Received November 24, 2018, accepted December 12, 2018, date of publication December 17, 2018,
date of current version January 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2887138

A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM

CHONGREN WANG[1], DONGMEI HAN[1,2], QIGANG LIU[3], (Student Member, IEEE), AND SUYUAN LUO[1,4]

[1]School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China
[2]Shanghai Key Laboratory of Financial Information Technology, Shanghai 200433, China
[3]SHU-UTS SILC Business School, Shanghai University, Shanghai 201899, China
[4]College of Management, Institute of Big Data Intelligent Management and Decision, Shenzhen University, Shenzhen 518060, China

Corresponding authors: Chongren Wang (chongrenw@163.com) and Dongmei Han (dongmeihan6@163.com)

ABSTRACT With the rapid development of the peer-to-peer lending industry in China, it has been a crucial task to evaluate the default risk of each loan. Motivated by the research in natural language processing, we make use of the online operation behavior data of borrowers and propose a consumer credit scoring method based on attention mechanism LSTM, which is a novel application of deep learning algorithm. Inspired by the idea of Word2vec, we treat each type of event as a word, construct the Event2vec model to convert each type of event transformation into a vector and, then, use an attention mechanism LSTM network to predict the probability of user default. The method is evaluated on the real dataset, and the results show that the proposed solution can effectively increase the predictive accuracy compared with the traditional artificial feature extraction method and the standard LSTM model.

INDEX TERMS P2P lending, credit scoring, machine learning, deep learning, LSTM, attention mechanism.

I. INTRODUCTION

With the development of computer technologies represented by big data, cloud computing, social networks and so on, various aspects of social life have undergone significant changes. Peer-to-peer (P2P) lending has recently been applied in the United States, Europe, and China. As a form of financial innovation and alternative finance, P2P lending has attracted many clients due to its convenience and comparatively low charges [1]. Recent years, the P2P lending industry is developing rapidly in China, which extends the field of financial services and solves the financing needs of many individuals and small businesses. According to the statistics from the Online Loan Home,[1] the number of operating platforms of P2P lending in China has reached 1931 by the end of December 2017. In 2017, the turnover of the online lending industry reached 280.48 billion yuan and the accumulated turnover of the online lending industry exceeded 6 trillion yuan.

While the P2P lending industry is developing at a high speed, various risk issues emerge. For example, the non-performing loan ratio of the platform rises, and the number of closed platforms increases. One of the crucial challenges that P2P lending platforms are facing is to accurately predict the default risk of each loan by tapping into consumer credit scoring models, thus effectively helping P2P online lending companies avoid credit risks [2].

Consumer credit scoring predicts individual repayment ability and repayment willingness based on individual historical records. The credit scoring model outputs the individual's default probability, and then divides the borrowers into "good" borrowers and "bad" borrowers. Compared with traditional consumer credit scoring of bank, credit scoring of P2P lending has the characteristics of high dimensional, diversified data format and large-sized [2], which brings many new challenges to the research field of consumer credit scoring.

The research methods on consumer credit scoring mainly includes expert models, statistical methods, machine learning methods and so on. Recently, some studies have tried to apply deep learning methods to the field of consumer credit scoring. Tomczak and Zięba [3] proposed a credit scoring method based on classification-based Boltzmann machine (Class RBM) and proved that the method got higher prediction performance. Yu *et al.* [4] proposed a new multi-level

[1]http://osscdn.wdzj.com/upload/2017wdnb.pdf

deep belief network (DBN) based on the Extreme Learning Machine (ELM) integrated learning method to improve the accuracy of credit risk assessment. However, the above studies are based on traditional feature data, whose features are design by experts.

With the development of P2P lending industry, a large amount of online operation behavior data is generated, including the record of user operating equipment, operating position, operation records and operating habits. The online operation behavior data is valuable for mining. With the improvement of deep learning technology, the application of online operation behavior data is becoming more and more practicable. At present, some researchers have noticed this problem, and some studies have tried to study user operation behavior data based on deep learning methods. Hidasi *et al.* [5] introduced RNN into the Internet recommendation system based on the user online operation behavior data, and built a recommendation system based on deep learning, experiments showed that this method is superior to the traditional recommendation method. Lang and Rettenmeier [6] used LSTM to predict the future behavior of consumers based on the consumer behavior sequence on the user's e-commerce website, empirical studies on a large European e-commerce platform show that the method has good predictive effect. Wang *et al.* [7] studied transaction fraud in the field of e-commerce based on the data of a large e-commerce website, the author proposed a novel method to capture user fraud in e-commerce websites. Based on the sequence of conversational behaviors on the business website, the LSTM method is used to model the user click sequence. The method proposed by the author is over three times better than the original fraud detection method. Although the deep learning model based on user online operation behavior data has been applied to the above fields, no research has been found to apply this method to the field of consumer credit scoring.

