International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:12, No:2, 2018

Prediction on Housing Price Based on Deep Learning

Li Yu, Chenlu Jiao, Hongrun Xin, Yan Wang, Kaiyang Wang

Abstract**—**In order to study the impact of various factors on the housing price, we propose to build different prediction models based on deep learning to determine the existing data of the real estate in order to more accurately predict the housing price or its changing trend in the future. Considering that the factors which affect the housing price vary widely, the proposed prediction models include two categories. The first one is based on multiple characteristic factors of the real estate. We built Convolution Neural Network (CNN) prediction model and Long Short-Term Memory (LSTM) neural network prediction model based on deep learning, and logical regression model was implemented to make a comparison between these three models. Another prediction model is time series model. Based on deep learning, we proposed an LSTM-1 model purely regard to time series, then implementing and comparing the LSTM model and the Auto-Regressive and Moving Average (ARMA) model. In this paper, comprehensive study of the second-hand housing price in Beijing has been conducted from three aspects: crawling and analyzing, housing price predicting, and the result comparing. Ultimately the best model program was produced, which is of great significance to evaluation and prediction of the housing price in the real estate industry.

Keywords**—**Deep learning, convolutional neural network, LSTM, housing prediction.

I. Introduction

THE homebuyers are most concerned about the ups and downs of housing prices. However, during the long process of comparison of the real estate, many buyers might miss the opportunities to purchase due to the rise in house prices. In order to solve this problem, the accurate prediction information for future fluctuations of the housing price is required to be provided to the buyers. Nowadays, at home and abroad, the prediction of housing price is in the research stage, and there is no practical business application [1]. The earliest analysis of housing price mainly analyzed the trend of housing price at the macroeconomic level [2]. In addition, from the statistical point of view, the time series method, which can better reflect the overall trend of housing price, is the most widely used method of housing price forecasting. In recent years, with the development of the big data, the deep learning method has become the important prediction method [3]-[5], because it can predict the price of each house more accurately by analyzing the feature attributes of the house [6], [7]. However, at present, the deep learning is rarely applied to house price prediction, and few studies have compared the performance of the prediction based on the deep learning with the prediction based on the time series [8]-[10].

In this paper, from the perspective of big data analytics, the prediction is implemented by automatically crawling the real estate information and using the deep learning method based on this information [3]. Besides, we focus on the comparison between the prediction method based on time series and the prediction method based on deep learning in the feature attributes of the house. The contributions of this paper mainly include the following three aspects: (1) By using the a web crawler, the real estate information of a city in China, including the location, area, surrounding facilities, was captured to analyze the influencing factors of the housing price. (2) We proposed a two-dimensional LSTM prediction model framework based on time and housing attributes, including CNN model and ARMA model. And we focused on the designation of two kinds of LSTM prediction models, namely the LSTM-1 model purely based on time series, as well as the model LSTM-2 based on the characteristic factors of the real estate as well as the time series; (3) Also, based on the price information of a city in China, a detailed and empirical prediction of the housing price in this city was made. The experiments, in which the better prediction was presented, showed that the method based on LSTM can not only capture the feature attributes of the house, but also reflect the time series attributes very well.

The paper is organized as follows: in the next section, we introduce the process of obtaining the real estate data by using the crawler technology and analyzing the influencing factors of housing price. In Section III, a housing price prediction model is designed from the perspective of deep learning and time series respectively. Section IV contains the detailed experiments as well as the comparison between the methods based on deep learning and the methods based on time series, and finally conclude the paper in Section V.

II. Crawling of Real Estate Information and Analysis of Housing Price Influencing Factors

A. Access to Real Estate Data with the Use of Web Crawler Technology

A large number of data samples are required to build a prediction and analyzing model of the housing price, so it is proposed to use the web crawler tool to obtain the required real estate data. In this paper, we improved the original crawler and build an efficient and scalable web crawler for the real estate data. The optimization of the original theme crawler tool is to add the XPath control module of the crawler, and use the method of using XPath and HtmlAgilityPack components in combination. Then, it is easier to implement the XPath dynamic

Li Yu is with Renmin University of China, Beijing, 100872 China (corresponding author, phone: +8610-82500907; e-mail: buaayuli@ruc.edu.cn).

Chenlu Jiao, Yan Wang, and Kaiyang Wang are with Renmin University of China, Beijing, 100872 China (e-mail: 379055151@qq.com, wangyan@ruc.edu.cn, wky_sky86@163.com).

Hongrun Xin is with Beijing University of Posts and Telecommunications, 1008762 China (e-mail: xinhongrun@bupt.edu.cn).

configuration to extract the data from the HTML tags, so that the crawler can better deal with the iterative updating problem of crawling objects.

We took fang.com as the main crawling object, and to mainly crawl three aspects of data in Beijing, namely the second-hand housing source data, real estate data and the trend of the housing price. Finally, we successfully obtained nearly 200,000 data, including approximately 13,000 data that were related to the relevant feature attributes of the house.

B. Analysis of Influencing Factors of Housing Price

After crawling the real estate information, data analysis should be conducted on the housing stock data, the real estate data and the trend of housing prices as well. The purpose is to extract the relevant characteristic factors that affect the housing price and to determine the data source of the housing price prediction model. Additionally, it is required to filter and remove noise information on the Beijing second-hand housing price data that have been obtained already.

Through data analysis, we summed up four main distinctive factors as the characteristic points of the housing price prediction model in the next stage. The four factors are: the building age, the number of subways around the building, the number of schools around the building and the location of the building. Moreover, we mainly present the relationship between the characteristic factors and the housing price in such graphic charts.

As can be clearly seen from Fig. 1, the average house price is gradually decreasing with the increase of building age. The more subways around the building, or the more schools around them, the higher the housing price will be. Furthermore, the average housing price is higher, where there are a relatively small number of real estates. Accordingly, these four characteristic factors, which directly affect the housing price, are ultimately determined to be the main characteristic factors of the housing price prediction model.

