Electronic Commerce Research (2020) 20:297–320
https://doi.org/10.1007/s10660-020-09409-0

Study on convolutional neural network and its application in data mining and sales forecasting for E-commerce

Hong Pan[1] · **Hanxun Zhou**[2]

Published online: 29 April 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

In recent years, the rapid development of e-commerce has brought great convenience to people. Compared with traditional business environment, e-commerce is more dynamic and complex, which brings many challenges. Data mining technology can help people better deal with these challenges. Traditional data mining technology cannot effectively use the massive data in the electricity supplier, it relies on the time-consuming and labour-consuming characteristic engineering, and the obtained model is not scalable. Convolutional neural network can effectively use a large amount of data, and can automatically extract effective features from the original data, with higher availability. In this paper, convolutional neural network is used to mine e-commerce data to achieve the prediction of commodity sales. First, this article combines the inherent nature of the relevant merchandise information with the original cargo log data that can be converted into a specific "data frame" format. Raw log data includes items sold over a long period of time, price, quantity view, browse, search, search, times collected, number of items added to cart, and many other metrics. Then, convolutional neural network is applied to extract effective features on the data frame. Finally, the final layer of the convolutional neural network uses these features to predict sales of goods. This method can automatically extract effective features from the original structured time series data by convolutional neural network, and further use these features to achieve sales forecast. The validity of the proposed algorithm is verified on the real e-commerce data set.

Keywords Convolutional neural network · E-commerce · Data mining · Sales forecasting

✉ Hanxun Zhou
lidachao223@sohu.com

[1] School of Economics, Liaoning University, Shenyang 110036, Liaoning Province, China

[2] School of Information, Liaoning University, Shenyang 110036, Liaoning Province, China

1 Introduction

E-commerce usually refers to a new business operation mode in which consumers conduct various business activities based on browser/server application on the open Internet platform [1]. On this platform, consumers can realize online shopping, transaction and online electronic payment without face-to-face contact with sellers. Since 2013, more and more e-commerce enterprises pay more attention to by using the Internet to provide users with quality service, and as the electronic commerce is more and more popular, more people choose in it, as a result, there is a large customer purchase behaviour data information, etc., it is more important and consumers' evaluation and feedback. Therefore, how to use the data information to analyse and mine the user behaviour law contained in it, to apply it to commodity sales forecasting has become one of the research hotspots [2].

With the popularity of e-commerce, more and more methods have been proposed and applied to commodity sales forecasting, such as logistic regression, decision tree, random forest, gradient ascending decision tree, neural network and so on. Logistic regression [3] is a generalized linear regression model, which uses logical functions to predict classification problems based on linear regression. Since logistic regression algorithm is good at explaining dichotomy problems and can better fit the functional relationship between independent variables and dependent variables, De Caigny et al. [4] used logistic regression algorithm to evaluate the stability level for telecom customers. Stripling et al. [5] used logistic regression algorithm to study customer classification problems in the telecommunications industry to prevent customer loss. The basic idea of decision tree [6] is that the classification rules of the representation form of decision tree can be deduced from a bunch of random and unordered instances according to some criteria in a top-down recursive way. Bell et al. [7] used decision tree to predict customers' purchase behaviour. Compared with other prediction methods, it can clearly and intuitively show logical classification. Sivasankar et al. [8] used decision trees to predict customers' shopping lists. Random forest [9] refers to the establishment of a forest with unrelated decision trees in a random way. After the forest is built, when a new sample is input, all the decision trees in the forest are asked to judge the category of the sample, and which category is the most selected, the predicted result will be the same. Mau et al. [10] used random forest to retain old customers in the insurance industry. Mahdavinejad et al. [11] applied the fusion model of perceptron vector machine, logistic regression and random forest to the project of predicting customers' repeated purchase behaviours on the e-commerce platform, and achieved good results. Like random forest, gradient boosting decision tree (GBDT) [12] is also a combinatorial model based on decision tree. Its idea is to build a decision tree each time in the direction where the loss function of the existing model decreases. Wang et al. [13] used GBDT to recommend personalized goods to users, and achieved good results in the feature project constructed. Convolutional neural network [14] is an intelligent information processing technology that mimics the information processing process of human brain. It has the characteristics of self-organization, self-adaptation, and

self-learning. Liberis et al. [15] used convolutional neural network algorithm to build a model to predict the possibility of repeated purchase by customers, and found that convolutional neural network algorithm has a strong learning ability for characteristic variables with complex non-linear relations. Pham et al. [16] represented the picture of each commodity as real number vectors through convolutional neural network, and generated commodity prediction information by solving nonlinear optimization problems based on these real number vectors. Qiu et al. [17] directly generated commodity prediction information by using the end-to-end model constructed based on convolutional neural network. Khaled et al. [18] obtained topic content and entity type through semantic analysis of user published content, and obtained more content related to the topic from external websites related to the topic for semantic enhancement to depict user behaviour, and finally built a model to make personalized recommendations to users. Kuzovkin et al. [19] proposed a new clustering factor to characterize user behaviour and predicted the popularity of newly launched products.

From the above analysis of relevant references, it can be seen that data mining technology relies on artificial feature engineering, which is not only time-consuming and laborious, but also requires that the personnel doing feature engineering have expertise in specific fields. This limits the scalability and availability of the model derived from traditional data mining techniques and makes it impossible to effectively utilize the large amount of available data in electricity suppliers. Because the convolutional neural network can automatically extract effective features from a large number of raw data, the model established by the convolutional neural network has stronger usability. The convolutional neural network model is more expressive and can effectively use a large amount of training data to learn more abundant pattern information. In order to solve the problems existing in the traditional data mining technology, this paper proposes an algorithm for predicting the sales volume based on the convolutional neural network. The algorithm can automatically extract effective features from the original structured data through the convolutional neural network, and further use the method to realize the forecast of commodity sales. Verification experiments on large data sets show that the proposed algorithm can effectively improve the accuracy of commodity sales forecasting.

2 E-commerce

2.1 Definition of e-commerce

E-commerce refers to business activities cantered on information exchange technology and commodity exchange can be understood, as the Internet, intranet and value-added online transactions in electronic transactions and related services activities, is the traditional business activities of all aspects of the electronic, network, information. At the same time, the Internet-based business practices belong to the scope of e-commerce.

If classified according to the payment situation of e-commerce, e-commerce can be divided into non-payment e-commerce and payment e-commerce [20].

Non-payment e-commerce refers to e-commerce that does not perform online payment and goods delivery. Its content includes information release, information inquiry, online negotiation, formation of contract text, etc., but does not include bank payments. In this kind of e-commerce, there is only the flow of material and information, and no flow of funds. Payment-type e-commerce refers to e-commerce for online payment and cargo delivery. In addition to the entire content of non-payment e-commerce, its content also includes bank payment, delivery activities, and goods delivery activities of suppliers. This includes both the flow of material and information and the flow of funds.

E-commerce is divided into business to business (B2B), business to consumer (B2C), business to government (B2G), and consumer to government (C2G) [21].

B2B refers to the traditional business-to-business transactions that tend to consume a lot of resources and time, whether it is sales and distribution or procurement costs of products. B2C refers to business-to-consumer transactions that are largely E retailing. B2G refers to the business transactions between enterprises and government agencies, including all transaction transactions between enterprises and government agencies. C2G refers to the personal affairs of government agencies mainly through the network to achieve the verification of personal identity, tax declaration, tax collection and other government transactions to individuals. Figure 1 is the flow chart of e-commerce commodity buying and selling.

E-commerce mainly refers to the first two types of e-commerce are B2B and B2C.

