째 데이터의 입력변수
 n : 변수의 고유번호

마. 최종 선정 하이퍼파라미터

최종 선정된 하이퍼파라미터는 경험적인 방법에 의해 선정하였으며, <표 4-6>과 같다. 유클리디안 거리 산정법으로 데이터 간의 거리를 계산하고, 거리 계산 시 변수들 마다 가중치를 달리하여 중요한 변수에는 가중치를 곱하여 거리를 계산할 수 있도록 하였다. 변수의 가중치는 랜덤 포레스트의 중요도 평가를 통해 얻게 되는 값을 이용하여 산정하였다. 그 중 유사도가 높은 즉, 거리 계산 시 거리가 짧은 데이터는 Kernel Function 중 Epanechnikov의 함수를 통해 가중하였으며, 가까운 데이터는 결과 값에 좀 더 영향을 많이 주고, 멀리 있는 데이터는 결과 값에 덜 영향을 미치도록 하였다. 개발 알고리즘의 결과 값은 훈련데이터 수 5%에 해당하는 k개의 데이터를 기반으로 산출된다.

<표 4-6> 최종 선정 하이퍼파라미터

구 분	설 명	
k값	$neighbors\ of\ x_{5\%}$	
거리 산정 방법	$Euclidean = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$	
거리 가중치 산정 방법	$Epanechnikov$	
	$K(u) = \frac{3}{4}(1 - u^2)$ $Range : u \leq 1$	
변수 가중치 산정 방법	RF model을 통해 산출한 각 변수의 중요도 사용	
	$Weight = \frac{ x_i }{\sum_{i=1}^n x_i } * 100 \quad (1)$	
	$D_{weight} = \sqrt{w_1(x_1 - y_1)^2 + w_2(x_2 - y_2)^2 + \dots + w_n(x_n - y_n)^2} \quad (2)$	

제 5 장 모형 검증

5.1 검증 방법

5.1.1 Confusion Matrix

Confusion Matrix(혼동행렬)는 두 개의 클래스가 범주형 속성일 때 구할 수 있으며, 머신러닝의 모델이 훈련을 통한 예측 성능을 측정하기 위해 예측된 값과 실제 값을 비교하기 위한 표이다. <표 5-1>의 속성에서 TRUE와 FALSE는 예측의 정확여부를 의미하고 예측이 맞는 경우에는 TRUE, 예측이 틀린 경우에는 FALSE라 한다. Positive와 Negative는 예측한 값을 의미하며 예측하는 대상이 'Yes'라고 예측할 경우에는 Positive, 'No'라고 예측할 경우에는 Negative를 사용한다. 예측의 정확여부와 예측한 값의 조합에 따라 True Positive, False Positive, False Negative, True Negative의 속성으로 나타낼 수 있다. 실제 사고가 난 곳에 '사고 위험이 있다'고 예측한 경우를 True Positive(TP), '사고 위험이 있다'고 예측하였지만 실제로 사고가 나지 않은 경우를 False Positive (FP), '사고 위험이 없다'고 예측하였지만 실제로 사고가 난 경우를 False Negative (FN), 실제 사고가 나지 않은 곳에 '사고 위험이 없다'고 예측한 경우를 True Negative (TN)라 하며, 이를 정리한 표는 <표 5-2>과 같다. 표 안에 있는 TP, FP, FN, TN의 속성 값들은 5.1.2 평가 메트릭을 계산할 때 사용한다.

<표 5-1> Confusion Matrix

		Actual	
		Y (acc)	N (non-acc)
Predicted	Y (acc)	True Positive (TP)	False Positive (FP)
	N (non-acc)	False Negative (FN)	True Negative (TN)

<표 5-2> Confusion Matrix의 속성

구분	내용
TP	모델의 예측 값 Positive, 실제 값 Positive
TN	모델의 예측 값 Negative, 실제 값 Negative
FP	모델의 예측 값 Positive, 실제 값 Negative
FN	모델의 예측 값 Negative, 실제 값 Positive

5.1.2 평가 메트릭(Evaluation Metric)

개발 알고리즘은 사고발생 확률 값으로 결과를 나타낸다. 연속된 값으로 결과가 나타나기 때문에 본 연구에서는 특정 기준 이상일 경우를 사고 위험이 있다고 판단하며, 이때의 특정 기준을 Threshold(임계값)라 한다. Threshold는 예측된 사고발생 확률 값의 사고 위험여부를 판단하기 위한 위험여부의 경계 값이라 할 수 있다. <표 5-3>과 같이 사고발생 확률 값이 Threshold 값보다 클 경우 ‘사고 위험이 있다’, Threshold 값보다 작을 경우 ‘사고 위험이 없다’로 판단할 수 있다. Threshold 값에 따라 결과 값은 ‘사고위험이 있다’와 ‘사고 위험이 없다’의 이진 값으로 나타낼 수 있기 때문에 5.1.1에 설명한 Confusion Matrix와 <표 5-4>의 TPR(True Positive Rate)과 FPR(False Positive Rate) 등 평가 메트릭을 구할 수 있다. 평가 메트릭 중 민감도(Sensitivity)는 실제 Y의 클래스에 속하는 레코드 중에서 Y로 예측한 레코드의 비율이고, 특이도(Specificity)는 실제 N의 클래스에 속하는 레코드 중에서 N으로 예측한 레코드의 비율을 의미한다.

<표 5-3> 본 연구에서 정의한 Threshold

구분	내용
Threshold < 사고발생 확률 값	사고 위험 있음
Threshold > 사고발생 확률 값	사고 위험 없음

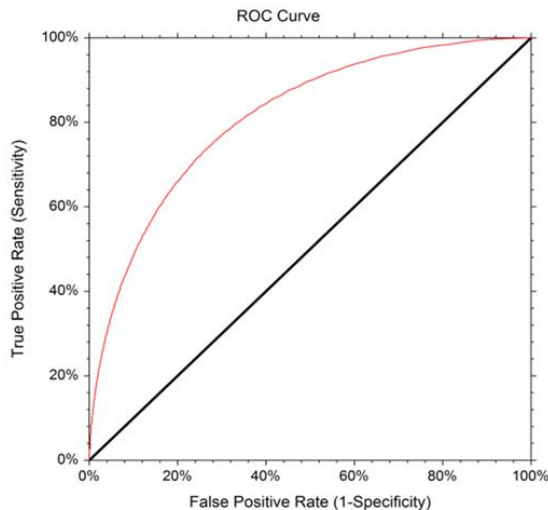
결과적으로 Threshold 값에 따라 여러 평가 메트릭을 산출할 수 있기 때문에 적정 Threshold를 결정하는 것이 필요하다.