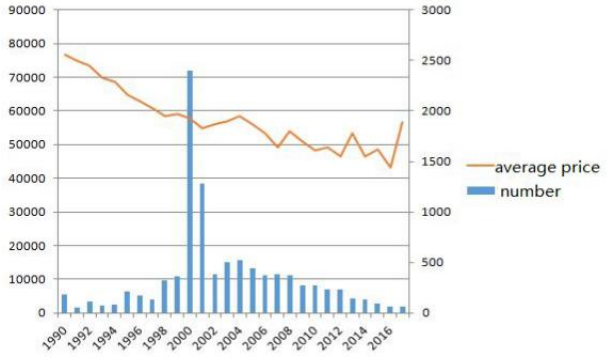

Fig. 1 Relationship between building age and housing price

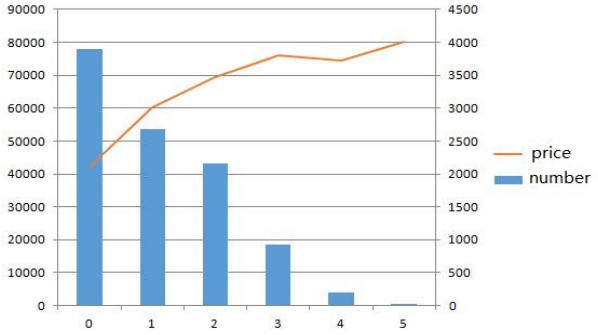

Fig. 2 Relationship between subway count and housing price

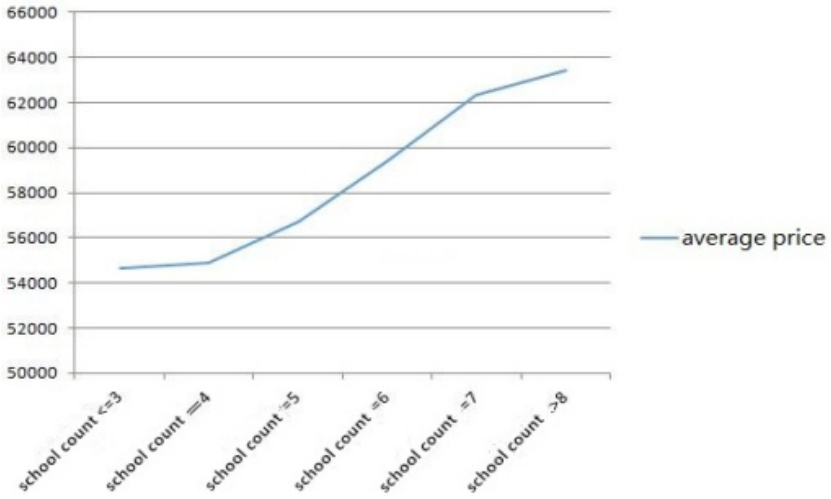

Fig. 3 Relationship between school count and housing price

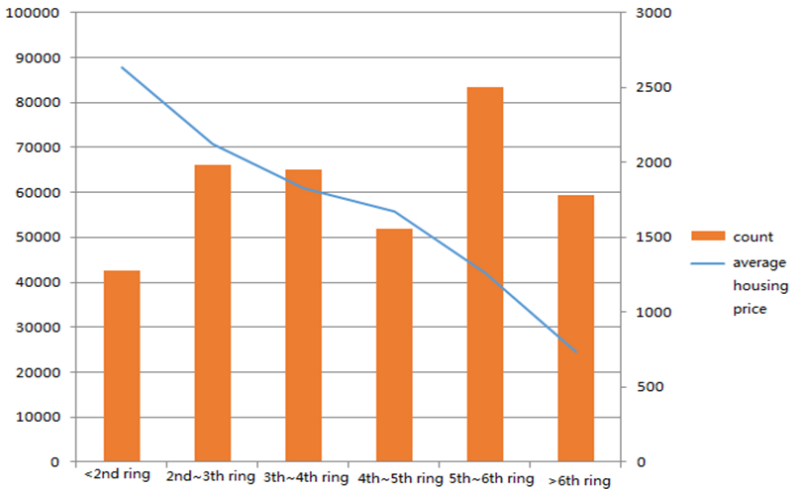

Fig. 4 Relationship between location and housing price

III. Housing Price Prediction Model Based on Deep Learning and Time Series

A. Prediction Model Based on Deep Learning of Characteristic Features of Housing Price

1. Housing Price Prediction Base on CNN

In fact, a real estate not only has four characteristic factors summarized in the previous section, but also has other characteristics, such as house age, the number of surrounding public transportation, etc. However, when there are too many characteristic factors, the predicting outcomes of the simple logistic regression model will be inaccurate, and there is not a well-established model to provide theoretical support for fixing the price of a house. What is even worse is that there is not a very scientific and reasonable basis to help sellers make scientific adjustments to a house price according to the market information. We found that CNN can well combine the processing and storage of the second-hand house information, and it can adaptively learn the sample data intelligently with the use of the relevant feature data and mathematical algorithm. In addition, CNN can deeply grasp the features of objects, so it is very suitable to solve the above problems.

Model Implementation - In this paper, the specific way to apply CNN to the housing price prediction model is as follows: First, the scientific computing library provided in Python is used to process the data of real estate. The next step is to train the different characteristic factors of real estate, for example, the building age, the surrounding facilities and the XY coordinates, etc. And finally, the function of predicting housing price is implemented according to the different characteristic factors of the real estate.

In terms of model structure, the housing price prediction model based on CNN is composed of the input layer, the convolutional layer, the activation function, the pooling layer and the full-connect layer.

