Research on the Prediction Model of House Rent Based on Machine Learning

Hongtao Zong[1,a], Jihong Song[1,b]

e-mail: zonghongtao@smail.sut.edu.cn[a]
e-mail: songjihong@sut.edu.cn[b]

Software School, Shenyang University of Technology, Shenyang, China[1]

Abstract—How to provide renters with rent reference based on housing characteristics is an urgent issue to be solved, and GBRT (Gradient Boosting Regression Tree) provides a solution to the rent prediction problem. However, in GBRT, there is a problem that the accuracy of model prediction is not ideal due to the fact that the average value is used as the output value for the subset of node samples when constructing a tree, and all training samples arriving at a certain leaf node are equally considered, which relies too heavily on data quality. This paper uses KNN algorithm to perform weighted averaging based on the contribution of neighboring points to the prediction results, and combines the advantages of SVR in processing high-dimensional data and small samples, and proposes SVR and KNN_GBRT fusion model. The improved fusion model has been validated in a housing rental datasets and has better prediction results compared to SVR model and GBRT model.

Keywords: machine learning; SVR model; GBRT model; KNN algorithm; rent prediction

1. INTRODUCTION

Over the past few years, more and more people have chosen the lifestyle of renting a house. In order to solve the problems exposed by the leasing market, such as unsystematic and standardized pricing, arbitrary price increases or decreases, and information inconsistency, machine learning algorithms are used to research and improve the housing rent prediction model.

Machine learning is one of the important directions in the field of artificial intelligence, which is divided into supervised learning, unsupervised learning, and reinforcement learning[1]. Currently, supervised learning can be subdivided into classification and regression problems, where regression is used to solve prediction problems. Common models include LR (Logical Regression Model), SVR (Support Vector Regression Machine Model), DT (Decision Tree Model), and so on. GBRT is an iterative decision tree algorithm that is widely used in machine learning[2]. Compared to other machine learning algorithms, it can be used for both classification and regression prediction, as well as for natural processing of mixed data types[3]. Despite this, the GBRT algorithm still has shortcomings. When constructing a tree, the average value is used as the output value for the subset of node samples, and all training samples

ICBBEM 2023, May 19-21, Hangzhou, People's Republic of China

DOI 10.4108/eai.19-5-2023.2334250

reaching a certain leaf node are equally considered. Due to excessive dependence on data quality, it is difficult to achieve ideal accuracy in model prediction.

Therefore, this paper uses KNN algorithm to perform weighted averaging based on the contribution of neighboring points to the prediction results, and combines the advantages of SVR in processing small samples of high-dimensional data, proposing SVR and KNN_GBRT fusion model. During model training, cross validation is used to avoid over fitting, and grid search is used to optimize model parameters, thereby reducing errors and improving prediction accuracy.

2. Principles of Common Machine Learning Models

2.1. SVR Model

For a given training sample, a regression function is trained by SVR, so that the deviation between the predicted value and the corresponding true value of each training sample based on the regression function does not exceed the error ε, at the same time, the regression function is required to be as smooth as possible[4],select loss function as in (1). The form of linear regression function as in (2). In order to make $f(x)$as smooth as possible, we need to find an x as small as possiblex', so the problem is described as an optimization problem $min\frac{1}{2}\left|\left|x'\right|\right|^2$, In addition, considering the possible error, two relaxation variables ξ_i and ξ_i^* are introduced, at this time, the optimization equation as in (3).

For support vector regression under nonlinear conditions, the idea is to first map the training sample nonlinear $\phi(x)$to a high-dimensional feature space H, in this high-dimensional feature space, the optimization problem is regressed to support vector regression under linear conditions. At the same time, the kernel function as in (4) is introduced to replace the dot product between samples in the input space to avoid the dimension disaster.

