# Estimating Ads' Click through Rate with Recurrent Neural Network

Qiao-Hong Chen[1,a], Shi-Min Yu[1], Zi-Xuan Guo[1] and Yu-Bo Jia[1]

[1]School of Information, Zhejiang Sci-Tech University, Hangzhou, Zhejiang Province, China

**Abstract.** With the development of the Internet, online advertising spreads across every corner of the world, the ads' click through rate (CTR) estimation is an important method to improve the online advertising revenue. Compared with the linear model, the nonlinear models can study much more complex relationships between a large number of nonlinear characteristics, so as to improve the accuracy of the estimation of the ads' CTR. The recurrent neural network (RNN) based on Long-Short Term Memory (LSTM) is an improved model of the feedback neural network with ring structure. The model overcomes the problem of the gradient of the general RNN. Experiments show that the RNN based on LSTM exceeds the linear models, and it can effectively improve the estimation effect of the ads' click through rate.

## 1 Introduction

In 2014, online advertising has overshadowed television advertising for the first time in terms of market size, reaching the sum of 154 billion yuan in China with a 40% increase year-on-year. Compared with 77.3 billion yuan in 2012, the market size of online advertising almost doubles in 2014, and in 2015 it is more than 200 billion yuan.

As an important research area in the field of computational advertising, ads' click through-rate estimation is one important way to increase the online advertising revenue. Based on rich historical data, the ads' click through-rate estimation model makes full use of the complex relation among a large number of nonlinear characteristics of historical data as much as possible for the sake of the estimation accuracy.

In combination with the advertising position, ad auction mechanism and other factors, the increasing ads' click-through rate estimation accuracy could make online advertising be put more accurately, so as to improve the real advertisement click-through rate. According to the online advertising payment mechanisms, most companies are using "click pay cost per click" (CPC), namely, the more the ads are clicked, the more profit it makes[2].

Ads' click-through rate estimation process can be divided into four steps: feature extraction, model building, model training, and model estimation. There are many tries to estimate ads' click through rate. Joachims[3] came up with online Bayesian probability regression(OBPR), but it was based on the specific characteristics of advertisement, which made it difficult to achieve personalized recommendation accurately. Chapelle et al.[5] proposed the dynamic Bayesian network model. Dave et al.[6] adopted gradient boosting decision tree (GBDT) as a regression model to extract the similar characteristics. Richardson[7]used logistic regression (LR) model,which is used in the nonlinear characteristics of learning, but it doesn't fully reflect the relationship among many features and with the increasing of the number of iterations and learning time, it could easily cause the over fitting problems. Agrawal et al.[8] presented spatio-temporal predicting models in 2009. Agarwal et al.[9] proposed to use sparse data pre-existing hierarchy to solve the sparse event and sparse data rate estimation problem. Zhang et al.[10] put forward the COEC (clicks over expected clicks) model, which set an expected figure in advance. Cheng et al.[11] matched the user's search terms and the content of the advertisement. Zhang et al.[12] proposed the use of RNN to predict search advertising click through rate, adopting back-propagation through time (BPTT) for model training.

According to experimental results, it is more accurate compared with the LR models and NN models. However, there would be problems of gradient disappearance or gradient outbreak when using gradient falling algorithm for the RNN algorithm. In order to solve this problem, this paper adopts RNN based on LSTM, and uses LSTM special structure to avoid problems of gradient disappearance or gradient outbreak and to improve the model's accuracy.

Advertisement data in this paper is from Avazu Company. The implicit features are extracted according to the dominant features and hidden features such as users' attribute characteristics. The hidden layer of our model adopts three layers of connection structure, which makes the model to be trained adequately. The experimental results show that our model is more accurate than the LR model, BP neural network (NN) model and RNN model.

[a] Corresponding author: chen_lisa@zstu.edu.cn

The main contributions of this paper are in three folds:

First we analyze large number of advertising data features and extract hidden users' attributes with mosaic characters. Mosaic characters are hashed to new characters, so the original different types of characteristic values are turned into the same type of characteristic values. In accordance with the NN model, the characteristic values are mapped to [0,1] by using the normalization process.

The improved RNN model based on LSTM has three layers of hidden layer, each layer possessing 256 nodes. This model is used to simulate the users' click behavior and to estimate the ads' click through rate.

