IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 January 2024, accepted 24 January 2024, date of publication 26 January 2024, date of current version 5 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3359052

RESEARCH ARTICLE

Prediction Method of O2O Coupon Based on Multi-Grained Attention Mechanism of CNN and Bi-GRU

LISHA YAO[1,2] AND MIDETH ABISADO[1], (Member, IEEE)
[1]College of Computing and Information Technologies, National University, Manila 1008, Philippines
[2]School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei, Anhui 230094, China

Corresponding author: Lisha Yao (jsjyaolisha@163.com)

This work was supported in part by the Key Research Project of Natural Science in Universities of Anhui Province under Grant KJ2020A0782, in part by the Key Scientific Research Project of Anhui Provincial Research Preparation Plan in 2023 under Grant 2023AH051806, in part by the University-Level Quality Engineering Demonstration Experiment and Training Center "Big Data Comprehensive Experiment and Training Center" under Grant 2020 sysxx01, and in part by the 2022 Anhui Province Quality Engineering Construction Project "Python Language Programming" Online and Offline Mixed Course under Grant 2022xsxx089.

ABSTRACT O2O (Online to Offline) can analyze users' behaviors according to data mining, realize personalized marketing and improve marketing effect. The "push" delivery method of O2O coupons ignores the active participation and user experience of users, and the pertinence and effectiveness of delivery are greatly affected. Aiming at the simple structure of a single network and mainly relying on artificial construction to extract features, in order to improve the utilization efficiency of deep-seated features in the model and effectively extract multi-level features, this paper introduces Bi-GRU according to the time-series characteristics of O2O consumption behavior, and proposes a new multi-grained attention mechanism. First, build a multi-dimensional consumer behavior feature project; Secondly, using convolutional neural network (CNN) and Bi-GRU to extract local and global features; Finally, multi-level and multi-grained information is extracted by using multi-grained attention to avoid the loss of hierarchical structure information, enrich feature vectors and further improve model performance. Using real O2O coupons and data sets, the CB-MA model proposed in this paper achieves 93.29% accuracy, 91.72% AUC and 0.0332 Loss. The results show that multi-grained attention mechanism can extract multi-level features more effectively than traditional attention mechanism. At the same time, CNN and Bi-GRU are combined to learn local and global features at the same time, and the correlation information of time and space is extracted.

INDEX TERMS Attention mechanisms, Bi-GRU architecture, Bi-GRU training, customer purchase patterns, convolutional neural network, multi-grained attention mechanism, O2O coupon prediction, optimization algorithms, redemption prediction, retail analytics.

I. INTRODUCTION

In the context of e-commerce, online shopping has become an important way of consumption. With the rapid development of the Internet, O2O (Online to Offline) consumption has become an important way of consumption. As a new Internet business model, O2O drives offline business and offline consumption through online marketing and online purchases and integrates the traditional business model with the Internet business model, combining "online payment and offline consumption experience" into different industries. This new model has realized the renewal of people's way of life and entertainment, and users trade more conveniently. O2O develops offline customers by personalized pushing coupons and merchant information to potential consumers.

Coupons, as a common consumer marketing tool, can increase customer loyalty or attract new customers. Among them, electronic coupons' production and dissemination cost

The associate editor coordinating the review of this manuscript and approving it for publication was Wu-Shiung Feng.

is low and the dissemination effect can be accurately quantified. It has been widely used in different industries, such as fitness, catering, and entertainment.

The traditional analysis of user coupon usage behavior mainly uses the methods of designing questionnaires, constructing statistics, and hypothesis testing to explore the various factors that affect coupon cancellation. In the big data environment of information explosion, it is difficult for traditional analysis methods to mine valuable information from massive data, mainly because: (1) When the amount of data increases sharply, the information obtained by means of manual feature acquisition such as designing questionnaires is inferior to that obtained by direct use of real transaction data in terms of quantity and quality; (2) It is difficult for a simple statistical model to explore the factors affecting the use and cancellation of coupons from multiple angles. Inspired by the success of big data, machine learning, and deep learning methods in the research fields of e-commerce recommendation systems and user purchase behavior prediction, as well as the application of deep learning in various fields of artificial intelligence, this paper takes coupon behavior prediction in e-commerce environment as the goal and applies machine learning and deep learning methods to solve this problem.

The prediction of O2O coupon usage adopts rule-based or traditional machine-learning methods. Such methods need to construct a large number of rules to extract text features and rely heavily on manual experience and knowledge in related fields. On the other hand, with the increase of training data, the limitations of the method itself, such as over-fitting and poor generalization ability, are more prominent. Using different types of deep neural networks, deep learning can automatically extract features from large-scale data and express learning and is very good at expressing the internal structure of large-scale complex data. Although deep learning can express the internal results of complex data, the data structure of O2O coupon consumption behavior is complex, with many influencing factors and strong time dependence, which challenges the accurate prediction of consumption behavior.

The research motivation is to extract the potential deep features effectively and accurately in the O2O coupon consumption behavior data, fully mine the time-influencing factors in the data, and optimize the prediction model to pay attention to the fine-grained features of sequence features. Attention mechanisms can focus on key features. However, the traditional attention mechanism lacks different levels of expression, leading to the problem of distraction.

The following is a summary of this paper's primary innovations and contributions:

(1) A CB-MA model based on CNN with multi-granularity input and combined with attention mechanism is proposed for O2O coupon consumption behavior prediction. The convolution part uses the convolution filter to extract the features of different levels, which enhances the ability of the model to capture the hidden local context features between the data. At the same time, multi-granularity features of original input data can be extracted based on the CNN model, such as useful features extracted from the description of coupons, historical usage and other relevant factors. In this way, the model is able to understand the various factors that influence coupon use, thus making the predictions more interpretable.

(2) Bi-gated cycle Unit (Bi-GRU) is introduced to propose a new multi-grained attention mechanism. Dynamic extraction of important contexts through local attention mechanisms maximizes information acquisition and suppresses irrelevant contexts. The global attention mechanism can extract context-related information of different subspaces to further improve feature recognition. Bi-GRU can capture patterns of coupon usage over time. Therefore, the model can take into account the influence of historical usage on future use, which not only improves the predictive accuracy of the model, but also makes the learned features more interpretable.

II. RELEVANT RESEARCH

O2O is a new e-commerce model of multi-channel marketing, which focuses on online promotions through platforms to increase sales in physical stores [1]. Scholars, both local and international have focused a great deal of emphasis on this new e-commerce model that integrates online and offline. As an important O2O marketing tool, coupons can revitalize old customers and attract new customers. Boone and Kurtz points out that coupons can sustain repeated purchases [2]. Jung and Lee pointed out the feature that coupons can stimulate consumption [3]. Paul et al. proposed to use coupon value, coupon quantity, brand loyalty, and net price range (NPR) as the driving factors of coupon use, divide consumer groups into brand-centered users and price-incentive-based users, and achieve user precision marketing through user segmentation, NPR and loyalty analysis [4]. Ma, based on the O2O model, the threshold of online coupons and the offline distance between consumers and merchants will affect consumers' consumption behavior and put forward corresponding suggestions for the design and distribution of O2O coupons [5]. Wan et al. discovered that consumers consider the perceived ease of use of coupons when making judgments about what to buy in addition to the discount rate [6]. Shi and Ji pointed out that e-coupons can realize the transfer of consumers from online to offline channels, and retailers can expand the market scale and make profits under certain conditions [7].

