한국ITS학회논문지 J. Korea Inst. Intell. Transp. Syst.

교통 빅데이터 활용 시 개인 정보 보호를 위한 연합학습 기반의 경로 선택 모델링

Federated Learning-based Route Choice Modeling for Preserving Driver's Privacy in Transportation Big Data Application

심 지 섭*

* 주저자 및 교신저자 : 국토연구원 국토인프라·공간정보연구본부 부연구위원

Jisup Shim*

* National Infrastructure & Geospatial Information Research Division, KRIHS

† Corresponding author : Jisup Shim, gis.up@krihs.re.kr

Vol. 22 No.6(2023)
December, 2023
pp.157~167

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2023.22.6.157

Received 31 October 2023
Revised 22 November 2023
Accepted 13 December 2023

요 약

본 연구에서는 분산 컴퓨팅 및 개별 디바이스 활용을 통해 개인 정보 보호에 특화된 학습 방법인 연합학습 방법론을 기반으로, 모바일 내비게이션 애플리케이션에서 수집된 대규모의 운전자 데이터를 이용하여 경로 선택 예측 모델을 수립하는 방법에 대해 고찰한다. 경로 선택 모델링에서 활용될 수 있는 운전자 데이터의 전처리 및 분석 방법을 수립하고, 서포트벡터머신(SVM) 및 다층 퍼셉트론(MLP)과 같이 기존에 널리 활용되는 학습 방법과 연합학습 방법의 성능과 특성을 비교한다. 분석 결과 연합학습을 통한 모델 성능은 중앙 서버 기반의 모델과의 비교에서 예측 정확도 측면의 차이가 거의 없는 것으로 나타났으나, 개별 데이터가 충분히 확보되는 경우 연합학습 모델과 같은 개인화 모델의 성능이 개선될 수 있다는 점을 확인하였다. 연합학습 모델은 본 연구의 경로 선택 모델링 사례와 같이 모빌리티 부문의 데이터 프라이버시 문제가 중요한 분야에서 대규모 데이터 처리를 필요로 하는 경우에 그 활용 가치가 매우 높을 것으로 기대된다.

핵심어 : 연합학습, 경로선택 모델링, 교통 빅데이터

ABSTRACT

The use of big data for transportation often involves using data that includes personal information, such as the driver's driving routes and coordinates. This study explores the creation of a route choice prediction model using a large dataset from mobile navigation apps using federated learning. This privacy-focused method used distributed computing and individual device usage. This study established preprocessing and analysis methods for driver data that can be used in route choice modeling and compared the performance and characteristics of widely used learning methods with federated learning methods. The performance of the model through federated learning did not show significantly superior results compared to previous models, but there was no substantial difference in the prediction accuracy. In conclusion, federated learning-based prediction models can be utilized appropriately in areas sensitive to privacy without requiring relatively high predictive accuracy, such as a driver's preferred route choice.

Key words : Federated Learning, Route Choice Modeling, Transportation Big Data

Ⅰ. 서 론

1. 연구 개요

연합학습(Federated Learning, FL)의 개념은 개별 데이터 수집 장치의 충분한 활용을 위해 2017년 구글을 통해 최초로 제안되었다(McMahan and Ramage, 2017). 일반적으로 분산 컴퓨팅에 기반한 데이터 처리 방법론과 애플리케이션들은 주로 추천 시스템, 자연어 처리 등과 같은 데이터 집약적 처리가 필요한 분야를 위해 활용되는데, 이러한 시스템 내에서 개인 정보와 같은 중요 데이터는 휴대폰 및 개인용 컴퓨터 등 말단 사용자의 로컬 장치에 우선 저장된다. 이 때, 대부분의 빅데이터 및 기계학습 기반 연구 방법론은 사용자의 모든 데이터를 중앙 서버에 집약하여 전체 데이터를 활용한 예측 모델 등을 수립한다. 반면 연합학습 기반 모델링은 말단 사용자의 로컬 장치에서 수집된 원시 데이터를 서버에 업로드하지 않고, 디바이스 간 정보 전달 역시 필요로 하지 않는다. 즉, 모든 데이터는 각 로컬 장치에서 데이터베이스화되며, 이는 전체 정보에 대한 중앙 서버의 접근 용이성을 저하시키므로 개인 정보 보호의 측면에서 훨씬 안전하고 효율적인 방식이 될 수 있다. 본 연구에서는 상기한 연합학습 방법론의 특성을 기반으로, 모바일 내비게이션 애플리케이션에서 수집된 대규모의 운전자 데이터를 이용하여 경로 선택 예측 모델을 수립하는 방법에 대해 고찰하고자 한다.

2. 선행 연구 사례

인공지능(Artificial Intelligence, AI) 모델 학습을 위한 연합학습 방법론은 데이터에 포함된 개인 정보의 민감도가 높은 분야에서 주로 활용해 왔다. 예를 들어, 인공지능을 이용한 원격 진료 등 의료 서비스를 위해 수집되는 개인의 데이터를 활용하는 연구(Brisimi et al., 2018), 모바일 기기에 입력한 단어 데이터를 이용해 자연어 처리 기반의 모델을 수립하는 연구(Hard et al., 2018; Chen et al., 2019), 인공지능 기반 추천 시스템을 위해 사용자의 구매 및 조회 이력을 이용하는 연구 등에 연합학습 방법론의 적용이 적합하다(Yang et al., 2020). 이러한 연합학습 기반 인공지능 연구의 최대 장점 중 하나는 다수가 참여하는 온라인 시스템 내에서의 개인 정보 보호 관련 문제들을 방지할 수 있다는 점이다.

