Integration learning oriented click anti-fraud prediction

LIANG Hua-xiong
School of Information Management, Beijing University of Information Technology
Beijing，China
huaxiongliang@gmail.com

ZHAO Gang
School of Information Management, Beijing University of Information Technology
Beijing，China
zhaogang@bistu.edu.cn

Abstract:Predicting whether an ad click is a normal click or a cheating click is of great importance to digital marketing. In order to solve the problem of information loss in the numerical process of small-scale data in feature engineering, this paper proposes a click anti-fraud prediction method based on SVD (Singular Value Decomposition) and integrated learning in the process of click anti-fraud data exploration. The method uses SVD to mine information on category features in the dataset while increasing the category data dimensionality; Xgboost is used as a feature converter for integrated learning, and the information of each leaf node in the tree is fed into a logistic regression model for click anti-fraud prediction as a feature vector. The experimental results show that both SVD and integrated learning methods can improve the accuracy of click-to-fraud prediction with an accuracy of 89.19%, and to a certain extent avoid the loss of data information in feature engineering.

Keywords:*singular value decomposition; data mining; click-to-counter fraud; integrated learning；feature transform*

I. INTRODUCTION

The act of clicking, either through manual clicks or manipulated device clicks, to maliciously increase the click-through rate of an advertisement and thereby disrupt the normal flow of marketing data is known as click fraud [1]. With the proliferation of mobile devices, more and more advertising marketers are using digital communication mediums to promote their products and communicate with consumers in a personalised, customised and cost-effective way. However, behaviours such as malicious swipes and deliberate clicking have seriously affected the promotion of advertisements on e-commerce platforms and the feedback profiling of audience demographic clicks. Illegal advertisers perform malicious human clicks on published ads to increase click-through rates and thus gain more media attention to the traffic, leading to media coverage and dissemination of the traffic's inflated click-through numbers.

In the current study, Gong Shangfu[2] collected and stored user behavior data as a data warehouse, analyzed users' legitimate behaviors and established rank evaluation criteria, and also used Bayesian methods to classify user click behavior prediction based on this criteria. In addition to the detection method for establishing evaluation criteria, Dong Yannan [3] first classified the click user features to obtain click fraud groups and users, and then used Bayesian algorithm for classification. Liu Guoqing [4] proposed a way to identify malicious clicks based on a combination of dwell time and number of clicks. Ren Yajin [5] proposed a click fraud prediction model based on user behaviour patterns in the mechanism of anomaly testing algorithm; however, the algorithm has the disadvantage of low efficiency in processing high-dimensional data. Lu-Ming Mao [6] improved click fraud accuracy by constructing a database of click fraud behavioural pattern features and using the comparison of behavioural pattern databases to identify click fraud. Zhang Xin [7] used an integrated Boosting-SVM model to solve the overfitting phenomenon generated by SVM in the click fraud dataset. Berrar used an improved random forest model and experimental results illustrated its detection accuracy over models such as SVM and logistic regression [8][9]. Perera [10] used machine learning models to construct six integrated learning schemes and demonstrated that integrated learning models are more effective than single models in click fraud. A number of researchers have used gradient boosting models such as Xgboost or LightGBM [11][12][13][14]to achieve better results than general machine learning models; Thejas [15]and others have combined random forest and Xgboost models to achieve better results than single gradient boosting models.

In click fraud prediction research, most researchers have devoted themselves to exploring more advanced algorithmic models and improving anti-fraud mechanisms, with less research on information mining and integrated learning [16]. And considering that integrated learning methods are characterised by strong generalisation and significant classification effects. In this paper, we will improve and predict from two perspectives: deeper mining of data information and integrated learning. The contributions of this paper are as follows:

1)To address the problems of insufficient data dimensionality and data samples with loss of implicit data information in feature engineering, SVD is used to perform singular value decomposition on the encoded data, and the decomposed matrix is stitched into the original data to achieve the effect of enhanced data.

2)The integrated learning method is used to solve the problem of low classification accuracy of a single model.

Xgboost was firstly chosen as the feature converter, and then the logistic regression model was used for rule and weight learning and prediction.

II. CATEGORY FEATURE INFORMATION MINING

A. Introduction to the data set

The data in this article is based on a dataset from the Baidu Platform Click Anti-Prediction Fraud Competition at https://aistudio.baidu.com/aistudio/competition/detail/52/0/data sets. The dataset contains fields summarized in the following table.

