Machine Learning Processing based on Blockchain Data Characteristics

Lu Liu
School of Economics and Management, Guangxi Normal University of Nationalities Chongzuo, Guangxi, China

Abstract: With the rapid development of Internet, big data and AI, the management of huge volumes of data is no longer a problem. The problem nowadays is how to collect enough and secure data to support the processing and analysis. Meanwhile, the blockchain has become a trustworthy data storage platform because of its sharing and unchangeability. By utilizing the combination of blockchain and AI technology, users can process decentralized learning, and optimize their decisions. Blockchain data has many characteristics, such as integrity, validity, traceability, etc. Integrity can provide more comprehensive data for research, and traceability allows users to track data transactions, while ensuring that data is not tampered with. In this paper, the deep learning method is used to process and analyze the blockchain data, and the advantages of the combination of blockchain data and AI are illustrated through experiments.

Keywords: Big Data; AI; Blockchain; Neural Network; Integrity; Traceability

1. Introduction

In the era of big data, artificial intelligence (AI) and blockchain are both one of the most hyped innovations these days. In 2008, the concept of blockchain was first proposed with Bitcoin [1]. The underlying technology of blockchain technology as the earliest bitcoin and cryptocurrency gradually attracted widespread attention from industry and academia. At the same time, for emerging technologies such as AI that require a large amount of trusted data, blockchain data with integrity and non-tamperability [2] provide just a reliable platform to store data. Many researchers have applied blockchain and AI to multi-modal machine learning problems [3], or combined with the Internet of Things [4] and other vertical fields and business affairs [5]-[9].With the continuous development and innovation of blockchain technology in recent years, its application has made great progress in the fields of financial services, credit and ownership management, resource sharing, trade management, and Internet of Things(IoT) [10] [11]. For instance, the blockchain technology has played an important role in the IoT. The IoT refers to the use of network technology to connect sensors, controllers, machinery and equipment, and realize the intelligent management and operation of machinery and equipment through the connection of objects. The integration of technology will greatly improve work efficiency and reduce costs [12] especially when it is related to big data, AI, blockchain, etc.

But so far, the relevant papers on blockchain lacks a comprehensive review and research on the role of its data in the AI environment. This paper illustrates the significance of the integrity and traceability of blockchain data in real life and the prospects for future applications.

2. Basic Principles and Methods

The blockchain is stored on a distributed network system composed of multiple nodes. Each complete node stores a copy of the entire blockchain, and the nodes share transaction information through the network. At the same time, the blockchain is also a transaction database that holds information shared by all nodes in the system. To achieve AI not only requires a complete database, but also the integrity, validity and authenticity of the data in the database. Machine learning is the core of AI, it is the fundamental way to make computers intelligent. Most of them use different depth neural network frameworks to learn, infer and predict massive data or information and apply them in many fields. Through the combination of AI and blockchain technology, more secure and trusted shared

data can be accessed. Especially when the idea of combining blockchain, AI and Internet of Things was put forward, a large number of scholars have been explored for theoretical research and application fields. This paper uses machine learning methods to discuss the validity and traceability of blockchain data. The validity means that the data collected through the blockchain is almost complete and unchangeable. The traceability means that it is not controlled by the centralized server, the number of assets can be tracked through the chain structure and trading activities. Figure 1 is a general idea of the combination of blockchain and AI: combining the effective data collected on the blockchain system with the deep neural network, the results can be applied to daily life through the Internet of Things.

