Postprint: XGBoost-Based Anomalous User Detection Technology in Social Networks

Authors: Yuan Lixin, Gu Yijun, Zhao Dapeng

Date: 2019-01-03T00:00:00+00:00

Abstract

To address the problems of low recall rate and low operational efficiency when traditional social network anomalous user detection algorithms are applied to real-world imbalanced datasets, we extract user content, behavior, attribute, and relationship features from social network datasets, apply the gradient boosting ensemble classifier XGBoost algorithm for feature selection, establish a classification model, construct imbalanced datasets, and identify three categories of spam advertising accounts. Experimental results show that, compared with traditional classification methods such as Random Forest, the proposed method achieves effective improvements in both recall rate and F1-value for anomalous user detection on both balanced and imbalanced datasets; furthermore, selecting only a small subset of features can still attain a high detection level, thereby demonstrating the effectiveness of the method.

Full Text

Preamble

Research on Abnormal User Detection Technology in Social Networks Based on XGBoost Method

Yuan Lixin, Gu Yijun†, Zhao Dapeng
(School of Information Technology & Network Security Enforcement, People' s Public Security University of China, Beijing 102600, China)

Abstract: Traditional social network abnormal user detection algorithms suffer from low recall rates and poor operational efficiency when applied to real-world imbalanced datasets. This study extracts user content, behavior, attribute, and relationship features from social network datasets, applies the gradient boosting ensemble classifier XGBoost algorithm for feature selection, establishes a classification model, constructs imbalanced datasets, and identifies three types of spam advertising accounts. Experimental results demonstrate that compared with

traditional classification methods such as random forest, the proposed method achieves effective improvements in recall rate and F1-score for abnormal user detection on both balanced and imbalanced datasets. Moreover, selecting only a small number of features can still achieve high detection performance, proving the effectiveness of the method.

Keywords: XGBoost; social networks; abnormal user detection; abnormal account detection; spam

---

0 Introduction

In recent years, social networks and social media have experienced vigorous development. However, abnormal users, primarily spam senders (also known as spam accounts), continuously pollute the social network environment [1]. These accounts are fake users created by attackers to publish advertisements, phishing links, pornographic content, and other malicious URLs, exhibiting distinct behavioral characteristics. They exploit online social networks to disseminate harmful information on a large scale, interfere with normal platform usage, and threaten Internet security [2]. Rapid and effective identification of spam accounts helps purify the social network environment from the source and safeguard Internet security, representing a key research focus in both public security public opinion monitoring and academia.

1.1 Existing Detection Technology

Current academic research on social network abnormal user detection generally involves extracting one or several types of features from social network nodes, including registration attributes, published content, activity behaviors, and connection relationships, to construct multi-dimensional feature vectors, followed by detection using machine learning methods. These approaches can be divided into supervised learning and unsupervised learning methods.

1.2 Unsupervised Learning Detection Methods

Unsupervised learning detection methods directly cluster samples based on their multi-dimensional features, thereby grouping normal users and spam users into different clusters. Since these methods do not require training samples, they can rapidly form detection systems. Miller et al. [3] utilized Twitter user profile information and text content features to cluster normal users and spam accounts into different categories. Chu et al. [5] clustered microblogs based on the final redirect addresses of URLs embedded in Twitter posts and determined whether accounts within each cluster were spam accounts.

1.3 Supervised Learning Detection Methods

Supervised learning detection methods leverage pre-labeled datasets to train classification models, which are then applied to predict unlabeled data. Zheng X [8] and Lyu Shaoqing [9] constructed classifiers using account creation time, message comment counts, and other content and behavioral features to detect spam accounts. Liu Chen [10] modeled and identified excessive forwarding, following behaviors, and fake followers based on user posting frequency and the number of "@" mentions in posts. Meng Jiang [11] and Xue [12] modeled the in-degree, out-degree, and influence of nodes in social network relationship graphs to detect fake accounts with mismatched follower and friend counts. F.B et al. employed random forest and SVM methods to detect spam users and published their dataset [1].

