Research on prediction of user advertising recommendation system based on NFM model

Yujia Chen[1,#], Siben Li[2,#], Ting Jiang[3,#]

[1]School of Economics and Management, Communication University of China, Beijing, China, 100024
[2]School of Data Science and Media Intelligence, Communication University of China, Beijing, China, 100024
[3]School of Computer and Cyber Sciences, Communication University of China, Beijing, China, 100024
[#]These authors contributed equally.

Abstract: *Click-through rate (CTR) predictions can impact business revenue and improve user experience. In large-scale enterprise-level advertising systems and recommendation systems, CTR prediction is a very important link, and it has always been one of the key issues of academic research. In the case of extremely unbalanced positive and negative samples, this paper uses as few positive samples as possible to predict the user's advertising click behavior. To solve the above problems, on the basis of data processing, this paper proposes a CTR prediction model that combines feature engineering and machine learning. In terms of feature engineering, this model selects the relevant features of users and videos from the original data based on the feature selection method, selects important features based on random forest (RF), and forms the interaction features of users and videos by factoring interaction parameters. In terms of machine learning, this model is based on the factorization machine (FM) and the integrated learning model of the deep neural network (DNN), the neural factorization machine (NFM) to realize the click rate prediction model. Based on the NFM model, this paper generates a suitable CTR prediction model through the training of the train data set. After continuous optimization and improvement, the parameters with the best prediction effect are selected. The NFM model established in this paper has a high improvement in performance. Compared with FM, DeepFM, FNN, IFM and DCN models, the AUC value of NFM model is improved by about 0.03 on average. The NFM model can better predict the click behavior of users on advertisements, and can provide decision-making reference for advertisement placement, and can be applied to content recommendation of different advertising systems.*

Keywords: *Click-through rate prediction; Random Forest; NFM model; AUC*

1. Introduction

With the rapid development of the times, the Internet has been integrated into the daily life of each of us, and the resulting large amount of data makes people live in a huge and complex information environment. The 2021 China Internet Advertising Data Report [1] shows that under the influence of factors such as the new coronavirus pneumonia epidemic and domestic demand growth, the Internet has performed strongly as a whole, with advertising revenue of 543.5 billion yuan (excluding Hong Kong, Macao and Taiwan), a year-on-year increase of 9.32%.

Click-through rate (CTR) predictions can impact business revenue and improve user experience. In large-scale enterprise-level advertising systems and recommendation systems, CTR prediction is a very important link, and it has always been one of the key issues of academic research. In recent years, with the rapid development of artificial intelligence technology, related technologies have been widely used in the problem of CTR prediction, and many breakthroughs have been made.

In the CTR prediction problem, there are four main challenges. The first is the problem of accuracy. User's click behavior is affected by many factors, and there are many uncertain factors, resulting in lower accuracy. The second is the problem of sample imbalance. Under normal circumstances, the CTR of advertisements is very low, only a few thousandths, which leads to the imbalance of the sample data learned by the model, which has a great impact on the accuracy of the prediction results. The third is the problem of data sparseness. In the training process of the model, there are very few valid sample data, and a large amount of data is negative sample data that the user has not interacted with, which can easily lead to difficulty in feature learning for the model. The fourth is the cold start problem. There is no interaction information for new ads or new users, making it difficult for the model to make accurate predictions [2].

The classic CTR prediction algorithm is the logistic regression algorithm. The algorithm model is simple and can classify binary problems well, but it cannot extract feature combinations of second order and above. Therefore, the logistic regression model was widely used in the early days [3].

The FM model decomposes the parameter matrix of the second-order combined feature into the click of the feature latent vector, which can be applied to the case where the feature is highly sparse and the sample size is huge. The FFM model introduces the idea of field, and improves it on the basis of the FM model, and achieves good results [4].

DNN models can automatically extract high-dimensional nonlinear combined features in the fields of natural language processing, computer vision, and image recognition. The DIN model introduces an attention mechanism, improves the pooling layer structure of DNN, and enhances the model's ability to represent user interest features [5].

When using multi-party data, there is a great risk of data leakage. In order to minimize the risk of data leakage, it is necessary to reduce the size of data transfers. This paper uses a highly imbalanced user click dataset to build an ad CTR prediction model based on users' real-time exposure, geographic location, mobile device attributes, portraits, and other information. In addition, AUC (i.e., the area under the ROC curve) is a performance indicator to measure the pros and cons of the learner. The higher the AUC, the better the performance of the model in distinguishing positive and negative classes.