Input Layer and Output Layer - To implement the input layer and the output layer of the convolutional neural network, first, the original data of existing real estate needs to be analyzed. According to the analysis, we know that the characteristics of a real estate are composed of various data characteristics such as the building age, the number of schools, supermarkets or banks within 2,000 meters around the location and so on. These data characteristics will affect the market, and thereby, the housing price. For example, the price of houses located in the center of the city will be higher, and houses will be more attractive with a greater number of surrounding subways and public transport. After deleting the illegal data, there were 9,600 sample data left. Also, the information about the surrounding facilities needs to be quantified according to the following examples, the information about the surrounding elementary schools "Haidian Primary School, Liuyi Kindergarten" is converted into a quantity of two. The specific data source information of the input layer is presented in Table I.

TABLE I
Data Source Information of Input Layer

Field name	Field data	Remarks
housing code	1010088440	
name	Run Qian Qiu Jia Yuan	
age	2006	
schools count	12	Peking University, 61 Kindergarton, Haidian primary school…
facilities count	2	The Summer Palace, Baiwangshan Forest Park
...	...	...
bus count	17	No. 328 bus…
subway count	1	Ma Lian Guisubway
conversed location	5	fifth to sixth ring road
X	116.269981...	
Y	40.0336799...	
housing price	84533	unit: yuan/m^2

The surrounding information of the real estate in Table I was mainly obtained through the API of Baidu, and the extraction range was: facilities within 2000 m. After observing and analyzing the data information in Table I, 13 kinds of data

characteristics of the real estate were obtained, but the problem was that it could not form a 4×4 matrix. Therefore, we replicated the three existing data characteristics, and changed the number of characteristic of the real estate to 16, and thus, the data characteristics could be split into a 4×4 matrix in the end.

To better calculate the matrix in the convolutional neural network, we use the data normalization method to process the data so that the variance of the characteristic data is 1, and the mean value is 0 after normalization, and this mean and variance are calculated. Afterwards, on the training set, the normalization of data is processed by calling the *fit_transform()* algorithm. The formula is as:

$$x' = \frac{x - \mu}{\sigma^2} \tag{1}$$

In this paper, the train_test_split function in the Sklearn scientific computation package, which is a commonly used function in cross validation, is used to divide the normalized data sets into a training set and a test set.

By the above methods, the data can be randomly selected, transformed and normalized, and finally, the input layer of the CNN housing price prediction model can be determined. And the output layer in this network is the normalized housing price that is obtained after a series of calculations.

Convolutional Layer - The convolutional layer is the core part of the housing price prediction model of convolutional neural network. The housing price prediction model mainly uses two layers of convolutional layer to calculate, so as to predict the housing price. What are the main source of the convolution layer are the input layer and the pooling layer.

The CNN predicts the housing price through a series of transformations and analyses to the input layer. First, it transforms the input layer into a 4×4 matrix that is served as the data source for the first layer of the convolutional layer. The input of the first convolutional layer is set to: length 2, width 2, height 1 and the output thickness of 32. A 2 × 2 × 32 matrix is output by the convolutional matrix operation and moving step is set to 1 in the first convolution operation.

The matrix data obtained after the operation in the first convolutional layer is used as the data source of the second convolutional layer. Similarly, the input of the second convolutional layer is set to a height of 32, the output thickness of 64, and a moving step length of 1, and then a 4 × 4 × 64 matrix data is output after the convolutional matrix operation.

Activation Function- The housing price prediction model based on CNN uses ReLU activation function, and its mathematical formula is:

$$f(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{2}$$

In this activation function, if the input is negative, the output is all 0; else if it is positive, it will be output as it used to be. Compared with the Sigmoid activation function, the ReLU is different in that: the Unilateral Inhibition and the Sparse Activation. The ReLU activation function takes 0 as the output of the function, resulting in the sparsity of the network, thus reducing the degree of the parameter interdependence and the over-fitting phenomenon.

As TensorFlow provides the encapsulation of the Relu activation function, we use the TensorFlow AI (Artificial Intelligence) scientific computing package in Python to write the program.

Pooling Layer - The input source of the pooling layer of the housing price prediction model based on CNN is the output of the second convolutional layer. The purpose is to introduce invariance, reduce the redundancy of parameter and avoid over-fitting problem. Moreover, the pool layer has no parameter influence in back propagation, and therefore, the weight will not be updated. In fact, the pooling layer is equivalent to again aggregating the characteristic matrix data obtained after the average pooling or the maximum pooling of the convolutional layer.

Full-Connect Layer - As a "classifier" in the computation of the convolutional neural network, the full-connect layer aims to map the trained sample characteristic parameter that have been learned and trained to the labeled sample space. When predicting, only in the fully-connect layer can the multidimensional matrix produced by the previous layer be converted into the convolution with a 1×1 convolution kernel. Then, the extracted characteristics are classified into modules by using the full-connect neural network model, and the classification results are obtained through the calculation of the full-connect layer in the end.

2. Housing Price Prediction Base on LSTM-2 Model

The LSTM neural network is one kind of the recurrent neural networks. However, when we use the RNN to achieve a long-term memory, it is required to relate the implicit calculation of the current state to the first n times calculation. And that will result in an exponential increase in computational complexity, leading to the problem of vanishing gradient problem or gradient explosion. LSTM is known as the Long short-term memory, which is a special model belonging to RNN (recurrent neural network), and it is suitable for dealing with and predicting the important tasks with longer interval and delay in time series. The difference between LSTM and RNN is that LSTM introduces the concept of gate, and it can use the “forget gate” to solve well the problem in RNN. In the algorithm of LSTM, a "processor" is added to determine whether information is useful. Furthermore, the LSTM hidden layer is composed of multiple recurrently connected subnets which are referred to as memory blocks, and each memory block consists of self-connected memory cells and three multiplicative gate units (input, output, and forget gates).

a) Model Implementation

Input Layer and Output Layer - According to the 13 characteristic factors of a real estate, the convolutional neural network can predict the second-hand housing price in Beijing. In addition, in order to further study whether the time factor has an impact on the price of second-hand houses, we introduce the

LSTM model and make a prediction again. Without changing the 13 characteristic factors, we add the time series into the model. The first step to predict in this model is to construct the input layer, whose structure is similar to CNN; depending on the monthly housing price trend information of a real estate extracted captured data from the captured data.

