데이터 마이닝을 이용한 교통사고 심각도 분류분석

Data Mining for Road Traffic Accident Type Classification

손소영 (연세대학교 산업시스템공학과 부교수)

신형원 (연세대학교 산업시스템공학과 석사과정)

목 차

ABSTRACT

본 연구는 교통사고 심각도와 관련된 중요변수를 찾고 이들 변수를 바탕으로 신경망, Decision Tree, 로지스틱 회귀분석을 이용하여 사고 심각도 분류 예측모형을 추정하였다. 다수의 범주형 변수로 이루어진 교통사고 통계원표상의 설명변수 들로부터 사고 심각도변화에 영향력 있는 변수 선택을 위하여 χ^2 독립성 검정과 Decision Tree를 이용하였고, 선택된 변수들은 신경망과 로지스틱회귀분석의 기초로 이용되었다. 분석결과 세가지기법간에 분류정확도에는 유의한 차이가 없는 것으로 나타났다. 그러나 Decision Tree가 설명변수 선택능력과 분석수행시간, 사고 심각도 결정요인 식별의 용이함 측면에서 범주형 종속변수인 사고 심각도의 분석에 적합한 것으로 보이며 사고 심각도에는 보호장구가 가장 큰 영향을 미치는 것으로 재입증되었다.

Ⅰ. 서론

우리나라의 산업발전과 더불어 증가된 교통량은 환경문제와 함께 많은 문제를 야기하고 있다. 지난 92년부터 조금씩 줄었던 교통사고 사망자가 95년 이후 다시 크게 늘어나기 시작해 96년 한 해 26만5천52건의 교통사고가 발생했고 그 중 1만2천6백53명은 사망자로 전년대비 22.6% 증가를 기록하고 있다.(경찰청, 1996) 이처럼 높은 교통사고 증가율은 강한 사회적 위기감을 일으키고 있으며, 막대한 규모의 사회·경제적 손실을 초래하고 있다.

교통사고 감소를 위한 대책을 세우기 위해서는 교통사고 발생원인 및 특성을 규명하는 일이 우선되어야 한다. 교통사고 발생과 관련된 도로시설구조, 교통이용실태, 기후조건, 운전자특성, 차량 특성 등 다양한 자료의 과학적이며 종합적인 조사 분석을 통하여 교통사고 예방을 위한 적절한 조치를 취할 수 있다. 이를 위하여 경찰청에서는 교통사고 조사보고서를 바탕으로 사고와 관련된 다양하고 상세한 정보를 사고 1건당, 79개 항목으로 이루어진 '통계원표'에 기록하고 있다. 그러나 대부분이 다수준 범주형으로 기록되는 '통계원표' 자료는 그 양이 매우 방대하며 자료간에 복잡한 상관관계가 있어 분석을 하는데 많은 시간과 비용이 소모된다. 따라서 본 연구에서는 자료

를 빠르고, 정확하고, 다양하게 분석하기 위하여 데이터 마이닝 기술을 이용한 교통사고 분석을 하고자 한다.

본 논문에서는 사고 심각도(통계원표의 '사고내용' 항목)를 통계원표상의 교통환경 변수로 분류 예측하기 위해 여러가지 데이터 마이닝기법중 신경망(Neural Network)분석, Decision Tree, 로지스틱 회귀분석을 이용하였으며 모형추정을 위해 96년 서울에서 발생한 교통사고 자료를 활용하였다. 본 연구를 통하여 사고 심각도 변화에 관련된 설명변수 및 수준(level)을 식별하고 사고 심각도에 대한 데이터마이닝 기법별 분류 정확성을 비교하고자 한다.

이를 위한 본 논문의 구성은 다음과 같다. 2장에서 데이터 마이닝에 대한 일반적 고찰과 기존 문헌을 정리하였다. 3장에서는 분류모형을 위한 변수선택법과 세 가지 기법의 분류정확도를 분석하였으며 마지막으로 결론에서는 논의된 내용을 종합하고 교통사고 분석에 적합한 데이터 마이닝 기법을 제시하였다.

