OBDII 데이터 기반의 회귀 분석을 통한 실시간 연료 소비량 예측

양희은*, 김도현**
*단국대학교 EduAI센터
**성균관대학교 데이터사이언스융합학과
yanghe@skku.edu, kimtj@me.com

Realtime Fuel Consumption Prediction using In-Vehicle Data from OBDII and Regression Methods

Hee-Eun Yang*, Do-Hyun Kim**
*EduAI Center, Dan-Kook University
**Dept. of Applied Data-science, Sungkyunkwan University

요 약

자율주행 차량이 많아지고 차량의 ECU가 고도화되면서 정확한 차량의 데이터를 획득하고 분석하여 활용하는 것이 중요해지고 있다. 현재에는 내연 기관 차량의 ECU 데이터를 얻기 위해서 OBDII 포트(규격)에 기반한 CAN통신을 주로 이용하고 있다. 하지만 OBDII 규격을 통해서 연비와 같은 중요한 차량 정보를 얻는 경우, 변환식(MAF 센서(흡입 공기량 센서)와 공기/연료 비율을 이용)의 오차 범위가 커서 데이터의 정확도가 낮다. 본 연구에서는 머신 러닝 기법 중에 하나인 회귀 기법을 통해서 기존의 계산보다 더 정확한 연비를 구할 수 있는 모델을 개발하였다. 이러한 모델 개발을 통하여 차량의 RAW 데이터를 기반으로 필요한 차량 데이터를 정확하게 구할 수 있게 되었으며 20회가 넘는 실 도로주행을 통해서 본 모델의 정확도를 검증하였다.

1. 서론

자율주행 차량이 점점 증가하면서 차량에서 정확한 데이터를 획득하고 이를 알맞게 분석하는 중요성이 대두되고 있다.[1] 현재는 자동차 제조사를 제외하고 내연기관 차량의 데이터를 얻기 위해서는 OBDII 규격을 활용하고 있다. 이러한 OBDII의 규격에는 표준 항목과 비표준 항목이 있는데 표준 항목은 법으로 명시되어 있어 비교적 획득하기 간단한 방면에 비표준 항목은 완성차 제조업체와 ECU 제조사에 따라서 일관된 통일성이 없어서 다양한 차량의 데이터를 상대적으로 얻기 힘든 실정이다.

이러한 비표준 항목 중에는 연비, 즉, 연료 소모량(순간 연료 분사량)변수도 속해 있다. 현재에는 연료 소모량을 예측하기 위해서 주로 MAF(Mass Air Flow)센서의 데이터를 이용하여 연비를 계산하는 방식[2]을 주로 사용하고 있다. 하지만 이러한 접근으로 연비를 계산하기 위해서는 공기/연료 비율(air/fuel ratio)을 계산해야 하며[3], 해당 비율은 각 엔진/제조사의 ECU 세팅에 따라서 차이가 있을 수 있어 정확한 계산이 어렵다. (일반적으로 가솔린 14.7, 디젤 14.6, 하이드로젠 34 등과 같은 비율로 계산)

본 연구에서는 MAF를 포함한 OBDII의 표준항목 데이터들을 이용하여 기존보다 정확하게 차량의 연료소모량을 예측하고자 한다. 이를 위해서 차량의 데이터를 추출 할 수 있는 OBDII 젠더와 단말기, 그리고 펌웨어를 제작하였다. 이를 바탕으로 데이터를 획득하고 가공한 데이터를 기반으로 순간 연료 분사량 변수를 통해 정확하게 연비를 예측하는 모델을 제시[4]하고자 한다.

2. 연구 방법

2.1 차량의 데이터 획득

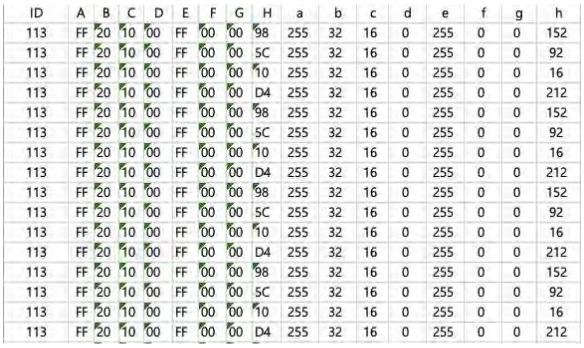
차량의 RAW 데이터를 얻기 위하여 자체적으로 OBDII 하드웨어 및 펌웨어를 제작하였다. OBDII 하드웨어는 ARM기반의 32비트 CPU와 데이터 저장을 위한 4GB의 롬 메모리가 있으며, 실시간 데이터를 확인할 수 있도록 3G/LTE 모뎀과 블루투스칩을 탑재하였다.