The prediction of specific users' purchasing behavior of specific goods mainly relies on the massive user purchasing behavior data of e-commerce platforms. The machine learning model is constructed to predict consumers' behavior by mining the specific user and product characteristics in the data. Yi et al. [8] constructed a random forest model to optimize the sampling method to solve the problem of unbalanced prediction samples in purchasing behaviors. Li et al. [9] constructed a feature combination by considering seven features, including commodity, user, location, and portfolio, and

predicted purchase behavior through an integrated gradient ascending decision tree (GBDT) model. Liu et al. [10] constructed a model to predict repeat purchase behavior and won the championship in the "Repeat Purchase Prediction" competition held by IJCAI in 2015. In 2016, Wepon [11] extracted the characteristics of merchants, users and coupons, and weighted fusion of probability prediction values of GBDT and XGBoost model was used to improve THE AUC value by 10%.In 2017, AaronChou [12] extracted the data of 7 days, 3 days and 1 day before and after the day users received coupons in the test set, and conducted model fusion based on rank method and GBDT, RandomForest and LR algorithms. In 2018, Liu constructed a feature project for users, merchants and coupons and completed the personalized delivery of coupons through the integrated learning Catboost algorithm, which is more accurate than other traditional algorithms [13]. In 2019, Xu et al. studied the targeted delivery of new retail coupons, used XGBoost algorithm to predict users' behavior of using coupons, and determined the characteristics with high contribution to consumers' use decision through k-folding cross-validation. Compared with random forest and GBDT algorithms, the superiority of XGBoost integrated learning algorithm was proved [14]. In 2019, Zhang et al. combined the three decision ideas with XGBoost integrated learning algorithm to consider the misclassification cost and learning cost, effectively improving the accuracy of prediction [15]. In 2021, Xiao et al. [16] used graph neural networks to predict the exchange rate of commodity vouchers, which is helpful to understand users' exchange preferences for vouchers.

The prediction of O2O coupon usage mainly adopts the method based on traditional machine learning. However, it necessitates a large number of artificially created data features, making it unable to understand the text's deep information elements more effectively. In order to compensate for the shortcomings of conventional approaches, this study uses deep learning to automatically finish feature extraction and data table features. Convolution neural network (CNN) model [17] and recurrent neural network (RNN) model [18] are currently the most commonly utilized network models. However, CNN is easy to lose important information in the convolution process, and can't make full use of the information correlation between networks at the same level, and lacks the ability to learn sequence characteristics. RNN is prone to gradient disappearance and lacks long-term correlation learning. Long Short Term Memory (LSTM) can solve the problem of RNN [19]. At the same time, the gated recurrent unit (GRU) has the advantages of fewer parameters and faster operation speed than the long LSTM [20]. The traditional network is indiscriminate feature extraction of data, which can not identify important information, so attention mechanism is introduced to solve the above problems [21]. However, the traditional attention mechanism lacks the expression of different levels of characteristics, leading to distraction. In order to eliminate noise interference and enhance the multi-granularity attention ability of the model, considering the deep feature extraction at local and global levels, this paper proposes an O2O coupon prediction method combining CNN and Bi-GRU with a multi-granularity attention mechanism. According to the temporal characteristics of O2O consumption behavior, Bi-GRU is introduced to learn the past and future bidirectional long-distance dependence information. The attention mechanism can focus on the key information of the model, but it ignores the deep features and contextual features in the information, and has limited ability to characterize and generalize the data. Therefore, this paper proposes a multi-grained attention mechanism of information. In this way, the internal structural characteristics of the data can be obtained from the local and global levels, and the local and global information can be extracted jointly with CNN and Bi-GRU to improve the description and identification ability of O2O consumption behavior.

III. DATA PROCESSING AND FEATURE ENGINEERING

A. DATA EXPLORATION

This paper used Alibaba company's O2O users' online consumption data at https://tianchi.aliyun.com/competition/entrance/231593/rankingList. The data contains both online and offline data sets. The data set provides users' online and offline consumption behaviors between January 1, 2016 and June 30, 2016. Table 1 shows the definitions of each attribute in the data set.

TABLE 1. Data feature description.

Attributes	Type	Explain
User_id	Discrete	User ID
Merchant_id	Discrete	Merchant ID
Coupon_id	Discrete	Coupon ID
Action	Discrete	User actions:0: click;1: Purchase;2: Get the coupon.
Discount_rate	Continuous	Discount rate:x∈[0,1] indicates the discount rate; x:y: means full x minus y, in units of yuan.
Distance	Continuous	Distance between consumer and business
Date_received	Continuous	Coupon collection date
Date	Continuous	Date of consumption

(1) Offline data set: Each consumption record contains a user ID, merchant ID, coupon ID, discount rate, the distance between user and merchant, coupon collection date, and consumption date attribute.

(2) Online data set: Each consumption record contains a user ID, merchant ID, user behavior (click, purchase, coupon), coupon ID, discount rate, coupon collection date, and consumption date attributes.

The offline data included 1,754,848 user behavior data, and the online data included 11,429,826 user behavior data. The offline data involved 539,483 users, 8,414 merchants, and 9,738 coupons, while the online data involved 762,858 users, 7,999 merchants, and 27,747 coupons.

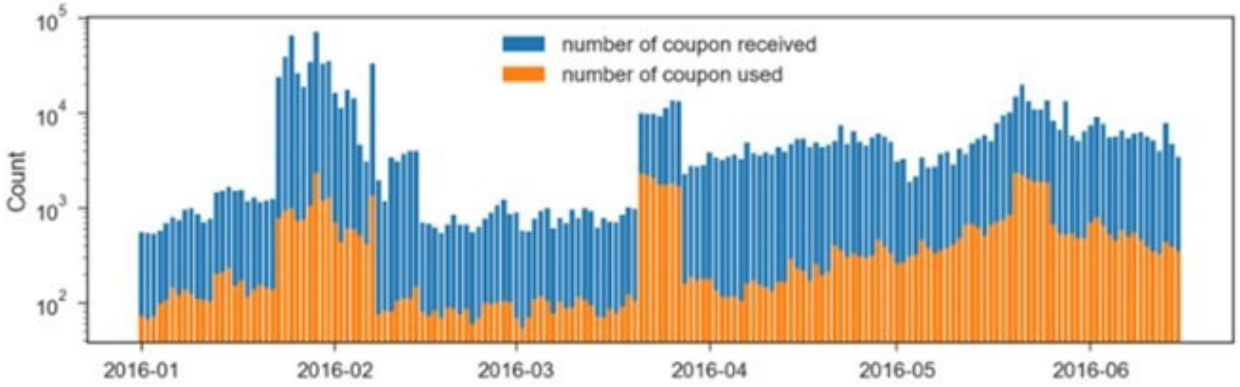

FIGURE 1. Time distribution of the number of coupons issued and used.

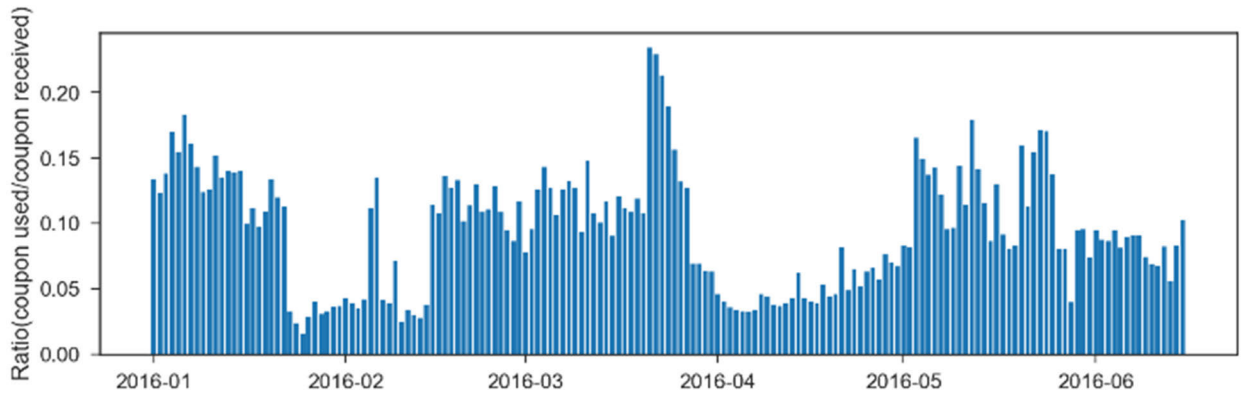

FIGURE 2. Time distribution of coupon usage.

