머신러닝을 활용한 버스 도착 시간 예측

김태진, 이영관, 태형만, 조유진, 김정홍, 박경규*
경북대학교 IT대학 컴퓨터학부,
㈜씨엘*
e-mail : taejin0323@gmail.com, dlghksah123@gmail.com,
better110@naver.com, dbwls2dpdy@gmail.com, jhk@knu.ac.kr,
zoology.park@ciel.co.kr

Prediction of Bus Arrival Time using Machine Learning

Tae-Jin Kim, Young-Gwan Lee, Hyeong-Man Tae, Yu-Jin Jo,
Jung-Hong Kim, Kyung-Gyu Park*
Dept of Computer Science and Engineering, Kyungpook National University,
Inc.Ciel*

요 약

버스는 1년간 약 62억명, 일평균 1천7백만명 이상 이용하는 우리나라 주요 대중교통 수단이다. 버스는 기차나 지하철과는 다르게 운행 정보가 시시각각 변하기 때문에 정확한 도착 시간을 예측이 어렵다. 효율적으로 시간을 사용하기 원하는 사람들은 버스의 정확한 시간정보를 알기를 원한다. 버스의 종류에는 시내버스 뿐만아닌 스쿨버스, 통근버스, 셔틀버스 등이 존재한다. 본 논문에서는 (주)씨엘에서 운영하는 셔틀 버스 운행정보를 기반으로 확률적 조건들을 반영한 머신러닝(Machine Learning)을 이용해 버스 도착 예정 시간을 알아낸다. 머신러닝으로 다층 퍼셉트론 모델을 사용해 1년치 데이터를 학습시켜 최대 2분 정도의 오차를 보이는 결과를 얻었다.

1. 서론

시간을 효율적으로 사용하기 위해서 사람들은 대중교통이나 셔틀버스의 정확한 도착정보를 알기 원한다. 전철 혹은 지하철과 같이 운행 시간표에 따른 대중교통은 정시성을 보장하기에 문제가 없지만 버스는 교통 상황, 신호등, 기상 상태 등과 같은 변수와 확률적 조건들에 의해 정시성을 보장하지 못한다.

기존 기술의 한계를 극복하여 셔틀 버스 도착 시각의 정시성을 보장하기 위해 관련 연구가 활발히 일어나고 기술들이 폭발적으로 발전하고 있는 인공지능 분야에서 방법을 찾고자 한다. 수집된 기존의 셔틀 버스 운행정보를 기반으로 한 머신러닝을 통하여 보다 정확한 버스 도착시간을 예측하고자 한다.

2. 기존 연구 동향

버스 도착시간 예측은 기존에도 많은 사람들에 의해 다음과 같이 연구되어 왔었다.

2.1 Historical Approach

이 유형은 동일한 기간의 이전 데이터에서 평균시간 혹은 평균속도를 분석해서 현재 및 미래의 이동시간을 예측한다. 예측 값이 신뢰성이 있기 위해서는 해당 지역의 교통 상황이 비교적 안정적 이여야 한다.

2.2 Real-time Approach

이 유형에서는 미래의 이동시간과 현재의 이동시간이 동일하다고 예측한다. 혼잡하거나 예측할 수 없는 교통 상황에서는 적용하기 힘들다.

2.3 Statistical Models

이 유형은 운전자의 습관, 교차로, 신호등과 같은 요인들을 독립적인 변수로 사용해 시계열 모형(time-series model) 혹은 회귀 모델(regression model)을 적용한다. 하지만 변수들 간의 관계를 연결하고 통합하는데 어려움이 있다.

2.4 Model-Based Approach

대표적으로 Kalman Filtering Model이 있다. 다른 모델에는 없는 동적 이동시간 예측을 제공하지만 측정 모델이 선형(linear)이고 가우스(Gaussian) 분포를 따라야 한다.

2.5 Machine Learning Models

최근 연구들에서는 복잡한 비선형 관계를 분석하고 예측하기 위해 인공신경망(Artificial Neural Network)을 활용한다.

표 1. 버스 도착시간 예측 방법 비교

방 법	요 약
역사적 접근	동일한 기간의 평균 이동시간을 분석해 특정 시간의 이동시간 예측
실시간 접근	현재 시간 간격의 이동시간과 다음 시간 간격의 이동시간이 동일하다고 가정한 예측
통계적 모델	독립 변수 집합에 의해 형성된 함수를 기반으로 버스 도착시간 예측
모델기반 접근 (Kalman filter)	Kalman filter model을 사용해서 위치와 이동시간을 예측
머신러닝 모델	대규모 데이터베이스를 학습하여 이동시간을 예측

3. 데이터 수집 & 정제

머신러닝 학습을 위한 데이터는 ㈜씨엘이 보유중인 셔틀버스 운행 데이터를 사용하며, 정제 과정은 아래와 같다.