<표 5-4> 평가 메트릭

구분	내용
민감도, 재현율 (True Positive Rate, Recall, Sensitivity)	실제 Y인 것들 중 Y로 예측한 경우의 비율
	$\frac{TP}{TP+FN}$ <식 5.1>
정밀도 (Positive Predictive Value, Precision)	Y로 예측한 것 중 실제로도 Y인 경우의 비율
	$\frac{TP}{TP+FP}$ <식 5.2>
특이도 (True Negative Rate, Specificity)	실제 N인 것들 중 N으로 예측한 경우의 비율
	$\frac{TN}{TN+FP}$ <식 5.3>
오차비율 (False Positive Rate)	실제 N인 것들 중 Y로 예측한 경우의 비율
	$\frac{FP}{FP+TN}$ <식 5.4>
정확도(Accuracy)	전체 예측에서 Y이든 N이든 무관하게 옳은 예측의 비율
	$\frac{TP+TN}{TP+FP+FN+TN}$ <식 5.5>

5.1.3 ROC Curve

ROC Curve(Receiver Operating Characteristic Curve)는 True Positive(= Sensitivity)와 True Negatives(= Specificity)를 동시에 나타낸 그래프이다. 5.1.2 평가 메트릭에서 설명한 민감도와 1-특이도는 <그림 5-1>의 x축과 y축이며, x축은 False Positive Rate(= 1 - True Negative), y축은 True Positive Rate라고도 할 수 있다. ROC Curve는 커브가 클수록 좋은 성능을 나타내며, 범주형 예측 모형의 성능을 평가할 때 평가 척도로 많이 이용된다.



<그림 5-1> ROC curve(예시)

ROC Curve의 아래면적을 AUC(Area Under the Curve)라고 하며, <식 5.6>과 같이 구한다. 이 값을 통해 모형의 성능을 수치화 하여 비교가 가능하고, 1에 가까울수록 모형의 예측 정확도가 높은 것으로 평가할 수 있다.

$$AUC = \int_a^b f(x)dx \quad \text{<식 5.6>}$$

5.1.4 종합

모형의 검증방법은 ROC Curve의 AUC, Accuracy, Recall, Precision 등의 여러 평가 지표들이 있다. AUC는 Threshold 설정과 별개로 Threshold를 설정하지 않고도 모형의 예측 성능을 파악할 수 있다. 이 외 지표들은 5.1.2 평가 에서 기술한 바와 같이 Threshold에 따라 여러 평가 메트릭의 산출이 가능하므로 적정 Threshold 결정이 필요하다. 특히, Accuracy의 경우 올바른 예측의 비율을 측정하는 것으로 대부분의 데이터가 하나의 클래스에 속하거나 두 클래스 중 하나의 성능에 대해 더 관심이 많은 경우에는 모형의 성능을 잘 나타내지 못하기 때문에 모형 성능을 평가할 때 Accuracy 뿐만 아니라 추가적으로 Recall과 Precision도 함께 고려해야 한다. Recall은 실제 사고가 난 데이터 중에서 사고 위험이 있다고 예측한 데이터의 비율이고, Precision은 사고 위험이 있다고 예측한 데이터 중 실제로 사고가 난 데이터의 비율을 의미한다. Recall과 Precision은 Recall 값이 올라가면 Precision은 내려가고, Precision이 올라가면 Recall 값이 내려가는 Trade off의 관계이다. Threshold의 변화에 따라 Recall과 Precision의 Trade off 조합이 변한다(Egan, 1975: 이은정, 2010에서 재인용). Recall과 Precision 둘 중 어떤 하나가 더 중요하다고 판단할 수 없기 때문에 두 가지 모두 적절히 조화를 이루는 것이 필요하고, 가장 적절한 조합을 찾을 수 있는 Threshold 값을 찾아야 한다.

5.2 사례 연구

5.2.1 사례 연구 대상지 개요

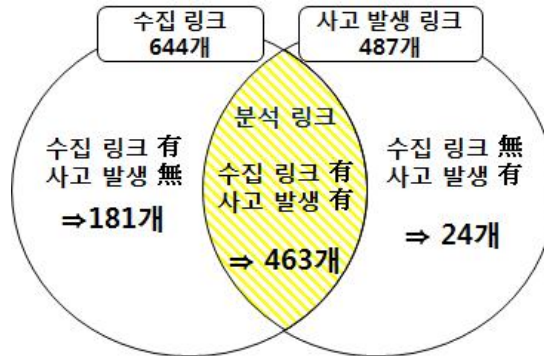
본 연구에서 개발한 교통사고위험 예측 모형은 대구광역시 간선도로에 초점을 맞춰 개발하였다. 대구광역시는 ATMS를 구축하여 운영 중이며, 단거리전용통신(DSRC)을 통해 하이패스 단말기가 설치된 차량인 OBU가 도

로의 노드에 설치된 RSE를 지날 때 정보를 수신하여 해당 OBU의 구간데이터를 생성하게 된다. 본 연구에서는 노드를 기준으로 설치된 RSE-RSE의 구간을 DSRC 링크로 칭하였으며, DSRC 링크 단위로 수집된 데이터를 활용하고 예측 모형에 대한 Dataset을 구축하였다. 따라서 개발모형의 성능을 검증하기 위한 대상은 대구광역시 DSRC 링크 상의 간선도로이며, DSRC 링크 상의 간선도로 중 검증에 적합한 대상지를 선정하기 위해 <그림 5-2>와 같이 각 링크별 사고수와 훈련데이터 수, 검증데이터 수, k값 등의 정보를 포함하여 하나의 Sheet로 구축하였다.

DSRC_LINK_SN	SECTION_NM	사고수 월일	사고수 주일	사고수 합계	NUM_DATA	TRAIN_WEEKDAY	TRAIN_WEEKEND	WEEKDAY_k개급	WEEKEND_k개급	WEEKDAY_k_0.05	WEEKEND_k_0.05	WEEKDAY_k_0.01	WEEKEND_k_0.01
391	두류내거리(북)-죽전내거리	133	133	133	17334	12376	4958	111	70	619	248	124	50
26	월촌역내거리-성당내거리	87	87	87	16314	11648	4666	108	68	582	233	116	47
25	성당내거리-월촌역내거리	76	76	76	16380	11695	4685	108	68	585	234	117	47
149	성당내거리-본리내거리	62	62	62	16692	11918	4774	109	69	596	239	119	48
122	반고개내거리-두류공원내거리(북)	67	67	67	12701	9069	3632	95	60	453	182	91	36
521	반고개내거리-두류공원내거리	57	57	57	15207	10858	4349	104	66	543	217	109	43
392	죽전내거리-두류내거리(북)	64	64	64	17338	12379	4959	111	70	619	248	124	50
181	울산내거리-성당내거리	58	58	58	17246	12314	4932	111	70	616	247	123	49
436	공평내거리-서성사거리	54	54	54	16890	12059	4831	110	70	603	242	121	48
29	상인내거리-진천역내거리	52	52	52	17154	12248	4906	111	70	612	245	122	49
22	두류공원내거리-두류내거리(남)	48	48	48	17184	12269	4915	111	70	613	246	123	49
187	백서산내거리-월촌내거리	57	57	57	16329	11516	4613	107	68	576	231	115	46
385	비산내거리-신평리내거리	57	57	57	16382	11697	4685	108	68	585	234	117	47
288	두류공원내거리-반고개내거리	55	55	55	14761	10539	4222	103	65	527	211	105	42
576	입석내거리-문고개거리	52	52	52	15216	10864	4352	104	66	543	218	109	44
148	본리내거리-성당내거리	48	48	48	16928	12087	4941	110	70	604	242	121	48
171	월곡내거리-상인내거리(고가)	48	48	48	16416	11721	4695	108	69	586	235	117	47
488	공천역상거리-황곡내거리	48	48	48	17164	12255	4909	111	70	613	245	122	49
331	반야월상거리-각산내거리	50	50	50	14337	10237	4100	101	64	512	205	103	41
367	백암로거리-문고개거리	45	45	45	9952	7106	2846	84	53	355	142	71	28
616	불미골내거리-용화대상거리	39	39	39	15127	10801	4326	104	66	540	216	108	43
117	죽전내거리-성서IC	51	51	51	10300	7354	2946	86	54	368	147	74	29

