Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Optimization of a Comprehensive Sequence Forecasting Framework Based on DAE-LSTM Algorithm

To cite this article: Yonghe Zhao *et al* 2021 *J. Phys.: Conf. Ser.* **1746** 012087

View the article online for updates and enhancements.

You may also like

- Elemental isomerization processes for a photochromic diarylethene film based on carrier injection toward all-electrically operable organic memory
Tsuyoshi Tsujioka and Kazuki Yamamoto
- Likelihood-free Cosmological Constraints with Artificial Neural Networks: An Application on Hubble Parameters and SNe Ia
Yu-Chen Wang, Yuan-Bo Xie, Tong-Jie Zhang et al.
- Metal pattern resolution for fine electrode formation using selective metal-vapor deposition using photochromic diarylethene
Tsuyoshi Tsujioka and Shoko Ishida

This content was downloaded from IP address 182.75.112.226 on 14/02/2024 at 11:38

Optimization of a Comprehensive Sequence Forecasting Framework Based on DAE-LSTM Algorithm

Yonghe Zhao[1,2,*], Xiaochen Ren[2] and Xingyi Zhang[3]

[1]Data Management Center, Netease Research and Development Centre, Beijing, China
[2]Mathmatical Modeling Center, North China Electric Power University, Baoding, China
[3]School of Statistics, Dongbei University of Finance and Economics, Dalian, China

*Corresponding author email: zhaoyonghe@corp.netease.com

Abstract. Sales forecasting is a multivariate time series forecasting problem. The main challenges of the forecasting task are the high-dimensional influence variables with noise and the complex time series relationships. In this paper, we propose a DAE-LSTM algorithm combined with denoising autoencoder(DAE) and long short-term memory(LSTM) to deal with this problem. The DAE is a non-linear dimensionality reduction method that enhances data robustness. The LSTM is a deep learning algorithm suitable for dealing with multivariate time series problem. Combining the advantages of the above two algorithms, DAE-LSTM is a new multivariate time series prediction algorithm, which transforms unsupervised DAE into supervised data feature extraction method, and adds feature selection function to LSTM algorithm. In addition, in order to build a comprehensive forecasting framework, the hyperparameters of the model are determined by Bayesian optimization method(BOM) and the parameters of the model are solved by adaptive moment estimation(ADAM) optimization algorithm. In the training process, due to the adaptability of the BOM and ADAM, the forecasting model has few parameters that need to be set artificially, only including dimensional scale and time steps determined by data. Finally, the proposed forecasting framework is applied to forecast the sales of online retailers on JD.COM. In comparison with other machine learning algorithms, including multiple linear regression(MLR), support vector regression(SVR), artificial neural network(ANN) and LSTM, the proposed DAE-LSTM achieves the best prediction accuracy.

Keywords: Sales forecasting; Denoising autoencoder; Long short-term memory; Bayesian optimization method; Adaptive moment estimation.

1. Introduction

The businesses are most concerned about the profit. Therefore, sales directly related to profits have been widely researched. Among which, sales forecasting is a challenging task. In practice, the combined action of complex influence factors, unstable business strategy and market rules adds difficulty to the assignment[1].

Fortunately, many experts and scholars have contributed to the forecasting method of sales over the years. Those based on time series forecasting methods are introduced in [2]. Although sales data is time series data, the impact of factors on sales can not be ignored. To solve this problem, the classical regression model is applied in [3] which is based on getting reasonable influence factors about sales. However, it is very difficult to get the factors that have a linear relationship with sales. And in [4], a method for small samples based on Support Vector Machine is proposed. It can effectively solve the

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

problems of small sample, nonlinear and high-dimensional pattern recognition, and can be applied to other machine learning problems such as function fitting. But regrettably the applicability of this method is reduced because most sales forecasting is based on a large number of sample data. Meanwhile, there is a limitation in the above statistical methods: the need to transform qualitative data into quantitative data. On this foundation, paper [5] puts forward the method combining sentiment analysis and Bass model which is aimed at processing online review data for sales forecasting. Moreover, the shallow neural network method has been applied to the sales forecasting in literature [6] and promising results have been obtained. However, the prediction accuracy of the above method is not satisfactory when the characteristics of the prediction problem are fuzzy influence factors, massive samples with complex structure and a large time interval. It is an urgent challenge that shallow learning is difficult to depict the relationship between variables of sales prediction problem. Based on the development of deep learning, R. G. Hiranya Pemathilake et al. provide a hybrid model with integrated moving average and recurrent neural network. They combined traditional statistical models with deep learning and achieved promising results. However, there is no description of how to deal with dynamic influence variables in [7]. To sum up, there are still three main difficulties in sales forecasting problem.

a. Massive high-dimensional data increases the difficulty of computing and modeling.

b. There is a complex non-linear relationship between influencing variables and sales.

c. Sales is influenced by time factor. But the influence of time factor on sales is difficult to quantify.

