PUD: Social Spammer Detection Based on PU Learning

Conference Paper · October 2017
DOI: 10.1007/978-3-319-70139-4_18

6 authors, including:

Yuqi Song
Chongqing University
8 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Junliang Yu
Chongqing University
9 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Wentao Li
University of Technology Sydney
10 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Qingyu Xiong
Chongqing University
30 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Neural Information Processing	
Series Title		
Chapter Title	PUD: Social Spammer Detection Based on PU Learning	
Copyright Year	2017	
Copyright HolderName	Springer International Publishing AG	
Author	Family Name	**Song**
	Particle	
	Given Name	**Yuqi**
	Prefix	
	Suffix	
	Division	Key Laboratory of Dependable Service Computing in Cyber Physical Society
	Organization	Chongqing University, Ministry of Education
	Address	Chongqing, China
	Division	School of Software Engineering
	Organization	Chongqing University
	Address	Chongqing, China
	Email	songyq@cqu.edu.cn
Corresponding Author	Family Name	**Gao**
	Particle	
	Given Name	**Min**
	Prefix	
	Suffix	
	Division	Key Laboratory of Dependable Service Computing in Cyber Physical Society
	Organization	Chongqing University, Ministry of Education
	Address	Chongqing, China
	Division	School of Software Engineering
	Organization	Chongqing University
	Address	Chongqing, China
	Email	gaomin@cqu.edu.cn
Author	Family Name	**Yu**
	Particle	
	Given Name	**Junliang**
	Prefix	
	Suffix	
	Division	Key Laboratory of Dependable Service Computing in Cyber Physical Society
	Organization	Chongqing University, Ministry of Education
	Address	Chongqing, China
	Division	School of Software Engineering

	Organization	Chongqing University
	Address	Chongqing, China
	Email	yu.jl@cqu.edu.cn
Author	Family Name	**Li**
	Particle	
	Given Name	**Wentao**
	Prefix	
	Suffix	
	Division	Faculty of Engineering and Information Technology, Centre for Artificial Intelligence, School of Software
	Organization	University of Technology Sydney
	Address	Ultimo, Australia
	Email	wentao.li@student.uts.edu.au
Author	Family Name	**Wen**
	Particle	
	Given Name	**Junhao**
	Prefix	
	Suffix	
	Division	Key Laboratory of Dependable Service Computing in Cyber Physical Society
	Organization	Chongqing University, Ministry of Education
	Address	Chongqing, China
	Division	School of Software Engineering
	Organization	Chongqing University
	Address	Chongqing, China
	Email	jhwen@cqu.edu.cn
Author	Family Name	**Xiong**
	Particle	
	Given Name	**Qingyu**
	Prefix	
	Suffix	
	Division	Key Laboratory of Dependable Service Computing in Cyber Physical Society
	Organization	Chongqing University, Ministry of Education
	Address	Chongqing, China
	Division	School of Software Engineering
	Organization	Chongqing University
	Address	Chongqing, China
	Email	xiong03@cqu.edu.cn

Abstract

Social networks act as the communication channels for people to share various information online. However, spammers who generate spam information reduce the satisfaction of common users. Numerous notable studies have been done to detect social spammers, and these methods can be categorized into three types: unsupervised, supervised and semi-supervised methods. While the performance of supervised and semi-supervised methods is superior in terms of detection accuracy, these methods usually suffer from the dilemma of imbalanced data since the labeled normal users are far more than spammers in real situations. To address the problem, we propose a novel method only relying on normal users to detect spammers. Firstly, a classifier is built from a part of normal and unlabeled samples to pick out reliable spammers from unlabeled samples. Secondly, our well-trained detector, which is based on the given normal users and

predicted spammers, can distinguish between normal users and spammers. Experiments conducted on real-world datasets show that the proposed method is competitive with supervised methods.