II. 데이터 마이닝

데이터 마이닝이란 대용량의 데이터 속에서 알려지지 않은 패턴을 발견하여 활용하기 적합한 정보로 변환하고 이를 바탕으로 다음에 무엇이 일어날지 예측하는 과정(기술)이다.(Berry & Linoff, 1997) 데이터 마이닝을 이용한 분석시에는 자료의 형태와 활용목적에 따라 어떤 기법을 적용시킬 것인지를 결정하여야 한다. 분류를 목적으로 하는 데이터 마이닝 기법에는 크게 '패턴추출', '데이터의 보유'로 나뉘어 진다. 패턴 추출은 데이터의 집합 속에 숨겨져 있는 정보를 찾기 위하여 데이터를 분석하고 패턴을 추출하는 반면, 데이터 보유는 데이터를 기존에 알고 있는 패턴과 연관 지어 분류한다. 즉, 새로 발생한 데이터와 기존의 데이터와의 관계성을 찾는 기법이다.(고재성, 1997) 패턴추출은 다시 적용기법에 따라 Decision Tree, 로지스틱 회귀분석, 신경망 등으로 나뉘어 지며 데이터 보유는 최근이웃법, 사례기반추론등 으로 나눌 수 있다. 본 연구에서는 여러 데이터 마이닝 기법중 일반적으로 예측능력에 높은 정확성을 가지고 있다고 평가되고 있는 신경망과 범주형 자료에 높은 분류 정확성을 가지고 있고 대상이 되는 결과에 대하여 그 원인을 나뭇가지 형태로 찾아가 사용자가 알아보기 쉬운 장점이 있는 Decision Tree, 전통적 통계분석 기법으로써 오랜 기간 이용 되어온 로지스틱 회귀분석을 교통사고 자료 적용시 사고 심각도 분류 정확성의 관점에서 비교해 보았다. 실제 자료분석에 앞서 각 기법을 간단히 소개하면 다음과 같다.

Decision Tree란 대상이 되는 집단을 설명변수를 기준으로 나뭇가지처럼 몇 개의 소집단으로 구분하여 분류(Classification)하고 예측하는 기법이다. Decision tree의 진행과정은 관측된 분류결과를 나타내는 종속변수에 영향을 끼칠 수 있는 모든 변수들을 반복적으로 검색하여 데이터의 분류에 가장 중요한 설명변수를 선택한 뒤 이 변수로 데이터를 분류하고 다음으로 중요한 변수를 선택하여 나머지 데이터를 다시 분류한다. Decision Tree는 독립성 검정을 목적으로 χ^2 test를 통하여 가장 낮은 유의도값을 갖는 설명변수를 이용하여 분류하는 CHAID와 Entropy 또는 GINI Index등에 의하여 분류하는 C4.5, CART알고리즘 등이 있다.(Bigus, 1996)

인공 신경망은 인간의 두뇌를 모방하여 인간 두뇌활동의 메카니즘을 수학적으로 재현한 인공지능의 한 분야이다. 신경망중 가장 널리 사용되는 역전파 신경망의 구성은 입력층(Input Layer), 하나 이상의 은닉층(Hidden Layer), 출력층(Output Layer)으로 이루어져 있으며 은닉층은 뉴런(neuron) 또는 노드(node)라 불리는 요소(element)로 구성되어 있다. 이들은 다음 층의 요소와 연결강도 또는 가중치(weight)를 갖는 링크(link)로 연결되어 있으며 전단계의 출력값을 입력값으로 받아 특정 활성함수(activation function)에 의해 출력값을 생성하게 되는데 이 값은 다시 다음 단계의 입력값으로 작용하게 된다.

로지스틱 회귀분석은 종속변수가 범주형일 때 그 변화를 설명변수(X)의 함수로 하여 예측하고자 할 때 사용되는 모수적인 방법이다. 일반적으로 로지스틱 회귀분석은 다음과 같이 정의된다.

$$P_{i(x)} = \frac{\exp[\alpha_i + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_n X_n]}{1 + \exp[(\alpha_i + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_n X_n)]}$$

여기서 $P_{i(x)}$는 주어진 설명변수 (X_1,.....,X_n) 하에서 종속변수의 분류수준이 i 보다 작거나 같을 확률을 말한다.

