IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED DATA MINING METHODS FOR SOCIAL COMPUTING

Received November 13, 2019, accepted November 26, 2019, date of publication December 2, 2019,
date of current version December 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957054

A Clickthrough Rate Prediction Algorithm Based on Users' Behaviors

XI XIONG[1,4], CHUAN XIE[1], RONGMEI ZHAO[1], YUANYUAN LI[2], SHENGGEN JU[3], AND MING JIN[5]

[1]School of Cybersecurity, Chengdu University of Information Technology, Chengdu 610225, China
[2]West China Hospital, Sichuan University, Chengdu 610041, China
[3]College of Computer Science, Sichuan University, Chengdu 610065, China
[4]School of Aeronautics and Astronautics, Sichuan University, Chengdu 610065, China
[5]School of Computing and Information Systems, University of Melbourne, Parkville, VIC 3010, Australia

Corresponding author: Shenggen Ju (jsg@scu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 81901389, in part by the China Postdoctoral Science Foundation under Grant 2019M653400, and in part by the Sichuan Science and Technology Program under Grant 2018GZ0253, Grant 2019YFS0236, Grant 2018GZ0182, Grant 2018GZ0093, and Grant 2018GZDZX0039.

ABSTRACT Besides the ID class features, the advertisement click log file contains many significant features, which make the study of the advertisement clickthrough rate prediction more difficult. In this study, we convert original features into numerical meaningful ones, which reduce the sparsity and redundancy. In order to solve the problem of class imbalance, we propose a downsampling algorithm based on K-means model to classify large samples, then divide them into some sensible and rational features by the heuristic methods. To further improve the feature representation, we finally select and combine features by the Gradient Boosting Decision Tree model and process high-dimensional features by the logistic regression method. We conducted experiments on the dataset of Tencent SOSO and demonstrated that our approach outperforms the state-of-the-art baseline methods by 0.05% on average in terms of R2 and by 50.5% on average in terms of RMSE.

INDEX TERMS Clickthrough rate, class imbalance, gradient boosting decision tree, downsampling.

I. INTRODUCTION

The rapid development of the Internet provides a broad platform for the advertising industry. Internet advertising [1] has the advantages of wide user, strong interaction, and real-time flexibility, which makes the advertising industry gradually tilt toward it. Today, most of search engines earn profit by placing text advertisements next to search results. According to eMarketer data, all online advertising revenue (including PC and mobile advertising revenue) reached $207.3 billion in the global advertising market in 2018, with an increase of 13.76%. In 2018, Google is still the dominant player in the mobile advertising market. Its revenue in the mobile Internet advertising market is ¥63.7 billion, accounting for 40.59% of the market share; followed by Facebook, its Internet advertising revenue is ¥50.1 billion and the market share was 31.92%; Alibaba with a market share of 14.49%, ranking third.

Internet advertising is mainly divided into two categories: search advertising and exhibition advertising. The target pages for search advertisements are returned by the search engine, and advertiser competes in the auction of a certain product information (such as keywords [2]), and then occupies several positions of the advertisement according to the competition result. Exhibition advertisements primarily appear in the form of texts or images, which are typically targeted at web pages, application interfaces and video media. Compared with traditional advertisements, both kinds of Internet advertisements rely on the netizens' behavior history(e.g. the clicks and the browsing) [3], and we can obtain valuable information from them for promotion. Internet advertising shows larger marketing capability through multiple channels for information delivery.

There are four typical categories of Internet advertising commercial billing models [4], Cost Per Mille(CPM), Cost Per Click(CPC), Cost Per Action(CPA) and Return on Investment(ROI). At present, most of the mainstream payment in the advertising market adopts the CPC mode, that is, the advertising revenue of the third-party media is determined by the probability that the user may click on the advertisement(CTR), the price per click (CPC) as well as the total num

The associate editor coordinating the review of this manuscript and approving it for publication was Shirui Pan.

ber of clicks (N). The revenue of the advertising media can be expressed as $N \times CTR \times CPC$.

Predicting the clickthrough rate of an advertisement not only increases the revenue of the advertising media, but also maximizes users' satisfaction. Therefore, clickthrough rate prediction is a key issue in the field of Internet advertising for both marketing and research. It is extremely important to increase the probability of clicking on the advertisement in the case of a constant price per click.

Despite the large commercial value, clickthrough rate prediction still suffers from several major drawbacks: [5]–[7]:

1) The amount of data in the advertisement click log file, which contain a large number of category features with higher values, is large and grows fast. Tencent Search is a marketing service provided by Tencent to display customer advertising information on search pages and other application pages. According to ALEXA statistics, SOSO daily average pageviews [8] (also called daily average hits) can reach 7 million. The size of China's Internet advertising market is expected to reach ¥441.48 billion in 2020. Therefore, it is a great challenge to collect and process the huge amount data in the advertisement click log files.

2) The advertising clickthrough rate is small and there is a long tail distribution with a problem of class imbalance. There are a lot of advertisements on the page, but we only click on a small number of advertisements we are interested in. Therefore, the clicked advertisements are much less than the unclicked ones.

Based on the above questions, we are going to convert ID features and other features into meaningful numerical features first. Then, a downsampling algorithm based on K-means model is proposed to alleviate the problem of class imbalance, and the fusion of gradient promotion tree and logistic regression is used to predict the clickthrough rate of advertisement. Using the level of detail of the user input query words to predict the probability of the user drifting in interest in real time.

In this paper, we make the following contributions:

- We first preprocessed the large amount of sparse data which are involved in the historical logs of advertisement clicks. Then we use three different types of methods to calculate text similarity, which help to reduce the redundancy and sparsity of features, further improve the accuracy of feature extraction.
- We propose a model named *Ad*vertising *Cli*ckthrough rate prediction (*Ad-Click*). In order to improve feature expression and process a large amount of sparse data, the fusion of gradient lifting tree and logistic regression is adopted to predict the advertising click-through rate. In the model, a downsampling algorithm based on K-means model is exploited to alleviate the problem of class imbalance.
- We conducted extensive experiments to evaluate the performance of the proposed *Ad-Click* model on the data set of advertising log file based on Tencent SOSO, and the results demonstrate that our approach outperforms the state-of-the-art baseline methods in an effective and efficient fashion.