Keywords (separated by '-')	Spammer detection - Social network - PU Learning

PUD: Social Spammer Detection Based on PU Learning

Yuqi Song[1,2], Min Gao[1,2(✉)], Junliang Yu[1,2], Wentao Li[3], Junhao Wen[1,2], and Qingyu Xiong[1,2]

[1] Key Laboratory of Dependable Service Computing in Cyber Physical Society, Chongqing University, Ministry of Education, Chongqing, China
{songyq,gaomin,yu.jl,jhwen,xiong03}@cqu.edu.cn
[2] School of Software Engineering, Chongqing University, Chongqing, China
[3] Faculty of Engineering and Information Technology, Centre for Artificial Intelligence, School of Software, University of Technology Sydney, Ultimo, Australia
wentao.li@student.uts.edu.au

Abstract. Social networks act as the communication channels for people to share various information online. However, spammers who generate spam information reduce the satisfaction of common users. Numerous notable studies have been done to detect social spammers, and these methods can be categorized into three types: unsupervised, supervised and semi-supervised methods. While the performance of supervised and semi-supervised methods is superior in terms of detection accuracy, these methods usually suffer from the dilemma of imbalanced data since the labeled normal users are far more than spammers in real situations. To address the problem, we propose a novel method only relying on normal users to detect spammers. Firstly, a classifier is built from a part of normal and unlabeled samples to pick out reliable spammers from unlabeled samples. Secondly, our well-trained detector, which is based on the given normal users and predicted spammers, can distinguish between normal users and spammers. Experiments conducted on real-world datasets show that the proposed method is competitive with supervised methods.

Keywords: Spammer detection · Social network · PU Learning

1 Introduction

With the popularity of the social network, users are taking delight in sharing. For example, users can send tweets and comments on Twitter [1]. However, spammers are also planning to benefit from the prosperity by means of advertising, posting nonsenses and spreading fake information. Series of security risks may be caused due to spammers. For instance, users' privacy information can be filched by phishing links and the recommended lists are polluted by spam. Hence, spammer detection has become a significant work in social service.

D. Liu et al. (Eds.): ICONIP 2017, Part V, LNCS 10638, pp. 1–9, 2017.
https://doi.org/10.1007/978-3-319-70139-4_18

By now, social spammer detection has attracted extensive attention from researchers and the industry. Existing efforts are categorized into unsupervised methods, supervised methods, semi-supervised methods, etc. Unsupervised spammer detection methods [2–4] do not need the labeled samples, which can save the cost of labeling. But the absence of labels may lead to the low accuracy. In contrast, supervised methods [1,5–7] and semi-supervised [8,9] methods perform better than unsupervised methods with the supervision of the labels. However, these methods relying on both positive and negative labels fail when there are only one class labels available. In addition, it is time-consuming to label numerous spammers in real situations. In order to resolve this problem, we propose a novel spammer detection method based on Positive and Unlabeled Learning (PU Learning) [10], named PUD. At first, we build a reliable negative (RN) classifier from normal users and unlabeled samples. Then some reliable negative samples are picked out. Secondly, the positive and unlabeled detecting (PUD) classifier is trained on positive and reliable negative samples. The main contributions of this paper are as follows:

- Propose a novel method PUD to detect spammers in social network;
- Evaluate and compare the performance of the proposed PUD method on real-world datasets with supervised methods;
- Discuss the effect of the proportion of positive samples in PUD, which proves PUD can achieve well result merely rely on a few positive samples.

The remainder of this paper is structured as follows. In Sect. 2, we introduce some related work. The problem statement and the illustration of PUD method are shown in Sect. 3. In Sect. 4, we conduct experiments on two real-world datasets. Finally, Sect. 5 concludes this paper and point out the potential future work.

2 Related Work

In this section, we review some related work from current research about social spammer detection and background knowledge about PU Learning.

2.1 Social Spammer Detection Methods

Generally speaking, the notable detection methods can be classified into unsupervised methods, supervised methods and semi-supervised methods according to the amount of needed labeled data.

Unsupervised Detection methods mainly utilize the social network topology to identify the abnormal nodes. The method of combining social relation graphs and user link diagrams was proposed in [3]. Zhang et al. [4] adopted 12 types of topological features in ego network to detect spammers.