일반적으로 설명변수가 분류 결정에 미치는 영향은 관련모수 β 에 대한 추론으로 가능하다. 그러나 통계원표에 의한 교통사고 자료와 같은 범주형 데이터에는 각 설명변수당 범주수에 비례하는 지시변수를 이용하여 모형화 함으로써 이와 같은 방법을 직접적으로 사용할 수 없다. 따라서 각 설명변수 수준들의 영향력은 참조집합과 비교하여 상대적으로 사고심각도에 미치는 영향을 비교 하도록 한다.

이상에서 제시된 세가지 데이터 마이닝 기법을 교통사고 자료에 적용하기 위해서는 데이터를 Training 과 Validation 두 가지 용도로 분할하는 과정이 필요하다. Training 데이터는 모델을 만드는데 사용되는 자료이며 Validation 데이터는 Training 자료로 만들어진 모델을 검증하는데 사용한다. 다양한 패턴추출 기법들의 분류능력을 비교한 기존 문헌은 다수 있으며 몇 개의 예는 다음과 같다. Brokett et al(1997)은 기업의 파산을 예측하기 위하여 파산에 영향을 미칠 수 있는 8개 변수들을 선택한 뒤, 이들로 로지스틱 회귀분석과 유사한, 판별 분석과 신경망 분석을 하여 파산과 비파산 분류모형을 수립하였다. 이 연구에서는 Texas 주의 44개 기업에 대한 자료를 가지고 신경망 분석을 한 결과, test 데이터의 분류 정확도는 91.11%이었으며 판별분석의 분류 정확도는 93% 이었다. 판별분석의 분류 정확도가 이처럼 높은 것은 판별분석시 training 데이터로 prediction을 했기 때문이다.

Leshno 와 Spector(1997)는 신경망과 판별분석의 분류 정확도를 비교하고 training sample size가 분류에 미치는 영향을 분석하였다. 이 연구에서는 두 개의 입력변수와 두 개 그룹으로 나뉘는 출력변수를 만들고 입출력변수간의 선형함수, Quadratic 함수등 네 가지 함수로 모의 데이터를 만들었다. 신경망은 은닉층과 뉴런의 수를 달리해가며 6개의 신경망 구조를 적용하여 분류하였다. 이상과 같은 연구결과, 신경망 분석이 판별분석보다 예측능력이 뛰어나고 은닉층에 있는 뉴런의 수가 많을수록 오분류율이 감소된 것으로 나타났다. 또 신경망은 선형함수를 갖는 모형보다 복잡한 모형에 예측력이 더 뛰어났으며 Training 데이터 크기가 클수록 일반화 능력이 뛰어난 것으로 나타났다.

Cherkassky et al(1996)은 Nearest Neighbors, Pursuit, Artificial Neural Networks 등 6가지 방법에 모의(simulated) 데이터를 사용하여 예측력 성능 비교를 하였다. 모의 데이터는 2차원 연결함수 8종류와 2개의 데이터의 분포, 3개의 데이터 크기, 3개의 잡음(noise) 수준을 조합하여(8×2×3×3=144) 자료를 만들고, 연결함수 5 종류와 1개의 데이터 분포, 3개의 데이터 크기, 3개의 잡음 수준을 조합하여(5×1×3×3=45) 총 189개의 treatment를 만들었다. 이 treatment들을 각 분류방법에 적용하여 방법별로 특징적인 실험결과를 요약하였으며 전반적으로 인공 신경망이 예측 정확도면에서 앞서는 것으로 나타났다.

