ATLANTIS PRESS
Advances in Computer Science Research (ACSR), volume 83
8th International Conference on Mechatronics, Computer and Education Informationization (MCEI 2018)

Forecasting Second-hand Housing Price using Artificial Intelligence and Machine Learning Techniques

Tao Fu[1,a]

[1]School of Economics, Shanghai University, P.R. China

[a]18818262366@163.com

Keyword: Artificial Intelligence; Support Vector Machine; Neural Network

Abstract. In this era of information explosion, the development of science and technology changes with each passing day. Therefore, the automatic analysis of scientific and technological trends aims to help scientists extract useful information from a large number of academic conferences and scientific and technological documents, which is of great practical significance. Artificial Intelligence give us an useful techniques, this paper we use using artificial intelligence and machine learning techniques to fit and forecast house price, From the point of view of model, the prediction effect of support vector machine is the best, with strong stability, while neural network is the worst. As a contrast, linear regression and random forest have little difference between them.

Introduction

Artificial Intelligence is a new science of technology to study, develop and extend human intelligence theory, method, technology and application system. Artificial intelligence is the study of computer to simulate certain thinking processes and intelligent behaviors (such as learning, reasoning, thinking, planning, etc.). Enable the computer to achieve a higher level of application. Artificial intelligence will involve computer science, psychology, philosophy and linguistics.

Corinna Cortes and Vapnik (1995) proposed support Vector Machine, which can improve the generalization ability of learning machine and minimize the empirical risk and confidence range by seeking the minimum structural risk[1].It can also be extended to other machine learning problems such as function fitting. Nello Cristianini and John Shawe-Taylor can be applied to support vector machine[2].A random forest is a classifier containing multiple decision trees, and the output categories are determined by the modes of the classes outputted by individual trees.Leo Breiman and Adele Cutler(1995) developed an algorithm to infer random forests[3], which was derived from random decision proposed by Tin Kam Ho of Bell Labs in 1995[4]. It improves the prediction accuracy of the model by summing up a large number of classification trees[5].Instead of traditional machine learning such as neural networks A new model of the. Artificial Neural Network (Ann), a hot research hot spot in artificial intelligence field since 1980s, Rumelhart, Hinton, William's(1986) developed BP algorithm has developed the radial basis function (RBF) basis function neural network for the first time[6].It has strong nonlinear fitting ability, can map any complex nonlinear relation, and the learning rules are simple, easy to realize by computer, and have strong robustness and memory ability. Nonlinear mapping ability and powerful self-learning ability.

Methodologies

Support Vector Machine. Support Vector Machine is based on the development of statistical learning theory. It systematically studies some fundamental problems in pattern recognition in the case of limited samples. The support Vector Machine regression (SVR) algorithm needs to define a loss function, which can ignore the errors in a range of real values, just like the SVM classification algorithm. This kind of function is insensitive loss. Figure 1shows the one-dimensional linear regression function with insensitive region.

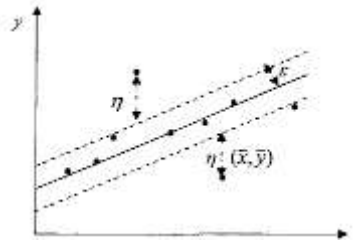


Figure 1 One-dimensional Linear regression function with insensitive region

The regression problem of support vector machine is similar to its classification problem, but the difference lies in the value of its output variable. The regression problem can be described in mathematical language as follows:

Set of given data samples $S = \{(x_i, y_i), \ldots, (x_m, y_m)\}$,where $x_i \in R^n$, $y_i \in R$, $i = 1,2,3,...,m$, looking for functions $f(x)$ on the R^n , $y = f(x)$.

The goal of regression is to find the following linear functions:

$$f(x) = \omega \cdot x + b \tag{1}$$

Where w is the vector of parameter column, $(w \cdot x)$ is the inner product, b is the threshold value. Support vector regression can be reduced to an optimization problem by minimizing the following functions:

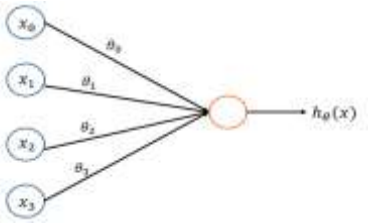
$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{m}(\xi_i + \xi_i^*) \tag{2}$$

$$\begin{aligned} s.t.\ & y_i - f(x_i) \le \xi_i^* + \varepsilon \\ & f(x_i) - y_i \le \xi_i + \varepsilon \qquad \xi_i^*, \xi_i \ge 0, i = 1,...,m \end{aligned} \tag{3}$$

Where c is the penalty coefficient, it is a compromise coefficient used to control model complexity, training error rate and generalization ability. ε is a default limit, ξ_i^* and ξ_i is a relaxation factor.

