A Distributed Approach of Big Data Mining for Financial Fraud Detection in a Supply Chain

Hangjun Zhou[1, *], Guang Sun[1, 2], Sha Fu[1], Xiaoping Fan[1], Wangdong Jiang[1], Shuting Hu[1] and Lingjiao Li[1]

Abstract: Supply Chain Finance (SCF) is important for improving the effectiveness of supply chain capital operations and reducing the overall management cost of a supply chain. In recent years, with the deep integration of supply chain and Internet, Big Data, Artificial Intelligence, Internet of Things, Blockchain, etc., the efficiency of supply chain financial services can be greatly promoted through building more customized risk pricing models and conducting more rigorous investment decision-making processes. However, with the rapid development of new technologies, the SCF data has been massively increased and new financial fraud behaviors or patterns are becoming more covertly scattered among normal ones. The lack of enough capability to handle the big data volumes and mitigate the financial frauds may lead to huge losses in supply chains. In this article, a distributed approach of big data mining is proposed for financial fraud detection in a supply chain, which implements the distributed deep learning model of Convolutional Neural Network (CNN) on big data infrastructure of Apache Spark and Hadoop to speed up the processing of the large dataset in parallel and reduce the processing time significantly. By training and testing on the continually updated SCF dataset, the approach can intelligently and automatically classify the massive data samples and discover the fraudulent financing behaviors, so as to enhance the financial fraud detection with high precision and recall rates, and reduce the losses of frauds in a supply chain.

Keywords: Big data mining, deep learning, fraud detection, supply chain, Internet of Things.

1 Introduction

Supply chain refers to the network structure composed of a set of entities directly involved in the upstream and downstream flows of products, services, finances, and/or information from a source to a customer [Mentzer, William, James et al. (2001); Sari (2018)]. Located at the intersection of logistics, management, collaboration, and finance,

[1] Hunan University of Finance and Economics, Changsha, 410205, China.

[2] College of Engineering, The University of Alabama, Tuscaloosa, USA.

[*] Corresponding Author: Hangjun Zhou. Email: zhjnudt@gmail.com.

Received: 21 January 2020; Accepted: 10 April 2020.

CMC. doi:10.32604/cmc.2020.09834

Supply Chain Finance (SCF) is an approach for entities in a supply chain to jointly create value through means of planning, steering, and controlling the flow of financial resources on an inter-organizational level [Hofmann (2005)]. To conduct SCF, Supply Chain Risk Management (SCRM) is indispensable, and it is defined by National Institute for Standards and Technology as multidisciplinary practice with a number of interconnected enterprise processes that, when performed correctly, will help departments and agencies manage the risk of using information technology products and services [Schlegel and Robert (2014)]. For the banks or financial institutions, one of the fundamental and pivotal mechanisms in SCRM is the improvement of the effective financial fraud detection to mitigate the financial fraud behaviors and reduce the losses in supply chain finance [Katz (2016); Patterson, Goodwin and McGarry (2018)].

In recent years, with the deep integration of supply chain and Internet, Big Data, Artificial Intelligence, Internet of Things, Blockchain, etc., the major entities connected in a supply chain become more diversified and forms a networked ecology. The new smart supply chain system has taken shape characterized by big data support, network sharing, and intelligent collaboration, which in turn provides conditions for the innovation of SCF mode. Compared with the financial modes of supply chain used by many commercial banks, an e-commerce platform can map the data of small and micro suppliers into the data of credit evaluation and develops the new mode of online supply chain finance by relying on the advantages of the Internet, using big data technology [Qi and Deng (2019)] and combining with third-party verification information. The Internet of Things (IoT) technology uses GPS, mobile sensors, biometrics, etc., to systematically and intelligently identify, locate, track and monitor the status of products or assets in supply chains, and then collect and analyze the status data for risk management, resource allocation and financial services. Also, the emergence of Blockchain technology has led to the distributed and decentralized mode of supply chain, making it more flexible and transparent.

However, accompanying with the rapid development of these new technologies, the volume, variety, velocity and value of SCF data has been increasing massively into big data level, which might create the new financial fraud behaviors or patterns that are more covertly scattered and easy to go unnoticed in SCRM processes. For example, the Camsing International Holding Limited (02662.HK) is on an accusation of alleged financial fraud levied by Noah Holdings Ltd., in 2019 that provides supply chain financing about 3.4 billion RMB ($494 million) of asset management products for Camsing. It is reported that through some big e-commerce platforms, Camsing expands sales data by forging supply chain contracts and round-trip transportation with business partners, and the asset-management products backed by Camsing's accounts receivable from big e-commerce enterprises were in danger of default. In this kind of cases, sometimes even with the hidden bribery between individual personnel of big e-commerce buyers and suppliers, the relatively little data of faked e-commerce contracts, consignment bills, financial reporting and profits, etc., were not easy to be discovered completely in the beginning because of the increasingly complicated supply chain networks and the generated massive normal data, which might cause huge losses in the end. In addition, in the inventory pledge financing, the IoT devices are deployed to obtain the supervision status data, but the IoT status data also has the risk of fraud.