TABLE 1 Description of data

feature	*type*	*Explanation*
sid	string	Session id
label	int	0 normal, 1 cheating
package	string	Media information
version	string	app version
android_id	string	Media information
media_id	string	Media information
apptype	int	Type of media
time	bigint	Timestamp
location	int	Location Coding
fea_hash	int	User Profile Code
ntt	int	Network type
carrier	string	Type of operator
os	string	Operating systems
osv	string	Opera-system version
lan	string	Language Type
dev_height	int	Equipment high
dev_width	int	Equipment width
dev_ppi	int	Screen resolution

As seen in Table 1, the dataset contains two types of feature data, numeric features and category features. By pre-processing the data, missing data for each feature is calculated and missing values are processed. Missing values for numerical features were filled using the mean value and missing values for categorical features were filled using missing; the categorical features were numerized using the DictVectorizer method.

B. Feature Importance Analysis

In order to dig deeper into the effective feature information, after a series of feature engineering such as feature construction, a tree model was used to do importance analysis on the data features. The results are shown in Figure 1.

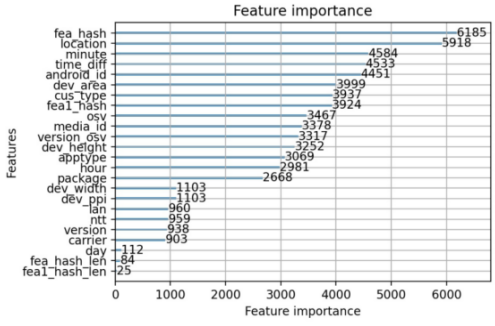

Figure 1 Feature contribution map

The histogram of feature contribution shows that the feature contribution of category features is generally higher than that of numerical features. In addition, category features lose a lot of important implicit information in the numerical process, ignoring the deep semantic information that the data itself has. Therefore, this paper uses SVD to data mine the category features with high feature contribution

C. SVD decomposition

SVD [17] is widely used in many fields because of its unique matrix decomposition. For example, in the field of machine learning, SVD is often used to downscale high-latitude data for visualization or easy computation; in the field of images, SVD is often used to compress images for storage. In this paper, SVD is used for information mining of category features while increasing the dimensionality of the features to improve the accuracy of click anti-fraud prediction. Steps of SVD decomposition of category features data are as follows:

1)Encode token2id for categories with high feature importance, such as user features, user geographic location, request arrival time, media category information, etc., so that each category feature has an id value corresponding to its value.

2) Construct the co-occurrence matrix A of the above category features, the data information in the co-occurrence matrix is the implicit information between the category features.

3) Use SVD to decompose the co-occurrence matrix A into singular values, and the decomposition is shown in equation (1). The idea of SVD is to decompose a large matrix into the product of three matrices.

4) In the matrix obtained by decomposition, the vertical dimension of the left odd matrix U is the same as the vertical dimension of the original matrix A, and can be directly spliced into the original matrix. The right odd vector matrix V needs to be transposed because of the dimensionality before it can be spliced into the original matrix according to token2id.

$$A_{m\times n} \approx U_{m\times r} \times \sum_{r\times r} \times V_{r\times n} \quad (1)$$

5) The left odd matrix and the transposed right odd matrix are spliced into the original data to realize the increase of feature dimensions while mining the hidden information of category features, so as to improve the accuracy of click anti-fraud prediction.

III. CLICK ON THE ANTI-FRAUD INTEGRATION MODEL

In this section, an integrated learning model structure based on Xgboost and Logistic is proposed for the click anti-fraud prediction problem. The model can extract deep information in features, complete feature screening and nonlinear transformation, thus improving the generalization ability and prediction accuracy of the classifier.

A. Integrated learning model

GBDT integrated with Logistic has been widely used in many fields, while in the Xgboost model, the regular term is explicitly added to control the complexity of the model, prevent overfitting and enhance the generalisation ability of the model. Compared to GBDT which uses the first order derivative

information of the loss function, Xgboost uses the first order derivative information of the loss function along with Taylor expansion to enhance the handling of the data. Therefore, it is more meaningful to use Xgboost instead of GBDT as the feature converter in integrated learning.

