Randomized Operating Point Selection in Adversarial Classification

Viliam Lisý, Robert Kessl, and Tomáš Pevný

Agent Technology Center, Department of Computer Science
Faculty of Electrical Engineering, Czech Technical University in Prague
{viliam.lisy,robert.kessl,tomas.pevny}@agents.fel.cvut.cz

Abstract. Security systems for email spam filtering, network intrusion detection, steganalysis, and watermarking, frequently use classifiers to separate malicious behavior from legitimate. Typically, they use a fixed operating point minimizing the expected cost / error. This allows a rational attacker to deliver invisible attacks just below the detection threshold. We model this situation as a non-zero sum normal form game capturing attacker's expected payoffs for detected and undetected attacks, and detector's costs for false positives and false negatives computed based on the Receiver Operating Characteristic (ROC) curve of the classifier. The analysis of Nash and Stackelberg equilibria reveals that using a randomized strategy over multiple operating points forces the rational attacker to design less efficient attacks and substantially lowers the expected cost of the detector. We present the equilibrium strategies for sample ROC curves from network intrusion detection system and evaluate the corresponding benefits.

Keywords: Game theory, operating point selection, receiver operating characteristic, adversarial machine learning, misclassification cost.

1 Introduction

Receiver operating characteristics (ROC) graph is a curve showing dependency of true positive rate (y-axis) and false positive rate (x-axis) of a classifier. The most attractive property of ROC curves is their insensitivity to changes in class distributions and costs of wrong decisions on different classes. Both these properties are almost certainly user specific and often non-stationary in security applications. For example, in spam detection the proportion of spam volume changes from month to month, and the cost of receiving spam can be different for different users. It is therefore important to determine operating points of classifiers (e.g., thresholds) systematically based on the costs in a specific deployment. A well-established result [20] shows that the threshold minimizing the (Bayesian) cost corresponds to the tangent of the ROC curve of slope defined by the ratio of mis-classification costs weighted by class probabilities. This method for selecting thresholds is routinely used in many domains, including steganalysis[11], watermarking [13, 6], and fraud detection[17].

Similarly to [3], we argue that the method is optimal only in non-adversarial setting without a rational attacker actively avoiding the detection. If knowledgeable attackers are present, such as in network intrusion detection, spam filtering, steganalysis, watermarking, etc., this standard operating point is sub-optimal.

Our approach is to model the canonical machine learning problem of setting the optimal operating point based on ROC in scenarios with a rational attacker as a two-player normal-form game. The goal of the defender is to detect a presence of an attack, for which she uses a publicly known classifier. The goal of the attacker is to design data samples (an attack) maximizing his benefit yet having a good chance of being undetected. The set of thresholds (or other parameters of the classifier) is the set of strategies for both players, as it is assumed the attacker can design data-samples not detected at a given threshold [4, 1].

We present, compute, and analyze two different solution concepts in this game. The first is the Nash equilibrium, which is the most standard solution concept for situations where players interact only once and decide about their strategies simultaneously. The second is the Stackelberg equilibrium, which has been recently very popular in security domains [21]. The latter assumes that one player, typically defender, computes its strategy and discloses it to the other player before the game is played. The other player (attacker) can then play optimally with respect to this strategy. This better describes the situation, when the classifier (detection system) is publicly known, as the attacker may even run his own copy of the classifier to verify undetectability of his attacks.

The main results of our analysis are that in adversarial setting, the defender can substantially reduce its expected cost by randomizing over a larger set of thresholds. We formally prove that in some games, no finite number of thresholds is sufficient for the optimal randomization, but a reasonably sparse discretization is often sufficient to guarantee strategies with performance close to the optimum.

Throughout the paper, we use a simple running example from the network security domain: The attacker tries to gain remote access to a server using a brute force password attack. The attacker knows that the defender deploys an intrusion detection system (IDS) and if she detects the attack, she will block attacker's IP address. The detector needs to decide how many passwords per second he will try. The defender has to decide how many login attempts per second is enough to manually inspect the incident. If she sets the threshold too high, the attacker has a good chance of succeeding in the attack. On the other hand, too low thresholds force her to inspect many false alarms caused by users who forgot their password.

2 Related Work

Only a few papers addressed operating point selection in game theoretic framework. Cavusoglu et al. [3] is one of the first papers advocating the importance of game theoretic models in configuration of detectors of malicious behavior. As in multiple similar papers, e.g.,[2], their models assume that the players can possibly make randomized decisions about whether to attack or whether to manually check a specific incident, but still allow choosing only a single fixed operating point as the optimal configuration based on an ROC curve. The main difference of our approach is that we propose randomization over multiple operating points, which determines strategies with respect to whole ROC curve resulting into lower defender's costs.

A game theoretic model of randomized threshold selection is presented in [8]. The rational attacker tries to hide its preferences by distributing his attacks between the preferred and nonpreferred target, while the defender sets a threshold for the number of attack attempts on the more valuable target. In contrast to our paper, this model is not connected to machine learning theory and general classifier characteristics, such as ROC. Furthermore, is requires a discrete set of thresholds and it analyses only Nash equilibria.

The use of Stackelberg equilibrium in our work has been inspired mainly by the recent progress in research and practical applications of resource allocation security games [21]. While there are several parallels, the class of games we study here is substantially different. The resource allocation games assume a specific utility structure, which causes the Nash and Stackelberg equilibria to prescribe the same strategies [12]. As we show in experiments, this is not always the case in our model. Also, there is no connection between these models and machine learning and all the studied models of resource allocation games are fundamentally discrete.

