

Learn To Detect Phishing Scams Using Learning and Ensemble Methods

Alireza Saberi, Mojtaba Vahidi, Behrouz Minaei Bidgoli

Department of Computer Engineering, Iran University of Science and Technology

{a_Saberi, mojtaba_vahidi}@comp.iust.ac.ir, minaeibi@cse.msu.edu

Abstract

Phishing attack is a kind of identity theft which tries to steal confidential data like on-line bank account information. In a phishing attack scenario, attacker deceives users by a fake email which is called scam. In this paper we employ three different learning methods to detect phishing scams. Then, we use ensemble methods on their results to improve our scam detection mechanism. Experimental results show that the proposed method can detect 94.4% of scam emails correctly, while only 0.08% of legitimate emails are classified as scams.

1. Introduction

Phishing is a branch of internet crimes. In these attacks users' sensitive information such as passwords and credit card details are captured. Phishers use social engineering in their attacks to masquerade themselves as legitimate servers [1]. In a usual phishing attack, the attacker spoofs a website, similar to a known and trusted one. Phisher then sends a fake e-mail which is called scam to the user. In most cases, users are encouraged to refer to the website immediately by clicking on a link embedded in the scam. By following the link, some unaware users arrive to the rogue website and they may enter the phisher desired information. In this stage, the phisher has gained the sufficient information for the fraudulent objectives. The phisher may forge user's identity or withdraw from victim's internet bank account [2].

Increasing number of scams, evolves demands for new methods for fighting them. Only in 2004, more than 57 million users in USA received scams of phishing and almost 2 million of them have become the victims of such attacks [2].

In this paper, we try to separate attackers' emails from legitimate ones using well known data mining approaches and ensemble their results. By preventing users from receiving scams, these attacks would fail.

This paper is organized as follows. Section 2 discusses previous approaches for filtering spam and scams. In section 3 we introduce email types and their relation with phishing. In section 4 we consider three

data mining algorithms used for scam detection. Section 5 includes the results of applying the algorithms on scam detection. We conclude with final remarks and portray future work in section 6.

2. Related work

Scams are subset of spams. A spam refers to a group of emails which are sent to users with the aim of marketing or advertising.

One of the methods for spam detection is using information embedded in the e-mail header [3]. M. Chandrasekaran et al [4] tries to detect phishing emails based on structural features of emails such as linguistic properties, email subject and "richness" of email vocabulary.

Another way for spam detection is employing data mining algorithms which have been of much interest such as [5]. The process is as follows: based on the email context, using text classification algorithms, it starts to classify emails into two categories of frauds and non-frauds. The closest related method to our work is [6]. In this solution data mining methods are used based on words in emails to detect the Nigerian 4-1-9 scam.

I. Fette et al [7] first extracts special features set from emails which are designed to deceive users. These features include IP based URL, non-matching URL and the number of domains in the emails. The second step in this solution is using data mining algorithm based on these features to detect phishing scams.

Our proposed method employs three data mining algorithms based on the e-mail's text and combines the results to improve the accuracy.

3. Email categorization

First step in scam detection is to know different types of emails. Emails can be classified into three categories: Spams, Scams and Hams. Figure 1 shows the relationship between these three categories.

Spam email, known as bulk or junk email involves sending nearly identical messages to numerous recipients by email. Scams are subset of spams. This group has been designed very intellectually and tries to deceive users in order to achieve illegal objectives. In

contrast to scams, there are legitimate emails or hams which are those emails exchange between users.

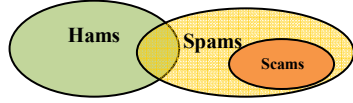


Figure 1. Email categorization [6]

4. Data mining algorithm for scam detection

In fact the problem on hand is a binary classification one. The aim is designing a message filter in order to detect fraudulent patterns in scams. Then filter is trained to make a boolean decision on a labeled dataset, where the labels are “scam” and “non scam”. After the filter has been trained, it can be applied to messages incoming to a mail server in real time. Our aim is to combine the results of data mining algorithms to achieve better results in email classification. In this paper we employ three algorithms for text classification: K nearest neighbor [8], Poisson probabilistic theory [6] and Bayesian probabilistic theory [9]. First, emails are classified into two categories of frauds and non-frauds using these algorithms. Then based on consensus method [10], combining the results of the proposed data mining algorithms, we improve the classification results.

