Feature Engineering for Search Advertising Recognition

Yue Sun[1], Guowei Yang[1,2]

1. School of electronic information, Qingdao University
2. College of information engineering, Nanjing Audit University

Qingdao, Nanjing, China
yue_er94@163.com, ygw_ustb@163.com

Abstract —Feature engineering is one of the key technologies in the research of search advertising recognition. Most of the existing search advertising methods are selected according to the prior knowledge, which is too subjective to be popularized. Taking the advertisement of Ali search advertising as the research object, a feature processing method based on the pre-analysis of store and user data is put forward, and then the conversion rate is predicted with XGBoost (eXtreme Gradient Boosting). Experiments show that compared with other priori Feature Engineering, the proposed method can significantly improve the prediction results.

Keywords—feature engineering; pre-analysis; XGBoost; search advertisement recognition

I. Introduction

Search advertising is a common way of Internet marketing. Businesses purchase specific keywords according to the characteristics of the goods. When users enter these keywords, the corresponding advertising goods will be displayed in the pages that the user sees. The conversion rate of search advertisements is used as an index to measure the effect of advertising transformation, that is, the probability of advertising products being bought by users after clicking. With the rapid development of Internet, search advertising has become more and more popular in Internet advertising, and has become one of the most important business models in the Internet industry.

In Feature Engineering, traditional feature processing methods are linear combination of original features, one-hot coding and so on. It is difficult to improve the recognition rate in traditional ways. This paper, taking Ali search advertising as the research object, proposes a feature processing method based on store and user data pre-analysis, which aims to pre-analyze the features, that is, the first prediction processing of the features of users and stores, and as a new feature. The results of this experiment take the size of Logarithmic Loss (Logless) as the evaluation standard. In general, we must correctly handle the features and reduce the Logless value as much as possible, which is the next problem we need to solve.

II. Theory of Feature engineering

A. Feature engineering concept

Feature engineering is the most important concept in machine learning field. It can generally be considered as the work of designing feature sets for machine learning applications. For feature-based machine learning methods, the selection of feature sets determines the extreme value of the algorithm to be iterative to the optimal condition. It is also the quality of the design of the feature system that determines the performance of the whole model in essence. Therefore, how to extract effective and highly correlated features from the existing historical data is the most important issue that we should consider next. The quality of feature design will directly affect the prediction effect of the prediction model, and also have a great impact on the spatio-temporal complexity and convergence speed of the model [1].

Data is the carrier of information, but the original data contains a lot of noise, and the expression of information is not concise enough. Therefore, the purpose of feature engineering is to use a series of engineering activities to express this information in a more efficient way of coding, and its purpose is to get better training data. Using the information represented by the feature, the information loss is less, and the law contained

978-1-5386-6243-4/19/$31.00 ©2019 IEEE

in the original data is still preserved. In addition, new coding methods also need to minimize the impact of uncertainties in raw data [4].

B. *Common method of Feature Engineering*

In general, feature selection refers to selecting a feature set that obtains the best performance of the corresponding model and algorithm. The commonly used methods in engineering are as follows:

- Calculate the correlation between each feature and response variable: The commonly used methods in engineering include Pearson coefficient and mutual information coefficient, Pearson coefficient can only measure linear correlation, and mutual information coefficient can measure various correlations well. The calculation is relatively complicated. Fortunately, many toolkits include this tool (such as sklearn's MINE). After getting correlation, you can sort the selection features [2-3].
- The model of a single feature is constructed, and the feature is selected by the accuracy of the model, and then the final model is trained when the target features are selected.
- The feature is selected by the L1 regular term: the L1 canonical method has the characteristics of sparse solution, so naturally has the characteristics of feature selection, but it should be noted that the features not selected by L1 do not represent unimportant because of two features with high correlation. Only one may be retained. If it is important to determine which features are important, cross-checking with the L2 regular method is again required.
- After feature selection, features are selected again: if the user id and user characteristics are combined to obtain a larger feature set and then select a feature, this practice is more common in recommendation system sand advertisement systems. This is also known as mega rating or even billion rating. The main source of the feature is that the user data is sparse and the combined features can take into account both the global model and the personalized model.
- Through depth learning to feature selection: at present, this means is becoming a means with the popularity of deep learning, especially in the field of computer vision, due to the ability of deep learning to have automatic learning features, which is also the reason for deep learning called unsupervised feature learning. After selecting the characteristics of a neural layer from the deep learning model, it can be used to train the final target model [5].