The remainder of this paper is structured as follows: Section II discusses the related works. Section III give some definitions in the paper and gives a description of the problem. Section IV introduces the method of advertising click rate prediction, which is divided into three parts. The experimental results are demonstrated in Section V. Lastly, the full study is concluded in Section VI.

II. RELATED WORKS

Accurate advertising clickthrough rate prediction will bring a good experience to users, and will bring greater economic benefits to website owners and advertisers [9], [10]. Thus, there are more and more researchers who calculate advertising in either industry or academia.

Effendi and Abbas [11] proposed a context-based clickthrough rate prediction algorithm based on linear regression. The algorithm uses context information to model the interaction between advertisements, and uses clustering algorithm to assist in the calculation of text similarity. The algorithm is simple, efficient, and easy to adjust parameters, but it is difficult to learn the complex relationship between features. Yin *et al.* [12] used the MapReduce-based coupled logistic regression model to predict the advertising clickthrough rate. The algorithm uses MapReduce's divide-and-conquer idea to process large amounts of sparse data. At the same time, the directional-based quasi-Newton optimization method is used to deal with non-convex and non-smooth data sets, but it is difficult for models to learn the complex relationships between features.

Kanagal *et al.* [13] proposed a focus matrix decomposition model to learn about the specific product preferences and related product information, and to solve the problem with sparse data caused by less user-product interaction. On the basis of the literature [13], Shan *et al.* [14] proposed a cubic matrix decomposition model, which decomposes the cubic matrix of the relationship among users, advertisements and web pages, and they use the value of the fitted matrix to predict CTR. Although the cubic matrix decomposition model adds a one-dimensional interaction, but the interactions depicted are still very limited, and the links between all the features of the advertisement cannot be fully exploited in the CTR prediction.

The research methods in the recommendation system [15] such as collaborative filtering are also applicable to the advertisement click rate prediction. Huo *et al.* [16] used collaborative filtering algorithm to find other neighboring pages similar to the page, and realized the click rate prediction, which was used as the basis for advertising recommendation. However, when the number of similar pages increased, the quality of the method would be seriously degraded.

The data in the advertisement history log data is sparsely severe, and the matrix decomposition algorithm, such as Singular Value Decomposition (SVD) [17], Factorization Machine (FM) [18] can solve problems such as

data sparseness. The FM algorithm was proposed in 2010, and it borrows the idea of implicit factor model and matrix decomposition, but expresses the inner product between hidden factors as the interaction between factors, which can predict the parameters that can be trusted in extremely sparse data. To this end, the literature [19] introduces the concept of the feature field based on the FM model, and proposed a domain-based factorization machine (FFM) model. The algorithm can solve the problem of data sparsity, but the number of model parameters is larger so that the model efficiency is lower and it is difficult to learn the high-order relationship between features. Guo*et al.* [20] used a fusion model based on factorization machine and neural network to predict the advertising clickthrough rate. The algorithm uses the characteristics of factorization machine and the architecture of neural networks to learn the complex relationship between features and improve the accuracy of the model prediction.

GBDT(Gradient Boosting Regression Tree), is a nonlinear model that can solve problems such as classification and regression, and can realize automatic mining and combination of features. Trofimov *et al.* [21] uses a gradient lifting tree to predict advertisement clickthrough rates. The gradient lifting tree is automatically combined, and a predicted value is formed from the root node of the tree to the leaf node. The gradient lifting tree can set the number of trees and the depth of each tree to solve the over-fitting problem. In the article [22] published by Facebook in 2014, the master's thesis of Harbin Institute of Technology proposed combining the gradient decision tree with the logistic regression model, and the output of the gradient decision tree as the input of the logistic regression model. Google proposed the FTRL model (Follow The Regularized Leader Proximal) [23], which is a generalized linear model that can update the parameters in the logistic regression model online and also shows good performance in clickthrough rate prediction.

At present, the CTR prediction model based on DNN fusion structure has gradually become a hot spot in academic and industrial research. Google researchers [24] have proposed a fusion structure, which subtly blends the linear model with the deep learning model. On this basis, the linear model is replaced by the FM model, and the DeepFM [25] is proposed. This structure combines the FM model and the DNN model to jointly train, and its advantage is that it does not require the support of feature engineering, and can also learn the interaction of low-order and high-order features at the same time. Considering that DNN is implicitly learning the interaction between features, not all feature intersections are valid. This model introduces Cross network [26], which can automatically, explicitly and limitly perform feature intersection, and do parallel training of cross network and common DNN network.

III. DEFINITIONS AND PROBLEM STATEMENT

In this section, we first present important notations and definitions, and then formalize the framework of *Advertising Clickthrough rate prediction (Ad-Click)*. To facilitate understanding, Table 1 describe the important notations used in this paper.

TABLE 1. Notations and their meanings.

SYMBOL	DESCRIPTION
t	a text
w	the query word
γ	the weight of the rational features
λ	the weight of the sensible features
comm(t_1, t_2)	the common part of text 1 and text 2
union(t_1, t_2)	the total number of words after text 1 and text 2 are deduplicated
count(w,t)	the number of times the query word appears in the text
size(t)	the total number of words of the text
tf(w,t)	the word frequency
idf	the frequency of the inverse text
$\overline{ctr}_{dis}$	the average advertising clickthrough rate displayed on the site
$\overline{ctr}_{pro}$	the average clickthrough rate of the advertisement served by the advertiser
ctr_{ij}	the advertising clickthrough rate of the site
clicks	the actual number of clicks for the advertisement
impression	the total number of impressions for the advertisement
wr_{ij}	the weight of the rational feature set
ws_{ij}	the weight of the sensible feature set
$qNum_i$	the number of the query words input by the user set
$kNum_j$	the number of keywords included in the advertisement j to be clicked set

A. QUERY RELEVANCE

Query relevance is used to describe the similarity between two attributes. When the content t_1 searched by the user is similar to the attribute t_2 of the delivered advertisement, the probability that the advertisement will be clicked will be greater.

B. RATIONAL FEATURES

We believes that each user is a mixture of sensibility and rationality, and the proportion of sensibility and rationality will change with time and place environment. Based on actual business analysis, rational features include user query relevance and advertisement placement.

C. SENSIBLE FEATURES

Sensible features include website attraction, advertisers' promotion, user gender, user age, and advertisement depth.