III. 교통사고 심각도 분류모형

교통사고 자료를 기록하는 통계원표에는 해당 사고 1건마다 사고 심각도를 알 수 있는 항목을 포함하여 79개의 항목을 기록하고 있다. 이들 79개 항목은 대부분이 범주형으로 이루어져 있으며 일부 항목은 가해자(1당), 피해자(2당)로 나누어 기록하고 있다. 본 연구에서는 이들 통계원표 자료중 1996년 서울에서 발생한 11564건의 자료를 표본 추출하여 사고 심각도를 3가지 데이터 마이닝 기법별로 예측분류 하였다. 사고 심각도를 나타내는 '사고내용' 항목은 <표-1>과 같이 5개

의 범주로 나뉘어져 있다. 그러나 이들중 '사망' 이나 '부상신고' 범주에 속하는 사고는 극히 소수이므로 사고 심각도의 분류 예측에 어려움이 예상된다. 따라서 본 장에서는 <표-2>에 나타난 바와 같이 종속변수인 사고 심각도를 3개와 2개의 범주로 나누고, 이러한 범주를 분류할 수 있는 모형 추정을 위한 변수 선택시 χ^2-test 와 Decision Tree를 이용하였다. 전체자료 중 모형추정을 위해 Training 데이터에 60%를 할당하였고 분류정확성 측정을 위해 40%에 해당하는 4626건을 Validation 데이터로 할당하였다. 선택된 변수를 바탕으로 신경망 분석시 설명변수와 종속변수간의 구조로는 동일하게 2개의 은닉층에 각각 3개와 2개의 뉴런을 설정하였으며 활성함수는 로지스틱 함수를 사용하였다. Decision Tree는 CHAID를 이용하여 분석하였다.

<표-1> 사고 심각도 항목의 범주구분과 내용

종속변수명	범주	범주별 비율
X4	사망	0.9%
	중상	26.8%
	경상	31.6%
	부상신고	0.6%
	물적피해	40.0%

1. 세 개 범주 분류

통계원표의 사고 심각도 항목은 <표-1>과 같이 5개의 범주로 이루어져 있으나 본 절에서는 <표-2>의 NX4와 같이 범주의 수를 3개로 줄여 분류 정확성을 측정하였다.

<표-2> 종속변수의 범주구분과 비율

종속변수명	범주
NX4	치명적 상해(사망+중상)
	경미한 상해(경상+부상신고)
	물적피해
NNX4	신체상해(사망+중상+경상+부상신고)
	물적피해

데이터 마이닝을 이용한 자료분석시 많은 사용가능 변수중 필요한 변수를 선택하여 신속한 분석을 하게 되면 데이터의 수집비용과 효율성을 증가시킬 수 있다. 특히 통계원표는 79개의 다양한 항목으로 이루어져 있고 각 항목마다 많은 범주로 구분되어 있어 교통사고 자료분석 대상을 중요항목 만으로 줄이는 것이 필수적으로 요구된다. 이를 위하여 <표-2>에 정리된 바와 같이 세 가지 수준의 사고 심각도(NX4)를 예측하는데 사용할 설명변수의 선택을 위해 통계원표상의 여타 변수와 사고심각도간의 독립성 검정을 위하여 χ^2 시험 (유의수준 5%)을 하였다. 분석결과, 79개 중 22개의 변수가 종속변수에 대하여 독립이라는 가설을 기각하였다. 위의 22개 변수를 모두 설명변수로 사용했을 때의 분류정확도는 3가지 데이터 마이닝 기법이 3개의 종속변수에 대하여 기본적으로 기대되는 분류정확성인 33%보다 다소 높은 약 55%의 분류 정확도를 보였다.

변수선택에 일반적으로 많이 사용되는 방법은 판별분석이나 로지스틱 회귀분석과 같은 모수분석을 하여 부분 상관계수로 종속변수의 분류에 높은 상관관계가 있는 변수를 찾는다. 그러나 본 연구에 사용된 통계원표 자료는 다차원 범주형 자료이므로 위의 방법을 사용할 수 없다. 따라서 독립성 검정에 의해 선택된 22개의 변수를 더욱 줄이면서 분류정확도를 유지하기 위한 한 가지 방법은 22개 변수를 바탕으로 추정된 Decision Tree가 종속변수의 분류에 최종적으로 사용한 변수를 다시 선택하는 것이다. Decision Tree가 분류에 사용한 변수는 <표-3>에 나타난 바와 같은 5개 변수이며 이를 설명변수로 이용한 신경망, Decision Tree, 로지스틱 회귀분석의 분류 정확도는 약 55% 로 나타났으며 자세한 결과는 <표-4>와 같다.