Supervised Detection methods usually extract relevant characteristics of users. Benevenuto et al. [1] extracted the user behavior characteristics and tweet

content characteristics to detect spammers. A group modeling framework was proposed in [7], which adaptively characterizes social interactions of spammers.

Semi-supervised Detection methods leverage labeled samples and massive unlabeled samples. A hybrid method that aimed to detect multiple spammers from user characteristics and user relationships was proposed in [8]. Li et al. [9] used the Laplace method to extract features, then used the semi-supervised method to train classifier.

Among these methods, Supervised methods outperform the unsupervised methods, but they need abundant labeled data. Semi-supervised methods require labeled and unlabeled data. Either supervised or semi-supervised methods rely on both positive and negative samples. Only a few positive labeled data and plenty of unlabeled data are required in our work.

2.2 Outline of PU Learning

The approach merely adopting positive and unlabeled data is called Positive and Unlabeled Learning or PU Learning. At the beginning, PU Learning mainly aimed to solve the task of text classification [10], then researchers extended this method to other areas. Such as the remote-sensing data classification, the disease gene identification, the Multi-graph learning, etc.

There are massive unlabeled user in real social networks, and the quantity of labeled spammers is much smaller than those of normal users. Furthermore, the cost of marking normal users is cheaper than marking spammers. These characteristics show that PU Learning can be applied in real situations.

PU Learning mainly consists of two steps [10]. Step 1: Identify the reliable negative samples (RN) from the unlabeled samples (U) according to the positive samples (P). Step 2: Construct the binary classifier by positive samples and reliable negative samples.

3 PUD Method

In this section, we will first state the problem of social spammer detection formally. Next, the main steps of the PUD method will be illustrated.

3.1 Problem Statement

Let $\mathbf{X} \in \mathbb{R}^{n \times t}$ be the t features of n users in a social network, and $\mathbf{Y} \in \{0, 1\}^n$ are corresponding labels of users, where $y_i = 0$ indicates the i^{th} account is a spammer and equals to 1 otherwise. U, P, RN represent the unlabeled samples, positive samples and reliable negative samples, respectively. Meanwhile μ, l, r represent the amount of users in the corresponding samples.

The task of the spammer detection can be summarized as follows: Given the features for all n instances and some positive labels, learning a model PUD with well performance to classify an unknown account.

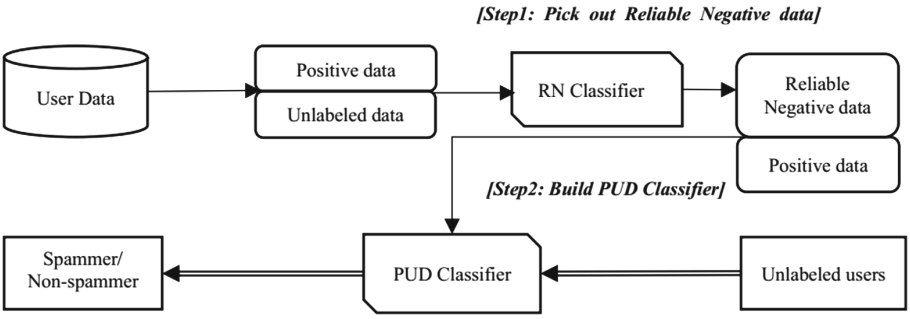


Fig. 1. The framework of PUD

3.2 PUD Framework

The framework of our proposed method consists of two steps, as described in Fig. 1, and each step will be illustrated in detail.

Step 1: Pick out Reliable Negative Samples. Picking out the reliable negative samples is a critical in PU Learning. Theoretically, maximizing the confidence of the negative samples and ensuring the positive samples are correctly classified, we can get a superior classifier [10]. Therefore, it is vital to find as many reliable negative samples as possible in the unlabeled dataset. In the following, we will describe the algorithm more specifically.