The data of borrower's online operation behavior on P2P lending website is a new data source of credit scoring, which records a series of online operations, such as user registration records, user login records, user click records, user browse records, user authentication records and other records. When generating credit scoring based on user online operation behavior data, the traditional method uses artificial design features method, which artificially designs features from the original data and generates feature data, then, trains model by statistical method or machine learning method. However, this kind of modeling method for extraction features manually requires high labor costs and professional domain knowledge. What is more, the features extracted manually are often incomprehensive and it is difficult to extract high-level features. Finally, as to online operation behavior data, the traditional method ignores the sequence information of user behavior.

Compared with the artificial feature extraction method, the deep learning method can automatically extract features, and even higher-level features without relying on expert knowledge and artificial design features.

Our research is inspired by the applications of deep learning methods in natural language processing (NLP), more specifically text classification. We propose a consumer credit scoring method based on attention mechanism LSTM by treating the sequence of behavior as sentences and event as words.

The main contributions of our research are as follows:

1.We introduce deep learning method into the field of credit scoring and put forward a method for credit scoring based on the user operation behavior data of P2P lending industry.

2. This paper use Word2vec's ideas, treat each type of event as a word, then build an Event2vec model that converts each type of event transformation into a vector.

3. Based on the LSTM neural network, the attention mechanism is introduced. The attention mechanism can calculate the importance of the LSTM output at different times and extract more critical information.

4.The method proposed in this paper is evaluated on the real dataset. Experiment results show that compared with traditional artificial feature extraction and standard LSTM model, the consumer credit scoring method based on the attention mechanism LSTM has significantly improved. consumer credit scoring

The structure of this paper is as follows. The second section introduce the techniques used in the proposed method, including Word2Vec, LSTM and attention mechanism. The third section introduces the consumer credit scoring method based on the attention mechanism LSTM proposed in this paper. Section 4 presents experimental setting and results analysis. The last section summarizes our research.

II. THEORY AND METHOD

A. Word2vec

In the field of NLP, when text data is processed, it is necessary to convert the text into a mathematical form firstly, and the word vector is an important method. The input event is text, and the output is the word vector which mainly includes two forms. One is sparse vector; the other is dense vector. Sparse vector is a simple and understandable representation method, often referred to as one-hot representation. This method uses a very long vector to represent a word. The length of the vector is determined by the size of the dictionary and often large. Only one dimension in the vector has a value of 1, and the value of other dimensions are zero. Because the vector length is too long, this representation method is prone to curse of dimensionality problem and does not represent semantic correlation between words. Dense vector is also known as distributed representation. Hinton [8] proposed the concept of distributed representation, which express words in the form of word vector. It is based on the hypothesis that "words with similar contexts should have similar semantics." The principle of this method is to map each word into a dense vector by model training. This representation method reduces

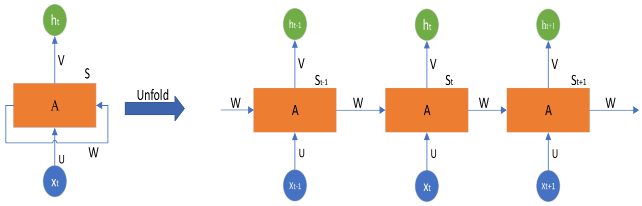

FIGURE 1. Structure of the RNN model.

the dimensions of the vector and show the semantic similarity between words at the same time.

There are many types of distributed representations, such as C&W [9], M&H [10], and the Neural Network Language Model (NNLM) [11]. In 2013, Mikolov *et al.* [12] proposed the CBOW (Continuous Bag of Words) and Skip-gram models. The tool for implementing CBOW and Skip-gram models is called Word2vec, by which word vectors can be obtained through fast and efficient training. The basic principle of Word2vec is to obtain the word vector by using the neural network model. The word vector can express the semantic similarity between words, so that the distance between the two words with similar meanings will be very close. In CBOW model, we need to predict the probability of the occurrence of the headword when we know the context. However, In Skip-gram model is when we know the headword, we need to predict the probability of the occurrence of the context word. Word2vec is simple, efficient, and widely used in the field of NLP and has achieved success.

