International Journal of Applied Sciences, Engineering and Management
ISSN 2320 – 3439, Vol. 04, No. 01, January 2015, pp. 46 – 50

Security Evaluation of Pattern Classifiers under Attack

P. Ramesh[1], M. Siva Anjaneyulu[2]

[1]M. Tech Student, Dept of CSE, Amrita Sai Institute of Science and Technology, Paritala, Krishna-521180.
[2]Assistant Professor, Dept of CSE, Amrita Sai Institute of Science and Technology, Paritala, Krishna-521180.

Abstract: Pattern classification systems are normally worn in adversarial applications, similar to biometric confirmation, system imposition uncovering, and spam filtering, in which data can exist intentionally control by humans to destabilize their process. As this adversarial situation is not engaged into explanation by traditional intend methods, pattern classification systems may display vulnerabilities, whose utilization may harshly influence their presentation, and subsequently limit their realistic usefulness. Extending pattern classification theory and intend methods to adversarial settings is thus a narrative and very applicable explore bearing, which has not yet been pursued in a methodical way. In this paper, we deal with one of the major release issues: evaluating at propose phase the sanctuary of example classifiers, namely, the routine degradation below possible attacks they may acquire throughout operation. We suggest a structure for experiential estimate of classifier sanctuary that formalizes and generalizes the major information planned in the creative writing, and give examples of its use in three real applications. Reported results show that security evaluation can provide a more absolute sympathetic of the classifier's performance in adversarial environments, and lead to enhanced intend choices.

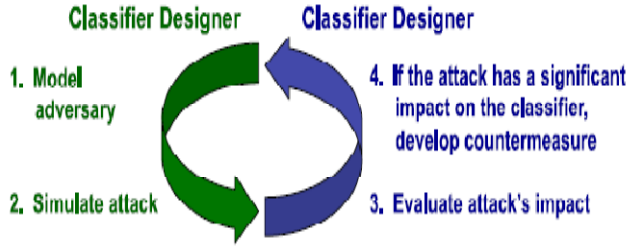
Index Terms*: Pattern classification, adversarial classification, performance evaluation, security evaluation, robustness evaluation.*

