International Journal of Informatics and Information Systems ISSN 2579-7069
Vol. 6, No. 2, March 2023, pp. 81-88

Application of Big Data Mining Technology in Blockchain Computing

Zejian Dong

Shandong Vocational College of Light Industry, Zibo, Shandong, China, 255300
hanxuebuji@sdlivc.com*
* corresponding author

(Received January 12, 2023 Revised February 06, 2023 Accepted February 23, 2023, Available online March 1, 2023)

Abstract

Big data in the modern science and technology and social activities play an important role, on the one hand, a large number of new applications and technology into our lives, in the use of these new technologies to produce a large amount of data, on the other hand, big data as one of the most important digital assets, many of the development of new technologies also rely on large data as support. This paper focuses on the research and application of big data mining technology in blockchain computing. Firstly, this paper extracts the corresponding transaction data according to the Bitcoin address and constructs the transaction features to get the Bitcoin data set. Then, the data features are processed. Then, three algorithm models, SVM, Adaboost and Random Forest, are selected to model and analyze the preprocessed data combined with different sampling strategies. According to the comprehensive performance of the model and its shortcomings, the model is selected and improved.

Keywords: Big Data, Data Mining, Virtual Currency, Blockchain Computing

1. Introduction

In the information age, new technologies are coming into being and gradually affecting our way of life. Such as mobile social networks and intelligent devices connected to everything, we are all actively or passively involved in the generation of big data when everything happens and every interaction. The development of these fields, in particular, needs the support of big data [1]. For example, in machine learning, when accuracy is improved to a certain extent, the bottleneck limiting its development is no longer the model and structure, but the quality of the data. For some high-tech companies, those who have higher quality big data will have the core competitiveness in innovation and development. In such an environment, the better connection between the owners and consumers of big data and the promotion of the sharing and use of big data can further release the value of big data, and this background also promotes the rapid development of the industries related to big data transactions [2]. Blockchain opens up a new computing paradigm for building trust at low cost in an untrusted competitive environment. Blockchain has emerged as a new security technology due to its ability to address the issue of "trustworthiness" in transactions. Blockchain is not an entirely new technology, but rather an innovative combination of existing technologies. Blockchain is based on peer-to-peer network, which uses a variety of technologies such as consensus mechanism, cryptography and smart contract to build a trusted environment and value transmission network, and has the characteristics of decentralization, immutability, openness, anonymity and traceability [3].

Many scholars have carried out research on the core technology of blockchain, including consensus and data consistency, smart contract, security and privacy, and scalability, etc., and some research achievements have been made in recent years [4]. Since the first COINS to consensus effort to prove that, under the impetus of the blocks in the chain network has spawned many variants of consensus and data consistency algorithm for the longest chain rules limit the transaction capacity problem, some scholars put forward the GHOST rules, the rules that the isolated block

also contribute to the safety of the main chain, for a given a block for root to creation of the block tree, the longest chain in the heaviest subtree serves as the main chain, which effectively reduces the impact of forking and allows shorter block spacing, thus increasing the transaction capacity [5]. Some scholars have proposed Scalable Consensus Protocol (SCP), which combines Byzantine fault tolerance and sharding into blockchain consensus. The core idea of SCP is to divide the network into subcommittees (i.e., sharding) through the POW mechanism. Each subcommittee controls limited computing power, and the number of subcommittees is proportional to the total computing power of the network [6].

This paper first introduces the research background of the topic and the relevant research at home and abroad, and then introduces the relevant theoretical basis needed in the research of this paper, including data mining, blockchain and so on. After the introduction of relevant theoretical basis, a detection model applied to Bitcoin is built based on data mining technology.

2. Application of data mining technology in blockchain computing

2.1. Data mining technology model

The advent of the Internet has ushered us into a new era of interconnectedness, where all kinds of data are digitized and stored in databases. While the data itself may not hold much intrinsic value, it contains vast amounts of information and application potential. As such, the information industry and related fields have seen a surge in demand for processing large amounts of data. Data Mining (DM) has emerged as a critical tool for collecting and analyzing big data to uncover valuable patterns and rules. DM technology aims to extract useful information from seemingly disorganized data through specific mining algorithms [7-8]. Today, data mining technology is widely used in various information and knowledge fields.

