SECURITY OF PATTERN RECOGNITION SYSTEMS IN ADVERSARIAL ENVIRONMENTS

Battista Biggio, Giorgio Fumera, Gian Luca Marcialis, and Fabio Roli
Pattern Recognition and Applications Group
Department of Electrical and Electronic Engineering, University of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy
WWW: http://prag.diee.unica.it
{battista.biggio, fumera, marcialis, roli}@diee.unica.it

Abstract Pattern recognition and machine learning techniques are also used in *adversarial* settings, like biometric authentication, network intrusion detection, and spam filtering, in which intelligent and adaptive adversaries may manipulate data to undermine their operation. This behaviour raises three main open problems: (i) understanding potential vulnerabilities of pattern recognition techniques; (ii) evaluating their security, namely, the performance degradation under the corresponding attacks; and (iii) developing pattern recognition systems robust to attacks. In this work we summarize our contributions to the field, in particular, to the security evaluation of classifiers in adversarial settings. We shortly discuss the main aspects of a framework which we have recently proposed to address this issue, and present three examples in the above mentioned applications. Further, we briefly discuss our experimental findings related to the security of multimodal biometric systems, where fake biometric traits can be used to mislead user verification. These results were obtained in the context of the FP 7 European Project "Tabula Rasa", in which our research group is involved.

Keywords: adversarial classification; security evaluation; performance evaluation; spam filtering; network intrusion detection; biometric verification; multimodal biometric systems; spoofing attacks.

Introduction

Pattern recognition and machine learning techniques are also used in *adversarial* settings, like biometric authentication, network intrusion detection, and spam filtering, in which intelligent and adaptive adversaries may manipulate data to undermine their operation. As the presence of malicious adversaries is not taken into account by classical theory and design methods, pattern classification systems are likely to exhibit vulnerabilities that can be exploited by carefully targeted attacks. Only few works have addressed this issue so far, mainly in the field of machine learning (see, e.g., Laskov and Kloft, 2009; Barreno et al., 2010; Huang et al., 2011; Nelson et al., 2011; and the references therein). A greater effort is thus required to extend the pattern recognition theory and design methods to adversarial enviroments. In particular, three main open issues have to be systematically investigated: (i) understanding potential vulnerabilities of pattern recognition techniques; (ii) evaluating their security, namely, the performance degradation under the corresponding attacks; and (iii) developing pattern recognition systems robust to attacks. In this contribution, we summarize the main results attained in this field by the Pattern Recognition and Application (PRA) Group of the University of Cagliari.

Under the methodological viewpoint, we are developing a framework for *empirical security evaluation* of pattern classifiers under design, that formalizes and generalizes the main ideas proposed in the literature (Biggio et al., 2012b). Our framework is based on *simulating* attacks and evaluating the corresponding performance degradation, using data collected for classifier design. We give examples of the use of our framework in the three real applications mentioned above, which highlight that proper security evaluation can provide a more complete understanding of classifier behaviour in adversarial settings, and results in design choices superior to classical performance evaluation, in terms of the trade-off between accuracy (in absence of attacks) and robustness to attacks.

One of the applications of pattern recognition system in adversarial settings we are currently focusing on through resarch projects, is biometric identity verification. In particular, we are involved in the EU FP7 research project "Trusted biometrics under spoofing attacks" (Tabula Rasa). This project is aimed at investigating the security of biometric systems against "spoofing" attacks, which consist of presenting to the sensor a fake biometric trait that resembles the "live" trait of a genuine client, and claiming the corresponding genuine identity. The following biometrics are considered in the Tabula Rasa project: fingerprint, iris, gait, face, voice, EEG and ECG, as well as the multimodal score-level fusion of two or more of them. The project is articulated in three basic steps: assessing biometric system performance on benchmark data sets; assessing the performance degradation under spoofing attacks; designing software-based liveness detection methods for improving robustness to spoofing attacks. This project involves four academic partners, three research institutions and five private companies. The PRA Group is the leader of "Dissemination, exploitation, and standards" work package.