D. PROBLEM STATEMENT

Given a database containing the user's advertisement click log file, which includes User ID, Advertiser ID, Query word ID, Purchase keyword ID, Headline ID, AD description ID, etc., a series of numerical features F are obtained after extraction by artificial features. F can be represented by

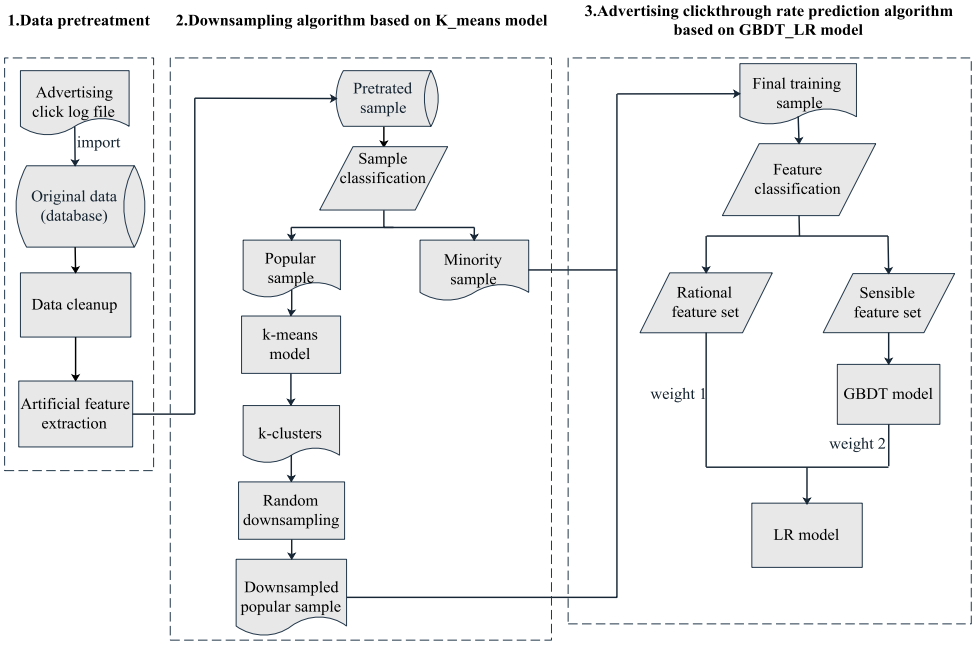


FIGURE 1. The framework of the proposed Ad-Click model.

a set with numerical features, such as Equation 1:

$$F\{sim, s_a, s_p, \cdots\} \tag{1}$$

where *sim* represents query relevance, s_a represents website attraction, and s_p is advertisers' promotion.

For easy training, we use the K-means model's downsampling algorithm to classify a large number of samples, and get a balanced positive and negative sample. During training, in order to better capture the user's features, we divide the obtained numerical features F into rational features(F_r) and sensible features(F_s). Then we input the sensible features(F_s) into the GBDT model and automatically construct new effective features and feature combinations(F_n). The rational features(F_r) and the new features(F_n) generated by GBDT model can be directly used as the input of LR to obtain the final advertisement click rate prediction(*CTR*). We define *CTR* prediction as Equation 2:

$$f(\gamma F_r, \lambda F_n) \rightarrow CTR \tag{2}$$

where γ represents the weight of the rational features, λ represents the weight of the sensible features, and *f(,)* is sigmoid function.

IV. METHOD

In order to achieve our goal presented in Section 3, we illustrate the framework of *Advertising Clickthrough rate prediction (Ad-Click)* in Fig.1.

Basically, the framework involves three steps:

1) We address the class imbalance problem by a downsampling-based algorithm.

2) We classify the features by the heuristic thinking and characterize the inductive features by the gradient tree.

3) We predict the advertising clickthrough rate via the logistic regression model with the input of combined characteristics and rational features.

A. FEATURE EXTRACTION

The feature extraction is utilized to reduce the redundancy and sparsity of features and further improve the effectiveness of feature expression. In this section, we describe the feature extraction in detail.

1) QUERY RELEVANCE

The text features of this paper belong to short text and are encrypted. At the same time, the advertising keywords, advertisement titles and advertisement descriptions are strongly related to each other.

Therefore, we uses Dice coefficient, Jaccard distance and TF-IDF to jointly calculate text similarity.

The Dice coefficient calculation formula is as shown in Equation 3:

$$dice(t_1, t_2) = \frac{2 * comm(t_1, t_2)}{size(t_1) + size(t_2)} \tag{3}$$

where $comm(t_1, t_2)$ represents the common part of text 1 and text 2, and $size(t_1)$, $size(t_2)$ respectively represents the total number of words of text 1,text 2.

The Jaccard distance [27] calculation formula is shown in Equation 4:

$$Jaccard(t_1, t_2) = \frac{comm(t_1, t_2)}{union(t_1, t_2)} \tag{4}$$

where $comm(t_1, t_2)$ represents the common part of text 1 and text 2, and $union(t_1, t_2)$ represents the total number of words after text 1 and text 2 are deduplicated.

The TF-IDF calculation formula is as shown in Equations 5-7:

$$tf(w,t) = \frac{count(w,t)}{size(t)} \tag{5}$$

$$idf = \log(\frac{size(t)}{count(w,t)+1}) \tag{6}$$

$$tf\_idf = tf * idf \tag{7}$$

Among them, *count(w,t)* indicates the number of times the query word appears in the text, *size(t)* represents the total number of words of the text, *tf(w,t)* represents the word frequency, and *idf* represents the frequency of the inverse text.

The final similarity calculation formula is shown in formula (8).

$$sim = \alpha * dice + \beta * Jaccard + \varphi * tf\_idf \tag{8}$$

2) WEBSITE ATTRACTION

Website attraction is the variance of the clickthrough rate of an advertisement displayed on a website. The calculation formula is as shown in formula (9).

$$s_a = \sqrt{\frac{\sum_{j=1}^{n}(ctr_{ij} - \overline{ctr}_{dis})^2}{n}} \tag{9}$$

where $\overline{ctr}_{dis}$ is the average advertising clickthrough rate displayed on the site, and ctr_{ij} represents the advertising clickthrough rate of the site.

