PUED: A Social Spammer Detection Method Based on PU Learning and Ensemble Learning

Yuqi Song[1,2], Min Gao[1,2(✉)], Junliang Yu[1,2], Wentao Li[3], Lulan Yu[1,2], and Xinyu Xiao[1,2]

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

Abstract. In social network, people generally tend to share information with others, thus, those who have frequent access to the social network are more likely to be affected by the interest and opinions of other people. This characteristic is exploited by spammers, who spread spam information in network to disturb normal users for interest motives seriously. 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 number of unlabeled normal users is far more than spammers' in real situations. To address the problem, we propose a novel method only relying on normal users to detect spammers exactly. We present two steps: one picks out reliable spammers from unlabeled samples which is imposed on a voting classifier; while the other trains a random forest detector from the normal users and reliable spammers. We conduct experiments on two real-world social datasets and show that our method outperforms other supervised methods.

Keywords: Spammer detection · Social network · PU Learning Ensemble Learning

1 Introduction

With the rapid development of internet, social network has become an excellent medium for both sharing information and delivering products and services. It is hardly a surprise that online users of social network are growing exponentially every year, and people incline to make their commercial decisions after checking

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
I. Romdhani et al. (Eds.): CollaborateCom 2017, LNICST 252, pp. 143–152, 2018.
https://doi.org/10.1007/978-3-030-00916-8_14

the online reviews. For example, users of YouTube tend to choose something to watch according to others' ratings and comments. However, in normal cases, social network is vulnerable to malicious information propagated by some users with special purpose [1,2]. Spammers as they were called, plan to benefit from advertising, posting nonsenses and spreading fake information. The existence of such spammers breaks down the ecological environment of network and affect the user experience of genuine users. Moreover, a diverse array of security risks might be caused as well, for instance, users' privacy information may be filched by phishing links and the recommended lists may be contaminated by spam. Hence, spammer detection has become a much-needed task in social service.

To the best of our knowledge, social spammer detection has attracted extensive attention from both academia and industry. As has been studied in previous literatures, spammer detections are categorized into unsupervised methods, supervised methods, and semi-supervised methods, etc. Unsupervised spammer detection methods [3–5] do not need the labeled samples, which can cut down the cost of labeling. But the absence of labels may lead to the low accuracy of the detection result. In contrast, supervised methods [6–8] and semi-supervised [9–11] methods perform better than unsupervised methods with the supervision of the labels. However, they might be exposed to a highly-risky situation where only one class label is available, because these methods highly rely on both positive and negative labels. In addition, it is time-consuming to label numerous spammers in real situations. In this work, we propose a novel spammer detection method based on Positive and Unlabeled Learning (PU Learning) [12] and Ensemble Learning [13], named PUED. The procedure of PUED contains two phases: first, a voting classifier is trained to pick out reliable negative samples (RN) from unlabeled samples; second, the detection classifier is constructed from positive and reliable negative samples. The main contributions of this paper are as follows:

- Propose a novel method PUED to detect spammers in social network;
- Evaluate and compare the performance of the PUED on two real-world datasets with supervised methods;
- Discuss the effect of the proportion of positive samples in our method, and demonstrate that PUED is capable of discriminating spammer effectively through merely a few positive samples.

The remainder of this paper is organized as follows. Section 2 provides related work. The problem statement and the illustration of PUED method are shown in Sect. 3. In Sect. 4, we conduct experiments on two real-world datasets. Finally, Sect. 5 concludes this research with potential direction for future work.

2 Related Work

2.1 Social Spammer Detection Methods

Generally speaking, the notable social spammer detection methods can be classified into three methods on the basis of labeled data as follows.

Unsupervised Detection methods mainly utilize the social network topology to identify the abnormal nodes. Gao et al. [3] exploited similarities of text content and URLs to cluster users in Facebook. The method of combining social relation graphs and user link diagrams was proposed in [4]. Zhang et al. [5] adopted 12 types of topological features in ego network to detect spammers.

Supervised Detection methods usually extract relevant characteristics of users. Benevenuto et al. [6] extracted the user behavior characteristics and tweet content characteristics to detect spammers. Wei et al. [7] explored characteristics of spammers and network stability on Twitter. A group modeling framework was proposed in [8], 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 [9]. In [10], the trust propagation which utilized PageRank to propagate labels was used to recognize spammers. Li et al. [11] used the Laplace method to extract features.