<그림 5-2> 링크 정보 Sheet

본 연구의 사고 예측은 과거 사고데이터를 이용하여 현재 상황과 유사한 과거의 데이터 k개를 이용한다. 그렇기 때문에 과거에 사고 발생한 링크를 사고 예측이 가능한 링크로 볼 수 있으며, 사고 발생 유무를 검증 대상지의 첫 번째 기준으로 설정하였다. 또한, 수집 장치에 의해 수집이 되지 않았던 링크도 존재한다. 교통소통 정보를 입력변수로 활용하기 때문에 데이터 수집 여부를 검증 대상지의 두 번째 기준으로 설정하였다. <그림 5-3>과 같이 두 가지 검증 대상지의 기준을 통해 총 654개의 링크 중 463개의 링크를 분석이 가능한 링크로 정의하였다. 본 연구에서는 대구광역시의 주요 간선도로 중 하나인 달구벌대로를 지정하여 달구벌대로의 최상위 사고다발링크를 사례 연구 대상으로 선정하였다. <표 5-5>는 달구벌대로의 최상위 DSRC 링크를 사고다발 순으로 정리한 표이다.



<그림 5-3> 검증 대상지 설정 기준

<표 5-5> 달구벌대로 DSRC 링크 - 사고다발 순

	NO.	링크번호	SECTION_NM	DIST	사고 수
달 구 벌 대 로	1	391	두류네거리(북)-죽전네거리	1997.36	133
	2	122	반고개네거리-두류네거리(북)	1415.11	67
	3	392	죽전네거리-두류네거리(북)	2017.37	64
	4	117	죽전네거리-성서IC	1410.43	51
	5	111	성서네거리-신당네거리	1382.54	35
	6	542	수성네거리-수성교남단(북)	1034.21	33
	7	423	봉산육거리-반월당네거리	584.64	33
	8	123	두류네거리(남)-반고개네거리	1412.02	31
	9	421	수성교남단(북)-봉산육거리	833.25	28
	10	540	수성교남단(북)-수성네거리	1071.6	28
	11	110	신당네거리-성서네거리	1385.77	25
	12	124	신남네거리-반고개네거리	984.39	25
	13	350	만촌네거리-담티고개 고가차도	1387.31	24
	14	116	성서IC-죽전네거리	1409.99	24
	15	342	고산역 1번 출구 앞-연호네거리	1909.9	24
	16	543	반월당네거리-계산오거리	553.73	23
	17	131	범어네거리-수성네거리	782.3	22

5.2.2 모형의 적용 결과

4.3.3 분석데이터에서 훈련데이터는 추출한 데이터 전체로 설정하고 검증 데이터는 각 DSRC 링크에서 발생한 사고건수와 동일한 비사고 데이터를 추출하여 사고 데이터와 비사고 데이터를 1:1로 구성한다고 하였다. 여기서 훈련데이터의 수는 DSRC 링크마다 다른데, 그 이유는 2014~2015년까지 1시간 단위로 구성된 통합테이블에서 수집 장비의 오류 발생으로 인한 결측과 차량이 지나다니지 않음으로 인해 데이터가 생성되지 않은 것이 원인이라 판단한다.

사례 연구 대상지의 훈련데이터 수는 총 12,383개이고 검증 데이터 수는 총 264개이다. 앞서 k의 수는 훈련데이터 수의 5%로 설정하였으므로, 개발 알고리즘을 통해 691개의 유사한 데이터를 찾고 그 데이터를 토대로 결과 값을 산출하였다.

사고 위험예측 모형의 적용 결과는 다음의 표들과 같다. 먼저, <표 5-6>은 검증 데이터 264개 각각의 데이터에 대해 유사한 691개 최근접 이웃의 Weight를 산출한 테이블이다. 행은 각각의 검증데이터를 의미하며 총 264행, 열은 최근접 이웃의 개수인 691열의 값을 산출하였다. <표 5-7>은 검증 데이터 264개 각각의 데이터에 대해 유사한 691개 최근접 이웃의 Distance를 산출한 테이블이다.

<표 5-6> Weight 산출 값

	V1	V2	V3	...	V616	V617	V618	V619
1	0.71251	0.68776	0.67648	...	0.00262	0.00162	0.0007013	0.0002962
2	0.70404	0.69681	0.66807	...	0.00249	0.00237	0.0015687	0.0004904
3	0.71485	0.65615	0.64576	...	0.00145	0.00114	0.0007493	0.0006443
⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	⋮
262	0.73773	0.73198	0.71214	...	0.00731	0.00679	0.0051294	0.0031824
263	0.68449	0.67687	0.67155	...	0.00281	0.0021	0.0002692	0.0000548
264	0.65949	0.65211	0.64622	...	0.00259	0.00239	0.0020031	0.0008431

<표 5-7> Distance 산출 값

	V1	V2	V3	...	V616	V617	V618	V619
1	0.36044	0.46441	0.50476	...	1.60931	1.61039	1.61137	1.611806
2	0.34007	0.36586	0.45405	...	1.37145	1.37157	1.372303	1.373291
3	0.63542	1.03832	1.09431	...	2.93244	2.93306	2.933816	2.934022
⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	⋮
262	0.19259	0.23339	0.33825	...	1.49815	1.49868	1.500348	1.502308
263	0.63608	0.67212	0.69613	...	2.14835	2.14936	2.151987	2.152294
264	0.99278	1.03251	1.06311	...	2.85296	2.85336	2.854099	2.856311

<표 5-6>과 <표 5-7>을 토대로 <표 5-8>와 같이 검증 데이터 264개에 대한 사고 발생 확률 값이 예측되며, 0은 사고가 발생하지 않을 확률, 1은 사고가 발생할 확률을 뜻한다.