Therefore, we use AUC as the metric to evaluate the model. This predictive model can help brand advertisers better carry out data marketing efforts.

2. Assumptions and notations

2.1 Assumptions

We use the following assumptions.

1). It is assumed that the sample can truly and objectively reflect the CTR of advertisements and the basic situation of users.

2). It is assumed that the CTR of advertisements is only related to the user's behavioral preferences, geographical factors and device models, ignoring the influence of other factors such as the financial environment, national policies and technological breakthroughs.

3). It is assumed that the clicks of different users do not affect each other, i.e., the click behaviors of different users are independent of each other.

2.2 Notations

The primary notations used in this paper are listed as Table 1.

Table 1: Notations

Symbols	Meaning
x_i, x_j	i-th and j-th dimensional features
n	Sample size
v_i, v_j	The hidden vector of the i-th and j-th features
w_0	Global bias
w_i	Weight of the i-th variable of the model
w_{ij}	Cross weights of feature i and i

3. Model construction and solving

3.1 Random Forest screening for important features

Random Forest (RF) is an ensemble learning algorithm whose basic idea is to combine multiple classifiers to achieve an ensemble classifier with better prediction effect.

The steps to generate RF are as follows:

1). Using the bootstrap method, randomly select K new autonomous sample sets from the train data set with replacement, construct K classification and regression trees, and construct K out-of-bag data (OOB) for the unselected sample groups.

2). Assuming there are n features, randomly extract mtry features at each node of each tree (mtry⩽ n). By calculating the amount of information contained in each feature, one feature with the most classification ability is selected, and node splitting is performed.

3). Do not make any cut to each tree, let it grow to the maximum.

4). Calculate the variable importance measure based on the classification accuracy.

If the accuracy outside the bag decreases significantly after adding noise randomly to a feature, it means that the feature has a high impact on the classification result of the sample, i.e., the importance of the feature is relatively high.

3.2 NFM Model

Whether it is FM model or DNN model, there are deficiencies in CTR prediction. For FM models, FM lacks the ability to learn higher-order feature interactions and nonlinearities. For DNN models, DNNs lack the ability to model low-order feature interactions. In this paper, the NFM model established by combining the FM model and the DNN model effectively solves the above problems, and has a significant improvement in the prediction ability of the click rate.

The NFM model is equivalent to an FM model with a hidden layer added, and the hidden layer uses the regularization technology of the DNN model, which can prevent overfitting during model training. The innovation of the NFM model is to use the feature cross-pooling part to replace the concatenation used by other models. The NFM model captures lower-level second-order feature interactions using feature cross-pooling. Compared with concatenation, which can provide more information, the way of feature cross-pooling greatly facilitates the subsequent hidden layers of the NFM model to learn useful higher-order feature interactions in a simpler manner.

The NFM model seamlessly combines the linearity of the FM model in modeling second-order feature interactions and the nonlinearity of the DNN model in modeling higher-order feature interactions. The objective function of this model is shown in formula (1).

$$\hat{y}_{NFM}(x) = w_0 + \sum_{i=1}^{n} w_i x_i + f(x) \tag{1}$$

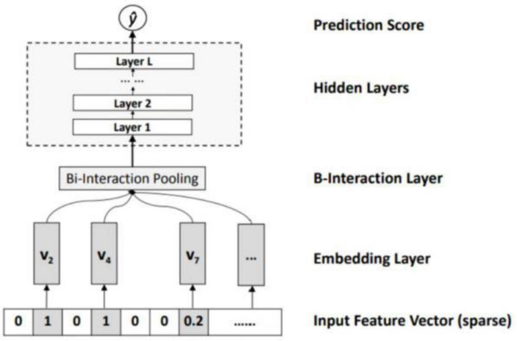

Figure 1: NFM Structure

It can be seen from the objective function that the NFM model replaces the inner product part of the second-order latent vector in the FM model with the DNN model. The DNN model in f(x) takes into account the cross-features in the FM model and is a more expressive function. The NFM network formed by this function is shown in Figure 1. In the NFM model, the forward propagation algorithm of the DNN model is used for boosting, i.e., the output layer is computed from the input layer after traveling to it.

As shown in the figure above, the NFM model is divided into five layers: input layer, embedding layer, Bi-interaction layer, hidden layer and prediction layer. The model uses a forward propagation algorithm to iterate through layers of the network.

Next, the paper introduces each layer, focusing on the Bi-interaction layer.