Traditional supervised learning classification methods include Support Vector Machines (SVM) and Random Forest (RF). SVM achieves sample classification by finding hyperplanes in high-dimensional vector space, offering low computational complexity and excellent performance on small-sample data, particularly suitable for binary classification tasks. Random Forest and other decision tree-based ensemble classification models select k most effective features from n-dimensional original features for splitting during training ($k < n$), generating multiple decision trees in parallel to determine classification results through voting. RF demonstrates excellent detection performance for multi-dimensional feature data.

1.4 Limitations of Current Detection Methods

Since unsupervised learning can only cluster users with similar intrinsic features but cannot directly determine cluster labels, supervised learning methods can effectively utilize multi-dimensional features of social network accounts to directly predict classification labels, generating classification models with higher accuracy. Therefore, supervised learning approaches are more effective for abnormal user detection. Although current commonly used supervised learning methods can achieve certain detection goals, their detection accuracy remains limited, primarily due to two aspects: feature selection and algorithm selection.

a) Feature Selection: Previous studies often selected only one type of feature, such as behavioral features, for detection. Since multiple categories of features of social network abnormal users differ from normal users, selecting only certain features easily overlooks information contained in other features, insufficiently describing the true data situation and resulting in poor detection performance. However, if all features are selected, due to correlations among various features of social network accounts, methods like SVM that directly project

non-orthogonal multi-dimensional features into orthogonal vector spaces using embedding approaches can cause deviations, yielding limited detection effectiveness for high-dimensional features. Therefore, a feature selection method is needed that utilizes all categories of features while avoiding noise caused by high-dimensional features.

b) Algorithm Selection: Although random forest can reduce dataset dimensionality and eliminate the impact of non-orthogonal features through its feature selection process, features not selected in each splitting round cannot participate in that iteration, causing feature information loss and generating errors. Moreover, since real-world social network datasets are imbalanced where normal users far outnumber abnormal users (exhibiting a long-tail effect), random forest encounters problems such as poor classification performance and increased generalization error when detecting on imbalanced datasets. Therefore, an algorithm is needed that can effectively utilize multi-dimensional features while remaining effective on severely imbalanced sample sets.

Currently, classifying imbalanced data represents one of the research challenges in social network abnormal detection. Academic solutions to imbalanced data classification mainly include resampling techniques [15-17] and improved classification algorithms [18, 19]. Resampling techniques reduce inter-class imbalance ratios by expanding smaller-class data scales or shrinking larger-class scales, but newly constructed datasets through under-sampling or over-sampling cannot completely match the true distribution of original datasets, easily causing information loss or overfitting. Improvements based on original algorithms by introducing incremental online learning [18], ensemble learning [19], and other methods can also reduce algorithm sensitivity on imbalanced datasets, but such approaches easily introduce new problems such as high computational complexity and low efficiency, and focusing on solving imbalanced classification problems at a single level can easily sacrifice model generalization.

---

2 XGBoost-Based Abnormal User Detection Method

Social network abnormal user detection essentially constitutes a multi-classification task that divides all samples in a dataset into normal users and various types of abnormal users. This study selects the XGBoost (extreme gradient boosting) [13] ensemble boosting method to construct classification models. Each sample in the training dataset corresponds to a user in the social network, consisting of an n-dimensional feature vector x_i containing content, behavior, attributes, relationships, etc., and corresponding p class labels y_i: $\{(x_i, y_i)\}$ where $x_i \in \mathbb{R}^n$ and $y_i \in \{class_1, class_2, ..., class_p\}$. The XGBoost-based user classification method constructs a classification model by learning from input training samples, 挖掘 the relationship $f(x_i) = y_i$ between feature values x_i and class labels y_i, thereby predicting the class of new samples.

The overall detection process is shown in Figure 1 [Figure 1: see original paper]. For the classification task proposed in this paper, each round of XGBoost training is iteratively generated based on the previous round. The objective function for tree construction in the t-th iteration is:

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + const$$

The tree model generated in each round is represented by both its structure part q and leaf node sample weights w as: $f_t(x) = w_{q(x)}$. Tree complexity is determined jointly by the number of leaves T and the L2 norm squared of sample weights w, where larger T and more uneven w values among samples indicate more complex tree structures. The regularization term $\Omega(f_t)$ controls model complexity, effectively preventing overfitting, defined as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Expanding the objective function using second-order Taylor series and rewriting yields the final objective function:

$$Obj^{(t)} \approx \sum_{j=1}^{T} [G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma T$$

where $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$ represent the sums of first-order and second-order derivatives of samples on leaf node j, respectively. Joint optimization using both first-order and second-order derivative information can achieve global optimality.

