MONTEREY, CALIFORNIA

ENHANCING TRUST IN BATTLE-MANAGEMENT SYSTEMS THAT USE ARTIFICIAL INTELLIGENCE

EXECUTIVE SUMMARY

Principal Investigator (PI): Dr. Neil C. Rowe, Computer Science (CS)

Additional Researcher(s): No additional researcher participated in this research project.

Student Participation: Mr. Gabriel Lipow, CIV, INT; Ms. Aroshi Ghosh, CIV, INT; and Ms. Lahari Yallapragada, CIV, INT

Prepared for:
Topic Sponsor Lead Organization: N2/N6 - Information Warfare
Topic Sponsor Name(s): Mr. William Treadway
Topic Sponsor Contact Information:
william.a.treadway.civ@us.navy.mil, (703) 695-8008

This research is supported by funding from the
Naval Postgraduate School, Naval Research Program (PE 0605853N/2098).

Project Summary

This work addressed ways of making machine learning more robust against adversaries manipulating its input data, as with sensor data an adversary can at least partially control. The approach we explored was training and comparing alternative artificial intelligence (AI) models for the same data. An AI model is a software structure that draws conclusions from data, such as a neural network, a random forest, or a Bayesian network. Usually adversaries assume a neural-network model is being used, and they manipulate the input to find cases in which incorrect conclusions are reached after standard training; they then try to supply similar data in operational scenarios. However, it is very difficult to find cases that fool more than one machine-learning model, so training more than one on the same data will often reveal the data manipulation by disagreements between conclusions. Our project tested this solution on real ship-tracking data from the public Automatic Identification System database available from the U.S. Coast Guard for U.S. coastal waters. We built ship tracks out of individual ship records and identified ten key features of tracks that help classify ship types, with the goal of identifying ships falsely reporting their identities. This identified some interesting suspicious anomalies and faulty reports in the data. Then we systematically perturbed the data to various degrees to see how that affected the ship-type classifications with eight standard machine-learning methods. We were particularly interested in effects that did not uniformly increase with the degree of perturbation, effects that varied significantly with the direction of perturbation, and in major differences in effects between the machine-learning methods. Our results did find some interesting weaknesses in the methods, and our methodology is general enough to be applied to many other kinds of machine-learning tasks.

Keywords: *machine learning, testing, adversary, data manipulation, deception, models, ships, tracking, ship types, artificial intelligence*

Background

The machine-learning subarea of artificial intelligence has become very popular for military applications recently due to its successes with computer vision and sensor data. However, these advances can be countered by an adversary that has at least partial access to the data. This could occur with aerial or satellite imagery of adversary territory falsified using conventional deception methods, with fake signals data provided by adversary emitters, or malware emplaced by adversaries on friendly systems to generate false information. Such data manipulation can cause learning of false rules and trends about an adversary as a kind of deception, and this can be a significant force multiplier for an adversary. Often it is applied to classification problems where friendly forces must decide whether they detect a threat and what kind it is, and the adversary can manipulate the threshold by small changes to parameters.

Countering these methods can be difficult. The popular neural-network models in particular are complex mathematically and are generally too hard for humans to understand, so understanding why and how an adversary manipulation works is usually impossible. Confirming the provenance of data is a challenging forensic problem due to the complexity of our digital systems and their hardware. Attempts to identify the parts of a neural network that are most contributing to a surprising conclusion ("salience") are sometimes possible but unreliable. Retraining the learned model in response to every new observation that might represent adversary manipulation requires considerable time for important neural networks.

One promising approach is to run multiple machine-learning models on some suspect data and compare their conclusions. If the models have been well-trained on trusted data, disagreements they have about conclusions should be rare, and large numbers of disagreements are suspicious. Despite the current tendency to make neural networks the default method for machine learning because of their often-high accuracy, other methods have advantages in features such as time and storage. In particular for classification problems, simple linear and logistic models, decision trees, and Bayesian inference can provide a baseline of performance for machine learning that can flag obvious failures of a neural network. This approach we investigate in this work.

