ISSN: 2664-9640

Research progress of feature importance evaluation in ensemble learning model

Shunan Wang *

School of Shanghai, Maritime University, Shanghai 201306, China.

* Corresponding Author

Abstract

In recent years, machine learning has developed rapidly, especially ensemble learning models. Ensemble learning models are used in many fields, and the most widely used is to predict various models. In the process of building a prediction model, there are often many influencing factors in the prediction task, and each influencing factor has a different impact on the prediction result. Therefore, it is necessary to evaluate the importance of each feature. The feature importance evaluations in different learning models are also different. Therefore, this paper will introduce the research progress of feature importance evaluation in ensemble learning model in detail, and provide a theoretical basis for the selection of the importance evaluation method for subsequent model prediction.

Keywords

Integrated learning model, Decision tree, Importance evaluation.

1. Introduction

Machine learning is one of the most important recent developments in artificial intelligence. With the great progress and development of computing speed and algorithms for programming, machine learning has grown rapidly. In the process of machine learning, in addition to using a single algorithm model, integrated learning model can also be used. Ensemble learning can better complete the prediction task by constructing multiple learners and combining them. It is also often called model fusion or committee based learning. In the prediction task, there are often many influencing factors, and each influencing factor has different influence on the prediction results. Therefore, it is necessary to evaluate the importance of each feature. The evaluation of feature importance in different learning models is also different. Therefore, the text will first introduce various integrated learning models, and then introduce the importance evaluation method in the integrated learning model.

2. Integrated learning model

Ensemble learning is a machine learning method. Its core idea is to integrate several "weak learners" into one "strong learner" to improve the performance of learners. Therefore, it has been paid more and more attention and used [1]. There are generally two types of integrated learning: boosting, random forest and bagging. In boosting ensemble learning, there is a dependency between learners, but there is no dependency between learners in bagging [2].

2.1. Decision tree

Because random forest is an algorithm based on decision tree, firstly, the research status of decision tree is introduced. In 1966, hunt et al. First used the decision tree algorithm for concept learning in their concept learning system CLS learning algorithm [3]. There are three main decision tree algorithms: ID3 and C4 5、CART。 Based on the initial empty decision tree, CLS

ISSN: 2664-9640

improves the performance of the decision tree by continuously adding decision nodes until the decision tree can correctly classify a specific test set. In 1984, Breiman et al. Proposed the classification and regression trees cart algorithm [4]. This method selects the attribute with the minimum Gini index as the extended attribute to segment the node, divides the sample set of the current node into two, and each non leaf node of the final decision tree has two branches. Quinlan proposed iterative dichotomizer 3-id3 algorithm in 1986 [5], which adopts divide and conquer strategy and the principle of maximum information gain in the selection of extended attributes of each node of the decision tree, so that the finally selected extended attributes can ensure that the samples of the test set can obtain the maximum class information. Then, in 1996, Quinlan proposed c4.0 based on ID3 algorithm 5 algorithm [6], C4 While inheriting the advantages of ID3 algorithm, the information gain rate is used as the selection standard of extended attributes, which effectively improves the deficiency of information gain as the selection standard of extended attributes and avoids the over fitting of constructing decision tree. In 2002, Ruggieri proposed a C4 Improved algorithm of ec4.5 5 algorithm [7], EC4 5 overcome C4 5 algorithm has the defect that the search threshold is too large when dealing with continuous value attributes, which makes the scale of the decision tree smaller. Its main defect is the excessive demand for memory.

2.2. Random forest

Random forest is a new algorithm based on decision tree and bagging algorithm, which belongs to the expansion of bagging algorithm. Random forest was first proposed by Leo Breiman and Adele Cutler to solve the over fitting problem and make the model more generalized. Although the decision tree can improve the over fitting problem by pruning, the decision tree is trained according to the data set, and the noise data still exists. In order to solve this problem, the decision tree and bagging are combined to solve this problem.

2.3. Boosting algorithm

Boosting is an algorithm that promotes weak learners to strong learners through iteration, which was first proposed by schapire [8] [9]. Weak learners are learners whose generalization performance is slightly better than random guess. The difference between boosting and bagging is that each base learner is associated. Firstly, the algorithm trains a weak learner according to the data set, changes the sample weight, increases the weight of the wrong prediction sample, reduces the weight of the correct prediction sample, and generates a new data set, and then continues to train the model until the training reaches the specified number of learners, and then weights and combines the number of learners into a new strong learner. The main representatives of boosting algorithms are AdaBoost, GB, GBDT and xgboost.

Here we mainly introduce GBDT algorithm. Its full name is gradient boosting decision tree, that is, gradient boosting decision tree. Here, we combine gradient boosting algorithm with decision tree. GBDT still combines the base learner through the addition model to form a strong learner. GBDT Ricky learner is the cart tree in the decision tree. Cart decision tree is a classification and regression tree, which can be used for both regression and classification problems in modeling problems. The difference between GBDT algorithm and AdaBoost algorithm lies in how to identify the problems of the trained weak learner. AdaBoost algorithm identifies samples with classification errors, while gbdt algorithm optimizes the model through negative gradient.

