CONTENT- BASED MESSAGE FILTERING IN ON-LINE SOCIAL NETWORKS

A. Dhinesh[1], *M.Prabhakaran*[2]

Abstract

On-line Social Networks (OSN's) have experienced tremendous growth in recent years and become a de facto portal for hundreds of millions of Internet users. These OSN's offer attractive means for digital social interactions and information sharing, but also raise a number of security and privacy issues. While OSN's allow users to restrict access to shared data, they currently do not provide any mechanism to enforce privacy concerns over data associated with multiple users. This project proposes a system enforcing content-based message filtering conceived as a key service for On-line Social Networks. The system allows the OSN's users to have a direct control on the messages posted on their walls. This is achieved through a flexible rule-based system, that allows a user to customize the filtering criteria to be applied to their walls, and Machine Learning based soft classifier automatically producing membership labels in support of content-based filtering. To this end, a new approach has been proposed to enable the protection of shared data associated with multiple users in OSN's. An access control model has designed to capture the essence of multiparty authorization requirements, along with a multiparty policy specification scheme and a policy enforcement mechanism. The existence of online social networks that include person specific information creates interesting opportunities for various applications ranging from marketing to community organization. On the other hand, security and privacy concerns need to be addressed for creating such applications. Improving social network access control systems appears as the first step toward addressing the existing security and privacy concerns related to online social networks. To address some of the current limitations, an experimental social network has been created using synthetic data which can used to test.

Keywords : Classifier, Attack, Content, Filters, Mechanism, Synthetic.

Introduction

A social networking service is a platform to build social networks or social relations among people, for example, share interests, activities, backgrounds, or real-life connections. A social network service consists of a representation of each user (often a profile), his/her social links, and a

[1]Research Scholar, Dept. Computer Science, Government Arts College, Ariyalur - 621713
e-mail : ayy.dhinesh@gmail.com
[2]Research Supervisor, Dept. Computer Science, Government Arts College, Ariyalur - 621713

variety of additional services. Pattern classification systems based on machine learning algorithms are commonly used in security-related applications like biometric authentication, network intrusion detection, and spam filtering, to discriminate between a "Legitimate" and a "Malicious" pattern class (e.g., Legitimate and Spam emails). Contrary to traditional ones, these applications have an intrinsic adversarial nature since the input data can be purposely manipulated by an intelligent and adaptive adversary to undermine classifier operation. This often gives rise to an arms race between the adversary and the classifier designer. Well known examples of attacks against pattern classifiers are: submitting a fake biometric trait to a biometric authentication system (spoofing attack) modifying network packets belonging to intrusive traffic to evade intrusion detection systems; manipulating the content of spam emails to get them past spam filters.

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 website 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. 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.

LITERATURE REVIEW

Many authors implicitly performed security evaluation as a what-if analysis, based on empirical simulation methods; however, they mainly focused on a specific application, classifier and attack, and devised adhoc security evaluation procedures based on the exploitation of problem knowledge and heuristic techniques. Their goal was either to point out a previously unknown vulnerability, or to evaluate security against a known attack. In some cases, specific countermeasures were also proposed, according to a proactive/security-by-design approach. Attacks were simulated by manipulating training and testing samples according to application-specific criteria only, without reference to more general guidelines; consequently, such techniques cannot be directly exploited by a system designer in more general cases. Few works proposed analytical methods to evaluate the security of learning algorithms or of some classes of decision functions (e.g., linear ones), based on more general, application-independent criteria to model the adversary's behavior.

Some of these criteria will be exploited in our framework for empirical security evaluation as high-level guidelines for simulating attacks.

Taxonomy of attacks against pattern classifiers

We will exploit it in our framework, as part of the definition of attack scenarios. The taxonomy is based on two main features: the kind of influence of attacks on the classifier, and the kind of security violation they cause. The influence can be either causative, if it undermines the learning algorithm to cause subsequent misclassifications; or exploratory, if it exploits knowledge of the trained classifier to cause misclassifications, without affecting the learning algorithm. Thus, causative attacks may influence Both training and testing data, or only training data, whereas exploratory attacks affect only testing data. The security violation can be an integrity violation, if it allows the adversary to access the service or resource protected by the classifier; an availability violation, if it denies legitimate users access to it; or a privacy violation, if it allows the adversary to obtain confidential information from the classifier.

Integrity violations result in misclassifying malicious samples as legitimate, while availability violations can also cause legitimate samples to be misclassified as malicious. A third feature of the taxonomy is the specificity of an attack, that ranges from targeted to indiscriminate, depending on whether the attack focuses on a single or few specific samples (e.g., a specific spam email misclassified as legitimate), or on a wider set of samples.

"Content-Based Book Recommending Using Learning for Text Categorization" Recommender systems improve access to relevant products and information by making personalized suggestions based on previous examples of a user's likes and dislikes.

"Machine Learning in Automated Text Categorization" This paper can improve the productivity of human classifiers in applications in which no classification decision can be taken without a final human judgment.

"Combining Collaborative Filtering with Personal Agents for Better Recommendations " This paper shows that a CF framework can be used to combine personal IF agents and the opinions of a community of users to produce better.

"Automatic Extraction of Facts from Press Releases to Generate News Stories" This paper uses a template-driven approach, partial understanding techniques, and heuristic procedures to extract certain key pieces of information from a limited range of text.