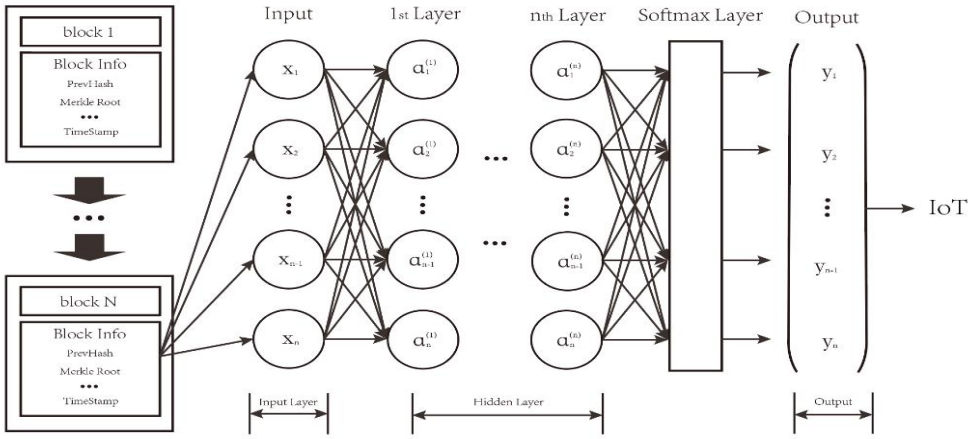


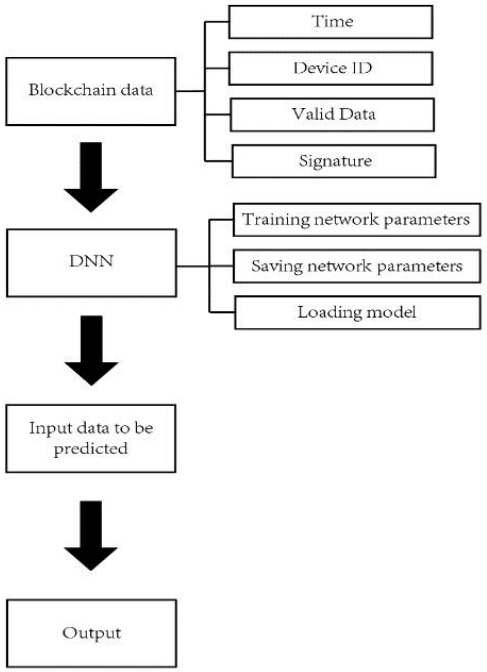
Figure 1. Blockchain Combined with AI

2.1 Data validity experiment

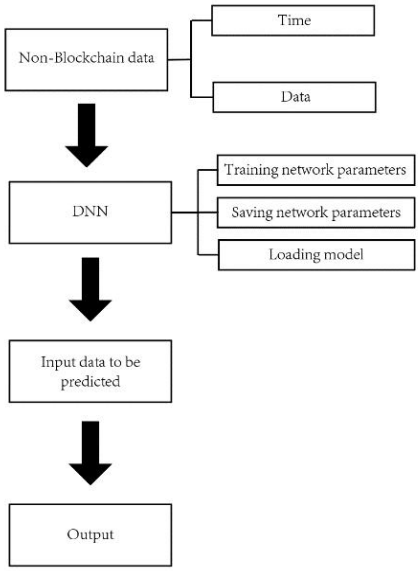
To illustrate the validity of the data, we used blockchain data and Non-blockchain data for comparison experiments.

The blockchain system was used to record the temperature and humidity of the closed room collected by the two sensors. The collection time was collected from midnight to 7:00 am and every half minute to ensure that no personnel flow affected the experimental data. Ignoring the collection error and environmental impact of the sensor itself, we can guarantee that the data obtained is completely accurate and has not been tampered with. We can build a model through a neural network and enter the corresponding temperature to predict the corresponding humidity at this temperature. Such data and models can be combined with the Internet of Things, used in the storage of fresh and perishable items, to ensure that temperature and humidity fluctuate within a small range, creating a good storage environment. The experimental process is shown in Figure 2.

(a) in Fig.2 is data obtained by using the blockchain technique, and the data includes time, device ID, signature, and temperature and humidity values, and the blockchain system determines whether the obtained data is legal (true/false). The illegal data is removed, and the legal data is input into the training network of the neural network. (b) in Fig.2 is non-blockchain data. The data includes values for time and temperature and humidity, allowing data to be entered directly into the neural network.



(a)



(b)

Figure 2. Effectiveness Experiment Flow Chart

When training a neural network, you need to set the learning rate to control the update speed of the parameters. The exponential decay learning rate is used in the model. First, a large learning rate is used to quickly obtain a better solution. Then, as the iteration continues, the learning rate is gradually reduced, making the model more stable in the later stage of training. The formula for exponential decay learning rate is as (1).

The Early stop method is used in the model training process to prevent over-fitting, when all training samples end a forward pass and a reverse pass, the accuracy of the validation set is calculated and the training is stopped when the accuracy is no longer increased.

In the process of optimizing the model, the Adam optimization algorithm is selected, which can design independent adaptive learning rate for different parameters by calculating the first moment estimation and second moment estimation of the gradient. The formula is as (2).

$$decayed\_learning\_rate = learning\_rate \cdot decay\_rate^{(global\_step/decay\_steps)} \tag{1}$$

$$\begin{aligned} m_t &= \mu * m_{t-1} + (1-\mu) * g_t \\ n_t &= \nu * n_{t-1} + (1-\nu) * g_t^2 \\ \hat{m}_t &= \frac{m_t}{1-\mu^t} \\ \hat{n}_t &= \frac{n_t}{1-\nu^t} \\ \Delta\theta_t &= -\frac{\hat{m}_t}{\sqrt{\hat{n}_t} + e} * \eta \end{aligned} \tag{2}$$

Where m_t , n_t are the first moment estimate and the second moment estimate of the gradient, respectively, which can be regarded as the estimate of the expected $E|gt|$ and $E|gt|^2$; $\hat{m}_t$, $\hat{n}_t$ is the correction of m_t , n_t , which can be approximated as the unbiased estimate of the expectation , $\Delta\theta_t = -\frac{\hat{m}_t}{\sqrt{\hat{n}_t} + e} * \eta$ is a dynamic constraint on the learning rate.