<그림-1> Decision Tree (CHAID)의 분석결과

<표-3> Decision tree가 선택한 변수들(Ⅰ)

종속변수(NX4 : 3개의 범주)	
설명변수명	내용
X49	사고유형
X50	사고직전속도
X56	난폭운전
X70	면허종류
X81	보호장구(2당)

Decision Tree가 분류에 사용한 5개 변수를 여타 분류기법의 설명변수로 사용할 경우, 5개 설명변수를 이용한 분류 정확도가 22개의 설명변수를 사용하는 모형의 분류 정확도와 별반 차이가 없다는 것을 알 수 있고 이는 사고 심각도를 예측하는데 필요한 항목의 수를 줄일 수 있다는 점에서 의의가 있다. 전반적으로 Decision Tree가 분류 정확성에서 다소 앞서는 것으로 나타났지만 큰 차이가 있다고 할 수 없다.

<표-4> NX4 데이터마이닝을 위한 기법별 분류 정확도

(자료1: 설명변수 22개 자료2: 설명변수 5개)

	자료1	자료2
Decision Tree	56.3%	56.1%
신경망	54.5%	55.2%
로지스틱 회귀분석	54.1%	54.0%

2. 두 개 범주 분류

이번에는 종속변수의 범주수가 분류정확도에 미치는 영향을 알아보기 위하여 종속변수 NX4의 치명적 상해와 경미한 상해를 하나의 범주로 묶어 <표-2>의 NNX4와 같이 신체상해와 물적피해로 나누어 보았다.

설명변수의 선택은 종속변수의 범주수가 3개일때와 마찬가지로 79개의 개개 변수와 사고 심각도변수간의 독립성검정을 위한 χ^2 테스트(α=0.05) 결과 가설을 기각한 23개 변수를 바탕으로 하였다. 이러한 변수들에 3가지 데이터 마이닝 기법을 이용하여 분류 정확성을 측정한 결과, 3가지 기법 모두 약 74%로 높아지는 것으로 나타났다. 그러나 이는 종속변수의 범주수를 2개로 줄인 경우, 설명변수 없이도 기대되는 분류정확도 50%에 비하면 그리 높은 편은 아니라 하겠다. 따라서 종속변수의 범주수를 줄이는 것이 실질적으로 분류정확도를 높이는데 크게 기여하지 못한 것으로 볼 수 있다.

23개 변수중 Decision Tree를 이용하여 설명변수를 선택 할 경우 <표-5>와 같이 6개 변수가 되며 이를 이용한 분류정확도는 3가지 기법이 평균적으로 73.6%로 나타났다 (<표-6> 참고).

<표-5> Decision tree가 선택한 변수들(Ⅱ)

종속변수(NNX4 : 2개의 범주)	
설명변수명	내용
X34	차도폭(1당)
X43	차체형상(1당)
X49	사고유형
X50	사고직전속도(1당)
X56	난폭운전(1당)
X81	보호장구(2당)

<표-6> NNX4의 데이터마이닝을 위한 기법별 분류 정확도

(자료3: 설명변수 23개 자료4: 설명변수 6개)

