거시적 동선 정보에 기반한 고객 재방문 예측

김범영[O1] 김선동[2] 김세진[3] 이재길[1]
[1]한국과학기술원 산업및시스템공학과 [2]기초과학연구원 데이터 사이언스 그룹 [3]한국과학기술원 지식서비스공학대학원
dglidgli@kaist.ac.kr, sundong@ibs.re.kr, {ksj614, jaegil}@kaist.ac.kr

Customer Revisit Prediction Using Macroscale Mobility Information

Beomyoung Kim[O1] Sundong Kim[2] Sejin Kim[3] Jae-Gil Lee[1]
[1]Department of Industrial and Systems Engineering, KAIST
[2]Data Science Group at Institute for Basic Science
[3]Graduate School of Knowledge Service Engineering, KAIST

요 약

거시적 동선 정보란 고객의 경로를 공간적, 시간적으로 표현한 것을 의미한다. 여러 회사들에서 무선 핑거프린팅 등의 기술을 통해 고객의 거시적 동선 정보를 확보하였고 그들은 다양한 분석 방법뿐만 아니라 재방문 예측의 필요성을 느끼고 있다. 본 연구에서는 서비스의 활용도를 더욱 높이기 위해서 단일 매장보다 특정 장소 범주(브랜드, 카테고리)에 관한 고객의 재방문 예측을 위한 모델을 제시한다. 하지만 이때 모델의 학습을 위해 특정 시점까지 만의 데이터를 활용하는 과정에서 그 이후 시점 고객의 재방문 여부를 알 수 없는 데이터 희소성 문제가 발생한다. 이를 해결하기 위해 생존 분석을 도입하고자 한다. 최근 WTTE-RNN이라는 방법론은 이러한 생존 분석과 함께 모델에 RNN을 접목함으로써 시간에 따라 변하는 데이터의 패턴을 반영하였다. 이에 본 연구는 거시적 동선 정보에서의 재방문과 관련된 특징과 임베딩 요소를 정의함으로써 WTTE-RNN 방법론을 확장하여 고객의 재방문 예측을 수행한다.

1. 서 론

거시적 동선 정보란 고객이 어떠한 순서로 움직였는지 위, 경도와 타임 스탬프를 통해 공간적 · 시간적으로 표현한 것을 의미한다. 이때 해당 위, 경도에 어떤 매장이 있는지와 같은 의미론적인 정보로부터 고객의 오프라인 방문 행태를 파악할 수 있다. 본 연구에서는 무선 핑거프린팅(WiFi-Fingerprinting) 기술을 통해 얻은 거시적 동선 정보를 통해 특정 장소 범주(카테고리)에 재방문하는 시간을 예측하고자 한다.

무선 핑거프린팅 기술은 가장 가까운 매장, 건물의 층수 등 매장 내부부터 외부까지 GPS보다 더 정확한 위치 정보를 제공한다. 덕분에 오프라인의 고객 움직임 기록을 활용한 매장의 퍼널 분석, 핫스팟 분석 등이 가능해졌고, 특정 매장 내에서 수집된 데이터만을 활용하여 성공적으로 고객의 재방문을 예측할 수 있게 되었다.[1] 이와 동시에 더 넓은 범위의 거시적 동선 정보를 확보한 회사들에서도 여러 분석 방법들과 재방문 예측에 관한 필요성을 느끼고 있다.

일례로, Loplat X[1)]는 고객사의 애플리케이션에 위치 정보 서비스를 제공하는 방법을 통하여 근처에 있는 매장에 대한 쿠폰이나 정보를 제공하여 결제를 유도하는 등 다양한 서비스를 제공하고 있다. 특히 거시적 동선 정보를 활용한 다음 방문 장소 예측 및 재방문 예측은 타겟 마케팅 및 고객군 분류 등의 다양한 분야로의 응용도가 높다. 예로, 특정 매장의 지점에 재방문할 고객군의 선별이 가능해지면, 근처에 입점하는 신규 매장의 초대권을 제공한다거나, 대량 구매 시 할인 기회를 제공하는 방법 등을 통해 매출을 증진할 수 있다. 서비스의 활용도를 높이기 위해서 단일 매장보다는 브랜드, 분류군에 해당하는 장소 범주의 재방문 예측 모델의 개발이 필요하다. 따라서 본 연구에서는 거시적 동선 정보를 활용하여 특정 장소 범주(브랜드, 카테고리)에 관한 고객의 재방문 예측을 위한 심층 생존 분석 모델을 제시한다.