교통 빅데이터의 활용에서도 운전자의 개인 정보 보호는 매우 중요한 문제 중 하나이다(Tian et al., 2022). 지능형 교통체계(ITS) 및 모바일 디바이스를 통한 교통 정보 수집체계가 확장되고, 운전자 궤적정보(Trajectory Data)를 비롯한 다양한 교통 관련 정보가 실시간으로 취득됨에 따라 개인 정보가 포함된 데이터를 다루는 방법론에 있어 세심한 주의가 필요해졌다. 현재까지 교통 분야에서 연합학습이 활용된 사례는 타 분야에 비해 드문 편이지만, 교통 데이터 내의 개인 정보 보호 중요성과 관심도가 점차 높아지면서 교통류 예측 시의 개인 정보 보호를 위한 연합학습의 도입(Liu et al., 2020), 대중교통 데이터 분석 시의 연합학습 적용 방법(Xu et al., 2022), 자율주행차량 통신 데이터와 같은 C-ITS 분야 내 연합학습 활용(Manias and Shami, 2021) 등 다양한 세부 분야에서의 적용 연구 사례가 나타나고 있다.

3. 연구의 필요성

인공지능 기반의 개인 맞춤형 모델은 삶의 편의성과 효율성을 높여주는 중요한 도구로 인식되고 있다. 그러나 인공지능의 발전은 방대한 데이터 처리를 기반으로 하는 까닭에 개인이 발생시키는 다양한 자료를 수집·활용해야만 하고, 이에 따라 개인 정보 보호 문제가 인공지능 기술 도입의 피할 수 없는 과제 중 하나로

여겨지고 있다. 최근에는 모빌리티 분야에서 수집되는 데이터와 인공지능 기술 역시 동일한 문제에 직면해 있다. 모바일 내비게이션을 통해 수집되는 경로 선택 정보, 운전자가 실제 주행한 이동 경로(GPS) 정보, 유료도로 통행 시 부과되는 결제 정보 등 개인의 이동과 관련된 정보는 매우 민감한 개인 정보들을 포함하는데, 이러한 유형의 데이터를 잘못 관리하면 개인의 프라이버시를 침해할 우려가 크다. 따라서 교통 빅데이터 분석 및 활용 시 대규모의 데이터를 처리할 수 있으면서도 개인 정보 보호 문제를 선결할 수 있는 효과적인 방법을 찾는 것은 중요한 과제 중 하나이다.

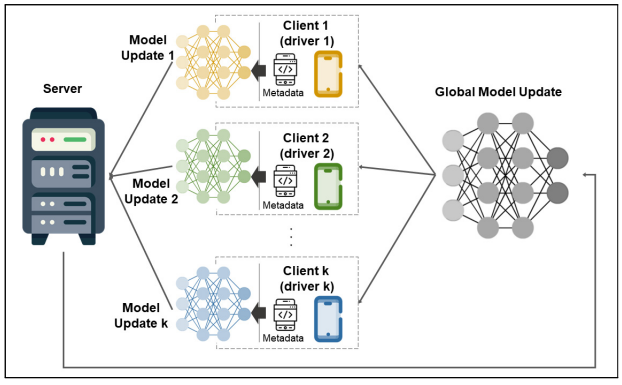
4. 연구의 목적

본 연구의 목적은 현실 세계에서 대규모로 수집되고 있는 내비게이션 정보 및 운전자의 경로 선택 데이터를 효율적인 활용 방법을 찾는 것이다. 스마트카드 및 각종 센서 자료 등 여러 분야에서 이미 다양하게 활용되고 있는 교통 빅데이터들과 달리, 내비게이션이나 운전자 데이터는 그 규모가 방대하고 잠재적인 활용 가치가 있음에도 불구하고 활용처를 모색하기 어려웠다. 아울러 전술한 바와 같이 운전자 주행 데이터는 민감정보를 포함하기 때문에 데이터의 분석·활용 시 개인 정보 보호 문제를 해결하기 위한 방법을 찾는 것이 중요한 문제로 여겨져 왔다. 본 연구에서는 내비게이션 데이터의 구체적인 활용법을 모색하기 위해 해당 데이터의 전처리·분석을 통한 모델 학습 방법을 제시하고, 특히 연합학습 방법론의 적용을 통해 개인 정보의 과다한 수집이나 서버 적재 없이 학습 모델을 구축하는 방법을 제안하고자 한다.

