The University of Texas at Austin
Center for Identity

ECONOMIC MACHINE LEARNING FOR FRAUD DETECTION

Maytal Saar-Tsechansky

2015

UT CID Report #1511

This UT CID research was supported in part by the following organizations:

LifeLock Guarantee Your Good Name
HID
TransUnion
Applied Fundamentals Consulting LLC
CSID
LexisNexis
SAFRAN MorphoTrust USA

ECONOMIC MACHINE LEARNING FOR FRAUD DETECTION

Introduction and Motivation

The daunting risk of health care, identity, and cyber fraud result in billions of dollars in losses each year and posit serious threats to both individuals and nations. In recent years, predictive machine learning models have emerged as critical for effective detection of fraud by automatically learning patterns of fraud from data and by adapting to new patterns of fraud as these emerge. Such models use data on the relevant domain to estimate the likelihood (and/or amount) of fraud and to effectively allocate auditing and enforcement resources.

However, learning effective predictive models from available data in these domains posit difficult challenges.

Consider for example the application of machine learning to detect health care fraud. In this case, supervised machine learning requires data in the form of prior claims that have been previously audited, such that it is known whether or not these cases are fraudulent. However, previously audited cases are not necessarily the most informative for machine learning to learn distinct characteristics fraud or illegitimate activities. Consequently, intelligent acquisition of particularly informative audits can cost-effective improve boost fraud detection and minimize losses through better allocation of auditing and enforcement resources. Towards that, it is imperative to develop methods that would enable machine learning techniques reason intelligently about opportunities to acquire particularly beneficial information for learning, and that in turn cost-effectively enhance key detection and enforcement goals, such as maximizing compliance or minimizing losses.

The cyber medium has generated major benefits to modern society in recent decades. However, over time it has become evident that cyber technologies are also the source of significant new risks and vulnerabilities. In effect, malicious conduct had become prevalent in the cyberspace in multiple forms, rendering effective defensive solutions highly desirable. Their search proposed here focuses on two critical aspects of such cyber defensive solutions: data-driven, autonomous detection technology and cost effectiveness. Specifically, we propose framework and specific methods for economically efficient allocation of labeling tasks required in supervised learning-based defensive solutions.

Automated technologies have emerged as a critical element of cyber defense applications. Because it is impossible for security experts to continuously monitor emerging patterns within data packets, transactions, or online interactions, cyber protection relies on automated solutions to continuously detect and prevent potential threats. One key technology that supports automated solutions is data-driven supervised learning. Supervised learning addresses security threats by capturing patterns in historical data that are characteristic of threats, and then detecting these patterns in the future. Unlike rule-based approaches that require human experts to specify rules that are indicative of

Copyright © 2015 Center for Identity at The University of Texas at Austin Confidential and Proprietary

security threats, supervised methods "learn" patterns that characterize such threats in a data-driven manner. Supervised learning has been successfully applied to a myriad of important security applications including inappropriate content filtering, intrusion detection, video-surveillance-based intention detection, internet bullying detection, and online fraud detection, among other tasks.

Another critical aspect of cyber defense is cost effectiveness. Because it is practically impossible to purchase perfect defense, companies struggle to provide the best possible defense given tightening security budget constraints (Lawrence and Loeb, 2002). In effect, driving defense costs higher to render adequate defense economically prohibitive is a likely strategy employed by attackers. Consequently, cost-effective defense becomes a key goal for sustainable cyber defense. This objective is also shared by off-the-shelf security solution providers, which aim to remain competitive by reducing development costs.

The rising costs of effective defense trigger the fundamental cyber challenge faced by companies, organizations and security solution providers: how to provide effective security given a limited budget. The research proposed here aims to address both the technological and economic aspects of this challenge in the context of supervised learning-based security defense technology. In what follows we describe the problem in more detail and outline our proposed approach.

Supervised learning is particularly suitable to efficiently achieve adequate cyber security. Supervised learning often surpasses human ability to detect patterns among vast amount of data and requires no significant guidance from humans. Perhaps more importantly, because many cyber security challenges emerge in an adversarial setting, where attackers continuously adapt their attacks to the target defense strategies, the ability of defense systems to efficiently "learn" new characteristics of attacks directly from the data eliminates the need for security domain experts to continuously update and maintain an ever growing number of predefined rules.