3) ADVERTISERS' PROMOTION

Advertiser promotion is the clickthrough rate variance of the advertisement served by the advertiser. The calculation formula is as shown in equation (10).

$$s_p = \sqrt{\frac{\sum_{j=1}^{n}(ctr_{ij} - \overline{ctr}_{pro})^2}{n}} \tag{10}$$

where $\overline{ctr}_{pro}$ represents the average clickthrough rate of the advertisement served by the advertiser, and ctr_{ij} represents advertising clickthrough rate of the site.

4) ADVERTISING LOCATION

The advertising position is the actual location of the advertisement. Based on data analysis, advertising clickthrough rate is negatively correlated with the actual position of the advertisement and is not related to the relative position of the advertisement.

5) ADVERTISING USER ANALYSIS

Advertisements are directional, meaning that each advertisement has its own target group. This article uses the age and gender with the most clicks for a particular advertisement as the age and gender of the advertising user.

6) ADVERTISING CLICKTHROUGH RATE

Advertising clickthrough rate is the probability that a user is predicted to click on an advertisement when given a user and an advertisement. The calculation formula is as shown in Equation 11.

$$ctr = \frac{clicks}{impression} \tag{11}$$

where *clicks* represents the actual number of clicks for the advertisement, and *impression* represents the total number of impressions for the advertisement.

We believes that each user is a mixture of sensibility and rationality, and the proportion of sensibility and rationality will change with time and place environment. In order to locate users more accurately, we divides the features into two disjoint feature sets(rational features(F_r) and sensible features(F_s)), and then measures the feature set weights based on the level of detail of the user input query words.

The formula for calculating the weight of the rational feature set is shown in Equation 12.

$$wr_{ij} = \frac{qNum_i}{qNum_i + kNum_j} \tag{12}$$

The formula for calculating the weight of the sensible feature set is as shown in Equation 13.

$$ws_{ij} = \frac{kNum_j}{qNum_i + kNum_j} \tag{13}$$

where $qNum_i$ represents the number of the query words input by the user, $kNum_j$ indicates the number of keywords included in the advertisement j to be clicked, wr_{ij} indicates the weight of the rational feature set when a given user i and the advertisement j to be clicked, ws_{ij} indicates that the weight of the sensible feature set when given user i and the advertisement j to be clicked.

B. DOWNSAMPLING BASED ON K-MEANS MODEL

According to the data analysis, the ratio of positive and negative samples in the training sample is 1:8, which is a class imbalance. We proposed a downsampling algorithm based on K-means model to solve the class imbalance problem from the data level, and at the same time, alleviate the problem of useful information loss caused by downsampling.

The algorithm first uses the K-means clustering method to divide the majority of the samples into k-clusters, where k is equivalent to the number of negative samples. Then, for each cluster, calculate their sample centers. Finally, each cluster sample center is replaced with the original majority sample as a new positive training sample. K-means algorithm is a typical greedy algorithm, the input is data set D = $\{x_1, x_2, \ldots x_m\}$, the number of clusters k(equivalent to the number of negative samples), and the maximum number of iterations N, and the output is cluster division C=$\{C_1, C_2, \ldots C_k\}$.

If represented by a data expression, assuming the cluster is divided into $(C_1, C_2, \ldots C_k)$, then our goal is to minimize the squared error E. The calculation formula is as shown in Equation 14.

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\| \tag{14}$$

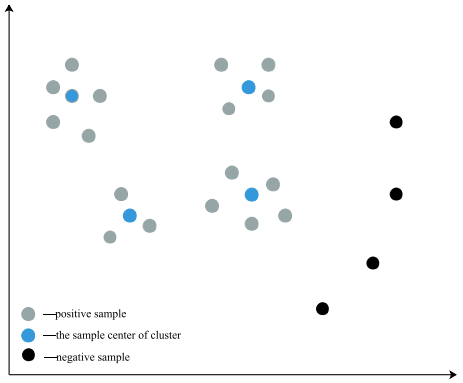


FIGURE 2. An illustration of K-means clustering based downsampling algorithm.

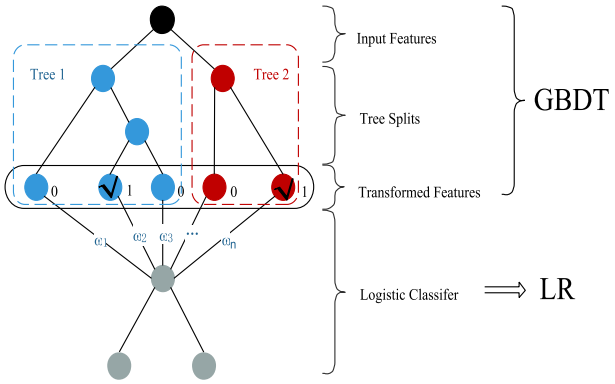


FIGURE 3. GBDT-LR model.

where μ_i is the sample center of cluster C, and the expression is as shown in Equation 15.

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{15}$$

Figure 2 shows a very tangible example of the method in which gray circle samples represent a majority of samples, for a total of 17 samples. The black circle sample represents a minority sample with a total of four samples. In order to reduce the imbalance, we uses the K-means clustering algorithm to divide the majority of the sample into 4 clusters, and calculate the sample center of each cluster, as shown by the blue circle sample. Finally, the four blue color samples and the black circle sample can be used as a new training set for classification learning.

The algorithm does not simply sample random samples of most classes, avoiding the loss of important samples. By clustering most of the sample sets, the method of finding the sample center not only preserves the distribution information of the original samples, but also reduces the number of majority samples and reduces the imbalance between the classes, so the algorithm can effectively improve classifier performance on unbalanced classification problems.

C. GRADIENT BOOSTING DECISION TREE + LOGISTIC REGRESSION

In this paper, the structural characteristics of GBDT are used to select and combine features, and the LR model is used to process a large amount of sparse data. The purpose is to improve the accuracy of model prediction and the efficiency of model training.

GBDT first trains the original training data to get a second classifier. Of course, we also need to use grid search to find the best combination of parameters. Different from the usual practice, when the GBDT is trained to make predictions, the output is not the final binary classification probability value, but the position of the leaf node to which the predicted probability value calculated by each tree in the model belongs is recorded.