Long Wang and Zijun Zhang elaborated that stack DAE model is capable to accurately forecast electricity prices in [8]. Meanwhile, LSTM takes account of historical information about sales and avoids the defects of gradient disappearance in the basic RNN[9]. For the sake of addressing the above three-mentioned difficulties, we propose the DAE-LSTM algorithm to deal with sales forecasting tasks. The advantage of DAE-LSTM is that it draws on the feature extraction mechanism of DAE and the cyclic structure of RNN. In addition, it is generally known that the machine learning algorithms, especially deep learning algorithm, have many hyperparameters that determine the structure of the model. However it is expensive for tuning these hyperparameters which is general a black-box operation or a rude global search[10]. BOM is a powerful strategy for searching the extrema of expensive cost object function[11]. We combine BOM with DAE-LSTM in this paper to seek the optimal model structure adaptively. The proposed prediction framework, including data feature extraction, adaptive modeling and parameter solving, is adaptive to actual data and greatly reduces the cost of tuning hyperparameters. Finally, We designed a comparative experiment including common machine learning methods, including MLR, SVR, ANN, LSTM and DAE-LSTM. The experimental results manifest that the DAE-LSTM algorithm can significantly improve the accuracy of sales forecasting problem.

Other parts of this paper are arranged as follows: The theories of DAE-LSTM, BOM and ADAM are introduced in section 2; A contrast experiment about sales forecast is provided in section 3; In section 4, we show the results and discuss them; The conclusion is given in section 5.

2. Method

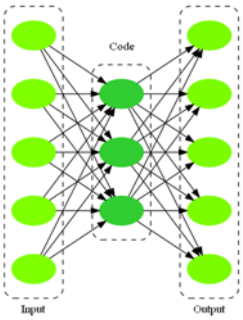
2.1. Denoising Autoencoder

DAE is a member of the class of autoencoder(AE) which can learn efficient representation of input data through unsupervised learning[20]. As shown in figure 1, the AE contains two processes of encoding and decoding. In the encoding process, the input vector $\boldsymbol{x}$, through a nonlinear mapping $h_{\theta_e}(x) = s(W_e x + b_t)$, $\theta_e = \{W_e, b_e\}$, translates into more representative feature $h_{\theta_e}(x)$. The reconstructing vector $\hat{\boldsymbol{x}}$ is decoded by mapping $g_{\theta_d}(x) = s(W_d x + b_d)$, $\theta_d = \{W_d, b_d\}$, in the next step. The contribution of AE is aimed at minimizing the reconstruction error showed as equation (1)[12].

$$L(x, \hat{x}) = -\sum_{i=1}^{n}[x_i \log(\hat{x}_i) + (1 - x_i)\log(1 - \hat{x}_i)] \quad (1)$$

Finally, through the backpropagation method, we solve the optimal parameters $\{\theta_e^*, \theta_d^*\}$:

$$\theta_e^*, \theta_d^* = \arg\min_{\theta_e, \theta_d} \frac{1}{N}\sum_{i=1}^{N} L(x_i, \tilde{x}_i) \quad (2)$$



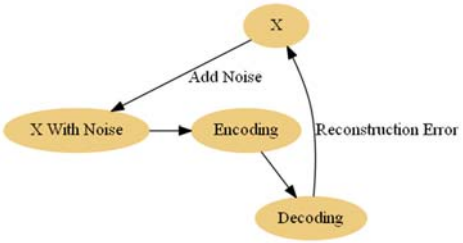


Figure 1. The Structure of AE. **Figure 2.** The Structure of DAE.

Since the AE only simply reconstruct the input data, it does not contain features of particularly useful information. In order to avoid the above situation and learn better feature representations, it is necessary to express certain constraints on the input data. The DAE learns more robust features by reconstructing the input data with additional noise[13]. As shown in figure 2, based on AE, a corrupted, partially damage process $C(\tilde{x}|x)$ is introduced into DAE. Then, the input data with additional noise $\tilde{\boldsymbol{x}}$ is treated as the input of the DAE and reconstruction error $L(\tilde{x}, \hat{x})$ is considered as optimization objective. Reference [12] details the surprising advantages of applying DAE-processed data for subsequent modeling.