The abundance of data generated by the Internet has given rise to a growing demand for technologies and techniques to process and extract useful insights from this data [10-11]. DM is one such technology that has gained popularity due to its ability to sift through massive amounts of data to find patterns that may have previously gone unnoticed. By using specific algorithms, DM enables the identification of relationships and trends in data that can be used for a wide range of applications.

The application of DM is not limited to any particular field but is rather a cross-disciplinary technology. From finance to healthcare, DM has become an integral part of many industries [12]. By analyzing data sets, DM can provide insights that can be used to make informed decisions, develop new products, and improve processes. The value of DM lies in its ability to extract information from data sets that may not be immediately apparent, and this information can be used to enhance businesses and improve outcomes in various sectors.

As the volume and complexity of data continue to grow, DM has become increasingly important. The ability to extract useful insights from data has become a critical component of business success in the 21st century [13]. With advances in technology, DM has evolved, and new algorithms and techniques are continually being developed to improve its effectiveness. Today, DM is no longer a luxury but a necessity, and its applications are only set to increase in the future.

In conclusion, the birth of the Internet has transformed the way we process and store data. DM has emerged as a valuable tool for extracting insights from vast amounts of data, and its applications span various industries. As the demand for data processing and analysis continues to grow, DM will play an increasingly critical role in helping businesses and organizations stay competitive and make informed decisions [14].

Generally, the theoretical model of data mining mainly includes the following parts.

1) Demand analysis The first step of a data mining task is to conduct a demand analysis and conduct a background investigation of the target domain. First of all, the problem to be solved and the required data should be clarified, and then a reasonable technical arrangement should be made.
2) Data acquisition After data requirements analysis, the next step is data acquisition. Data can be obtained through various channels, such as applying to relevant departments, collecting by oneself, cooperating with data owners and sharing, or using web crawler technology to obtain data, etc. [9,15-17].
3) Data preprocessing The main methods of data preprocessing include three aspects: data cleaning, data conversion and data simplification. The commonly used methods of data cleaning include Binning, regression

and clustering. Data conversion involves normalization, discretization, attribute selection and other technologies. Data simplification includes dimension-reduction, attribute subset selection, quantity specification and other technologies [10,18].

4) Data mining As the most critical part of the whole model, its quality directly affects the performance of the final mining analysis. Currently, the most frequently used methods are Anomaly Detection, classification, clustering, association analysis, automatic text summary, etc. [11,19]. In addition, processing methods in other fields can also be migrated to data mining, such as decision tree, genetic learning and case based learning.
5) Result analysis and expression The final process of data mining should be measured by evaluation criteria to analyze whether the original technical objectives have been achieved. At the same time, data tables and graphs can be used to analyze whether the mining results are reasonable. Finally, these results need to be presented to analyze whether the specific requirements proposed in the first step have been solved. If not, it is necessary to improve the algorithm and continue mining [12,20].

2.2. Bitcoin detection model based on data mining

Data mining technology has undergone extensive development over the years and has been applied in various fields, including finance, industrial production, and telecommunications [21]. One area that has benefited from data mining technology is the digital finance field, where Bitcoin system, a digital currency, operates. In recent times, the use of data mining technology has been increasingly applied to the analysis of Bitcoin networks, including the profiling of the network, the identification of transaction patterns, and the detection of illegal behaviors.

To construct a Bitcoin Ponzi scheme detection model using data mining technology, three primary steps are necessary [22,23]. The first step involves extracting Bitcoin transaction data from the addresses. Next, relevant transaction features are calculated based on the transaction data, and the constructed characteristic data set is preprocessed. Finally, modeling and analysis are carried out on the processed data set using different sampling technologies. The effectiveness of the model is then evaluated and analyzed using the call-back rate and AUC value.

