A CHANGEPOINT METHOD FOR OPEN-WORLD NOVELTY DETECTION

Matthew D. McLure, David J. Musliner

Smart Information Flow Technologies (SIFT), Minneapolis, MN 55401 USA

ABSTRACT

Novelty detection in open worlds is a valuable endeavor for improving the robustness of autonomous systems. Open worlds are characterized by a lack of constraints on the types of novelties that might occur. Here we describe recent progress on detecting novelties in open-world numerical properties in our OpenMIND agent, a domain-independent planning-based AI architecture. Our approach has three elements: (1) Feature construction by crossing a tractable number of domain-dependent sensors with some basic domain-independent statistical derivations, (2) online, univariate changepoint detection on each signal, and (3) negative feature selection by screening for false positives in non-novel scenarios. We report on the effectiveness of this approach in a single-blind evaluation in which OpenMIND plays a game and encounters novelties never observed by it or its developers.

***Index Terms*—** novelty detection, changepoint detection, open world, Mann-Whitney, negative selection

1. INTRODUCTION

Modern AI has demonstrated powerful learning capabilities in controlled settings, but operating in the real world presents novel circumstances that can derail normal operation and learning, if not actively addressed. This is true of remote sensing systems, for which weather, geographical diversity, shifts in perspective, and adversarial agents might introduce radical, unfamiliar changes. Autonomous systems could benefit from an ability to detect diverse novelties in open-world settings, so their operations and learning algorithms may be prepared or modified to better adapt.

OpenMIND is a planning-based architecture that detects novelties in diverse, open-world domains, and responds by modifying its goals and planning models on the fly [1]. Here we describe our recent work on a domain-independent method for monitoring numerical data streams for changepoints that indicate domain novelty. We refer to the resulting software module as the Romulan Novelty Detector (RND), inspired by the notion of a modular, bolt-on Romulan cloaking device from Star Trek[1]. The RND takes as input a tractable set of domain-dependent, univariate sensor data streams. The RND approach has three elements:

Feature construction: Expand the set of numeric signals by applying domain-independent operators such as min/max/mean.

Changepoint detection: Hypothesize recent changepoints across all signals and apply a hybrid changepoint detection technique, from which a confidence or probability estimate can be extracted for each changepoint hypothesis. Apply a high confidence threshold for novelty detection.

Negative feature selection: Trim the set of signals by screening for false positive detections in non-novel scenarios.

2. BACKGROUND

Funded by the DARPA SAIL-ON program, this work is part of a multi-team effort to explore ways that autonomous agents can be robust to a wide variety of types of novelty, in a scientifically sound research approach based on uniform metrics and single-blind evaluations [2]. The research groups developing domain simulations that include novelty are separate from the research groups developing agents, and only very limited exchange of novelty examples is allowed. Agents must detect and respond to novel aspects that arise during their interactions with the domain simulations; their performance on both detection and response is measured over many different interactions and different novelties. Evaluations consist of a set of "trials," each of which is a series of games; at the beginning of each trial the agent restarts with its baseline knowledge (trained or hand-coded), and throughout the trial the agent can learn new information. Evaluations are conducted every six months by the domain developers, using a variety of "unrevealed" novelty examples that have not been shown to the agent developers.

2.1. OpenMIND

OpenMIND is a planning-based agent that creates and chooses goals, makes plans to achieve its goals, executes those plans,

Thanks to Larry Holder and Brian Thomas at WSU for their hard work in running blind evaluations. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0041. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA.

[1] https://en.wikipedia.org/wiki/The_Enterprise_Incident

monitors their results and other features of the open world, and when unexpected situations occur, modifies its goals and planning models on the fly. OpenMIND has contributed several domain-independent model-modification and goal-reasoning heuristics for novelty handling in open worlds, which have proven effective in past rounds of SAIL-ON evaluation [1].