The new financial fraud behaviors and the lack of enough capability to deal with the hugely growing data volumes increase the cost of the supply chain management and lead to the serious consequences, which brings the new challenges for supply chain finance to detect potential risks and frauds more intelligently and timely. In this article, a distributed approach of big data mining is proposed for financial fraud detection in a supply chain. Based on the four major modules in the SCRM procedure, the comprehensive approach is designed to implements the distributed deep learning model of Convolutional Neural Network (CNN) on big data infrastructure of Apache Spark and Hadoop to speed up the processing of the large dataset in parallel and reduce the processing time significantly. By training and testing on the continually updated SCF dataset, the approach can intelligently and automatically classify the massive data samples and discover the financing applications of fraudulent affiliate enterprises, faked operation profits, faked contracts, etc., so as to enhance the financial fraud detection with high precision and recall rates, and mitigate the loss of frauds in a supply chain.

The rest of the article is organized as follows. Literature of related works is described in Section 2. Section 3 demonstrates the intelligent fraud detection in risk management procedure of supply chain finance. A distributed approach of big data mining for financial fraud detection in a supply chain is proposed in Section 4. In Section 5, groups of experiments are implemented to evaluate the efficiency of the proposed approach. Conclusions and future works are summarized in Section 6.

2 Related works

Fraud detection is one of the essential parts of a financial risk management system and has been researched for a long time. In order to supervise the financial frauds in supply chain and decrease the potential losses, the researchers and industry experts are devoted to find out and improve the anti-fraud methods to prevent the occurrences of fraudulent business deals as much as possible. Shearer et al. [Shearer and Sarah (1998)] have pointed out although the risk evaluation methods play an important role in traditional bank financing modes, they cannot identify new business fraud risks effectively and accurately with the rise of supply chain finance and logistics finance. Coulter et al. [Coulter and Onumah (2002)] have analyzed the role of warehouse receipt pledge in accelerating the operation of agricultural products, and discussed the implementation process of the financing model and the measures to prevent fraud risks in supply chain finance. In order to reduce risks and avoid frauds, Mahmut et al. [Mahmut and Kevin (2003)] have analyzed the risk of inventory management in inventory pledge financing model and calculates the optimal target of inventory by using the expected utility. For the credit risk pricing, it should be adopted to identify the frauds in supply chain by utilizing the price risk of pledges as the standard, and the specific forms of credit risk pricing models under different circumstances is described [Cossin and Hricko (2003)]. Through quantitative research, Buzacott et al. [Buzacott and Zhang (2004)] have demonstrated the impact of the determination of interest rates and loan quotas on the risk of the development of supply chain financial services. It has been concluded that the interest rates are one of the effective tools for risk management. Barsky et al. [Barsky and Catanach (2005)] have found that the financial risk management in supply chain should consider the entire supply chain transaction process, and a risk assessment model including information control, business process, manpower, basic structure and macro environment has been constructed. Rosenberg et al. [Rosenberg and Schuermann (2006)] have considered the various risks of supply chain finance as a whole part, and built a model through the VaR method to measure the comprehensive risks of credit risk, market risk and operational risk. In the context of economic globalization, the detection of financial frauds has been also researched as the premise of the effective implementation of supply chain financial risk management by identifying risks and taking active measures to deal with risks [Manuj and Mentzer (2008)]. Darrell et al. [Darrell and Kenneth (2009)] have categorized the risks of banks or financial institutions in the supply chain financial business into four categories: credit risk, market risk, operational risk, and systemic risk, and proposed that the credit fraud and risk is the main research focus of supply chain financial risk management. Josef et al. [Josef and Arne (2009)] have conducted risk management research by constructing a system-oriented model and classifying the relevant factors, which stimulate the financial risk of the supply chain, into three factors: financing enterprises, supply chain and environment. Dyckman [Dyckman (2011)] has indicated that the impairment risk is one of the main risks of supply chain finance and there is the risk of the collaborating of buyer and supplier companies to hedge bank funds in supply chain finance. Hu et al. [Hu, Zhang and Zhang (2012)] have combined supply chain relationship and core enterprise credit information to construct an evaluation index system, and then used BP neural network algorithm and support vector machine (SVM) for empirical results comparison. Ghadge et al. [Ghadge and Dani (2012)] have pointed out that the financial frauds of supply chain finance mainly come from the capital flow in the process of supply chain enterprise cooperation and also believed that the management of cash flow in the supply chain is an effective way to control risk. Leon [Leon (2014)] has found that supply chain finance improves the operational efficiency of the entire supply chain through the financing modes of accounts receivable, prepayment and inventory pledge. Kraus et al. [Kraus and Raul (2014)] have indicated that large data volumes and the inability to analyze them enables fraudulent activities to hide in supply chain management, and then developed a data warehouse design supporting generic and reusable store procedure for analytics by using the Benford's law to detect frauds. Vollmer [Vollmer (2015)] has shown that nearly one-third of about 2,600 supply chain professionals said their companies experienced supply chain fraud, and the steps to avoid overpaying has been described. Sharma et al. [Sharma, Bhavna and Vipin (2016)] have demonstrated that Big Data is playing a very significant role to automated fraud detection in supply chain finance, and a theoretical framework has been proposed to analyze how individual enablers of unstructured data impacts supply chain management. In order to build a supply chain traceability system, Tian [Tian (2016)] has utilized the RFID (Radio-Frequency IDentification) and Blockchain technology to gather and transfer the trusted information for reducing agri-food risk. Hackius et al. [Hackius and Moritz (2017)] have conducted an online survey on use case exemplars, barriers, facilitators, etc., of Blockchain in logistics and supply chain management, and found that it should be more carefully to get a rather conservative industry because of the risks and frauds. Kara et al. [Kara, Seniye and Abhijeet (2018)] have embraced data-driven approaches in supply chain risk management and developed a DM-based framework for the identification of risks and frauds in supply chains. For the growing problem of frauds in international shipping, a Bayesian network is developed to investigate whether intelligent fraud detection systems can improve the efficiency by analyzing large sets of historical shipment data [Triepels, Hennie and Ad (2018)]. DuHadway et al. [DuHadway and Steven (2019)] have indicated that the supply chain risk management faces the challenges of fraudulent behaviors, and a framework is proposed for aiding managerial decision to cope with the intentional risks in a supply chain. However, due to the application of new technologies, the fast growing volume, variety, velocity and value of supply chain data makes the existing methods not always effective and fast enough for the anti-fraud classification and prediction in supply chain finance.

