



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

제 127 회 석사학위논문
지도교수 손 기 민

딥러닝 기법을 이용한 버스 하차지점 예측 모형에 관한 연구

Predicting Smart-Card Holders' Bus-Alighting Locations Using a
Deep Learning Technology

중앙대학교 대학원
토목공학과 교통공학전공
정 재 영
2017년 8월

딥러닝 기법을 이용한 버스 하차지점 예측 모형에 관한 연구





Predicting Smart-Card Holders' Bus-Alighting Locations Using a
Deep Learning Technology

이 논문을 석사학위논문으로 제출함

2017년 8월

중앙대학교 대학원
토목공학과 교통공학전공
정 재 영

정재영의 석사학위논문으로 인정함

심사위원장	류 중 석	
심사위원	李 亨 宰	
심사위원	김 태 완	
심사위원	손 기 민	

중앙대학교 대학원

2017년 8월

목 차

제 1 장 서론.....	1
---------------	---

제 1 절 연구의 배경·목적 및 내용.....	1
---------------------------	---

1. 연구의 배경 및 목적.....	1
2. 연구의 내용.....	5

제 2 장 선행연구 고찰.....	7
--------------------	---

제 1 절 기존 하차지점 예측 모형.....	7
--------------------------	---

1. 기존 하차지점 예측 모형 분석.....	7
2. 모형 검증 방법론 분석.....	19
3. 연구의 공간적 배경 비교.....	21
4. 모형 검증 데이터 비교 및 분석.....	23

제 3 장 데이터 설명 및 전처리 과정.....	27
----------------------------	----

제 1 절 스마트카드 데이터.....	27
----------------------	----

1. 데이터 설명.....	27
2. 데이터 전처리.....	27

제 2 절 토지이용 데이터	30
1. 데이터 설명	30
2. 데이터 전처리	30
제 4 장 하차지점 예측 모형	33
제 1 절 다항로짓 모형을 통한 하차지점 예측	33
1. 다항로짓 모형 개요	33
2. 입력 변수 구축	36
제 2 절 기계학습을 이용한 하차지점 예측	40
1. 기계학습 모형의 개요	40
2. 입력 및 출력변수 구축	43
제 5 장 분석결과 도출 및 비교	46
제 1절 분석결과 도출	46
1. 다항로짓 모형 분석결과	46
2. 기계학습 모형 분석결과	51
3. 종합결과 비교 및 분석	57
제 6 장 결론 및 향후 연구과제	61

참 고 문 헌	64
국 문 초 록	68
ABSTRACT	70

〈표 차례〉

표 2-1. 각 선행 연구의 공간적 배경 및 대중교통 규모 비교	21
표 2-2. 각 선행 연구의 하차지점 추정 대상 및 검증 관련 정보	24
표 3-1. 건축물 용도별 연상면적 데이터의 코드 및 분류	31
표 5-1. Limdep 3.0을 이용한 일반 모형 정산 결과	47
표 5-2. 일반 모형 정산 결과 해석	48
표 5-3. Limdep 3.0을 이용한 상호작용 모형 정산 결과	49
표 5-4. 상호작용 모형 정산 결과 해석	50
표 5-5. 심층 신경망 하이퍼 파라미터 정리	55
표 5-6. 선행 연구 모형 및 제안 모형의 성능 비교	59

〈그림 차례〉

그림 1-1. 기·종점 정보가 완전한 스마트카드 데이터를 보유한 국가	2
그림 2-1. Barry et al. (2002)의 첫 번째 가정	7
그림 2-2. Barry et al. (2002)의 두 번째 가정	8
그림 2-3. 일반화 시간 모형의 개념도	13
그림 3-1. 통행사슬(Trip chain) 결정 과정 예시	28
그림 3-2. 스마트카드 데이터 전처리 과정	29
그림 3-3. 서울시 버스 정류장과 건축물 용도별 연상면적 GIS 데이터	30
그림 4-1. 개인-특유 변수와 대안-특유 변수를 가진 각 정류장	36
그림 4-2. 다항로짓 모형에 사용되는 개인-특유 변수와 대안-특유 변수	38
그림 4-3. 서포트 벡터 머신의 개념도	41
그림 4-4. 인공 신경망과 심층 신경망의 차이	42
그림 4-5. 기계학습 모형의 입력 및 출력 변수 형태	44
그림 5-1. 하차지점 예측을 위한 최종 심층 신경망 모형	54
그림 5-2. 'ReLU' 활성화 함수	56
그림 5-3. 드랍아웃(Dropout) 개념도	56

제 1 장 서론

제 1 절 연구의 배경 · 목적 및 내용

1. 연구의 배경 및 목적

대중교통 체계와 단말기 기술의 발달로 인하여 스마트카드를 이용한 자동요금징수체계(Automated Fare Collection System, AFC)의 보급이 세계 여러 도시에 보편화되었다. 대중교통 자동요금징수체계가 보편화되면서 방대한 양의 스마트카드 데이터가 축적되었으며, 이는 대중교통을 운영 및 관리하는 교통 운영자의 대중교통 정책 결정에 중요한 역할을 한다(Bagchi et al, 2005). 효율적 대중교통 정책 결정을 위해서는 대중교통 이용자의 통행 행태를 그대로 보여주는 스마트카드 데이터를 통해서 분석해야 할 필요가 있다. 따라서 여러 도시에서 스마트카드 데이터를 이용하여 통행 행태를 파악하기 위한 연구가 진행되어 왔다.

그러나 대부분의 도시에 설치된 대중교통 자동요금징수체계는 승차지점에 대한 정보만을 수집하도록 구축되어 있으며(Entry-only), 대중교통 중 특히 버스 이용자의 하차지점에 대한 정보가 결핍되어 있다. 그러므로 하차지점 정보가 없는 대다수의 도시에서, 하차지점을 파악하기 위한 시도는 해당 도시의 버스 이용자의 통행 패턴을 분석하기 위해 필수적이라고 할 수 있다. (Barry et al., 2002; Zhao et al., 2007; Trépanier et al., 2007; Zhang et al., 2007; Farzin, 2008; Nassir et al., 2011;

Wang et al., 2011, Munizaga et al., 2012; Munizaga et al., 2014, Zhang et al., 2015; Nunes et al., 2016).



[그림 1-1] 기·종점 정보가 완전한 스마트카드 데이터를 보유한 국가

오직 대한민국의 서울, 싱가포르, 호주의 퀸즈랜드 주 남동부 (SEQ, South East Queensland)에 위치한 브리즈번 등 소수의 도시만이 버스 승차 및 하차지점에 대한 완전한 기·종점 정보를 가진 자동요금징수체계를 구축하고 있다. 싱가포르의 경우, 국가 자체가 도시 규모인 도시국가이기 때문에 하차지점에 대한 예측의 필요성 보다는 버스 이용자의 통행 패턴, 통행 목적, 기·종점 (OD) 통행량 등을 파악하는 연구가 주를 이루었다. 특히 서울을 포함한 수도권의 자동요금징수체계는 가장 완전한 기·종점 정보를 보유하고 있다고 할 수 있으나, 하차지점 추정에 대한 연구는 거의 이루어지지 않았다.

다양한 연구에서 기점에 대한 정보만을 가진 스마트카드 데이

터를 이용하여 버스 이용자의 하차지점을 예측하기 위한 모형을 제안하였다. 사용된 주요 데이터로는 해당 도시의 자동요금징수 체계로(AFC)부터 수집된 스마트카드 거래내역 데이터와 버스의 실시간 위치 정보(AVL, Automated Vehicle Location) 데이터, 버스 정류장의 좌표에 대한 정보를 포함하고 있는 지리정보체계(GIS, Geographic Information System), 그리고 버스의 운행 정보를 포함하는 버스 일정(Schedule) 데이터 등이 해당된다. 위 데이터를 바탕으로, TransCAD, SQL(Structured Query Language) 소프트웨어 등의 교통 분석 및 데이터베이스 프로그램을 통해 해당 연구에서 제안하는 특정 논리를 검증하는 방법으로 연구가 이루어졌다.

또한 하차지점 예측 방법론의 기본 논리는 특정 스마트카드 거래내역의 하차는 다음 거래내역의 승차지점의 주변에서 이루어진다는 가정과 하루의 마지막 통행은 하루의 통행이 시작된 지점의 주변에서 이루어진다는 Barry et al. (2002)에 의해 제안된 두 가지 가정에 기초하고 있다. 그러나 위 선행 연구에서 제안한 방법론들의 근본적인 한계는 정확한 하차지점의 예측이 아닌 추정(Inference)을 모형의 정확도로 제시하였다는 것과 통행일지(Travel diary) 조사, 설문조사, 가구통행 실태조사 등의 외생(Exogenous) 데이터를 통한 검증이 대부분 이루어지지 않았다는 것이다. 설령 외생 데이터를 통한 검증이 이루어졌다고 하더라도, 그 데이터의 수가 현저히 부족한 한계점을 갖고 있다.

두 번째로 기·종점 정보를 모두 가진 스마트카드 데이터를 이용한 연구의 경우, 이미 하차지점에 대한 정보를 보유하고 있기 때문에 추가적으로 하차지점을 예측하기 위한 노력이 필요하지

않게 된다. 이 경우, 환승에 대한 정보가 부재한 경우가 대부분이며, 따라서 환승 여부 및 해당 통행의 목적을 추정하기 위한 연구가 진행되었다. 또한 완전한 스마트카드 데이터를 이용하여 선행 연구에서 제안한 다양한 방법론을 검증하는 연구가 이루어졌으나, 개별 거래내역의 하차지점에 대한 검증이 아닌 집계(Aggregate)된 단위에서의 검증이었으며, 따라서 하차지점 추정 방법론을 검증한 연구는 전무한 실정이다.

앞서 언급한 선행 연구의 배경을 종합하면, 기존 하차지점 예측의 주요 목적은 스마트카드 데이터에서 결핍된 하차지점에 대한 정보를 특정 논리를 통해 추정한 뒤, 외생의 검증 데이터가 존재하는 경우 이를 통하여 방법론을 검증하는 것이다. 이 과정에서 사용된 방법론은, Barry et al. (2002)에 의해 제안된 하차지점 예측 방법론의 두 가지 기본 원리를 바탕으로, 해당 연구에서 제안한 논리를 추가하여 하차지점을 추정한다. 또한 특정 연구의 방법론으로 하차지점을 추정한 거래내역이 전체 거래내역에 대해서 차지하는 비율, 즉 추정율(Inference rate)을 제안 방법론의 성능으로 제시한다. 대부분 기·종점 정보가 완전한 외생 데이터를 통한 검증이 이루어지지 않았으며, 이루어진 경우도 데이터의 수가 현저히 부족하다는 것을 확인할 수 있다.

따라서 본 연구에서는 완전한 기·종점 정보를 가진 서울시의 스마트카드 데이터를 이용하여 개인의 하차지점을 정류장 단위에서 예측하는 다항로짓 모형 및 기계학습 모형을 제안한다. 또한 각 제안 모형을 개인의 통행 행태를 반영한 완전한 데이터를 이용하여 검증하고자 한다. 이렇게 구축된 모형은 기점에 대한 정보만을 보유한 대한민국의 지방도시 및 세계의 여러 도시에 적용

할 수 있는 일반화된 하차지점 예측 모형을 제안할 수 있는 중요한 의미를 지닌다. 다음으로, 3% 내외의 표본에 대해서 실시하는 가구통행실태조사를 통해 구축한 기존의 기·중점 행렬(O-D matrix)을 전수에 가까운 스마트카드 데이터를 통해 구축한 기·중점 행렬로 대체할 수 있는 방법론을 제안하고자 한다. 이는 막대한 비용과 시간이 소요되는 기존의 가구통행실태조사를 대체하여 사회·경제적 비용을 최소화 하는 효과를 얻을 수 있다. 본 연구는 앞서 언급한 두 개의 연구 목적을 달성함으로써, 하차지점 예측에 다항로짓 및 기계학습 모형(심층 신경망)을 최초로 사용한 연구라는 점과 완전한 기·중점 정보를 가진 스마트카드 데이터를 통해 모형을 검증했다는 점에서 그 의미를 가진다.

2. 연구의 내용

본 연구에서는 하차에 대한 정보가 결핍된 스마트카드 데이터로부터 하차지점을 예측하기 위해서 다항로짓 모형(일반적(Conventional) 모형 / 상호작용(Interaction) 모형)과 기계학습 모형(서포트 벡터 머신(SVM, Support Vector Machine), 심층신경망(DNN, Deep Neural Network))을 사용하여 새로운 하차지점 예측 모형을 제안한다.

2장에서는 Barry et al. (2002)이 제안한 두 가지 기본 가정을 기본으로 하차지점을 추정한 여러 선행연구에 대해서 고찰한다. 선행연구 고찰은 두 주요 항목인 검증 방법론에 대한 분석과 검증에 사용된 데이터에 대한 분석에 대해서 이루어진다.

3장에서는 앞서 고찰한 선행연구의 내용을 바탕으로, 본 연구에서 제안하는 모형의 분석 방법론 및 분석 절차에 대한 틀을 제

시한다. 또한 제안 모형의 두 가지 분석 모형인 다항로짓 모형과 기계학습 모형의 개요 및 하차지점 예측에 대한 적용 여부에 대해서 자세히 논의한다.

4장에서는 3장에서 소개한 모형을 이용하여 하차지점 예측이 어떻게 이루어지는지에 대한 내용을 소개한다. 특히, 각 모형이 기존의 방법론에 비해 가질 수 있는 장점에 대해서 언급하며, 모형을 정산하기 위해 필요한 데이터의 설명 및 전처리 과정에 대해서 다항로짓 모형, 기계학습 모형의 순으로 상세히 소개한다.

5장에서는 4장에서 구축한 두 가지 모형의 분석 결과를 설명하며 이를 기존의 방법론과 비교하는 과정이 이루어진다.

최종적으로, 6장에서는 본 연구를 통해 맺을 수 있는 결론을 언급하고 향후 연구과제에 대해서 논의한다.

제 2 장 선행연구 고찰

제 1 절 기존 하차지점 예측 모형

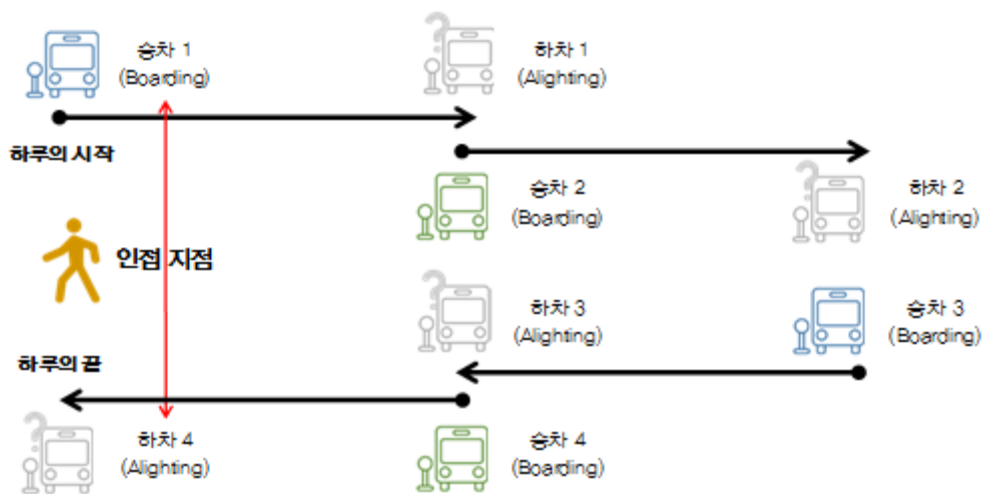
1. 기존 하차지점 예측 모형 분석



[그림 2-1] Barry et al. (2002)의 첫 번째 가정

기점에 대한 정보만을 가진 스마트카드 데이터를 이용한 하차지점 예측에 관한 연구는 Barry et al.(2002)를 시작으로 본격적으로 이루어졌다. 하차지점 예측을 위해 해당 연구에서 제안한 논리는 두 가지이다. 첫 번째로, 대부분의 대중교통 이용자는 이전 통행의 하차지점 근처에서 다음 통행을 위해 승차하는 행태를 보인다는 것이다(그림 2-1). 두 번째는, 하루 중 대중교통 이용자의 마지막 통행의 하차는 그 날의 최초 통행이 시작된 정류장의 근처에서 이루어진다는 것이다(그림 2-2). 이는 Barry et al.(2002) 이후에 진행된 논리 중심의 하차지점 예측 관련 연구

에서 필수적으로 설정되는 기본 가정으로, 하차지점 예측 연구의 기본적인 틀을 마련하였다. 위 논리를 NYCT(New York City Transit Authority)에서 소수의 도시철도 이용자를 대상으로 실시한 통행일지 설문 데이터를 이용하여 검증한 결과, 약 90%의 정확도를 얻을 수 있었다.