The process of data mining starts with the extraction of the basic transaction data of the address from the Bitcoin network. This data is then used to calculate relevant transaction features that are essential in identifying patterns that could be indicative of a Ponzi scheme. Preprocessing of the data set is then carried out to ensure that the data is clean, consistent, and usable. Preprocessing is a crucial step in the data mining process because it helps eliminate errors and inconsistencies that could adversely affect the accuracy of the analysis.

1) SVM algorithm

SVM, also known as support vector machine, is a binary classification algorithm with high performance. The basic model of the algorithm is defined as a linear classifier with the largest interval in the feature space. The basic idea of SVM is to find a plane in the space that can divide different samples in the data set D. This plane is usually called the "hyperplane", and its equation description is shown in Equation (1).

$$w^T x + b = 0 \tag{1}$$

In the training data, there are always some positive and negative examples in which the distance between samples and the hyperplane is the shortest. The line that crosses these sample points and is parallel to the hyperplane is called the support vector, and the distance between these two support vectors is called the interval. Its calculation formula is shown in Equation (2) :

$$r = \frac{2}{|w|} \tag{2}$$

The optimization goal of support vector machine is to find the partition hyperplane of "maximum interval", which is expressed in Equation (3):

$$max_{wb} \frac{2}{|w|}$$

$$s.t\, y_i(w^T x + b) \geq 1, i = 12 \ldots . m \tag{3}$$

2) Adaboost algorithm

AdaBoost is a Boosting ensemble learning algorithm. The algorithm trains the weak classifier by constantly adjusting the weight distribution of the training sample data, and the weak classifier can be combined into a stronger classifier with better classification effect by linear superposition.

3) Random forest algorithm

Random forest is a Bagging type ensemble learning algorithm. The algorithm integrates multiple decision trees in a parallel way, and the final result is produced by each base learner through voting. The basic steps of the algorithm are as follows:

Random forest is a popular ensemble learning algorithm that uses a collection of decision trees to make predictions. The algorithm works by randomly selecting samples from a dataset and building a decision tree for each sample. This process is repeated multiple times to create a forest of decision trees. In this method, sampling with put back is used, where n samples are selected from the dataset as a training set.

After the training set is sampled, a decision tree is generated from it. This process is repeated m times, and each time a new decision tree is generated. Finally, the collection of decision trees forms a random forest with m decision trees. The random forest is then used to predict the test set, and the prediction result is decided by voting.

To ensure that the sampling process is done correctly, the final number of subset samples should be the same as the total input sample data. This is because the samples are selected with replacement, meaning that the same sample can be selected multiple times. This helps to increase the diversity of the training set, and thus improve the performance of the model.

During the training process of each decision tree, the number of selected features is less than or equal to the total number of features. This is done to ensure that the trained decision tree can train models with stronger generalization ability from different angles. By randomly selecting a subset of features for each tree, the algorithm reduces the correlation between the trees, which improves the accuracy of the predictions made by the random forest.

In summary, the random forest algorithm is a powerful machine learning technique that uses a collection of decision trees to make predictions. It works by randomly selecting samples from a dataset and building a decision tree for each sample. By repeating this process multiple times and using voting to make predictions, the algorithm improves the accuracy of its predictions. By sampling with put back and selecting a subset of features for each tree, the algorithm increases the diversity of the training set and reduces the correlation between the trees, thus improving the generalization ability of the model.

3. System Test

3.1. Data sources

The data set of Bitcoin addresses used in this study was selected from artificially identified fraud addresses in "Hyip" cryptocurrency forums abroad and other well-known Ponzi scheme addresses that have been verified. Massimo's dataset contains 32 Ponzi addresses for model training and 20 for model testing.

3.2. Transaction data extraction

In Massimo's research data set, the transaction address and some transaction features of Bitcoin are given. In order to construct more useful features for modeling and analysis, the transaction data corresponding to the Bitcoin address must be extracted. There are many ways to extract the transaction data corresponding to the address. You can directly download the entire Bitcoin client to analyze the corresponding transaction information, or you can obtain it through the public API provided by the blockchain browser. Because the Bitcoin client contains all the block and transaction information of the entire Bitcoin blockchain, the amount of data is relatively large, coupled with the complexity of the blockchain network itself, it is quite difficult to parse the transaction data of the Bitcoin address from the complete Bitcoin blockchain. So this article chose to use the API provided by Blockchain.info to get the list of transactions for the address.

