한국정보전자통신기술학회논문지(jkiiect)'16-04, Vol.9 No.2
논문 16-09-02-243 http://dx.doi.org/10.17661/jkiiect.2016.9.2.243

운전자 행동자료 및 고위험군 군집 분석

김용철*

Drivers Driving Habits Data and Risk Group Cluster Analysis

Yong-Chul Kim*

요 약 본 논문은 급가속, 급 감속, 급제동, 급출발, 그리고 과속 등과 같은 여러 운행 이벤트 데이터는 운전자의 운행습관과 운전자의 사고위험성을 예측 또는 분석하는데 중요한 정보를 제공한다. 일반적인 자료의 분포는 정규분포, 로그정규분포, 감마분포 등을 이용하지만 운전자 운행습관을 나타내는 자료에서 사고위험성을 추정 할 수 있는 극단적인 부분에서는 언급한 분포로 적합하지 않은 경우가 발생한다. 특히 왜도가 발생하여 정규분포에 적합하지 않은 영역이 생겨난다. 본 논문에서는 이 영역에서 적합한 분포 함수와 사고를 유발하는 위험군을 분리 할 수 있고 운전자 운행 시 사전경고로 사고율을 줄 일수 있는 임계점을 분포함수의 quantile값과 군집분석 결과와 비교하여 제시하였다.

Abstract Driving Event Data such as the rapid acceleration, the rapid deceleration, the sudden braking, and the sudden departure, and over speeding provide important information to predict or analyze the driving habits and accident risk of a driver. Most of the data that represent the driver's driving habits generally fit to the parametric distribution, whereas extreme parts of the data to estimate the accident risk of a driver may not. This paper presents an empirical distribution that is divided into two regions, one is from the normal distribution, and the other is from the general pareto distribution for the driving habits of a driver.

Key words : Cluster analysis, Drivers driving habits, Event Recorder, Normal distribution, Pareto distribution

1. 서론

운전자 행동 분석 및 진단 시스템은 현재 미국 혹은 일본에서 많은 상용 차량들에 장착되어 있는 영상 기록 장치 혹은 이벤트 데이터 레코더 혹은 주행기록 장치로부터 얻어지는 급 가속, 급 감속, 급제동, 급출발, 그리고 과속 등과 같은 여러 이벤트 데이터를 바탕으로 운전자의 운전 성향을 파악하고 진단하는 장치 및 시스템이다. 이를 위하여 여러 센서(주로 가속 센서)의 데이터들을 수집하고 이 데이터들이 미리 정해진 임계값을 초과할 경우 이벤트로서 등록되고 분석할 수 있도록 하는 텔레매틱스 모니터링을 하고 있다. 이러한 이벤트의 등록을 위한 임계값을 추정하고 이 임계값을 초과하는 데이터들의 초과 횟수를 이용하여 운전점수를 계산 하고 이를 운전자의 행동 분석 및 진단에 사용하는데 이는 보험 텔레매틱스와 연결되어 많은 보험 회사들이 상품을 출시하고 있는 상황이다. 예를 들어 미국의 GreenRoad는 이벤트 수에 운행 시간의 가중치를 곱하여 점수화하고 있고 일본 YAZAKI METER는 급정거 (1.62g 이상) 및 급가속, 급출발 (1.51g 이상) 횟수를 산술적으로 누적 덧셈하여 점수화 한다. 이와 같은 시스템을 설계하기 위해서는 운전자 행동 자료의 분포를 추정하여야 한다.

일반적으로 대부분의 다량의 자료는 정규분포를 따른다고 한다. 그러나 운전자의 행동 자료는

* Corresponding Author : Department of Logistics and Statistical Information, YongIn University(yckim@yongin.ac.kr)
Received April 4, 2016 Revised April 14, 2016 Accepted April 21, 2016

다량의 자료임에도 불구하고 일부분에서 정규성을 갖지 못한다. 따라서 운전자 행동 자료에 적합한 분포함수를 필요로 한다. 특히, 자료의 꼬리 부분인 극단 치 값에서 정규성을 갖지 못하므로 이 부분에 대하여 분포함수의 조정이 필요하다.

