Analysis of the Click through Rate of Chinese Netizens through Social Software based on DeepFM Algorithm

Li Bo¹², Zhang Xiao¹², Li Kewei¹² and Yan Jingyi ³

¹ School of Computer Science, Northwestern Polytechnical University, Xi’an Shaanxi 710129, China;
² Ministry of communications Key Laboratory of big data storage and management, Xi’an Shaanxi 710129, China;
³ School of Management, Northwestern Polytechnical University, Xi’an Shaanxi 710129, China

Corresponding author: e-mail: zhangxiao@nwpu.edu.cn

Abstract. Advertising as an important means of marketing, plays an important role in the promotion of commercial services, but its profit efficiency has not been optimistic. How to reduce the push cost and improve the accuracy rate of advertising is the algorithm problem of various enterprises. Chinese netizens have a large base and a large proportion use social software. It is a novel idea to analyze the click-through rate of Chinese netizens' advertisements through the data provided by social software. Based on the idea of lookalike crowd expansion, this paper uses DeepFM algorithm and designs the corresponding experimental process to analyze and predict the clicks of Chinese netizens through social software.

1. Introduction

With the development of life, we are surrounded by advertising, but people are indifferent to the overwhelming advertising. It has become an important research direction that how to put the right advertisement to the right crowd or find the right potential customers through advertising data mining. When Facebook began using the Facebook Custom Audiences feature in the advertising space in 2012, audiences found that the concept was really being used on a large scale. Audience discovery can find potential users according to existing users, carry out appropriate advertising promotion, reduce the cost of promotion[1]. A recommendation algorithm like Facebook that discovers and extends other users through a group of existing users is called Lookalike. Each company has its Lookalike algorithm designed and kept secret. For example, Tencent hosted the Tencent Advertising Algorithms Competition in 2017 and 2018 to discover talents and more accurate algorithms in this area. Aiming at the problem of advertising click-through rate in Lookalike algorithm, this paper analyzes the complicated social scene, various advertising forms, and huge crowd data, and explores the current click-through situation of Chinese netizens through social software[2].

2. Research status

Lookalike, also known as similar crowd extension, seeks similar audiences based on seed user portraits and social links by opening up rich data capabilities. This algorithm requires a relatively large order of magnitude, suitable for the current era of data blowout development. In other words, it’s a...
technology that seeks out similar groups of people through seed users, which can improve the accuracy of crowd orientation, as shown in the Figure 1.

**Figure 1.** The picture of population expansion

Ads lookalike crowd expansion algorithm, in Weixin, Weibo, Facebook, Google has a very wide range of applications. Advertising, as a very important part of all kinds of services, can be analyzed as a separate part, because even close friends may be interested in different advertisements. Delivering corresponding advertisements to specific users can reduce the operation and maintenance costs of the company. What’s more, clustering analysis of clicks on certain advertisements can lead to people with similar interests and then recommend similar advertisements[3-4].

3. Algorithm

The analysis of advertisement click rate algorithm in lookalike algorithm can be regarded as CTR problem. CTR (Click-Through-Rate) refers to the click-through rate of an online advertisement divided by the actual click-through rate of the advertisement by the amount of display of the advertisement[5]. CTR prediction is to predict whether users click on the ads. In fact, it can be seen as a two-class problem, that is, point and no point. Aiming at this problem, the following algorithms are mainly studied at present.

3.1. Existing algorithms

3.1.1. FM algorithm. FM (Factorization Machine) is mainly to solve the problem of how to combine features when data is sparse. In the problem of advertising classification, according to some characteristics of users and advertising space, to predict whether users will click on the advertisement[6]. The objective function of the two element cross FM (2-way FM) algorithm is as follows:

$$
\hat{y} := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle V_i, V_j \rangle x_i x_j
$$

(1)

Among them, $w_0 \in \mathbb{R}$, $w \in \mathbb{R}^n$, $V \in \mathbb{R}^{nk}$, $W$ is the parameter of input feature, $\langle v_i, v_j \rangle$ is the cross-parameter between input features I and j, and V is the k dimensional vector $\langle V_i, V_j \rangle := \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f}$.