3.3. Data preprocessing

In the process of building a model, it is essential to perform correlation analysis. This is because high correlation between variables in the data can lead to the model repeatedly iterating through the same information during the training stage. This can result in slower convergence of the model during the training process. Additionally, the introduction of noise into the data may cause the information to be unstable and could potentially affect the stability of the model. As a result, it is important to measure the correlation of each characteristic variable after constructing transaction data features of Bitcoin addresses in order to carry out further analysis.

Correlation analysis is crucial as it helps to identify the strength and direction of the relationship between variables in the data. If there is a strong positive correlation between two variables, for instance, then changes in one variable are likely to correspond to changes in the other. On the other hand, a strong negative correlation indicates that changes in one variable correspond to changes in the opposite direction in the other variable. By understanding the correlation between variables, one can make better decisions in building the model.

When building a model, it is important to ensure that the model is stable and does not introduce noise into the data. This is because noise can affect the stability of the model and lead to inaccurate predictions. Correlation analysis helps to identify which variables are correlated and which are not, thereby allowing the model builder to focus on those variables that are relevant to the model's performance.

By measuring the correlation of each characteristic variable, it is possible to determine which variables are most relevant to the model's performance. This allows the model builder to focus on those variables that are most important to the model's accuracy and stability. Additionally, by understanding the correlation between variables, the model builder can identify potential issues with the data and make adjustments to ensure that the data is as accurate and stable as possible.

In conclusion, correlation analysis is a critical step in model building. By measuring the correlation of each characteristic variable, it is possible to determine which variables are most relevant to the model's performance and make adjustments to ensure the data is as accurate and stable as possible. This helps to ensure that the model is able to make accurate predictions and is reliable for use in real-world applications.

4. Simulation Experiment Results

4.1. Recall values of the model under different sampling strategies

Table. 1. Recall values of the model under different sampling strategies

	No sample	Random up sampling	random subsampling
SVM	0.06	0.51	0.78
Adaboost	0.34	0.62	0.81
Random Forest	0.31	0.31	0.82

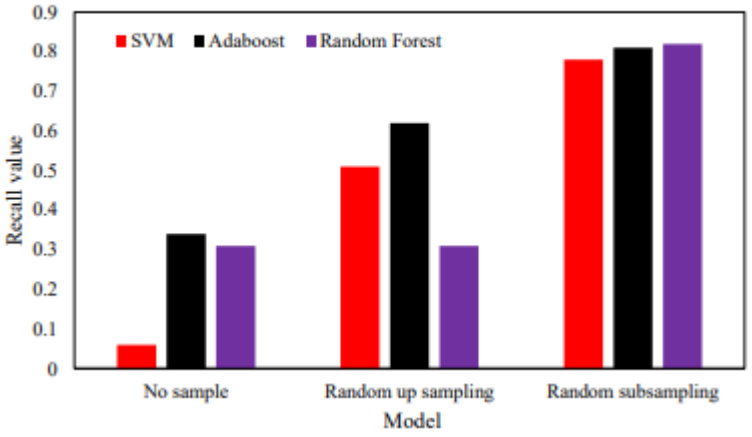

Fig. 2. Recall values of the model under different sampling strategic

As shown in Figure 1 and Table 1, under the random upsampling strategy, the Recall value of SVM model and Adaboost model was improved, while the Recall value of random forest model was not improved. Under the strategy of random subsampling, the Recall values of all three models were improved.

