머신러닝을 이용한 이동통신 데이터 기반 교통량 추정 모형 개발

A Study on the Development of Traffic Volume Estimation Model Based on Mobile Communication Data Using Machine Learning

오 동 섭* · 윤 소 식** · 이 철 기*** · 조 용 성****

* 주저자 : 한국지능형교통체계협회 산업진흥본부 기술지원센터 센터장
** 교신저자 : 경찰청 교통국 국장
*** 공저자 : 아주대학교 교통시스템공학과 교수
**** 공저자 : 한국지능형교통체계협회 기획조정본부 본부장

Dong-seob Oh* · So-sig Yoon** · Choul-ki Lee*** · Yong-Sung CHO****

* Industry Promotion Office, Center for Supporting Innovative Business, ITS Korea
** Traffic Bureau of Korea Police Agency
*** Dept. of Transportation Systems Eng., Univ. of Ajou
**** Planning and Strategy Office, ITS Korea

† Corresponding author : Dong-seob Oh, ods76@itskorea.kr

Vol. 22 No.4(2023)
August, 2023
pp.01~13

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2023.22.4.1

Received 25 April 2023
Revised 12 May 2023
Accepted 4 July 2023

요 약

본 연구는 이동통신 로그 데이터를 통해 산출된 교통량 정보를 활용하여 기존 검지기에 준하는 교통량 정보를 추정하기 위해, 머신러닝의 앙상블 기법을 기반으로 하는 최적의 이동통신 기반 교통량 추정 모형을 개발하는 것이다. 이동통신 데이터를 통해 계측된 교통량 등의 정보와 VDS 실측 데이터를 활용하여 머신러닝 모형들을 통해 비교·분석한 결과, LightGBM 모형이 교통량 추정의 최적모형으로 선정되었다. 국도 1, 3, 6호선 검지영역 96개소를 대상으로 교통량 추정 모형의 성능을 평가한 결과, 전체 검지영역의 경우 MAPE 8.49로 교통량 추정 정확도가 91.51%로 분석되었다. VDS가 설치되지 않은 구간의 경우 교통량 추정 정확도는 92.6%로, VDS 설치가 어려운 구간에서도 LightGBM 교통량 추정 모형이 적용 가능하였다.

핵심어 : 이동통신 데이터, 교통량, 추정, 머신러닝, LightGBM

ABSTRACT

This study develops an optimal mobile-communication-based National Highway traffic volume estimation model using an ensemble-based machine learning algorithm. Based on information such as mobile communication data and VDS data, the LightGBM model was selected as the optimal model for estimating traffic volume. As a result of evaluating traffic volume estimation performance from 96 points where VDS was installed, MAPE was 8.49 (accuracy 91.51%). On the roads where VDS was not installed, traffic estimation accuracy was 92.6%.

Key words : Mobile communication data, Traffic volume, Estimation, Machine learning, LightGBM

Ⅰ. 서 론

1. 개요

ITS(Intelligent Transport Systems)를 통한 교통정보 수집 및 교통운영 방식은 루프 및 영상검지기, DSRC (Dedicated Short Range Communication), CCTV(Closed Circuit Television) 등 현장 시설물을 통해 주기적으로 수집된 교통량, 속도 자료를 활용하여 5분, 15분 단위 등의 교통정보를 생성하고, 이를 통해 교통운영·분석 및 교통정보 서비스를 제공하는 것이다. 교통정보센터는 주로 속도값을 교통정보로 사용하는데, 이는 소통상태를 정량화된 수치(km/h)로 직관적으로 표현할 수 있기 때문이다. 이로 인해 교통량과 속도와의 상관관계를 이용한 도로 소통상태 및 돌발상황 판단은 미흡한 실정이다.

국내외 연구들은 검지기를 통해 수집된 속도정보를 기반으로 통행속도와 통행시간을 예측하는 것에 중점을 두고 있다. 교통량 정보와 관련된 대부분의 국내 연구의 경우 교통량 결측 상황에 대응하기 위해, 기 운영 중인 검지기 데이터를 활용하여 교통량 및 AADT(Annual Average Daily Traffic) 추정하고 있다. 최근에는 이동통신 데이터를 활용한 교통정보 추정 및 예측을 위한 분석이 시도되고 있는데, 휴대전화와 이동통신 기지국 간 신호를 데이터화하고, 공간정보와 머신러닝 등 기법을 접목하여 빅데이터 분석을 통해 교통정보를 산출하기 위한 노력이 진행되고 있다.

이에 본 연구는 이동통신 데이터를 통해 교통량 정보를 산출하고, 이동통신 기반 교통량과 VDS(Vehicle Detection System) 교통량을 머신러닝 기법에 적용하여 교통량 추정 모형을 개발함은 물론, 선정된 최적모형을 현장 적용하기 위한 모형 평가를 수행하여 기존 검지기에 준하는 교통량 정보의 정확도[1]를 확보하도록 한다.

Ⅱ. 선행연구 고찰

1. 교통량 추정의 정의

교통량 추정(estimation)은 기 산출된 교통량 정보 간의 관계 또는 규칙을 모델링하여 주요 도로 지점 또는 구간을 주행한 교통량을 산출하는 것으로, 특정 시점에서 수집 또는 추정된 교통량으로부터 미래 시점의 교통량 변화를 산출하는 예측(prediction)과는 다르다. 대표적인 교통량 추정 기법으로 교통량 결측 지점의 교통량 변동 패턴이 인접한 검지기에서 수집된 교통량과 상관성이 높은 경우 회귀분석기법(다중회귀, 공간회귀 등 포함) 등을 사용한다.