Figure 3 is a simple GBDT-LR model structure that GBDT has two weak classifiers, which are represented by blue and red parts respectively. The number of leaf nodes of the blue weak classifier is 3, and the leaf nodes of the red weak classifier is 2, and the prediction result of 0-1 in the blue weak classifier falls to the second leaf node, and the prediction result of 0-1 in the red weak classifier falls to the second leaf node. Then we mark that the prediction result of the blue weak classifier is [0 1 0], and the prediction result of the red weak classifier is [0 1]. And since each weak classifier has one and only one leaf node outputs the prediction result, in a GBDT with n weak classifiers and a total of m leaf nodes, each training data is converted into 1*m dimensional sparse vector, and there are n elements of 1, and the remaining m-n elements are all 0.

Each sensible features(F_s) can be transformed by GBDT into new feature vectors (Transformed Features [0 1 0 0 1] in Fig.3). After a well-trained GBDT, it can be regarded as a series of high-quality feature processing rules (transformation, combination, filtering, etc.), and finally realizes the mapping of new and old features. All output leaf nodes of GBDT constitute a new feature vector(F_n) which is more conducive to the learning and training of LR model.

The new features(F_n) obtained through GBDT are used as the input of the LR model. LR linearly combines the features, and then maps the combined result into 1 or 0 through the sigmoid function layer is the probability that a user clicks on this advertisement(*CTR*).

V. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

1) DATASETS

The data in this article is the advertisement click log file of Tencent SOSO. The specific data description is shown in Table 2-3.

From the data analysis and actual business analysis, there are unknown users and errors in the original data. We makes a statistical analysis of the data after data cleansing. The statistical table of the number of samples under different categories is shown in Table 3.

TABLE 2. Field description.

Training center field description

User ID	Query word ID	Headline ID
Ad description ID	Advertising link	Advertiser ID
Purchase keyword ID	Ad depth	Ad placement
Ad clicks	Ad impressions	

TABLE 3. Sample number under different class.

Category	Number of samples
Positive class(click)	38491
Negative class(non-click)	241027

2) BASELINES

In order to verify the effectiveness of the proposed algorithm, we selects two traditional algorithms and three new algorithms for progressive comparison experiments. The comparison algorithm and parameter settings are as follows.

Logistic regression algorithm: The maximum number of iterations is 700 and the learning rate is 0.05.

Gradient lifting tree: the maximum number of sub-models is 700, the learning rate is 0.05, the loss function is the mean square error function, the maximum number of features considered in the division is the square root of the total feature number, the maximum depth of the decision tree is 8, and the internal nodes are subdivided. The minimum number of samples required is 20, and the minimum number of samples for leaf nodes is 25.

Document [10] (FNN algorithm): 12 nodes in the input layer, 300 nodes in the first layer, 100 nodes in the second layer, and 1 node in the output layer. The learning rate is 0.05. The activation function of the input layer node is linear. Function, hidden layer activation function is sigmoid function.

Document [22] (GBDT-LR algorithm): the maximum number of sub-models is 700, the learning rate is 0.05, the loss function is the mean square error function, the maximum number of features considered in the division is the square root of the total feature number, and the maximum depth of the decision tree is 8, internal nodes The minimum number of samples required for subdivision is 20, and the minimum number of samples for leaf nodes is 25.

Document [19] (FFM algorithm): The learning rate is 0.05, the loss function is Logloss, and the number of iterations is 700.

3) EVALUATION METRICS

Davies-Bouldin Index(DMI), is the maximum ratio of the sum of the distances within the class of any two categories and the distance between the centroids of the two clusters. The DMI indicator calculation formula is shown in Equation 16.

$$DBI = \frac{1}{N}\sum_{i=1}^{N}\max_{j\neq i}\left(\frac{\overline{S_i}+\overline{S_j}}{\|\omega_i-\omega_j\|_2}\right) \tag{16}$$

where $\overline{S_i}+\overline{S_j}$ represents the sum of the average distances from all points in the cluster to the centroid of the cluster,

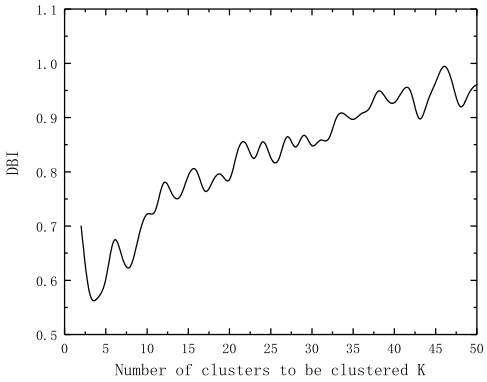


FIGURE 4. The DBI under different k values.

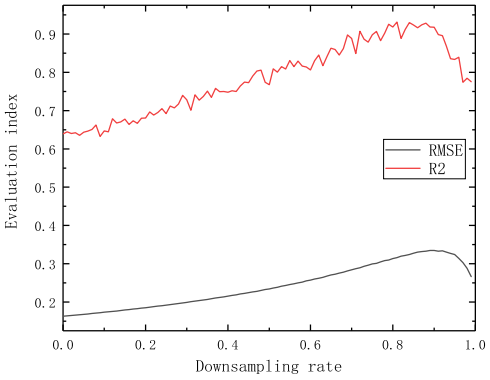


FIGURE 5. The RMSE and R2 under different sampling ratios.

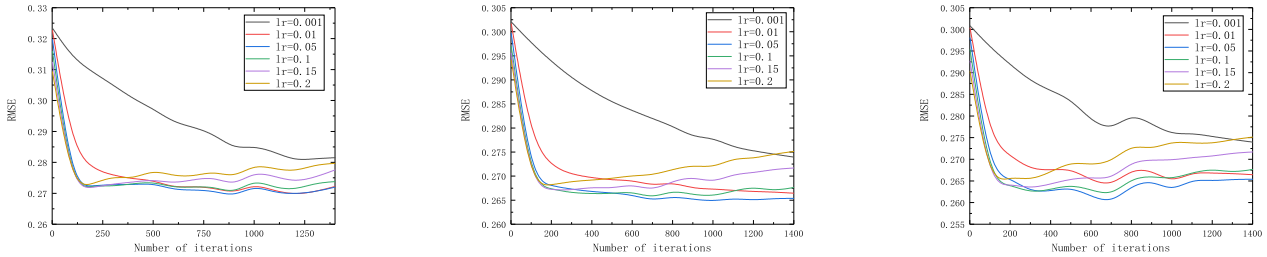
$\omega_i-\omega_j$ represents the distance between the centroids of the two categories. The maximum value in each set of proportions is selected (i.e. the worst set is selected) and finally divided by the number of categories. The smaller the DBI value obtained, the smaller the clustering result is, and the different clusters are far apart. That is, the smaller the distance within the class, the greater the distance between classes.