다양한 OBDII 프로토콜 중에서 SAEJ1850과 ISO 15765를 지원하며 현재에도 다양한 차량의 규격과 요구에 맞도록 지원 프로토콜 및 차종을 늘려나가고 있다.

(그림 1) OBDII Device(Left) / OBDII Gender PCB(Right)

2.2 획득한 데이터 파악 (리버스엔지니어링)

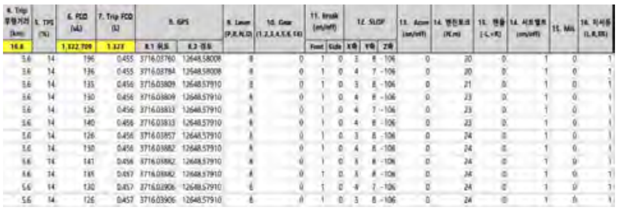
OBDII 단말기를 통하여 데이터를 얻는다고 해서 바로 모델링에 적용할 수 있는 것은 아니다. RAW 바이너리 데이터를 분석하여 각 데이터가 의미하는 수치를 파악해야 한다. 제작한 하드웨어에서 데이터를 기록하면 (그림 2)와 같이 각 데이터 인덱스의 바이너리 값을 얻을 수 있다.



ID	A	B	C	D	E	F	G	H	a	b	c	d	e	f	g	h
113	FF	20	10	00	FF	00	00	98	255	32	16	0	255	0	0	152
113	FF	20	10	00	FF	00	00	5C	255	32	16	0	255	0	0	92
113	FF	20	10	00	FF	00	00	10	255	32	16	0	255	0	0	16
113	FF	20	10	00	FF	00	00	D4	255	32	16	0	255	0	0	212
113	FF	20	10	00	FF	00	00	98	255	32	16	0	255	0	0	152
113	FF	20	10	00	FF	00	00	5C	255	32	16	0	255	0	0	92
113	FF	20	10	00	FF	00	00	10	255	32	16	0	255	0	0	16
113	FF	20	10	00	FF	00	00	D4	255	32	16	0	255	0	0	212
113	FF	20	10	00	FF	00	00	98	255	32	16	0	255	0	0	152
113	FF	20	10	00	FF	00	00	5C	255	32	16	0	255	0	0	92
113	FF	20	10	00	FF	00	00	10	255	32	16	0	255	0	0	16
113	FF	20	10	00	FF	00	00	D4	255	32	16	0	255	0	0	212
113	FF	20	10	00	FF	00	00	98	255	32	16	0	255	0	0	152
113	FF	20	10	00	FF	00	00	5C	255	32	16	0	255	0	0	92
113	FF	20	10	00	FF	00	00	10	255	32	16	0	255	0	0	16
113	FF	20	10	00	FF	00	00	D4	255	32	16	0	255	0	0	212

(그림 2) Binary Data from CAN Communication

OBDII에서 얻은 바이너리 데이터를 리버스 엔지니어링하여 각 항목들이 의미하는 변수를 역으로 알아내는 과정이 필요하다. 각 인덱스가 의미하는 데이터를 차량 직접 주행 및 기능 동작을 통해 찾아내었다. 그림 3은 바이너리 데이터를 변수에 맞춰 변환한 값이다.



(그림 3) Reverse Engineered Feature Data

또한 GPS 수신기를 탑재하여 위/경도를 포함하였고, 자이로센서를 포함하여 X,Y,Z 축의 기울기 정보도 포함하였다.

3. 데이터 가공 및 실험방법

3.1 데이터 범위

본 논문에서는 제작한 소프트웨어를 이용하여 2019년 1월부터 2019년 2월의 주행 데이터를 수집하였다. 전체 데이터 53,580건(초당 Data Set)에서 80%에 해당하는 데이터를 학습데이터(Training data), 20%에 해당하는 데이터를 테스트데이터(Validation/Test Data)로 사용하였다.