The first contribution of this work, also achieved thanks to some results of the Tabula Rasa project, consists of a thorough vulnerability analysis of multimodal biometric systems based on face and fingerprint traits, when they are subject to real spoofing attacks (Biggio et al., 2012a). This analysis was carried out on different data sets, using different spoofing techniques (e.g., using different materials to fabricate the fake traits). Another ongoing work consists of exploiting our framework for security evaluation (Biggio et al., 2012b) to develop methods for assessing robustness of biometric systems to spoofing attacks, during design, and without the need of fabricating any fake trait.

1. Security evaluation of pattern classifiers

In this section we summarize our work in Biggio et al., 2012b. We start from a short description of previous work dealing with performance evaluation of classifiers in security applications.

The evaluation of classifier performance in adversarial environments can not be done according to the classical paradigm of performance evaluation. The adversary may in fact manipulate training and/or testing data to achieve her goals (Huang et al., 2011), violating the stationarity assumption of classical performance evaluation. Further, it is not possible to know in advance how many and which kinds of attacks a classifier will be subject to, and, thus, how the data distribution will change. The solution used in all previous work is implicitly based on a well-known approach in the security field: the so-called *what-if* analysis (Rizzi, 2009). It consists of evaluating classifier performance under one or more *simulated* attack scenarios. To this end, many authors used empirical simulation techniques, based on specific attack simulation techniques and model of adversaries (Wittel and Wu, 2004; Globerson and Roweis, 2006; Fogla et al., 2006; Nelson et al., 2008; Cretu et al., 2008; Rodrigues et al., 2009; Kolcz and Teo, 2009; Rubinstein et al., 2009; Kloft and Laskov, 2010; Dekel et al., 2010; Barreno et al., 2010; Johnson et al., 2010; Biggio et al., 2011b; Biggio et al., 2011c). Basically, one or more attack scenarios of interest are defined, and the training and testing sets are modified accordingly; for example, a testing set of collected spam emails can be manipulated to simulate evasion attempts against spam filters during operation, making different assumptions on the adversary's knowledge and capabilities (Wittel and Wu, 2004; Globerson and Roweis, 2006; Kolcz and Teo, 2009; Biggio et al., 2011c). Classifier performance is then evaluated by using the modified training and testing sets. This allows one to assess the degradation of classifier performance when it processes samples explicitly targeted to evade the classifier. Clearly, the outcome of this process is different from the classical evaluation of the *generalization error* under the stationarity assumption. Various authors used accordingly the term *security* or *robustness* evaluation to denote it (Cardenas et al., 2006; Kolcz and Teo, 2009). Qualitatively, the less the performance degradation, the more secure (robust) the classifier.

Building on previous work, we propose a quantitative and general-purpose basis for the application of the *what-if analysis* to classifier security evaluation, based on the definition of potential attack scenarios. In particular, we define:

(i) a model of the adversary, that allows us to define any attack scenario; (ii) a model of the data distribution that allows us to implement the chosen attack scenario, keeping into account that training and testing data may follow different distributions; and (iii) a method for the creation of training and testing sets that are representative of the data distribution, and are used for empirical performance evaluation.

The definition of attack scenarios is ultimately an application-specific issue. Nevertheless, it is possible to give some general guidelines that can help the designer of a pattern recognition system. In Biggio et al., 2012b we propose to specify the attack scenario in terms of the adversary model and the attack strategy. Building on Laskov and Kloft, 2009; Huang et al., 2011 and on the taxonomy in Barreno et al., 2006; Huang et al., 2011, we define the conceptual model of the adversary in terms of her *goal*, *knowledge*, and *capability*. The *attack strategy* is then defined accordingly, by assuming that the adversary *rationally* chooses the optimal attack strategy. Details can be found in Biggio et al., 2012b.

Once the *attack scenario* is defined in terms of the adversary model and the resulting attack strategy, our framework proceeds with the definition of the corresponding data distribution, that is used to construct training and testing sets for security evaluation.

To this end, we exploit a data model that defines different training and testing data distributions, and a corresponding resampling algorithm. We do not describe them here due to lack of space. Details can be found in Biggio et al., 2012b.

2. Application examples

We consider here three different application examples of our framework in spam filtering, network intrusion detection, and biometric authentication. Our aim is to show how the designer of a pattern classifier can use our framework, and what kind of additional information he can obtain from security evaluation. Such information can be exploited for several purposes; for instance, to choose the best model among the ones under evaluation at the model selection phase, in terms of both classification accuracy and security.