To simplify the program, we merge the 13 characteristic factors. Assuming that regardless of the influence of the weight factor, we add the number of surrounding schools, the number of surrounding subways, the number of surrounding buses and other factors to form main factor called the surrounding information summary. And the data source is then transformed into the surrounding information, the location and the housing price, and they are used as the data source for the input layer of the prediction model. Finally, the prediction of the second-hand housing price based on the 13 factors and the time factor is achieved in the output layer, and the error is output as well.

Input Gate, Output Gate and Forget Gate - Using the algorithm to implement the structure of the input gate is:

$$i_t = g(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (3)$$

where h denotes the output vector from the upper layer, W denotes the weight matrix between the corresponding index and b stands for the bias. Similarly, we use the same formula to implement the structure of the output gate o_t and the forget gate f_t

Training Set and Method Determination - After building the input layer, output layer, input gate, output gate and forget gate of LSTM, the training set and test set of the housing price prediction model based on LSTM also need to be defined.

By analyzing the data set, 90% of the data set is determined to be the training set and the remaining 10% is the test set. However, during the training and testing of LSTM model, not all parameters are reasonable, so training error will inevitably occur. To reduce this error, we use the inverse method to predict. The inversion method deduces different parameters and known conditions according to the known rules or the hypothetical rules, and then it compares the different errors produced by different models to adjust the operational parameters affecting the error in the model and selects the model with the smallest error. After adjusting the parameters, the Relu function is set to be the activation function. The application flow chart of the inversion method is as follows:

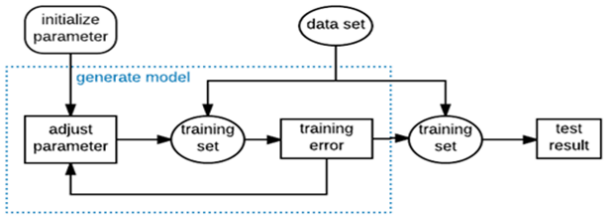

Fig. 5 Application flow chart of Inverse Method

3. Housing Price Prediction Base on Logistic Regression

Nowadays, according to individual demand, homebuyers often compare and analyze several properties they focus on. In this process, they are more concerned about whether the price of the house they choose is up or down; so, the majority of them would lose the opportunity to purchase due to the rise in the housing price during the long-term comparison of houses. In order to solve this problem, we propose the use of logistic regression in the housing price prediction model.

Model Implementation - When building this prediction model, we first used relevant characteristic factors of the housing prices to set up the characteristic data, such as the building age, the real estate location, the number of surrounding facilities and so on. Then we used the dummy variables to quantify some variables that could not be quantitatively dealt with, for example, the change trend of housing prices, location and so on. We programmed the model in Python, and the related algorithm packages are Numpy, Pandas, Statsmodels and Pylab.

The specific calculation and implementation of the algorithm, as well as the prediction results of this model have been described in Section VI.

B. Housing Price Prediction Model Based on Time Series

1. Prediction Model Based on ARMA

The full name of the ARMA model is the autoregressive moving-average model, which is based on the autoregressive model and the moving average model, and is mostly used to study the time series problem. The formula is:

$$Z_t = \varphi_1 Z_{t-1} + \varphi_2 Z_{t-2} + \ldots + \varphi_1 Z_{t-1} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \quad (4)$$

We proposed to use the logarithmic transform method to deal with the data source stability. The purpose is to reduce the vibration amplitude of the data, so that the linear law is more obvious and the time series of the second-hand housing prices trend in Beijing can harden. In addition, we took advantage of the differencing techniques to eliminate the effect the periodic factors brought, and then implemented linear reduction of the data with the same periodic interval.

After the twelfth order difference and first order difference of the data source, the results are presented in Fig. 6.

Test Statistic	-1.945271
p-value	0.311073
#Lags Used	12.000000
Number of Observations Used	66.000000
Critical Value (5%)	-2.906444
Critical Value (10%)	-2.590724
Critical Value (1%)	-3.533560

Fig. 6 Difference statistics data

In this model, the transaction needs predicting by sample fitting. It should be noted that the data fitted in ARMA must have been pre-processed already, so the predicted values are required to be restored through relevant inverse transformation.

After fitting the sample the error is calculated to be 3.104%, and the prediction result is obtained. The comparison between

the prediction data and the actual data is shown in Fig. 7. The actual trend is marked in red while and the prediction trend is the blue one.

2. LSTM-1 Prediction Model Purely Based on Time Series

Considering that, in real life, the impact of housing price trend is not just limited to its characteristic factors. What are closely related as well are the housing price and the time. And after analyzing of data of second-hand housing price in Beijing prices that we crawled, we found that that the price has risen linearly with time since 2010, the results are shown below. Therefore, in order to predict the housing price in a future time in Beijing, in this paper, one time factor is set to be the factor affecting the price trend, and use the LSTM neural network model purely based on the time series to predict second-hand housing prices in Beijing.