INTRODUCTION

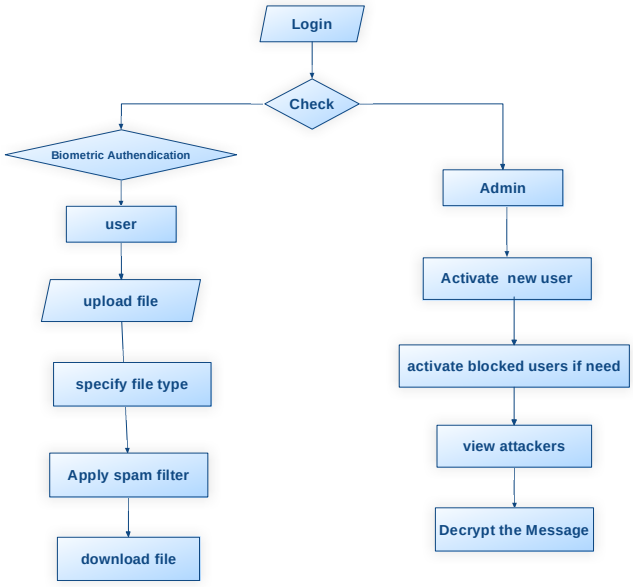
Pattern classification systems based on mechanism learning algorithms are normally used in safety related applications like biometric authentication, network interruption detection, and spam filtering, to differentiate among a "legitimate" and a "malicious" prototype class (e.g., legitimate and spam emails). Opposing to conventional ones, these applications have a basic adversarial natural history since the input data can be intentionally manipulated by a bright and adaptive challenger to weaken classifier operation. This repeatedly gives rise to an arm race connecting the opponent and the classifier designer. Well known examples of attacks next to outline classifiers are: submitting a false biometric trait to a biometric validation system (spoofing attack) modifying complex packets belonging to disturbing traffic to evade interference detection systems; manipulating the content of spam emails to get them past spam filters (e.g., by misspelling common spam words to avoid their detection). Adversarial scenarios can also occur in intelligent data analysis and information retrieval; e.g., a malicious webmaster may manipulate search engine rankings to artificially promote her1 web site. It is now acknowledged that, since pattern classification systems based on classical theory and design methods do not take into account adversarial settings, they exhibit vulnerabilities to several potential attacks, allowing adversaries to undermine their effectiveness. A systematic and unified treatment of this issue is thus needed to allow the trusted adoption of pattern classifiers in adversarial environments, starting from the theoretical foundations up to novel design methods, extending the classical design cycle of. In particular, three main open issues can be identified: (i) analyzing the vulnerabilities of classification algorithms, and the corresponding attacks. (ii) developing novel methods to assess classifier security against these attacks, which is not possible using classical performance evaluation methods (iii) developing novel design methods to guarantee classifier security in adversarial environments. Although this emerging field is attracting growing interest the above issues have only been sparsely addressed under different perspectives and to a limited extent. Most of the work has focused on application-specific issues related to spam filtering and network intrusion detection, e.g., while only a few theoretical models of adversarial classification problems have been proposed in the machine learning literature however, they do not yet provide practical guidelines and tools for designers of pattern recognition systems. Besides introducing these issues to the pattern recognition research community, in this work we address issues (i) and (ii) above by developing a framework for the empirical evaluation of classifier security at design phase that extends the model selection and performance evaluation steps of the classical design cycle. In Sect. 2 we summarize previous work, and point out three main ideas that emerge from it. We then formalize and generalize them in our framework. First, to pursue security in the context of an arms race it is not IEEE Transactions on Knowledge and Data Engineering sufficient to react to observed attacks, but it is also necessary to proactively anticipate the adversary by predicting the most relevant, potential attacks through a what-if analysis; this allows one to develop suitable countermeasures before the attack actually occurs, according to the principle of security by design. Second, to provide practical guidelines for simulating realistic attack scenarios, we define a general model of the adversary, in terms of her goal, knowledge, and capability, which encompasses and generalizes models proposed in previous work. Third, since the presence of carefully targeted attacks may affect the distribution of training and testing data separately, we propose a model of the data distribution that can formally characterize this behavior, and that allows us to take into account a large number of potential attacks; we also propose an algorithm for the generation of training and testing sets to be used for security evaluation, which can naturally accommodate application-specific and heuristic techniques for simulating attacks. In we give three concrete examples of applications of our framework in spam filtering, biometric authentication, and network intrusion

detection., we discuss how the classical design cycle of pattern classifiers should be revised to take security into account. Finally, we summarize our contributions, the limitations of our framework, and some open issues.

SYSTEM ARCHITECTURE



1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.



SYSTEM ANALYSIS

EXISTING SYSTEM: Pattern classification systems based on classical theory and design methods do not take into account adversarial settings; they exhibit vulnerabilities to several potential attacks, allowing adversaries to undermine their effectiveness. A systematic and unified treatment of this issue is thus needed to allow the trusted adoption of pattern classifiers in adversarial environments, starting from the theoretical foundations up to novel design methods, extending the classical design cycle of. In particular, three main open issues can be identified: (i) analyze the vulnerabilities of classification algorithms, and the corresponding attacks. (ii) Developing novel methods to assess classifier security against these attacks, which are not possible using classical performance evaluation methods. (iii) Developing novel design methods to guarantee classifier security in adversarial environments. Poor analyzing the vulnerabilities of classification algorithms, and the corresponding attacks. A malicious webmaster may manipulate search engine rankings to artificially promote website.

PROPOSED SYSTEM: In this exertion we deal with issues above by initial a support for the experiential assessment of classifier defence at design phase that extends the model collection and presentation assessment steps of the traditional intend cycle .We review preceding work, and point out three main ideas that come out from it. We then celebrate and generalize them in our structure. First, to follow refuge in the situation of an artillery race it is not enough to react to experimental attacks, but it is also required to proactively await the opposition by predicting the most related, impending attacks from side to side a what-if breakdown; this allows one to enlarge proper countermeasures ahead of the do violence to actually occurs, according to the attitude of safety by devise. Second, to present no-nonsense plan for simulating reasonable bother scenarios, we define a common model of the antagonist, in terms of her goal, knowledge, and capability, which encompass and generalize models proposed in previous work. Third, since the being there of suspiciously targeted attacks may involve the allotment of guidance and testing data individually, we recommend a model of the data allocation that can properly distinguish this activities, and that allows us to take into account a large number of potential attacks; we also propose an algorithm for the generation of instruction and testing sets to be worn for security evaluation, which can logically contain application-specific and heuristic techniques for simulating attacks. Planned system prevents initial novel methods to assess classifier sanctuary against these attacks. The number present of an quick and adaptive adversary makes the classification problem highly non-stationary.

