Detecting Anomalous Agent Decision Sequences Based on Offline Imitation Learning

Extended Abstract

Chen Wang
The University of Melbourne
Melbourne, Australia
chenwang4@unimelb.edu.au

Sarah Erfani
The University of Melbourne
Melbourne, Australia
sarah.erfani@unimelb.edu.au

Tansu Alpcan
The University of Melbourne
Melbourne, Australia
tansualpcan@unimelb.edu.au

Christopher Leckie
The University of Melbourne
Melbourne, Australia
caleckie@unimelb.edu.au

ABSTRACT

Anomaly detection in decision-making sequences is a challenging problem due to the complexity of normality representation learning, the sequential nature of the task and the difficulty of real-world implementation. In this work, we propose extracting two behaviour features: *action optimality* and *sequential association* to detect anomalous behaviour. Our offline imitation learning model is an adaptation of behavioural cloning with a transformer policy network, where we modify the training process to learn a Q function and a state value function from normal trajectories.

KEYWORDS

Anomaly Detection; Offline Imitation Learning; Sequential Decision-making

ACM Reference Format:
Chen Wang, Sarah Erfani, Tansu Alpcan, and Christopher Leckie. 2024. Detecting Anomalous Agent Decision Sequences Based on Offline Imitation Learning: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Anomaly detection for decision-making sequences generated by goal-oriented agents has not been widely studied. A goal-oriented agent is capable of making decisions based on its environment and desired goal. An anomalous decision-making sequence may indicate a different goal compared to other agents. Examples of this include malicious taxi drivers who take detours or go through congested streets to increase their fare [9, 12], or faulty robots that perform unexpected actions resulting in safety issues [4].

There are three significant challenges in this area. (1) *Normality representation learning*: The success of anomaly detection algorithms largely depends on a suitable data representation that can separate anomalous data from normal data. (2) *Sequential nature of the decisions*: Most anomaly detection methods fail to consider the sequential nature of the decision-making process [10, 16]. To predict if one decision is normal or not, a good algorithm needs to take past observations and actions into consideration [8]. (3) *Real-world implementation*: Most existing methods [4, 10, 16] based on Reinforcement Learning (RL) or Inverse RL [9, 11, 14, 17] are difficult to implement in the real world due to unrealistic assumptions, such as having access to environment dynamics, reward signals, and online interactions with the environment.

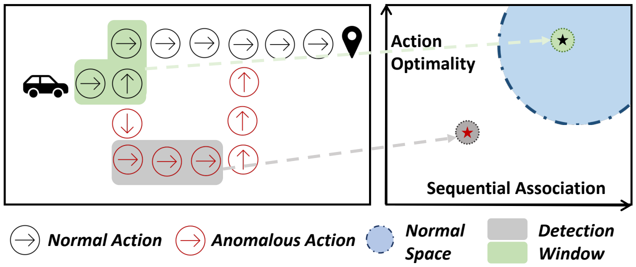

Figure 1: A demonstration of our anomaly detection method. Each decision sequence in the detection window is transformed into a novel two-dimensional feature space: action optimality and sequential association.

In this paper, we propose an **online** anomaly detection framework (i.e., online detection using a sliding window technique) targeted at decision-making sequences in a realistic **offline** imitation learning setting, i.e., without access to the reward function, environment dynamics or online interactions with the environment, and solely relying on previously recorded sequences. This paper provides a preliminary view of the problem, please see [13] for a more extensive study. The source code is available on https://github.com/chenwang4/OILAD .

2 BACKGROUND

Markov Decision Process (MDP) provides a mathematical framework for RL. MDP can be defined as a tuple $(S, A, \mathcal{P}, R, \gamma, b_0)$, where S is a state space, A is an action space, $\mathcal{P}^a_{ss'}$ is the transition probability, R^a_s is the reward, γ is the discount factor and $b_0(s)$ is the

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

probability of starting at state s. The goal of MDP is to find an optimal policy π^* to maximize the expected value of discounted future rewards. A policy π is a distribution over actions given a state. Each policy is related to one state-value function $v_\pi(s)$ and one action-value function $Q_\pi(s,a)$.