2. 차량검지기 교통량 정확도 평가

1) 개요

국내에서 진행된 차량검지기 정확도 평가와 관련된 연구는 차량검지기 성능 및 유지관리 등을 기반으로 한 내용이 대부분으로, 검지기 정보의 정확도 판단 기준 정의와 관련된 연구 사례는 없다. 이에 반하여 국외의 경우 비매설식 검지기의 활용도 검증을 위해 매설식 검지기를 기준으로 교통량, 속도 등에 대한 정확도

1) 「자동차·도로교통분야 ITS 성능평가기준」의 차량검지기 성능

평가를 수행하였다. 국내외 모두 검지기 정확도 평가를 위해 퍼센트오차(% error)와 평균절대오차백분율(MAPE, Mean Absolute Percentage Error) 등을 주요 지표로 활용하였다.

2) 연구사례

Kim and Kim(2002)은 내부순환로에 설치 운영 중인 영상식 검지기의 성능평가를 위해 교통량과 속도의 정확도가 MAPE 10% 미만이 되도록 튜닝하는 과정을 거쳐 오차발생 원인을 파악하고 정확도 제고를 위한 성능 유지관리 방안을 검토하였다.

Jang et al.(2011)은 검지기 성능평가 체계를 개선할 수 있도록, 차량검지기 성능평가를 위한 평가척도(R, MAPE, 상관계수)를 기반으로 평가척도 선정절차를 수립하고, 합리적 표본추출을 위한 구간추정 방법론으로 층화추출법을 제안하였으며, 통계적 신뢰성 확보를 위해 표본 평균과 표본 표준편차를 활용한 모집단 오차에 대한 신뢰구간 추정 방법론을 제시하였다.

국토교통부의 「자동차·도로교통분야 ITS 성능평가기준」에 따르면, 교통량, 속도를 평가항목으로 하여 차량검지기의 성능을 평가하도록 하고 있다. 차량검지기의 평가를 위해 “100(%)－MAPE”를 지표로 하여, 차량검지기의 평가등급(상, 중, 하)을 정의한다.

Texas Transportation Institute(2002)는 루프검지기를 대체할 수 있는 비매설식 검지기의 정확도 확인을 위해 루프검지기를 기준으로 영상검지기와 레이더검지기의 속도, 교통량, 점유율에 대해 퍼센트오차를 활용하여 분석하였다. 당시 기준으로 비매설식의 경우 자유교통류상황(free flow condition)에서의 교통량 정확도는 95%, 속도 정확도는 5mph 이내였으나 혼잡상황이 진행될수록 교통량은 70~90%, 속도는 10~30mph의 오차를 보이는 것으로 확인되었다.

미국의 각 교통 관련 기관 등에서 수행한 검지기 정확도 평가의 경우, 비매설식 검지기를 대상으로 교통량, 속도에 대해 MAPE, MAPD(Mean Absolute Percentage Difference), MPD(Mean Percentage Difference), RMSE(Root Mean Squared Error) 등을 주요 활용하였다.

3. 교통량 등 교통변수 추정 및 예측

1) 개요

교통변수 추정 등과 관련된 연구는 속도 및 통행시간의 추정 또는 예측을 중심으로 이루어졌다. 교통량 추정의 경우 회귀모형을 이용한 방식이 주를 이루고 있으며, 속도 및 통행시간 예측의 경우 딥러닝 등을 활용한 예측 모형 개발 연구 사례가 있다. 이동통신 데이터를 기반으로 하는 연구의 경우 VDS의 설치·운영비 및 정보 수집의 공간적 범위에 대한 문제 해결을 위해 머신러닝 등을 적용한 연구를 진행하였다.

2) 연구사례

Oh(2001)는 일반국도상에 설치된 상시교통량조사 장비를 통해 수집된 자료를 활용하여 일교통량과 시간교통량과의 관계에 대해 상관분석을 수행하고 시간 관측 교통량(독립변수)과 일교통량(종속변수)의 관계를 회귀분석으로 추정하는 모형을 개발하였다. 상시교통량 조사지점의 교통특성을 3개 그룹으로 구분하여 그룹별 모형을 구축 후 MAPE를 통해 각 추정모형의 정확도를 평가하였다.

Yoon(2016)은 고속도로 TCS(Toll Collection System) 자료와 자동차 내비게이션의 GPS(Global Positioning System) 통행자료를 이용하여 고속도로 연결로 구간에서의 AADT를 추정하는 연구를 수행하였다. 고속도로 연결로 구간의 AADT 관측값과 추정값의 정확도 평가를 위해 MAPE와 MAE(Mean Absolute Error)를 사용하

였으며, 비교 결과 TCS와 GPS 관측값을 통해 AATD의 추정이 가능함을 제시하였다.

Lee(2019)은 교통변수 추정 등과 관련된 대부분의 연구가 통행속도와 통행시간을 대상으로 이루어진 점에 착안하여, 고속도로 내 하이패스 기반 DSRC 프로브 자료와 일정 간격으로 설치된 검지기의 관측 교통량과의 관계를 이용하여 교통량 미관측 지점에서의 동적 교통량 추정 모형을 개발하였다. 개발 모형은 프로브 자료를 통한 미관측지점의 전수계수를 산정하는 전수화계수모형, 프로브 자료를 관측교통량과 유사한 패턴으로 조정하는 변동성 축소모형, 관측 교통량과 프로브 자료의 관계를 통해 미관측지점의 프로브 자료를 교통량으로 전환하는 교통량전환모형으로, 개발된 모형은 MAPE, MAE 등으로 평가하였다.

Chung et al.(2018)은 기존 설문조사 기반의 통행시간예산(TTB, Travel Time Budget) 산출 방식을 개인휴대전화(CP, Cellular Phone)와 통신 기지국 간에 발생하는 신호인 MPSD(Mobile Phone Signaling Data)를 기반으로 산출하는 연구를 진행하였다. 그 결과 개인휴대전화가 통신 기지국의 범위 안에 있을 때 수집되는 이동통신 로그 중 통신시작 및 종료시간 그리고 각 통신 시간에 위치한 개인휴대전화 기반의 시공간적 정보를 기반으로 TTB를 산출할 수 있음을 확인하였다.

