교통사고 데이터의 패턴 분석과 Hybrid Model을 이용한 피해자 상해 심각도 예측

(Pattern Analysis of Traffic Accident data and Prediction of Victim Injury Severity Using Hybrid Model)

주영지*, 홍택은**, 신주현***

(Yeong Ji Ju, Taek Eun Hong, Ju Hyun Shin)

요 약

우리나라의 경제 성장과 도로 환경의 변화를 통해 국내 자동차 시장이 성장하였으나, 이로 인해 교통사고율 또한 증가하였고, 인명 피해가 심각한 수준이다. 이에 따라, 정부에서는 교통사고 데이터를 개방하고 문제를 해결하기 위한 정책을 수립 및 추진 중이다. 본 논문에서는 교통사고 데이터를 이용하여 클래스의 불균형을 해소하고, Hybrid Model 구축을 통한 교통사고 예측을 위해 원본 교통사고 데이터와 Sampling을 수행한 데이터를 학습 데이터로 사용한다. 두 학습데이터에 연관규칙 학습기법인 FP-Growth 알고리즘을 이용하여 교통사고 상해 심각도와 연관된 패턴을 학습한다. 두 학습 데이터의 연관 패턴을 분석을 통해 같은 연관된 패턴을 추출하고 의사결정트리와 다항 로지스틱 회귀분석기법에 연관된 속성에 가중치를 부여하여 융합형 Hybrid Model을 구축하고 교통사고 피해자 상해 심각도를 예측하는 방법에 대해 제안한다.

■ 중심어 : 피해자 상해 심각도 예측; FP-Growth 알고리즘; 로지스틱 회귀분석; 의사결정트리; Hybrid Model

Abstract

Although Korea's economic and domestic automobile market through the change of road environment are growth, the traffic accident rate has also increased, and the casualties is at a serious level. For this reason, the government is establishing and promoting policies to open traffic accident data and solve problems. In this paper, describe the method of predicting traffic accidents by eliminating the class imbalance using the traffic accident data and constructing the Hybrid Model. Using the original traffic accident data and the sampled data as learning data which use FP-Growth algorithm it learn patterns associated with traffic accident injury severity. Accordingly, In this paper purpose a method for predicting the severity of a victim of a traffic accident by analyzing the association patterns of two learning data, we can extract the same related patterns, when a decision tree and multinomial logistic regression analysis are performed, a hybrid model is constructed by assigning weights to related attributes.

■ keywords : Prediction of Victim Injury Severity Using Hybrid Model; FP-Growth Algorithm; Logistic regression analysis; Decision Tree; Hybrid Model

Ⅰ. 서 론

우리나라의 경제 성장과 도로 환경의 변화에 따라 국내 연간 차량 보급률이 높아지면서 2015년을 기준으로 자동차 등록 대수는 약 2099만 대이며, 전년 대비 4.3%의 증가율을 보인다[1].

자동차 이용의 증가에 따라 교통사고율 또한 증가하였으며, 교통사고는 물적 피해, 인명 피해, 사회 비용을 발생시킨다. 우리나라의 교통사고로 인한 사망자 수는 2013년을 기준으로 인구 10만 명당 사망자수 10.1명, 자동차 1만 대당 사망자 수 2.2명으로 OECD 국가 중 사망률이 상위권에 속하고, 교통사고로 인한 인명 피해액은 2013년에는 약 135억 원에서 2014년에는 156억 원으로 전년 대비 1.16% 증가하여 인명 피해가 심각한 수준이다[2]. 따라서 정부에서는 교통사고로 인한 피해를 줄이기 위해 교통사고 빅 데이터를 구축하고 데이터를 개방하여 교통사고 문제를 해결하기 위한 '교통사고 사상자 줄이기 종합대책(2013~2017)' 정책을 추진 중이며, 이에 따라, 빅 데이터 연구를 활용하는 연구가 활발히 진행 중이다[3].

교통사고 데이터는 사고와 연관된 다양한 속성으로 이뤄져있으며, 교통사고로 인한 인명 피해를 줄이기 위해서는 피해자 상해 심각도와 관련된 다른 속성간의 패턴 분석 및 사고 발생 요인을 찾아 개선하여 교통사고 피해에 따른 후속조치가 필요하

*,**준회원, 조선대학교 소프트웨어융합공학과

*** 정회원, 조선대학교 제어계측로봇공학과

본 논문(저서)은 교육과학기술부의 재원으로 한국연구재단의 지원을 받아 수행된 산학협력 선도대학(LINC) 육성사업의 연구결과이며, 2016년 교육부와 한국연구재단의 지역혁신창의인력양성사업의 지원을 받아 수행된 연구입니다.(2015H1C1A1035823)

접수일자 : 2016년 12월 15일 게재확정일 : 2016년 12월 26일

수정일자 : 2016년 12월 21일

교신저자 : 신주현 e-mail : jhshinkr@chosun.ac.kr

다[4]. 따라서 본 연구에서는 교통사고 데이터를 이용하여 데이터 마이닝 기법을 Hybrid Model로 구축하고 교통사고 피해자 상해 심각도 예측을 위한 방법을 제안한다. 교통사고 데이터에서 클래스 불균형의 문제를 해결하기 위해 기존 데이터와 Sampling 기법을 적용한 교통사고 데이터를 학습 데이터로 사용한다. 각 학습데이터에 FP-Growth 알고리즘을 이용하여 교통사고 피해자 상해 심각도와 연관된 패턴을 추출하고 의사결정트리와 다항 로지스틱회귀 알고리즘이 결합된 Hybrid 기법을 활용하여 교통사고 피해자 상해 심각도를 예측한다.