3 PROBLEM STATEMENT

We consider RL agents in an MDP environment, defined by a tuple $(S, A, \mathcal{P}, R, \gamma, b_0)$. Goal-oriented agents take actions to maximize their sum of future rewards $\sum_t^{t\to\infty} \gamma^t r(s_t, a_t)$. Each decision-making sequence τ with a variable length T is composed of state-action pairs $\{(s_1,a_1),\ldots,(s_T,a_T)\}$. In our anomaly detection setting, let C_{normal} and $C_{anomaly}$ denote the sequence sets for normal and anomalous agents respectively. We consider that anomalous behaviours are performed because of agents' anomalous intention (reward function) $R_{anomaly}$ or unexpected transition $\mathcal{P}_{anomaly}$. In this work, we do not know R or $\mathcal{P}$ from both C_{normal} and $C_{anomaly}$, and we can only observe decision-making sequences of agents from C_{normal}. The goal is to extract an appropriate feature representation $\phi(\tau)$ from sequences, and find a suitable detection model to predict the anomalous level p such that

$$p(\phi(\tau)) < p(\phi(\tau')) \quad \forall \tau \in C_{normal}, \forall \tau' \in C_{anomaly} \tag{1}$$

4 METHODOLOGY

We use a variant of traditional behavioural cloning that adds a transformer block in the policy network to maintain the sequential nature of decision-making. This model can be used in continuous state space and discrete action space. In order to construct an expected Q function, we consider the values before the softmax layer as Q values $\{q(s,a)\}_{a\in A}$ for one state s. By computing the dot product between $q(s,a)$ and $\pi(a|s)$ as Eq. 3, we can derive the state values $v(s)$. We propose two objectives: (1) the Q function can select the optimal/normal action in a given state, as described in Eq.2, and (2) the state values for optimal/normal trajectories tend to be monotonically increasing [12], as described in Eq.4.

$$q(s,a^*) > q(s,a), \quad \forall a \neq a^*, \forall s \in S \tag{2}$$

where a^* is the optimal action.

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a) \tag{3}$$

$$v_\pi(s_1) \le v_\pi(s_2) \le \cdots \le v_\pi(s_T), \quad \{s_i\}_{i=1}^T \in \tau^* \tag{4}$$

where τ^* is a sequence that follows the optimal policy. We use two training objectives to satisfy Eq.2 and Eq.4 respectively. The first training objective, named action loss, is formulated as:

$$\min_{\boldsymbol{\theta}} H(\pi^*(a|s)|\pi(a|s,\boldsymbol{\theta})) - \alpha H(\pi(s,a,\boldsymbol{\theta})) \tag{5}$$

where H represents the entropy function. The first term is the cross entropy of the predicted actions $\pi(a|s)$ and true actions $\pi^*(a|s)$. The second term is the self-entropy of the predicted actions $\pi(a|s)$, and $\alpha > 0$ here is a hyper-parameter to decide the level of regularisation. We add the second term to prevent overfitting. The second objective, named monotonicity loss, is to maximize Spearman's rank correlation coefficient between $\boldsymbol{v}_\tau$ and $\boldsymbol{t}_\tau$:

$$\max_{\boldsymbol{\theta}} \frac{cov(rank(\boldsymbol{v}_\tau), rank(\boldsymbol{t}_\tau))}{\sigma_{R(\boldsymbol{v}_\tau)}\sigma_{R(\boldsymbol{t}_\tau)}} \tag{6}$$

where $rank(\boldsymbol{v}_\tau)$ and $rank(\boldsymbol{t}_\tau)$ are the ranks of $\boldsymbol{v}_\tau$ and $\boldsymbol{t}_\tau$ respectively, cov is the covariance of the ranks, and $\sigma_{R(\boldsymbol{v}_\tau)}$ and $\sigma_{R(\boldsymbol{t}_\tau)}$ are the standard deviations of the ranks.

We first train the policy neural network based on Eq.5 to learn normal agents' policy as a good foundation for the next stage. This step is similar to the regular training in Behavioural Cloning. To recover the temporal relationship, we then train the model based on Eq.5 and Eq.6 simultaneously. After training, we can extract two features to represent the behavioural normality based on the learned Q function and the state value function. The feature *action optimality* characterises whether the agent is selecting the normal/optimal action at each state and the feature *sequential association* characterises whether the agent is making the right sequential decisions given the context of desired normal goals.

At the detection stage, we first generate the boundary from the normal features and then we can transform any new windowed trajectory to the latent space, as shown in Figure 1. If the behaviour features in the latent space exceed the pre-trained boundary, the trajectory is identified as an anomaly; or otherwise, as normal.