To sum up, SVR is the learning process of transforming the input space into a high-dimensional space through the nonlinear transformation defined by the inner product function, and solving the regression function in the high-dimensional space[5]. Its constructed hyperplane makes all data points as close to the hyperplane as possible.

$$L_\varepsilon(f(x),y)=\begin{cases}|f(x)-y|-\varepsilon, |f(x)-y|\geq\varepsilon\\ 0, |f(x)-y|<\varepsilon\end{cases}. \tag{1}$$

$$f(x)=(x'\cdot x)+b. \tag{2}$$

$$min\frac{1}{2}\left|\left|x'\right|\right|^2+C\frac{1}{n}\sum_{i=1}^{n}(\xi_i+\xi_i^*). \tag{3}$$

$$K(x_i\cdot x_i)=\phi(x_i)\cdot\phi(x_j). \tag{4}$$

2.2.GBRT Model

GBRT algorithm is a representative algorithm in Boosting algorithm. Boosting is a very excellent integration strategy[6]. It is a process of linear combination of multiple weak learners to form a strong learner. CART is used as a weak learner in GBRT. The gradient lifting tree model is formed after the optimization of the boosting tree model with gradient descent. Gradient lifting decision tree is an additive model. The forward distribution algorithm is used for greedy learning during training. Each iteration learns a CART tree to fit the residual between the prediction result of the previous tree and the real value of the training sample.

Assume that the strong learner obtained in the previous iteration is$f_{t-1}(x)$,The loss function as in (5),so the goal of this iteration is to find a weak learner $h_t(x)$ in the CART regression tree model, Minimize the loss function as in (6) of this round[7].That is, this iteration finds the decision tree to minimize the loss of samples.

The generation process of CART algorithm: in the input space where the training datasets is located, each region is recursively divided into two sub-regions and the output value of each sub-region is determined to build a binary tree.

Step 1: find out the optimal segmentation feature j and segmentation point s to ensure the minimum loss function;

Step 2: calculate the average error of all segmentation points of all features, select the smallest to split, and determine the output values corresponding to the left and right subtree regions and nodes after splitting;

Step 3: continue to call Step 1 and Step 2 for the two sub-regions until the stop conditions are met;

Step 4: divide the input space into M areas to generate the final CART tree[8].

$$L = \left(y_i, f_{t-1}(x)\right). \tag{5}$$

$$L\left(y_i, f_t(x)\right) = \left(y_i, f_{t-1}(x) + h_t(x)\right). \tag{6}$$

3. GBRT Model Improvement and Fusion

3.1.GBRT Model Improvement (KNN_GBRT)

Because the CART tree is used as a weak learner in GBRT, the gradient lifting regression tree algorithm divides the input data sample into multiple sub sample spaces when building a regression tree. Each sub sample space equally considers all training samples arriving at the node, and uses the sample average value as the prediction value for model regression. Such discrete fixed values greatly reduce the prediction accuracy of the decision regression tree. This article improves it based on the idea of KNN algorithm. KNN algorithm will perform weighted averaging based on the contribution of the nearest neighbor points to the prediction results, in order to fully utilize the information contained in the distance between data, thereby obtaining a relatively high accuracy result.

The process of constructing a gradient lifting regression tree: Given a datasets D as in (7), assume that the input space is divided into M regions, $R_1, R_2, \ldots, R_M$, calculate the distance between all training samples reaching a certain node and the prediction sample, find the K training samples closest to the prediction sample, record the output variable value of the K training samples as $(y'_1, y'_2, \ldots, y'_k)$, and calculate the distance $(d_1, d_2, \ldots, d_k)$ from the prediction sample. Using the Gaussian kernel function $\frac{1}{\sqrt{2\pi}} exp\left(-\frac{d^2}{2}\right)$, obtain the weight as in (8) of each training sample.

The predicted value as in (9)of the prediction sample at this node, which is the output value c_m of the leaf node region, is traversed through all input variables to find the optimal segmentation variable and segmentation point. Divide each region into two sub regions and determine the output value on each sub region until the stop condition is satisfied to build a regression tree.