본 논문의 구성은 다음과 같다. 2장에서는 교통사고 데이터를 이용한 기존 연구에 관해서 설명한다. 3장에서는 교통사고 데이터를 이용하여 교통사고 피해자 상해 심각도를 예측하는 모델에 대해서 기술하며, 4장에서는 본 논문에서 제안하는 예측 모델에 대한 성능을 평가한다. 마지막으로 5장에서 결론 및 향후 연구를 기술하며 마무리한다.

Ⅱ. 관련 연구

국내 교통사고로 인한 피해가 증가됨에 따라 교통사고 관련 공공데이터가 개방되었다. 교통사고 데이터가 개방되기 이전의 연구는 미국 NASS의 GES Data Set를 주로 활용하였으며, 데이터를 기반으로 예측을 위한 연구에서는 데이터 마이닝 기법을 융합한 Hybrid Model을 이용하여 예측성능을 높일 수 있게 하였다[5].

Chong et al.은 미국 GES 데이터에서 1995년부터 2000년 사이에 발생된 정면충돌사고에 해당하는 데이터를 이용하여 의사결정트리, 인공신경망, Support Vector Machines를 이용한 Hybrid Model을 구축하여 운전자의 상해 심각도를 예측하는데 활용하였다[6].

이재식의 연구에서는 Chong et al.논문의 모델 구축 방식에서 새로운 사고 데이터가 발생하였을 때, 모든 모델에서 상해 심각도를 판단하지 못하면 LOOP에 빠지는 단점을 지적하면서 인공신경망, 의사결정트리, 로지스틱 회귀분석 모델을 모두 적용하고 적중률이 너무 낮은 경우에 사례기반 추론 기법을 이용하여 '치명적 상해'의 경우 95.9%의 높은 적중률을 보였다[7].

홍성은의 연구에서는 국내 공공데이터를 활용하여 교통사고 상해 심각도 예측을 위한 연구를 하였다. 기존의 연구에서는 미국의 데이터를 활용하여 국내 현실을 반영하지 못한다는 단점을 지적하였으며, 공공데이터를 이용하여 분류 예측 모델 생성 시, 분류의 정확도나 순수도를 낮추는 속성을 제거하기 위해 전처리 기능을 수행하는 역할로써, CART 의사결정트리 알고리즘을 이용하여 변수를 선택한 후 랜덤 포레스트 알고리즘에 선택 변수를 사용하는 Hybrid Model을 구축하였다[8].

이은정은 미국 GES 데이터를 이용하여 여러 개의 의사 결정트리를 결합하여 앙상블 모델을 구축하고 적중률의 불균형을 해소하였으며, 후속 연구로 인공신경망을 이용한 모델을 적용하고 Under-Sampling방식을 적용하여 데이터 수가 적은 상해 심각도의 적중률도 향상시켰다[9,10].

기존의 연구에서는 클래스의 불균형을 해소하기 위해 Up-Sampling, Over-Sampling 혹은 Under-Sampling 방식을 수행하여 학습 데이터의 셋으로 활용하였으나, 이는 데이터의 변형으로 인해 분류 알고리즘을 수행할 때, 다수의 클래스 혹은 소수의 클래스의 적중률은 높일 수 있으나, 다른 클래스의 적중률이 낮아지는 문제를 발생시켜, 분류 알고리즘의 성능을 저하시킬 수 있다[11]. 따라서 본 논문에서는 기존의 원본 교통사고 데이터와 Sampling을 수행한 데이터를 학습 데이터로 모두 사용하여, 클래스의 불균형으로 인한 오·분류를 낮출 수 있도록 한다. 의사결정트리를 이용하여 교통사고 피해자 상해 심각도의 주요 속성을 선정하고, FP-Growth 알고리즘을 이용하여 연관된 패턴 추출 및 로지스틱 회귀분석을 결합한 Hybrid Model을 이용하여 교통사고 피해자 상해 심각도를 예측하는 방법을 제안한다.