[그림 2-2] Barry et al. (2002)의 두 번째 가정

이후 하차지점 예측 연구는 도시철도와 버스 간, 그리고 버스 간 통행에서의 하차지점 예측을 중심으로 연구되었다. (Zhao et al. 2007; Trépanier et al. 2007; Zhang et al. 2007; Farzin, J. 2008). Zhao et al. (2007)은 Barry et al. (2002)의 기본 가정에 최대 도보 거리 기준(400m, 도보 5분)과 연속된 통행 사이에는 사적인 교통수단(자가용, 오토바이, 자전거 등)을 이용한 통행이 없다는 가정을 추가한 뒤, 개별 대중교통 통행자의 통행사슬을 구성하여 하차지점을 예측하는 모형을 구축하였다. 이를 바탕으로 도시철도 통행 후 후속 통행이 도시철도인 경우뿐만 아니

라 버스인 경우까지를 연구의 대상으로 설정하였다. 데이터베이스 관리 시스템(DBMS, Database Management System)과 지리 정보체계를 사용하여 하차지점을 추정한 결과, 180만 개의 거래 내역에 대해 71.2%의 하차지점 추정율을 보였다.

앞선 연구와 동일한 가정을 바탕으로, Trépanier et al. (2007)은 버스와 버스 간 통행에 대해 하차지점 예측 모형을 구축하였다. 특히 이전 통행의 예상 하차지점과 다음 통행의 승차 지점까지의 거리를 최대 2km 범위에서 추가적으로 고려하였다. 이를 바탕으로 TOOM(Transportation Object-Oriented Modeling) 알고리즘을 제안하였다. 대중교통 공급을 설명하는 정태적(Static) 데이터와 교통 및 대중교통 이용자의 행태를 대표하는 동적(Kinetic) 데이터, 교통의 주체를 설명하기 위한 역동적(Dynamic) 데이터, 그리고 대중교통 네트워크 및 스마트카드 체계를 설명하기 위한 체계적(Systematic) 데이터로 분류하여 데이터베이스를 구축하고, 이를 정해진 논리에 대입하여 하차지점을 추정하는 형태의 연구를 진행하였다. 그 결과 통행사슬이 구성된 표본 데이터를 대상으로 약 66%(침두시 80%)의 추정율을 보였다.

Zhang et al. (2007)은 스마트카드 거래내역에 기록되어 있는 승차 일시 정보와 버스 운행 데이터에 기록되어 있는 정류장 도착 및 출발 시간 정보를 서로 매칭하여 승차지점을 추정한 뒤, 실제 버스에 탑승하여 확보한 정류장별 하차 시간을 통해 승차지점 추정과 동일한 방법으로 하차지점을 추정하였다. 또한 모든 시간대가 아닌, 출·퇴근 시간대(오전: 06:00 ~ 07:00 / 오후: 16:30 ~ 17:30)에 발생한 데이터에 대해서만 연구를 진행하였

다. 추가적인 데이터를 통한 검증은 이루어지지 않았으며, 앞서 제안한 방법론을 통해 O-D 행렬을 구축하는 방법에 대해 기술하였다.

Farzin (2008)의 경우, 브라질 상파울루의 스마트카드 데이터를 활용하여 O-D 행렬을 추정하는 과정에 대한 연구를 진행하였다. 상파울루의 자동요금징수체계를 통해 수집된 스마트카드 거래내역 데이터, 차량의 위치 데이터, 그리고 버스 정류장 GIS 데이터를 이용하여 전체 데이터를 구성하였다. 총 630만 개의 거래내역 데이터 중, 차량의 위치 데이터와 버스 정류장 GIS 데이터와 매칭되는 데이터는 505,000개로 확연히 줄어들었다. 이를 기반으로 O-D 행렬을 구축하여 기존의 O-D 행렬 구축 방법론과 비교한 결과 유사한 결과를 보임을 확인할 수 있었다.

앞서 언급한 Barry et al. (2002)의 후속 연구로 진행된 Barry et al. (2009)는 대상이 되는 대중교통 수단의 범위를 기존 지하철에서 버스, 노면전차(TRAM), 그리고 페리(Ferry)로 확장하여 하차지점을 추정하였다. 해당 연구에서 사용된 가정 및 알고리즘은 Barry et al. (2002)와 동일하지만, 여러 선행 연구에서 제외한 단일 통행¹⁾에 대해서도 하차지점 추정을 진행했다는 점이 이전과 다른 점이라고 할 수 있다. 이러한 경우는 해당 통행의 승차지점에서 발생한 여러 통행의 하차지점 중 무작위 추출하여 단일 통행의 하차지점으로 결정하는 방법을 사용하였다.

미네소타 주의 메네아폴리스-세인트폴(Minneapolis-Saint Paul)에서 발생한 ‘Go-To Card’ 데이터를 이용하여 하차지점을 추정한 Nassir et al. (2011) 연구에서는 차량의 위치 데이터와

1) 당일 특정 ID에 의해 발생한 통행이 오직 하나인 경우

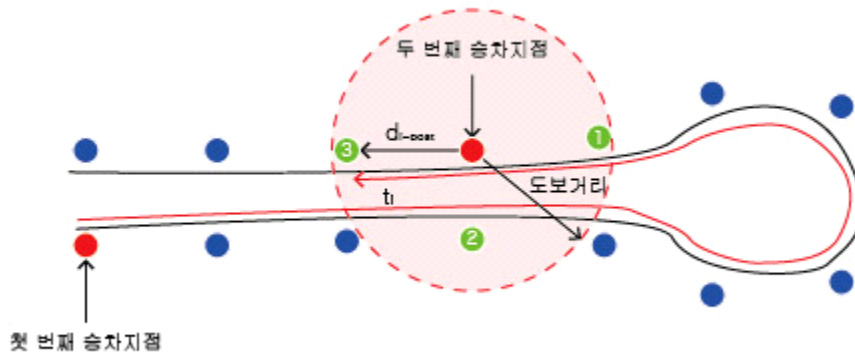
승객 수 측정 데이터를 결합한 데이터를 통해 방법론을 검증하였다. 승객 수 측정 데이터는 미네아폴리스에서 운행하는 버스의 30%에 탑재된 기기로부터 수집되었으며, 이는 승차 또는 하차 활동이 이루어진 시점의 차량 위치에 대한 정보를 포함하고 있다. 또한 예정된 정차 시간과 실제 도착 시간이 기록되어 있으며, 기존 선행 연구에서 거의 사용되지 않았던 GTFS(Google's General Transit Feed Specification)를 통해 각각의 통행 ID에 대한 세부적인 버스 일정 정보를 지도에 표현하여 활용하였다. 마찬가지로 Barry et al. (2002)에서 제안한 두 개의 가정을 기본으로 하였으며, 개인이 특정 활동을 하는 시간을 30분 초과로, 환승을 위해 기다리는 시간을 90분 미만으로 설정하였다. 또한 버스 이용자의 일반적인 도보 속도는 3mph(=4.8km/h)로 가정하여 모형을 구축하였다. 추가적으로 두 개의 연속된 통행의 하차 지점과 승차지점의 유클리디언 거리(Euclidean distance)를 사용하지 않고, 두 지점을 연결하는 길이 직선이 아니라는 점에 입각하여 $\sqrt{2}$ 를 곱한 값을 사용하였다. 즉, 승차가 이루어지는 최대 거리와 하차가 이루어지는 최대 거리, 그리고 실제 승차시간과 버스의 예정 도착시간 차이의 평균 지체에 대해서 민감도 분석을 진행하였다. 민감도 분석의 결과, 다음 거래내역의 승차지점으로부터 이전 거래내역의 하차지점까지의 거리가 0.25 또는 0.5 마일(400m 또는 800m)인지의 여부는 하차지점 추정의 정확도에 큰 영향을 미치지 않음을 보였다.

Wang et al.(2011)의 연구 역시 마찬가지로 Barry et al.(2002)의 두 기본 가정을 중심으로 모형을 구축하였으며, 이전 하차지점과 다음 승차지점까지의 도보 거리를 1km(1,000m, 12

분 도보 거리)로 설정하였다. 영국의 자동요금징수체계로부터 수집된 ‘Oyster Smart Card’ 데이터에 대해서 하차지점을 추정하였으며, 이를 영국의 대중교통 관련 기관인 TfL(Transport for London)에 의해 수집된 버스 통행자의 O-D 조사(BODS: Bus passenger Origin and Destination Survey)와 비교하여 그 추정율을 검증하였다. 또한 5개의 특정 버스 노선에 대해서 57.5~78.5%에 해당하는 추정율을 보였으며, 특히 185번 노선의 북향과 남향에 대해서 각각 66%, 65%의 추정율을 보였다. 그러나 본 연구에서 검증의 용도로 사용한 BODS 데이터가 개별 승객의 통행 패턴을 충분히 설명하지 못하는 집계된 데이터라는 것과 버스 네트워크에 해당하는 추정 및 검증이 아닌 특정 노선에 대해서만 추정 및 검증이 이루어졌다는 것이 한계라고 할 수 있다.

칠레 산티아고의 자동요금징수체계인 ‘Bip!’ 카드 데이터를 이용하여 하차지점을 추정한 Munizaga et al.(2012)에서는 하차지점을 추정하기 위해서 ‘일반화 시간(Generalized time, Tg_i)’ 모형을 제안하였다. 일반화 시간 모형의 개념은 다음 통행의 승차지점에서 이전 통행의 하차지점이 될 수 있는 인접 정류장까지(이하 후보 정류장)의 도보시간, 이전 승차지점으로부터 후보 정류장까지의 통행시간의 합으로 이루어진다. (수식 1)에서, t_i 는 이전 승차지점으로부터 각 후보 정류장까지의 통행 시간, d_{i-post} 는 다음 승차지점에서 각 후보 정류장까지의 직선거리를 의미한다.

$$Tg_i = t_i + f_w \cdot \frac{d_{i-post}}{s_w} \quad (\text{수식 1})$$



[그림 2-3] 일반화 시간 모형의 개념도

f_w 는 도보시간의 비효용(the disutility of walking-time)을 차내 시간의 비효용(the disutility of in-vehicle-travel-time)으로 나눈 패널티 요소, s_w 는 평균 보행속도를 나타낸다. 이때 하차지점 후보가 될 수 있는 정류장은 다음 승차지점을 기준으로 도보거리 1km(1,000m) 이내에 존재하는 경우에 한하며, 평균 보행속도는 구체적인 수치가 주어지지 않았다. 일반적으로 하차지점 추정 모형에 필요한 도보거리, 보행속도 등의 값은 연구의 공간적 배경이 되는 도시의 특성에 따라 상이하지만, 선행 연구를 분석해 본 결과 큰 차이가 없음을 알 수 있다. Munizaga et al. (2012)의 제안 모형은 하차지점이 될 수 있는 여러 후보 정류장 중 가장 작은 일반화 시간을 갖는 정류장이 하차지점으로 추정되는 원리로 사용이 매우 간단하며, 그 추정을 또한 80.77~83.01%로 우수한 성능을 보인다. 따라서 본 연구에서는 Munizaga et al. (2012)가 제안한 일반화 시간 모형을 다항로짓 모형 및 기계학습 모형의 성능을 비교하기 위한 기본 모형으로 선택하였다.

Munizaga et al. (2012)의 후속 연구인 Munizaga et al.

(2014)에서는 선행 연구에서 제안한 하차지점 추정 방법론을 내생적인(Endogenous) 측면과 외생적(Exogenous) 측면에서 검증하였다. 사용한 데이터는 선행 연구와 동일하며, 추가적으로 지원자를 통한 검증 데이터를 확보하여 검증을 진행하였다. 내생적인 측면에서의 검증은 도보거리, 하루의 첫 번째 통행에 대한 정의, 그리고 단일 통행에 대해 이루어졌다. 도보거리의 경우, 선행 연구에서 설정한 1km를 초과하는 통행이 전체 통행의 7%임을 확인하였으며, 이 경우 대부분 상업지역이 밀집한 중심업무지구(CBD, Central Business District)의 근처에서 발생한 통행임을 증명하였으며, 이를 통해서 중심 업무지구의 경우에는 1km보다 큰 거리를 기준으로 설정해야 함을 주장하였다. 다음으로 하루의 첫 번째 통행을 당일 00:00:00부터 23:59:59까지로 정의한 기존의 방법을 당일 04:00:00부터 다음날 03:59:59로 변경하였다. 이는 거래내역의 분포를 확인한 결과, 당일 통행의 마지막 통행이 자정(Midnight) 부근에서 이루어진다는 것을 확인하였기 때문이다. 단일 통행의 경우, 전체 데이터의 5%에 해당하는 양으로 해당 데이터의 시간대를 분석한 결과, 오전 및 오후 침투 시간에 가장 빈번하게 발생하는 것을 확인하였다. 이 경우, Barry et al. (2009)에서 제안한 단일 통행의 하차지점 추정 방법과 동일한 방법을 사용하였으나, 5%의 데이터 중 7%에 해당하는 데이터에 대해서만 하차지점을 추정할 수 있었다. 외생적 측면에서의 하차지점 추정 모형 검증은 연구 지원자를 통해서 확보한 검증 데이터를 이용하여 진행하였다. 총 882개의 통행(1,350개의 거래내역) 데이터를 확보하였으며, 이 중 715개의 거래내역에 대해서 하차지점 추정을 검증하였다. 그 결과, 84.2%의 정확도를 확보할

수 있었으며, 이를 통해 Munizaga et al. (2012)에서 제안한 모형의 성능을 검증할 수 있었다. 그러나 검증에 사용된 통행조사 데이터가 버스 또는 지하철에서 버스로 이동하는 경우에 국한되었으며, 지원자의 대부분이 학생으로 이루어져 표본이 하나의 계층으로 편향되어 유사한 특성을 가진 통행이 발생할 수 있다는 한계가 있다. 또한 연구에 지원하였기 때문에 실험 목적에 부합하기 위한 의도를 가지고 통행하는 문제도 발생할 수 있다.

Zhang et al. (2015)의 연구에서는 스마트카드 데이터뿐만 아니라 금액 충전 데이터, 도로 네트워크 데이터를 이용하여 하차지점 추정 방법론을 제안하였다. 데이터는 버스에 승차 또는 하차함으로써 지불하는 금액과 스마트카드의 잔액에 대한 정보를 의미하는 금전적 공간(M, Monetary space), 스마트카드 거래내역의 태그 시간을 나타내는 시간적 공간(T, Temporal space), 마지막으로 정류장의 위치와 도로 네트워크를 구성하는 지리적 공간(G, Geospatial space)로 나뉜다. 해당 연구에서 사용한 스마트카드 데이터는 대부분 기점에 대한 정보만을 갖고 있으나(고정 요금제 노선의 경우), 특정 노선(거리비례 요금제 노선)의 경우 기점 및 종점에 대한 정보를 모두 갖고 있다. 즉, 종점에 대한 정보가 있는 데이터(Labeled data)와 없는 데이터(Unlabeled data)를 모두 보유하고 있기 때문에 기계학습 방법론 중 하나인 준-지도형(Semi-supervised) 조건부 무작위장(CRF: Conditional Random Field) 방법을 사용하여 하차지점 추정 모형을 구축하였다. 그 결과 약 80%의 정확도를 확보할 수 있었으며, 추가로 102명의 지원자를 통해서 4개월 동안 수집한 거래내역 데이터를 이용하여 검증을 진행한 결과 역시 80%에 가까운

정확도를 보였다. 해당 연구는 기계학습 방법론을 하차지점 추정
에 적용했다는 점에서 기존의 방법론과 다른 접근 방법을 보여주
었으며, 검증의 결과 역시 기존의 방법보다 좋은 결과를 보였다.
그러나 검증에 사용된 데이터가 거리비례 요금제를 도입한 특정
노선으로부터 확보한 제한된 데이터라는 점과 102명이라는 소수
의 참여자를 통해 수집된 데이터를 사용했다는 한계가 있다.

서울, 싱가포르 이외에 완전한 기·종점 정보를 가진 호주의 퀸
즈랜드 남동부 스마트카드 데이터를 바탕으로 진행된 Alsger et
al. (2015)의 연구에서는 선행 연구에서 사용된 변수들의 타당성
에 대해서 민감도 분석을 진행하였다. 기존의 연구에서는 도보거
리 및 도보로 이동한 시간에 대한 변수를 추가하여 연구를 진행
했으나, 도보로 이동한 외의 시간(Non-walking time)에 대한 연
구는 거의 진행된 바가 없었다. 해당 연구에서는 대중교통을 타
기 위해서 기다리는 대기 시간과 특정 활동을 하는 시간을 구분
하기 위해서 30분, 60분, 90분으로 나누어 모형의 정확도를 비교
하였다. 총 473,525개의 거래내역에 대해서 연구를 진행하였으
며, 환승 시간의 기준을 60분으로 설정한 경우에 가장 높은 하차
지점 추정율을 보였다. 또한 도보거리의 경우, 하루의 첫 승차지
점 반경 400m 이내에서 하루의 마지막 하차가 이루어졌으며, 하
루의 시작 승차지점 반경 800m 이내에서 95%에 해당하는 대중
교통 이용자의 마지막 하차가 이루어진 것을 증명하였다. 앞서
도보거리의 민감도에 대해서 연구한 선행 연구의 결과를 종합하
면, 도보거리가 750~1,000m 사이에서는 하차지점 추정의 정확
도에 큰 영향을 미치지 않음을 확인할 수 있다. 해당 연구는 별
도의 하차지점 추정 모형을 제안하지 않았으며, 선행 연구에서

사용된 변수 이외의 고려사항에 대한 민감도 분석을 진행했다는 점에 그 의의가 있다.