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}. \tag{7}$$

$$w_i = \frac{e^{-\frac{d_i^2}{2}}}{e^{-\frac{d_1^2}{2}} + e^{-\frac{d_2^2}{2}} + \ldots + e^{-\frac{d_k^2}{2}}} (i = 1,2,\ldots,k). \tag{8}$$

$$\hat{y}_0 = WM, WM = \sum_{i=1}^{k} w_i y'_i. \tag{9}$$

3.2.SVR and KNN_GBRT Fusion

Based on the advantages of SVR in processing small samples of high-dimensional data, this paper proposes SVR and KNN_GBRT fusion model. The main idea is to divide the training samples into "large error" samples and "small error" samples based on the prediction error on the regression leaf node. The "large error" samples are trained using the SVR algorithm, while the "small error" samples are trained using the improved KNN_GBRT algorithm is trained using cross validation during model training to avoid over fitting. The KNN algorithm is used to calculate the distance between the predicted data and the divided two categories of data, and then the division of data categories and the prediction of the algorithm are obtained, resulting in the final prediction results. The algorithm block diagram is shown in Figure 1.

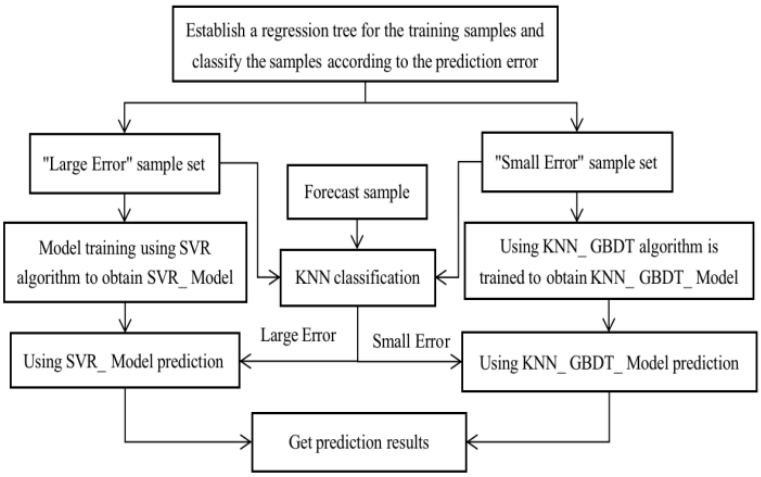

Figure 1. SVR and KNN_GBRT Fusion Model Algorithm Block Diagram

4. EXPERIMENT AND ANALYSIS

4.1.Data Source and Data Preprocessing

The housing rental information comes from the Tianchi Big Data Competition, which contains 41440 pieces of data and 30 characteristic variables. The initial data set contains a large amount of abnormal information such as missing values and noise, which is not conducive to the training of algorithm models. In order to improve data quality and facilitate subsequent model training, it is necessary to clean the data before use.

In order to better handle missing values, the missing rate is calculated for the feature information of the dataset, as shown in Figure 2. Delete and fill in the missing values of the data according to the statistical chart.

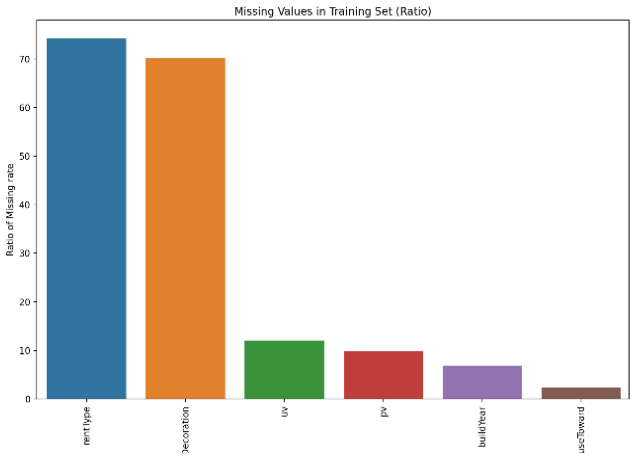

Figure 2. Statistics of Data Missing Rate

Delete the lease type and decoration style fields with a missing rate greater than 70%; The construction year is grouped by the name and price of the community. Find the construction year with the same name and similar price of the community, and fill in it with the median value; The KNN interpolation method is used to fill PV and UV. Delete records where the missing rate is less than 5% and the value of the house orientation is empty.