	자료3	자료4
Decision Tree	73.9%	73.5%
신경망	73.0%	73.4%
로지스틱 회귀분석	74.5%	73.7%

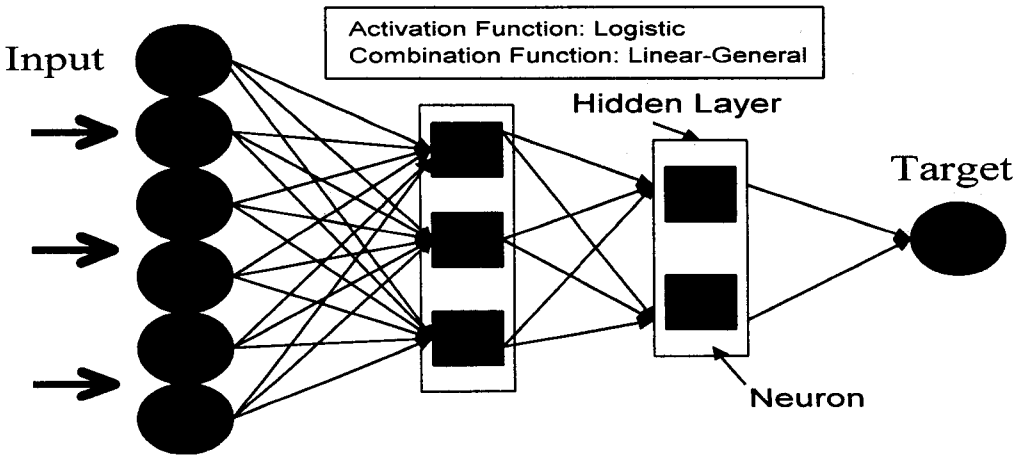
종속변수를 NNX4 하고 설명변수를 <표-5>에 나타난 변수로 하여 Decision Tree, 신경망, 로지스틱 회귀분석을 적용한 분석결과를 정리하면 다음과 같다.

<그림-1>에 나타난 Decision Tree는 CHAID를 이용한 결과로 나뭇가지 형태로 예측된 분류와 함께 나타나 있으며 진한 색 나뭇잎의 경우 통계적으로 유의한 분류 상태를 보여주고 있다. 이들 결과에 의하면 X81.보호장구가 사고 심각도를 결정하는 가장 중요한 변수로 나타났다. 보호장구를 기준으로 운전자가 '안전벨트 비착용(12)' 이거나 '헬멧 부적합 착용(21)' 일때 X49.사고유형이 차량 대 사람 사고로, 보행자가 '등지고 통행중(2)' 또는 '횡단보도 횡단중(3)' 발생했으면 그러한 사고의 99.2%가 신체상해 사고(Predict1)인 것으로 나타났다. 또한 보호장구가 위와 같은 상황에서 사고유형이 '차대차 정면충돌(22)' 또는 '차대차 회전시(27)' 사고일 때 X56.난폭운전의 상태가 '난폭운행중(2)'이면 신체상해를 유발하는 사고일 가능성이 높은 것으로 추정되었다. 반편, 사고유형이 '차량단독 표지판충돌(42)'이거나 '차량단독 담장충돌(45)'일 때 X43.차체형상이 '본네트 있음(1)'이면 물적피해 사고(Predict2)가 발생하는 것으로 나타났다. 또한 보호장구가 2점식(1) 또는 3점식(2) 착용상태에서 X50.사고직전속도가 20km/hour(2)(3) 이하이거나 '당사자불명(99)' 사고이면 물적피해 사고(Predict2)가 발생할 확률이 높은 것으로 나타났다. 이 밖에 다양한 분류 결과가 <그림-1>에 나타나 있으며 이를 이용하여 사고 심각도 분류의 근거를 찾아 갈 수 있다.

신경망 분석을 위한 망의 구성은 종속변수와 설명변수의 연결을 위해 2개의 은닉층(Hidden Layer)에 각각 3개와 2개의 개의 뉴런(neuron)을 사용했으며 활성함수(Activation function)로는 Logistic 함수를 사용하였다(<그림-2참고>). 활성함수는 Arctangent, Hyperbolic tangent등을 사용했을 때 보다 Logistic 함수를 이용했을 때의 분류 정확도가 다소 높았다.

신경망 분석이 분류결과를 설명할 수 없는 치명적 단점에도 불구하고 데이터 마이닝의 도구로써 널리 활용되는 것은 높은 분류 정확도 때문이다. 그러나 validation 자료에 적용된 본 연구의 신경망 결과는 다른 기법들에 비해 다소 떨어지는 정확도를 보였다.