<표 5-8> 사고 발생 확률 결과 값

No.	0	1
1	0.996395	0.003605
2	0.98563	0.01437
3	0.977716	0.022284
4	0.987957	0.012043
⋮	⋮	⋮
131	0.979125	0.020875
132	0.979157	0.020843
133	0.978961	0.021039
134	0.994224	0.005776
135	0.983702	0.016298
⋮	⋮	⋮
261	0.979036	0.020964
262	0.988675	0.011325
263	0.986243	0.013757
264	0.973985	0.026015

모형의 결과 값은 연속적인 수치이고 확률 값이기 때문에 실제 값과 비교하여 정확하게 예측 한 것인지 평가할 수 없다. 또한 발생확률이 아주 낮은 경우에는 사고의 위험이 있다고 판단하기 어렵기 때문에 본 연구에서는 적정 확률 이상의 값이 나올 경우를 사고의 위험이 있다고 판단하였다.

본 연구에서 Threshold는 모형의 결과 값인 사고발생 확률 값에서의 임계값을 의미하며, 사고발생 확률 값의 사고 위험여부를 판단하기 위한 위험여부의 경계 값으로 정의한다. 5.1.2 평가 메트릭에서 기술하였듯이 Threshold에 따라 여러 평가 메트릭의 값이 변화한다. 사고 발생 확률 값의 최대치를 100으로 나누어 100개의 각 절사(cut-off value)기준 즉, Threshold에 따른 평가 메트릭의 값을 산출하였다. 산출 결과는 <표 5-9>와 같다.

<표 5-9> Threshold에 따른 각 평가 메트릭의 변화

No.	Threshold	TPR	FPR	Precision	Accuracy	TPR-FPR
1	0	1	1	0.5	0.5	0
2	0.006688	1	1	0.5	0.5	0
3	0.013377	1	1	0.5	0.5	0
4	0.020065	1	1	0.5	0.5	0
5	0.026754	1	1	0.5	0.5	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
52	0.341112	0.916667	0.712121	0.562791	0.602273	0.204545
53	0.3478	0.886364	0.659091	0.573529	0.613636	0.227273
54	0.354489	0.871212	0.636364	0.577889	0.617424	0.234848
55	0.361177	0.848485	0.606061	0.583333	0.621212	0.242424
56	0.367866	0.833333	0.583333	0.588235	0.625	0.25
57	0.374554	0.818182	0.553033	0.596685	0.632576	0.265152
58	0.381242	0.810606	0.530303	0.60452	0.640152	0.280303
59	0.387931	0.787879	0.522727	0.601156	0.632576	0.265152
60	0.394619	0.780303	0.477273	0.620482	0.651515	0.30303
61	0.401308	0.772727	0.44697	0.63354	0.662879	0.325758
62	0.407996	0.765152	0.424242	0.643312	0.670455	0.340909
63	0.414685	0.75	0.416667	0.642857	0.666667	0.333333
64	0.421373	0.734848	0.401515	0.646667	0.666667	0.333333
⋮	⋮	⋮	⋮	⋮	⋮	⋮
96	0.635404	0.015152	0	1	0.507576	0.015152
97	0.642093	0.007576	0	1	0.503788	0.007576
98	0.648781	0.007576	0	1	0.503788	0.007576
99	0.02987	0.007576	0	1	0.503788	0.007576
100	0.030175	0.007576	0	1	0.503788	0.007576

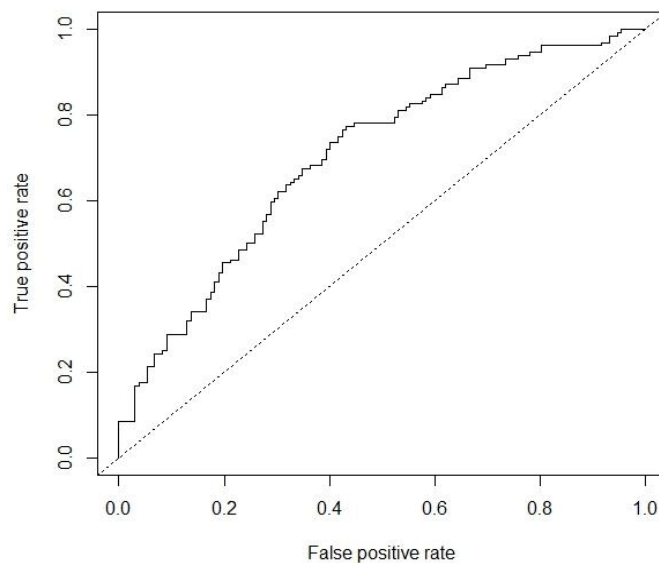
5.2.3 결과 해석

예측 결과에 대한 모형의 성능을 파악하기 위해 <표 5-9>의 사고가 발생할 확률과 실제 사고 발생여부를 비교할 수 있도록 비교 테이블을 생성하였고, <표 5-10>과 같다. 실제 값과 예측 값을 비교한 표인 <표 5-10>과 Threshold에 따른 각 평가 메트릭의 값을 나타낸 <표 5-9>를 이용하여 1) ROC Curve 분석 2) 예측 결과 값의 확률밀도분포 비교 3) Threshold에 따른 정확도 비교의 총 3가지 분석을 수행하였다.

<표 5-10> 실제 값과 예측 값 비교

No.	사고가 발생할 확률(예측 값)	실제 사고 발생 여부
1	0.003605	1
2	0.01437	1
3	0.022284	1
4	0.01204	1
5	0.014475	1
⋮	⋮	⋮
130	0.006202	1
131	0.020875	1
132	0.020843	1
133	0.021039	0
134	0.05776	0
135	0.016298	0
⋮	⋮	⋮
260	0.006903	0
261	0.020964	0
262	0.011325	0
263	0.013757	0
264	0.026015	0

첫 번째, 검증 모형의 ROC Curve는 <그림 5-4>와 같다. 커브의 아래면적인 AUC는 0.704로 나타났다. 1에 가까울수록 좋은 모형임을 나타내며 0.7 이상의 경우 좋은 모형이라 평가하고 있다. AUC에 따른 모형의 성능은 <표 5-11>과 같다. ROC Curve는 (0,1)의 점과 가장 가까운 점이 Trade off 관계인 Recall과 Precision의 조합이 가장 좋은 것을 나타낸다. 그래프에서 (0,1)과 가장 가까운 점의 TPR은 0.7 중·후반이고 <표 5-9>를 통해 TPR이 0.7 중·후반일 때의 Threshold는 0.394619 ~ 0.414685 범위이다.