Use the box diagram to form a graphical description through the quartiles of the datasets, view the data distribution of housing area and rental amount, and visually and clearly identify abnormal values in the data.

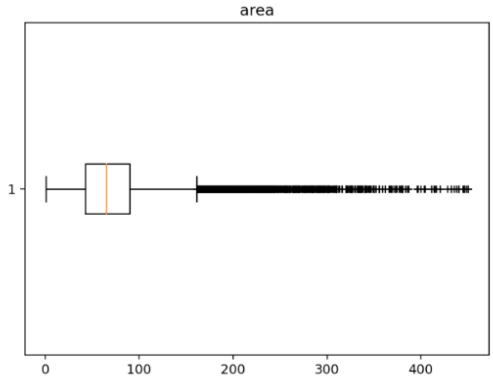

Figure 3. Box Diagram of House Area

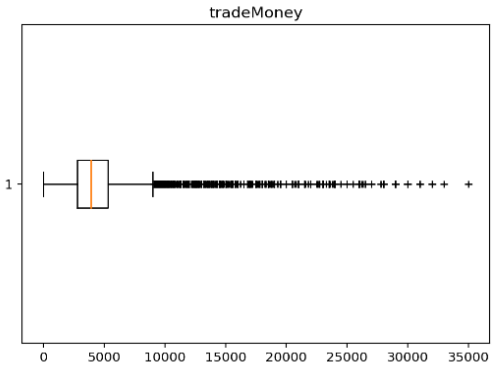

Figure 4. Box Diagram of Rental Amount

As shown in Figure 3, both data with a housing area of less than 15 square meters and data with a housing area of more than 280 square meters are regarded as outliers. The method of processing missing values is used to group data with a housing area of less than 15 square meters according to the community to which they belong, and fill in the corresponding housing area based on data with similar rental amounts. Delete all data with an area exceeding 280 square meters, and use the deleted data to draw a box graph of the rental amount. As shown in

Figure 4, data with an amount greater than 25000 can be considered as outliers and should be deleted.

Because the characteristic variables in the datasets include not only numerical types but also textual types. It is necessary to digitize text type data to facilitate subsequent model establishment. This article uses tag codes for attributes such as rental type and floor number of the house, as shown in Table I.

TABLE I. QUANTITATIVE CRITERIA FOR TEXTUAL VARIABLES

Attribute Name	Quantitative criteria
rent type	cotenancy-0,entire tenancy-1
house floor	low-0,middle-1,high-2
house toward	north-0,southeast-1,southwest-2,east-3,north-4,northeast-5,west-6,northwest-7
house decoration	roughcast-0,paperback-1,hardcover-2

After digitizing the features, logarithmize the rental amount, use the algorithm to perform KS test, and compare the QQ chart to find that the rental amount data meets the normal distribution. Min-max standardization is used to eliminate the dimensional impact between each indicator, so that the converted data falls within the range of [0,1], thereby eliminating the bias caused by different sizes of numerical variables on mining results.

The house ID and city in the data have no practical significance for predicting the house rent, so they will be deleted. Finally, through Pearson correlation coefficient, it is found that the correlation between UV and PV is as high as 90.4%, which may lead to multiple collinearity. Therefore, the characteristic variable UV is deleted. At the same time, delete the plate attribute that is less than 0.01 in importance to the characteristics of the house rent.

4.2.Model Construction and Parameter Optimization

The 32912 samples after data preprocessing were allocated using a 9:1 ratio. The training samples are divided into training sets and verification sets using a 5 fold cross validation. Building SVR models, GBRT models, and SVR and KNN_ GBRT using default parameters fused the model and used grid search to optimize the parameters of the three models. Finally, the model was compared using the model evaluation indicators MSE, RMSE, MAE, and R^2 to select the optimal model[9]. Where MSE is the mean square error, and the greater the error, the greater the MSE value; RMSE is a root mean square error that solves the problem of different output units in MSE by taking square roots, which can better describe data; MAE is the average absolute error, and the absolute value can effectively avoid the problem of error mutual cancellation, which can better reflect the prediction error; R^2 is the determination coefficient, using the mean value as the error benchmark, and observing whether the prediction error is greater than or less than the mean value benchmark error.