In our method, the reliable negative classifier is constructed by Naive Bayes, because it is a mature and popular classified algorithm, while other algorithms are alternative. Naive Bayes learns the joint probability distribution $P_r(X, Y)$ from the training dataset. Before that, it needs to learn priori probability distribution in Eq. (1) and conditional probability distribution in Eq. (2).

$$P_r(Y = c_k), \quad k = 0, 1 \tag{1}$$

$$P_r\left(X = x \,|\, Y = c_k\right) = P_r\left(X^{(1)} = x^{(1)}, \ldots, X^{(d)} = x^{(d)} \,|\, Y = c_k\right), \quad k = 0, 1 \tag{2}$$

where c_0 and c_1 denote the labels of positive samples and unlabeled samples, respectively. And then the joint probability distribution $P_r(X, Y)$ is learnt by Eqs. (1) and (2).

Given X, a set of user features, Naive Bayes algorithm calculates the posterior probability distribution in Eq. (3) by the learnt model. The label of x is the one having the highest posterior probability.

$$P_r\left(Y = c_k \,|\, X = x\right) = \frac{P_r\left(X = x \,|\, Y = c_k\right) P_r\left(Y = c_k\right)}{\sum_k P_r\left(X = x \,|\, Y = c_k\right) P_r\left(Y = c_k\right)} \tag{3}$$

The reliable negative classifier can be define as

$$y = f(x) = \arg\max_{c_k} P_r\left(Y = c_k\right) \prod_{j=1}^{d} P_r\left(X^{(j)} = x^{(j)} \,|\, Y = c_k\right). \tag{4}$$

The process of identifying the reliable negative samples RN from the positive sample P and unlabeled samples U is as follows: first of all, we assign each normal user label 1 to constitute P and some unlabeled users label 0 to form U. Secondly, RN classifier is learnt from αP and βU by Naive Bayes, where l is the amount of αP and r is the amount of RN. We set $\beta = 0.5$, because training and predicting both need plenty of unlabeled samples. Note that, α is an important parameter will be discussed in experiment. Thirdly, we exploit the classifier to identify other unlabeled users, $(1-\beta)U$. Finally, reliable negative users are pick out from $(1-\beta)U$ until $r = l$, whose predicted labels are spammer. These reliable negative samples will be utilized in PUD classifier.

Step 2: Build PUD Classifier. A binary classifier is build in step 2. from positive and reliable negative samples by Random Forest algorithm to detect spammers. Random Forest is an ensemble algorithm that constructs a multitude of many decision trees at training time and outputs the class that is the mode of the classification of the individual tress. It is efficient for estimating missing data and maintains accuracy when a large proportion of the data are missing. Thus, Random Forest meets the requirements for our methods.

The procedures are as follows: firstly, the PUD classifier is trained by the predicted negative samples RN from RN classifier and given positive samples P. Then the PUD classifier can be utilized to detect spammers: the user is a spammer if the predicted label is negative, otherwise the user is legitimate.

The complete process of PUD method which integrates step 1 and step 2 is shown in Table 1.

Table 1. The complete process of PUD method

Input:
User Feature Matrix $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_n\} \in \mathbb{R}^{t \times n}$
User Labels $\mathbf{Y}$
Parameter α, β
Output:
A spammer detection classifier PUD

Step:
$P = \emptyset, \quad U = \emptyset, \quad RN = \emptyset$
for $\mathbf{x}_i$
    if $\mathbf{y}_i == 1$
        $P = P \cup \mathbf{x}_i$
    else
        $U = U \cup \mathbf{x}_i$
RN$\leftarrow clf.learn(\alpha P, \beta U)$
RN.predict$((1-\beta)U)$
while $r < l$
    $RN = RN \cup \{\ predict == 0\ \}$
PUD $\leftarrow clf.learn(P, RN)$

4 Experiments

In this section, we conduct experiments to evaluate the effectiveness of the proposed PUD method. We first introduce the datasets and metrics. Then we compare the performance of our method with other detection methods. Finally the sensitivity of parameter α will be discussed.