5 EXPERIMENTS

We apply our method and other baselines to three different datasets including two real-life datasets [1, 2] and one simulated dataset from the Gym environment [3]. To provide a comprehensive evaluation, we generate two types of anomalies, policy anomalies and perturbed anomalies, for each dataset. Our experimental results show that the anomaly detection performance of our method achieves an average improvement of around 14.15% in F_1 score compared to existing state-of-the-art methods including [5–7, 15].

6 DISCUSSION

Effectiveness of Behavioural Cloning. Behavioural cloning has massive advantages in terms of simplicity and efficiency, but it also suffers from the difficulty of recovering from distribution shifts. However, the robustness to distribution shifts is not part of the performance evaluation when we switch our perspective from offline imitation learning to the problem of anomaly detection. Furthermore, the compounding error can even potentially help to detect abnormal patterns (out-of-distribution) from training/normal data.

Monotonicity. The reason why we use monotonicity loss to train the model is that we try to recover the true state value function. However, in the offline imitation learning setting, it is not feasible to recover the true state value function without knowing the transition probability or having access to the environment. Our prior work proposed the monotonicity property of state values based on the Bellman equation: that state values of trajectories performed by the optimal policy should tend to be monotonically increasing [12]. We believe this property can be beneficial to approximate the true state value function.

ACKNOWLEDGMENTS

This work was supported in part by the Australian Research Council Linkage Project under the grant LP190101287, and by Northrop Grumman Mission Systems' University Research Program. Sarah Erfani is in part supported by Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) DE220100680.

REFERENCES

[1] 2016. DiDi GAIA Open Dataset. https://outreach.didichuxing.com/en/.

[2] 2020. Australian Maritime Safety Authority Digital Data. https://www.operations.amsa.gov.au/Spatial/DataServices/DigitalData.

[3] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. arXiv:arXiv:1606.01540

[4] Tom Haider, Karsten Roscher, Felippe Schmoeller da Roza, and Stephan Günnemann. 2023. Out-of-Distribution Detection for Reinforcement Learning Agents with Probabilistic Dynamics Models. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 851–859.

[5] Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George Chen. 2022. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[6] Yiding Liu, Kaiqi Zhao, Gao Cong, and Zhifeng Bao. 2020. Online anomalous trajectory detection with deep generative sequence modeling. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 949–960.

[7] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. 2016. LSTM-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148* (2016).

[8] Robert Müller, Steffen Illium, Thomy Phan, Tom Haider, and Claudia Linnhoff-Popien. 2022. Towards anomaly detection in reinforcement learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 1799–1803.

[9] Min-hwan Oh and Garud Iyengar. 2019. Sequential anomaly detection using inverse reinforcement learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1480–1490.

[10] Andreas Sedlmeier, Robert Müller, Steffen Illium, and Claudia Linnhoff-Popien. 2020. Policy entropy for out-of-distribution classification. In *Artificial Neural Networks and Machine Learning–ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part II 29*. Springer, 420–431.

[11] Mahan Tabatabaie, Suining He, and Xi Yang. 2021. Reinforced Feature Extraction and Multi-Resolution Learning for Driver Mobility Fingerprint Identification. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*. 69–80.

[12] Chen Wang, Sarah Erfani, Tansu Alpcan, and Christopher Leckie. 2023. Online Trajectory Anomaly Detection Based on Intention Orientation. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[13] Chen Wang, Sarah Erfani, Tansu Alpcan, and Christopher Leckie. 2024. OIL-AD: An Anomaly Detection Framework for Sequential Decision Sequences. arXiv:2402.04567 [cs.LG]

[14] DaeYoung Yoon and Simon S Woo. 2020. Who is Delivering My Food? Detecting Food Delivery Abusers using Variational Reward Inference Networks. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2917–2924.

[15] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. 2022. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 8980–8987.

[16] Hongming Zhang, Ke Sun, Bo Xu, Linglong Kong, and Martin Müller. 2021. A Simple Unified Framework for Anomaly Detection in Deep Reinforcement Learning. *arXiv preprint arXiv:2109.09889* (2021).

[17] Shixiang Zhu, Henry Shaowu Yuchi, Minghe Zhang, and Yao Xie. 2023. Sequential adversarial anomaly detection for one-class event data. *INFORMS Journal on Data Science* (2023).