Adversaries manipulating data are performing a kind of deception, so it is important to them that their manipulations remain undiscovered. Thus their standard method is to perturb the data just enough to cause misclassifications. Thus to provide a good test of effects of data manipulations we should perturb data minimally and look for unusual effects. This is what we have done.

This research used data from ship tracking provided by the U.S. Coast Guard from their Marine Cadastre site (www.marinecadastre.gov). Tracking data provides much useful data about the purposes and missions of ships, some of it quite subtle. Though this data includes few military ships, the principles we have developed should be applicable to them to detect military reconnaissance and operations, and enabling such classified analysis is the ultimate goal of this work.

Findings and Conclusions

We tested the prediction of ship type from its properties and tracks, something important in detecting illegal activity at sea. The results showed significant differences between not only ship-type classification accuracy because of perturbation, but also in what attributes each model considers most important; coordinate popularity (how much the ship moves between transponder reports) and ratio of width to length were the most important overall. Four ship classes were harder to classify: military ships, passenger ships, "unknown" ships, and "other" ships. We removed the "unknown" ships from the final tests since they were so variable, but predicting their true types can be done with our methods. Some models were clearly more resistant to perturbation than others, with Bayesian networks and the decision tables being good in this property, and random-forest and nearest-neighbor models were entirely unaffected. The most severely affected were logistic neural networks, Naïve Bayes, and simplified logistic models, and the J48 decision trees ranked approximately average for these tests.

Tests with evenly spaced perturbations gave us the best data about the sensitivity of each model. We observed and documented several unusual phenomena in our data, with some tests showing that classification accuracy jumped at some points, sometimes even above the baseline accuracy without any data perturbation. This was especially noticeable in the Bayesian models (net and Naïve). Other unusual results included large dropoffs in accuracy for the simple logistic and logistic neural-network models, and our inability to cause the random forest model to significantly misclassify ships, since we could only cause a drop in accuracy from 100% to 99.7% accuracy with large negative perturbations. In many models, the importance of each attribute seemed to change with every perturbation in a nonlinear way. For instance, the performance of the Naïve Bayes model with a perturbation of -1.0 was 70% worse than with a perturbation of 1.0.

We observed and documented cases where either classification accuracy dropped inconsistently and asynchronously with increasing perturbations, or less often, where classification accuracy actually rose with increasing perturbations. Different models had different tipping points in accuracy. Our models struggled to classify several kinds of ships, in particular military vessels, passenger ships, "unknown" ships, and "other" ships.

We produced detailed summaries of the effects of eight machine-learning methods on ten attributes of ship tracks for future guidance of designers of machine-learning methods. Overall, machine-learning methods are shown effective in identifying ship types in real data from their track data, which should be helpful in detecting military activities, smuggling, or illegal fishing. Increasing input perturbations usually mean increasing effects, but not always; this means that some judgment is needed in applying machine-learning models in critical situations involving tracking data. All models were vulnerable to sufficiently large perturbations by demonstrating greatly reduced classification accuracy, but these cases are highly noticeable on data inspection and should not be a threat.

Recommendations for Further Research

We did not find any major vulnerabilities in the machine-learning methods we tested in regard to perturbations of their input data. We thus confirm their effectiveness in future work in analysis of tracking data. Nonetheless, future work needs to investigate further the sensitivities we found in some of the machine-learning methods, and analyze further how an adversary might exploit them. It also needs to be expanded to cover some of the more complex neural-network methods popular today, especially those that are slow and difficult to train although very accurate. The more complex a model, the more there is for adversaries to exploit. More data should be collected, especially outside of U.S. coastal waters, to allow testing on the rare but important kinds of suspicious maritime activity like military reconnaissance, illegal fishing, and smuggling.

Acronyms

AI artificial intelligence