Research on the factors influencing the price of commercial housing based on support vector machine (SVM)

To cite this article: Zhong Xiaoyang *et al* 2018 *IOP Conf. Ser.: Earth Environ. Sci.* **128** 012067

Research on the factors influencing the price of commercial housing based on support vector machine (SVM)

Zhong Xiaoyang, Ren Hong, Gao Jingxin*

(Chongqing University, Chongqing 400045)

Corresponding author's e-mail:jxin25@qq.com

Abstract. With the gradual maturity of the real estate market in China, urban housing prices are also better able to reflect changes in market demand and the commodity property of commercial housing has become more and more obvious. Many scholars in our country have made a lot of research on the factors that affect the price of commercial housing in the city and the number of related research papers increased rapidly. These scholars' research results provide valuable wealth to solve the problem of urban housing price changes in our country. However, due to the huge amount of literature, the vast amount of information is submerged in the library and cannot be fully utilized. Text mining technology has been widely concerned and developed in the field of Humanities and Social Sciences in recent years. But through the text mining technology to obtain the influence factors on the price of urban commercial housing is still relatively rare. In this paper, the research results of the existing scholars were excavated by text mining algorithm based on support vector machine in order to further make full use of the current research results and to provide a reference for stabilizing housing prices.

1. Introduction

Starting from 1998, along with the housing industry to become a new economic growth point in the system reform[1],the introduction and implementation of various policy measures , the real estate market and real estate industry of our country has entered a new period of development. Subsequently, various preferential policies such as tax exemption for the secondary market of living housing were introduced. Housing construction accelerated, housing finance developed rapidly, and the secondary market of housing became active. After decades of development, China's real estate market system has gradually improved and the market mechanism is more robust. But under the condition of overall is good, the phenomenon of "overheating" and "cold winter" appear alternately [2]. Starting form 2009, the housing sales and housing prices of our country have experienced from low to slowly rise to explosive growth stage [3]. Although the government has rolled out the intensive and strict regulation policy since the second half of 2009, the housing prices in the first and second-tier cities remain high [4]. Therefore many scholars in our country do a lot of research on factors affecting the urban commodity residential house prices, the number of papers related to this problem grow rapidly. Input keywords "residential prices" and "impact factor" into the CNKI to search related literatures, get 692 results, and the number of related literature is increasing every year. These scholars' research provides valuable wealth for solving the problem of the urban housing price changes. However, due to the huge amount of literature, the vast amount of information is submerged in the library and cannot be fully utilized. Text mining technology has been widely concerned and developed in the field of Humanities and Social Sciences in recent years. But through the text mining technology to obtain the influence

factors on the price of urban commercial housing is still relatively rare, the existing results are obtained by manual study of some of the literatures.

Luo Ping et al used the System dynamics model to conduct a comprehensive study on the impact factors of housing prices, got a well simulation results[5]; Shen Yue, Liu Hongyu analyzed and qualitatively elaborated the relationship between the housing prices with household income, GDP and other macroeconomic factors[6]; Chen Duochang and Zong Jiafeng pointed out that the levy of property tax, residential transfer income tax and deed tax impact housing prices in different way, and did detailed analysis[7]. Wang Jinming, Gao Tiemei used the data from 1995 to 2002 to analyze the income, price and interest rate elasticity of the residential demand and supply function, but did not give the relationship between the factors[8]; Shao Feibo, Zhang Xin analyzed the influencing factors of residential price form a micro perspective by using the Hedonic model, and pointed out that the factors such as the distance from the city center, the property cost, the volume rate and the educational level are the main factors affecting the housing price[9]. Zhang Jitong, Lan Hao used the Generalized pulse corresponding analysis method to analyze the impact of these factors on housing prices, concluded that interest rates and down payment ratio have greater impact on housing prices [10].

However, these results are widely dispersed in semi-structured or unstructured literature texts and are expressed in natural language, so cannot be processed directly by the system. Salton, G, Wong, A et al proposed that the Vector Space Model method can transform natural language into Vector Space to process text data[11-12]. Asbjornsen, Heidi et al proposed to extract feature words by constructing vocabulary database, and classify different texts by feature vector method[13]; Arenas-garcia, J, perez-cruz, F proposed to classify the literature texts by the method of Support Vector Machine and succeeded[14].This paper builds an impact factors vocabulary database of city commodity housing price by Vector Space Model, uses Support Vector Machine method to classify related texts, and identifies the current scholars research achievements on the impact factors of urban commodity residential house prices.