Spam filtering

Assume that a classifier has to be designed to discriminate between legitimate and spam emails on the basis of their textual content, and that the *bag-of-words* feature representation has been chosen, where each feature is a Boolean random variable denoting whether a given word occurs in an email. This kind of classifier has been considered by several authors Drucker et al., 1999; Nelson et al., 2008; Kolcz and Teo, 2009, and it is included in several real spam filters, like SpamAssassin, Bogofilter, and SpamBayes.

In this example, we focus on the model selection phase. We assume that the designer wants to build a linear classifier, choosing between a support vector machine (SVM) with linear kernel, and a logistic regression (LR) classifier. He also wants to choose a feature subset, among all the words occurring in training emails. A set of legitimate and spam emails is available for this design phase. We assume that the designer wants to evaluate not only classifier accuracy in the absence of attacks, as in the classical design scenario, but also its security against the well-known bad word obfuscation (BWO) and good word insertion (GWI) attacks, which consist of modifying spam emails by inserting "good words" that are likely to appear in legitimate emails, and by obfuscating "bad words" that are typically present in spam Kolcz and Teo, 2009.

We assume that the adversary aims to get the highest percentage of spam emails misclassified as legitimate, and that she has perfect knowledge of the classifier, as done in Dalvi et al., 2004; Kolcz and Teo, 2009. Further, we assume that the adversary can modify up to a maximum number $n_{\max}$ of features (words) in each spam, as in Dalvi et al., 2004; Kolcz and Teo, 2009. This allows us to evaluate how gracefully the classifier performance degrades as an increasing number of words is modified, by repeating the security evaluation procedure for increasing values of $n_{\max}$.

Under the above assumptions, the goal of maximizing the percentage of misclassified spam emails can be attained by modifying up to $n_{\max}$ words in each spam, to cause the largest change in the discriminant function of the classifier, in order to classify the considered spam as legitimate. Accordingly, the most discriminant words are modified first, based on the weight assigned by the linear classifier. If a good (bad) word is missing, it will be inserted (obfuscated).

We report experiments on a subset of the TREC 2007 data set. In particular, we select the first 10,000 emails in chronological order to create the training (TR) set, and the next 10,000 emails to create the testing (TS) set. Each spam email in TS is modified according to the attack strategy, for each given $n_{\max}$ value. The features (words) are extracted from TR using the SpamAssassin tokenization method. Four feature subsets with size 1,000, 2,000, 10,000 and 20,000 have been selected using the information gain criterion Sebastiani, 2002. The AUC$_{10\%}$ is used as performance measure, since, in spam filtering, FP errors (i.e., misclassifying a legitimate email as spam) are much more harmful than false negative (FN) ones (Kolcz and Teo, 2009).

The above design choices (two classifiers, SVM and LR, and four feature subsets) lead to the evaluation of eight different classifier models. The AUC$_{10\%}$ value attained by each classifier on TS, is computed for different values of $n_{\max}$. In this case, it is a decreasing function of $n_{\max}$. The more graceful

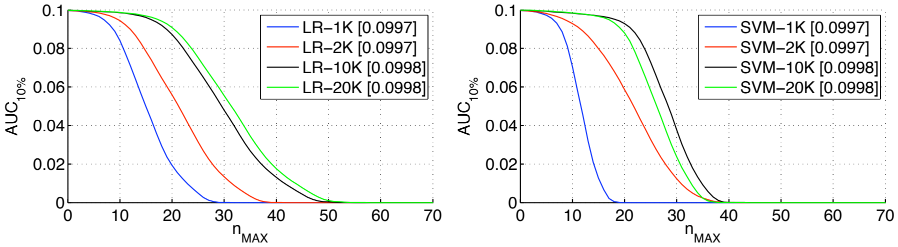

Figure 1. $\mathrm{AUC}_{10\%}$ attained on TS as a function of $n_{\max}$, for the LR (left) and SVM (right) classifier, with 1,000 (1K), 2,000 (2K), 10,000 (10K) and 20,000 (20K) features. The $\mathrm{AUC}_{10\%}$ value for $n_{\max} = 0$, corresponding to classical performance evaluation, is also reported in the legend between square brackets.

its decrease, the more robust the classifier is to the considered attack. Note that, for $n_{\max} = 0$, the spam emails in TS are not modified at all; thus, the corresponding $\mathrm{AUC}_{10\%}$ value corresponds to the one attained by classical performance evaluation methods.