This dataset contains actual consumption behavior from January 1, 2016 to June 30, 2016. The attention mechanism observes the model's salient qualities but ignores the data's contextual and deep elements. FIGURE 1 is a distribution diagram of the amount of coupons issued and used. FIGURE 2 shows the usage rate of corresponding coupons. From the graph analysis, it can be seen that the utilization rate of O2O coupons is mostly within 20%. From the overall distribution, the utilization rate of O2O coupons is greatly affected by time. For example, around February 2016, the corresponding time period is the Spring Festival holiday, and the amount of coupons is the highest, but the utilization rate of coupons is the lowest. Secondly, around March 20th, when the amount of coupons is normal, the utilization rate of coupons is the highest. Through the visual analysis of the overall user consumption characteristics, we can find the hidden information in the data, which provides the basis for the subsequent analysis.

B. FEATURE EXTRACTION AND CONSTRUCTION

Because there are too few features of data, only the basic description of data can be given, and the problem cannot be described comprehensively and accurately. It is necessary to construct new and more effective features according to business logic and experience to express the problem in depth. Feature construction is to transform the original features of training data into effective features of model construction, and its purpose is to extract valuable information hidden behind the data, obtain effective features better, and improve the performance of the training model.

The following briefly introduces the main results of feature construction:

(1) Characteristics of user

- The frequency with which customers obtain coupons;
- The frequency with which they receive coupons but do not utilize them;
- The frequency with which they write off coupons;
- Write-off rate subsequent to coupon recipients;
- The percentage of user-cancelled coupons that have complete refunds among all cancelled coupons;
- Average consumption discount rate of coupons written off by users.

(2) Characteristics of merchants

- The number of times the merchant's coupons were collected;
- The write-off rate after the coupon of the merchant is received;

- The average discount rate of coupons written off by merchants;
- The maximum discount rate of the merchant's coupon verification;
- The minimum discount rate of the merchant's coupon verification.

(3) Characteristics of coupon

- Discount rate of coupons;
- The quantity of times the user has previously accepted the coupon;
- The number of times the user used the coupon in the past;
- The user's write-off rate of this coupon in the past.

IV. MULTI-GRAINED ATTENTION PREDICTION MODEL COMBINING CNN AND BI-GRU

A. DATA EXPLORATION

Using CNN can increase generalization ability, minimize parameters, mitigate over-fitting, and extract local continuous information. Convolution, pooling, and fully connected layers make up the fundamental building blocks of the CNN model. The output from one layer is the input for the layer that follows it. The convolution layer, that is, convolution operation, completes the extraction of input features through the set convolution kernel, and obtains the feature graph c. The specific expression formula is:

$$c_i = f(w \cdot x_{i+h-1} + b) \tag{1}$$

where $\boldsymbol{w}$ is the convolution kernel with height $\boldsymbol{h}$ and width $\boldsymbol{d}$. $\boldsymbol{b}$ is the offset vector and $\boldsymbol{f}$ is the activation function ReLu.

Pool layer, that is, using pool operation to simplify convolution layer information. Generally, the maximum pooling operation is adopted, and its expression is:

$$\hat{c} = max(c) \tag{2}$$

Fully connected layer, that is, the extracted local features are fused into global features, and finally the probability is calculated by using classification function.

In addition to extracting local continuous features and reducing parameters and over-fitting, CNN can also enhance generalization ability. In the O2O coupon consumption behavior data, the data are all time series data. The choice between one-dimensional CNN and two-dimensional CNN depends on the nature of the data under consideration. 1D CNN is primarily applied to time-series data, where the kernel moves in a single direction, resulting in both input and output data being two-dimensional. This work processes input data using a one-dimensional convolution layer to extract data attributes more rapidly and correctly for consumption prediction. The one-dimensional convolution layer is similar to the two-dimensional convolution layer, but the difference is that both convolution layer and convolution kernel are one-dimensional. One-dimensional CNN requires fewer training cycles and is easier to train than two-dimensional CNN for efficient feature extraction.

B. BI-GRU

RNN is a neural network specially used for time series data processing. Unlike other neural networks, RNN contains a memory mechanism in addition to taking into account the input from the previous instant. Unfortunately, during the training phase, RNN is prone to gradient expansion and disappearance, which makes it impossible to transmit gradients in a lengthy sequence. The gated cycle unit (GRU) was developed to address this issue. Its unique control learning mechanism makes it one of the best network topologies for handling time series data by resolving the gradient problem of RNN.

GRU and LSTM are both commonly used gating algorithms of cyclic neural networks. Compared with the structure of LSTM, GRU uses the linear relationship between hidden state and candidate hidden state instead of the complex connection between cell state and hidden state in LSTM, which is simple in structure and easier to calculate and train. The structure of GRU control information transmission is also relatively simple, mainly consisting of update gate and reset gate. The reset gate determines the ratio of the hidden state information from the previous moment to the current candidate hidden state, while the update gate determines the ratio of the hidden state information from the previous moment to the current hidden state and the current candidate hidden state information to the current hidden state. At a certain time, t, the update formulas of GRU's hidden state, update gate, standby hidden state and reset gate are as follows:

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t \tag{3}$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{4}$$

$$\tilde{h}_t = tanh(W_h x_t + r_t \cdot (U_h h_{t-1}) + b_h) \tag{5}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{6}$$

where $h_t, \tilde{h}_t$ and h_{t-1} represent the hidden state and candidate hidden state at time t and the hidden state at time $t-1$. x_t represents the input vector. z_t and r_t represent the update gate and reset gate, and $tanh$ and σ represent two nonlinear activation functions tanh and sigmoid, respectively. $W_z, W_h, W_r, U_z, U_h, U_r$ represent the deviation parameters corresponding to GRU neurons. $*$ represents Hadamard product.

Considering that different context compositions may also cause different feature expressions, Bi-GRU is used for feature extraction, and the relationship between contexts is fully considered. Bi-GRU network is composed of a reverse GRU network and an output state connection layer of a forward and backward CRU network. FIGURE 3 displays the model structural diagram.

C. MULTI-GRAINED ATTENTION

The feature vector is encoded by Bi-GRU (Deep Bidirectional Gated Cyclic Unit), and the local context features are extracted to obtain the hidden layer representation with context information. In order to extract hierarchical features with different granularity to a greater extent, this paper combines Bi-GRU network with multi-granularity attention network to process local and global multi-granularity

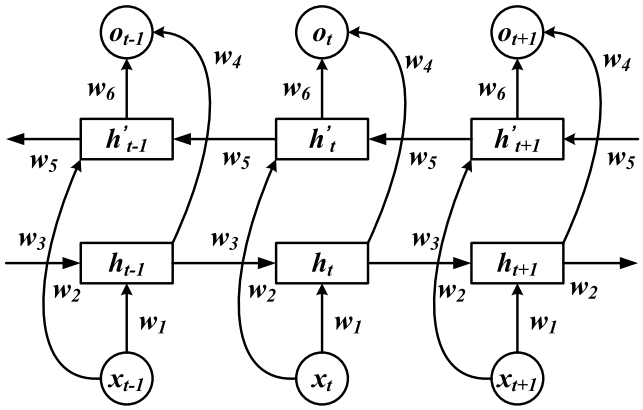

FIGURE 3. Bi-GRU model.

feature information. The Multi-grained attention module is mainly divided into local encoder layer, local attention layer, global encoder layer and global attention layer. The structure diagram of multi-grained attention module is shown in FIGURE 4.

1) Local encoder layer: First, the GRU of the local encoder layer is used for vectorization feature extraction to obtain the hidden layer feature vector h_{it}. The calculation process is as follows:

$$\vec{h}_{it} = G\vec{R}U(x_{it}), t \in [1, T] \tag{7}$$

$$\overleftarrow{h}_{it} = G\overleftarrow{R}U(x_{it}), t \in [T, 1] \tag{8}$$

$$h_{it} = [\vec{h}_{it}, \overleftarrow{h}_{it}] \tag{9}$$

$\vec{h}_{it}$ and $\overleftarrow{h}_{it}$ indicate the forward and backward hiding state of local feature w_{it}, respectively. T represents the number of local features.

2) Local attention layer: h_{it} uses the local attention layer to obtain local feature weights. The calculation process is as follows:

$$u_{it} = tanh(W_w h_{it} + b_w) \tag{10}$$

$$a_{it} = softmax\,(u_w, u_{it}) \tag{11}$$

$$s_i = \sum\nolimits_t a_{it} h_{it} \tag{12}$$

where u_{it} is a nonlinear transformation of the eigenvector h_{it} obtained by GRU through the tanh function. a_{it} is the local attention weight factor. W_w, u_w and b_w are the corresponding weight parameters and deviations of the local vector layer respectively. u_w is a randomly initialized context vector.