Ⅲ. 교통사고 상해 심각도 예측 모델

1. 시스템 구성도

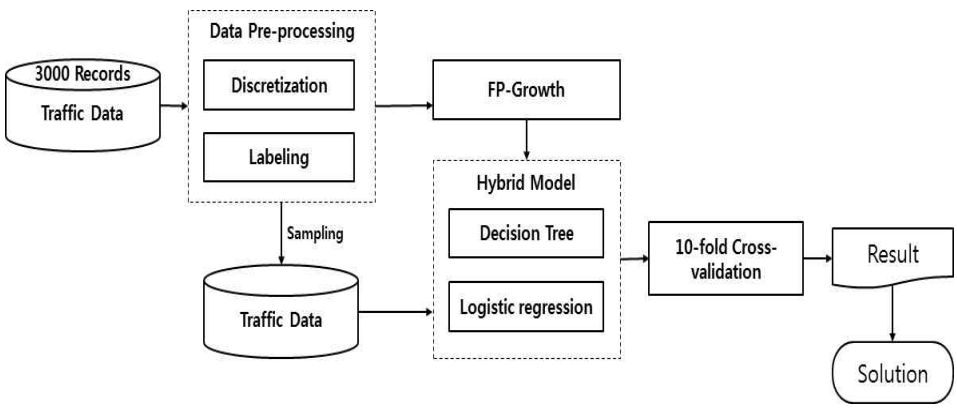

그림 1. 교통사고 피해자 상해 심각도 예측을 위한 시스템 구성도

본 절에서는 교통사고 데이터를 추출하고 연관규칙 학습기법인 FP-Growth알고리즘과 의사결정트리, 로지스틱 회귀분석 알고리즘을 이용하여 교통사고 피해자 상해 심각도 예측을 위한 방법을 제안한다.

그림 1은 교통사고 데이터를 이용한 피해자 상해 심각도 예측을 위한 시스템 구성도이다. 기존의 교통사고 데이터에서 피해자 상해 심각도와 관련 없는 속성은 제거하였고, 피해자 상해 심각도 속성에서 '사망'의 경우 데이터가 많고 '부상 신고'의 경우 데이터가 적어 분류 모델의 성능을 저하시키는 데이터 불균형 현상이 발생한다. 따라서 기존 데이터와 Sampling기법을 적용한 데이터를 학습데이터로 이용하고, 전처리 과정으로 이산화 및 Labeling 과정을 수행한 후 FP-Growth알고리즘을 수행하여 교통사고 상해 심각도와 연관된 패턴을 추출하고 의사결정트리와 다항 로지스틱 회귀를 10-fold Cross-validation을 수행하여 분류 및 예측을 수행한다.

2. 교통사고 데이터

가. 교통사고 데이터 재정의

본 연구에서는 TAAS(Traffic Accident Analysis System)에서 2012년부터 2015년까지 서울특별시, 경기도 일대와 광역시에서 발생한 약 3000여개의 교통사고 데이터를 수집하여 활용한다. 표 1은 원본 교통사고 데이터의 속성을 나타낸다.

표 1. 교통사고 데이터 속성

유형	속성
사고	사고 번호, 발생 일시, 발생 요일, 발생시군구, 사고유형, 법규위반, 노면상태, 기상상태, 도로형태, 사고 장소
차량	가해 운전자 차종, 피해 운전자 차종
사람	인명 피해(사망, 중상, 경상, 부상 신고)자수, 가해 및 피해 운전자 성별, 연령, 상해정도

교통사고 데이터는 크게 3가지 유형으로 분류되고, 속성은 총 23개로 구성되어 있지만 데이터의 일반화를 위해 피해자 상해 심각도와 관련이 없는 '사고 번호', '부상자 수', '피해운전자의 성별 및 연령'과 지역적인 특성인 '발생 시군구', '사고 장소'의 속성은 제외하였다.

FP-Growth 알고리즘 및 분류 알고리즘을 수행하기 위해 이산화 및 Labeling 과정을 통해 교통사고 데이터를 재정의 하였으며, 표 2는 재정의 한 교통사고 데이터이다.

표 2. 교통사고 데이터 재정의

속성	값
발생요일	평일/주말
발생시간	01~06시, 07~12시, 13~18시, 19~24시
사고유형	car-car, car-human, caronly (기타, 통행, 횡단 중, 정면충돌, 추돌, 추돌-주정차 중, 추돌-진행 중, 측면 직각 충돌)
법규위반	과속, 보행자보호 의무위반, 불법유턴, 신호위반, 안전운전불이행. 중앙선침범, 교차로운행방법위반, 직진우회전진행방해, 안전거리미확보, 차로위반, 기타
노면상태	건조, 결빙 및 적설, 습기, 기타
기상상태	맑음, 비, 흐림, 눈, 안개, 기타
도로형태	교차로(교차로 부근, 교차로 안), 단일로(횡단보도 부근, 횡단보도 상, 교량, 터널, 기타)
가해운전자 차종	건설기계, 승용, 승합, 원동기, 이륜, 자전거, 특수, 화물, 농기계
가해운전자 성별	남/여
가해운전자 연령	0~19, 20~29, 30~39, 40~49, 50~59, 60~69, 70~79, 80+
가해운전자 상해정도	상해 없음, 경상, 중상, 사망, 부상신고
피해운전자 상해정도	상해 없음, 경상, 중상, 사망, 부상신고
피해운전자 차종	건설기계, 승용, 승합, 원동기, 이륜, 자전거, 특수, 화물, 농기계, 보행자