The results are reported in Fig. 1. The $\mathrm{AUC}_{10\%}$ of each classifier drops to zero for $n_{\max}$ values between 30 and 50 (depending on the classifier): this means that all testing spam emails got misclassified as legitimate, after adding or obfuscating from 30 to 50 words.

The SVM and LR classifiers perform very similarly when they are not under attack (i.e., for $n_{\max} = 0$), regardless of the feature set size. Accordingly, they all appear identical under the viewpoint of "classical" classification performance, and the designer could choose any of the eight models. However, they exhibit a very different robustness to the considered attack, since their $\mathrm{AUC}_{10\%}$ value decreases at very different rates as $n_{\max}$ increases. In particular, the LR classifier with 20,000 features clearly outperforms all the other ones, for all $n_{\max}$ values. This result clearly suggests the designer a very different choice than the one coming from the classical performance evaluation.

Network intrusion detection

Intrusion detection systems analyse network traffic to prevent and detect malicious activities like intrusion attempts, port scans, and denial-of-service attacks. When suspected malicious traffic is detected, an alarm is raised by the IDS and subsequently handled by the system administrator. *Anomaly-based* detectors build a statistical model of the normal traffic using machine learning techniques, usually one-class classifiers (e.g., PAYL Wang and Stolfo, 2004), and raise an alarm when anomalous traffic is detected. Their training set is constructed, and periodically updated to follow the changes of normal traffic, by collecting unsupervised network traffic during system operation, assuming that it is normal. This kind of IDS may thus be vulnerable to attacks during the learning phase. In fact, an attacker may inject carefully designed malicious traffic during the collection of training samples, to force the IDS to learn a wrong model of the normal traffic Barreno et al., 2006; Laskov and Kloft, 2009; Kloft and Laskov, 2010; Rubinstein et al., 2009.

In this application example, we assume that an anomaly-based IDS is being designed, using a one-class ν-SVM classifier with radial-basis function (RBF) kernel and the feature vector representation proposed in Wang and Stolfo, 2004. Each network packet is considered as an individual sample to be labeled as normal (legitimate) or anomalous (malicious), and is represented as a 256-dimensional feature vector, defined as the histogram of byte frequencies in its payload (this is known as "1-gram" representation in the IDS literature). We then focus on the model selection stage. For the sake of simplicity, we assume that the parameter ν is set to 0.01 (to keep the FP rate lower than 1%), and that only the γ parameter of the RBF kernel has to be chosen.

We show how the IDS designer can select a model (the value of γ) based also on the evaluation of classifier security against an attack staged at the training phase, similar to the ones considered in Cardenas et al., 2006. In particular, we consider an attack aimed at forcing the learned model of normal traffic to include samples of intrusions to be attempted during operation. To this end, the attack samples should be carefully designed such that they resemble the statistical profile of the intrusions to be attempted at operation phase, without performing necessarily an intrusion. This can be simply obtained by shuffling the payload bytes of each intrusive testing packet.

We use here a subset of the data set of Perdisci et al., 2006. We initially set TR as the first 20,000 legitimate packets of day one, and TS as the first 20,000 legitimate samples of day two, plus all the malicious samples. We then consider different attack scenarios by increasing the number of attack samples in the training data, up to 40,000 samples in total, that corresponds to $p_{\max} = 0.5$.

Classifier performance is assessed using the $\mathrm{AUC}_{10\%}$ measure, for the same reasons as in the previous section. The performance under attack is evaluated as a function of $p_{\max}$, as in Laskov and Kloft, 2009; Nelson et al., 2008, which reduces to the classical performance evaluation when $p_{\max} = 0$. For the sake of simplicity, we consider only two values of the parameter γ, which clearly point out how design choices based only on classical performance evaluation methods can be unsuitable for adversarial environments.

