IETI Transactions on Engineering Research and Practice
http://ieti.net/TERP/

2020, Volume 4, Issue 1, 1-9, DOI 10.6723/TERP.202006_4(1).0001

VISUALIZATION ANALYSIS OF REAL-TIME BIDDING DATA OF ONLINE ADVERTISING BASED ON HADOOP AND PYTHON

Xie Linjun, Xie Xiaolan
College of Information Science and Engineering, Guilin University of Technology, Jiangan Road 12, Guilin, China

Abstract with the development of the information age and the change of the market environment, the demand for accurate advertising delivery technology is increasing day by day. At the same time, the diversification of new media results in fragmentation of information transmission. Therefore, how to conduct more efficient advertising marketing is a problem faced by advertisements in the current Internet era. This paper explains the relationship between real-time bidding mode and big data as well as the specific operation mode. Advertising real-time bidding is a new algorithm technology, based on which the optimal placement strategy can be generated through analyzing and calculating the constantly updated and iterative large data set and combining many objective factors and other constraints. Combined with the several cases, this paper makes a reasonable analysis of various factors affecting click-through rate in accurate advertisements, and puts forward appropriate suggestions for improvement and effective advertising strategies.

Keywords: big data; real-time advertising bidding; data cleaning; visualization analysis.

1. INTRODUCTION

The digital advertising world is developing rapidly, and a lot of resources are being invested in relevant fields of research. CTR prediction is an important step in locating user tag technology. The main content of this research project is to carry out big data analysis on the data set related to advertising clicks, and build the prediction model for Avazu data released by kaggle and RTB advertising click-forecast data provided by CluBear.

Online advertising, also known as online marketing or Internet advertising, is a new way of advertising. In has been mentioned in the journal of The Digital News Project that the traditional advertising display on the Internet is now suffers from declined click rates and the revenue of advertisers is static or even decreased, so many users completely choose to turn off the advertising.

2. ANALYSIS OF ADVERTISING BIDDING DATA SETS

2.1. Research Object

The data researched in this paper came from the kaggle and CluBear. The data set provided by kaggle is the CTR prediction data. In Internet advertising, click-through rate prediction is an indispensable one of the important indicators and factors in advertising strategy, which determines actual sponsored search or real-time bidding. The data set provided by the CluBear is used for RTB advertising click prediction. Real-time advertising bidding is an extremely important advertising delivery mode. The main content of this chapter will be the data preprocessing and exploratory data analysis of the above two data sets, so as to prepare for the modeling and predictive analysis in chapter 4.

IETI Transactions on Engineering Research and Practice
http://ieti.net/TERP/

2020, Volume 4, Issue 1, 1-9, DOI 10.6723/TERP.202006_4(1).0001

2.2. Data Exploration and Analysis

Correlation analysis is an analysis method to measure the distribution trend or variation trend of two sets of data or two sets of samples. Correlation coefficient is the most commonly used in correlation analysis, among which Pearson correlation coefficient used in research analysis is a measure to reflect the degree of linear correlation between two variables, and its coefficient directly measures the magnitude of correlation. The formula used is as follows:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

The following correlation coefficients can be obtained from the dataset studied in this paper, as shown in figure 1. According to the figure, the dc attribute is positively correlated with ["instl","mfr"] attribute and negatively correlated with ["atype","nt","period"]attribute. In the heat map, the correlation index of ["bidf","isp"]attribute approaches 0, which means that there is no correlation between it and the annotation attribute dc, therefore, the ["bidf","isp"]column can be removed in the correlation analysis.

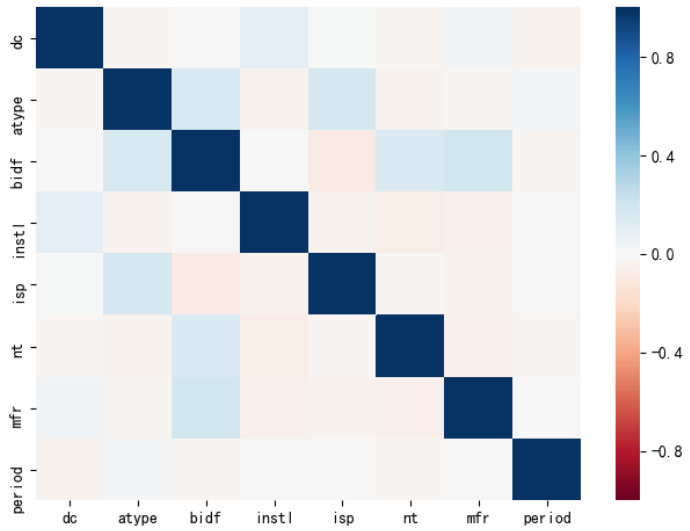

Figure 1. Thermal diagram of correlation coefficient of each attribute

The correlation between the attribute data contained in the analysis data column ["mfr"]is shown in figure 2 (significance = 0.5). Although combined with the actual situation analysis, the correlation coefficient presented between handset manufacturer isp under CTR dc constraint is not necessarily meaningful, but it can pave the way for the following single-factor and multi-factor analyses.