2.2. Long Short-Term Memory

LSTM belongs to an improved RNN which is suitable for processing problems closely related to sequential nature[14]. The advantage of LSTM over RNN is that it can handle long-term dependency problems. In the standard RNN model, there is a cyclic structure as shown in figure 3. Where, cell A of the neural network reads some input $\boldsymbol{x_i}$ and outputs a value $\boldsymbol{h_t}$. This cell allows information to be passed from the current step to the next[15]. In a standard RNN, the repeating cell has only a very simple structure: a tanh activation function.

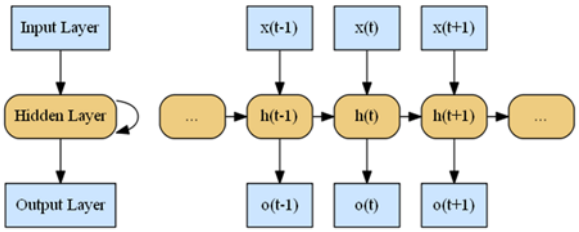


Figure 3. The Structure of RNN.

Unlike RNN, the LSTM repeating cell has there gates that interact in a very special way. The LSTM adds or removes information to the cell through a structure called gate. The gate consists of a sigmoid activation function and a linear operation. Its function is to allow information to be passed selectively[9]. Three gates are placed in a cell, called the input gate, the forgetting gate and the output gate respectively. Figure 4 shows the overall structure of LSTM. And the LSTM runs with the following three main processes. The algorithm steps of the LSTM are shown as follows.

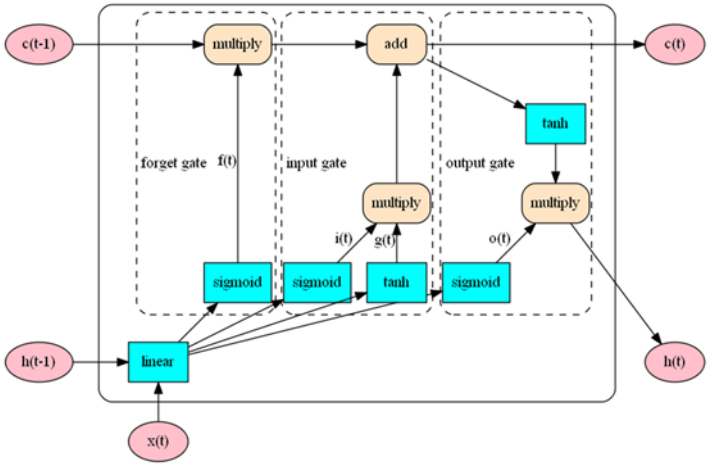


Figure 4. The Structure of LSTM.

Step1: The forgetting gate can forget unnecessary information. When the sigmoid operation takes input $\boldsymbol{X_t}$ which is the new input to this cell and $\boldsymbol{h_{t-1}}$ which is the output from the previous cell, it determines which parts are removed from the old output. The output of this step is shown as equation (3).

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{3}$$

Where $\boldsymbol{f_t}$ stands for sigmoid operation, $\boldsymbol{W_f}$ stands for weight and $\boldsymbol{b_f}$ represents bias.

Step2: Determine the new information stored in the cell. The sigmoid operation is located in the input gate, which determines which values will be updated. Then, a tanh operation creates a new candidate value vector, which is added to the cell. figure 4 shows where we actually discard the gender information of the old pronoun and add new information according to the previously determined goals.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4}$$

$$\widetilde{C_t} = \tanh(W_C \cdot [h_t, x_t] + b_C) \tag{5}$$

Where $\boldsymbol{i_t}$ and $\boldsymbol{\widetilde{C_t}}$ represents new outputs and tanh represents tanh operation.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t \tag{6}$$

Step3: Finally, decide what the network want to output. The sigmoid operation determines which parts of the information we want to output. Then, the information is obtained by tanh, and the output of all possible values multiplied by sigmoid operation is generated, so that the network can only output the values determined by it.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_0) \tag{7}$$

$$h_t = o_t \tanh(C_t) \tag{8}$$

Where $\boldsymbol{o_t}$ represents the output of the sigmoid operation in the output gate.

2.3. DAE-LSTM

Considering the robust data features extracted by DAE and the excellent properties of LSTM network for time series modeling, a new network structure named DAE-LSTM network is obtained by combining the two algorithms. The network structure is shown in figure 5.

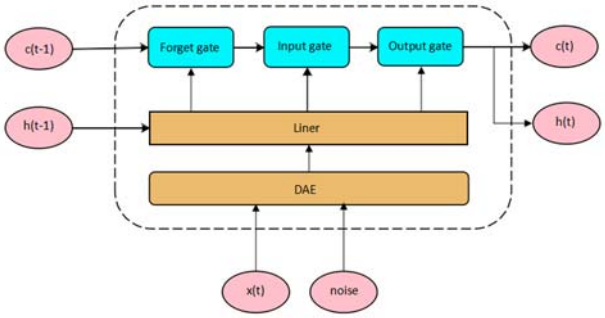


Figure 5. The Structure of DAE-LSTM.