본 논문에서는 자료를 이분화 하여 극단 치 값에 대하여 일반적 파레토 분포(Generalized Pareto Distribution)를 적용하고 나머지 자료에 대해서는 정규분포를 적용한 경험적 분포를 제시하고 위험군과 비위험군을 k-평균 군집분석을 이용하여 분리하였고 분리된 임계값과 파레토 분포의 quantile값과 비교하였다. 임계값은 운전자의 운행 시 위험신호 경계점으로 임계값이상으로 운전 시 운전자에게 사전 경고함으로써 사고율을 줄일 수 있다. 다음 절에서는 왜도가 발생할 때 추정에 사용되는 분포함수로써 로그정규분포, 감마분포, 와이블분포와 일반 파레토 분포함수에 대하여 서술하고 일반적인 k-군집분석에 대하여 논의 하였다. 또한, 3절에서는 운전자 행동습관 자료 중 급 가속 자료를 혼합 모형의 적용 예제로 논의하였고 사고를 발생시킬 수 있는 위험 군을 분리하고 의미 있는 임계점을제시하였다. 마지막 4절에서는 관련된 결론에 대해서 논의하였다.

2. 모형분포 및 군집분석

2.1 운전자 행동 분포 추정을 위한 분포

연속형 분포함수로써 일반적으로 알려진 정규분포를 자료가 충분히 확보된 경우에 자주 모형의 추정에 사용되어진다. 하지만 적합성 검정에 적당하지 않은 경우가 발생된다. 이러한 경우에는 대칭성을 고려하여 함수의 변환에 의한 로그정규분포 또는 왜도가 있는 감마 분포, 와이블 분포를 이용한다. 다음은 각분포의 특성과 적용범위에 대하여 언급하고자 한다.

로그정규분포는 정규분포의 함수에 로그 변환하여 척도의 변환을 줄여 대칭성을 만들어 사용하는데 적합하다. 감마분포는 함수의 형태는 왜도가 있는 경우에 적용하며 다음과 같다. $\alpha > 0,\ \beta > 0$ 모수 값에 대하여

$$f(x|\alpha,\beta)=\begin{cases}\dfrac{\beta^{\alpha}}{\Gamma(\alpha)}\ x^{\alpha-1}e^{-\beta x} & x>0\\ 0 & x\le 0\end{cases}$$

이다. (1)

와이블 분포는 실패비율에 적용되어지며 관측값이 충분히 큰 지역에서 정규분포보다 급격히 낮아지는 함수의 특성을 나타낸다. 역시 비대칭성을 가지고 있다. 함수의 형태는 다음가 같다. $\alpha > 0,\ \beta > 0$ 모수 값에 대하여

$$f(x|\alpha,\beta)=\begin{cases}\dfrac{\beta}{\alpha^{\beta}}\ x^{\beta-1}e^{-(\frac{x}{\alpha})^{\beta}} & x>0\\ 0 & x\le 0\end{cases}$$

이다. (2)

파레토 분포(Pareto Distribution)는 $x_0 > 0,\ \alpha > 0$ 모수 값에 대하여

$$f(x|\xi,\beta)=\begin{cases}\dfrac{\beta\xi^{\beta}}{x^{\beta+1}} & x\ge\xi\\ 0 & x<\xi\end{cases}$$

이다. (3)

일반화된 특히 파레토 분포 함수는 다음과 같이 표현된다.

$$G(x)=\begin{cases}1-\left(1+\xi\dfrac{x}{\beta}\right)^{-1/\xi} & \text{if } \xi\neq 0\\ 1-e^{-x/\beta} & \text{if } \xi=0\end{cases}$$

이며, (4)

ξ 는 형태 변수, β는 크기 변수이다. $\beta>0$일 때, $\xi\ge 0$이면 $x\ge 0$이며 $\xi<0$ 이면, $0\le x\le \dfrac{-\beta}{\xi}$ 인 관계를 만족한다.