Recent works [4, 1, 10] from different domains show that for a fixed detector an attacker can devise an invisible attack just below the detection threshold. This paper uses the following generalization: against every detector from a set of detectors, an attacker can plant an invisible attack, providing he knows the detector. Since every detector has certain false positive and false negative rate, the set can be described by ROC curve, which can be parameterized by a single parameter – threshold (more on this in the next section). Hence, thresholds used in discussions of operating point selection serve here as an abstraction linking a single parameter to a particular classifiers from a possibly rich set. With respect to cited works on evasion attack, this simplification does not decrease generality of the presented approach.

3 Background

A two player *normal form game* is defined by a set of players $\mathcal{I}$; set of actions for each player $\mathcal{A}_i, i \in \mathcal{I}$; and utility functions $u_i : \mathcal{A} \to \mathbf{R}$ for each player and *action profile* from $\mathcal{A} = \Pi_{i\in\mathcal{I}}\mathcal{A}_i$. A *(mixed) strategy* of a player $\sigma_i \in \Sigma_i$ is a probability distribution over her actions and a *pure strategy* is a strategy playing only one of the actions. The utility functions can be extended to mixed strategies by taking expectation over players' randomization. For a strategy profile $\sigma \in \Sigma = \Pi_{i\in\mathcal{I}}\Sigma_i$, we denote σ_i the strategy of player i and σ_{-i} the strategy of the other player. A strategy profile σ^* is an *ϵ-Nash Equilibrium*

$$\text{if } u_i(\sigma_i, \sigma^*_{-i}) - u_i(\sigma^*) \le \epsilon \; \forall i \in \mathcal{I}, \sigma_i \in \Sigma_i.$$

A strategy profile is an exact Nash equilibrium (NE) if $\epsilon = 0$.

A *Stackelberg Equilibrium* (SE) assumes that one of the players is a leader who commits to a strategy and discloses it to the other player (termed follower). The other player then plays the action that maximizes her utility. For a two player game with leader i is (σ_i^*, a_{-i}^*) a SE if

$$\begin{aligned} u_i(\sigma_i^*, a_{-i}^*) &\geq u_i(\sigma_i, a_{-i}^*) \; \forall \sigma_i \in \Sigma_i \\ \& \; u_{-i}(\sigma_i^*, a_{-i}^*) &\geq u_{-i}(\sigma_i^*, a_{-i}) \; \forall a_{-i} \in \mathcal{A}_{-i}. \end{aligned}$$

The first line says that the leader does not have an incentive to change the strategy and the second line says the follower plays the best response to the leader's strategy. The *value* of the equilibrium for player i is $v_i = u_i(\sigma_i^*, a_{-i}^*)$. If the follower breaks ties in favor of the leader, it is a *Strong* Stackelberg Equilibrium (SSE). Breaking ties in favor of the leader is generally not a restrictive assumption, because minimal perturbation of any SE can ensure this choice is the only optimal for the follower [22].

Receiver Operating Characteristic (ROC) of a classifier is a parametric curve describing dependency of true positive rate (rate of successfully detected attacks) on the false positive rate (rate of benign events flagged as alarms). Each value of the parameter corresponds to a single point on the curve with specific true and false positive rates, which is also called operating point. Without loss of generality we assume the curve to be parameterized by a detection threshold $t \in \mathbf{T}$, but other parameterizations such as different penalties on error on different classes during training of the classifier are indeed possible. ROC curve is non-decreasing in the false positives rate, but we do not assume it to be necessarily concave.

In reality the operating point of a classifier can be controlled by more than one parameter, for example by varying costs of errors on different classes during training. Nevertheless, for every false positive rate the rational defender always chooses a classifier with the highest detection accuracy. Consequently a particular false positive rate is linked to a particular classifier which is in this paper abstracted by a threshold. By similar reasoning it can be assumed that the ROC curve is non-decreasing, because in the defender does not have any incentive to use classifier with higher false positive rate and smaller detection accuracy.

With respect to the above arguments it is assumed that there is a bijective decreasing mapping between false positive rate and the threshold, which means that the higher threshold implies smaller false positive rate. $R_{FP} : \mathbf{T} \rightarrow \langle 0, 1 \rangle$ maps thresholds to false positive rate.

4 Game Model

We formalize the operating point selection in presence of adversary as a two-player non-zero-sum normal form game with continuous strategy spaces.

Players: The two players in the game are the defender (denoted d) and the attacker (denoted a). In reality the game will be played between one defender and many different attackers, but since at this point we assume all attackers to share the same costs and penalties, they can be represented as a single player. We plan to generalize the model to the Bayesian game setting [18] in future work.

Actions: The action sets of the players are identical. Each player selects a threshold from a set $\mathbf{T}$, which can be mapped to $\langle 0,1 \rangle$ without loss of generality. If the defender selects a threshold $t_d \in \mathbf{T}$, all attacks stronger than this threshold are detected. If the attacker selects a threshold $t_a \in \mathbf{T}$, he plays an attack of maximal intensity undetected by the detector with threshold t_a. The attacker is detected if $t_a > t_d$.