4.2. G-mean value of the model under different sampling strategies

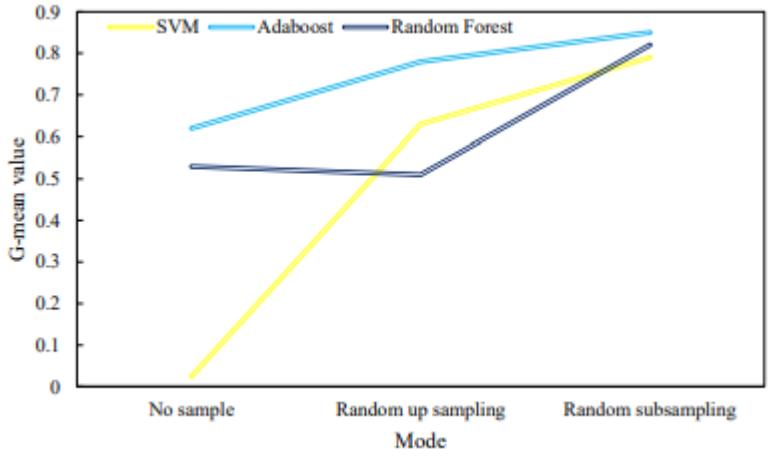

Fig. 3. G-mean value of the model under different sampling strategies

As shown in Figure 2, as for Recall, random subsampling improved G-mean of the three models, while up-sampling only improved G-mean of SVM and Adaboost models. It shows that the lower sampling strategy is better than the upper sampling strategy to solve the problem of data imbalance in the Ponzi scheme of Bitcoin.

In conclusion, the overall performance of SVM algorithm is worse than that of Adaboost and Random Forest. The AUC value of random forest is higher than that of Adaboost in the case of no sampling, random up-sampling and random down-sampling, but the performance of Adaboost algorithm combining Recall value and G-mean value is better. This paper pays more attention to Recall value. Therefore, the Adaboost model is more suitable for the detection of Bitcoin compared with the Random Forest algorithm. In view of the imbalance of Bitcoin data, the up-sampling and down-sampling processing methods both play a certain role in the performance improvement of Adaboost model. In particular, the down-sampling strategy has a significant effect on the performance improvement of Adaboost, but the effect of up-sampling cannot be ignored.

5. Conclusion

Bitcoin, a novel digital currency, has revolutionized the world of finance with its innovative blockchain technology, which has also influenced other Internet industries. As a result, there has been an increase in research on Bitcoin from

various perspectives. This article aims to investigate the theories of blockchain and data mining and their potential applications in the development of Internet finance. The first section of this paper focuses on the fundamental principles of blockchain and data mining. It discusses how blockchain technology works and its impact on Internet finance. Additionally, it explains the concept of data mining, which refers to the process of extracting valuable information from large datasets.

The second section of this paper outlines the construction of a Bitcoin dataset using transaction data. By capturing the transaction data that corresponds to each Bitcoin address in the dataset through Requests, the transaction characteristics of each address are constructed. This process is crucial in understanding how Bitcoin works and how transactions take place in the Bitcoin network. In the third section of this article, correlation analysis is conducted on each characteristic of the Bitcoin dataset. The evaluation results of the model indicate that the Adaboost model shows better performance overall. This section highlights the importance of understanding the transaction characteristics of Bitcoin in predicting its future performance. In conclusion, this paper provides a comprehensive overview of blockchain technology and data mining, their applications in Internet finance, and the construction and analysis of Bitcoin transaction data. The study of Bitcoin's transaction characteristics is critical in understanding its potential as a digital currency and predicting its future performance. Overall, this article contributes to the ongoing discussion on the role of Bitcoin in the development of Internet finance.

References

[1] K. Tulkinbekov and D.-H. Kim, “Blockchain-enabled Approach for Big Data Processing in Edge Computing,” IEEE Internet Things J., vol. 9, no. 19, pp. 18473–18486, 2022.

[2] R. Rahim, R. Patan, R. Manikandan, and S. R. Kumar, “Introduction to blockchain and big data,” in Blockchain, Big Data and Machine Learning, CRC Press, 2020, pp. 1–23.

[3] F. A. Acheampong, “Big data, machine learning and the BlockChain technology: an overview,” Int. J. Comput. Appl., vol. 975, p. 8887, 2018.

[4] E. Karafiloski and A. Mishev, “Blockchain solutions for big data challenges: A literature review,” in IEEE EUROCON 2017-17th International Conference on Smart Technologies, 2017, pp. 763–768.