3. Feature importance evaluation based on ensemble learning model

3.1. Random assessment method of forest importance

There are mainly two methods to evaluate the importance of random forest features in the integrated learning model: one is the mean decrease probability, which is commonly measured by Gini, entropy and information gain. Now this method is used in sklearn; The other is the

ISSN: 2664-9640

mean decrease accuracy, which is often measured by the out of pocket error rate (00B). Ge bin [10] according to the data characteristics of SCADA historical data of wind farm, the random forest algorithm is used to classify and predict the sample data, so as to improve the accuracy of classification and prediction. At the same time, the importance evaluation of the reduction of average accuracy is used to calculate the attribute importance of the input parameters. When the dimension of the input factors is reduced, the weight value of each important factor is obtained and the decision table is given. Lin Xia [11] et al. Used decision tree, random forest and gradient lifting regression tree to predict the monthly oil production. For the three algorithms, the characteristic importance method was used to calculate the main controlling factors of production indicators affecting oil production. The characteristic importance analysis showed that water content, production days and dynamic liquid level were the main controlling factors of oil production. Song Bowei [12] used machine learning to detect the security of Android applications in response to the security threat of Android applications with rapid iteration. Logistic regression algorithm and random forest algorithm are used to train and detect the data set. For random forest algorithm, the contribution of different features in building decision tree is expounded based on Gini coefficient. Li Huan [13] et al. Proposed a random forest model integrating factor analysis, taking the accuracy and running time of classification prediction, regression fitting and feature importance analysis of the model as evaluation indicators, in which OOB data is used to estimate the generalization error of RF and evaluate the importance of features. Jia Hanxi [14] and others evaluated the importance of the influencing factors of fire loss in order to improve the accuracy of post earthquake fire loss assessment. The random forest algorithm is used to model and analyze it, and the importance ranking of seven factors is obtained.

3.2. Feature importance evaluation method of xgboost and gbdt algorithms

Xgboost and gbdt in the integrated learning model have five main methods for evaluating the importance of features:

(1) Weight (split): indicates the total number of times a feature is used to split nodes in the whole algorithm. The more it appears, the greater the "contribution" of the feature to the whole algorithm.

(2) Cover (average coverage): the weighted result of the number of times a feature is used to divide data in all decision trees and the number of samples involved in each division.

(3) Gain (average gain): the average of the training loss and benefit reduction caused by a feature when it is used to divide data.

(4) Tree SHAP: A Novel Local Approach [15][16]. The first three methods have certain limitations and cannot compare the magnitude and direction of the influence of each variable. The SHAP framework is introduced to solve the visualization problem of the model. SHAP (SHAPley Additive exPlanations) is a framework based on additive feature attribution methods. It was first proposed by Lloyd Shapley in game theory in 2017, and the method is named after the proposer. The method initially mainly studied the value of each participant in the field of game theory. In recent years, some scholars have also explained the value of variables in complex models based on this idea. That is, for the output of a model, each variable becomes a participant in the final result of the model [17]. The SHAP explanatory framework measures the value of the participants. For a sample, each variable is its participant, so each variable of a sample will each correspond to a SHAP value. In order for the SHAP interpretive framework to meaningfully relate the actual value of a variable to the SHAP value, the SHAP value must have three properties, namely local precision, tolerance for missingness, and consistency. Local accuracy ensures that the sum of the contributions of all features is equal to the output of the model, which meets the basic requirements of the additivity interpretation framework; allowable deletion ensures that if a feature is missing, the contribution of the feature is 0; consistency

ISSN: 2664-9640

ensures the robustness of the model and The robustness of the SHAP calculation results, that is, if the model changes, and the effect of a feature on the output is not reduced, then its contribution will not be reduced. SHAP values allow for variable-level visualization and sample-level visualization. Because the calculation of the SHAP value adopts a sample-level exhaustive simulation method to calculate the average influence of the existence of each variable in the sample on the model prediction, the SHAP value can represent the contribution of each variable in a sample; Statistical analysis of SHAP values provides variable-level visualization. The effects that can be achieved by the SHAP framework are shown in formulas 1 and 2, that is, the real value Xi of each variable will be converted into a SHAPi value, and accumulated to become the final SHAPy value, and SHAPy will be mapped to the predicted probability score of the model through the sigmoid function.

$$SHAP_{base} + \sum_{i=1}^{N} SHAP_i = SHAP_y \quad (1)$$

$$\hat{y} = \frac{1}{e^{-SHAP_y}+1} \quad (2)$$

(5) Saabas algorithm: a heuristic local method. From a blog named Saabas, the concept involved is the internal value of the node, the basic value. Its internal logic is: for the predicted value of a sample in a decision tree, it is obtained from the sample starting from the root node and going through a series of feature splits in turn, so the features in the entire decision path will determine The final predicted value of the sample, the contribution value of the feature is obtained by calculating the difference between the internal values (Value) of the front and rear features in the path in the tree, which is recorded as the Saabas value of the feature, and the sample predicted value = base value + the value of each feature in the path Saabas value, it is worth noting that the Saabas value of the same feature may be different on different paths.

$$A_i(MP) = \sum_{k=1}^{k} b_k + \sum_{i=1}^{M}\sum_{k=1}^{K} C(MP, i) \quad (1)$$

Among them: $A_i(MP)$ represents the importance of the ith feature; b_k represents the basic score of the root node of the kth decision tree; $C(MP, i)$ represents the contribution of the ith feature to the sample MP, and the ith feature The feature contribution value is the difference between the parameter value of the next node on the path and the parameter value of this node.