In a limited time step, random noise is added to the original data and encoded. The encoded data features are used as the input of the LSTM network, and the connectivity of the hidden layer is retained. Finally, the prediction results are obtained through the full connection layer.

2.4. HyperParameter Optimization: Bayesian Optimization Method

For most machine learning algorithms, especially deep learning, it is expensive to evaluate their hyperparameters every time[16]. In this case, it is worthwhile to spend some extra time searching for the hyperparameter combination that is most likely to be improved[10]. BOM provides an elegant strategy for searching the extrema of expensive black-box function. Firstly, we summarize the task at this stage: finding the hyperparameter combination $\boldsymbol{x}^*$ that makes the objective function $f(x)$ optimal in the hyperparameter space. The objective function $f(x)$ is the black-box function which specific form is not clear. The BOM treats this tricky problem as a Gaussian Process(GP). A GP is a kind of extension of the multivariate Gaussian distribution[11]. In order to understand GP, as showed in figure 6, the GP is used to represent the distribution of function $f(x)$.

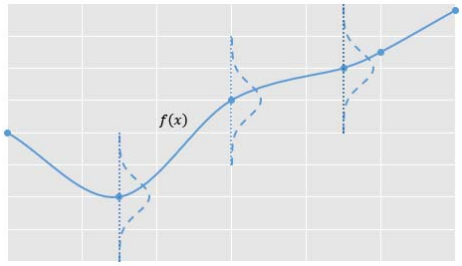


Figure 6. Gaussian Process.

Similar to the Gaussian distribution, which is determined by the mean and the variance, the GP is determined by the mean function $m(x)$and the covariance function $k(x, x')$. For convenience, we assume here that the prior means is a function $m(x) = 0$. The key to GP is to choose the suitable covariance function[17]. The automatic correlation determination(ARD) squared exponential kernel and ARD Matern 5/2 kernel are two popular forms of $k(x, x')$ as shown in equation (10) and equation (11).

$$r^2(x, x') = \sum_{d=1}^{D}(x_d - x'_d)^2/\theta_d^2 \tag{9}$$

$$K_{SE} = \theta_0 exp\left\{-\frac{1}{2}r^2(x, x')\right\} \tag{10}$$

$$K_{M52} = \theta_0\left(1 + \sqrt{5r^2(x, x')} + \frac{5}{3}r^2(x, x')\right) exp\left\{-\sqrt{5r^2(x, x')}\right\} \tag{11}$$

where, θ_0 represents the covariance amplitude and $\theta_{1:D}$ represent that different hyperparameters affect covariance differently.

Based on the determination of the covariance function, it's easy to get the expression for predicting distribution of f_{t+1} by applying the observations $D_{1:t} = \{x_{1:t}, f_{1:t}\}$[18]:

$$P(f_{t+1}|D_{1:t}, x_{t+1}) = N\left(\mu_t(x_{t+1}), \sigma_t^2(x_{t+1})\right) \tag{12}$$

Where

$$\mu_t(x_{t+1}) = k^T K^{-1} f_{1:t} \tag{13}$$

$$\sigma_t^2(x_{t+1}) = k(x_{t+1}, x_{t+1}) - k^T K^{-1} k \tag{14}$$

$$k = [k(x_{t+1}, x_1)k(x_{t+1}, x_2) \cdots k(x_{t+1}, x_t)] \tag{15}$$

$$k = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_t) \\ \vdots & \ddots & \vdots \\ k(x_t, x_1) & \cdots & k(x_t, x_t) \end{bmatrix} \tag{16}$$

In conjunction with the GP, An acquisition function $u(\cdot)$ needs to be constructed to search the optimal next evaluation $\left(x_{t+1}, f(x_{t+1})\right)$ based on current observations $D_{1:t}$. If the GP is compared to a function, the acquisition function $u(\cdot)$ is analogous to the derivative function of this function. As we all know, the point where the derivative function is equal to zero is called the stationary point of the original function. Similarly, as showed in figure 7, the point where the $u(\cdot)$ function reaches its maximum is the point needs to be evaluated in the next iteration:

$$x_{t+1} = \arg\max_x u(x|D_{1:t}) \tag{17}$$

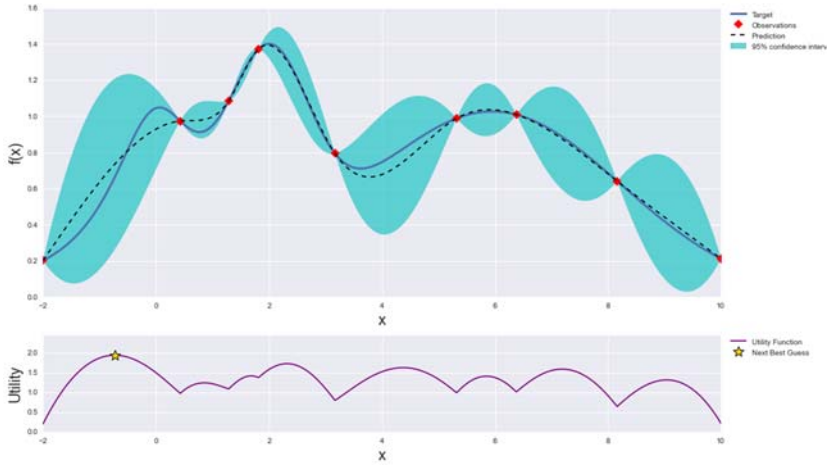


Figure 7. The Iterative Process of BOM.

According to the missions of $u(\cdot)$, a popular acquisition function is the probability of improvement(PI) over the incumbent $f(x^+)$[11], where $x^+ = arg\,max_{x_i \in x_1}$, as showed as equation (18). The more intuitive representation of $PI(x)$ is showed as figure 7. In fact, there are many other acquisition functions such as expected improvement and GP upper confidence Bound.

$$PI(x) = P\left(f(x) \geq f(x^+)\right) = \Phi\left(\frac{\mu(x) - f(x^+)}{\sigma(x)}\right) \tag{18}$$

Considering the GP and acquisition function, the iteration process of the BOM are shown in Algorithm 1.

Algorithm 1 Bayesian Optimization Method

Require:

X: hyperparameter set;
$u(\cdot)$: acquisition function;
f: objective function to be evaluated;
D: initialization of observation $(f(x_0), x_0)$.

Ensure:

for $t = 1, 2, \cdots$ **do**

Find x_t according to acquisition function $u(\cdot)$: $x_t = \arg max_x\, u(x|D_{0:t-1})$;
Evaluate the objective function: $y_t = f(x_t)$;
Update the ovservation $D_{0:t}$ and the GP.
End for

The advantage of DAE-LSTM is not only to combine the characteristics of DAE and LSTM, but also to transform the processing mechanism of DAE aiming at prediction data, which is the transformation from unsupervised feature extraction to supervised feature extraction. The data feature obtained in this way is more targeted, which is also called target oriented feature analysis. LSTM network itself does not have the function of feature selection. The encoded features make it easier to describe the rules between data, and then get a good performance prediction model. As the main model of this paper, DAE-LSTM is the core of the whole prediction framework.

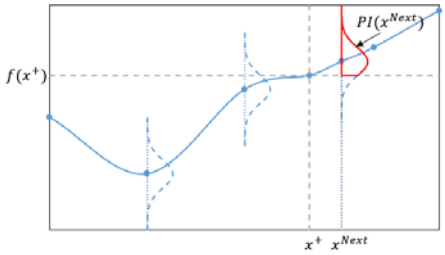


Figure 8. The Intuitive Show of PI.

3. Contrast Experiment

3.1. Experimental Data

The experimental data consists of 24000 sales records from 3000 online retailers of the Jingdong which was collected from the Jingdong Global Financial Data Explorer Contest. Each sales record contains 13 input data including the influencing factors of sales, and 1 output data that is next month sales. Descriptive statistics data of these variables are shown in table 1.

Table 1. Descriptive Statistics of the Variables.

Statistic	1st-Qu.	Mean	3rd-Qu.	Max.
Critical Review	1	10	9	1095
Mediocre Review	1	11	11	610
Positive Review	51	999	762	86989
Image Review	3	29	26	3462
Review Quantity	57	1020	800	87017
Advertise Amount	0	187885	69261	20887767
Order Quantity	159	848	1074	22073
Customer Quantity	159	846.6	1073	22073
The Month Sales	16474	60872	81807	2479758
Refund Quantity	4	19	22	1643
Refund Amount	161	1105	1102	71503
Preferential Quantity	21	472	577	21682
Preferential Amount	145	16375	12476	5233502
Next Month Sales	17102	62408	84447	2479758

Due to the large dispersion of variables, as shown in table 1, it is difficult to build a model of strong generalization ability. Meanwhile, the phenomenon illustrates that there are great discrepancies in the commodities operated by different stores in the dataset. Therefore, it is necessary to standardize data before the comparative experiment. In addition, some obvious wrong in raw data should be handled correctly. For example, the number of critical review in one month is 1095. However, the numbers of

other months are between 80 and 100, and the evaluation quantity has not changed significantly. The number of negative evaluation in this month would be replaced by the mean value of other months.