B. LSTM

The Recurrent Neural Network (RNN) introduces the recurrent circuit structure into the traditional neural network. As shown in Figure 1, structure of RNN consists of three layers: the input layer, the hidden layer, and the output layer.

Suppose x is a vector, it represents the value of the input layer. We denote S as the value of the hidden layer. h indicates the value of the output layer. U represents the weight matrix of the input layer to the hidden layer, and V refers to the weight matrix of the hidden layer to the output layer. W represents the weight matrix for the previous time point to the current time point of the hidden layer. t represents the current time point and $t-1$ represents the previous time point. $t+1$ represents the next time point. The value of the hidden layer of the RNN not only depends on the input of the current time point, but also depends on the value of the hidden layer at the previous time point. The formulas of the RNN is shown as below:

$$\mathrm{h}_t = g(V\mathrm{s}_t) \tag{1}$$

$$\mathrm{s}_t = f(U\mathrm{X}_t + W\mathrm{s}_{t-1}) \tag{2}$$

Equation 1 and equation 2 are the calculation formulas of the output layer and the hidden layer respectively. Both g and f represent the activation function. The output layer is a fully connected layer. Each node of the output layer is connected to the hidden layer. The hidden layer is a recurrent layer. Main difference between the recurrent layer and the fully connected layer is that the recurrent layer adds weight matrix W to the equation. After repeatedly introducing the equation 2 into the equation 1, we can derive:

$$\begin{aligned}
\mathrm{h}_t &= g(V\mathrm{s}_t) \\
&= gVf(U\mathrm{x}_t + W\mathrm{s}_{t-1}) \\
&= gVf(U\mathrm{x}_t + Wf(U\mathrm{x}_{t-1} + W\mathrm{s}_{t-2})) \\
&= gVf(U\mathrm{x}_t + Wf(U\mathrm{x}_{t-1} + Wf(U\mathrm{x}_{t-2} + W\mathrm{s}_{t-3}))) \\
&= gVf(U\mathrm{x}_t + Wf(U\mathrm{x}_{t-1} \\
&\quad + Wf(U\mathrm{x}_{t-2} + Wf(U\mathrm{x}_{t-3} + \ldots))))
\end{aligned} \tag{3}$$

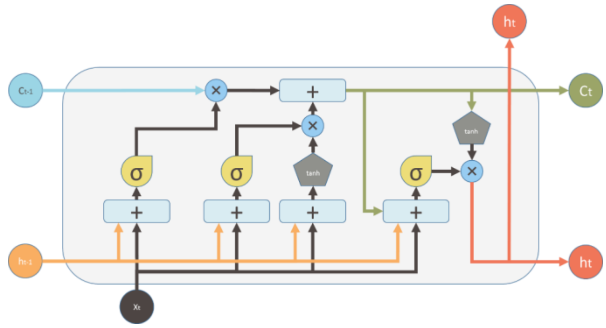

FIGURE 2. Structure of the LSTM model.

For the traditional neural network, the layers are fully connected, and the nodes in the same layer are not connected. For the sequence data, the network structure cannot mine the sequence information. For the RNN, the nodes in the same layer are connected. When calculating the output of the current hidden layer, the state of the hidden layer of the input layer and the previous moment can be utilized at the same time. As a result, the RNN has the ability to remember.

Because RNN is prone to gradient explode or gradient vanish during training, it is difficult to learn long-term dependencies. In order to solve this problem, Hochreiter and Schmidhuber (1997) [13] proposed a long short-term memory network (LSTM) based on RNN. In LSTM network, a special structural unit and three kind of unique "gate" structures are designed. The information passing through the unit is selectively added or removed. The "gate" structure is implemented using the sigmoid function. The value of sigmoid ranges from 0 to 1, which denotes how much information can be allowed to pass. The structure of the LSTM is shown in Figure 2.

Firstly, the LSTM unit processes the information of the previous memory state through the forget gate to determine which information to be forgotten from the memory state.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{4}$$

Next, the LSTM unit decides what information is stored in memory. On one hand, the input gate decides which information to update. On the other hand, the candidate vector is updated by a tanh layer.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{5}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{6}$$

Next, the LSTM unit combines the two parts above to update the memory state.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{7}$$

Finally, the LSTM unit uses the output gate to control the memory state which needs to be output.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{8}$$
$$h_t = o_t * tanh(C_t) \tag{9}$$

The main training method for RNN model is the Back-Propagation Through Time (BPTT) [14]. The BPTT algorithm is an improvement of the classical backpropagation (BP) algorithm. BPTT firstly expands the LSTM into a deep network in chronological order, and then uses the BP algorithm to train the network. Because the BPTT algorithm is clear and computationally efficient, this paper uses the BPTT algorithm to train the LSTM network.