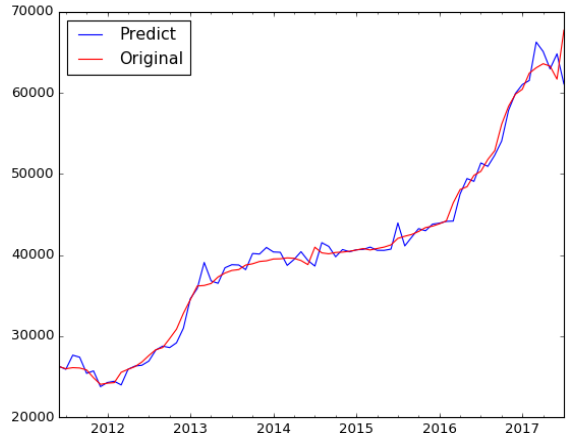

Fig. 7 Prediction result of ARMA

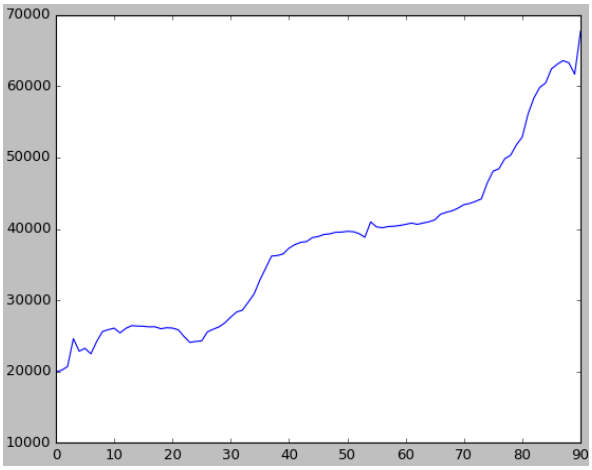

Fig. 8 The trend of second-hand housing price in Beijing

The LSTM-1 prediction model uses the LSTM neural network to firstly learn the average price of second-hand housing in Beijing in the past period, and then to predict that average price in the future based on the time characteristics. It mainly aims at the characteristic data of house price changing with time. Once again, the house price information is trained, tested and validated to calculate the house price forecast trend in the coming months. According to the change value in housing price over time, it again trains, tests, verifies and finally calculates the prediction housing price trend in coming months.

When building the LSTM-1 model, we first build the data source of LSTM. Specifically speaking, we take the time series from different cities (the city is Beijing in this paper) as the main source data to build the housing price prediction model. Taking this data source as the first input layer of LSTM-1, and then enter the input gate, the data needs to be processed a function conversion in the input gate. Finally, we use the inverse method to adjust the parameters of the LSTM functions, which is equivalent to adding the time series as one of the characteristic factors to the original LSTM.

IV. Experiments and Analysis

In this paper, by investigation and observation to crosswise compares, we determined to take Fang.com to be the crawling object of the real estate data network crawler. Then based on the filtering rules and data arrangement method, the crawled data was consolidated and reorganized into the data source information which is necessary in analyzing the real estate data and predicting the housing price. Afterwards through the comparative experiments, we analyzed the second-hand housing real estate data in Beijing and predicted the price. In addition, we used quantitative qualitative analysis method to predict the second-hand housing prices in Beijing, and used the inverse method to adjust the parameters, and ultimately determined the experimental results.

A. Prediction Based on Learning Characteristic Attributes of Housing Price

1. Logistic Regression Prediction

The second-hand housing price data source in Beijing was sorted into CSV file, which then was read into the Python data source template. In order to better look at the distribution of each characteristic factor, we designed a bar graph with Pylab and the graph was shown below:

The results of the logistic regression prediction model are: 850 samples were hit among the total 1000 samples, so the accuracy was 85% and the error was 24.7458814165. Parts of the prediction data were presented in Table II. It was noted that if the prediction value is less than 0.5 means fall while more than 0.5 means rise. In Table II, Flag denotes the fluctuation situation of the housing price, Local presents the location of the property, CreateTime is the construction age, SubWayCount and SchoolCount, respectively, denotes the number of surrounding subway and schools.

We can see from the experiment, using logistic regression model to classify part of the characteristic factors of the real estate can finally implement the housing price prediction model more accurately. The main advantage of this model is that the algorithm is fast in training, accurate in prediction and can predict the accuracy because the results are probability values.

However, the disadvantage is that the data with a large number of characteristic factors cannot be accurately predicted, and the results are under fitting because the algorithm may be affected by the samples. Moreover, due to the Sigmoid function, there is no discrimination on the impact of the target probability, or the threshold cannot be determined.

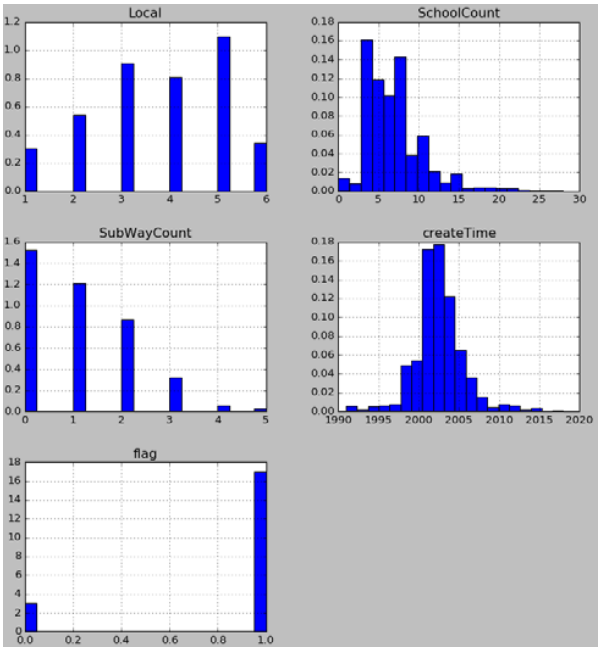

Fig. 9 Distribution of characteristic factors

2. Prediction Result Analysis Based on CNN Convolutional Neural Network

After building the whole experimental model based on the CNN, the sample training times was set to 2000, and the error was output in the end. Parts of the results are as follows:

TABLE II
PREDICTION RESULT OF HOUSING PRICE FLUCTUATION BASED ON LOGICAL REGRESSION

Flag	Local	CreateTime	SubWay	School	Intercept	Predict
0	2	2001	3	4	1.0	0.8584
1	3	2000	2	7	1.0	0.8580
1	2	2000	2	9	1.0	0.8514
1	1	1996	1	5	1.0	0.6683
1	1	1999	2	5	1.0	0.7734

TABLE III
ANALYSIS INFORMATION OF OPERATIONAL PROCESS

Count	Error
0	117.725
1	34.0834
2	8.82918
…	…
1997	0.260717
1998	0.260676
1999	0.260435