Xgboost is essentially a decision tree fitted iteratively to the data, by training different decision trees and combining them linearly into a strong classifier. This is shown in equation (2).

$$f_M(x)=\sum_{m=1}^{M} \mathrm{W}_m \cdot T(x;\Theta_m)=f_{M\_1}(x)+T(x;\Theta_m) \quad (2)$$

Where $T(x;\Theta\_m)$ is the decision tree, $\Theta_{\_m}$ denotes the decision tree parameters, W denotes the weight of the tree, M is the number of decision trees, $f_{(M\_1)}(x)$ is the current decision tree classifier model, and the parameters of the next decision tree, $T(x;\Theta\_m)$, are determined by minimizing the loss function.The Xgboost model fits the training data well, but does not produce a feature vector by itself. During the Xgboost fitting of the data, the values of the sample features falling into the leaf nodes are calculated, as shown in equation (3).

$$T=\sum_{m=1}^{M} leaf\left(\mathrm{T}(x;\Theta_m)\right) \quad (3)$$

The *leaf()* function in the above equation serves to encode the values of the sample features at the leaf nodes in the tree. In this paper, the path taken from the root node to the leaf node of each tree of Xgboost is used as a transformation rule for a category of features, and the leaf node of each tree is encoded with one_hot, and all the encoding vectors in the tree are stitched together to obtain a particular transformation encoding, and then the feature vectors are fed into a logistic model for prediction. The model structure is shown in Figure 2.

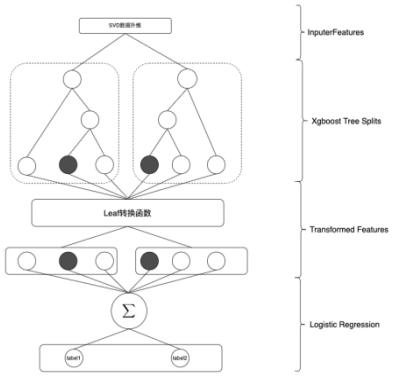

Figure 2 Structure of the integrated learning model

In the integrated learning model structure, the data after SVD dimension upgrading is first sent to Xgboost model for learning. In Xgboost's tree structure, each category occupies a decision tree, and 0/1 of the path from the root node to the leaf node represents each feature vector. Finally, these feature vectors are sent to the Logistic model for click anti-fraud prediction.

B. 3 Click Anti-Fraud Forecasting Process

There are four steps in the click-to-fraud prediction process based on SVD and integrated learning, and the process is shown in Figure 3.

1) In the process of data exploration, fill in missing values, filter numerical features and category features according to feature categories, and analyze the importance of category data features using tree model.

2) In feature engineering, a co-occurrence matrix between class features with high feature contribution is constructed. SVD is used to decompose the co-occurrence matrix, and the decomposed left odd matrix and transposed right odd matrix are connected to the co-occurrence matrix to make up for the hidden information lost by class features in the token2id process.

3) Send the dimensioned data to Xgboost for feature transformation and get the transformed one_hot encoding.

4) The one_hot encoding is used as input to the logistic model for click anti-fraud prediction.

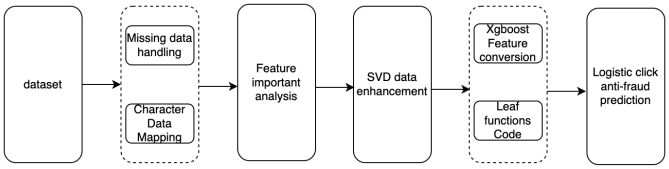

Figure 3 Flowchart of clickback prediction

IV. EXPERIMENT

A. Experimental evaluation indicators

The total number of experimental data samples is 500,000, including 242,240 positive samples and 257,760 negative samples. The positive and negative samples are relatively balanced, and there is no need to do the positive and negative sample balancing process. The binary classification prediction confusion matrix is shown in Table 2.

TABLE 2 Dichotomous confusion matrix

	Positive	***Negative***
True	TP	FN
False	FP	TN

The accuracy (Acc), ROC curve and AUC values were used for the experimental evaluation. the Acc value was calculated as in equation (4); the ROC curve was calculated as equation (5) for the horizontal coordinate of FPR and as equation (6) for the vertical coordinate of TPR.

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \quad (4)$$

$$FPR=\frac{FP}{FP+TN} \quad (5)$$

$$TPR=\frac{TP}{TP+FN} \quad (6)$$

In the ROC curve, the larger the value of the vertical coordinate TPR the better, and the lower the horizontal coordinate FPR the better. the AUC is defined as the area enclosed by the coordinates under the ROC curve and takes values between 0.5 and 1. The closer the value of the AUC is to 1, the better the method of model detection.