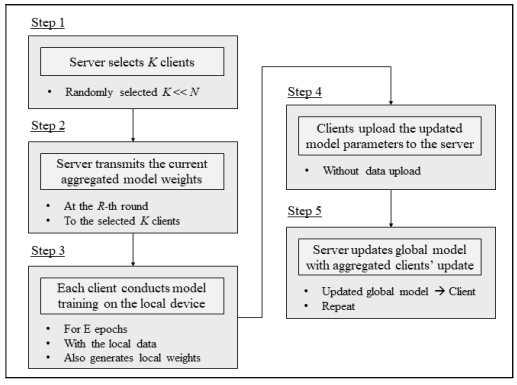
Ⅱ. 연합학습 방법론 및 데이터 전처리

1. 연합학습(Federated Learning) 개요

연합학습은 스마트폰과 같은 모바일 장치를 이용하여 로컬 수준의 기계학습 모델을 먼저 구성하고, 각 장치에서 추출된 로컬 기계학습 모델의 매개변수만을 서버로 전송하여 전역적인 기계학습 모델을 구축한다. 이 때, 로컬 기기에서 수집된 원본 데이터는 서버에 전달하지 않는다(<Fig. 1>).



<Fig. 1> Architecture of Federated Learning



<Fig. 2> FedAvg Algorithm

연합학습 방법론에서의 학습 방법은 주로 FedSGD와 FedAVG 방식으로 구분된다(McMahan et al., 2017).

FedSGD는 로컬 기기에서 학습하는 SGD(Stochastic Gradient Descent, 확률적 경사 하강법) 기반의 학습 방식을 연합학습 프레임워크에 적용시킨 것이다. FedSGD를 적용하는 연합학습 프레임워크에서의 핵심 하이퍼파라미터는 전체 클라이언트 숫자 대비 활용되는 기기의 임의 비율을 나타내는 C (Random fraction of clients)이다. 즉 클라이언트의 비율이 전역 모델의 배치 크기(batch size)로 여겨진다. 반면 FedAVG는 FedSGD를 개선한 방법론으로서, 각 사용자의 모델이 독립적이면서 보다 효율적으로 연합학습 프레임워크를 구성하기 위한 방법이다. FedSGD에서 하이퍼파라미터 C만 설정하는 것과 달리, FedAVG에서는 세 가지 하이퍼파라미터 C (Random fraction of clients), E (local epoch size), B (local batch size)를 설정한다. 즉 클라이언트의 비율, 로컬 업데이트에 사용한 로컬 에포크 크기(E) 및 로컬 배치 크기(B)에 대한 하이퍼파라미터를 설정하는데, 여기서 C는 전역(global) 모델의 배치 크기로 간주되고, E는 로컬 모델 학습의 반복 횟수를 나타낸다.

본 연구에서는 운전자의 경로 선택 행동 예측을 위한 연합학습 프레임워크로서 FedAvg 방법을 적용한다. FedAVG 방법에서는 먼저 <식 (1)>에 나타난 바와 같이 N개의 클라이언트(i.e.모바일 기기)가 존재하는 연합학습 시스템에서 각 라운드마다 모델 훈련에 참여할 K개의 클라이언트를 무작위로 선택한다. 즉 이 단계에서 하이퍼파라미터 C를 설정하고, $K = NC$ 가 된다. R번째 라운드에서, 서버는 선택된 K개의 클라이언트들에게 현재 모델의 가중치인 w_R을 전송한다. 각 클라이언트는 E번의 에포크(epoch)로 모델 트레이닝을 수행하되, 이 때 각 로컬 기기의 데이터를 사용하여 가중치를 업데이트하고 해당 클라이언트 k에서의 w_{R+1}^k을 계산한다. 클라이언트들은 다시 업데이트된 모델 매개변수를 서버에 업로드하고, 서버는 각각의 클라이언트들에서 업데이트된 매개변수를 통합하여 전역 모델의 가중치 w_{R+1}을 계산한다. 이 때 n_k와 n은 각각 클라이언트 k에서의 데이터 포인트 숫자와 전체 데이터 포인트의 숫자를 나타낸다. 본 연구에서는 FedAVG의 하이퍼파라미터를 $C = 0.02$, $B = 100$, $E = 5$로 설정하였다.

$$w_{R+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_{R+1}^k \quad \cdots\cdots (1)$$

한편 연합학습 기반의 학습 모델 설계 과정에서는 사용자 및 데이터 수와 관련한 몇 가지 고려 사항이 존재한다. 연합학습의 방법에는 본 연구에서 사용된 내비게이션 데이터와 같이 수많은 사용자를 대상으로 하는 Cross-device 방식과 소수의 신뢰 높은 사용자를 대상으로 하는 Cross-silo 방식이 존재한다(Kairouz et al., 2019). Cross-device 방식은 사용자 및 데이터 수가 매우 방대하기 때문에 비용이 많이 들고, 불특정 다수의 사용자로 인한 신뢰성 문제가 발생할 수도 있다는 단점이 존재한다. 반면 Cross-silo 방식은 표본의 수가 줄어든다는 제한점이 존재하나 학습 모델의 오류 가능성 적기 때문에 의료기관·금융기관과 같이 적은 수의 사용자가 높은 수준의 개인 정보 보호와 신뢰성을 요하는 경우에 적합하다.

