Title: Addressing the class imbalance problem in twitter spam detection using ensemble learning

Author: Shigang Liu, Yu Wang, Jun Zhang, Chao Chen, Yang Xiang

PII: S0167-4048(16)30175-4
DOI: http://dx.doi.org/doi: 10.1016/j.cose.2016.12.004
Reference: COSE 1076

To appear in: *Computers & Security*

Please cite this article as: Shigang Liu, Yu Wang, Jun Zhang, Chao Chen, Yang Xiang, Addressing the class imbalance problem in twitter spam detection using ensemble learning, *Computers & Security* (2016), http://dx.doi.org/doi: 10.1016/j.cose.2016.12.004.

Addressing the Class Imbalance Problem in Twitter Spam Detection Using Ensemble Learning

Shigang Liu, Yu Wang, Jun Zhang, Chao Chen, and Yang Xiang
School of Information Technology, Deakin University, Geelong, Australia
{shigang, y.wang, jun.zhang, chao.chen, yang}@deakin.edu.au

Abstract

In recent years, microblogging sites like Twitter have become an important and popular source for real-time information and news dissemination, and they have become a prime target of spammers inevitably. A series of incidents have showed that the security threats caused by Twitter spam can reach far beyond the social media platform to impact the real world. To mitigate the threat, a lot of recent studies apply machine learning techniques to classify Twitter spam and promising results are reported. However, most of these studies overlook the class imbalance problem in real-world Twitter data. In this paper, we experimentally demonstrate that the unequal distribution between spam and non-spam classes has a great impact on spam detection rate. To address the problem, we propose FOS, a fuzzy-based oversampling method that generates synthetic data samples from limited observed samples based on the idea of fuzzy-based information decomposition. Moreover, we develop an ensemble learning approach that learns more accurate classifiers from imbalanced data in three steps. In the first step, the class distribution in the imbalanced data set is adjusted by using various strategies, including random oversampling, random undersampling and FOS. In the second step, a classification model is built upon each of the redistributed data sets. In the final step, a majority voting scheme is introduced to combine the predictions from all the classification models. We conduct experiments real-world Twitter data for the purpose of evaluation. The results indicate that the proposed learning approach can significantly improve the spam detection rate in data sets with imbalanced class distribution.

Keywords

Online Social Networks; Twitter; Spam; Machine Learning; Class Imbalance.

1. Introduction

Twitter has gained significantly in popularity in recent years and become an important source for real-time information sharing and news dissemination. Inevitably, the growth of Twitter is accompanied by a significant increase of spamming activities targeting on the platform. Twitter spam is usually referred to as the unsolicited tweets that contain malicious links directing victims to external sites with malware downloads, phishing, drug sales, or scams, etc. [1]. A series of incidents showed that Twitter spam not only affects user experience, but also poses significant threats of damages beyond the social networking platform itself. As an example, in September 2014, a nationwide Internet meltdown in New Zealand was caused by a Twitter spam campaign that spread DDoS attack malware in the guise of leaked nude photos of Hollywood celebrities [2].