The experiment gradually generates optimal tree structures by attempting to add splits to existing leaf nodes at each step. The gain from splitting is:

$$Gain = \frac{1}{2}[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}] - \gamma$$

When the splitting gain continuously falls below a fixed value or the number of splits reaches the specified maximum depth, splitting stops and the final classification model is obtained.

Regarding the feature selection problem mentioned above, this study retains all features including user content, behavior, attributes, and relationships when constructing classification models, fully utilizing effective information from various features to avoid information loss. It optimizes tree structure by finding

optimal values for the loss function through serial iterative operations, eliminating the impact of non-orthogonal features. After initial training, XGBoost statistics are used to count how many times each feature is used for decision tree splitting, calculating the correlation degree between sample features and classification results, thereby performing feature selection according to feature importance and reducing dimensionality.

For the imbalanced data reality where spam users are far fewer than normal users, this study performs multiple ensemble iterative operations and controls XGBoost' s `max_delta_step` parameter to limit each tree' s weight and change the maximum step size, thereby avoiding excessive influence from instance samples in small-quantity categories on classification results and reducing errors caused by training data imbalance.

3.1 Dataset and Comparison Methods

This study employs the Apontador dataset [1] to validate method effectiveness. This dataset was collected from a famous Brazilian location-based social network and is a balanced dataset containing both normal users and spam users. Spam users include three categories: product marketing advertisers (LM), content polluters posting content inconsistent with topic tags (PL), and attackers posting abusive speech (BM), accounting for 31%, 48.5%, and 21.4% of abnormal users, respectively.

Each record contains 59 feature fields (Table 1) and 2 classification fields. The original authors used Support Vector Machine (SVM) and Random Forest (RF) methods to perform: direct classification on the dataset's four user types, and two-stage classification that first distinguished whether samples were abnormal then classified abnormal user categories. They verified that RF outperformed SVM in both classification tasks (in direct classification, RF' s recall rates for three spam types were 3.2%, 4.5%, and 5.8% higher than SVM respectively; in two-stage classification, improvements were 1.7%, 3.9%, and 6.3% respectively). To demonstrate the rationality of our method, we reproduced the RF classification experiment from reference [1] under optimal parameters in a Python environment as our experimental comparison baseline.

3.2 Experimental Steps and Parameter Selection

Experiments were conducted on macOS 10.13.4 system with 2.9 GHz Intel Core i5 processor and Python 3.6.4 environment, following these steps:

a) **Data Loading and Preprocessing:** Load data, check data format and distribution, use XGBoost to calculate feature importance rankings, and perform feature importance analysis.

b) **Dataset Partitioning:** Partition the original dataset using 5-fold cross-validation to create experimental and test sets, cyclically evaluating model classification performance. Random stratified sampling is applied to samples from each category during partitioning to ensure the distribution of various sample types in training and test sets matches the original dataset, avoiding sampling errors.

c) **Model Training and Parameter Tuning:** For each training set obtained in step b), further partition into training subsets and validation subsets using 5-fold cross-validation. Iteratively train models using XGBoost on training subsets, employ CV grid search to select optimal values for each parameter, tune parameters gradually, and validate model classification performance on validation subsets to select the optimal parameter set.

d) **Prediction and Evaluation:** Select the model trained with optimal parameters to predict classification results on test sets, output confusion matrices, and calculate evaluation metrics including accuracy (P), recall rate (R), and F1-score.

Empirical verification shows that XGBoost achieves optimal classification performance when parameters are set to `max_depth=3`, `n_estimators=100`, and `n_threshold=None`, as shown in Figures 2 [Figure 2: see original paper] and 3 [Figure 3: see original paper].

3.3.1 Balanced Dataset Detection Results

All classification experiments in this paper repeat the same experiment 5 times and calculate average values to avoid 偶然性 in results. The Random Forest (RF) method used as the control group operates in the same environment and experimental steps as XGBoost. In direct classification and binary classification experiments, confusion matrices and classification reports from both algorithms are shown in Tables 2 through 4 (NS denotes not spam).

