Automation and Machine Learning(2019) Vol. 1: 1-4
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2019.11001
ISSN 2516-5003

Prediction of purchasing power of Google store based on deep ensemble learning model

Wang Peng[1,a,*]

[1]*School of Economics and Management, Dalian University, No.10, Xuefu Avenue, Economic & Technical Development Zone, Dalian, Liaoning，The People's Republic of China（PRC）*
a. email：wangpeng1@dlu.edu.cn

Keywords: *machine learning; data mining; deep model; ensemble learning; feature engineering*

Abstract: Aiming at the defect that the ensemble learning model such as Light Gradient Boosting Machine only mines the data information once, which can not automatically refine the granularity of data mining and dig into the more potential internal correlation information of data, the ensemble learning model is made into a deep form by sliding window and deepening, and the deep ensemble learning is proposed. Sliding window enables the ensemble learning model to automatically refine the granularity of data mining, so as to dig deeper into the potential internal correlation information in the data, and at the same time endue it with certain representation learning ability. Based on the sliding window, the deepening step further improves the representation learning ability of the model. Finally, the results show that the prediction accuracy of the deep ensemble learning model is 6.16 percentage points higher than that of the original ensemble learning model.

1. Introduction

The integrated learning models widely used in data mining tasks today are Random Forest, LightGBM, and Catboost. Random forest belongs to bagging in integrated learning, and LightGBM and Catboost belong to the integrated learning. They are all based on the integrated learning model of gradient-driven decision tree, and have the advantages of high prediction accuracy, fast training speed and low memory consumption. Mainly explore the mainstream solution of the problem; but at present, most of the application of the integrated learning model only puts the obtained feature set into the model for training and learning. It only mines the data association once, and the mining granularity is rough. With the ability to dig deep, it is not possible to automatically mine deeper relationships between data[1].

In order to obtain more data-related information, the current mainstream methods rely on the artificially designed feature engineering to process the original feature set, and then put the processed data into the integrated learning model for training; the artificially designed feature engineering is very Relying on the experience of the designer and the large amount of analytical data work, how to enable models such as LightGBM to automatically obtain depth information is key[2].

For how to automatically obtain depth information in the data, the deep network has achieved excellent results in both image processing and natural language processing tasks. For example, Long Short Term Memory (LSTM) and its variants are widely used in tasks such as natural language processing, and deep convolutional networks are widely used in image processing tasks. Among them, the ability to learn is recognized as an integral part of the deep network. If the integrated learning model can be given a certain ability to express learning, it will be able to automatically mine deeper relationships in the data.

The introduction of deep forests in 2017 brought a new direction of thinking to the depth model – models that can be made into deep forms are not just neural networks. Drawing on the idea of deep forests, this paper makes the LightGBM and Catboost models into a deep form, so that it has the ability to mine deep information in data. Firstly, this paper divides the feature set by adding sliding window, and refines the data mining granularity, so that the model can potentially have situational awareness or structural awareness. On the basis of the sliding window, the model's representation learning ability is further enhanced by deepening operation. Compared with the original integrated learning model, the deep integrated learning model proposed in this paper can automatically explore the potential relationship between more features and dig deeper into the potential information of the data instead of just staying at the existing feature level. Moreover, the basic model used is an integrated tree model. The tree model has better interpretability than other models. The partitioning idea of the tree nodes is very similar to the human thinking process, which is helpful for the analysis of the model. And research.

2. Basic model

LightGBM is an efficient implementation of GBDT, although there are already some algorithms for GBDT implementation, such as XGBoost (eXtreme GradientBoosting, XGBoost) pGBRT (parallel Gradient Boosted Regression Trees, pGBRT), scikit-learn (machine learning in Python). But when the characteristic dimension of the data is high and the number of samples is large, their performance is not satisfactory. The main reason for this is that the above implementation algorithms need to traverse all data samples and then estimate the information gain of all possible partition points, which is very time consuming. Therefore, LightGBM proposes two solutions: 1 Gradient-based One-Side Sampling (GOSS); 2 Exclusive Feature Building (EFB).

The GOSS algorithm (Algorithm 1) excludes a large number of samples with small gradients and uses only the remaining samples for information gain estimation. LightGBM paper proves that samples with large gradients play a role in calculating information gain. A more important role, GOSS can get very accurate information gain calculations with smaller data[3].