The e-commerce model refers to the many elements that make up e-commerce, various combinations and methods and methods of e-commerce operation and management. Different combinations of the components of e-commerce have different models. According to the different combination elements and their combined effects, the e-commerce model can be divided into e-commerce space

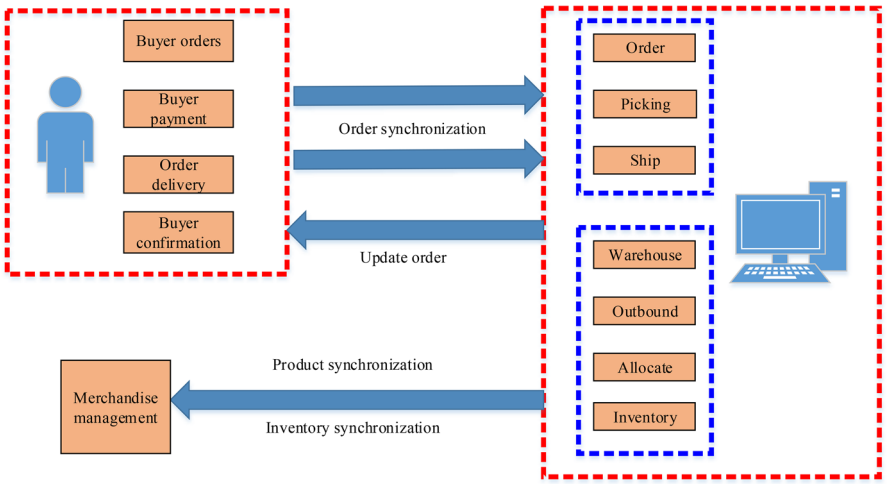

Fig. 1 Commodity buying and selling process in e-commerce

model, e-commerce scale model, e-commerce hierarchy model, e-commerce business model, and e-commerce operation management model.

With the continuous development of computer hardware and software technology, the complexity of society is getting higher and higher, and at the same time, the types of data will be more numerous. In this process, the development of data mining technology has important practical value. In the current data analysis, regression analysis, classification, association rule domain clustering, etc. are widely used methods in e-commerce mining at present, and these methods focus on different angles.

(1) Classification: This method is the most basic in e-commerce mining. It mainly classifies the massive data based on the characteristics of the data, divides the numerous data into several categories, and then uses a certain data analysis model for analysis. In the application of classification methods, the most important thing is to find a suitable data classification model. Using the classification model can complete the classification of data quickly and accurately. By analyzing the classified data, certain trends can be predicted, and accurate analysis of the data can effectively increase the sales of goods.
(2) Regression analysis: This method is to find the data attribute characteristics in a large amount of data, and use the function's mapping relationship to find the attribute association rules between the data. This method plays an important role in data series prediction and correlation.
(3) Clustering: This method is similar to classification, but the ultimate purpose of the two methods is different. Clustering divides the data more precisely. It divides the data into multiple categories according to the characteristics and similarity between the data. The similarity between the data in the same group is relatively large, but the data between different classes the correlations are very small.
(4) Association rules: This method can more accurately find the hidden associations between different data, and use association rules to push another data group from one data group. There are two main processes in the application of association rules. First, we must find the data group with the highest frequency among a large number of original data. Based on these high-frequency data, association rules between the data are generated. Data mining using association rules is widely used in the financial industry. It can accurately obtain the needs of customers, and according to the user's preferences, launch the business that customers are interested in on the human–computer interaction interface, so that they can formulate better marketing strategies to provide customers with high-quality services.
(5) Web data mining: This is a comprehensive technology. It can discover some hidden patterns through the WEB data collection. It can be regarded as a function mapping to the data mining process, and it is an analysis of data from input to output.
(6) Convolutional neural network method: This method is an advanced artificial intelligence technology, which can complete the processing and storage of data

by itself, and has high fault tolerance. It has a very strong advantage in dealing with incomplete and modulus data.

2.2 Analysis of consumer behavior in the E-commerce environment

Analysis of consumer behavior in the e-commerce environment can start from the following aspects [22]:

(1) Segmentation of the consumer market
In the traditional sales model, the target of market segmentation is a specific customer group, and to provide them with specific products and services, the pursuit of market share. In the mode of e-commerce, with the further development of technology, the traditional marketing activities of "market segmentation" are more differentiated. The market is divided into individual consumers, and the company is seeking customer share.
(2) Mainstreaming service demand
According to Maslow's hierarchy of needs theory, after low-level needs (such as physiological needs) are met, people often pursue higher-level needs. When ordering online, consumers not only need product features, but also more importantly, get emotional satisfaction, that is, the understanding and respect of the merchant. Corresponding to this, changes in marketing concepts, 4PS (product, price, channel, promotion) to 4CS (idea, convenience, cost, communication), developed to the current 4RS (association, reaction, relationship, return), fully reflects the consumer demand for services under the e-commerce model has become mainstream.
(3) Expansion of the scope of choice and perceptualization of consumer behavior
Under the traditional sales model, the consumer's choice range is to choose a limited number of goods in a limited space (a city). The principle of selection is "shopping around the goods". Consumers value the utility of the product (the product's Function, value, after-sales service, etc.), is a rational-oriented behavior: In the context of e-commerce, consumers are the products that 'choose to buy globally, and in the case of relatively transparent prices, it can be estimated.
(4) Direct participation of consumers in the production and circulation cycle
Under the traditional business model, the products and services that consumers choose are already produced by the company, and then reach consumers through various channels. In this model, consumers are just a passive receiving container. However, in the e-commerce model, the situation of consumption has changed, and they can directly participate in the production and circulation cycle of the enterprise according to their own needs. For example, Dell's direct sales model in the United States and IBM's user engagement design practices.

2.3 Data mining process in e-commerce

The most important part in the e-commerce management system is the database, in which information can be stored. In addition, the use of data mining can

analyse related data from a deeper level. Its main purpose is to find valuable and potentially useful information in massive data. These data are generally incomplete, fuzzy, accompanied by noise, and random. Because the data at the beginning is huge, to achieve the goal of data mining needs to be accomplished by computer technology. Generally, statistics, information retrieval, online analysis, artificial intelligence, fuzzy learning and machine learning are used to achieve it [23].

With the continuous development of computer hardware and software technology, the complexity of society is getting higher and higher, and at the same time, the data types will be more numerous. In this process, the development of data mining technology has important practical value. In the current data analysis, regression analysis, classification, association rule domain clustering, etc. are widely used methods in data mining at present, and these methods focus on different angles.

The data mining process in e-commerce can be subdivided into several steps such as determining business objects, data collection and extraction, data preprocessing, mining model construction, data mining, result analysis, and usage results [24], as shown in Fig. 2.

(1) Identify business issues. The results of data mining in everyday applications are often unpredictable, but the problems that need to be addressed are targeted and predictable. If you only blindly conduct data mining, it takes many labour and material resources, and the results are often unsatisfactory.
(2) Collection and selection of data. To clarify the specific objectives of data mining, it is necessary to collect all the data related to the business. The purpose of data selection is to extract information suitable for data mining from the collected data to improve the quality of mining.
(3) Data pre-processing. The main steps of data pre-processing include data integration, data cleansing, data specification, and data transformation. The purpose of data pre-processing is to prepare for subsequent data mining. It is the basis of building a model. Its quality directly affects the data mining effect. The most difficult thing is to reduce the deviation.