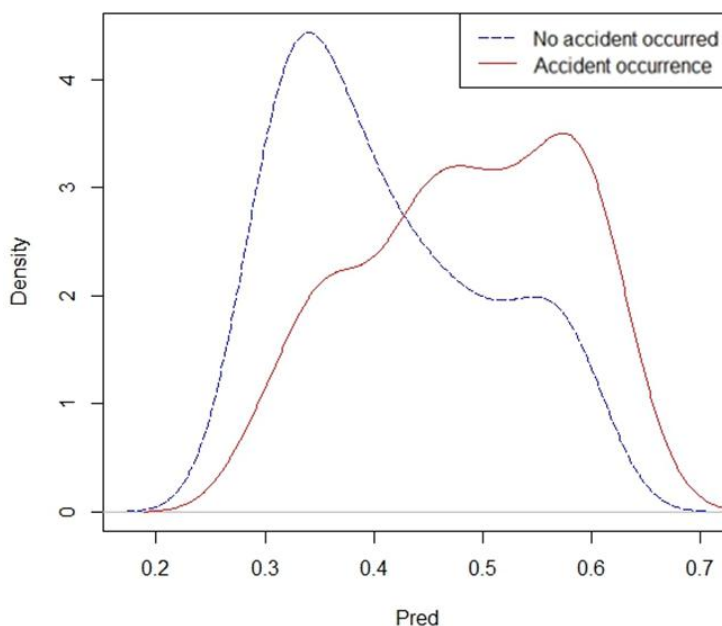


<그림 5-4> ROC curve (AUC:0.704)

<표 5-11> AUC에 따른 모형의 성능

AUC	Accuracy
0.9 ~ 1.0	Excellent
0.8 ~ 0.9	Very good
0.7 ~ 0.8	Good
0.6 ~ 0.7	Sufficient
0.5 ~ 0.6	Bad
< 0.5	Not useful

두 번째, 예측결과 값의 확률밀도분포 비교는 <표 5-10>을 이용하여 이 실제 사고가 난 데이터의 사고발생 확률 값과 사고가 나지 않은 데이터의 사고발생 확률 값의 각 확률밀도 분포를 살펴본 결과이다. <그림 5-5>는 실제 사고가 난 데이터와 실제 사고가 나지 않은 데이터를 비교한 확률밀도분포 그래프이다.



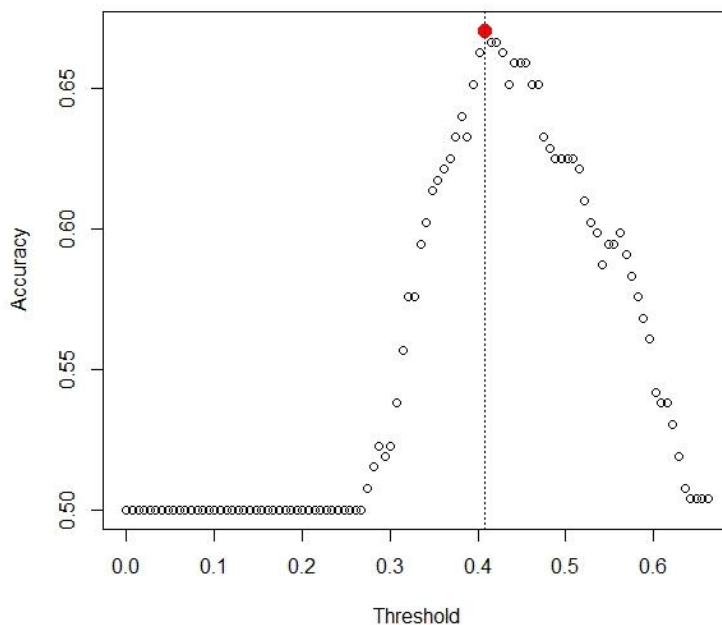
<그림 5-5> 예측결과 값의 확률밀도분포

실제 사고가 난 데이터의 확률밀도 분포는 오른쪽으로 치우쳐 사고발생 확률 값이 높게 예측되었으며, 실제 사고가 나지 않은 데이터는 왼쪽에 치우쳐져있고 사고발생 확률 값이 낮게 예측되었다는 것을 확인할 수 있었다. 이는 어느 정도 오차는 있겠지만 예측 확률 값이 클수록 사고위험이 있다는 것을 나타내며, 두 분포의 교차점을 통해 예측된 확률 값 중 실제 사고가 난 것과 나지 않은 것을 확실히 구분할 수 있을 것이라 판단하였다. 그래프를 통해 두 분포의 교차점은 확률 값이 0.4 초반으로 나타났다.

세 번째, Threshold에 따른 정확도 비교는 <그림 5-6>의 그래프와 같다.

정확도가 가장 높은 Threshold는 0.4 초반이며, <표 5-9>를 통해 정확한 값을 살펴보면 정확도가 가장 높은 0.670455에서의 Threshold는 0.407996으로 나타났다.

전술한 바와 같이 일반적으로 머신러닝의 예측 모형 성능은 Accuracy를 높이는 것을 목적으로 모형을 학습한다. 하지만 Accuracy는 하나의 클래스에 대부분의 데이터가 속한 경우이거나 하나의 클래스에 대한 성능에 더 관심이 많은 경우에는 실제 모형의 성능을 잘 나타내지 못하기 때문에 Accuracy 뿐만 아니라 다른 지표도 함께 고려해야한다. 따라서 본 연구에서는 Accuracy와 더불어 Precision과 Recall, AUC를 종합하여 최적의 Threshold값을 통해 개발 예측 모형의 성능을 평가하고자 하였다.



<그림 5-6> Threshold에 따른 정확도

앞서 세 가지 분석에서 각 Threshold의 범위는, 첫 번째 분석의 0.394619 ~ 0.414685, 두 번째 분석의 0.4 초반, 세 번째 분석의 0.407996으로 나타났다. 이 결과들을 종합하여 적정 Threshold의 값의 공통 범위 내에

Accuracy가 가장 높았을 때의 값인 0.407996이 포함된다. 따라서 전체적인 Threshold는 0.407996으로 설정하고, 이 값을 경계로 예측된 결과 값의 사고 위험 여부를 판단하였다. Threshold가 0.407996일 때 Accuracy는 0.670455, Recall은 0.765152, Precision은 0.643312로 나타났다.

개발 모형은 Precision 값을 통해 사고 위험이 있다고 예측한 것 중 64.33%가 실제로 사고가 발생한다고 예측이 가능한 성능이 있다는 것을 확인할 수 있다. 또한, Recall 값을 통해 실제 사고가 발생한 것 중 76.52%가 사고가 위험이 있다 예측하였고, 실제 사고 발생의 70%이상 예측이 가능한 모형임을 나타낸다.