We use a large number of training data and test data to verify the validity of the model. The experimental results show that the model is better than the LR model and the general RNN model.

The first section of this paper gives the introduction of the researches for Ads' CTR prediction. In the second section, we propose the definition of the RNN model based on LSTM, including the training process and evaluation function of the model. After that, we analyze the advertising data and introduce experiments and results of each model in section four. Section five presents the conclusions and prospects for the future work.

## 2 Recurrent Neural Network Model Based on Lstm

### 2.1 Model Definition

Recurrent neural network based on LSTM , which uses LSTM structure to replace the hidden layer nodes of the general recurrent neural network, and the LSTM structure increases the input gates, output gates, forget gates and an internal unit (Cell).

The input gate indicates whether the input layer is allowed to enter the hidden layer node. The door is opened to allow the input layer of the output signal to enter, the door is closed to refuse to enter the signal ,input gate is denoted as $\iota$ . The output gate indicates whether the output value of the current node is output to the next layer. The door is opened to allow the hidden layer node of the signal output, the door is closed to refuse to signal output, output gate is denoted as $\omega$ . The forget gate decides whether to retain the current hidden layer node storage of historical information. The door is opened to keep the history information of the node of the hidden layer, and the door shut does not keep the history information of the node of the hidden layer,forget gate is denoted as $\phi$ . $s_c^t$ represents the value of the information stored at the t time. And the input and output layers of the model are consistent with the RNN model , as shown in figure1:

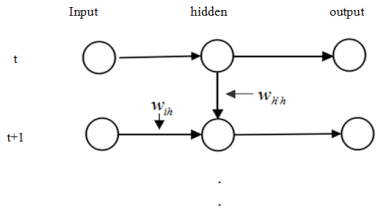


**Figure 1.** Recurrent Neural Network Structure

Where $w_{h\tau}$ is expressed as the weights between the hidden layer and the input gate unit.In the following part of this paper,the different subscript of $w$ indicates the weights of different nodes.

The hidden layer nodes are replaced as shown in figure 2:

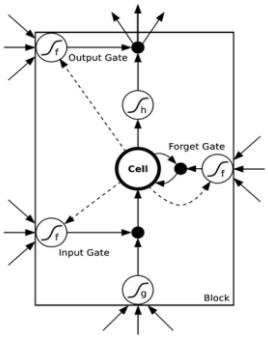


**Figure 2.** Long-Short Term Memory Structure

So, the structure of the recurrent neural network model based on LSTM is shown in figure 3.

### 2.2 Model Training

With the general recurrent neural network of the two values of the source as the input is different, The input of the input gate consists of three values. Including the input layer node of the output vector $x_i^t$ , the first hidden layer of the cell's output vector $b_h^{t-1}$ , the previous time cell of the retention of information $s_c^{t-1}$ , the input vector of input gate at time t $a_\tau^t$ is as follows:

$$a_\iota^t = \sum_{i=1}^{I} w_{i\iota} x_i^t + \sum_{h=1}^{H} w_{h\iota} b_h^{t-1} + \sum_{c=1}^{C} w_{c\iota} s_c^{t-1} \quad (1)$$

Through the activation function $f$ of the input gate, the output vector at t time $b_\tau^t$ is as follows:

$$b_\iota^t = f(a_\iota^t) \quad (2)$$

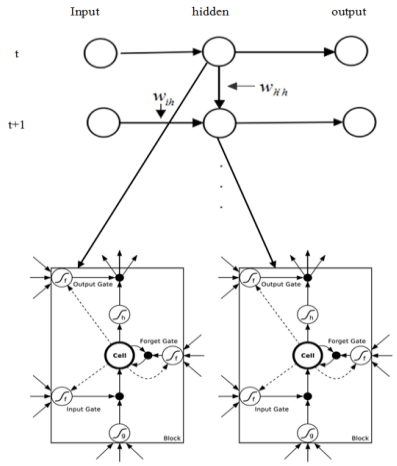


**Figure 3.** Recurrent neural network base on LSTM

The input of the forget gate is also made up of three input vectors, which is the same as the source of the input gate.the input vector of the forget gate at time t $a_\phi^t$ is as follows:

$$a_\phi^t = \sum_{i=1}^{I} w_{i\phi} x_i^t + \sum_{h=1}^{H} w_{h\phi} b_h^{t-1} + \sum_{c=1}^{C} w_{c\phi} s_c^{t-1} \quad (3)$$

