ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940www.ijasem.org

Vol 8, Issuse.4 Oct 2020

A Big Data Mining Approach of PSO-Based BP Neural Network for Financial Risk Management With IoT

M.Arya Bhanu[1],N.Ch.Ravi[2],B.Venkateswarlu[3],M.Naveen[4],

ABSTRACT

In recent years, the technology about IoT (Internet of Things) has been applied into finance domain, and the generated data, such as the real-time data of chattel mortgage supervision with GPS, sensors, network cameras, mobile devices, etc., has been used to improve the capability of financial credit risk management of bank loans. Financial credit risk is by far one of the most significant risks that commercial banks have to face, however, when confronting to the massively growing financial data from multiple sources including Internet, mobile networks or IoT, traditional statistical models and neural network models might not operate fairly or accurately enough for credit risk assessment with those diverse data. Hence, there is a practical need to establish more powerful risk prediction models with artificial intelligence based on big data analytics to predict default behaviors with better accuracy and capacity. In this article, a big data mining approach of Particle Swarm Optimization (PSO) based Backpropagation (BP) neural network is proposed for financial risk management in commercial banks with IoT deployment, which constructs a nonlinear parallel optimizationmodelwithApacheSparkandHadoopHDFStechniquesonthedatasetofon-balancesheetitem and off-balance sheet item. The experiment results indicate that this parallel risk management model has fast convergence rate and powerful predictive capacity, and performs efficiently in screening default behaviors. In the meanwhile, the distributed implementation on big data clusters largely reduces the processing time of model training and testing.

INDEX TERMS

Big data, artificial intelligence, financial risk management, Internet of Things, particle swarm optimization,BPneural network.

INTRODUCTION

With the growing utilization of Internet of Things technology, many IoT-based applications have been developed and deployed in a broad range of fields, such as finance, healthcare, resource management, industry, etc [1]–[3]. For banks and financial organizations, IoT solutions can help them to gain real-time data on their own and their clients' assets, which would lead to more effective evaluation algorithmThe associate editor coordinating the review of this manuscript and approving it for publication was Tie Qiu.of financial risk management [4], [5]. For example, chattel mortgage loans based on traditional financial data and real-time data from IoT equipments like GPS, sensors, network cameras, mobile devices, etc., and relative financial risk evaluation services, have been developed into management standards in many countries like China and South Korea.When confronting to the massively growing financial data with mixing-structured or unstructured formats from multiple sources including Internet, mobile networks or IoT, the risk management and prevention has become more important on research and operation in commercial banks [6]. Before the 1990s, commercial banks mainly evaluate the credit risk of enterprises applying loans based on financial indicator ratios. Commonly used analytical methods are Z-score model, Lgit model, Probit model, etc. In these methods, analytical models are constructed based on various key financial ratios to find out the mapping relationship between financial ration data and credit risk, then the critical value of the financial ratios is obtained according to the occurrences of credit risk so as to decide whether a loan has risks. After the 1990s, many commercial banks use mathematical methods and financial theory to construct statistical models for quantitative analysis of credit risk.

Professor[1,2,3,4], Assistant Professor[1,2,3,4], ,Associate Professor[1,2,3,4],
Department of CSE Engineering,
Pallavi Engineering College,
Mail ID:mabhanuu@gmail.com, Mail ID:ravi@saimail.com,
Kuntloor(V),Hayathnagar(M),Hyderabad,R.R.Dist.-501505.

The mainstream models include KMV method, Credit Metrics, Credit Risk+, etc. Nevertheless, due to the shortcomings of these statistical models like strict financial assumptions, and that credit risk analysis of bank loan itself is a nonlinear problem, many researchers consider applying nonlinear models such as neural network to conduct the classification and prediction. These neural network models are usually running on a single machine and successfully applied to the management of relatively small sample dataset without strict financial assumptions.However, in the last decade due to the prevailing of Internet, mobile network and IoT, the availability and amount of financial data has been increasing dramatically in the volume, variety, velocity and value [7]. Moreover, with the application of IoT solutions, huge amount of data also generates in target tracking, environmental monitoring and information collecting [8], [9]. For financial risk management with IoT-based chattel mortgage loans, the possible default behaviors could scatter more covertly and easily with normal ones than before in this kind of diverse financial data, so that traditional methods seem not to function fairly or accurately enough when confronted to this new scenario. That is why recently there is a practical need for more powerful risk prediction models of artificial intelligence based on big data mining to predict default behaviors with better accuracy and capacity. In this article, a big data mining approach of PSO based BP (Back-Propagation) neural network for financial risk management is proposed to construct large-scale nonlinear parallel optimization models by training, validating and testing on the dataset obtained from a large commercial bank with IoT-based services in China. Through evaluating the data of on-balance sheet item and off-balance sheet item on Apache Spark and Hadoop HDFS, the experiment results indicate that this parallel risk management model has fast convergence and powerful predictive capacity, and performs efficiently in screening default behaviors. In the meanwhile, the distributed implementation on big data clusters could largely reduce the processing time of model training, validating and testing.The rest of the article is organized as follows. Literature of related works is described in Section 2. Section 3 introduces the models of PSO based BP Neural Network. A big data mining approach of PSO based BP neural network for financial risk management is proposed in Section 4. In Section 5, groups of experiments are implemented to evaluate the classification and prediction efficiency of the proposed model. Conclusions are summarized in Section 6.