포르투갈의 포르투의 자동요금징수체계인 STCP(Sociedade de Transportes Colectivos do Porto)에서 수집된 ‘Andante’ 스마트카드 데이터를 이용한 Nunes et al.(2016)의 연구에서는 총 3백 만 개의 거래내역에 대해서 62.44%에 해당하는 추정율로 하차지점을 추정하였다. 하차지점 추정에 사용된 모형은 기존의 연구와 비교하여 추가적인 내용은 확인할 수 없었으며, 별도의 데이터를 통한 검증 역시 이루어지지 않았다. 다만 다음 승차지점과 이전 하차지점의 최대 이동 거리를 4.8km/h의 속력으로 8 분 동안 이동한 거리인 640m를 사용하여 하차지점을 추정하였다.

위 선행 연구에서 하차지점 추정을 위해 제안한 모형은 주로 구조화 질의 언어(SQL), 교통 분석 소프트웨어(e.g. TransCAD), 그리고 데이터베이스 관리 시스템(DBMS) 등을 통해서 구축되었다. 즉, 특정 모형을 사용한 것이 아닌, 연구에서 설정한 하차지점 추정 논리를 구현한 것에 불과하다. 또한 기계학습 모형을 사용한 경우가 있으나(Zhang et al. 2015), 제한된 데이터를 검증에 사용했다는 한계가 있다.

도보거리, 환승 시간, 보행속도 등 하차지점 추정 논리의 각 단계에서 필요한 계수에 대해서 많은 연구가 진행되었다. 도보거리의 경우, 400m부터 1,000m까지 다양한 값에 대해서 연구가 진행되었으며, 그 결과 하차지점 추정의 정확도는 도보거리의 영향을 크게 받지 않음을 확인할 수 있다. 또한 환승 시간의 경우, 하차지점 추정 시 환승에 대한 고려가 필요한 경우에 한하여 사

용되며, 환승에 대한 정보가 없는 데이터를 사용하는 연구의 경우에 필수적인 요소라고 할 수 있다. 그러나 본 연구에서 사용하는 서울시 스마트카드 데이터는 환승에 대한 완전한 정보가 있기 때문에, 이에 대한 고려는 필요하지 않다. 보행속도의 경우, 각 연구에서 사용된 보행속도가 서로 상이하나 크게 차이나지 않으며, 이는 연구가 이루어진 도시의 인구·지형적 특성에 따라 다르게 설정할 수 있음을 선행 연구를 통해서 확인할 수 있다.

마지막으로 선행 연구에서 하차지점을 예측하기 위해 사용한 데이터는 스마트카드 데이터, 대중교통 일정(Schedule) 정보, 도로 네트워크 정보, 차량 GPS 데이터 등이 있다. 그런데 교통은 기점과 종점을 통행하는 대중교통 이용자가 기점과 종점이 갖는 어떠한 특성으로 인해 해당 지역을 통행하게 되는 원리를 통해서 발생하게 된다. 따라서 이러한 특성을 반영하기 위해서는 정류장 또는 지하철역 주변의 정보를 반영해야 한다. 그러나 본 연구를 통해 조사한 바에 의하면, ‘교통 수요는 파생수요’라는 근본적인 원리를 반영하는 데이터를 사용하여 하차지점 예측 모형을 연구한 경우는 없었다. 본 연구에서는 정류장 또는 지하철역 주변의 특성을 반영하는 데이터로 28개의 용도로 이루어진 토지이용(Land-use) 데이터를 사용하였다. 다음으로 각 정류장 또는 지하철역에 대해서 주거, 상업, 업무, 문화, 공업, 기타 여섯 개의 용도로 집계하여 변수화한 뒤 이를 하차지점 예측 모형에 사용하였다. 건축물 용도별 연상면적 데이터의 사용은 기존의 선행 연구에서 사용하지 않은 데이터를 사용했다는 점과 정류장 또는 지하철역 주변의 특성을 반영함으로써 모형의 설명력을 얻을 수 있다는 이점이 있다. 또한 다양한 변수를 사용함으로써 모형의 성능

을 향상시킬 수 있는 가능성을 보여준다고 할 수 있다.

2. 모형 검증 방법론 분석

하차지점을 추정하기 위한 모형의 성능을 확인하기 위해서 가장 중요한 것은 해당 연구에서 제안한 모형을 별도의 데이터를 통해 검증하는 것이다. 그러나 대부분의 선행 연구에서 기점에 대한 정보만을 가진 스마트카드 거래내역 데이터를 사용하기 때문에, 해당 데이터를 통한 자체적인 검증을 하기는 불가능하다. 따라서 몇몇 연구에서 별도의 외생 데이터를 통해 모형을 검증하는 시도를 하였다. (Barry et al., 2002; Farzin, 2008; Wang et al., 2011; Munizaga et al., 2014; Zhang et al., 2015). 외생 데이터를 통한 검증에 앞서, 사용된 검증 데이터는 크게 통행일지 조사(Travel diary survey), 가구통행 실태조사(Household survey), 연구 참여자의 실제 통행 자료로 나뉜다. 세 가지 데이터 모두 표본 조사를 통해서 확보한 데이터이며, 해당 연구의 공간적 배경인 특정 도시의 대중교통 관련 기관에 의해서 조사된다.

이러한 데이터를 바탕으로 검증을 진행하는 방법에는 가구통행 실태조사를 통한 집계된 단위에서의 검증, 개인의 통행 데이터를 통한 검증이 있다. 먼저 집계된 단위에서 하차지점을 추정하는 모형은 정류장 단위에서의 하차지점 추정이 아닌, 교통분석의 기본 단위인 교통분석 존(TAZ, Traffic Analysis Zone) 단위의 O-D 통행량에 대한 모형을 제안하였으며, 검증 역시 교통분석 존 단위 O-D 통행량에 대해 이루어졌다. (Farzin, 2008;

Wang et al., 2011; Munizaga et al., 2014). 이때 사용된 데이터는 대중교통 관련 기관에서 조사한 표본을 통해 전수로 확대한 O-D 통행량을 사용하는데, 대부분의 데이터가 실제 스마트카드 데이터를 사용한 시점과 일치하지 않는 경향을 보인다. 그 이유는 표본 조사를 통한 O-D 통행량 데이터는 특정 주기로 확보되기 때문에 대부분 연구에서 사용한 스마트카드 데이터 이전의 시점에 구축된 데이터이기 때문이다.

하차지점 추정의 궁극적인 목적은 특정 지역과 다른 지역에 대한 O-D 통행량을 확보하기 위한 것이지만, 보다 세부적인 관점에서 정류장-정류장 단위의 통행량 및 대중교통 이용자의 통행 행태를 분석하기 위한 목적 또한 존재한다. 일반적으로 교통분석준 단위에서의 검증이 높은 정확도를 보이며, 따라서 세부적인 단위에서의 검증보다 용이하다는 것을 선행 연구를 통해서 확인할 수 있다. 이러한 면에서 개인의 통행 행태 기반의 하차지점 추정에 대한 연구는 중요한 의미를 지니며, Barry et al. (2002); Munizaga et al. (2014); Zhang et al. (2015)이 개인의 통행 데이터를 이용하여 하차지점 추정의 검증을 진행하였다. 개인의 통행 데이터는 해당 연구의 실험에 참여한 참가자를 대상으로 특정 기간 동안의 스마트카드 거래내역을 수집한 뒤, 가용한 데이터를 추출하여 가공된다. 다음으로 해당 연구에서 제안한 모형을 개인의 통행 데이터에 적용한 뒤, 모형을 통해서 실제 하차지점을 정확하게 예측한 데이터가 전체 데이터에 대해 차지하는 비율을 정확도로 제시하였다. 이러한 검증 방법론은 본 연구에서 제안하는 하차지점 추정 모형의 검증에 있어 가장 정확하고 신뢰성 있는 방법론이다. 그러나 선행 연구에서 사용한 개인 기반의 통행 데

이터는 그 수가 현저히 부족한 문제를 보이며, 연구의 목적에 부합하는 통행으로 유도될 수 있다는 한계를 갖는다.

선행 연구의 하차지점 추정 모형 검증 방법론 분석을 정리하면, 대부분의 연구가 개인의 통행 행태를 반영한 외생적 데이터를 통한 검증이 이루어지지 않았다는 것이다. 또한 검증이 이루어졌다고 하더라도, 검증에 사용된 데이터의 수가 현저히 부족하다는 한계를 가지고 있다. 따라서 본 연구에서는 이러한 선행 연구의 검증 방법론의 한계를 극복하기 위한 방향으로 연구를 진행하였다.

3. 연구의 공간적 배경 비교

선행 연구의 하차지점 추정 모형을 고찰하고 모형의 검증 방법론을 분석한 결과, 다양한 도시를 대상으로 스마트카드 데이터를 포함한 다양한 데이터를 이용한 하차지점 추정 모형이 개발된 것을 알 수 있다. 하지만 각 연구의 공간적 배경이 되는 도시가 서로 상이하기 때문에 이를 있는 그대로 비교하기는 타당하지 않다. 따라서 각 연구의 공간적 배경이 되는 도시와 해당 도시의 규모에 대해서 비교함으로써, 하차지점 추정 모형의 비교를 어느 정도 일반화 할 수 있다.

[표 2-1] 각 선행 연구의 공간적 배경 및 대중교통 규모 비교

선행연구	대상 도시	도시 면적	도시의 인구	대중교통 체계의 규모
본 연구	서울 대한민국	605.2 km^2	10,204,057명	7,482대 버스 차량 (415개 노선) 311개 지하철 역

				(9개 노선)
Barry et al. (2002)	뉴욕 뉴욕주 미국	789 km^2	8,491,079명	4,525대 버스 차량 (230개 노선) 6,344대 지하철 차량 (25개 노선)
Zhao et al. (2007)	시카고 일리노이주 미국	606.1 km^2	2,720,546명	1,879대 버스 차량 (140개 노선) 1,190대 지하철 차량 (8개 노선)
Trépanier et al. (2007)	가티노 퀘벡주 캐나다	598 km^2	240,000명	302대 버스 차량 (66개 노선)
Lianfu et al. (2007)	창춘 지린성 중국	4,738 km^2	4,193,073명	100개 이상의 버스 노선 49개 지하철 역
Farzin (2008)	상파울루 브라질	1,521 km^2	12,038,175명	1,400대의 버스 차량 (GPS를 장착한)
Barry et al. (2009)	뉴욕 뉴욕주 미국	789 km^2	8,491,079명	4,525대 버스 차량 (230개 노선) 6,344대 지하철 차량 (25개 노선)
Nassir et al. (2011)	미네아폴리스-세인트폴 미네소타주 미국	2,646 km^2	3,112,117명	1,010대 버스 차량 (186개 노선)
Wang et al. (2011)	런던 영국	1,572 km^2	8,538,689명	8,000대 버스 차량 (700개 노선)
Munizaga et al. (2012)	산티아고 칠레	641 km^2	6,158,080명	6,591대 버스 차량 (391개 노선) 108개 지하철 역 (5개 노선)
Munizaga et al. (2014)	산티아고 칠레	641 km^2	6,158,080명	6,591대 버스 차량 (391개 노선) 108개 지하철 역 (5개 노선)
Zhang et al. (2015)	-	-	-	-
Alsger et al. (2015)	퀸즈랜드 남동부 퀸즈랜드 호주	22,420 km^2	3,400,000	-
Nunes et al. (2016)	포르투 포르투갈	41.42 km^2	240,000	472대 버스 차량 (83개 노선)

[표 2-1]에서 Zhang et al. (2015)의 경우 연구의 공간적 배경에 대한 설명이 명시되지 않았으며, Alsger et al. (2015)의 공간적 배경인 호주 퀸즈랜드 남동부 지역에 대한 대중교통 네트워크의 구체적인 규모 역시 언급되지 않았다. 이 두 경우를 제외한 나머지 연구에 대해서 대상 도시의 규모를 확인한 결과, 본 연구의 대상 도시인 서울이 면적 대비 인구와 대중교통 체계가 가장 높은 수치를 보임을 확인할 수 있었다. 즉, 타 도시에 비해 조밀한 대중교통 체계가 구축되어 있으며, 높은 인구 밀도를 보이고 있다. 본 연구의 목적인 정류장 단위에서의 하차지점 추정은, 위와 같이 조밀하고 혼잡한 대중교통 체계를 가진 도시에서 더욱 어려운 경향을 보인다. 따라서 이러한 점을 고려하여 본 연구의 하차지점 추정 모형의 성능을 비교 및 분석하는 관점이 필요하다.

4. 모형 검증 데이터 비교 및 분석

앞서 선행 연구의 한계점으로 모형 검증에 사용된 데이터의 수가 부족하다는 점을 언급하였다. 이를 구체적으로 확인하기 위해서 본 연구에서 고찰한 선행 연구의 검증 데이터를 비교할 필요성이 있다. 각 선행 연구에서 제안한 하차지점 추정 모형의 대상 교통수단, 검증에 사용된 데이터의 종류와 양, 그리고 각 연구의 하차지점 추정을 또는 정확도에 대한 내용을 [표 2-2]와 같이 정리하였다.

[표 2-2] 각 선행 연구의 하차지점 추정 대상 및 검증 관련 정보

선행연구	대상 교통수단	검증 데이터	검증 데이터의 수	정확도 (추정율)
Barry et al. (2002)	지하철	통행일지 (Travel diary survey)	795개 거래내역 (2개 거래내역: 100명 / 3개 이상 거래내역: 150명)	90%
Zhao et al. (2007)	지하철-지하철 지하철-버스	-	검증 X	(71.2%)
Trépanier et al. (2007)	버스	-	검증 X	(66%) (첨두시 80%)
Lianfu et al. (2007)	버스	-	검증 X	-
Farzin (2008)	버스	-	검증 X	(100%)
Barry et al. (2009)	버스 지하철 트램 페리	승객수 측정 데이터	언급 X	언급 X
Nassir et al. (2011)	버스	승객수 측정 데이터와 차량의 위치 데이터를 결합한 데이터	10,886 거래내역	95.4%
Wang et al. (2011)	버스	-	검증 X	(65~66%)
Munizaga et al. (2012)	버스 지하철	-	검증 X	(80.77~ 83.01%)
Munizaga et al. (2014)	버스-지하철 지하철-버스	53명의 지원자 (주로 학생)	715개 거래내역	84.2% (602 / 715)
Zhang et al. (2015)	버스	102명의 지원자 (무료 교통카드)	102 ID (4개월 기간)	75%
Alsger et al. (2015)	버스 지하철 페리	-	검증 X	-
Nunes et al. (2016)	버스	-	검증 X	(62.4%)

* 괄호() 내에 기입된 %는 본 연구에서 정의한 추정율(Inference rate)이며, 괄호()가 없는 %는 정확도(Accuracy)

먼저 대상 교통수단에 대해 살펴보면, 버스에 대한 하차지점 추정이 상당한 비율을 차지함을 확인할 수 있다. 이는 일반적으

로 버스의 하차지점 추정이 지하철 등의 다른 교통수단의 하차지점 추정보다 변칙적인 특성을 가지며, 따라서 하차지점 추정보다 어려워지기 때문이라고 할 수 있다. 이러한 이유로, 본 연구에서도 지하철이 포함되지 않은, 오직 버스로만 구성된 통행에 대해서 하차지점 추정을 진행하였다.

검증 데이터의 종류는 앞서 언급한 바와 같이 개인의 통행 행태를 보여주는 조사 또는 연구 참가자의 데이터와 버스 또는 지하철에서 확보되는 승객 수 측정 데이터가 존재한다. 먼저 승객 수 측정 데이터(APC, Automated Passenger Counter)의 경우, 버스 또는 지하철 개찰구에 설치된 기기를 통해서 해당 교통수단에 탑승하는 승객의 수를 자동으로 세어 측정된다. 그러나 특정 지하철 역 또는 특정 버스 차량에만 탑재되기 때문에 전체 데이터를 대표할 수 있는 당위성을 충분히 확보하지 못한다. 해당 데이터를 사용한 선행 연구에서도 이러한 한계로 인해 충분한 양의 검증 데이터를 확보하지 못했으며, 확보된 데이터 역시 차량의 위치 데이터의 시간에 대해 일치하는 경우에 한해서만 구축된다. 즉, 데이터 자체가 개인의 통행 행태를 내포하기 보다는, 차량의 위치와 시간을 통한 데이터 전처리 과정을 거쳐서 확보한 데이터이므로 완전한 검증 데이터라고 할 수 없다. 그런 의미에서, 통행 일지 또는 연구 참가자를 통해서 수집한 데이터가 가장 완전한 의미의 데이터라고 할 수 있다. 그러나 해당 데이터는 수집하는 비용이 크기 때문에 많은 양을 확보할 수 없으며, 연구의 의도가 반영된 통행 행태를 보일 수 있다는 한계가 있다. 또한 Munizaga et al. (2014)의 경우, 대부분의 참가자가 학생으로, 하나의 통행 행태에 편향되어 전체 통행 행태를 대표할 수 없는

문제점을 가진다. Zhang et al. (2015)의 경우, 102명의 참가자에게 무료 교통카드를 배분한 뒤, 4개월 동안 통행에 관한 데이터를 추적하였는데, 이 역시 마찬가지로 102명의 반복적인 통행 행태를 반영할 뿐, 전체 데이터의 통행 행태를 반영한다고 할 수 없다.