Through the activation function $f$ of the forget gate, the output vector formula of the gate is obtained at time t $b_\phi^t$ is as follows:

$$b_\phi^t = f(a_\phi^t) \quad (4)$$

Cells unit from the figure2, it can be known that its input is composed of two parts, one is the input vector of the input layer $x_i^t$ , one is the output of the first hidden layer of the output gate $b_h^t$ , the input vector of cell unit at time t $a_c^t$ is as follows:

$$a_c^t = \sum_{i=1}^{I} w_{ic} x_i^t + \sum_{h=1}^{H} w_{hc} b_h^{t-1} \quad (5)$$

According to the Forget door to determine whether to retain the information value $s_c^t$ of the past. $s_c^t$ is as follows:

$$s_c^t = b_\phi^t s_c^{t-1} + b_i^t g(a_c^t) \quad (6)$$

The input of output gate is composed of three parts, the output vector of the input layer, a hidden the output vector of the layer of the cell and the cell retention of information, the input vector formula of the output gate at time t $a_\omega^t$ is as follows:

$$a_\omega^t = \sum_{i=1}^{I} w_{i\omega} x_i^t + \sum_{h=1}^{H} w_{h\omega} b_h^{t-1} + \sum_{c=1}^{C} w_{c\omega} s_c^t \quad (7)$$

The output vector $b_\omega^t$ of the output gate is obtained by using the activation function of the output gate unit at time t:

$$b_\omega^t = f(a_\omega^t) \quad (8)$$

The output vector of the Cell unit $b_c^t$ is as follows:

$$b_c^t = b_\omega^t h(s_c^t) \quad (9)$$

The output vector $a_k^t$ of cell unit, that is, the hidden layer of the output vector, as the input vector of the output layer, the formula is as follows:

$$a_k^t = \sum_{c=1}^{H} w_{ck} b_c^t \quad (10)$$

The resulting vector $b_k^t$ of output from the output layer is as follows:

$$b_k^t = f(a_k^t) \quad (11)$$

The weights $w_{ij}$ between the i node and the j node are updated as:

$$w_{ij} = w_{ij} - \eta \nabla L(w_{ij}) \quad (12)$$

Where $\eta$ is expressed as Learning step, $\nabla L(w_{ij})$ is expressed as follows:

$$\nabla L(w_{ij}) = \frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial a_j^t} \frac{\partial a_j^t}{\partial w_{ij}} = \delta_j^t b_j^t \quad (13)$$

Where $\delta_j^t$ is expressed as residual error of the j node, $b_j^t$ is expressed as output vector of the j node.So ,the formula 12 can be simplified using formula 13:

$$w_{ij} = w_{ij} - \eta \delta_j^t b_j^t \quad (14)$$

### 2.3 Loss Function and Evaluation Function

We need a criterion to judge whether the model is good or bad. Area Under roc Curve (AUC) is a common criterion. However, AUC focuses on the sorting of the CTR estimation, and logloss focuses on the accuracy of th e CTR estimation.When the click through rate all increase s in a certain proportion, it will not cause change in AUC. However, it will cause a change in logloss. Logloss is a r eflection of differences between the estimated click rate t hrough the model and the true click rate. The logloss is s

maller, the estimation results of the ads' CTR is more acc urate. This paper uses logloss in the scikit-learn, and loglo ss is defined as follows:

$$\log loss = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1-y_i)\log(1-p_i)) \quad (15)$$

Where $y_i$ is expressed as the i-th click true value, $p_i$ is expressed as the i-th click value estimated by the model.

## 3 Experimental Result

### 3.1 Data Analysis

In this paper we use avazu company's advertising data (on ly the training set)to validate the model we proposed can enhance the CTR accuracy. The training set samples are made of 40428967 records. Each record consists of 24 fea tures , including 15 explicit features and 9 hidden encrypt ion features.We divide the training set samples into four p arts, the three parts as the training data, one part for the test data. As shown in the table 1, the click rate of test data set and the real hit rate of the training data set is similar, so the data will not affect the model's prediction.