One disadvantage of the standard LSTM is that it only uses the forward information without considering the reverse information. The bidirectional LSTM (BLSTM) has a forward LSTM and a reverse LSTM in the hidden layer. Compared to the standard LSTM, BLSTM can capture more information.

C. Attention Mechanism

The Attention Mechanism in the deep learning model is a model that simulates the attention of the human brain. In the mid-1990s, attention mechanisms were firstly proposed in the field of image research. But the real emphasis on attention mechanisms began with the study of image recognition guided by the Google DeepMind team [15]. When people observe images, they don't carefully look at every pixel of the image. Instead, they focus their attention selectively on some important parts of the image, ignoring other unimportant parts. So, the DeepMind team introduced the attention mechanism into RNN for image recognition research.

After that, Bahdanau *et al.* [16] applied the attention mechanism to the field of machine translation. This study was the first to apply the attention mechanism to the field of NLP. The author proposed an Encoder-decoder machine translation model based on the attention mechanism. Traditional Encoder-decoder neural network machine translation uses two RNN. One encodes the input text into intermediate hidden vector, and then uses the other one to decode according to the hidden vector. However, the drawback of this method is that the weight of each word in the input text is same. So, the author adds the attention mechanism to the Encoder-decoder model and calculates the importance of the correlation between the input and output of the translation model by the attention mechanism.

III. PROPOSED MODELS

This research is motivated by research in the field of NLP based on deep learning, and more specifically inspired by text classification research based on deep learning.

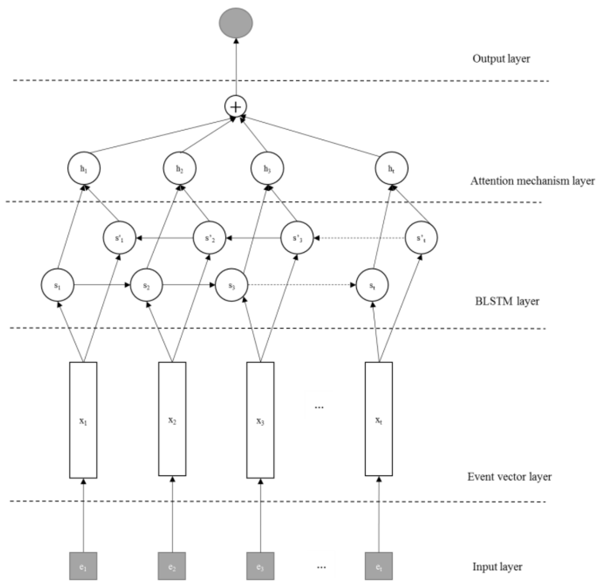

FIGURE 3. Structure of the AM-LSTM model.

Text classification is a basic NLP task. With the rapid development of deep learning technology in recent years, it has shown advantages in the field of text classification. Zhu *et al.* [17] used the RNN model to solve the problem of text sentiment classification, the author proposed a LSTM model to solve the problem of text emotion classification. Yang *et al.* [18] combined the Attention mechanism with RNN and applied it to the English text classification, the author proposed a two levels of attention mechanisms at the level of words and sentences.

By treating the sequence of behavior as sentences and event as words, we propose a consumer credit scoring method based on attention mechanism LSTM(AM-LSTM). Although there are similarities between natural language and user behavior sequences, but there are still differences. The number of event categories of user behavior is much smaller than vocabulary size of language. As is shown in Figure 3. the structure of our method is divided into five layers from the bottom to the top: input layer, event vector layer, BLSTM layer, attention mechanism layer, and output layer.

Input layer:

The borrower's online operation behavior data can record the user's click behavior, browse behavior, input behavior, etc. These behaviors can be treated as events. We process the raw online operation behavior record data, then convert those behavior actions into a sequence of events in chronological order.