기존 연구의 검증 데이터를 비교·분석한 결과, 본 연구가 기존 연구의 모형 검증과 비교하여 차별성을 갖기 위해서는 ① 개인의 통행 행태를 있는 그대로 반영할 수 있는 데이터를 사용하며, ② 보다 많은 검증 데이터를 확보하여 검증의 신뢰성을 향상시켜야 할 것이다. 또한 ③ 특정 논리에 입각한 하차지점 추정율이 아닌, 정확한 하차지점을 추정하는 정확도를 모형의 성능으로 제시하고자 한다. 따라서 본 연구에서는 기·종점에 대한 완전한 정보를 보유하고 개인의 통행 행태를 그대로 반영하고 있는 서울시의 스마트카드 데이터를 사용하고, 검증에 사용되는 데이터를 기존 연구에 비해 많이 확보함으로써 하차지점 추정 모형에 대한 연구 분야에 위 세 가지 항목에 대해서 작은 기여를 할 수 있을 것으로 기대한다.

제 3 장 데이터 설명 및 전처리 과정

제 1 절 스마트카드 데이터

1. 데이터 설명

본 연구에서 사용한 스마트카드 데이터는 서울시 내부에서 하루 동안 발생한 교통카드 거래내역을 축적한 데이터로, 하루 동안 총 12,258,466개의 교통카드 거래내역을 원본 데이터로 한다. 하나의 거래내역은 암호화된 카드의 ID와 승차시간 및 승차 정류장 ID, 하차 시간 및 하차 정류장 ID가 모두 기록되어있다. 또한 해당 거래내역의 통행 수단, 승차 인원, 환승 순서, 그리고 사용자 유형 등의 특성을 갖는다. 스마트카드 데이터의 다른 여러 특성이 존재하나, 본 연구의 목적에 필요한 데이터의 속성은 앞서 언급한 특성 중, 카드 ID, 승차 및 하차 시간, 승차 및 하차 정류장 ID, 통행 수단, 그리고 환승 순서이다.

2. 데이터 전처리

본 연구의 하차지점 예측에 필요한 단일 통행사슬(an trip chain)은 두 개의 연속된 거래내역을 이용하여 결정된다. 그 과정은 (그림 3-1) 예시를 통해서 설명할 수 있다. 특정 ID의 카드를 소유한 대중교통 이용자가 하루에 세 번의 거래내역을 생성한 경우, 총 세 번의 승차 태그에 세 번의 승차 정류장이 기록된다.

첫 번째 거래내역의 하차지점은 해당 거래내역과 두 번째 거래내역의 승차 정류장의 정보를 통해서 결정된다. 즉, 연속된 두 개의 거래내역의 승차 정보를 이용하여 하차 정류장을 예측하기 때문에 k 개의 거래내역으로 $(k-1)$ 개의 통행 사슬을 결정하게 된다. 일반적인 통행 사슬의 정의는 하나의 통행 목적을 달성하기 위해 발생한 연속된 통행의 묶음이라고 할 수 있는데, 이때는 활동을 구분하기 위한 시간적 기준이 필요하다. 따라서 환승 통행은 하나의 통행 사슬의 일부를 결정하게 되며, 두 개의 연속된 통행 사슬은 환승 가능 시간을 초과한 경우에 나뉘게 된다. 이는 본 연구에서 정의한 통행 사슬과 시간적 기준을 사용하지 않았다는 점에서 차이를 보이며, 이 점을 유의하여 본 연구의 하차지점 예측 모형을 고려해야 한다.

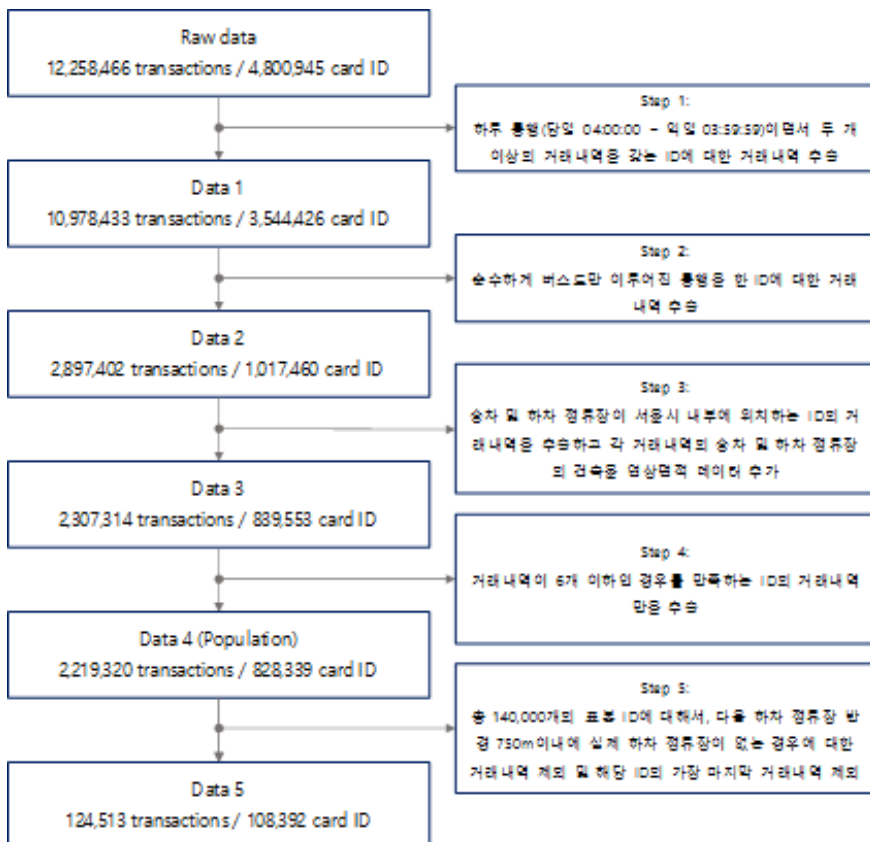


[그림 3-1] 통행사슬(Trip chain) 결정 과정 예시

위와 같은 통행 사슬의 결정을 위해서는 단일 통행(Single trip)은 제외되어야 한다. 다음으로, 본 연구의 하차지점 예측 대상 교통수단은 오직 버스로만 이루어진 통행이기 때문에 총 데이터 중 버스로만 이루어진 통행을 갖는 카드 ID에 대한 거래내역을 추

출하였다. 하루의 통행 중, 한 번이라도 도시철도를 이용한 거래내역을 보유한 카드 ID의 모든 거래내역은 대상에서 제외하였다.

또한 본 연구의 하차지점 예측은 서울시 내부에서 발생한 거래내역에 한하며, 거래내역에 기록된 승차 정류장 및 하차 후보 정류장 중 토지이용 데이터가 가용하지 않은 경우를 제외하였다. 위 데이터로부터 140,000개의 카드 ID의 거래내역을 임의로 추출한 뒤, 다음 거래내역의 승차 정류장 반경 750m 이내에 실제 하차 정류장이 존재하지 않는 경우를 제외하여 최종 데이터를 구축하였다. 데이터의 전처리 과정은 (그림 3-2)으로 정리할 수 있다.

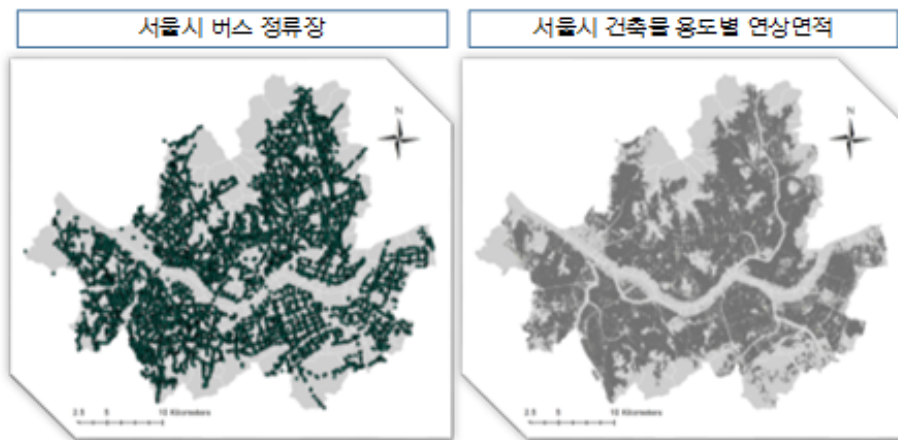


[그림 3-2] 스마트카드 데이터 전처리 과정

제 2 절 토지이용 데이터

1. 데이터 설명

본 연구의 공간적 배경인 서울시 내부에 위치한 버스 정류장과 건축물 용도별 연상면적 GIS 데이터를 이용하여 토지이용 데이터를 구축하였다 (그림 3-3). 버스 정류장 데이터의 경우, 본 연구에서 사용한 스마트카드 데이터에 기록된 모든 승차 및 하차 정류장 ID를 추출한 결과를 사용하였다.



[그림 3-3] 서울시 버스 정류장과 건축물 용도별 연상면적 GIS 데이터

2. 데이터 전처리

버스 정류장의 좌표를 기준으로 반경 500m 버퍼(Buffer)를 설정한 뒤, 해당 버퍼 내에 위치한 건축물의 용도별 연상면적을 집계하였다. 건축물 용도별 연상면적 데이터는 총 28개의 용도로

구분되어 있으며(표 3-1), 본 연구에서는 이를 주거, 업무, 상업, 공업, 문화, 기타 여섯 개의 상위 토지이용으로 집계하였다.

[표 3-1] 건축물 용도별 연상면적 데이터의 코드 및 분류

코드	소분류	대분류(집계)
BDS000	미분류	제외
BDS001	주택	주거
BDS002	근린생활시설	상업
BDS005	판매시설	
BDS006	운수시설	
BDS013	숙박시설	
BDS014	위락시설	
BDS015	공장	공업
BDS003	문화 및 집회시설	문화
BDS004	종교시설	
BDS007	의료시설	
BDS009	노유자(노인 및 어린이) 시설	
BDS010	수련시설	
BDS011	운동시설	
BDS016	창고시설	기타
BDS017	위험물 저장 및 처리시설	
BDS018	자동차 관련시설	
BDS019	동물 및 식물 관련시설	
BDS020	분뇨 및 쓰레기 처리시설	
BDS021	교정 및 군사시설	
BDS022	방송통신시설	
BDS023	발전시설	
BDS024	묘지 관련시설	
BDS025	관광휴게시설	
BDS026	장례시설	
BDS999	기타시설	
BDS008	교육연구시설	업무
BDS012	업무시설	

이렇게 집계된 건축물 용도별 연상면적 데이터를 본 연구에서

는 토지이용 데이터로 정의하였다. 토지이용 데이터는 본 연구의 하차지점 예측 모형에서 필요한 연속된 두 통행의 승차 정류장과 다음 거래내역의 승차 정류장 반경 750m 이내의 후보 정류장이 갖는 특성 변수로 사용된다. 즉, 각 정류장별로 주거, 업무, 상업, 공업, 문화, 기타 토지이용 면적을 변수로 갖게 되며, 이는 각각의 통행 사슬을 구성하는 정류장 ID와 연결되어 하나의 표본 벡터를 형성하게 된다. 이렇게 구축된 스마트카드 및 토지이용 데이터는 본 연구에서 제안하는 다항로짓 모형과 심층 신경망 모형에 사용된다. 각 모형에 대한 데이터의 입력 및 출력 변수의 적용은 다음 장에서 자세하게 다룬다.

제 4 장 하차지점 예측 모형

제 1 절 다항로짓 모형을 통한 하차지점 예측

1. 다항로짓 모형 개요

제 2장에서 이루어진 선행연구 고찰 결과, 하차지점 예측에 관한 기존 연구에서는 구체적인 모형이 아닌 논리 과정에 의한 하차지점 추정을 중심으로 한 연구가 주를 이루었다. 특정 모형을 사용한 연구의 경우에도, 다항로짓 모형을 사용하여 하차지점 예측을 시도한 사례 역시 전무하다. 또한 앞서 언급한 내용들을 바탕으로, 실제 특정 거래내역의 하차지점을 예측하기 위해서는 하차지점이 될 수 있는 여러 정류장 가운데 가장 선택될 가능성이 큰 정류장을 하차지점으로 선택한다는 것은 자명하다. 그런데 이는 여러 수단 선택 대안 중, 선택 결정자의 통행 시간, 비용 등의 효용(Utility)을 최대화하는 대안을 선택한다는 개념과 유사하다고 할 수 있다. 따라서 본 연구에서는 기존 교통계획 분야에서 통행의 수단을 선택하는 수단 선택(Mode choice) 모형에 사용되어온 다항로짓 모형을 통한 하차지점 예측 모형을 제안한다.

다항로짓 모형은 개인 또는 가구를 대상으로 여러 교통수단 중 특정 수단을 선택할 확률을 유도하기 위한 이산선택 모형(Discrete choice model)의 한 종류이다. 이산선택 모형의 특징은 개인의 수단 선택에 대한 여러 설명 변수(Explanatory variables)가 미치는 영향을 고려할 수 있으며, 통계적 유의성

(Statistical significance)을 확인할 수 있다는 것이다. 이러한 모형의 특성은 선행 연구의 하차지점 예측 모형에서는 증명할 수 없는 부분이며, 따라서 다항로짓 모형을 통항 하차지점 예측 모형 구축은 영향 변수에 대한 설명과 통계적 유의성 부분에서 중요한 의미를 갖는다.

다항로짓 모형은 수단을 선택하는 개인이 가용한 여러 수단 중 효용을 최대화 하는 수단을 선택한다는 효용극대화(Utility maximization) 이론을 기반으로 한다. 이때 개인 i 가 특정 수단 j 를 선택하는 경우의 효용(U_{ij})은 (수식 2)로 계산되며, 효용 U_{ij} 은 관측 가능한 효용(Systematic utility)인 V_{ij} 와 관측이 불가능한 임의의 효용(Random utility)인 ϵ_{ij} 으로 이루어져 있다.

일반적으로 다항로짓 모형에서 관측되지 않는 임의 효용 ϵ_{ij} 은 유형-2 검벨 분포(Type-II Gumbel distribution, 또는 EVI, Extreme Value Type I distribution)를 따른다고 가정한다. 이때 비관련 대안간의 독립성(I.I.A, Independence of Irrelevant Alternatives)이 선행되는데, 이는 대안의 추가나 제외가 남아있는 다른 대안의 결정에 영향을 미치지 않는다는 것을 의미한다. 비관련 대안간의 독립성이 보장되지 않는 문제는 다항로짓 모형의 한계로 알려져 있다. 그러나 본 연구의 다항로짓 모형은 동일한 노선을 지나지만, 다음 거래내역의 승차지점으로부터의 직선 거리가 서로 다르고 주변의 토지이용 정보가 고유한 독립된 정류장을 대안으로 설정함으로써 비독립 대안간의 독립성을 만족한다고 가정하였다.

$$U_{ij} = V_{ij} + \epsilon_{ij} \quad (\text{수식 2})$$

$$\epsilon_{ij} \sim EVI$$

관측되는 효용 V_{ij} 는 개인 i 의 대안 j 가 갖는 효용을 개인이 갖는 고유한 변수인 개인-특유 변수(Individual-specific variables)와 각 대안이 갖는 변수인 대안-특유 변수(Alternative-specific variables)로 이루어진 설명변수(x_{ijk} , Explanatory variables)를 통해서 결정되며, (수식 3)과 같이 표현된다. θ_{j0} 는 개인 i 의 선택 대안 중 기준이 되는 수단을 제외한 다른 수단이 갖는 고유한 상수(Alternative-specific constant)를 의미하며, $\theta_{j1}, \dots, \theta_{jk}$ 는 k 개의 설명 변수에 대한 계수(Coefficients associated with explanatory variables)를 나타낸다.

$$V_{ij} = \theta_{j0} + \theta_{j1}x_{ij1} + \theta_{j2}x_{ij2} + \dots + \theta_{jk}x_{ijk} \quad (\text{수식 3})$$

(수식 2)와 (수식 3)을 통해서 계산된 각 대안의 효용 중, 가장 큰 효용을 갖는 대안을 선택할 확률은 (수식 4)와 같이 계산된다.