제 6 장 결 론 및 향후연구

6.1 결론 및 의의

본 연구에서는 도시부 도로에서 기 수집된 대용량의 교통정보 이력 데이터를 데이터 마이닝 하여 사고 위험에 대해 예측하는 기법을 제시 하였다. 대용량의 데이터로부터 유용한 패턴을 찾는데 적합한 머신러닝의 지도학습 알고리즘을 사용하고자, 지도학습 알고리즘 중에서 문헌고찰을 통해 예측 변수 값들이 최적의 조합으로 결정지어질 때 다른 예측 모형보다 뛰어난 예측 성능을 낼 수 있는 k-NN 알고리즘을 채택하였다.

k-NN 알고리즘은 비모수모형으로 종속변수와 설명변수간의 관계를 정의하지 않고, 예측하고자 하는 관측치가 있을 때 이와 가장 가까운 거리에 있는 k개의 데이터를 결정한 후 이들의 특성을 이용해 관심 관측치를 예측하는 과정을 거치는 알고리즘이다. 단순하고 효율적이며 훈련데이터의 양이 많을수록 효과적이기 때문에 대용량의 데이터를 활용하는 본 연구에서 적합한 알고리즘이라 판단한다.

본 연구를 위해 교통 소통 정보를 알 수 있는 DSRC 데이터와 교통사고 이력데이터, 기상데이터, 링크정보데이터를 수집하였으며, 데이터는 모두 2014~2015년까지 대구광역시의 자료를 활용하였다. 수집한 데이터들은 전처리·가공 과정을 거쳐 하나의 테이블로 통합하였고, 사고위험예측의 기본 분석데이터로 사용하였다.

k-NN 알고리즘을 이용한 사고위험예측의 설계는 전체적으로 변수의 선정과 불균형데이터 처리, 데이터 정규화, 분석 데이터 설정, k-NN 알고리즘에 필요한 하이퍼파라미터 선정의 과정으로 이루어졌다. 본 연구에서 사용하는 분석데이터는 클래스가 불균등한 데이터인 점을 고려하여 이를 해결하기 위해 SMOTE 기법을 적용하였다. 또한, 거리 기반 알고리즘에서 중요한 데이터의 정규화 과정을 거쳐 분석데이터를 설정하였다. k-NN 알고

리즘에서 설정이 필요한 하이퍼파라미터는 k 값 설정, 거리산정방법, 거리가중치 산정 방법, 변수 가중치 산정 방법으로 정의하였고, 그 중 k 값 설정과 거리 가중치 산정 방법은 경험적인 방법을 통해 적합한 하이퍼파라미터를 선정하였다.

개발 알고리즘은 사례 연구를 통해 검증하였으며, 사례 연구 대상지는 대구광역시 DSRC 링크 중 최상위 사고다발링크로 선정하였다. 검증모형의 성능을 평가하는 메트릭은 ROC Curve의 AUC와 Recall, Precision, Accuracy를 활용하였다. ROC Curve에서 (0,1)과 가장 가까운 점의 TPR 값, 예측 결과 값의 확률밀도분포 교차점, 정확도가 가장 높은 Threshold 값 등 총 3가지를 통해 적정 Threshold를 설정하였다. 적정 Threshold를 이용하여 각 평가 메트릭을 산출하였으며, 검증 결과 AUC는 0.7이상으로 “Good”한 모형으로 나타났고, Recall과 Precision, Accuracy는 각각 0.77, 0.64, 0.67로 나타났다. Precision 값은 사고위험이 있다고 예측한 것 중 64.33%가 실제로 사고가 발생한다고 예측이 가능하다는 것을 의미하며, Recall 값은 실제 사고가 발생한 것 중 76.52%가 사고위험이 있다고 예측한 것을 의미한다. 최종적으로 본 연구의 사고위험 예측은 실제 사고 발생의 70%이상 예측이 가능함을 나타낸다.

본 연구의 의의를 학술적 측면과 정책적 측면, 실용적 측면의 총 세 가지 측면으로 설명하면 다음과 같다.

첫째, 국외에서는 k -NN 알고리즘을 활용한 교통사고 발생 예측 연구가 진행되고 있지만 국내에서는 교통사고 대응시간 예측과 통행시간 예측, 교통량 예측에 대한 연구로 한정되어있다. k -NN 알고리즘은 풍부한 데이터가 있을 경우 모수적 기법을 이용한 예측결과를 능가하며, 간단하고 모형의 갱신이 실시간으로 이루어질 수 있다는 점에서 대용량 데이터를 활용한 단기 사고 위험 예측의 기법으로서 유망하다. 본 연구는 국내에서 시도되어지지 않은 k -NN 알고리즘을 활용하여 교통사고 위험 예측 기법을 제시한 것에 연구의 의의가 있다.

둘째, 정책적 측면으로는 다음과 같다. ITS의 정보수집체계를 통해 수집된 방대한 양의 이력 데이터가 쌓이고 있으며, 이 데이터는 4차 산업혁명과

지능정보사회에서 아주 큰 역할을 하고 있다. 이로 인해 정부에서는 방대한 이력자료에 대한 참여기반 조성 및 사회적 현안 해결을 위한 데이터 활용 강화 등 다양한 정책을 내세우고 있다. 본 연구는 우리 사회의 문제로 꾸준히 제기되어온 교통사고 문제에 대해 교통사고 감소를 위한 하나의 대책으로 대용량의 이력 데이터를 활용한 교통사고 위험 예측 기법을 제시하였다. 이는 대용량의 이력 데이터의 활용가치를 높이는 데 큰 도움이 될 것이라 생각한다.

마지막으로 실용적 측면은 다음과 같다. 본 연구의 사고 위험 예측 기법은 실제 운영되고 있는 시스템의 탑재를 목표로 개발하였다. 연구에서 사용한 데이터는 대구광역시 ATMS에서 수집하고 있는 데이터를 활용하였으며, DSRC 데이터의 수집단위와 동일하게 예측모형을 설계하였기 때문에 대구광역시 ATMS에 본 연구의 결과를 탑재 및 운영을 할 경우 모형의 적용성이 높을 것이다. 또한, 향후 이식성 평가를 통해 적합하다면 대구광역시 이외에 구축되어 있는 ITS 관련 교통 시스템에서도 직관적인 활용이 용이할 것이라 예상한다.

6.2 향후 연구

본 연구에서 제시한 교통사고위험 예측 기법의 성능 향상과 적용성을 확대시키기 위해서는 다음의 향후 연구가 필요하다.

첫째, 제시한 교통사고위험 예측 기법은 k-NN 알고리즘을 이용하였다. k-NN 알고리즘은 대표적인 지도학습의 비모수모형으로서 대용량 이력자료에 내재된 교통사고 패턴을 찾아 사고위험을 예측하고자하는 본 연구의 목적에 부합하는 방법론이다. 그러나 k-NN의 알고리즘의 경우 지도학습이기 때문에 사고 위험을 예측하고자 하는 곳에 사고 발생 이력이 있어야 예측이 가능하다는 한계점을 가지고 있다.

하이브리드 모델(Hybrid Model)은 다른 여러 개의 모형을 결합하여 더 좋은 성능을 낼 수 있는 모델이다. 향후 k-NN 알고리즘과 조합이 좋은 예

측 알고리즘을 찾아 개발모형을 하이브리드 모델로 업그레이드 한다면 사용한 각 모형의 한계점을 서로 보완할 수 있으며, 기존에 제시한 예측 기법이 보다 향상된 성능을 지닐 수 있을 것이라 생각한다.