3.2.1 Input layer and embedding layer

For the train data, this paper first performs data preprocessing, and performs data dimension reduction after re-encoding the data, which greatly reduces the data dimension. In order to reduce the impact of duplicate data of the same user on the prediction results of CTR, the same features of the same user are combined in this paper, and different features are represented by one hot, which greatly reduces the amount of data.

After reducing the amount of data, in order to better play the predictive role of the model, one hot processing is performed for the features of discrete data. One hot data format can detect illegal states and is also helpful for machine learning algorithms to identify. For the features of continuous data, normalization is carried out in this paper. Normalization processing can speed up the convergence speed of the program when it is running, and it is also convenient to quickly find the optimal solution. Finally, this paper inputs all the features into the mid-linear part and the deep part of the model for learning.

Because there is no correlation between feature domains, the embedding operation needs to be performed separately for each feature domain, while numerical features do not need to perform this operation.

3.2.2 Bi-interaction layer

Assuming that vx is the set of all feature embedding vectors, this paper defines a Bi-interaction layer to perform second-order interactions of embedded spatial features.

$$f_{BI}\left(v_x\right)=\sum_{i=1}^{n}\sum_{j=i+1}^{n}x_iv_i\odot x_jv_j=\frac{1}{2}\left[\left(\sum_{i=1}^{n}x_iv_i\right)^2-\sum_{i=1}^{n}\left(x_iv_i\right)^2\right] \tag{2}$$

The above formula represents the element-wise product operation of two vectors, i.e., the element-wise product vector obtained by multiplying the corresponding dimensions of the two vectors, where the operation of the kth dimension is shown in formula (3).

$$\left(v_v\odot v_j\right)_k=v_{ik}v_{jk} \tag{3}$$

The output of this layer is a k-dimensional vector, which is the second-order output in the FM model. The upper layer continues to connect the DNN model, which can enhance the expressive ability of the FM model. There are no additional parameters in this layer, and the time complexity is linear O(kn). Converting multiple vectors into one vector realizes the seamless connection between the FM model and the DNN model.

3.2.3 Hidden layer

The operation of the hidden layer in the NFM model is the same as the operation of the hidden layer in the DNN model, that is, the L weight coefficient matrix W, the bias vector b and the vector obtained from Bi-interaction layer are used to perform linear activation operations, one layer to one layer. It is then calculated to the output layer.

$$\begin{gathered}z_1=\sigma_1\left(W_1f_{BI}\left(v_x\right)+b_1\right)\\ z_2=\sigma_2\left(W_2z_1+b_2\right)\\ \cdots\cdots\\ z_L=\sigma_L\left(W_Lz_{L-1}+b_L\right)\end{gathered} \tag{4}$$

3.2.4 Prediction layer

The result from the last hidden layer zL to the output layer is shown in formula (5).

$$f(x)=h^Tz_L \tag{5}$$

To sum up, each layer of the NFM model is introduced to complete the forward propagation process of the model, which is summarized as formula (6).

$$\hat{y}_{NFM}(x) = w_0 + \sum_{i=1}^{n} w_i x_i + h^T \sigma_L \left(W_L \left(\cdots \sigma_1 \left(W_1 f_{BI}(v_x) + b_1 \right) \cdots \right) + b_L \right) \tag{6}$$

3.3 Model solving

NFM model solving process is shown in Table 2.

Table 2: NFM model solving process

Steps	Operation
STEP 1	Set various parameters and import train dataset
STEP 2	Import data for one hot coding
STEP 3	Perform feature cross-pooling
STEP 4	Import hidden layer for training and start iteration
STEP 5	Train the data and generate neuron weights
STEP 6	Output compute the loss and update the neuron weights in the hidden layer
STEP 7	Repeat steps 4, 5 and 6 until the set number of iterations is reached
STEP 8	Store the trained model for testing

Based on the NFM model, this paper generates a suitable CTR prediction model by training the train dataset, and then selects the best parameters for prediction after continuous optimization and improvement. The model predicts the CTR results of the test dataset, and the best AUC is obtained in this paper.

3.4 Model Optimization

When the number of samples and the number of features of the data set is different, the parameters of the NFM model that can predict the best result will also be different. Among them, batch_size is a relatively important parameter, which represents the amount of data samples captured in one training. The number of training times for each Epoch is related to the value of batch_size, and the relationship is the number of training times per Epoch = the number of samples in the training set/batch_size. In addition, the value of batch_size also affects the training speed and optimization results of the model.

A suitable value of batch_size can make the CPU or GPU run at full load, improve the training speed of the model, and make the direction of gradient download more accurate. In order to maximize the utility of the computer, the value of batch_size is generally N times 2, i.e., 2,4,8,16,32,...,2N.

