Research on market consumption prediction based on machine learning

Xu Zhao[1]
[1]monash university Guangzhou, China

Abstract—With the rapid development of artificial intelligence industry and big data technology in recent years, the traditional financial industry has gradually transformed into fintech. China Merchants Bank Credit Card Center proposes to rely on data to predict whether users will buy the Pocket Life APP coupons as a practical business scenario. Based on this practical problem, a variety of machine learning methods are used, including logistic regression, random forest. Xgboost, LightGBM, to explore this problem. Finally, an integrated learning method is used to fuse the final result. This paper uses the above several algorithm models for prediction. The model principle is analyzed, and the performance of each model is measured on multiple evaluation indicators. The advantages and disadvantages of different models are compared horizontally, and the reasons for the difference in results are summarized.

1 INTRODUCTION

With the rapid development of the Internet in recent years, the accumulation of data and the continuous improvement of computer computing power, how to mine the value behind the data has become an important topic of common concern for academia and industry. As a statistical learning method, machine learning uses models to automatically learn the hidden rules behind data, which has become an important means of tapping the potential value of data. It has penetrated into all aspects of people 's lives, and the application scenarios of machine learning have gradually penetrated into the financial field.

The field of machine learning can be further divided into three categories: supervised learning, unsupervised learning, and reinforcement learning. This article focuses on the research of supervised learning with labeled training data [1]. Based on the credit card data of China Merchants Bank users, it is predicted whether users will purchase handheld life APP coupons. In this study, we explored traditional single-model machine learning algorithms, such as logistic regression and decision trees, and also explored integrated learning algorithms, including random forest, GBDT, Xgboost, LightGbm and other applications in the field of financial consumption [2].

2 RESEARCH STATUS OF MACHINE LEARNING

Machine learning is a common research hotspot in the fields of artificial intelligence and pattern recognition. Its theories and methods have been widely used to solve complex problems in engineering applications and science. The winner of the Turing Award in 2010 was Professor Leslie vlliant of Harvard University. One of his award-winning work was the establishment of Probably Approximate Correct (PAC) learning theory; Professor Judea Pearll, whose main contribution is the establishment of artificial intelligence methods based on probability statistics. These research results have promoted the development and prosperity of machine learning.

Machine learning is a science that studies how to use computer simulation or realize human learning activities. It is one of the most intelligent and cutting-edge research fields in artificial intelligence. Since the 1980s, as a way to achieve artificial intelligence, machine learning has aroused widespread interest in the artificial intelligence community. Especially in the past decade, research in the field of machine learning has developed rapidly, and it has become an important part of artificial intelligence. One of the topics. Machine learning is not only applied in knowledge-based systems, but also in many fields such as natural language understanding, non-monotonic reasoning, machine vision, and pattern recognition. Whether a system has learning ability has become a sign of whether it has "intelligence". The research of machine learning is mainly divided into two types of research directions: the first category is the research of traditional machine learning, this type of research is mainly to study the learning mechanism, focusing on exploring the learning mechanism of the simulated person; Research, this type of research is mainly to study how to effectively use information, focusing on obtaining hidden, effective, and understandable knowledge from huge amounts of data.

zhaoxudehaizi@gmail.com

After 70 years of tortuous development, machine learning takes deep learning as a representative to learn from the multi-layered structure of the human brain and the layer-by-layer analysis and processing mechanism of neuron connection and interactive information. The powerful parallel information processing capabilities of adaptive and self-learning It has achieved breakthrough progress, the most representative of which is the field of image recognition.

A. Research status of traditional machine learning

The research directions of traditional machine learning mainly include decision trees, random forests, artificial neural networks, and Bayesian learning.

