

# 그라디언트 부스팅을 활용한 GPS 기반 이상 궤적 탐지

조원희<sup>1‡</sup>, 김시은<sup>2‡</sup>, 김범수<sup>3</sup>, 양형정<sup>1\*</sup>

<sup>1</sup>전남대학교 인공지능융합학과

<sup>2</sup>전남대학교 정보보안협동과정

<sup>3</sup>댓츠잇 주식회사

<sup>‡</sup>공동 1저자, <sup>\*</sup>교신저자

{206465, \*hjyang}@jnu.ac.kr

## Anomaly Trajectory Detection based on GPS using Gradient Boosting

Won-Hee Jo<sup>1‡</sup>, Si-Eun Kim<sup>2‡</sup>, Bum-Soo Kim<sup>3</sup>, Hyung-Jeong Yang<sup>1\*</sup>

<sup>1</sup>Department of Artificial Intelligence Convergence, Chonnam National University

<sup>2</sup>Interdisciplinary Program of Information Security, Chonnam National University

<sup>3</sup>Thatzit Co., Ltd.

<sup>‡</sup>co-first author, <sup>\*</sup>corresponding author

### 요약

위치정보 산업의 시장 규모는 매년 빠르게 증가하고 있으며 이러한 서비스에서 일부 사용자들은 GPS 조작 프로그램을 악용하여 부당한 이득을 취한다. 이와 같은 부정행위를 탐지하기 위해 기존의 방법은 규칙 기반 시스템을 사용하지만 이러한 시스템은 새로운 패턴이나 대량의 데이터에서 성능이 하락하는 문제가 있다. 본 연구에서는 GPS 궤적 데이터의 통계적 정보를 활용하여 그라디언트 부스팅 모델을 통한 이상 궤적 탐지 방법을 제안한다. 모델의 성능 평가를 위해 SVM, MLP 등의 모델과 비교하였으며 이상 궤적에 대해 정밀도 91.3%, 재현율 90.0%, F1-Score 90.4%로 다른 모델들에 비해 높은 성능을 보였다.

### 1. 서론

위치정보 산업의 시장 규모는 매년 빠르게 증가하고 있다. 그 중 생활 및 엔터테인먼트 서비스는 두 번째로 높은 점유율을 차지하고 있으며[1] 모바일 앱과 GPS(Global Positioning System) 기술을 결합한 형태로 다양한 서비스를 제공한다[2, 3, 4]. 이러한 서비스에서 일부 사용자는 이익을 얻기 위해 조작 프로그램을 사용하여 가짜 GPS 정보를 생성한다.

우버나 포켓몬고와 같은 인기 있는 위치기반 서비스에서는 부정 사용자를 탐지하기 위해 규칙 기반 시스템[5]을 사용한다. 기존의 규칙 기반 시스템은 특정 패턴에 따라 이상 행위를 탐지하는 방법으로 새로운 패턴이 출현하거나 데이터가 많아지면 성능이 떨어지는 한계를 가진다[6].

본 연구에서는 GPS 궤적 데이터의 통계적 정보를 활용하여 그라디언트 부스팅(Gradient Boosting) 모델을 통해 이상 궤적을 탐지한다. 탐지 알고리즘의 성능은 정밀도(Precision), 재현율(Recall), F1-Score를 사용하여 평가한다. 본 연구에서 제안한 그라디언트 부스팅 모델이 다른 모델에 비해 F1-Score에서 좋은 성능을 보였다.

### II. 스탬프투어 GPS 이상 궤적 탐지

스탬프투어 서비스는 GPS 기반으로 관광지에 방문하면 스탬프가 찍히고, 스탬프의 개수에 따라 소정의 상품을 지급하는 비대면 관광 특화 서비스이다[4]. 서비스를 제공하는 회사는 특정 조건에 따라 GPS 포인트의 이상 여부를 체크하여 각 궤적에서 발생하는 이상 포인트의 빈도를 통해 이상 궤적을 식별한다. 이 방법은 수작업을 통한 방법이므로 시스템의 효율화를 위해 이상 궤적 탐지를 자동화할 방법이 필요하다.