둘째, 본 연구에서 사용하는 데이터는 OBU의 교통정보를 수집한 DSRC 데이터이다. DSRC 데이터는 프로브 카(Probe Car)로부터 수집되는 정보이기 때문에 도로 상 모든 차량의 데이터를 대표하기엔 부족할 수 있다. 향후 교통량 자료와 택시나 화물차의 실시간 주행데이터인 DTG 자료를 수집하여 분석데이터에 추가함으로써 DSRC 데이터를 보완을 한다면 데이터의 신뢰성을 높일 수 있을 것이다.

셋째, 본 연구에서는 제시한 예측 기법을 사례 연구를 통해 검증하였으며, 검증 결과 우수한 예측 성능을 나타냈다. 그러나 제시한 예측 기법의 적용성을 확대시키기 위해서는 사례 연구 대상지뿐만 아니라 공간적 범위를 확대·적용하여 제시한 예측 기법이 공간적으로 강건한(robust) 모형임을 증명할 필요가 있다.

넷째, 본 연구에서는 세 가지 분석내용을 통해 적정 threshold 값을 설정하였다. 일반적으로 알고리즘에서 Threshold의 기본 값(default value)은 0.5로 설정되어있다. 또한, 아직까지 Threshold 설정에 대한 명확한 기준을 제시하는 객관적인 가이드라인이 없어 Threshold를 조정할 경우, 연구자의 주관적인 판단에 의해 선택을 하게 된다. 그러나 주관적인 Threshold의 설정은 잘못된 판단으로 결정될 가능성이 있고, 이에 대한 검증 방안도 없기 때문에 Threshold를 설정하는 객관적인 가이드라인이 필요하다고 생각한다. 본 연구에서 적용한 적정 Threshold 설정 기준을 심도 있는 분석 과정을 거쳐 다른 비슷한 연구 분석에서도 적용이 된다면 Threshold 설정에 대한 객관적인 가이드라인을 만드는 데 기초가 될 수 있을 것이다.

다섯째, 본 연구에서의 교통사고 위험 예측 결과는 Threshold에 따라 ‘사고 위험이 있다’와 ‘사고 위험이 없다’로 구분하였다. 보다 효율적인 교통사고 발생 예방을 위해서는 위험 정보를 단순히 위험 여부로 구분하는 것보다 더 세부적으로 분류할 필요가 있다고 생각한다. 따라서 향후에는 예측 결과를 ‘매우 위험’, ‘위험’, ‘주의’, ‘안전’ 등의 방식으로 여러 개의 위험 등

급으로 나눠 등급화 함으로써 최종적인 결과 값으로 단계별 위험도를 나타내 고자 한다.

마지막으로 본 연구의 결과가 실제 반영이 될 수 있도록 실용적인 측면에서도 추가적인 연구를 할 필요가 있다.

참 고 문 헌

- 강필성, 이형주, 조성준. (2004). 데이터 불균형 문제에서의 SVM 양상블 기법의 적용. **한국정보과학회 학술발표논문집**, 31(2Ⅱ), 706-708.
- 국토교통부. (2011). **ITS 사업시행지침 수립연구(첨단교통관리시스템(ATMS)을 중심으로)**. 세종: 국토교통부.
- 국토교통부. (2016). **제8차 국가교통안전기본계획**. 세종: 국토교통부.
- 국토교통부. (2017년 2월). **2021년까지 교통사고 사망자수 2,700명대 목표**. 2017년 10월 20일 검색, http://www.molit.go.kr/USR/NEWS/m_71/dtl.jsp?lcmepage=1&id=95078854
- 기상청. (2001년 5월). **기상이 교통사고에 미치는 영향**. 2017년 09월 15일 검색, <http://www.kma.go.kr/kma15/2001.5/%BC%DB%BB%F3%B1%D4%B3%B2%B1%C3%C1%F6%BF%AC.htm>
- 김은미, 홍태호. (2015). 불균형 데이터 환경에서 변수가중치를 적용한 사례 기반추론 기반의 고객반응 예측. **한국지능정보연구**, 21(1), 29-45.
- 김한용, 이우주. (2017). 불균형적인 이항 자료 분석을 위한 샘플링 알고리즘들: 성능비교 및 주의점. **응용통계연구**, 30(5), 681-690.
- 김형주, 박신형, 장기태. (2016). k-NN 알고리즘을 활용한 단기 교통상황 예측: 서울시 도시고속도로 사례. **대한교통학회지**, 34, 158-167.
- 김혜원, 이영인. (2015). 고속도로 교통사고 발생 시 사고 대응시간 예측모형 개발. **대한교통학회 학술대회지**, 73, 197-202.
- 도로교통공단. (2016). **2016년판 OECD 회원국 교통사고 비교**. 원주: 도로교통공단.
- 신강원, 심상우, 최기주, 김수희. (2014). KNN 알고리즘을 활용한 고속도로 통행시간 예측. **대한토목학회논문집**, 34(6), 1873-1879.
- 안병탁. (2015). **지식발견 기반 고속도로 영업소**. 석사학위논문. 인천대학교 대학원.

- 이승봉, 한동희, 이영인. (2015). 사고등급별 고속도로 교통사고 처리시간 예측모형 개발. **대한교통학회지**, 33(5), 497-507.
- 이경준, 정임국, 노윤환, 윤상경, 조영석. (2015). 도로위의 기상요인이 교통사고에 미치는 영향-부산지역을 중심으로. **한국데이터정보과학회지**, 26(3), 661-668.
- 이은정. (2010). 데이터 마이닝 앙상블 모델을 이용한 교통사고 분석. 석사학위논문. 아주대학교 대학원.
- 이재명, 김태호, 이용택, 원제무. (2008). CART 분석을 이용한 교통사고예측모형의 개발. **한국도로학회논문집**, 10(1), 31-39.
- 박준태, 이수범, 이동민. (2011). 도시부 도로구간 사고예측모형 개발. **교통연구**, 18(1), 63-73.
- 백승걸, 장현호, 강정규. (2005). 교통량과 통행길이를 고려한 고속도로 교통사고 예측 연구. **대한교통학회지**, 23(2), 95-105.
- 최새로나. (2012). 기상 및 교통조건이 고속도로 교통사고 심각도에 미치는 영향분석. 석사학위논문. 한양대학교 대학원.
- 최재익. (2016). 고속도로 통행정보를 활용한 KNN 기반 교통정체 예측 기법. 석사학위논문. 서강대학교 정보통신대학원.
- 한국에스리. (2012년 6월). ArgGIS Tool 소개. 2017년 11월 18일 검색, <http://blog.naver.com/esrikr/110140892222>
- 한상진, 김근정, 오순미. (2008). 전통적 사고예측모형의 한계 및 개선방안: Hauer 사고예측모형의 소개 및 적용. **한국도로학회논문집**, 10(1), 19-29.
- Data Science Central. (2014). *About the Curse of Dimensionality*. Retrieved November 12, 2017, from <https://www.datasciencecentral.com/profiles/blogs/about-the-curse-of-dimensionality>
- Data Science Central. (2017). *Handling imbalanced dataset in supervised learning using family of SMOTE algorithm*. Retrieved November 23, 2017, from <https://www.datasciencecentral.com/profiles/blogs/handling-imbalanced-data-sets-in-supervised-learning-using-family>