As can be seen from the results in Table III, with the increasing of training times, the error of prediction housing price decreased and converged to zero. In order to view the prediction results more intuitively, through the Python data visualization technology, output part of the data prediction graph as shown below, the real solid line represents the real data, CNN dot line represents the prediction data: to display the prediction results more intuitively, part of the prediction graphs were produced with the use of the Python data visualization technology. As shown below, the solid line represents the real data and the dot line represents the predicted data:

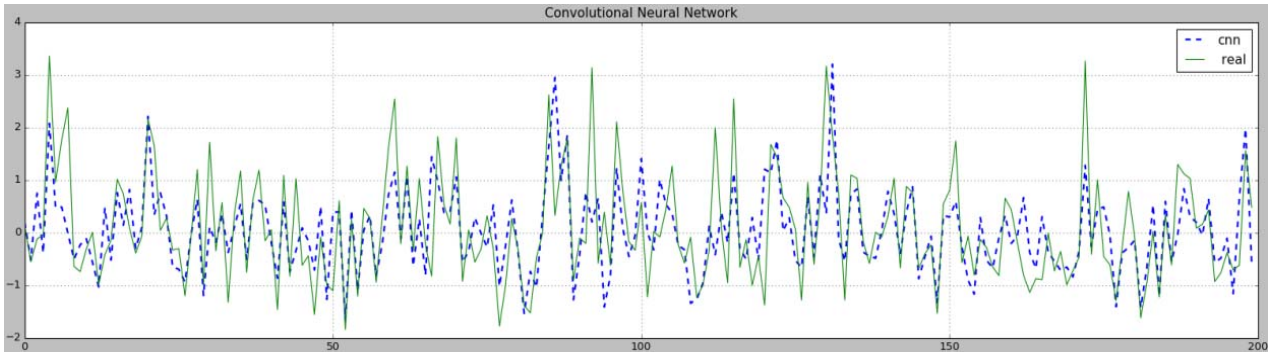

Fig. 10 Comparison of predicted results and actual results

It can be seen from the graph that the prediction housing price was similar to the r eal housing price, but there were also some discrepancies. In terms of the error, it decreased with the increase of training times, so the training times of 2000 can also be increased to increase the accuracy of the model. Finally, we found that after training 2000 times that the CNN model could obtain a relatively accurate prediction of the housing price with a final error of 0.260435. We can achieve the accurate prediction of second-hand house price, with a final error of 0.260435. However, the error of 5000 training times was reduced to 0.089548, which means that it is to increase the training times that can help reduce the error.

Compared with the logistic regression model, the CNN-based model can share the convolution kernel very well, so that its 13 data features can be computed in high dimensional without any pressure. Also, CNN applied a part of the

operational principle of the logistic regression to classify and select the characteristic data, which achieved the multi-dimensional regression prediction, so it can accurately predict the price of a real estate with different characteristic factors. In addition, the results will be more accurate with the increase in the number of iterations.

3. Prediction Result Analysis Based on LSTM-2

Similarly, we first built the whole LSTM-2 prediction model, and set the sample training times to be 2000, and finally, output the error. Part of the error data and the comparison between the prediction price and the real price are shown in Fig. 11. However, we calculated the specific error that was up to 10.35%, which means that the LSTM prediction model is not ideal for the 13 characteristic factors and the time characteristic factors. This is mainly because the 13 characteristic factors do not belong to those factors that will change over time. But, the fundamental reason is that the characteristic factors that can change over time are too few among the characteristic factors obtained by the data analysis, such as monthly prices, monthly sales, and monthly volume. So, the error of this model trained 200 times is a bit big.

```
Epoch 110/200
 1/80 [..............................] - ETA: 2s - loss: 0.2385
 3/80 [>.............................] - ETA: 2s - loss: 0.1304
 5/80 [>.............................] - ETA: 2s - loss: 0.1221
 7/80 [=>............................] - ETA: 2s - loss: 0.1071
 9/80 [==>...........................] - ETA: 2s - loss: 0.1339
```

Fig. 11 Specific data of the operational process

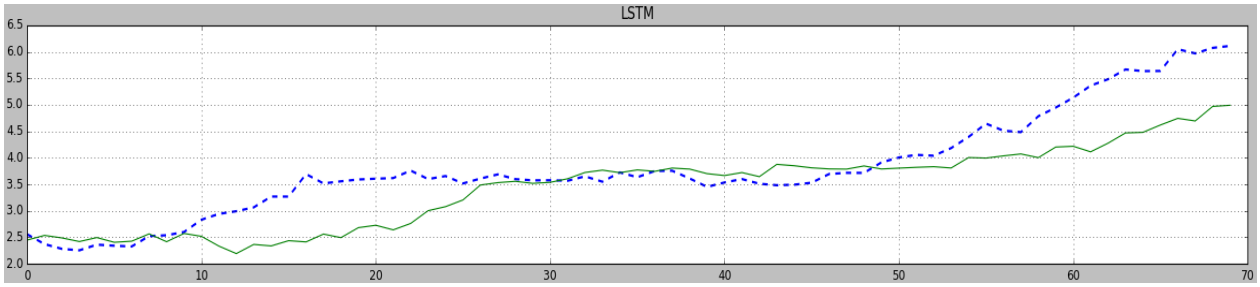

Fig. 12 Comparison of predicted results and actual results

B. Prediction Based on Time Series

In order to better analyze the accuracy of the model based on LSTM time series, we take the ARMA model as a baseline and compare it with the LSTM model. The building process and prediction results of the ARMA model have been introduced in the last section; in the following, we will introduce the experimental results of LSTM-2, as well as a comparison results with the baseline.

1. Experimental Results of LSTM-2 Model and Comparison

In this paper, some data from January 2010 to July 2017 were set as the training set. The total number of samples was 91, of which, the first 80% were the training set and the latter 20% were the testing set. We used a Keras model to invoke the TensorFlow artificial intelligence algorithm library, and finally, obtained the prediction prices of the past five months of second-hand housing in Beijing.

