Neurocomputing xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

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

Discovering social spammers from multiple views

Hua Shen[a,d], Fenglong Ma[c], Xianchao Zhang[b,*], Linlin Zong[a,b], Xinyue Liu[b], Wenxin Liang[b]

[a] School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China
[b] School of Software, Dalian University of Technology, Dalian 116620, China
[c] SUNY at Buffalo, Buffalo, NY 14221, USA
[d] College of Mathematics and Information Science, Anshan Normal University, Anshan 114007, China

ARTICLE INFO

Communicated by Peng Cui

Keywords:
Social spammer detection
Multi-view learning
Social regularization term

ABSTRACT

Online social networks have become popular platforms for spammers to spread malicious content and links. Existing state-of-the-art optimization methods mainly use one kind of user-generated information (i.e., single view) to learn a classification model for identifying spammers. Due to the diversity and variability of spammers' strategies, spammers' behavior may not be completely characterized only by single view's information. To tackle this challenge, we first statistically analyze the importance of considering multiple view information for spammer detection task on a large real-world Twitter dataset. Accordingly, we propose a generalized social spammer detection framework by jointly integrating multiple view information and a novel social regularization term into a classification model. To keep the completeness of the original dataset and detect more spammers by the proposed method, we introduce a simple strategy to fill the missing data for each view. Experimental results on a real-world Twitter dataset show that the proposed method outperforms the existing methods significantly.

1. Introduction

Online social networks (OSNs), such as Twitter and Facebook, have become popular platforms to disseminate and share information [1]. Unfortunately, social spammers take advantage of those platforms to spread phishing scams, publish malicious content and links, and promote commodity information [2–4]. According to a study by Nexgate [5], the number of social spam grew more than 355% from January to July of 2013, which means that one in two hundred social messages was a spam, and 15% of all spams contained URLs linking to risky websites. Spammers are so sophisticated and concealed that they change spamming strategies irregularly and try to disguise as legitimate users. Moreover, to increase their influence and be undetected, spammers collude with each other to construct the criminal communities [6]. The malicious behavior of spammers has not only hindered the OSNs' development largely [7], but also threatened information security and personal privacy [8]. Therefore, it is crucial to design effective and novel spammer detection methods for the development of social systems.

Traditional approaches for combating spammers mainly focus on analyzing and extracting users' features, and then applying the existing classification methods to detect spammers or spam campaigns [3,9–11]. As the spamming strategies evolve, these methods only relying on the features could not effectively detect spammers with new spamming strategies. Ranking schemes are also employed in some anti-spam measures using social network information, which can decrease the spammers' impact on legitimate users [12,13]. However, these ranking methods are hard to distinguish legitimate users and spammers only depending on network information.

The state-of-the-art approaches employ supervised machine learning techniques to train an optimization model using both user-generated content and network structures [14–16,8], which identify spammers more accurately than the traditional approaches. However, these optimization methods only rely on one kind of user-generated information, such as text features, URLs or hashtags (i.e., *single view*). As we all known, spam strategies are diverse and protean so that the single view information may not characterize spammers' behavior completely. For example, some spammers may post legal text but adding unfriendly shorten URLs or tempting hashtags to achieve their malicious purposes. Consequently, these spammers would not be correctly identified by existing approaches. Thus, it is more reasonable and challenging to take multiple perspectives of spammers' behavior (i.e., *multiple views*) into consideration when detecting spammers.

To tackle the aforementioned challenges, we propose a generalized spammer detection framework, named **Multi-View Learning for Social Spammer Detection** (MVSD), by taking advantage of multiple view information of users and network information as shown in Fig. 1. We first statistically analyze the distribution differences between

* Corresponding author.
E-mail address: xczhang@dlut.edu.cn (X. Zhang).

http://dx.doi.org/10.1016/j.neucom.2016.11.013
Received 23 March 2016; Received in revised form 18 October 2016; Accepted 7 November 2016
Available online xxxx
0925-2312/

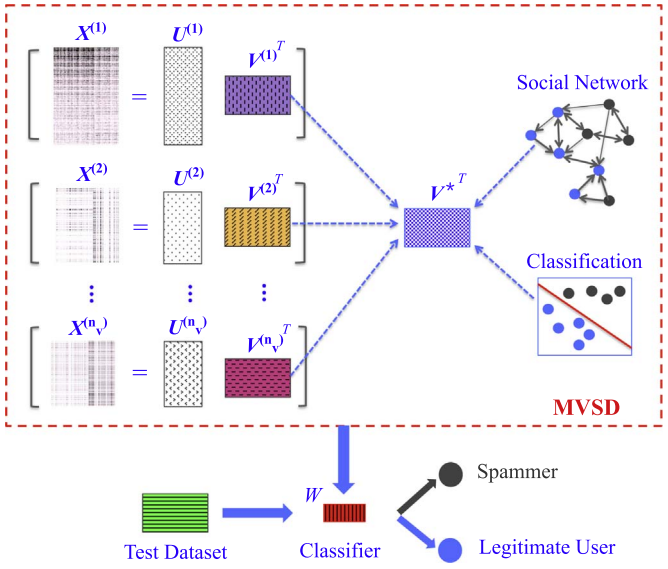

Fig. 1. The Framework of MVSD.

