Exposing Statistical Errors Made by Machine Learning Algorithms: Revisiting the Boy Who Cried Wolf in the Context of Phishing Detection

(Authors' names blinded for peer review)

Grown out of the quest for artificial intelligence (AI), machine learning (ML) is today's most active field across disciplines with sharp increase in applications ranging from criminology to fraud detection and to biometrics. Machine learning and statistics both emphasize the importance of model estimation / training and thus share the inescapable type I and II errors. Extending the concepts of statistical errors into the domain of machine learning. we devise aground-breaking pH scale (in chemistry)-like alpheta ratio and intend it as a litmus test of decision risk bias supplementing the established criterion of Accuracy. Using publicly available phishing data set, we conduct experiments on a series of single-feature models using the CHAID package in R. Based on the results, we recommend practitioners match the risk over / under estimation cost ratio with the error ratio associated with each machine learning model in order to mitigate potential losses in their specific decision-making environments

Key words: machine learning, Type I and II errors, CHAID, alpheta ratio, CRAN-R, Phishing Websites Classification

1. Introduction

Aesop's fable, the The Boy who Cried Wolf, is about a young shepherd boy found his life dull in the pasture as he sat on the hillside tending his masters sheep. To amuse himself, he ran toward the village shouting at the top of his voice, "Wolf! Wolf! The Wolf is chasing the sheep!". The villagers dropped their work and rushed up the hill to help the boy scare the wolf away. But upon their arrival, they found no wolf and only the grinning face of the shepherd boy. "Don't cry 'wolf', when there's no wolf!" The grumbling villagers told the boy and returned to their village. A few days later, the boy felt bored and yell out again, "Wolf! Wolf! The wolf is chasing my sheep!" To his wicked delight, he viewed the villagers dash up the hill and fall for his trick gain. The villagers sternly warned "Don't cry 'wolf' when there is NO wolf!". Later, he saw a real wolf prowling about his herd. In terror, he leaped to his feet and sang out as loudly as he could, "Wolf! Wolf!". To no avail, the villagers thought he was trying to repeat his foolish game, so they stay away. At sunset, everyone wondered why the shepherd boy hadn't returned to the village with their sheep. They went up the hill and found him weeping. The price the village paid is costly: the killing of a great many of the flock.

At the end of the story, an old man comforted the boy. If you tell too many lies, no one believes you when you tell the truth. The moral message the tale conveys is that liars are not trustworthy even when they are telling the truth. Undoubtedly, the shepherd boy was not even close to a pure rational decision maker, perhaps a naughty one. Considering his maturity level and working environment, are we being too harsh on the young boy entrusted with the task of guarding the villages most valuable asset? Roulston and Smith (2004) delineate the classical tale in a decision-making matrix as shown in table 1. Based on their cost-loss analysis, the villagers in the tale were unprepared to tolerate a reasonably high false-alarm rate (cries of Wolf! when there is no wolf looming). As revealed by our ensuing investigation, even machine learning (ML) models trained on large data sets are not immune from sounding false alarms.

Our research was motivated by what we observed as a slight disconnect between the computing side (predictive accuracy) and statistical side (model bias) of ML. ML should induce more objective and evidence-based decision-making, since machines are supposedly free from human prejudice. As a demonstration effort, one major point to mention at the start is that our project did not focus on any testable hypotheses per se. Instead, our study was undertaken with three overall goals:

1. To alert the decision analysis community to the increasing popularity of AI-powered machine decision making (MDM) and automation, which pose to render vast swaths of the working professionals literally redundant.
2. To propose and formulate an alpheta ratio (akin to the pH scale in chemistry) for the assessment of decision risk bias of machine learning models.
3. To put old wine (statistical errors) into new bottles (machine learning framework) and examine how the decision-making environment (phishing detection) influences model selection.

The paper proceeds as follows. In the Related Works section, we describe phishing detection, which serves as the context of our experiments. Also, discussed in that same section are machine learning and statistics, statistical errors, and Chi-squared Automatic Interaction Detection (CHAID) - the machine learning algorithm we used for our project. Amid the Related Works section, we proposed a unique alpheta ratio for the purpose of assessing machine learning models. In the Methodology section, we describe the Phishing data set, the set of 12 models to be evaluated, as well as the K-fold cross validation used to assess models. The Results section presents the results we acquired from out experiments. In the Implications section, we discuss how to match a machine learning models alpheta ratio to the risk overestimation cost and underestimation loss profile dictated by the decision-making environment. In the Concluding Remarks section, we re-examine the Boy Who Cried Wolf fable via alternative lens of decision making and spotlight the importance of human intelligence in the face of wide-spreading artificial intelligence.

Table 1 The Villagers Benefit and Cost Matrix

	Perception	
Reality	Wolf - Yes	Wolf - No
Wolf -Yes	Respond and Rescue Time and Efforts well spent	Stay Put Loss of Assets
Wolf -No	Opportunity Cost Waste of Time and Efforts	No Harm Zero Cost

Adapted from Roulston and Smith (2004)

2. Related Works

2.1. Phishing Detection

Mohammad (2015)

2.2. Classification using ML Algorithms

Machine learning is the manifestation of statistical learning (statistics) algorithms implemented via software (computing) applications.

2.3. Evaluation of ML Algorithms

3. The Proposed Alpheta Ratio

4. Concluding Remarks

George Box, "one of the great statistical minds of the 20th century", made famous the aphorism - "all models are wrong, but some are useful" (Box and Draper 1987, p. 424). offers an insightful testament to

References

Box GE, Draper NR (1987) *Empirical model-building and response surfaces.* (John Wiley & Sons).

Mohammad RM (2015) Phishing websites features. URL `http://dx.doi.org/10.13140/rg.2.1.2595.6000`.

Roulston MS, Smith LA (2004) The boy who cried wolf revisited: The impact of false alarm intolerance on costloss scenarios. *Weather and Forecasting* 19(2):391–397, URL `http://dx.doi.org/10.1175/1520-0434(2004)019<0391:TBWCWR>2.0.CO;2`.