**XAI-Driven Explainable Multi-view Game Cheating Detection**

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9231843>

**0. INTRODUCTION**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 게임에서의 부정행위 탐지는 다음의 5가지 view를 통해 이루어진다.   |  |  | | --- | --- | | Portrait view | MLP neural network를 이용한 **time series technique를 이용하여 사람의 행동인지 bot의 행동인지 구분**한다. | | Behavior view | Supervised 및 unsupervised method를 이용하여 **사용자의 행동 순서에 기반하여 bot인지의 여부**를 알아낸다.   * 이때 사람과 bot 간의 서로 다른 행동 패턴을 이용한다. | | Graph view | **Real money trading (RMT)를 탐지하기 위하여 in-game transaction을 조사**한다.   * 이때 가상 상품 교환 및 RMT transaction을 이용하여 social graph를 구축한다. | | Image view | **Convolutional Neural Network**를 이용하여 **실시간 자동 wallhack detection system**을 개발하여 사용한다.   * Class Activation map을 이용하면 네트워크는 스크린샷에서 모델의 신뢰도를 높일 수 있는 suspicious area를 찾고, debugging dataset을 더 효율적으로 생성한다. | | Multi-view | **Multi-view attention network (MVAN)**을 이용하여 온라인 게임에서의 real money trading을 탐지한다.   * 이것은 **Portrait view, Behavior view와 Graph view data**에 대한 강력한 expressiveness를 가능하게 한다. |   부정행위 탐지를 위한 설명 가능한 AI (eXplainable AI)는 다음 그림과 같이 여러 종류의 사용자에게 유용하다. |

**1. GAME DATA**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **<온라인 게임의 로그>**  온라인 게임의 log는 다음과 같은 정보들로 구성된다.   |  |  | | --- | --- | | Timestamp | 특정한 activity 또는 event의 발생 시점 | | Log ID | Player activity의 identifier (ID) | | Source Player’s Information | Player의 role ID, 성별, 레벨, 클래스, 가상 재화 등 | | Target Player’s Information | 플레이어가 상대 플레이어와 상호작용할 때 지정됨 | | Detailed Information | 각 activity 또는 event의 type에 기반하여 결정됨 |   **<캐릭터의 portrait 구조>**  **Real-time game log로부터 universal한 게임 캐릭터 portrait을 생성**한다.   * 이것은 계정 정보, 디바이스 정보, social 정보, task 정보 등 여러 가지의 서로 다른 topic을 커버한다.   **<Behavior Sequence 구조>**  각 플레이어의 behavior sequence는 timestamp에 따라 정렬된 event list로 구성되어 있으며, 다음의 4가지 feature를 포함한다.   |  |  | | --- | --- | | Timestamp | 특정한 activity 또는 event의 발생 시점 | | Event ID | 플레이어에 의해 수행된 현재 이벤트 (특정 스킬의 사용, 아이템 획득 등) | | Interval | 지난 게임 이벤트와 현재 게임 이벤트 사이의 간격 | | Level | 플레이어의 현재 게임 레벨 |     오른쪽 그림은 **normal player와 cheating player의 behavior sequence를 시각화**한 것이다. |
| **NEXT PAGE** |

**1. GAME DATA**

|  |
| --- |
| **PREVIOUS PAGE** |
| **<Client Image 구조>**  예를 들어, Knives Out Plus 게임의 부정행위 탐지 기술에서 **게이머가 적을 처치하는 순간의 게임 클라이언트의 스크린샷을 캡쳐**한다.  **<Social Graphs 구조>**  Justice Online의 게임 로그에서 4가지의 서로 다른 종류의 graph를 구성하였는데, 이것은 **transaction graph, friendship graph, teaming graph, chatting graph**이다.   * 이것은 **multi-relational graph**를 생성한다.  |  |  | | --- | --- | | Transaction graph | 플레이어 간의 **assets exchange relation**을 보여준다.   * Edge는 플레이어 간 생성된 트랜잭션의 virtual currency를 나타낸다. | | Friendship graph | 온라인 게임에서의 **unidirectional한 friendship**을 보여준다. | | Teaming graph | **플레이어 간의 협력적인 관계**로부터 만들어지며, 팀은 임시적으로 서로 같은 목표를 가지고 행동한다. | | Chatting graph | 플레이어 간의 **communication relationship**을 나타낸다.   * 캐릭터는 다른 플레이어에게 개인적인 소통을 위해 private message를 전송할 수 있다. | |