3) Global encoder layer: Encodes the global feature s_i output by the local attention layer using the same operations as the local encoder layer. s_i then input the global encoder layer GRU for depth feature extraction to obtain the global hidden layer vector h_i. L represents the number of global features. The calculation process is as follows:

$$\vec{h}_i = G\vec{R}U(s_i), i \in [1, L] \tag{13}$$

$$\overleftarrow{h}_i = G\overleftarrow{R}U(s_i), i \in [L, 1] \tag{14}$$

$$h_{it} = [\vec{h}_i, \overleftarrow{h}_i] \tag{15}$$

4) Global attention layer: h_i obtains the final feature d through the global attention layer. The calculation process is as follows:

$$u_i = tanh(W_s h_i + b_s) \tag{16}$$

$$a_i = softmax(u_s, u_i) \tag{17}$$

$$d = \sum\nolimits_i a_i h_i \tag{18}$$

where, W_s, u_s and b_s represent the weight parameters and deviation corresponding to the global attention layer respectively.

D. CB-MA MODEL

The use behavior of O2O coupons has a hierarchical structure. Considering that different levels in the data contain different hierarchical information and the structural information is highly dependent on the context, it plays different roles. Therefore, to further express the structural information of O2O consumption behavior, this paper introduces the multi-grained attention mechanism and proposes the CB-MA model in combination with CNN and Bi-GRU model to capture the internal structural information from the global and local levels. In this paper, multi-grained attention is used to present the key local and global internal structure information in information, and it is used to capture the correlation between contexts, instead of simply filtering feature sequences through context information to obtain global information. The model structure diagram is shown in FIGURE 5. The CB-MA model is mainly divided into three parts: CNN module, multi-grained attention module, and Softmax classifier module. The input data learns spatial local features through four layers of CNNS of different sizes. Then, the multi-grained attention module is input to learn the multi-scale and multi-grained features. Finally, the prediction of O2O coupon usage is completed through Softmax classifier module.

1) CNN MODULE

Firstly, learn the spatial local characteristics of O2O coupon usage behavior through CNN module. Stacked multi-layer CNN has better feature learning ability than single-layer network. Therefore, this paper adopts the convolution layer stack of four convolution windows with different sizes. The features extracted by this network are better than the traditional convolutional neural network. The first two layers adopt one-dimensional convolution with 64 convolution kernels of 7 × 7, and the last two layers adopt one-dimensional convolution with 128 convolution kernels of 10 × 10. Then we use the dropout mechanism to prevent over-fitting, and the parameter setting is 0. 5. The activation function adopts ReLu. The structure diagram of CNN module is shown in FIGURE 6.

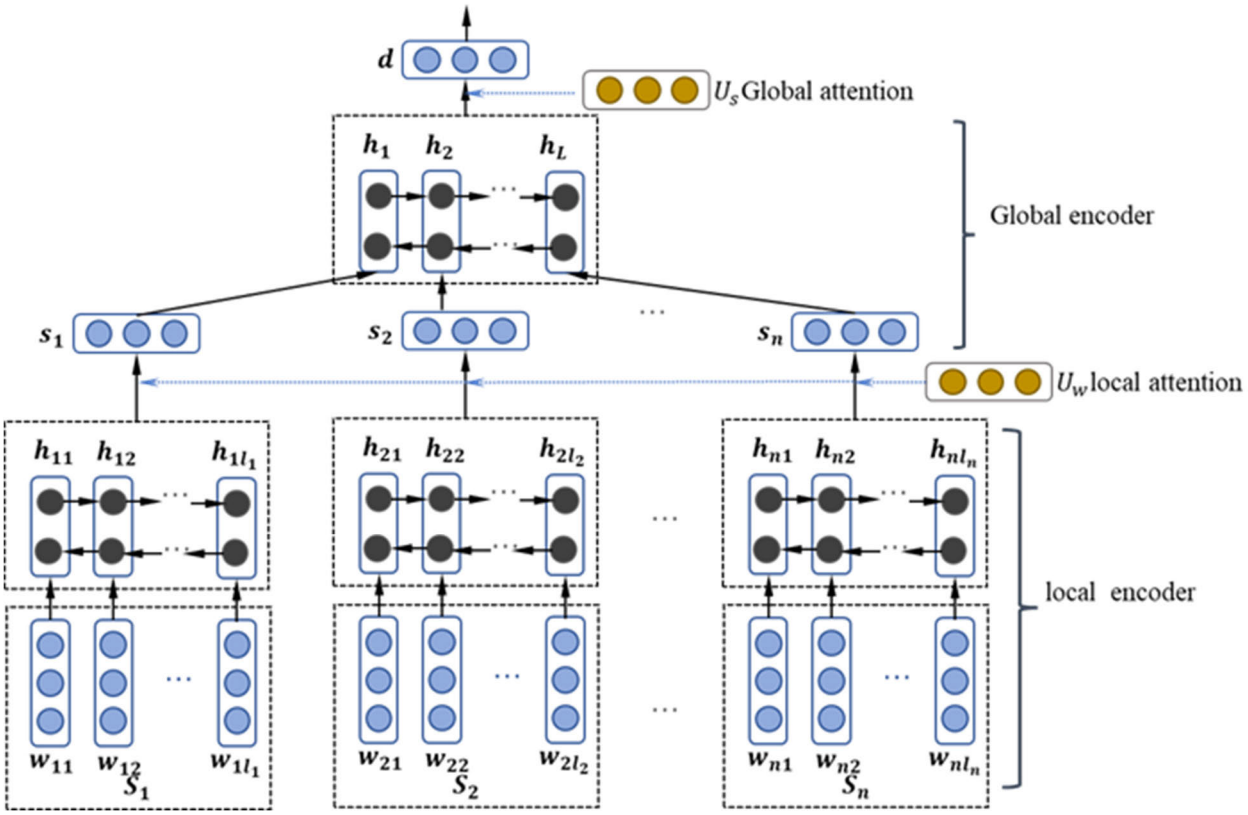

FIGURE 4. Multi-grained attention module structure diagram.

2) MULTI-GRAINED ATTENTION MODULE

When a single CNN module extracts features, it will cause problems such as losing location information and not considering context dependence. However, the single Bi-GRU model lacks the learning ability of long-term correlation. Therefore, the multi-grained attention module is proposed. Bi-GRU extends it in time to extract the global features of the data. Secondly, it is considered that local features in data play a key role in the representation of global features. In this paper, local ''attention'' is used to extract the key information of local features in global features, and the weight of local features in global features is calculated. Then the local features are weighted and merged. At the same time, global ''attention'' is introduced to describe the representation of global features. The local Bi-GRU dimension of the first layer is set to 64. The local Bi-GRU dimension of the second layer is set to 32.

3) SOFTMAX CLASSIFIER

Finally, the possibility of each feature belonging to different categories is calculated by Softmax classifier, and the prediction of O2O coupon usage is completed. It can be expressed as follows: the multi-grained attention module's output is used as the classification layer's input, and the Softmax function is utilized to determine the likelihood that each piece of data falls into a distinct category.

$$p(y = k|D) = \frac{exp(w_k^T D + b_k)}{\sum_{k=1}^{n} exp(w_k^T D + b_k)} \tag{19}$$

where the layer's weight matrix and bias vector are denoted by the symbols w_k and b_k. Back propagation is used in the model training described in this paper to update parameters. The parameter settings are shown in TABLE 2.

TABLE 2. Network configuration.