Event vector layer (Event2vec):

Motivated by the idea of Word2vec, we treat each type of event as a word, and then construct the Event2vec model to convert each type of event into a vector. Since event is a categorical variable, the traditional way of processing the categorical variable is One-hot encoding. Compared with One-hot encoding, using Event2vec to vectorize events has two advantages. on one hand, the Event2vec model can convert each event into a low-dimensional dense vector, which is more conducive to the subsequent processing of the neural network. On the other hand, the vector generated by the Event2vec model has semantic properties that can represent similarities between events. The original event sequence can be transformed into an event vector sequence after being transformed by Event2vec. Each time point of the sequence is a feature vector. The vector at the time point i is x_i.

BLSTM layer:

As shown in Figure 3, the BLSTM network is implemented by extending a second hidden layer LSTM on a standard LSTM. The two hidden layers process input data from different directions.

Attention mechanism layer:

We propose an attention mechanism for the BLSTM network to apply to the credit scoring field. As shown in Figure 3, the attention mechanism layer is based on the output of BLSTM layer, and the output of the forward and reverse hidden layers of the BLSTM is concatenated together as an input for the attention layer.

We use the attention mechanism to calculate the importance of the BLSTM output at each moment, and then weight and sum the output results of each moment. The attention mechanism formulas are below:

$$u_i = tanh(Wh_i + b) \tag{10}$$

$$a_i = softmax(u_i) \tag{11}$$

$$c = \sum_{i=1}^{t} a_i h_i \tag{12}$$

where h_i represents the output of the i time point of BLSTM; t represents the length of the event sequence; a_i represents the weight of the output of the i time point; c refers to the weighted total of the BLSTM output at each time point.

Output layer:

We use the Sigmoid function to receive the output of the attention layer and get the output of the borrower's default probability result.

$$y = sigmoid(Wc + b) \tag{13}$$

To evaluate the effect of the model more comprehensively and consider the need to evaluate the default probability of the model output, we chose three indicators which are commonly used in credit scoring to evaluate the performance of the model: ROC (Receiver Operating Characteristic) curve, AUC (Area under Curve) and KS (Kolmogorov-Smirnov).

We define:

FN: False Negative, which represents that the predicted result is a negative sample but is actually a positive sample.

FP: False Positive, which represents that the predicted result is a positive sample but is actually a negative sample.

TN: True Negative, which represents that the predicted result is a negative sample and is in fact a negative sample.

TP: True Positive, which represents that the predicted result is a positive sample, and in fact it is also a positive sample.

At first, the values of true positive rate (TPR) and false positive rate (FPR) are calculated. Then, FPR and TPR were used as coordinates to form a line graph, namely ROC curve. The formula for calculating TPR and FPR are as follows:

$$TPR = \frac{TP}{TP + FN} \tag{14}$$

$$\mathrm{FPR} = \frac{FP}{TN + FP} \tag{15}$$

The closer the ROC curve to the upper left corner of the graph, the higher the accuracy of the classification model. AUC is the area under the ROC curve. The larger the AUC, the better the performance of the model.

KS (Kolmogorov-Smirnov) is a commonly used discrimination evaluation indicator in the field of credit scoring. Firstly, the data samples are sorted according to the predicted default probability from low to high, and then the cumulative TPR value and the cumulative FPR value under each default rate are calculated. Finally, the maximum value of the difference between the two values is obtained as the KS value. The larger the KS value, the stronger the model's ability to distinguish between default borrowers and on-time repayment borrowers.

In order to prove the advantages of the AM-LSTM method, we make a detailed comparative analysis. we selected four comparison methods, including the following methods:

1. BOA-XGBoost: the traditional method uses artificial design features. We select the number of times of each type of event as the feature, that is, each event is a feature. The value of the feature is the number of times the user operates the event.

extreme gradient boosting using Bayesian hyper-parameter optimization for credit scoring(BOA-XGBoost) was proposed by Xia *et al.* [19] , this method make use of extreme gradient boosting (XGBoost) [20] as the classification algorithm, which has been used most times in the data competition platform of Kaggle and got best performance. meanwhile this method chooses Bayesian optimization algorithm (BOA) to optimize the hyperparameters of XGBoost. BOA-XGBoost have be proved that this method outperforms baseline models, therefore, we choose this method as the credit scoring model of artificial design features.

2. LSTM: standard LSTM network is used, and the state vector of the last moment is used as the output for the LSTM network.

3. BLSTM: A bidirectional LSTM network is used, and the state vector of the last moment is used as the output for the bidirectional LSTM network.

4. BLSTM-Meanpool: a bidirectional LSTM network is used, and the intermediate state of all moments is output for the LSTM network, and the output variables of all moments are averaged.

TABLE 1. Parameter setting of AM-LSTM model.