The root mean square error is the sum of the squared values of the difference between the predicted and the true value of the click rate, divided by the square root of the size of the test set, which can measure the degree of dispersion of the predicted value, and thus can measure the stability of the prediction of the algorithm. The RMSE indicator calculation formula is shown in Equation 17.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(pctr_i - tctr_i)^2}{n}} \tag{17}$$

The R2 index is a method of comparing the predicted values of the model with the prediction without using the model and using the mean as the predicted value to measure the prediction ability of the model. The formula for calculating the R2 index is as shown in Equation 18.

$$R2 = 1 - \frac{\sum_{i=1}^{n}(pctr_i - tctr_i)^2}{\sum_{i=1}^{n}(\overline{ctr} - tctr_i)^2} \tag{18}$$

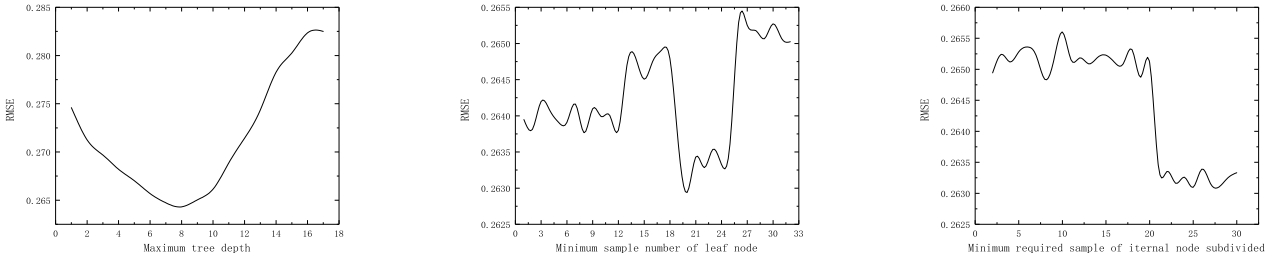


(a) The loss function is huber and the maximum number of features is sqrt

(b) The loss function is ls and the maximum number of features is all

(c) The loss function is ls and the maximum number of features is sqrt

FIGURE 6. The RMSE under different learning rates and iterations.



(a) The RMSE under different tree depths

(b) The RMSE under different minimum samples of leaf node

(c) The RMSE under different minimum sample number required for redivision of internal nodes

FIGURE 7. The RMSE under different minimum sample number required for redivision of internal nodes.

where $tctr_i$ represents the predicted click rate of the sample i, $pctr_i$ indicates the true click rate of the sample i, and n indicates the number of samples of the test set.

B. STUDY ON THE PARAMETERS EXPERIMENTS

We first uses the K-means model to cluster the mass class samples, and the purpose is to learn the distribution characteristics of the mass class samples. The variation curve of DBI under different k values is shown in Figure 4.

From Fig.4, when K is 3, the DBI value is at least 0.551. Therefore, in the subsequent random downsampling rate parameter experiment, the number of clusters to be clustered is set to 3.

The downsampling rate parameter experiment is performed by random sampling algorithm and GBDT-LR model. The variation curves of RMSE and R2 at different sampling ratios are shown in Figure 5.

According to the definition in the evaluation metrics, only when R2 takes a larger value and RMSE takes a smaller value, the GBDT-LR model works best at this time. From Fig.5, when the downsampling rate is 0.28, the above conditions are met simultaneously. And our experiment proves that when the downsampling rate is 0.28, the ratio of positive and negative samples in the training set is 1:2. At this time, the number of positive samples is 38491, and the number of negative samples is 76982.

In the training part of the GBDT model parameters, we first carries out the training of the process parameters, and then the training of the base classifier parameters.

The RMSE curve of the test set at different learning rates and iterations is shown in Figure 6.

From Fig. 6, when the loss function of the GBDT model is ls, the maximum number of features is sqrt, the learning rate is 0.05, and the number of base classifiers is 700, the model works best.

The loss function is ls, the maximum number of features is sqrt, the learning rate is 0.05, the number of base classifiers is 700, and the variation curve of RMSE at different tree depths, the least sample of leaf nodes, and the minimum number of samples required for internal nodes are further divided. Figure 7 shows the figure.

From Fig. 7, when the maximum tree depth of the GBDT model is 8, the minimum number of samples for the leaf nodes is 20, and the minimum number of samples required for the internal nodes is 25, the model works best.

The parameters of this paper are obtained by the downsampling algorithm parameter experiment and GBDT parameter experiment based on K-means model. The parameter list in this paper is shown in Table 4.

C. EXPERIMENTAL RESULTS

In this paper, the effectiveness of our work is verified from two aspects: feature extraction and class imbalance. Finally, the effectiveness of the algorithm-based advertising clickthrough rate prediction algorithm based on user behavior is verified.

TABLE 4. Parameter settings.

Parameter name	Value
Number of clusters to be clustered K	3
Downsampling rate	0.28
Number of submodels	700
learning rate	0.05
Downsampling rate	0.8
Loss function	ls
The maximum number of features to consider when dividing	sqrt
Decision tree maxumum depth	8
The minimum number of samples required for internal node subdivision	25
Leaf node minimum sample number	20

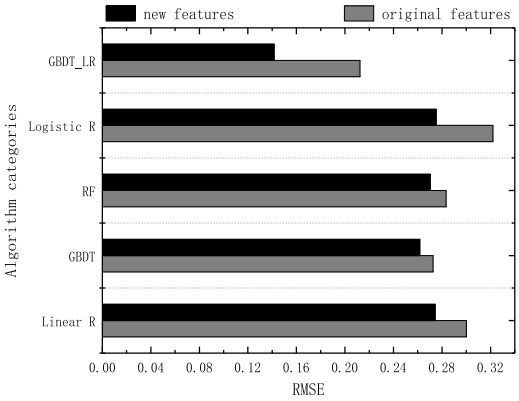


FIGURE 8. The RMSE under different features.