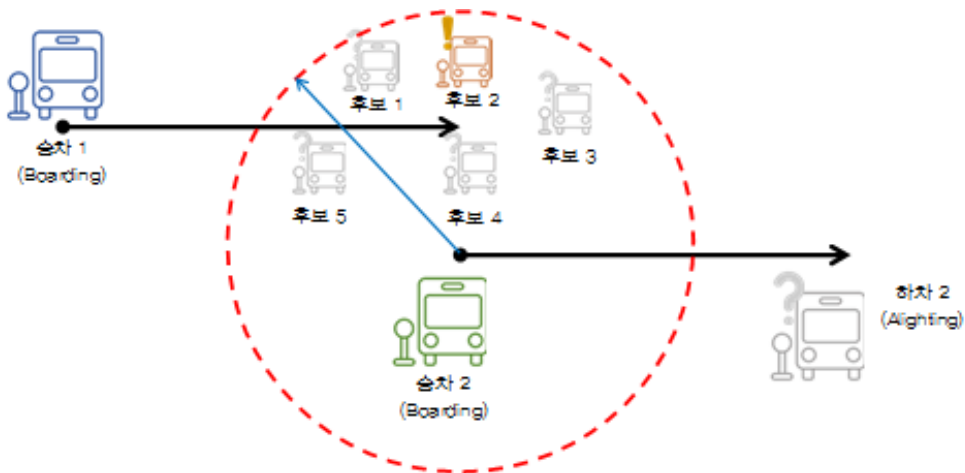
$$P_{ij} = \frac{\text{Prob}\{U_j \geq U_q, \forall A_q \in A(i)\}}{\text{Prob}\{\epsilon_{iq} \leq \epsilon_{ij} + (V_{ij} - V_{iq}), \forall A_q \in A(i)\}} \quad (\text{수식 4})$$

최종적으로 개인 i 가 수단 j 를 선택할 확률(P_{ij})은 (수식 5)와 같이 표현된다. 개인 i 의 모든 선택 가능한 대안($A_q \in A(i)$)의 효용의 지수함수 값의 합에 대해서 특정 수단 j 의 효용의 지수함수 값이 차지하는 비율로, 모든 수단 선택 확률의 합은 1이 된다.

$$P_{ij} = \frac{\exp(U_{ij})}{\sum_{A_q \in A(i)} \exp(U_{iq})} \quad (\text{수식 5})$$

2. 입력 변수 구축

다항로짓 모형에 사용되는 설명 변수는 모든 대안이 공통적으로 갖는 개인-특유 변수와 각 대안별로 다른 값을 갖는 대안-특유 변수로 이루어져 있다. 일반적으로 수단 선택 모형에서 개인-특유 변수는 수단을 선택하는 결정자의 개인적인 속성으로 이루어져 있으며, 대안-특유 변수는 결정자가 선택할 수 있는 모든 수단이 각각 갖는 속성으로 이루어진다. 그렇다면, 다항로짓 모형의 개인-특유 변수와 대안-특유 변수는 어떠한 속성으로 이루어지는가?



[그림 4-1] 개인-특유 변수와 대안-특유 변수를 가진 각 정류장

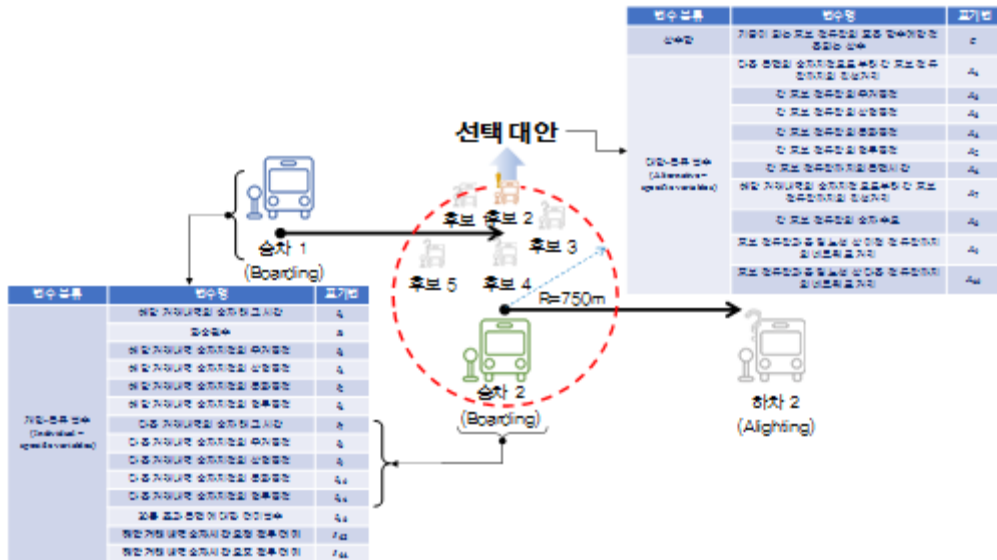
하차지점 예측은 [그림 4-1]의 단일 통행사슬에서 두 개의 승차 정보 및 정류장과 하차지점이 될 수 있는 다섯 개의 후보 정류장으로 결정된다. 즉, 하나의 통행사슬 내에서 승차에 대한 정보 및 승차 정류장의 특성은 변하지 않으며, 각 후보 정류장의 대한 정보는 서로 다른 특성을 갖는다. 따라서 승차에 대한 정보와 승차 정류장의 특성을 나타내는 변수는 개인-특유 변수로 설

정되며, 후보 정류장의 특성 변수는 대안-특유 변수로 결정할 수 있다. 여기서 해당 거래내역의 승차가 이루어진 노선과 동일한 노선에 위치하면서 다음 거래내역의 승차 정류장 반경 750m 이내에 위치한 여러 후보 정류장 중, 다음 거래내역의 승차 정류장으로부터 직선거리가 가장 짧으며 실제 하차 정류장을 포함하는 다섯 개의 정류장을 추출하였다. 다음으로 해당 거래내역의 승차 지점으로부터 각 정류장까지의 네트워크 통행 시간이 가장 짧은 순서로 다섯 개의 대안을 정렬하였다. 본 연구에서 사용한 스마트카드 데이터에는 차량의 진행 방향이 결핍되었기 때문에, 해당 거래내역의 진행 방향과 반대 방향에 위치한 정류장이 후보 정류장으로 선택되는 문제를 방지하기 위해 위와 같은 순서로 대안을 선택하였다. 또한 모든 표본의 선택 가능한 대안 수를 다섯 개로 동일하게 설정하였으며, 이는 어느 정도 합리적인 수의 대안을 설정하고자 함에 그 목적이 있다.

각 정류장 및 승차에 대한 정보는 제 3장에서 서술한 과정을 통해 처리한 스마트카드 및 토지이용 정보를 사용하였다. 그 결과, [그림 4-2]와 같이 해당 거래내역 및 다음 거래내역의 승차 정류장에 대한 정보 및 승차 정보를 통해 14개의 개인-특유 변수, 다섯 개의 후보 정류장 각각의 특성으로 10개의 대안-특유 변수로 총 24개의 변수가 결정된다.

이렇게 구축된 변수를 갖는 다항로짓 모형의 단일 표본은 다섯 개의 각 대안으로 구성된다. 즉, 하나의 표본은 다섯 개의 대안의 변수를 의미하는 행벡터(Row vector)로 구성되어 있기 때문에 (전체 데이터의 수 \times 5)의 최종 표본을 갖게 된다. 이러한 입력 변수는 본 연구에서 사용한 계량경제학 및 통계학 소프트웨어

인 ‘Limdep 3.0’에 필요한 데이터 구조를 충족하게 된다.



[그림 4-2] 다항로짓 모형에 사용되는 개인-특유 변수와 대안-특유 변수

이렇게 설정된 개인-특유 변수와 대안-특유 변수는 다음과 같이 각 모형의 특성에 따라 서로 상이한 형태로 모형에 반영된다.

1) 일반적 모형(Conventional model)

일반적 모형은 대안-특유 변수와 개인-특유 변수의 상호 작용을 고려하지 않은 단순 모형으로, (수식 6)으로 표현된다. $j=1$ 인 경우인 Alt_1 은 다섯 개의 대안 중, 다음 거래내역의 승차지점으로 부터 가장 가까운 거리에 있는 정류장으로 설정된다. 상수 C 는 다른 대안과 비교하여 첫 번째 대안(Alt_1)이 갖는 고유한 효용을 확인하기 위한 고유 상수이다. 일반적으로 기준이 되는 대안을 제외한 나머지 대안에 설정되지만, 본 연구에서는 기준이 되는

정류장, 즉 다음 거래내역의 승차지점으로부터 가장 가까운 거리에 있는 정류장이 다른 대안에 비해서 갖는 고유한 효용을 확인하기 위해 첫 번째 대안에 대해서만 고유 상수를 설정하였다. C_A 는 대안-특유 변수 A 에 대한 계수를 의미한다. C_I 는 개인-특유 변수 I 에 대한 계수를 의미하는데, 기준이 되는 대안을 제외한 나머지 대안의 효용 함수에 설정됨으로써 나머지 대안이 개인-특유 변수를 통해서 갖는 효용을 확인할 수 있다.

$$U(Alt_j) = \begin{cases} C + C_A A & \text{if } j = 1 \\ C_A A + C_I I & \text{otherwise} \end{cases} \quad (\text{수식 6})$$

2) 상호작용 모형(Interaction model)

상호작용 모형은 대안-특유 변수와 개인-특유 변수의 상호 작용을 고려하는 모형으로, (수식 7)과 같이 표현된다. 일반적 모형과 동일하게, 상수 C 는 다른 대안과 비교하여 첫 번째 대안(Alt_1)이 갖는 고유한 효용을 확인하기 위한 고유 상수이다. C_{AI} 는 대안-특유 변수 A 와 개인-특유 변수 I 의 상호작용 변수 $A \cdot I$ 가 효용에 미치는 영향을 나타내는 계수를 의미한다. 상호작용 변수의 해석은, 대안-특유 변수 A 가 효용에 미치는 영향이 C_A 이며, 이때 A 와 상호작용하는 변수 I 가 추가됨으로써 효용에 미치는 영향이 C_{AI} 만큼 변화한다고 할 수 있다. 즉, 두 변수의 상호작용은 $A(C_A + C_{AI} \cdot I)$ 와 같이 표현될 수 있다.

$$U(Alt_j) = \begin{cases} C + C_A A + C_{AI} A \cdot I & \text{if } j = 1 \\ C_A A + C_{AI} A \cdot I & \text{otherwise} \end{cases} \quad (\text{수식 7})$$

제 2 절 기계학습을 이용한 하차지점 예측

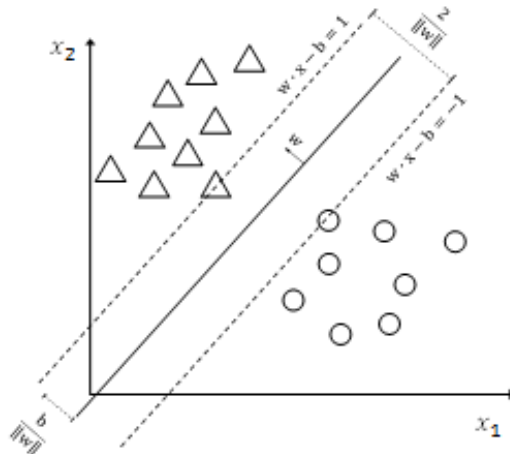
1. 기계학습 모형의 개요

1) 서포트 벡터 머신(Support Vector Machine)

서포트 벡터 머신은 기계학습의 여러 모형 중, 패턴 인식 및 데이터 분석을 위한 지도형 학습의 하나로 주로 분류(Classification)와 회귀 분석(Regression)에 사용되었다. 서포트 벡터 머신은 그래픽 처리 장치(GPU, Graphic Processing Unit)를 포함한 하드웨어의 발달과, 이로 인한 컴퓨터 연산 능력의 향상으로 비약적인 발전을 이룬 기계학습 방법론 중 하나인 심층 신경망을 포함한 딥러닝(Deep learning) 기법이 등장하기 이전에 가장 활발히 사용되었다. 즉, 서포트 벡터 머신은 뛰어난 성능을 보인 얇은 학습(Shallow learning) 기계학습 기법의 하나라고 할 수 있다.

서포트 벡터 머신은 k 개의 클래스 중 하나를 갖는 정답이 있는 데이터를 학습한 뒤, 새로운 데이터를 추가하였을 때 해당 데이터가 어느 클래스에 속하는지 여부를 판단하는 분류 모형이다. 각 데이터는 n 개의 특성(x_1, x_2, \dots, x_n)을 가진 경우, n 차원의 특성 공간(Feature space)에서 k 개의 클래스를 구분할 수 있는 $(n-1)$ 차원의 최적 서포트 벡터(=초평면, Hyperplane)을 구하는 것을 목표로 한다. 이때의 초평면은 특성 공간에서 데이터 사이의 경계로 표현되는 여러 초평면 중, 가장 큰 폭(마진, $\text{margin} = \rho = \frac{2}{\|w\|}$)을 갖도록 설정되며, 이는 (그림 4-3)과 같이 표현된다. (그림 4-3)은 특성의 차원이 2차원임을 가정할 때, 원과 삼각형으로 구분되는 두 데이터를 구분하기 위한 최

적의 초평면을 결정하는 개념을 예시로 표현한 것이다.



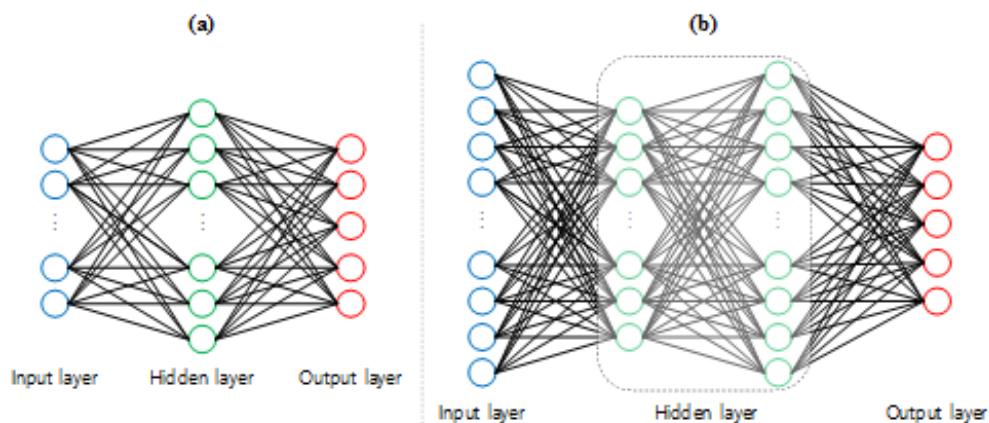
[그림 4-3] 서포트 벡터 머신의 개념도

만약 초평면의 폭이 작으면, 데이터에 포함된 노이즈(Noise)로 인한 모형의 성능 저하가 관측될 수 있으며, 따라서 일반화된 모형을 얻지 못하는 문제가 발생하게 된다. 그러므로 서포트 벡터 머신의 목적은 가장 큰 폭을 가지는 초평면을 찾아냄으로써 데이터를 분류하는 과정에서 발생하는 오차(Classifier error)를 줄이기 위해 모형을 최적화 하는 것이라고 할 수 있다.

2) 심층 신경망(Deep Neural Networks)

심층 신경망은 인간의 뇌를 모방한 인공 신경망(ANN, Artificial Neural Network)을 보다 깊게 쌓아 구축한 모형으로, 모형의 한계를 해결하는 다양한 수학적 기법과 하드웨어를 통한 연산 속도의 향상으로 인해 현실화되었다. 인공 신경망은 (그림 4-4)의 (a)와 같이 입력 층(Input layer), 은닉 층(Hidden layer), 그리고 출력

층(Output layer)이 하나로 이루어진 얇은(Shallow) 구조를 가지고 있다. 각 층은 여러 개의 노드(Node)로 이루어져 있으며, 이는 특정 층을 구성하는 특성(Feature)을 의미한다. 반면 심층 신경망은 두 개 이상의 은닉 층을 갖는 깊은(Deep) 구조이며, 특히 빅데이터의 등장과 컴퓨터 연산 능력의 발달로 입력 층의 노드 수가 많아지는 특징을 보인다. 따라서 보다 많은 데이터를 보다 깊게 학습할 수 있는 모형의 구조적 특징을 가진다.



[그림 4-4] 인공 신경망과 심층 신경망의 차이

심층 신경망에서 각 층의 노드는 서로 연결되어 링크(Link)를 형성하고, 각 링크는 모형의 학습을 통해서 최적화되는 가중치(θ)를 갖는다. 즉, 은닉 층의 노드는 이전 층의 노드가 갖는 값의 선형 결합을 통해 계산된 어떠한 값이 활성화 함수(Activation function)를 거쳐 최종적으로 출력된 값을 갖게 된다. 마지막에 위치한 출력 층의 노드 값은 분류(Classification)와 실수 값(Real value) 예측에 따라 별도의 함수를 통해서 최종적으로 계산된다. 이때 분류의 경우, 출력 층의 노드는 모형을 통해 구분하고자 하는 클래스

의 수와 동일하며 각 노드의 최종 값은 입력 데이터가 각 클래스에 속할 확률을 의미한다. 실수 값 예측의 경우 출력 노드의 수는 하나(단일 노드)이며, 출력 노드의 값은 이전 은닉 층의 노드 값과 가중치의 회귀식(Regression)을 통해 계산된다. 본 연구에서는 다섯 개의 후보 정류장에 대한 분류의 문제이므로 실수 값 예측의 경우는 고려하지 않는다.

