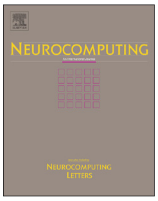
Contents lists available at ScienceDirect

Neurocomputing

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



Spam filtering for short messages in adversarial environment

Patrick P.K. Chan *, Cheng Yang, Daniel S. Yeung, Wing W.Y. Ng

School of Computer Science and Engineering, South China University of Technology, China

ARTICLE INFO

Article history:
Received 15 May 2014
Received in revised form
21 September 2014
Accepted 14 December 2014
Communicated by Lijun Tang

Keywords:
Spam filtering
Short messages
Good word attack
Feature reweighting
Adversarial learning

ABSTRACT

The unsolicited bulk messages are widespread in the applications of short messages. Although the existing spam filters have satisfying performance, they are facing the challenge of an adversary who misleads the spam filters by manipulating samples. Until now, the vulnerability of spam filtering technique for short messages has not been investigated. Different from the other spam applications, a short message only has a few words and its length usually has an upper limit. The current adversarial learning algorithms may not work efficiently in short message spam filtering. In this paper, we investigate the existing good word attack and its counterattack method, *i.e.* the feature reweighting, in short message spam filtering in an effort to understand whether, and to what extent, they can work efficiently when the length of a message is limited. This paper proposes a good word attack strategy which maximizes the influence to a classifier with the least number of inserted characters based on the weight values and also the length of words. On the other hand, we also proposes the feature reweighting method with a new rescaling function which minimizes the importance of the feature representing a short word in order to require more inserted characters for a successful evasion. The methods are evaluated experimentally by using the SMS and the comment spam dataset. The results confirm that the length of words is a critical factor of the robustness of short message spam filtering to good word attack.

1. Introduction

Electronic spamming, which refers to the unsolicited bulk messages distributed via the electronic messaging systems, is a serious problem for the Internet users. Spam is an unwanted message sent to a recipient who has not requested it. Examples include advertisement, scam, phishing, *etc.* Due to the low operation cost, electronic spamming has steadily grown and spread to many media, like SMS, web search engine, blog and forum recently. Over 25 billion comment spams have been detected by Akismet in Wordpress blogs between 2009 and 2012 [1], and 69.6% of the total email traffic was generated by spams in 2013.[1]

Although the current classification methods [2–6] have achieved high accuracy of separating the legitimate from the unsolicited message, they are not adequate in providing an effective defence against an intelligent adversary such as the spammer who consciously manipulates the data to mislead the decisions of the system. The spam filters implemented by traditional classification algorithms which do not consider the presence of an adversary may be downgraded significantly in an adversary environment. Spam filtering has become an arms race between the spammers and the defenders countering each other's tactics.

The good word attack [7,8] is one of the well-known examples of the evasion attack in spam filtering. It aims to mislead the decision of a classifier on the junk mails by adding good words. Good words are common in the legitimate e-mails, but rare in junk e-mails. One example is the affiliation of a recipient. The camouflaged junk mails can then pass through a spam filter and successfully deliver unsolicited messages to clients. Most studies on the good word attack in spam filtering focus on the analysis of the vulnerability of a classifier [7,8] and the development of adversary-aware algorithms [9–11]. A few studies focus on data contamination [12–14]

Short messages are commonly used in many communication media which includes SMS, instance message, blog and forum riding on the popularity of the smart phone. However, to the best of our knowledge, until now the problem of adversarial learning for the short message spam filtering has not been investigated. A short message often consists of only a few words with a maximum length limit. For example, a traditional SMS message is limited to 140 bytes and a twitter comment has at most 140 characters. The current adversary attacks and counterattack strategies may not work efficiently for short messages since they do not consider the

* Corresponding author.
E-mail addresses: patrickchan@ieee.org (P.P.K. Chan), orangeyoung@foxmail.org (C. Yang), danyeung@ieee.org (D.S. Yeung), wingng@ieee.org (W.W.Y. Ng).

[1] http://www.securelist.com/en/analysis/204792322/Kaspersky_Security_Bulletin_Spam_evolution_2013

http://dx.doi.org/10.1016/j.neucom.2014.12.034
0925-2312/© 2014 Elsevier B.V. All rights reserved.

length limitation. In this paper, we investigate the good word attack for short message spam filtering from the perspective of an attacker and a defender in an effort to understand whether, and to what extent, the spam filtering for short messages may be vulnerable to an adversarial attack. In this study a short message is formulated as a Boolean vector while each Boolean feature indicates the presence of a word in the message.

The worst scenario of the good word attack assumes that the adversary has the complete knowledge of the classifier and it inserts the good words according to the weights of the classifier [15]. However, as the length of a short message is limited, adding a long good word to a short message may not be effective. Therefore, this paper introduces a good word attack model for short message spam filter which considers not only the weights but also the lengths of words. A short good word with a heavy weight is preferable in the proposed model. The practical implementation with low time complexity for a linear classifier based on a greedy algorithm is also described.

The feature reweighting method is proposed [16] to avoid a classifier over- and under-emphasizes on some features by rescaling the feature values according to their importance defined by the weight values of the initially trained classifier. The classifier with more evenly distributed feature weights is obtained by using the rescaled dataset. It has been shown that a classifier with more evenly distributed feature weights is more robust since it requires more manipulation on a sample to evade the detection [15,17]. We examine whether this observation is still valid under the length limitation in spam filtering for short messages. Then, the feature reweighting method with a new rescaling function is proposed to defend against the good word attack in short message spam filter. The new rescaling function adjusts the feature values based on the weight value and also the length of words. It punishes the short word with larger weight in order to reduce the influence of adding a short word to the output of the classifier. As a result, more words should be inserted into a message for a successful evasion.