Ji and Hong(2019)은 딥러닝 LSTM(Long Short-Term Memory) 모형에 LTE(Long Term Evolution) 기지국과 개인휴대단말의 통신에 따른 시공간적 정보 즉, 이동통신 로그 데이터를 활용하여 실시간으로 속도를 예측하는 연구를 수행하였다. 서울 남부순환로, 강변북로 및 올림픽 대로를 대상으로 분석 결과, 기존 과거 패턴 자료를 활용하는 방식의 속도 예측 정확도가 70.5%인데 반하여 제안 모형의 경우 84.3%로 높아, 이동통신 로그를 활용한 속도 예측이 가능함을 도출하였다.

Caceres et al.(2012)은 교통량 수집을 위해 모든 도로에 VDS를 설치하는 것은 경제적 등의 이유로 비현실적이기 때문에, 휴대전화의 발신통화(outgoing call) 이력 자료를 통해 교통량을 수집 또는 추정할 수 있다면, 인프라 기반의 교통정보 수집 방안의 대안이 될 수 있을 것으로 보았다. 연구결과 Spanish Traffic Management Centre에서 운영 중인 검지기로부터 수집되는 교통량과 연구를 통해 추정된 교통량 간 오차가 20% 미만으로, 이동통신 데이터를 통한 교통량 추정이 가능함을 보여주었다.

Snowdon et al.(2018)[2]은 PVD(Probe Vehicle Data)로 통칭하는 내비게이션 또는 개인휴대단말의 GPS 정보 등을 활용하여 교통량을 예측하는 것은, 특정 지점에서만 교통량을 수집할 수 있는 기존 검지기를 대체하는 방법이라고 제안하였다. 이에 개별 차량의 내비게이션으로부터 수집되는 정보를 활용하여 교통량을 예측하고, 예측값과 Virginia Department of Transportation에서 운영하는 검지기 정보[3]를 비교하는 연구를 진행하였다. 그 결과 MAPE가 20.62~21.94%로 도출되어 고정식 ITS 설비를 이용하지 않고도 교통량 예측이 가능함을 확인하였다.

4. 머신러닝 기반 교통 분석

1) 개요

머신러닝을 이용한 국내외 교통분석의 경우 교통수단선택, 사고유형구분, 교통혼잡상태, 통행속도 예측, 교통사고 사망자 추정과 같이 다양한 분야에 사용되고 있다. 머신러닝 분석 시 앙상블 기반의 모형을 주로 사용하고 있으며, 학습 및 검증을 위한 데이터셋은 7:3에서 8:2의 비율을 주로 적용하였다. 최적모형의 경우 연구 주제와 무관하게 앙상블 모형 중 LightGBM이 주로 활용되었다.

2) 해당 연구는 교통량을 추정(Traffic Volume Estimation)하는 모형을 개발하는 것으로 하였으나, 세부 내용은 교통량 예측(Traffic Volume Prediction)으로 진행하였음

3) 참값 또는 타켓 데이터를 ground truth data로 칭함

2) 연구사례

Lee and Hong(2019)는 개인별 교통수단 선택 예측을 위해 2016년 서울시 가구통행실태조사 데이터를 활용하여 다항로짓, 의사결정나무, 서포트벡터 머신 기반의 연구를 진행하였다. 개인별 교통수단선택 예측모형으로 서포트벡터머신이 가장 좋은 결과를 보였으며, 해당 모형을 이용하여 개인 교통수단 예측은 물론 교통정책 추진을 위한 의사결정 도구로 활용할 수 있을 것으로 분석했다.

Lee et al.(2019)은 고속도로 사고 위험등급 예측모형 개발을 위해 Random Forest, LightGBM, CatBoost, XGBoost, AdaBoost 모형을 비교 분석하였으며, 사고 위험등급 예측에 LightGBM을 최종 모형으로 선택하였다. 학습과 검증 데이터셋을 8:2로 하여 예측모형을 검토하였으며, 머신러닝 기법으로 사고 위험등급을 예측한 사례가 적음에 있어 LightGBM 모형의 효과성을 확인하였다.

Choen et al.(2021)은 고속도로 VDS 교통정보를 이용하여 CatBoost 모형을 통해 교통상황을 원활, 다소 지체, 지체, 다소 정체, 정체로 예측하였다. Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost, 선형회귀 및 결정트리를 대상으로 머신러닝 모형들과 비교하였으며, 데이터셋은 7:3의 비율로 학습과 검증용으로 구분하였다. 분석 결과 교통상황 예측의 경우와 같이 범주형 데이터를 기반으로 하는 경우 CatBoost 모형이 적합한 것으로 도출되었다.

Nam et al.(2021)은 교통사고 데이터와 기상환경 등의 자료를 기반으로 머신러닝 모형(의사결정트리, Random Forest, XGBoost, LightGBM)을 통해 서울시 교통사고 사상자를 예측하고자 하였다. 2015~2018년 데이터를 학습용으로, 2019년 데이터를 검증용으로 사용한 결과 단일 모형인 의사결정트리에 비해 앙상블 모형, 특히 LightGBM의 추정력이 우수하였다.

Yaghoubi et al.(2021)은 LTE 통신 주파수 특성 정보와 도로 및 교통변수를 활용하여 교통량을 추정하기 위해 회귀모형과, Random Forest, LSTM 등 머신러닝 기법을 활용하여 검증(학습 및 검증 데이터셋 비율 8:2)하였다.

5. 본 연구의 차별성

본 연구는 이동통신 데이터 기반 교통량 등 정보와 VDS 교통량 정보를 활용하여 교통량을 추정하는 모형을 개발하는 것으로, 기존 연구와 다음과 같은 차별성을 가진다.