Among these methods, supervised methods outperform the unsupervised methods, but they are limited by abundant labeled data while unsupervised methods suffer from low accuracy. Semi-supervised methods also need a part of labeled data. Either supervised or semi-supervised methods depend on both positive and negative samples. In our work, only a few positive labeled data and plenty of unlabeled data are exploited in particular.

2.2 PU Learning

The approach merely utilizing positive and unlabeled data is called Positive and Unlabeled Learning or PU Learning. In the initial research, PU Learning mainly aimed at the text classification [12], then researchers applied this method to other field, such as the web page classification, the disease gene identification, and the Multi-graph learning.

PU Learning mainly consists of two steps [12]. 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.

In real-world situation, despite that there are enormous unlabeled users and a large scale of the labeled normal users, the number of labeled spammers is still quite small. In addition, the expense of manually marking spammer is exceedingly higher than labeling normal user. Compared with the traditional methods, PU Learning has definite advantage whether in labeling time, labor force or the amount of labeled samples.

2.3 Ensemble Learning

Ensemble Learning [13,14], which integrates multiple learning algorithms, is a powerful method to obtain better performance than one learning classifier. Currently, it is almost being used in every latest research, from text mining to image

processing. Commonly-used Ensemble Learning techniques include Bagging and Boosting [15].

Bagging tries to implement similar learners on small sample populations and then takes the mean value of all the predictions. In generalized bagging, we can use different learners on different populations to reduce the variance error. As a most common example, the Random Forest (RF) algorithm integrates bagging with random decision trees to make great progress in accuracy.

Boosting is an iterative method which adjusts the weight of an observation based on the last classification. If a sample was discriminated wrong, the method would increase the weight of the sample and vice versa. In general, Boosting decreases the bias error and builds strong predictive models, however, they may sometimes over fit on the training data. During the many typical algorithm of Boosting, Adaboost is a frequently-used one and Gradient Boosting Decision Tree (GDBT) is a novel one which achieve better performance.

3 PUED Method

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 those users. $y_i = 0$ indicates that the i^{th} user account is a spammer and $y_i = 1$ otherwise. U, P, RN represent the unlabeled samples, positive samples and reliable negative samples, respectively. Meanwhile μ, l, r represent the number of users in the corresponding samples. Since only a few positive samples and plenty of unlabeled samples are used, we assume that $l \ll \mu$. In order to balance the scale of P with RN, we set $r \approx l$.

The task of the spammer detection can be summarized as follows: Given the features of all n instances and some positive labels, learning a model PUED with well performance, and then classifying the unknown user account.

3.2 PUED Framework

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

Step1: Pick out Reliable Negative Samples Recursively. Picking out reliable negative samples well and truth is a critical stage in PU Learning. Theoretically, after maximizing the confidence of the negative samples and ensuring the positive samples are correctly classified, we can get a superior classifier. It is vital to find as many reliable accurately-classified negative samples as possible from the unlabeled dataset. So we utilize Ensemble Learning to construct a multi-classifier. The common combination strategy of bagging for classification task is voting. More specifically, the trained classifier h_i predict a tag from the label set $\{C_1, C_2, ...C_N\}$ or the unlabeled sample $\mathbf{x}_i$. And then the predictive output on the sample $\mathbf{x}_i$ is expressed as an N dimensional vector $(h_i^1(x), h_i^2(x), ..., h_i^N(x))$, where $h_i^j(x)$ is the output of h_i on the label C_j. The absolute-majority-voting

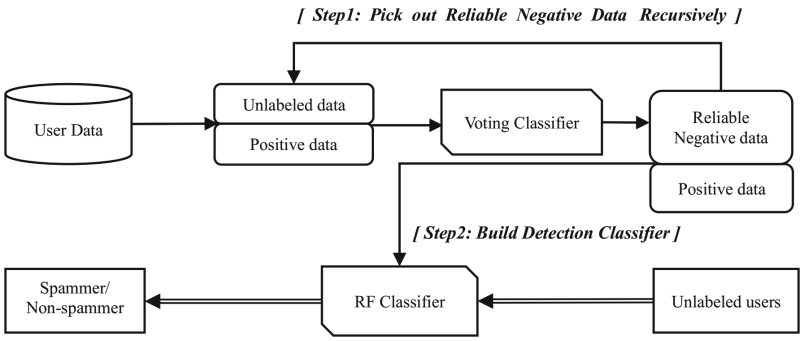


Fig. 1. The framework of PUED

accepts the predicted label whose occupancy is more than half, otherwise it rejects the prediction, as is shown in Eq. (1).