Utility functions: The utility functions of the players depend on ROC curves, defender's costs for processing false positives and cost of missed detection, and attackers reward for successful attack and penalty for the attack being detected. Formal definitions of all quantities are following: $ROC : \langle 0,1 \rangle \to \langle 0,1 \rangle$ is the receiver operation characteristic of the classifier; $C^{FP} \in \mathbf{R}_0^+$ is the defender's cost of processing a false positive and $C^{FN} : \mathbf{T} \to \mathbf{R}$ is a non-decreasing defender's cost of missing an attack of certain intensity; $r_a : \mathbf{T} \to \mathbf{R}_0^+$ is the non-decreasing attacker's reward for performing an undetected attack and $p_a \in \mathbf{R}_0^+$ is the attacker's penalty for being detected while performing an attack.We allow the attacker to choose not to attack for zero reward and penalty. We further assume not all attackers being rational, as there is $A_r \in \mathbf{R}_0^+$ times more rational attackers than non-rational, who attack with the same intensity regardless of the classifier's setting. Strategies of irrational attackers are reflected in the true positives of the ROC.

In our running example, the attacker's reward $r_a(t)$ can be the number of passwords he tries per second without being detected; $C^{FN}(t)$ can be $c \cdot r_a(t)$ for some scaling factor c and p_a being the penalty the attacker suffers if his attack IP address is blocked. The notion of attack intensity in other domains could represent the entropy of attack sources in a DDoS attack, the amount of information injected to a media file in steganalysis, or the negative of distortion caused to the media file in watermarking.

Based on the inputs above, we define the defender's background cost for irrational attackers as the standard classification cost used in non-adversarial setting:

$$c_d^b(t) = R_{FP}(t) \cdot C^{FP} + (1 - ROC(t)) \cdot C^{FN}(t) \tag{1}$$

For rational attackers playing a threshold t the defender suffers an additional penalty for the undetected attacks:

$$c_d^r(t) = A_r \cdot C^{FN}(t). \tag{2}$$

The utility function of the defender is the negative of the sum of the background cost and the cost for rational attacks if undetected:

$$u_d(t_d, t_a) = \begin{cases} -c_d^b(t_d) - c_d^r(t_a) \text{ if } t_d \geq t_a \\ -c_d^b(t_d) \text{ otherwise.} \end{cases}$$

The utility of the attacker is his reward in case of being undetected and the negative penalty when he is detected:

$$u_a(t_d, t_a) = \begin{cases} r_a(t_a) \text{ if } t_d \geq t_a \\ -p_a \text{ otherwise.} \end{cases}$$

5 Game Model Properties

The ROC curves from real problems are usually estimated from data samples without clear analytical formulations. For this reason we base our study on a discretized version of the game, which means that optimal strategies are only approximated. We therefore first derive approximation bounds of Nash (NE) and Stackelberg equilibria (SSE) between discretized and continuous version of the problem. Then we show that even if we can get arbitrarily good approximations with finite sets of thresholds, creating the exact optimal randomized strategy may require using infinitely many thresholds. Finally, we show that some subsets of thresholds will never be used by a rational defender and can be disregarded in the strategy computation.

Proposition 1. *Let v_d be the value of SSE of the continous game for the defender; $(t_i) = t_0 < t_1 < \cdots < t_n$; $t_i \in \mathbf{T}$; $t_0 = \min(\mathbf{T})$; $t_n = \max(\mathbf{T})$ be a discretization of the set of applicable thresholds and*

$$\Delta = \max_{i \in \{0,\dots,n-1\}} \{\max\{c_d^r(t_{i+1}) - c_d^r(t_i), \max_{t \in (t_i, t_{i+1}\rangle} c_d^b(t) - \min_{t \in (t_i, t_{i+1}\rangle} c_d^b(t)\}\},$$

be the maximal difference between the highest and the lowest point in the defender's cost functions within one interval. Then there is a mixed strategy selecting only the thresholds from (t_i), such that its expected value for the defender is at least $v_d - 2\Delta$.

Proof. Assume (D, t_a) are the cumulative distribution function[1] (CDF) of the defender's strategy and threshold selection of the attacker in a SSE of the continuous game. Let $t_j \in (t_i)$ be a threshold in the discretization, such that $t_a \in (t_j, t_{j+1}\rangle$. We construct a CDF D' lower than D in the interval $(t_j, t_{j+1}\rangle$ and higher then D outside, so that the attacker still plays in $(t_j, t_{j+1}\rangle$ and the cost of the defender is not increased substantially.

$$D'(t) = \begin{cases} D(t_{i+1}) & \forall t \in (t_i, t_{i+1}\rangle \quad i \neq j \\ D(t_j) & \forall t \in (t_j, t_{j+1}\rangle \end{cases} \tag{3}$$

The expected utility of the attacker for playing threshold t in response to distribution D is

$$u_a(D, t) = (1 - D(t))r_a(t) - D(t)p_a \tag{4}$$

[1] Probability that randomly selected threshold is below the input parameter, i.e., $D(t_d) = P(t \leq t_d)$.