4.1. K Nearest Neighbor

K Nearest Neighbor (KNN) is a simple and common method which is used in different data mining applications [8]. Assume the set $M = \{m_1, m_2, \dots, m_p\}$ consists of p messages and each message belongs to one of the two classes of scam and non scam, $C = \{Scam, Non-Scam\}$. So that $M = \bigcup_{c \in C} M_c$ is the union of disjoint sets of messages (M_c) in different categories. A vector V is derived from the set M which has x words and x is the total number of different words in the set M . The vector $V = \{v_1, v_2, \dots, v_x\}$ with x words is constructed for each message m in the learning phase. In case there is word v_i in message m the respective word in the corresponding vector is set to 1.

$$\forall_{i \in \{1, \dots, x\}} v_{mi} = 1 \quad \text{if} \quad v_i \in m \quad (1)$$

In the test phase for each test message m , we construct vector v_m similar to the training phase. Then distance between test vector v_m and all training vectors is computed as follow:

$$\forall_{i \in m} \sin(m_{input}, m_i) = \frac{\vec{m}_{input} \cdot \vec{m}_i}{\|\vec{m}_{input}\| \cdot \|\vec{m}_i\|} = \frac{\sum_j m_{inputj} m_{ij}}{\sqrt{\sum_j m_{inputj}^2} \sqrt{\sum_j m_{ij}^2}} \quad (2)$$

It is clear that the nearest message to input is the one with highest value in relation 2.

4.2. Poisson probabilistic theory

One of important algorithms which is used in text classification is Poisson model. The output of this

classifier is a number which represents the dependency of a message to one of the two classes (fraud or non-fraud).

The sets M , C and V are similar to KNN. Assume that X_{mv} is frequency of word v in message m . Assuming a Poisson process model for X_{mv} , the probability distribution of X_{mv} is as follows:

$$p(x_{mv} | w_m, \mu_{vc}) = \frac{e^{-w_m \mu_{vc}} (w_m \mu_{vc})^{x_{mv}}}{x_{mv}!} \quad (3)$$

Where W_m is message size in scale of 1000 and μ_{vc} is Poisson rate for word v in class c that is equal to the expectation of the occurrence of word v among the words of class c . The value of parameter μ_{vc} for each class c is determined by following relation during the learning phase:

$$\mu_{vc} = \frac{\sum_{m \in M_c} x_{mv}}{\sum_{m \in M_c} W_m} \quad (4)$$

In this method after computing the probability of words frequency in each message of each class, the probability that this message belongs to one class rather than the other is obtained from the following relation:

$$r_m = \frac{\prod_{v \in V} p(X_{mv} | \mu_{v-Scam})}{\prod_{v \in V} p(X_{mv} | \mu_{v-NonScam})} \quad (5)$$

With regard to relation 5, if r_m is greater than 1, message belongs to Scam class and otherwise to Non-Scams. Of course because of high importance of detecting Scam messages from Non-Scams, the threshold should be specified less than 1.

4.3. Naïve Bayes probabilistic theory

Naïve Bayes theory is one of the strongest information classification methods. The idea is based on naïve bayes probabilistic theory which selects most probable class [9].

Assume $V = \{v_1, v_2, \dots, v_d\}$ as a vector of input random message with d distinct words and each message belongs to one of two classes: $C = \{Scam, Non-Scam\}$. The aim of this algorithm is calculating the probability of belonging each message to one of the two classes. The algorithm is calculating as follows:

$$P(C = Scam, Non-Scam | v_1, \dots, v_d) = \frac{P(C = Scam, Non-Scam) \cdot P(v_1, \dots, v_d | C = Scam, Non-Scam)}{P(v_1, \dots, v_d)} \quad (6)$$

Such as Poisson algorithm the probability of belonging to two classes are divided together to create a ratio, in order to specify winner class.

$$\frac{P(C = Scam | v_1, \dots, v_d)}{P(C = Non - Scam | v_1, \dots, v_d)} = \frac{P(C = Scam)}{P(C = Non - Scam)} \times \frac{P(v_1, \dots, v_d | C = Scam)}{P(v_1, \dots, v_d | C = Non - Scam)} \quad (7)$$