4.1 Datasets and Metrics

Two real datasets provided by Benevenuto [1,5] are used for evaluation. The one is from YouTube [5], includes 188 spammers and 641 legitimate users. Each user has 60 features which are derived from video attributes, individual characteristics of user behavior, and node attributes. The other is from Twitter [1]. This dataset contains 1650 labeled users, 355 of them are spammers. Each user has 62 features which are derived from tweet content and user social behavior.

The experiments are conducted by 5-fold cross validation 10 times, and average value are used to represent the results. We adopt the three frequent used evaluation metrics, i.e., *Precision*, *Recall* and *F-measure* for performance evaluation.

4.2 Experimental Results

Table 2 reported the performance of PUD method on both datasets. We apply Naive Bayes algorithm to pick out reliable negative samples on YouTube dataset while Logistic Regression is utilized on Twitter. The results show the validity of PUD and prove it is a general and base method.

Table 2. Performance of PUD

	Precision	Recall	F-measure
YouTube	0.786	0.662	0.71
Twitter	0.85	0.69	0.756

In order to further show our proposed method has competitive performance, it is compared with traditional supervised methods which exploit various proportion of labeled spammer in training. Traditional methods include Naive Bayes (NB), Logistic Regression (LR), Decision tree (DT), Random Forest (RF) and Gradient Boosting Decision Tree (GBDT). The results of different methods are displayed in Table 3, and we bold the best values in each dataset.

Based on the results, we make following observations. Firstly, the F-measure of PUD is quite close to the best values in Twitter while Random Forest and Gradient Boosting Decision Tree both need 30% labeled spammers. In YouTube, the F-measure of our method can reach to 71%, it increases over 4.7% than other methods. Secondly, it can be seen that PUD are superior to tradition methods whose labeled spammers are less than 20%. Therefore, the proposed method can relieve the dilemma of imbalanced data.

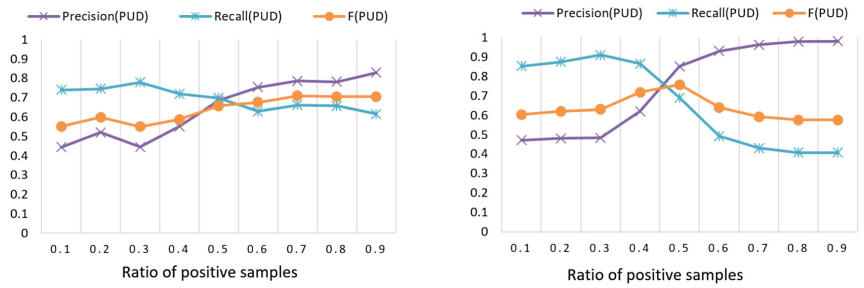
Table 3. F-measure comparison between PUD and other methods

	Spammer ratio	LR	NB	DT	RF	GBDT	PUD
YouTube	0%	\	\	\	\	\	**0.71**
	1%	0.232	0.269	0.218	0.27	0.25	\
	2%	0.262	0.314	0.246	0.276	0.262	\
	5%	0.39	0.418	0.422	0.53	0.37	\
	10%	0.416	0.432	0.538	0.624	0.478	\
	20%	0.542	0.434	0.618	0.65	0.562	\
	30%	0.644	0.44	0.646	0.678	0.674	\
Twitter	0%	\	\	\	\	\	0.768
	1%	0.214	0.14	0.376	0.24	0.38	\
	2%	0.296	0.21	0.558	0.45	0.548	\
	5%	0.35	0.426	0.644	0.612	0.586	\
	10%	0.36	0.49	0.69	0.706	0.654	\
	20%	0.38	0.51	0.71	0.736	0.72	\
	30%	0.45	0.542	0.716	0.776	**0.78**	\

4.3 Parametric Sensitivity Analysis

Now, we discuss the sensitivity of the parameter α which determines the proportion of positive samples chosen. The experimental results are shown in Fig. 2.