충분히 큰 값 u에 대하여 초과하는 x가 y+u를 초과 할 수 없을 때의 초과 누적 분포는

$$F(y+u)=\Pr(X<y+u|X>u)=\frac{F(y+u)}{1-F(u)}$$

이고, (5)

위의 식에서 충분히 큰 u값에 대하여 $F(y+u)\rightarrow G(y+u)$이므로 식 (5)를 이용하면 $F(x)=[1-F(u)]G(x-u)+f(u)$ 이다.

그러므로 충분히 큰 수 u 에 대하여, 운전자의 습관의 극단적 위험치의 한계를 결정할 수 있는 q%-quantile을 추정하는 식은 다음과 같이 나타낼 수 있다.

q%-quantile은

$$\hat{x_q}= u+\frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{N}{N_u}(1-q)\right)^{-\hat{\xi}}-1\right)$$

이고 N은 전체관찰수, N_u는 u보다 큰 관찰수 이다.

2.2 k-평균 군집분석

본 논문에서는 위험군을 분리하고자 비계보적 군집분석으로 가장 널리 사용되어지는 k-평균 군집방법을 사용하였고 다음과 같다.

1 단계 : 자료를 초기에 k개의 군집으로 나누어서 k-개의 seed를 무작위로 선택한다.

2 단계 : 각각의 군집에 속한 자료의 평균값을 계산한다.

3 단계 : 각 자료에 대하여 평균까지의 거리를 계산한다. 계산된 결과가 자장 인접한 군집에 속하도록 알고리즘을 작성한다.

4 단계 : 일정한 기준 하에서 각 자료가 만족할 때 까지 3 단계를 반복한다.

2.3 군집분석 결과

사례 분석 자료는 하루 동안 운전자 행동분석 서비스를 제공받는 사업체의 트럭을 이용하여 4개의 이벤트 타입(급가속, 급 감속, 급 좌회전, 급우회전)에 대하여 각각 3,887,376개의 가속센서 샘플 데이터를 수집하였고 4개의 이벤트 타입 중에 속도 변위량에 대한 데이터를 분석사례로 제시하고자한다.

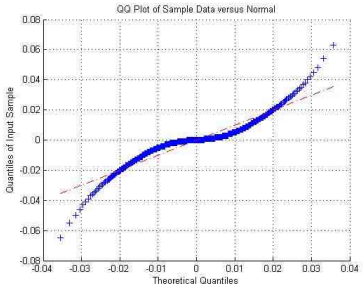

그림 1. 차량 진행방향 속도 변위량에 대한 QQ Plot
Fig. 1. speed rate QQ Plot for the vehicle traveling direction

위의 [그림 1]에서와 같이 자료의 꼬리 부분인 극단 치 값에서 정규성을 갖지 못하므로 꼬리부분에 대하여 분포함수의 조정이 필요하다. 그러므로 본 논문에서 제시한 여러분포로 적합하기 위해 자료를 이분화 하고 꼬리부분에 대하여 로그정규, 감마분포, 와이블분포, 일반적 파레토 분포를 적용하고 나머지 자료에 대해서는 정규분포를 적용하였다.

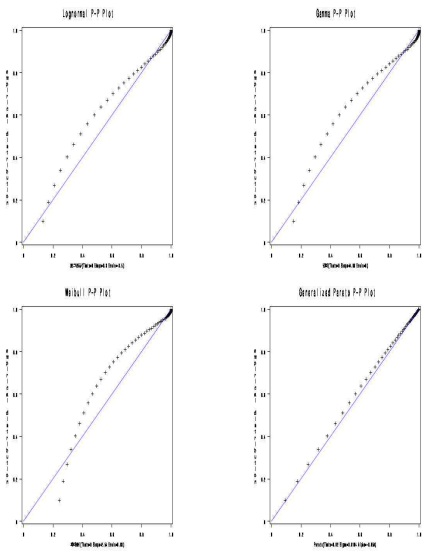

그림 2. u=0.02 km/h 보다 큰 자료에 대한 4개의 분포 pp plot
Fig. 2. pp plot for data greater than u=0.02 km/h