$$f(n) = \begin{cases} c_j, & if \sum_{i=1}^{T} h_i^j(x) > 0.5 \sum_{k=1}^{N} \sum_{i=1}^{T} h_i^k(x); \\ reject, & otherwise \end{cases} \tag{1}$$

In this work, five sub-classifiers predict the 1 or 0 label of unlabeled samples, and then vote for spammer label. The accepted threshold of spammer increases to 0.75 for higher precision, in Eq. (2). The five sub-classifiers include Logistic Regression classifier, Naive Bayes classifier, Decision Tree classifier, Random Forest Classifier and Gradient Boosting Decision Tree classifier. The powerful effect of the ensemble method will be presented in our experiment.

$$f(n) = \begin{cases} 0, if \sum_{i=1}^{5} h_i^0(x) > 0.75 \sum_{k=0}^{1} \sum_{i=1}^{5} h_i^0(x); \\ 1, otherwise \end{cases} \tag{2}$$

Note that, each new predicted spammer would be added into the training set and these spammers form the reliable negative group.

Step2: Build Detection Classifier. A binary classifier is built in step 2, which differentiates between normal users and spammers by Random Forest algorithm. Random Forest is an ensemble algorithm which constructs a multitude of many decision trees at training time and outputs the class and it has a slight advantage in the individual tress. It can fill the gap in unbalanced data and maintain accuracy.

The procedures are as follows: firstly, the classifier is trained by the reliable negative samples and positive samples. Then the detection classifier can be utilized to distinguish the labels of samples: the user is a spammer if the predicted label is negative, otherwise the user is legitimate.

The whole process of PUED method which combines step 1 and step 2 is shown in Table 1. The parameter α determines the quantity of positive samples, and we will analyze it in the experiment part. The parameter β means the proportion of unlabeled samples in training stage, it is set to 0.5 in our work.

Table 1. The complete Process of PUED 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 PUED

Step:
$P = \emptyset, \quad U = \emptyset, \quad RN = \emptyset$
for $\mathbf{x}_i$ in trainingSet
  if $\mathbf{y}_i == 1$
    $P = P \cup \mathbf{x}_i$
  else
    $U = U \cup \mathbf{x}_i$
Vote $\leftarrow mulclf.learn(\alpha P, \beta U)$
for $\mathbf{x}_i$ in $(1 - \beta)U$
  if Vote.predict($\mathbf{x}_i$) == 0
    $\beta U = \beta U \cup \mathbf{x}_i$
    Vote $\leftarrow mulclf.learn(\alpha P, \beta U)$
    $RN = RN \cup \mathbf{x}_i$
PUED $\leftarrow rfclf.learn(P, RN)$
for $\mathbf{x}_i$ in testSet
  userLabel = PUED.predict ($\mathbf{x}_i$)

4 Experiments

4.1 Datasets and Metrics

Two real datasets provided by Benevenuto [6,16] were used for evaluation. The one is from Twitter [6] which contains 1650 labeled users, and 355 spammers in those labeled users. Each user has 62 features which are derived from tweet content and user social behavior. The other one is from YouTube [16], which 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 experiments were conducted by 5-fold cross validation 10 times, where an average values of each set of trials were generated to represent the final results. We adopt the three frequently-used evaluation metrics, i.e., *Precision*, *Recall* and *F-measure* for performance evaluation.

4.2 Experimental Results

Credibility of Reliable Negative Samples. To draw statistical valid conclusions, we implemented several traditional methods to carry out the voting in PUED, and otherwise compare its results with each of them.

