의사결정나무모형을 이용한 교통사고 유형 분석

김유진[1] · 최종후[2] · 이의용[3]

요 약

본 연구에서는 의사결정나무모형을 이용하여 교통사고 유형 분석을 시도한다. 분석에 이용된 자료는 도로교통안전관리공단에서 수집한 교통사고 정밀조사 자료이다. 본 연구에서 목표변수는 '사고내용'이며, 설명변수는 '인적 요인', '차량적 요인', '도로환경적 요인' 관련 변수이다. 목표변수에 주요한 기여를 하는 주요 설명변수를 도출하였으며, 얻어진 의사결정나무모형을 토대로 하여 교통사고를 유형화하였다.

주요용어 : 의사결정나무모형

1. 개요

본 연구는 도로교통안전관리공단에서 수집한 1999. 1. 1 ~ 2000. 6. 30 사이에 전국에서 발생한 인명피해를 수반한 교통사고자료를 토대로 하여 교통사고 유발요인을 찾고, 이를 토대로 하여 교통사고의 유형을 분석하는데 연구의 목적이 있다.

따라서 본 연구에서 목표변수는 '사고내용'이 되며, 이를 설명하는 설명변수는 '인적 요인', '차량적 요인', '도로환경적 요인'의 관련 변수이다. 분석을 위하여 데이터마이닝 의사결정모형을 적용하였다.

2. 분석자료

도로교통안전관리공단에서 수집한 최초의 원시자료는 '인적 요인', '차량적 요인', '도로환경적 요인'의 관련변수 212개이며, 관찰값의 수는 2,869개이다. 이를 대상으로 데이터 스크리닝 절차에서 자료의 구조탐색과 결측항목을 가진 관찰치를 제거하여 분석용 자료를 생성하였는데, 그 결과 얻어진 분석용 자료의 설명변수는 105 개이며, 관찰치는 2,310개이다.

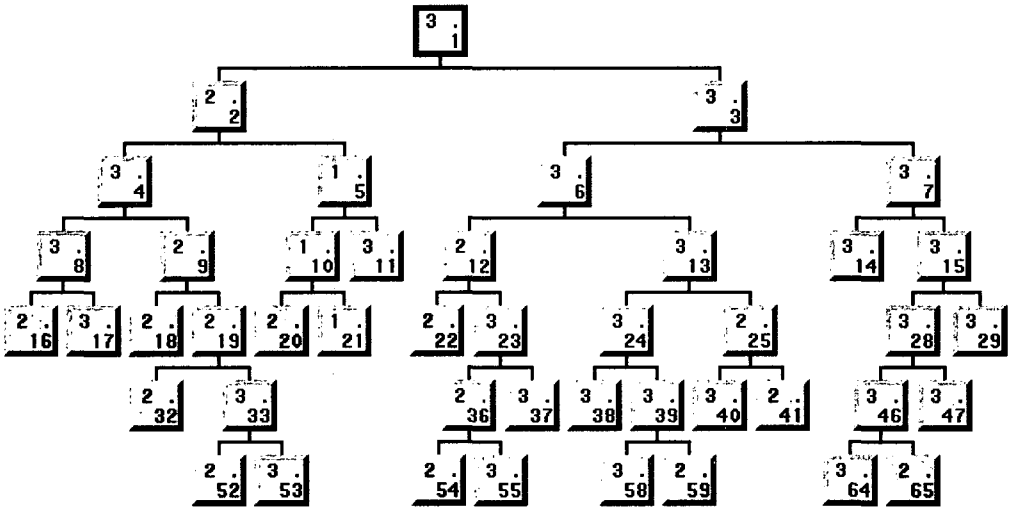
최초 원시자료의 목적변수 '사고내용'은 5개의 범주인데 제4범주와 제5범주인 중상사고와 경상사고는 사고건수가 각각 284(9.9%), 14(0.49%)에 불과하여 이를 제3범주인 부상사고와 통합하였다. 따라서 목표변수인 '사고내용' 변수는 총 3개의 범주이다. 따라서 분석용 데이터는 목표변수의 범주에 따라 범주1: 대형사고(97건, 5%), 범주2: 사망사고(631건, 27%), 범주3: 부상사고(1582건, 68%)로 재정리된다.

1) (339-700) 충남 연기군 조치원읍 고려대학교 정보통계학과. sasw@naver.com
2) (339-700) 충남 연기군 조치원읍 고려대학교 정보통계학과 부교수. jchoi@tiger.korea.ac.kr
3) (100-713) 서울특별시 중구 신당5동 171 도로교통안전관리공단. eylee20@chollian.net

3. 의사결정나무모형

<그림 1>은 분석용 자료에 대한 의사결정나무모형의 구조이며, <그림 2>는 그 일부에 대한 출력 결과이다. 의사결정나무모형은 SAS Eminer 3.0에 의한 결과이며 사용된 분리기준(Splitting Criterion)은 Entropy Reduction이다(강현철 외, 2000). 적용된 선택사양은 마지막 노드의 최소 관찰값의 수는 20, 가지치기를 위한 관찰값의 수는 60, 노드에서의 최대 가지수는 2이다.

