Real-World Evaluation and Deployment of Adversary Attack Prediction Models

(Doctoral Consortium)

Benjamin Ford
University of Southern California, Los Angeles, CA, 90089, benjamif@usc.edu

CCS Concepts

•**Computing methodologies** → **Multi-agent systems;**

Keywords

Innovative Applications; Human Behavior Modeling; Wildlife Conservation; Deployed Applications

1. INTRODUCTION

Wildlife crime continues to be a global crisis as more animal species are hunted toward extinction [10, 8]. Species extinction has dire consequences on ecosystems and the local and national economies that depend on them (e.g., eco-tourism, ecosystem services). To combat this trend, wildlife conservation organizations send well-trained rangers to patrol in protected conservation areas to deter and capture poachers and also to confiscate any tools used for illegal activities that they find. At many sites, rangers collect observation data on animals, poachers, and signs of illegal activity. Given the magnitude of wildlife poaching and the difficulty of the patrol planning problem, patrol managers can benefit from tools that analyze data and generate forecasts of poacher attacks - a key focus of this work. In working with real-world wildlife crime data, we illustrate the importance of research driven by data from the field and real-world trials. This work potentially introduces a paradigm shift in showing how adversary modeling ought to be done for deployed security games [9, 2], particularly in domains such as green security games [3, 5], where data is sparse compared to settings such as urban crime [12]. Security games have received significant attention at AAMAS [4, 6], and past work in security games has often focused on behavioral models that are learned from and tested in human subject experiments in the laboratory, which provides a large amount of attacker choice data over a small number of targets [11]. The Quantal Response model is one example that models boundedly rational attackers' choices as a probability distribution via a Logit function [11]. However, the wildlife crime domain introduces a set of real-world challenges (e.g., rangers collect limited, noisy data over a large number of targets with rich target features) that require behavior modeling efforts to not only focus more on real-world data and less on laboratory data, but also not rely on plentiful attack data.

Outperforming previous laboratory-developed models [11], CAPTURE [7] is a two-layered model, developed using real-world wildlife poaching data, that incorporates key insights and addresses the challenges present in wildlife crime data. CAPTURE's top layer attempts to predict the "attackability" of different targets, essentially providing predictions of poacher attacks. The bottom observation layer predicts how likely an attack that has occurred would be observed given the amount of patroller coverage (also known as effort). CAPTURE models the attackability layer as a hidden layer and uses the Expectation Maximization (EM) algorithm to learn parameters for both layers simultaneously. Moreover, CAPTURE also contains a Dynamic Bayesian Network, allowing it to model attacker behavior as being temporally dependent on past attacks. The CAPTURE model, the current state-of-the-art in the wildlife crime domain, represents a level of complexity not previously seen in behavior modeling in the security game literature.

While the focus of CAPTURE is on the observation layer's performance (i.e., "Where will patrollers observe past poaching attacks given their patrol effort?"), our focus is on forecasting where future attacks will happen and thus we are interested in the attackability layer's predictions and performance (e.g., "Where will poachers attack next?"). However, CAPTURE's attackability predictions would sometimes predict too many targets to be attacked with a high probability and would thus have poor performance, as discussed in more detail later in the paper. Given that CAPTURE embodied the latest in modeling adversary behavior in this domain, our first attempt focused on three different enhancements to CAPTURE: replacement of the observation layer with a simpler layer adapted from [1] (CAPTURE-LB), modeling attacker behavior as being dependent on the defender's historical coverage in the previous time step (CAPTURE-PCov), and finally, exponentially penalizing inaccessible areas (CAPTURE-DKHO). Unfortunately, all of these attempts ended in failure.

While poor performance is already a significant challenge, there are two additional, important shortcomings of CAPTURE and other models in this same family. First, CAPTURE's learning process takes hours to complete on a high-performance computing cluster - unacceptable for rangers in Uganda with limited computing power. Second, CAPTURE's learned model is difficult to interpret for domain experts since it makes predictions based on a linear combination of different decision factors; the values of all its parameters' feature weights (i.e., 10 weights and a free parameter for the attack layer) need to be simultaneously accounted for in a single interpretation of poacher preferences. These limitations and CAPTURE's poor performance, the most recent in a long line of behavioral game theory models, drove us to seek an alternative modeling approach.

This paper presents INTERCEPT (INTERpretable Classification Ensemble to Protect Threatened species), a new adversary behavior modeling application, and its three major contributions. (1) Given the limitations of traditional approaches in adversary behav-

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

Figure 1: Ashes and snare found by rangers directed by INTERCEPT. Photo credit: Uganda Wildlife Authority ranger.