Such traditional methods include Naive Bayes (NB), Logistic Regression (LR), Decision tree (DT), Random Forest (RF) and Gradient Boosting Decision Tree (GBDT). Precision, Recall and F-measure should all be taken into consideration when judging the credibility of classifying work. And the parameter is set as 0.5, which will be explained later. Table 2 reports the credibility of reliable negative samples on both datasets. The best values are bolded in each dataset. In Twitter dataset, the precision reached 0.876 and F-measure achieved 0.826 of PUED method while sacrificing a little recall. Similarly, in YouTube dataset, precision and F-measure of our method were higher than those generated by any other single classifier. Generally, the experimental results verify the validity of Ensemble Learning. In addition, the obtained higher credibility guarantee the accuracy in the next step of our experiment to a certain extent.

Table 2. Credibility of Reliable Negative Samples

	Metrice	LR	NB	DT	RF	GBDT	PUED
Twitter	Precision	0.722	0.612	0.654	0.662	0.864	**0.876**
	Recall	0.352	0.45	0.79	**0.864**	0.746	0.792
	F-measure	0.46	0.384	0.716	0.75	0.792	**0.826**
YouTube	Precision	0.702	0.686	0.516	0.53	0.81	**0.862**
	Recall	0.756	0.624	0.74	**0.87**	0.689	0.688
	F-measure	0.722	0.618	0.608	0.75	0.75	**0.76**

Compare PUED with Other Methods. To further demonstrate that our proposed method has competitive performance, we especially compared the F-measure results of PUED with the same five traditional supervised methods we used in the last experiment, which exploit various proportion of labeled spammers in training. Similarly, two sets of trials on two datasets are conducted. In each sets of trials, we set the spammer ratio as 0%, 1%, 2%, 5%, 10%, 20%, and 30% respectively. The results of different methods are described in Table 3, with the best value in each dataset bolded as well. Note that, the results of PUED did not change at all, due to the fact that no labeled spammer was used in our method.

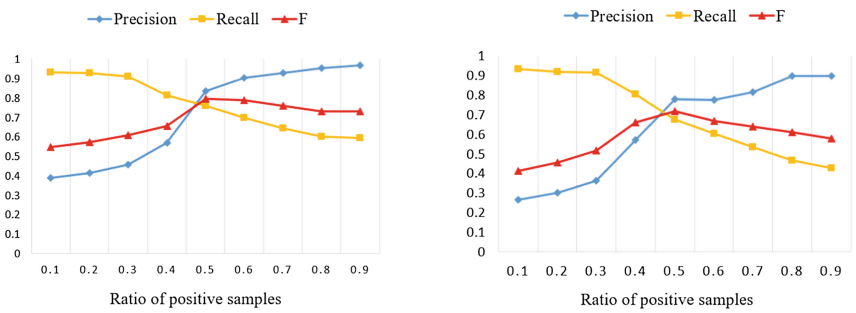
According to the experimental results, we could make several conclusions. First of all, in comparison with tradition methods, the F-measure of PUED, reaching 0.795, outperformed any other ones in Twitter dataset; likewise, increasing at least by 6.2% (compared with the best performer RF) in YouTube. Secondly, since the performance of supervised classifiers highly depended on labeled spammers, when the proportion of labeled spammers was low, they almost did not work. Thirdly, PUED, who only utilized positive samples, significantly outperforms other traditional methods whose labeled spammers are less than 30% in both datasets. Therefore, facing with the dilemma of imbalanced data in supervised learning, our proposed method can effectively address the problem.

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

	Spammer ratio	LR	NB	DT	RF	GBDT	PUED
Twitter	0%	0	0	0	0	0	**0.795**
	1%	0.214	0.14	0.376	0.24	0.38	0.795
	2%	0.296	0.21	0.558	0.45	0.548	0.795
	5%	0.35	0.426	0.644	0.612	0.586	0.795
	10%	0.36	0.49	0.69	0.706	0.654	0.795
	20%	0.38	0.51	0.71	0.736	0.72	0.795
	30%	0.45	0.542	0.716	0.776	0.78	0.795
YouTube	0%	0	0	0	0	0	**0.72**
	1%	0.232	0.269	0.218	0.27	0.25	0.72
	2%	0.262	0.314	0.246	0.276	0.262	0.72
	5%	0.39	0.418	0.422	0.53	0.37	0.72
	10%	0.416	0.432	0.538	0.624	0.478	0.72
	20%	0.542	0.434	0.618	0.65	0.562	0.72
	30%	0.644	0.44	0.646	0.678	0.674	0.72

4.3 Parametric Sensitivity Analysis

In this subsection, we will discuss the sensitivity of the parameter α which determines the proportion of positive samples chosen. The experimental results on both datasets are shown in Fig. 2.