재정의를 한 교통사고 데이터는 총 13개의 속성으로 구성되었다. 상해정도의 속성은 '상해 없음', '경상', '중상', '사망', '부상신고', '원인 불명'으로 구성되어 있는데, 원인 불명은 상해 정도를 알 수 없으므로 해당 데이터를 제거하였다. 연속치 데이터인 발생 시간은 '01~06시', '07~12시', '13~18시', '19~24시'로 구분하였으며, 연령의 경우 10년 단위로 구분하였다. 순서적 범주형 데이터인 발생 요일은 평일과 주말로 단순화하여 구분하는 이산화 및 Labeling을 통해 전처리 과정을 수행하여 교통사고를 재정의 하여 분류 학습 데이터의 모델로 사용한다.

나. 불균형 데이터 집합

교통사고 데이터의 '피해자 상해 심각도'는 전체 데이터 중, 다수 클래스인 사망은 40.83%, 소수 클래스인 부상신고는 8.67%로 데이터의 불균형을 나타낸다. 교통사고 데이터를 학습 분류모델로 사용하기 위해 불균형 데이터에 Sampling 과정이 필요하다. 표 3은 원본데이터의 피해자 상해 심각도에 따른 레코드 수 분포도이다.

표 3. 원본 피해자 상해 심각도 레코드 수 분포도

피해자 상해 심각도	레코드 수 (개)	비율 (%)
상해 없음	638	21.42
경상	476	15.98
중상	389	13.06
사망	1220	40.95
부상 신고	256	8.59
합계	2979	100

교통사고 데이터의 '피해자 상해 심각도'의 경우 전체 데이터에서 상해 없음, 경상, 중상, 사망, 부상 신고 클래스가 차지하는 비율은 모두 다르다. 소수와 다수의 클래스의 분포로 나타나지는 것을 불균형 데이터라고 하며, 불균형 데이터의 경우 분류 모델을 수행 시 소수 클래스가 다수의 클래스로 분류되는 등 오·분류의 문제가 발생하게 된다. 이러한 문제점을 해결하기 위해 기존 데이터에서 Sampling 과정을 수행하여 '피해자 상해 심각도'의 모든 클래스 분포가 20%인 균형적인 데이터를 구축하고 불균형 데이터와 균형적인 데이터를 분류 모델을 수행하기 위한 학습 데이터로 사용한다.

3. FP-Growth 알고리즘을 이용한 피해자 상해 심각도 연관 규칙패턴 추출

본 절에서는 전처리 과정이 수행된 데이터에서 FP-Growth 알고리즘을 이용하여 '피해자 상해 심각도'와 연관된 규칙 패턴에 관해서 기술한다. FP-Growth 알고리즘은 후보 패턴을 만들지 않고 데이터 필드를 두 번에 걸쳐 객체를 삽입하고 트리를 만드는 구조로 시간과 메모리를 절약하기 위해 제안한 연관규칙학습기법의 대표적인 알고리즘이다[12]. '피해자 상해 심각도'를 대상으로 FP-Growth 알고리즘의 최소 지지도를 적용하면, 만족하지 않은 패턴 규칙은 제외된 빈발 데이터 집합이 생성된다. 최소 지지도가 높을 경우 신뢰성이 높은 패턴 규칙이 생성될 수 있지만, 유용한 연관 패턴 추출을 할 수 없고, 최소 지지도가 낮을 경우, 많은 패턴이 생성되어 신뢰성이 낮은 패턴이 추출되거나 연관 규칙 패턴 결과를 종합하기 어렵다. 표 4는 최소 지지도별 피해자 상해 심각도의 연관규칙 패턴을 추출한 개수이다.

표 4와 같이 최소 지지도별 추출되는 패턴의 개수는 모두 다르며, 최소 지지도가 낮을수록 추출된 연관규칙패턴의 개수가 많아진다. '피해자 상해 심각도'에서 부상신고의 클래스가 가장 적은 레코드를 가지고 있기 때문에 적은 개수의 연관규칙이 추출되고, 사망 클래스의 레코드 개수가 다른 클래스에 비해 많기 때문에 많은 연관규칙패턴이 추출된다. 최소 지지도가 50%를 넘어갈 경우, 다른 클래스의 유용한 패턴 규칙이 추출되지 않는다. 따라서 본 논문에서는 각 클래스의 최소 지지도를 50%로 적용하고 연관규칙을 추출한다. 표 5는 불균형 클래스 집합을 가지는 데이터에서 피해자 상해 심각도와 연관된 패턴을 추출한 결과이다.