본 연구에서 다루는 연합학습은 Cross-device 방식을 상정하고 있으므로 발생할 수 있는 몇 가지 문제점에 대한 고려가 필요하다. 데이터 가용성의 측면에서 특정 순간에 일부 클라이언트만 가용 가능한 제한적 상황, 각 클라이언트의 네트워크 환경에 따른 정보 유실 상황, 신뢰성이 떨어지는 일부 유저의 악의적인 공격 상황 등이 그 예이다. 이러한 경우에 발생할 수 있는 문제로는 기기 및 통신 상황에 따라 업데이트된 모델 적용에 지연이 발생하거나, 일부 기기에서 발생하는 오염된 데이터(Data poisoning) 등으로 인해 학습 모델 자체에 성능 저하가 발생할 수 있다는 점이다. 따라서 연합학습 방법을 통해 수립된 모델이 모든 사용자에게 실시간 업데이트 정보를 제공하는 데에는 어려움이 따르며, 데이터 정화(Data sanitization) 기법 등을 통해 연합학습 모델의 성능 및 신뢰성이 확보하는 방법이 지속적으로 개발되고 있다(Bagdasaryan et al., 2020; Liu et al., 2018).

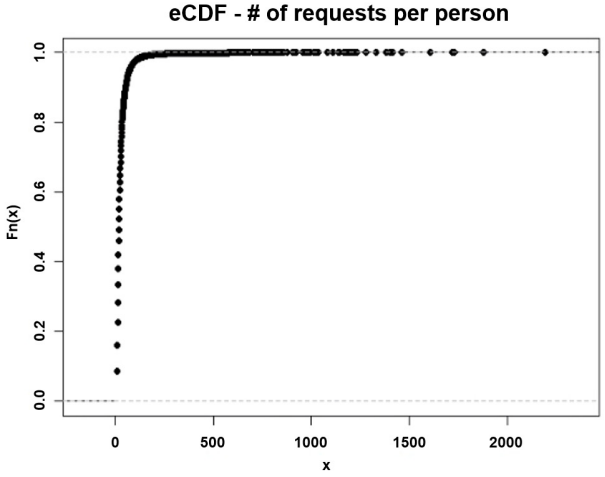
2. 연구 데이터

1) 데이터 명세

상용 모바일 내비게이션에서 수집된 운전자의 경로 선택 및 이동궤적 자료는 개인의 사생활 정보가 포함된 매우 민감한 자료이다. 이에 따라 본 연구에서는 연합학습 기반의 기계학습 방법론을 사용하여 운전자의 개별 데이터가 서버로 전송되지 않는 학습 모델을 구축한다. 다만 본 연구에서 사용한 기수집 자료의 경우 중앙 데이터 서버에 저장되어 있던 자료이며, 연합학습 모델 수립 시 해당 데이터는 각 운전자의 모바일 장치에 저장되어 있다는 점을 상정하고 분석을 수행하였다.

본 연구에서는 상용 모바일 내비게이션 서비스인 카카오 내비를 통해 수집된 운전자의 경로 선택 자료를 이용하여 운전자의 경로 선택 모델을 수립하였다. 데이터 수집 기간은 Covid-19의 영향 시기 이전인 2019년 7월 17일 부터 8월 23까지 총 40일이며, 출발지 혹은 도착지가 대구시, 제주시, 분당구 내에 포함되는 내비게이션 길 안내 데이터를 지역적 범위로 국한하였다. 해당 시공간 범위 내 데이터는 전체 운전자 수 907,003명으로부터 발생한 8,335,687건의 길안내 요청을 포함하며, 40일간 1일 평균 9회의 길 안내 요청을 발생시켰다. 전체 데이터는 운전자의 경로 선택 데이터 및 각 경로 내 세부 링크 데이터를 포함하고 있어, 약 1.8TB의 대용량 데이터를 분석 대상으로 한다.

<Fig. 3>은 한 명의 운전자가 요청하는 길 안내 횟수에 따라 경험적 누적분포함수(eCDF) 및 도수분포표를 나타낸 것이다. 분석 기간 내 길 안내 요청 데이터 중 전체의 78.3%에 해당하는 710,424명의 운전자가 10건 이하의 길안내 요청을 발생시키고, 89.4%에 해당하는 810,918명이 20건 이하의 요청을 발생시키는 것으로 나타났다. 길 안내 요청을 정기적으로 발생시키는 운전자(54건 이상, 데이터 수집기간인 40일 중 주말 및 휴일을 제외한 27일×2)는 약 34,000명으로 전체 운전자의 약 2.3%를 차지한다.