Table 3 and Figure 4 [Figure 4: see original paper] show that diagonal blocks in confusion matrices have darker colors, indicating both methods can effectively monitor abnormal users. The above tables demonstrate that our method achieves an overall detection recall rate of 93.22% for abnormal users in binary classification tasks (Table 5), representing a 4 percentage point improvement over Random Forest' s 89.11%. In multi-classification tasks (Table 3), recall rates for various spam users reach 78.96%, 68.66%, and 58.68% respectively, with stable improvements in both recall rates and F1-scores compared to Random Forest (recall rates improve by approximately 1%, 4%, and 5% respectively; F1-scores improve by over 1%). This indicates our method holds greater practical significance for public security operations targeting abnormal user detection.

3.3.2 Imbalanced Dataset Detection Results

This study constructs imbalanced datasets by retaining all normal users while randomly removing abnormal users proportionally, creating datasets where abnormal users account for 10%-40% of all users (50% represents the balanced dataset), while maintaining the same proportional relationships among the three types of abnormal users as in the original dataset.

Table 5 presents results from training and testing XGBoost and RF on these imbalanced datasets. Numbers in the table represent accuracy, recall rate, and F1-score for detecting three types of abnormal users and normal users in corresponding imbalanced datasets. Comparing each corresponding data point between XGBoost and RF reveals that both ensemble methods demonstrate even stronger capability in detecting abnormal users in imbalanced datasets than in balanced datasets, proving the excellent ability of ensemble learning to handle imbalanced data. Moreover, XGBoost shows significantly higher recall rates than RF for detecting BM and LM abnormal users, indicating greater effectiveness on imbalanced datasets.

This occurs because XGBoost possesses strong generalization capability under the combined effect of quadratic terms and regularization terms in its objective function, yielding superior performance on imbalanced datasets compared to RF. Additionally, XGBoost' s slightly lower accuracy than Random Forest for detecting PL users in imbalanced datasets may be explained by: PL being the most numerous abnormal user type and still representing a relatively high proportion in imbalanced datasets; and during data collection and annotation, content polluters (PL) are accounts posting content inconsistent with topic tags, which correlates more strongly with content features, making the full-feature XGBoost approach potentially less precise than RF using partial features for this specific category.

3.3.3 Feature Selection Detection Results

Social network user features can be divided into four categories: text, location, user, and relationship features. To investigate the impact of different feature categories on classification results and validate the effectiveness of XGBoost feature selection, this round of experiments trains models using each feature category separately, then selects the top 10 and top 20 features by influence ranking from XGBoost to train XGBoost and RF classifiers separately, repeating experiments 5 times and averaging results. Classification performance is shown in Table 6 .

Experimental results demonstrate that while using partial feature categories alone can achieve certain classification effects—for instance, using 32 content features alone yields 73% recall rate—XGBoost feature selection using only 20 features can achieve over 80% average recall rate across both classification algo-

rithms, approaching results obtained using all features. Using only the top 10 important features still achieves 73.3% recall rate, higher precision than selecting all features from any single category alone. This proves that in social network abnormal user detection, comprehensively selecting various feature categories achieves more effective results than selecting the same number of features from a single category, validating the effectiveness of XGBoost feature selection. In public security operations, effective feature selection can reduce the number of features required for sample collection, thereby improving detection efficiency. Furthermore, XGBoost achieves higher recall rates than RF in all scenarios, again demonstrating the superiority of the XGBoost classification algorithm.

---

4 Conclusion

Social network abnormal user detection can essentially be 归结为 classification or clustering problems. During decision tree construction, the XGBoost algorithm performs quadratic optimization on the loss function in its objective function, possessing greater global search capability than other boosting methods that only consider first-order derivatives. Simultaneously, the introduced regularization term enhances model generalization performance, while the node weight update strategy preserves complete feature information while eliminating the impact of non-orthogonal features, achieving outstanding results in both binary and multi-classification detection tasks for social network spam users. XGBoost demonstrates even more excellent performance when detecting spam users on imbalanced datasets that more closely reflect real social network conditions. In the process of identifying spam using public datasets, XGBoost-based feature selection that retains only one-third of features can achieve detection performance similar to using all features, improving data collection efficiency. Moreover, regardless of whether all features or partial features are selected, XGBoost achieves improvements in recall rate and F1-score compared to the Random Forest ensemble classifier, holding important practical significance for public security work.

