Missing Value Filling Research Based on Ensemble Learning

Jianqiao Sun

Yanshan University, Qinhuangdao 066000, China.

Abstract: This paper studies missing value filling and compares the filling effects of five methods: Mean, KNN, Random Forest, GBDT, and Stacking under different missing proportions, proving the superiority of ensemble learning algorithms in filling performance when multiple feature values are missing. Then the missing value filling method of KNN+integrated learning is proposed to further improve the filling performance.

Keywords: Missing Value Filling; Ensemble Learning; KNN

1. Introduction

Faced with the problem of missing data, the research of experts and scholars in the missing value filling is also enriched. Zheng Zhiquan and others [1] weighted the general K-nearest neighbor algorithm by using Gaussian function and proved that the weighted K-nearest neighbor algorithm is superior to the traditional K-nearest neighbor algorithm in terms of filling effect regardless of the size of the K value. Du Yingkui and others [2] used mean value method, regression method and multiple interpolation method to process missing values to improve the prediction level of LSTM model. Zhang Xiaoqin and Cheng Yuying [3], due to the special geometric properties of component data, cannot directly apply traditional filling methods to such data, so they proposed an iterative filling method for missing values of component data based on random forest. Atiq and others[4] combined KNN filling with deep learning classifier to predict coupon acceptance and obtained relatively good results.

Integrated learning has a long history and is difficult to trace back. It has been continuously enriched and perfected and applied to all fields of human production and life. Zhang Mingwei and others [5] proposed a new diabetes detection method based on Stacking integrated learning. Shi Yuntao and others [6] used the method of Stacking to establish models in order to analyze and warn the security risks of meat and meat products in major activities, and achieved good results.

On the basis of the above research, this paper compares a variety of filling methods, and improves the method with better performance under different loss rates, in order to improve the filling performance and restore the real data to the maximum extent.

2. Knowledge preparation and filling experiment

The most commonly used missing value filling methods are statistic filling and regression filling. In this paper, Mean in statistic filling, KNN (k-Nearest Neighbor) and Random Forest in regression filling are used. GBDT (Gradient Boosting Decision Tree) and Stacking are compared with five methods. Among them, Random Forest, GBDT and Stacking are integrated learning algorithms.

The statistical filling method usually uses the mean, median and mode of the feature of the missing variable to fill in the missing value, while the regression filling method usually uses various algorithms to perform regression prediction on the missing value to achieve the filling of the missing value. This paper mainly selects the following algorithms for regression prediction. These include a single learner KNN and integrated learning methods Random Forest, GBDT, and Stacking.

(1) KNN

KNN is one of the most common supervised learning methods, through a given test sample, find the nearest K samples, and make predictions according to the nearest K “neighbors”. Usually, in the classification task, the category that will appear the most is taken as the result by voting, and in the regression task, the average value is calculated as the result, and the result can be obtained by weighted voting or weighted average according to the distance. The distance in this paper is measured using Euclidean distance with a K value of 5, and all “near neighbors” are treated equally without weighting.

(2) Random Forest

Random Forest is a typical representative of Bagging in ensemble learning. Bagging uses self-sampling to generate different basis learners, and Random Forest adds random feature selection on the basis of it. That is, when selecting segmentation points in each decision tree, a feature subset is randomly selected first, and then the segmentation points are selected on this basis. This article allows each tree to choose among all the features.

(3) GBDT

GBDT is a kind of ensemble learning Boosting. The loss function in GBDT can be an arbitrarily differentiable function that influences the establishment of the next base learner by fitting the residual, and samples and features can be sampled before the tree is built.

(4) Stacking

Unlike the Random Forest algorithm and GBDT algorithm that are homogenous integrators above, Stacking is heterogeneous integration. It can integrate multiple different types of learners. It can be regarded as both an algorithm and an integration strategy. When Stacking is applied, this paper selects Random Forest, GBDT (Gradient Boosting Decision Tree), XGBoost, and Stacking is selected in this paper. LightGBM(Light Gradient Boosting Machine), SVM(Support Vector Machine), KNN are first-level learners, and linear regression are second-level learners.

This paper uses Aliyun Tianchi industrial-steam volume forecast data set to carry out the missing value filling experiment. In order to compare the filling effect of different filling methods, five features were randomly hollowed out with 5%, 10%, 20% and 30% missing proportions respectively.