	Start date	End date	Count
Training data	2019.01.09	2019.01.29	42,864
Validation/Test data	2019.01.30	2019.02.12	10,716

<표 1> Data Classification for Model

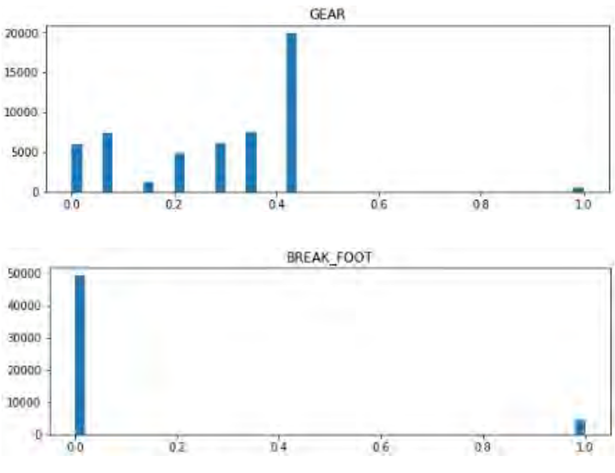
OBDII에서 추출한 데이터는 다음과 같이 구성되어 있다.

칼럼	설명	타입	칼럼	설명	타입
TIMS (iso8601)	측정시각	연속형	Break_foot (0, 1)	페달 브레이크	논리형
RPM (rpm)	엔진회전수	연속형	Break_side (0, 1)	사이드 브레이크	논리형
Vss (km/h)	차량속도	연속형	Slop_x (°)	x축 기울기	연속형
Lever (P,R,N,D)	레버 위치	범주형	Slop_y (°)	y축 기울기	연속형
TPS (%)	가속페달량	연속형	Slop_z (°)	z축 기울기	연속형
FCO (uL)	순간 연료소모량	연속형	ACON (0, 1)	에어컨	논리형
Mil (0, 1)	엔진경고등	논리형	Torque (N.m)	순간 토크	연속형
Latitude (°)	GPS위도	연속형	Handle (°)	핸들 조향각	연속형
Longtitude (°)	GPS경도	연속형	Belt (0, 1)	안전벨트 착용	논리형
Gear (1,2,3,4,5,6,14)	기어 단	범주형	Light (L,R,ER)	방향지시등	범주형

<표 2> Data Feature List

3.2 데이터 가공

학습에 사용할 feature들을 살펴보기 위해 모든 수치형 feature들의 분포를 시각화하여 확인해 본 결과 일부 항목들은 매우 불균형 분포를 가진다는 것을 알 수 있었다. 또한, feature 간의 단위가 달랐는데 예를 들어 'Gear(기어 단)' 데이터는 값이 1~14인 반면, 'Break' 데이터는 0~1 사이의 값을 포함하고 있다. 이러한 feature값들의 편차를 줄이기 위해 Scaling 과정을 통하여 데이터를 가공하였으며, Scaling 방법 중 MinMax Scaler를 이용하여 값이 0 ~ 1 사이가 되도록 데이터를 재조정하였다[5].



(그림 4) Data Histogram (Gear/Footbreak)

3.3 실험방법

실험은 회귀분석을 기반으로, Linear regression, Ridge, Gradient Boosting, XGBoost, Adaboost 모델 5가지를 구성하여 성능을 측정하였다. 실험은 모델의 학습(Training), 평가(Evaluation), 예측(Prediction) 세 부분으로 나누어 진행하였다. 성능 측정 지표는 평균절대오차, MAE(Mean Absolute Percentage Error)와 평균제곱근 오차인 RMSE(Root Mean Square Error)를 이용하였다. 두 개의 지표는 회귀문제의 성능 지표 이며 오차가 커질수록 RMSE값은 커지며, MAE 또한 이상치로 보이는 값이 많을 경우 사용하는 지표로써 값이 클수록 예측에 오차가 많은 것을 나타낸다. 모델 별로 파라미터 값은 가장 최적의 성능을 보이는 값으로 설정하였다.