생존 분석은 환자의 죽음이나 기계의 고장 등 관심 있는 사건의 발생 시간을 분석하는 통계학의 한 분야이다. 생존 분석을 활용한 모델을 제시하는 이유는 다음과 같다. 한정된 데이터 수집 기간을 가진 애플리케이션에서는 데이터 희소성 문제가 존재하는데, 이는 연구 기간 내에 해당 사건이 일어나지 않아 정확한 정답을 알 수 없는 경우를 일컫는다. 일례로, 모델의 학습을 위해 특정 시점까지 만의 데이터를 활용하는 경우에, 모델을 학습하는 과정에서는 그 이후 시점 고객의 재방문 여부를 알 수 없으며, 생존 분석에서 이를 중단 자료(censored data)라고 한다.

생존 분석의 대표적인 모델로는 Cox proportional hazards model[2]이 있다. 해당 사건이 일어나는 시간에 대해 준 모수적 방법으로 공변량(covariate)의 로그-선형 결합이 사건 발생에 영향을 준다. 더 나아가 최근 머신러닝, 딥러닝의 발전과 함께 공변량의 관계를 더욱 복잡하게 표현한 Random Survival Forest[3]와 Deepsurv[4]등이 존재한다. 하지만 이와 같은 방법은 사건의 발생과 관련된 특징을 제안해야 하고 동선 정보의 순서상 관계를 정확히 반영하지 못한다. 딥러닝의 RNN 모델을 활용하여 이러한 관계를 반영하고자 하는 연구도 있었다.[5] 그중에서도 WTTE-RNN[6]은 확률분포 함수를 통해 직관적인 해석이 가능하다. 이에 본 연구는 거시적 동선 정보에서의 재방문과 관련된 특징과 임베딩 요소를 정의함으로써 WTTE-RNN 방법론을 확장하여 고객의 재방문 예측을 수행한다.

제시하는 심층 생존 분석 모델의 우수성을 입증하기 위해 Loplat X가 한국 전역에서 6개월간 수집한 25,038명의 500만 건의

1) http://loplat.com/

매장 방문 기록 데이터를 활용하였다. 이는 재방문 예측을 하기에 충분히 많은 로그이며, 각 매장은 카테고리로 범주화할 수 있으며, 카테고리는 커피숍, 음식점, 도서관 등 총 243개로 세분되어 있다.

2. 본 론

2.1 문제 정의

앞서 언급한 것과 같이 동선 정보란 고객별로 어떤 장소에 방문했는지 위치와 시간 정보의 집합을 의미한다. 이때 U를 고객, L를 매장, T를 시간 C를 카테고리 집합이라고 하였을 때, 각 방문 v는 $(u,l,t,c)\in U\times L\times T\times C$로 표현할 수 있고 이는 고객 u가 t시간에 카테고리가 c인 매장 l에 방문했음을 나타낸다. 이때 여러 매장이 하나의 카테고리 c에 포함될 수 있는데 이를 다음과 같이 표현할 수 있다. $\{l_1,l_2,\dots l_k\}\subseteq c$. 특정 장소 범주(카테고리) $c\in C$와 시간 $t\in T$가 주어졌을 때 t시간 이후 해당 범주에 처음 재방문할 시간 간격을 구하는 것이 문제의 목표이다.

2.2 중단 자료(Censored data)

각 고객마다 동선 정보의 시작부터 관찰 종료 시점까지를 관찰 기간 o^u라고 하면 고객이 그 기간 안에 특정 카테고리 c에 방문하지 않을 수 있다. 이러한 정답이 없는 데이터는 학습에 문제가 된다. 이때 특정 카테고리 c에 방문하는 시간을 정확히는 모르지만, o^u이후라는 것만 알 수 있다. 즉 실제 관찰 시간 y를 이렇게 표현할 수 있다. 여기서 y^{u^*}는 각 고객의 실제 사건 발생 시간이다.

$$y^u=\begin{Bmatrix} y^{u^*} & \text{if} & uncensored\,data \\ o^u & \text{if} & censored\,data \end{Bmatrix}$$

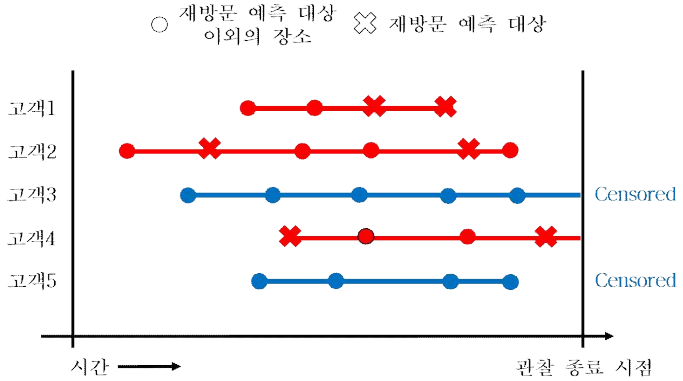

그림 1 거시적 동선 정보와 중단 자료

위 그림에서 파란색의 경우 중단 자료이므로 $y^u=o^u$이고 빨간색의 경우는 완전 자료(uncensored data)이므로 $y^u=y^{u^*}$이다.