4.2.1 SVR Model

Establish an SVR model and optimize the C and gamma parameters in SVR using grid search and 5 fold cross validation. First, determine the value range of the two parameters, set the

search step, and establish a two-dimensional grid. The nodes in the grid are parameter pairs composed of two parameters. Then, perform cross validation on each parameter pair to calculate the error, and finally call the best_params_ of GridSearchCV obtaining the best combination of hyperparameters is the optimal parameter [10].

The optimal parameter combination obtained using grid search parameter adjustment is C 3000 and gamma 0.06. Comparing the error before and after parameter optimization, in the model evaluation indicators, MSE decreased by 6%, RMSE decreased by 3%, MAE decreased by 31, and R^2 increased from 0.629 to 0.652. It is proved that the error decreases after parameter optimization, and the prediction ability of the model is improved. The comparison diagram between the actual value and the predicted value is shown in Figure 5.

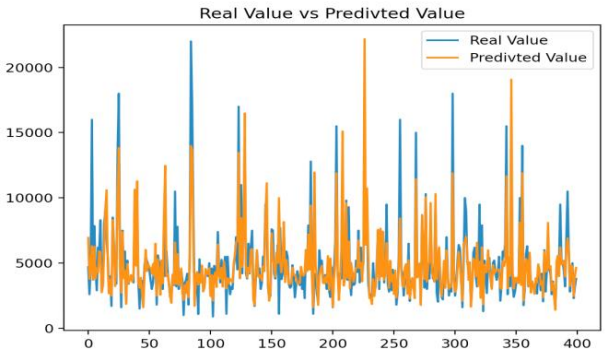

Figure 5. SVR Prediction Fitting Diagram

4.2.2. GBRT Model

After establishing the GBRT model, first adjust the number of iterations and learning rate within the liftingframework. Next, adjust the parameters of the regression tree, where, first adjust the max_depth, and then adjust the minimum sample number min_samples_split required for internal node division and the minimum sample number min_samples_leaf for leaf nodes together. See if the optimal value of both is on the boundary. If it is on the boundary, further change the parameter range and perform a grid search again.

Using grid search to obtain optimal parameters, n_estimators is 350, learning_rate is 0.06, max_depth is 7, min_samples_split is 10, min_samples_leaf is 2. Comparing the error before and after parameter optimization, the model evaluation indicators include a 5.7% decrease in MSE, a 2.9% decrease in RMSE, a 25% decrease in MAE, and an increase in R^2 from 0.652 to 0.672. It is proved that the error decreases after parameter optimization, and the prediction ability of the model is improved. The comparison diagram between the actual value and the predicted value is shown in Figure 6.

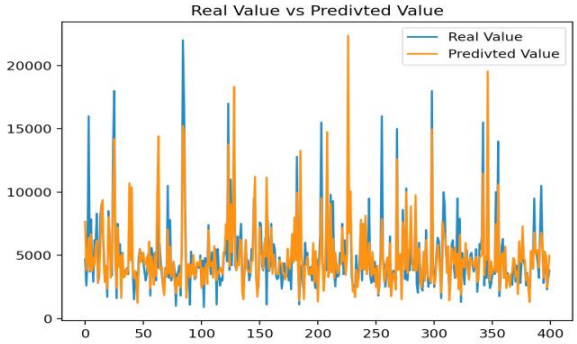

Figure 6. GBRT Prediction Fitting Diagram

4.2.3 SVR and KNN_GBRT Fusion Model

Using SVR algorithm and improved KNN_GBRT algorithm for training "large error" samples and "small error" samples, and using grid search for parameter optimization. The C in SVR is 150, and the gamma is 0.01. In the improved KNN_GBRT model, n_estimators is 150, learning_rate is 0.04, max_depth is 7, min_samples_split is 8, min_samples_leaf is 2. Comparing the error before and after parameter optimization, the model evaluation indicators include a 2.6% decrease in MSE, a 1.3% decrease in RMSE, a 13% decrease in MAE, and an increase in R^2 from 0.697 to 0.713. It is proved that the error decreases after parameter optimization, and the prediction ability of the model is improved. The comparison diagram between the actual value and the predicted value is shown in Figure 7.