While CDFs and r_a are non-decreasing, the attacker's expected utility with D' cannot increase outside $(t_j, t_{j+1}\rangle$ and cannot decrease in the interval. Hence, he will keep playing to interval $(t_j, t_{j+1}\rangle$ and even if he modifies his strategy within this interval, it will not increase the costs of the defender by more than Δ in the c_d^r component of her utility by definition of Δ. Furthermore, any time the defender would play $t \in (t_i, t_{i+1}\rangle$ with distribution D, she plays one of the bounds instead with D'. For each of these bounds, she has the cost c_d^b at most Δ more than with t. In the worst case, the defender will suffer the increased cost in both components. □

Proposition 2. *Let (D, A) be CDFs of strategies in NE of the operating point selection game discretized to $(t_i) \subseteq \mathbf{T}$ and Δ defined as in Proposition 1, then (D, A) is a 2Δ-NE of the continuous game.*

Proof. Attacker: The attacker's expected utility cannot be increased by playing thresholds not included in the discretization. For any $t \in (t_i, t_{i+1})$ the attacker might consider, he can only improve his payoff by playing t_{i+1} instead. Recall that playing the same threshold as the defender results to an undetected attack. The probability of detection by D is the same on the whole interval and $r_a(t)$ is non-decreasing.

Defender: The best response to any mixed strategy can always be found in pure strategies. Assume that $t_d \in (t_i, t_{i+1})$ is the best response of the defender to the attacker's strategy A. From definition of Δ, $u_d(t_d, A) \leq u_d(t_i, A) + 2\Delta$, because it can differ by Δ in each component of the utility function. The defender has no incentive to deviate to t_i from a discrete NE strategy, because this threshold was considered in its computation; hence, $u_d(t_i, A) \leq u_d(D, A)$. Combining the two inequalities gives us $u_d(t_d, A) \leq u_d(D, A) + 2\Delta$. □

It is important to realize that Δ is not a parameter of the problem, but rather a guide for creating a suitable discretization. The goal is to select a discretization, such that Δ is small. Δ can even be selected in advance and then the algorithm to compute a matching discretization could just swipe through the interval of possible thresholds and add a new threshold to the discretization always when one of the relevant functions changes its value by more than Δ. For a function with range [0,1] and $\Delta = \frac{1}{n}$, a monotonic function will require at most n thresholds; convex/concave function at most $2n$ thresholds. If we want to guarantee less than 5% error from the optimum in the worst case, we can always choose 40 thresholds for the monotonous c_d^r function to keep $\Delta = 2.5\%$ for this component of its definition. If the c_d^b function is convex (which seems to be the case in the real world examples presented in our experiments), we will need at most additional 80 thresholds to guarantee even this component of definition of Δ to be 2.5%. Moreover, as we explain later, we can remove some portions of the thresholds completely form consideration.

The above discussion shows that the error in the quality of the produced solutions caused by discretization is bounded and we can always choose a relatively small discretization of set $\mathbf{T}$ that guarantees a low error. We further show that there are instances of the game, in which the optimal solution of the discretized version of the game is always worse than the optimal solution of the continuous game.

Proposition 3. *There are continuous operating point selection games, in which the optimal strategy requires the defender to use infinitely many thresholds.*

Proof. Let the mapping from thresholds $\langle 0,1\rangle$ to false positive rate be $R_{FP}(t) = (1-t)$; $ROC(t) = \min(2(1-t), 1)$; the misclassification costs $C^{FN}(t) = C^{FP} = 1$; and twice as much rational as background attackers ($A_r = 2$). Then

$$c_d^b(t) = 1 - t \text{ on } \langle 0, \tfrac{1}{2}\rangle \text{ and } t \text{ on } \langle \tfrac{1}{2}, 1\rangle \tag{5}$$

$$\text{and } c_d^r(t) = 2 \text{ for } t \in (0,1\rangle. \tag{6}$$

In this case, the rational defender prefers to prevent the rational attacker from attacking at all, even if it meant setting detection threshold to 0. $c_d^r(t_a)$ is always larger than $c_d^b(t_d)$ for any $t_a > 0$.

Assume the rational attacker's penalty $p_a = 1$ and reward $r_a(t) = 1 + t$. The attacker will not attack if $u_a(D, t) \leq 0$ for all $t \in \mathbf{T}$, because it can always get zero utility by not attacking. If we assume this is an equality, we can derive

$$D(t) = \frac{t+1}{t+2} \Rightarrow D(0) = \tfrac{1}{2}, D(\tfrac{1}{2}) = \tfrac{3}{5}. \tag{7}$$

If the rational attacker does not attack at all, the defender prefers to play $t = \frac{1}{2}$, as it minimizes c_d^b. Therefore, the optimal continuous strategy for this situation is to play D on $\langle 0, \frac{1}{2}\rangle$ and set $D(t) = 1$ for all larger thresholds. D is strictly increasing and c_d^b strictly decreasing on $\langle 0, \frac{1}{2}\rangle$. If the defender wants to prevent the rational attack with a discrete distribution, her CDF has to be larger or equal to D for each threshold. If it is lower, the attacker has positive utility for attacking. Hence, she plays $D'(t_j) = D(t_{j+1}); \forall t_j \in (t_i)$; i.e., threshold t_j with probability $\pi(t_j) = D(t_{j+1}) - D(t_j)$. Her cost $u_d(D', 0) = \sum_i \pi(t_i) c_d^b(t_i)$ can always be decreased by adding any new threshold in $\langle 0, \frac{1}{2}\rangle$. The monotonicity of the involved functions implies that for any new threshold $t_m \in (t_i, t_{i+1})$ holds

$$(D(t_{j+1}) - D(t_j))\, c_d^b(t_j) > (D(t_{j+1}) - D(t_m))\, c_d^b(t_j) + (D(t_m) - D(t_j))\, c_d^b(t_m).$$

□

Besides proving that it might not be possible to play optimally with a finite number of thresholds, the previous proposition also demonstrates how the model motivates the attacker to perform weaker attacks. The mechanism is the same even if it is beneficial only to reduce the attack strength and not to prevent it completely.

