Expert Systems with Applications 41 (2014) 5948–5959

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Phishing detection based Associative Classification data mining

Neda Abdelhamid [a,*], Aladdin Ayesh [a], Fadi Thabtah [b]

[a] Computing and Informatics Department, De Montfort University, Leicester, UK
[b] Ebusiness Department, Canadian University of Dubai, Dubai, United Arab Emirates

ARTICLE INFO

Keywords:
Classification
Data mining
Forged websites
Phishing
Internet security

ABSTRACT

Website phishing is considered one of the crucial security challenges for the online community due to the massive numbers of online transactions performed on a daily basis. Website phishing can be described as mimicking a trusted website to obtain sensitive information from online users such as usernames and passwords. Black lists, white lists and the utilisation of search methods are examples of solutions to minimise the risk of this problem. One intelligent approach based on data mining called Associative Classification (AC) seems a potential solution that may effectively detect phishing websites with high accuracy. According to experimental studies, AC often extracts classifiers containing simple "If-Then" rules with a high degree of predictive accuracy. In this paper, we investigate the problem of website phishing using a developed AC method called Multi-label Classifier based Associative Classification (MCAC) to seek its applicability to the phishing problem. We also want to identify features that distinguish phishing websites from legitimate ones. In addition, we survey intelligent approaches used to handle the phishing problem. Experimental results using real data collected from different sources show that AC particularly MCAC detects phishing websites with higher accuracy than other intelligent algorithms. Further, MCAC generates new hidden knowledge (rules) that other algorithms are unable to find and this has improved its classifiers predictive performance.

1. Introduction

The internet is not only important for individual users but also for organisations doing business online. Many of the organisations offer online trading and online sales of services and goods (Liu & Ye, 2001). Nevertheless, internet-users may be vulnerable to different types of online threats that may cause financial damages, identity theft, and loss of private information. Therefore, the internet suitability as a channel for commercial exchanges comes into question.

Phishing is considered a form of online threat that is defined as the art of impersonating a website of an honest firm aiming to acquire user's private information such as usernames, passwords and social security numbers (Dhamija, Tygar, & Hearst, 2006). Phishing websites are created by dishonest individuals to imitate genuine websites. These websites have high level of visual similarities to the legitimate ones in an attempt to defraud honest internet-users. A report published by "Gartner Co." (Gartner Inc., 2011), which is a research and advisory company shows that phishing attacks are increasing rapidly. Gartner estimated that theft through phishing attacks costs U.S. banks and credit card companies $2.8 billion annually. In 2011, the Director of Cisco's security-technology-business-unit issued his concerns that today's main attacks focus on gaining access to corporate accounts that contain valuable financial information.

Social engineering which is the act of manipulating people to obtain sensitive information, can be combined with computerised technical tricks in order to start a phishing attack (Aburrous, Hossain, Dahal, & Thabtah, 2010a). Fig. 1a depicts the general steps conducted in phishing. Phishing websites have become a serious problem not only because of the increased number of these websites but also the intelligent strategies used to design such websites. Therefore, users that have extensive experience and knowledge in computer security and internet might be deceived (Sanglerdsinlapachai & Rungsawa, 2010).

Typically, a phishing attack begins by sending an e-mail that seems to be from an authentic organisation to victims. These emails ask them to update their information by following a URL link within the e-mail. Other methods of distributing phishing URLs include, Black Hat search engine optimization (Black Hat SEO) (Seogod, 2011), Peer-to-peer file sharing, blogs, forums, instant messaging (IM) and Internet Relay Chat (IRC) (Kirda & Kruegel, 2005).

* Corresponding author. Tel.: +44 (0)116 2 50 60 70.
E-mail addresses: P09050665@myemail.dmu.ac.uk (N. Abdelhamid), aayesh@dmu.ac.uk (A. Ayesh), fadi@cud.ac.ae (F. Thabtah).

http://dx.doi.org/10.1016/j.eswa.2014.03.019
0957-4174/

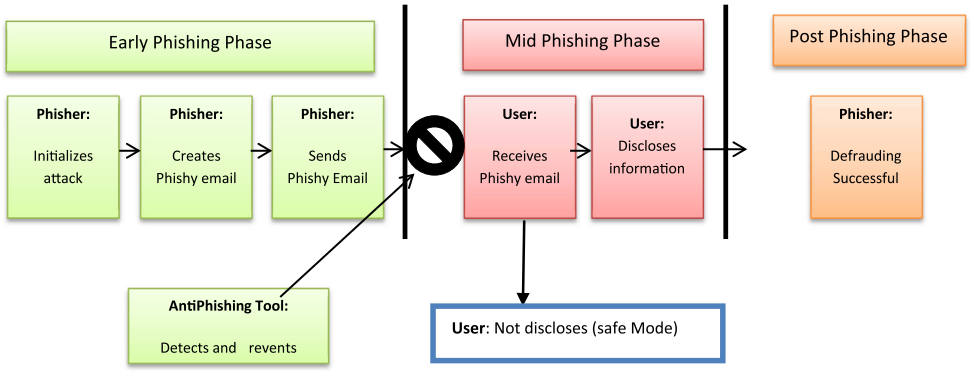

Fig. 1a. Phishing life cycle.