Layer	Output Shape
Conv1D	64, 7×7
Conv1D	64, 7×7
Maxpool	3×3
Conv1D	128, 10×10
Conv1D	128, 10×10
Dropout	0.5
Bi-GRU1	64
Bi-GRU2	32

V. EXPERIMENTAL ANALYSIS

In order to verify the performance of the CB-MA model proposed in this paper, this paper adopts real online and offline consumption data of O2O users, and divides the data

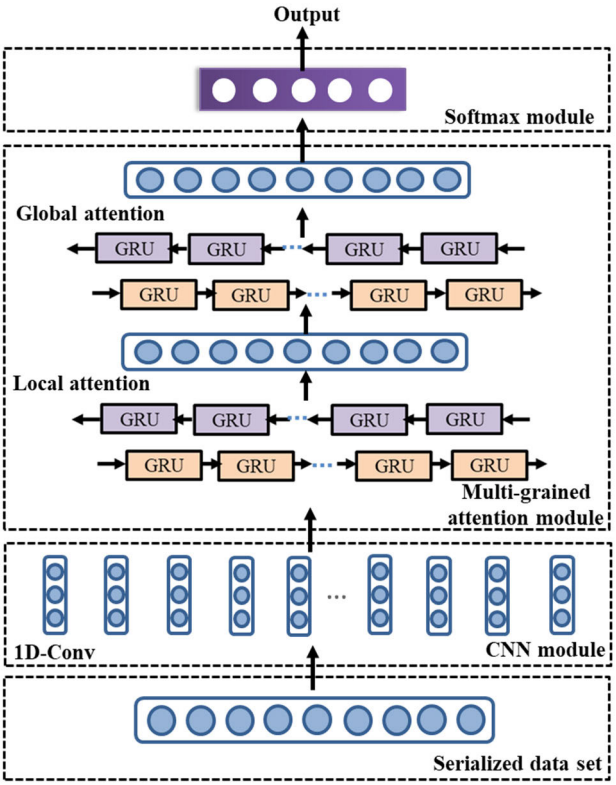

FIGURE 5. CB-MA model structure diagram.

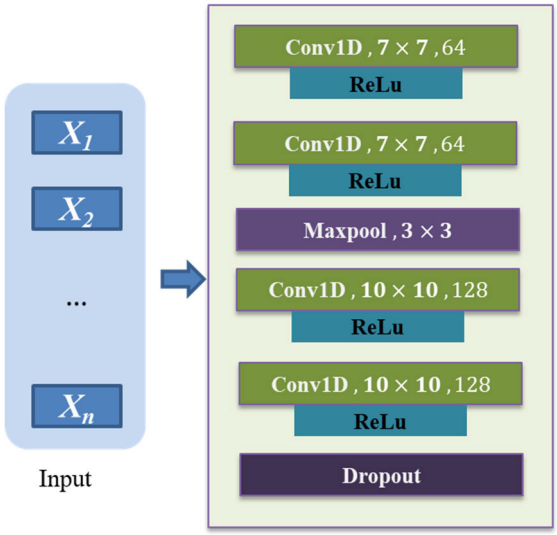

FIGURE 6. CNN module structure diagram.

set into training set, verification set and test set according to 8:1:1.

A. LEARNING RATE ANALYSIS

In order to find the optimal learning rate, CB-MA model is used to train data, and the dynamic learning rate is set to find the optimal learning rate that can fit the data. After training, the line graphs of loss value and accuracy changing with learning rate are obtained and visualized, as shown in FIGURE 7 and FIGURE 8. The abscissa in the figure is the Learning Rate, the ordinate in FIGURE 7 is the loss value, and the ordinate in FIGURE 8 is the acc value.

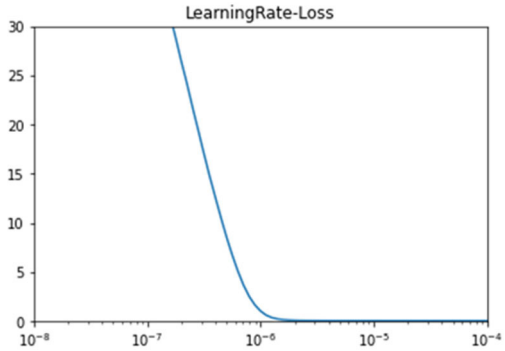

FIGURE 7. The influence of learning rate change on LOSS value.

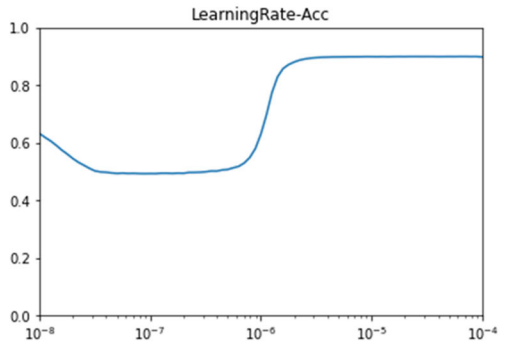

FIGURE 8. The influence of learning rate change on ACC value.

As can be seen from FIGURE 7 and FIGURE 8, when the learning rate is less than 10-6, the loss value decreases and the acc value increases with the increase of the learning rate, which shows that the model fitting effect becomes better and better with the increase of the learning rate. However, when the learning rate is greater than 10-6, the loss value and acc value have no obvious change. Therefore, the initial learning rate is set to 9e-5.

B. EXPERIMENTAL PARAMETER SETTING

The selection of optimization methods in model training has an important influence on model training. SGD and Adam are the most commonly used neural network optimization methods used in current research. SGD optimizer can be used for both classification calculation and regression calculation, which is more suitable for dealing with the learning problems of large-scale and sparse data, but its disadvantages are that it is easy to converge to the local optimal solution and strongly depends on the selection of learning rate. SGD and Adam optimizer are selected for training in the experiment, and the comparison is made based on the CNN module set

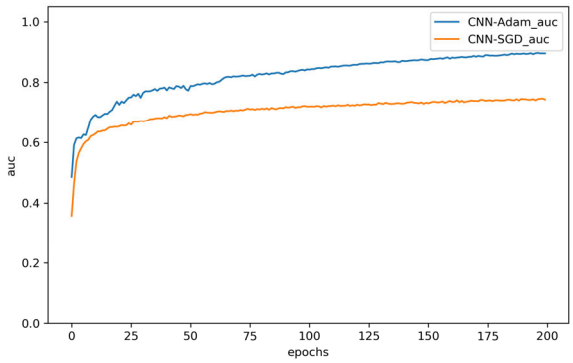

FIGURE 9. AUC comparison diagram of different optimizers.

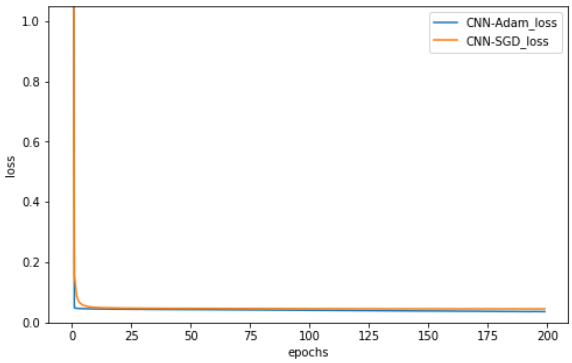

FIGURE 10. LOSS comparison diagram of different optimizers.

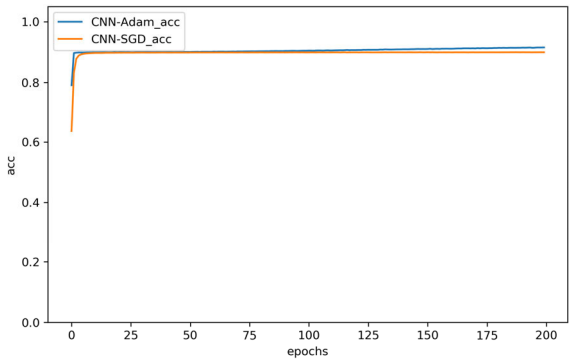

FIGURE 11. ACC comparison diagram of different optimizers.

TABLE 3. Comparison results of different optimizers.

Model	AUC	Loss	ACC
CNN-Adam	0.8957	0.0363	0.9148
CNN-SGD	0.7433	0.0454	0.8988

in this paper. Other parameters are set in the same way. FIGURE 9-11 shows the comparison curves of AUC, LOSS and ACC of two different optimizers based on CNN model.

The specific comparison results of the two optimizers based on CNN module are shown in TABLE 3.