The above method has been applied by many scholars. Youngman [2] used ensemble learning to study the movie box office, and verified that the movie box office prediction based on the XGBoost algorithm was better than the random forest algorithm and the GBDT algorithm through two datasets. The XGBoost algorithm is used for feature screening, reducing the complexity of the model, and retraining the data to build the model to achieve the purpose of simplifying the model. The selected feature importance indicator is "gain", and a function plot_importance is used when training the XGBoost model in Python to obtain the feature importance ranking, and it is obtained that factors such as actors, movie release dates, and movie types will have an important impact on movie box office factors. Zhong Minhui [18] and others proposed a railway accident type prediction and cause analysis algorithm based on the Grandient Boosting Decision Tree (GBDT). Aiming at the problem of the imbalance of railway accident record data categories, an integrated GBDT model was proposed. , to complete the robust prediction of accident types. On this basis, according to the feature importance ranking in the GBDT prediction model, the global importance of a feature xj is measured by the average importance of the feature in a single decision tree. That is (average gain), to realize accident cause analysis.

4. Conclusion

To sum up, feature importance evaluation in ensemble learning models has been involved in various fields in recent years, including movie box office prediction, railway accident prediction, traffic prediction, diabetes prediction, etc. many. Among them, the first three methods for

evaluating feature importance of XGBoost and GBDT appeared earlier, so they were used more by predecessors. It solves the situation when the results obtained by the current three algorithms are inconsistent, and the corresponding contribution degrees can be obtained for different samples. Therefore, the ensemble performance of applying decision tree is not only better than other ensemble methods, but also can find out the key factors that affect the prediction results. However, in the field of estuary back-silting prediction, the research results of introducing decision tree method are still blank. Therefore, by using the decision tree method and taking advantage of its importance assessment, the ability to assess the importance of channel back silting influencing factors under high-dimensional and multi-factor coupling can be improved.

References

[1] Zhihua Z:Ensemble Methods: Foundations and Algorithms.(Taylor & Francis,2012)

[2] M Yang:Film box office prediction based on xgboost algorithm (MS., Lanzhou University, China ,2020.)

[3] E. B. Hunt, J. Marin, P. J. Stone:Experiments in induction. (Academic Press,New York, 1966)

[4] L. Breiman, J. Friedman, C. J. Stone, et al:Classification and regression trees. (Chapman & Hall,1984)

[5] J. R. Quinlan: Induction of Decision Trees,Machine Learning (1986) No.1, p.81-106.

[6] J. R. Quinlan: Improved Use of Continuous Attributes in C4.5, Journal of Artificial Intelligence Research (1996) No.4,p.77-90.

[7] S.Ruggieri:Efficient C4.5 [classification algorithm],Transactions on Knowledge and Data Engineering, Vol.14(2002) No.2,p. 438-444.

[8] Schapire R E, Singer Y:Improved Boosting Algorithms Using Confidence-rated Predictions, Machine Learning,(1999)No.37,p.297-336.

[9] Freund Y,Schapire R E:A decision-theoretic generalization of on-line learing and an application to Boosting,Journal of Computer and System Science,Vol.55 (1997) No.1,p.23-37.

[10] B Ge:Data preprocessing method based on importance analysis and research on fan active power prediction,Technology and Innovation(2020) No.14,p.10-12.

[11] X Lin, Z S Lin,Y Gao, B Y Wu: Analysis of main controlling factors of oil production based on machine learning, Information System Engineering,(2019)No.12,p.94-97+99.

[12] B W Song:Research on Android software security based on feature importance (MS,Xidian University,China,2019)

[13] W H Li:Research on defect report recommendation method based on feature importance (MS,Qingdao University of Science and Technology,China,2020)

[14] H X Jia,J Q Lin,J L Liu:Research on the influencing factors of fire loss based on random forest,Fire Science and Technology,Vol.38(2019)No.11,p.1642-1644.

[15] Lundberg S M , Lee S I :Consistent feature attribution for tree ensembles. (2017)

[16] Lundberg S , Lee S I : A Unified Approach to Interpreting Model Predictions(Nips), 2017.

[17] W H Chen:XGBoost-based overdue identification and model representation of Internet financial loans. (MS,Harbin Institute of Technology, China,2019)

[18] M H Zhong, W L Zhang, Y R Li, Z F Zhu, Y Zhao. Prediction and Cause Analysis of Railway Accident Types Based on GBDT. Journal of Automation(2020), p.1-9.