Related Works

Finical risk management generally refers to a comprehensive evaluation of the borrower's current financial status, credit status, and future development status. Credit risk assessment is one of core contents of financial risk management. The evaluation result of credit risk of enterprises applying loans directly affects the work of banks, such as how to prevent frauds and risks, avoid financial loss, reduce the cost of risk control, etc.The evaluation method based on financial ratios is firstly proposed for prediction through analyzing the effects of single financial ratio and multivariate discriminant analysis[10].Afterthat,AZScoreModelisconstructedto analyze the bank loan cases based on extracting the most effective financial ratios and evaluate the financial status and credit risks [11]. In commercial fields, statistical models based on mathematics and financial theory, such as KMV and Credit Metrics [12], are designed for quantitative analysis of credit risk, but they have too strict financial assumption.With the development of computer technology and artificial neural network theory, many researchers pay attention to use neural network to establish nonlinear models to evaluate the credit risk of commercial bank loans. A neural network model developed for bankruptcy prediction by testing financial data from various companies [13]. Through a comparison of the predictive abilities of both the neural network and the discriminant analysis method, the results show that neural networks might be more applicable and effective. A comparison is made between traditional statistical methodologies for distress classification and prediction with neural networksAnalyzeover1000healthy,vulnerableandunsoundindustrial Italian firms from 1982-1992 [14], and the results indicate a balanced degree of accuracy and other beneficial characteristics. In order to screen potential defaulters on consumer loans, a study compares the performance of artificial neuro-fuzzy inference systems and multiple discriminant analysis models, and finds that the neuro-fuzzy system performs better than the multiple discriminant analysis approach to identify bad credit applications [15]. Support vector machines (SVM) is evaluated with BP neural network [16] to conduct a market comparative analysis on the differences of determining factors in the United States and Taiwan markets and the interpretability of the Artificial Intelligence based models is improved. Three ensemble strategies of cross validation, bagging, and boosting are investigates based on the multilayer perceptron neural network [17] and the generalization ability of the neural network ensemble is found to be superior to the single best model for three real world financial decision applications. A comparative analysis of artificial neural network and linear logistic regression with panel data is introduced based on a database of 1434 files of credits granted to industrial Tunisian companies by a commercial bank from 2003 to 2006 [18]. The results show that the multilayer neural network model is better and the best information set is the one combining accrual, cash-flow and collateral variables. Several non-parametric credit-scoring models are built on the multilayer perceptron with a sample dataset of almost 5500 borrowers from a Peruvian microfinance institution [19] and the results reveal neural network models outperform the other three classic techniques both in AUC and misclassification costs. A credit scoring model is proposed using artificial neural networks in classifying peer-to-peer loan applications into default and non-default groups [20] to demonstrate that the neural network-based credit scoring model performs effectively in screening default applications. A hybrid model of discriminant neural networks is designed to study the risk of failure of Moroccan firms with the data availability and reliability [21]. The dynamic model considers the firms' three-year behavior to predict risk failure and adapts to financial environment. A method of adaptive Particle Swarm Optimization (PSO) based Fuzzy Support Vector Machines is developed to minimize the influence of outlier in finding the best hyper plane to give the highest accuracy for each process of risk analysis [22].As the processing volume of financial risk assessment databases increasing so greatly, it has become necessary to apply the solution of big data techniques for the classification and prediction of massive financial datasets. A linear mixed model with big data techniques and algorithms is implemented to calculate the credit risk of financial companies [23]. The results show that faster and unbiased estimators could be archived with big data techniques to extract the value of data and thus better decisions can be made without the runtime component, which would be less risk for financial companies when predicting which clients will be successful in their payments. Moreover, a study is conducted to leverage alternative big data source and make unique combination of datasets, including call-detail records, credit, and debit account information of customers, to create scorecards for credit card applicants. The results show that combining call-detail records with traditional data in credit scoring models increases their performance when measured in AUC [24].In the era of artificial intelligence based on big data analytics, traditional statistical methods and neural network approaches running on single machine may not be sufficient in processing large-scale datasets [25]. Therefore, in this article a nonlinear and parallel PSO-BP neural network approach is proposed and implemented on a distributed cluster to process the big data set of on-balance sheet item and off-balance sheet item for better prediction and efficient risk management.

THE MODELS OF PSO BASED BP NEURAL NETWORKBP

(Back-Propagation) Neural Network is one of the classic fully-connected neural network structures, which is usually composed of input layer, hidden layer and output layer. Generally, BP neural network uses gradient descent optimization algorithm to adjust the weight of neurons by calculating the gradient of the loss function.