첫째, 본 연구는 ITS 정보수집 인프라를 기반으로 교통량을 추정하는 것이 아니라, 이동통신 로그 데이터를 통해 계측한 교통량을 기반으로 교통량을 추정한다. 즉, 활용 데이터 특성에 있어 해외의 경우 이동통신 데이터로 발신통화 이력 또는 이동통신 주파수 특성, PVD GPS 정보를 활용하여 교통량을 추정하고 있으나, 본 연구는 현재 이동통신 로그 데이터를 통해 교통량, 속도 등 교통변수를 집계 및 활용하여, VDS 수집정보와 직접 비교 평가한다. 둘째 기존 연구가 특정 구간에 한정된 모형을 개발했다면, 본 연구는 이동통신 기지국 내 또는 기지국 간 이동체에 대한 교통량 추정 모형 개발 및 관련 모형의 추정 정확도를 확인함으로써 전국 단위 모형으로 확대할 수 있는 특징이 있다. 셋째, 회귀모형 및 한정된 머신러닝 기법에서 탈피하여, 최근 활용도가 높은 앙상블 기반의 머신러닝 기법을 검토하고 최적모형을 선정함으로써 이동통신 데이터를 통한 교통량 추정 모형을 검증, 평가한다.

즉 본 연구는 도로 전역을 커버하는 이동통신 데이터를 통해 VDS 운영 여부와 상관없이 모든 도로구간의 교통량을 추정할 수 있는 머신러닝 모형을 개발함으로써, 기존 속도 중심의 정보활용 체계에서 교통량을 접목한 교통운영·분석과 교통량을 활용한 서비스의 가능성, 그리고 VDS 설치가 어려운 구간에 해당 모형을 적용할 수 있는지 확인한다는 측면에서 차별성이 있다.

Ⅲ. 머신러닝 기반 교통량 추정모형의 개발

1. 교통량 추정모형 개발 절차

머신러닝 기법을 활용한 교통량 추정모형은 이동통신 데이터를 통한 교통량 자료와 참값에 해당하는 VDS 교통량을 기반으로 개발한다. 데이터 수집을 위해 VDS 운영지점과 이동통신 기반의 교통량 생성 단위 구간을 중심으로 데이터셋을 정의하고, 수집된 자료로 데이터셋을 구성한다. 이후 데이터셋은 머신러닝 모형 개발을 위한 학습 및 학습 모형에 대한 검증용으로 분리한다.

이동통신 교통량(입력값)을 VDS 교통량(참값)과 비교하여 학습하는 지도방식 기반의 최적모형 개발 및 선정을 위해 비교·분석 대상 모형을 선정하고, 비교·분석 대상 모형들에 대해 학습 및 검증용으로 분리된 데이터셋을 적용하여 모형의 학습 및 검증을 수행한다. 최적모형 선정을 위해 MAPE 및 RMSE를 활용하여 모형별 비교·분석을 수행한다. 선정된 최적모형은 추후 이동통신 데이터를 통해 추정된 교통량과 VDS 실측값을 비교하여 MAPE 및 RMSE 등을 기반으로 교통량 추정모형의 성능(정확도)을 검증 및 평가한다.

2. 데이터 전처리 및 데이터셋 정의

최적 머신러닝 모형 개발을 위해서는 관련 데이터의 수집 및 처리가 필요하다. 즉 머신러닝의 지도학습을 통한 교통량 추정을 위해서는 기본적으로 전처리 과정 즉 이동통신 로그를 통해 수집된 교통량(입력값)과 참값(레이블)에 해당하는 VDS 교통량은 물론, 이동통신 및 VDS 기반 속도 등 교통변수 데이터와 교통량 변화에 영향을 미치는 시간적 요소 그리고 차로수 및 구간 길이 등 도로환경요소를 변수로 하는 데이터셋을 정의하고, 이를 통해 일반 회귀분석 등으로는 파악하기 어려운 변수 간 특성을 머신러닝을 통해 확인할 수 있도록 하였다.

전처리 데이터는 이동통신을 통해 수집되는 개별 단말기 단위의 연속적인 위치 추적 정보를 통해 산출된 단말기 ID, 정보 수집 시간, 이동시간과 이동거리를 기반으로 교통량과 속도 등을 산출하여 도로구간에 매핑(교통량 생성 구간으로 본 연구 및 데이터셋에서는 LID로 정의)하고 그 결과를 데이터셋화하였다. VDS 교통정보의 경우, 현장 제어기를 통해 30초 주기로 수집된 자료를 기반으로 센터(또는 제어기)에서 1분 집계 및 보정을 수행하고, 이를 가공하여 최종 5분 단위 정보로 생성한다. 이에 교통량 생성 구간의 이동통신 데이터도 VDS 정보와 동일하게 30초 단위 1분 집계, 5분 단위 교통량 정보를 생성하도록 하였다.

교통량 추정모형의 개발을 위해 필요한 데이터는 데이터셋의 정의에 따라 일시, 평일 및 휴일여부, 차로수(2~4차로) 및 전용차로 등의 특성을 반영할 수 있도록 하였으며, 일반국도 1, 3, 6호선 내 운영 중인 VDS 제어기 기준 총 73개 지점(방향별 103개)을 대상으로 자료를 수집하였다. 특히 데이터셋 중 시간(또는 기간) 특성을 반영한 교통량 추정 모형 개발을 위해, ‘21년 12월 13일~31일, ‘22년 1월 1일~9일, ‘22년 2월 10일~24일, ‘22년 3월 1일~24일, 총 62일 동안 수집된 이동통신 기반 교통량과 VDS 교통량 등 자료를 활용하였다.

데이터셋의 정의에 따라 교통량, 속도 등의 정보를 수집 및 집계하였으며, 이상치 제거를 위해 교통량-속도의 관계에 따라 교통량이 0대임에도 불구하고 속도가 0이 아닌 경우 또는 속도가 0임에도 교통량이 집계된 경우, 데이터셋 구성 시 수집 데이터에서 제거하였다. 그 결과 총 2,762,252개의 데이터셋을 구성하였다.