IETI Transactions on Engineering Research and Practice
http://ieti.net/TERP/

2020, Volume 4, Issue 1, 1-9, DOI 10.6723/TERP.202006_4(1).0001

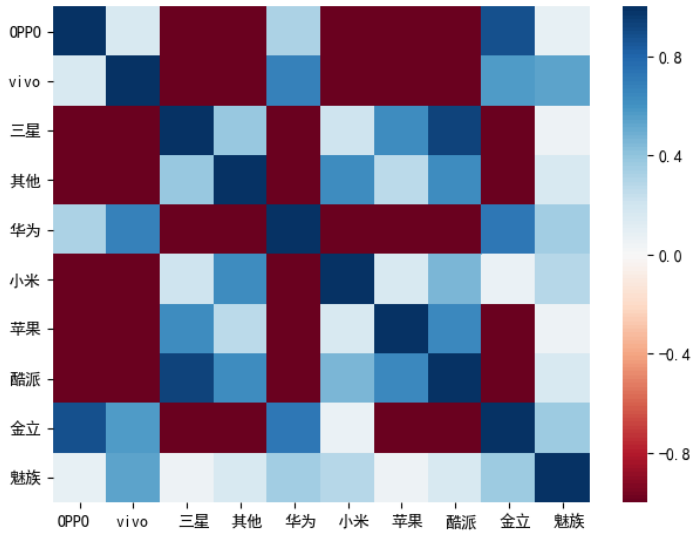

Figure 2. Correlation coefficient of mfr data column

Through observing above two thermal diagrams ,the Pearson correlation coefficient used in the study provides a good guidance for the data preprocessing part and feature extraction part in the feature engineering below, and also shows the relationship between feature attributes clearly in the graph, providing referene for the research on screening valuable attribute relationships.

3. MODELING ANALYSIS AND PREDICTIVE EVALUATION OF ADVERTISING CLICK-THROUGH RATE

3.1. Decision-Making Tree

If there are no other child nodes under any child node in the decision tree, then this node is the annotation, otherwise it is the feature. The concept of entropy was mentioned in the previous paper. In the information gain, the greater the value of entropy, the greater the uncertainty of the system.

$$\mathrm{H(X)} = -\sum p_i \log(p_i) \tag{2}$$

$$\mathrm{I(X,Y)} = \mathrm{H(Y)} - \mathrm{H(Y|X)} = \mathrm{H(X)} - \mathrm{H(X|Y)} \tag{3}$$

The formula above shows that the difference between the entropy of the Y event and the entropy of the other X event represents the magnitude of the effect of the X event on the Y event. When the resulting entropy gain is large, there is a greater effect of the X event on the Y event, so the the attribute segmentation of the event is needed, alternatively the information gain rate and Gini coefficient are introduced according to the requirements. When using a decision tree, you may encounter three problems that need to be solved, including continuous value segmentation, rule exhaustion, and overfitting. These issues will be addressed and code implementation is shown as below:

```
from sklearn.tree import DecisionTreeClassifier,export_graphviz
from sklearn.externals.six import StringIO
```

2020, Volume 4, Issue 1, 1-9, DOI 10.6723/TERP.202006_4(1).0001

```
models = []
******************

……

******************
models.append(("DecisionTreeGini",DecisionTreeClassifier()))
for classifierName,clf in models:
    clf.fit(dataSet_train,dataLabel_train)
    setAndLabelList=[(dataSet_train,dataLabel_train),
            (dataSet_valid,dataLabel_valid),(dataSet_test,dataLabel_test)]
    for i in range(len(setAndLabelList)):
        setPart=setAndLabelList[i][0]
        labelPart=setAndLabelList[i][1]
        label_pred=clf.predict(setPart)
        print(i)
        print(classifierName, "-ACC",accuracy_score(labelPart,label_pred))
        print(classifierName, "-REC", recall_score(labelPart, label_pred))
        print(classifierName, "-F-Value", f1_score(labelPart, label_pred))
        dot_data=StringIO()
        export_graphviz(clf,out_file=dot_data,
                        feature_names=feat_names,
                        class_names=["NotClick","Click"],
                        filled=True,
                        rounded=True,
                        special_characters=True)
        graph=pydotplus.graph_from_dot_data(dot_data.getvalue())
        graph.write_pdf("tree3.pdf")
```