For a classic three layer BP neural network, let $X = (x_1,x_2,...,x_M)^T$ denotes a input vector, where M is the number of features of a input vector; $W = (w_1,w_2,...,w_M)$ denotes the weight vector between input layer and hidden layer; $\theta = \theta_1,\theta_2,...,\theta_q{}^T$denotes the bias vector, where q is the number of neurons in hidden layer; $\emptyset(x)$ denotes the transfer function of hidden layer; $\alpha = (\alpha_1,\alpha_2,...,\alpha_L)^T$ denotes the bias vector of output layer, where L is number of output $(o_1,o_2,...,o_L)$ denotes the output vector; $\phi(x)$ denotes the transfer function of output layer. The input of hidden layer is as in (1), where $1 \le i \le q$ and $1 \le j \le M$,

$$Net_i = \sum_{j=1}^{M} w_{ij}x_j + \vartheta_i. \tag{1}$$

The output of hidden layer is as in (2),

$$y_i = \emptyset(Net_i). \tag{2}$$

The input of output layer is as in (3), where $1 \le k \le L$,

$$Net_k = \sum_{i=1}^{q} w_{ki}y_i + \alpha_k. \tag{3}$$

The output of output layer is as in (4),

$$o_k = \phi\,(Net_k). \tag{4}$$

It could suppose the E_p is the Mean Squared Error (MSE) of a single sample p, then E_p is denoted as in (5), where T_k is the expected output,

$$E_p = \frac{1}{2}\sum_{k=1}^{L}(T_k - o_k)^2. \tag{5}$$

The total MSE is denoted as in (6), where P is the total number of samples,

$$P \quad L \quad = \frac{1}{2}\sum\sum \quad - o_k^p)^2$$

Suppose $1w_{ki}$ denotes the weight increment of output layer, $1\alpha_k$ denotes the threshold increment of output layer, $1w_{ij}$ denotes the weight increment of hidden layer, $1\theta_i$ denotes the threshold increment of hidden layer. The equations are calculated as in (7) - (10),

$$1w_{ki} = -\eta \frac{\partial E}{\partial w_{ki}} = -\eta \frac{\partial E}{\partial o_k} \cdot \frac{\partial o_k}{\partial Net_k} \cdot \frac{\partial Net_k}{\partial w_{ki}},$$

$$1\alpha_k = -\eta \frac{\partial E}{\partial \alpha_k} = -\eta \frac{\partial E}{\partial o_k} \cdot \frac{\partial o_k}{\partial Net_k} \cdot \frac{\partial Net_k}{\partial \alpha_k},$$

$$1w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} = -\eta \frac{\partial E}{\partial o_k} \cdot \frac{\partial o_k}{\partial Net_k} \cdot \frac{\partial Net_k}{\partial w_{ij}},$$

$$1\vartheta_i = -\eta \frac{\partial E}{\partial \vartheta_i} = -\eta \frac{\partial E}{\partial o_k} \cdot \frac{\partial o_k}{\partial Net_k} \cdot \frac{\partial Net_k}{\partial \vartheta_i}.$$

The calculation of the partial differential equations are as in (11) - (18),

$$\frac{\partial E}{\partial o_k} = -\sum_{p=1}^{P}\sum_{k=1}^{L}\left(T_k^p - o_k^p\right) \quad (11)$$

$$\frac{\partial Net_k}{\partial w_{ki}} = y_j, \quad (12)$$

$$\frac{\partial Net_i}{\partial w_{ij}} = x_j, \quad (13)$$

$$\frac{\partial Net_k}{\partial \alpha_k} = 1, \quad (14)$$

$$\frac{\partial Net_i}{\partial \vartheta_i} = 1, \quad (15)$$

$$\frac{\partial E}{\partial y_i} = -\sum_{p=1}^{P}\sum_{k=1}^{L}\left(T_k^p - o_k^p\right) \cdot \varphi'(Net_k) \cdot w_{ki}, \quad (16)$$

$$\frac{\partial y_i}{\partial Net_i} = \emptyset^0(Net_i), \quad (17)$$

$$\frac{\partial o_k}{\partial Net_k} = \phi^0(Net_k). \quad (18)$$

With the above equations, we can obtain results as in (19) - (22),

$$1w_{ki} = \eta\sum_{p=1}^{P}\sum_{k=1}^{L}\left(T_{kp} - o_k^p\right) \cdot \varphi'(Net_k) \cdot y_i,$$

$$1\alpha_k = \eta\sum_{p=1}^{P}\sum_{k=1}^{L}\left(T_k^p - o_k^p\right) \cdot \varphi'(Net_k),$$

$$1w_{ij} = \eta\sum_{p=1}^{P}\sum_{k=1}^{L}\left(T_{kp} - o_k^p\right) \cdot \varphi'(Net_k) \cdot w_{ki} \cdot \varphi_0(Net_i) \cdot x_j,$$

$$1\vartheta_i = \eta\sum_{p=1}^{P}\sum_{k=1}^{L}\left(T_{kp} - o_k^p\right) \cdot \varphi'(Net_k) \cdot w_{ki} \cdot \varphi_0(Net_i).$$

B. PSO MODEL

In PSO optimization algorithm, the system is initialized with a population of random solutions and searches for optima by updating generations. PSO maintains a swarm of particles which are flying with some velocities in the n-dimensional search space. The particles have no weight and no volume.