Class Interval	Frequency
[0, 10]	710424
(10, 20]	100494
(20, 30]	41526
(30, 40]	20288
(40, 50]	11260
(50, 60]	6857
(60, 70]	4591
(70, 80]	2872
(80, 90]	1964
(90, 100]	1401
(100, 110]	951
(110, 120]	795
⋮	

<Fig. 3> Empirical CDF for the number of trips by user

2) 데이터 전처리

경로 선택 데이터는 데이터 취득 시점을 기준으로 [운전자 식별 ID], [출/도착지 위치정보], [추천경로의 예상소요시간, 거리, 통행요금], [대안경로의 예상소요시간, 거리, 통행요금] 정보를 포함한다. 현재 위치에서 운전자가 목적지를 설정하면, 해당 출발지-목적지에 해당하는 두 가지 추천경로를 표시한다. 추천경로와 대안

경로의 예상소요시간, 거리, 통행요금은 다른 경우가 대부분이나 일부 동일한 경우도 발생한다. 요청 직후에는 기본 옵션으로 추천경로가 선택되어 있고, 운전자가 의식적으로 경로를 선택하지 않는 경우엔 자동으로 추천경로로 길 안내를 수행한다. 본 연구에서는, 추천경로와 대안경로 중 하나의 경로를 운전자가 의식적으로 선택하는 경우를 상정하기 위해, 데이터 수집 기간 동안 2회 이상 경로 요청을 한 운전자 중 적어도 한 번 이상 대안경로를 선택한 데이터만을 필터링하였다. 아울러 전체 기간 중 대안경로를 선택하는 전체 비율이 10% 미만이거나 90% 이상인 경우 역시 이상값으로 처리하고 제외하였다. 즉, 추천경로와 대안경로를 의식적으로 선택하는 운전자는 전체 기간동안 각각의 경로를 모두 선택한 적이 있어야 하며, 한 가지 경로에 대한 선택 비율이 과도하게 높은 경우는 분석에서 제외되었다(Sun et al., 2020; Sun et al., 2023). 필터된 데이터에서 운전자가 추천 경로를 선택한 비율은 약 76%, 대안경로를 선택한 비율은 약 24% 였다.

3. 학습 모델 및 훈련

1) 데이터셋 구성 및 학습 모델

모델 훈련을 위한 학습 데이터는 아래 <Table 1>에 정리된 바와 같이, 추천경로와 대안경로에 따른 예상소요시간, 거리, 통행요금 정보와 이를 기반으로 계산된 두 경로의 특성을 나타내는 항목을 포함한다.

<Table 1> Data scheme for model training

Attribute	Type	Description
Time_Default	Numeric	travel time / fare / distance of default route
Fare_Default		
Dist_Default		
Time_Alter		travel time / fare / distance of alternative route
Fare_Alter		
Dist_Alter		
Time_Per		the ratio of the travel time / fare / distance of the default route over that of the alternative route
Fare_Per		
Dist_Per		
Time_Diff		difference between default and alternative route
Fare_Diff		
Dist_Diff		
Average_Dist		min-max normalized trip distance for distinguishing long/short trip [0,1]
Day_of_Week	Categorical	Monday: 0 to Sunday: 6
Time_of_Day		time of day when the route guidance request occurred (0-23)
Morning_Peak	Binary	Morning Peak: 1 (7AM to 9AM) / Otherwise: 0
Evening_Peak		Evening Peak: 1 (5PM to 7PM) / Otherwise: 0

경로의 특성을 나타내는 항목은 크게 두 가지로 구분되는데, 추천경로 대비 대안경로의 시간, 거리, 요금 비율 값 및 차이값을 이용하여 학습 데이터를 구성한다. 또한 통행 선택 시 장거리-근거리 통행을 구분하기 위해, 하나의 출발지-목적지에 대한 두 개 경로의 평균 거리를 전체 데이터셋의 최대-최소값을 기준으로 정규화하는 변수도 추가하였다. 그 외 통행 선택에 영향을 미칠 수 있는 통행 요일 및 시간대, 첨두 시간 여부

도 범주형 변수로 구분하였다.

요일, 시간대, 오전/오후 첨두시간 여부 등의 범주형 변수에 대해서는 원핫인코딩(one-hot-encoding)을 수행하였다. 그 외 수치형 변수에 대해서는 평균을 0, 표준편차를 1로 가지도록 표준화(stnadardization) 하였다. 전술한 데이터 전처리 방법을 통해 필터링된 데이터셋을 학습 및 검증용 데이터로 분리할 때에는 Train/Test/Validation 데이터셋의 비율을 6:2:2로 하였다.