C. *Alternative proof of shop features and user needs*

Classification tree is used to classify in CART（Classification And Regression Tree）. Gini index is used to select the best segmentation features, and each time is a two-class classification. The Gini index is a concept similar to entropy, which indicates the probability of a random selected sample in the set of samples. The smaller the Gini exponent is, the smaller the probability of the sample being selected in the set. For a random variable X with probability of K state, its Gini exponent is defined as follows:

$$Gini(x) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2 \qquad (1)$$

According to the formula, the Gini coefficient of Bernoulli distribution can be obtained as follows:

$$Gini(x) = \sum_{k=1}^{K} p_k(1-p_k) = 2p(1-p) \qquad (2)$$

For the training data set D, if there is a total of K categories, C_k represents the sample subset of class k, $|C_k|$ is the size of C_k, and $|D|$ is the size of D, then the Gini coefficient of the set D is:

$$Gini(D) = \sum_{k=1}^{K} \frac{|C_k|}{|D|}\left(1 - \frac{|C_k|}{|D|}\right)$$

$$= 1 - \sum_{k=1}^{K} \frac{|C_k|}{|D|} \qquad (3)$$

Suppose that the data is now segmented by feature A. If feature A is a discrete feature, then D is divided into D1 and D2 according to a possible value a of A.

$$D1=\{D|A=a\}，D2=\{D|A\neq a\} \qquad (4)$$

Then we get a quantity Gini (D, A), which is similar to conditional entropy, that is, the Gini index of D under the condition of known characteristic A:

$$Gini(D,A)=\frac{|D1|}{|D|}Gini(D1)+\ \frac{|D2|}{|D|}Gini(D2) \qquad (5)$$

The greater the Gini (D, A) value, the greater the uncertainty of the sample, which is similar to entropy, so the criterion for selecting feature A is the smaller the Gini (D, A) value, the better.

So we can get the Gini coefficients of advertisement information, user information, context information and store information to judge the optionality of features. Finally, the Gini coefficient of customer demand and store characteristics is relatively low, which is 0.32 and 0.35 respectively. For this reason, we can pre_analyze the data of shop and user, and produce two new features after XGBoost, that is, the conversion rate of different users and different stores, which is predicted_shop_score and predicted_user_score.

III. EXPERIMENT AND ANALYSIS

A. *experimental data*

In order to ensure the rigorousness and accuracy of the experimental results, the experimental data we use comes from the Alibaba International Advertising Algorithm Competition. The data used this time includes five categories (basic data, advertising product information, user information, context information, and store information). The basic data table provides the most basic information about search advertising, as well as the "whether to trade" tag. Four types of data, such as advertising product information, user information, contextual information, and store information, provide supplementary information that may help in the conversion rate estimate. The experimental code is implemented in python and optimized with XGBoost. Each piece of data has its own instance_id, a total of 478,138 pieces of data. The characteristics of specific data are as follows:

TABLE I. BASIC DATA

Fields
instance_id
is_trade
item_id
user_id
context_id
shop_id

TABLE II. ADVERTISING PRODUCT INFORMATION

Fields
item_id
item_category_list
item_property_list
item_brand_id
item_city_id
item_price_level
item_sales_level
item_collected_level
item_pv_level

TABLE III. USER INFORMATION

Fields
user_id
user_gender_id
user_age_level
user_occupation_id
user_star_level

TABLE IV. CONTEXT INFORMATION

Fields
context_id
context_timestamp
context_page_id
predict_category_property

TABLE V. STORE INFORMATION

Fields
shop_id
shop_review_num_level
shop_review_positive_rate

shop_star_level
shop_score_service
shop_score_delivery
shop_score_description

B. *Feature processing of experimental data*

The real data collected from the experimental data processing often result in unpractical data due to the low degree of correlation. In order to make data analysis or data prediction work scientific and reliable, these "dirty data" often can not be used directly. Before this, we need to extract the original data from these data.