표 4. 최소 지지도 별 상해심각도 연관패턴 개수

피해자 상해 심각도	최소 지지도		
	50%	60%	70%
상해 없음	53	39	21
경상	41	27	13
중상	39	15	7
사망	112	65	25
부상신고	33	25	7

표 5. 50%의 최소 지지도를 적용한 피해자 상해 심각도의 연관규칙패턴

피해자 상해 심각도	연관규칙패턴
상해 없음	법규위반 = 안전운전불이행
	운전자 성별 = 남 & 기상상태 = 맑음
	기상상태 = 맑음
	기상상태 = 맑음 & 노면상태 = 건조
	⋮
경상	도로형태 = 단일로 - 기타
	기상상태 = 맑음 & 노면상태 = 건조
	발생요일 = 평일
	운전자 성별 = 남 & 기상상태 = 맑음
	⋮
중상	운전자 차종 = 승용
	피해자 차종 = 승용
	도로형태 = 단일로-기타
	운전자 성별 = 남 & 도로형태 = 단일로-기타
	⋮
사망	안전운전 불이행 & 발생요일 = 평일
	운전자 상해 심각도 = 상해 없음 & 발생요일 = 평일 & 법규위반 = 안전운전불이행
	피해자 차종 = 보행자
	운전자 성별 = 남 & 피해자 차종 = 보행자
	운전사 상해 심각도 & 노면상태 = 건조
	운전자 성별 = 남 & 운전자 상해 심각도 = 상해 없음 & 노면상태 = 건조
	⋮
부상신고	법규위반 = 안전운전불이행 & 노면상태 = 건조
	도로형태 = 단일로-기타
	운전자 나이=50
	운전자 차종 = 승용
	법규위반 = 안전운전불이행 & 기상상태 = 맑음
	법규위반 = 안전운전불이행 & 운전자 상해심각도 = 사망
	⋮

'피해자 상해 심각도'에서 주로 운전자의 성별이 남자이거나, 기상상태가 맑고, 노면상태가 건조한 경우는 모든 클래스에 공통적으로 나온 연관규칙패턴들이다. 이는 운전자의 성별비가 남자인 경우가 높고, 날씨가 좋은 날에 차량을 이용하는 경우가 많아 교통사고 데이터의 비율 자체가 높기 때문에 모든 클래스에서 공통적으로 추출된다. 모든 클래스에 공통적으로 나온 연관패턴들은 분류모델을 수행 할 때, 오·분류를 수행할 가능성이 높은 속성들로, 해당 속성들을 제외하고 분류모델을 수행한다. 또한, 다르게 나온 연관규칙패턴들은 해당 클래스에 높은 빈도수를 가지고 있는 속성으로 분류모델을 수행할 시, 가중치를 주어 분류모델을 수행한다.

4. 피해자 상해 심각도 예측을 위한 Hybrid Model

불균형 데이터와 균형 데이터에서 '피해자 상해 심각도'를 예측하기 위한 데이터 마이닝 기법을 적용한다. 피해자 상해 심각도의 예측성능을 높이기 위해 데이터마이닝 기법을 Single Model이 아닌, 다양한 모델을 결합한 Hybrid Model을 이용한다. Hybrid Model은 입력 데이터를 처리하는 방법과 모델을 통해 나온 결과를 처리하는 방법 등에 따라 여러 유형이 있다. 본 연구에서 사용된 Hybrid Model은 Whole Data Approach 접근법의 4가지 유형 중 W3 Type의 모델을 이용한다[13]. 그림 2는 W3 Type의 Hybrid Model을 나타낸다.

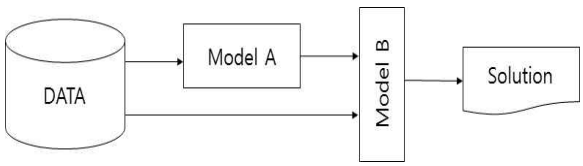

그림 2 Whole Data Approach 접근법의 W3 Type

W3 Type은 데이터에서 모델 A를 수행한 결과와 기본 데이터를 모델 B의 입력데이터로 사용한다. 본 연구에서는 Model A는 의사결정트리이며, Model B는 다항 로지스틱회귀 기법을 적용하였다. 의사결정트리는 입력 변수를 바탕으로 목표변수의 값을 예측하는 모델을 생성하는 것으로, C4.5 알고리즘을 이용하여 의사결정트리를 구축하고 중요 속성을 결정하는 모델로 사용하였다. 다항 로지스틱 회귀기법은 예측변수의 값에 따라서 대상을 분류할 때 사용되는 기법으로 일반적인 로지스틱 회귀기법의 경우, 종속변수가 이항형 문제를 지칭할 때 사용되므로 다항 로지스틱 회귀기법을 사용하여 '피해자 상해 심각도'를 예측한다. 의사 결정트리에서 중요 변수를 추출하기 위해 Entropy와 Information Gain 계산식을 이용한다. 식 1은 Entropy 계산식이며, 식 2는 Information Gain 계산식이다.

$$Entropy(S) = -\sum_{i=1}^{m} p_i \log_2 \left(\frac{freq(C_i, S)}{|S|}\right) \quad (1)$$