Under sparse conditions, the independence of features is broken through the learning method of vector v, which can better mine the correlation between features. FM takes into account the relationship between features through learning a hidden variable for each feature. Its advantages are: considering the relationship between features, enhance the generalization ability of the model; through clever decomposition and merging of the objective function, can be completed under $o(n)$ time complexity; suitable for processing sparse data.

3.1.2. FFM algorithm. FFM (Field-aware Factorization Machine) improves the FM model by classifying the same features into one field. FFM. In the FM model, each feature corresponds to a hidden variable. But in FFM model, the feature should be divided into multiple fields, each feature corresponds to a hidden variable for each field. The model of FFM is:
\begin{equation}
  f(x) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{j_1=1}^{n-1} \sum_{j_2=j_1+1}^{n} (v_{j_1}^T v_{j_2}) x_{j_1} x_{j_2}
\end{equation}

\(j_1, j_2\) represent the index of features.

The FFM algorithm is based on the FM model, and the performance of the model is further improved. It is also suitable for dealing with sparse data. But FFM can not decompose and merge the objective function like FM and complete the calculation of the function in linear time. Time complexity is \(o(k \cdot n^2)\) which more complex than FM. For large data, the time complexity of FFM will increase significantly, and the performance of computer CPU will be tested very much.

3.2. Algorithm analysis

Most CTR problems are consistent with the Long Tail Effect. "Head" and "tail" are two statistical nouns. The middle part of the normal curve is called "head", and the relatively gentle part on the two sides is called "tail". From the point of view of people's needs, most of the demand will be concentrated in the head, which we can call popular, and distributed in the tail of the demand is personalized, scattered small amount of demand. The relationship between users and advertisements can be abstracted as the relationship between point and point. A pink circle with a green triangle and a black line indicate that the user clicked on the ad, resulting in Figure 2. But the fact is that everyone doesn't have to click on just one ad, so it turns out that Figure 2 down Figure 3.

\begin{figure}[h]
  \centering
  \includegraphics[width=0.5\textwidth]{basic_relationship.png}
  \caption{The chart of the basic relationship between users and advertisements.}
  \end{figure}

\begin{figure}[h]
  \centering
  \includegraphics[width=0.5\textwidth]{complex_relationship.png}
  \caption{The chart of the complex relationship between users and advertisements.}
  \end{figure}

In the current situation, one person can correspond to more than one advertisement, one advertisement can correspond to more than one person, but there is no relationship between advertising and advertising, between people. The ultimate goal is to explore the relationship between people and then recommend the right ads according to the group or the person's preferences. We can add the weight of the degree of association between advertisements, such as the new yellow line in the Figure 4, and the thickness of the line represents the degree of association between advertisements. The thicker the yellow line represents the greater the degree of association between the two advertisements.
The previous three scenarios assumed that the number of clicks per person on an advertisement was fixed, but the actual number of clicks per person was significantly different, and according to the long tail effect, the number of clicks per person on an advertisement would be significantly different, more than 90% of the advertisement clicks would not exceed five, and the remaining 10% of the advertisement clicks would be 90%. Advertising hits several times. In Figure 4, it is not for the first time to consider the difference between people clicking on advertisements and the relevance of advertisements. At this point, to recommend potential ads to the same group of people, you only need to recommend ads that the user does not pay attention to, but are most relevant to the ads that user has already paid attention to.

Similarly, by introducing the difference in click-to-click advertising rates, the thicker the black line, the greater the user's attention to the advertisement, and this leads to Figure 5. According to the analysis of this situation, only need to recommend to the user that he did not click on the ads, and he clicked on the ads associated with the greatest degree of click, and the same crowd clicked the greatest degree of advertising.

In Figure 4 and Figure 5, because of a lot of restrictions, the problem will be simplified. At this point, through data pre-processing, we can classify advertising, and introduce the related functions of image recognition and character recognition. But there are actually two methods:

- Advertisers in advertising when the tag, so that advertising can be directly classified so that it can be completed in accordance with the above analysis(Figure 4).
- According to the relevance between users and advertisements, the relationship between users and advertisements is analyzed by data mining.