3.1 원시 데이터

표 2. 차량 위치 정보 DB Schema

Feature Name	Description
idx	DB 인덱스
msg_type	메시지 타입
get_date	측위 시간
get_method	측위 방법 1=GPS, 2=CELL
lat	위도
lng	경도
dir	방향
speed	속도
dist	누적 거리
input_date	DB 등록 시간
line_no	노선 번호

표 3. 정거장 진출입 정보 DB Schema

Feature Name	Description
idx	DB 인덱스
line_no	노선 번호
inout_type	정거장 진/출입 타입 ENTER_BASE, LEFT_BASE, ENTERED, LEFT, LEFT_LAST
stop_no	정거장 번호
seq	정거장 순번
stop_name	정거장 이름
x	위도
y	경도
get_date	정류장 진/출입 시간
input_date	DB 등록 시간

(주)씨엘은 표 2.와 같은 데이터베이스 스키마로 2018-01-01부터 2018-12-31까지 3초마다 버스로 부터 운행정보를 수집해 월 단위로 평균 230,000개, 총 2,760,000개의 데이터를 수집했다.

또한 표 3.과 같은 데이터베이스 스키마로 똑같은 기간인 2018-01-01부터 2018-12-31 동안 9889개의 정거장 진출입 데이터를 수집했다.

3.2 가공 데이터

표 4. 정제한 데이터

Feature Name	Description
get_date	측위 시간
lat	위도
lng	경도
speed	속도
dist	남은 거리
line_no	노선 번호
y	최종 도착시간

버스 도착시간 예측을 머신러닝으로 학습시키기 위해 원시 데이터에서 표 4.와 같은 특성(feature)를 뽑아내었다. 지도학습을 위한 도착시간 값(y)는 정거장 진출입 정보 DB의 정거장으로 들어온 시간과 차량 위치 정보 DB의 측위 시간 값을 뺀 값으로 가공을 하였다.

남은거리(dist)는 경위도 값들을 계산하여 버스 노선길이에서 뺀 값들이다.

표 5. Sample Data

Feature Name	Description
get_date	2018-01-01 6:23
lat	35.21159
lng	129.0111
speed	54
dist	1095
line_no	13395
y	0:17:48

표 5.는 training data set 중 하나로, 실제로 학습을 하기 위해 넣은 데이터이다.

3.3 버스 운행정보 분석

총 6대의 버스 운행정보를 보유하고 있지만 대표적으로 하나의 버스 경로만 그림 1.과 같이 분석하였다. Python의 matplotlib와 pandas를 이용하여 월 별로 버스의 경로를 표시했다. 버스의 이동데이터에서 경도와 위도정보를 이용해 scatter로 경로를 푸른 선으로 표시하였고, 각 정류장 진출입정보를 이용하여 정류장위치를 붉은 점으로 표시하였다.

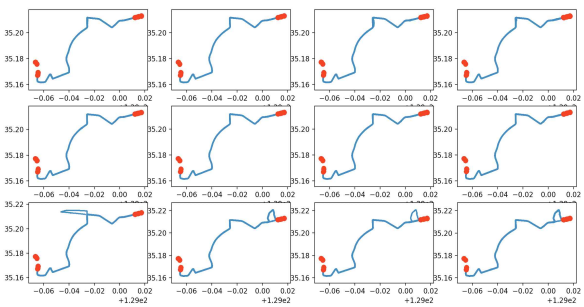

그림 1. 월별 버스 운행정보

4. 머신러닝 모델

머신러닝의 핵심구조 모델은 그림 2.와 같다. 2개의 입력 값인 input layer 1개, 예측시간인 output layer1개 그리고 hidden layer 3개로 총 5개의의 레이어로 구성하였다. 입력 계층에서 거리, 속도 데이터를 입력받는다.

숨겨진 계층에는 총 3개의 레이어로 Rectified Linear Unit(ReLu) 활성 함수를 사용한 출력뉴런(혹은 노드)이 32개인 레이어, 이전과 같은 종류의 출력뉴런이 16개인 레이어와 Linear 활성 함수를 사용한 뉴런 1개로 된 레이어로 구성되어있다. 마지막 출력 층에는 하나의 노드로 도착시간 예측 값을 출력한다.