3 The intelligent fraud detection in risk management procedure of supply chain finance

With the rapid development of Cloud Computing, Big Data, Artificial Intelligence, Internet of Things, etc., the risk management procedure to detect the financial fraud in supply chain finance is gradually improved to confront new challenges, and the main procedure is depicted in Fig. 1. There are majorly four modules in the procedure.

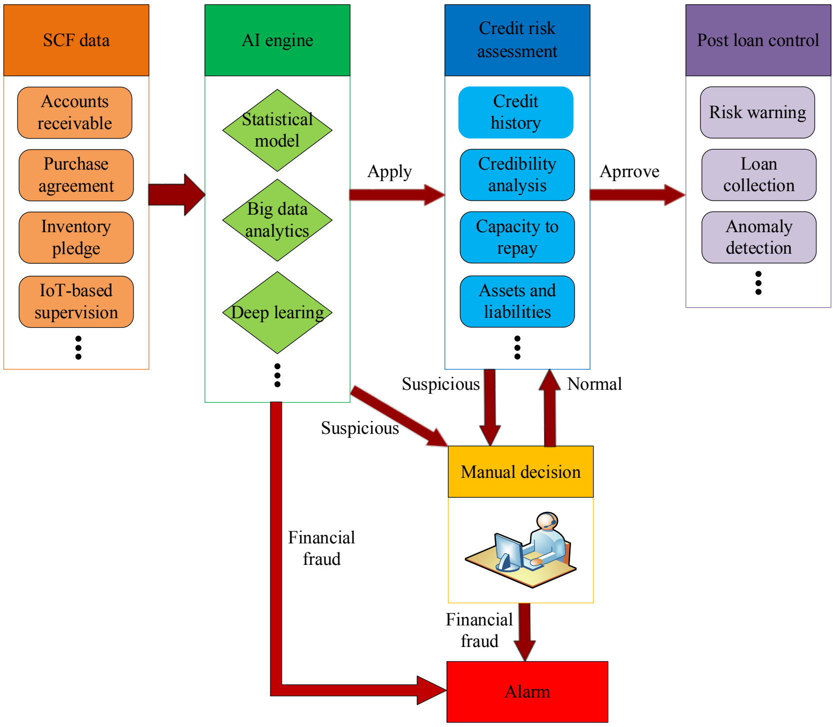

Figure 1: Risk management to detect the financial fraud in supply chain finance

(1) SCF data module

Through using the technology of Cloud Computing and Internet of Things, many platforms of supply chain management, such as O2O supply chain management platforms in China, are established to connect the buyer enterprises, supplier enterprises, banks or trusted financial institutions. The SCF data of each deal, which is mainly about accounts receivable, purchase agreement, inventory pledge, logistics, IoT supervision status, etc., is recorded and stored in a platform. Based on the SCF data, many online or offline financial value-added services could be provided by banks or trusted financial institutions, such as solving the financing problem of medium-sized and small enterprises in China.

(2) AI engine module

On the supply chain management platforms, because of the possible financial fraud transactions, it might not be that all the SCF data are generated from the genuine business deals. Particularly, with the volume, variety, velocity and value of the massive amounts of SCF data becoming big data level, it makes the fraudulent behaviors or patterns more covertly scattered among genuine ones than before. Therefore, confronting the new challenge, the AI engine using traditional statistical models, big data analytics, and deep learning algorithms, is developed and deployed to enhance the capability of risk management systems for serving supply chains. By training and testing on the continually updated SCF dataset, the AI engine is aimed to intelligently and automatically detect the fraudulent business deals in time from the vast transaction data and improve the accuracy and efficiency of anti-fraud, anti-money laundering, anti-bribery, and other aspects in supply chain to a large extent. If the AI engine definitely identifies a financial fraud, it'll raise the alarm immediately in the risk management system. If the result of identification is suspicious, the AI engine will select the suspicious business deal for manual decision.

(3) Credit risk assessment module

After the processing of AI engine, the applications of loans for supply chain enterprises will be evaluated by credit risk assessment module. The enterprise applicants are assessed and classified by algorithms through many dimensions like credit history, credibility analysis, capacity to repay, cash flow ability, assets and liabilities, corporate tax certificate, period of repayment, etc. If a suspicious business deal in supply chain is discovered, usually it'll be sent to manual decision for further assessment. Then, if the result of manual decision is definitely a fraudulent behavior, the alarm will be raised to end the application.