3.2. Experimental Design

Firstly, in order to eliminate the difference of magnitude between variables, the raw data is standardized by Z-Score method. Data processed by Z-Score method is recorded as standard data.

Secondly, before training, we should do correlation analysis to confirm this experiments is meaningful. Then, for contrast experiment of the prediction models, DAE-LSTM and other machine learning algorithms including MLR, SVR, ANN and LSTM are compared for sales forecasting. Among which, MLR is a typical type of multivariate prediction model that attempts to obtain the linear relationship between response variables and influence variables. In addition, as representatives of shallow learning, SVR and ANN attempt to characterize the non-linear relationship between variables[19]. However, the difference between the SVR and ANN is: SVR uses kernel function transformation technology to process data, ANN applies nonlinear activation function to train model. Furthermore, LSTM attempts to solves the vanishing gradient problem or exploding gradient problem in the process of long sequence training. It is worth mentioning that, due to only eight months of sales record, it is not suitable to apply classic time series methods.

Finally, the core indicators we focus on are the generalization of the model. To avoid contingency, we applied the coefficient of determination R^2 showed as equation (19) and mean square error(MSE) showed as equation (20) of the cross-validation experiment to measure the predicted accuracy. Meanwhile, to ensure that each experiment achieves the best results, the BOM and grid search method are applied to obtain the hyperparameters in the experiment. For each back propagation process, the parameters of the model are solved by adaptive moment estimation(ADAM) optimization algorithm which could dynamically adjust the learning rate for each parameter according to the first-order moment estimation and the second-order moment estimation of each parameter gradient of the loss function [20].

$$MSE(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{19}$$

$$R^2(y,\hat{y}) = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y}_i)^2} \tag{20}$$

4. Result and Discussion

4.1. Correlation Analysis

In this paper, Pearson correlation coefficient(PCC) defined as equation (21) is used to analyze the correlation properties between variables[21]. The PCC indicates the degree of correlation between variables. In this experiment, the PCC diagram between variables is shown in figure 9.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X\sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} \tag{21}$$

Firstly, figure 9 describes that the PCC between the sales of next month and influence variables is distributed between 0.15 and 0.75. Among which, the coefficient of the sales of this month and the next month is the maximum which is up to 0.75. The phenomenon shows a strong correlation between sales in different months. Secondly, there are significant correlations between the influence variables, especially between the positive evaluation and evaluation quantity, which is close to 1. Businesses are highly concerned about the positive evaluation and carry out various activities to improve it. Meanwhile, the PCC between the counts of discount and order is 0.8 that illustrates that discounts can greatly boost turnover. Finally, as can be seen, the next month sales are positively correlated with the negative evaluation number which is due to that the growth rate of negative evaluation is far less than positive evaluation.

4.2. Hyperparameters Selection

The model parameters refer to variables that are constantly updated in order to make the model more suitable for training data. However, the hyperparameters are parameters that set values before starting

the learning process which play a decisive role for the accuracy of models. In this experiment, we use grid search and 3-fold cross-validation to determine the hyperparameters of ANN and SVR. And the BOM is used to find the best hyperparameters of LSTM and DAE-LSTM which is cost expensive. table 2, 3, 4 and 5 separately display the set values of the hyperparameters of the SVR, ANN, LSTM and DAE-LSTM.

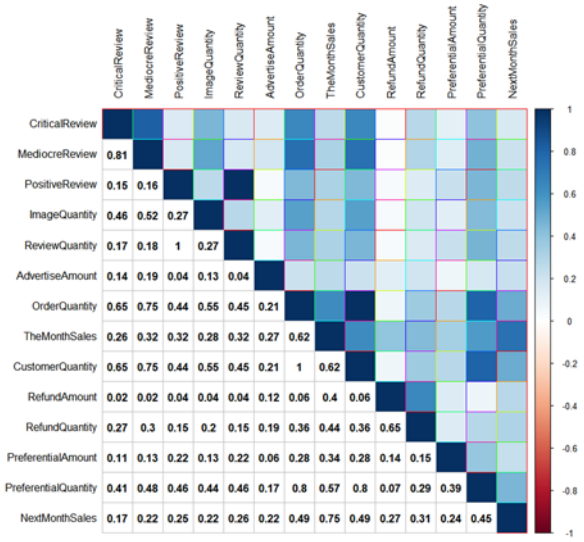


Figure 9. Pearson Correlation Coefficient Diagram between Variables.

