◎ 논 문

http://dx.doi.org/10.22680/kasa2022.14.1.026

VIMS와 DTG 데이터를 이용한 창원시 시내버스 머신러닝 분석 연구

박지양* · 정재환** · 윤진수*** · 김성철**** · 김지연*****
이호상****** · 류익희******* · 권영문********

A Study on the Analysis of Bus Machine Learning in Changwon City Using VIMS and DTG Data

Jiyang Park*, Jaehwan Jeong**, Jinsu Yoon***, Sungchul Kim****, Jiyeon Kim*****,
Hosang Lee******, Ikhui Ryu*******, Yeongmun Gwon********

Key Words: VIMS(자동차검사관리시스템), DTG(운행기록데이터), Machine Learning(기계학습), Citybus(시내버스), Correlation Analysis(상관성 분석)

ABSTRACT

Changwon City has the second highest accident rate with 79.6 according to the city bus accident rate. In fact, 250,000 people use the city bus a day in Changwon, The number of accidents is increasing gradually. In addition, a recent fire accident occurred in the engine room of a city bus (CNG) in Changwon, which has gradually expanded the public's anxiety. In the case of business vehicles, the government conducts inspections with a short inspection cycle for the purpose of periodic safety inspections, etc., but it is not in the monitoring stage. In the case of city buses, the operation records are monitored using Digital Tacho Graph (DTG). As such, driving records, methods, etc. are continuously monitored, but inspections are conducted every six months to ascertain the safety and performance of automobiles. It is difficult to identify real-time information on automobile safety. Therefore, in this study, individual automobile management solutions are presented through machine learning techniques of inspection results based on driving records or habits by linking DTG data and Vehicle Inspection Management System (VIMS) data for city buses in Changwon from 2019 to 2020.

1. 서 론

국내적으로 시내버스사고는 매우 많이 일어나고 있다. 특히 경상남도 창원시에는 Fig. 1과 같이 사고율이 무려 79.6%로 전국 지역별 시내버스 사고 중 2위를 차지하고 있다.(1) 실제로 창원시 시내버스 이용자는 하루 25만 명이 이용하고 있고 연도별 사고도 점차적으로 증가 하고 있다.(2) 최근에는 창원시 시내버스 엔진룸에서 화재도 발생하고 전기버스에서도 화재가 발생했다.(3) 이에 따라

* 한국교통안전공단 첨단안전연구처, 과장
** 한국교통안전공단 첨단안전연구처, 팀장
*** 한국교통안전공단 빅데이터센터, 선임연구원
**** THE IMC, 소장
***** THE IMC, 선임연구원
****** 한국교통안전공단 첨단안전연구처, 처장
******* 한국교통안전공단 자동차검사본부, 본부장
******** 한국교통안전공단 첨단안전연구처, 과장
E-mail: pjy2049@kotsa.or.kr

자동차안전학회지: 제14권, 제1호, pp. 26~31, 2022
논문접수일: 2021.5.31, 게재확정일: 2021.12.20

국민의 불안감이 점차적으로 증가하는 것을 확인 할 수 있다.

시내버스에 대한 운수업체에서도 지속적인 관리 등 점검을 하고 있으나, 아직 많이 부족한 실정이다.

따라서 버스 등 자동차 검사 부적합 및 결함여부를 사전에 미리 발견하고 사고가 일어나기 전에 미리 점검을 받도록 하는 시스템 개발이 필요 하다.

미국과 영국 등은 전체 교통사고의 2%가 자동차 결함으로 발생하는 것으로 나타난다.(4) 전체에서 차지하는 비중은 낮지만, 불특정 다수를 수송하는 대중교통의 경우 대형사고로 이어질 가능이 매우 높다.

차량결함 여부 판단을 위한 사고기록장치(EDR)의 도입 여부 등을 검토하고 있으나 이는 사후 대처방안으로 사고 예방을 하기 위한 사전 대처방안 마련이 필요하다.