Entropy는 주어진 데이터 집합의 혼잡도를 의미하는 것으로, 0에서 1사이의 값을 갖는다. 1에 가까울수록 다른 클래스들로 구성되어있고, 0에 가까울수록 같은 클래스의 레코드들이 많이 있다. 식 1에서 S는 주어진 데이터의 집합을 나타내며, C는 클래스 값들의 집합을 나타낸다. $freq(C_i, S)$는 S에서 클래스 C_i에 속하는 레코드 개수를 말하며, $|S|$는 데이터 집합의 데이터 개수를 나타낸다.

$$Gain(A) = I(s_1, s_2, \cdots, s_m) - E(A) \quad (2)$$

Information Gain은 특정 속성을 기준으로 데이터를 구분하게 될 때, 감소되는 Entropy의 양을 의미하며, 값이 클수록 정보 이득이 크고, 해당 속성의 변별력이 좋다는 것을 의미한다. 식 2에서 $I(s_1, s_2, \cdots, s_m)$는 상위노드의 Entropy를 의미하며, $E(A)$는 A 속성을 선택 시, 하위 노드의 Entropy를 이용하여 노드에 속한 레코드의 개수를 가중치로 하여 엔트로피를 평균한 값이다. 표 6과 7은 Information Gain을 이용하여 의사결정트리를 구축하고 10겹 교차방식을 통해 불균형 데이터와 균형데이터에서 각 피해자 상해 심각도를 구분하는 중요 속성들을 나타낸다.

표 6. 의사결정트리를 이용한 불균형 데이터 중요속성

피해자 상해 심각도	중요 속성
상해 없음	사고유형, 법규위반, 운전자 나이, 운전자 차량 종류, 피해자 차량 종류
경상	운전자 나이, 피해자 차량 종류, 도로 형태
중상	운전자 성별, 사고발생시간, 사고유형, 노면상태
사망	피해자 차량 종류, 운전자 나이. 사고 발생시간, 도로 형태, 사고 유형
부상신고	운전자 차량 종류, 노면상태, 사고유형, 사고 발생시간, 기상상태

불균형 데이터에서 의사결정트리를 통해 나온 중요 속성 중 상해 없음은 피해자 차량종류와 운전자 나이, 경상은 도로형태, 중상은 법규위반사항, 사망은 피해자 차량 종류, 부상신고는 노면상태와 사고 발생시간, 사고유형이 가장 중요한 속성으로 작용한다.

균형 데이터에서 중요 속성 중 상해 없음은 운전자 차량 종류, 경상은 운전자 차량종류와 사고 발생시간, 운전자 상해 심

각도, 중상은 노면 상태와 운전자 나이, 사망은 운전자 상해 심각도, 부상신고는 사고 발생시간, 운전자 나이, 도로 형태가 가장 중요한 속성으로 작용한다.

표 7. 의사결정트리를 이용한 균형 데이터 중요속성

피해자 상해 심각도	중요 속성
상해 없음	기상 상태, 법규 위반, 운전자 차량 종류, 운전자 나이, 사고 유형
경상	운전자 차량 종류, 사고 발생시간. 운전자 상해 심각도, 운전자 나이, 사고유형, 도로 형태
중상	사고 유형, 사고 발생시간, 운전자 성별, 노면상태, 도로 형태, 법규위반, 운전자 나이
사망	운전자 상해 심각도, 피해자 차량 종류
부상신고	사고 발생시간, 사고 발생요일, 운전자 상해 심각도, 운전자 나이, 도로형태

표 6과 7의 결과에 따라 불균형 데이터와 균형 데이터에서 분류를 위한 가장 중요한 속성이 다른 점을 알 수 있다. 의사결정트리는 입력 속성에 따라 목표 값을 예측하는 것으로 하향식 기법을 통해 적합한 속성을 기준으로 분류된다. Sampling은 분류를 수행하기 위해 목표 변수를 기준으로 데이터를 조절하기 때문에 원본 데이터와 Sampling을 한 후의 데이터의 중요속성이 달라질 수 있고 모든 속성의 탐색이 어렵다. 따라서 본 논문에서는 표 5에서 나온 연관규칙패턴의 결과와 표 6과 7을 통한 의사결정트리를 이용한 중요속성에 따라 가중치를 부여하고, 다항 로지스틱회귀기법을 이용하여 불균형 데이터인 원본데이터와 균형데이터를 통해 피해자 상해 심각도를 예측한다.

IV. 실험 결과 및 고찰

본 연구에서는 혼동 행렬방식을 이용하여 Single Model과 Hybrid Model을 적용한 정밀도를 통해 성능을 비교 평가하였다. 혼동 행렬은 총 4개의 요소로 구성되어 있으며, 표 8은 혼동행렬의 예를 나타낸다[12].