Assume that the position of i-th particle in the n-dimensional space is vector $X = (x_1, x_2, \cdots, x_n)$ and the velocity of it is vector $V = (v_1, v_2, \cdots, v_n)$, a particle updates its velocity and positions with equations in (23) - (24).

$$v_{k+1} = v_k + c_1 * rand() * (pbest_k - present_k) + c_2 * rand() * (gbest_k - present_k), \quad (23)$$

$$present_{k+1} = present_k + v_{k+1}. \quad (24)$$

In every iteration, each particle is updated by two best values. *pbest*is the best solution (fitness) each particle has achieved so far. *gbest*is the best value tracked by the particle swarm optimizer, obtained so far by any particle in the population. $v_k$is the current particle velocity and $persent_k$is the current particle (solution). *rand*() function generates a random number between (0,1). c_1, c_2 are learning factors.

PSO-BP NEURAL NETWORK MODEL

As we know, usually BP neural network uses gradient descent method to adjust the connection weights and thresholds. However, these iterations of training procedure are easily immersed in getting local minimum and have slow convergence rate. The PSO algorithm is a global algorithm, which has a strong ability to find global optimal result and a good convergence rate. The idea for PSO based BP neural network model is that at the beginning stage of searching for the optimum, the PSO is applied to accelerate the training speed. When the fitness function value has not changed for some iterations, or value changed is smaller than a predefined number, the searching process is switched to gradient descending searching according to the heuristic knowledge.

In PSO based BP neural network model, the dimension of a particle in the population is denoted as in (25)

$$n = Y_1 \times (A + 1) + (Y_1 + Y_2),$$

where Y_1 is the number of neurons in the hidden layer, Y_2 is the number of neurons in the output layer, A is the number of neurons in the input layer.

A BIG DATA MINING APPROACH ON APACHE SPARKAND HADOOP HDFS

This article proposes a novel big data mining approach of PSO based BP neural network models for financial risk management, which uses Apache Spark on Yarn as the infrastructure to distributedly implement the machine learning algorithms with big dataset so as to improve the efficiency of risk control. As we could see in Figure 1, 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) [26] 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 procedure can be implemented in parallel with MapReduce mechanism on Spark cluster and the general run-time is reduced to obtain better performance and efficiency. In order to acquire the global search ability and advanced optimization with high efficiency, the distributed algorithm of PSO based BP neural network is designed to train data and adjust the connection weights in parallel on an Apache Spark and HDFS cluster. The particle swarm is randomly divided into *N*equal subgroups to generate Spark RDD datasets of particle swarm, which would be processed by Map API to calculate and update the fitness value of each particle, and then thenewRDDdatasetsofparticleswarmareformedaccording to the update of particle swarm position in relative searching direction. The specific steps of distributed algorithm of PSO based BP neural network are depicted in Figure 2. The dataset 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.The initial connection weights of BP neural network is globally optimized by PSO algorithm: setting the value of weights and threshold, obtaining the initial particle position and speed vectors, getting the fitness function with MSE (Mean Squared Error), etc.Spark uses MapReduce mechanism to implement the program in parallel. Spark Context broadcasts the relevantcodesandMaptaskstoeachcorrespondingExecutor in the clusters and constructs the neural network.In each Executor, local samples are read to train the current BP neural network with iterative learning process.Each Executor collects the precision rate and scale of sample records.

The <key, value>pairs are generated and transferred to the Reduce stages, where key denotes the connection weight and value denotes the change amount of the weight.

Reduce tasks are run to group the <key, value> pairs of Map stage and merge the change of weight in each group.