본 연구는 경상남도 창원시 시내버스 중 일부업체에 대해서 자동차 안전과 관련이 있는2019년부터 2020년 검사결과데이터(Vehicle Inspection Management System, VIMS)5)와 자동차 운행 습관 등의 정보를 제공할 수 있는 운행기록데이터(Digital Tacho Graph, DTG)의 데이터(6)를 표준화 하고 이에 따른 예측에 우수한 랜덤포레스트 알고리즘을 이용하여 머신러닝 분석을 진행할 것이다.

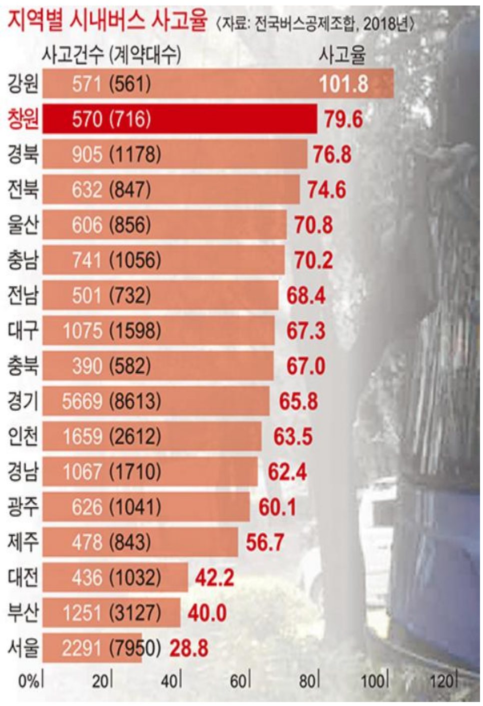


Fig. 1 Local city bus accident rate

2. 데이터 보유항목

2.1. VIMS

VIMS 자료의 경우 매우 많은 자동차에 대한 정보를 가지고 있으나, 대표적으로 전조등, 배출가스, ABS 측정결과, 육안검사, 사진정보 등이 있다.

Fig. 2(7)와 같이 자동차 검사를 받을 경우 얻게되는 검사결과표에 해당하는 내용은 전부 데이터화 되고 있고, 보유 관리를 하고 있다.

또한 Fig. 3과 같이 자동차에 대한 기본적인 제원정보 또한 관리를 하고 있다.

2200006 / 2021-03-20 14 : 42

접수번호 : 4440-20200219-0074-1 (자동차소유자용)

자동차종합검사 결과표

① 자동차등록번호	66서6147	② 성명(상호)	박**	③ 생년월일	-
④ 차 명	SM3	⑤ 차종(용도)	승용자일반형중형 자가용	⑥ 제작연식	2008
⑦ 연 료 종 류	휘발유	⑧ 누적주행거리	127,966	⑨ 최초등록일	2008-02-26

검 사 결 과

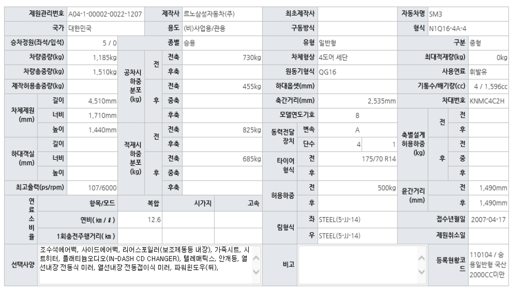
⑩ 종합판정	적합	관능·기능검사	적합	안전검사	적합	배출가스정밀검사	적합
⑪ 검사유효기간	2020-02-26 ~ 2022-02-24			⑫ 재검사기간	~		

⑬ 세부검사내용

1부(관능 및 기능검사) / 2부(기계/기구검사)

