Identifying Precursors to Anomalies Using Inverse Reinforcement Learning*

Vijay Manikandan Janakiraman† Santanu Das‡ Bryan Matthews§ Nikunj Oza¶

Abstract

In this paper, we consider the problem of discovering candidate precursors to anomalies in a set of time sequenced data. Typical scenarios involving time sequential data include dynamical systems and general monitoring systems. In such scenarios, a precursor could be any event that frequently precedes a given event of interest. Anomalies are rare but significant events in time series data and identifying precursors to anomalies is vital in proactive management. In this work, an inverse reinforcement learning (IRL) based method is formulated to succinctly represent the nominal behavior and identify sequences that preceded the anomalous events. A preliminary evaluation is performed on flight recorded data identifying challenges and future directions for application.

1 Introduction

In many applications including finance, study of natural calamities and extreme weather, network security [1] etc., finding precursors to an event of interest (a phenomenon) is a task of high importance. The knowledge about precursors to these phenomena can be vital to proactive management of risk. If precursor events could be identied, appropriate alarming mechanisms can be designed to either prevent or at least minimize the deleterious consequences of the phenomenon. Anomalous events are rare but significant events which in many cases, lead to an abnormal behavior or a risky situation. In such cases, it is important to analyze and identify precursors that lead to anomalies for proactive risk management. This paper considers anomalies in time sequenced data and attempts to discover candidate precursors to the anomalous events.

2 Discovering Precursors to Anomalies

In this section, an algorithm using inverse reinforcement learning is proposed to identify candidate precursors to anomalies in time series data. The section proceeds by introducing some background in inverse reinforcement learning, using its solution to perform value function estimation and using the optimal value function, discover precursors.

2.1 Inverse Reinforcement Learning The goal of inverse reinforcement learning (IRL) is to determine the underlying reward function using observed behavior of the agent making decisions in a Markov Decision Process (MDP). A finite MDP is a tuple ($\mathcal{S}$, $\mathcal{A}$, $P_{s,a}$, γ and $R(s)$) where $\mathcal{S}$ is a state space with n states, $\mathcal{A}$ is an action space with k actions, $\{P_{s,a}\}$ are the state transition probabilities corresponding to an action a at state s, $\gamma \in [0, 1)$ is the discount factor, $R(s) \in \mathbb{R}$ is the underlying reward function. A policy π can be defined as any map $\pi : \mathcal{S} \mapsto \mathcal{A}$ and the corresponding value function at any state s_1 can be given by

$$V^{\pi}(s_1) = E[R(s_1) + \gamma R(s_2) + \gamma^2 R(s_3) + ...|\pi] \tag{2.1}$$

where the expectation is over the distribution of state sequences $(s_1, s_2, s_3, ...)$ following the policy π starting from s_1.

Given the setting above, the goal of standard reinforcement learning is to determine a policy π^* that maximizes $V^{\pi}(s)$ among all policies for all $s \in \mathcal{S}$. When the agent's reward function is known, this task can be achieved using existing techniques for value function estimation [2]. However, in several situations, the agent's behavior is not completely known, i.e., the reward function cannot be defined easily. In such situations, the expert's observed behavior can be used to either reconstruct the underlying reward function as in the case of inverse reinforcement learning [3] or construct optimal policies directly as in the case of apprenticeship learning [4].

Assuming availability of sampled trajectories (relevant to the problem involving time series in this paper), the IRL problem can be posed as in [3]. The sampled trajectories can be considered as demonstrations of both the expert and non-expert acting in the MDP. Using the trajectories, the value functions of the expert and non-expert policies can be determined as follows. Let the unknown reward function be parameterized as

$$R(s) = \alpha_1\phi_1(s) + \alpha_2\phi_2(s) + .. + \alpha_d\phi_d(s) \tag{2.2}$$

*Supported by the NASA System-wide Safety and Assurance Technologies (SSAT) Project.