(4) Post loan control module

If a loan is approved for a supply chain enterprise, the post loan control is also indispensable to prevent the occurrence of financial fraud and reduce the repayment risk. It is necessary to timely and irregularly visit the enterprise to clearly be aware of its current operation and finance status, and be certain about whether it can pay back on time. Risk warning, loan collection, anomaly detection, etc., could be conducted according to the enterprise status.

Based on the business characteristics and trade relationship of the supply chain, banks or financial institutions mainly provide three financing modes for supply chain enterprises: accounts receivable, inventory pledge financing and prepayment financing. However, due to the rapid development of technology, new financial fraud behaviors emerge and scatter more covertly than before, and they might not be identified by the traditional black-white list and rule strategy. For instance, the 8 typical factoring steps in supply chain finance are described in Fig. 2, which would function well in normal situations. But considering the current Internet economy, if the buyer is a big e-commerce enterprise, some medium-sized or small supplier fakes a lot of e-commerce contracts, consignment bills, financial reporting and profits, and then asset-management products backed by supplier's accounts receivable could be possibly used to deceive the financing fund. Moreover, sometimes there exist hidden bribery cases between individual personnel of big e-commerce buyer and small supplier. In addition, in the inventory pledge financing, the IoT devices are deployed to obtain the supervision status data, but the IoT status data also has the risk of fraud. Hence, the new emerging financial fraud behaviors bring new challenges for supply chain finance to detect potential risks and frauds more intelligently and timely.

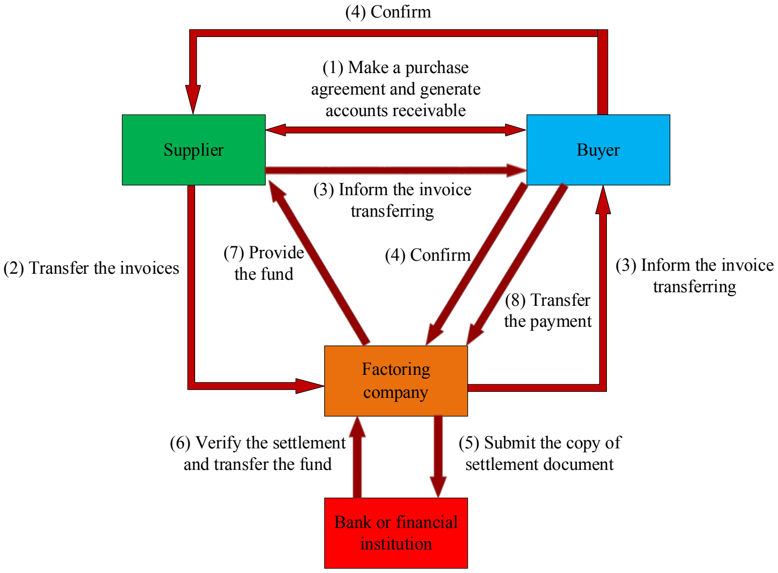
(4) Confirm
(1) Make a purchase agreement and generate accounts receivable
Supplier
Buyer
(3) Inform the invoice transferring
(2) Transfer the invoices
(7) Provide the fund
(4) Confirm
(8) Transfer the payment
(3) Inform the invoice transferring
Factoring company
(6) Verify the settlement and transfer the fund
(5) Submit the copy of settlement document
Bank or financial institution

Figure 2: Factoring in supply chain finance

4 A distributed approach of big data mining for financial fraud detection in a supply chain

4.1 Big data infrastructure of apache spark and Hadoop

When confronting the challenges of new technology and massively growing data, distributed big data clusters are utilized to establish the intelligent risk management platforms to detect the financial frauds as capable and fast as possible. In this article, Apache Spark on Yarn is deployed as the big data infrastructure to run the machine learning algorithms distributedly so as to improve the efficiency of fraud detection in supply chain finance. As can be seen in Fig. 3, at first the Hadoop HDFS is initiated on the cluster of data nodes where the dataset is distributedly stored. Then Spark environment is created and client node uses SparkContext to transform the processing request into Directed Acyclic Graph (DAG) in driver program. The DAG is analyzed into stage tasks and sent to the Resource Manager that has initiated a Node Manager on each Spark worker node. Each Node Manager receives one or several computing tasks and initiates Executor containers to run the tasks, so that the whole data processing can be implemented in parallel on Spark cluster and the general run-time is reduced greatly compared to the traditional serial processing.