**2. THE EMGCD**

|  |
| --- |
| 이 논문에서는 다음과 같은 **EMGCD (explainable multi-view game cheating detection)** 모델을 제안한다. |
| **NEXT PAGE** |

수능

**2. THE EMGCD**

|  |
| --- |
| **NEXT PAGE** |
| **<Game Data Construction>**  Game Data는 **game log에서 추출**되며, 다양한 방법으로 표현된다. (tabular, sequential, visual, graphical 등)  **<Cheating Classifiers and Explainers>**  분류 및 설명에 사용한 알고리즘은 **use-case, model type, data format 등에 의해 결정**된다. 여기서는 주어진 classifier에 대한 post-hoc explaining을 선택한다.   * 이때 **black-box model과 확률로부터 설명을 생성하기 위한 companion model**을 이용한다.   **<Explanation>**  Explainer는 다음 표에서 A의 예측을 B의 결과로 본다.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | A | Tabular Model | Sequential Model | Visual Model | Graphical Model | | B | **features** | **개별적인 event** | **이미지의 pixel** | **subgraph** |   한 가지 explanation은 적절하지 않으며, 따라서 다음과 같은 3가지 explanation을 이용한다.   |  |  | | --- | --- | | Individual explanation | **개별적인 예측**과 관련된 설명 | | Local explanation | **Individual explanation으로부터 local groupness를 요약**한다.   * 여기서는 2000개의 표본을 수집하여 supervised clustering으로 클러스터링했다. | | Global explanation | **전체적인 predictive model을 사용자에게 설명**하고 사용자들이 이해할 수 있게 한다. (global feature importance 등) |   **<Applications for audiences>**  블랙박스 모델은 정확한 예측을 제공하려고 하는 반면, explanation은 인간 상호작용 및 인간의 모델 검토를 돕는다. 성공적으로 적용된 application은 다음과 같다.   |  |  | | --- | --- | | Evidence and reason generation | 게임 운영팀과 고객 서비스 팀은 개별적인 설명을 읽고 **부정행위의 증거 또는 의도가 있었는지 파악**할 수 있다. | | Model debugging | 데이터 과학자들은 모델의 설명을 이용하여 **비논리적인 연관성과 label leak**을 추출할 수 있다. | | Model compression | **Feature attribution value**는 모델의 예측에 있어서 가장 중요한 feature들을 identify하는 데 사용된다. | |

**3. EXPERIMENTS**

|  |
| --- |
| 여기서는 획득한 dataset을 **서로 중복되지 않는 training set(80%)과 test set(20%)**으로 각각 나누어 평가하였다. **(5-fold cross validation 이용)**   * 테스트의 정확도는 **AUC** (area under the curve), **ACC** (accuracy), 시간은 **training time** (for all training set), **inference time** (for all test set), **explaning time** (for 1 sample on average)을 측정하였다.   **<Game Cheating Detection Performance>**  **<Portrait Explanations>** |
| **NEXT PAGE** |

**3. EXPERIMENTS**

|  |
| --- |
| **PREVIOUS PAGE** |
| **<Behavior Explanations>**        **<Image Explanations>**    **<Graph Explanations>** |

**4. APPLICATIONS**

|  |
| --- |
| **<Evidence and Reason Generation>**  설명이 가장 자연스럽고 직관적이라는 것을 검증하기 위해, **게임 운영자 및 게임 디자이너에게 user study**를 요청하였다. 이때 참여자들은 인공지능에 대한 경험이 없다.   * 설명에 대한 평가 기준은 **진실성, 유용성, 관련성, 이전의 믿음과의 coherence, 일반성**이다.     **<Model Debugging and Compression>**  부정행위 탐지 모델은 일반적인 그림을 Figure 12b와 같이 인식한다.   * 여기서 image explainer의 설명을 분석하여, **플레이어의 red crosshair가 부정행위로 오인될 수 있는 plug-in element와 유사하다**는 것을 발견하였다.   Figure 13은 **같은 classifier를 이용한 예측을 비교**한 것을 나타낸다.   * Figure 13a에서 볼 수 있듯이, portrait explainer가 선택한 **portrait feature가 있는 예측**은 random sample에서 선택된 것보다 좋다. * Figure 13b는 **behavior view에서의 서로 비슷한 결론**을 보여준다. 여기서 **Transformer Based Explanation이 가장 좋은 성능**을 보여준다. |