Analysis and Prediction of Multi-type Illicit Accounts on Ethereum Based on Transaction Features (Postprint)

Authors: Zhou Jian, Yan Shi, Zhang Jie, Huang Shihua

Date: 2022-06-06T00:00:00+00:00

Abstract

The increasingly frequent illegal transaction activities undermine secure transactions on Ethereum, and the anonymity inherent to cryptocurrencies makes such illegal activities difficult to track and analyze. Using Ethereum platform transaction data as the data source, employing datasets of labeled illegal accounts and unlabeled legitimate accounts as the training set, and utilizing the characteristic attributes of transaction data as the foundation for construction, the CatBoost algorithm is employed to perform holistic prediction of the various types of illegal accounts contained therein. In this process, the T-SNE algorithm is employed to achieve dimensionality reduction and visualization of transaction features, multiple-fold cross-validation is adopted, and SHAP Value factors are introduced to determine the positive or negative influence attributes of features. The prediction accuracy of the established model reaches 94.29%, and the evaluation metric of the Area Under the Receiver Operating Characteristic Curve (AUC) reaches 0.9846. The proposed scheme can relatively accurately predict illegal activities existing on the Ethereum transaction platform, thereby effectively improving the blockchain-based transaction environment.

Full Text

Analysis and Prediction of Multi-Type Illegal Accounts on Ethereum Based on Transaction Characteristics

Zhou Jian[1],[2], **Yan Shi**[1]†, **Zhang Jie**[1], **Huang Shihua**[1] [1]College of Management Science & Engineering, Anhui University of Finance & Economics, Bengbu Anhui 233040, China
[2]College of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract

The increasing frequency of illegal transactions undermines secure trading on Ethereum, and the anonymity of electronic currency makes such activities difficult to track and analyze. This study utilizes Ethereum platform transaction data as its data source, employing both labeled illegal accounts and unlabeled normal accounts as the training set. Based on the characteristic attributes of transaction data, we construct a predictive framework using the CatBoost algorithm to forecast multiple types of illegal accounts collectively. The process employs T-SNE for dimensionality reduction and visualization of transaction features, utilizes multi-fold cross-validation, and introduces SHAP Value factors to determine the positive and negative impacts of features. The established model achieves a prediction accuracy of 94.29%, with an AUC evaluation metric reaching 0.9846. The proposed scheme accurately predicts illegal behavior on the Ethereum trading platform and effectively improves the blockchain-based trading environment.

Keywords: blockchain; machine learning; ethereum; illegal account; transaction features

Introduction

Blockchain technology was proposed by Nakamoto in 2008, subsequently giving rise to electronic virtual currencies based on blockchain, such as Bitcoin, Ethereum, and Ripple, which have been applied in electronic transactions. However, the anonymity of blockchain technology makes illegal transactions difficult to track and analyze, attracting criminals and triggering more illicit activities including dark web transactions, prohibited item trading, and financial fraud. The first network crime cases utilizing blockchain smart contract technology, "BigGame" and "MDF Project," involved over 1.3 million digital currencies with a market value of approximately 26 million RMB. Blockchain-based crime has become a new form of high-tech crime that seriously undermines the security and stability of electronic virtual currency transactions.

Analyzing illegal transactions on blockchain represents a challenging problem. Currently, detection of illegal transaction behavior primarily relies on on-chain data analysis. Research methods and objectives can be divided into three categories: (a) Machine learning-based data analysis approaches. Fan et al. proposed a privacy-preserving DML model for permissioned blockchains to address security performance issues, though deployment and implementation remain inadequate. Subsequently, a typical fraud activity—smart Ponzi schemes—emerged in both Bitcoin and Ethereum, prompting in-depth research using different machine learning algorithms to predict illegal accounts. However, these approaches still suffer from insufficient accuracy and incomplete problem exploration. CHUN WEI et al. used metaheuristic algorithms to identify appropriate deep learning hyperparameters in blockchain environments to explore inter-blockchain communication issues, yet challenges remain regarding addi-

tional communication costs and synchronization waiting. (b) Analysis based on complex networks and transaction logic structures. Chen, Weili analyzed smart Ponzi scheme contracts on Ethereum to identify a healthy blockchain trading environment, though the study was limited in terms of quantity, data, and types of illegal accounts examined. Dan Lin, Jiajing Wu et al. modeled and understood transaction information on Ethereum through complex network methods to mine potential value transaction analysis, but their method was not a holistic inductive approach and could not incorporate the latest node representations. (c) Feature value-based analysis. Early research analyzed Bitcoin wallets to parse challenges facing Bitcoin in large-scale crime and fraud activities in the Bitcoin environment, yet these studies had shortcomings in feature extraction and supervised methods.