Novelty detection in OpenMIND occurs at conceptually distinct stages of processing: (1) "Cognitive proprioception" of novelty when a plan cannot be created (e.g., initial domain conditions are novel and will not support the usual plan), (2) Failure of executed planning operators: novelty is detected when OpenMIND perceives a total action failure or mismatch between expected and actual operator post-conditions (for operators that are expected to be reliable). (3) Object recognition: novelty is detected when an individual object's properties are perceived to be substantially outside the learned distributions of all known classes. (4) Perception of numerical properties: novelty is detected when there is an observed shift in a numerical property of a known class, or of relationships between known classes. Our focus in this paper is the last of these aspects of novelty detection.

2.2. Vizdoom

Vizdoom is a modified form of the vision-oriented first-person shooter game [3], provided in top-down 2D symbolic-sensing form to the SAIL-ON program by Washington State University (WSU) [4]. In this domain, the player moves through a map dotted with enemies who shoot at it, health packs, ammo packs, and obstacles. The agent has limited ammo to shoot at enemies, and the game is over when all enemies are dead (a win!) or the agent dies or runs out of moves (a loss). The agent receives a feature vector of sensor information on every turn, including the locations of other objects/agents, health, etc. We were given the simulator to run locally, so we could run many non-novel games, varied by a random seed. We also were given the ability to run five sample novelties:

1. Change in number of health packs.
2. Enemies have increased speed.
3. Enemies move toward player.
4. Enemy damage increases when closer to player.
5. Enemies spread out.

3. APPROACH

Our goal was to develop a domain-independent technique for detecting the introduction of novelties, as quickly as possible, while limiting false positive detections, across any data streams that could represent a property of a class of objects (or set of classes or relationships). For a given stream of numeric values, we assigned a probability estimate to a hypothesized changepoint (i.e. introduction of novelty) by treating the values before the changepoint as one sample and those after the change point as another, and feeding these samples to the Mann-Whitney U test. To bootstrap detection in early games, we stored sensor data from 200 non-novel games and prepended this data to the pre-changepoint sample.

Just before the end of each game in a trial, the RND hypothesizes 15 change points— one at the beginning of each of the last 15 games. For each changepoint, a novelty assessment is performed on each of 22 sensors, described below. A probability threshold of 99.9% was required before the RND accepted a changepoint hypothesis and declared that novelty was detected. Additionally, the RND only accepted a changepoint hypothesis if two other necessary conditions were met: The absolute value of the difference between the mean of the post-changepoint sample and the pre-changepoint sample must be both

i) greater than twice the standard deviation of the pre-changepoint sample, and
ii) greater than 5% of the pre-changepoint mean.

Note we faced a common challenge when trying to detect novelties in open worlds: the number of sample novelties that we had to test on was small (five), while our ability to test on randomly generated non-novel games was virtually unlimited. Thus our ability to test for false positives was much greater than our ability to test for false negatives. We therefore initially cast a wide net with the signals we monitored— leveraging domain-independent modifiers like min/max/mean to have a multiplicative effect on a modest set of domain-specific sensors (quantitative domain aspects)— and then we disabled individual sensors that were observed to cause false positives over many hundreds of non-novel test games.

3.1. Sensors

The development of sensors was partially guided by the five novelties revealed by the makers of the Vizdoom domain. The list below enumerates families of domain-dependent sensors that were implemented in the Vizdoom domain.

- Item counts (ammo packs, health packs, and obstacles) and enemy counts.
- Player's health change, ammo change.
- Player's distance moved, orientation changed.
- Average health change in enemies.
- Average distance moved by enemies.
- Average change in distance between player and each enemy, and average change in distance between enemies (pairwise).
- Distance moved by items.

For each sensor, our domain-independent novelty detector optionally tracks all the values, the first value in a game, the maximum value per game, the minimum value per game, and/or the mean value per game. Given those five modifiers applied to each of the families of domain-specific quantities listed above, 65 or more individual sensor were created. After

eliminating redundant sensors or those that empirically created false positives, we were left with 22 monitored sensors in this domain.

4. EVALUATION

Here we describe two experiments that were conducted. The first was the first formal SAIL-ON single-blind evaluation of the OpenMIND agent in the Vizdoom domain, conducted by WSU. The second was a small ablation study conducted by the OpenMIND team to assess the effects of the two additional necessary conditions layered on top of the Mann-Whitney U test (conditions *i*) and *ii*) in section 3).