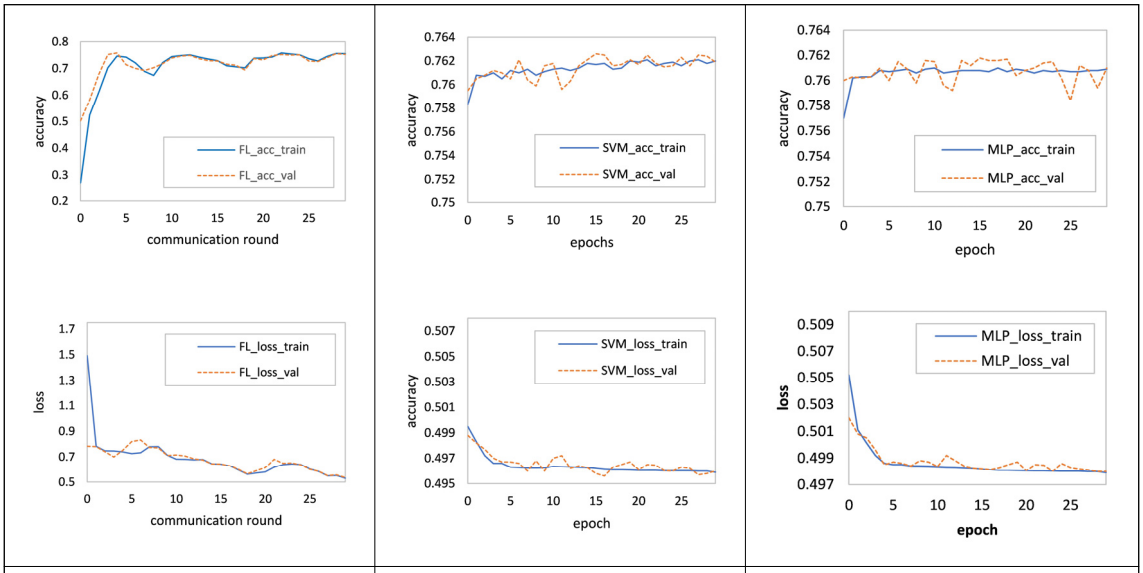
Ⅲ. 학습 모델별 경로 선택 모델링 성능 평가

본 연구에서 제시하는 연합학습 모델(FL)과 기존의 서버 기반 학습 모델의 성능 및 차이점을 분석하기 위해 비교 모델로서 서포트 벡터 머신(Support Vector Machine, SVM), 다층 퍼셉트론(Multi-Layer Perceptron, MLP)을 활용하였다. SVM은 경로 선택 모델링에서 자주 사용되는 머신러닝 기법으로서, 운전자의 행동이나 선택을 모델링하는데 좋은 성능을 보여주는 것으로 알려져 있다(Sun and Park, 2017; Yuksel and Atmaca, 2021; Zhang and Xie, 2008). 경로 선택 모델링에서의 SVM 활용은 운전자 전체를 활용하는 통합(Aggregate) 모델과 개별(Individual) 모델로 구분될 수 있는데, 본 연구에서의 SVM 학습 모델은 서버 기반 학습 모델의 프레임워크 구조를 차용하기 위해 Aggregate SVM 모델을 적용하였다. 즉, 모든 운전자의 데이터가 동일한 학습데이터셋 안에 포함되어 모델 학습을 수행한다. 최종적으로는 Aggregate SVM 모델을 통해 운전자의 입력 데이터에 따른 경로 선택 결과를 이진 분류하여 예측할 수 있다. MLP에서도 서버 기반 학습 모델 및 이진 분류 모델을 구조화할 수 있도록 하이퍼 파라미터를 설정하였다(Kweon et al., 2021). 본 연구에서 적용된 서버 기반 학습 모델 유형의 MLP 모델은 16개 노드로 구성된 2개의 히든레이어와 하나의 출력레이어를 구성하고, 활성화 함수로 시그모이드를 사용하였으며, 경사 하강 알고리즘으로는 Adam, 학습률(learning rate) = 0.001, epochs = 100, batch = 128로 설정하였다.

<Table 2>와 <Fig. 4>는 FL 모델과 SVM, MLP 모델을 적용한 경우의 예측 성능 및 오차를 비교하여 보여주고 있다. 예측 성능 평가를 위해 데이터셋에서 무작위 선택된 데이터를 활용하는 과정을 여러 번에 걸쳐 시행하였으며, <Table 2>에 나타난 결과는 그 평균 값을 나타낸다. FL 모델의 예측 성능은 75.8%로, 각각 76.2%와 76.1%의 예측 성능을 보이는 SVM과 MLP에 비해 낮게 나타난다. 즉 로컬 디바이스를 많이 활용하는 연합학습 모델은 중앙 집중형 서버 기반 학습 모델에 비해 예측 성능 면에서 우위를 점하지 못하는 양상을 보인다. 그러나 이를 달리 해석하면, 연합학습 방법은 다른 학습 모델과 비교했을 때 개인 정보 보호에 강점을 가지면서도 예측 성능 면에서 크게 뒤떨어지지 않는다는 점을 확인할 수 있다.

<Table 2> Comparison of model performances

	FL (Federated Learning)	SVM	MLP
Accuracy	0.758	0.762	0.761
Loss	0.534	0.496	0.498



<Fig. 4> Training(train)-validation(val) loss and accuracy over the epochs from each model

한편, 연합학습 방법론의 개념을 차용한 아이디어로 Aggregated SVM 모델과 Individual SVM 모델을 동시에 활용하는 학습 모델을 생각해 볼 수 있다. 본 연구에서는 해당 모델들의 적용 성능을 직접 비교하지 않았지만, Sun et al.(2023)의 연구에서는 본 연구에서 활용한 데이터와 동일한 데이터를 가지고 Individual SVM과 Aggregated SVM 모델의 성능을 비교한 바 있다. 해당 연구에서는 Individual SVM 모델을 이용하는 경우 데이터 수집 규모에 따라 큰 폭의 예측 성능 저하가 발생하는 점을 지적하였다. Aggregated SVM 모델이 데이터의 크기와 관계 없이 일정한 수준의 모델 예측 성능을 보이는 반면, Individual SVM 모델은 제한된 양의 개인 데이터 중에서도 일부 데이터를 이용한 학습을 수행하기 때문에 모델 성능이 저하되는 특성을 보였다. 즉 개인의 주행 이력 데이터가 충분하지 않아 학습 시 테스트 데이터 셋에 새로운 시나리오가 포함되는 경우가 발생하면 Individual SVM 모델의 유용성이 떨어지게 된다. 본 연구를 통해 도출된 결과에서 연합학습 모델의 성능이 다른 전역 모델에 비해 다소 낮게 나오는 점도 이러한 모델 구조 및 데이터 수집상의 한계점과 결부하여 해석할 수 있다.