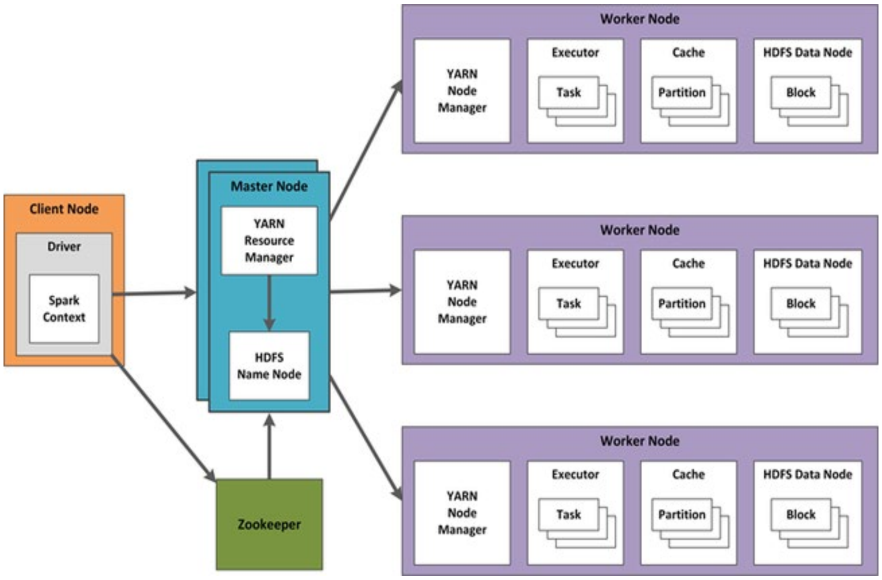

Figure 3: The architecture of Apache Spark on Yarn

4.2. Distributed convolutional neural network

On the Apache Spark infrastructure, distributed Convolutional Neural Network (CNN) model is programmed and run in parallel to use deep learning pipelines archiving large-scale data mining. Convolutional Neural Network (CNN) is one kind of deep machine learning algorithms based on artificial neural network [Shin, Ahn, Lee et al. (2019)]. With the use of local connection and weight sharing, a CNN can maintain the deep structure of the neural network and meanwhile decrease the number of network parameters, which brings about good generalization ability, easy training advantage and better classification effect. As depicted in Fig. 4, a CNN network model is usually composed of input layer, convolution layer, pooling layer, fully connected layer and output layer.

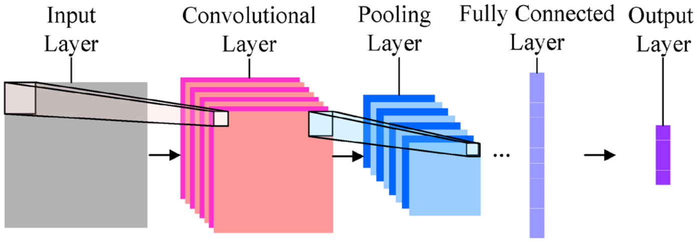

Figure 4: A CNN network model

In a CNN network model, convolutional layer is one of the core layers and always has the highest cost of calculation and complexity. However, by using local connectivity, each neuron node in one layer is merely connected to neuron nodes in the previous and adjacent layer so that the scale of parameters of CNN is reduced. Moreover, the implementation of shared weights also decreases the amount of parameters of CNN model. The convolution of feature maps in previous layer is calculated through the convolution kernel and then the output feature map is computed and obtained with activation functions**.** The formula of convolutional layer is as follows:

$$x_j^l = f\left(\sum_{i\in M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right) \quad (1)$$

where $\mathrm{x}_\mathrm{j}^\mathrm{l}$ represents the j-th feature graph in l-th convolution layer, the symbol "*" represents the operation of convolution, $\mathrm{M_j}$ represents the set of all the input feature maps, $\mathrm{k}_\mathrm{ij}^\mathrm{l}$ represents the weight of the convolution kernel j in the layer l, $\mathrm{b}_\mathrm{j}^\mathrm{l}$ represents the bias, f represents the activation function.

Pooling layer usually has two kinds of functions: Max Pooling and Average Pooling, which are the methods to perform down-sampling operation and decrease the complexity. The formula of pooling operation is:

$$s_j = \frac{1}{|R_j|}\sum_{i\in R_j} a_i \quad (2)$$

where s_j represents the pooling value of j-th pooling region R_j, a_i represents the value in the pooling region outputted from activation function of i-th activation value.

As CNN training on large dataset is computationally expensive and time-consuming in single machine, the distributed environment Apache Spark is used to attain better speedup and economical results compared to supercomputers or GPUs. The implementation of distributed CNN model is aimed to intelligently classify the massive data samples and discover the financing applications of fraudulent affiliate enterprises, faked operation profit, faked contracts, etc., so as to enhance the financial fraud detection in the supply chain with high precision and recall rates. Firstly, the high dimension SCF data is splitting stored in form of HDFS files on the cluster nodes, and Spark Resilient Distributed Dataset (RDD) is initialized through the Spark Context.textfile() function by periodically reading interval sample data from the HDFS files. Then, the preprocessing RDD is created to connect data containers with the pipeline like filtering, mapping, transforming, or shuffling. The number of the processed RDD partitions should match that of worker executors. In feature extraction and pooling step, features are computed according to the convolution kernels through mapper and reducer functions, and max pooling function is applied to perform down-sampling operation. After that, the weight and bias values are randomly initialized and the Spark master node distributes these parameters and network configuration to Spark worker nodes. The backpropagation algorithm is employed to train data on each work node and Stochastic Gradient Descent (SGD) is run in parallel with subset of data for a fixed number of iterations or a fixed length of time. After iterations, the parameters and states on each worker node are sent back to the master node that will average the values of parameters and states to update the trained neural network. Then, the parameters are distributed to worker nodes again for further training until the end.