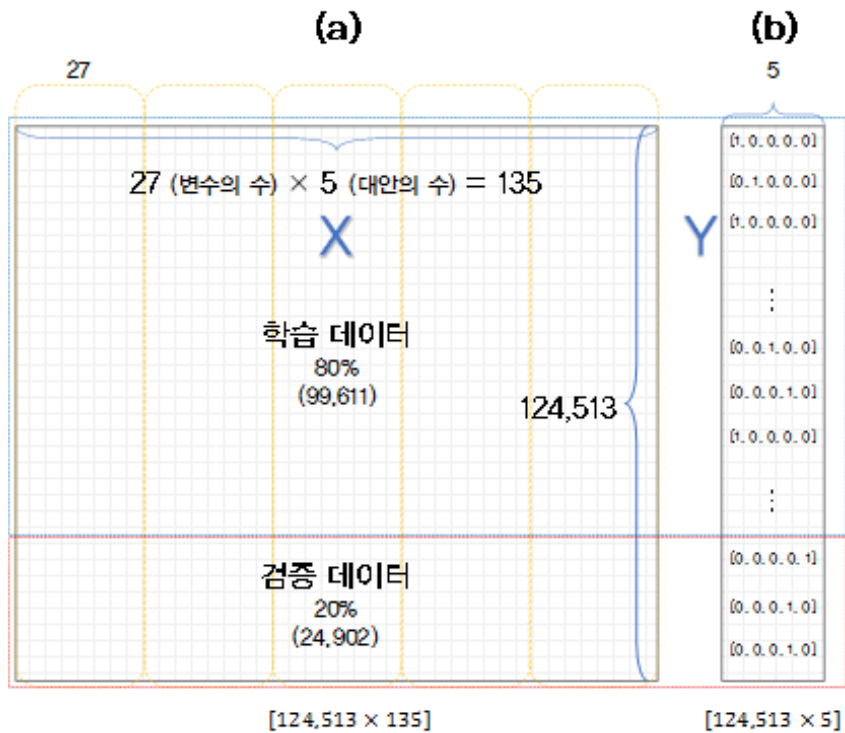
2. 입력 및 출력변수 구축

본 연구에서 제안한 기계학습 모형인 서포트 벡터 머신과 심층 신경망 모형의 입력 및 출력변수는 다항로짓 모형의 입력 변수와 달리 여러 표본 벡터로 이루어진 행렬의 형태를 갖는다. 다항로짓 모형의 입력 변수는 단일 표본에 대해 다섯 개의 행으로 이루어진 반면, 기계학습 모형의 입력 변수는 다섯 개의 행이 하나의 열벡터(Column vector)로 이루어지며 (그림 4-5)의 (a)와 같은 형태를 갖는다. 여기서 각 대안별 변수의 수가 다항로짓 모형에 비해 세 개가 추가되어 27개의 변수가 되었는데, Munizaga et al. (2012)이 제안한 일반화 시간 모형을 통해 계산된 변수와 체류시간(다음 거래내역의 승차시간과 해당 거래내역의 승차시간의 차에서 네트워크 통행 시간을 차감한 시간), 그리고 활동(Activity)을 포착하기 위해 60분 초과 통행에 대한 더미 변수가 추가되었다.

기계학습 모형의 가장 큰 특징은 데이터를 통해 모형의 여러 계수를 최적화(Optimization)하는 것이며, 특히 본 연구에서 제안한 두 개의 기계학습 모형은 정답이 있는 데이터를 통해 학습하는 지도형 학습이기 때문에 종속 변수(정답, 레이블)를 함께 구축

해야 한다. 각 표본의 정답은 다섯 개의 대안 중, 실제 하차지점에 해당하는 대안이 1의 값(선택)을 갖고 나머지 대안은 0의 값(미선택)을 갖는 (1×5) 열벡터로 표현되며 (그림 4-5)의 (b)와 같은 형태를 갖는다. 이러한 레이블의 형태는 다른 표현으로 단일 핫 벡터(One-hot vector)라고도 명명된다.

학습 및 검증 데이터의 구분은 다음 장에서의 모형 성능 비교를 위해 다항로짓 모형의 정산 및 검증에 사용된 데이터와 동일한 데이터를 사용한다.



[그림 4-5] 기계학습 모형의 입력 및 출력 변수 형태

또한 기계학습 모형의 학습을 보다 효율적으로 진행하기 위해 서, 총 135개의 변수를 각 열에 대해서 표준화(Normalization)

하는 과정을 거친다. 표준화는 통계학 및 기계학습 분야에서 서로 다른 척도(Scale)의 데이터를 동일하게 변형해주는 기법으로 다양한 방법이 존재한다. 본 연구에서는 기계학습 모형의 입력변수 표준화를 위해 (수식 8)의 표준 점수(Standard score) 방법을 사용하였다.

$$Standard\ score = \frac{X_j^{(i)} - \mu_j}{s_j} \quad (\text{수식 8})$$

(수식 8)에서 $X_j^{(i)}$ 는 i 번째 데이터의 j 번째 변수의 값을 의미하며, μ_j 는 모든 데이터를 통해 계산된 j 번째 변수의 평균을, s_j 는 표준 편차를 의미한다. 이렇게 표준화된 변수는 여러 변수들의 서로 다른 척도를 통일함으로써 기계학습 모형의 최적화 과정에서의 계산을 보다 안정적으로 만들어 주는 효과적인 방법이라고 할 수 있다.

제 5 장 분석결과 도출 및 비교

제 1절 분석결과 도출

1. 다항로짓 모형 분석결과

1) 일반적 모형(Conventional model)

본 연구에서는 총 124,513개의 표본을 99,611개의 학습 데이터와 24,902개의 검증 데이터로 나누어 두 가지 다항로짓 모형을 정산 및 검증하였다. 먼저 99,611개의 학습 데이터를 통해 정산한 각 다항로짓 모형의 결과를 분석한다.

[그림 4-2]에서 설정한 설명변수를 바탕으로, 일반 모형은 (수식 9)과 (수식 10)로 정리된다. 다음 거래내역의 승차지점으로부터 가장 짧은 지점에 위치한 정류장을 기준으로 하여(수식 9), 다른 대안이 개인-특유 상수(해당 거래내역의 고유한 값들)가 증가함으로써 얻는 효용(수식 10)을 분석하기 위해 설정되었다.

$$U(Alt_1) = C + C_{A_1}A_1 + C_{A_2}A_2 + C_{A_3}A_3 + C_{A_4}A_4 + C_{A_5}A_5 + C_{A_6}A_6 + C_{A_7}A_7 + C_{A_8}A_8 + C_{A_9}A_9 + C_{A_{10}}A_{10} \quad (\text{수식 9})$$

$$U(Alt_{other}) = C_{A_1}A_1 + C_{A_2}A_2 + C_{A_3}A_3 + C_{A_4}A_4 + C_{A_5}A_5 + C_{A_6}A_6 + C_{A_7}A_7 + C_{A_8}A_8 + C_{A_9}A_9 + C_{A_{10}}A_{10} + C_{I_1}I_1 + C_{I_2}I_2 + C_{I_3}I_3 + C_{I_4}I_4 + C_{I_5}I_5 + C_{I_6}I_6 + C_{I_7}I_7 + C_{I_8}I_8 + C_{I_9}I_9 + C_{I_{10}}I_{10} + C_{I_{12}}I_{12} + C_{I_{13}}I_{13} + C_{I_{14}}I_{14} \quad (\text{수식 10})$$

제 4장 다항로짓 모형의 개요에서 언급한 바와 같이, 다항로짓 모형의 특징은 각각의 설명변수가 종속변수(효용)에 미치는 영향(계수, coefficient)을 정량화 할 수 있으며, 통계적 유의성 역시

확인할 수 있다는 것이다. Limdep 3.0을 통해 위 일반 모형을 정산한 결과, [표 5-1]의 결과를 얻을 수 있었다.

[표 5-1] Limdep 3.0을 이용한 일반 모형 정산 결과

변수	계수 (Coefficient)	표준 편차 (Standard error)	t-ratio	P-value
C	0.198926	0.055986	3.55311	0.000381***
A_1	-0.00425	2.27E-05	-187.085	2.89E-15***
A_2	-2.96E-07	3.01E-08	-9.81867	2.89E-15***
A_3	1.22E-07	2.06E-08	5.90671	3.49E-09***
A_4	1.30E-08	1.31E-07	0.099151	0.921019
A_5	-8.03E-08	1.11E-07	-0.72084	0.471009
A_6	-0.00141	2.00E-05	-70.1911	2.89E-15***
A_7	-0.00059	2.26E-05	-26.2485	2.89E-15***
A_8	0.00031	1.28E-05	24.1427	2.89E-15***
A_9	-0.00041	1.98E-05	-20.6327	2.89E-15***
A_{10}	-1.85E-05	9.09E-07	-20.3132	2.89E-15***
I_1	0.025741	0.00339	7.59309	3.13E-14***
I_2	0.086072	0.023102	3.72578	0.000195***
I_3	7.69E-08	3.31E-08	2.32253	0.020204**
I_4	-1.43E-07	2.27E-08	-6.31117	2.77E-10***
I_5	-6.08E-07	1.93E-07	-3.14817	0.001643***
I_6	-1.36E-07	1.45E-07	-0.94046	0.346981
I_7	-0.0052	0.00297	-1.74949	0.080206*
I_8	8.93E-08	3.33E-08	2.6849	0.007255***
I_9	-6.81E-08	1.80E-08	-3.78252	0.000155***
I_{10}	-7.78E-07	1.63E-07	-4.77308	1.81E-06***
I_{11}	-5.91E-07	1.30E-07	-4.52975	5.91E-06***
I_{12}	0.041169	0.026267	1.56736	0.11703
I_{13}	0.218741	0.024963	8.76255	2.89E-15***
I_{14}	-0.10442	0.036434	-2.866	0.004157***

* 유의수준 10%이내($\alpha=10\%$), ** 유의수준 5% 이내($\alpha=5\%$), *** 유의수준 1% 이내($\alpha=1\%$)

[표 5-1]에서 유의수준 10% 이내에서 통계적 유의성을 검증한 결과를 볼드체로 표기하였다. 그 결과, 대부분의 변수가 10%

유의수준 하에서 통계적으로 유의함을 보였다. 이 중, 통계적으로 유의하고(Statistically significant) 직관적으로 설명 가능한(Intuitively accountable) 변수를 [표 5-2]와 같이 정리한다.

[표 5-2] 일반 모형 정산 결과 해석

변수	부호	해석
C	+	먼저 기준이 되는 대안의 효용에만 영향을 미치는 상수항 C는 양의 값을 가지며, 이는 가장 짧은 거리에 위치하는 정류장이 다른 대안 정류장에 비해서 어느 정도 높은 효용을 가짐을 의미함
A_3, A_4	+	각 후보 정류장의 상업면적(A_3)과 문화면적(A_4)이 넓을수록 해당 정류장을 하차지점으로 선택할 확률이 높아지며, 이는 해당 지역이 상업 또는 문화 중심지인 경우 많은 통행이 발생할 것이라는 일반적 통념과 일치하는 결과임
A_6	-	해당 거래내역의 승차지점으로부터 각 후보 정류장까지의 네트워크 통행 시간(A_6)이 늘어나면 해당 후보 정류장이 실제 하차지점이 될 확률이 줄어드는 것 의미하며, 이는 교통계획에서 일반적으로 알려진 바와 일치함
A_8	+	각 후보 정류장에서의 승차 수요(A_8)가 많을수록 해당 정류장의 효용이 높아지는데, 이는 승차 수요와 하차 수요가 연관성을 가지며, 해당 지역이 사람들의 통행이 빈번한 곳일수록 하차지점이 될 확률이 높다는 것을 반영함
I_{13}	+	해당 거래내역의 승차 시간이 오전 첨두에 발생한 경우(I_{13}), 대중교통 이용자는 가장 가까운 정류장이 아닌 다른 대안 정류장을 선택할 확률이 다소 증가함을 보임
I_{14}	-	해당 거래내역의 승차 시간이 오후 첨두에 발생한 경우(I_{14})에는 기준이 아닌 정류장을 선택하는 효용이 감소하며, 이는 가장 가까운 정류장을 선택하게 됨을 의미함

2) 상호작용 모형(Interaction model)

모형의 기본적인 설정은 일반 모형과 동일하게 설정한 뒤, 상호작용 변수를 추가하여 (수식 11)과 (수식 12)의 상호작용 모형을 설정하였다. 그 결과, [표 5-3]에서 확인할 수 있듯이 대부분의 계수가 통계적으로 유의함을 보였다.

$$U(Alt_1) = C + A_1(C_{A_1} + C_{A_1I_1}I_1 + C_{A_1I_2}I_2 + C_{A_1I_4}I_4 + C_{A_1I_8}I_8 + C_{A_1I_{10}}I_{10} + C_{A_1I_{11}}I_{11} + C_{A_1I_{12}}I_{12}) + C_{A_2}A_2 + C_{A_3}A_3 + C_{A_4}A_4 + C_{A_5}A_5 + A_6(C_{A_6} + C_{A_6I_3}I_3 + C_{A_6I_5}I_5 + C_{A_6I_6}I_6 + C_{A_6I_7}I_7 + C_{A_6I_9}I_9) + C_{A_7}A_7 + C_{A_8}A_8 + C_{A_9}A_9 + C_{A_{10}}A_{10} \quad (\text{수식 11})$$

$$U(Alt_{other}) = A_1(C_{A_1} + C_{A_1I_1}I_1 + C_{A_1I_2}I_2 + C_{A_1I_4}I_4 + C_{A_1I_8}I_8 + C_{A_1I_{10}}I_{10} + C_{A_1I_{11}}I_{11} + C_{A_1I_{12}}I_{12}) + C_{A_2}A_2 + C_{A_3}A_3 + C_{A_4}A_4 + C_{A_5}A_5 + A_6(C_{A_6} + C_{A_6I_3}I_3 + C_{A_6I_5}I_5 + C_{A_6I_6}I_6 + C_{A_6I_7}I_7 + C_{A_6I_9}I_9) + C_{A_7}A_7 + C_{A_8}A_8 + C_{A_9}A_9 + C_{A_{10}}A_{10} \quad (\text{수식 12})$$

[표 5-3] Limdep 3.0을 이용한 상호작용 모형 정산 결과

변수	계수 (Coefficient)	표준 편차 (Standard error)	t-ratio	P-value
C	-0.02182	0.012937	-1.68705	0.091594*
A_1	-0.00482	8.23E-05	-58.5404	2.89E-15***
A_1I_1	-1.94E-05	4.31E-06	-4.50076	6.77E-06***
A_1I_2	0.000705	4.40E-05	16.0342	2.89E-15***
A_1I_4	4.29E-10	4.28E-11	10.0209	2.89E-15***
A_1I_8	-4.01E-10	6.30E-11	-6.36237	1.99E-10***
A_1I_{10}	1.92E-09	3.53E-10	5.44538	5.17E-08***
A_1I_{11}	9.50E-10	2.72E-10	3.49641	0.000472***
A_1I_{12}	0.001136	3.88E-05	29.2528	2.89E-15***
A_2	-3.34E-07	3.25E-08	-10.2707	2.89E-15***
A_3	1.83E-07	2.38E-08	7.70682	1.29E-14***
A_4	1.01E-07	1.34E-07	0.753776	0.450984
A_5	-2.02E-07	1.36E-07	-1.48835	0.136659
A_6	-0.00165	8.98E-05	-18.4049	2.89E-15***
A_6I_3	5.93E-11	5.91E-11	1.00336	0.315689
A_6I_5	-8.44E-10	4.41E-10	-1.9127	0.055786*
A_6I_6	-5.70E-10	2.88E-10	-1.97908	0.047807**
A_6I_7	2.91E-05	4.22E-06	6.88051	5.96E-12***
A_6I_9	-3.66E-10	3.31E-11	-11.0382	2.89E-15***
A_7	-0.00062	2.26E-05	-27.4459	2.89E-15***
A_8	0.000295	1.29E-05	22.9467	2.89E-15***
A_9	-0.00041	1.96E-05	-20.8808	2.89E-15***
A_{10}	-1.82E-05	9.16E-07	-19.8896	2.89E-15***
A_3I_{13}	-1.49E-07	4.09E-08	-3.64839	2.64E-04***

$A_5 I_{13}$	1.00E-06	2.30E-07	4.34481	1.39E-05***
$A_2 I_{14}$	-1.83E-07	1.07E-07	-1.70216	0.088725*

* 유의수준 10%이내($\alpha=10\%$), ** 유의수준 5% 이내($\alpha=5\%$), *** 유의수준 1% 이내($\alpha=1\%$)

마찬가지로 상호작용 모형의 결과 해석을 [표 5-4]과 같이 정리하였다.

