



An Adaptation of Bayesian Theorem for Passenger Destination Estimation

승객 목적지 추정을 위한 베이지안 이론 적용

Sanghyuck Lee

tkdgur658@gmail.com

2021. 01. 05.

Outline

- Introduction
- Related Works
- Proposed Method
- Further Research

Introduction

- **Intelligent Transportation Systems (ITS)** is an innovative idea in the field of transportation and the transit **Origin-Destination matrix (OD matrix)** is an essential input for some important ITS applications
- The powerful source data for generating the OD matrix is **Automated Fare Collection (AFC)** data. However, many AFC systems around the world are **entry-only systems** and do **NOT include destination information** so that essentially we have to go through the **Passenger Destination Estimation (PDE)** process..
 - Moreover, in the real world, even within the datasets of AFC system that contains alighting information, some of them is often omitted due to reasons such as passengers leaving without tagging.
- Therefore, **improving the PDE algorithm** based on AFC data can be a great contribution to the entire ITS field.

Related Works

Barry et al. (2002), using **two assumptions**, proposed the **Trip-Chain method** for PDE. Then, several subsequent researchers, based on these **two assumption**, studied PDE (Zhao and Jinhua, 2004).

However, these past studies are mainly based on data from **entry-only AFC systems**, so it has been pointed out that **verification is necessary** and based on AFC data including **both boarding and alighting information**, some studies for **validating the algorithms** based on entry-only AFC data (Alsger et al., 2016).

Card ID	Card style	Route	Stop	Veh. ID	Time	Date	Fare
-1760518818	Monthly card	115	156	2798	06:27:16	2009-07-07	-
-1760518818	Monthly card	115	3	2346	18:47:21	2009-07-07	-
-1760517522	Common card	115	19	2810	08:34:53	2009-07-07	0.9
-1760517522	Common card	115	151	2808	17:47:31	2009-07-07	0.9
-1760515746	Monthly card	115	133	2794	11:56:22	2009-07-07	-
-1760515746	Monthly card	115	133	2814	18:52:29	2009-07-07	-
-1760515714	Monthly card	115	23	2794	17:32:51	2009-07-07	-

AFC data NOT including alighting information

Transaction Number	Boarding Time (yyyy/mm/dd/hh/mm/ss)	ID of Boarding Location	Alighting Time (yyyy/mm/dd/hh/mm/ss)	ID of Alighting Location
1	20150318092455	4196066	20150318092714	4196117
2	20150318214734	4196061	20150318215035	4196065
3	20150318225205	12174	20150318225855	8001017
4	20150318075726	70647	20150318082805	10585
5	20150318070948	216	20150318084718	4101317
...	...	...	...	...

AFC data including alighting information

Related Works

Moreover, recently, some studies based on **Machine Learning** methods, including Deep learning approaches, have been conducted:

Jung et al. (2017) first proposed a PDE approach using a deep learning approach.

Assemi et al. (2020) constructed a deep learning structure with the Trip-Chain method.

Yan et al. (2019) used Trip-Chain method as the first step and 6 Machine Learning model as the second step.

Cheng et al. (2020) proposed a three-dimensional latent Dirichlet allocation model to extract latent topics of each trips.

Shi et al. (2017) applied Trip-Chain method first, and then used semi-supervised self-training model for PDE.

=> There are **some limitations**: Low explanatory power, Lack of scalability, Dependence on the performance of the Trip-Chain method

Proposed Method

An Adaptation of Bayesian Theorem for Passenger Destination Estimation

- Motivation:
 - A system that can access not only boarding but also **alighting information** has been developed, so that the **posterior probability** can be calculated through Bayesian theory, which shows successful results in many inference problems.

$$p(y|x) = \frac{p(x|y) p(y)}{p(x)} \quad \longrightarrow \quad p(s|c) = \frac{p(c|s) p(s)}{p(c)} \quad (1)$$

Terms	Meaning
$\mathcal{C}$	The set of conditions for determining passenger destinations, $\mathcal{C} = \{c_1, \dots, c_n\}$, where n is the number of all the conditions.
$\mathcal{S}$	The set of destination candidate stations, $\mathcal{S} = \{s_1, \dots, s_m\}$, where m is the number of all the transit station.

Proposed Method

$$p(s|c) = \frac{p(c|s) p(s)}{p(c)} = \frac{p(c, s)}{p(c)} \quad (1)$$

$$s^* = \operatorname{argmax}_{s \in S} p(s|c) = \operatorname{argmax}_{s \in S} \frac{p(c, s)}{p(c)} \quad (2)$$

$$= \operatorname{argmax}_{s \in S} p(c, s) = \operatorname{argmax}_{s \in S} p(c_1, \dots, c_n, s) \quad (3)$$

Using chain rule:

$$\begin{aligned} & p(c_1, \dots, c_n, s) \\ &= p(s) p(c_1|s) p(c_2, \dots, c_n|s, c_1) \\ &= p(s) p(c_1|s) p(c_2|s) p(c_3, \dots, c_n|s, c_1, c_2) \\ &= p(s) p(c_1|s) p(c_2|s) \cdots p(c_n|s, c_1, c_2, c_3, \dots, c_{n-1}) \end{aligned}$$