The loss function is a criterion for evaluating the performance of the model. In this paper, the mean square error (MSE) is used as the loss function of the model. The formula is as follows:

$$MSE(y, y') = \frac{\sum_{i=1}^{n} (y_i - y_i')^2}{n} \tag{3}$$

Train the network parameters through the input data, load the trained parameters into the network and save the network model. When we input the prediction data, we can get the corresponding prediction results.

2.2 Data Traceability Experiment

Because the blockchain data records the user's identity information, we can regard the machine model in the data as its identity. Each machine collects different data because of its own differences, so if we know its identity information It is possible to have a general understanding of the source of the data through the characteristics of the data.

Firstly, the order of the data is scrambled and divided into two parts: the test set and the verification set. The test set data accounts for 80% of the total data, and the verification set data accounts for 20% of the total data. Secondly, the data is tagged according to its identity information. Finally, the identity information of the data is classified by the method of neural network through the correspondence between temperature and humidity. The training set is used to train the network model, and the verification set is used to verify the classification accuracy of the network model. If you know the data corresponding to the temperature and humidity, you can guess which machine is collecting the

probability of the data. The experimental process is shown in Fig.3.

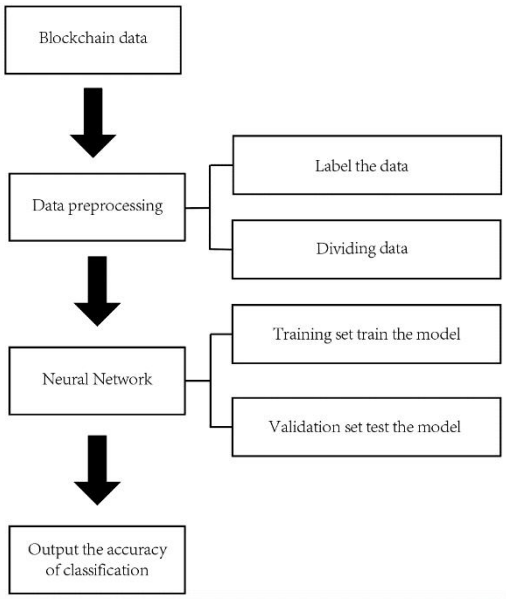


Figure 3. Traceability Experiment Flow Chart

The experiment shuffles the order between the data through the shuffle operation, randomizing the data to avoid overfitting. In the model, the exponential decay learning rate and the Adam optimization algorithm are also used, and the Batch Normalization (BN) layer is added to avoid gradient disappearance or gradient explosion, and the network convergence speed can be accelerated. Use the logarithmic loss function. The loss function expression is as follows:

$$L\left(Y, P\left(Y \mid X\right)\right) = -\log P\left(Y \mid X\right) \tag{4}$$

The final experimental results will output the network structure accuracy rate and the classification accuracy of the verification set.

3. Results and analysis

Through the principles and methods of the previous section, the experimental results are as follows.

The experimental results of 2.1 are shown in Table 1.

The experimental results show that when the predicted temperature is 31.2℃, the predicted humidity of the blockchain data is 62.4%, and the predicted humidity of the Non-blockchain data is 72.18%;when the predicted temperature is 28.3℃, the predicted humidity of the blockchain data is 73.15%, the Non-blockchain data is 61.07%. It can be seen from the above experimental results that the results obtained by the two different data sources are quite different. When the Non-blockchain data is at 28.3℃ and 27.7℃, the corresponding humidity value is abnormal. This may be due to the fact that the Non-blockchain data contains illegal data or the data has been tampered with. The experimental results show that the blockchain technology can guarantee the authenticity, accuracy and integrity of the data, and the experimental results obtained through the blockchain data are also credible.

Table 1. Part of the experimental Results in Section 2.1

Data Sources	Predicted temperature value/℃	redicted umidity alue/%
Blockchain data	31.2	62.4
	28.3	73.15
	27.7	81.21
	26.5	84.4
	24.8	87.32
Non-blockchain data	31.2	72.18
	28.3	61.07
	27.7	59.38
	26.5	80.05
	24.8	81.43

The experimental results in section 2.2 are shown in Fig. 4:

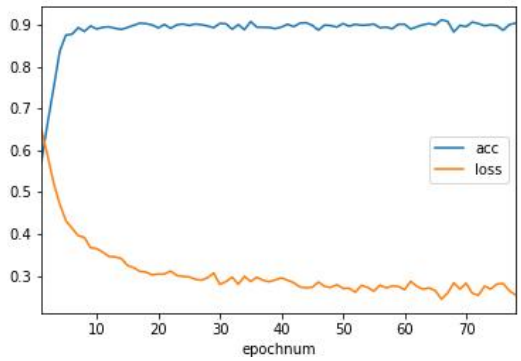


Figure 4. Section 2.2 Experimental Results

By observing the loss function curve of Fig. 4, it can be found that after 30 epochs, the total loss tends to be stable and stabilizes at around 0.2, the network model is basically fitted, and the training set accuracy is stable at about 90%. The experimental results show that the accuracy of the verification set data on the network model is 92.39%.