Bartoletti et al. conducted a more detailed comprehensive analysis of numerous Ponzi schemes on Ethereum to summarize the impact of various perspectives, though their work lacked application-level implementation and broader investigation of illegal account types. In summary, existing methods either fall short of practical accuracy requirements, have room for performance optimization, or focus only on specific types of illegal accounts such as Ponzi schemes, phishing nodes, or illegal money laundering. Therefore, research on illegal account behavior on blockchain remains insufficient.

Addressing these limitations in methodology, performance, and prediction scope, this paper innovates feature construction based on original transaction characteristics and employs the K-Means clustering algorithm in machine learning to first cluster attribute features in the dataset, followed by CatBoost for illegal account prediction.

Methodology

Data Collection and Feature Construction

The dataset primarily originates from two sources: publicly labeled illegal account datasets from the Ethereum platform and normal account datasets crawled from the platform. The illegal account dataset provided by the Ethereum community mainly includes addresses that attempt to imitate other contracts to provide tokens, scam lotteries, fake initial coin offerings, imitation of other transaction users, smart Ponzi contract scams, and phishing nodes. The selected illegal accounts in our dataset contain multiple types of the above categories in an irregular arrangement.

For normal account selection, we randomly chose 4,024 normal accounts between Ethereum blocks 1,209,500 and 12,010,000. Illegal accounts were selected from Ethereum' s publicly labeled dataset, comprising 4,300 accounts. After comparing and screening the collected addresses to ensure no duplication between the two categories, we obtained a total of 4,024 normal accounts and 4,300 illegal accounts. We crawled transaction data using Ethereum' s API by passing the combined data to obtain relevant transaction records for each account.

After examining the 43 feature attributes identified by Steven Farrugia et al., we innovatively constructed a new attribute by integrating "min_{val}*{sent}*" *and* " *max*{val}*{sent}*" *to create* " *Sent*{Diff}*{betweent}*{max}*{and}*{min}," resulting in 44 transaction features for model construction, as shown in [Figure 1: see original paper].

Data Cleaning

Since the dataset is derived from real-time Ethereum transaction data, complex transaction data requires checking for missing values, invalid values, and null values to ensure model accuracy and applicability. Using Python for preprocessing, we identified missing and invalid values in certain attribute features, as shown in .

The cleaning principle involves calculating the mean of the overall data for problematic attribute features and using this value to fill in missing or invalid entries. To ensure comprehensive data cleaning, we observed the specific distribution of the dataset. Based on the characteristics of noisy data—where over 90% of values are "0," measurements show significant errors compared to true values, or the feature burden affects model performance—we removed features such as "total_{ether}*{sent}*{contracts}" and "ERC20_{avg}*{val}*{sent}_{contract}" that were deemed performance-burdening, excluding them from final model construction.

CatBoost Algorithm

As the latest research algorithm among gradient boosting trees, CatBoost has not been previously applied to predicting illegal accounts on blockchain trading platforms. It effectively handles categorical features and reduces overfitting problems. According to equation (1), categorical feature values are converted to numerical results:

$$\sigma_{p,j} = \frac{\sum_{k=1}^{K}[x_{p,k} = j] \cdot Y_k + a \cdot P}{\sum_{k=1}^{K}[x_{p,k} = j] + a}$$

where P is the added prior term, a is a weight coefficient greater than 0, j represents the coefficient of categorical feature values, k is the training sample coefficient, and i counts the number of times the categorical feature value equals the label value.

As a gradient boosting tree algorithm, CatBoost employs symmetric trees as base learners, forming a strong learner through serial iteration of a set of classifiers. The objective of the k-th iteration is to find h_k, that is:

$$h_k = \arg\min_{h \in H} \sum_{i=1}^{m} L(y_i, F_{k-1}(x_i) + h(x_i))$$

where L is the loss function and F_{k-1} is the current learner formed after $k-1$ iteration steps. To obtain an unbiased gradient estimate, CatBoost' s specific modeling process for our dataset is as follows: (a) For each sample x_i in the illegal account dataset X, CatBoost uses all training samples except x_i to obtain model M_i; (b) It employs ordered boosting to calculate gradient estimates using M_i; (c) The new model re-evaluates sample x_i and forms a base learner; (d) The base learner is further processed to ultimately form a strong learner. This process continuously reduces prediction errors on the training set, ultimately forming the CatBoost model.