Adapt Naïve assumption:

$$\begin{aligned} &= p(s) p(c_1|s) p(c_2|s) p(c_3|s) \cdots \\ &= p(s) \prod_{i=1}^n p(c_i|s) \end{aligned} \quad (4)$$

Proposed Method

By (3) and (4):

$$\begin{aligned} s^* &= \operatorname{argmax}_{s \in S} p(s|c) \\ &= \operatorname{argmax}_{s \in S} p(s) \prod_{i=1}^n p(c_i|s) \end{aligned} \tag{5}$$

Further research

필드명	필드 내용	필드명	필드내용
카드번호	연번으로 암호화 되었음	환승횟수	
운행출발일시		버스노선ID	
트랜잭션ID	- 환승 거래를 그룹핑하는 단위에 부여하는 id (환승 거래가 있는 경우 최초 승차부터 최종 하차까지를 하나의 거래로 그룹핑하기 위한 id) - 일자내 카드번호별로 같은 환승 거래 내역이 있는 경우 같은 ID를 부여 (5/5 일자 A라는 카드가 버스 에서 지하철을 환승한 경우 버스, 지하철에 거래내역 01라는 id를 부여)	교통사업자ID	
		차량ID	
		교통카드발행사ID	
		교통카드사용자구분코드	
		과금기준사용자구분코드1	
		과금기준사용자구분코드2	
		과금기준사용자구분코드3	
		승차일시	
교통수단CD	100-버스 101-도시형버스 102-일반좌석버스 ... 202-한국철도공사 203-도시철도 204-인천지하철	승차정류장ID	
		하차일시	
		하차정류장ID	
		승객수1	
		승객수2	
		승객수3	
		이용거리	

<Seoul AFC data description>

Further research

필드명	필드 내용
교통운영기관ID	
교통운영기관명	
노선ID	
정류장ID	
정류장순서	
버스노선명	
버스정류장명	
정류장수	
버스정류장누적거리	
정류장 X좌표	※ 좌표계 : WGS84, EPSG : 4326
정류장 Y좌표	※ 좌표계 : WGS84, EPSG : 4326
행정동ID	
버스정류장ARS번호	

<Bus stop data description>

필드명	필드 내용
교통운영기관ID	
교통운영기관명	
호선코드	
호선명	
역ID	
역명	

<Subway station data description>

Appendix

Preliminary: a simple PDE example

- Use 2 simple assumptions:
 - 1) The destination of all the trips is the origin of the next trip.
 - 2) The last destination is the first origin.

Terms	Meaning
X	The AFC dataset consisting of card ID, boarding station and boarding time, $X = \{x_1, \dots, x_n\}$, $x = (x^c, x^s, x^t)$, where n is the number of all the transactions.
S	The set of all the transit station in X , $S = \{s_1, \dots, s_m\}$, where m is the number of different transit station in X .
C	The set of all the unique card ID in X , $C = \{c_1, \dots, c_l\}$, where l is the number of unique card ID in X .
x^c, x^s, x^t	Card ID, boarding station and boarding time of the transaction x respectively, $x^c \in C$, $x^s \in S$, $x^t = [0, 1440]$.
T	The set of all the trips that should be estimated from X , $T = \{t_1, \dots, t_n\}$, $t = (o, d)$.
o	The origin of the trip that should be estimated at t of T .
d	The destination of the trip that should be estimated at t of T .
X^c	The subset with x 's in X that have the same card ID c , $X^c = \{x x^c = c\}$.
T^c	The set of the trips that will be estimated from X^c , that should consist of trips derived from the single card ID.

Preliminary: a simple PDE example

Algorithm 1 A simple PDE algorithm			Meaning
Input: X, S ; Output: T ; 1: Initializing $T=\{\}$; 2: Generate C containing unique card ID from X ; 3: for c in C do 4: $T^c = \text{Estimate}(X^c)$; 5: concat(T, T^c); 6: end for 7: return T		X	The AFC dataset consisting, $X = \{x_1, \dots, x_n\}$, $x=(x^c, x^s, x^t)$.
		S	The set of all the transit station in X , $S = \{s_1, \dots, s_m\}$, where m is the number of different transit station in X .
		C	The set of all the unique card ID in X , $C = \{c_1, \dots, c_l\}$, where l is the number of unique card ID in X .
		x^c	Card ID of x , $x^c \in C$.
		x^s	boarding station of x , $x^s \in S$.
		x^t	boarding time of x , $x^t = [0,1440]$.
		T	The set of all the trips that should be estimated from X , $T = \{t_1, \dots, t_n\}$, $t = (o, d)$.
		o	The origin of the trip
		d	The destination of the trip
		X^c	The same card ID subset of X , $X^c = \{x x^c = c\}$.
Algorithm 2 Procedure of the Estimate function		T^c	The set of the same card ID trips that will be estimated from X^c .
Input: X^c ; Output: T^c ; 1: Ascend X^c according to boarding time x^t ; 2: for ($i = 1$; $i < X^c $; $i++$) do 3: $o_i \leftarrow x_i^s$; $d_i \leftarrow x_{i+1}^s$; 4: end for 5: $o_{ D^c } \leftarrow x_{ D^c }^s$; $d_{ D^c } \leftarrow x_1^s$; 6: return T^c			