In this paper, the group-experiment method was used to divide experiment into three groups according to different training times of namely 200, 400 and 500. The three charts following show the prediction results.

The blue solid line represents the real trend of the housing price, the "X" represents the training set, the "origin" represents the verification set, and the green solid line in the last section represents the prediction set.

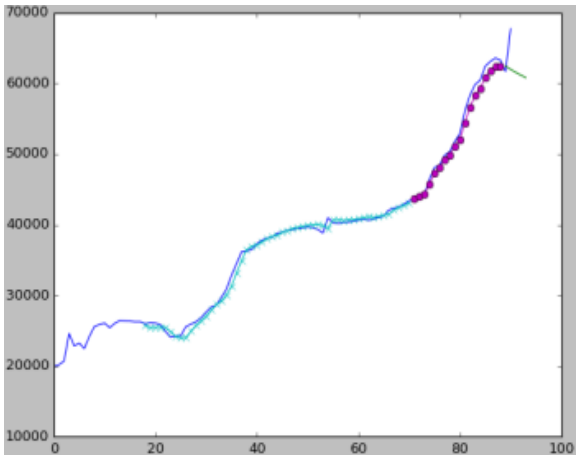

Fig. 13 Result of training 200 times

Then, we compared the three groups of results to summarize the prediction of this model. We can see from the results that the final error is 2%, meaning that the prediction error of the LSTM model is smaller than that of the ARMA model. In addition, this model can also predict the housing price for the second-hand housing price information with a certain regularity. As seen from the prediction results of the second-hand housing price in Beijing, the model can predict the approximate second-hand housing price information and trends in the next five months. In addition, through the comparison of different training times, it can be seen that more than 500 times of

training can get more accurate prediction results. The information of comparison between ARMA and LSTM is shown in Table IV.

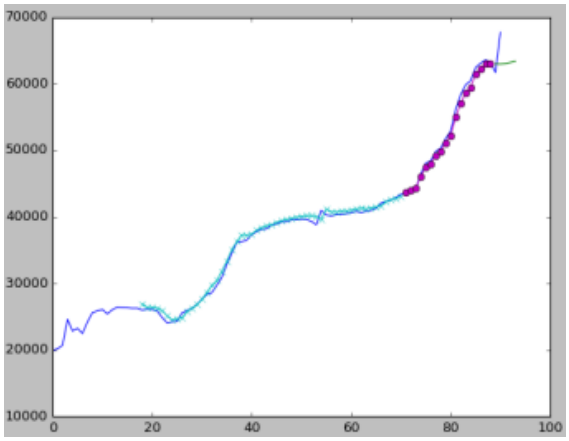

Fig. 14 Result of training 400 times

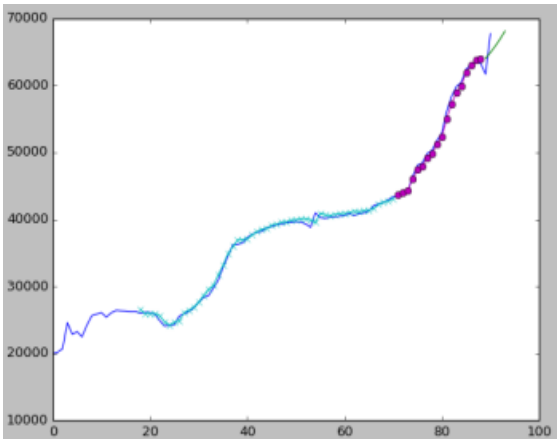

Fig. 15 Result of training 500 times

TABLE IV
COMPARISON INFORMATION OF ARMA AND LSTM

Model	Error	Application Scenarios and Factors	Parameters
logistic regression	24.74%	4 kinds of characteristic factors	sigmoid function
CNN	8.95%	13 kinds of characteristic factors	5000 iterations ReLU function
LSTM-2	10.35%	13 kinds of characteristic factors+time characteristic factors	2000 iterations tanh hyperbolic function
ARMA	3.104%	time characteristic factors	
LSTM-1	2%	time characteristic factors	500 iterations tanh hyperbolic function

From Table IV, we know that the prediction model based on LSTM is more suitable for the prediction with time factor. And compared with ARMA, its error range is reduced and the result is more accurate. However, the prediction result of ARMA differs greatly from the actual price, as it only highlights time series and ignores the influence of external factors. Therefore, when the sample or external factors are changed, errors often occur. On the contrary, the prediction result of LSTM is in good agreement with the actual price trend, which can be used as a prediction reference In addition, the housing price prediction model based on LSTM-2 is more suitable for some real estates with the price fluctuating regularly, such as housing price did not increase or decline significantly. However, it should be noted that prediction of the recent trend of housing prices based on the LSTM-2 model will be more accurate than the prediction of long-term trend. Compared to the LSTM-1 model, the LSTM-2 model considers the time series and is therefore more suitable for predicting data with time series, making it easier to predict the change of housing prices for a city in the future.

C. Comparison of Five Kinds of Housing Price Prediction Model

In order to compare the five types of housing price prediction model, we used the formula:

$$| \textit{prediction value} - \textit{real value} |/\textit{real value} \quad (5)$$

to calculate the relative error, and to reflect the accuracy of the each model. The errors of each prediction model are presented as follows:

TABLE V
COMPREHENSIVE ANALYSIS AND COMPARISON INFORMATION

	ARMA Prediction Model	LSTM Prediction Model
Theory Origin	Time Series	RNN Recurrent Neural Network
Applied Functions		*tanh* hyperbolic function
Influencing Factors of Prediction Model	Cyclical changes in data	training times and. hyperbolic function
Error	3.104%	2%
Whether the error is controllable	Uncontrollable	Controllable, and related to the number of trainings

From the error chart, we can see that the prediction of housing price trend purely based on time factor is more accurate than the prediction of housing price based on multiple characteristic factors. The main reason is that the monthly house price information is a very core characteristic factor for the housing price prediction model. However, in the prediction under multiple characteristic factors, the crawler does not crawl such core factors like monthly sales volume and monthly trading volume. Additionally, the macroeconomic factors and the related policies also have an obvious impact on housing price, so it is hard to reduce the error in prediction under multiple characteristic factors when the characteristic factors are not obvious. However, the prediction model based on multiple characteristic factors can predict the specific price of a real estate according to different characteristic factors, while the model based on time factor can only predict the trend of housing price for a city. We can also see from Fig. 16 that CNN performed the most outstandingly among the housing price prediction models based on multiple characteristic factors. And among the housing price trend prediction models, the LSTM-1 model, which is purely based on time series, has the smallest error.