5 Experimental results

Groups of experiments are carried out on a cluster consisting of 20 identical machines, where one of them is designated as the master node and the rest are designated as worker nodes. Each machine has 8 physical cores and 32 GB of RAM. The operating system is CentOS 7 with Java Development Kit 10.0.2 and Scala 2.12. The stable release version of Apache Spark 2.4.2 is running on top of the cluster resource negotiator Hadoop Yarn and storage file system HDFS [Chan and Thein (2018)].

For the two-category classification problem, confusion matrix is usually selected to evaluate the results. The confusion matrix is a cross-count contingency table to determine the performance of the model based on real categories and prediction categories. For the classification model to detect financial fraud transactions in supply chain, the confusion matrix is described in Tab. 1. The true positive (TP) means the number of correctly predicted financial fraud transactions among all the true fraud transactions in supply chain finance; the false negative (FN) means the number of fraud transactions which are incorrectly predicted as normal transactions; the false positive (FP) means the number of normal transactions which are incorrectly predicted as fraud transactions; the true negative (TN) means the number of correctly predicted normal transactions among all the true normal transactions.

Table 1: The confusion matrix of financial fraud detection in supply chain

Real categories	Prediction categories	
	Fraud transaction	Normal transaction
Fraud transaction	TP	FN
Normal transaction	FP	TN

The experiment results are evaluated in terms of Recall, Precision, and F1-Score, which are defined in Tab. 2. The Precision is the synonym for the positive predictive value. The Recall is synonym for the true positive rate. The F1-Score is the harmonic mean of precision and recall. In addition, the computation time is also evaluated in the experiments.

Table 2: Performance metrics table

Performance metrics	Formulas
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-Score	$2\times\frac{Precision\times Recall}{Precision+Recall}$

The original experimental dataset is obtained from a large O2O supply chain management platform in China. Because of the characteristics of multidimensional data, the raw dataset might have the possibility to be affected by some abnormal data, so it should be preprocessed before the neural network model training and testing. After the data preprocessing, there are 136820 samples of financing applications over almost last 4 years in the dataset and each sample has over 100 dimensions including the data of enterprise name, legal representative, balance sheet, commerce contracts, payment records, consignment bills, financial profits, accounts receivable, purchase agreement, inventory pledge, logistics information, IoT supervision status, etc. For the reason to maintain data confidentiality of sensitive information, not all the data dimensions are mentioned. Among all samples, there are 10524 fraud ones that is less than 10% of the entire dataset, and usually the dataset for fraud detection is an imbalanced dataset. In order to evaluate the classification results of different machine learning models, the dataset is divided into 4 subsidiary datasets to conduct the cross-validation. Each time the ration of training data and testing data is nearly 2:1.

The distributed machine learning algorithms of CNN, SVM, Decision Tree are programed and run in the Apache Spark to compare the results of groups of experiments. The experimental results of precision comparison on 4 sub datasets are demonstrated in Fig. 5. After model training, all of the precision test results of fraud detection in supply chain finance are over 62% on 4 datasets. The results of SVM model are better than those of Decision Tree model, but they are still under 85% because of the high dimensions of datasets. Through the iterations of training, the test results of CNN model show that it detects more financial fraud samples as true positives or less normal samples as false positives. The highest precision rate of CNN model is close to 93% on dataset 3 and the average precision rate of it is over 91%. Although the lowest precision rate of CNN model is under 90% on dataset 1, it still performs better than the other 2 models.

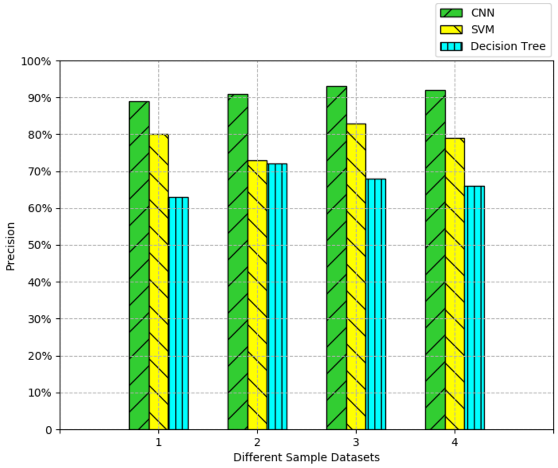

Figure 5: Precision rates on different datasets

The experimental results of recall rate tests are described in Fig. 6. As shown in the figure, the Decision Tree model on Spark has the lowest recall rates on 4 datasets, most of which are under 70%. The recall rates of SVM model are higher than those of Decision Tree on all datasets, but all of them are under 90%. Only one of them is close to that of CNN model on dataset 3. The recall rates of CNN model are not less than 90% on 4 datasets and highest rate of it is nearly 94% on dataset 2, which means most of the financial frauds are detected and identified as true positives in the experiments.