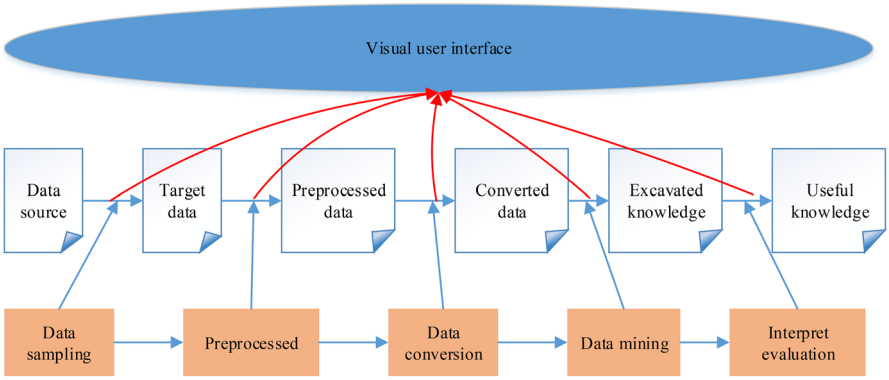

Fig. 2 Data-mining process in e-commerce

(4) Build a data mining model. Use data mining algorithms to build appropriate mining models based on the problems that need to be solved.
(5) Data mining. Use the model to mine processed data to obtain effective, potentially, and interpretable knowledge and information.
(6) Analysis and evaluation of results.

The traditional sales forecasting method is mainly based on time series analysis technology, which predicts future sales by analysing historical sales. The time series analysis mainly uses the linear combination of historical data points to predict future sales. According to the combination method, it can be divided into three categories: AR (Auto regressive), I (Integrated) and MA (Moving average). Combining these three categories can form a more general model ARMA (Autoregressive moving average) and ARIMA (Autoregressive integrated moving average) [25], etc., they can often achieve better prediction results. When using time series analysis to predict sales volume, because the model only uses historical sales data as input and uses a linear combination of historical sales to predict future sales, it is only applicable to those products with stable sales or obvious sales changes.

The business environment of e-commerce is more dynamic and complex, and the law of product sales is less obvious. The accuracy of timing analysis technology in this kind of scenario prediction is very limited. From another point of view, a large amount of data in e-commerce can be easily collected and utilized, such as number of visits (PV), number of viewers (UV) and price (PAY), etc., and these data can be effectively considered in the model to improve the accuracy of sales forecasts. However, they all rely on artificial feature engineering to extract relevant features from the data. The extraction of features is often a time-consuming, labour-intensive task and requires expertise in a particular field. Therefore, traditional data mining methods cannot automatically reorganize raw data.

Feature learning can automatically extract effective features from raw data, eliminating the dependence on artificial feature engineering, and convolutional neural network is one of the most commonly used feature learning methods. Guo et al. [26] proposed a prediction model based on the combination of mathematical model and GBRT. The algorithm can use mathematical models to model very complex functions. When using it for sales forecasting, first manually extract 523-dimensional features for each item, including the UV of the item for the past day, the average UV for the past three days, the average UV for the past week, the average UV for the past month, and whether it has recently been reduced. These features are then used as input to predict the next sales of the item using the GBRT model. Borkar et al. [27] proposed a predictive model for DNN. The algorithm consists of multiple layers of full connections. First, flatten the data frame to a vector. Then, the final feature representation vector is obtained by four-layer full join, where the dimension of each layer fully connected is set to 1024 dimensions. Finally, linear regression is applied to the representation vector to achieve sales forecast.

3 Sales forecasting algorithm based on convolutional neural network

Existing methods mainly focus on automatic extraction of effective features from unstructured data such as image, voice, and text [28]. In this paper, a novel method is proposed to automatically extract effective features from structured time series data by using convolutional neural networks. First, this paper converts the original log data related to commodities into a specific "data frame" format by combining the inherent attribute information of commodities. The original log data includes the sales volume, price, and number of visits, number of visitors, number of searches, number of searches, number of collectors, and other indicators of commodities in a long period in the past. Then, convolutional neural network is applied to extract effective features on the data frame. Finally, the final layer of the convolutional neural network uses these features to predict sales of goods. In addition, this paper also uses such techniques as sample weight attenuation and transfer learning to improve the accuracy of commodity sales forecasting. The method in this paper takes the original log data as input and the final sales forecast results as output, hardly requiring any manual intervention.

Next, this paper first introduces the basic knowledge of convolutional neural network. Then, the algorithm proposed in this paper is introduced. In addition, the network structure diagram is given.

3.1 The basics of convolutional neural networks

Convolutional neural network [29, 30] can express complex applications and scenarios through relatively simple and intuitive representation, and solve the problems such as the inability to extract high-level and abstract features from the original data in representation learning [31]. As shown in Fig. 3, the differences between convolutional neural network and traditional machine learning on the process are shown.

A typical convolutional neural network [32] consists of multiple convolutional layers, pooled layers, and fully connected layers. Usually, the convolutional layer and the pooled layer appear alternately in the front part of the network, while the latter part is the fully connected layer. Usually the convolutional layer is used to

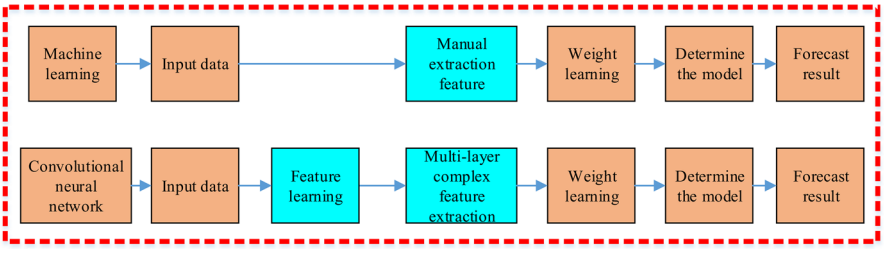

Fig. 3 Comparison of the flow of traditional machine learning and convolutional neural network

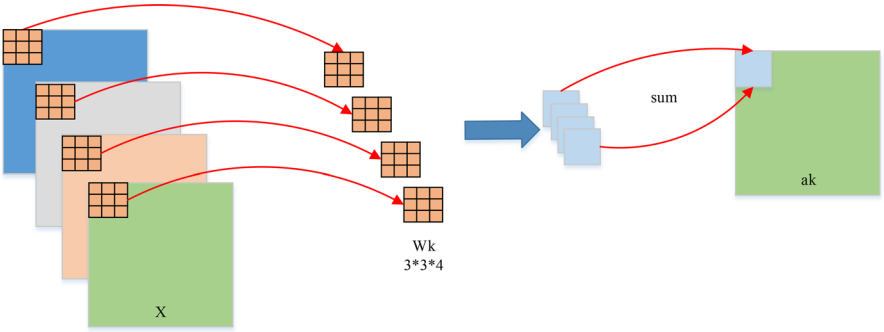

Fig. 4 Calculation process of convolution layer

detect local features, as shown in Fig. 4. For the feature map a_k corresponding to the same convolution kernel, the equation is as Eq. (1).

$$a_k = b_k + \sum_{i=0}^{C} W_k \otimes X_i \tag{1}$$

where b is the offset term corresponding to the convolution kernel. C is the number of channels of the previous feature map. The variable W_k represents the weight matrix corresponding to the kth convolution kernel. The symbol $\otimes$ indicates the convolution operation, and X is the activation value of the previous layer, which is the current layer input value. As can be seen from Fig. 4, the number of input feature map channels is 4, indicating that the thickness of the convolution kernel is also 4. The convolution kernel size is 3×3, and the number of parameters of the convolution kernel is 36. It should be noted that the number of convolution kernels is artificially set according to the task characteristics, and the weights are self-learning adjustments during the training process.