In summary, we can see that the blockchain data can not only ensure the validity and accuracy of the data, but also ensure that the original data can be traced.

4. Conclusion

As an emerging distributed framework

protocol with decentralization, trust, anonymity, traceability, and non-tamperability, blockchain can be applied in many fields, but the application of combination of blockchain and AI method still in its infancy. The research in areas related to smart contract security, privacy, and so on still remains many issues to be solved. In this paper, the deep learning method is used to illustrate the advantages of combining AI and blockchain data together and make the best of the integrity, validity and traceability of data. With the continuous development of the IoT, the demand for information sharing between different systems is increasing. Combining AI, blockchain technology and the IoT is definitely a future development trend.

The combination of AI, blockchain technology, and the Internet of Things can be applied in various fields. For instance, enterprises can use blockchainbased IoT technology to provide creditable information to customers and win their trust in the transactions. All data would be valid and traceable, such as whether the milk powder sold is imported or whether the vegetables are organic, etc. Also the enterprises can implement the predictive analysis such as predict when the store has the most guests.

References

[1]Nakamoto S.Bitcoin: a peer-to-peer electronic cash system[EB/OL]. 2018-11-24.

[2]Ping Z .China Blockchain Technology and Application Development White Paper[M].Bejing, Ministry of Industry and Information Technology,2016. (in Chinese)

[3]Baltrusaitis T, Ahuja C, Morency L P .Multimodal Machine Learning: A Survey and Taxonomy[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018:1-1.

[4]Yeow K, Gani A, Ahmad R W, et al. Decentralized Consensus for Edge-Centric Internet of Things: A Review, Taxonomy, and Research Issues[J].IEEE Access, 2018,6:1513-1524.

[5]Christidis K, Devetsikiotis M.Blockchains and Smart Contracts for the Internet of Things[J] IEEE Access,2016,4:2292-2303.

[6]Tara S,Maede Z, Aiman E, et al.Security Services Using Blockchains: A State of the Art Survey[J].IEEE Communications Surveys & Tutorials,2018:1-1.

[7]Uriarte R B, Nicola R D. Blockchain-Based Decentralized Cloud/Fog Solutions: Challenges, Opportunities, and Standards[J]. IEEE Communications Standards Magazine, 2018, 2(3):22-28.

[8]Li X, Jiang P, Chen T, et al.A Survey on the security of blockchain systems[J]. Future Generation Computer Systems, 2017:S0167739X17318332.

[9]Salah K, Rehman M H, Nizamuddin N, et al.Blockchain for AI: Review and Open Research Challenges[J].IEEE Access, 2019:1-1.

[10]Liu A D, Du X H, Wang N, Li S Z. Research progress of blockchain technology and its application in information security [J]. Journal of Software, 2018, 29(7): 2092-2115. (in Chinese)

[11]Fang W D, Zhang W X, Pan T,et al. Cyber security in blockchain: threats and countermeasures [J]. Journal of Cyber Security, 2018, 3(2): 87-104.

[12]Zhang Y L, Xu C, Xu Z. Blockchain + Open a new era of intelligence.[M].China Machine Press,2019.

[13]Sha Z ,Wu Y W, Zhao G D,et al.Blockchain and Big Data: Building a Smart Economy[M].Bejing**:**People Post Press,2017,43-56. (in Chinese)

[14]Tikhomirov S,Voskresenskaya E,Ivanitskiy I,et al.SmartCheck:static analysis of ethereum smart contracts[J].International Workshop.IEEE Computer Society, 2018:5(6): 9–16.

[15]Yuan Y, Wang Y F. Blockchain: the state of the art and future trends [J]. Acta Automatica Sinica, 2016, 42(4): 481-494.

[16]Min Y, Zhang S B, Hang Z , et al. User trust negotiation model based on two-layer blockchain in heterogeneous alliance system[J]. Journal of Applied Science, 2019, 37(2). (in Chinese)

[17]Huckle S, Bhattacharya R, White M, Beloff N.Internet of things, blockchain and shared economy applications [J].Procedia Computer Science, 2016, 98(C): 461-466.