3. Model of this paper

3.1. Sliding window

In order to enable the integrated learning model to automatically refine the data mining granularity and potentially have situational awareness or structural awareness on the dataset, this paper proposes to cascade the sliding window (corresponding to the inner and outer dashed boxes in Figure 1) and the original feature vector. See Algorithm 4 for the specific algorithm flow as a new feature vector. Moreover, K-fold cross-validation is used for training during training to avoid over-fitting.

It can be clearly seen from Figure 1 and Algorithm 4 that the outermost dashed box in Figure 1 corresponds to step 2) of Algorithm 4: full feature training and its prediction results. The inner dashed box in Figure 1 corresponds to the fourth in Algorithm 4. 4) Step: sliding window training

and its prediction results, the rightmost output vector of Figure 1 corresponds to the output of Algorithm 4; whether it is a full feature or a sliding window can be regarded as a sliding window, but the window_size is different.

3.2. Deepen

In order to make the proposed algorithm also have a certain ability to represent learning in the depth network, the model hierarchy is deepened on the basis of the sliding window, and the algorithm representation ability is further enhanced by deepening the model hierarchy; the specific algorithm See algorithm 5

3.3. Feature engineering

This section will perform feature engineering operations by analyzing the importance of each feature in the Google Store customer dataset in LightGBM, where each row of the dataset represents a session and each column corresponds to a feature. Table 1 explains some of the more important features in the original feature set.

4. Experiment and result analysis

4.1. Data set

The dataset used in this article is the Google Analytics Custmer Revenue Prediction dataset from the Kaggle website around September 2018. It analyzes the customer dataset of Google Merchandise Store (also known as GStore swag is sold) to predict each customer in the future. Purchasing power. It can ensure the authenticity and validity of the data source, and it can also prove that the method proposed in this paper has practical application value. The original data set stated in this article.

Through a simple analysis of the data, it can be found that the data given by the competition does meet the 80/20 criteria. Figure 6 shows that most of the users who purchase the goods are concentrated above the number 70000, and the users before this almost produced nothing. profit.

4.2. Result analysis

It can be found that after the operations of adding features and deleting features, the final feature engineering obtained has improved the prediction accuracy of the model, especially the LightGBM model has a significant increase of 2%; indicating that the semi-automatic feature engineering proposed in this paper is for LightGBM. Effective. Compared to the original LightGBM and Catboost models that have not been improved, the resulting depth model consists of a sliding window and a deepening. In Table 5, LightGBM represents the prediction accuracy of the original LightGBM model after training on the original data set; the addition feature and the deletion feature correspond to the two-step feature engineering in the previous section.

Compared with the Google Store customer data set without semi-automatic feature engineering processing, the original LightGBM model is directly put into training prediction. Finally, the Google Store customer data set is processed by semi-automatic feature engineering and then placed into the deep LightGBM integrated learning model. The accuracy of the medium training prediction is 6.16 percentage points higher; thus, the deep LightGBM integrated learning model and semi-automatic feature engineering proposed in this paper can indeed obtain more potential information in the data set.

5. Conclusions

The experimental results show that the deep LightGBM integrated learning model proposed in this paper has deep mining ability, and can find more in-depth relationship between customer data of e-commerce platform, and also shows that semi-automatic feature engineering is also effective. Through the deep LightGBM integrated learning model and semi-automatic feature engineering, the prediction accuracy has been significantly improved. Not only can neural networks be made into deep forms, other machine learning models can be made into deep forms, and most of the current depth models are neural networks; and these non-neural networks have fewer depth model parameters and smaller model footprints. The accuracy is almost the same as or even higher than the deep neural network. I hope that more non-neural network depth models will emerge in the future.

Acknowledgements

This article was specially funded by Dalian University's 2019 Ph.D. Startup Fund (20182QL001) and 2019 Jinpu New District Science and Technology Project.

References

[1] Pouria Sadeghi-Tehran, (2017) Multi-feature machine learning model for automatic segmentation of green fractional vegetation cover for high-throughput field phenotyping, Plant methods, 1，96-108

[2] Sung Kyun Park. (2015) Construction of environmental risk score beyond standard linear models using machine learning methods: application to metal mixtures, oxidative stress and cardiovascular disease in NHANES, Environmental Health, 1, 356-374.

[3] Tanchanok Wisitponchai. (2018) AnkPlex: algorithmic structure for refinement of near-native ankyrin-protein docking. BMC Bioinformatics, 2, 119-136.