From Data to Decisions:

Placing Machine Learning Challenges In Context

David J. Stracuzzi, Maximillian G. Chen, Michael C. Darling, and Matthew G. Peterson

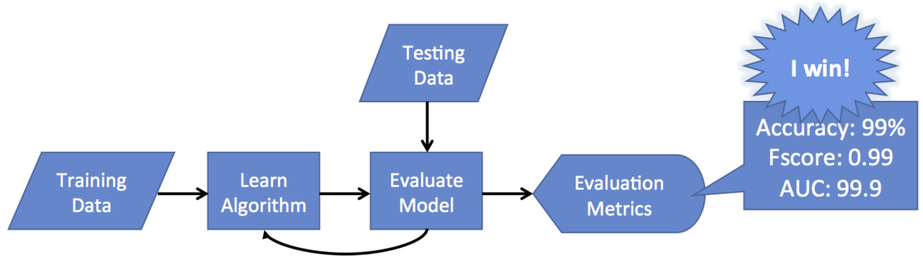
Sandia National Laboratories

Background

Challenges advance both the state of the art and the state of practice by requiring participants to apply modeling and prediction techniques to real world data. The difficulties in dealing with noisy, messy, or insufficient data requires participants to adapt and generalize algorithms in novel, exciting ways. However, **challenges in their current form neglect the larger decision-making context in which machine learning algorithms are often used.** In particular, uncertainty plays an important role in data-driven decision making.

Current Process For Winning Machine Learning Challenges: Maximize Evaluation Metrics

- **Current challenge problems lack application-specific context; criteria are accuracy-based metrics.**
- **ML systems are part of larger pipeline; challenges should include post-analysis decision making.**
- **Contest evaluation metrics should relate to application goals, not just standard performance criteria.**

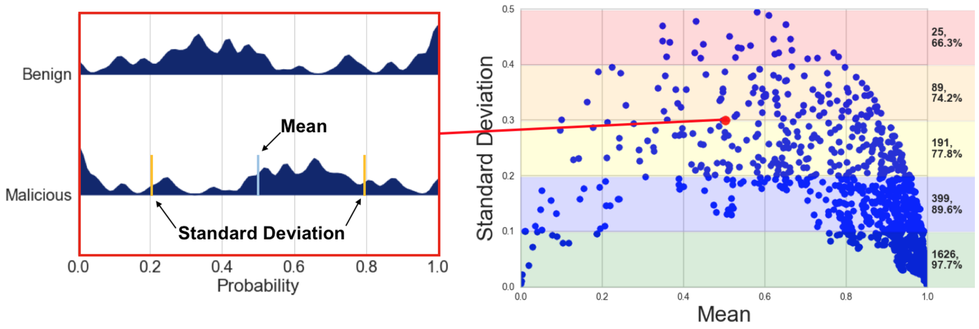


Example: Classifying Malicious Websites Given URL Text and Structure

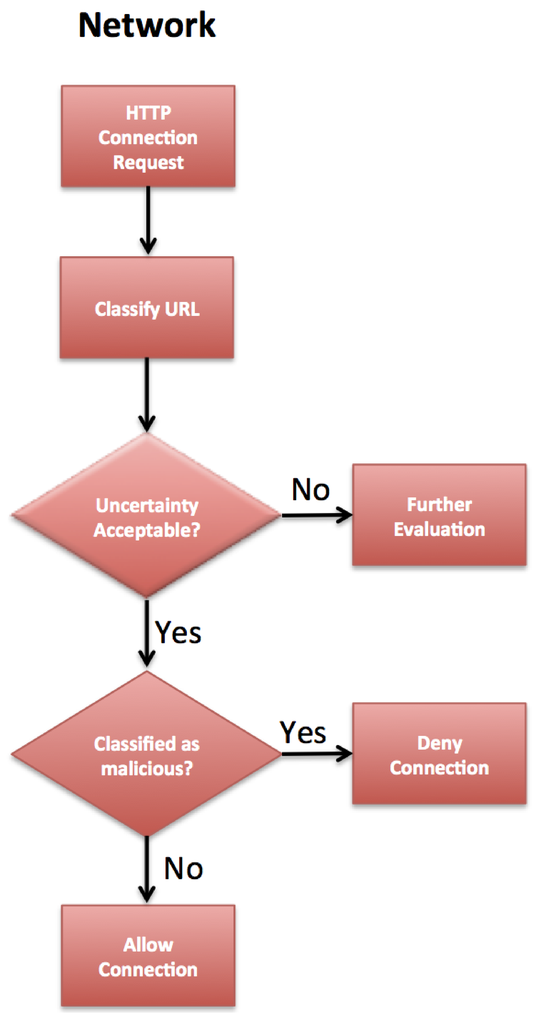
The Challenge: To avoid blacklists, adversaries create and abandon URLs rapidly. The expectation of concept drift means that the system needs to recognize when errors are more likely.

Without Context: Challenge winner determined using traditional metrics. The system shows high performance with respect to evaluation data. However, due to the size of the population, there will always be sets of URLs which are outliers with respect to the system's hypothesis.

With Context: Participants must consider false negatives as extreme risk to network and recognize that a human may play a role in decision making, such as by overriding an automated browser decision.



The graphs show that prediction distributions with higher means and lower standard deviations tend to be classified correctly. Therefore, a decision-maker can have increased trust in such predictions. The color bars annotate the averaged prediction accuracy of the samples they contain. For example, the green bar contains 1626 predictions whose average accuracy is 97.7%.



Conclusion

Machine learning algorithms are one part of a larger pipeline, and future challenge problems should include the post-analysis decision making. This means that evaluation metrics must relate to application goals instead of on traditional machine learning performance criteria. A possible avenue is to provide uncertainty information above and beyond yes/no classifications or even point-estimate probabilities. The criteria by which machine learning research is judged often differs from the criteria by which applications are judged. As a simple illustration, the review criteria for the research-focused AAAI conference is vastly different from the application-centric Innovative Applications of AI conference [1, 2]. Machine learning competitions are well-positioned to establish more nuanced and application-driven evaluation measures in the community.

References

AAAI-18 call for papers. `http://www.aaai.org/Conferences/AAAI/2018/aaai18call.php`, 2018. Accessed: 2017-10-03.

IAAI-18 call for papers. `http://www.aaai.org/Conferences/IAAI/2018/iaai18call.php`, 2018. Accessed: 2017-10-03.

Acknowledgements

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Contact: djstrac@sandia.gov