[표 5-4] 상호작용 모형 정산 결과 해석

변수	부호	해석
C	+	일반 모형의 결과와 동일한 결과를 보임
$A_1 I_1$	-	일반적으로 후보 정류장까지의 직선거리(A_1)가 먼 경우 후보 정류장의 효용이 감소하는데, 해당 거래내역의 승차시간(I_1)이 늦을수록 거리의 효용에 대한 민감도가 더욱 낮아짐
$A_1 I_2$, $A_1 I_4$	+	환승 횟수가 증가하면 거리의 효용에 대한 민감도가 다소 완화되며($A_1 I_2$), 마찬가지로 해당 거래내역의 승차지점의 상업면적이 큰 지역에서 승차한 경우 거리의 효용에 대한 민감도를 다소 완화시킴($A_1 I_4$)
$A_6 I_6$	-	해당 거래내역의 승차지점으로부터 각 후보 정류장까지의 통행시간이 증가할수록 효용이 낮아지는데(A_6), 해당 승차지점의 업무면적이 증가할수록 효용에 대한 민감도가 더욱 낮아짐($A_6 I_6$)
$A_6 I_7$	+	다음 거래내역의 승차 태그 시간(I_7)이 늦어질수록 민감도는 완화되는 것을 확인할 수 있음($A_6 I_7$)
$A_2 I_{13}$	-	후보 정류장의 주거면적이 증가하면 하차지점이 될 확률이 감소하는데(A_2), 해당 거래내역이 오후 침투시간에 발생했다면 그 확률이 더욱 감소하는 것을 나타냄($A_2 I_{13}$)
$A_3 I_{13}$	-	각 후보 정류장 주변의 상업면적이 증가하면 해당 거래내역이 선택될 확률은 증가하는데(A_3), 해당 거래내역이 오전 침투시간에 발생한 경우 후보 정류장 주변의 상업면적이 효용에 미치는 영향이 다소 줄어들게 됨($A_3 I_{13}$)
$A_5 I_{14}$	+	후보 정류장 주변의 업무면적이 늘어날수록 효용은 줄어드는데(A_5), 오전 침투시간의 경우에는 효용에 대한 민감도가 다소 증가함을 확인할 수 있음($A_2 I_{14}$)

2. 기계학습 모형 분석결과

기존의 기계학습 분야에서 데이터를 학습하기 위한 기계학습 모형을 구현하기 위해서 다양한 프로그래밍 언어를 사용해왔다. 가장 대중적으로 사용되는 상업용 프로그래밍 언어로는 매트랩(MATLAB)이 있으며, 무료 오픈소스 프로그래밍 언어로는 R, 파이썬(Python) 등이 있다. 그러나 최근 하드웨어의 발달로 인하여 딥러닝 기술이 활발히 연구되고 있으며, 딥러닝을 포함한 기계학습 모형을 구현하기 위해 파이썬 프로그래밍 언어의 프레임워크를 활발히 사용되고 있다. 본 연구에서 제안한 두 기계학습 모형 역시 파이썬 프로그래밍 언어를 사용하였으며, 서포트 벡터 머신은 파이썬에서 제공하는 기계학습 패키지인 ‘skit-learn (sklearn)’을 활용하였다. 또한 심층 신경망의 경우, 파이썬 기반의 딥러닝 프레임워크 중 가장 활발히 사용되는 프레임워크 중 하나인 ‘keras’를 사용했으며, 행렬 연산 패키지인 ‘theano’를 후위 엔진(Backend²⁾ engine)으로 하여 구현하였다.

많은 양의 데이터를 이용한 기계학습 모형의 학습은 충분한 컴퓨터 연산 능력이 충족되어야 가능하다. 서포트 벡터 머신 및 심층 신경망 모형의 학습에 사용된 연구 장비는 HP사의 Z620 모델로, Intel(R) Xeon(R) CPU E5-2697 V2를 프로세서, 64GB RAM을 기본 사양으로 한다. 또한 심층 신경망의 학습은 그래픽 처리 장치의 병렬 연산으로 이루어지며, 이때 사용된 그래픽 처리 장치는 NVIDIA사에서 제작한 워크스테이션용 그래픽 처리 장

2) ‘keras’는 상위-레벨(high-level)의 프로그래밍 언어이기 때문에 데이터를 가공, 처리 및 계산하는 하위 레벨(low-level)의 단계가 필요하며, 이를 행렬의 형태로 최적화 및 계산하는 후위 엔진 ‘theano’가 필요함. 본 연구에서 사용한 ‘keras’는 ‘theano’ 라이브러리만을 후위 엔진으로 사용하였으나, 최근 ‘theano’와 ‘tensorflow’를 모두 사용할 수 있게 최신화 되었음.

치인 Quadro K5200이다. 위와 같은 사양을 기반으로 본 연구에서 제안한 두 가지의 기계학습 모형은 합리적인 시간 내에 학습되었으며, 기존의 하차지점 예측 모형보다 향상된 성능을 보였다.

1) 서포트 벡터 머신(SVM, Support Vector Machine)

[그림 4-5]의 입력 변수 형태를 통해서, 본 연구에서 제안한 서포트 벡터 머신의 목적은 135차원으로 구성된 특성 공간에서 총 5개의 클래스로 이루어진 데이터를 구분하는 134차원(R^{134})의 초평면을 찾는 것임을 확인할 수 있다. 초평면을 찾는 최적화 과정은 선형 제약조건(Linear constraints)을 가진 라그랑지 승수(Lagrange multipliers), 쌍대 문제(Dual problem) 등의 문제를 해결하는 과정을 통해 진행된다. 본 연구에서 서포트 벡터 머신의 최적화 과정에 대한 자세한 설명은 생략하며, 이는 기계학습 모형을 하차지점 예측 모형의 응용에 대한 본 연구의 목적을 벗어남에 그 이유가 있기 때문이다.

파이썬의 ‘skit-learn’ 패키지에서 제공하는 서포트 벡터 머신 함수를 설정함에 있어, 가장 중요한 설정 사항은 커널이다. 일반적으로 낮은 차원의 데이터를 분류하는 초평면을 찾는 것은 간단하지만, 다차원 특성 공간에서 이를 찾는 것은 쉽지 않다. 이때 데이터를 분류하는 초평면을 효율적으로 찾을 수 있도록 사용하는 방법이 커널 방법(Kernel trick)이다. ‘skit-learn’ 패키지에서는 선형(Linear), 다차원 비선형(Polynomial), 그리고 가우시안 방사 기저 함수(Gaussian radial basis function)을 제공한다. 일반적으로 가우시안 방사 기저 함수가 특성 공간의 차원이 큰 경우에 사용된다. 따라서 본 연구의 135차원의 특성 공간을 효율적

으로 학습하기 위해서 가우시안 방사 기저 함수를 사용하였다.

기본적인 모형 설정 후에 앞서 분류한 학습 데이터를 입력하여 모형을 학습(Fitting) 시킨 뒤, 학습된 모형의 결과를 이용하여 검증 데이터를 예측(Predict)하였다. 이때 모형의 예측을 확률로 설정하여 각 대안별 선택 확률을 계산한 뒤, 실제 정류장과 일치하는지 여부를 분석하여 모형의 성능을 계산하였다.

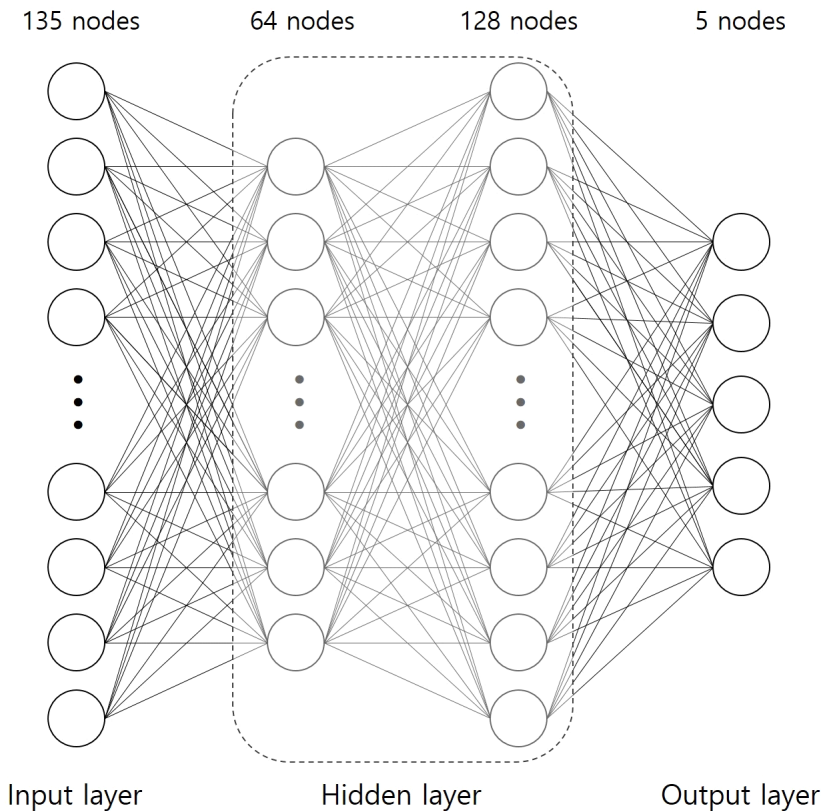
2) 심층 신경망(DNN, Deep Neural Networks)

심층 신경망 모형은 얇은 학습의 서포트 벡터 머신과 다르게, 여러 층으로 이루어진 심층 학습 모형이다. 따라서 하차지점 예측에 최적화된 심층 신경망 모형을 찾기 위해서는 조정해야 할 많은 하이퍼 파라미터(Hyper parameters)가 존재한다. 이러한 최적 하이퍼 파라미터의 발굴은 여러 번의 모형 설정을 통한 결과 비교를 통해서 진행된다. 그러나 이러한 여러 번의 실험을 통해 도출된 결과가 과연 최적해(Optimal solution)인지 의문을 가질 수 있다. 최근 Kawaguchi. (2016)에 의해 연구된 내용에 따르면, 심층 신경망의 학습에 의해 도달한 국지 최소값(Local minima)은 항상 전역 최소값(Global minima)와 같다는 것이다³⁾. 따라서 본 연구에서 여러 번의 시행착오를 거쳐 찾아낸 하이퍼 파라미터의 조합으로 구성된 심층 신경망 모형의 결과는 최적해를 보일 수 있다.

심층 신경망을 구축하기 위해서는 먼저 모형 전체를 구성하는 은닉 층의 수를 결정한 뒤, 각 층의 노드 수, 활성화 함수, 그리고

3) 심층 신경망은 모형의 예측 값과 실제 값의 차이인 손실 값(Loss)을 최소화(Minimization) 하는 가중치(Weights)를 찾는 최적화 과정임

해당 층에 변형을 주거나 추가적인 정보를 추출하기 위한 다양한 기법들이 설정된다. 심층 신경망의 층과 관련된 하이퍼 파라미터를 조정해가면서 최적의 성능을 보이는 모형을 찾는 과정을 거쳐, [그림 5-1]의 최종 심층 신경망 모형을 찾을 수 있었다.



[그림 5-1] 하차지점 예측을 위한 최종 심층 신경망 모형

본 연구에서 제안한 하차지점 예측 심층 신경망 모형은 총 네 개의 층으로 구성되어 있다. 데이터가 입력되는 입력 층과 결과를 출력하는 출력 층, 그리고 두 개의 은닉 층으로 구성되어 있으며, 각 층의 하이퍼 파라미터는 [표 5-5]과 같이 정리된다.

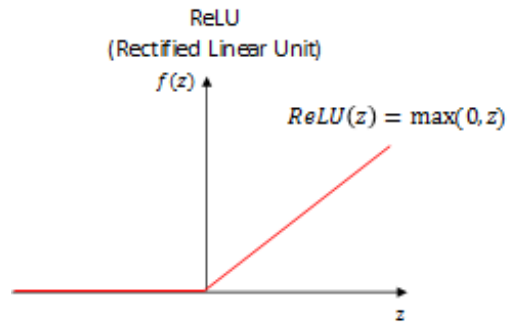
[표 5-5] 심층 신경망 하이퍼 파라미터 정리

	입력층 (Input layer)	은닉층 1 (Hidden layer 1)	은닉층 2 (Hidden layer 2)	출력층 (Output layer)
노드 수 (Number of nodes)	135 (fixed)	64	128	5 (fixed)
활성화 함수 (Activation function)	-	ReLU	ReLU	softmax
드롭아웃 비율 (Dropout ratio)	-	0.3	0.3	-

입력층과 출력층의 노드 수는 데이터의 형태에 의해 결정되는 고정된 값이다. 또한 입력층은 입력 변수의 형태만을 의미하기 때문에 활성화 함수, 드롭아웃과 같은 별도의 장치가 설정되지 않는다. 출력층은 드롭아웃은 설정되지 않지만, 활성화 함수는 ‘softmax’ 함수로 설정된다. ‘softmax’ 함수는 분류를 위한 심층 신경망 모형의 마지막 층인 출력층에서 사용되는 함수로, 분류하기 위한 K개의 클래스에 대한 선택 확률을 계산하는 함수이며, (수식 13)과 같이 표현된다. 모형의 가중치가 θ 일 때 데이터가 특정 정류장 j에서 내렸을 확률 $P(y=j|\theta)$ 는 특정 정류장 j의 가중치(θ_j)와 이전 층의 입력 값(x^T)의 곱의 지수 값이 전체 정류장의 지수 값의 합에서 차지하는 비율로 계산된다.

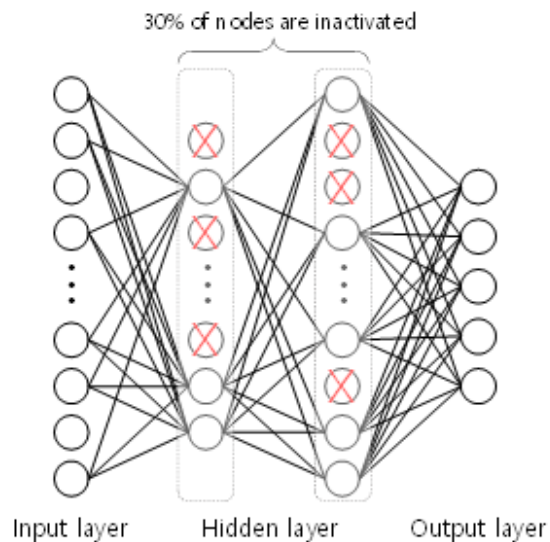
$$P(y=j|\theta) = \frac{e^{x^T \theta_j}}{\sum_{k=1}^K e^{x^T \theta_k}} \quad (\text{수식 13})$$

두 은닉층에 설정된 활성화 함수는 ‘ReLU(Rectified Linear Unit)’으로, 일반적으로 가장 많이 사용되는 활성화 함수이다. ReLU 함수는 입력변수와 가중치가 곱해져 계산된 값(z)이 0보다 작은 경우 0, 0보다 큰 경우에는 z 값이 출력되도록 구성된다(그림 5-2).



[그림 5-2] 'ReLU' 활성화 함수

'ReLU' 활성화 함수는 학습 데이터에 심층 신경망 모형이 지나치게 최적화되어 별도의 데이터에 대한 모형의 성능이 저조한 과적합(Overfitting)의 문제를 방지하기 위해 드랍아웃(Dropout)과 함께 사용된다(Srivastava, N., 2014). 드랍아웃은 특정 층의 노드 중, 일정 비율(Dropout ratio)에 해당하는 노드를 비활성화 시킴으로써 학습을 시키지 않도록 하는 장치이다.



[그림 5-3] 드랍아웃(Dropout) 개념도

각 층에 대한 하이퍼 파라미터를 위와 같이 설정한 뒤, 모형의 최적화 과정에서 설정되는 하이퍼 파라미터를 결정한다. 심층 신경망의 최적화는 모형을 통해 예측한 값과 실제 정답과의 차이를 통해 계산된 ‘Loss’ 값을 최소화 하는 방향으로 진행된다. ‘Loss’를 계산하기 위한 손실 함수(Loss function)는 ‘Categorical cross entropy’로, 본 연구의 하차지점 예측과 같은 다중 클래스 분류(Multi-class classification)에 주로 사용되는 함수이다. 이는 앞서 출력층에 설정한 활성화 함수인 ‘softmax’ 함수와 같이 사용된다.

손실 함수의 최적화에 사용되는 방법(Optimizer)에는 여러 가지가 있으나, 본 연구에서는 Kingma et al. (2014)에 의해 연구된 최적화 방법이자 높은 성능을 보인 ‘Adam(Adaptive Moment Estimation)’을 사용하였다.

앞서 설정된 모형은 전체 데이터를 학습하는 횟수를 의미하는 ‘Epoch’와 가중치를 최신화하기 위한 과정에서 사용되는 데이터의 수인 ‘Batch size(=mini batch size)’를 설정함으로써 최종 학습된다. 본 연구에서는 총 500번(Epochs=500)의 반복을 통해 모형을 학습하였으며, 512개의 데이터를 임의로 추출(Batch size=512)하여 가중치를 최신화 하도록 설정하였다.

3. 종합결과 비교 및 분석

위와 같이 정산된 모형과 분석 결과를 바탕으로, 각 방법론을 통해 학습 및 검증 데이터에 대한 정확도를 산출하였다. 제안 방법론의 정확도 산출에 앞서, 본 연구의 제안 모형의 성능을 비교

하기 위한 Munizaga et al. (2012)의 일반화 시간 모형에 대한 성능을 측정하였다. 또한 Barry et al. (2002)의 기본 원리를 통한 모형의 성능을 추가하였는데, 이는 하차지점 예측의 가장 초석을 마련한 연구라는 점에 그 의의가 있으며, 따라서 해당 원리를 통한 하차지점 예측 정확도와의 모형 성능 비교는 중요한 의미를 갖는다.