In case that occurrence probability of each word is independent from the others, we have

$$P(v_1, \dots, v_d | C = Scam, Non - Scam) = \prod_{i=1}^d P(v_i | C = Scam, Non - Scam) \quad (8)$$

Therefore:

$$\log \frac{P(C = Scam | v_1, \dots, v_d)}{P(C = Non - Scam | v_1, \dots, v_d)} = \log \frac{P(C = Scam)}{P(C = Non - Scam)} + \sum_{i=1}^d \log \frac{P(v_i | C = Scam)}{P(v_i | C = Non - Scam)} \quad (9)$$

In equation 9 the value of first logarithm is always constant. If the right side logarithm result becomes higher than zero, it would belong to scam class and otherwise to non scams.

4.4. Ensemble classification

The key idea in ensemble methods is to achieve higher accuracy by combining results of different algorithms. In this paper we applied the ensemble method on the outputs of different classifiers.

There are many ways which do the ensemble in output level of classification algorithms such as Majority voting [10], Weighted majority voting [11], Naïve Bayes combination [12] and N Dimensional naïve Bayes sampling [13]. Using ensemble classification has made considerable results in spam detection [14]. In this article majority voting method has been used.

5. Experimental results

In this section we study the results. First we describe how the data is gathered. Then we consider the results of using each algorithm mentioned in section 4. Finally in this section we ensemble the algorithms to achieve higher accuracy.

5.1 Sample data set

Spams and hams have been provided from a standard data set called Enron-spam [15]. We used the scam samples from a web repository of phishing [16]. The data set consists of 4500 spams, 1500 legitimate emails and 529 scams of phishing type with 70,000 different words. 2400 most frequent words have been selected from the data set.

5.2 Poisson filter results

First all messages are categorized into two groups of frauds and non-frauds. Non fraud group consists of spams and hams. Frauds include phishing messages.

We employed Poisson classifier with balanced 5-fold cross-validation mechanism [9]. The results of Poisson algorithm is shown in table 1.

Table 1: Confusion matrix of Poisson algorithm

		<i>Poisson Algorithm Output</i>	
		<i>Scam</i>	<i>Spam and Ham</i>
<i>Real Classes</i>	<i>Scam</i>	88%	12%
	<i>Spam and Ham</i>	0.1%	99.9%

Although Poisson filter is well done in spam and ham detection, in case of scam the results have considerable mistake.

5.3 K Nearest Neighbor filtering results

K nearest neighbor algorithm has been studied for different number of neighbors. The output result is shown in table 4. A key point in this filter is growth in fault occurrence according to increase in the number of Neighbors. The presented results show that, since the number of samples in each class is not balanced, decrease in the number of Neighbors may improve the result respectively.

5.4 Naïve Bayes filter results

Naïve Bayes results are shown in table 2. Considering the number of selected scam and non scam messages, the filter threshold is chosen -1.079 during the training phase. In other words the messages which their score in formula 7 is more than the threshold are detected as scam. Although naïve Bayes filter does not act well in non scam detection, but because its mistake results have no intersection with other filters, it would have effective role in ensemble process.

Table 2: Confusion matrix of Naïve Bayes algorithm

		<i>Naïve Bayes Algorithm Result</i>	
		<i>Scam</i>	<i>Spam and Ham</i>
<i>Real Classes</i>	<i>Scam</i>	90.6%	9.4%
	<i>Spam and Ham</i>	1.8%	98.2%

5.5 Ensemble classification

The aim of the ensemble classification is to increase the accuracy of other filter results. In this paper the Majority voting approach has been applied. Ensemble method results are shown in table 3. As expected, Ensemble algorithm improved detection results in both classes considerably.

Table 3: Confusion matrix of ensemble classification

		<i>Majority Vote Results</i>	
		<i>Scam</i>	<i>Spam and Ham</i>
<i>Real Classes</i>	<i>Scam</i>	94.4%	5.6%
	<i>Spam and Ham</i>	0.08%	99.92%

6. Conclusion and future work

Considering the importance of spam and scam detection, various data mining algorithms have been employed. Three of these classification algorithms have been used: Naïve Bayes, Poisson and K Nearest Neighbor. Then by using majority voting ensemble classification algorithm, their results were merged in order to increase the accuracy.