2. Text mining method of impact factors on urban commodity house price

In order to extract the influencing factors of urban commercial housing price from the current related texts, this paper adopts the following four steps: (1)Extract key words artificially from some text downloaded from the CNKI ,and establish "factors dictionary"; (2) Construct the eigenvector of the text by the Vector Space Model; (3) Use text eigenvector as input to train the text classifier; (4) Select the text with high confidence and extract relevant opinions.

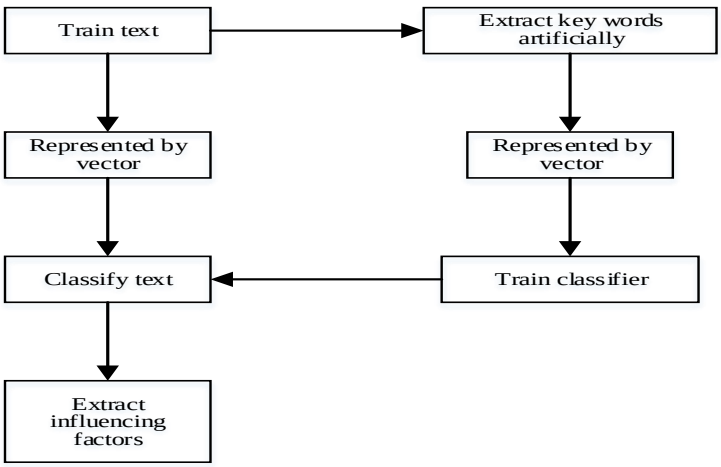

Figure 1. Urban commodity housing price influencing factors mining flow chart

2.1. Obtain the text and construct "factor dictionary"

The characteristics of commodity housing price changes in China appears under the specific background of China, and the characteristics of the transformation from traditional construction to the industrialization is also different from other countries. What's more, the research results about urban commodity residential house prices in China mainly made by the domestic scholars, so the text this paper selected is mainly Chinese text. Currently, the CNKI is the most comprehensive Chinese journals database search engine, including Chinese Journal Full-text Database and Chinese PhD Dissertation Database, Chinese Excellent Master Degree Dissertations Full-text Database , Chinese Newspapers Full-text Database and Chinese Important Conference Literature Full-text Database, etc., so this paper uses the CNKI to search related literature. Search literature with "housing price" and "impact factors" as key words in the CNKI, 692 references were obtained. In this paper, 200 references are selected from 692 references as the trained text, and the key words are extracted from these references to construct the "factor dictionary".

In the process of extracting keywords, because there is no intrinsic space between Chinese words as English, the related text should be segmented into words at first. This paper uses the R language to segment words, then constructs the "factor dictionary". The steps are as follows:

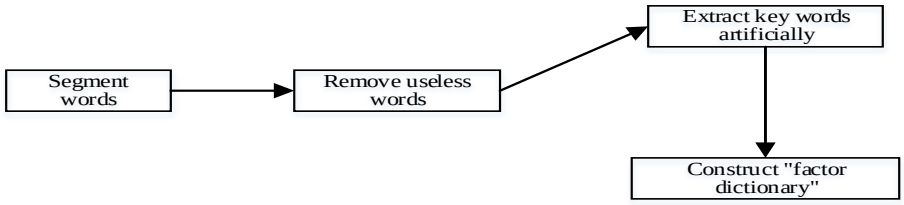

Figure 2. "Factor dictionary" construction process

Use the Chinese word segment package Rwordseg and Rjava of the R language. There are many "useless words" in the word segmentation, which have no real meaning in the actual analysis, such as "what", "even", "but" and so on. Through the "remove common words" command of R language can manually add useless vocabularies and delete these "useless words." Then extract key words artificially of urban commodity prices influencing factors, and establish "factor dictionary." In this paper, 301 key words were obtained through text processing, and finally 152 keywords were left through artificial extraction.