검사항목			기준값	측정값	판정
⑭사이드슬립			1m 주행에 5mm 이내	(OUT)0.1 mm/m	적합
⑮ 배출가스	휘발유 가스	일산화탄소(CO):	0.62 / 1.0%	0.0 / 0.0	적합 / 적합
		탄화수소(HC)	49.0 / 120.0ppm	3.0 / 1.0	적합 / 적합
		질소산화물(NOx)	350.0ppm	0.0	적합
		공기과잉률(λ):	1 ± 0.1	1.01	적합

Fig. 2 Vehicle inspection results



제원관리번호	A04-1-00002-0022-1207	제작사	르노삼성자동차(주)			최초제작사			자동차명	SM3	
국가	대한민국	용도	(비)사업용/관용			구동방식			형식	N1Q16-4A-4	
승차정원(좌석/입석)	5 / 0	종별	승용			유형	일반형		구분	중형	
차량중량(kg)	1,185kg	공차시 하중분포(kg)	전	전축	730kg	차체형상	4도어 세단		최대적재량(kg)	0kg	
차량총중량(kg)	1,510kg		전	후축		원동기형식	QG16		사용연료	휘발유	
제작허용총중량(kg)			후	전축	455kg	하대옵셋(mm)			기통수/배기량(cc)	4 / 1,596cc	
차체제원(mm) 길이	4,510mm		후	중축		축간거리(mm)	2,535mm		차대번호	KNMC4C2H	
차체제원(mm) 너비	1,710mm		후	후축		모델연도기호	8		축별설계 허용하중(kg)	전	전
차체제원(mm) 높이	1,440mm	적재시 하중분포(kg)	전	전축	825kg	동력전달장치 변속	A			전	후
하대객실(mm) 길이			전	후축		동력전달장치 단수	4	1		후	전
하대객실(mm) 너비			후	전축	685kg	타이어형식 전	175/70 R14			후	중
하대객실(mm) 높이			후	중축		타이어형식 후				후	후
최고출력(ps/rpm)	107/6000		후	후축		허용하중 전	500kg		윤간거리(mm)	전	1,490mm
연료소비율 항목/모드		복합	시가지	고속		허용하중 후			윤간거리(mm)	후	1,490mm
연료소비율 연비(㎞/ℓ)		12.6				림형식 좌	STEEL(5-JJ-14)		접수년월일	2007-04-17	
연료소비율 1회충전주행거리(㎞)						림형식 우	STEEL(5-JJ-14)		제원취소일		
선택사양	조수석에어백, 사이드에어백, 리어스포일러(보조제동등 내장), 가죽시트, 시트히터, 플래티늄오디오(IN-DASH CD CHANGER), 텔레매틱스, 안개등, 열선내장 전동식 미러, 열선내장 전동접이식 미러, 파워윈도우(뒤),					비고			등록현황코드	110104 / 승용일반형 국산 2000CC미만	

Fig. 3 Vehicle specifications information

2.2. DTG

DTG데이터의 경우 초단위와 트립단위로서 사업용차량에 대한 위험운전행동 등을 관리 하기 위해 자료를 Table 1과 같이 보유하고 있다. Table 1은 초단위 위험운전행동 데이터로서 대표적으로 급가감속, 과속 및 GPS 등을 통해 실시간의 데이터를 수집하고 있다.

Table 1 Digital Tacho Graph Table

Column	Information
INDEX	Number
Date	Date
Company Code	Company Code
Vehicle Number	Vehicle Number
Trip	Travel distance
Operating time	Vehicle operating time
GPS X	Location information
GPS Y	Location information
Driver code	Vehicle driver
Drive Velocity	Vehicle Velocity
Speeding 20KM(case)	Provision of statistical data based on risk driving behavior analysis standards
Speeding 20KM(time)	
Speeding 40KM(case)	
Speeding 40KM(time)	
Speeding 60KM(case)	
Speeding 60KM(time)	
Long term speeding(case)	
Long term speeding(time)	
Rapid acceleration	
Rapid start	
Rapid decline	
Rapid stop	
Rapid left turn	
Rapid right turn	
Rapid U turn	
Rapid change of course	
Rapid overtaking	
Date and time of data	
Creation date and time	