Research Outlook:
a) The XGBoost algorithm can only process numerical features, requiring additional data preprocessing steps to convert non-numeric features into numerical values during practical detection.
b) XGBoost and RF perform better on different feature categories in imbalanced data, suggesting that different classification methods can be selected based on specific detection targets.
c) Future research can integrate multiple algorithms into abnormal user detection models to enhance the robustness of social network abnormal user detection systems.

References

[1] Helen Costa, Luiz H. C. Merschmann, et al. Pollution, bad-mouthing, local marketing: The underground of location-based social networks [J]. Information Sciences, 2014, 279: 123-137.

[2] Zhang Yuqing, Lyu Shaoqing, Fan Dan. Anomaly Detection in Online Social Networks [J]. Chinese Journal of Computers, 2015, 38(10): 2011-2027.

[3] Miller Z, Dickinson B, Deitrick W, et al. Twitter spammer detection using data stream clustering [J]. Information Sciences, 2014, 260(1): 1-15.

[4] Henderson K, Gallagher B, Li Lei, et al. It' s who you know: graph mining using recursive structural features [C]//Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2011: 663-671.

[5] Chu Zi, Widjaja I, Wang Haining. Detecting social spam campaigns on Twitter [C]//Proc of International Conference on Applied Cryptography and Network Security. Berlin: Springer, 2012: 455-472.

[6] Hao Yazhou, Zheng Qinghua, Chen Yanping, et al. Recognition of abnormal behavior based on data of public opinion on the Web [J]. Journal of Computer Research and Development, 2016, 53(3): 611-620.

[7] Chen Tianqi, Tong He. Higgs boson discovery with boosted trees [J]. Proc of International Conference on High-Energy Physics and Machine Learning. 2014: 69-80.

[8] Zheng Xianghan, Zeng Zhipeng, Chen Zheyi, et al. Detecting spammers on social networks [J]. Neurocomputing, 2015, 159(2): 27-34.

[9] Lyu Shaoqing. Research on anomaly detection in online social networks [D]. Xi' an: Xidian University, 2016.

[10] Liu Chen. The detection of anomaly accounts based on behavioral analysis for social networks [D]. Beijing: Beijing Jiaotong University, 2017.

[11] Jiang Meng, Cui Peng, Beutel A, et al. CatchSync: catching synchronized behavior in large directed graphs [C]//Proc of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2014: 941-950.

[12] Yang Zhi, Xue Jilong, Yang Xiaoyong, et al. VoteTrust: leveraging friend invitation graph to defend against social network sybils [J]. IEEE Trans on Dependable & Secure Computing, 2016, 13(4): 488-501.

[13] Chen Tianqi, Carlos Guestrin. XGBoost: A scalable tree boosting system [C]//Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 785-794.

[14] Zhang Qingqing. Anomaly detection research for imbalanced classes [D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2010.

[15] Laurikkala J. Improving identification of difficult small classes by balancing class distribution [C]//Proc of Conference on AI in Medicine in Europe: Artificial Intelligence Medicine. Springer-Verlag, 2001: 63-71.

[16] Yen S J, Lee Y S. Cluster-based under-sampling approaches for imbalanced data distributions [J]. Expert Systems with Applications, 2009, 36(3): 5718-5727.

[17] Chawla N V, Bowyer K W, Hal L O, et al. SMOTE: synthetic minority oversampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[18] Ertekin S, Huang J, Bottou L, et al. Learning on the border: active learning in imbalanced data classification [C]//Proc of the 16th ACM Conference on Information & Knowledge Management. New York: ACM Press, 2007: 127-136.

[19] Li Kewen, Yang Lei, Liu Wenying, et al. Classification method of imbalanced data based on RSBoost [J]. Computer Science, 2015, 42(9): 249-252.

Note: Figure translations are in progress. See original paper for figures.