†UARC, Nasa Ames Research Center, Moffett Field, CA
‡Verizon, Palo Alto, CA
§SGT Inc., Nasa Ames Research Center, Moffett Field, CA
¶Nasa Ames Research Center, Moffett Field, CA

where the ϕ_i represent the features of the reward function. The expert value function following policy π_E at state s_1 can be given by

$$\begin{aligned}(2.3)\quad V^{\pi_E}(s_1) &= E[R(s_1)+\gamma R(s_2)+...|\pi] \\ &= E[\alpha_1\phi_1(s_1)+\alpha_2\phi_2(s_1)+..+\alpha_d\phi_d(s_1) \\ &+\gamma\alpha_1\phi_1(s_2)+\gamma\alpha_2\phi_2(s_2)+..+\gamma\alpha_d\phi_d(s_2)+..|\pi_E] \\ &= E[\alpha_1(\phi_1(s_1)+\gamma\phi_1(s_2)+..)+\alpha_2(\phi_2(s_1)+\gamma\phi_2(s_2)+..) \\ &+...+\alpha_d(\phi_d(s_1)+\gamma\phi_d(s_2)+..)|\pi_E] \\ &= \alpha_1 E[(\phi_1(s_1)+\gamma\phi_1(s_2)+..)|\pi_E]+ \\ &+\alpha_2 E[(\phi_2(s_1)+\gamma\phi_2(s_2)+..)|\pi_E] \\ &+...+\alpha_d E[(\phi_d(s_1)+\gamma\phi_d(s_2)+..)|\pi_E] \\ &= \alpha_1\lambda_1+\alpha_2\lambda_2+..+\alpha_d\lambda_d\end{aligned}$$

where λ_i represent the feature expectations, i.e., the value function if the reward function is composed of $\phi_i(s)$ only. After calculating the feature expectations knowing the state sequences, the value function can be defined as a function of the unknown $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, .., \alpha_d]^T$ as follows.

$$(2.4)\qquad V^{\pi_E}(\boldsymbol{\alpha}) = \sum_{i=1}^{d}\alpha_i\lambda_i$$

Similarly, by knowing the sequence of states (trajectories) for J sub-expert/non-optimal policies, the $V^{\pi_j}(\boldsymbol{\alpha})$ can be calculated. The objective of IRL is to determine the coefficients α_i so that $V^{\pi_E}(\boldsymbol{\alpha}) \geq V^{\pi_j}(\boldsymbol{\alpha})$ for $j = 1, 2, ..J$. A linear programming problem can be solved for α_i as follows

$$(2.5)\qquad \min_{\alpha}\sum_{j=1}^{J}\zeta_j$$

$$(2.6)\qquad \text{subject to}\begin{cases} V^{\pi_j}(\boldsymbol{\alpha}) - V^{\pi_E}(\boldsymbol{\alpha}) - \zeta_j \leq 0 \\ \zeta_j \geq 0, j = 1, 2, .., J \\ |\alpha_i| \leq 1, i = 1, 1, .., d \end{cases}$$

2.2 Value Function Estimation The IRL problem gives an optimal $\boldsymbol{\alpha}$ which gives a model of the underlying expert's reward function. The reward function can then be used to determine the expert's value function using a regular reinforcement learning algorithm. Any of the methods described in [2] such as dynamic programming, monte carlo or temporal difference depending on availability of the system model, ability to sample etc. In this paper, considering sample time series from a policy as monte carlo samples, the value function is approximated as follows. For each policy π_j including the expert policy π_E, a sample trajectory is used to identify the state sequences and using the reward function obtained above, the values of every state in $\mathcal{S}$ is updated. This is repeated for several trajectories from the selected policy and the average returns are stored as state values.

2.3 Precursor identification The value function of the expert policy π_E obtained above can be used to compare a non-expert behavior to identify a possible precursor sequence. V^{π_E} can be thought of as the expert's value function and any action that is greedy with respect to the expert's value function gives the optimal policy π_* [2]. Let the greedy action at s be $a^*(s)$ and the corresponding value be $V^{\pi_*}(s)$. In our problem involving time series, the time sequence and the physics of the problem can be used to restrict the state space for searching optimal actions in some cases. A given test time series can be analyzed as follows. Using the state sequences of the test data and the obtained reward function, the state values $V^{\pi_{test}}$ can be estimated. By comparing the $V^{\pi_{test}}$ with V^{π_*}, we can indirectly evaluate the actions taken by the agent in the test trajectory. Let

$$(2.7)\qquad \Delta V = V^{\pi_{test}}(s) - V^{\pi_*}(s)$$

and if $\Delta V \leq 0$, then it would mean that a sub-optimal action has been taken by the agent executing the test policy and by comparing over the state sequence, we can identify a sequence of bad actions by the agent. As defined earlier, an optimal action is one that corresponds to a nominal time series while a non-optimal action would correspond to an anomalous sequence as defined in the IRL problem. It should however be noted that the test policy is evaluated just based on one time series and hence not an expectation. However, the goal is to identify the level of sub-optimality in the state sequences specifically executed by the test trajectory to identify the precursor and not for the policy in general. This assumption needs to be analyzed more in detail and will be considered in the future. Further, if the action space is well defined, instead of comparing the value functions as above, the actions of the test agent can be directly compared against the optimal actions of the expert and precursors can be identified by noting their difference.