3. 데이터 수집 및 표준화

3.1. 데이터수집

창원시로부터 시내버스를 운영하는 업체는 총 10개의 운수업체가 있으며 총 시내버스 대수는 758대이다. 이중 모집단 수를 어느 정도 보유하고 있고 버스의 종류가 일정한 버스회사 하나를 선정하였다. 이 운수업체의 경우 총 82대의 시내버스를 보유하고 있으며, 주기적인 관리를 하고 있는 버스 업체이다.

이 버스업체에 대한 Table 2와 같이 한국교통안전공단에서 보유중인 2019년-2020년 일부 운행기록데이터 및 차량검사결과 데이터를 활용하여 사전 분석하고 모델링을 시행했다.

Table 2 DTG and VIMS data collection tables

Data	Information	Period	Data number
DTG data	Date	2019.03.01. ~ 2020.12.31.	119,185
	Company Code		
	Vehicle Number		
	Mileage		
	Average velocity		
	Braking count		
	MAX RPM		
	Average RPM		
	Rapid acceleration		
	Rapid decline		
	Rapid line change		
VIMS dats	Inspection date	2019.01.03. ~ 2020.11.14	131
	Vehicle information		
	Braking test result		
	Headlight test result		
	Visual test result		
	Result		

3.2. 데이터 표준화

운행기록 데이터와 차량검사 데이터 셋으로는 트립별 운행기록에 대한 차량상태를 평가할 수 있는 자동차통신(CAN)데이터가 없어 현재는 차량의 최종 판정결과를 각 트립별 결과로 대체 후 분석을 진행하였다.

운행기록데이터(DTG)와 차량검사결과(VIMS) 데이터를 통합하기 위해 연결키로 자동차등록번호를 통해 연결하였고 인과관계를 고려해 운행기록데이터가 차량검사결과 이전 일 까지만 고려하여 데이터를 통합하였다.

차량별 여러 개의 트립 단위로 데이터가 수집되고 있으며, 최종 차량검사결과의 판정결과를 각 트립의 목표 값으

로 설정하였다.

인과관계의 타당성을 확인하기 위해 원인이 되는 운행기록데이터와 결과가 되는 차량검사 결과 데이터의 인과성을 회귀분석을 통해 확인했다.

모델의 타당성 여부를 확인하기 위해 회귀모델의 P-value값이 0.05보다 작을 경우 유의하다고 판단했을 때 두 데이터 셋 간의 인과관계가 성립하는 것을 Table 3과 같이 확인 할 수 있다.

Table 3 Causal relationship of DTG and VIMS

Result(y)	Braking force (front)	Braking force (rear)	Total brake	Parking brake	Speedo meter
DTG(x)	TRIP distance, average RPM, idling count, rapid start, etc				
F-score	412.6	109.3	128.1	683.9	151.7
Degree of freedom	56589	56589	56589	56589	56589
P-value	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16
Validity	Reasonable	Reasonable	Reasonable	Reasonable	Reasonable

3.3. 정제 및 전처리

결측치/이상값 보정을 위해 운행기록데이터와 차량검사결과 데이터를 연결 시 어느 한쪽 데이터에서만 존재하는 차량번호의 경우 결측값으로 보고 해당 차량번호를 갖는 행을 분석에서 모드 제외 하였다.

정규화를 위해 운행기록데이터의 경우 수집된 단위에 따라 데이터 분포 간의 차이가 많이 났으며, 데이터 범위를 0~1사이로 정규화(Normalization) 하였다.

중복제거를 위해 운행기록데이터 내 총 119,185개 중 10,364개의 중복을 제거한 108,821개의 데이터를 분석에 사용하였다.