spammers and legitimate users from three views (text, URL and hashtag information) extracted from users' tweets. Based on the statistical analysis, we observe that different information has different ability to characterize users. Thus, we employ multi-view learning, which assigns different weight to each view. Since the values of features are non-negative, multi-view NMF can be used to learn a consensus matrix representing users' features. The proposed MVSD also takes social relationships into consideration, which measures the pairwise interactions among users. Different from existing work [14–16], we model all types of social relationships between legitimate users and spammers. Furthermore, in order to keep the completeness of the dataset, we use a simple method to fill in all the missing feature values. Finally, we jointly integrate users' multiple features and social relationships into a classification model to learn the classifier iteratively on the complemented dataset.

In summary, the contributions of this paper are as follows:

- The paper makes in-depth empirical analysis on a large real-world Twitter dataset, and the statistical results show that different views' feature distributions between legitimate users and spammers are different. Thus, it is reasonable to consider multiple view information of users when detecting spammers.
- The novelty of the paper is proposing a generalized spammer detection framework[1] by jointly modeling multiple view information and a novel social regularization term into a classification model. Through iteratively learning among multiple view information, social regularization and classification model, the proposed MVSD can train an accurate classifier.
- The experimental results on a real-world dataset show that the proposed framework can identify more spammers compared with baseline approaches. We conduct experiments to demonstrate the significance of taking multiple view information into consideration and validate the effectiveness of the new social regularization term. Finally, the importance of complementing missing values for spammer detection task is illustrated by the experimental results.

The remained of this paper is organized as follows: Section 2 analyzes different information from multiple views in OSNs. Section 3 formally defines the problem of multi-view spammer detection. In Section 4, we propose a multi-view learning model based on joint nonnegative matrix factorization. Section 5 demonstrates our evaluation process and experimental results. Section 6 reviews the related work on social spammer detection. The last section concludes the discussion.

2. Data analysis

We first introduce the dataset used in this paper, then statistically demonstrate the rationality of using multi-view method, and finally illustrate the importance of handling missing values for spammer detection task.

2.1. Dataset

In order to validate the proposed method impersonally and fairly, we select a standard and public dataset, Twitter Social Honeypot Dataset [17]. It has been used in [16,18,9], which provides the ground truth data, i.e., labeling users as spammer or legitimate ones, but only a *part* of following relationships. In order to complete the whole following relationships among users, we use the other public dataset, the Kwak's dataset [19][2]. We filter the non-English tweets and the users who posted less than one tweet. We also parse the shortened URLs to the original formats (i.e., long URLs) and only leave the hostname for each long URL. After processing, the final dataset contains 10,080 users (4,414 spammers and 5,666 legitimate users). For each user, we extracted features of text (9,749), URL (4,410) and hashtag (3,491) information.

2.2. Importance of multiple view information

Previous researches have shown that text information can be used to detect spammers effectively [15,16,20,21]. However, we find that not only text information but URLs and hashtags can also help to identify spammers as well. Moreover, they have different abilities to characterize spammers' behavior. For example, some spammers try to post duplicate or similar tweets to increase the probability of successfully alluring legitimate users, i.e., text spamming. Though some spammers publish normal text, the malicious URLs are embedded in the tweets, which is difficult to be detected. This is URL spamming. The third type of spamming is using hashtags, which adds the trending hashtags in the tweets to lure legitimate users to read or retweet them. These three spamming strategies are commonly used by spammers. Thus, we select text, URL and hashtag features as different views to describe users.

To further demonstrate the rationality of applying multiple view information to detect spammers, we first give an intuitive analysis by matrix graph [22] shown in Fig. 2. For each subfigure, X-axis represents users' features that we extracted for each view and Y-axis denotes users' classes (spammers and legitimate users). Each point in the matrix graph represents a feature's attribute value of each user in a single view. Ideally, the greater the color difference of points between the two classes of users, the better the classification performance of the view. We can observe that the distributions of spammers and legitimate users in text view are more similar compared with those in the other two views. It means that URLs and hashtags can be used to characterize the difference between spammers and legitimate users more effectively and separate the users into different classes more easily.

From statistical perspective, a non-parametric method, Spearman rank test, is employed to measure the difference of different views between spammers and legitimate users. The coefficient of spearman rank test r is from −1 to +1. Ideally, the closer the coefficient is to −1, the easier it is to separate users into different classes. Let n_i denote the number of the i-th attribute (word/URL/hashtag) posted by legitimate users and s_i denote the number of the i-th attribute published by spammers. We randomly select 10,000 pairs of such attributes (i.e.,

[1] The code can be downloaded from http://www.acsu.buffalo.edu/fenglong/.

[2] Note that this dataset is only used to complete the following relationships as [18].