본 연구에서는 그라디언트 부스팅 모델을 사용하여 위치기반 서비스의 이상 궤적 탐지 방법을 제안한다. 그라디언트 부스팅 모델은 다양한 머신러닝을 이용한 대회에서 많은 우승을 차지한 모델로 특성의 범주가 다양한 GPS 궤적 데이터에서도 잘 동작한다[7].

그라디언트 부스팅 모델은 다수의 결정트리(Decision Tree)가 순차적으로 구성된 모델로 각 트리의 오차를 다음 트리가 보완하는 방식으로 구축된다[7]. 손실함수를 통해 각 트리의 오차를 수치화하고 기울기(gradient)를 통해 오차를 줄여가는 방식으로 학습되며 식은 (1)과 같이 표현할 수 있다.

$$D(x) = d_{tree1}(x) + d_{tree2}(x) + d_{tree3}(x) + \dots \quad (1)$$

그라디언트 부스팅 모델에서는 Least Squares, Least Absolute Deviation 등의 다양한 손실함수를 적용할 수 있다. 이상 탐지는 이진 분류 작업이므로 Negative Binomial Log-Likelihood를 사용하여 (2)와 같이 손실함수를 계산한다[8].

$$L(y, F) = \log(1 + \exp(-2yF)), \quad y \in \{-1, 1\},$$
$$\text{where } F(x) = \frac{1}{2} \log \left[ \frac{\Pr(y = 1|x)}{\Pr(y = -1|x)} \right] \quad (2)$$

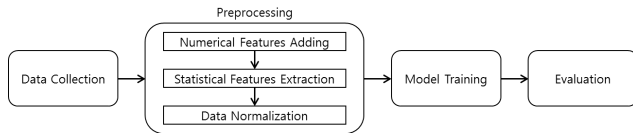
(2)에서 구한 손실함수 값은 트리의 오류로 이를 최소화하기 위해 Negative Gradient를 구한다. Negative Gradient의 식은 (3)과 같으며 이렇게 계산된 값을 통해 그라디언트 부스팅 모델  $D(x)$ 를 훈련한다.

$$\tilde{y}_i = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} = \frac{2y_i}{1 + \exp(2y_i F(x_i))} \quad (3)$$

스탬프투어 서비스의 부정 사용자에 대한 이상 궤적을 탐지하기 위해 본 연구에서 제안한 전체적인 알고리즘은 그림 1과 같다.

<표 1> 실험결과

| Model               | Anomaly      |              |              | Normal       |              |              |
|---------------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                     | Precision    | Recall       | F1-Score     | Precision    | Recall       | F1-Score     |
| Gradient Boosting   | 0.913        | <b>0.900</b> | <b>0.904</b> | <b>0.970</b> | 0.973        | <b>0.971</b> |
| Random Forest       | <b>0.933</b> | 0.843        | 0.879        | 0.955        | <b>0.982</b> | 0.968        |
| Decision Tree       | 0.838        | 0.871        | 0.851        | 0.961        | 0.947        | 0.953        |
| MLP                 | 0.908        | 0.857        | 0.877        | 0.958        | 0.973        | 0.965        |
| SVM                 | 0.873        | 0.886        | 0.876        | 0.965        | 0.960        | 0.962        |
| Logistic Regression | 0.790        | 0.886        | 0.832        | 0.964        | 0.924        | 0.943        |



<그림 1> GPS 기반 이상 궤적 탐지 알고리즘

### III. 실험

#### 1. 데이터셋

실험에 사용한 데이터는 스탬프투어 모바일 앱[4]에서 수집한 데이터로 971,612개의 GPS 포인트와 295개의 고유한 궤적 ID를 포함한다. 레이블의 경우 기업에서 선별한 블랙리스트 유저 70명과, 관광지 방문수 100회 이상인 유저 225명이 각각 이상(Anomaly) 클래스와 정상(Normal) 클래스로 이루어져 있다. 본 실험에서는 GPS 데이터에 공통으로 포함된 위도, 경도, 타임스탬프 속성을 사용하여 새로운 특성인 포인트 사이의 거리, 시간차, 평균속도, 방향 등의 속성을 추가한다. 모델의 입력으로 사용하기 위해 궤적 ID 별로 GPS 포인트의 수와 각 특성들의 최솟값, 최댓값, 평균, 표준편차 등의 통계적인 특성을 계산한 후 Z-Score 정규화를 수행한다.