Figure 2(a) shows the fluctuant performance of PUD with the different values of α on the YouTube dataset. It can be observed that the precision increases while the recall reduces as a result of imbalanced data. In order to balance the performance of PUD, we take $\alpha = 0.7$ in experiment, and then the F-measure can reach the optimal state. Figure 2(b) shows the performance on Twitter, and α is set to 0.5 to make the precision and recall balance in experiment.



(a) YouTube (b) Twitter

Fig. 2. Performance of PUD with varying α on datasets

In summary, the proposed method is not always outstanding in F-measure compared with supervised methods, but it can achieve competitive performance without labeled spammers. In addition, the effect of the parameter is analyzed as well. It proves that PUD can get ideal result merely using a few positive samples which reduces the cost of labeling.

5 Conclusion and Future Work

In this paper, we proposed a novel method PUD based on PU Learning, it aims to construct a detection classifier by a few positive samples and plenty of unlabeled data. Our method includes two steps: at first, we pick out reliable negative samples from unlabeled users. After that, the PUD classifier is trained by positive and reliable negative samples. Experimental results on the two real-world datasets show that our approach has competitive performance and prove it is a general and base method. Furthermore, PUD shows its merits in detecting spammers. Thus the proposed method can be applied extensively.

A few possible works remain to be done. We will combine PUD with various state-of-the-art supervised methods to improve the accuracy of spammer detection. Besides, our method can be used to detect fake comments in social networks.

Acknowledgments. The work is supported by the Basic and Advanced Research Projects in Chongqing under Grant No. cstc2015jcyjA40049, the National Key Basic Research Program of China (973) under Grant No. 2013CB328903, the Guangxi Science and Technology Major Project under Grant No. GKAA17129002, and the Graduate Scientific Research and Innovation Foundation of Chongqing, China under Grant No. CYS17035.

References

1. Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detecting spammers on Twitter. In: Collaboration, Electronic Messaging, Anti-abuse and Spam Conference (CEAS), vol. 6, p. 12 (2010)
2. Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., Zhao, B.Y.: Detecting and characterizing social spam campaigns. In: Proceedings of 10th ACM SIGCOMM conference on Internet measurement, pp. 35–47. ACM (2010)
3. Tan, E., Guo, L., Chen, S., Zhang, X., Zhao, Y.: Unik: unsupervised social network spam detection. In: Proceedings of 22nd ACM international conference on Information & Knowledge Management, pp. 479–488. ACM (2013)
4. Zhang, B., Qian, T., Chen, Y., You, Z.: Social spammer detection via structural properties in ego network. In: Li, Y., Xiang, G., Lin, H., Wang, M. (eds.) SMP 2016. CCIS, vol. 669, pp. 245–256. Springer, Singapore (2016). doi:10.1007/978-981-10-2993-6_21
5. Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., Gonçalves, M.: Detecting spammers and content promoters in online video social networks. In: Proceedings of 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 620–627. ACM (2009)

6. Hu, X., Tang, J., Zhang, Y., Liu, H.: Social spammer detection in microblogging. In: IJCAI, vol. 13, pp. 2633–2639. Citeseer (2013)
7. Wu, L., Hu, X., Morstatter, F., Liu, H.: Adaptive spammer detection with sparse group modeling. In: ICWSM, p. 319–326 (2017)
8. Wu, Z., Wang, Y., Wang, Y., Wu, J., Cao, J., Zhang, L.: Spammers detection from product reviews: a hybrid model. In: 2015 IEEE International Conference on, Data Mining (ICDM), pp. 1039–1044. IEEE (2015)
9. Li, W., Gao, M., Rong, W., Wen, J., Xiong, Q., Ling, B.: LSSL-SSD: social spammer detection with laplacian score and semi-supervised learning. In: Lehner, F., Fteimi, N. (eds.) KSEM 2016. LNCS, vol. 9983, pp. 439–450. Springer, Cham (2016). doi:10.1007/978-3-319-47650-6_35
10. Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.S.: Building text classifiers using positive and unlabeled examples. In: 3rd IEEE International Conference on Data Mining, ICDM 2003, pp. 179–186. IEEE (2003)