1) User information processing

This part of the user's natural attribute features and the user's registration information can be extracted directly from the original data, but the user's registration information is not perfect, some users' sex, age data are unknown, and some users are home users. So we mainly solve the conversion rate according to the data given, that is, in the case of all given data and labels, we can find the conversion rate of the purchase behavior of different sex and the conversion rate of purchase behavior at different age grades, so as to produce new characteristics for us to use.

2) Advertising goods information processing

As for the attribute list of advertising goods, because of its classification, we need to calculate the number of attributes and the weights of attributes. For the list list of items, the first column of the list of items calculated is exactly the same, so we only have the latter two and divide them into two features, and we are prepared for the later code. We count the number of advertising products, the brand number of advertising products, and the city code of advertising products, respectively, and calculate the conversion rate.

3) Context information processing

Whether users buy goods or not has a great relationship with time, for example, double eleven Shopping Festival, Spring Festival, Christmas, Valentine's day and so on will use the consumer to produce the purchase behavior. Therefore, the display time of the commodity is also an important attribute. In this regard, we convert the time to the normal format based on the original data given, and add the weekly attributes to add new features. For the prediction category attributes of commodities, because of their missing values, we can count the number, maximum and sum of category attributes.

4) One_hot coding processing

In many machine learning tasks, features are not always continuous values, but may be classified values. It will be much more efficient if the characters are represented by numbers. But when converted to digital representation, data can not be directly used in classifiers. Because classifier often acquiesce data is continuous and orderly. In order to solve these problems, a feasible solution is to adopt One-Hot Encoding. Also called an effective encoding, the method is to use the N bit state register to encode the N state, each state is made by his independent register bit, and at any time, only one of them is valid.

5) Initial prediction processing

In order to solve the drawbacks of traditional feature processing, new innovation points are added. For this purpose, we perform prediction processing on the relevant data of shop and user and generate two new features, namely predicted_shop_score and predicted_user_score.

User data is very important, and conversion is closely related to the user's wishes, so it is very important to deal with user-related features. The user's gender, age, occupation, and star rating are all relevant characteristics of the user. These characteristics directly affect the probability of the user's willingness to purchase an advertisement commodity. So for these characteristics of the user, we perform the following processing: Combine the relevant features of each user, and then use XGBoost to calculate the conversion rate of the combined data and then regenerate the one-dimensional features.

The pros and cons of a store also have a significant impact on the conversion rate of advertising products. For the same product, the conversion rates in different stores are definitely different. The number of shops evaluated, the rate of shops, the store's star rating, and the store's the characteristics of service attitude, store's logistics service, and store's description match all affect the conversion rate of the store. The conversion rate of the store is closely related to the conversion rate of the advertising product. Therefore, research on the conversion of search advertising by the relevant characteristics of the store the rate has a significant impact. For these features of the store, we perform the following processing: Combine the relevant characteristics of each store, then use XGBoost to find the conversion rate of the combined data and regenerate the one-dimensional features.

C. Evaluation index

The logarithmic loss (Logloss) is used to evaluate the model effect formula as follows:

$$Logloss=-\frac{1}{N}\sum_{i=1}^{N}(y_i log(p_i)+(1-y_i)log(1-p_i)) \qquad (6)$$

Where N represents the number of test set samples, y_i represents the true tag of the i sample in the test set, and p_i represents the estimated conversion rate of the i sample.

D. Experimental model and result

XGBoost is a tool for massively parallel boosted tree. It is the fastest and the best open source Boosted tree toolkit at present. It is more than 10 times faster than the common toolkit. The most basic component of Boosted tree is called the regression tree (regression tree), also called CART, and CART will assign the input to each leaf node according to the input property, and each leaf node corresponds to a real number score[6].