4. 학습데이터셋 구축 및 예측모델구축

4.1. 학습데이터셋 구축

전체 데이터 셋을 랜덤으로 7:3으로 분할하여 70%를 학습데이터로 사용하였고, 변수 간 다중공선성이 존재하는 14개(TRIP운행시간, 운행 중 정지시간, 공회전시간, 과속시간(20, 25, 30, 35, 40, 45, 50, 55, 60Km)초과, 장기과속시간, 중립기어시간)의 변수에 대해서 분석은 제외 하였다.

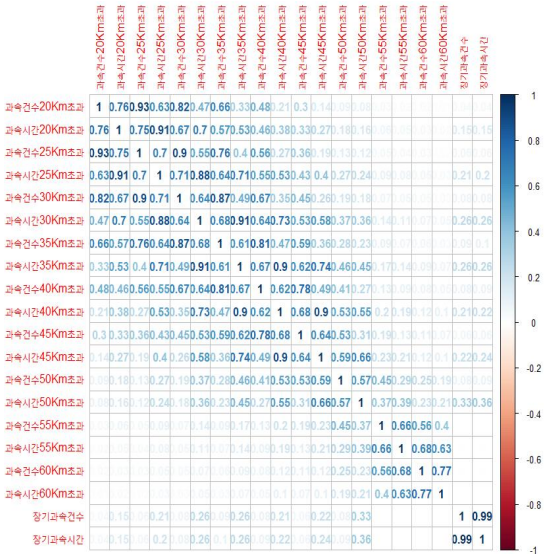


Fig. 4 Check for multicollinearity between variablesl

4.2. 예측모델 구축 및 모델 성능평가

4.2.1. 랜덤포레스트[8]

랜덤포레스트(RF)는 무작위로 선택된 데이터 하부집합(subsets)과 자질집합(feature stes)으로 학습시킨 의사결정 트리에 기반한 대포적인 앙상블 분류 알고리즘으로 Breiman이 개발하였다. 랜덤포레스트에서는 각각의 노드를 나타낼 때 설명변수를 무작위로 선택하고 선택된 설명변수의 집합 중에서 가장 최적의 결과를 내는 방법을 이용하며 Fig. 5와 같다.[9]

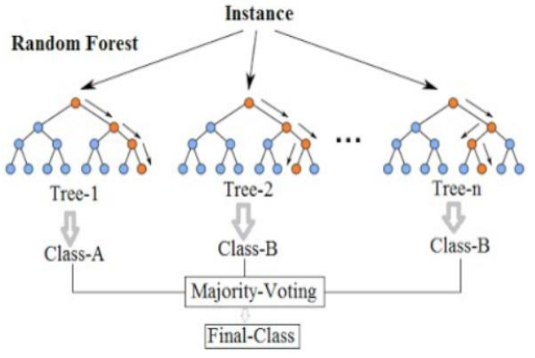


Fig. 5 Random Forest Simplified

4.2.2. 예측모델

예측모델 구축을 위해 기계학습(머신러닝)을 활용하였고 예측에 우수한 랜덤포레스트 알고리즘을 적용하였다.

학습데이터에서 30%인 Test 데이터 셋으로 모델 검증 성능평가를 진행하였고 최종 분류 결과는 Table 4와 같다.

Table 4 Build a prediction model

Actual / Predicted	Unsuitable	Conformity	Recommendation for correction
Unsuitable	1,808	163	142
Conformity	635	8,165	1,104
Recommendation for correction	142	1,104	4,073

4.2.3. 모델성능평가[10]

최종 모델의 성능평가를 위해 분류 문제의 4가지 성능평가 지표인 정확도, 재현율, 정밀도, F1-score를 이용하여 모델의 성능평가 결과를 확인하였다.

Table 5는 4가지 성능평가 지표에 대한 내용을 의미하며 식 (1), 식 (2), 식 (3)은 각각 지표에 대한 계산식이다.

$$\text{정확도}(Accuracy) = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{정밀도}(\text{Pr}ecision) = \frac{TP}{TP+FP} \quad (2)$$

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

F1-score는 정밀도와 재현율을 하나로 요약한 값이며 식 (4)와 같이 조화평균의 방법을 이용하여 계산한다.