#### 2. 평가지표

평가지표로는 분류 모델에서 널리 사용되는 정밀도, 재현율, F1-Score를 사용한다. 각 지표를 계산하기 위해 각 샘플의 실제 분류에 대한 긍정(Positive)과 부정(Negative)의 예측 결과가 있는 혼동행렬이 만들어지며 그림 2와 같이 나타낼 수 있다. 정밀도는 모델이 긍정으로 분류한 샘플 중 실제 긍정인 샘플의 비율로 식은 (4)와 같다. 재현율은 실제 긍정인 샘플 중 모델이 긍정으로 분류한 샘플의 비율로 식은 (5)와 같다. F1-Score는 정밀도와 재현율의 조화평균으로 식은 (6)과 같다. 각 지표는 5-Fold 교차 검증을 통해 이상 클래스와 정상 클래스에 대해 각각 계산한다.

|       |          | 실제 정답              |                    |
|-------|----------|--------------------|--------------------|
|       |          | Positive           | Negative           |
| 분류 결과 | Positive | TP(True Positive)  | FP(False Positive) |
|       | Negative | FN(False Negative) | TN(True Negative)  |

<그림 2> 혼동행렬

$$Precision = \frac{TP}{TP + FP} \quad (4) \quad Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1\_Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

#### 3. 실험결과

본 실험에서는 그래디언트 부스팅(Gradient Boosting) 모델과 다른 기계 학습 모델인 로지스틱회귀(Logistic Regression), SVM(Support Vector Machine), MLP(Multi-Layer Perceptron), 결정트리(Decision Tree), 랜덤포레스트(Random Forest)를 사용하여 성능을 비교한다. 표 1은 각 평가지표 별로 가장 높은 성능을 보이는 모델(볼드체)과 두 번째로 높은 성

능을 보이는 모델(밑줄)을 나타낸다. 그래디언트 부스팅 모델을 적용하였을 때 다른 모델들에 비해 F1-Score에서 높은 성능을 보인다.

### IV. 결론

본 연구에서는 GPS 궤적 데이터의 통계적인 특성을 추출한 후 그래디언트 부스팅 모델을 적용함으로써 이상 궤적을 효과적으로 식별할 수 있는 알고리즘을 제안하였다. 이를 통해 GPS 데이터를 활용하는 위치기반 서비스에서 자동화된 이상 궤적 탐지 방법을 적용할 수 있을 것이다.

본 논문에서 제안한 방법은 하이퍼 파라미터에 민감하며 특성 선택에 따라 성능의 편차가 크다는 한계를 가지고 있다. 향후 연구방향으로 추가적인 성능 향상을 위해 새로운 특성 추가 및 모델의 하이퍼 파라미터 조정 작업 등이 있다.

### ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT). (NRF-2020R1A4A1019191)

### 참고 문헌

- [1] 김훈, "7월 위치정보 산업 동향보고서 <스마트시티와 위치정보 기술 연구동향>", KISA 한국인터넷진흥원, (<https://www.kisa.or.kr/20204/form?postSeq=174>)
- [2] Farhanna Mar'i, Gusti Pangestu, "Classification of Fake GPS in GOJEK Application using Logistic Regression", SIET '21: 6th International Conference on Sustainable Information Engineering and Technology, pp.94-99, September 2021
- [3] 한재혁, 이상진, "모의 위치 서비스를 이용한 온라인 게임 악용 탐지 방안", 한국정보보호학회논문지, vol.27, no.5, pp.1069-1076, 2017
- [4] 댓츠잇 주식회사, "스탬프투어 모바일 앱", (<https://stamptour.modoo.at/>)
- [5] "Mastermind: Using Uber Engineering to Combat Fraud in Real Time", Uber Blog, (<https://eng.uber.com/mastermind/>)
- [6] Yufeng Kou, Chang-Tien Lu, S. Sirwongwattana, Yo-Ping Huang, "Survey of fraud detection techniques", IEEE International Conference on Networking, Sensing and Control, vol.2, pp.749-754, 2004
- [7] 안드레아스 뮐러, 세라 가이드, 박해선 (역), "파이썬 라이브러리를 활용한 머신러닝 (번역개정판)", 한빛미디어, pp.122-125, January 2020
- [8] Jerome H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine" The Annals of Statistics, vol.29, no.5, pp.1189-1232, October 2001