RELATED WORK

We address the security of multimodal biometric systems when one of the modes is successfully spoofed. We propose two novel fusion schemes that can increase the security of multimodal biometric systems. The first is an extension of the likelihood ratio based fusion scheme and the other uses fuzzy logic. Besides the matching score and sample quality score, our proposed fusion schemes also take into account the intrinsic security of each biometric system being fused.

Experimental results have shown that the proposed methods are more robust against spoof attacks when compared with traditional fusion methods

In biometric systems, the threat of "spoofing", where an imposter will fake a biometric trait, has lead to the increased use of multimodal biometric systems. It is assumed that an imposter must spoof all modalities in the system to be accepted. This paper looks at the cases where some but not all modalities are spoofed. The contribution of this paper is to outline a method for assessment of multimodal systems and underlying fusion algorithms. The framework for this method is described and experiments are conducted on a multimodal database of face, iris, and fingerprint match scores.

A very effective means to evade signature-based intrusion detection systems (IDS) is to employ polymorphic techniques to generate attack instances that do not share a fixed signature. Anomaly-based intrusion detection systems provide good defense because existing polymorphic techniques can make the attack instances look different from each other, but cannot make them look like normal. In this paper we introduce a new class of polymorphic attacks, called polymorphic blending attacks, that can effectively evade byte frequency-based network anomaly IDS by carefully matching the statistics of the mutated attack instances to the normal profiles. The proposed polymorphic blending attacks can be viewed as a subclass of the mimicry attacks. We take a systematic approach to the problem and formally describe the algorithms and steps required to carry out such attacks. We not only show that such attacks are feasible but also analyze the hardness of evasion under different circumstances. We present detailed techniques using PAYL, a byte frequency-based anomaly IDS, as a case study and demonstrate that these attacks are indeed feasible. We also provide some insight into possible countermeasures that can be used as defense.

The efforts of anti-spammers and spammers has often been described as an arms race. As we devise new ways to stem the flood of bulk mail, spammers respond by working their way around the new mechanisms. Their attempts to bypass spam filters illustrates this struggle. Spammers have tried many things from using HTML layout tricks, letter substitution, to adding random data. While at times their attacks are clever, they have yet to work strongly against the statistical nature that drives many filtering systems. The challenges in successfully developing such an attack are great as the variety of filtering systems makes it less likely that a single attack can work against all of them. Here, we examine the general attack methods spammers use, along with challenges faced by developers and spammers. We also demonstrate an attack that, while easy to implement, attempts to more strongly work against the statistical nature behind filters.

Unsolicited commercial email is a significant problem for users and providers of email services. While statistical spam filters have proven useful, senders of spam are learning to bypass these filters by systematically modifying their email messages. In a good word attack, one of the most common techniques, a spammer modifies a spam message by inserting or appending words indicative of legitimate email. In this paper, we describe and evaluate the effectiveness of active and passive good word attacks against two types of statistical spam filters: naive Bayes and maximum entropy filters. We find that in passive attacks without any filter feedback, an attacker can get 50 % of currently blocked spam past either filter by adding 150 words or fewer. In active attacks allowing test queries to the target filter, 30 words will get half of blocked spam past either filter.

IMPLEMENTATION

Attack Scenario and Model of the Adversary:

Although the definition of attack scenarios is ultimately an application-specific issue, it is possible to give general guidelines that can help the designer of a pattern recognition system. Here we propose to specify the attack scenario in terms of a conceptual model of the adversary that encompasses, unifies, and extends different ideas from previous work. Our model is based on the assumption that the adversary acts rationally to attain a given goal, according to her knowledge of the classifier, and her capability of manipulating data. This allows one to derive the corresponding optimal attack strategy.