Table 5 Meaning of performance evaluation indicators

Mean	Content
TN (True Negative)	Right actual, predicted value Negative
FN (False Negative)	Wrong actual predicted value Negative
TP (True Positive)	Right predicted predicted value Positive
FP (False Positive)	Wrong predicted predicted value Positive

$$F1-Score = 2\times\frac{\text{정밀도}\times\text{재현율}}{\text{정밀도}+\text{재현율}} \quad (4)$$

다음 식을 이용하여 계산을 하면 Table 6과 같이나오며 성능평가 결과 75%의 정확도를 구현하는 것을 확인할 수 있다.

Table 6 performance evaluation results

Accuracy	Precision	Reproduction rate	F1-score
0.8269	0.8557	0.6825	0.7593

5. 결 론

본 연구는 창원시 시내버스 회사 중 82대를 보유하고 있는 특정 운수업체에 대해 2019~2020년의 기간 동안 자동차검사결과와 운행기록데이터의 데이터를 가지고 머신러닝기법을 통해 분석을 해보았다.

본 연구의 결과를 요약하면 다음과 같다.

1) 대표적인 두개의 데이터의 표준화를 위해 인과관계의 원인이 되는 운행기록데이터와 결과가 되는 검사결과데이터의 회기분석 결과 P-value 값이 2.2e-16으로 인과관계가 성립함을 확인할 수 있었다.
2) 운행기록데이터와 검사결과데이터 중 한쪽데이터에서만 존재하는 차량번호의 경우 결측값으로 사용하였으며, 중복제거 등을 통해 총 108,821개의 데이터에 대해 7:3으로 분할하여 70%를 학습데이터로 사용하였고, 30%를 예측모델로 구성하였다.
3) 예측 또는 실험의 지표로서 사용되는 대표적인 F1-score를 사용하여 모델에 대한 성능평가 결과 75%를 구현하였으며, 각 지표별 정확도 82%, 정밀도 85%, 재현율 68% 확인 할 수 있다.
4) 2019년도의 검사결과의 바탕으로 2020년의 검사결과를 예측해보았을 때 75%의 정확도를 가지는 모델을 개발 하였으며, 집단 수 그리고 실시간 데이터의 추가를 통하여 미래의 검사결과도 예측이 가능할 것으로 보인다.
5) 트립데이터인 운행기록 데이터에 대한 차량상태를 평가 할 수 있는 데이터가 없어 본 연구에서는 차량의 최종판정 결과를 각 트립별 결과로 대체 후 분석하였다. 추 후 CAN데이터를 확보하여 현재의 1단계 알고리즘이 아닌 2단계 알고리즘을 통

해 차량상태를 예측할 수 있는 모델이 제시될 것이라 판단된다.

6) 또한 현재는 82대의 시내버스에 대한 평가를 진행하였으나, 창원시 전체에 대한 758대의 평가를 넘어서 전국의 시내버스에 대한 분석을 통해 버스차량상태 예측을 할 수 있을 것이라 판단된다.

참고문헌

(1) 류민기, 2019, "'쾅'끊이지 않는 사고".
(2) 창원시 버스정보시스템 BIS.
(3) 최호영, 2021, "운전기사 덕분에...승객 내린 시내버스에 불 치솟아".
(4) www.nhtsa.gov
(5) Vehicle Inspection Management System Manual.
(6) Digital Tacho Graph Manual.
(7) 자동차 관리법.
(8) Kwon, A., 2013, "Variable selection using Random Forest", unpublished master's thesis, Inha University.
(9) Breiman, L., 2002, "Random forests. Machine Learning", 45(1), 5~32.
(10) Davide Chicco, Giuseppe Jurman, 2020, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation", 21(1), ISSN 1471~2164.