The rest of this paper is organized as follows. The overview of the literature on the adversarial classification in spamming, including the good word attack and the feature reweighting, is given in Section 2. Assuming that the adversary has the complete knowledge of the classifier, the proposed good word attack model to spam filter for short messages and one of its implementation methods for linear classifiers are described in Section 3. Section 4 introduces the feature reweighting method with a revised rescaling function. The proposed good word attack model and the feature reweighting method are compared with the existing methods experimentally using the SMS and the comment spam datasets in Section 5. The conclusion is given in Section 6.

2. Background of adversarial classification in spamming

Machine learning techniques are widely used in security problems which include network spam filtering [18,19], intrusion detection [20,21] and malware detection [22,23]. It aims to separate the malicious from the legitimate samples. A major characteristic of these problems is the presence of an adversary who misleads the detection system by modifying the data. Since the traditional classification systems do not consider the influence of an adversary, they may be vulnerable to adversary attacks. Typically the traditional classifiers assume that data distributions of the training and the unseen samples are the same. Hence their detection effectiveness are undermined due to the non-stationary data generated by adversarial attacks [24–27].

Adversarial attacks can be categorized according to their properties in the taxonomy [26,28,27]. *Causative attack* manipulates the training data to influence the learning process, while decisions of a classifier are misled by modified testing samples in *exploratory attack*. The security violation of an attack can be separated into three types: *availability violation* which downgrades the general performance of a classifier; *integrity violation* which increases the classification error on malicious samples; and *privacy violation* which invades the system by stealing its information.

Spam filtering is a classic application of the adversarial learning. Evasion attack [26,28,27], in which an attacker attempts to evade the detection by manipulating the malicious samples, is a well-known attack. It aims to increase the false negative rate of the classifier, *i.e.* the accuracy of classifying junk mails. Good word attack [7,8] is a kind of evasion attack. The malicious samples are manipulated by inserting the good words which appear frequently in legitimate messages but rarely in spam messages. The ability of an adversary is usually constrained by the cost of manipulation on feature values, *e.g.* the number of manipulation.

Depending on the level of knowledge on the classifier, the attack strategies can be categorized into the passive and the active attack [7]. In a passive attack, the adversary has no knowledge on the classifier. The inserted good words are chosen according to the prior information, *e.g.* the given dataset, which is independent of the classifier.

By contrast, partial or complete information of the classifier can be acquired by the adversary in an active attack [7]. In the worst scenario [15] the weights ($\mathbf{w}$) of a linear classifier ($f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x}+b)$) are known. Assume that $f(\mathbf{x}) = +1$, $\mathbf{x}$ belongs to malicious; otherwise, it is a legitimate sample. The weights and features sorted according to the absolute values of the weights are denoted as $w_{(1)}$, $w_{(2)}$, …, $w_{(m)}$ and $x_{(1)}$, $x_{(2)}$, …, $x_{(m)}$, *i.e.* $|w_{(1)}| \geq |w_{(2)}| \geq \ldots \geq |w_{(m)}|$, where m is the number of features. The manipulated sample $\mathbf{x}'$ is set to $\mathbf{x}$ initially. For $i = 1, 2, \ldots, m$, if $x_{(i)} = 0$ and $w_{(i)} < 0$, $x'_{(i)}$ is set to 1; otherwise, $x'_{(i)}$ is left unchanged. The algorithm stops when $f(\mathbf{x}') = +1$ or the number of feature manipulation is equal to its maximum value. As an active attack requires the knowledge of the classifier, unsurprisingly it is more efficient than the passive one. However, since the existing algorithms of good word attack do not consider the length limitation of inserted words, they may not be suitable for the short message spam filtering.

On the other hand, many counterattack methods have been proposed to increase the robustness of a classifier to an evasion attack. For example, multiple instance learning [29] splits a message in order to increase the difficulty of an attack. It has been shown that multiple classifier systems [15] are more robust to the evasion attack than single classifiers. Feature reweighting, which is investigated in this paper, is a method to construct a robust linear classifier to defend against an evasion attack [16]. It aims to reweight a trained classifier which may overemphasizes the highly relevant features and underemphasizes the less informative features to avoid unevenly distributed weight values. After an initial linear classifier is trained, each sample, which is denoted as $\mathbf{x} = [x_1, x_2, \ldots, x_m]$ with m features in the training set, is rescaled according to its weight vector ($\mathbf{w} = [w_1, w_2, \ldots, w_m]$)

$$\overline{x_i} = x_i / s(w_i) \tag{1}$$

where $s(u) = \log\ (e + |u|)$, which is a positive and monotonically increasing function mapping R to $(0, +\infty)$. The weights of the classifier which is trained by using the rescaled training set are more evenly distributed in terms of the weight evenness measure [16]. It also shows that the retrained classifier is more robust to evasion attack experimentally. However, the short good word may be more preferable to the good word attack in spam filtering for short messages due to the length limitation. Therefore, a classifier with evenly distributed weights may not be efficient in defence against the good word attack in short message spam filtering.

Please cite this article as: P.P.K. Chan, et al., Spam filtering for short messages in adversarial environment, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2014.12.034