2.3 Loss 함수

Loss 함수의 경우 중단 자료인지 여부에 따라 다르게 설정된다. 먼저 카테고리 c에 방문한 확실한 정답이 있는 완전 자료인 경우 그 시간에 일어날 확률을 최대화한다. 그리고 중단 자료의 경우 카테고리 c이외에 마지막 방문이 있었던 시간 이후에 사건이 일어날 확률을 최대화하는 방식으로 학습을 한다.

$$\sum_{n=1}^{N}\sum_{t=0}^{T_n} u_t\log\left[\Pr\left(Y_t^n=y_t^n\,|\,x^n_{t_{0:t}}\right)\right] + (1-u_t^n)\log\left[\Pr\left(Y_t^n>y_t^n\right)|\,x^n_{0:t}\right)]$$

$$u=\begin{Bmatrix} 1 & \text{if} & uncensored\,data \\ 0 & \text{if} & censored\,data \end{Bmatrix}$$

y는 위와 같이 사건 관찰 시간을 의미하고 $x_{0:t}$는 t까지의 동선 정보로부터 얻은 특징이다. T_n는 총 경로를 부분 경로로 나눈 횟수인데 본 연구에서는 T_n= 1로 학습을 진행한다.

2.4 Weibull Distribution & Model

와이블 분포는 기계의 고장 등 사건의 발생과 관련된 분포로 많이 사용되었다. 본 연구에서도 재방문 시간 함수가 와이블 분포를 따른다고 가정하여 Loss 함수를 확률 분포 함수로 직관적으로 나타낸다. 와이블 분포의 경우 두 가지 변수 α,β로 이루어진다. 와이블 분포의 이산적 형태로 가정하였을 때 앞서 설명한 Loss 함수는 이처럼 표현할 수 있다.

$$L=\log\left(p(y)^u S(y+1)^{1-u}\right)$$
$$=u\log\left(e^{(\frac{y+1}{\alpha})^\beta-(\frac{y}{\alpha})^\beta}-1\right)-(\frac{y}{\alpha})^\beta$$

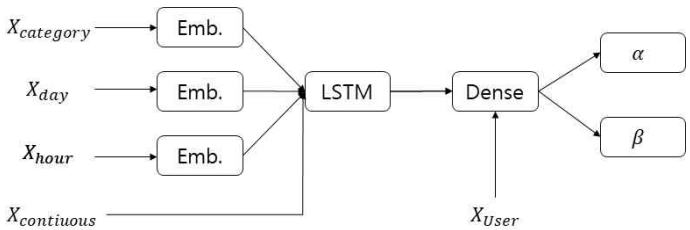

그림 2 WTTE-RNN을 참고한 Model

변수 α,β를 학습하기 위해 본 연구에서는 방문 v마다 생기는 특징과 동선 전체 기간에 대한 특징을 모두 이용한다. 그림 2에서 임베딩 특징($X_{category}$, X_{day}, X_{hour})들과 연속적인 특징($X_{continuous}$)은 각 방문을 나타내는 특징들로 RNN을 통해 동선의 순서를 학습한다. 이때 미래에 어떤 방문을 할지는 가까운 과거일수록 영향이 크다는 가정하에 마지막으로부터 일정 시점 이전까지의 방문만을 선택하여 임베딩 한다. 임베딩 하는 과거 방문의 개수는 실험에 따라 조절할 수 있다. X_{user}는 동선 전체 기간에 대한 전반적인 특징들을 나타낸다. 사용된 특징들은 아래와 같다.

- 임베딩 특징
1) 카테고리 (유일한 243개의 카테고리)
2) 요일 (총 7개의 요일)
3) 시간대 (1시간 단위로 나눈 24개의 시간대)
4) 고객 Id (유일한 24646명의 고객)

- 연속적인(Continuous) 특징
1) 다음 방문까지의 시간

- 사용자 특징(전체 구간의 전반적인 특징)
* 이 특징들은 Cox Model에서도 동일하게 사용한다.
1) 전체 기간 대비 목표 카테고리(c) 방문 횟수
2) 방문이 있는 일수 대비 방문 횟수
3) 마지막 목표 카테고리 방문 이후 지난 시간
4) 전체 기간 대비 주말 횟수
5) 목표 카테고리 사이의 평균 시간 간격
6) 모든 방문 기록 사이의 평균 시간 간격
7) 전체 방문 횟수
8) 방문이 있는 날짜 수