[그림 2]는 u=0.02 km/h 값을 기준으로 자료들을 이분화여 극 단치 값에 대하여 로그정규분포, 감마분포, 와이블 분포, 그리고 일반 파레토 분포에 대한 pp-plot을 나타낸다. 실험적 분포에 대한 각각의 분포를 적합한 결과 일반 파레토 분포함수가 다른 분포함수보다 적합성이 높다는 것을 알 수 있다. 따라서 본 논문에서 제시한 혼합 분포함수는 충분히 큰 수 u=0.02 km/h에 대하여, G(x)는 일반 파레토 분포이며 형태 변수 ξ = -0.059, 크기 변수 β= 0.0104 이다.

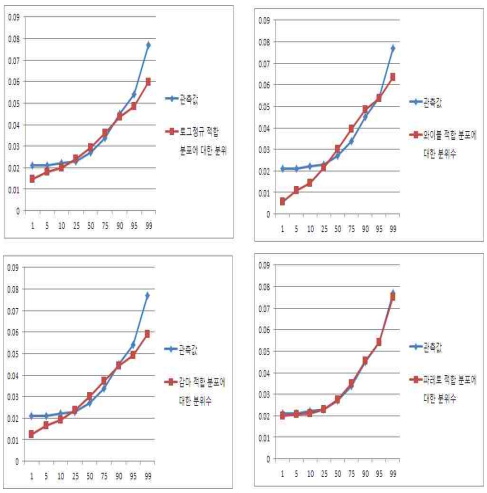

그림 3. 경험적 관측치와 비교분포의 q%-quantile 비교
[Fig. 3. compare q%-quantile with empirical observation and other distributions

관측치의 q%-quantile과 로그정규분포, 감마분포, 와이블 분포, 그리고 일반 파레토 분포에 대한 각각의 q%-quantile을 비교한 [그림 3]에서와 같이 파레토 분포가 관측값과 거의 일치함을 알 수 있다. 운전자습관의 자료를 이용하여 사고 위험군을 분리할 예측할 경우 파레토 분포의 임계점을 중심으로 분리 운전자습관에 관련된 시스템에 적용하면 효율성을 제공 할 수 있다.
자료의 파레토 분포의 q%-quantile을 표로 표시하면 다음과 같다.

표 1. 파레토 q%-quantile 추정값
Table 1. pareto q%-quantile estimate value

q%-quantile	pareto q%-quantile estimates
0.01	0.021
0.05	0.0205
0.1	0.0211
0.25	0.023
0.5	0.0274
0.75	0.035
0.9	0.0456
0.95	0.0541

합리적인 임계점을 찾기 위해서 위의 자료를 통계프로그램 SAS 10.1을 이용하여 k-평균 군집분석을 한 결과 다음과 같다.

표 2. k-평균 군집분석 결과
Table 2. k-means cluster analysis results

Cluster	Non-Risk	Risk
Frequencies	140628	16402
mean	0.027	0.058
Standard Deviation	0.0059	0.0151
Maximum	0.044	0.029
Minmum	0.021	0.045

[표 1]에서 90%-quantile 값은 0.045로 나타났고 [표 2]에서의 군집분석에서 위험 군에서의 최소가 0.045로 근사적으로 일치됨을 알 수 있고 95%-quantile 값은 0.0541이고 위험 군에서의 평균값이 0.058로 유사하게 나타났다. 또한 사고를 유발 할 수 있는 임계점을 90%-quantile이나 95%-quantile 로 추정하여 운전자 운행 시 사전 경고를 한다면 사고율을 줄일 수 있을 것이다.