The a_k obtained by the convolutional layer also needs to be fed to a nonlinear function, the activation function. A neural network with a nonlinear processing unit can approximate an arbitrary function, so it helps to improve the expressiveness of the network. The most commonly used activation functions are sigmoid function, tanh function, rectified linear units (ReLU) [33], leaky ReLU [34], etc., as shown in Fig. 5. However, sigmoid and tanh also have their own flaws, the most obvious being saturation, which causes the biggest problem in deep convolution networks—the gradient disappears. As can be seen from Fig. 5, the derivatives on both sides gradually approach 0.

(1) Sigmoid function: It is the most widely used type of activation function. It has the shape of an exponential function. It is closest to biological neurons in the physical sense. In addition, the output of (0, 1) can also be expressed as a probability or used for normalization of the input.

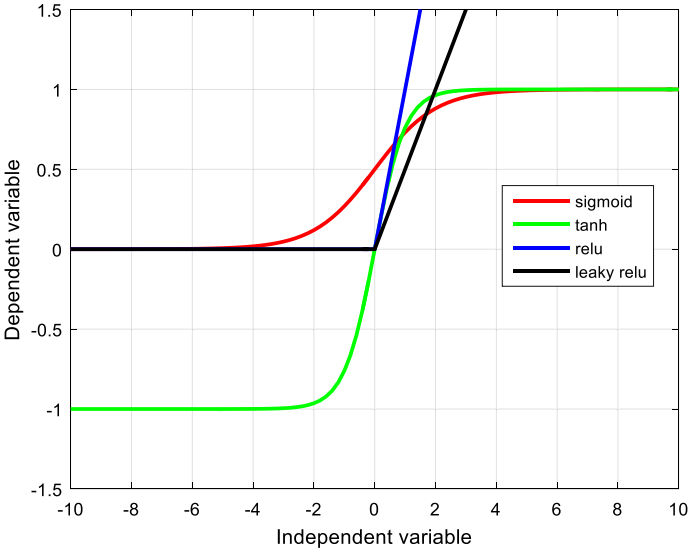

Fig. 5 Common activation functions

(2) Tanh function: The value range is (− 1, 1). Compared with sigmoid, its output mean is 0, making it faster than sigmoid, which can reduce the number of iterations.

(3) ReLU: Cut any negative value input to zero. Its role is to reduce the dependence between parameters, so that the network has a certain degree of sparseness, which is more in line with the sparseness of human brain neurons working in response to stimuli.

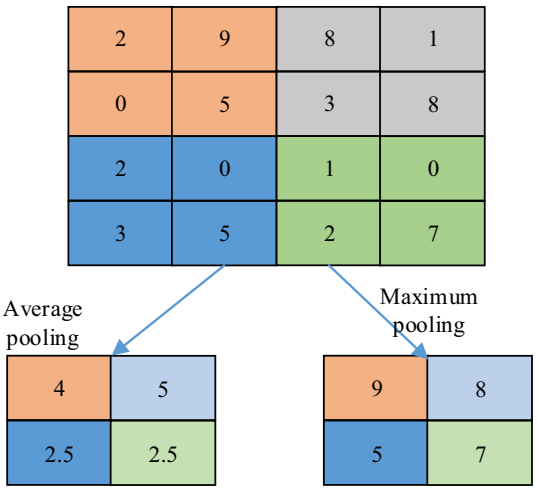

Fig. 6 Calculation process of pooling layer (The number represents the pixel value in the 4 by 4 area of the image. The left side shows the result after average pooling. On the right side is the result of maximum pooling)

(4) Leaky ReLU: It is an improved method for ReLU. It scales the negative value of the input to a certain proportion. If part of the ReLU input is negative, the weights cannot be updated.

The pooling operation actually replaces the area information by calculating the average value or the maximum value in the local area of the feature map, thereby implementing the function of down sampling. As shown in Fig. 6, the 4×4 feature graph is divided into four 2×2 small regions, and then the maximum value or draw value of each region is calculated to obtain a 2×2 feature graph. Even if the target object in the image has a small translation or scaling, the pooling operation can still obtain the same pooling features as before the change. Therefore, the pooled feature map information has certain rotation, translation, and telescopic invariance.

A fully connected layer is a tiled structure composed of a large number of neurons, the essence of which is matrix multiplication, usually placed behind the convolutional layer and the pooled layer. The convolutional neural network takes the loss layer as the terminal point to calculate the error between the actual network output and the target output and calculate the updated value of the error. The updated value of the error of each layer is calculated through the back propagation algorithm. The ownership weight is adjusted after the end of the back propagation. The purpose of training a convolutional neural network is to move the weight from the initial value to the optimal value. This process can be transformed into a global optimization problem that minimizes the loss function. Assuming a training set D, the average value of the overall loss function on data set D is shown in Eq. (2).

$$J(W,b) = \frac{1}{|D|}\sum_{i=0}^{|D|} S(X^{(i)}, Y^{(i)}) + \lambda r(W) \tag{2}$$

Among them, the symbol S is a loss function, which is usually defined according to specific tasks. The variable $(X^{(i)}, Y^{(i)})$ is the i-th sample in data set D, X is the data value, Y is the true value label, r(W) is the regular item, and λ is the weight of the regular item. Since the training set D is usually very large, if the error of all samples needs to be calculated for each iteration, the calculation amount is very large and the training process is slow. Therefore, in actual engineering, the random function of the objective function is generally used in each iteration. Approaching, that is, using small batches of data instead of datasets. Therefore, the function is converted as shown in Eq. (3).

$$J(W,b) = \frac{1}{N}\sum_{i=0}^{N} S(X^{(i)}, Y^{(i)}) + \lambda r(W) \tag{3}$$

N is the batch size of the input samples per iteration.

Let us say I have a neural network. The neural network has three input units, three hidden units and one output unit. The forward propagation process can be expressed by formula (4).

$$\begin{cases} z^{(2)} = W^{(1)}x + b^{(1)} \\ a^{(2)} = f(z^{(2)}) \\ z^{(3)} = W^{(2)}a^{(2)} + b^{(2)} \\ a^{(3)} = f(z^{(3)}) \end{cases} \tag{4}$$

Among them, z^l represents the intermediate result of layer l, a^l represents the activation value or output value of layer l, W is the weight matrix, b is the bias term, and f is the activation function. In formula (4), the activation value of the input layer can be expressed by $a^1 = x$. Then, the definition for each layer of the network becomes as shown in formula (5).

$$\begin{cases} z^{l+1} = W^l a^i + b^l \\ a^{l+1} = f(z^{l+1}) \end{cases} \tag{5}$$

Given a single sample (x, y), let the L-th layer be the last output layer, and set $S = \frac{1}{2}||a^{(L)} - y||_2^2$ in Eq. (3). That is, one-half the variance cost function, the residual calculation formula is shown in formula (6).

$$\delta^L = \frac{\partial J(W, b, x, y)}{\partial z^L} = -\frac{1}{N}\sum_{i=0}^{N}\left(y^{(i)} - a^{L(i)}\right)f'\left(z^L\right) \tag{6}$$

Then, for the hidden layer $k(L-1, L-2, \ldots, 2)$, the residual is shown in formula (7).

$$\delta^k = \frac{\partial J(W, b, x, y)}{\partial z^k} = \frac{\partial J(W, b, x, y)}{\partial z^{(k+1)}}\frac{\partial z^{(k+1)}}{\partial a^{(k)}}\frac{\partial a^{(k)}}{\partial z^{(k)}} = \left((W^{(k)})^T\delta^{(k+1)}\right)f'(z^{(k)})\big) \tag{7}$$

The weighted partial derivative of the corresponding layer is calculated as shown in formula (8) and formula (9).

$$\frac{\partial J(W, b)}{\partial W^k} = \frac{\partial J(W, b)}{\partial z^{(k+1)}}\frac{\partial z^{(k+1)}}{\partial W^k} = \delta^{(k+1)}(a^{(k)})^T \tag{8}$$

$$\frac{\partial J(W, b)}{\partial b^k} = \frac{\partial J(W, b)}{\partial z^{(k+1)}}\frac{\partial z^{(k+1)}}{\partial b^k} = \delta^{(k+1)} \tag{9}$$

Among them, formula (8) only calculates the partial derivative of the first formula of formula (3) on the weight.