Table 2. Hyper-parameters of the SVR.

Hyper-parameter	Value
Input Size	13
C	10
Gamma	0.0005
Epsilon	0.3
Kernel	rbf

Table 3. Hyper-parameters of the ANN.

Hyper-parameter	Value
Number of Layers	3
Input Size	13
Hide Size	23
Learning Rate	0.01
Activation Function	relu
Optimization Algorithm	ADAM

In SVR model, four hyperparameters including penalty parameter C, gamma, epsilon and Kernel deserve attention. The value of penalty parameter C represents the importance of outliers in the model. A larger C value represents a greater penalty for training error, but it is easy to appear over-fitting at this time. Gamma is the coefficient of the kernel function and its value must be greater than 0. With the increase of gamma, there is a situation where the test error is large, but the training error is smaller, and the model is easy to appear over-fitting[22]. Epsilon specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value. Finally, the transformation of the kernel function enables the model to describe the non-linear relationship.

As for hyperparameters setting for LSTM network, several illustrations are explicated in the next. Firstly, the lag time steps of the network is set at 7, which is according to the sales records of the past 7 months

to predict the sales volume of the next month. The choice of lag time steps is based on the time range of the dataset. Secondly, in order to avoid over fitting phenomenon, dropout method is applied when training the network. Thirdly, this paper sets up the gradually attenuated learning rate. That is, the initial learning rate is set to a larger value, so that the LSTM network converges quickly, and then attenuates the learning rate to further achieve the optimal solution. And the hyperparameters of DAE-LSTM is similar except the different network structure.

Table 4. Hyper-parameters of the LSTM Network.

Hyper-parameter	Value
Number of Layers	4
Input Size	13
H_1 Size	18
H_2 Size	16
Time Step	7
Learning Rate	0.005
Batch Size	200
Dropout Coefficient	0.43
Optimization Algorithm	ADAM

Table 5. Hyper-parameters of the DAE-LSTM Network.

Hyper-parameter	Value
Number of Layers	5
Input Size	13
Decode Size	7
H_1 Size	16
H_2 Size	15
Time Step	7
Learning Rate	0.07
Batch Size	200
Dropout Coefficient	0.29
Optimization Algorithm	ADAM

4.3. Model Performance

After determining the optimal hyperparameters of each model, they are verified by 10-fold cross-validation. Finally, we calculate the mean of MSE and R^2 for cross-validation as the evaluation criteria for each model. The mean of MSE and R^2 of the each cross-validation experiment are shown as table 6 and figure 10. table 6 and figure 10 demonstrate that DAE-LSTM has the best generalization performance.

Table 6. The Mean MAE and EVS of 10-Fold Cross-Validation of Each Model.

Model	Mean MSE	R^2
MLR	0.4344	57.10%
SVR	0.4205	58.37%
ANN	0.4246	57.68%
LSTM	0.2702	65.22%
DAE-LSTM	0.2594	66.74%

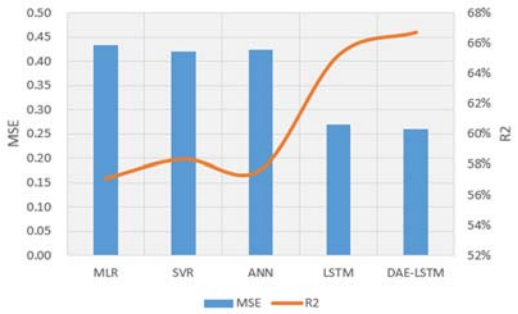


Figure 10. The Mean MAE and EVS of 10-Fold Cross-Validation of Each Model.

The prediction effect of DAE-LSTM in validation set is shown as figure 11. As shown in figure 11, it is satisfactory for the prediction effect of the DAE-LSTM model. The proposed sales forecasting framework improves the accuracy of forecasting under acceptable computing cost and the entire process of automated modeling is achieved.

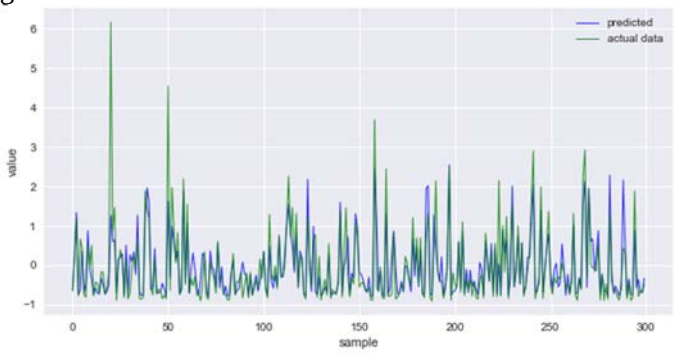


Figure 11. Prediction Effect of DAE-LSTM in Validation Set.