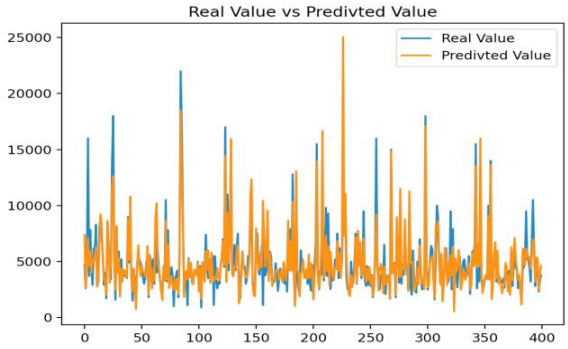

Figure 7. Fusion Model Prediction Fitting Diagram

4.3.Experimental Results and Analysis

The test set samples are respectively brought into the SVR model, GBRT model, and SVR and KNN_GBRT fusion model, after parameter optimization. Summarize and compare the MSE, RMSE, MAE, and R^2 of the three prediction models, as shown in Table II. It can be found that the SVR and KNN_GBRT fusion model has the smallest error, proving that the improved fusion model has higher prediction accuracy.

TABLE II. COMPARISON OF INDICATORS OF THREE PREDICTION MODELS

Model	Regression Model Indicators			
	MSE	***RMSE***	***MAE***	R^2
SVR Model	3008833	2833820	2243680	3008833
GBRT Model	1734	1683	1498	1734
SVR and KNN_GBRT Fusion Model	988	958	880	988

5. CONCLUSION

In order to reduce the error of GBRT in regression prediction, a SVR and KNN_GBRT fusion model is proposed using KNN algorithm based on the weighted average of the contribution of the nearest neighbor points to the prediction results, and combined with support vector machines. However, it takes a long time to optimize parameters using grid search during the model construction process. The next step is mainly focused on how to improve the calculation speed, and consider using genetic algorithms to optimize grid search to improve model performance.

REFERENCES

[1] Zhou Z H. Ensemble methods:foundations and algorithms[M].Chapman and Hall/CRC,2012:20-118.

[2] Leo G. Gradient boosting trees for auto insurance loss cost modeling and predictions[J]. Expert Systems with Applications,2012,39:3659-3667.

[3] Meng S. Research and Application of Personal Credit Evaluation Model Based on the Integration of GBDT and LR [D]. Beijing: Beijing University of Technology, 2018.

[4] Hu Z.Y. Research on support vector machine parameter optimization based on MAs algorithm [D]. Huazhong University of Science and Technology, 2011.

[5] Liu Y. X, Yang H, Su H.L, Liu Q, Chen T.X, Long C.Y. Quality evaluation model of 3D point cloud without reference based on v-SVR[J].Journal of Qingdao University,2021,34(04):30-39.

[6] Chen H, Deng F.M. Power electronic circuit fault diagnosis based on gradient lifting decision tree [J]. Measurement and control technology, 2017 (05): 14-17.

[7] Hong S, Li H, Duan X.C. A trend prediction method based on gradient boosting method[P]. China patent: CN108984893B. 2021.05.07

[8] Huan F, Jiang J, Yu C J, Xu H Y. Design of a two-stage octane number prediction algorithm based on fusion dimensionality reduction and ensemble learning [J].Natural Gas Chemical-C1 Chemistry and Chemical Engineering, 2022,47 (02): 95-102.

[9] Cheng Y Y, Chen R, Sun J Q. Technology Foresight Model Based on Multiple Function Fitting [J]. Intelligence Theory and Practice, 2021,44 (04): 185-188+168.

[10] Chen J H, Guo Y Y, Zheng Y, Lin Z Q, Sun C Y. Vascular elasticity detection based on characteristic parameters of photocapacitance product pulse wave [J] Journal of Electronic Measurement and Instrumentation, 2021,35 (03): 11-17.