(a) Twitter (b) YouTube

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

Figure 2(a) shows the fluctuant performance of PUED with the different values of α in the Twitter dataset. Note that, with the increasing of α, the number of positive samples became large. It can be observed that the precision increases while the recall reduces as a result of existing imbalanced data. In order to

balance the performance, we took $\alpha = 0.5$ in experiment, where F-measure can reach the optimal state. Figure 2(b) shows the performance in YouTube, and thus, α was set as 0.5 to make the precision and recall balance in experiment as well.

In summary, our proposed method PUED significantly outperforms several state-of-the-art supervised methods in credibility and F-measure. To the best of our knowledge, it can achieve competitive performance without sufficient labeled spammers. In addition, we studied the performance tradeoffs in various schemes of parameter, and an optimization was obtained. Experimental studies indicate that PUED can obtain favorable result merely using a few positive samples, significantly reducing the cost of labeling.

5 Conclusion and Future Work

In this paper, we propose a novel method PUED which integrates PU Learning and Ensemble Learning. It aims to construct a detection classifier under the circumstances of a few positive samples and sufficient unlabeled data. In general, PUED is consist of two steps: (1) picking out reliable negative samples from unlabeled users with the voting strategy; (2) utilizing the Random Forest classifier which is trained from positive and reliable negative samples to distinguish spammer. Experimental results on the two real-world datasets demonstrate that our approach, as a general and base method, has highly competitive performance. Furthermore, PUED shows its computational merits in detecting spammers. Thus, the proposed method has a reasonable overhead in recognizing spammers in social networks. This provides the foundation for further enhancement in terms of improving its accuracy, combining PUED with various state-of-the-art supervised methods, and detecting spurious comments.

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. Hu, X., Tang, J., Zhang, Y., Liu, H.: Social spammer detection in microblogging. In: IJCAI, vol. 13, pp. 2633–2639 (2013). Citeseer
2. Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R.: Exploiting burstiness in reviews for review spammer detection. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, pp. 175–184. AAAI (2013)
3. Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., Zhao, B.Y.: Detecting and characterizing social spam campaigns. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, pp. 35–47. ACM (2010)

4. Tan, E., Guo, L., Chen, S., Zhang, X., Zhao, Y.: UNIK: unsupervised social network spam detection. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 479–488. ACM (2013)
5. 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). https://doi.org/10.1007/978-981-10-2993-6_21
6. 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)
7. Wei, W., Joseph, K., Liu, H., Carley, K.M.: Exploring characteristics of suspended users and network stability on twitter. Soc. Netw. Anal. Mining **6**(1), 51 (2016)
8. Wu, L., Hu, X., Morstatter, F., Liu, H.: Adaptive spammer detection with sparse group modeling (2017)
9. 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)
10. Li, Z., Zhang, X., Shen, H., Liang, W., He, Z.: A semi-supervised framework for social spammer detection. In: Cao, T., Lim, E.-P., Zhou, Z.-H., Ho, T.-B., Cheung, D., Motoda, H. (eds.) PAKDD 2015. LNCS (LNAI), vol. 9078, pp. 177–188. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18032-8_14
11. 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 (LNAI), vol. 9983, pp. 439–450. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47650-6_35
12. Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.S.: Building text classifiers using positive and unlabeled examples. In: Third IEEE International Conference on Data Mining, ICDM 2003, pp. 179–186. IEEE (2003)
13. Polikar, R.: Ensemble learning. In: Zhang, C., Ma, Y. (eds.) Ensemble Machine Learning, pp. 1–34. Springer, Heidelberg (2012)
14. Sun, Y., Tang, K., Minku, L.L., Wang, S., Yao, X.: Online ensemble learning of data streams with gradually evolved classes. IEEE Trans. Knowl. Data Eng. **28**(6), 1532–1545 (2016)
15. Bühlman, P.: Bagging, boosting and ensemble methods. In: Gentle, J., Härdle, W., Mori, Y. (eds.) Handbook of Computational Statistics, pp. 985–1022. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-21551-3_33
16. Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., Gonçalves, M.: Detecting spammers and content promoters in online video social networks. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 620–627. ACM (2009)