표 8. 혼동 행렬의 예

	예측된 ⊕	예측된 ⊖
실제 ⊕	TP	FN
실제 ⊖	FP	TN

혼동 행렬에서 TP(True positive)는 실제 1인 값을 1로 예측하여 올바르게 판단된 경우이며, TN(True Negative은 실제 0인 값을 0으로 올바르게 판단된 경우를 말한다. 이와 반대로 FP(False Positive)는 실제 0인 값이나 1로 예측을 한 잘못된 판단이며, FN(False negative)는 실제 1이나 0으로 예측한 잘못된 판의 경우를 말한다. 식 3은 혼잡행렬을 이용하여 민감도를 구하는 식이다.

$$Recall = \frac{(TP)}{(TP+FN)} \quad (3)$$

민감도는 실제 값이 1인 것 중 예측 값이 1인 경우를 나타내며, 민감도가 1에 가까울수록 분류가 잘 이루어졌음을 알 수 있다. 본 논문에서 제안한 Hybrid Model을 이용하여 피해자 상해 심각도를 예측하는 방법이 Single Model을 사용하는 방법보다 정확한지 확인하기 위해 3.2절에서 재정의한 데이터를 이용하여 식 3을 이용하였다. 표 9는 균형데이터와 불균형 데이터에서 Single Model로써, 각각 10겹 교차방식을 이용하여 의사결정트리와 로지스틱회귀기법을 사용했을 경우 나타나는 민감도이다.

표 9. Single Model 민감도 결과

피해자 상해 심각도	의사결정트리		로지스틱회귀	
	불균형 데이터	균형 데이터	불균형 데이터	균형 데이터
상해 없음	0.73	0.62	0.72	0.69
경상	0.56	0.59	0.46	0.46
중상	0.63	0.67	0.47	0.58
사망	0.91	0.88	0.92	0.89
부상신고	0.89	0.95	0.86	0.98

표 9는 각 Single Model의 특성을 보여준다. Single Model에 따라 피해자 상해 심각도별 예측 정확도는 모두 다르다. 의사결정트리는 레코드 개수가 적은 클래스의 민감도가 높은 것을 확인할 수 있으며, 로지스틱 회귀는 레코드 수가 많은 클래스의 분류 정확도가 높다. 레코드의 수가 적은 부상신고가 민감도가 높게 나온 이유는 활용한 기존 데이터에서 부상신고가 발생하였을 때 나타나는 중요속성이 많은 비중을 차지하여 민감도가 높게 나왔다. 따라서 Single Model은 모든 클래스에서 예측 정확도가 높은 것은 아니기 때문에 Single Model만을 사용할 수 없다. 또한, 표 9의 결과에 따라 Sampling한 데이터가 모든 분류 결과에서 좋게 나오지 않는다는 것을 확인할 수 있다. 균형 데이터는 소수 클래스의 민감도는 향상되나 다수 클래스의 민감도는 감소되는 것을 볼 수 있다. 이와 같이, 기존 연구에서 Sampling 한 데이터만을 사용하여 분류 모델을 수행 시, 모든 클래스에서 높은 예측 성능을 보이는 것은 아니다. 따라서 본 연구에서는 불균형 데이터와 균형 데이터를 모두 사용하고, 데이터 마이닝 기법을 융합한 Hybrid Model을 구축하였다.

표 10은 본 연구에서 구축한 Hybrid Model을 이용하여 예측을 진행한 민감도의 결과이다.

표 10. Hybrid Model 민감도 결과

피해자 상해 심각도	Hybrid Model
상해 없음	0.76
경상	0.50
중상	0.64
사망	0.93
부상신고	0.92

기존 연구에서 다른 클래스에 비해 레코드의 수가 많은 상해 없음과 사망 클래스의 경우 균형데이터만 이용하여 Single Model 수행 시 불균형 데이터를 이용할 때보다 예측 성능이 저하되었으나, 본 연구에서 구현한 Hybrid Model은 분류 모델의 성능을 높여준다. 두 클래스의 예측 성능은 불균형 데이터, 균형 데이터를 이용하여 Single Model을 수행했을 때보다 높은 결과가 나왔으며, 레코드 수가 적은 경상과 중상의 경우 로지스틱회귀 기법만을 사용할 때보다 높게 나왔으나, 의사결정트리보다는 낮게 나온 결과를 확인할 수 있다. 부상신고의 경우 불균형 데이터에서 Single Model 수행 시 보다 높게 나왔으나 균형 데이터를 이용한 예측 결과에 비해 낮게 나오는 것을 확인할 수 있다.