**Table 1.** Data Set.

| Data | Expression | Click | Click rate |
|---|---|---|---|
| Train | 30000000 | 5102362 | 17% |
| Test | 10428967 | 1762704 | 16.9% |

### 3.2 Feature Processing

The analysis of the characteristics of advertising data indi cates that in 24 features there are most two sets of the feat ures beginning with "device_ip" or "device_id", and the value of the features beginning with "device_ip" or "devi ce_id" are too much small numbers, so that advertising da ta has a lot of the long tail of the characteristics. The anal ysis results are shown in table 2:

**Table 2.** The Feature Sets Of Device_Ip and Device_Id.

| Feature | Sets |
|---|---|
| device_ip | 2686408 |
| device_id | 6729486 |

In order to make the model more stable to learn, so we filter out some of the long tail characteristic values. In th e characteristics of samples, we filter the device_ip freque ncy less than 10 times and device_id frequency less than 10 times. Thus the size of training samples are reduced to 23548762 records.

We merge the characteristics of device_ip, device_id, device_ model and C14 into mosaic characters. Then mos aic characters are hashed to new characters, from which t he first 8 characters are denoted as the user_id. We merge feature C15 and feature C16 into a new feature, which is denoted as banner_size. Banner_size also is hashed. Rem ove the C15 and C16 features, add the features of banner_ size and user_id, the number of features is still 24.
The obtained features are normalized, and the characterist ic values are mapped to [0,1].

### 3.3 Result

Most machine learning models are adapted to find a line, plane or higher dimensional space to reach the trend of infinite approximation of data character.

Table 3 shows the logloss values of each model under different iteration times. As shown in the table3, the logloss value of LR reaches the minimum value 0.388364 in the fortieth iteration, and then the value begins to increase, which indicates that the model has been fully studied. The logloss value of BP reaches the minimum value 0.461315 in the fiftieth iteration. BP neural network model is essentially a gradient descent algorithm, so there are local minimum problems, which will lead to early termination. Compared with the LR model and the BP neural network model, the logloss value of RNN reaches the minimum value 0.386299 in the sixth iteration. Therefore, RNN has a better ads' click-through rate estimation.

In comparison with BP, RNN model has the input of the hidden layer nodes not only from the output of the input layer nodes, but also from last moment output of Hidden layer nodes, so our model can learn the relationships among the more complex characteristics.

Although the RNN can remember the historical infor mation, yet with the increasing of the depth of learning, RNN results in problems of vanishing gradient. We us e LSTM instead of the normal neuron, so that improved RNN model Based on LSTM can memorize history information as much as possible to prevent the disappeara nce of the gradient. The experimental results show that th e minimum logloss value of improved RNN mode based on LSTM is 0.383213, thus the model is better than other models in the estimation of ads 'click-through rate. The rectangular chart in Figure 4 is more intuitive to draw the logloss values of each model.

**Table 3.** Experimental Result.

| | 10 | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|---|
| LR | 0.396062 | 0.391579 | 0.389702 | 0.388364 | 0.389496 | 0.391109 |
| BP | 0.467845 | 0.464784 | 0.462847 | 0.461937 | 0.461315 | 0.461674 |
| RNN | 0.410213 | 0.408456 | 0.402215 | 0.394515 | 0.390126 | 0.386266 |

| | | | | | | |
|---|---|---|---|---|---|---|
| LSTM | 0.401 235 | 0.395 651 | 0.38 526 3 | 0.383 213 | 0.383 756 | 0.38 345 6 |

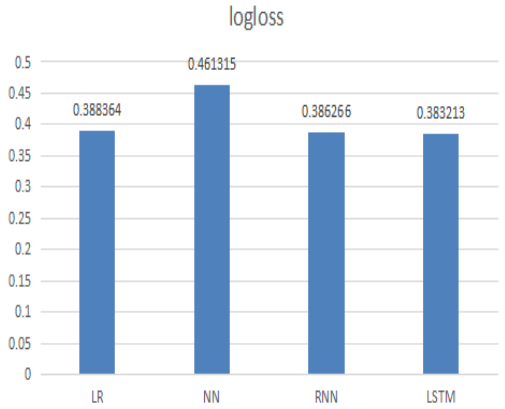


**Figure 4.** Experimental Result.

## 4. Conclusions and Prospects For Future Work

The problem of online advertising CTR estimation is one of the most popular fields of computational advertising, which has been drawing more and more attention. This paper is based on improved LSTM recurrent neural network model for the estimation of the advertising history data. The experimental results show that our model is better than the contrast models in accuracy concerning estimated advertising CTR. It turns out that the work done in this paper is effective.