The Gini decision tree of DecisionTreeClassifier with default parameters is obtained through code implementation, as shown in figure 3. The significance of each end node in the decision tree is analyzed.The first line in a node is the child node jump that determines whether the condition is True or False in a particular case of feature compounding. The value included in other contents of the node refers to the proportion of each value marked, samples represents the total number of samples, and gini coefficient represents the current impurity. In order to facilitate the business analysis, post-pruning is required. When some nodes show extreme proportion (for example, greater than 30:1) while node segmentation is still going on, the maximum depth max_depth, minimum sample segmentation min_samples_split, minimum number of leaves min_samples_leaf should be set in the parameters. Parameter debugging is a complicated process, which requires constant debugging and comparison so as to obtain the accuracy rate, recall rate and F value under different calculation standards (Table 1) as well as the appropriate decision tree (Figure 4).

Table 1. Two values returned by bayesian naive algorithm

Calculation standard	ACC accuracy	REC recall rate	F-Value
Gini Impurity coefficient	0.8118	0.0659	0.1153
Entropy production	0.8118	0.0769	0.1320

IETI Transactions on Engineering Research and Practice
http://ieti.net/TERP/

2020, Volume 4, Issue 1, 1-9, DOI 10.6723/TERP.202006_4(1).0001

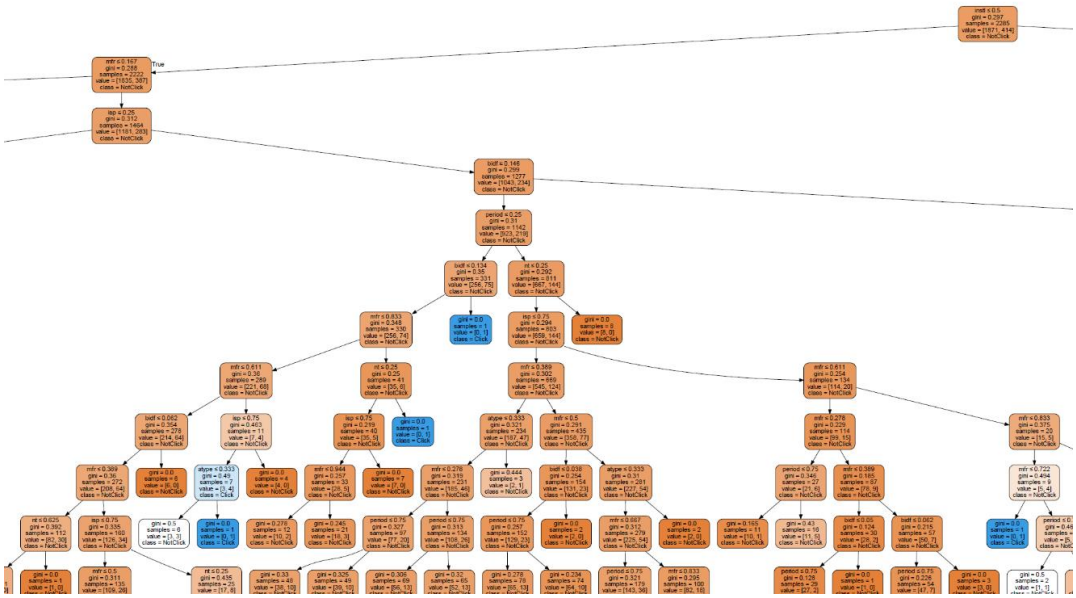

Figure 3. Part legend of the decision tree generated by debugging without parameters

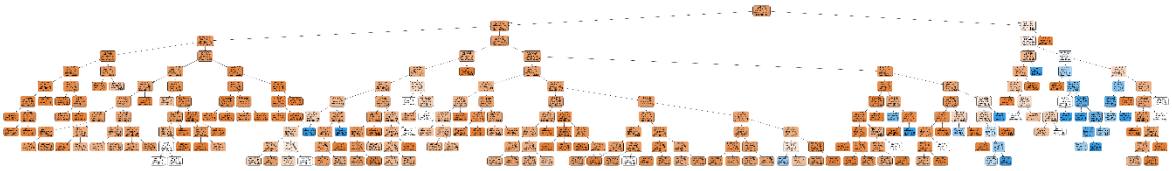

Figure 4. Complete legend of the decision tree generated by debugging with optimized parameters

In the case of the decision tree with entropy gain as the calculation standard, F value reached the highest 0.132, so it seems that the data set studied in this paper is suitable for building the model with this algorithm.