분석결과 목적변수에 주요한 기여를 하는 설명변수는 파손정도1당, 승차인원1당, 법규위반내용1당, 교차로의 구체적위치, 면허경과년수1당, 승차인원1당, 위험인지시속도1당(사고직전), 충돌후상태1당, 도로종류 등으로 나타났다. 여기서 '1당'이라 표기된 변수는 사고시 가해자에 대한 변수임을 뜻한다.



<그림 1> 의사결정나무모형

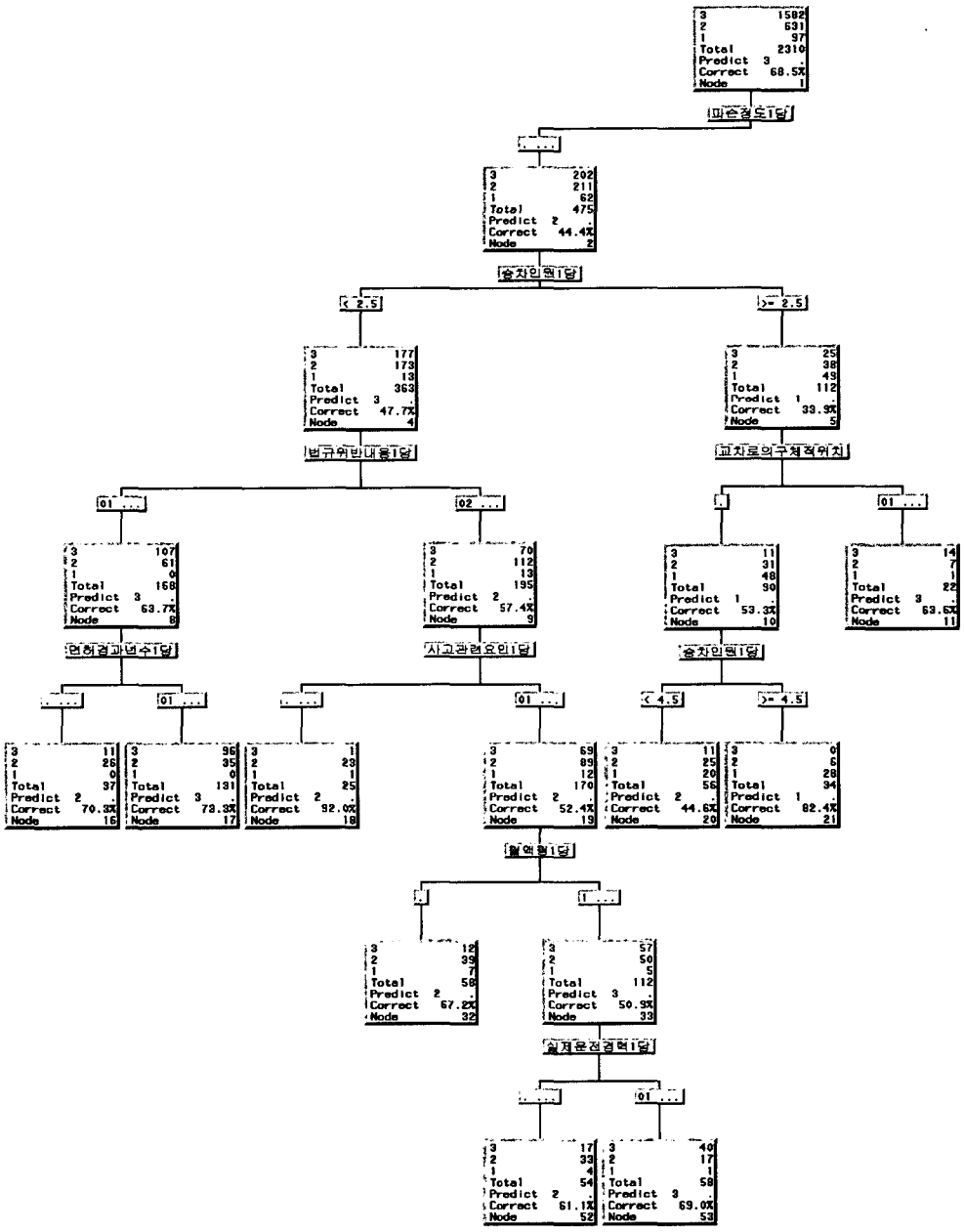
범례 : 3 . ← 노드의 판단 항목 (1: 대형사고, 2: 사망사고, 3: 부상사고)
1 ← 노드 번호

4. 오분류표

<표 1> 오분류표

		예측결과			
		부상사고(3)	사망사고(2)	대형사고(1)	전체
실제결과	부상사고(3)	1488(64%)	94(4%)	0(0%)	1582(68%)
	사망사고(2)	329(14%)	296(13%)	6(0%)	631(27%)
	대형사고(1)	1(1%)	38(2%)	28(1%)	97(4%)
	전체	1848(80%)	428(19%)	34(1%)	2310(100%)
오분류율		0.215			

<표 1>에서 보면 모형의 오분류율은 21.5% 이다.