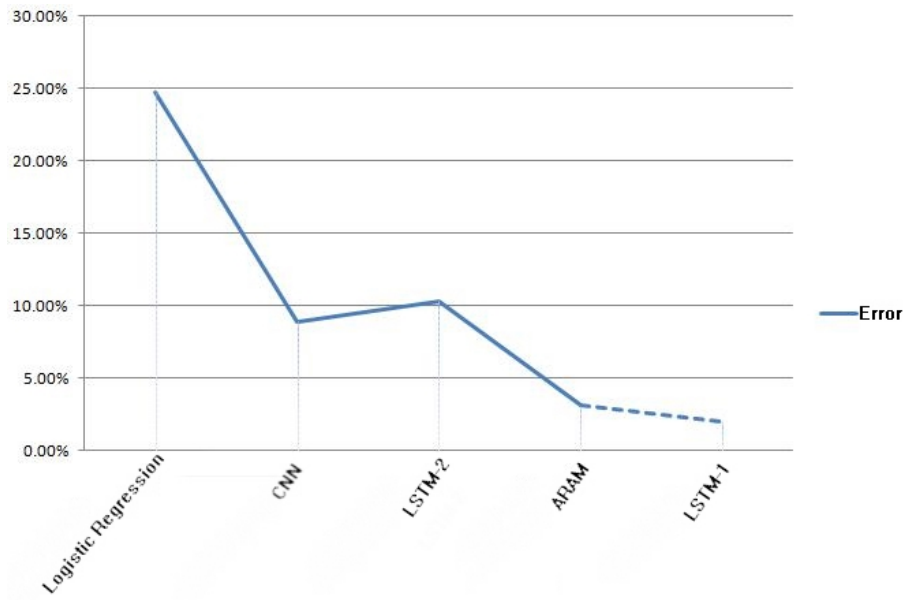

Fig. 16 Error trend of all prediction models

In summary, the following conclusions can be drawn: the selection of the characteristic attributes of the prediction model directly affects the accuracy of the prediction results. The more obvious the influence of the characteristic attributes on the housing price is, the more accurate the result will be. Also, the more realistic and effective the training data of the prediction model is, the more accurate the prediction result will be. By contrast, the convolutional neural network (CNN) is more accurate in predicting the characteristic factors obtained by deep crawling, while the LSTM is more accurate in the predicting of the time series data. The prediction results of CNN and LSTM are affected by their own parameters and training times, and increasing training times can reduce the error.

V. Conclusion

Nowadays, real estate has a variety of characteristic factors, such as building age, location, surrounding facilities and so on. In this paper, we mainly used a web crawler to crawl the real estate data, and through the preliminary analysis of the data statistics to obtain the required characteristic factors of our prediction models. In the implementation of prediction models, we built three kinds of models, namely logistic regression, CNN and LSTM, and to predict according to a variety characteristics of the real estate. Finally, we compared these models by specific experiments. However, the experimental results showed that the prediction error was large. By further analysis, we found that the reason was that the characteristic data crawled by crawler tools cannot represent all factors, and the housing prices may also be affected by macroeconomic or policy-related factors. In addition, the housing prices are also related to time factors. Therefore, we further studied how the time factors such as monthly housing prices influence the trend of housing prices in a city. We transferred the research object to the monthly housing price trend information, and then use them to predict the housing prices in Beijing in the future. Finally, the prediction results of the five models were compared and analyzed through experiments.

References

[1] Siqin L I, Lin L, Sun C. Click-Through Rate Prediction for Search Advertising based on Convolution Neural Network (J). Intelligent Computer & Applications, 2015.
[2] Akbar N M A, Ali F M, Maryam Z. The Dynamic Effect of Macroeconomic Factors on Functions of Housing Price in Iran (1990-2007) (J). 2010.
[3] Heydon A, Najork M. Mercator: A scalable, extensible Web crawler (J). World Wide Web-internet & Web Information Systems, 1999, 2(4):219-229.
[4] Jain A, Zamir A R, Savarese S, et al. Structural-RNN: Deep Learning on Spatio-Temporal Graphs (J). 2015:5308-5317.
[5] Wöllmer M, Marchi E, Squartini S, et al. Multi-stream LSTM-HMM decoding and histogram equalization for noise robust keyword spotting (J). Cognitive Neurodynamics, 2011, 5(3):253.
[6] Sundermeyer M, Schlüter R, Ney H. LSTM Neural Networks for Language Modeling (C)// Interspeech. 2012:601-608.
[7] Xian Z B, Qiang L. Arma-based Traffic Prediction and Overload Detection of Network (J). Journal of Computer Research & Development, 2002, 39(12):1645-1652.
[8] Wang L. Analysis of Non-steady Time-series Forecast for Economy Based on ARMA Model (J). Journal of Wuhan University of Technology, 2004.S
[9] Shao D, Zhang T, Mannar K, et al. Time Series Forecasting on Engineering Systems Using Recurrent Neural Networks (M)// Advanced Data Mining and Applications. Springer International Publishing, 2016.
[10] Balluff S, Bendfeld J, Krauter S. Meteorological Data Forecast using RNN (J). International Journal of Grid & High Performance Computing, 2017:61-74.