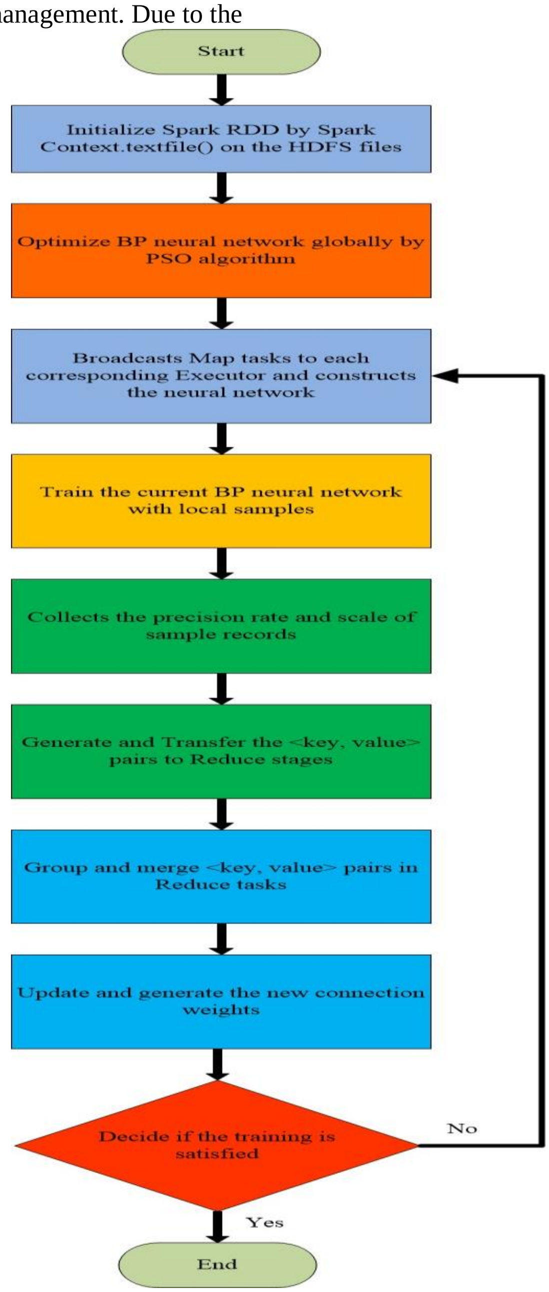
BP neural network is updated and the new connection weights are generated.

The decision is made with two situations: if the result is satisfied, the training procedure is ended; if not, the training procedure is carrying on with looping back to the step (3).

EXPERIMENTS

Groups of experiments are implemented on a cluster consisting of 20 machines, each with 8 cores and 32 GB of RAM. The operating system is CentOS 7 with Java Development Kit 10.0.2 and Scala 2.12.7. The stable release version of Apache Spark2.3.3isrunningontopoftheclusterresourcenegotiator Hadoop Yarn and storage file system HDFS.

The experiment sample dataset is obtained from a large commercial bank in China, where the samples are randomly selected from 1000 companies that had applied bank loans with IoT-based chattel mortgage management. Due to the



Steps of distributed algorithm of PSO based BP neural network.

characteristics of multidimensional data [27], the raw dataset might have the possibility to be affected by the some abnormal data, so it should to be preprocessed before the neural network model training and testing. The dataset covers many industries like information technology, manufacturing, real estate, construction, medicine, etc., and after preprocessing it approximately includes 10000 samples including 6000 samples of on-balance sheet item and 4000 samples of off-balance sheet item. There is also a dataset of 100 samples applying bank loans and it will be predicted and evaluated by the constructed PSO based BP neural network models in this article. The credit score of an evaluated sample is set between 0 and 1. If a score is in the area [0, 0.5), it means it's a default sample or negative sample, and if a score is in the area [0.5, 1], it means it's a normal sample or positive sample. The PSO-BP neural network model is set up with 2 hidden layers containing 8 neurons in each layer based on trial rule. The activation function of hidden layers and output layer is sigmoid function. The maximum iteration number is set 30000 and the iteration error accuracy is 0.63×10^{-3}. The network performance function is set with MSE (Mean Squared Error).

EVALUATIONS ON DATA OF ON-BALANCE SHEET ITEMS

In the dataset, there are 6000 sample data of on-balance sheet item, which has over 30 dimensions in each sample. These 6000 sample data are divided into two subsets: training dataset and testing dataset, and there are 3000 samples in each subset. Among 3000 samples, the number of default samples is 1200 and that of the normal samples is 1800. After the training samples being feed into the PSO based BP neural network model for iterative computing, the network iteration performance reaches the least error 0.612×10^{-3} with 8 neurons in each hidden layer, which satisfies the preset limit 0.63×10^{-3}. The training result is showed in Table 1.

The training result of on-balance sheet item data.

Actual	Predicted		Total
	Positive	Negative	
Positive	1786(99.22%)	14(0.78%)	1800(100%)
Negative	7(0.58%)	1193(99.42%)	1200(100%)

From the table it could be seen that for one hand the precision rate of the training result is 99.22% and the number of wrongly predicted as default samples is 14. For another hand, the number of wrongly predicted as normal samples is 7 among 1200 default samples. The experiment shows that the constructed PSO-BP neural network model has effective result on training data.