5. Conclusion

In this paper, we put forward a sales forecasting framework based on DAE-LSTM. The proposed framework, together with other machine learning methods such as MLR, SVR, ANN and LSTM, is applied to forecast the sales of online retailers on JD.COM. Firstly, the correlation analysis of experimental data shows that there are strong linear correlations between the influential variables. Then, in the training process, the BOM and ADAM method are used to determine the hyperparameters and parameters of the model respectively. And the modeling process is close to automated completion. The empirical results reveal that, compared with other models, the DAE-LSTM which takes account of past seven months' sales records to predict future sales shows the best generalization ability. In the future, adding business strategy disturbance to the model and exploring more robust features of dynamic sales data will be the next focus of work.

References

[1] Chi-Jie Lu. Sales forecasting of computer products based on variable selection scheme and support vector regression. *Neurocomputing*, 128:491–499, 2014.

[2] A. Cyril, R. H. Mulangi, and V. George. Modelling and forecasting bus passenger demand using time series method. In *2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 460–466, Aug 2018.

[3] G. T, R. Choudhary, and S. Prasad. Prediction of sales value in online shopping using linear regression. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pages 1–6, Dec 2018.

[4] Dalibor Petkovi, Shahaboddin Shamshirband, Hadi Saboohi, Tan Fong Ang, Nor Badrul Anuar, and Nenad D. Pavlovi. Support vector regression methodology for prediction of input displacement of adaptive compliant robotic gripper. *Applied Intelligence*, 41(3):887–896, 2014.

[5] Zhi-Ping Fan, Yu-Jie Che, and Zhen-Yu Chen. Product sales forecasting using online reviews and historical sales data: A method combining the bass model and sentiment analysis. *Journal of Business Research*, 74:90–100, 2017.

[6] Y. Qin and H. Li. Sales forecast based on bp neural network. In *2011 IEEE 3rd International Conference on Communication Software and Networks*, pages 186–189, May 2011.

[7] R. G. Hiranya Pemathilake, S. P. Karunathilake, J. L. Achira Jeewaka Shamal, and G. U. Ganegoda. Sales forecasting based on autoregressive integrated moving average and recurrent neural network hybrid model. In *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 27–33, July 2018.

[8] L. Wang, Z. Zhang, and J. Chen. Short-term electricity price forecasting with stacked denoising autoencoders. *IEEE Transactions on Power Systems*, 32(4):2673–2681, July 2017.

[9] Huan-huan Wang, Long Yu, Sheng-wei Tian, Yong-fang Peng, and Xin-jun Pei. Bidirectional lstm malicious webpages detection algorithm based on convolutional neural network and independent recurrent neural network. *Applied Intelligence*, 2019.

[10] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. *arXiv e-prints*, page arXiv:1206.2944, Jun 2012.

[11] Eric Brochu, Vlad M. Cora, and Nando de Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *arXiv e-prints*, page arXiv:1012.2599, Dec 2010.

[12] Pascal Vincent, Hugo Larochelle, Y Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. *In Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103, 01 2008.

[13] Chunhong Zhang, Tiantian Li, Zhibin Ren, Zheng Hu, and Yang Ji. Taxonomy-aware collaborative denoising autoencoder for personalized recommendation. *Applied Intelligence*, 2019.

[14] Sepp Hochreiter and Jrgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[15] Alex Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*, volume 385. 01 2012.

[16] Rafael Alcal, Jess Alcal-Fdez, Mara Jos Gacto, and Francisco Herrera. Improving fuzzy logic controllers obtained by experts: a case study in hvac systems. *Applied Intelligence*, 31(1):15–30, 2009

[17] J. Mockus, V. Tiesis, and A. Zilinskas. The application of bayesian methods for seeking the extremum. *Towards Global Optimization*, 2:117–129, 1978.

[18] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts, 2006.

[19] Karthik Nadig, Walter Potter, Gerrit Hoogenboom, and Ronald Mcclendon. Comparison of individual and combined ann models for prediction of air and dew point temperature. *Applied Intelligence*, 39(2):354–366, 2013.

[20] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[21] Hongliang Zhu, Wenhan Liu, Maohua Sun, and Yang Xin. A universal high-performance correlation analysis detection model and algorithm for network intrusion detection system. *Mathematical Problems in Engineering*, 2017.

[22] P. Tsirikoglou, S. Abraham, F. Contino, C. Lacor, and G. Ghorbaniasl. A hyperparameters selection technique for support vector regression models. *Applied Soft Computing*, 61:139 – 148, 2017.