3 Application to Flight Anomalies

In this section, the IRL based precursor discovery algorithm is evaluated on flight time series data sets ob-

tained from a FOQA (Flight Operations Quality Assurance) archive. Typical FOQA parameters consist of both continuous and discrete (categorical) data from the avionics, propulsion system, control surfaces, landing gear, the cockpit switch positions, and other critical systems. Each flight record can have up to 500 parameters in the form of time sequences and are sampled at 1 Hz.

Flight anomalies are of significant interest within the NASA System-wide Safety and Assurance Technologies (SSAT) project to assess the health of large commercial fleets of aircraft. In this paper, flights that violated exceedance thresholds on computed air-speed are considered as operational anomalies. A specific exceedance defined as computed air-speed above a certain threshold (in knots) at an altitude of 1000 feet is considered an operationally significant high-energy approach. The goal of this study is to discover precursors to such high-energy approach flights [5] for use in proactive flight management. The data set consists of about 20000 nominal flights (flights that did not violate the exceedance and considered optimal with respect to the exceedance) and about 250 anomalous flights.

3.1 Discovery of candidate precursor sequences

The FOQA raw data consists of more than 400 parameters recorded as time sequences during the flight. However, to overcome the curse of dimensionality in solving the Markov decision process in the IRL, the FOQA data is abstracted to represent the various events happening in a flight using a high level parameter such as the aircraft energy. With the given definition of an anomaly, the flight data is considered as a sample from an expert policy (π_E) if it doesn't flag the exceedance or a sample from a non-expert policy (π_j) if it flags the exceedance. A reward function $R(s)$ can be defined as a linear combination of several Gaussian functions defined with respect to the states s. It has to be noted that the state definition is given by $s = [E \quad D]^T$ where E represents the kinetic energy of the aircraft while D represents the distance in nautical miles to touchdown. The reward function $R(s)$ can be represented as in equation (2.2) where ϕ_i could represent Gaussian functions with mean μ_i and spread σ_i and d represents the total number of Gaussian functions in the state space. Using the reward function with unknown coefficients α_i the value function of each trajectory is calculated and the IRL problem is solved as in section 2.1. The optimal value of $\boldsymbol{\alpha}$ gives a model of the underlying reward function that when used to solve the associated MDP, results in maximum possibility of avoiding the given exceedance. The model hyper-parameters including d, μ_i, σ_i are determined based on cross-validating the learned model on a hold-out data set. Following section 2.1, the ΔV for a given test flight is calculated. A negative value for ΔV indicates that the given test flight performs inferior to the optimal policy π_* and a negative rate of ΔV indicates a sequential inferior behavior. These two features are used in defining precursor candidates for the given test flight. It should be noted that the problem in hand uses FOQA data that only records the state of the flight and no explicit information about the intentions/actions of the agent (a pilot) is available and hence we were restricted to comparing the value functions as mentioned in section 2.3

Using the identified precursor sequences of a given test flight in terms of the states s, the FOQA historical data can be used to identify the flight parameters that are abnormal. The identified precursor sequence points to a section of the flight prior to the adverse event where interesting precursor events can be discovered. By modeling a nominal distribution of the FOQA parameters, any abnormality can be detected by comparison against the nominal. The identified abnormal parameters may contain information about possible factors that lead to the adverse event. This is algorithmically analyzed and validated by a domain expert.

4 Results and Discussion

In this section, a high-energy approach flight is analyzed for precursors from 35 nautical miles until touchdown. Figure 1 shows the evolution of the fight in terms of the parameters reported by the algorithm as candidate precursors. The blue shaded region represents the nominal distribution of that parameter (99 percentile of the non-exceedance flights) The green shaded regions of the figure represent the sequence of precursors (a precursor window) as identified by comparing the flight's state values to V^{π_*}.