2.2. Text vector representation

In this paper, the text is represented as the eigenvector by Vector Space Model method. The principle is to think of the text as the multidimensional vector space D consisted of a set of orthogonal terms vectors, each of which can be represented as a corresponding normalized eigenvector.

$$V(d) = \{t_1, t_2, \dots \dots t_n\} \quad (1)$$

In formula (1), t_i is the key word extracted, V(d) is the text vector space, n is the dimension of vector space. The keyword t_i in each text d_j corresponds to a weight w_{ij}, of which $d_j \in D$, $w_{ij} > 0$. If $t_i \notin d_j$,then $w_{ij} = 0$. The text can be represented as the corresponding weight characteristic vector through the method of Vector Representation.

$$w_j = \{w_{1j}, w_{2j}, \dots \dots w_{nj}\} \quad (2)$$

This paper uses TF/IDF to calculate the weight vector of text terms [15]:

$$TF/IDF(t_i, d_j) = TF(t_i, d_j) \lg {|D|}/{|DF(t_i)|} \quad (3)$$

In formula (3), t_i is the characteristic word; d_i is the text that contains t_i; $TF(t_i, d_j)$ is the number of t_i occurrences in d_j; $|D|$ is the training text number; $|DF(t_i)|$ is the number of text containing the characteristic word t_i; $TF/IDF(t_i, d_j)$ is the weight of corresponding special words in a particular text. But the obtained weight values of high-frequency words through formula (3) will inhibit the obtained weight values of low-frequency words. Therefore, the weight values obtained need

to be normalized and processed, namely:

$$w_j = \frac{\mathrm{TF/IDF}(t_i, d_j)}{\sqrt{\sum_{j=1}^{|D|}\left(\mathrm{TF/IDF}(t_i, d_j)\right)^2}} \quad (4)$$

In this paper, the process of vectorized representation of the text related to urban commodity house price influencing factors is shown in figure 4

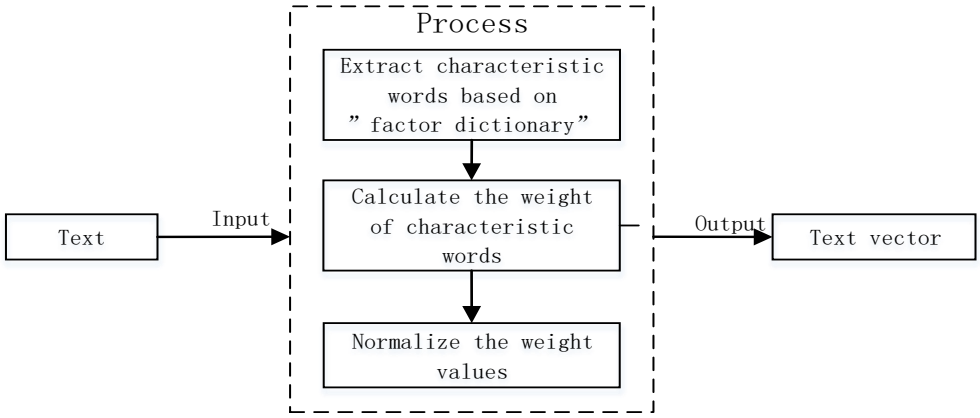

Figure 3. The process of vectorized representation of the text

2.3. Train the classifier

A lot of text classification machine learning algorithms used in text mining, commonly used include the Naive Bayes theorem, K-Nearest Neighbor classification (kNN) ,Support Vector Machines (SVM), C4.5, etc, this paper selects the Support Vector Machine (SVM) and R implementation classifier after comparing. The function package e1071 of R provides libSVM interface. Using e1071 function SVM () can get the same results as libSVM and the function write. This paper selects Gaussian Radial Basis Function to train and apply classifier, and optimizes the two important parameters C and γ of kernel function. The training flow of SVM classifier is shown in figure 5.