B. Experimental results and analysis

The experiment is divided into three parts. The first part consists of four models and eight comparison experiments to verify the impact of SVD dimensionality increase on model accuracy; the second part verifies the effectiveness of the integrated learning approach by comparing a single model with the integrated learning approach; the third part verifies the effectiveness of the click-to-fraud model based on SVD and integrated learning.

1) Effect of SVD data up-dimensioning on experimental results

The SVD was used to increase the dimensionality on the original data. The experimental models GBDT, Xgboost, LightGBM and Logistic were chosen to validate the SVD dimensionalised data, with Acc and AUC used as evaluation metrics. the experimental results are shown in Table 3 below.

TABLE 3 Comparison of forecast results

model	*acc*
Logistic	55.16%
SVD+Logistic	55.17%
Xgboost	89.06%
SVD+Xgboost	**89.09%**
GBDT	88.05%
SVD+GBDT	89.07%
LightGBM	88.97%
SVD+LightGBM	89.06%

Based on Table 3, it can be concluded that the accuracy of each model improved after the data dimension was increased using the SVD approach. The accuracy of the Logistic Regression model increased by 0.01%, the accuracy of the Xgboost model increased by 0.03%, the accuracy of the GBDT increased by 0.02% and the accuracy of the LightGBM model increased by 0.09%. The reason is that after using SVD to increase the dimension of data, the loss of information in the numerical process of category features is compensated to a certain extent, so that the accuracy rate is improved.

2) Results and analysis of integrated learning experiments

In order to verify the effectiveness of the integration of Xgboost and Logistic algorithms, GBDT+Logistic and Xgboost+Logistic integration methods were selected for comparison. The experimental results are shown in Table 4.

TABLE 4 COMPARISON OF EXPERIMENTAL RESULTS

model	*Acc*	*AUC*
Logistic	55.17%	0.6099
GBDT	89.05%	0.9436
GBDT+Logistic	89.13%	0.9501
Xgboost	89.09%	0.9442
Xgboost+Logistic	**89.16%**	**0.9506**

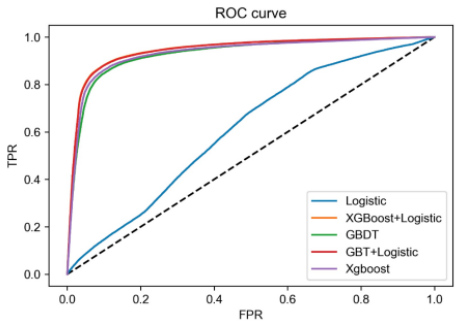

Figure 4 Figure 4 Comparison of ROC curves

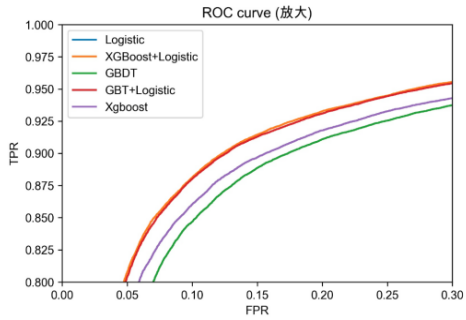

Figure 5 ROC local zoom comparison

From Table 4 and Figure 5, the method using the Xgboost model as a feature converter and the Logistic model integrated with the Logistic model performed better in terms of accuracy than the method using the Xgboost model alone; the method using the GBDT model as a feature converter and the Logistic model integrated with the Logistic model performed better in terms of accuracy than the method using the GBDT model alone. The accuracy of the method using the Xgboost model combined with the Logistic model was 89.16%, with an AUC value of 0.9506. The accuracy of the method using the GBDT model combined with the Logistic model was 89.13%, with an AUC value of 0.9501. It further illustrates the effectiveness of Xgboost model and Logistic integration method.