After the training stage, the testing dataset, which also has 1800 normal samples and 1200 default samples, is applied into the trained PSO-BP neural network model to verify the effect of the classification. For the testing dataset, the error between the predicted value and the corresponding actual one of each sample is depicted in the Figure 3. The result shows that most of the errors are less than 1%, and only

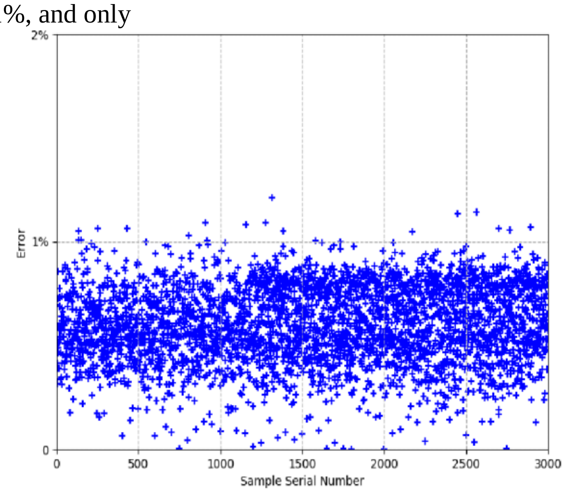


FIGURE 3. The error of each on-balance sheet item sample.

approximate 20 errors among 3000 is bigger than 1% but also less than 1.3%.

Among 1800 normal samples, 1765 samples are predicted as positive so the precision rate reaches 98.06%, and the number of wrongly predicted as normal samples is 13 among 1200 default samples.

B. EVALUATIONS ON DATA OF OFF-BALANCE SHEET ITEMS

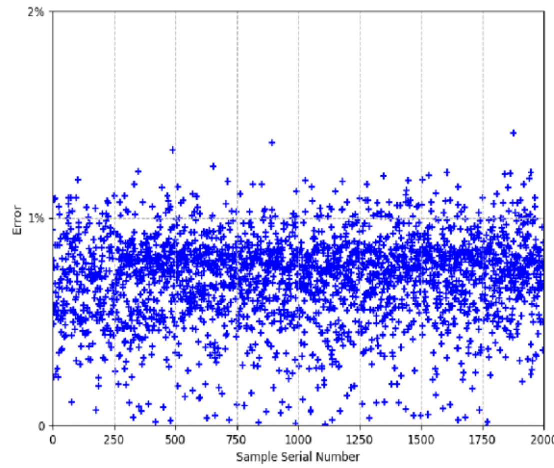
There are 4000 sample data of on-balance sheet items, which has nearly 20 dimensions in each sample. These off-balance sheet items sample data are also divided into two subsets: training dataset and testing dataset. There are 2000 samples in each subset respectively composing of 1200 normal samples and 800 default samples. After the training samples being applied into the PSO-BP neural network model for iterative computing, the network iteration performance reaches the least error 0.627×10^{-3} with 8 neurons in each hidden layer, which satisfies the preset limit 0.63×10^{-3}. The training result is showed in Table 2.

The training result of off-balance sheet item data.

Actual	Predicted		Total
	Positive	Negative	
Positive	1183(98.58%)	17(1.42%)	1200(100%)
Negative	11(1.37%)	789(98.6%)	800(100%)

It could be seen that the precision rate of the training result is 98.58% and the number of wrongly predicted as default samples is 17. For negative ones, the number of wrongly predicted as normal samples is 11 among 800 default samples. The result shows that the constructed PSO-BP neural network model off-balance sheet item is effective for classification.

In the testing stage, the 1200 normal samples and 800 default samples of testing dataset are used into the trained PSO-BP neural network model to verify the effect of the classification. For the testing dataset, the error between the predicted value and the corresponding actual one of each sample is depicted in the Figure 4. The result shows that most of the errors are less than 1%, and merely a small part among 2000 is bigger than 1% but also less than 1.5%.

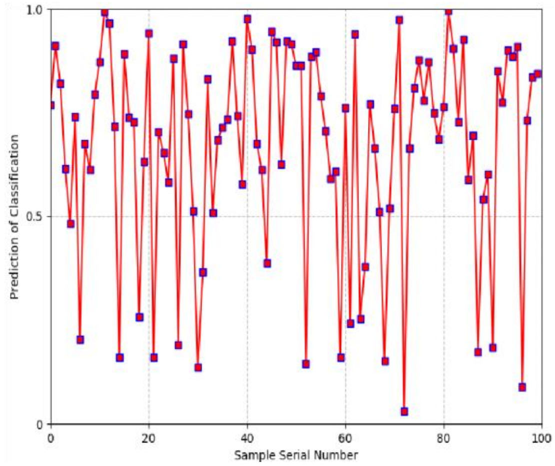


The error of each off-balance sheet item sample.

Among 1200 normal samples, 1177 samples are predicted as positive so the precision rate reaches 98.08%, and the number of wrongly predicted as normal samples is 16 among 800 default samples.