The algorithm can establish the model by distribution, and select the direction of gradient descent in the continuous update iteration to ensure the optimal prediction results. The algorithm flow is as follows:

```
（1）  $F_0(x) = argmin_\rho \sum_{i=1}^{N} L(y_i, \rho)$
（2）  $for\ m = 1\ to\ M\ do$
（3）  $\tilde{y}_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right] F(x) = F_{m-1}(x),\ i = 1, \dots, N$
（4）  $\alpha_m = argmin_{\alpha,\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(x_i, \alpha)]^2$
（5）  $\rho_m = argmin_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i)) + \rho h(x_i, \alpha_m)$
（6）  $F_m(x) = F_{m-1}(x) + \rho_m h(x, \alpha_m)$
（7）  $end\ for$
```

For example, if we predict whether a person will like a computer game advertisement to buy the CART of the game or not, enter the various attributes of each user, such as age, occupation, sex, because different user needs will produce different results for the same advertisement, and we can understand the leaf score as this How many people are likely to like computer games.

The XGBoost model can be expressed as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \qquad (7)$$

Among them, i=1, 2,..., n, n are the number of samples; F is the set of all regression trees[7], and f_k is a function in F. In the two-classification problem, $\hat{y}_i$>0.5is generally divided into one class, and $\hat{y}_i$<0.5 is divided into another category. In this experiment, we divide the experimental results into two categories, namely, successful transformation and unsuccessful transformation.

The essence of Boosted Tree is understood, which helps to apply this method directly to solve practical classification problems. The essence of using XGBoost method in this paper is to improve the prediction accuracy of Boosted Tree by parallel Boosted Tree on a single multi-CPU computer. We call the train_test_split function and randomly divide our experimental data into a training set and a test set, all of which are 242 dimensional and are used to train our models.

Run the python program and record the experimental results as follows:

TABLE VI. EXPERIMENTAL RESULT

Feature Evaluation Index	*Without initial forecast*	*Add initial forecast*
logloss	0.09238	0.08245

From the above table, it can be concluded that the loss function obtained is higher when the data of the primary prediction item is not added, and the result of the loss function is significantly decreased after the initial prediction is added.

E. Application scenario requirements

This paper discusses the prediction method of advertising conversion rate based on store characteristics and user needs, and makes a reasonable prediction for search advertising conversion rate. This method can be used to predict the conversion rate of commercial network platforms (Jingdong, Tmall, Taobao, etc.) due to the universality of the two features of store features and user needs.

IV. Conclusion and Prospect

In this paper, the prediction of advertising conversion rate is studied by experiments. For training data, the selection of features has a crucial impact on the prediction performance of the model. In the prediction of advertising click rate, there are many features that can be used, including basic data, advertising commodity information, user information, context information and store information. To achieve a good accuracy rate, the model should be fully excavated as a feature, and the better the combination of these features, the better the performance of the model. The feature learning method proposed in this paper estimates the ad click rate, only considering that the full advertising data is not considered to show inadequate advertising. In the next work, how to estimate the click rate of sparse advertising from the point of view of characteristic learning is a problem worthy of study, and it is also an urgent problem to be solved at present. At the same time, we should also pay attention to the research of different models integration.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No.61772277 、National key research and development plan "intelligent evaluation and prevention of occupational crime" under Grant No.2017YFC080-4000 and A robust sparse representation model based on L1 norm and its application in feature extraction in Jiangsu provincial basic research program "under Grant No. BK20171494.

REFERENCES

[1] Zeng D S, Huang F L, Pan C D. Feature Engineering for Product Review Spam Identification[J]. Journal of Fujian Normal University, 2017.

[2] Cormack G V. Feature engineering for mobile (SMS) spam filtering[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2007:871-872

[3] Li J, Liang X, Ding W, et al. Feature engineering and tree modeling for author-paper identification challenge[C]// Kdd Cup 2013 Workshop. 2013:1-8.

[4] Stankevich M, Isakov V, Devyatkin D, et al. Feature Engineering for Depression Detection in Social Media[C]// International Conference on Pattern Recognition Applications and Methods. 2018:426-431.

[5] Chen J H, Li X Y, Zhao Z Q, et al. A CTR prediction method based on feature engineering and online learning[C]//International Symposium on Communications and Information Technologies. IEEE, 2018.

[6] Chen T, Tong H, Benesty M, et al. xgboost: Extreme Gradient Boosting[J]. 2015.

[7] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016:785-794.