Based on the previous propositions, we can choose a discretization of set $\mathbf{T}$ that guarantees a low error. Below we show that considering only a subset of $\mathbf{T}$ for discretization is sufficient.

Proposition 4. *If* $t_d^* = \arg\min\{c_d^b(t) + c_d^r(t)\}$, *then a rational defender will never play a threshold* t *for which*

$$c_d^b(t) > c_d^b(t_d^*) + c_d^r(t_d^*)$$

Proof. Threshold t_d^* is the best pure strategy for the Stackelberg setting and the maximin strategy for the defender. The defender can always guarantee payoff at least $c_d^b(t_d^*) + c_d^r(t_d^*)$ for any strategy of the attacker. Hence, she will not play a threshold that certainly induces a higher cost. □

5.1 Concavity of ROC Curves

The machine learning literature often assumes that the ROC curves are concave [9]. If an ROC curve is not concave then there is a false positive rate b so that

$$ROC(b) < \frac{b-a}{c-a} ROC(c) + \frac{c-b}{c-a} ROC(a) \quad (8)$$

for some $a < b < c$. If we use, instead of the threshold corresponding to b (t_b), the threshold for c (t_c) with probability $\frac{b-a}{c-a}$ and the threshold for a (t_a) with probability $\frac{c-b}{c-a}$, then the expected false positive rate is still b, but the true positive rate is the right hand side in equation 8. This randomization creates a classifier with strictly better expected performance than the classifier described by the original ROC; therefore, all ROC curves are assumed to be concave.

We argue that this well-known procedure is correct only for the traditional settings without rational attackers. In their presence, playing the probability distribution on t_a and t_c is not strategically equivalent to playing t_b. Recall that playing t_b motivates the rational attacker to play t_b as well; however, playing the randomization over t_a and t_c will motivate the attacker to play either t_a or t_c (depending on his costs), but generally not to play t_b. The attacker playing one of t_a or t_c may induce a substantially different cost to the defender compared to playing t_b. Consequently the widely adopted procedure to make ROC curves convex is not applicable in presence of rational attackers, as it results into misrepresenting the actual costs for the defender. We are not aware of any existing work presenting a similar observation.

6 Experimental Evaluation

This section experimentally demonstrates that the proposed game-theoretic approach randomizing among multiple detector's operating points forces rational attackers to attack with lower intensity than with a single threshold optimized against non-rational attackers. Consequently, the expected cost of the classifier is reduced. The use-case is an intrusion detection system (IDS) where results of this paper can be readily applied. We show three kinds of experiments: (i) experiments with multiple thresholds showing the strategies computed for specific ROC curves and the cost reduction they provide; (ii) experiments with only two thresholds which provide better insight into the rationale behind the model; and (iii) an experiments varying attacker's penalty showing the effect of this parameter on computed strategies.

In our experiments, we use ROC curves of a real world IDS from [19]. Each ROC curve has 100 points representing thresholds used by the detector. As explained in the previous section, we do not assume ROC curves to be concave, because the stochastic concave envelope changes the solution space. We have chosen the cost of false positives to be $C^{FP} = 15$, to represent that typical IDS faces much more benign traffic than actual attacks; and the amount of rational attackers to be the same as the background attackers $A_r = 1$. We set the attacker's penalty $p_a = 2$, as the cost for having the IP address blocked. The attacker's reward is a linear function in terms of attack intensity, i.e., ith point on the ROC curve, in the order of increasing false positive rate, is assigned the reward $\frac{(100-i)}{100} \cdot 10$. Choosing lower attack strength is equivalent to choosing a lower threshold to attack. Note that if the defender lowers the detection threshold, then it increases its false positive rate. The graphs in this section show false positive rate increasing on x-axis from left to right to have the ROC curves in their standard form. In all these graphs, the threshold grows on the same axes from right to left.