Parameter	*Parameter Description*	*Parameter Value*
maxlen	event sequence length	100
n_events	Number of _events type	200
event_dim	dimension of event vector	16
lstm_units	neurons number of LSTM model	50
lstm_dropout	Dropout ratio	0.2
mini_batch_size	Mini-batch size	64

IV. EXPERIMENTAL SETUP AND RESULT ANALYSIS

A. DATASET

The dataset of our research is provided by an anonymous P2P lending platform in China. When borrowers apply for loan for the first time on the P2P lending platform, they need to submit the required application materials, which will generate a large amount of online operation data. Therefore, the user group selected in this paper includes the new borrowers of the P2P lending platform, that is, the users who applied for loan for the first time on the platform. Our dataset contains the online operation behavior data and other credit data of 100,000 new borrowers. The user's online operation behavior data contains all the online operation behavior event records before the user's first loan. The label of the data is whether the user defaults. If the user defaults, the label is defined as 1, otherwise the label is defined as 0.

B. PARAMETER SETTING AND MODEL TRAINING

We chose Keras,[2] a deep learning library based on Python, as the deep learning framework, which is an open source high-level neural network API. We chose Tensorflow[3] as the back end of Keras, which is an open source deep learning tool developed by Google. We use the Gensim[4] library to train the Event2vec model. Gensim is an NLP toolkit based on Python.

We randomly select 80% of the dataset as training set, and 20% of the data as the test set, and randomly select 10% of the training set as the verification the parameters of the deep learning models are selected by grid search, and the parameters of the AM-LSTM model proposed in this paper are shown in table 1. We process the original data and convert each user's original operation behavior data into an event sequence. Considering that length of each user's event sequence is different, we convert the sequence into a fixed length sequence, sequences with event sequence length greater than fixed length are truncated from the back to the front, and sequences with sequence length less than fixed length are padding processed from the back to the front.

Based on sequence data, we construct Event2vec model, which transforms each event into a vector. We adopt Adam (Adaptive Moment Estimation) method to train deep

[2]https://keras.io/
[3]https://www.tensorflow.org/
[4]https://radimrehurek.com/gensim/

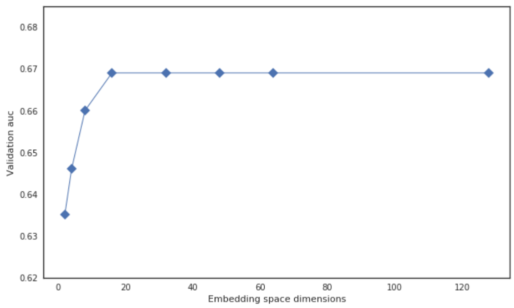

FIGURE 4. The influence of event vector space dimensions on the performance of AM-LSTM model.

TABLE 2. Model results of operation behavior dat.

Models	KS	AUC
BOA-XGBoost	0.099	0.561
LSTM	0.196	0.623
BLSTM	0.215	0.643
BLSTM-Meanpool	0.223	0.649
AM-LSTM	0.246	0.669

learning models. Dropout is an effective method to prevent over-fitting and improve the model effect in deep learning. It is a method of discarding neural units from the network according to certain probability during training of neural network [21]. In this paper, Dropout mechanism is added to BLSTM layer. At the same time, to prevent over-fitting we adopt the strategy of early stopping.

C. RESULT ANALYSIS

In order to identify the effect of dimensionality of the event vector space on prediction performance, we repeat our experiments using event vector spaces of 2, 4, 8, 16, 32, 48, 64 and 128 dimensions, keeping the other parameters unchanged. The AUC results for the validation set for different numbers of dimensions are shown in Figure 4. In the figure, the low dimension of the embedded space will have an adverse effect on the prediction effect. As long as the dimension of the embedded space is greater than 16, the effect will tend to be stable. Therefore, this paper finally sets the dimension of the embedded space to 16, so that it can achieve the best model performance and prevent the model from significant overfitting.