The aforementioned findings demonstrate how the Adam optimization algorithm can determine the adaptive learning rate of each model parameter. Because of this algorithm's faster convergence and superior learning effect, the Adam optimization algorithm fits the model.

C. EXPERIMENTAL PARAMETER SETTING

The experimental parameter settings are shown in TABLE 4. The maximum number of rounds of network training is 200, and the batch size is 500. Using Adam optimization function to update weights, the initial learning rate is 9e-5, the impulse factor is 0.9, and the loss function is Huber function.

TABLE 4. Experimental parameters.

Epochs	200
Batch_size	500
Optimizer	Adam
Learning rate	9e-5
Momentum	0.9
Loss	Huber

D. EVALUATION STANDARD

In this paper, AUC (Area Under roc Curve) and Accuracy(Acc) are used as the evaluation indexes of the prediction model. The calculation method of AUC value is to assume that there are $(M + N)$ samples in total, in which there are M positive samples and N negative samples, so there are $M \times N$ sample pairs in total. The likelihood that the negative sample is anticipated to be false, which is recorded as 1, is less than the probability that the positive sample is predicted to be genuine. These counted values are accumulated, and finally the accumulated counted values are divided by $(M \times N)$, and the final result is the AUC value.

$$AUC = \frac{\sum_{i \in positiveclass} rank_i - \frac{M(1+M)}{2}}{M \times N} \tag{20}$$

where $rank_i$ represents the probability that i samples are positive samples. The greater the AUC value, the more effective the classification algorithm we choose, and the better the prediction effect of the classifier. To calculate the Acc value, it is necessary to consider the confusion matrix, which is shown in the following TABLE 5.

TABLE 5. Comparison results of different optimizers.

	Positive	Negative
True	True Positive(TP)	True Negative(TN)
False	False Positive(FP)	False Negative(FN)

Among them, the column (Positive, Negative) represents the predicted result of the model, while the row (True, False) represents the actual result. Acc represents the classifier's ability to judge the whole sample, that is, the proportion of the

number judged to be correct in the total number of samples. Its calculation formula is:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{21}$$

E. EXPERIMENTAL ANALYSIS

1) COMPARED WITH MACHINE LEARNING ALGORITHM

Currently, machine learning is the primary basis for O2O coupon prediction. The experiment compares the following conventional machine learning techniques:

① GBDT [9]: The feature combination is constructed by considering seven features such as product, user, location and combination, and the integrated gradient lifting decision tree (GBDT) model is used to predict O2O coupon usage;

② RandomForest [12]: RandomForest model is used to predict the use of O2O coupons;

③ GBDT+XGBoost [11]: Extract the characteristics of merchants, users and coupons, and use the weighted fusion of probability prediction values of GBDT and XGBoost models to predict the use of O2O coupons;

④ GBDT+ RandomForest +LR [12]: Based on rank method and model fusion of GBDT, RandomForest and LR algorithm;

⑤ XGBoost [15] : Combining three decision ideas with XGBoost integrated learning algorithm to consider misclassification cost and learning cost, effectively improve the accuracy of prediction and predict the use of O2O coupons.

AUC is used as the evaluation index. The result of AUC comparison is shown in FIGURE 12.

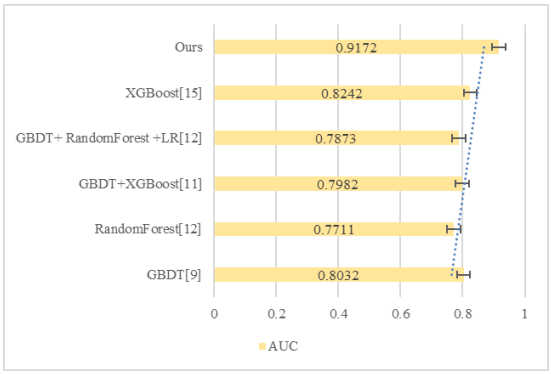

FIGURE 12. AUC comparison results.

At present, XGBoost, the best forecasting method, has achieved an AUC score of 0.8242. The AUC score of this model is 0.9172, which is 9.3% higher than the best XGBoost model.

2) ABLATION EXPERIMENT

In order to verify the validity of the CB-MA model proposed in this paper, the CB-MA model proposed in this paper is compared with the following models.

① CNN model: It adopts the same settings and parameters as the CNN module in the CB-MA model proposed in this paper.

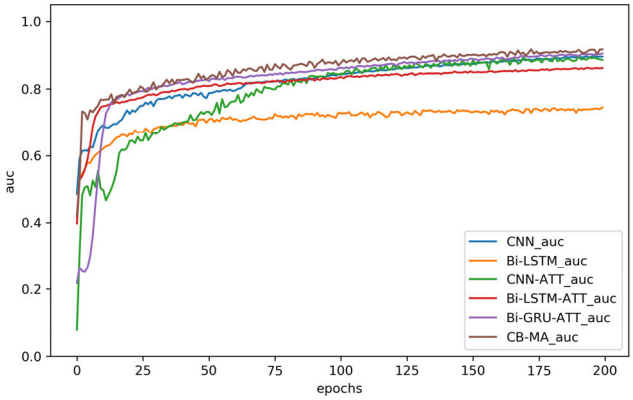

FIGURE 13. AUC comparison diagram of ablation experiment.

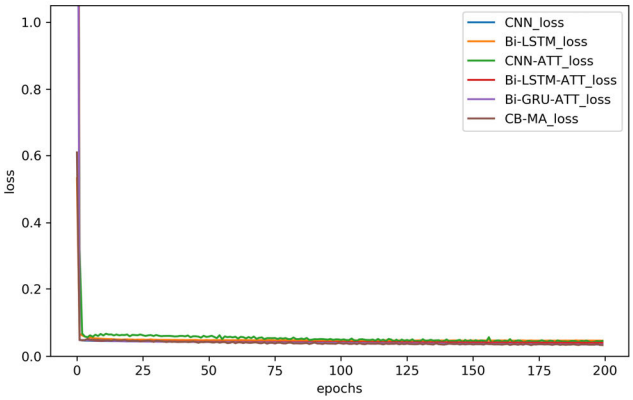

FIGURE 14. LOSS comparison diagram of ablation experiment.

TABLE 6. Comparative results of ablation experiments.

Model	AUC	LOSS	ACC
CNN	0.8957	0.0363	0.9148
Bi-LSTM	0.7456	0.0462	0.8980
CNN-ATT	0.8857	0.0455	0.9296
Bi-LSTM-ATT	0.8621	0.0401	0.9066
Bi-GRU-ATT	0.9048	0.0349	0.9189
CB-MA	**0.9172**	**0.0332**	**0.9329**

② Bi-LSTM model: A bidirectional LSTM model with 32 neurons in each layer.

③ CNN-ATT model: Add the traditional attention layer to the original CNN model.

④ Bi-LSTM-ATT model: Add the traditional attention layer to the original Bi-lSTM model.

⑤ Bi-GRU-ATT model: Bi-LSTM in Bi-LSTM-ATT model is replaced by Bi-GRU model with the same configuration.

⑥ CB-MA model: The model proposed in this paper.

The comparison curves of AUC, LOSS and ACC of the above models are shown in FIGURE 13-15. The specific contrast results of ablation experiments are shown in TABLE 6.

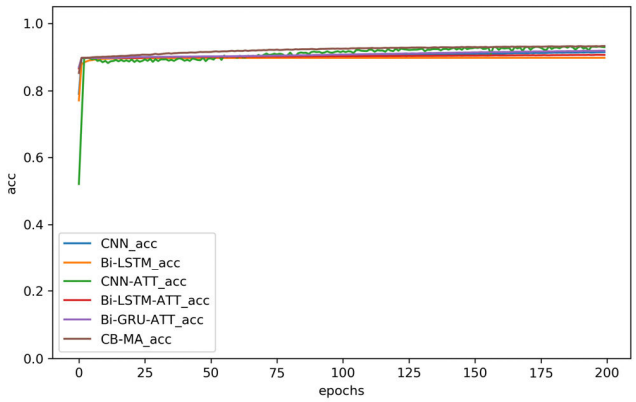

FIGURE 15. ACC comparison diagram of ablation experiment.