3.2. Model Evaluation

The model of dichotomy refers to the annotation classification discrete as only two values。 For example, in this study, CTR 1 and CTR 0 are a binary attribute, the click value 1 is a positive class, and the non-click value 0 is a negative class. In the ROC curve, the horizontal axis is the FPR misjudgment rate and the vertical axis is the recall rate. It is easy to see the influence of any threshold value as the threshold value on the model's estimated performance recognition ability. In the evaluation, the ideal state is to increase the recall rate of TPR without expanding the misjudgment rate of FPR， so the curve should be inclined to the upper left. AOC refers to the area covered under the ROC curve, which can more obviously show the deviation degree of the curve from the ideal curve. The code implementation is as follows:

```
models.append(("SVM Classifier",SVC(C=C)))
for classifierName,clf in models:
    clf.fit(dataSet_train,dataLabel_train)
    setAndLabelList=[(dataSet_train,dataLabel_train),
                     (dataSet_valid,dataLabel_valid),(dataSet_test,dataLabel_test)]
```

2020, Volume 4, Issue 1, 1-9, DOI 10.6723/TERP.202006_4(1).0001

```
for i in range(len(setAndLabelList)):
    setPart=setAndLabelList[i][0]
    labelPart=setAndLabelList[i][1]
    label_pred=clf.predict(setPart)
    foo.add_subplot(1,3,i+1)
    fpr,tpr,threshold=roc_curve(labelPart,label_pred,pos_label=1)
    plt.plot(fpr,tpr)
    print("AUC",auc(fpr,tpr))
    print("AUC_Score",roc_auc_score(labelPart,label_pred))
plt.show()
```

The output results in the console are shown in figure 5 and figure 6.

```
 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0]
AUC 0.5199719101123595
AUC_Score 0.5199719101123595
[1000]
```

Figure 5. Output AUC results in console

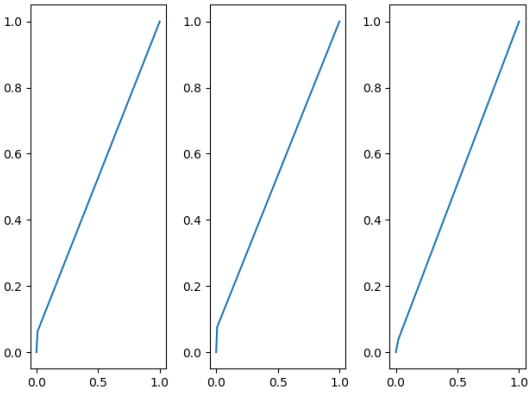

Figure 6. ROC curves of training, verification and test data sets

According to the analysis, the AUC value is only 0.56. As shown in the ROC curve in the figure above, the three data sets correspond to the training set, the verification set and the test set, all of which are diagonal, meaning that in the model estimation, the results of all positive and negative classes are close to 1:1.

4. SET UP OF A FULLY DISTRIBUTED HADOOP CLUSTER

4.1. Objectives of System Design

Hadoop is a very popular distributed cluster platform for processing big data in the field of data analysis. In order to provide a larger data set for data modeling and analysis, Hadoop's open source components need to be used to build a more efficient storage system on multiple computers or cloud platforms in this study.

2020, Volume 4, Issue 1, 1-9, DOI 10.6723/TERP.202006_4(1).0001

4.2. System Technology and Architecture

Hadoop mainly consists of three core components, including HDFS distributed file system, MapReduce distributed computing framework, and resource scheduling management system YARN. HDFS has three features, including mass storage amounts of data, excellent compatibility and fault tolerance, and excellent extensibility. This can solve the problem that only limited data can be stored in a single machine environment. In HDFS, data can be segmented and distributed on each machine in the cluster for storage and backup with blocks, and at the same time the loss of data can be avoided when uploading data. The basic calculation flow of MapReduce is shown in figure 7. The third component, YARN, is specially used to manage the resource scheduling of the whole cluster system, and meanwhile unifies the scheduling of various framework resources like Hive and Spark. The general yarn-based system structure diagram is shown in figure 8. At the bottom is the distributed cluster system of Hadoop, on which is the data operation and cluster management resource system. At the top are the Sciprt, SQL, JavaScala and other framework components that can be directly used based on YARN. Both batch processing and real-time statistics are available, which greatly improves the convenience of resource scheduling.