The results are reported in Fig. 2. In the absence of attacks ($p_{\max} = 0$), the choice $\gamma = 0.5$ appears slightly better than $\gamma = 0.01$. Under attack, the performance for $\gamma = 0.01$ remains almost identical as the one without attack, and starts decreasing very slightly only when the percentage of attack samples in the training set exceeds 30%. On the contrary, for $\gamma = 0.5$ the performance

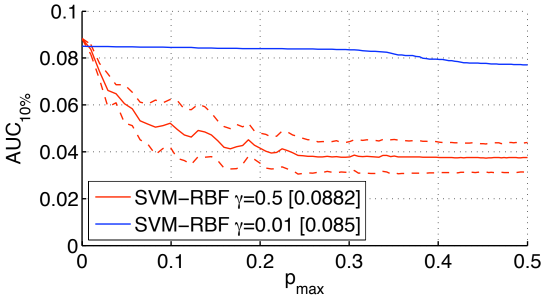

Figure 2. $\mathrm{AUC}_{10\%}$ as a function of $p_{\max}$, for the ν-SVMs with RBF kernel. The $\mathrm{AUC}_{10\%}$ value for $p_{\max} = 0$, corresponding to classical performance evaluation, is also reported in the legend between square brackets. The standard deviation (dashed lines) is reported only for $\gamma = 0.5$, since it was negligible for $\gamma = 0.01$.

suddenly drops as $p_{\max}$ increases, becoming lower than the one for $\gamma = 0.01$ when $p_{\max}$ is as small as about 1%. Accordingly, the choice of $\gamma = 0.5$, suggested by classical performance evaluation for $pmax = 0$, is clearly unsuitable from the viewpoint of classifier security under the considered attack. In fact, the designer should select $\gamma = 0.01$, trading a small decrease of classification accuracy in the absence of attacks for a higher security.

Biometric identity verification

Here, we only shortly describe the main results obtained by performing security evaluation of multimodal biometric verification systems under spoofing attacks. We refer the reader to Biggio et al., 2012a; Biggio et al., 2011a for further details, and experimental evidences.

First, our experimental findings confirmed that multimodal biometric systems can be misled with high probability by spoofing even a single biometric trait. This evidence was already highlighted in previous work (Rodrigues et al., 2009; Johnson et al., 2010), although without experimenting on real fake biometric traits. In fact, the matching score hypothetically assigned to each fake trait was assumed to be the same assigned to the corresponding live trait, under a worst-case scenario (perfect replication). Accordingly, the matching scores of fake traits were numerically simulated by resampling the genuine matching scores. Our experiments, involving face and fingerprint fakes, fabricated with several different materials, showed that, whereas in the case of fake faces it is possible to obtain matching score distributions very close to the genuine ones, this is much more difficult in the case of fake fingerprints. Thus, the worst-case assumption of perfect replication does not always hold, and may even poorly approximate the distributions of matching scores of fake traits. Shortly, it may be too pessimistic.

Second, our analysis also highlighted that the main techniques proposed thus far to mitigate the effect of spoofing attacks, designed under the perfect replication assumption, can lead to a higher vulnerability to *real* spoofing attacks. The reason is that, as mentioned above, such assumption may provide a too pessimistic approximation of real fake distributions. This raises the problem of providing more appropriate models for the distributions of spoofing attacks, in order to build more secure and reliable classification systems, which is currently part of our ongoing research.

Tabula Rasa. The experimental results discussed in this section were obtained in the context of the F7 Research Project "Tabula Rasa - Trusted biometrics under spoofing attacks", that addresses the issues raised by spoofing attacks against trusted biometric systems. This project thus not only involves companies in the security field but also emerging small and medium sized enterprises that wish to sell biometric technologies in emerging fields. In particular, the TABULA RASA project: (1) addresses the need for a draft set of standards to examine the problem of spoofing attacks; (2) proposes countermeasures such as combining biometric information from multiple sources; and (3) examines novel biometrics that may be inherently robust to spoofing attacks.