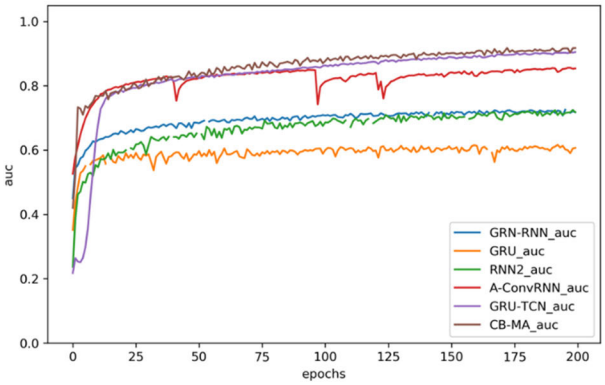

FIGURE 16. AUC comparison diagram of contrast experiment.

TABLE 7. Contrast experimental results.

Model	AUC	LOSS	ACC
RNN2	0.7177	0.0466	0.8988
GRU	0.6068	0.0485	0.8973
GRU-RNN	0.7176	0.0468	0.8976
A-ConvRNN	0.8538	0.0403	0.9057
GRU-TCN	0.9043	0.0349	0.9190
CB-MA	**0.9172**	**0.0332**	**0.9329**

The results show that compared with other composite models, CNN and Bi-LSTM, the two single models, all perform poorly. It shows that CB-MA model plays a complementary and positive role in combining CNN and Bi-GRU, enriching feature vectors, extracting more sufficient feature vectors and improving the accuracy of model sentiment analysis. CNN has a strong learning ability in feature learning, which can better extract deep-seated features. Through Bi-GRU, it can strengthen the learning of serialized information, thus optimizing the model and improving the accuracy of the model. Compared with CNN model, CNN-ATT with attention mechanism did not improve AUC and LOSS, but ACC increased by 1.48%. Compared with Bi-LSTM model, Bi-LSTM-ATT with attention mechanism increased AUC by 11.65%, LOSS decreased by 0.0061 and ACC increased by 0.86%. This shows that after abandoning the attention mechanism layer, although the feature fusion information is richer, there are a lot of redundancy and noise interference in the fusion information at different levels, which has a great impact on the final classification results. In terms of extracting the whole feature of data, the intentional mechanical system has more advantages than the traditional CNN or Bi-LSTM model. Compared with Bi-LSTM-ATT, Bi-GRU-ATT model uses Bi-LSTM network instead of Bi-LSTM. The AUC of the model increases by 4.27%, the LOSS decreases by 0.0052, and the ACC increases by 1.23%. It is proved that Bi-GRU has better ability to process time series features and capture more feature context information than Bi-LSTM network. Compared with CNN model, CNN-ATT model and Bi-GRU-ATT model, CB-MA model has better performance, with AUC increased by 2.15%, LOSS decreased by 0.0031 and ACC increased by 1.81%. This verifies the effectiveness of the multi-grained attention module. Adding the multi-grained attention module to obtain vector features considers multi-granularity information and deep features, which has higher accuracy than introducing the traditional attention model. Both the experimental results and the visualization results verify that the local attention mechanism is effective for the acquisition of adjacent information. Global attention can obtain global context information from the long-distance relationship between local features, which overcomes the limitations of recursive neural network.

3) CONTRAST EXPERIMENT

In order to further analyze the contribution of the multi-granularity attention mechanism combined with CNN and Bi-GRU proposed in this paper to the model performance, this paper sets up several comparison models, which are introduced as follows:

① RNN2 model: Time series prediction based on RNN [22]. In this paper, two layers of RNN are set.

② GRU model: StockNet model based on GRU [23]. This paper sets 2 layers of GRU.

③ GRU-RNN model: An improved stackable gated recurrent unit and recurrent neural network (GRU-RNN) model [24].

④ A-ConvRNN model: A recursive neural network prediction model combining convolutional neural network and attention mechanism [25].

⑤ GRU-TCN model: a prediction model combining GRU and Temporal convolutional network (TCN) [26].

The contrast curves of AUC, LOSS and ACC of different models are shown in FIGURE 16-18 respectively. The specific results are shown in TABLE 7. The above experimental parameters are set the same.

Through the above comparison, we can find that the model proposed in this paper has a significant improvement over other network models in AUC, LOSS and ACC, which proves the superiority of this model. In this paper, CNN is used to

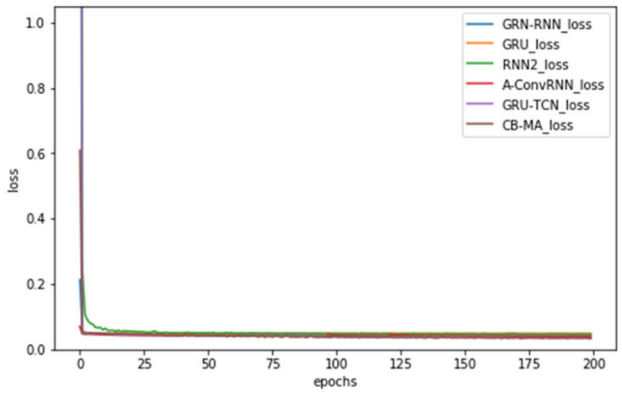

FIGURE 17. LOSS comparison diagram of contrast experiment.

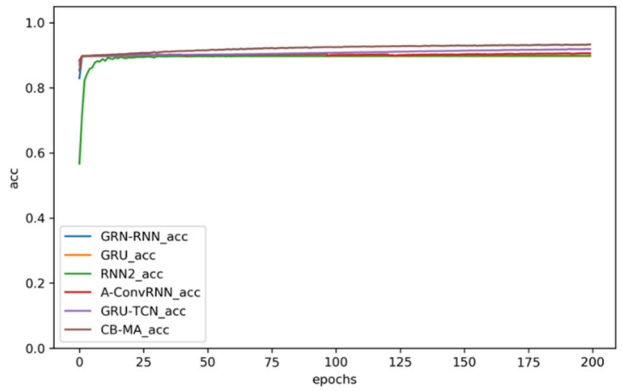

FIGURE 18. ACC comparison diagram of contrast experiment.

learn the deep-seated local characteristics of O2O coupon consumption behavior data, and Bi-GRU is used to learn the time series relationship of more dependent features from front to back, and more attention weight is given to key features through multi-grained attention layer.

VI. DISCUSSION

In this paper, a new CB-MA model is proposed by combining the multi-granularity attention mechanism of CNN and Bi-GRU to predict the use of O2O coupons. The assumptions of this method are:

1) Local features are extracted by CNN and global features are extracted by Bi-GRU to capture dependencies of sequence data. The combination of CNN and Bi-GRU can effectively capture temporal dependencies and local features in O2O coupon data.

2) A new multi-grained attention mechanism is proposed to extract multi-level and multi-grained information, thereby enhancing the accuracy of prediction.

The advantage of this method is that the combined use of CNN and Bi-GRU can better deal with complex nonlinear relationships and time series data, and the multi-granularity attention mechanism can better capture patterns and correlations at different granularity, providing more accurate predictions.

The difficulty with this approach is that:

1) Data sparsity: there may be high sparsity in the use of O2O coupons, which makes it difficult for the model to learn useful patterns.

2) Timing dependence: Coupon usage can be affected by time and other factors, and the model needs to be able to capture this timing dependence.

3) Feature selection: Coupon prediction requires selecting features that are relevant to the predicted outcome, which can be a challenge.

In the actual application scenario of O2O coupon prediction, this method faces the limitation that it is suitable for coupon issuance prediction on large-scale O2O platforms, but it may not be effective when the data is sparse or there are many outliers.