Decision tree is a common method of machine learning. At the end of the 20th century, machine learning researcher J. Ross Quinlan introduced Shannon's information theory into the decision tree algorithm and proposed the ID3 algorithm. In 1984, I. Kononenko, E. Roskar and I. Bratko proposed the AS-SISTANT Algorithm based on the ID3 algorithm. This algorithm allows the intersection of the values of the categories. In the same year, A. Hart proposed the Chi-Squa statistical algorithm, which uses a statistic based on the degree of association between attributes and categories. In 1984, L. Breiman, C. Ttone, R. Olshen and J. Freidman proposed the concept of decision tree pruning, which greatly improved the performance of decision trees. In 1993, Quinlan proposed an improved algorithm based on ID3 algorithm, namely C4.5 algorithm. The C4.5 algorithm overcomes the problem of the attribute bias of the ID3 algorithm and increases the processing of continuous attributes through pruning to avoid the "overfit" phenomenon to a certain extent. However, when the algorithm discretizes continuous attributes, it needs to traverse all the values of the attribute, which reduces efficiency, and requires the training sample set to reside in memory, which is not suitable for processing large-scale data sets. In 2010, Xie proposed a CART algorithm, which is a flexible method to describe the conditional distribution variable Y for a given prediction vector X, which has been applied in many fields. The CART algorithm can process out-of-order data and uses the Gini coefficient as the selection criterion for test attributes. The accuracy of the decision tree generated by the CART algorithm is high, but when the complexity of the decision tree generated by the CART algorithm exceeds a certain level, as the complexity increases, the classification accuracy will decrease, so the decision tree established by the algorithm should not be too complicated. In 2007, Fang Xiangfei expressed a algorithm called SLIQ (decision tree classification). The classification accuracy of this algorithm is comparable to other decision tree algorithms, but its execution speed is faster than other decision tree algorithms. There is no limit to the number of samples and the number of attributes. The SLIQ algorithm can handle large-scale training sample sets and has good scalability; it has a fast execution speed and can generate a small binary decision tree. The SLIQ algorithm allows multiple processors to process attribute tables simultaneously, thereby achieving parallelism. But the SLIQ algorithm still cannot get rid of the limitation of main memory capacity. In 2000, RajeevRaSto et al proposed the PUBLIC algorithm, which pruned the decision tree that has not been completely generated, thus improving efficiency. Fuzzy decision trees have also flourished in recent years. In consideration of the correlation between attributes, researchers have proposed a hierarchical regression algorithm, a constrained hierarchical induction algorithm, and a function tree algorithm. These three algorithms are based on a combination of multi-classifier decision tree algorithms, and they may have correlations between attributes. Some experiments and studies have been carried out, but these studies have not explained how the correlation between attributes affects the performance of decision trees in general. In addition, there are many other algorithms, such as a rough set-based optimization algorithm proposed by Zhang.J in 2014, and an extreme learning tree-based algorithm model proposed by Wang.R in 2015.

Random forest (RF), as one of the important algorithms of machine learning, is a method of classification and prediction using multiple tree classifiers. In recent years, the development of random forest algorithm research has been very rapid and has been applied research in many fields such as bioinformatics, ecology, medicine, genetics, remote sensing geography.

Artificial Neural Networks (ANN) is an algorithm with non-linear adaptive information processing capabilities, which can overcome the defects of traditional artificial intelligence methods for intuition, such as pattern, speech recognition, and unstructured information processing. As early as the 1940s, artificial neural networks have received attention, and have since been rapidly developed.

Bayesian learning is the earliest research direction of machine learning. Its method originated from the British mathematician Thomas, a special case of Bayes' theorem proved by Bayes in 1763. After the joint efforts of many statisticians, Bayesian statistics was gradually established after the 1950s and became an important part of statistics.

A) Research status of machine learning in big data environment

The value of big data is mainly concentrated on the turning of data and the information processing capabilities of data. In today's industry development, the arrival of the era of big data brings better technical support for data conversion, data processing, data storage, etc. Industrial upgrading and the birth of new industries have formed a driving force to enable big data to Programs that can discover things are automatically planned to achieve coordination between human users and computer information. In addition, many existing machine learning methods are based on memory theory. If big data cannot be loaded into the computer's memory, many algorithms cannot be processed. Therefore, new machine learning algorithms should be proposed to meet the needs of big data processing. Machine learning algorithms in a big data environment can ignore the importance of learning results based on certain performance standards. Using distributed and parallel computing to implement the divide-and-conquer strategy can avoid interference caused by noisy data and redundancy, reduce storage costs, and improve the efficiency of learning algorithms.

As the demand for data analysis continues to increase in various industries in the big data era, efficiently acquiring knowledge through machine learning has gradually become the main driving force for the development of machine learning technology today. Machine learning in the era of big data puts more emphasis on "learning itself is a means". Machine learning has become a support and service technology. How to conduct in-depth analysis of complex and diverse data based on machine learning and use information more efficiently has become the main direction of machine learning research in the current big data environment. Therefore, machine learning is more and more developing in the direction of intelligent data analysis and has become an important source of intelligent data analysis technology. In addition, in the era of big data, with the continuous acceleration of data generation, the volume of data has increased unprecedentedly, and new types of data that need to be analyzed are also emerging, such as text understanding, text sentiment analysis, and images Search and understanding, graphics and network data analysis, etc. It makes intelligent computing technologies such as big data machine learning and data mining play an extremely important role in the application of intelligent analysis and processing of big data. In December 2014, the Chinese Computer Society (CCF) Big Data Expert Committee voted through hundreds of scholars and technical experts in the field of big data to vote for the "Top Ten Big Data Hotspot Technologies and Development Trends in 2015", which combined machine learning Big data analysis technology such as intelligent computing technology has been selected as the largest research hotspot and development trend in the field of big data.