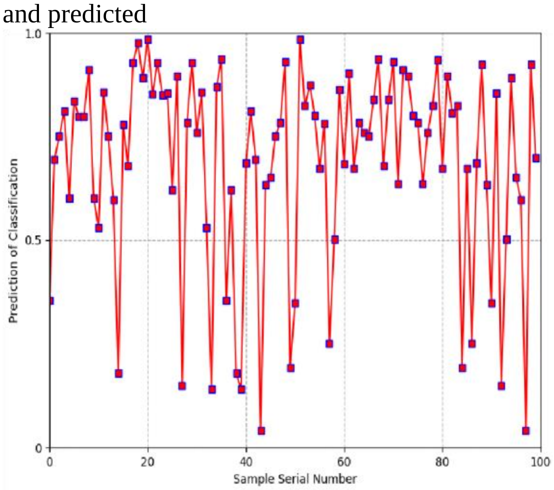
EVALUATIONS ON DATA OF APPLYING LOAN SAMPLES

In this section, 100 samples of applying bank loans are respectively evaluated in the two PSO-BP neural network models of on-balance sheet item and off-balance sheet item, so as to predict the classification and provide the references and suggestions for the risk management of bank loans. The evaluated results are depicted in the Figure 5 and Figure 6.



The prediction of on-balance sheet item model.

From the prediction results of classification, it could be seen that 19 samples are evaluated under 0.5 and predicted



The prediction of off-balance sheet item model.

as default samples by the PSO-BP neural network model of on-balance sheet item, and 16 samples are evaluated under 0.5 and predicted as default samples by the PSO-BP neural network model of off-balance sheet item. Among these predicted default samples, there are 15 same samples that means 15 companies applying bank loans are both classified as high risk ones by the two models, and totally there are 20 companies are evaluated as high risk ones. From the Chinese industries where these 20 companies belong, the proportion of the raw material production and processing industry and real estate industry are 75% with 7 coal mining companies, 3 nonferrous metal smelting and processing companies, 4 real estate companies and 1 metal mining company. Nowadays, China is further promoting the supply-side structural reforms and many companies, such as ones in the industries of the raw material production and processing and real estate, are in the difficult situation for the lack of strong demand, advanced technology, production capacity, etc. Therefore, the bank loans to these kinds of companies are probably with relatively higher risk.

In summary, the big data processing approach with parallel PSO-BP neural network models of on-balance sheet item and off-balance sheet item is effective and supportive for the risk management and approval of bank loans. Moreover, the traditional serial PSO-BP neural network models need several days to handle the dataset, while the parallel PSO-BP neural network models distributedly running on big data clusters could compute almost 90 times faster than the traditional serialonesforlargelyreducingtheprocessingtimeoftraining and testing.

CONCLUSION

With the processing volume of multi-source financial data for risk assessment increasing so dramatically, such as extra data from Internet, mobile network, IoT, etc., it has become necessary to design the big data mining solution for the classification and prediction of massive financial datasets. In this article, a nonlinear and parallel PSO-BP neural network approach is proposed and distributedly implemented on a Spark and HDFS cluster for mining the big dataset of on-balance sheet item and off-balance sheet item. The results of groups of experiments show that the proposed approach can discriminate the default sample and predict the financial risk with high accuracy and capacity. By using big data parallel framework, the running time of model training and testing is greatly reduced.

REFERENCES

I. U. Din, M. Guizani, S. Hassan, B.-S. Kim, M. K. Khan, M. Atiquzzaman, and S. H. Ahmed, ''The Internet of Things: A review of enabled technologies and future challenges,'' IEEE Access, vol. 7, pp. 7606–7640, 2018. doi: 10.1109/ACCESS.2018.2886601.

T. A. Ahanger and A. Aljumah, ''Internet of Things: A comprehensive study of security issues and defense mechanisms,'' IEEE Access, vol. 7, pp. 11020–11028, 2018. doi: 10.1109/ACCESS.2018.2876939.

T. Qiu, J. Liu, W. Si, and D. O. Wu, ''Robustness optimization scheme with multi-population co-evolution for scale-free wireless sensor networks,'' IEEE/ACM Trans. Netw., vol. 27, no. 3, pp. 1028–1042, Jun. 2019.

R. Wang, C. Yu, and J. Wang, ''Construction of supply chain financial risk management mode based on Internet of Things,'' IEEE Access, vol. 7, pp. 110323–110332, 2019. doi: 10.1109/ACCESS.2019.2932475.

C. Shepherd, F. A. P. Petitcolas, R. N. Akram, and K. Markantonakis, ''An exploratory analysis of the security risks of the Internet of Things in finance,'' in Proc. Int. Conf. Trust Privacy Digit. Bus. (ICTPDB), Jul. 2017, pp. 164–179.

M. Kim, W. Jo, J. Kim, and T. Shon, ''Visualization for Internet of Things: Power system and financial network cases,'' Multimedia Tools Appl., vol. 78, no. 3, pp. 3241–3265, Feb. 2019.

P. Centonze, ‘‘Security and privacy frameworks for access control big data systems,’’ Comput., Mater. Continua, vol. 59, no. 2, pp. 361–374, May 2019.