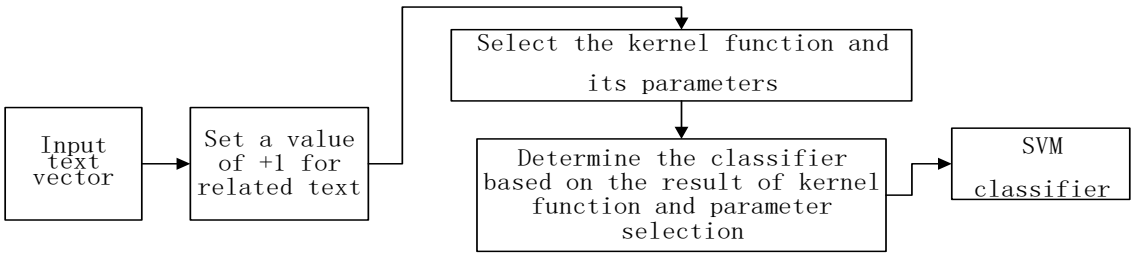

Figure 4. Classifier training flow chart

Support Vector Machine is a machine learning algorithm based on statistical learning, which classifies vectors by constructing optimal hyperplane. Let D be the training set, $D = \{(x_1, y_1)(x_2, y_2) \ldots (x_k, y_k)\}$, where $x \in R^n$, $y \in \{+1, -1\}$. +1 and -1 represent two different categories, The optimal hyperplane equation of support vector machine is $(w * x) + b = 0$, the classification discriminant is:

$$y_i[(w * x_i) + b] \geq 1, i = 1,2, \ldots k \quad (5)$$

The vector that makes formula (5) established is the support vector, $(w * x_i)$ is Ou inner product. For the optimal hyperplane, the solution of its decision function is to find the optimal solution of the vector w and the offset b according to the specific sample, and to minimize the weight cost function $\varphi(w)$.

$$\varphi(w) = {\|\omega\|^2}/{2} \quad (6)$$

For the formula (6), use the Lagrange multiplier $\alpha_i \geq 0, (i = 1,2, \ldots k)$ to transform the problem into the maximum of the target function $L(\alpha)$:

$$L(\alpha) = \sum_{i=1}^{k} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{k} \alpha_i \alpha_j y_i y_j (x_i * x_j) \quad (7)$$

In the formula (7), $\sum_{i=1}^{k} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C$, C is the punishment factor. Assuming that $\alpha' = (\alpha_1', \alpha_2' \ldots \alpha_k')$ is the solution of the problem, then the optimal solution for w and b is as follows:

$$w' = \sum_{i=1}^{k} \alpha' y_i x_i \quad (8)$$

$$b' = \frac{1}{k} \sum_{i=1}^{k} w' * x_k \quad (9)$$

x_k is the support vector and the classification function is

$$f(x) = \operatorname{sgn}\left\{ \sum_{i \in \text{support vector}} y_i \alpha' (x * x_i) + b' \right\} \quad (10)$$

If the formula (10) return +1, $f(x)$ is positive. If return -1, $f(x)$ is negative, and the principle is shown in figure 6.

2.4 Determine the optimal value of C and γ by Cross-Validation

CV(Cross-Validation)is also called loop estimation, is a useful way of statistically cutting data samples into smaller subsets. The basic idea is to group raw data (dataset) into training set and validation set. Firstly, use the training set to train the classifier, then use the verification set to test the training model, as the best performance index(C and γ) for the calculation of classifier (SVM).

3. Experimental results

3.1.Experimental results and evaluation indicators

This paper chooses the accuracy rate a and the recall rate r for the text classification evaluation index of the house price influencing factors.

$$a = \frac{\sum TP}{|D|} \quad (11)$$

$$r = \frac{TP}{TP + FP} \quad (12)$$

TP represents the number of documents belonging to the class and being classified correctly; D represents the number of all documents; FP represents the number of documents belonging to this category but being classified incorrectly.In this paper, the text vector results of different dimensions are compared and analyzed. The results are shown in Table 1.

Table 1. Text classified evaluation index

Dimension	Accuracy rate a	Recall rate r
20	0.851	0.864
30	0.865	0.863
40	0.889	0.889
50	0.872	0.871

As can be seen from the results in Table 1, when the dimension is 30 or more, the accuracy and

recall rate of the results are more than 0.86. This paper chooses the top 30 keywords with the highest χ^2 statistics. Most of these words are urban quality of life, urban economic strength, residents' income, consumption structure, quality of residence, traffic conditions, education investment, social security, medical and health care, Life and health, public safety, living environment, cultural and leisure, employment, GDP, per capita GDP and so on,these words are the key words that well describe the factors that affect the price of urban housing.

3.2. Optimization of classifier parameters and classification results

The main purpose of the text categorization algorithm based on support vector machine is to determine the two parameters of the algorithm: C and γ. In this paper, we compare the correctness of text classification when C and γ take different values. Finally, we determine that C is 5 and γ is 0.0525, and the correct rate of classification is 0.900.