T-SNE Visualization

The experimental dataset' s feature space is multi-dimensional. We use the T-SNE algorithm for non-linear transformation to visualize dataset account label types in a 2D plane. T-SNE is a statistical method that visualizes high-dimensional data by assigning each data point a position in a two- or three-dimensional map. Its two main advantages are: (a) For dissimilar points, a small distance produces a large gradient to repel these points; (b) The repulsion is not excessive, avoiding placing dissimilar points too far apart.

The algorithm logic is as follows:
Algorithm 1 T-SNE Algorithm Pseudocode
Input: Experimental dataset and Python third-party libraries
Output: T-SNE visualization graph
a) Import the experimental dataset
b) Remove irrelevant columns ("address") and separate the "FLAG" label column as y
c) Standardize target features using StandardScaler and set n_{components}=2 for 2D visualization
d) Use scatterplot function for final visualization

[Figure 2: see original paper] shows that: (a) labeled and unlabeled data form several distinguishable clusters scattered in a point-like distribution; (b) clusters of labeled and unlabeled data types still have substantial overlap; (c) a small cluster of "illegal" account labels is clearly present in the lower right corner. These results emphasize the importance of using machine learning to distinguish between the two binary types.

Experiments

Parameter Tuning

Through K-Means clustering algorithm classification, we further cluster and analyze attribute feature categories to observe whether different account categories have the same practical effects. To determine the optimal K value for K-Means, we use the Elbow Rule, which clearly shows that the optimal centroid K for our dataset is 2—when K=2, the decline curve significantly slows, as shown in [Figure 3: see original paper].

The clustering results in [Figure 4: see original paper] indicate that except for a small portion showing different category features, most selected account categories share the same features, further emphasizing the importance of our machine learning approach.

All modeling experiments were implemented in Python 3.9. When building models with CatBoost, we considered three critical factors: learning rate, number of estimators (n_{estimators}), and maximum depth (max_{depth}). Since optimal parameters cannot be determined during training, we used grid search optimization to adjust CatBoost parameters. In addition to these three factors, parameters include iterations, border_{count}, leaf_{estimation}*{method}, and* *l2*{leaf}_{reg}, as defined in .

During tuning, we kept other parameters constant, first optimizing border_{count}, then iterations and learning_{rate}, followed by tree depth, ultimately determining the model parameter structure. The optimized results are shown in .

Based on these tuning results, we evaluated CatBoost under different cross-validation folds, with results shown in . The 10-fold cross-validation significantly outperformed 3, 4, 5, 11, and 12-fold validations, effectively enhancing model construction and improving prediction accuracy for illegal accounts on Ethereum.

Model Evaluation

Using grid search optimization, we identified three crucial elements: learning rate, tree depth, and number of estimators. We further refined the experiment based on AUC values and Logloss. Under 10-fold cross-validation, we evaluated model prediction capability across different tree depths and estimator quantities, with results shown as trend lines in [Figure 5: see original paper].

When tree depth is 2, AUC values are significantly lower than other depths as the number of estimators increases, though showing a typical growth trend. As depth increases, AUC evaluation metrics gradually rise. The optimal parameter adjustment results show that when depth=5 and n_{estimators}=300, the model' s AUC value is slightly higher than other parameter combinations.

Based on Logloss, CatBoost executes iterations on training and test sets. [Figure 6: see original paper] shows that after 100 iterations, both datasets begin to converge, proving the model's strong predictive adaptability. To verify reliability, we expanded the random sampling range, as shown in [Figure 7: see original paper].

To evaluate prediction capability, we randomly extracted 20 samples from the test set. shows that only one prediction differed from the actual label, demonstrating strong predictive ability. As the random sample size expands, prediction capability gradually stabilizes, with final overall dataset prediction accuracy reaching 0.9407, further confirming the model' s strong adaptability and

accuracy for predicting illegal accounts on Ethereum.

Feature Importance

We ranked attribute feature importance for Ethereum illegal account prediction, identifying the top ten influential features shown in [Figure 8: see original paper]. The results indicate that "total_{ether}*{balance},*" "*sent*{tnx}," and "Unique_{Received}*{from}*{Address}"are significantly important variables, while "Time_{Diff}*{between}*{first}*{and}*{last}(Mins),""Avg_{min}*{between}*{sent}*{tnx},*" "*min*{value}*{received},*""*Unique*{Sent}*{To}*{Addresses},""avg_{val}*{sent},*""*max*{val}*{sent},*" *and* "*min*{val}_{sent}" are moderately important. The remaining 33 predictor variables have importance ratios approaching or equal to zero, with complete rankings shown in [Figure 9: see original paper].