The first issue is addressed by analyzing the effectiveness of spoofing attacks on a number of biometrics, to provide insights on the vulnerability of each biometric trait. The second issue is explored in two lines: combining multiple biometric traits to build a single system that is robust to spoofing attacks; and investigating novel methods for liveness detection (namely, the discrimination between "live" and "fake" traits). The third issue is addressed by investigating novel biometrics that might be inherently robust to spoofing attacks, such as gait and vein or electro-physiological signals.

The project considers several biometrics, including fingerprint, iris, gait, face (2d and 3d representations), voice, EEG and ECG; and multimodal score-level fusion of different biometrics. Its basic steps are: the assessment of the average performance of several biometric systems on benchmark data sets; the parallel fusion of multiple biometrics through several fusion rules; the assessment of robustness of the same systems to spoofing attacks; and the design of countermeasures for improving robustness of unimodal and multimodal systems by software-based liveness detection methods. The first step has been completed and the second one is nearly finishing. Besides the above steps, a novel proof-of-concept system integrating the countermeasures proposed in the project will be developed.

Twelve partners are involved in the FP7 Tabula Rasa project: four academic partners, three research institutes (one of them has the leadership of the project), and five private companies. This witnesses the transversal interest about this relevant research topic. Among partners, our research group leads the "Dissemination, exploitation, and standards" Work Package (WP), explicitly devoted to the individuation of targets and actions for disseminating and exploiting the research project results at best. With regard to this WP, we already released a "Dissemination plan", with the list of targets and actions to be performed, and an "Exploitation plan", with the help of private partners, in order to bring to the market the main achievements of the project.

3. Conclusions and future work

In this paper, we reported our main results related to the security of pattern recognition systems in adversarial settings, such as spam filtering, network intrusion detection, and biometric authentication, in which malicious and intelligent adversaries manipulate data and their behaviour to mislead classification. We discussed the main methodological issues related to performance and security evaluation of such systems, which we addressed in a recently submitted work (Biggio et al., 2012b), currently under review, and provide three examples of use of our framework in real applications, simulating potential attacks. We also discussed some recent results obtained in the context of the Tabula Rasa project, and published in Biggio et al., 2012a; Biggio et al., 2011a, related to the vulnerability of multimodal biometric systems to spoofing attacks, highlighting novel open problems raised from our extensive experimental analysis, such as providing more reliable models for the probability distributions of matching scores of fake traits (spoofing attacks).

References

Barreno, M., Nelson, B., Joseph, A., and Tygar, J. (2010). The security of machine learning. *Machine Learning*, 81:121–148. 10.1007/s10994-010-5188-5.

Barreno, M., Nelson, B., Sears, R., Joseph, A. D., and Tygar, J. D. (2006). Can machine learning be secure? In *ASIACCS '06: Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, pages 16–25, New York, NY, USA. ACM.

Biggio, B., Akhtar, Z., Fumera, G., Marcialis, G., and Roli, F. (2011a). Robustness of multimodal biometric verification systems under realistic spoofing attacks. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–6.

Biggio, B., Akhtar, Z., Fumera, G., Marcialis, G. L., and Roli, F. (2012a). Security evaluation of biometric authentication systems under real spoofing attacks. *IET Biometrics*, 1(1):11–24.

Biggio, B., Corona, I., Fumera, G., Giacinto, G., and Roli, F. (2011b). Bagging classifiers for fighting poisoning attacks in adversarial environments. In Sansone, C., Kittler, J., and Roli, F., editors, *10th International Workshop on Multiple Classifier Systems (MCS)*, volume 6713 of *Lecture Notes in Computer Science*, pages 350–359. Springer-Verlag.

Biggio, B., Fumera, G., and Roli, F. (2011c). Design of robust classifiers for adversarial environments. In *IEEE Int'l Conf. on Systems, Man, and Cybernetics (SMC)*, pages 977–982.

Biggio, B., Fumera, G., and Roli, F. (2012b). Security evaluation of pattern classifiers under attack. *Submitted to TPAMI (under review)*.

Cardenas, A. A., Baras, J. S., and Seamon, K. (2006). A framework for the evaluation of intrusion detection systems. In *SP '06: Proceedings of the 2006 IEEE Symposium on Security and Privacy*, pages 63–77, Washington, DC, USA. IEEE Computer Society.