For future work other types of Naïve Bayes implementation could be employed in order to find the

most efficient of them. Furthermore, the manner of choosing word set could be studied later. Using other classification methods such as SVM may be studied later. Other features in emails such as hex format IP address and matched URL can be used to train learning algorithm. Furthermore using different ensemble methods would improve scam detection accuracy and can be considered as a novel approach in the battle against spams and scams.

Table 4: Confusion matrix resulted from K Nearest Neighbor

		<i>KNN algorithm Result</i>							
		K=1		K=3		K=5		K=7	
		<i>scam</i>	<i>Non scam</i>	<i>scam</i>	<i>Non scam</i>	<i>scam</i>	<i>Non scam</i>	<i>scam</i>	<i>Non scam</i>
<i>Real classes</i>	<i>Scam</i>	91.5%	8.5%	87.5%	12.2%	83.9%	14.1%	86.6%	13.2%
	<i>Non scam</i>	0.3%	99.7%	0.3%	99.7%	0.3%	99.7%	0.3%	99.7%

7. References

- [1] N. Chou, R. Ledesma, Y. Teraguchi and J.C. Mitchell, "Client-Side Defense against Web-Based Identity Theft", 11th Annual Network and Distributed System Security Symposium, San Diego, USA, February 2004.
- [2] R. Dhamija, J. D. Tygar, M. Hearst, "Why Phishing Works", CHI Conference on Human Factors in Computing Systems, Montreal, Canada, April 2006.
- [3] P. Pfleeger, S. Bloom, "Canning Spam: Proposed Solutions to Unwanted Email", Security & Privacy Magazine, IEEE, vol. 3, no. 2, March-April 2005, pp. 40-7.
- [4] M. Chandrasekaran, K. Karayanan, S. Upadhyaya. "Towards phishing e-mail detection based on their structural properties", New York State Cyber Security Conference, USA, 2006.
- [5] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz, "A bayesian approach to filtering junk e-mail", AAAI Workshop on Learning for Text Categorization, Madison, Wisconsin, USA, July 1998.
- [6] E. Airoldi, B. Malin, "Data Mining Challenges For Electronic Safety: The Case of Fraudulent Intent Detection in E-Mails", In proceedings of the IEEE Workshop on Privacy and Security Aspects of Data Mining, 2004, pp. 57-66.
- [7] I. Fette, N. Sadeh, A. Tomasic, "Learning to Detect Phishing Emails", Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada 2007.
- [8] B. V. Dasarathy. "Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques". IEEE Computer Society Press, Los Alamitos, California, 1990.
- [9] E. Airoldi and W. Cohen. "Bayesian models for frequent terms in text", Technical Report No. CMU-CALD-04-106, Carnegie Mellon University, July 2004.
- [10] E. Day, "Consensus methods as tools for data analysis", In H. Bock, editor, Classification and Related Methods for Data Analysis, Elsevier Science Publishers B.V. (North Holland), 1988, pp. 317-324.
- [11] N. Littlestone and M. Warmuth, "Weighted majority algorithm", Information and Computation, vol. 108, 1994, pp. 212-261.
- [12] P. Domingos and M. Pazzani. "On the optimality of the simple Bayesian classifier under zero-one loss" Machine Learning, 29:103-130, 1997
- [13] Y. S. Huang and C. Y. Suen. "A method of combining multiple experts for the recognition of unconstrained handwritten numerals." IEEE Transactions on Pattern Analysis and Machine Intelligence, 17:90-93, 1995.
- [14] S. Hershkop, S. Stolfo, "Combining email models for false positive reduction", Proceeding of the eleventh ACM international conference on Knowledge discovery in data mining, Chicago, Illinois, USA, 2005.
- [15] V. Metsis, I. Androustopoulos, G. Paliouras, "Spam Filtering with Naive Bayes - Which Naive Bayes?", Third Conference on Email and AntiSpam, Mountain View, California, USA, July, 2006.
- [16] The Internet Defence Phishery. Repository of phishing emails, September 2006
<http://phishery.internetdefence.net/data/>