To further investigate feature impacts, we introduced SHAP Value variables to show positive and negative influences, as visualized in [Figure 10: see original paper]. Combined with feature importance results, the top ten features are displayed where color represents feature value magnitude and width represents feature distribution. SHAP Value less than zero negatively impacts label features; otherwise, it has positive impact. "Unique_{Received}*{from}*{Address}" shows significantly higher positive than negative influence, while "sent_{tnx}" and "Time_{Diff}*{between}*{first}*{and}*{last}(Mins)"show higher negative influence. SHAP Value variables thus intuitively determine the positive and negative impacts of different attributes on model construction.

Model Comparison

By creating new transaction features and combining other algorithms with CatBoost, we compared our approach with other algorithm models under 10-fold cross-validation to verify effectiveness, as shown in .

The improved CatBoost algorithm, through feature creation and K-Means clustering integration before model construction, outperforms other algorithms in both accuracy and AUC. LightGBM, another gradient boosting tree algorithm, shows good AUC but lower accuracy than CatBoost. Compared with standard CatBoost, our improved version shows slight improvements in both metrics, further proving CatBoost' s excellent predictive capability for Ethereum illegal account datasets and making our algorithm model more convincing.

Conclusion

Based on collecting and organizing publicly available Ethereum account datasets, we constructed a model using the CatBoost algorithm in a Python development environment. The model demonstrates strong capability in predicting future illegal accounts on Ethereum, showing high standards in accuracy and AUC values while revealing predictor importance rankings and positive/negative influence capabilities. The proposed algorithm model has strong adaptability

and the methodology is foundational, making it applicable to other domains. Further optimization and improvement can explore deeper illegal behaviors on blockchain trading platforms.

Our work contributes to detecting illegal accounts based on transaction histories among massive Ethereum accounts, with the proposed algorithm model being usable by relevant economic institutions and departments. Future work will apply this architecture to other domains to ensure model applicability and practicality, analyzing and exploring illegal account prediction not only on Ethereum but also on other blockchain-based trading platforms. We will further attempt to extract novel features from blockchain transactions to improve prediction capabilities. While CatBoost has certain limitations, we will optimize algorithms to enhance prediction accuracy and efficiency.

References

[1] Nakamoto S. Bitcoin: A peer-to-peer electronic cash system. White Paper.
[2] Tschorsch F, Scheuermann B. Bitcoin and Beyond: A Technical Survey on Decentralized Digital Currencies [J], IEEE Communications Surveys and Tutorials. 2016, 18(3): 2084-2123.
[3] Chipere M. Virtual currency as an inclusive monetary innovation for the unbanked poor [J], Electronic Commerce Research and Applications. 2018, 28: 28-37.
[4] Bohme R, Christin, N, Edelman, B, et al. Bitcoin: Economics, Technology, and Governance [J], Journal of Economic Perspective. 2015, 29(2): 213-238.
[5] Sajana P, Sindhu M, Sethumadhavan M. On blockchain applications: Hyperledger fabric and Ethereum [J], International Journal of Pure and Applied Mathematics. 2018, 118(18): 2965-2970.
[6] Wang S, Ouyang, LW, Yuan Y, et al. Blockchain-Enabled Smart Contracts: Architecture, Applications, and Future Trends [J], IEEE Trans on Systems Mam Cybernetics-Systems. 2019, 49(11): 2266-2277.
[7] Hasan M, Naeem MA, Arif M, et al. Higher moment connectedness in cryptocurrency market [J], Journal of Behavioral and Experimental Finance. 2021, 32.
[8] Kristoufek L, Vosvrda M, et al. Cryptocurrencies market efficiency ranking: Not so straightforward [J], Physica-Statistical Mechanics and Its Applications. 2019, 531.
[9] ElBahrawy A, Alessandretti L, Rusnac L, et al. Collective dynamics of dark web marketplaces [J], Scientific Reports. 2020, 10(1).
[10] Lehdonvirta V. Virtual item sales as a revenue model: Identifying attributes that drive purchase decisions [J], Electronic Commerce Research. 2009, 9(1-2): 97-113.
[11] Hyvarinen H, Risius M, Friis G, et al. A Blockchain-Based Approach Towards Overcoming Financial Fraud in Public Sector Services [J], Business & Information Systems Engineering. 2017, 59(6): 441-456.
[12] Bayramova A, Edwards DJ, Roberts C, et al. The Role of Blockchain Tech-