T. Qiu, B. Li, W. Qu, E. Ahmed, and X. Wang, ‘‘TOSG: A topology optimization scheme with global small world for industrial heterogeneous Internet of Things,’’ IEEE Trans. Ind. Informat., vol. 15, no. 6, pp. 3174–3184, Jun. 2019.

V. Dineshreddy and G. R. Gangadharan, ‘‘Towards an ‘Internet of Things’ framework for financial services sector,’’ in Proc. 3rd Int. Conf. Recent Adv. Inf. Technol. (RAIT), Mar. 2016, pp. 177–181.

W. H. Beaver, ‘‘Financial ratios as predictors of failure,’’ J. Accounting Res., vol. 4, no. 1, pp. 71–111, Mar. 1966.

E. I. Altman, ‘‘Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,’’ J. Finance, vol. 23, no. 4, pp. 589–609, Sep. 1968.

E. I. Altman and A. Saunders, ‘‘Credit risk measurement: Developments over the last 20 years,’’ J. Banking Finance, vol. 21, nos. 11–12, pp. 1721–1742, May 1997.

M. D. Odom and R. Sharda, ‘‘A neural network model for bankruptcy prediction,’’ in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jun. 1990, pp. 163–168.

E. I. Altman, G. Marco, and F. Varetto, ‘‘Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the italian experience),’’ J. Banking Finance, vol. 18, no. 3, pp. 505–529, May 1994.

R. Malhotra and D. K. Malhotra, ‘‘Differentiating between good credits and bad credits using neuro-fuzzy systems,’’ Eur. J. Oper. Res., vol. 136, no. 1, pp. 190–211, Jan. 2002.

Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, and S. Wu, ‘‘Credit rating analysis with support vector machines and neural networks: A market comparative study,’’ Decis. Support Syst., vol. 37, pp. 543–558, Sep. 2004.

D. West, S. Dellana, and J. Qian, ‘‘Neural network ensemble strategies for financial decision applications,’’ Comput. Oper. Res., vol. 32, no. 10, pp. 2543–2559, Oct. 2005.

H. Matoussi and A. K. Abdelmoula, ‘‘Credit-risk evaluation of a tunisian commercial bank: Logistic regression vs neural network modelling,’’ Int. J. Accounting Inf. Manage., vol. 19, no. 2, pp. 92–119, Jun. 2011.

A. Blanco, R. Pino-Mejías, J. Lara, and S. Rayo, ‘‘Credit scoring models for the microfinance industry using neural networks: Evidence from Peru,’’ Expert Syst. Appl., vol. 40, no. 1, pp. 356–364, Jan. 2013.

A. Byanjankar, M. Heikkilä, and J. Mezei, ‘‘Predicting credit risk in peerto-peer lending: A neural network approach,’’ in Proc. IEEE Symp. Series Comput. Intell. (SSCI), Dec. 2015, pp. 719–725.

F. Z. Azayite and S. Achchab, ‘‘Hybrid discriminant neural networks for bankruptcy prediction and risk scoring,’’ Proc. Comput. Sci., vol. 83, no. 2, pp. 670–674, 2016.

M. D. Murjadi and Z. Rustam, ‘‘Fuzzy support vector machines based on adaptive particle swarm optimization for credit risk analysis,’’ J. Phys., Conf. Ser., vol. 1108, no. 1, pp. 1–6, Dec. 2018.

A. Pérez-Martín, A. Pérez-Torregrosa, and M. Vaca, ‘‘Big data techniques to measure credit banking risk in home equity loans,’’ J. Bus. Res., vol. 89, no. 1, pp. 448–454, Aug. 2018.

M. Óskarsdottir, C. Bravo, C. Sarraute, J. Vanthienen, and B. Baesens, ‘‘The value of big data for credit scoring: Enhancing financial inclusion usingmobilephonedataandsocialnetworkanalytics,’’Appl.SoftComput., vol. 74, no. 1, pp. 26–39, Jan. 2019.

B. Wang, P. Liu, Z. Chao, W. Junmei, W. Chen, N. Cao, G. M. P. O’Hare, and F. Wen, ‘‘Research on hybrid model of garlic short-term price forecasting based on big data,’’ Comput., Mater. Continua, vol. 57, no. 2, pp. 283–296, Nov. 2018.

X. Zhang, Z. Li, G. Liu, J. Xu, T. Xie, and J. P. Nees, ‘‘A spark scheduling strategy for heterogeneous cluster,’’ Comput., Mater. Continua, vol. 55, no. 3, pp. 405–417, Jun. 2018.

T. Qiu, H. Wang, K. Li, H. Ning, A. K. Sangaiah, and B. Chen, ‘‘SIGMM: A novel machine learning algorithm for spammer identification in industrial mobile cloud computing,’’ IEEE Trans. Ind. Informat., vol. 15, no. 4, pp. 2349–2359, Apr. 2019.