VII. CONCLUSION

In this work, a model to forecast the usage of O2O coupons is constructed using the deep learning method. In view of the problem that a single RNN cannot effectively extract deep-seated and hierarchical features, and traditional CNN cannot obtain the feature representation of sequential sentences, this paper proposes a multi-grained attention mechanism combining CNN and Bi-GRU network to solve the problem of O2O coupon use prediction. This method fully perceives the local features and global context information through CNN and GRU, and then combines the multi-grained attention mechanism to pay attention to key information with different granularity to learn consumption features, and extracts hierarchical features with different granularity to a greater extent, thus completing the use prediction of O2O coupons. On the O2O coupon data of Tianchi Contest on Alibaba Cloud platform, AUC score is 0.9172, LOSS value is 0.0332, and ACC reaches 93.29%. Compared with other methods, this method is significantly improved. The experiment demonstrates this method's superiority and effectiveness. The following step will continue to investigate the model in detail and take into consideration a finer-grained examination of the data because neural networks have complicated structural makeups. At the same time, other O2O coupon data were collected to test and strengthen the generalization ability of this model.

REFERENCES

[1] P. Yao, S. Osman, M. F. Sabri, and N. Zainudin, "Consumer behavior in online-to-offline (O2O) commerce: A thematic review," *Sustainability*, vol. 14, no. 13, p. 7842, Jun. 2022, doi: 10.3390/su14137842.

[2] L. E. Boone and D. L. Kurtz, *Contemporary Marketing*. HinsdaleIll, IL, USA: Dryden Press, 1989, pp. 89–93.

[3] K. Jung and B. Y. Lee, "Online vs. offline coupon redemption behaviors," *Int. Bus. Econ. Res. J.*, vol. 9, no. 12, pp. 23–36, Dec. 2010.

[4] P. Mills and C. Zamudio, "Scanning for discounts: Examining the redemption of competing mobile coupons," *J. Acad. Marketing Sci.*, vol. 46, no. 5, pp. 964–982, Sep. 2018.

[5] X. J. Ma, "Research on the impact of online coupons on consumer behavior-based on the analysis of O2O model," *Price, Theory Pract.*, vol. 2, pp. 117–120, Jun. 2019.

[6] Q. Wan, S. Yang, Y. Liao, and Y. Xia, "Group-buying coupons considering consumers' perceived ease of use," *Int. Trans. Oper. Res.*, vol. 27, no. 3, pp. 1638–1663, May 2020.
[7] W. R. Shi and Y. Q. Ji, "Channel integration strategy of retailers based on e-coupon from the perspective of mental accounting—A case study of Uniqlo," *Oper. Res. Manag. Sci.*, vol. 30, no. 3, pp. 137–143, 2021.
[8] Z. Yi, D. Wang, K. Hu, and Q. Li, "Purchase behavior prediction in M-commerce with an optimized sampling methods," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 1085–1092.
[9] D. Li, G. Zhao, Z. Wang, W. Ma, and Y. Liu, "A method of purchase prediction based on user behavior log," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 1031–1039.
[10] G. Liu, T. T. Nguyen, G. Zhao, W. Zha, J. Yang, J. Cao, M. Wu, P. Zhao, and W. Chen, "Repeat buyer prediction for e-commerce," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 155–164.
[11] Z.-F. Yan, Y.-L. Shen, W.-J. Liu, J.-M. Long, and Q. Wei, "An e-commerce coupon target population positioning model based on random forest and eXtreme gradient boosting," in *Proc. 11th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Beijing, China, Oct. 2018, pp. 1–5.
[12] AaronChou. *O2O Coupons Enable Pre-Testing [EB/OL].* Accessed: Mar. 2017. [Online]. Available: https://blog.csdn.net/shine19930820/article/details/53995369?spm=a2c4e.11153940.blogcont408173.7.48e97eb7Bxj1Xf
[13] J. S. Liu, "Personalized coupon delivery based on CatBoost algorithm," *Electron. World*, vol. 23, pp. 31–32, Jun. 2018.
[14] N. Xu and L. La, "On XGBoost-based prediction of new retail coupon usage behavior," *J. Southwest Normal Univ.*, vol. 44, no. 3, pp. 101–105, 2019.
[15] W. W. Zhang, D. Liu, and X. Y. Jia, "Three classified coupon prediction based on XGBoost algorithm," *J. Nanjing Univ. Aeronaut. Astronaut.*, vol. 51, no. 5, pp. 643–651, 2019.
[16] F. Xiao, L. Li, W. Xu, J. Zhao, X. Yang, J. Lang, and H. Wang, "DMBGN: Deep multi-behavior graph networks for voucher redemption rate prediction," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Singapore, Aug. 2021, pp. 3786–3794, doi: 10.1145/3447548.3467191.
[17] E. Elbasani and J.-D. Kim, "AMR-CNN: Abstract meaning representation with convolution neural network for toxic content detection," *J. Web Eng.*, vol. 21, no. 3, pp. 677–692, Feb. 2022.
[18] A. Datta, D. J. Wu, W. Zhu, M. Cai, and W. L. Ellsworth, "DeepShake: Shaking intensity prediction using deep spatiotemporal RNNs for earthquake early warning," *Seismol. Res. Lett.*, vol. 93, no. 3, pp. 1636–1649, May 2022.
[19] C. Liu, Y. Zhang, J. Sun, Z. Cui, and K. Wang, "Stacked bidirectional LSTM RNN to evaluate the remaining useful life of supercapacitor," *Int. J. Energy Res.*, vol. 46, no. 3, pp. 3034–3043, Mar. 2022.
[20] Z. Sun, Y. Hu, W. Li, S. Feng, and L. Pei, "Prediction model for short-term traffic flow based on a K-means-gated recurrent unit combination," *IET Intell. Transp. Syst.*, vol. 16, no. 5, pp. 675–690, May 2022.
[21] S. Li, J. Cao, J. Yao, J. Zhu, X. He, and Q. Jiang, "Adaptive aggregation with self-attention network for gastrointestinal image classification," *IET Image Process.*, vol. 16, no. 9, pp. 2384–2397, Jul. 2022.
[22] I. Amalou, N. Mouhni, and A. Abdali, "Multivariate time series prediction by RNN architectures for energy consumption forecasting," *Energy Rep.*, vol. 8, no. 9, pp. 1084–1091, doi: 10.1016/j.egyr.2022.07.139.
[23] U. Gupta, V. Bhattacharjee, and P. S. Bishnu, "StockNet—GRU based stock index prediction," *Expert Syst. Appl.*, vol. 207, Nov. 2022, Art. no. 117986.
[24] M. Xia, H. Shao, X. Ma, and C. W. de Silva, "A stacked GRU-RNN-based approach for predicting renewable energy and electricity load for smart grid operation," *IEEE Trans. Ind. Informat.*, vol. 17, no. 10, pp. 7050–7059, Oct. 2021, doi: 10.1109/TII.2021.3056867.
[25] H. Wang and R. Tian, "A-ConvRNN: A prediction model for e-commerce page views based on convolutional neural network and attention mechanism," in *Proc. IEEE 3rd Int. Conf. Electron. Technol., Commun. Inf. (ICETCI)*, Qingdao, China, May 2023, pp. 823–826.
[26] J. Jeon, S. Baek, B. Jeong, and Y. S. Jeong, "Early prediction of ransomware API calls behaviour based on GRU-TCN in healthcare IoT," *Connection Sci.*, vol. 35, no. 1, Art. no. 2233716, doi: 10.1080/09540091.2023.2233716.

LISHA YAO was born in Anhui, China, in 1986. She received the master's degree in applied computer technology from Anhui University, in 2011. She is currently pursuing the Ph.D. degree in computer science with National University, Philippines, in 2020.

She has been a Teacher with Anhui Xinhua University, since 2011. She is also an Associate Professor. She presided over six scientific research projects, published more than 20 papers in domestic and foreign academic journals and international conferences, and obtained one national invention patent.

MIDETH ABISADO (Member, IEEE) received the master's degree in information technology from the Technological University of the Philippines, in 2004, the Master of Science degree in computer science from Mapúa University, in 2016, and the Ph.D. degree in information technology from the Technological Institute of the Philippines, in 2019. She was with the Artificial Intelligence Research, with a particular interest in affecting computing, natural language processing, and image processing. She is currently a Professor with the Graduate School Department, College of Computing and Information Technologies, National University, Manila. She is also the Principal Investigator of a major research project on Philippine AI powered disease surveillance using social media analytics, funded by the Philippine Department of Science and Technology. She has contributed to over 50 research papers in various respected journals and conferences. She is a National Board Member of the Philippine Computing Society Special Interest Group of Women in Computing.