nology in Augmenting Supply Chain Resilience to Cybercrime [J], Buildings, 2021, 11(7).
[13] Kim H, Kim SH, Hwang JY, et al. Efficient Privacy-Preserving Machine Learning for Blockchain Network [J], IEEE ACCESS. 2019.
[14] Fan SH, Fu SJ, Xu HR, et al. Anti-leakage smart Ponzi schemes detection in blockchain [J], Information Processing & Management. 2021, 58(4).
[15] Tsai CW, Chen YP, Tang TC, et al. An efficient parallel machine learning-based blockchain framework [J], ICT Express. 2021, 7(3), 300-307.
[16] Chen, Weili. Detecting ponzi schemes on ethereum: Towards healthier blockchain technology [C]// Proceedings of the 2018 World Wide Web Conference. 2018.
[17] Dan Lin, Jiajing Wu, Qi Yuan, et al. Modeling and Understanding Ethereum Transaction Records via a Complex Network Approach [J], IEEE Trans on Circuits And Systems. November 2020.
[18] Meiklejohn S, Pomarole M, Jordan, G, et al. A fistful of bitcoins: Characterizing payments among men with no names [C]// Proceedings of The 2013 Conference On Internet Measurement Conference. 2013.
[19] Monamo, P, Vukosi M, et al. Unsupervised learning for robust Bitcoin fraud detection [C]// 2016 Information Security for South Africa (ISSA). IEEE, 2016.
[20] Abad-Segura E, Infante-Moro A, Gonzalez-Zamar MD, et al. Blockchain Technology for Secure Accounting Management: Research Trends Analysis [J], Mathematics, 2021, 9(14).
[21] Bartoletti, M, Carta S, Cimoli T, et al. Dissecting Ponzi schemes on Ethereum: Identification, analysis, and impact [J], Future Generation Computer Systems. 2020, 101: 259-277.
[22] Bartoletti M, Carta S, Cimoli T, et al. Dissecting Ponzi schemes on Ethereum: Identification, analysis, and impact [J], Future Generation Computer Systems-The International Journal Of Escience. 2020, 102.
[23] Gurun UG, Stoffman N, Yonker SE. Trust Busting: The Effect of Fraud on Investor Behavior [J], Review of Financial Studies, 2018, 31(4).
[24] Chen L, Peng JY, Liu Y, et al. Phishing Scams Detection in Ethereum Transaction Network [J], ACM Trans on Internet Technology. 2021, 21(1).
[25] Lin GY, Liu BW, Xiao PC, et al. Phishing Detection with Image Retrieval Based on Improved Texton Correlation Descriptor [J], Cmc-Computers Materials & Continue. 2018, 57(3): 533-547.
[26] Campbell-Verduyn M. Bitcoin, crypto-coins, and global anti-money laundering governance [J], Crime Law And Social Change. 2018, 69(2).
[27] Ducas E, Wilner A. The security and financial implications of blockchain technologies: Regulating emerging technologies in Canada [J] International Journal. 2017, 72(4): 538-562.
[28] Bradley, AP. The use of the area under the roc curve in the evaluation of machine learning algorithms [J], Pattern Recognition. 1997.
[29] 周建, 张杰, 闫石. 基于链上数据的区块链欺诈账户检测研究 [J]. 计算机应用研究, 2022, 39(04): 992-997. (Zhou Jian, Zhang Jie, Yan Shi. Research on detection of fraudulent accounts in blockchain based on on-chain data [J], Application

Research of Computers. 2022, 39(04): 992-997.)
[30] 俞莎莎, 牛保宁. 基于交易不可信度的比特币非法交易检测 [J]. 计算机工程, 2021. (Yu Shasha, Niu Baoning. Bitcoin illegal transaction detection based on transaction unreliability [J]. Computer Engineering, 2021.)
[31] Huang GM, Wu LF, Ma X, et al. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions [J], Journal of Hydrology. 2019. 574: 1029-1041.
[32] Punmiya R, Choe S. Energy Theft Detection Using Gradient Boosting Theft Detector With Feature Engineering-Based Preprocessing [J], IEEE Trans on Smart Grid, 2019, 10(2): 2326-2329.
[33] Steven F, Joshua E, George A. Detection of illicit accounts over the Ethereum blockchain[J], Expert Systems with Applications. 2020, 15.

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