Artificial Neural Network. Multi-layer neural network is a feedforward neural network with one or more layers between the input and output layers. BP algorithm for multi-layer neural network includes forward propagation and back propagation. In the forward propagation process, the sample input value is transferred from the input layer through the hidden unit layer to the output layer, and then the error back propagation process is carried out, that is, the error is transferred to the input layer one by one.

Neural network is a nonlinear learning algorithm. The most basic component of neural network is neuron . So the basic model of neuron is given below:



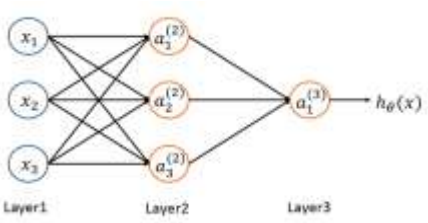


Figure 2 Basic model of neurons Figure 3 Basic neural network model

Where $\{x_1, x_2, x_3\}$ is input unit, x_0 is bias unit, $\{\theta_0, \theta_1, \theta_2, \theta_3\}$ is the connection weight, and $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$. The first layer1 is called input layer,layer2 called the hidden layer(the neural network may have more than one hidden layer, but in this case there is only one),layer3 is called the

output layer.α is called the output layer, and the excitation of the i unit is called the j layer, θ represents the weight from the j layer to the j+1 layer.

Random Forest. Random forest is an algorithm based on classification tree proposed by Breiman and Cutler in 2001.It improves the prediction accuracy of the model by summing up a large number of classification trees and is a new model instead of traditional machine learning methods such as neural networks. The calculation speed of random forest is very fast, and it is excellent when dealing with big data. The stochastic forest does not need to worry about the general regression analysis but the problem of multivariate linearity, does not have to make the variable choice. The existing random forest software package gives the importance of all variables. In addition, the random forest is convenient to calculate the non-line of variables. Sexual interaction, and can reflect the interaction between variables.

Empirical Research and Results

This paper takes second-hand housing in Shanghai as the research object, crawls 11805 samples from the platform, removes missing and abnormal value data and leaves individual property right housing and commercial housing data, a total of 7023 samples. It includes 31 variables, According to the research needs, the research variables covered in this paper are as follows: price is the unit price, subway is distance from nearest subway station, payment is down payment amount, area is house area, year is house age, landscape is green rate, fee is the property management fee, and some dummy variable, there are tax_2, tax_5, school_good, type_11, type_21, type_22, type_3, orientation, high floor, middle floor, comprehension, open or plain house, decoration, slab or tower house, private or commerce house, house has facility and regional clustering factor

In this paper, four models such as OLS, support vector machine, neural network and random forest are used to fit and predict the data. In the model estimation, a linear regression equation is constructed. At the same time, there is no over-fitting phenomenon in the random forest, and the importance measure of independent variables can also be output. The results are shown in the following table (Table 1).

Table 1 The results of the importance of the variables under random forest model

Variable	Inc MSE	Node Purity	Variable	Inc MSE	Node Purity
subway	12.194	5084.896	comprehension	9.137	46.992
payment	33.8	12778.904	structure_open	5.232	29.236
area	5.085	4039.182	structure_plain	1.34	61.321
year	6.307	1646.886	decoration	10.938	129.487
landscape	9.118	697.413	building_slab	10.185	167.642
fee	16.343	3893.911	building_concrete	12.368	132.581
tax_2	4.614	203.101	building_tower	3.865	56.448
tax_5	1.537	217.988	type_private	6.808	213.959
school_good	4.018	226.346	type_commercial	6.98	179.293
type_11	13.948	204.738	facility	0.539	146.242
type_21	12.838	245.445	cluster_1	15.02	2175.329
type_22	8.335	213.04	cluster_2	37.537	8725.161
type_32	11.633	230.073	cluster_3	3.721	1160.049
orientation	15.25	220.303	cluster_4	7.647	1172.437
floor_high	4.494	285.992			
floor_middle	4.018	329.33			

*** p<0.01, ** p<0.05, * p<0.1

The regression results show that the overall fitting effect is very good, and most of the variable coefficients have passed the significance test. Through the importance table of each variable under the stochastic forest model,In the rank of "Inc MSE" ,the variables are the second cluster factor, down payment amount and property fee in the area and the importance in the rank of "Inc Node Purity".In the few variables are the down payment amount, second-hand housing in the area of the second cluster factor to and distance from the nearest subway station.

Support vector machine, neural network model,random forest will be presented with OLS fitting and forecasting effect for RMSE(Root Mean Square Error),MAE(Mean Absolute Error) and MAPE(Mean absolute percentage error).