<그림-2> 인공 신경망 기본구성



로지스틱 회귀분석의 장점은 모형 추정을 위한 시간이 다른 기법에 비해 짧다는

것이다. 모형이 비교적 단순하기에 분류정확성 측면에서 좀 떨어지는 면이 있다. 로지스틱 회귀분석 결과를 유의수준 5% 에서 살펴보면, 6개 변수중 4개 변수의 각 수준은 모두 유의하지 않았으며 X49.사고유형과 X50.사고직전 속도 변수의 일부 수준만이 참조 집합에 비교할 때 유의하였다. <표-7>은 로지스틱 회귀분석의 결과중 유의수준 5% 에서 유의한 결과만을 보인 것이다. X49와 X50 의 참조 집합은 각각 '차량단독 사고기타' 와 '당사자 불명'이다. 이들을 기준으로 <표-7>을 이용하여 분석하면 사고유형이 22(차대차정면충돌), 23(차대차 진행중 추돌), 24(차대차 주정차중 추돌, 27(차대차 회전시)인 경우는 치명적 사고일 확률이 높다. 반면 41(전주에 충돌), 42(표지에 충돌), 46(교량에 충돌), 47(기타 공작물에충돌), 50(주차차량에 충돌) 사고는 비교 그룹에 비하여 단순 물적 피해사고일 확률이 높다. X50.사고직전 속도는 '당사자 불명' 사고에 비하여 2(0~10km), 3(10~20km), 4(20~30km) 일 때 단순 물적피해 사고일 확률이 높다. 본 장에서 사용된 모든 데이터마이닝 분석은 SAS Enterprise Miner 소프트웨어를 사용하였다.

<표-7> 로지스틱 회귀분석의 결과

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-square	Pr > Chi-square
X49	22	1	3.2201	1.6301	3.90	0.0482
X49	23	1	3.1892	1.6163	3.89	0.0485
X49	24	1	3.2307	1.6114	4.02	0.0450
X49	27	1	3.9748	1.6679	5.68	0.0172
X49	41	1	-6.4072	1.8485	12.01	0.0005
X49	42	1	-5.6092	1.8879	8.83	0.0030
X49	46	1	-5.0121	1.9517	6.60	0.0102
X49	47	1	-5.9747	1.5807	14.29	0.0002
X49	50	1	-6.2225	1.9217	10.48	0.0012
X50	2	1	-1.0926	0.2269	23.19	0.0001
X50	3	1	-0.7734	0.2240	11.92	0.0006
X50	4	1	-0.4848	0.2230	4.73	0.0297

III. 결론

본 논문은 교통사고의 심각도 변화와 관련된 변수를 선택하고 사고 심각도 분류 예측모형을 추정하기 위하여 Decision Tree, 신경망분석, 로지스틱 회귀분석을 이용하여 분류 정확도를 비교분석하였다. 종속변수의 범주수를 3개 혹은 2개로 하고 설명변수를 줄여가며 분석한 결과 종속변수의 범주수는 데이터가 종속변수의 특정 수준에 지나치게 치우치지 않는 한 실질적으로 분류 정확도에 큰 영향을 미치지 않았으며 Decision Tree가 분류에 사용한 변수를 설명 변수로 사용하면 분류 정확도를 저해하지 않으면서 수집해야할 데이터의 범위를 줄일 수 있어 효과적인 변수선택법이 되는 것으로 나타났다. 또한 종속변수의 범주수가 3개 일때와 2개일때 모두 사고유형, 사고직전속도(1당), 보호장구(2당) 항목이 사고 심각도를 구분하는 변수로 선택되어 이들 변수가 사고 심각도와 큰 관련이 있는 것으로 나타났다.

데이터 마이닝의 3가지 기법이 모두 비슷한 분류 정확도를 보였으며 이중 종속변수 범주수가 3개 일때는 Decision Tree가, 종속변수의 범주수가 2개 일때는 로지스틱 회귀분석이 다소 높은 분류 정확도를 나타냈다.

데이터 마이닝 각 기법별 분석결과를 간단히 살펴보면 신경망의 경우, 사고 심각도에 각 설명변수가 미치는 영향 정도를 파악할 수는 없으나 주목할 사항은 설명변수 수를 줄이면 오히려 Validation 데이터의 분류 정확성이 높아지는 현상을 보였다. Decision Tree의 분석결과를 살펴보면, 운전자의 보호장구 착용상태가 안전벨트 비착용이나 헬멧 기준 부적합 착용일때 사고유형이 차대차 정면충돌 또는 차대차 회전시 사고이면서 운전자가 난폭운행중이면 90.9%의 확률로 신체