When Mean is used to fill in the missing value, the mean of each feature is directly calculated to fill in the missing value. When the regression filling method is used to fill in the missing values, the proportion of missing values of each feature is sorted first, the feature with the least missing proportion is filled first, and the missing values of other columns except the selected filling column are filled with Mean. The set of rows with non-missing data in the filling column is the training set, and the set of rows with missing data in the filling column is the test set. Use the selected algorithm to train using the training set. The test set is then input to generate data of missing positions and return the original data set to fill in the missing values. And so on, the missing values of the next feature are filled until all the missing values of the feature are filled.

3. Experimental results and method improvement

In this paper, the average absolute error (MAE) is used as a measure of filling effect. The calculation formula is as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \quad (1)$$

MAE results with different deletion ratios are shown in Table 1. It can be seen from the table that under the four missing proportions of 5%, 10%, 20% and 30%, the filling effect of heterogeneous integrated Stacking is the best, followed by Random Forest. On the whole, the filling effect of integrated learning algorithm is better than that of mean filling and single learner filling.

Table 1. Comparison of filling effect of each algorithm under different missing ratio.

Missing ratio	Mean	KNN	RandomForest	GBDT	Stacking
5%	0.725	0.318	0.254	0.289	0.211
10%	0.753	0.347	0.253	0.282	0.215
20%	0.745	0.345	0.268	0.288	0.228
30%	0.735	0.352	0.279	0.297	0.238

In the missing value fill above, the quick fill uses traditional mean fill. In this paper, KNN is added for the first round interpolation (KNNI) when the missing value filling process is optimized. KNN is one of the most common supervised learning methods in which the mean value is calculated as a result in a regression task. This paper presents a KNNI-Stacking method based on weighted average of distance and stacking to further improve the accuracy of missing value interpolation.

Figure 1 shows the comparison of MAE before optimization and the optimized KNNI-Stacking method. It can be seen from the figure that KNNI-Stacking methods with different missing proportions achieve lower MAE and higher filling accuracy than before optimization.

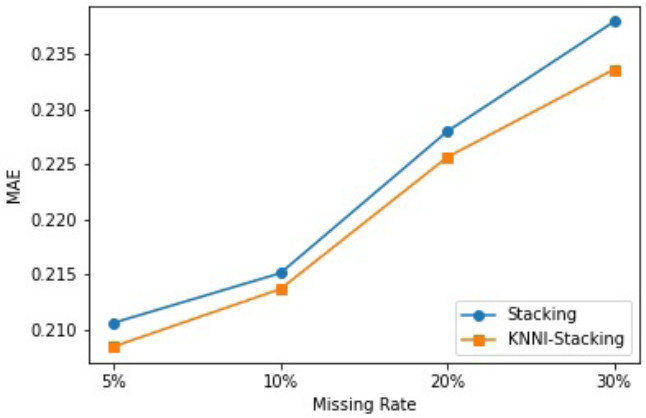

Figure 1. Comparison of interpolation effect before and after optimization of missing value interpolation method.

4. Conclusions

Based on the data missing problem, this paper compares different filling algorithms and draws the conclusion that ensemble learning algorithm has advantages in filling missing values. Meanwhile, KNN+ ensemble learning method is proposed to further improve the filling performance.

References

[1] Zheng Zhiquan, Wang Mengmeng, Tian Weiqi. Research on Missing data filling based on weighted K-nearest neighbor algorithm [J]. Intelligent Computers and Applications, 2021, 11(11).

[2] Du Yingkui, Zhang Yifang, Yuan Zhonghu, et al. Analysis of air pollution prediction accuracy of LSTM network by data Preprocessing [J]. Computer and Digital Engineering, 2021, 49(7).

[3] Zhang Xiaoqin, Cheng Yuying. Missing value filling method of component data based on random forest model [J]. Applied Probability and Statistics,2017,33 (1).

[4] Atiq R, Fariha F, Mahmud M, et al. A Comparison of Missing Value Imputation Techniques on Coupon Acceptance Prediction[J]. International Journal of Information Technology and Computer Science(IJITCS), 2022, 14(5).

[5] Zhang Mingwei, Zhang Tianyi, Zhong Ming, et al. The significance of arterial damage in the early detection of diabetes mellitus verified by the integrated learning algorithm Stacking [J]. Chinese Journal of Medical Physics,2022,39(8).

[6] Shi Yuntao, Ren Peng, Li Shuqin, et al. Safety risk analysis and prediction of active meat and meat products based on Ensemble learning [J]. Journal of Food Safety and Quality Inspection, 2019,13(16).