<Table 1> Definition of a dataset

Category		Description	note
representative ID	Lid ID	Mobile based traffic data generation segment ID (representative segment when estimating traffic volume using input and label value)	Lid
traffic variable	LID volume	Lid volume(5min)	input value
	LID speed	Lid speed(5min)	km/h
	LID density	Lid density(5min)	veh/km/5min
	LID aggregated volume	5min, 15min, 60min aggregated volume	confirmation of traffic volume changes
	VDS volume	VDS volume(5min)	lable value
	VDS speed	VDS speed(5min)	km/h
road environment	route number	National Highway route number	1, 3, 6
	bus lane	confirmation of the existence of a bus lane	none/opt/non-opt
	number of lanes	number of lanes in the data generation segment	-
	length	length of the data generation segment	m
time element	yyy/mm/dd	data generation year/month/day	yyyy, mm, dd
	hh/mm	data generation time	hh, mm
	number of weeks	number of weeks	1, 2, 3, 4, 5
	day	monday~sunday	mon:0, tue:1 ~ sun:6
	type of day	confirmation of traffic characteristics according to the distinction between holidays and weekdays	-
	holiday	confirmation of traffic characteristics on holidays	-

<Table 2> Dataset for machine learning analysis

Mobile based traffic data generation segment ID(Lid ID)	LID volume	LID speed	VDS volume	VDS speed	route number	number of lanes	…
003000N002	18	58.92	20	76	3	4	…
003000N002	17	44.96	17	78	3	4	…
003000N002	15	52.84	31	74.8	3	4	…
003000N002	14	55.09	21	78.2	3	4	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3. 교통량 추정모형의 개발

1) 데이터셋의 분리

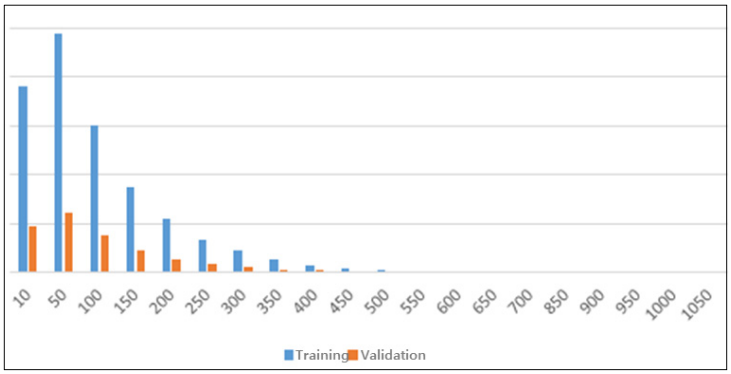
교통량 추정모형 선정 및 개발에 앞서, 수집 데이터셋 2,762,252개에 대해 학습 데이터셋과 검증 데이터셋의 분리를 수행하였다. 데이터셋 분리 목적은 주어진 전체 데이터로만 학습이 진행될 경우 머신러닝 모형 성능이 떨어지기 때문에, 머신러닝 모형 개발을 위해서는 학습과 검증용 데이터셋의 분리 과정은 필수적이다. 데이터 분리 비율의 경우 경험치를 기반으로 학습과 검증 데이터를 나누고 있으며, 머신러닝 관련 연구의 경우 주로 7:3 또는 8:2 비율을 사용(Lee et al., 2019; Lee, 2020; Yaghoubi et al., 2021)하고 있다.

이에 본 연구에서는 '21년 12월부터 '22년 3월 중 총 62일 동안 수집된 데이터셋 2,762,252개에 대해 학습

데이터(2,209,800개)와 검증 데이터(552,452개)를 8:2의 비율로 분리하였다.

<Table 3> Basic statistics values depending on the dataset split

Category	Training Dataset	Validation Dataset
Sample size (%)	2,209,800 (80%)	552,452 (20%)
Average	70.05	67.25
Standard deviation	78.26	74.16



<Fig. 1> Traffic volume distribution according to the training-validation dataset split

2) 모형별 최적화 수행에 따른 최적모형의 선정

데이터 2,209,800개를 활용하여 학습을 진행하였으며, 모형 최적화를 위해 Bayesian Optimization 방법을 적용하여 최적 하이퍼파라메터를 도출하였다. 회귀모형의 경우 하이퍼파라메터는 y 절편을 계산하고, 학습 시 원본 데이터의 수정을 방지함은 물론 계수의 부호를 특정 값에 한정하지 않도록 하였다. 앙상블 모형의 경우 이터레이션 수, 학습률, 트리의 깊이 등이 하이퍼파라메터로 사용되었다. 이터레이션 수의 경우 Random Forest는 학습을 위해 사용될 총 트리 개수를, 부스팅 계열인 XGBoost, LightGBM, CatBoost 모형은 총 부스팅 횟수를 의미하며, 학습률의 경우 GBM(Gradient Boost Machine) 모형을 기반으로 하는 특징에 따라 최적값을 찾아가는 비율을, 트리 깊이는 모형의 과적합을 제어하는 하이퍼파라메터로 설정하였다.

학습을 통해 도출한 최적 하이퍼파라메터를 각 모형에 반영하고 검증용 데이터 552,452개를 사용하여, 모형별로 추정한 교통량과 검증용 VDS 교통량을 기반으로 모형 간 교통량 추정 결과를 비교하였다. 최적모형검토 결과 LightGBM의 MAPE는 6.35 즉, 교통량 추정 정확도(100-MAPE)가 93.65%로「자동차·도로교통분야 ITS 성능평가기준」VDS 성능평가의 교통량 정확도 기준, 상급 이상(90%)으로 분류되는 성능을 보인다. Random Forest의 경우 6.63으로 두 번째로 좋은 성능을 보였으며, CatBoost, XGBoost, 회귀모형 순으로 교통량 추정 정확도가 좋은 것으로 확인되었다. RMSE 역시 추정 교통량(대)과 실제 값(대)의 차이가 가장 작은 LightGBM을 최적모형으로 선정하였다.