Finally, according to the above formula, it is possible to calculate the weight update value of the weight layer through back propagation, and update these weights together after the back propagation is completed.

3.2 Network improvement

The algorithm proposed in this paper takes the original log data as input and the final sales forecast result as output, hardly requiring any manual intervention. Innovations are as follows:

(1) Combined with the inherent attribute information of the item, the raw log data-related item is converted into a specific "data frame" format.
(2) Apply convolutional neural network to extract effective features on the data frame. Finally, these characteristics are used in the final layer of the convolutional neural network to predict the sales volume of goods.
(3) Using techniques such as sample weight attenuation and migration learning to improve the accuracy of commodity sales forecast.

Next, elaborate on these three aspects.

(1) Convert the timing data into a data frame

The deep learning network in literature [28], literature [29] and literature [30] takes text data directly as input without processing. However, there is a lot of noise and useless text in this unfiltered data. As a result, for a given commodity i in a specific region r, this paper hopes to use the log data x_{ir} associated with it for a period [1, T] to predict the overall sales volume within y_{ir} with the next time [$T+1$, $T+l$] in the region r. In this document, the commodity vector of the commodity i at a specific time point t within a specific region r represented by x_{ir}. This vector contains the d-dimensional information of item i, such as sales volume, number of views (PV), number of searches (SPV), number of viewers (UV), number of searchers (SUV), total amount of transactions (GMV), and price (PAY). Then, x_{ir} at a plurality of time, points are combined to form a commodity matrix $x_{ir} = [x_{ir_1}, x_{ir_2}, \ldots, x_{ir_T}]$. In addition, the vector a_i is used to represent the intrinsic attribute set of the item i, including categories, brands, suppliers, and the like.

The goal of this paper is to construct a mapping function $f(\bullet)$, with x_{ir} and a_i as input to predict.

$$y_{ir} = f(x_{ir}, a_i, \theta) \tag{10}$$

Among them, the variable θ is a parameter vector that needs to be optimized during training.

For each commodity i, this paper needs to construct the log data related to it into a data frame according to its inherent attribute information in the manner shown in Fig. 7.

First, at the time t in the region r, for each brand b, category c and supplier s, the brand vector x_{br_t}, the category vector x_{cr_t} and the supplier vector x_{sr_t} are respectively calculated. The calculation equation is as shown in Eq. (11), Eq. (12), and Eq. (13).

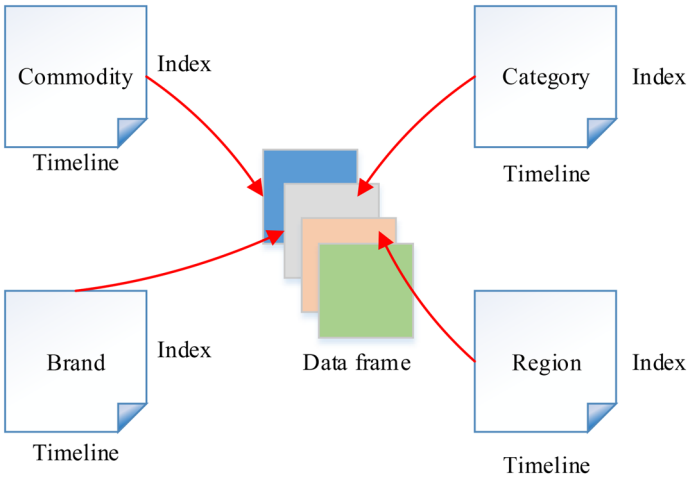

Fig. 7 Constructing a data box

$$x_{br_t} = \sum_{brand(i)=b} x_{ir_t} \tag{11}$$

$$x_{cr_t} = \sum_{category(i)=c} x_{ir_t} \tag{12}$$

$$x_{sr_t} = \sum_{spplier(i)=s} x_{ir_t} \tag{13}$$

The above vectors are then combined into a matrix $X_{br} = [x_{br_1}, x_{br_2}, \ldots, x_{br_T}]$, a category matrix $X_{cr} = [x_{cr_1}, x_{cr_2}, \ldots, x_{cr_T}]$, and a vendor matrix $X_{sr} = [x_{sr_1}, x_{sr_2}, \ldots, x_{sr_T}]$, respectively.

Secondly, for each region r, this paper calculates the region vector x_{rt} at time t.

$$x_{rt} = \sum_{i=0}^{T} x_{rt_t} \tag{14}$$

Then, the region vectors are combined into an area matrix $X_r = [x_{r_1}, x_{r_2}, \ldots x_{r_T}]$.

Finally, for each item i for the region r, the data frame DF_{ir} is constructed as shown in Eq. (15).

$$DF_{ir} = [X_{ir}, X_{brand(i)r}, X_{category(i)r}, X_r] \tag{15}$$

(2) Feature extraction network based on data frame

This paper predicts the sales of goods through the function $f(\bullet)$, which is a convolutional neural network, and the structure is shown in Fig. 8. Different from the

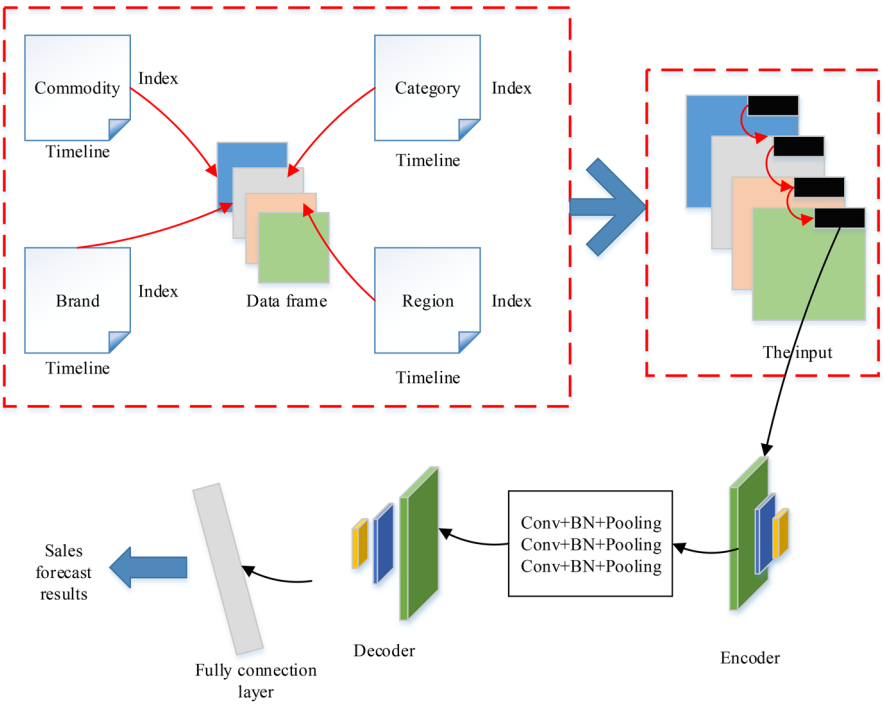

Fig. 8 Sales forecast with convolutional neural networks

traditional CNN, the network designed in this paper is a way of encoding and decoding. In the encoding process of convolution, the input information bits are block coded, and the encoded output bits of each code group are related not only to the information bits of the packet, but also to the information bits of other packets at the previous moment. Similarly, in the decoding process of convolution, decoding information is not only obtained from the packets received at the current time, but also extracted from the related packets. It is because the correlation of each group is fully utilized in the encoding process of convolution that the data has a good performance gain.