Below, we briefly explain the two most popular approaches in designing technical anti-phishing solutions (Aaron & Manning, 2012; Sadeh, Tomasic, & Fette, 2007).

- Blacklist approach: where the requested URL is compared with a predefined phishing URLs. The drawback of this approach is that the blacklist usually cannot cover all phishing websites since a newly created fraudulent website takes a considerable time before it can be added to the list.
- Search approach: the second approach is based on search/heuristic methods, where several website features are collected and used to identify the type of the website. In contrast to the blacklist approach, the heuristic-based approach can recognise newly created fake websites in real-time (Miyamoto, Hazeyama, & Kadobayashi, 2008).

The numbers of phishing websites are expected to increase over time. Thus, smart solutions are needed to keep pace with the continuous evolution of this problem. Smart solutions are the subject of our interest in this article. They can be combined with the heuristic-based approach as long as historical data exists. In fact, the accuracy of the heuristic-based solution mainly depends on a set of discriminative features extracted from the website. Hence, the way in which those features are processed plays an extensive role in accurately classifying websites. Therefore, an effective intelligent based method when merged with the heuristic method can be essential for making a good decision.

Associative Classification (AC) in data mining is one of the promising approaches that can make use of the features extracted from phishing and legitimate websites to find patterns among them (Costa, Ortale, & Ritacco, 2013; Thabtah, Cowling, & Peng, 2005). This approach normally devises classifiers (set of rules) that are simple yet accurate. The decision-making process becomes reliable because these decisions are made based on rules discovered from historical data by the AC algorithm. Although plenty of applications are available for combating phishing websites few of them make use of AC data mining (Jabbar, Deekshatulu, & Chandra, 2013).

Phishing is a typical classification problem (Abdelhamid, Ayesh, & Thabtah, 2013) in which the goal is to assign a test data (a new website) one of the predefined classes (phishy, legitimate, suspicious, etc.). Once a website is loaded on the browser a set of feature values will be extracted. Those features have a strong influence in determining the website type by applying the rules that have been previously found by the AC algorithm from the historical data (already labelled websites). Then, the chosen rule's class will be assigned to the browsed website and an appropriate action will take place. For instance, a message or an alarm will be fired to alert the user of the risk.

In this paper, the problem of phishing detection is investigated using AC approach in data mining. We primarily test a developed AC algorithm called MCAC and compare it with other AC and rule induction algorithms on phishing data. The phishing data have been collected from the Phishtank archive (PhishTank, 2006), which is a free community site. In contrast, the legitimate websites were collected from yahoo directory. The evaluation measures used in the comparison are accuracy, number of rules, any label, and label-weight (Thabtah, Cowling, & Peng, 2004). More details are given in Section 5.

We show that MCAC is able to extract rules representing correlations among website's features. These rules are then employed to guess the type of the website. The novelty of MCAC is its ability not only to discover one class per rule, but rather a set of classes bringing up the classifier performance in regards to accuracy. This is unlike current AC algorithms that only generate a single class per rule. Thus, the new classes connected with the rules that have been revealed by MCAC correspond to new knowledge missed by the majority of the existing AC algorithms. More details of MCAC processes are given in Section 4.

This paper is divided into different sections where Section 2 surveys common related learning approaches to phishing detection. Section 3 sheds the light on AC as well as its advantages, and MCAC algorithm. The features related to the phishing problem that have been utilised in the experimental section are discussed in Section 4. Section 5 is devoted to experiments where we demonstrate the data collection process, the evaluation measures, the compared algorithms, the results, and the analysis of the results. Lastly, conclusions are given in Section 6.

2. Related works

In this section, we review common intelligent phishing classification approaches from the literature, after shedding the light on the general steps required to solve the phishing problem and its general combating approaches. Further, the section starts by showing the phishing life cycle.

2.1. Phishing lifecycle

Fig. 1a depicts the general steps conducted in the phishing life cycle. According to Fig. 1a, a phishing attack begins by sending an e-mail that seems to be from an authentic organisation to users urging them to change their data by selecting a link within an e-mail. E-mails remain a spreading channel for phishing links since 65% of phishing attacks start by visiting a link received within an e-mail (Kaspersky Lab, 2013).