It can be observed from Figure 1 that the computed air-speed of the flight is high compared to the nominal distribution of the non-exceedance flights indicating that the test flight is indeed an example of a high-energy approach. Further, out of the 56 chosen parameters from the FOQA list, only 11 were listed as possible precursor parameters as these parameters were out of the nominal distribution in the precursor window. The algorithm also reported ground speed which is correlated with the computed air-speed, vertical speed, stabilizer position, engine speed, flight director specified speed etc. However, a close look at the discrete parameters reported by the algorithm gives a clear picture of the actions responsible for the anomalies, i.e., the landing gear has been deployed a little earlier compared to nominal flights and the flaps were deployed very late causing the aircraft to slow down late leading

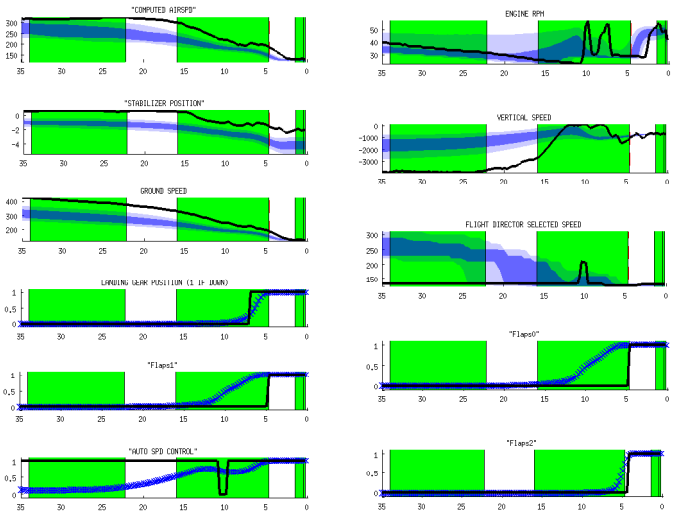

Figure 1: Figure showing the test flight trajectory (black curve) along with the precursor window (green region) as identified by the IRL algorithm. The nominal distribution of the continuous parameters such as computed air-speed, engine RPM, stabilizer position, vertical speed and ground speed are shown in blue - light blue region represents 0 - 99 percentile while dark blue region represents 25 - 75 percentiles. The nominal distribution of discrete variables including landing gear, flaps, auto speed control are shown by blue curve with markers indicating the probability of the variable having a value of 1.

to the exceedance (In the discrete plots, the marked blue curve represents the probability that a nominal discrete event takes a value of 1). Both these factors were validated by domain experts as probable precursors to the high speed exceedance. The initial high computed air-speed followed by a lack of optimal action (which is to deploy landing gear and flaps on time) in this case can be concluded as a valid precursor for flights violating the high-speed exceedance at 1000 feet altitude.

5 Conclusions

In this paper, a novel method to discover precursors to anomalies has been formulated using inverse reinforcement learning. It is argued that a value function of a non-expert, if compared against the optimal value function of an expert, can be used to identify instances of a "bad" or sub-optimal actions/situations in time series data. A high dimensional FOQA time series data has been abstracted and used for preliminary evaluation of the algorithm. The results indicate that the algorithm indeed finds the precursors that were validated by a domain expert. Although the analysis on a couple of flights gave us promising results, the algorithm is at infancy and requires extensive validation for which data sets with ground truth information about the precursors and anomalies are required. Also, for precursor identification, an appropriate performance metric will be identified for evaluation of this algorithm in future. Finally, some of the underlying hypotheses/assumptions of the algorithm will be tested in future.

References

[1] J. B. D. Cabrera, L. Lewis, X. Qin, W. Lee, R. Prasanth, B. Ravichandran, and R. Mehra, "Proactive detection of distributed denial of service attacks using mib traffic variables-a feasibility study," in *Integrated Network Management Proceedings, 2001 IEEE/IFIP International Symposium on*, 2001, pp. 609–622.

[2] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[3] A. Y. Ng and S. Russell, "Algorithms for inverse reinforcement learning," in *in Proc. 17th International Conf. on Machine Learning*. Morgan Kaufmann, 2000, pp. 663–670.

[4] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the Twenty-first International Conference on Machine Learning*, ser. ICML '04. New York, NY, USA: ACM, 2004, pp. 1–.

[5] S. Das, L. Li, A. Srivastava, and R. J. Hansman, "Comparison of algorithms for anomaly detection in flight recorder data of airline operations," in *12th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*. American Institute of Aeronautics and Astronautics, 2012.