Firstly, the original data is processed and the corresponding data box is obtained. After that, the convolutional layer is adopted for coding. After the encoding of the data box is obtained, three convolution operations, regularization operations and maximization pooling operations are adopted. After the above operations are applied three times, the highest order representation of the original log data is obtained and the data is decoded. Finally, all the highest-order expressions are aggregated into the final feature representation vector $\hat{x}_{ir}$ by the full join operation. After the feature representation vector is obtained, it is used as an input to the linear regression to predict the sales volume y_{ir} of the commodity i on the region r, as shown in Eq. (16).

$$y_{ir} = [1, \hat{x}^T]w \tag{16}$$

Among them, the vector w is a parameter that needs to be optimized during the model training process.

(3) Sample weight attenuation and migration learning

This article builds a model for each region. For each region r, a model is trained to minimize the mean square error MSE (Mean Squared Error) on the training set D_r, as shown in Eq. (17).

$$L_r = \frac{1}{|D_r|} \sum_{ir \in D_r} (y_{ir} - \hat{y}_{ir})^2 \tag{17}$$

Among them, the variable y_{ir} is the real total sales volume of the commodity i at time $[T+1, T+l]$ in the region r, and $\hat{y}_{ir} = f(x_{ir}, a_i, \theta)$ is the corresponding predicted sales volume. The parameter that needs to be optimized in the whole model is the variable θ in the Eq. (10), as shown in the Eq. (18).

$$\theta = \{F, B, H, w\} \tag{18}$$

They are filter group F, offset group B, transformation matrix H, and linear regression parameter W, respectively.

In this paper, many training samples can be constructed by sliding the time window of data, but different training samples have different degrees of importance: the closer the prediction interval is, the more weight should be. Let sp be the starting point of the prediction interval, variable l be the length of the prediction interval, and ep_{ir} is the termination point of the data frame DF_{ir}, obviously $ep_{ir} \leq sp - l$ is established. For each region r, this paper assigns the following weights to each sample in the training set D_r.

$$\text{Weight}_{ir} = e^{\beta(ep_{ir} - sp + l)} \tag{19}$$

Among them, the variable β is the hyperactive parameter of the model.

Then for each region r, this paper minimizes the weighted mean square error of the model on the training set D_r:

$$L_r^w = \frac{1}{|D_r|} \sum_{ir \in D_r} \text{weight}_{ir} (y_{ir} - \hat{y}_{ir})^2 \tag{20}$$

Migration learning aims to move the knowledge learned from the model in one problem to another. This paper hopes to migrate the patterns of change learned by a model in one region to another. Although the pattern of change in product sales varies from region to region, the patterns of change in certain indicators related to sales have commonalities in different regions. For example, although the northern part of China sells cotton coats earlier than the southern part, the number of searches for cotton coats will increase significantly before the sales of cotton coats increase. Based on this, the predictive model by this paper can first learn the common pattern features with all the data, and then use the data on the specific region to learn the special pattern features.

First, this paper trains a neural network model on the entire data set D, where:

$$D = \bigcap_r D_r \tag{21}$$

Then, for each region r, the training set is replaced with D_r and then the training is continued, and a specific model suitable for the region r is obtained.

In this paper, the stochastic gradient descent algorithm is used to optimize the model, and the parameters are updated by backward conduction. The update method uses the Adam rule. Each time the model is trained, 128 samples are read. After reading all the samples, it is recorded as 1 cycle. First, 10 cycles are trained on the total data set D, and then the different models are respectively trained on the data set D_r for 10 cycles. The corresponding model is obtained for each region r.

4 Results and discussion

4.1 The experimental data

The experimental data set, provided by Alibaba group, contains 1814,892 records. We selected records for five of these areas. The selected records span from 2017-10-10 to 2018-12-27.In addition, each log data records 25-dimensional indicators, including sales volume, number of views, number of searches, number of views, number of searches, total transaction amount and price, etc.

For each region r, the goal of this paper is to use the data frame spanning [2018-10-28, 2018-12-20] to predict the sales volume of each product in the region in the period [2018-12-21, 2018-12-27], that is, these data samples constitute the test set. The end span of the sample data box used in the training model was [2018-01-01, 2018-12-13]. After more training data samples were obtained through the sliding time window, these data samples were randomly divided into two parts, training set and verification set, with a ratio of 4:1.

4.2 Setting of hyper parameters

In order to further improve the performance of this algorithm, the optimal super parameter is determined through experiments.

The mean square error (MSE) is the expected value of the squared difference between the estimated value of the parameter and the true value of the parameter. MSE can evaluate the degree of data change, and the smaller the value of MSE, the better accuracy the prediction model has in describing the experimental data. The calculation formula is as follows:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{observed}_t - \mathrm{predicted}_t)^2\theta \tag{22}$$

Figure 9 shows the change of mean sales mean square deviation when the predicted interval length changes in the data set of five regions. It can be seen that when

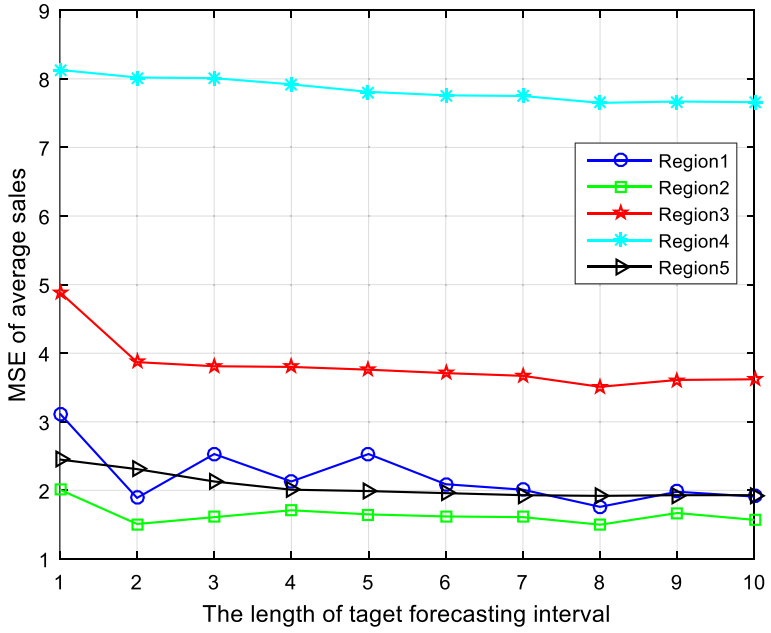

Fig. 9 The change of mean variance of average sales volume when the prediction interval length changes

the prediction length is one at the beginning, the mean variance of sales data in the five regions is relatively large, indicating that the prediction ability of the algorithm is poor. With the increase of the length of the forecast interval, it can be seen that the mean square deviation of the sales data in the five regions changed little and gradually stabilized. When the length of the forecast interval is eight, MSE reaches the minimum value, indicating that the average sales volume over this forecast interval is more stable. When the predicted interval length is greater than 8, MSE grows slowly. Therefore, in order to improve the prediction performance of the algorithm, the length of the prediction interval is chosen to be eight.