Figure 7. Basic calculation flow chart of MapReduce

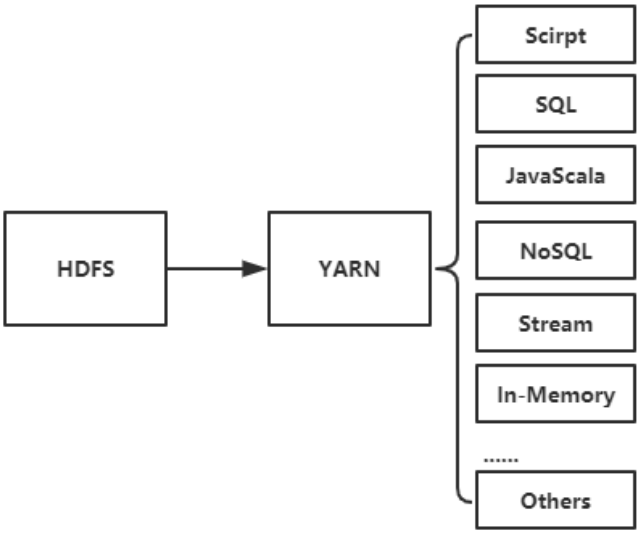

Figure 8. Basic system structure of YARN

2020, Volume 4, Issue 1, 1-9, DOI 10.6723/TERP.202006_4(1).0001

Based on these components, the next section will discuss the planning and functionality of the system modules.

4.3. System Module

The planning of the cluster system is shown in table 2, and the usage of distributed cluster nodes is shown in table 3. The node IP represents an example that is only done when the local virtual machine cluster is set up, which needs to be changed in the real machine environment.

Table 2. Hadoop cluster planning I

Component	Usage
HDFS	NameNode、DataNode
YARN	ResourceManage、NodeManage

Table 3. Hadoop cluster planning Ⅱ

Name of node	Node IP	Usage
hadoop000	192.168.1.234	NameNode、DataNode、ResourceManage、NodeManage
hadoop001	192.168.1.235	DataNode、NodeManage
hadoop002	192.168.1.236	DataNode、NodeManage

There are three modules in the cluster, including data cleaning module, feature engineering module and data modeling module. The data set (6 gb) provided by Kaggle is ideal for analysis on a clustered platform. After analyzing the CTR record data set, it is found that there are 24 columns of fields, as shown in table 4.

Table 4. Kaggle dataset field explanations

Field name	**Field meaning**
id	Advertising identifier ID
click	Click or not
hour	Recording duration
banner_pos	Banner price
site_id	Site ID
site_domain	Site domain
site_category	Site type
app_id	Application ID
app_domain	Application domain
app_category	Application category

2020, Volume 4, Issue 1, 1-9, DOI 10.6723/TERP.202006_4(1).0001

device_id	Device ID
device_ip	Device IP
device_model	Device model
device_type	Device type
device_conn_type	Device network type
C1、C14-C21	Anonymous data

Considering the privacy and security of user data, the field at C1 and fields C14 and C21 have been anonymized and can be removed. In addition, in ["hour"], abnormal values that do not conform to the common knowledge range are identified and excluded, and four-point abnormal values are identified for other discrete data columns. Because of the huge amount of data, the threshold K is set at 1.5.

5. CONCLUSION

This paper systematically combs the real-time bidding system of advertisements under the big data. In the context of the current trade war between China and the United States, in order to maximize the revenue of the advertising industry and optimize advertising strategies, improving financial models have become an indispensable part. According to the background information in the first chapter, the major revenue of domestic e-commerce platforms is still from the advertising industry, which indicates that Internet advertising is increasingly important and in a sense can promote China's financial development before facing such a powerful economic power as the United States.

However, there are only 4969 pieces of data in the data set used in this study, which means that there may be overfitting phenomenon in the data, the possibility of overfitting in the training set becomes larger, and there may also be overfitting phenomenon in the verification set. Moreover, abnormal values are more obvious and more harmful to the data, because in this case the noise becomes a real problem that cannot be ignored, no matter the abnormal values exist in the annotation or in the feature.

References

[1] He Qiang., 2016, Big data prediction. *China Statistics*, (03), pp.18-20.
[2] Lv Benfu, Chen Jian., 2014, Research on big data prediction and related issues. *Science & Technology for Development*, (01), pp. 60-65.
[3] Li Na, Li Ai'jun., 2010, Research on accurate advertising based on user characteristics classification. *Computer Knowledge and Technology*, 6(01), pp. 196-198.
[4] Yang Zhihong., 2017, Analysis of real-time bidding advertising mode under the background of big data. *Journal of Popular Science*, (10), pp. 158.
[5] Ye J , Janardan R, Li Q., 2005, Two-dimensional linear discriminant analysis. *Advances in Neural Information Processing Systems*, 17(6), pp. 1431-1441.