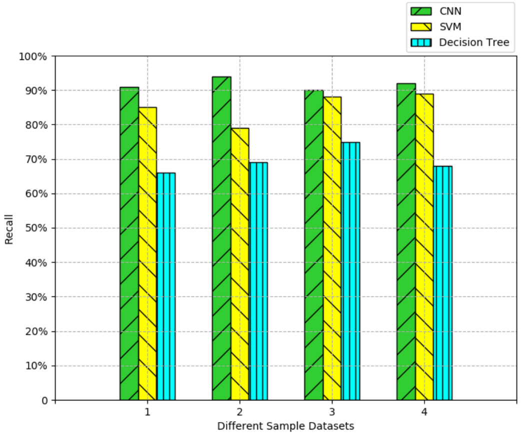

Figure 6: Recall rates on different datasets

The F1-Score considers both the precision rate and the recall rare of the test to compute the score. In Fig. 7, the results of F1 score are the harmonic mean of corresponding precision rates and recall rates, which are above 64% on all datasets. The F1-Score results of CNN model on Spark are better than those of the other 2 models since they are not less than 90% and the highest one is almost 92.5% on dataset 2.

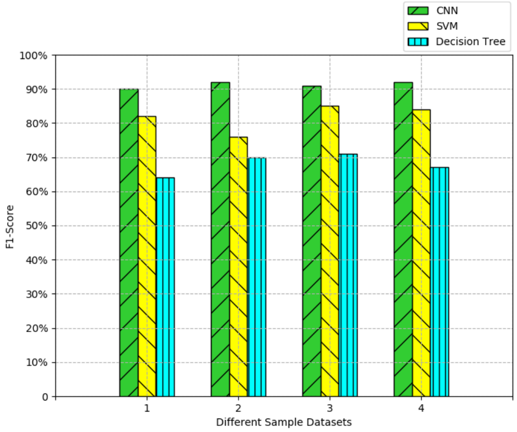

Figure 7: F1-Score on different datasets

In Fig. 8, the experimental results are shown about the training computation time differences of traditional CNN model on one node and distributed CNN on the Spark cluster. In traditional fraud detection of supply chain finance, the CNN model is usually programmed in R language and trained on one node. Here, the distributed CNN model is programed in Scala language and trained on the Spark cluster of 20 machines. Both training experiments are fed with samples of dataset 2. The results show that the training computation time of the CNN model on one node is over 60 hours, while that of the CNN model on Spark is greatly reduced to less than 1 hour. The parallel processing pipelines significantly speeds up the training of the data classification model.

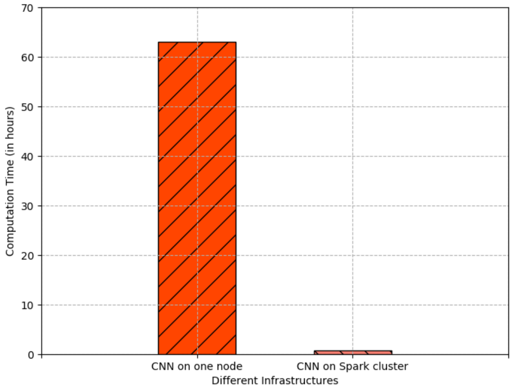

Figure 8: Computation time on different infrastructures

6 Conclusions

Confronting with the new emerging challenges of financial fraud behaviors of SCF, in this article, the four major modules of intelligent fraud detection in risk management procedure of SCF are described: SCF data module, AI engine module, Credit risk assessment module and Post loan control module. Based on the SCRM procedure, a distributed approach of big data mining is proposed for financial fraud detection in a supply chain, which deploys the big data infrastructure of Apache Spark and Hadoop to process the dataset obtained from a large O2O supply chain management platform in China. On the Apache Spark clusters, distributed Convolutional Neural Network (CNN) model is programmed to run in parallel to intelligently classify the massive data samples and discover the fraudulent financing behaviors so as to enhance the efficiency of fraud detection in supply chain finance. The groups of experimental results show that the proposed approach performs better with a higher level of detection precision rate, recall rate, F1-score than the SVM and Decision Tree models. In the meanwhile, compared to the CNN processing on one node, the computation time of the large dataset processed in parallel on Apache Spark is speeded up significantly. In future work, besides the experimental research, the practical application in the field would be developed in a collaborative way. Also, the research work is going to study the distributed machine learning algorithms to optimize the loss rate of assets of financial frauds in supply chains and obtain better performance of reducing the fraud damages.

Acknowledgement: The authors would like to thank the 2011 Collaborative Innovation Center of Big Data for Financial and Economical Asset Development and Utility in Universities of Hunan Province, and also the anonymous reviewers for their valuable comments and insightful suggestions.

Funding Statement: This research work is supported by Hunan Provincial Education Science 13th Five-Year Plan (Grant No. XJK016BXX001, Zhou, H., http://jyt.hunan.gov.cn/jyt/sjyt/jky/index.html), Social Science Foundation of Hunan Province (Grant No. 17YBA049, Zhou, H., https://sk.rednet.cn/channel/7862.html). The work is also supported by Open Foundation for University Innovation Platform from Hunan Province, China (Grand No. 18K103, Sun, G., http://kxjsc.gov.hnedu.cn/).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Barsky, N. P.; Catanach, A. H. (2005): Evaluation business risks in the commercial lending decision. *Commercial Lending Review*, vol. 20, no. 3, pp. 3-10.

Buzacott, J. A.; Zhang, R. Q. (2004): Inventory management with asset-based financing. *Management Science*, vol. 50, no. 9, pp. 1274-1292.

Chan, W. N.; Thein, T. (2018): Sentiment analysis system in big data environment. *Computer Systems Science and Engineering*, vol. 33, no. 3, pp. 187-202.