3 DATA

A. Data set description

The data set used in this project is mainly divided into the following three parts: one is personal attributes and credit card consumption data; the second is APP operation behavior log; the third is annotated data.

B. Data preprocessing

In order to mine the hidden information behind the palm life APP data, the following pre-processing was done to extract features:

1) Traditional feature engineering

Traditional features are mainly based on the following two subcategories: ① Basic statistical features. The total number of user clicks, the number of user clicks on each day (week), the number of days the user clicks, the average number of user clicks per day (week), maximum, minimum, mode, variance, sharpness, skewness, etc. ② Timing related features. The time interval between user clicks, the maximum number of consecutive user clicks, and the interval between the last click of the user and the last day

2) word2vec features

The TF-IDF feature fails to consider the order of user behavior, so we use word2vec to capture local co-occurrence features of user behavior. Word2vec uses shallow neural networks to embed high-dimensional sparse word vectors into a low-dimensional (100) dense space. This vector is used to represent user behavior features that contain sequence information.

3) Data set division

The China Merchants Bank Credit Card Center provides data for 31 days in March [3]. In order to fit the actual application scenario, the data is divided according to time series-the data of the first 28 days is used as training data, and the data of the last 3 days is used as test data.

4 MACHINE LEARNING MODEL

A. Logistic regression

Logistic Regression is a very widely used classification model in machine learning. It fits the data to the sigmoid function to complete the prediction of the probability of occurrence.

B. Random Forest

The two main methods in the integrated learning method are Bagging and Boosting. The Bagging model can learn multiple base models in parallel, and the results of the base model are averaged to obtain the final result of the model. Random Forest is a typical Bagging algorithm based on CART decision tree. In order to reduce the variance of the model and reduce the overfitting, the integrated learning algorithm needs to increase the difference of the base model. Random forest algorithm mainly uses bootstrap sampling to increase the difference of training data and feature sampling to increase the difference of features [4].

C. Xgboost, LightGBM

Xgboost model and LightGBM model are both typical boosting algorithms, both are algorithms and engineering improvements to GBDT model. Different from the Bagging model, the base learners can be parallel, and there is a dependency between the base learners of the Boosting model. GBDT is a lifting tree model. In the m-th round, a CART regression tree is used to fit the negative gradient of the previous m-1 round loss, which reduces the model's bias. Compared with GBDT, Xg-boost optimizes the loss function, introduces second derivative information, and adds regular terms to control the complexity of the model; in addition, although the training of the base model exists in order, there is a tree inside each base learner Node splitting can be done in parallel, and Xgboost optimized it in parallel. Compared with Xg-boost, LightGBM proposes the Histogram algorithm to bucket features and reduce the complexity of querying split nodes. In addition, the GOSS algorithm is proposed to reduce small gradient data. At the same time, the EFB algorithm is proposed to bundle mutually exclusive features and reduce feature dimensions To reduce model complexity.[5]

5 Comparison of experimental results

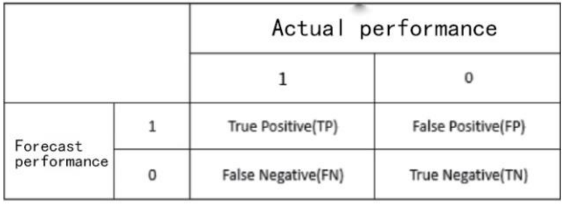
A. Evaluation index

In this experiment, comprehensive use of accuracy. precision, recall, f1_score, AUC as measurement indicators.

1) Precision rate, recall rate, F1

The confusion matrix is a visual tool for comparing the prediction results with the real results in the supervised learning classification task, as shown in Figure 1.

The confusion matrix (Figure 1) contains four values of TP, FN, FP, and TN: TP represents the true example, that is, the number of samples where the prediction result and the true result are 1; FP represents the false positive example, that is, the prediction result is 1, However, the number of samples with a true result of 0; FN represents a false negative example, that is, the number of samples with a predicted result of 0, but a true result of 1; TN represents a true negative example, that is, the number of samples with a predicted result and a true result of 0.