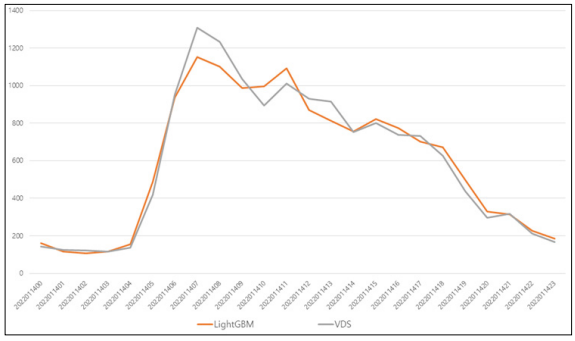
최적모형으로 선정된 LightGBM의 교통량 추정에 따른 데이터셋의 변수 중요도(feature importance)를 확인한 결과, 교통량 생성 구간 ID가 이동통신 로그 데이터를 연속된 도로 구간 상에 매칭하여 교통량 등을 추정함은 물론, 참값에 해당하는 VDS가 설치된 구간의 교통량을 참조하는 정보생성 대푯값으로써 가장 중요도가 높았다.

다음으로 일자 및 시간 변수, 평일과 휴일을 구분하는 날짜 유형, 교통량, 속도, 밀도 및 과거 교통량 등이 중요한 것으로 도출되었다. 이에 반하여 도로환경 변수의 경우 대부분의 변수가 교통량 추정에 있어 중요하지 않은 것으로 나타났다.

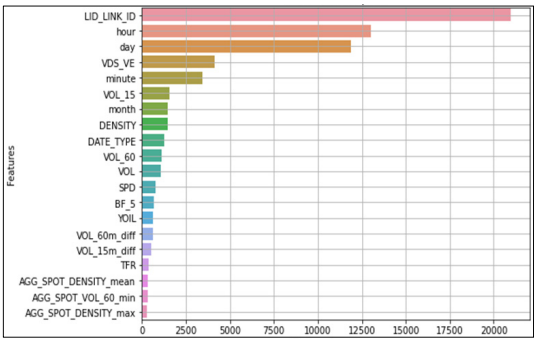
결과적으로 LightGBM의 교통량 추정을 위해서는 차로수, 구간 길이 등 도로환경 변수보다는 교통량 추정을 위해 연속적으로 정보를 생성하는 구간 대푯값(교통량 생성 구간 ID), 시간 미 날짜 유형 그리고 교통량, 속도 등 교통변수들이 LightGBM의 교통량 추정에 있어 중요한 변수 들인 것으로 분석되었다.

<Table 4> Performance comparison of models for selecting the optimal mode

Model	MAPE	100-MAPE(%)	RMSE
Random Forest	6.63	93.37	14.89
XGBoost	7.38	92.61	15.97
LightGBM	6.35	93.65	14.42
CatBoost	7.20	92.80	16.59
Liner Regression	12.32	87.68	36.74



<Fig. 2> Validation results of traffic volume estimation using LightGBM model



<Fig. 3> Feature Importance

Ⅳ. 교통량 추정모형의 평가

1. 모형의 평가 개요

1) 평가 지표

LightGBM 기반 교통량 추정 모형 평가는 VDS 운영구간과 VDS 미운영구간을 대상으로 교통량 추정 정확도를 확인하였다. 평가 지표로 MAPE, RMSE를 사용하였으며, MAPE의 경우 정확도 분석을 위한 기본 지표로, 「자동차·도로교통분야 ITS 성능평가기준」중 VDS 성능평가 부문의 교통량 정확도와 비교하였다.

2) 평가 범위

VDS 운영구간 및 미운영구간을 대상으로, LightGBM 기반 교통량 추정 결과와 VDS 실측 교통량을 비교하여 평가하였다. 평가는 일반국도 1, 3, 6호선에서 운영 중인 제어기 73개소, 103개 검지영역을 대상으로

‘22년 4월 5일 하루 동안으로 하되, 해당 기간 중 VDS 이상 등에 따라 교통량이 수집되지 않는 제어기 또는 검지영역은 제외하였다. 그 결과 68개소 96개 VDS 검지영역에 대해 평가를 진행하였다. VDS 미구축 구간의 경우 ‘22년 4월 7일 10시부터 17시 사이에 1시간 간격으로 2회씩 총 2시간 동안 데이터를 수집하였다.

2. 모형의 평가

1) LightGBM 추정 교통량과 VDS 교통량의 비교

전체 96개 검지 영역에서 ‘22년 4월 5일 하루 동안 시간대별로 집계된 추정 교통량 분석 결과 MAPE 8.49, RMSE 129.97, RMSE 정규화 0.028이 도출되었다. 개별 검지영역을 시간당 하루 집계 결과로 그룹화하여 분석 시 MAPE 9.66, RMSE 111.12, RMSE 정규화 결과 0.081로 확인되었다.

각각의 결과를「자동차·도로교통분야 ITS 성능평가기준」의 VDS 성능평가의 교통량 정확도를 기준으로 볼 때 91.51%와 90.34%로, LightGBM 교통량 추정 모형의 정확도는 상급 이상(90%)으로 분류되어, 알고리즘 성능이 우수하다. 특히 RMSE 정규화 결과의 경우 0.028, 0.081로 0에 근사하여 LightGBM은 교통량 추정 모형으로 적합하다고 볼 수 있다.

즉 LightGBM 교통량 추정 모형은 VDS가 운영되는 지점을 기반으로 VDS가 운영되지 않는 인접구간의 교통량의 추정은 물론, VDS 장애 등으로 인해 교통량이 수집되지 않는 경우에서도 이동통신 데이터를 활용하여 VDS에 준하는 교통량 정보를 수집 및 활용 가능한 성능을 가진다.