The length of the data frame used by the model is one of the important hyperactive parameters in the model, which determines how much historical data the model uses to predict future sales. Although the model is relatively robust to this parameter, if the data box is too short, the information contained in it is insufficient and the prediction effect is poor. If the data box is too long, it contains useless information, and the prediction will be poor. In addition, longer data frames require more computing resources. In conclusion, it is necessary to select the shortest data box in practical application on the premise of ensuring accuracy. As can be seen from Fig. 10, with the change of the length of the data box, the MSE score first decreases and then increases. The MSE score is related to the stability of the algorithm. The lower the MSE score, the better the stability of the algorithm. When the data box length was 40, the MSE of the region 1 was 121, the MSE of the region 2 was 86, the MSE of the region 3 was 139, the MSE of the region 4 was 311, and the MSE of the region 5 was 68. While region 3 and region 5 did not have the lowest MSE,

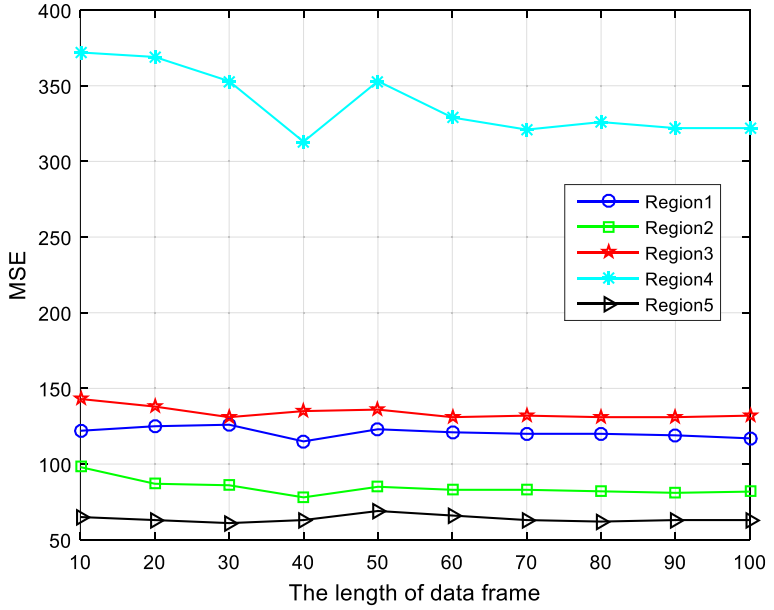

Fig. 10 Change of MSE score as the length of data frame changes

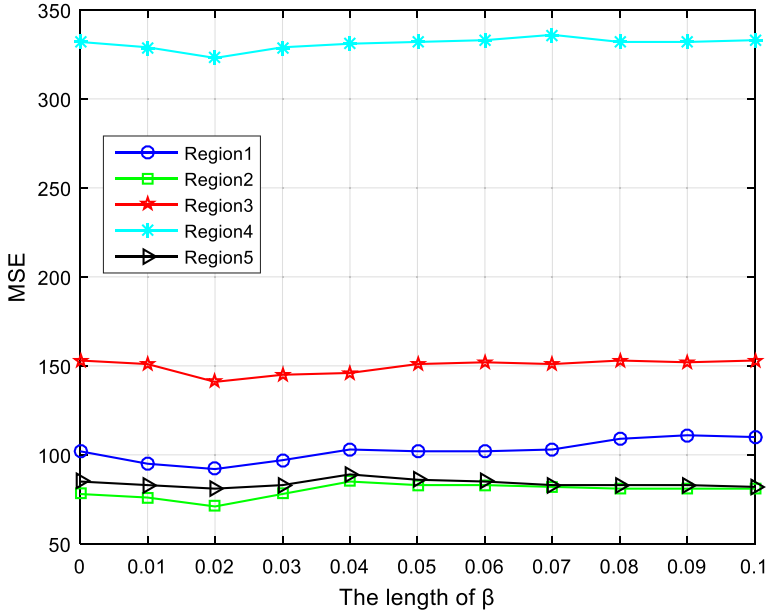

Fig. 11 Change of MSE when parameter β changes in weight attenuation

the other three regions had the lowest MSE. In order to realize the robustness of the algorithm, this article chooses the length of the data box to be 40.

In this paper, weight is assigned to the training samples through Eq. (20). The point closer to the prediction interval has more weight. In the formula, the parameter is used to adjust the decay rate of sample weight. If the value of values of values is large, the model is more inclined to the pattern reflected by the recent samples, which is not representative. When the value of values is small, the model considers all samples more evenly, and the algorithm at this time has good representativeness. As can be seen from Fig. 11, the value of drunk days does not increase, but MSE decreases first and then increases slowly. When the value of values of values is 0.02, the MSE value obtained from the dataset of five locales is the smallest. It shows that the deviation of the data is small and the stability of the algorithm can be better demonstrated. Therefore, the value of values selected for this article is 0.02. In this way, the algorithm can consider the existence of all samples, so that the results obtained by the algorithm are more representative.

4.3 This paper compares the algorithm with other algorithms

In order to further verify the effectiveness of the proposed algorithm, the proposed algorithm is compared with the following algorithm.

(1) CNN. The application of convolution neural network on data frame is the basic version of the method in this paper.
(2) CNN + WD. After assigning weights to training samples according to Eq. (13), the training model minimizes the weighted mean variance, which can improve the accuracy of sales forecasting.
(3) Single-CNN. Learn the generic model on the data set D containing all training samples, and use the generic model directly to predict the sales volume of goods in the region for each region r. Heavy attenuation and transfer learning techniques are not used here.
(4) ARIMA algorithm proposed in literature [25]. ARIMA is a classical time series analysis. In the sales forecast of commodities, it takes the historical sales data of commodities as the input to predict the subsequent sales of commodities.

Table 1 MSE scores of different algorithms on data sets tested in five regions

Methods	Region 1	Region 2	Region 3	Region 4	Region 5	Mean
ARIMA [25]	103.12	97.23	189.12	398.15	86.35	174.794
MGBRT [26]	98.78	83.78	188.09	321.77	83.81	140.246
PDNN [27]	98.93	73.22	182.88	346.21	85.15	157.278
CNN	96.98	72.22	152.89	327.18	81.09	146.072
CNN + WD	89.01	57.98	143.27	302.87	76.98	134.022
Single-CNN	102.87	78.12	161.01	342.78	86.77	154.31
This paper	83.31	53.65	131.92	289.34	71.19	125.882

(5) Algorithm proposed in literature [26]. The algorithm can model very complex functions. When using it for sales forecasting, it first manually extracts 523 dimensional features for each item. These characteristics are then used as input models to predict subsequent sales of the item.
(6) DNN algorithm proposed in literature [27]. DNN is the simplest neural network structure, which consists of multiple layers of full connections. First, flatten the data box to a vector. Then, the final feature representation vector is obtained through four-layer full connection. Finally, linear regression is applied on the representation vector to achieve sales forecast.

Table 1 shows experimental results on test data sets in five regions. Compared with the time series data analysis ARIMA proposed in literature [25], the algorithm proposed in literature [26] can take into account more information and achieve better prediction effect. Although DNN proposed in literature [27] is the simplest neural network architecture, it can automatically extract features. In some cases, features extracted by DNN are more effective than those extracted manually are, so the prediction effect of DNN in regions numbered Region 2, Region 3 and Region 5 proposed in literature [27] is better than the algorithm proposed in literature [26]. Convolutional neural network can make better use of the prior information of locality of time dimension in data, to extract features more effectively, and the prediction effect is greatly improved. It can be seen that the performance of CNN applied only on the data frame is better than that of literature [27]. After assigning weights to training samples according to Eq. (13), the weighted mean variance can be reduced. As can be seen from Table 1, the accuracy of sales forecast can be significantly improved. Single-CNN algorithm does not use heavy attenuation technology and transfer-learning technology, its effect is not good. In this paper, weight attenuation technology and transfer learning technology are adopted. The improvement brought by sample weight attenuation technology and transfer learning technology is very considerable, and the prediction effect after integrating all technologies is very competitive. Therefore, the algorithm in this paper can significantly improve the accuracy of sales forecasting.