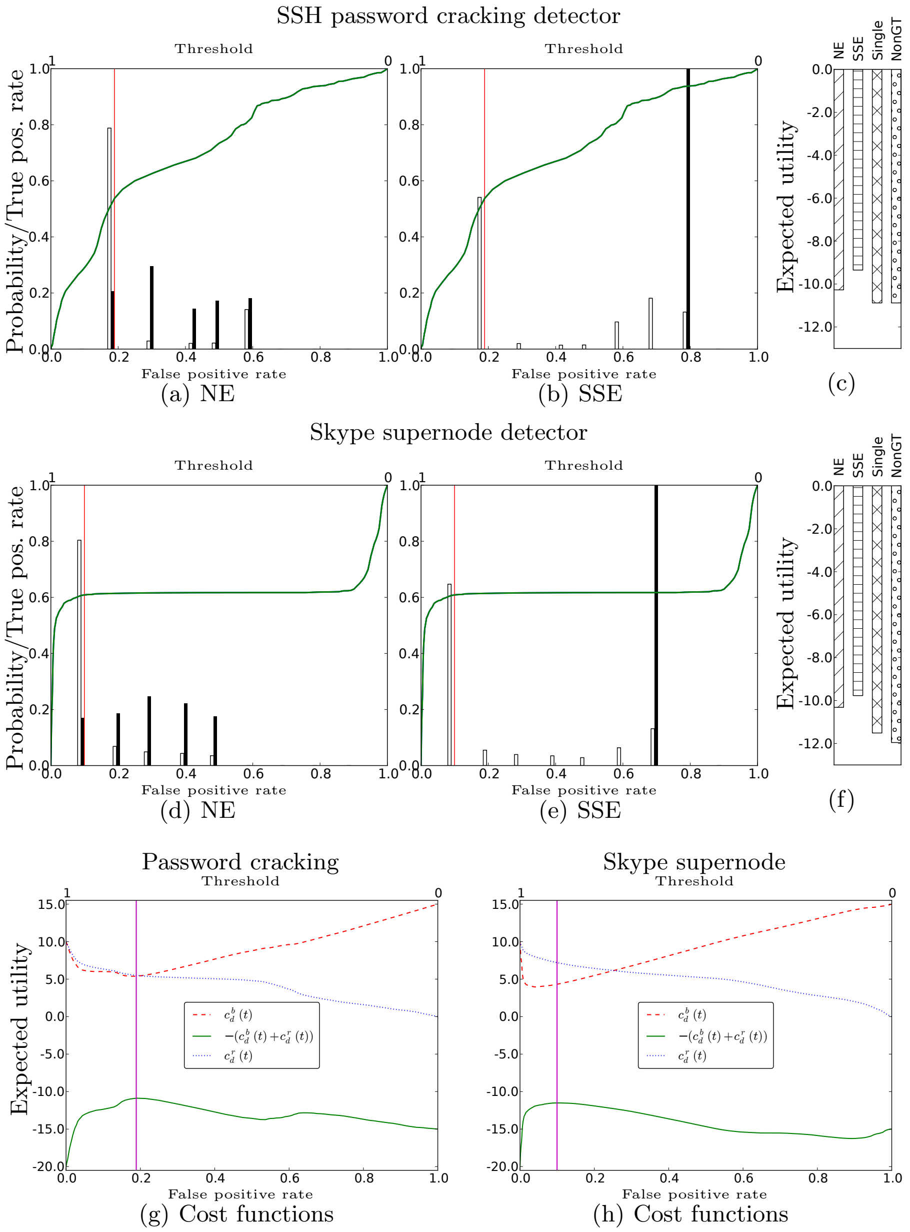
6.1 Computing the Equilibria and Scalability

Computing a NE may be computationally expensive as it belongs to the PPAD complexity class [7]. However, we do not reach the scalability limits of the standard equilibrium computation tools with our model. In our experiments, we use the Gambit [16] implementation of an algorithm for computing all Nash equilibria [15]. On standard Intel i5 2.3GHz laptop computer, the computation takes up to 1 minute for 12 thresholds, up to 5 minutes for 13 and up to an hour for 14 thresholds. Even though the algorithm computes all NE, it always found only a single NE in our experiments. It indicates that these games generally have a unique NE. We intend to formally study this property in our future research. In practical application, a more scalable algorithm computing only one NE can be used [14]. The gambit implementation of this algorithm is able to compute the strategy for 100 thresholds in less than one minute. Since the calculation of NE can be done off-line on a computer cluster, we do not consider the computational complexity here to be an issue.

Computing the SSE is a polynomial problem and we used the multiple linear programs (LP) method described in [5] with IBM CPLEX 12.4 as the LP solver. Computation of SEE even for 100 thresholds takes up to 10 seconds.

6.2 Multiple Thresholds

Figure 1 presents the results of the games defined based on ROCs of detectors of Secure Shell (SSH) password cracking and Skype supernodes. The games are discretized to contain 11 thresholds. Ten thresholds are chosen to be equidistant on the range of false positive values $[0, 1.0]$ and the last one is the optimal fixed threshold t_d^* defined in Proposition 4. This threshold is marked by the vertical lines in the graphs. White / black bars in Figures 1(a,b,d,e) correspond to the defender's / attacker's probability of selecting the threshold at their position. The curve in these figures is the ROC. In Figures 1(b,e) we can see that the attacker is forced to pick the lowest threshold played by the defender in the case



(a) NE (b) SSE (c)

(d) NE (e) SSE (f)

(g) Cost functions (h) Cost functions

Fig. 1. The top two rows show the main results for ROC curves corresponding to detection of SSH pasword cracking and Skype supernodes. Figures (a-b) and (d-e) show the probability distributions of using thresholds for the defender (white bars) and the attacker (black bars) in the Nash and Stackelberg equilibria of the game. Figures (c) and (f) show the expected utility of these strategies in comparison to the optimal fixed thresholds selection considering the rational opponent (Single) and the standard Bayesian threshold disregarding the rational opponent (NonGT). Figures (g) and (h) show the relevant cost functions and the optimal fixed thresholds for reference.