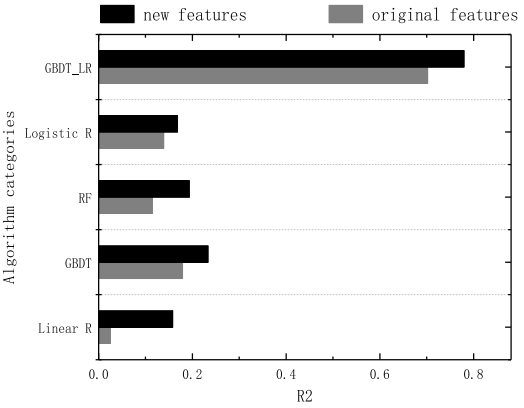


FIGURE 9. The R2 under different features.

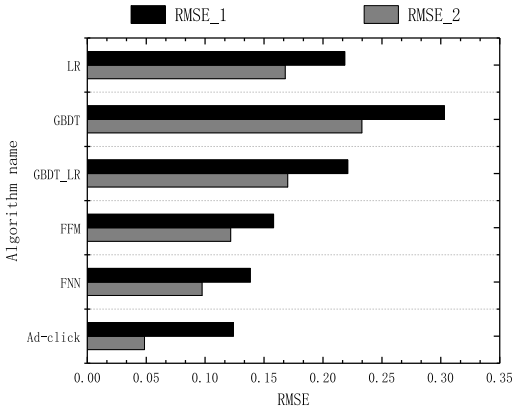


FIGURE 10. The RMSE under different data sets.

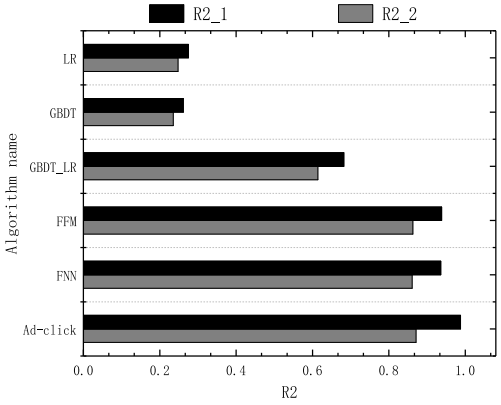


FIGURE 11. The R2 under different data sets.

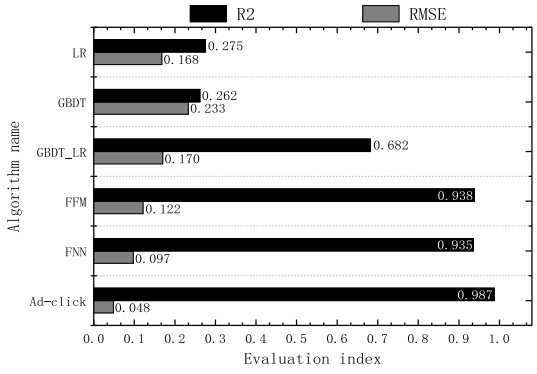


FIGURE 12. The RMSE and R2 under different algorithms.

1) FEATURE EXTRACTION

In this paper, logistic regression algorithm, random forest algorithm, gradient lifting tree algorithm, linear regression algorithm and GBDT-LR algorithm are used to verify the feature extraction efficiency from RMSE and R2.

From Fig.8 and Fig.9, the new features are superior to the original features in terms of RMSE and R2, indicating the effectiveness of the feature processing in this paper.

2) CLASS IMBALANCE

In this paper, logistic regression algorithm, gradient lifting tree algorithm, GBDT-LR algorithm, FFM algorithm and FNN algorithm are used to verify the effectiveness of the K-means model based downsampling algorithm from RMSE and R2.

From Fig.10 and Fig.11, the data set processed by the K-means model based downsampling algorithm is superior to the original data set in terms of RMSE and R2, indicating the effectiveness of the downsampling algorithm based on K-means model.

3) ALGORITHM OF THIS PAPER

In this paper, logistic regression algorithm, gradient lifting tree algorithm, GBDT-LR algorithm, FFM algorithm and

FNN algorithm are used to verify the effectiveness of the proposed algorithm from RMSE and R2.

From the Fig.12, the proposed algorithm*(Ad-Click)* outperforms the predecessor algorithm and the classical algorithm in terms of RMSE and R2, and illustrates the effectiveness of the proposed algorithm.

VI. CONCLUSION

Based on the current research difficulties,we first performs feature extraction based on experimental data and actual business analysis, with the aim of reducing feature redundancy and feature sparsity and improving feature expression. Next, based on user behavior analysis, the user inputs the query word. The degree of detail is used to predict the probability of user drift of interest in real time. Then, an downsampling algorithm based on K-means model is proposed to alleviate the problem of useful information loss and class imbalance caused by downsampling. In addition, further use GBDT model to choose and combine features to improve feature expression. Finally, a large number of high-dimensional data are processed by logistic regression model.

Because the GBDT model is used to learn the complex relationship between features, the time performance of the proposed algorithm is poor when dealing with large-scale training data. Based on the above analysis, future work will focus on two parts, one is feature extraction, mainly considering the user's historical click information, which can reflect the user's preferences, and the accuracy of the click rate prediction will be significantly improved. The second is time performance, mainly considering limiting the input characteristics of LR model - the number of leaf nodes of GBDT.

REFERENCES

[1] D. Liu and V. Mookerjee, "Advertising competition on the Internet: Operational and strategic considerations," *Prod. Oper. Manage.*, vol. 27, no. 5, pp. 884–901, May 2018.

[2] H. Zhou, M. Huang, Y. Mao, C. Zhu, and X. Zhu, "Domain-constrained advertising keyword generation," Tsinghua Univ., Beijing, China, Tech. Rep. 10.1145/3308558.3313570, 2019.

[3] L. Bing, W. Lam, T.-L. Wong, and S. Jameel, "Web query reformulation via joint modeling of latent topic dependency and term context," *ACM Trans. Inf. Syst.*, vol. 33, no. 2, Feb. 2015, Art. no. 6.

[4] L. Miralles-Pechu, D. Rosso, F. Jimez, and J. M. Garc, "A methodology based on Deep Learning for advert value calculation in CPM, CPC and CPA networks," *Soft Comput.*, vol. 21, no. 3, pp. 651–665, Feb. 2016.