4. 결론

운전자의 급가속, 급 감속, 급제동, 급출발, 그리고 과속 등과 같은 여러 이벤트 데이터는 운전자 운행습관을 나타내는데 이 자료의 대부분은 모수적 분포함수에 적합하지만 운전자의 사고위험성을 추정 할 수 있는 극단적인 부분은 적합하지 않은 경우가 발생한다. 본 논문에서는 운전자 운

행 습관과 관련한 자료를 두 개의 영역으로 나누어서 경험적 분포함수를 제시하였다. 이 분포는 정규분포와 일반 파레토분포를 u값을 기준으로 구분하여 자료를 이분화 하여 적용한 혼합분포이며 운전자의 운행에 관련된 자료의 극단 치 임계값을 추정 할 수 있도록 하고 관련된 통계 변수들을 바탕으로 운전자의 운전 위험도를 산출할 수 있다. 운전자의 운행관련 빅 자료가 수집되어 운전습관에 대한 경고 시스템에 결과를 접목하면 사고율을 줄이고 시스템의 효율성을 향상할 수 있을 것이다. 본 논문의 분석은 자료의 제한점 즉 운행 시 부가적인 자료 조건인 도로조건(아스파트 도로, 콘크리트 도로, 자갈길, 흙길), 날씨조건(정상, 비, 눈, 강풍), 그리고 도로경사도(오르막길, 정상, 내리막길)등의 자료를 확보 할 수 없는 제한된 분석결과이다. 부가적인 자료가 확보된다면 사고율 임계값의 예측은 보다 추정의 정도를 높힐 수 있을 것이다.

REFERENCES

[1] Yiilbyeong, imheonyeon, "A Study on traffic accidents prediction model developed in Korea," Korean Society of Transportation, Vol. 8, No. 1, pp. 73-88, 1990.

[2]Oeo, Gee Young, Do-Gyeong Kim, and Yuhwa Lee. "The Characteristics of Secondary Crashes Occurred on Expressways in Korea." International Journal of Highway Engineering 15.2 (2013): 139-147.

[3] Jeongilyoung, "traffic accident types and causes safety breach Analysis", pp. 1-107, Transportation Safety Authority, 2012.

[4] Sonsoyoung, "traffic accident statistics analysis study," Statistical Analysis, Vol. 2, No. 2, pp. 181-201, 1997.

[5] Johnson, R.A and Wichern, D.W., "Applied Multivariate Statistical Analysis", <Prentice Hall>, pp. 1-400. 1992.

[6] Masaru Ueyama and Hideo Chikasue and Kizuki Muramatu, RELATIONSHIP BETWEEN DRIVING BEHAVIOR AND TRAFFIC ACCIDENTS -ACCIDENT DATA RECORDER AND DRIVING MONITOR RECORDER", Paper Number 98-S2-O-06), 1998.

[7] Schmidt-Cotta, R R, "Vehicle Event Recording based on Intelligent Crash Assessment: VERONICA - II", pp. 1-203, 2009.

[8] Shanker, V. and Mannering, F., "An Exploratory Multiomial Logit Analysis of Single- Vehicle Motorcycle Accident Severity", Journal of Safety Research, Vol. 27. No.3. pp.183-194, 1996.

[9] Yasuhiro Yamai and Toshinao Yoshiba, "Comparitive analyses of expected shortfall and value-at-risk under market stress", IMES Discussion Paper No 2002-E-2,Bank of Japan, 2002.

[10] Yay M., Madrid N.M., Ramirez J.A.O., "Using an improved rule match algorithm in expert system to detect broken driving rules for an e4nergy-efficiency and safety relevant driving system", Procedia Coumputer Science 35, pp. 127-136, 2014.

저자약력

김 용 철(Yong-Chul Kim) **[회원]**

- 1985년 2월 : 경희대학교 수학과(학사)
- 1994년 6월 : 미국 미주리 주립대 통계학과 (통계학박사)
- 1995년 3월 ~ 1996년 2월 : 통계청 사무관
- 1996년 3월 ~ 현재 : 용인대학교 물류통계정보학과 교수

<관심분야> 산학융합, IT 융합기술