However, supervised learning requires labeled training data, which are inherently costly to acquire in many important cyber settings. Particularly, for supervised learning, the dependent variable value of each training example (e.g., whether or not a certain transaction is fraudulent, or whether a certain website is a "spoof"), must be known so as to induce predictive patterns from data.

More specifically, security applications commonly require high levels of performance in order to detect most actual threats while minimizing undesirable false detections. To achieve such high levels of performance, data-driven security applications require large volumes of highly expensive labeled training instances to learn accurate models. For example, inappropriate content filtering requires costly labeling of large volumes of textual, image, and video content as appropriate or inappropriate. Similarly, video-surveillance-based intention detection requires labeling of numerous videos of body gestures and movements as threating/non-threatening. Additionally, a variety of online fraud detection tasks require hiring expensive fraud specialists to determine whether a large number of

Copyright © 2015 Center for Identity at The University of Texas at Austin Confidential and Proprietary

transactions are in fact fraudulent or not, before these transactions can be used for training data-driven models.

Furthermore, because of the adversarial nature of many security domains, hackers and fraudsters often adapt their attacks based on current patterns of detection so as to decrease the likelihood of detection of future attacks. In such settings in particular, re-learning patterns of cyber attackers must occur continuously, requiring a constant flow of costly labeled training instances to maintain high level of detection performance. Similarly, for effective detection of offensive Internet content (such as for online bullying detection), the rapid evolution of Internet content, such as emerging themes, expressions or slang, can render a supervised learning model obsolete unless continuous flow of labeled training examples is available to adapt the models.

Overall, the challenge to achieve the great promise of automated, data-driven learning lies in the ability to effectively manage mounting labeling costs.

Recently, online marketplaces for human workforce intelligence tasks, such as Amazon's Mechanical Turk, or freelancers' websites (e.g., Freelancer.com) have presented exciting opportunities for bringing to bear human intelligence to support data-driven learning (Brynjolfsson et al., 2014).1Particularly relevant for this research, online marketplaces such as Amazon Mechanical Turk present new opportunities for "automating" the labeling procedure by programmatically allocating data instances for labeling. These abilities also provide opportunities for cost-effective labeling. However, achieving these benefits is non-trivial. For this promise to materialize, it is imperative to characterize and address several key challenges.

First, given the nature of the work as well as the incentive structure, key challenges include inaccurate (noisy) labeling (Imperators et al., 2014), timely completion of tasks, and meeting strict budget constraints. This is in addition to the challenge of carefully selecting informative data items so as to improve the accuracy of the model learned from the acquired labeled instances. Addressing all these challenges simultaneously is a novel and difficult problem that has not been addressed in prior work.

Accuracy-centric algorithms for cost-effective economic labeling

We begin by considering labeling markets in which the cost of labels can reflect arbitrary relationships. In particular, research thus far has documented conflicting evidence of relationships between the quality of labels, namely the proportion of correct labels obtained by outsourcing platforms, and the cost of label acquisitions. These include for example evidence of no apparent change in quality for different pays, or hyperbolic relationship such that increasing costs first yield increasingly higher quality followed by dropping quality for increasing cost. Indeed, it is possible this evidence suggests, that the relationship between cost and label quality varies across different domains.

We consider a setting in which labels can be acquired at different levels of costs, each level potentially yielding different quality of labeling. Specifically, the quality of labeling is

Copyright © 2015 Center for Identity at The University of Texas at Austin Confidential and Proprietary

reflected by the probability of the label being correct. We aimed to develop an acquisition policy that is entirely data-driven and agnostic to the prevalent (though unknown) relationship between the cost of labels and the ensuing label quality. In addition, label acquisition is done sequentially, such that at each phase, N labels are acquired. All the algorithms we evaluated begin with a.