Ⅳ. 결 론

본 연구에서는 교통 빅데이터 활용 시의 개인 정보 보호 문제를 고려한 인공지능 기술 도입을 위해 연합학습을 활용하는 방법을 고찰하였다. 연합학습은 분산 컴퓨팅 기반 학습 방법론 중 하나로, 원시 데이터의 전송을 필요로 하지 않는다는 점에서 개인 정보 보호에 특히 큰 장점을 갖고 있다. 따라서 운전자의 경로 선택 문제와 같이 민감한 데이터(GPS 궤적, 출발지·목적지 정보 등)를 포함하는 분야에 활용하기 적합하다. 이에 본 연구에서는 FedAVG 기반의 연합학습 방법론을 운전자 경로 선택 모델링 분야에 적용하기 위해 필요한 데이터 전처리 과정을 고찰하고, 기존의 중앙 서버 집약형 학습 모델 등과 비교하여 그 성능과 장·단점을 비교 분석하였다.

분석 결과 연합학습 기반의 학습 모델은 전체 데이터를 한 번에 이용하는 글로벌 모델과 비교하여 예측 성능 면에서의 우위를 보이지는 못하였다. 그 이유 중 하나는 개개인의 주행 이력 데이터가 충분하지 못한 상황에서 개인화 모델을 구축하는 경우, 학습 모델의 테스트 셋에 예측하기 어려운 시나리오가 포함될 수 있기 때문이다. 이는 3장의 결과에서 보인 바와 같이 학습의 반복 횟수가 적을 때 다른 두 모델에 비해 예측 성능이 낮은 연합학습의 특성을 통해서도 확인할 수 있다. 따라서 모델의 실시간 적용 등 빠른 업데이트가 필요하거나, 로컬 장치 및 데이터 리소스의 제한이 있는 경우 연합학습 적용은 적절하지 않을 수 있다. 즉, 연합학습과 같은 개인화 모델의 예측 성능을 제고하기 위해서는 충분한 양의 개별 데이터 확보가 필수 요소이다. 그러나 전반적인 예측 성능에 있어 연합학습 모델과 기존 모델 간의 정확도는 1% 미만의 차이를 보였으며, 이는 향후 개별 데이터의 수집 규모가 증가할수록 개인화 모델의 정확도가 전역적 모델에 비해 증가할 수 있음을 의미한다(Sun et al., 2023).

아울러 본 연구에서 도출된 결과 해석 시에는 몇 가지 유의점이 존재한다. 추천경로 대비 대안경로의 선택 집합을 정의하는 데 있어 본 연구에서 사용한 카카오 내비게이션 알고리즘의 경로 샘플링(path sampling) 규칙이 일정하다는 가정을 필요로 한다. 대안경로의 선택 집합이 일정하게 생성되지 않는 경우, 경로 선택 모형의 추정 및 학습 모델 상에 왜곡이 발생할 수 있기 때문이다. 나아가, 연합학습은 대부분의 경우 데이터 이질성(Data Heterogeneity) 문제로 인해 우수한 정확도 성능을 달성하는 것이 어렵다. 즉, 각 사용자의 다양한 사회인구학적 특성 및 시공간의 이질성 등으로 인해 클라이언트가 생성한 훈련 데이터는 불균형하며, 독립적이고 동일하게 분산되는 데이터 성질(IID: Independent Identically Distributed)을 갖지 못한다(Non-IID). 연합학습 내 각 클라이언트의 데이터 이질성은 전체 데이터의 수렴 속도를 느리게 할 뿐만 아니라 정확도 성능 자체를 떨어뜨린다. 동적 시공간 환경에서 이러한 데이터 이질성 문제를 해결하기 위해서는 데이터 수에 따라 학습에 사용하는 데이터의 분포를 조정하는 다양한 데이터 엔지니어링 기술을 통해 훈련 데이터를 필터링하거나(Alain et al., 2015; Katharopoulos and Fleuret, 2018; Loshchilov and Hutter, 2015), 데이터 숫자에 따라 과대·과소 샘플링을 통해 데이터의 분포를 조정하는 방법을 수행할 수 있다(Sarkar et al., 2020).

본 연구에서 도입한 연합학습 기반 학습 모델의 가장 큰 장점은 앞서 여러 차례 강조한 바와 같이 개인정보 보호 측면에 있다. 연합학습 방법론은 본 연구와 같이 경로 선택 모델링 문제에서의 데이터 프라이버시 문제를 해결할 수 있고, 분산 컴퓨팅을 통한 대규모 데이터 처리가 용이하다는 점에서 향후 활용 가치가 매우 높다. 향후 연구에서는 개별 데이터의 확대 적용을 통한 모델 성능 제고뿐만 아니라, 소수의 신뢰성 있는 집단을 대상으로 하는 Cross-silo 방식의 모델 수립을 통해 개인화에 더욱 초점을 맞춰 예측 모델의 성능을 높이는 방법을 고려해 볼 수 있다. 또한 필드 테스트를 통해 지연 시간 감소 측면에서의 연합학습 성능 등을 실측하는 연구도 수행되어야 할 필요가 있다.