3. 결 과

3.1 실험 결과 및 비교

앞서 소개한 중단 자료를 학습에 사용할 수 있는 Cox 알고리즘과의 비교를 통해 재방문 시간 예측 성능을 비교한다. 측정 항목으로는 평균 제곱근 오차(Root Mean Square Error; RMSE), C-index [7](Concordance Index), Non-returning Recall, Non-returning F1 score를 사용한다. 첫 번째 평균 제곱근 오차는 일반적인 회귀 문제에서 가장 자주 쓰이는 평가 기준으로 예측값과 실제 값의 오차의 정도를 나타낸다. 이는 방문 시간이 존재하는 경우에만 계산할 수 있는데 즉, 완전 자료의 오차의 크기를 의미한다. 두 번째로 C-index는 생존 분석에서 많이 사용되는 평가 기준으로 아래와 같다.

$$c = \frac{1}{|\varepsilon|} \sum_{T_i\, uncensored\, T_j > T_i} 1_{f(x_i) < f(x_j)}.$$

$|\varepsilon|$ 는 인스턴스의 개수, $f(x)$는 예측된 사건 발생 시간을 의미한다. 실제로 발생한 사건 시간의 순위와 예측값의 순위의 유사한 정도를 나타낸다. 마지막으로 Non-returning Recall과 Non-returning F1 score는 주어진 관찰 기간 안에 실제 방문 값이 없을 때를 참이라 하였을 때 이진 분류의 정확성을 나타내는 평가 기준이다. 재방문 모델은 관찰 기간 안에 실제 정답이 존재하는 경우 실제 값과의 오차를 적게 예측해야 하는 것뿐만 아니라 그렇지 않은 경우는 연구 기간 이상으로 예측하여 재방문이 해당 기간 안에 없다고 판단할 수 있어야 한다.

표 1 'Coffee shop'에 대한 실험결과

	Cox	WTTE-RNN
C-index	0.698	**0.726**
RMSE	205.24	**199.49**
non-returning recall	**0.488**	0.438
non-returning f1	0.543	**0.555**

'Coffee shop'이라는 카테고리를 기준으로 실험을 한 결과이다. 기존의 생존 분석의 모델 중 하나인 Cox보다 실제 재방문 시간과의 오차도 적고 재방문이 없는 고객에 대해서도 알맞게 분류하는 것을 볼 수 있다.

3.2 결론 및 추후 연구

앞서 언급하였듯 재방문 예측은 타겟 마케팅 및 고객군 분류 등의 다양한 분야로의 응용도가 높아 연구의 필요성이 있다. 본 연구에서는 생존 분석의 방법을 통해 데이터 희소성 문제를 해결하고 RNN 모델을 통해 방문의 순서상 관계를 반영하였다. 또한, 거시적 동선 정보에서의 재방문과 관련된 특징과 임베딩 요소를 정의함으로써 WTTE-RNN 방법론을 확장하여 고객의 재방문 예측을 성공적으로 수행하였다.

이후에는 데이터를 전처리하여 의미 있는 방문만을 이용하여 성능을 낮추지 않으며 빠르게 예측할 수 있는 모델을 만들 수 있을 것이다. 또한, 거시적 동선으로부터 공간적, 시간적 정보를 잘 반영하는 고객, 장소 임베딩을 구한다면 X_{user}와 같은 종합적인 특징 없이도 방문 간 순서상의 정보를 통해 좋은 결과를 얻을 수 있을 것이다.

참 고 문 헌

[1] Sundong Kim and Jae-Gil Lee, "Utilizing In-store Sensors for Revisit Prediction," *IEEE International Conference on Data Mining*, pp. 217-226, 2018.

[2] David Roxbee Cox, "Regression Models and Life Tables (with Discussion)," *Journal of the Royal Statistical Society: Series B*, Vol. 34, No. 2, pp. 187-220, 1972.

[3] Hemant Ishwaran et al., "Random Survival Forests," *The Annals of Applied Statistics,* Vol. 2, No. 3, pp. 841-860, 2008.

[4] Jared Lee Katzman et al., "DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network," *BMC Medical Research Methodology,* Vol. 18, No. 1, p. 24, 2018.

[5] Guolei Yang et al., "Spatio-Temporal Check-in Time Prediction with Recurrent Neural Network based Survival Analysis," *International Joint Conference on Artificial Intelligence,* pp. 2976-2983, 2018.

[6] Egil Martinsson, "WTTE-RNN : Weibull Time to Event Recurrent Neural Network," Chalmers University of Technology, Gothenburg, Sweden, 2016.

[7] Vikas Chandrakant Raykar et al., " On Ranking in Survival Analysis: Bounds on the Concordance Index," *Advances in Neural Information Processing Systems 20*, pp. 1209-1216, 2008.