- Edwards, J. B. (1999). The relationship between road accident severity and recorded weather. *Journal of Safety Research*, 29(4), 249-262.
- Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. New York: Academic Press.
- Goutte, C. and Larsen, J. (2000). Adaptive metric kernel regression. *Journal of VLSI signal processing systems for signal, image and video technology*, 26(1-2), 155-167.
- Lin, L., Wang, Q. and Sadek, A. W. (2015). A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. *Transportation Research Part C: Emerging Technologies*, 55, 444-459.
- Linkedin. (2016). *Random Forest Algorithm, An Interactive Discussion*. Retrieved December 3, 2017, from <https://www.linkedin.com/pulse/random-forest-algorithm-interactive-discussion-niraj-kumar>
- Liu, Y., An, A., Huang, X. (2006). Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles. *In PAKDD*, 6, 107-118.
- Lv, Y., Tang, S., and Zhao, H. (2009). Real-time highway traffic accident prediction based on the k-nearest neighbor method. *International Conference on Measuring Technology and Mechatronics Automation*, 3, 547-550.
- Oh, C., Oh, J. S., Ritchie, S. and Chang, M. (2001). Real-time estimation of freeway accident likelihood. In *80th Annual Meeting of the Transportation Research Board, Washington, DC*.
- Pirdavani, A., Magis, M., De Pauw, E., Daniels, S., Bellemans, T., Brijs, T., et al. (2014). Real-time crash risk prediction models using loop detector data for dynamic safety management system applications. *In Second International Conference on Traffic and Transport Engineering (ICTTE)*.
- Sun, J. and Sun, J. (2016). Real-time crash prediction on urban

expressways: identification of key variables and a hybrid support vector machine model. *IET intelligent transport systems*, 10(5), 331-337.

Shin, K. S., and Han, I. (1999). Case-based reasoning supported by genetic algorithms for corporate bond rating. *Expert Systems with applications*, 16(2), 85-95.

A Study on Traffic Accident Risk Prediction Method based on k-NN Algorithm Using DSRC Data

Kang, Min Ji

Department of Urban Planning and Transportation Engineering
Graduate School

Keimyung University

(Supervised by Professor Park, Shin Hyoung)

(Abstract)

Despite the steady efforts of the government and related organizations to reduce traffic accidents, Korea is still the lowest among OECD member countries in the field of traffic safety. In order to reduce traffic accidents, the government has pursued policies that focus on post-response measures such as the safety improvement project of high collision concentration locations, but the accident prevention is limited to activities other than traffic safety education and campaigns. Therefore, it is necessary to establish a proactive response system. This study aims to develop a traffic accident risk prediction method which can identify the roads with accident risk in real time and enable them to respond in advance. Data mining technique is used based on massive traffic historical data of Daegu Metropolitan City and k-Nearest Neighbors

(kNN) algorithm, which is one of the machine learning techniques, is applied as a prediction algorithm. The analytical data were obtained by integrating traffic historical data, traffic accident historical data, and weather historical data. In addition, SMOTE technique is used to solve the imbalance between data, and a method of verifying the Hyper parameter selection criteria for k-NN algorithm and the prediction result using the appropriate threshold is proposed. The test site was selected as the top accident links in the main arterial road. and the evaluation metric was AUC of ROC Curve, Recall, Precision, Accuracy. As a result of the verification, AUC was found to be "Good" with 0.7 or more, Recall, Precision and Accuracy were 0.77, 0.64, and 0.67, respectively, which confirms that the model can predict more than 70%. Since the data collected from the ATMS are utilized and the algorithm suitable for real-time operation is used, the prediction method of this study is considered to be easy to use directly in the urban traffic management system. Also, it is expected that the developed accident risk prediction information will play a central role in the proactive system in the

future.

DSRC 데이터를 활용한 k-NN 알고리즘 기반 교통사고 위험예측 기법 연구

강 민 지

계명대학교 대학원
도시계획 및 교통공학과
(지도교수 박 신 형)

(초록)

우리나라는 교통사고 감소를 위한 정부와 관련기관의 꾸준한 노력에도 불구하고 교통안전분야에서는 여전히 OECD 회원국 중 최하위권에 머물고 있다. 정부에서는 교통사고를 감소시키기 위해 교통사고가 잦은 곳 개선사업 등과 같은 사후적인 대응 위주로 정책을 추진한 반면 사전적인 대응은 교통안전 교육과 캠페인 등의 활동 이외에는 미미한 수준이기 때문에 체계적인 교통사고 예방을 위한 사전 대응 시스템의 필요성이 크다고 할 수 있다. 본 연구는 실시간으로 사고위험이 있는 도로를 판별하여 사전 대응을 가능토록 하는 교통사고 위험 예측 기법을 개발하는 것을 목적으로 한다. 대구광역시 도시부 도로의 대용량 이력자료를 기반으로 데이터 마이닝 기법을 활용하였고, 예측 알고리즘으로는 머신러닝 기법 중 하나인 k-NN(k-Nearest Neighbors) 알고리즘을 적용하였다. 분석 데이터는 교통소통 이력데이터와 교통사고 이력데이터, 기상 이력데이터 등을 통합하여

활용하였다. 또한, 데이터 간의 불균형을 해소하기 위해 SMOTE 기법을 사용하고, k-NN 알고리즘에 필요한 하이퍼파라미터 선정 기준과 적정 Threshold를 이용한 예측 결과의 검증 방법을 제시하였다. 검증 대상지는 주요 간선도로의 최상위 사고다발링크로 선정하였으며, 평가 메트릭은 ROC Curve의 AUC와 Recall, Precision, Accuracy을 활용하였다. 검증 결과, AUC는 0.7이상으로 “Good”한 모형으로 나타났고, Recall과 Precision, Accuracy는 각각 0.77, 0.64, 0.67로 나타나 최종적으로 70% 이상 예측이 가능한 모형임을 확인하였다. 본 연구의 예측 기법은 현재 ATMS에서 수집하고 있는 데이터를 활용하고 실시간 운영에 적합한 알고리즘을 사용하였기 때문에 도시부 교통관리시스템에 직접적인 활용이 용이할 것이라 판단된다. 또한, 개발한 사고위험예측 정보는 향후 사전 대응적인 시스템에서 중심적인 역할을 할 수 있을 것으로 기대한다.