In this paper, we analyze the correlation between users and advertisements. In view of this problem, such a huge amount of data for general graphics, are very inappropriate, so this paper chose the DeepFM algorithm to Chinese netizens through social software advertising clicks.

3.3. DeepFM algorithm
DeepFM is an algorithm proposed by Google to solve the CTR problem. This paper applies this algorithm to analyze the click-through rate of Chinese netizens'social media ads and uses the data set
provided by Tencent Advertising Algorithms Competition to verify the algorithm. The core idea of the DeepFM algorithm is:

- Using FM Component + Deep Component. FM extracts low order combination features, and Deep extracts higher order combination features. DeepFM is end-to-end training without artificial feature engineering.
- Analysing feature embedding. FM and Deep share input and feature embedding to make training faster and more accurate.

As can be seen, the whole model is divided into two parts: FM and DNN, which are responsible for extracting low-order features and high-order features respectively. With the idea of FNN, FM is used to embedding, and then the wide and deep models share the results after embedding. The input of DNN is exactly the same as that of FNN. When combined in a certain way, the model simulates the effect of FM completely on the wide. Finally, the results of DNN and FM are combined to activate the output. DeepFM's prediction results are:

\[ y = \text{sigmoid}(y_{FM} + y_{DNN}) \]  

Compared with other algorithms, DeepFM does not pre-train the hidden vector V with FM and initializes the neural network with V. The FM module is not independent, it is trained and learned with the whole model. No feature engineering is needed, and the training efficiency is high. No pre-training shares Feature Embedding, no feature engineering, and captures both low-high-order interaction features. The DeepFM algorithm can effectively reduce the singularity and uncertainties in the data set, and get more accurate prediction results.

4. Experiment

In order to verify the click-through rate of Chinese netizens'social software advertisements discussed in this paper, the data set is provided by Tencent Advertising Algorithms Competition in 2018. The dataset has no time information, so it is difficult to construct time-dependent features. This dataset mainly includes information such as users and advertising categories, so we explore these data.

4.1. Data set analysis

4.1.1. Evaluation indicators. For the extended similar users, if there is relevant effect behavior (click or transform) in the advertising, it is considered as a positive case; if there is no effect behavior, it is considered as a negative case. Each seed pack to be evaluated will provide the following information: the ad aid corresponding to the seed pack and its characteristics, and the corresponding candidate user set (userid and its characteristics). So the accuracy of the algorithm is tested.

4.1.2. Data problem. According to the five features of the data set, click, ratio, conversion rate, special conversion rate, multi-value length, tagging tags are needed for each class feature. It is noteworthy that the conversion rate uses the pre-processed block tags to separate a block verification set without adding statistics, the remaining blocks do drop out cross-statistics, and the test set uses all the training set data for statistics.

4.2. Characteristic structure

According to the five features of the data set, click, ratio, conversion rate, special conversion rate, multi-value length, tagging tags are needed for each class feature. It is noteworthy that the conversion rate uses the pre-processed block tags to separate a block verification set without adding statistics. We found that the importance of some multi-value fields is very high and extract the top 20 coding features for storage and analysis, as shown in the Figure 6.
As the structure of the data set, this paper selects the following features: the original features of the data set (user characteristics, advertising features), sparse features, unique and click features, as well as proportional features. The proportional feature is used to characterize the preferences of a class of users, and better distinguish users according to preferences. Because no time feature is given in this data set, it is easy to generate data traversal and over-fitting when constructing the conversion rate feature. In order to solve data traversing, we can use the block conversion rate or Bias smoothing method.

4.3. Feature selection
After the above process of constructing features, the data set can achieve hundreds of features, but it is impossible to train all features, because it may contain many redundant features, and we need to achieve the effect of multiple features in the case of few features. The most commonly used method is the correlation coefficient method and the method of modeling the importance of output characteristics. This paper divides the data set into 10 groups, screening each group of features, sorting the importance of the characteristics within the group.