For both experiments, we report results for the three primary SAIL-ON novelty detection metrics. These metrics are functions of false positives – games before the novelty has been introduced and after the agent has (erroneously) detected novelty – and false negatives – games after the novelty has been introduced, in which the agent fails to report that novelty has been introduced. True positives are games after novelty has been introduced and the agent has correctly detected novelty. A *correctly detected trial* (CDT) is a trial which contains no false positive games and at least one true positive game. The three primary SAIL-ON detection metrics are:

FN_CDT The average number of false negatives in CDTs (less is better).

CDT% The percentage of trials that are CDTs (more is better).

FP% The percentage of trials that contain at least one false positive (less is better).

4.1. Blind evaluation

The SAIL-ON evaluation run by WSU consisted of multiple trials for each of four unrevealed novelties and each of 3 difficulty levels. Each trial consisted of 200 games, some non-novel followed by some novel. The nature of the novelties is unknown to the authors. The number of non-novel games is also unknown, and likely variable.

The OpenMIND agent correctly detected novelty in 84.2% of the trials (CDT%), while only detecting novelty erroneously early in 1.1% of the trials (FP%). In the trials where OpenMIND correctly detected novelty, that detection, on average, came 33.3 games after the novelty was introduced (FN_CDT). The median was 16, indicating that there were some outlier, very late detections. When separated by novelty, the medians for this metric were 3, 4, 21, and 33.5, but the means were much more similar, ranging from 29 to 36.

4.2. Ablation study

We ran 75 trials, five trials for each of the five revealed novelties described above, configured to medium difficulty, using each of three versions of our novelty detection: (1) As described, (2) without the required mean-difference of two standard deviations, and (3) without the required mean-difference of 5%. We used the same five random seeds across conditions to produce a paired comparison. In each of the two ablated conditions, we expected to see an increase in false positives, potentially in addition to a more modest decrease in false negatives.

In the case of ablating the required two-standard-deviation change in mean, CDT% plummeted from 96% to 4%, due to the expected increase in false positives. FP% increased from 0% to 96%. The FN_CDT also decreased as expected (5.11 to 1), but due to the explosion in false positives, in the ablated condition this was only able to be measured in a single correctly detected trial (the only one), where the false negatives dropped from three to one. In the case of ablating the required 5% change in mean, CDT% and FP% were unchanged, and the FN_CDT decreased marginally (5.11 to 5.07), only dropping from three false negatives to two in a single trial.

5. DISCUSSION

The results were relatively strong for the first and only blind evaluation in the Vizdoom domain, the only domain in which the RND has been deployed thus far. The high percentage of trials containing correct detections and the extremely low percentage of trials containing false positives, alongside the high number of games required on average to detect novelty, indicate that our approach has a stark yet unsurprising conservative bias toward waiting for a very strong novelty signal to avoid false positives. It was particularly promising that a small collection of domain-specific sensors inspired by the revealed novelties generalized to (eventually) detect such a high proportion of unrevealed novelties, and it suggests that some signals have unforeseen sensitivity to indirectly related novelties. For example, a deeper look at detections of revealed novelty four ("Enemy damage increases when closer to player."), which OpenMIND's RND was slow and inconsistent to detect in our own testing, showed that sometimes it was detected based on the minimum ammo-change signal. This happened when an enemy happened to spawn immediately next to the agent, and due to the novelty, was able to kill the agent before the agent was able to fire a shot, resulting in a minimum ammo change of zero as opposed to the usual -1.

It was evident from the ablation study that parametric requirement *i*), a required mean change of 2 standard deviations, played a crucial role in preventing false positives from the Mann-Whitney U test. The catastrophic explosion in false positives caused by dropping that requirement was not worth the small decrease in false negatives, for which the evidence was weak anyway. On the other hand, parametric requirement *ii*) of a 5% change in mean did not demonstrate any value in preventing false positives, and may be worth removing, contingent on further testing.