KS value and AUC value of the experimental results of the five models are shown in Table 2. As can be seen from the table, four different LSTM-based models perform better than the XGBoost method. On the one hand, this indicates that the LSTM-based deep learning models can find some sequence information in the field of credit risk assessment based on P2P lending user online operation behavior data. On the other hand, this indicates that the performance of LSTM-based deep learning models is better than traditional artificial feature extraction method. As seen from the table, compared with standard LSTM, KS value of BLSTM increases 9.69% and AUC value increases 3.21%. Therefore, performance

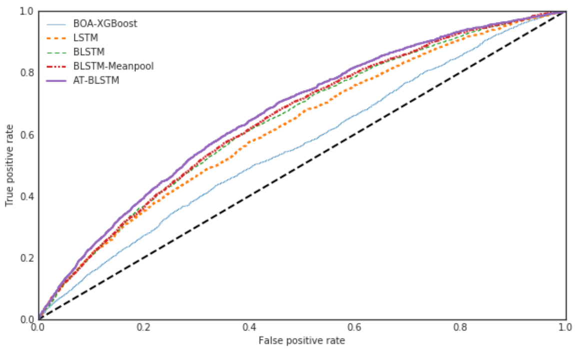

FIGURE 5. ROC curve (models of operation behavior data).

of BLSTM is better than standard LSTM, indicating that the bidirectional LSTM can obtain sequence information in different directions.

As seen from the experimental results, compared with BLSTM, KS value of BLSTM-Meanpool has improved by 3.72% and the AUC value improved by 0.93%, indicating that the performance of BLSTM-Meanpool improved. The BLSTM model uses the state of the last moment as the final output, thus discarding the output information of the previous moment. Because the BLSTM-Meanpool averages the outputs at all t moments, the drawback of BLSTM-Meanpool is that it simply summarizes the output at each moment. However, the importance of different moments may be different, so the attention mechanism is introduced. AM-LSTM is the attention-based LSTM model proposed in this paper. Compared with BLSTM-Meanpool, KS and AUC values of AM-LSTM are improved. KS value of AM-LSTM improved by 10.3%, and AUC value of AM-LSTM improved by 3.08%. This indicates that the attention mechanism enables the model to pay close attention to the specific target characteristics during the training process. At the same time, as can be seen from Figure 5, the ROC curve of AM-LSTM is always at the top left of the graph, which proves that AM-LSTM has better user credit default prediction performance.

In order to better demonstrate the superiority of applying user operation behavior data to the field of consumer credit scoring, we perform an additional experiment. We construct two credit scoring models. Data source of one model is credit data other than user behavior data, this model is called model without the operation behavior data. The other model is called model without operation behavior data, on the basis of the first model, data source of this model adds user operation behavior score, which refers to the output result of AM-LSTM model based on user operation behavior data. Both models choose BOA-XGBoost as the credit scoring method. Experimental results of above models are shown in Table 3. Compared with model without the operation behavior data, KS value of model with operation behavior data has improved by 10.7% and the AUC value improved by 5.65%. At the same time, as can be seen from Figure 6, the ROC curve of model with operation behavior data is always at the top left of the graph, which indicates that the performance of model with the operation behavior data outperforms model without the operation behavior data. This fully proves the advantages of operation behavioral data in consumer credit scoring field.

TABLE 3. Model results of models with and without operation behavior data.

Models	KS	AUC
Model Without Operation Behavior Data	0.28	0.672
Model With Operation Behavior Data	0.31	0.71

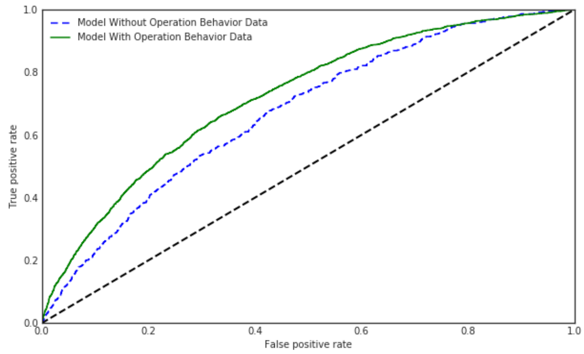

FIGURE 6. ROC Curve (Models With And Without Operation Behavior Data).

In summary, we analyzed the influence of the dimension of the event vector space on the prediction performance. At the same time, the experimental results show that the performance of LSTM based models is better than the traditional artificial feature extraction method. The bidirectional LSTM is better than standard LSTM, because bidirectional LSTM can capture more feature information in different directions than standard LSTM. LSTM based on attention mechanism is better than standard BLSTM and BLSTM-Meanpool. At last, experimental results indicate that the performance of model with the operation behavior data outperforms model without the operation behavior data.

V. CONCLUSION

Based on the user operation behavior data of the P2P lending industry, we proposed a consumer credit scoring method based on the attention mechanism LSTM. This is a novel application of deep learning methods. This method is evaluated on real dataset and the research results show that the consumer credit scoring method based on the attention mechanism LSTM proposed in this paper has obvious improvement effect compared with the traditional artificial feature extraction method and the standard LSTM model.