Our first policy, Max Ratio aims to evaluate the expected improvement in AUC per unit cost from acquiring labels at each possible cost, chi, and select the cost with the highest expected ratio. In particular, key to this evaluation is that we begin with an initial labeled set composed by drawing a random sampled from UL and acquiring the corresponding instances labels at each of the cost levels chi. subsequently, the expected improvement in performance per unit cost is estimated for each labeling cost. Specifically, we draw a random subset of instances acquired at cost chi and omit it from the training set. The different between the performance of the model induced from the labeled set L and the performance of a model induced from the "reduced" set is used as a proxy to the expected improvement in performance if an additional set of instances is acquired at cost chi. This expected improvement is then divided by the cost chi to estimate the expected improvement per unit cost. Furthermore, we repeat the evaluation of the expected improvement for different random draws of a subset of instances acquired at cost chi to accommodate the model variance. The pseudo code for the Max Ration procedure is shown below.

Copyright © 2015 Center for Identity at The University of Texas at Austin Confidential and Proprietary

Pseudo-code: Max Ratio

Given:
Ci : Cost of label at quality i. i:1,2,..,k
p(c_i) : the probability of yielding a correct label given cost c per label
Inducer I to induce a predictive model

An initial training set of labeled training instances L, a representative test set T of instances, and a large set of unlabeled instances UL

1. Populate L such that it includes T instances acquired at each cost c_i, i:1,2,..,k
2. Repeat V times (acquisition phases)
   a. Induce a model from L and evaluate its performance P(L) via cross validation
   b. For each cost i= 1,2,..,k
      For j = 1 to N do:
         Draw a sample Sj of size t << (T*K) from L of instances acquired at cost ci
         Remove sample Sj from L, induce a model M(L-Sj) from the reduced set L-Sj, and evaluate its performance via cross validation.
         Bj(Ci)=P(L)-P(L-Sj)
      End For j
   Compute average improvement per unit cost from labeling at cost ci:
   Effect_B(Ci)= Average_j [Bj(Ci)]/ci, for j=1,2,...,N
   End for i
3. Select cost Ci such with the maximum average benefit Avg_B(Ci)
4. Draw a random set W from UL, acquire labels for instances in P at cost Ci, and augment L with the instances: UL←UL+W
5. Induce a model M from the augmented UL and evaluate M_i's performance on the test set T

   End Repeat

6. Produce learning curve

Max ratio may suffer from several potential weaknesses, most of which stem from the difficulty in correctly estimating the correct improvement in performance. First, due to estimation variance this estimation is likely to be imprecise, particularly early on the learning curve, during the initial acquisition phases, and when the training set L is small. A key challenge is when all expected improvements is estimated to be negative. This is likely to be an artifact of the estimation. In such cases we employ a heuristic and use the least costly labels to minimize the risk.

MAX RATIO VARIANTS

We examined several variations to the Max ration algorithms. First we explore a policy which does not consider the improvement per unit cost, but only aims to estimate the expected improvement in performance from acquiring labels at different costs.

Max Quality: If the differences in cost are small, given the variance in estimating the expected improvement dividing the expected change in performance by the corresponding

Copyright © 2015 Center for Identity at The University of Texas at Austin Confidential and Proprietary