The experimental data are from Kaggle data, provided by Avazu. The data has been encrypted and sampling processed. In the light of the data analysis, sampling data click rate is 17%, which is significantly higher than that in real life. Actually advertisement data is distributed extremely unevenly, so we are still committed to finding true advertisement data for training and evaluation of each model.

Gradually we begin to use deep learning neural network model for ad CTR, such as convolutional neural network and deep network learning model. This paper adopts improved RNN model based on LSTM for estimation of ads' click-through rate. In view of the experimental results, it is better than LR model and general NN model. The NN model is more sensitive to the input feature, and high dimensional features affect the model training, even result in its failure. So it would become a hot research prospect in the future to work more effectively among the sea quantity data for feature extraction, feature selection and feature reduction,

There are a lot of neural network optimization algorithm and the objective function. The optimization algorithm used in RNN in this paper is SGD algorithm. The new optimization algorithms have been constantly appearing, such as adaptive algorithm Adadelta[14], Adagrad[15] algorithm. Objective function in this paper is cross entropy function, and common objective function also includes MSE (mean squared error), MAE (mean absolute error) and categorical cross entropy as well. Optimization algorithm and objective function are also problems worthy of research in the future.

## References

1. Iresearch. Analysis of the overall development of China's online advertising market[EB/OL], http://www.iyunying.org/seo/sjfx/12636.html
2. Zhou A Y, Zhou M Q, Gong X Q. Computational advertising: A data-centric comprehensive web application[J]. Jisuanji Xuebao(Chinese Journal of Computers), 2011, 34(10):pp.1805-1819.
3. Joachims T. Optimizing search engines using click through data[C]//Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002: pp.133-142.
4. Graepel T, Candela J Q, Borchert T, et al. Web-scale bayesian click-through rate estimation for sponsored search advertising in microsoft's bing search engine[C]//Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010: pp.13-20.
5. Chapelle O, Zhang Y. A dynamic bayesian network click model for web search ranking[C]//Proceedings of the 18th international conference on World wide web. ACM, 2009: pp.1-10.
6. Dave K, Varma V. Predicting the Click-Through Rate for Rare/New Ads[R]. Center for Search and Information Extraction Lab International Institute of Information Technology Hyderabad, INDIA, April 2010.
7. Richardson M, Dominowska E, Ragno R. Predicting clicks: estimation the click-through rate for new ads[C]//Proceedings of the 16th international conference on World Wide Web. ACM, 2007: pp.521-530.
8. Agarwal D, Broder A Z, Chakrabarti D, et al. estimation rates of rare events at multiple resolutions[C]//Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007: pp.16-25.
9. Agarwal D, Agrawal R, Khanna R, et al. estimation rates of rare events with multiple hierarchies through scalable log-linear models[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010:pp. 213-222.
10. Zhang W V, Jones R. Comparing click logs and editorial labels for training query rewriting[C]// WWW 2007 Workshop on Query Log Analysis: Social And Technological Challenges. 2007.
11. Cheng H, Cantu-Paz E. Personalized click estimation in sponsored search.[C]// Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010:pp.351-360.
12. Zhang Y, Dai H, Xu C, et al. Sequential Click estimation for Sponsored Search with Recurrent Neural Networks[C]. AAAI 2014:pp.1369-1375.
13. LogarithmicLoss[EB/OL].https://www.kaggle.com/wiki/LogarithmicLoss.

14. Zeiler M D. ADADELTA: An adaptive learning rate method[J]. arXiv preprint arXiv:1212.5701, 2012.
15. Wager S, Wang S, Liang P. Dropout Training as Adaptive Regularization[J]. Advances in Neural Information Processing Systems, 2013:pp.351-359.