[5] R. Zheng, J. Jiang, X. Hao, W. Ren, F. Xiong, and Y. Ren, “bcBIM: A blockchain-based big data model for BIM modification audit and provenance in mobile cloud,” Math. Probl. Eng., vol. 2019, 2019.

[6] L. Yue, H. Junqin, Q. Shengzhi, and W. Ruijin, “Big data model of security sharing based on blockchain,” in 2017 3rd International Conference on Big Data Computing and Communications (BIGCOM), 2017, pp. 117–121.

[7] J. Li, M. S. Herdem, J. Nathwani, and J. Z. Wen, “Methods and Applications for Artificial Intelligence, Big Data, Internet-of-Things, and Blockchain in Smart Energy Management,” Energy AI, p. 100208, 2022.

[8] F. Li, X. Yu, R. Ge, Y. Wang, Y. Cui, and H. Zhou, “BCSE: Blockchain-based trusted service evaluation model over big data,” Big Data Min. Anal., vol. 5, no. 1, pp. 1–14, 2021.

[9] H. Yu, Z. Yang, and R. O. Sinnott, “Decentralized big data auditing for smart city environments leveraging blockchain technology,” IEEE Access, vol. 7, pp. 6288–6296, 2018.

[10] J. Zhang, “Interaction design research based on large data rule mining and blockchain communication technology,” Soft Comput., vol. 24, no. 21, pp. 16593–16604, 2020.

[11] C. Xu et al., “Making big data open in edges: A resource-efficient blockchain-based approach,” IEEE Trans. Parallel Distrib. Syst., vol. 30, no. 4, pp. 870–882, 2018.

[12] H. Hassani, X. Huang, and E. Silva, “Banking with blockchain-ed big data,” J. Manag. Anal., vol. 5, no. 4, pp. 256–275, 2018.

[13] K. R. Devi, S. Suganyadevi, S. Karthik, and N. Ilayaraja, “Securing Medical Big data through Blockchain technology,” in 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), 2022, vol. 1, pp. 1602–1607.

[14] F. Muheidat, D. Patel, S. Tammisetty, A. T. Lo’ai, and M. Tawalbeh, “Emerging concepts using blockchain and big data,” Procedia Comput. Sci., vol. 198, pp. 15–22, 2022.

[15] V. K. Chattu, “A review of artificial intelligence, big data, and blockchain technology applications in medicine and global health,” Big Data Cogn. Comput., vol. 5, no. 3, p. 41, 2021.

[16] P. Sharma, M. D. Borah, and S. Namasudra, “Improving security of medical big data by using Blockchain technology,” Comput. Electr. Eng., vol. 96, p. 107529, 2021.

[17] P. Raj, K. Saini, and C. Surianarayanan, Blockchain technology and applications. CRC Press, 2020.

[18] H. Hassani, X. Huang, and E. Silva, “Big-crypto: Big data, blockchain and cryptocurrency,” Big Data Cogn. Comput., vol. 2, no. 4, p. 34, 2018.

[19] N. Deepa et al., “A survey on blockchain for big data: approaches, opportunities, and future directions,” Futur. Gener. Comput. Syst., 2022.

[20] H. Honar Pajooh, M. A. Rashid, F. Alam, and S. Demidenko, “IoT Big Data provenance scheme using blockchain on Hadoop ecosystem,” J. Big Data, vol. 8, pp. 1–26, 2021.

[21] N. Tariq et al., “The security of big data in fog-enabled IoT applications including blockchain: A survey,” Sensors, vol. 19, no. 8, p. 1788, 2019.

[22] A. S. Bataineh, J. Bentahar, O. Abdel Wahab, R. Mizouni, and G. Rjoub, “A game-based secure trading of big data and iot services: Blockchain as a two-sided market,” in Service-Oriented Computing: 18th International Conference, ICSOC 2020, Dubai, United Arab Emirates, December 14–17, 2020, Proceedings 18, 2020, pp. 85–100.

[23] F. Zhang and Y. Zhang, “A big data mining and blockchain-enabled security approach for agricultural based on Internet of Things,” Wirel. Commun. Mob. Comput., vol. 2020, pp. 1–8, 2020.