3.3. Extraction and Evaluation of Factors Affecting Price of Urban Commercial Housing

The most common method of analyzing text content is to extract the words in the text and to count the frequency. Frequency can reflect the importance of words in the text. In general, the more important words appear more frequently in the text. After the word extraction, made visual cloud. Show the frequency of different subject words visually can be more intuitively and clearly. In this paper, we randomly selected 85 in the literatures related to the influencing factors of urban commercial housing price downloaded from the CNKI, and 53 related documents were collected by the classifier. Through the word segmentation and word frequency processing, extract urban commodity housing prices influencing factors, use cloud toolkit of R language to draw the word cloud. Obtain the city commercial housing influencing factors cloud.

4. Conclusion

This paper proposes a classification and recognition method for the research of influencing factors of urban commercial housing based on the theory of text mining and support vector machine (SVM). Through constructing of "dictionary" to reduce the dimension of the text vector, and calculate the weight of the feature word by tf-idf, then obtain the text vector. For the classification training of support vector machine, according to the initial sample training to find the support vector to determine the decision function.

In this paper, the support vector machine (SVM) is well used to classify the housing price. It shows that the method can identify the relevant text of the house price influencing factors. It can make full use of the decentralized semi-structured and non-structured text data, provide an important reference for control of housing prices and promote real estate market in China develops healthily. However, after the text classification for the city commodity housing price, factors are extracted artificially, how to automatically extract the relevant views through the text mining need further study.

Reference:

[1] Zheng Siqi. Cao Yang. Liu Hongyu. Urban value determines urban housing price - An empirical study on housing price of 35 cities in China [D]. 2007. 8
[2] Zhang Bo.Spatial distribution and urban value of commercial housing price - Taking Shenyang cty as an Example [J]. Price Theory and Practice. 2006. (10). 49-50
[3] Ren Hong. Wen Zhao. Lin Guangming. "Urban value determines the price" argument analysis and macro-control recommendations [J]. Construction economy. 2007. (08). 22-26
[4] Li Huiping. Study on the Relationship between the Overall Price and Urban Value of Urban Residential [D] .Guangdong University of Technology.
[5] Luo Ping. He Sufang. Niu Huien. An empirical study on the dynamic model of market price system in urban residential market [J]. Humanities geography. 2001. 16(2). 57-61
[6] Sheng Yue. Liu Hongyu. Research on the relationship between real estate prices and macroeconomic indicators [J]. Price theory and practice. 2002. 25 (8).20-22

[7]Cheng Duochang. Zong Jiafeng. The price of real estate tax and residential property: theoretical analysis and policy evaluation [J]. Financial research. 2004. 15 (1).57-60
[8] Wang Jinming. Gao Tiemei Dynamic analysis of demand and supply function of real estate market in China [J]. China soft science. 2004. (4).69-74
[9] Shao Feibo. Zhang Xin. Analysis on the influence factors of Shanghai residential price based on Hedonic model [J]. Economic BBS. 2007. (23). 9-13
[10] Zhang Jitong. Lan Hao. An empirical analysis of the influence factors of real estate prices [J]. China price. 2007. (11).40-42
[11] Calvo Hiram.Méndez Oscar.Moreno-Armendáriz Marco A. Integrated concept blending with vector space models[J]. 2016. 40. 79-96
[12] Salton G.Wong A.Yang C. S. A vector space model for automatic indexing[J]. Communications of the Acm. 1975. 18(11). 273-280
[13] Asbjornsen Heidi.Goldsmith Gregory R.Rebel Karin.Van Osch Floortje P.Rietkerk Max.Chen Jiquan.Gotsch Sybil.Tobón Conrado.Geissert Daniel R. Ecohydrological Advances and Applications in Plant-Water Relations Research: A Review[J]. Journal of Plant Ecology. 2011. 4(1-2). 3-22
[14] Arenas-Garcia J.Perez-Cruz F. Multi-class support vector machines: a new approach[J]. 2003. 2(2). 751-756
[15] He JianMin. Xu Yanling. The bayesian analysis and processing methods of customer complaints in network environment [J]. Journal of HeFei university of technology: natural science edition. 2010. 33 (6).929-933