4.4. Model training and assessment
DeepFM model method is used in this paper. After AUC sorting according to the multi-group prediction values of the verification set, the weights are traversed in turn (list (range (0,101)*0.01) to obtain the best weights, and then the same weights are applied to the prediction results of the test set, so that each of them is more. By weighting a sub model, the AUC of the validation set will only be greater than or equal to the AUC before the weighted sub model. In fact, the whole weighting process is similar to a linear fitting. It can also use the verification set of each sub-model and the prediction results of the test set as the characteristics, use the label of the verification set as the real label, and use the xgboost and other models to train, so the effect is similar to the previous ergodic weighting. According to aid, a set of 20% training sets is used as validation set to verify the accuracy of the algorithm.

4.5. Experimental results
Through the algorithm described above, and the corresponding experimental steps, rent the Ali cloud server, configure the Python environment under the server Linux, and then do the experimental analysis. Although the research in this paper is still inadequate, but through analysis, it can greatly improve the accuracy of the GTR problem. Comparing the top ten results of the 2018 Tencent Advertising Algorithms Competition, all above 0.7728, with the highest accuracy of 0.781422. In this paper, the DeepFM algorithm is used to do three experiments, the prediction accuracy is 0.74456, 0.75375, 0.75221. The algorithm presented in this paper is less accurate than that of the Tencent Advertising Algorithms Competition in 2018, but it provides a way to analyze the click-through rate of Chinese netizens on social software. Moreover, in the competition, if the excessive collection of features, may also make a higher accuracy, but this is not desirable, because it is not conducive to generalization of research, so the results of the 2018 Tencent advertising algorithm contest can only be a reference result, can not be fully judged according to this result.
5. Conclusion
The current era belongs to the era of data, mainstream social software companies, such as Tencent, Facebook have a strong interest in lookalike algorithm, and in the search for a better algorithm. GTR problem is a problem that appears in recent years. Its precision is generally low, and it has great research prospects. In this paper, DeepFM algorithm is used to analyze and predict the potential interest of Chinese netizens in social software advertisements. Through the algorithm analysis, the relationship between users and advertising is still in line with the long tail distribution, Chinese netizens are still mainly young people, so for games, entertainment software advertising click rate is higher than other advertising, if further excavation, there will be great potential. In most experimental scenarios of recommended categories, if only to achieve good prediction rate, there is a relatively simple feature mining idea: the more fine-grained features are important, the closer the features are, the more important the features are, and the feature weights about the future are the largest. This is also a drawback of data mining in recent years.

Acknowledgments
This work is supported by the MIIT Key Laboratory of Big Data Storage and Management, the NFS of China under Grant No.61502392 and No.61272323 and the Ministry of Science and Technology of China, National Key Research and Development Program (No.2018YFB1004401).

References
[1] Jiesi Guo,Herbert W. Marsh,Philip D. Parker,Theresa Dicke. Cross-cultural generalizability of social and dimensional comparison effects on reading, math, and science self-concepts for primary school students using the combined PIRLS and TIMSS data[J]. Learning and Instruction,2018,58.
[2] Daisuke Ichikawa,Toki Saito,Waka Ujita,Hiroshi Oyama. How can machine-learning methods assist in virtual screening for hyperuricemia A healthcare machine-learning approach[J]. Journal of Biomedical Informatics,2016,64.
[3] Colin E. Vize,Katherine L. Collison,Joshua D. Miller,Donald R. Lynam. Using Bayesian methods to update and expand the meta-analytic evidence of the five-factor model's relation to antisocial behavior[J]. Clinical Psychology Review,2018.
[4] M. Vulfin,K. F. Tagirova. Enhancement of accuracy of deep-pumping equipment based on data mining[J]. Optical Memory and Neural Networks,2015,24(1).
[5] Dzyuba, Vladimir, and M. V. Leeuwen. "Learning What Matters – Sampling Interesting Patterns." Pacific-Asia Conference on Knowledge Discovery and Data Mining Springer, Cham, 2017:534-546.
[6] Dzyuba, Vladimir, and M. V. Leeuwen. "Learning What Matters – Sampling Interesting Patterns." (2017):534-546.