<Table 5> Summary of LightGBM estimation model evaluation results

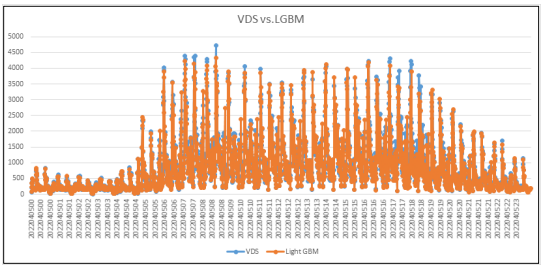
Evaluation index	Aggregation of the entire detection area by time period	Daily aggregation per detection area
MAPE (100-MAPE)	8.49 (91.51)	9.66 (90.34)
RMSE	129.97	111.2
Normalized RMSE	0.028	0.081

2) 일반국도 노선별 LightGBM 교통량 추정 모형 평가

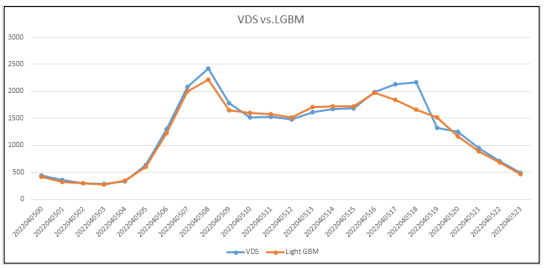
일반국도 노선별 MAPE 분석 결과, 1호선은 12.47, 3호선은 7.29, 6호선은 10.89로 나타났다. RMSE의 경우 1호선 134.53, 3호선 108.12, 6호선 94.33, RMSE 정규화는 1호선 0.108, 3호선 0.068, 6호선 0.077로 나타났다. MAPE 기준 최소 7.29에서 최대 12.47의 범위이며, 이는 정확도(100-MAPE) 87.53~92.71% 수준으로 LightGBM 교통량 추정 모형은 적정 성능을 확보하고 있다.

<Table 6> Summary of LightGBM estimation model evaluation results for national highway routes

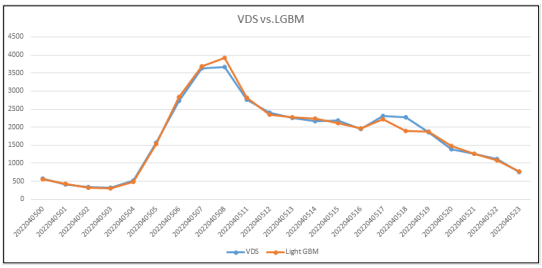
Evaluation index	Route-based aggregation		
	Route 1	Route 3	Route 6
MAPE (100-MAPE)	12.47 (87.53)	7.29 (92.71)	10.89 (89.11)
RMSE	134.53	108.12	94.33
Normalized RMSE	0.108	0.068	0.077



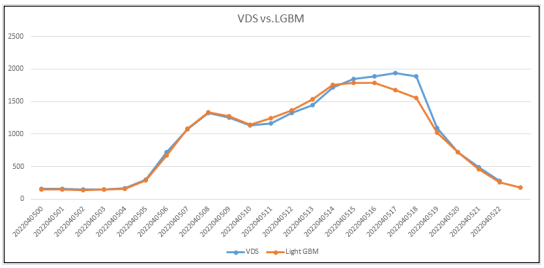
<Fig. 4> Daily estimated volume and VDS volume



<Fig. 5> Route 1(VDS 015 case)



<Fig. 6> Route 3(VDS 443 case)



<Fig. 7> Route 6(VDS 488 case)

3) VDS 미운영 구간에 대한 LightGBM 교통량 추정 모형 평가

LightGBM 모형을 통해 추정된 교통량이 VDS가 운영되지 않는 구간에도 활용 가능한지 확인하기 위해 VDS가 설치 운영되지 않는 지점을 선정하고, LightGBM 모형에서 추정한 교통량을 이동식 검지기를 통해 집계한 교통량과 비교 분석하였다. 대상 장소는 일반국도 1, 3, 6호선 각 1곳씩 총 3곳으로, '22년 4월 7일 10시부터 17시 사이에 1시간 간격으로 장소별 2회씩 총 2시간 자료를 수집하였다.

3개 대상지에 대한 교통량 추정값과 실제 집계 교통량을 비교한 결과, 노선 별 MAPE는 1호선 9.0, 3호선 2.8, 6호선 10.4이며, 전체 평균 MAPE 7.4로 교통량 추정 정확도는 92.6%로 분석되었다. 이는 VDS가 운영되지 않는 구간에서도 이동통신 로그 데이터 기반 LightGBM 모형의 교통량 추정력이 우수하며, VDS가 설치되지 않은 구간에서도 해당 모형을 적용할 수 있음을 의미한다.

<Table 7> LightGBM model evaluation results for unoperated segments of VDS

Route No	Location	Time	Volume(veh/h)		MAPE	Accuracy (%)
			LightGBM	Portable VDS		
1	Daeja-dong, Goyang-si	10~11	971	905	7.3	92.7
		13~14	967	874	10.7	89.3
	Average		-	-	9.0	91.0
3	Bongyang-dong, Yangju-si	15~16	1,077	1,030	4.6	95.4
		16~17	1,133	1,121	1.1	98.9
	Average		-	-	2.8	97.2
6	Yangpyeong-eup, Yangpyeong-gun	10~11	910	1,026	11.4	88.6
		11~12	960	1,059	9.4	90.6
	Average		-	-	10.4	89.6
Total Average			-	-	7.4	92.6