<그림 2> 의사결정나무모형의 일부

5. 교통사고내용 패턴

분석결과를 토대로 교통사고내용을 패턴화한 결과는 <표 2>와 같다(교통사고내용 패턴에서 각 변수에 대한 자세한 설명은 김기홍 외(1999), 교통사고종합분석센터(1999)를 참고).

<표 2> 교통사고내용 패턴

패턴번호	해당 노드	패턴 내용
1	16, 17	사고시 파손정도가 '대파'인 경우로 면허경과년수가 '10년 이상'인 경우가 부상사고보다 사망사고에 더 가까움을 볼 수 있다.
2	18, 32, 52, 53	52와 53번 노드를 비교해 보면 실제운전경력이 '5년이상~15년 미만'에서 사망사고로의 분류비율이 이 범위 밖의 운전경력에 비해 더 높음을 볼 수 있다. 이 패턴에서는 주로 사망사고로 분류되는 패턴인데 주요변수의 선택이 법규위반내용, 사고관련요인 등의 변수가 중요함을 볼 수 있다.
3	11, 20, 21	교차로의 구체적 위치라는 변수에 의해 대형사고와 사망사고, 부상사고가 분류되고 있음을 볼 수 있다. 교차로의 정의는 정지선으로부터 30M 이내의 범위를 말하는 것으로 교차로와 인근에서는 사망사고 및 대형사고로 이어질 확률이 상대적으로 낮음을 보여주고 있다.
4	22, 54	파손정도는 '대파'가 아니며, 사고유형이 주로 '차대사람' 이나 도로상에서의 '차량단독' 의 사고로 분류된 패턴으로 이는 주변토지이용 변수의 '공장지대' '농장지대' '산간/무인지대'로 분류된 것과 관련이 있음을 알 수 있다.
5	37, 55	부상사고로 분류된 패턴중에서는 비교적 발생건수가 작은 패턴이나, 패턴 4와 비교해 볼 때 사고회피/예방조치1당변수에서 더 적극적인 조치를 취한 사고가 사망사고보다 부상사고로 분류되었음을 볼 수 있다.
6	38, 58, 59	패턴5와의 차이는 주변토지이용변수에서 '상가지대' '주택지대' '학교/놀이터'로 분류된 사고유형으로 사고직전속도는 30~60km로 낮은 편이며 이 때의 야간조명이 '주간' 및 '야간(밝음/양호)'의 경우가 부상사고로 분류되고, '야간(어둠/불량)'의 경우는 사망사고로 분류되고 있음을 볼 수 있다.
7	40, 41	패턴 6과 같이 주변토지이용변수에서 '상가지대' '주택지대' '학교/놀이터'로 분류된 사고유형으로 사고직전속도에서 60km 이상으로 분류되는 패턴이다. 이 경우는 총배기량이 2500~3000cc 와 10000cc 이상의 차량에 의한 사고가 사망사고로 분류되고 있고 그 이외의 차량들은 부상사고로 분류되고 있다.
8	14	도시지역 도로에서 주로 발생한 사고의 패턴으로 90%가 부상사고 이며, 파손정도도 '대파' 이외의 범위이다. 사고유형은 '차대차'와 '차량단독' 항목을 보이고 있다. 전체 데이터에서 약 31%를 차지하고 있다.
9	29, 47, 64, 65	패턴8과 비교되는 패턴으로 도로종류가 '일반국도' '지방도' '고속국도'에서 발생한 사고들로 주요하게 우선적으로 용도-세분류 변수에 의해 구분이 되어지는데 29번 노드에서 전체사고의 약 17%의 사고 건수가 일반적인 비사업용 차량들과 '사업용 컨테이너'와 '택시'에 포함되었다. 반면 사업용 '버스'들과 화물차 등은 다시 곡선반경 변수로 사망사고와 부상사고로 분류되었는데 곡선반경이 '100m~500m' 이상의 조건에서 사망사고로 분류되었다.

참고문헌

강현철 외(2000), 데이터마이닝 - 방법론 및 활용, 서울:자유아카데미.

교통사고종합분석센터(1999), 정밀조사 통계업무 전산처리 지침, 서울:도로교통안전관리공단.

김기홍 외(1999), 교통사고요인분석, 서울:도로교통안전관리공단.