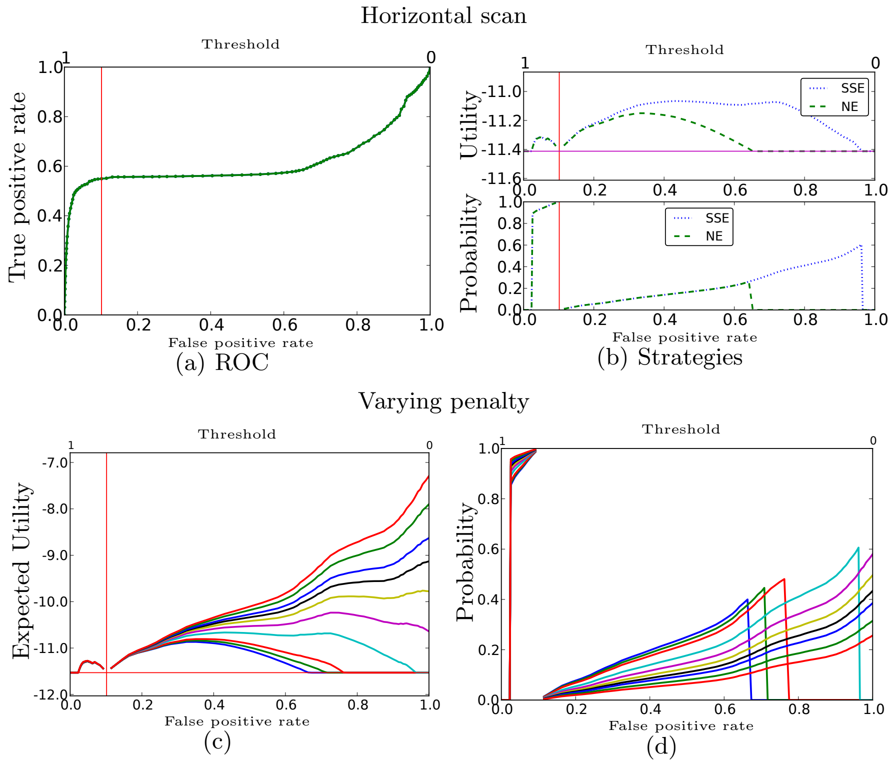
of Stackelberg equilibria (SSE). Although the defender plays most often the high threshold optimizing her background cost (54% of time), she plays the lower thresholds sufficiently often to force rational attackers to use weaker attacks. The threshold corresponding to the false positive rate just below 0.8 is played 13.5% of time. In these cases, the rational attacker uses less than half the attack intensity it would use without the randomization, i.e., the threshold marked by the vertical line. In the Nash equilibria presented in Figures 1(a,d), the attacker uses all the thresholds played by the defender with almost uniform probability. If the defender does not commit to a strategy in advance, the attacker also needs to randomize to prevent exploitation of his strategy by the defender. This is the main source of the defender's higher cost with NE compared to SSE.

Figures 1(c,f) present the expected costs of the Nash (NE) and Stackelberg (SSE) equilibrium strategies compared to the single threshold maximizing the utility defined as t^*_d in Proposition 4 (Single), and the standard Bayesian cost minimizing threshold disregarding the rational attackers (NonGT). The value of the SSE is better than the value of the NE. In both graphs, it is more than 10% better than the fixed operation point selection (t^*_d). The difference between the expected utility for fixed threshold selection considering and disregarding the rational attacker is quite low. Figures 1(g,h) present the utility function components computed based on the ROCs. The vertical line marks the optimal fixed game theoretic threshold (t^*_d) and the optimal threshold disregarding the rational opponents would be the minimum of $c^b_d(t)$.

Besides the results presented in this figure, we computed the expected utility values for 34 other ROC curves from [19]. The improvements of using multiple thresholds against the optimal fixed threshold (t^*_d) is between 5% and 20% (average 15%) for the Stackelberg equilibria and between 0.5% and 14% (average 9%) for the Nash equilibria.

6.3 Two Thresholds

Figure 2(a) shows the ROC curve of the horizontal scan detector we used for experiments with two thresholds. The first threshold is fixed in the optimal static thresholds selection t^*_d from Proposition 4 and the second varies over the x-axis of the graphs. Figure 2(b) presents the expected value of both equilibria and the probability of playing the threshold other than t^*_d if we optimally randomize only among these two thresholds. The vertical lines denote the position of the optimal fixed threshold and the horizontal line denotes its utility. The graph shows that substantial reduction of cost is possible already with two thresholds (top), and that even though values differ, the different equilibria suggest playing the same threshold with the same probability on large portion of possible thresholds (bottom). Furthermore, the probability of playing a low threshold (high false positive rate) quickly drops to zero at a point when it would no longer increase the defender's utility. Recall that the NE and SSE strategies overlap completely in well studied resource allocation security games [12]. Better understanding of when this happens in our model could enable reuse of many interesting results, such



(a) ROC

(b) Strategies

(c)

(d)

Fig. 2. Randomizing over two thresholds: first fixed at the optimal static threshold and the second varying on the x-axis. The graphs show (a) ROC curve; (b) upper: defender solution values, lower: second threshold probability; (c) the defender's SSE values for penalty for detected attack varying in the range $[0.1, 20]$ – curves from bottom to top; and (d) are the probabilities corresponding to (c) – curves from top to bottom.

as efficient computation of strategies for Bayesian games with different player types.

6.4 Varying Penalty

Figures 2(c,d) present the effect of attacker's penalty set to 0.1, 0.5, 1, 3, 5, 7, 9, 11, 15, 20 on the defender's payoff and probabilities in the scenario with two thresholds. We use the same setting as in the previous subsection, i.e., using one fixed thresholds t_d^* and changing the other threshold. Figure 2(c) shows that increasing the penalty increases the defender's payoff: the lines (from bottom up) represents the defender's utility with increasing penalty. Figure 2(d) with the probabilities of the alternative threshold selection shows that increasing the penalty decreases the probability of playing the second threshold: the lines (from top down) correspond to increasing attacker's penalty. At $p_a = 5$, detecting the IP address by the defender has so high penalty for the attacker that if the defender chooses the lowest threshold sufficiently often, the attacker stops attacking at all.

