제목 : 부스팅 기법을 이용한 게임봇 탐지

김경환[1], 김현우[2]

Ⅰ. 서론

본 대회의 목적은 특정 계정의 게임 내역을 분석하여, 인간에 의해 조작 되었는지 혹은 게임 봇에 의해 조작 되었는지 판별하는 모델을 세우는 것이다. 게임 내에는 플레이 시간, 경험치 획득, 파티 활동 내역을 비롯하여 유저의 행동을 설명하는 많은 지표가 있을 수 있다. 본 팀은 제공받은 로그 데이터로 유저의 행동을 유저정보, 게임 활동내역, 사회적 활동내역 등과 같은 카테고리로 나누고 변수를 생성하였다. 생성한 변수 중에서 의미가 없는 변수들을 null importance 방법에 의해 제거하였다. 유저의 수가 8000명을 적은 점을 고려해서 인공신경망 기법보다는 트리 기반의 부스팅 기법을 사용하여 0.86의 F1 Score를 얻을 수 있었다.

Ⅱ. 방법론

[1] 변수 생성 : 데이터의 용량이 매우 큰 관계로 pyspark를 통하여 전처리 및 변수 생성을 실시하였다. 본 팀은 유저의 특성을 설명할 수 있는 요소로 다음의 3가지로 분류하였다.

(1) 유저정보 - 게임의 접속 시간, 접속 횟수. 접속한 아이피의 개수. 레벨의 최댓값.

(2) 게임 활동내역 -경험치/ 게임 머니 획득, 부활횟수, 포탈 이용횟수, captcha 관련 내용 등을 기록

(3) 사회적 활동내역 - 친구와의 혹은 파티/길드 내에서의 활동내역이나 거래내역. 사람들과의 결투내역.

1. 서강대학교 경제학과
2. 한양대학교 산업공학과

유저의 행동을 위의 기준에 따라 유저 정보, 개인 활동, 사회적 활동으로 나누어 변수를 생성하였다.

[2] 변수 선별 : 모델에 투입될 최종 변수는 [1]을 통해 생성된 변수들 중 null-importance-feature-elimination 방법을 통해 최종적으로 선별된 변수를 사용하였다. 본 방법은 특정 칼럼의 값을 임의로 shuffle 후 모델 학습을 진행할 시, 원래의 값을 통해 학습한 결과보다 성능이 좋아질 경우, 해당 칼럼은 모델에 도움이 되지 않는다고 판단하고 제거하는 식으로 작동한다.

[3] 모델링 : 학습에 사용할 수 있는 데이터의 수가 그리 많지 않은 관계로 Neural Network 모델을 사용하는 것 보다는 Tree-based 모델을 사용하는 게 적합하다고 생각하였다. 따라서 비교적 빠른 속도로 학습이 가능하면서 성능이 비교적 우수하다고 알려져 있는 boosting 모델 중, lightgbm과 catboost모델을 사용하였다. Validation 방법으로는, 학습에 사용할 다소 불균형적인 모습을 보인다고 판단하였고, 각 Fold마다 동일한 target의 분포를 배치시키기 위해 Stratified KFold(K=5)를 사용하였다.

III. 탐지결과 및 평가

[1] 실험 환경 : 실험은 개인 컴퓨터와 캐글 서버를 이용해서 진행하였다. 전처리와 변수의 생성은 메모리를 많이 잡아서 램64GB의 컴퓨터로 진행하였고, 모델링의 경우 캐글 서버를 통해서 진행하였다.

[2] 분석 방법 : 학습 시 모델의 평가 방식을 roc_auc_score로 규정하였다. roc_auc_score는 FPR가 한 단위 변화할 때, TPR가 어느정도로 변화하는 가에 대한 정보를 알려줌과 동시에, Binary Classification의 경우 서로 다른 타켓값을 얼마 잘 분리되게 모델을 학습시켰는가를 평가하기 때문이다. 학습한 결과를 바탕으로 본 대회의 평가함수인 f1_score의 값을 최대화 할 수 있는 threshold를 찾는 과정을 거쳤다.

[3] 탐지 결과 : F1_score에 대하여, 본 팀이 사용한 모델 중 catboost는 0.8614, lgbm은

0.8601 정도의 점수를 기록하였다. 최종적으로 두 가지 모델을 결합한 결과 0.864의 점수를 얻을 수 있었다.

[4] 변수 중요도 : 변수 중요도가 높은 피쳐들로는

"sum_playtime"(총 플레이 시간),

"Day_unique_count"(제공된 데이터셋 기간 중 총 몇일을 접속했는지),

"mean_playtime"(접속한 날들에 평균적으로 얼마나 플레이를 했는지),

"avg(Exp_get_day_count)"(접속한 날들에 평균적으로 얼마의 경험치를 획득하였는지),

"avg(sit_per_day_count)"(접속한 날들에 sit이 몇개 관찰되는지),

"quest_sucess_count"(퀘스트 성공 횟수),

"quest_try_count"(퀘스트 도전 횟수) 가 있었다.

위의 피쳐들은 공통적으로 많은 시간을 플레이 하면 수치가 높아지는 특성을 가지고 있다. 우리의 모델은 게임을 많이 이용하면 많이 이용할 수록 인간이 아닌 봇에 의한 플레이라고 판별을 하는 것으로 보인다.

Ⅳ. 프로그램 설명

[1] 개발환경 : Windows10, Ram 64GB

[2] 프로그램 : Python

[3] 패키지 : pandas, numpy, tqdm, sklearn, catboost, math, lightgbm, glob, matplotlib, os, gc, pyspark, findspark, functools, networkx

Ⅴ. 결론

우리는 대용량의 로그데이터에서 게임봇을 탐지하는 방법으로, pyspark를 이용한 데이터 전처리와 boosting기법을 이용한 모델링을 제안했다. 이러한 방법은 112GB의 파일을 64RAM의 컴퓨터에서 돌아가는 수준이었고, 변수의 생성까지 24시간 이하의 시간이 걸렸다. 유저와 봇의 행동을 구분하기 위해서 유저정보, 게임활동 내역, 사회활동 내역으로 3가

지 카테고리를 만들어서 구분했고, 변수의 중요도를 살펴본 결과 인게임 플레이를 얼마나 많이 했느냐가 중요하다는 것을 확인할 수 있었다. Ligbgbm과 catboost의 F1_score는 각각 0.8601과 0.8614 이었으며 둘을 앙생블한 결과 0.864의 F1_score를 얻을 수 있었다.

[참고문헌]

[1] Kang, A. R., Jeong, S. H., Mohaisen, A., & Kim, H. K. (2016). Multimodal game bot detection using user behavioral characteristics. *SpringerPlus,* 5(1), 1-19.