5 Conclusion

In this paper, convolutional neural network is used to automatically extract effective features from structured timing data, which can effectively avoid artificial feature engineering, which is time-consuming, labor consuming, and requires the field knowledge of the personnel carrying out the feature engineering. This paper uses this method to predict the sales volume of commodities, taking commodity attribute information and relevant original log data as input and the total sales volume of commodities in a period of time in the future as output, hardly requiring any manual intervention. Firstly, this paper converts the original log data related to the commodity into a specific “data box” format by combining the attribute information of the commodity. Then, the convolution neural network is applied to the data box to extract effective features. Finally, the final layer of the convolutional neural

network uses the linear regression with these characteristics as input to predict the sales volume of goods. In addition, this paper also uses such techniques as sample weight attenuation and transfer learning to improve the accuracy of commodity sales forecasting. Verification experiments on large-scale data sets show that the proposed algorithm can effectively improve the accuracy of sales forecasting. Our next job is to continue to collect a large number of samples for training and testing. Then, we optimize and improve the network model based on a large number of samples.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Lin, J., Luo, Z., Cheng, X., et al. (2019). Understanding the interplay of social commerce affordances and swift guanxi: An empirical study. *Information & Management, 56*(2), 213–224.
2. Ramos, A. L., Mazzinghy, D. B., Barbosa, V. S. B., et al. (2019). Evaluation of an iron ore price forecast using a geometric Brownian motion mode. *REM-International Engineering Journal, 72*(1), 9–15.
3. Yao, Y., & Wang, H. Y. (2019). Optimal subsampling for softmax regression. *Statistical Papers, 60*(2), 235–249.
4. Vijaya, J., & Sivasankar, E. (2019). An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing. *Cluster Computing, 22*(5), 10757–10768.
5. Stripling, E., vanden Broucke, S., Antonio, K., et al. (2018). Profit maximizing logistic model for customer churn prediction using genetic algorithms. *Swarm and Evolutionary Computation, 40,* 116–130.
6. Kannadath, B. S., Cen, P., Rowe, J., et al. (2018). Decision tree analysis of pancreatic cyst fluid data for the detection of mucinous cysts: 73. *American Journal of Gastroenterology, 113,* S41–S42.
7. Bell, D., & Mgbemena, C. (2018). Data-driven agent-based exploration of customer behavior. *Simulation, 94*(3), 195–212.
8. Sivasankar, E., & Vijaya, J. (2019). A study of feature selection techniques for predicting customer retention in telecommunication sector. *International Journal of Business Information Systems, 31*(1), 1–26.
9. Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228–1242.
10. Mau, S., Pletikosa, I., & Wagner, J. (2018). Forecasting the next likely purchase events of insurance customers: A case study on the value of data-rich multichannel environment. *International Journal of Bank Marketing, 36*(6), 1125–1144.
11. Mahdavinejad, M. S., Rezvan, M., Barekatain, M., et al. (2018). Machine learning for internet of things data analysis: A survey. *Digital Communications and Networks, 4*(3), 161–175.
12. Rao, H., Shi, X., Rodrigue, A. K., et al. (2019). Feature selection based on artificial bee colony and gradient boosting decision tree. *Applied Soft Computing, 74,* 634–642.
13. Wang, J., Lin, L., Zhang, H., et al. (2017). A novel confidence estimation method for heterogeneous implicit feedback. *Frontiers of Information Technology & Electronic Engineering, 18*(11), 1817–1827.
14. Hanson, J., Paliwal, K., Litfin, T., et al. (2018). Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics, 34*(23), 4039–4045.
15. Liberis, E., Veličković, P., Sormanni, P., et al. (2018). Parapred: antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics, 34*(17), 2944–2950.

16. Pham, D. H., & Le, A. C. (2018). Learning multiple layers of knowledge representation for aspect based sentiment analysis. *Data & Knowledge Engineering, 114,* 26–39.
17. Qiu, X., Suganthan, P. N., & Amaratunga, G. A. J. (2019). Fusion of multiple indicators with ensemble incremental learning techniques for stock price forecasting. *Journal of Banking and Financial Technology, 3*(1), 33–42.
18. Khaled, A., Ouchani, S., & Chohra, C. (2019). Recommendations-based on semantic analysis of social networks in learning environments. *Computers in Human Behavior, 101,* 435–449.
19. Kuzovkin, D., Pouli, T., Meur, O. L., et al. (2019). Context in photo albums: Understanding and modeling user behavior in clustering and selection. *ACM Transactions on Applied Perception (TAP), 16*(2), 1–20.
20. Tong, Wu, Liu, Xinwang, & Qin, Jindong. (2017). A linguistic solution for double large-scale group decision-making in E-commerce. *Computers & Industrial Engineering, 116,* 97–112.
21. Xu, S. X., & Huang, G. Q. (2017). Efficient multi-attribute multi-unit auctions for B2B E-commerce logistics. *Production & Operations Management, 26*(2), 292–304.
22. Wang, Dong, Zha, Yong, & Bi, Gongbing. (2018). A meta-analysis of satisfaction-loyalty relationship in e-commerce: sample and measurement characteristics as moderators. *Wireless Personal Communications, 103*(1), 941–962.
23. Zhu, L., Li, M., Zhang, Z., et al. (2018). Big data mining of users' energy consumption patterns in the wireless smart grid. *IEEE Wireless Communications, 25*(1), 84–89.
24. Wu, P. J., & Lin, K. C. (2018). Unstructured big data analytics for retrieving e-commerce logistics knowledge. *Telematics and Informatics, 35*(1), 237–244.
25. Ortega, J. A., Losada, E., Besteiro, R., et al. (2018). Validation of an autoregressive integrated moving average model for the prediction of animal zone temperature in a weaned piglet building. *Biosystems Engineering, 174,* 231–238.
26. Guo, Z., Zhao, X., Chen, Y., et al. (2019). Short-term passenger flow forecast of urban rail transit based on GPR and KRR. *IET Intelligent Transport Systems, 13*(9), 1374–1382.
27. Borkar, T. S., & Karam, L. J. (2019). DeepCorrect: Correcting DNN models against image distortions. *IEEE Transactions on Image Processing, 28*(12), 6022–6034.
28. Rout, J. K., Choo, K. K. R., Dash, A. K., et al. (2018). A model for sentiment and emotion analysis of unstructured social media text. *Electronic Commerce Research, 18*(1), 181–199.
29. Gysel, P., Pimentel, J., Motamedi, M., et al. (2018). Ristretto: A framework for empirical study of resource-efficient inference in convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 29*(11), 5784–5789.
30. Wan, S., Liang, Y., Zhang, Y., et al. (2018). Deep multi-layer perceptron classifier for behavior analysis to estimate parkinson's disease severity using smartphones. *IEEE Access, 6,* 36825–36833.
31. Manogaran, G., & Lopez, D. (2018). Health data analytics using scalable logistic regression with stochastic gradient descent. *International Journal of Advanced Intelligence Paradigms, 10*(1–2), 118–132.
32. Jiang, X., Pang, Y., Li, X., et al. (2018). Deep neural networks with elastic rectified linear units for object recognition. *Neurocomputing, 275,* 1132–1139.
33. Yadav, S., & Bist, A. S. (2018). Learning overcomplete representations using leaky linear decoders. *International Journal of Digital Information and Wireless Communications, 8*(3), 174–180.
34. Seinen, C., & Khouider, B. (2018). Improving the Jacobian free Newton–Krylov method for the viscous–plastic sea ice momentum equation. *Physica D: Nonlinear Phenomena, 376,* 78–93.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.