[5] X. Ling, W. Deng, G. Chen, H. Zhou, L. Cui, and S. Feng, "Model ensemble for click prediction in bing search ads," in *Proc. Int. Conf. World Wide Web Companion*, Apr. 2017, pp. 689–698.

[6] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, 2000.

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

[8] Baidubaike. *Soso*. Accessed: Oct. 2019. [Online]. Available: https://baike.baidu.com/item/%E8%85%BE%E8%AE%AFsoso%E6%8E%A8%E5%B9%BF/15398552

[9] K. Gai, X. Zhu, H. Li, K. Liu, and Z. Wang, "Learning piece-wise linear models from large scale data for ad click prediction," Apr. 2017, *arXiv:1704.05194*. [Online]. Available: https://arxiv.org/abs/1704.05194

[10] W. Zhang, T. Du, and J. Wang, "Deep learning over multi-field categorical data," Univ. College London, London, U.K., Tech. Rep. 10.1007/978-3-319-30671-1_4, 2016.

[11] M. J. Effendi and S. A. Ali, "Click through rate prediction for contextual advertisment using linear regression," *Int. J. Comput. Sci. Inf. Secur.*, vol. 14, no. 11, pp. 1–8, 2017.

[12] N. Yin, H. Li, and H. Su, "CLR: Coupled logistic regression model for CTR prediction," in *Proc. ACM Turing 50th Celebration Conf.-China*, May 2017, Art. no. 21.

[13] B. Kanagal, A. Ahmed, S. Pandey, V. Josifovski, L. Garcia-Pueyo, and J. Yuan, "Focused matrix factorization for audience selection in display advertising," in *Proc. IEEE Int. Conf. Data Eng.*, Apr. 2013, pp. 386–397.

[14] L. Shan, L. Lei, S. Di, and X. Wang, *CTR Prediction for DSP With Improved Cube Factorization Model from Historical Bidding Log*. Cham, Switzerland: Springer, 2014.

[15] Z. Li, F. Xiong, X. Wang, H. Chen, and X. Xiong, "Topological influence-aware recommendation on social networks," *Complexity*, vol. 2019, Feb. 2019, Art. no. 6325654.

[16] X. Huo, H. E. Liang, and Y. Yang, "An advertisement collaborative recommendation algorithm without position bias," *Comput. Eng.*, vol. 12, pp. 39–44, Dec. 2014.

[17] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, *A Multilinear Singular Value Decomposition*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2000.

[18] S. Rendle, "Factorization machines," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 995–1000.

[19] Y. Juan, D. Lefortier, and O. Chapelle, "Field-aware factorization machines in a real-world online advertising system," Int. World Wide Web Conf. Steering Committee, Perth, WA, Australia, Tech. Rep. 10.1145/3041021.3054185, 2017.

[20] H. Guo, R. Tang, Y. Ye, and X. He, "Holistic neural network for CTR prediction," in *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp. 787–788.

[21] I. Trofimov, A. Kornetova, and V. Topinskiy, "Using boosted trees for click-through rate prediction for sponsored search," in *Proc. Int. Workshop Data Mining Online Advertising Internet Economy*, Aug. 2012, Art. no. 2.

[22] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, and J. Q. Candela, "Practical lessons from predicting clicks on ads at facebook," in *Proc. 8th Int. Workshop Data Mining Online Advertising*, Aug. 2014, pp. 1–9.

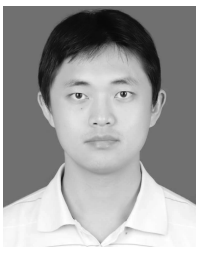
[23] H. B. Mcmahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, and D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos, and J. Kubica, "Ad click prediction: A view from the trenches," in *Proc. Acm SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2013, pp. 1222–1230.

[24] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, R. Anil, Z. Haque, L. Hong, V. Jain, and X. Liu, and H. Shah, "Wide & deep learning for recommender systems," in *Proc. 1st Workshop Deep Learn. Recommender Syst.*, Sep. 2016, pp. 7–10.

[25] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, "DeepFM: A factorization-machine based neural network for CTR prediction," Shenzhen Graduate School, Harbin Inst. Technol., Harbin, China, Tech. Rep. 10.24963/ijcai.2017/239, 2017.

[26] R. Wang, B. Fu, G. Fu, and M. Wang, "Deep & cross network for ad click predictions," in *Proc. ADKDD*, Aug. 2017, Art. no. 12.

[27] C.-M. Hwang, M.-S. Yang, and W.-L. Hung, "New similarity measures of intuitionistic fuzzy sets based on the jaccard index with its application to clustering," *Int. J. Intell. Syst.*, vol. 33, no. 8, pp. 1672–1688, Aug. 2018.



XI XIONG received the B.S. and M.S. degrees from the Beijing Institute of Technology and the Ph.D. degree in information security from Sichuan University, Chengdu, China, in 2013. He is currently an Assistant Professor with the School of Cybersecurity, Chengdu University of Information Technology, Chengdu, and also a Postdoctoral Fellow with the School of Aeronautics and Astronautics, Sichuan University. His research interests include social computing, web data mining, and natural language processing.

CHUAN XIE was born in Chongqing, China, in 1994. She is currently pursuing the master's degree with the School of Cyberspace Security, Chengdu University of Information Technology. Her research concerns natural language processing, recommendation systems, and sentiment analysis.

RONGMEI ZHAO received the degree from the School of Cyberspace Security, Chengdu University of Information Technology, with a focus on natural language processing and recommendation systems.

YUANYUAN LI received the Ph.D. degree in psychiatry from Sichuan University, China, in 2012. She is currently an Associate Professor with the Mental Health Center, West China Hospital, Sichuan University, Chengdu, China. Her research interests include data mining and sentiment analysis.

SHENGGEN JU received the B.Sc., M.E., and Ph.D. degrees in computer science from Sichuan University, Chengdu, China, in 1997, 2005, and 2010. He is currently a Full Professor with the College of Computer Science, Sichuan University, China. His research interests include natural language processing, big data mining, and recommender systems.

MING JIN received the master's degree from the School of Computing and Information Systems, University of Melbourne, Parkville, Melbourne, Australia, in 2019. Since November 2018, he has been working in the field of industrial data mining and software development. His research focuses on data mining and machine learning.