In this study, we only consider the sequence of events in the online user operation behavior data without considering the time interval at which the event occurs. However, the time interval of the event also contains important information. In future research, we will continue to conduct in-depth research on this issue.

REFERENCES

[1] Y. Xia, X. Yang, and Y. Zhang, "A rejection inference technique based on contrastive pessimistic likelihood estimation for P2P lending," *Electron. Commerce Res. Appl.*, vol. 30, pp. 111–124, Jul./Aug. 2018.

[2] W. Li, S. Ding, Y. Chen, and S. Yang, "Heterogeneous ensemble for default prediction of peer-to-peer lending in China," *IEEE Access*, vol. 6, pp. 54396–54406, 2018.

[3] J. M. Tomczak and M. Zięba, "Classification restricted Boltzmann machine for comprehensible credit scoring model," *Expert Syst. Appl.*, vol. 42, pp. 1789–1796, Mar. 2015.

[4] L. Yu, Z. Yang, and L. Tang, "A novel multistage deep belief network based extreme learning machine ensemble learning paradigm for credit risk assessment," *Flexible Services Manuf. J.*, vol. 28, no. 4, pp. 576–592, 2016.

[5] B. A. Z. Hidasi, M. Quadrana, A. Karatzoglou, and D. Tikk, "Parallel recurrent neural network architectures for feature-rich session-based recommendations," in *Proc. 10th ACM Conf. Recommender Syst.*, 2016, pp. 241–248.

[6] T. Lang and M. Rettenmeier, "Understanding consumer behavior with recurrent neural networks," in *Proc. Int. Workshop Mach. Learn. Methods Recommender Syst.*, 2017, pp. 1–8.

[7] S. Wang, C. Liu, X. Gao, H. Qu, and W. Xu, "Session-based fraud detection in online e-commerce transactions using recurrent neural networks," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2017, pp. 241–252.

[8] G. E. Hinton, "Learning distributed representations of concepts," in *Proc. 8th Annu. Conf. Cogn. Sci. Soc.*, Amherst, MA, USA, 1986, p. 12.

[9] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.

[10] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 641–648.

[11] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Feb. 2003.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[15] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[16] D. Bahdanau, K. Cho, and Y. Bengio. (2014). "Neural machine translation by jointly learning to align and translate." [Online]. Available: https://arxiv.org/abs/1409.0473

[17] X. Zhu, P. Sobhani, and H. Guo, "Long short-term memory over recursive structures," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1604–1612.

[18] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2017, pp. 1480–1489.

[19] Y. Xia, C. Liu, Y. Li, and N. Liu, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," *Expert Syst. Appl.*, vol. 78, pp. 225–241, Jul. 2017.

[20] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

CHONGREN WANG received the B.S. and M.S. degrees from the Department of Information Management and Information System, Shandong University of Finance and Economics, Jinan, China, in 2006 and 2013, respectively. He is currently pursuing the Ph.D. degree with the School of Information Management and Engineering, Shanghai University of Finance and Economics, China. His research interests include business intelligence, machine learning, and deep learning.

DONGMEI HAN received the Ph.D. degree in quantitative economics from Jilin University, China, in 2002. She is currently a Professor and the Doctoral Advisor with the School of Information Management and Engineering, Shanghai University of Finance and Economics, China. Her research interests include finance engineering and management science and engineering.

QIGANG LIU received the B.S. degree in computer science from Bohai University, Liaoning, China, in 2004, and the M.S. degree in network engineering from Shanghai Normal University, Shanghai, China, in 2009. He is currently pursuing the Ph.D. degree with the School of Information Management and Engineering, Shanghai University of Finance and Economics. Since 2009, he has been a Lecturer on the Faculty of the Sydney Institute of Language and Commerce, Shanghai University. His research interests include machine learning and computer vision. He has published more than six papers on these subjects. He is a Student Member of the IEEE. He was a recipient of the National Scholarship, in 2015.

SUYUAN LUO was born in Shenzhen, China. She received the double degrees from the Shanghai University of Finances and Economics, Shanghai, China, with a major in management and a minor in finance, in 2012. She is currently pursuing the Ph.D. degree in management science.

From 2015 to 2016, she was a Visiting Scholar with the Foster School of Business, University of Washington, Seattle, WA, USA. Her research interests include electronic commerce, machine learning in the financial services industry, and reputation systems in online business.