Pattern Classification:

Multimodal biometric systems for personal identity recognition have received great interest in the past few years. It has been shown that combining information coming from different biometric traits can overcome the limits and the weaknesses inherent in every individual biometric, resulting in a higher accuracy. Moreover, it is commonly believed that multimodal systems also improve security against Spoofing attacks, which consist of claiming a false identity and submitting at least one fake biometric trait to the system (e.g., a "gummy" fingerprint or a photograph of a user's face). The reason is that, to evade multimodal system, one expects that the adversary should spoof all the corresponding biometric traits. In this application example, we show how the designer of a multimodal system can verify if this hypothesis holds, before deploying the system, by simulating spoofing attacks against each of the matchers.

Adversarial classification:

Assume that a classifier has 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, with binary features denoting the occurrence of a given set of words

Security modules:

Intrusion detection systems analyze network traffic to prevent and detect malicious activities like intrusion attempts, ROC curves of the considered multimodal biometric system under a simulated spoof attack against the fingerprint or the face matcher. 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. Two main kinds of IDSs exist: misuse detectors and anomaly-based ones. Misuse detectors match the analyzed network traffic against

a database of signatures of known malicious activities. The main drawback is that they are not able to detect never-before-seen malicious activities, or even variants of known ones. To overcome this issue, anomaly-based detectors have been proposed. They build a statistical model of the normal traffic using machine learning techniques, usually one-class classifiers, 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 operation, assuming that it is normal (it can be filtered by a misuse detector, and should).

INPUT DESIGN AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things: What data should be given as input? How the data should be arranged or coded? The dialog to guide the operating personnel in providing input.? Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.2. Select methods for presenting information.3. Create document, report, or other formats that contain information produced by the system. The output form of an information system should accomplish one or more of the following objectives. Convey information about past activities, current status or projections of the Future. Signal important events, opportunities, problems, or warnings. Trigger an action. Confirm an action.

CONCLUSION

In this paper we discussed on experiential security evaluation of pattern classifiers that have to be deployed in adversarial environments, and proposed how to change the conventional concert valuation design step, which is not appropriate for this purpose. Our main giving is a construction for experimental security evaluation that formalizes and generalizes ideas from before work, and can be functional to different classifiers, learning algorithms, and classification tasks. It is stranded on a proper model of the challenger, and on a model of data allotment that can symbolize all the attacks measured in before works; provides a systematic method for the creation of direction and difficult sets with the purpose of enables safety evaluation; and can provide accommodation application-specific techniques for attack simulation. This is a clear advancement with respect to previous work, since without a wide-ranging structure most of the proposed techniques (often tailored to a given classifier model, attack, and application) could not be directly applied to other problems. An inherent restraint of our work is that security evaluation is carried out empirically, and it is thus data dependent; on the other hand, model-driven analyses require a full investigative model of the difficulty and of the adversary's behavior that may be very not easy to expand for real-world applications. Another inherent limitation is due to fact that our method is not application-specific, and, therefore, provides only high-level guidelines for simulating attacks. Indeed, detailed guidelines require one to take into account application-specific constraints and opponent models. Our future work will be devoted to develop techniques for simulating attacks for different applications. Although the design of secure classifiers is a different crisis than security estimate, our framework could be also exploited to this end. For instance, simulated attack samples can be included into the training data to look up security of discriminative classifiers (e.g., SVMs), even as the planned data model can be broken to design more secure generative classifiers. We obtained encouraging introduction consequences on this topic.

REFERENCES

[1] R.N. Rodrigues, L.L. Ling, and V. Govindaraju, "Robustness of Multimodal Biometric Fusion Methods against Spoof Attacks," J. Visual Languages and Computing, vol. 20, no. 3, pp. 169-179, 2009.

[2] P. Johnson, B. Tan, and S. Schuckers, "Multimodal Fusion Vulnerability to Non-Zero Effort (Spoof) Imposters," Proc. IEEE Int'l Workshop Information Forensics and Security, pp. 1-5, 2010.

[3] P. Fogla, M. Sharif, R. Perdisci, O. Kolesnikov, and W. Lee, "Polymorphic Blending Attacks," Proc. 15th Conf. USENIX Security Symp., 2006.

[4] G.L. Wittel and S.F. Wu, "On Attacking Statistical Spam Filters," Proc. First Conf. Email and Anti-Spam, 2004.

[5] D. Lowd and C. Meek, "Good Word Attacks on Statistical Spam Filters," Proc. Second Conf. Email and Anti-Spam, 2005.

[6] A. Kolcz and C.H. Teo, "Feature Weighting for Improved Classifier Robustness," Proc. Sixth Conf. Email and Anti-Spam, 2009.