ior modeling, INTERCEPT takes a fundamentally different modeling approach, decision trees, and delivers a surprising result: although decision trees are simpler and do not take temporal correlations into account, they perform significantly better than CAPTURE (a complex model that considers temporal relationships), its variants, and other popular machine learning models (e.g., Logistic Regression, SVMs, and AdaBoost). Furthermore, decision trees satisfy the fundamental requirement of interpretability; without an interpretable model, relevant authorities would not test INTERCEPT in the field, thus completely defeating the spirit of innovative applications research. However, decision trees do not take into account the spatial correlations present in this dataset, and we introduce a spatially aware decision tree algorithm, BoostIT, that significantly improves recall with only modest losses in precision. To further augment INTERCEPT's performance, we construct an ensemble of the best classifiers which boosts predictive performance to a factor of 3.5 over the existing CAPTURE model. (2) These surprising results raise a fundamental question about the future of complex behavioral models (e.g., Quantal Response based security game models [11, 7]) in real-world applications. To underline the importance of this question, we conduct the most extensive empirical evaluation to date of the QENP dataset with an analysis of 41 different models and a total of 193 model variants (e.g., different cost matrices) and demonstrate INTERCEPT's superior performance to traditional modeling approaches. (3) As a first for adversary behavior modeling applications applied to the wildlife crime domain, we present the results of a *month long* real-world deployment of INTERCEPT: compared to historical observation rates of illegal activity, rangers that used INTERCEPT observed 10 times the number of findings than the average. In addition to many signs of trespassing, rangers found a poached elephant, a roll of elephant snares, and a cache of 10 antelope snares before they were deployed (pictures in Figure 1). Each confiscated snare represents an animal's life saved; while the rangers' finding of a poached elephant carcass is a grim reminder that poachers are active, these successful snare confiscations demonstrate the importance of real-world data in developing and evaluating adversary behavior models.

2. ONGOING DEPLOYMENTS

Although the initial field test of the proposed predictions yielded promising results, it is necessary to conduct a more systematic real-world evaluation of our models. Starting in November 2016, we have been conducting a larger-scale deployment in order to test the predictive performance of our decision tree ensemble.

For every patrol post (the starting and ending point for ranger patrols) in the park, we generated a set of 3x3 sq. km patrol areas and compute the attack prediction rate for that area (i.e., $\frac{\#positivePredictions}{areaSize}$). Then, each of the patrol areas were divided into three experiment groups via k-means clustering on the attack prediction rate; cluster 1 will correspond to the areas with the highest attack prediction rates, cluster 2 for medium attack prediction rate areas, and cluster 3 for the lowest attack prediction rates. Based on these groups, we will then test the following hypotheses: (1) "More attacks are observed in areas with higher concentrations of attack predictions." and (2) "Less attacks are observed in areas with small concentrations of attack predictions." In addition to testing these hypotheses and analyzing our model's deployed performance (e.g., hit rate comparison), we will evaluate existing models' predictive performance on the new dataset: does our deployed model still outperform all other models?

Acknowledgments: This research was supported by MURI grant W911NF-11-1-0332 and a subcontract from Cornell University for NSF grant CCF-1522054. We are grateful to the Wildlife Conservation Society and the Uganda Wildlife Authority for supporting data collection in Queen Elizabeth National Park (QENP). We thank all the rangers and wardens in QENP for their contributions in collecting and providing patrolling data in SMART.

REFERENCES

[1] R. Critchlow, A. Plumptre, M. Driciru, A. Rwetsiba, E. Stokes, C. Tumwesigye, F. Wanyama, and C. Beale. Spatiotemporal trends of illegal activities from ranger-collected data in a ugandan national park. *Conservation Biology*, 29(5):1458–1470, 2015.

[2] F. M. Delle Fave, A. X. Jiang, Z. Yin, C. Zhang, M. Tambe, S. Kraus, and J. P. Sullivan. Game-theoretic patrolling with dynamic execution uncertainty and a case study on a real transit system. *Journal of Artificial Intelligence Research*, 50:321–367, 2014.

[3] F. Fang, P. Stone, and M. Tambe. When security games go green: Designing defender strategies to prevent poaching and illegal fishing. In *International Joint Conference on Artificial Intelligence*, 2015.

[4] D. Korzhyk, V. Conitzer, and R. Parr. Solving stackelberg games with uncertain observability. In *International Conference on Autonomous Agents and Multiagent Systems*, 2011.

[5] S. Mc Carthy, M. Tambe, C. Kiekintveld, M. L. Gore, and A. Killion. Preventing illegal logging: Simultaneous optimization of resource teams and tactics for security. In *AAAI*, 2016.

[6] E. Munoz de Cote, R. Stranders, N. Basilico, N. Gatti, and N. Jennings. Introducing alarms in adversarial patrolling games. In *International Conference on Autonomous agents and Multiagent systems*, 2013.

[7] T. H. Nguyen, A. Sinha, S. Gholami, A. Plumptre, L. Joppa, M. Tambe, M. Driciru, F. Wanyama, A. Rwetsiba, R. Critchlow, et al. Capture: A new predictive anti-poaching tool for wildlife protection. In *International Conference on Autonomous Agents & Multiagent Systems*, 2016.

[8] Phys.org. More tigers poached so far this year than in 2015: census. Web, April 2016. `http://phys.org/news/2016-04-tigers-poached-year-census.html`.

[9] E. Shieh, B. An, R. Yang, M. Tambe, C. Baldwin, J. DiRenzo, B. Maule, and G. Meyer. Protect: A deployed game theoretic system to protect the ports of the united states. In *International Conference on Autonomous Agents and Multiagent Systems*, 2012.

[10] Traffic.org. South africa reports small decrease in rhino poaching, but africa-wide 2015 the worst on record. Web, January 2016. `http://www.traffic.org/home/2016/1/21/south-africa-reports-small-decrease-in-rhino-poaching-but-af.html`.

[11] R. Yang, C. Kiekintveld, F. Ordonez, M. Tambe, and R. John. Improving resource allocation strategy against human adversaries in security games. In *International Joint Conference on Artificial Intelligence*, 2011.

[12] C. Zhang, A. X. Jiang, M. B. Short, P. J. Brantingham, and M. Tambe. Defending against opportunistic criminals: New game-theoretic frameworks and algorithms. In *International Conference on Decision and Game Theory for Security*, 2014.