"Learning and Revising User Profiles: The Identification of Interesting Web Sites" Algorithms for learning and revising user profiles that can determine which World Wide Web sites on a given topic would be interesting to a user. Describe the use of a naive Bayesian classifier for this task, and demonstrate that it can incrementally learn profiles from user feedback on the interestingness of Web sites.

EXISTING METHOD

In existing there is no content-based preferences are supported and therefore it is not possible to prevent undesired messages, such as political or vulgar ones, no matter of the user who posts them. Providing this service is not only a matter of using previously defined web content mining techniques for a different application, rather it requires to design ad hoc classification strategies. This is because wall messages are constituted by short text for which traditional classification methods.

We summarize here the three main concepts more or less explicitly emerged from previous work that will be exploited in our framework for security evaluation.

1) Arms race and security by design: since it is not possible to predict how many and which kinds of attacks a classifier will incur during operation, classifier security should be proactively evaluated using a what-if analysis, by simulating potential attack scenarios.

2) Adversary modeling: effective simulation of attack scenarios requires a formal model of the adversary.

3) Data distribution under attack: the distribution of testing data may differ from that of training data, when the classifier is under attack.

PROPOSED METHOD

To propose and experimentally evaluate an automated system, called Filtered Wall (FW), able to filter out unwanted messages from social network user walls. The key idea of the proposed system is the support for content based user preferences. This is possible thank to the use of a Machine Learning (ML) text categorization procedure able to automatically assign with each message a set of categories based on its content.

We believe that the proposed strategy is a key service for social networks in that in today social networks users have little control on the messages displayed on their walls. We can construct the three tier architecture such as social network manager, social network applications and Graphical user interface. The core components of the proposed system are the Content-Based Messages Filtering (CBMF) and the Short Text Classifier (STC) modules. The first aspect is related to the fact that, in OSN's like in everyday life, the same message may have different meanings and relevance's based on who writes it. As a consequence, filtering rules should allow users to state constraints on message creators. Thus, creators on which a filtering rule applies should be selected on the basis of several different criteria; one of the most relevant is by imposing conditions on user profile's attributes. In such a way it is, for instance, possible to define rules applying only to young creators, to creators with a given religious/ political view, or to creators that we believe are not expert in a given field (e.g. By posing constraints on the work attribute of user profile).Implement pattern classifiers to enables the users to block the malicious messages from friends.

To build a robust Support vector machine is concentrated in the extraction and selection of a set of characterizing and discriminate features.

We propose here a framework for the empirical evaluation of classifier security in adversarial environments that unifies and builds on the three concepts.

Our main goal is to provide 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. To this end, we propose:

(i) A model of the adversary, that allows us to define any attack scenario;

(ii) A corresponding model of the data distribution; and

(iii) A method for generating training and testing sets that are representative of the data distribution, and are used for empirical performance evaluation.

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.

Results And Discussion

Spam Filtering

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. This kind of classifier has been considered by several authors, and it is included in several real spam filters.

Biometric Authentication

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 a multimodal system, one expects that the adversary should spoof all the corresponding biometric traits.

Network Intrusion Detection

Intrusion detection systems (IDS's) analyze 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's and subsequently handled by the system administrator. Two main kinds of IDS's exist: misuse detectors and anomaly-based ones. Misuse detectors match the analyzed network traffic against a database of signatures of known malicious activities (e.g., Snort).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 (e.g.,PAYL), 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.

This kind of IDS's is vulnerable to causative attacks, since an attacker may inject carefully designed malicious traffic during the collection of training samples to force the id's to learn a wrong model of the normal traffic.

CONCLUSION

In this project, we are using the software system to filter unwanted messages from social network walls. Exploiting a flexible language to specify Filtering

Rules (FR's), by which users can state what contents, should not be displayed on their walls. We can extend our framework to analyze unwanted images which are posted in online social network. Improve accuracy prevent malicious messages from unwanted users.

References

1. P. Johnson, B. Tan, and S. Schuckers, "Multimodal fusion vulnerability to non-zero effort (spoof) imposters," in IEEE Int'l Workshop on Inf. Forensics and Security, 2010.

2. P. Laskov and R. Lippmann, "Machine learning in adversarial environments," Machine Learning, vol. 81, pp. 115–119, 2010.

3. R. N. Rodrigues, L. L. Ling, and V. Govindaraju, "Robustness of multimodal biometric fusion methods against spoof attacks," J. Vis. Lang. Comput., vol. 20, no. 3, pp. 169–179, 2009.

4. P. Johnson, B. Tan, and S. Schuckers, "Multimodal fusion vulnerability to non-zero effort (spoof) imposters," in IEEE Int'l Workshop on Inf. Forensics and Security, 2010, pp. 1–5.

5. P. Fogla, M. Sharif, R. Perdisci, O. Kolesnikov, and W. Lee, "Polymorphic blending attacks," in Proc. 15th Conf. On USENIX Security Symp. CA, USA: USENIX Association, 2006.

6. G. L. Wittel and S. F. Wu, "On attacking statistical spam filters," in 1st Conf. On Email and Anti-Spam, CA, USA, 2004.

Authors Biography

A.DHINESH have completed MCA, M.Phil. At present he is perusing Doctoral Degree in Computer Science, in Gov. Arts College, Ariyalur. His area of interest is web mining and security.