Because the size of batch_size will have a certain impact on the prediction results, determining the appropriate value will help improve the accuracy of the prediction results. In this paper, the data set is divided into training set and test set according to the ratio of 9:1, 8:2 and 7:3, and the AUC value of advertisement click rate prediction is obtained under different batch_size, as shown in Table 3.

Table 3: AUC values of NFM models with different batch size

	NFM							
	64	128	256	512	1024	2048	4096	8192
9:1	0.6352	**0.6405**	0.6369	0.6367	0.6390	0.6387	0.6378	0.6384
8:2	0.6322	0.6370	0.6367	**0.6379**	0.6334	0.6368	0.6337	0.6133
7:3	0.6358	0.6351	0.6364	0.6354	0.6359	0.6363	**0.6386**	0.6339

AUC values of NFM model with different batch_size is shown in Figure 2.

As can be seen from Figure 2, as the batch_size increases, the AUC value will gradually increase. When the batch_size value reaches 128, the AUC value begins to fluctuate to a certain extent. In Figure 2, when the ratio of training set to test set is 9:1, 8:2 and 7:3, the range of AUC values for advertisement CTR prediction is 0.6352-0.6405 and 0.6322-0.6379, 0.6351-0.6386. When the value of batch_size is between 128 and 512, the AUC value results are better.

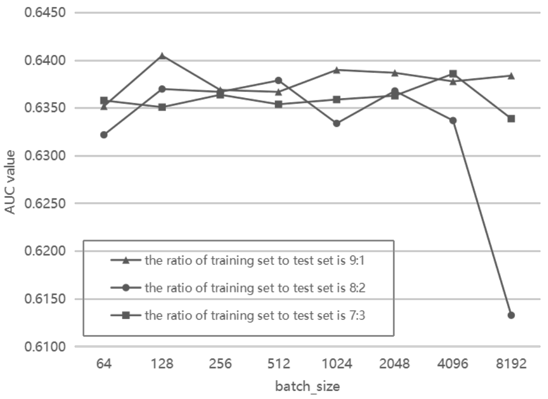

Figure 2: AUC values of NFM model with different batch size

4. Conclusion

The purpose of this question is to improve the accuracy of ad CTR predictions. AUC (i.e. the area under the ROC curve) is a performance metric that measures the quality of a learner. The higher the AUC, the better the performance of the model in distinguishing positive and negative classes, indicating that the prediction effect of the model is better.

It can be seen from the experimental data that compared with the FM, DeepFM, FNN, IFM and DCN models, the NFM model established in this paper has a higher performance improvement, and its AUC value is increased by about 0.03 on average.

The NFM model has the same time complexity as the FM model, making higher-order interactions easy. This model introduces feature intersection in the Bi-interaction layer, making the model very suitable for dealing with sparse data. The robustness of the model can effectively prevent the occurrence of overfitting. However, the performance of this model increases as the dimension of the embedding layer vector increases, but it leads to an increase in the amount of computation. The model is relatively insensitive to parameter initialization, and it is difficult to optimize the model.

In order to improve the accuracy of advertising CTR prediction results, this paper combines random forest algorithm and NFM algorithm to establish a suitable model. Screening user characteristics makes the trained ad click rate prediction model not affected by irrelevant information. This measure greatly improves the accuracy of the prediction results. The experimental results verify the feasibility and efficiency of the model. The NFM model can better predict the user's click behavior on advertisements. In addition to being suitable for classification tasks, it also has excellent performance in regression and ranking tasks. The model can be applied not only to this dataset, but also to other datasets, which can provide decision-making reference for advertisement placement.

References

[1] Ma Jia. China Internet Advertising Data Report 2021" published [N]. China Market Monitor, 2022-01-20(004).

[2] Yan Jinyao, Zhang Hailong, Su Yumin. Research on click-through rate and conversion rate prediction in computational advertising [J]. Journal of Communication University of China (Natural Science Edition), 2021, 28(02):54-60.

[3] McMahan H B, Holt G, Sculley D, et al. Ad click prediction: a view from the trenches[C]. Proceedings of the 19 th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013:1222-1230.

[4] Rendle S. Factorization machines[C]. In Proc Int Conf Data Mining, Sydney, NSW, AUS, 2010: 995-1000.

[5] Covington P, Adams J, Sargin E. Deep neural networks for YouTube recommendations[C]. RecSys'16: Proceedings of the 10th ACM Conference on Recommender Systems, 2016:191-198.