7 Conclusions

We analyze the problem of classifier operation point selection in presence of rational adversaries, applicable in various real world domains, such as network intrusion detection, spam filtering, steganalysis, or watermarking. We formalize it in game theoretic framework and focus on two well-known solution concepts: the more standard Nash equilibrium and the Stackelberg equilibrium commonly used in security domains. While it is not clear how to find these solutions exactly for the exact (continuous) games, we formally prove that we can create a discretized version of the game, which is solvable by standard techniques and its solution is a good approximation of the optimal solution of the original game.

We have experimentally evaluated the benefits of the model on a set of ROC curves originating from a real-world intrusion detection system. Using game theoretic randomization over multiple thresholds improves the defender's expected cost by up to 20% for some types of attacks, compared to using just single optimal threshold. This cost reduction is caused by the rational attacker selecting more than two times smaller attack strength in response to the randomization. While randomizing among larger number of thresholds is generally better, we show that substantial improvements can be achieved also by using only two different thresholds. We analyze this simplified case showing the main mechanisms by which the randomized strategies operate. Furthermore, we show that the Nash and Strong Stackelberg equilibrium strategies overlap on some subsets of threshold selections as in the resource allocation security games, but it is not true in general. This motivates more detailed study of the relation of these two models.

The future work on the proposed model may include generalization of the model to allow optimizing the thresholds against multiple different types of adversaries with different reward and cost functions. We would also like to generalize the model to allow multiple classifiers with correlated outputs and further analyze the relation to resource allocation games and other formal properties of the proposed model.

Acknowledgments. This work was supported by Ministry of the Interior of the Czech Republic, grant number VG20122014079. The work of T. Pevný on this paper was supported by the Grant Agency of Czech Republic under the project P103/12/P514.

References

1. Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim rndi, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and

Filip elezn, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pages 387–402. Springer Berlin Heidelberg, 2013.

2. Alvaro A Cárdenas, John S Baras, and Karl Seamon. A framework for the evaluation of intrusion detection systems. In *Security and Privacy, 2006 IEEE Symposium on*, pages 15–77. IEEE, 2006.
3. Huseyin Cavusoglu and Srinivasan Raghunathan. Configuration of detection software: A comparison of decision and game theory approaches. *Decision Analysis*, 1(3):131–148, 2004.
4. Pedro Comesana, Luis Pérez-Freire, and Fernando Pérez-González. Blind newton sensitivity attack. In *Information Security, IEE Proceedings*, volume 153, pages 115–125. IET, 2006.
5. Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 82–90. ACM, 2006.
6. Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. *Digital Watermarking and Steganography.* Cambridge University Press, 2008.
7. Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.
8. Lemonia Dritsoula, Patrick Loiseau, and John Musacchio. Computing the nash equilibria of intruder classification games. In *Decision and Game Theory for Security*, pages 78–97. Springer, 2012.
9. Peter A. Flach and Shaomin Wu. Repairing concavities in roc curves. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pages 702–707, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.
10. Prahlad Fogla and Wenke Lee. Evading network anomaly detection systems: Formal reasoning and practical techniques. In *Proceedings of the 13th ACM Conference on Computer and Communications Security*, CCS '06, pages 59–68, New York, NY, USA, 2006. ACM.
11. Jessica Fridrich. *Steganography in Digital Media: Principles, Algorithms, and Applications.* Cambridge University Press, 2009.
12. Dmytro Korzhyk, Zhengyu Yin, Christopher Kiekintveld, Vincent Conitzer, and Milind Tambe. Stackelberg vs. nash in security games: An extended investigation of interchangeability, equivalence, and uniqueness. *Journal of Artificial Intelligence Research*, 41(2):297–327, 2011.
13. Martin Kutter and Fabien AP Petitcolas. Fair benchmark for image watermarking systems. In *Electronic Imaging'99*, pages 226–239. International Society for Optics and Photonics, 1999.
14. Carlton E Lemke and Joseph T Howson, Jr. Equilibrium points of bimatrix games. *Journal of the Society for Industrial & Applied Mathematics*, 12(2):413–423, 1964.
15. Olvi L Mangasarian. Equilibrium points of bimatrix games. *Journal of the Society for Industrial & Applied Mathematics*, 12(4):778–780, 1964.
16. Richard D. McKelvey, Andrew M. McLennan, and Theodore L. Turocy. Gambit: Software tools for game theory, version 13.1.1. http://www.gambit-project.org, 2013.
17. Francisca Nonyelum Ogwueleka. Data mining application in credit-card fraud detection system. *Journal of Engineering Science and Technology*, 6(3):311–322, 2011.

18. Praveen Paruchuri, Jonathan P. Pearce, Janusz Marecki, Milind Tambe, Fernando Ordóñez, and Sarit Kraus. Efficient algorithms to solve bayesian stackelberg games for security applications. In *AAAI*, pages 1559–1562, 2008.
19. T. Pevny, M. Rehak, and M. Grill. Detecting anomalous network hosts by means of pca. In *Information Forensics and Security (WIFS), 2012 IEEE International Workshop on*, pages 103–108, Dec 2012.
20. Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Mach. Learn.*, 42(3):203–231, March 2001.
21. Milind Tambe. *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned.* Cambridge University Press, 2011.
22. Bernhard Von Stengel and Shmuel Zamir. Leadership with commitment to mixed strategies. Technical Report LSE-CDAM-2004-01, Centre for Discrete and Applicable Mathematics, London School of Economics and Political Science, 2004.