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(observed_t - predicted_t)^2 \tag{4}$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|(observed_t - predicted_t)\right| \tag{5}$$

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{(observed_t - predicted_t)}{observed_t}\right| \tag{6}$$

Table 2 RMSE,MAE and MAPE and Rank of machine learning Model

	RMSE	MAE	MAPE	RANK
OLS	1.119	0.713	0.189	3
Support vector machine	0.865	0.380	0.093	1
Neural network	6.255	5.600	0.989	4
Random forest	1.164	0.697	0.161	2

From the point of view of the fitting effect of the model, the result of neural network model is worse than that of the other three models on the value of RMSE, MAE and MAPE. Linear regression, support vector machine and random forest are the three models with better effect, among them, support vector machine has the best effect, and these three models have better fitting effect.

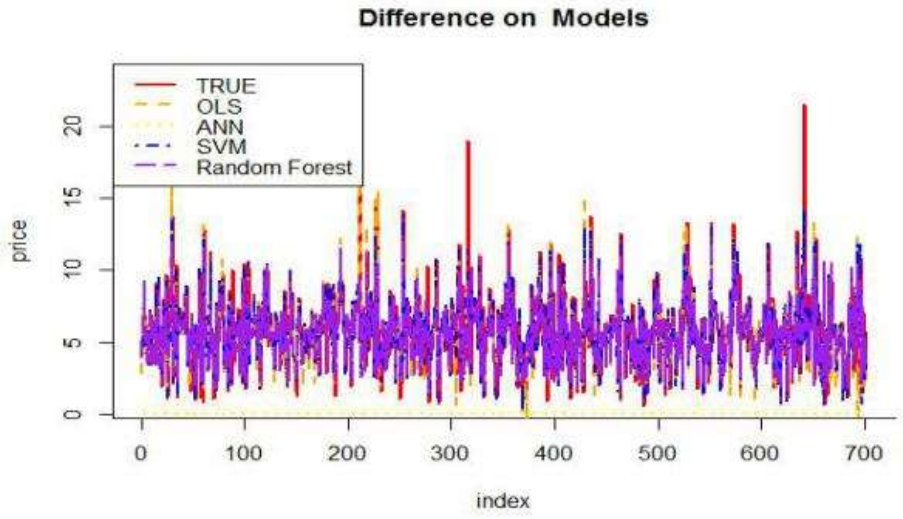


Figure 4 Fitting effect of several models

The fitting curve of SVM model is closer to the true value from the close degree of curve. Finally, each model is used to predict the data out of the sample period, and compared with the real value. This paper draws samples of the total sample to compare, calculates the average error, the results

are shown in Table 3.

Table 3 Comparison between predicted and true values of several models

id	OLS	Support vector machine	Neural network	Random forest	True value
6491	2.878927022	3.930551113	0.048921119	3.932338698	4.1176
3773	6.644618361	5.905528063	0.049145624	6.41629874	6.0714
3295	7.380855847	7.853376958	0.048921119	9.254368422	7.6471
2027	7.304843336	6.842796379	0.066164404	7.000604951	6.9231

Conclusions

Based on the characteristic price theory widely used in foreign countries, this paper constructs the characteristic price model of Shanghai second-hand house market by using the data of crawling search network, analyzes the influencing factors and leaves a part to forecast, and obtains the following conclusions: As far as the model is concerned, support vector machine has the best prediction effect and has strong stability. RMSE is 0.865, and MAE is 0.380, the MAPE is 0.093, and the neural network is the worst. These two models are black box models, but the difference is very big. As a contrast, linear regression and random forest have little difference between them. The characteristic is that linear regression is very sensitive to outliers, but random forest is not. But the whole can depict the main characteristic.

In this era of information explosion, the development of science and technology changes with each passing day. Therefore, the automatic analysis of scientific and technological trends aims to help scientists extract useful information from a large number of academic conferences and scientific and technological documents, which is of great practical significance. The good forecast performance of support vector machine is expected to be used as a reference for house price forecasting and to provide policy support for relevant departments.

References

[1] Cortes C, Vapnik V. Support Vector Networks. Machine Learning, 1995, 20: 273-297.
[2] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press,2000.
[3] Breiman, Leo . "Random Forests". Machine Learning. 45 (1): 5 – 32.
[4] Ho, Tin Kam. Random Decision Forests Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14 – 16 August 1995. pp. 278 – 282.
[5] Ho, Tin Kam. "The Random Subspace Method for Constructing Decision Forests". IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832 – 844.
[6] D. E. Rumelhart, G. E. Hinton and R. J. Williams. Learning representation by back-propagating errors. Nature, 1986, 323(6088): 533-536.