[7] D.B. Skillicorn, "Adversarial Knowledge Discovery," IEEE Intelligent Systems, vol. 24, no. 6, Nov./Dec. 2009.

[8] D. Fetterly, "Adversarial Information Retrieval: The Manipulation of Web Content," ACM Computing Rev., 2007.

[9] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification. Wiley-Interscience Publication, 2000.

[10] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial Classification," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 99-108, 2004.

[11] M. Barreno, B. Nelson, R. Sears, A.D. Joseph, and J.D. Tygar, "Can Machine Learning be Secure?" Proc. ACM Symp. Information, Computer and Comm. Security (ASIACCS), pp. 16-25, 2006.

[12] A.A. C_ardenas and J.S. Baras, "Evaluation of Classifiers: Practical Considerations for Security Applications," Proc. AAAI Workshop Evaluation Methods for Machine Learning, 2006.

[13] P. Laskov and R. Lippmann, "Machine Learning in Adversarial Environments," Machine Learning, vol. 81, pp. 115-119, 2010.

[14] L. Huang, A.D. Joseph, B. Nelson, B. Rubinstein, and J.D. Tygar, "Adversarial Machine Learning," Proc. Fourth ACM Workshop Artificial Intelligence and Security, pp. 43-57, 2011.

[15] M. Barreno, B. Nelson, A. Joseph, and J. Tygar, "The Security of Machine Learning," Machine Learning, vol. 81, pp. 121-148, 2010.

[16] D. Lowd and C. Meek, "Adversarial Learning," Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 641-647, 2005.

[17] P. Laskov and M. Kloft, "A Framework for Quantitative Security Analysis of Machine Learning," Proc. Second ACM Workshop Security and Artificial Intelligence, pp. 1-4, 2009.

[18] NIPS Workshop Machine Learning in Adversarial Environments for Computer Security, http://mls-nips07.first.fraunhofer.de/, 2007.

[19] Dagstuhl Perspectives Workshop Mach. Learning Methods for Computer Sec., http://www.dagstuhl.de/12371/, 2012.

[20] A.M. Narasimhamurthy and L.I. Kuncheva, "A Framework for Generating Data to Simulate Changing Environments," Proc. 25th Conf. Proc. the 25th IASTED Int'l Multi-Conf.: Artificial Intelligence and Applications, pp. 415-420, 2007.

[21] S. Rizzi, "What-If Analysis," Encyclopedia of Database Systems, pp. 3525-3529, Springer, 2009.

[22] J. Newsome, B. Karp, and D. Song, "Paragraph: Thwarting Signature Learning by Training Maliciously," Proc. Ninth Int'l Conf. Recent Advances in Intrusion Detection, pp. 81-105, 2006.

[23] A. Globerson and S.T. Roweis, "Nightmare at Test Time: Robust Learning by Feature Deletion," Proc. 23rd Int'l Conf. Machine Learning, pp. 353-360, 2006.

[24] R. Perdisci, G. Gu, and W. Lee, "Using an Ensemble of One-Class SVM Classifiers to Harden Payload-Based Anomaly Detection Systems," Proc. Int'l Conf. Data Mining, pp. 488-498, 2006.

[25] S.P. Chung and A.K. Mok, "Advanced Allergy attacks: Does a Corpus Really Help," Proc. 10th Int'l Conf. Recent Advances in Intrusion Detection (RAID '07), pp. 236-255, 2007.

[26] Z. Jorgensen, Y. Zhou, and M. Inge, "A Multiple Instance Learning Strategy for Combating Good Word Attacks on Spam Filters," J. Machine Learning Research, vol. 9, pp. 1115-1146, 2008.

[27] G.F. Cretu, A. Stavrou, M.E. Locasto, S.J. Stolfo, and A.D. Keromytis, "Casting out Demons: Sanitizing Training Data for Anomaly Sensors," Proc. IEEE Symp. Security and Privacy, pp. 81-95, 2008.

[28] B. Nelson, M. Barreno, F.J. Chi, A.D. Joseph, B.I.P. Rubinstein, U. Saini, C. Sutton, J.D. Tygar, and K. Xia, "Exploiting Machine Learning to Subvert Your Spam Filter," Proc. First Workshop Large- Scale Exploits and Emergent Threats, pp. 1-9, 2008.