Ⅴ. 결 론

본 연구는 이동통신 로그 데이터를 통해 산출된 교통량 정보에 대해 기존 검지기에 준하는 정확도를 확보하고, 교통정보 수집이 어려운 구간 등을 대상으로 교통량을 추정하기 위해, 최근 활용도가 높은 앙상블 기반의 머신러닝 기법을 검토하여 최적의 이동통신 기반 교통량 추정 모형을 개발하였다. 이동통신 데이터를 통해 계측된 교통량 등의 정보와 VDS 실측 데이터를 기반으로 최적모형 선정 결과 LightGBM이 MAPE 6.35, RMSE 14.42로 다른 모형보다 우수하여, 교통량 추정의 최적 모형으로 선정하였다. LightGBM을 활용하여 국도 1, 3, 6호선의 VDS 검지영역 96개소를 대상으로 교통량 추정 성능을 평가한 결과, MAPE 8.49, RMSE 129.97, RMSE 정규화 0.028로 분석되었다. 이는 전체 검지영역에 대한 교통량 추정 정확도(100-MAPE)로 분석 시 91.51% 수준으로, 「자동차·도로교통분야 ITS 성능평가기준」의 VDS 성능평가의 교통량 정확도 상급에 해당하는 성능이다. LightGBM 모형을 통해 추정된 교통량이 VDS가 운영되지 않는 구간에도 활용 가능한지 분석한 결과, MAPE는 1호선 9.0, 3호선 2.8, 6호선 10.4, 평균 MAPE 7.4로 교통량 추정 정확도가 92.6%로 분석되어 VDS 설치가 어렵거나 장비 고장 등으로 정보수집이 불가능한 구간에 충분히 적용 가능함을 확인하였다. 이에 본 연구를 통해 개발된 모형을 사용할 경우, 도로 전역을 커버하는 이동통신 데이터를 이용하여 VDS 설치 여부와 무관하게 적정 수준의 교통량 추정 정확도를 확보할 수 있음을 확인하였다.

특히 본 연구는 국내외 교통량 추정 모형 개발 과정과 다르게 이동통신 기반 데이터를 직접 활용한 점, 이동통신 데이터를 통해 산출된 교통량에 대해 앙상블 기반 머신러닝 기법 간 비교로 최적모형을 개발한 사례가 적은 여건에서 최적모형을 개발 및 적용하여 모형의 활용 가능성을 평가했다는 측면에서, 향후 ITS 분야의 교통정보 수집 체계 개선 및 보완에 기여할 수 있다고 판단된다.

향후 연구에서는 수집 및 적용한 데이터 측면에서 다양한 교통류 변동요인을 반영할 수 있는 변수인 특수일, 도로등급별 교통류 형태, 주요 교통시설 주변의 교통 특성, 사고 및 공사 등 다양한 도로교통 이벤트 등 도로교통 상황별 특성을 반영한 교통량 추정 모형 개발이 필요하다. 이와 함께 실시간 기반의 교통량 예측모형으로 고도화를 병행함으로써, 교통정보의 품질 향상은 물론 교통정보 분석 및 제공체계의 고도화에 기여할 수 있을 것으로 기대된다.

ACKNOWLEDGEMENTS

본 논문은 오동섭의 박사학위로 아주대학교에 제출되었던 논문을 수정·보완하여 작성하였습니다.

REFERENCES

Caceres, N., Romero, Luis, M., Benitez, F. G. and Castillo, J. M.(2012), “Traffic Flow Estimation Models Using Cellular Phone Data”, *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, vol. 13, no. 3, pp.1430-1441.

Cheon, M. J., Choi, H. J., Park, J. W., Choi, H. Y., Lee, D. H. and Lee, O.(2021), “A study on the Traffic Prediction throught Catboost Algorithm”, *Journal of the Korea Academia-Industrial Cooperation Society*, vol. 22, no. 3, pp.58-64.

Chung, Y. S., Nam, S. G. and Song, T. J.(2018), "A Travel Time Budget Estimation Using a Mobile Phone Signaling Data", *Journal of he Korean Society of Civil Engineers*, vol. 38, no. 3, pp.457-465.

Jang, J. W., Park, C. S., Baek, N. C. and Lee, M. Y.(2011), "Analysis on Video Image Detection System Performance by Vehicle Speed", *Journal of Korea Society of Transportation*, vol. 23, no. 5, pp.105-112.

Ji, B. S. and Hong, E. J.(2019), "Deep Learning Based Real Time Road Traffic Prediction Using LongTerm Evolution Access Data", *Sensors*, vol. 19, 5327, pp.1-16.

Kim, D. H. and Kim, S. G.(2002), "A study on Testing the Ability of Vehicle Detection System in ATMS", *Journal of Korea Society of Transportation*, vol. 20, no. 5, pp.231-244.

Lee, E. J.(2020), *A study on road speed of Seoul Gangnam-gu and Jung-gu predictions based on Machine Learning Methods*, Ewha Womans University.

Lee, H. M., Jeon, G. S. and Jang, J. A.(2019), "Predicting of the Severity of Car Traffic Accidents on a Highway Using Light Gradient Boosting Model", *Journal of the Korea Institute of Electronic Communication Sciences*, vol. 15, no. 6, pp.1123-1130.

Lee, S. B.(2019), *Development of Dynamic Traffic Volume Estimation Model Using DSRC Probe Data*, Seoul National University.

Lee, Y. H. and Hong, S. Y.(2019), "A Machine Learning Approach to the Prediction of Individual Travel Mode Choices", *Journal of the Korea Data Information Science Society*, vol. 30, no. 5, pp.1011-1015.

Nam, M. W., Park, D. S., Jang, Y. J. and Lee, H. C.(2021), "Prediction of Traffic Accident Casualties Using Machine Learning", *Korea Society of Computer Information*, vol. 29, no. 1, pp.27-30.

Oh, J. S.(2001), "Development of a Daily Traffic Volume Estimation Model Using Regression Analysis", *Journal of the Korean Official Statistics*, vol. 6, no. 2, pp.161-179.

Snowdon, J., Gkountouna, O., Züfle, A. and Pfoser, D.(2018), *Spatiotemporal Traffic Volume Estimation Model Based on GPS Samples*, Association for Computing Machinery.

Texas Transportation Institute(2002), *VEHICLE DETECTOR EVALUATION*, pp.1-77.

Yaghoubi, F., Catovic, Armin., Gusmao, A., Pieczkowski, J. and Boros, P.(2021), *Traffic Flow Estimation using LTE Radio Frequency Counters and Machine Learning*, Cornell University, pp.1-9.

Yoon, H. S.(2016), *AADT estimation of connection road section in highway using large-scale GPS trip data*, Seoul National University.