		Actual performance	
		1	0
Forecast performance	1	True Positive(TP)	False Positive(FP)
	0	False Negative(FN)	True Negative(TN)

Fig. 1 Confusion matrix

$$Accurracy=\frac{TP+TN}{TP+FP+FN+TN}$$

$$Precision=\frac{TP}{TP+FP}$$

$$Recall=\frac{TP}{TP+FN}$$

$$F_1=\frac{2*P*R}{(P+R)}$$

In statistics and machine learning, AUC is often used to evaluate the performance of binary classification models. The full name of AUC is areaunderthecurve, which is the area under the curve.

For binary classification problems, the prediction model predicts a probability p for each sample. Then, a threshold t can be selected to make the sample with score p> t predict positive, and the sample with score p <t predict negative. In this way, according to the predicted results and the actual label, the sample can be divided into 4 categories: TP, FN, FP, TN.

As the threshold t changes continuously, the values of TP, FN, FP and TN also change continuously. Define the true case rate TPR and false positive rate FPR as:

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

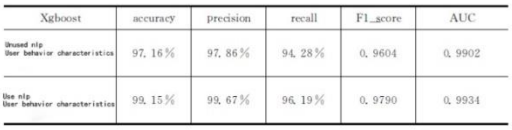
$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$$

Adjust the threshold value p to get different TPR and FPR values. This curve is the ROC curve. The area under the ROC curve is AUC.

B. Experimental results

Comparing nlp user behavior feature effect improvement

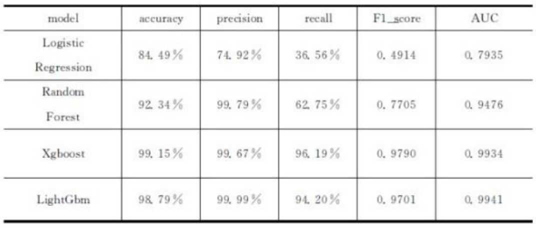
TABLE 1 NLP CHARACTERISTICS COMPARISON



Xgboost	accuracy	precision	recall	F1_score	AUC
Unused nlp User behavior characteristics	97. 16%	97. 86%	94. 28%	0. 9604	0. 9902
Use nlp User behavior characteristics	99. 15%	99. 67%	96. 19%	0. 9790	0. 9934

Comparison of the effects of various models

TABLE 2 COMPARISON OF VARIOUS MODELS



model	accuracy	precision	recall	F1_score	AUC
Logistic Regression	84. 49%	74. 92%	36. 56%	0. 4914	0. 7935
Random Forest	92. 34%	99. 79%	62. 75%	0. 7705	0. 9476
Xgboost	99. 15%	99. 67%	96. 19%	0. 9790	0. 9934
LightGbm	98. 79%	99. 99%	94. 20%	0. 9701	0. 9941

C. Analysis of experimental results

①Comparing Table 1 and Table 2, it can be seen that by introducing TFIDF features and Word2Vec features to obtain user behavior features, it helps the model to better mine the laws of data and improves the performance of the model. ② Random forest, Xg-boost, Lightgbm and other ensemble-based models are superior to logistic regression in accuracy, precision, recall, f1, and AUC, indicating that the tree model may be more suitable for this data set and use the ensemble method To fuse weak classifiers, its performance is better than a single classifier. ③ Boosting-based integrated learning algorithm (Xgboost, Lightgbm) is better than Bagging-based integrated learning algorithm (RandomForest), indicating that for the data and features, the importance of reducing bias is better than reducing variance.

6 Summary

In this project, the financial scenario data is modeled through machine learning methods to predict whether China Merchants Bank credit card users will purchase Pocket Life APP coupons. According to the experimental results, the GBDT-based Xgboost model and LightG-BM model exceed 0.9 in each evaluation index, which verifies the superiority of the machine learning model. It can be used in actual CTR scenarios to enhance the user experience of China Merchants Bank's handheld life APP and help companies to obtain more profits.

References

[1] Russell Man, Han Lu, Xu Qin, etc. Explore the application of commercial banks in the field of big data mining technology [J].

Computer Application and Software, 2017, 34 (9): 43 ～ 45 + 81.

[2] Kleinbaum D G, Dietz K, Gailil M, et al. Logistic REGRESSION[M]. New York: Springer-Verlag, 2002.

[3] Liaw A, Wiener M. Clasififacion AND REGRESSION BY RANDOMFOREST[J]. R news, 2002, 2 (3): 18 ～ 22.

[4] Liow A, Wiener M. Classifiance and end registration by randomForest [J]. Rnews, 2002, 2 (3): 18 ~ 22.

[5] Steinberg D, Colla P. CART: Classifiance and Andregresions [J]. The top ten algoritms in indata mining,2009, 9: 179