Cossin, D.; Hricko, T. (2003): A structural analysis of credit risk with risky collateral: a methodology for haircut determination. *Economic Notes*, vol. 32, no. 2, pp. 243-282.

Coulter, J.; Onumah, G. (2002): The role of warehouse receipt systems in enhanced commodity marketing and rural livelihoods in Africa. *Food Policy*, vol. 27, no. 4, pp. 319-337.

Darrell, D.; Kenneth, J. (2009): *Credit Risk: Pricing, Measurement and Management*. Shanghai University of Finance & Economics Press.

DuHadway, S.; Steven, C. (2019): Malicious supply chain risk: a literature review and future directions. *Revisiting Supply Chain Risk*, vol. 7, no. 1, pp. 221-231.

Dyckman, B. (2011): Supply chain finance: risk mitigation and revenue growth. *Journal of Corporate Treasury Management*, vol. 4, no. 2, pp. 168-173.

Ghadge, A.; Dani, S. (2012): Supply chain risk management: present and future scope. *International Journal of Logistics Management*, vol. 23, no. 3, pp. 313-339.

Hackius, N.; Moritz, P. (2017): Blockchain in logistics and supply chain: trick or treat? *Proceedings of Hamburg International Conference of Logistics*, pp. 3-18.

Hofmann, E. (2005): Supply chain finance: some conceptual insights. *Beiträge Zu Beschaffung Und Logistik*, vol. 3, no. 2, pp. 203-214.

Hu, H.; Zhang, L.; Zhang, D. (2012): Research on the credit risk evaluation of medium-sized and small enterprises in supply chain-the comparative study based on SVM and BP neural network. *Management Review*, vol. 24, no. 11, pp. 70-80.

Josef, O.; Arne, Z. (2009): System-oriented supply chain risk management. *Production Planning & Control*, vol. 20, no. 4, pp. 343-361.

Kara, M.; Seniye, F.; Abhijeet, G. (2018): A data mining-based framework for supply chain risk management. *Computers & Industrial Engineering*, vol. 126, no. 1, pp. 105-117.

Katz, N. A. (2016): *Detecting and Reducing Supply Chain Fraud*. Routledge Press.

Kraus, C.; Raul, V. (2014): A data warehouse design for the detection of fraud in the supply chain by using the Bedford's law. *American Journal of Applied Sciences*, vol. 11, no. 9, pp. 1507-1518.

Leon, B. (2014): Supply chain finance: the next big opportunity. *Supply Chain*

Management Review, vol. 3, no. 1, pp. 57-60.

Mahmut, P.; Kevin, W. (2003): Balancing desirable but conflicting objectives in the newsvendor problem. *IIE Transactions*, vol. 35, no. 2, pp. 131-142.

Manuj, I.; Mentzer, J. T. (2008): Global supply chain risk management strategies. *Journal of Business Logistics*, vol. 29, no. 1, pp. 133-155.

Mentzer, J. T.; William, D.; James, S. K.; Soonhong, M.; Nancy, W. N. et al. (2001): Defining supply chain management. *Journal of Business Logistics*, vol. 22, no. 2, pp. 1-25.

Patterson, J. L.; Goodwin, K. N; McGarry, J. L. (2018): Understanding and mitigating supply chain fraud. *Journal of Marketing Development and Competitiveness*, vol. 12, no. 1, pp. 70-83.

Qi, E. N.; Deng, M. (2019): R & D investment enhance the financial performance of company driven by big data computing and analysis. *Computer Systems Science and Engineering*, vol. 34, no. 4, pp. 237-248.

Rosenberg, J. V.; Schuermann, T. (2006): A general approach to integrated risk management with skewed, fat-tailed risks. *Journal of Financial Economics*, vol. 79, no. 3, pp. 569-614.

Sari, K. (2018): Modeling of a fuzzy expert system for choosing an appropriate supply chain collaboration strategy. *Intelligent Automation and Soft Computing*, vol. 24, no. 2, pp. 405-412.

Schlegel, G. L.; Robert, J. T. (2014): *Supply Chain Risk Management: An Emerging Discipline*. CRC Press.

Sharma, V.; Bhavna, P.; Vipin, K. (2016): Importance of big data in financial fraud detection. *International Journal of Automation and Logistics*, vol. 2, no. 4, pp. 332-348.

Shearer, A. T.; Sarah, K. D. (1998): Shortcomings of risk ratings impede success in commercial lending. *Commercial Lending Review*, vol. 14, no. 1, pp. 21-22.

Shin, H. K.; Ahn, Y. H.; Lee, S. H.; Kim, H. Y. (2019): Digital vision based concrete compressive strength evaluating model using deep Convolutional Neural Network. *Computers, Materials & Continua*, vol. 61, no. 3, pp. 911-928.

Tian, F. (2016): An agri-food supply chain traceability system for China based on RFID & blockchain technology. *Proceedings of 2016 13th International Conference on Service Systems and Service Management*, pp. 1-6.

Triepels, R.; Hennie, D.; Ad, F. (2018): Data-driven fraud detection in international shipping. *Expert Systems with Applications*, vol. 99, no.1, pp. 193-202.

Vollmer, S. (2015): Monitoring fraud risks in the supply chain. *Journal of Accountancy*, vol. 219, no. 4, pp. 25-26.