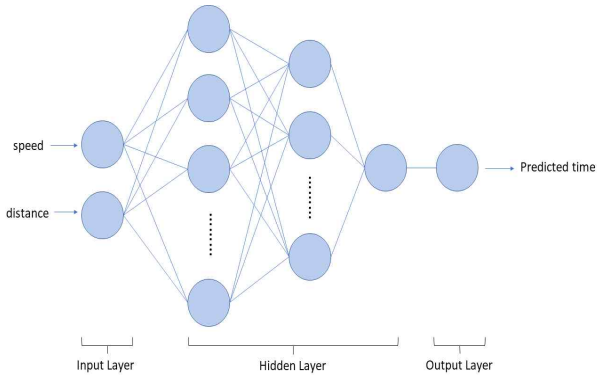

그림 2. 딥러닝 머신러닝 모델 구조

5. 구현 결과

가공된 데이터는 1월부터 10월까지는 훈련 세트(training set), 11월은 검증 세트(validation set) 그리고 마지막 12월은 시험 세트(test set)으로 구분하여 사용하였다. 앞서 제시된 머신러닝 모델을 텐서플로우(tensorflow)와 케라스(keras)로 구현하였다. 훈련 세트를 이 모델에서 100회 반복 학습시켰을 때 손실과 정확도가 그림 3.과 같이 나타나며 손실 값은 줄고 정확도는 증가하는 것을 볼 수 있다.

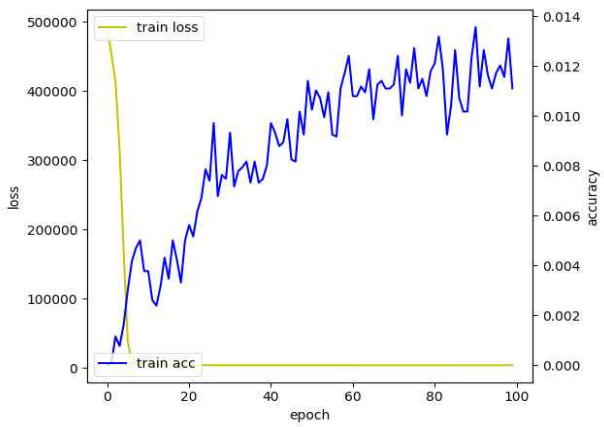

그림 3. train "loss" & "accuracy" 그래프

그림 4.는 matplot 라이브러리를 이용하여 실제 데이터의 분포를 시각화 하였다. 직관적인 이해를 돕기 위해서 본 논문에서 제시된 여러 층의 레이어를 가진 머신러닝 모델이 아닌 하나의 숨겨진 레이어만 가지는 모델을 같이 시각화하면 그림 4.의 중앙에 나타나는 직선으로 나타나게 된다. 이를 통해 머신러닝 모델이 실제 데이터 패턴과 유사하게 나타남을 알 수 있다.

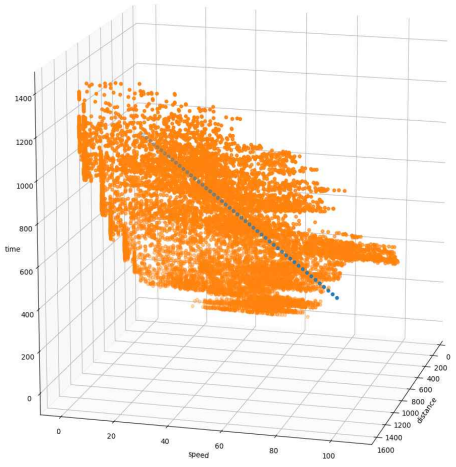

그림 4. 실제 데이터 패턴과 1차 머신러닝 모델의 상관관계 그래프

6. 결론 및 향후 연구

본 논문에서는 하나의 버스 운행정보만을 기반으로 과거 데이터를 학습시켜 버스 도착시간 예측 모델을 만들었다. 실제 데이터로 계산된 버스 도착시간과 머신러닝을 통해 예측된 시간 값에는 최대 2분의 오차를 보이고 있다. 약 8.32% 오차율을 보였다. (주)씨엘에서 수집한 다른 버스의 운행정보를 추가로 분석하면 머신러닝 모델의 정확도를 더 높일 수 있을 것으로 예상된다.

향후 연구로는 버스 도착시간에 영향을 줄 수 있는 변수들인 교통량, 강수량, 교차로 구간, 신호등 등을 적용한 새로운 머신러닝 모델을 구축하여 확장성을 넓혀갈 것이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음(2015-0-00912)

참고문헌

[1]M. Zaki, I. Ashour, M. Zorkany, B. Hesham, "Online Bus Arrival Time Prediction Using Hybrid Neural Network and Kalman filter Techniques", IJMER, Vol.3, Issue.4, Jul-Aug. 2013 pp-2035-2041

[2] Mehmet Altinkaya. Metin Zontul, "Urban Bus Arrival Time Prediction: A Review of Computational Models", IJRTE, Volume-2, Issue-4, September 2013

[3] 고두환, 황순연, "우리나라 대중교통요금 추이와 국가간 요금 비교" Vol. 29

[4] J.Patnaik, S.Chein,and A.Bladihas "Estimation of Bus Arrival Times Using APC Data", Journal of Public Transportation, Vol. 7, No. 1, pp.1-20, 2004