The RND's general approach - to put out a tractable number of domain-dependent univariate numerical feelers, multiply them using domain-independent operations, monitor those data streams with a statistical test that can output a probability estimate, threshold it at a conservatively high level (here, 99.9%), then to weed out the overly sensitive signals through negative selection process on non-novel data - seems an empirically promising one to novelty detection in open worlds, especially where immediate novelty detection (minimizing false negatives) is less of a priority than eventually detecting as many novelties as possible while avoiding false positives.

6. RELATED WORK

Our combination of feature construction and feature downselection is remarkably analogous to the negative selection approach to novelty detection [5, 6]. Like in our approach, the invention of negative selection was explicitly driven by the imbalance between the abundance of non-novel data and scarcity of novel examples, but their sensors ("detectors") were string-based and randomly generated. Here the analogous sensors are a matrix of signals derived from a set of domain-dependent input signals, then subjected to a statistical changepoint test. The core of our changepoint detection step is the Mann-Whitney U test, a.k.a Wilcoxon rank-sum test. Therefore a closely related approach is [7], who use this test with control charts for changepoint detection. As pointed out in [8], Mann-Whitney is insensitive to changes in a distribution that do not change the mean. Their proposed Klyushin—Petunin test may be a promising alternative for its sensitivity to variance and its fast detection properties. The Kolmogorov-Smirnov statistical test, recently used for changepoint detection in [9], is another test that is sensitive to scale. It may be worth experimenting with these tests as alternatives or as additions to our hybrid approach. Sensitivity to scale could also be achieved by simply adding variance to our set of domain-independent modifiers used for feature construction.

7. FUTURE WORK

The highest-priority future work is to apply the RND to OpenMIND's other SAIL-ON domains, an effort that is underway.

It seems that our use of the Mann-Whitney U test is inherently limited at fast detection, even given extreme values. We plan to experiment with a "bypass" condition — an alternative sufficient condition under which to declare novelty — to trigger fast detection in extreme cases and reduce false negatives. We also plan to go beyond detecting changes in distribution mean by incorporating detection for changes in variance, as discussed in the previous section.

In addition to the existing trailing window of changepoint hypotheses, we plan to experiment with monitoring the overall leading (highest probability) hypothesis, in order to continue to evaluate it in light of the latest data. We also plan to fully automate the negative selection training process.

8. REFERENCES

[1] David J. Musliner, Michael J. S. Pelican, Matthew McLure, Steven Johnston, Richard G. Freedman, and Corey Knutson, "OpenMIND: Planning and adapting in domains with novelty," in *Proc. Ninth Conf. on Advances in Cognitive Systems*, November 2021.

[2] Pat Langley, "Open-world learning for radically autonomous agents," in *Proc. National Conf. on Artificial Intelligence*, 2020.

[3] Marek Wydmuch, Michał Kempka, and Wojciech Jaśkowski, "Vizdoom competitions: Playing doom from pixels," *IEEE Transactions on Games*, 2018.

[4] Siddharth Vodnala, "IQ test for artificial intelligence systems," `https://news.wsu.edu/2019/12/12/iq-test-artificial-intelligence-systems/`, December 2019, Retrieved September 6, 2021.

[5] Stephanie Forrest, Alan S Perelson, Lawrence Allen, and Rajesh Cherukuri, "Self-nonself discrimination in a computer," in *Proceedings of 1994 IEEE computer society symposium on research in security and privacy*. Ieee, 1994, pp. 202–212.

[6] Dipankar Dasgupta and Stephanie Forrest, "Novelty detection in time series data using ideas from immunology," in *Proceedings of the international conference on intelligent systems*. Citeseer, 1996, pp. 82–87.

[7] Douglas M Hawkins and Qiqi Deng, "A nonparametric change-point control chart," *Journal of Quality Technology*, vol. 42, no. 2, pp. 165–173, 2010.

[8] Dmitriy Klyushin and Irina Martynenko, "Nonparametric test for change-point detection in data stream," in *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*. IEEE, 2020, pp. 281–286.

[9] Oscar Hernan Madrid Padilla, Yi Yu, Daren Wang, and Alessandro Rinaldo, "Optimal nonparametric change point detection and localization," *arXiv preprint arXiv:1905.10019*, 2019.