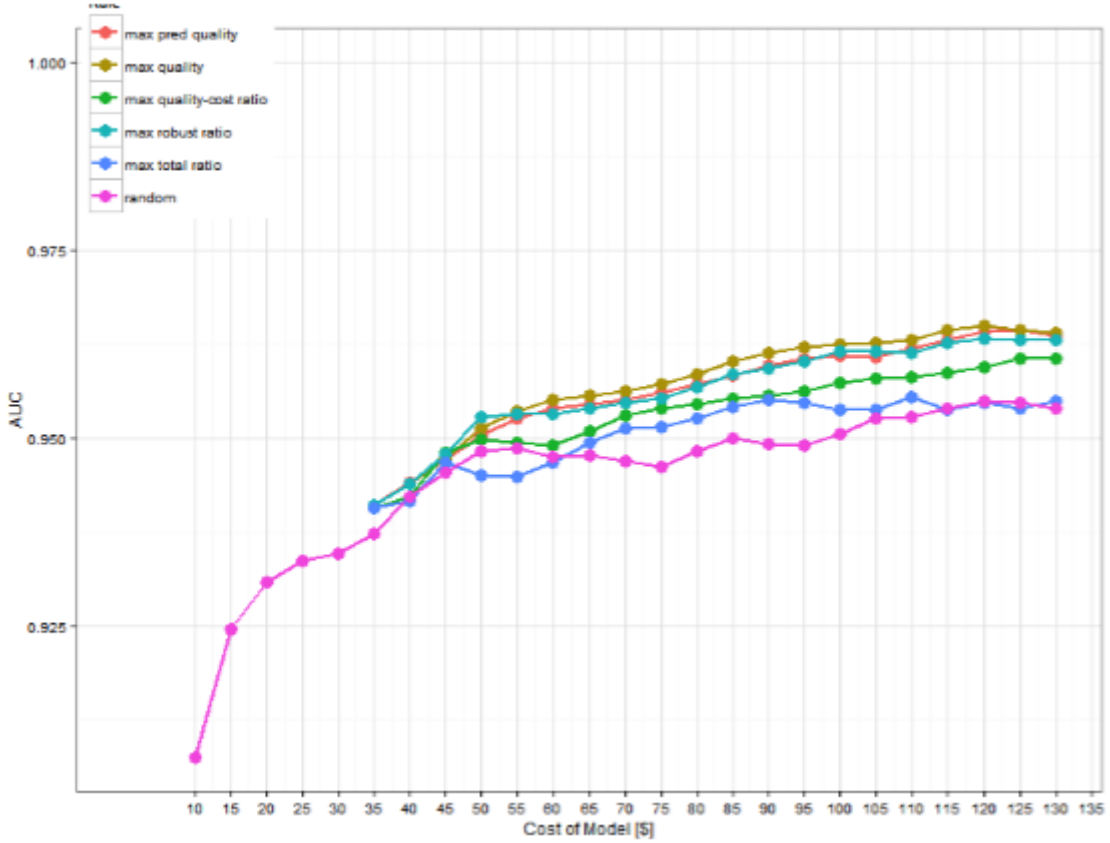
cost has the potential to augment the error in estimation. We consider several variants to this policy:

Max Robust Ratio: Instead of entirely ignoring the cost, two other variants aim to reduce the effect of cost on selecting the best labeling cost to choose. In particular, as before, one variance considers the marginal improvement in predictive performance. However, rather than divide this improvement by the marginal cost per label we divide it by the cumulative cost.

Max Predicted Quality: This variant does not aim to estimate the expected improvement in performance by removing previously acquired instance, but rather but estimating the trend of the learning curve and projecting the expected performance

RANDOM: As a benchmark, we consider acquisition of labels at a randomly selected cost.

The figure below presents the area under the AUC curve obtained after each acquisition phase as a function of the cost incurred for label acquisition.



Copyright © 2015 Center for Identity at The University of Texas at Austin Confidential and Proprietary

Conclusions and Discussion

Our results demonstrate that evaluating the expected improvement in performance yield an ability to select generally good acquisitions in a cost-effective manner. Several policies yield comparable performance. These results suggest that our policies are able to identify acquisition costs that yield labeling quality to produce the desired improvement in performance. Yet, it would be desirable to further improve the estimation of these improvements so as to identify the best cost-effective acquisitions.

One possible explanation for these results is that the differences in costs are very small and insignificant to yield meaningful differences in benefits. Towards that, we plan to experiment with settings in which the differences in costs are more significant. Another direction for improvement is to further improve the estimation in expected predictive performance. Towards that we aim to explore two directions. The first is to conduct the cross validation multiple times so as to reduce the estimation variance further. A second strategy we aim to explore is consider adding new instances to the training set rather than evaluate the loss in omitting instances acquired at a given cost. Specifically we aim to draw instances from the training set L acquired at a given cost chi and create copies of these instances that will be added to the training set. Evaluating the difference in performance between the augmented set and the current one may yield a better estimate of the expected change model performance.

Copyright © 2015 Center for Identity at The University of Texas at Austin Confidential and Proprietary

The University of Texas at Austin
Center for Identity

© 2015 Proprietary, The University of Texas at Austin, All Rights Reserved.

For more information on Center for Identity research, resources and information, visit **identity.utexas.edu.**