본 연구에서 모형의 성능으로 제시하는 정확도는 하차지점이 될 확률이 가장 높은 정류장이 실제 하차지점과 일치하는지 여부를 측정한 정확도와, 하차지점이 될 확률이 가장 높은 정류장과 두 번째로 높은 정류장 중에 실제 하차지점이 있는지 여부를 측정한 정확도 두 가지로 나뉜다. 전자의 경우, 문장 그대로 정확하게 하차지점을 예측하는 경우를 모형의 성능으로 제시한다. 후자의 경우, 확률이 두 번째로 높은 정류장까지 기준을 완화하여 (Second best relaxation) 모형의 성능을 평가하였다. 완화된 기준을 모형의 성능으로 제시한 이유는 대중교통 이용자가 자신의 목적지가 아닌 지점, 즉 실제 하차지점 이전에 하차 태그를 하는 사전태그의 경우를 포함하기 위해서이다. 김경태 외. (2014)에 따르면, 버스 승객의 약 40.3%에 해당하는 인원이 실제 하차지점 이전에 하차 태그를 실시하는 것으로 분석되었다. 학습 및 검증 데이터에 대해서 각각의 정확도를 계산한 결과, [표 5-6]의 최종 결과를 확인할 수 있었다.

먼저 Barry et al. (2002)에 의해 제안된 기본 원리를 통해 하차지점을 예측한 경우가 가장 낮은 정확도를 보였다. 그 다음으로 Munizaga et al. (2012)에 의해 제안된 일반화 시간 모형이 정확히 하차지점을 예측한 경우 54.835%의 정확도를, 완화된

기준에서 하차지점을 예측한 경우 81.522%의 정확도를 보였다. 본 연구에서 제안한 다항로짓 모형인 일반화 시간 모형과 상호작용 모형은 모두 선행 연구의 모형보다 높은 정확도를 보였는데, 상호작용 모형이 일반 모형보다 높은 정확도를 보일 것이라는 기대와 달리 일반 모형이 가장 높은 모형 성능을 확보하였다. 그러나 두 모형의 정확도에는 큰 차이는 없었으며, 이는 1차원 또는 2차원의 관계만을 반영한 다항로짓 모형의 구조적 제한으로 인한 한계라고 생각할 수 있다.

[표 5-6] 선행 연구 모형 및 제안 모형의 성능 비교

		하차지점을 정확히 예측		1개 정류장 완화 (Second best relaxation)	
		학습 데이터 (Training set)	검증 데이터 (Test set)	학습 데이터 (Training set)	검증 데이터 (Test set)
Barry et al. (2002) 기본 원리		47.106	47.165	54.763	54.855
Munizaga et al. (2012) 일반화 시간 모형		55.274	54.835	81.516	81.522
다항로짓 모형 (Multinomial Logit Model)	일반 모형 (Conventional)	57.069 (+1.795)	57.184 (+2.349)	83.329 (+1.813)	83.487 (+1.965)
	상호작용 모형 (Interaction)	57.068 (+1.794)	57.192 (+2.357)	83.325 (+1.809)	83.560 (+2.038)
기계학습 모형 (Machine learning)	서포트 벡터 머신 (SVM)	62.994 (+7.720)	57.670 (+3.835)	88.776 (+7.260)	86.162 (+4.640)
	심층 신경망 (DNN)	61.708 (+6.434)	60.104 (+5.269)	88.850 (+7.334)	87.491 (+5.969)

다음으로 기계학습 모형 정확도의 경우, 학습 및 검증 데이터에 대해서 모두 다항로짓 모형보다 높은 성능을 보였다. 특히 학습 데이터에 대해 정확히 하차지점을 예측한 경우, 서포트 벡터

머신이 62.994%로 가장 높은 정확도를 보였다. 그러나 검증 데이터의 경우, 심층 신경망 모형이 60.104%로 서포트 벡터 머신에 비해 1.434% 높은 결과를 얻었다. 이는 학습한 데이터에 지나치게 최적화되는 과적합의 문제가 서포트 벡터 머신에서는 발생하였으나, 심층 신경망 모형에서는 상대적으로 약하게 발생하였음을 보여준다. 완화된 기준의 성능의 경우 역시 심층 신경망이 가장 높은 정확도를 얻었다. 학습 데이터의 경우, 최대 88.850%에 해당하는 데이터에 대해서 실제 하차지점을 예측했는데, 이는 사전 태그를 고려한 경우 약 89%에 가까운 확률로 하차지점을 예측할 수 있음을 시사한다. 기계학습 모형의 정확도가 다항로짓 모형의 정확도 보다 전체적으로 높은 것은 방대한 차원의 변수들의 비선형 상관관계를 기계학습 모형이 보다 효율적으로 학습하여 모형을 최적화 했다는 의미로 해석할 수 있다.

또한 본 연구에서 제안한 다항로짓 모형 및 기계학습 모형 모두 기존의 연구에 비해 높은 성능을 보였다. 이는 기존 연구에서 사용하지 않은 별도의 데이터인 토지이용 데이터를 사용한 결과라고 할 수 있다.

본 연구에서 심층 신경망 모형을 통해 선행 연구의 모형 대비 5.269% 향상시켜 60.104%의 하차지점 예측 정확도를 달성하였다. 5.269%의 향상은 큰 성능 향상으로 여겨지지 않을 수 있다. 그러나 이러한 성능 향상은 기계학습 분야에서 일반적으로 달성하기 어려운 정도로 여겨지며, 특히 이미지 또는 음성 인식 분야에서는 Krizhevsky, A. (2012)에 의해 이러한 내용이 언급된 바 있다.

결과 분석 내용을 종합하였을 때, 본 연구에서 제안한 하차지점 예측 모형은 성능과 새로운 방법론의 사용 측면에서 그 공헌이 있다고 할 수 있다.

제 6 장 결론 및 향후 연구과제

본 연구에서는 승차에 대한 정보만을 이용하여 하차지점을 예측할 수 있는 하차지점 예측 모형을 제안하였다. 이는 자동요금징수체계를 구축한 전 세계의 많은 도시 중에서, 기·종점이 완전한 정보를 보유한 서울시의 스마트카드 데이터를 사용함으로써 가능했다. 또한 기존에 이루어지지 않았거나 소수의 데이터만을 사용해 이루어진 모형의 검증 역시, 보다 많은 별도의 데이터를 통해 이루어졌다.

하차지점 예측에 일반적으로 통용되어 오던 논리 위주의 하차지점 추정은 Barry et al. (2002)에 의해 본격적으로 시작되었으며, 이후 많은 연구들이 진행되었다. 그러나 하차지점을 예측할 수 있는 특정한 모형을 제시한 연구는 소수였으며, 이는 다른 도시의 스마트카드 데이터를 이용해 하차지점을 예측하기(Time & space transferability) 어렵다는 한계를 보였다. 따라서 본 연구에서는 선행 연구에서 사용하지 않았던 다항로짓 모형과 기계학습 모형을 하차지점 예측에 적용하여 구체적인 모형을 제시했다는 점에 의의가 있다. 이는 본 연구를 통해 학습한 기계학습 모형에 타 도시에 적용함으로써 기존의 방법론 보다 높은 확률로 정확한 하차지점을 예측할 수 있음을 의미한다. 또한 기존의 하차지점 예측 연구에서 사용하지 않은 토지이용 데이터를 추가함으로써 모형의 성능 향상에 기여했다는 점을 언급할 수 있다.

다항로짓 모형의 경우 기존 방법론보다 큰 성능 향상을 보이지 못하였으나, 여러 변수들이 하차지점 예측에 미치는 영향과

각 변수의 통계적 유의성을 확인할 수 있는 장점을 보였다. 하지만 모형의 구조적 특성으로 인해, 변수간의 단순한 선형 관계만을 반영하는 한계를 보였다.

두 개의 기계학습 모형은 모두 다항로짓 모형보다 높은 모형의 성능 향상을 보였다. 서포트 벡터 머신은 학습 데이터에 대해 62.994%의 높은 정확도를 보였으나, 학습 데이터에 지나치게 최적화된 과적합의 문제로 인해 검증 데이터에 대해 57.670%의 정확도를 보이는데 그쳤다. 반면 심층 신경망의 경우, 기존의 방법론 보다 검증 데이터의 경우 5.269% 향상된 60.104%의 성능을 보였으며, 이는 기계학습 분야에서 쉽게 달성할 수 없는 성능 향상임을 확인하였다. 또한 정확하게 하차지점을 예측한 경우와 더불어 완화된 기준에 대한 정확도를 보임으로써, 버스 통행에서 빈번하게 일어나는 사전 태그에 의한 하차지점 예측의 변동성까지 고려하였다. 다음으로 기계학습 모형의 특성 상, 새로운 데이터를 학습함으로써 모형을 지속적으로 최신화할 수 있는 특성을 가지고 있으며, 이러한 특성을 활용하여 하차지점 예측 모형을 관리 및 운영할 수 있을 것으로 기대한다.

앞서 언급된 내용을 종합하여 본 연구에서 최종적으로 달성하고자 하는 바는 다음과 같다. 서울시의 기·종점 정보가 완전한 방대한 스마트카드 데이터를 이용해 개인별 정류장 단위의 하차지점 예측 모형을 구축하였으며, 이를 통해서 기점에 대한 정보만을 보유한 대한민국의 지방 도시 및 세계 여러 도시에 적용할 수 있는 일반화된 모형을 제공하는 가능성을 확인하였다. 이러한 모형은 기존에 막대한 시간과 비용을 들여 조사되어온 가구통행 실태조사를 통한 기·종점 행렬(O-D matrix)을 대체할 수 있는 방

법론이 될 수 있을 것이다.

위와 같은 결론을 통해서 본 연구가 하차지점 예측 분야에 작은 기여를 했다고 할 수 있으나, 본 연구를 진행하는 과정에서 발생한 한계를 인지하고 차후 연구 과제를 언급함으로써 앞으로 하차지점 예측에 관한 연구가 나아갈 방향을 제시할 수 있을 것이다.

첫째로 다항로짓 및 기계학습 모형의 정산과 학습에 사용된 데이터의 수를 표본이 아닌 하루 전체의 데이터를 대상으로 진행해야 한다는 점이다. 이는 보다 많은 데이터를 학습함으로써 기계학습 모형의 성능 향상을 기대할 수 있기 때문이다.

둘째로, 본 연구에서는 하차지점 예측이 가장 어려운 것으로 알려진 버스 통행만을 모형의 대상으로 설정하였으나, 향후 연구에서는 도시철도를 포함한 대중교통 전반에 걸친 하차지점 예측 모형을 연구할 필요성이 있다. 이는 대중교통 전체를 포함함으로써 모형의 일반화를 도모할 수 있으며, 나아가 타 도시에 대한 모형 적용의 효율성을 높일 수 있을 것이다.

마지막으로 토지이용 데이터 이외에 정류장 주변에서 발생하는 활동을 유발하는 파생수요를 반영할 수 있는 추가적인 데이터의 필요성을 이야기할 수 있다. 정류장 주변의 활동을 파악하기 위해서는, 해당 정류장 주변에 위치한 상업 시설의 매출 등과 같은 데이터가 필요할 것이다. 또한 주거지의 경우, 정류장 주변의 상주인구에 대한 정보를 추가함으로써 그 특징을 효율적으로 반영할 수 있을 것이다.

참 고 문 헌

국 내 문 헌

1. 김경태, 민재홍, 이인묵, (2014). 교통카드 데이터를 활용한 사전태그 행태 분석. 2014년 한국철도학회 추계학술대회, pp. 635-638

국 외 문 헌

1. Alsger, A. A., Mesbah, M., Ferreira, L., & Safi, H. (2015). Use of smart card fare data to estimate public transport origin-destination matrix. Transportation Research Record: Journal of the Transportation Research Board, Vol. 2535, pp. 88-96
2. Barry, J., Newhouser, R., Rahbee, A., Sayeda, S. (2002). Origin and Destination Estimation in New York City with Automated Fare System Data. Transportation Research Record: Journal of the Transportation Research Board, Vol. 1817, pp. 183-187
3. Bagchi, M., White, P.R. (2005). The Potential of Public Transport Smart Card Data. Transport Policy, Vol. 12(5), pp. 464-474
4. Barry, J., Freimer, R., & Slavin, H. (2009). Use of entry-only automatic fare collection data to estimate linked transit trips in New York City. Transportation Research

- Record: Journal of the Transportation Research Board, Vol. 2112, pp.53-61
5. Farzin, J. (2008). Constructing an automated bus origin-destination matrix using farecard and global positioning system data in Sao Paulo, Brazil. Transportation Research Record: Journal of the Transportation Research Board, Vol. 2072, pp. 30-37
 6. Kawaguchi, K. (2016). Deep learning without poor local minima. In Advances in Neural Information Processing Systems, pp. 586-594
 7. Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412.6980
 8. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pp. 1097-1105
 9. Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. Transportation Research Part C: Emerging Technologies, Vol. 24, pp.9-18
 10. Munizaga, M., Devillaine, F., Navarrete, C., & Silva, D. (2014). Validating travel behavior estimated from smartcard data. Transportation Research Part C: Emerging Technologies, Vol. 44, pp.70-79

11. Nassir, N., Khani, A., Lee, S., Noh, H., & Hickman, M. (2011). Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2263, pp.140-150
12. Nunes, A. A., Dias, T. G., & e Cunha, J. F. (2016). Passenger Journey Destination Estimation From Automated Fare Collection System Data Using Spatial Validation. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17(1), pp. 133-142
13. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, Vol. 15(1), pp. 1929-1958.
14. Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, Vol. 11(1), 1-14
15. Wang, W., Attanucci, J. P., & Wilson, N. H. (2011). Bus passenger origin-destination estimation and related analyses using automated data collection systems. *Journal of Public Transportation*, Vol. 14(4), 7.
16. Zhao, J., Rahbee, A., Wilson, N. H. (2007). Estimating a Rail Passenger Trip Origin-Destination Matrix Using

Automatic Data Collection Systems. Computer-Aided Civil and Infrastructure Engineering, Vol. 22(5), pp. 376-387

17. Zhang, L., Zhao, S., Zhu, Y., & Zhu, Z. (2007, September). Study on the method of constructing bus stops OD matrix based on IC card data. In Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on IEEE, pp. 3147-3150
18. Zhang, F., Yuan, N. J., Wang, Y., & Xie, X. (2015). Reconstructing individual mobility from smart card transactions: a collaborative space alignment approach. Knowledge and Information Systems, Vol. 44(2), pp. 299-323

국 문 초 록

딥러닝 기법을 이용한 버스 하차지점 예측 모형에 관한 연구

정 재 영

토목공학과 교통공학전공

중앙대학교 대학원

스마트카드 데이터는 대중교통 이용자의 통행 행태를 반영하는 중요한 정보이며, 전 세계 여러 도시에서 스마트카드 데이터를 이용한 통행 행태 분석에 관한 연구가 활발히 진행되고 있다. 그러나 대부분의 도시에서 승차에 대한 정보만을 가진 스마트카드 데이터를 이용하여 연구가 진행되고 있으며, 따라서 대중교통 이용자의 승차와 하차 정보를 모두 포함한 별도의 데이터를 통한 모형의 검증이 어려운 실정이다.

본 연구에서는 서울시의 기·종점 정보가 완전히 반영된 스마트카드 데이터를 이용하여 기존에 연구되지 않은 다항로짓 및 기계학습 모형을 통한 하차지점 예측 모형을 제안하였다. 또한 토지이용 데이터를 추가함으로써 모형의 설명력을 높이고 성능을 향상시킬 수 있었다. 대다수의 선행 연구에서 시행하지 못한 별도의 데이터를 통한 모형의 검증을 통해서 본 연구의 하차지점 예측 모형의 성능을 입증하였다.

완전한 정보를 보유한 서울시의 스마트카드 데이터와 토지이용 데이터를 이용해 효율적이며 새로운 일반화된 하차지점 예측 모형을 구축한다. 구축한 모형을 기점에 대한 정보만을 보유한 (Entry-only) 전 세계의 여러 도시에 도입함으로써, 대중교통 이용자의 통행 행태를 보다 정확하게 파악할 수 있는 밑거름이 될 수 있음에 본 연구의 의의가 있다.

핵심어 : 스마트카드, 딥러닝, 기계학습, 하차지점, 버스, 토지이용 데이터, 다항로짓 모형, 심층 신경망

ABSTRACT

Predicting Smart-Card Holders' Bus-Alighting Locations Using a Deep Learning Technology

Jaeyoung, Jung

Dept. of Civil Engineering

The Graduate School,

Chung-Ang University

Smart-card data is crucial information for transportation analysis since it reflects the behaviour of public transportation users; thus numerous researchers around the world actively conduct researches with smart-card data to analyze the transit users' behavior. However, most of the researchers conducted researches based on entry-only smart-card data, which has only passengers' boarding information of public transit, not the alighting information. Therefore, it was difficult to predict the alighting location and validate its prediction model with exogenous data that reflects complete information.

In this study, I suggest novel models to forecast the

alighting locations by adopting Multinomial Logit and supervised Machine Learning methodologies that have not been studied before. Moreover, by supplementing the land-use data, I could improve that the models' power of explanation and performance. The performance of the proposed models to predict alighting stops has been validated through the exogenous data acquired from smart-card data in Seoul, Korea. The validation process is meaningful, since such approach using large amount of exogenous data has not rarely been attempted from previous researches.

In conclusion, present study proposes the novel, efficient and generalized models to predict the alighting locations of bus users, due to the complete smart-card and land-use data of Seoul. Therefore, if the proposed models could be introduced to other cities that has entry-only smart-card data, the prediction of the alighting stops can be improved.

Keywords : Smart-card, Deep learning, Machine learning, Alighting location, Bus, Land-use data, Multinomial logit model, Deep Neural Network