Cretu, G. F., Stavrou, A., Locasto, M. E., Stolfo, S. J., and Keromytis, A. D. (2008). Casting out demons: Sanitizing training data for anomaly sensors. In *IEEE Symposium on Security and Privacy*, pages 81–95, Los Alamitos, CA, USA. IEEE Computer Society.

Dalvi, N., Domingos, P., Mausam, Sanghai, S., and Verma, D. (2004). Adversarial classification. In *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 99–108, Seattle.

Dekel, O., Shamir, O., and Xiao, L. (2010). Learning to classify with missing and corrupted features. *Machine Learning*, 81:149–178. 10.1007/s10994-009-5124-8.

Drucker, H., Wu, D., and Vapnik, V. N. (1999). Support vector machines for spam categorization. *IEEE Transaction on Neural Networks*, 10(5):1048–1054.

Fogla, P., Sharif, M., Perdisci, R., Kolesnikov, O., and Lee, W. (2006). Polymorphic blending attacks. In *USENIX-SS'06: Proceedings of the 15th conference on USENIX Security Symposium*, Berkeley, CA, USA. USENIX Association.

Globerson, A. and Roweis, S. T. (2006). Nightmare at test time: robust learning by feature deletion. In Cohen, W. W. and Moore, A., editors, *ICML*, volume 148 of *ACM International Conference Proceeding Series*, pages 353–360. ACM.

Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B., and Tygar, J. D. (2011). Adversarial machine learning. In *4th ACM Workshop on Artificial Intelligence and Security (AISec 2011)*, pages 43–57, Chicago, IL, USA.

Johnson, P., Tan, B., and Schuckers, S. (2010). Multimodal fusion vulnerability to non-zero effort (spoof) imposters. In *Information Forensics and Security (WIFS), 2010 IEEE International Workshop on*, pages 1–5.

Kloft, M. and Laskov, P. (2010). Online anomaly detection under adversarial impact. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 405–412.

Kolcz, A. and Teo, C. H. (2009). Feature weighting for improved classifier robustness. In *Sixth Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, USA.

Laskov, P. and Kloft, M. (2009). A framework for quantitative security analysis of machine learning. In *AISec '09: Proceedings of the 2nd ACM workshop on Security and artificial intelligence*, pages 1–4, New York, NY, USA. ACM.

Nelson, B., Barreno, M., Chi, F. J., Joseph, A. D., Rubinstein, B. I. P., Saini, U., Sutton, C., Tygar, J. D., and Xia, K. (2008). Exploiting machine learning to subvert your spam filter. In *LEET'08: Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, pages 1–9, Berkeley, CA, USA. USENIX Association.

Nelson, B., Biggio, B., and Laskov, P. (2011). Understanding the risk factors of learning in adversarial environments. In *4th ACM Workshop on Artificial Intelligence and Security*, AISec '11, pages 87–92, Chicago, IL, USA.

Perdisci, R., Gu, G., and Lee, W. (2006). Using an ensemble of one-class svm classifiers to harden payload-based anomaly detection systems. In *International Conference on Data Mining (ICDM)*, pages 488–498. IEEE Computer Society.

Rizzi, S. (2009). What-if analysis. *Encyclopedia of Database Systems*, pages 3525–3529.

Rodrigues, R. N., Ling, L. L., and Govindaraju, V. (2009). Robustness of multimodal biometric fusion methods against spoof attacks. *J. Vis. Lang. Comput.*, 20(3):169–179.

Rubinstein, B. I., Nelson, B., Huang, L., Joseph, A. D., Lau, S.-h., Rao, S., Taft, N., and Tygar, J. D. (2009). Antidote: understanding and defending against poisoning of anomaly detectors. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, IMC '09, pages 1–14, New York, NY, USA. ACM.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34:1–47.

Wang, K. and Stolfo, S. J. (2004). Anomalous payload-based network intrusion detection. In Jonsson, E., Valdes, A., and Almgren, M., editors, *RAID*, volume 3224 of *Lecture Notes in Computer Science*, pages 203–222. Springer.

Wittel, G. L. and Wu, S. F. (2004). On attacking statistical spam filters. In *First Conference on Email and Anti-Spam (CEAS)*, Microsoft Research Silicon Valley, Mountain View, California.