3) Experimental results and analysis based on SVD and integrated learning

In the click anti-fraud prediction model, two innovations and applications are proposed in this paper, one is to use SVD to mine the implicit information of category features and elevate the feature dimension of category feature data; the other is to use Xgboost integrated with logistic, both methods achieve the improvement of click anti-fraud prediction accuracy and prove the effectiveness of these two methods. By combining the two methods to build a click-to-fraud model based on SVD and integrated learning, the experimental results are shown in Table 5.

table 5 Comparison of experimental results

model	*Acc*	*AUC*
Xgboost+Logistic	89.16%	0.9506
SVD+Xgboost+Logistic	**89.19%**	**0.9535**

The accuracy of the integrated method based on SVD and Xgboost with logistic phase reached 89.19%, which is 0.03% higher than the integrated method without SVD. It proves the

advantages of SVD and integrated learning in click anti-fraud models, which has some has some research significance.

V. 6. CONCLUDING REMARKS

In the feature engineering of click anti-fraud prediction, this paper uses the SVD method to expand the dimensionality of the data in order to compensate for the loss of deep information of the data in the numerical process of feature engineering, and also uses Xgboost and Logistic regression to integrate with each other to improve the accuracy rate. The experiment shows that the accuracy of the prediction is improved to a certain extent using the above method, which has certain reference value.

Considering the relatively small size of the data in this experiment, the neural network model was not selected for the experimental model selection. In addition, subsequent work will go deeper into the feature construction of the data, and also consider the application of the idea of recommendation algorithms in cyber security areas such as click anti-fraud prediction.

REFERENCES

[1] Lin Hongwei. Research on some key issues of online advertising operation [D]. Sichuan: University of Electronic Science and Technology, 2013. doi:10.7666/d.D763593.

[2] Gong, Shangfu, Jiang, Xiaoxu. Ad fraud click detection based on user behavior analysis [J]. Computer Applications and Software, 2011, 28(4):3.

[3] Dong YAN, Liu XJ, Li B, et al. Click fraud group detection and discovery[J]. Computer Application Research, 2016, 33(6): 1771-1774.

[4] Liu Guoqing. A new click fraud prevention algorithm [J]. Computer Engineering,2011,37(S1):160-161+167.

[5] Ren Yajin. Research on fraudulent click prevention techniques in online advertising [D]. Lanzhou Jiaotong University, 2014.

[6] Lu-Ming Mao. Research on ad click fraud detection technology in mobile World Wide Web [D]. Southwest Jiaotong University,2016.

[7] Zhang X, Liu XJ, Li B, Guo H. An SVM integration method for online advertising click fraud detection[J]. Small Microcomputer Systems,2018,39(05):951-956.

[8] Berrar D. Random forests for the detection of click fraud in online mobile advertising[C]//Proceedings of the 1st International Workshop on Fraud Detection in Mobile Advertising. 2012: 1-10.

[9] SHAOHUI D, QIU G W, MAI H, et al. Customer transaction fraud detection using random forest[C]//2021 IEEE International Con-ference on Consumer Electronics and Computer Engineering (IC-CECE). 2021: 144-147.

[10] Perera K S, Neupane B, Faisal M A, et al. A novel ensemble learning-based approach for click fraud detection in mobile advertising[M]//Mining Intelligence and Knowledge Exploration. Springer, Cham, 2013: 370-382.

[11] Gohil N P, Meniya A D. Click ad fraud detection using XGBoost gradient boosting algorithm[C]//International Conference on Computing Science, Communication and Security. Springer, Cham, 2021: 67-81.

[12] VIRUTHIKA B, DAS S S, KUMAR E M, et al. Detection of Ad-vertisement Click Fraud Using Machine Learning[J]. International Journal of Advanced Science and Technology, 2020.

[13] MINASTIREANU E A, MESNITA G. Light gbm machine learning algorithm to online click fraud detection[J]. J. Inform. Assur. Cy-bersecur, 2019.

[14] ZHANG Y, TONG J, WANG Z, et al. Customer transaction fraud detection using Xgboost model[C]. 2020 International Conference on Computer Engineering and Application (ICCEA). 2020: 554-558.

[15] THEJAS G S, DHEESHJITH S, IYENGAR S S, et al. A hybrid and effective learning approach for Click Fraud detection[J]. Machine Learning with Applications, 2021, 3: 100016.

[16] Li, X.I.. Click fraud prediction based on data mining [D]. Dalian University of Technology, 2021. doi:10.26991/d.cnki.gdllu.2021.001330.

[17] Anwar T, Uma V, Srivastava G. Rec-cfsvd++: Implementing recommendation system using collaborative filtering and singular value decomposition (svd)++[J]. International Journal of Information Technology & Decision Making, 2021, 20(04): 1075-1093.