Ⅴ. 결론

본 논문에서는 균형 데이터와 불균형 데이터를 이용하여 피해자 상해 심각도를 예측하는 Hybrid Model에 관해 제안하였다. 원본 데이터에서 전처리 과정을 수행한 후 분류 모델을 수행할 수 있도록 교통사고 데이터를 재정의하였다. 원본 데이터에는 총 23개의 속성으로 이루어져있으나 데이터의 일반화를 위해 피해자 상해 심각도와 관련이 없는 속성은 제거되어 총 13개의 속성으로 재정의 되었으며, 연속치 데이터와 순서적 범주형 데이터는 일정 구간으로 구분하였다. 원본 데이터에는 피해자 상해 심각도 클래스의 레코드 수가 동일하지 않기 때문에 데이터 불균형이 발생되며, Sampling과정을 수행하여 균형적인 데이터를 생성하였다. 원본 데이터를 이용하여 분류모델을 수행할 때 고려해야할 사고 패턴을 확인하기 위해 FP-Growth 알고리즘을 이용하여 피해자 상해 심각도와 연관된 패턴을 추출하였고, 불균형 데이터와 균형데이터에서 의사결정트리를 이용하여 중요 속성을 추출하여 비교하였다. 비교된 중요 속성은 로지스틱회귀에서 분류를 수행할 때, 가중치의 역할로 사용하였다. 기존 연구에서는 분류 모델을 수행하기 위해 Sampling 데이터만 이용하였으나 이는 원본 데이터의 손실을 발생시켜 다수 클래스의 예측 정확도가 낮아지는 것을 알 수 있다. 따라서 본 논문에서는 원본 데이터의 불균형 데이터와 Sampling된 균형데이터를 모두 이용하여 Hybrid Model을 수행하였으며, Single Model과 균형 데이터만을 이용한 예측 모델에 비해 소수 클래스의 예측 역시 비교적 좋은 결과를 얻을 수 있고, 다수 클래스의 예측 성능은 불균형 데이터, 균형 데이터만을 이용할 때보다 좋은 예측 성능 결과를 확인할 수 있다. 향후 연구로는 본 연구를 통한 피해자 상해 심각도 패턴들을 이용하여 지형과 교통사고 피해에 관한 연관 패턴을 분석하고 이를 예방하기 위한 연구를 진행할 계획이다.

References

[1] Ministry of Land, Infrastructure and Transport
[2] TAAS Traffic Accident Analysis System
[3] C.K. Lee, “A Study of Big Data Information Systems Building and Cases,” Journal of the KISM Smart Media, Vol.4, No.3, pp. 56-61, 2015.
[4] S.S. Han and B.H. Park, “Comparative Analysis of Traffic of Cheongju,” Korea Planning Association, Vol. 46, No. 2, pp. 183-192, 2011.
[5] S.Y. Sohn and S.H. Lee, “Data Fusion, Ensemble and Clustering for the Severity Classification of Read Traffic Accident in Korea,” Safety Science, Vol. 41, No. 1, pp. 1-14, 2013. 5.
[6] Chang, M., A. Abraham and M. Paprzycki, “Traffic Accident Analysis Using Machine Learning Paradigms,” Informatica, Vol. 29, pp. 89-98, 2005.
[7] J.S. Lee and K. Huh, “Injury Severity Prediction of Traffic Accident using Data Mining,” General Autumn Conference of Korea Intelligent Information System Society, pp. 199-206, 2011.
[8] S.E. Hong, G.Y. Lee and H.J Kim, “A Study on Traffic Accident Injury Severity Prediction Model Based on Public Data,” Jornal of KIIT, Vol. 13, No. 5, pp. 109-118, 2015.
[9] J.S. Lee and E.J. Lee, “Analysis of Traffic Accident using Decision Tree Ensemble Model,” General Autumn Conference of Korea Intelligent Information System Society, Vol. 11, pp. 211-218, 2009.
[10] E.J. Lee, “Analysis of Traffic Accidents using Data Mining Ensemble Models,” Master’s Thesis, Ajou University, 2010.
[11] J.S. Lee, J.G. Kwon, “A Hybrid SVM Classifier for Imbalanced Data Sets,” Journal of Intelligent Information Systems, No. 19, Vol. 2, pp. 125-140, 2013.
[12] Jason Bell, “machine Learning: Hands-On for Developers and Technical Professionals,” John Wiley & Sons, pp. 1-408, 2014
[13] J.S. Lee and J.C. Lee, “Customer Chum Prediction by Hybrid Model,” Advanced Data Mining and Applications, Vol. 4093, pp. 959-966, 2006.

저 자 소 개

주영지(준회원)
2016년 조선대학교 제어계측로봇공학과 공학사
2016년~현재 조선대학교 소프트웨어융합공학과 석사과정
<주관심분야 : 데이터 마이닝, 빅 데이터 처리, 병렬처리 시스템>

홍택은(준회원)
2015년 조선대학교 컴퓨터공학부 공학사
2015년~현재 조선대학교 소프트웨어융합공학과 석사과정
<주관심분야 : 소셜 네트워크, 오피니 마이닝, 감성정보 분석>

신주현(정회원)
1986년~2011년 ㈜청전정보 팀장, ㈜투루텍 기술이사
2007년 조선대학교 전자계산학과 이학박사
2011년~현재 조선대학교 제어계측로봇공학과 산학협력중점교수
<주관심분야 : 멀티미디어 데이터베이스, 빅 데이터 처리, 텍스트마이닝, 감성정보 처리 등>