REFERENCES

Alain, G., Lamb, A., Sankar, C., Courville, A. and Bengio, Y.(2015), *Variance reduction in sgd by distributed importance sampling*, arXiv preprint arXiv:1511.06481.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D. and Shmatikov, V.(2020), "How to backdoor federated learning", *International Conference on Artificial Intelligence and Statistics, PMLR*, pp.2938-2948.

Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. C. and Shi, W.(2018), "Federated

Learning of Predictive Models from Federated Electronic Health Records", *International Journal of Medical Informatics*, vol. 112, pp.59-67.

Chen, M., Mathews, R., Ouyang, T. and Beaufays, F.(2019), *Federated learning of out-of-vocabulary words*, arXiv preprint arXiv:1903.10635.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S. and Ramage, D.(2018), *Federated learning for mobile keyboard prediction*, arXiv preprint arXiv:1811.03604.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N. and Oliveira, R. G.(2019), *Advances and open problems in federated learning*, arXiv preprint arXiv:1912.04977.

Katharopoulos, A. and Fleuret, F.(2018), "Not all samples are created equal: Deep learning with importance sampling", *International Conference on Machine Learning*, pp.2525-2534.

Kweon, Y., Sun, B. and Park, B. B.(2021), "Preserving privacy with federated learning in route choice behavior modeling", *Transportation Research Record*, vol. 2675, no. 10, pp.268-276.

Liu, K., Dolan-Gavitt, B. and Garg, S.(2018), "Fine-pruning: Defending against backdooring attacks on deep neural networks", *International Symposium on Research in Attacks, Intrusions, and Defenses*, Springer International Publishing, pp.273-294.

Liu, Y., James, J. Q., Kang, J., Niyato, D. and Zhang, S.(2020), "Privacy-preserving traffic flow prediction: A federated learning approach", *IEEE Internet of Things Journal*, vol. 7, no. 8, pp.7751-7763.

Loshchilov, I. and Hutter, F.(2015), *Online batch selection for faster training of neural networks*, arXiv preprint arXiv:1511.06343.

Manias, D. M. and Shami, A.(2021), "Making a case for federated learning in the internet of vehicles and intelligent transportation systems", *IEEE Network*, vol. 35, no. 3, pp.88-94.

McMahan, B. and Ramage, D.(2017), "Federated learning: Collaborative machine learning without centralized training data", *Google Research Blog*, 3.

McMahan, B., Moore, E., Ramage, D., Hampson, S., Blaise AgueraAg, H. and Arcas, A.(2017), "Communication-Efficient Learning of Deep Networks from Decentralized Data", *20th International Conference on Artificial Intelligence and Statistics*, vol. 54, pp.1273-1282.

Sarkar, D., Narang, A. and Rai, S.(2020), *Fed-focal loss for imbalanced data classification in federated learning*, arXiv preprint arXiv:2011.06283.

Sun, B. and Park, B. B.(2017), "Route choice modeling with support vector machine", *Transportation Research Procedia*, vol. 25, pp.1806-1814.

Sun, B., Gong, L., Shim, J., Jang, K., Park, B. B., Wang, H. and Hu, J.(2023), "A human-centric machine learning based personalized route choice prediction in navigation systems", *Journal of Intelligent Transportation Systems*, vol. 27, no. 4, pp.523-535.

Sun, B., Gong, L., Shim, J., Jang, K., Park, B., Wang, H. and Hu, J.(2020), "MT-LinAdapt: A Human Centric Machine Learning Based Individual Drivers' Route Choice Model for Personalized Route Recommendation", *Presented at 99th Annual Meeting of the Transportation Research Board*, Washington, D.C.

Tian, Y., Wang, J., Wang, Y., Zhao, C., Yao, F. and Wang, X.(2022), "Federated vehicular transformers and their federations: Privacy-preserving computing and cooperation for autonomous driving", *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp.456-465.

Xu, C., Qu, Y., Luan, T. H., Eklund, P. W., Xiang, Y. and Gao, L.(2022), “An efficient and reliable asynchronous federated learning scheme for smart public transportation”, *IEEE Transactions on Vehicular Technology*, vol. 72, no. 5, pp.6584-6598.

Yang, L., Tan, B., Zheng, V. W., Chen, K. and Yang, Q.(2020), “Federated recommendation systems”, *Federated Learning: Privacy and Incentive*, Springer, pp.225-239.

Yuksel, A. S. and Atmaca, S.(2021), “Driver’s black box: A system for driver risk assessment using machine learning and fuzzy logic”, *Journal of Intelligent Transportation Systems*, vol. 25, no. 5, pp.482-500.

Zhang, Y. and Xie, Y.(2008), “Travel mode choice modeling with support vector machines”, *Transportation Research Record*, vol. 2076, no. 1, pp.141-150.