상해를 유발하는 사고가 발생되는 것을 알 수 있다. 이 밖에 차량 진행 방향에 보행자가 등지고 통행중 발생한 사고나 보행자가 횡단보도 횡단중 발생한 사고는 신체상해를 유발할 확률이 높은 것으로 나타났다. Decision Tree는 중요변수 선택, 분석 수행시간, 사고 심각도 결정요인의 유추가능 측면 등에서 교통사고 분석에 가장 적합한 기법으로 볼 수 있다. 로지스틱 회귀분석 역시 Decision Tree와 마찬가지로 사고유형과 사고직전속도 항목이 사고심각도와 유의하게 관련 있는 것으로 나타났다. 사고유형의 관점에서는 차량단독 사고에 비하여 차대차 정면 혹은 회전시 사고가 신체상해를 유발하기 쉬우며 사고직전 속도의 경우, 당사자 불명인 경우에 비하여 사고직전 속도가 10~30 km/hour 이면 단순 물적 피해 사고를 유발할 확률이 높은 것으로 나타났다. 이러한 사고를 줄이기 위하여 차량의 관점에서는, 차대차 사고가 심각한 피해를 유발함으로 중앙분리대등을 추가설치할 필요가 있겠으며 보행자의 입장에서는 횡단보도 횡단중 사고가 치명적 상해를 유발하므로 횡단보도 이용중에도 주의를 요해야 한다. 또한 보행자는 차량진행 방향에 마주보고 통행을 하는 것이 상대적으로 안전하다고 할 수 있겠다.

전반적으로 분류 정확도가 낮은 이유는 X37.도로선형 항목이 '평지'로 처리되거나 X45.용도별 항목이 '기타'로 처리된 정도가 70%~90%에 이르는 등, 일부 항목이 어느 한 수준에만 치중되고 있어 분류정확도를 저해하고 있기 때문이다. 이는 교통사고 당사자간의 민·형사상의 책임을 규명하기 위하여 작성된 교통사고 조사보고서를 바탕으로 통계원표가 작성되고 있으므로 일부항목이 활용할 수 없는 정보를 제공하고 있는데 그 원인이 있는 것으로 보인다. 본 논문에서 연구된 사고 심각도 분석외에 Decision Tree를 이용하여 인적요인 변수(연령,직업,학력, 면허경과년수,인적원인구분코드)를 사용하여 교통사고의 발생원인을 분석할 것을 향후 과제를 남겨놓고 있다.

Reference

[1]강상규(1996), "인공 신경망 모형을 이용한 우리나라 주가의 비선형적 규칙성에 관한 연구", 서 울대 석사학위 논문.

[2]경찰청(1996), 교통사고통계, pp103-115.

[3]경찰청(1996), 도로교통안전백서, pp62-76.

[4]고재성(1997), 데이터 웨어하우징과 데이터마이닝 기법을 이용한 의사결정 사례에 관한 연구", 성균관대 석사학위 논문.

[5]Berry, M. J.A. and Linoff, G.(1997), Data Mining Techniques, John Wiley & Sons, pp243-285.

[6]Bigus, J. P.(1996), Data Mining with Neural Networks, Mcgraw-Hill, pp61-97.

[7]Brokett, P.L., Cooper, W.W., Golden, L.L. and Xia, X.(1997), "A Case Study in Applying Neural Networks to Predicting Insolvency for Property and Casualty Insurers", *Journal of the Operational Research Society*, 48, pp 1153-1162.

[8]Cherkassky, V., Ghring, D., and Mulier, F.(1996), "Comparison of Adaptive Methods for Function Estimation from Sample", *IEEE Transaction on Neural Networks*, Vol. 7, No 4, pp969-984.

[9]Leshno, M. and Spector, Y.(1997), "The Effect of Training Data Set Size and the Complexity of the Separation Function on Neural Network Classification Capability: The Two-Group Case", *Naval Research Logistics*, Vol 44, pp 699-717.

[10]Michie D., Spiegelhalter, D.J. and Taylor, C.C.(1994), Machine Learning, Neural and Statiscal Classification, Ellis Horwood, pp84-96.