4. 실험 결과 분석

Statsmodel 라이브러리의 OLS클래스를 이용하여 계수에 대한 분석내용을 <표 3>와 같이 살펴보았다. 결정계수(R-square) 는 0.408로 회귀분석으로 추정한 모델이 주어진 데이터를 얼마나 잘 설명하는지에 대한 점수로 값이 1에 가까울수록 데이터를 잘 설명하는 모델이다. 다음으로 F-통계량에 대한 수치를 확인하였을 때, Prob(F-statistic)을 살펴보았다. 각 입력 변수의 p-value값을 확인한 결과 [GEAR] p-value값이 0.619 로 유의미하지 않았다.(p-value>0.05) [GEAR] 칼럼을 제외한 나머지 변수는 0.05 미만으로 회귀분석에서 유의미한 것으로 나타났다.

Dep. Variable	MPG	R-squared	0.408
Model	OLS	Adj. R-squared	0.408
Method	Least Squares	F-statistic	1556.
Date	Thu, 09 Apr 2020	Prob (F-statistic)	0.00
Time	15:01:07	Log-Likelihood	10522.
No. Observations	42864	AIC	-2.100e+04
Df Residuals	42844	BIC	-2.083e+04
Df Model	19	Covariance Type	nonrobust

<표 3> OLS Regression results

표 5는 3장에서 설명한 실험 모델에 대한 성능을 나타내고 있다. 회귀분석 모델을 적용한 결과, MAE 값이 0.009, RMSE 값이 0.178 으로 AdaBoost 모델의 성능이 제일 좋은 것으로 나타났다.

	Linear regression	Ridge	Gradient Boosting	XGBoost	AdaBoost
MAE	0.14	0.143	0.035	0.049	**0.009**
RMSE	0.192	0.192	0.062	0.077	**0.178**

<표 4> Prediction Results

5. 결 론

최근 연료 값의 폭등과 환경적인 문제에 있어서 자동차 연비가 차량 구매에 있어 결정적인 요소 중 하나가 되고 있다. 본 연구는 MAF를 포함한 OBDII의 표준항목을 이용하여 추출한 데이터를 통하여 가장 효과적인 연비 예측 모델을 개발하고자 scaler 및 기계학습 알고리즘 등을 이용하여 변수와 연비와의 관계를 예측하는 연구를 진행하였다. 기계학습을 이용하여 최대 0.009의 MAE값을 갖는 연비 예측 모델을 만들 수 있었다.

더 나아가 신경망 모델을 적용하여 예측 모형의 성능을 고도화[7]하는 연구를 진행할 것이며, 다양한 차종의 데이터를 OBDII 단말기로부터 획득하여 모델에 적용한다면 범용적인 모델을 활용한 실제 응용에 활용될 수 있을 것으로 기대된다.

References

[1] Dimitrios Rimpas & Andreas Papadakis & Maria Samarakou(2020). OBD-II sensor diagnostics for monitoring vehicle operation and consumption. Energy Reports, Vol 6, Iss , Pp 55-63 (2020)

[2] https://www.windmill.co.uk/fuel.html

[3] https://www.researchgate.net/figure/Scheme-of-the-different-possibilities-of-MAF-calculation_fig4_285614280

[4] Aliyu, Abdullateef & Adeshina, Steve & Siraj, Fadzilah. (2014). Classifying Auto-MPG Data set using Neural Network. Proceedings of the 11th International Conference on Electronics, Computer and Computation, ICECCO 2014. 10.1109/ICECCO.2014.6997582.

[5] Shaheen, H. Agarwal, S. Ranjan, P. Minmaxscaler binary pso for feature selection(2020). In: 1st International Conference on Sustainable Technologies for Computational Intelligence- Proceedings of ICTSCI 2019. (Advances in Intelligent Systems and Computing, 2020, 1045:705-716)

[6] Jamala, Mohammed & Abu-Naser, Samy. (2018). Predicting MPG for Automobile Using Artificial Neural Network Analysis. Information Systems Research. 2. 5-21.

[7] Han, Chang-Wook. (2017). Auto MPG Prediction using Tree Architectures of Fuzzy Neural Networks. 39-43. 10.14257/astl.2017.145.08.