# Application of SVM Based on Rough Set in Real Estate Prices Prediction

Ting Wang Yan-qing Li
Department of Economic Management
North China Electric Power University 071003
Baoding Hebei China
Liyanqing2002@163.com

Shu-fei Zhao
Department of Applied Mathematics
Hebei University of Technology 300401
Tianjin China

***Abstract:*** **In an increasingly competitive real estate market, establishing a reasonable price has become an important guarantee for the survival of a real estate Developer in the fierce competition. In order to distinguish from the traditional method of formulating prices, this paper selects rough set (RS) and Support Vector Machine (SVM) algorithms to establish a new mathematical model of pricing on the basis of hedonic price. Selecting the rough set to reduce numbers of price indicators, thus reducing the dimensions of the input space of SVM. When treating the reduced data as the input space of SVM, we find that both the convergence speed and the forecast accuracy are enhanced and the result is fairly good prediction.**

***Keywords: Rough Set, Support Vector Machine, Hedonic Price***

## Ⅰ INTRODUCTION

The real estate market is one of the most competitive markets in domestic. Pricing have become the core of real estate business. High prices improve the profits while may affect the sales volume; Low prices, although can expand sales, they may reduce higher profits at the same time. The current formulation of prices uses the cost-oriented pricing, the demand-oriented pricing, the competition-oriented pricing, and the comparable real estate pricing, etc. It's the real estate hedonic price model that gain well promotion and application abroad. Real estate hedonic price model [4] is such a pricing model which conducts multiple regression analysis to the selected hedonic price equation through acquisition of the prices of real estate sample and the hedonic factor data (we refer to it as hedonic regression). Firstly, pricing implicitly for all the hedonic of a residential and ultimately determine the hedonic price function, then pricing and making analysis according to the hedonic price function.

Using the hedonic pricing model to establish the sales price of new residential projects is a new attempt. Due to amount of price attributes, there must exist linearly correlated redundant information which results in the reduction of forecast accuracy and speed of prediction. This paper attempts to use rough set to make a reduction of the hedonic attributes, obtaining a minimal attributes set, then use support vector machine to conduct a regression prediction. We found that the result is fairly good prediction.

## Ⅱ THEORY OF ROUGH SET

The basic idea of rough set(RS)[1] is that any knowledge system S can be expressed by four parts : $S=(U,R,V,F)$ [1], where $R=(X_1,X_2\cdots,X_n)$ is a collection of all samples, $R=C\cup D$ is a attribute set, C is a condition attribute set, reflecting characteristics of object, D is the subset of the attributes for decision-making, reflecting the type of object, $V=\bigcup_{r\in R}V_r$ is a collection of attribute values, $V_r$ is the rang of attribute $r$; $f:U\times R\to V$ is a information function, used to determine the attribute value of $x$ in $U$, i.e. for any $x_i\in U$, $r\in R$, we have $f(x_i,r)=V_r$.

For any attribute set that describes the expression system, it may contain redundant information. An important part of reducing rough set is to reduce the attributes in the decision-making table, making the reduced decision-making table with fewer attributes but still accurately describes objects. Now, the basic concepts of rough set are described below:

*1) Reduction and Kernel:* Assume that $Q\subseteq P$, if $Q$ is independent, and $Ind(Q)=Ind(P)$, then we say that $Q$ is a reduction of equivalent relationship group $P$, written as $\mathrm{Re}\,d(P)$. The collection of all non-omitted relations is called the kernel of equivalent relationship group $P$, written as $Core(P)$, and we have the following relation $Core(P)=\bigcap \mathrm{Re}\,d(P)$ [1].

*2) Discernable Matrix :* Assume that $a_k(x_j)$ is the value of sample $x_j$ on attribute $a_k$, discernable atrix $M(s)=\left[m_{ij}\right]_{n\times n}$, its element of row and column is

$$m_{ij}\begin{cases}a_k\in C, & a_k(x_i)\neq a_k(x_j)\wedge D(x_i)\neq D(x_j)\\ \phi, & D(x_i)\neq D(x_j)\end{cases} \tag{1}$$

### *A Reduction of rough set*

There are two ways for rough set reduction, when the attributes needed to deal with in the decision-making table are less, we can reduce it using discernable matrix directly. Each discernable matrix corresponds to a unique distinguish function $f_{M(s)}$, it's a Boolean function with m variables $a_1,a_2\cdots a_m$ $(a_i\in C,i=1,2\cdots m)$, $f_{M(s)}=\wedge\left\{\vee m_{ij},1\le j\le i\le n,m_{ij}\neq\phi\right\}$. When we compute the discernable matrix using distinguish function, each conjunction in the disjunction we get is just a reduction, and

the intersection of them is kernel.

When the discernable matrix is more complicated, using this method will be very difficult for reduction, then we need to use the heuristic attribute reduction algorithm based on discernable matrix. From the definition of discernable matrix, we can see that, the shorter the length of discernable matrix, the bigger role the attributes will play on classification, and the more frequent it is used, the more important it is. So we use the weighted method to calculate the frequency of attributes, the formula is as following.

$$f(a)=\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\lambda_{ij}}{|m_{ij}|} \tag{2}$$

where $\lambda_{ij}=\begin{cases}0, a\notin m_{ij}\\ 1, a\in m_{ij}\end{cases}$ $|m_{ij}|$ is the number of attributes

Steps of this method are as follows

① Let the set of condition attributes Reduct=R;

② Compute discernable matrix M, find out all the attribute combination S which doesn't contain the kernel attributes.

③ Expressing all the attributes combination that doesn't contain the kernel attributes as disjunctive normal form, $P=\wedge\{\vee m_{ik}, i=1,2,\cdots s, j=1,2\cdots m\}$, calculating the importance of attributes.

④ Choose one of attributes with the smallest importance, let Reduct=Reduct$-\{a\}$, and judge whether the operations is valid, if valid, delete this sample and cycle of the above steps, otherwise, recover data

⑤ The condition for determine whether deletion is $\frac{p_1}{p_0}<\alpha$, where $p_0$ is the sample size in the table of information before reduction, $p_1$ is the inconsistency sample size introduced after reducing. $\alpha$ is a threshold value, usually $\alpha=0.5$.

## III Support Vector Machine

Support Vector Machine (SVM) was formed by Vapnik and others on statistical principle by combining structural risk minimization instead of the traditional experience risk minimization with Huber robust regression theory and Wolfe dual program theory. The basic idea of SVM is to find a nonlinear mapping from input space to output space $\varphi(x)$, mapping the data to high-dimensional space, construct a hyper-plane for data classification. After Vapnik introduced the non-sensitive loss function $\varepsilon$, SVM can solve the nonlinear problems in the high-dimensional space, conducting linear regression forecast [6]. Assume that there is a training set of samples $(x_i, y_i)$, where $x_i=R^n$ is input space, $y_i$ is the output value, then construct a linear function.

$$f(x)=\omega\varphi(x)+b \tag{3}$$

Where $\omega$ is the weight vector of the input space, $b$ is a threshold value, the non-sensitive loss function $\varepsilon$ of SVM take the following form

$$|y_i-f(x_i)|=\begin{cases}0 & f|y_i-f(x_i)|<\varepsilon\\ |f(x_i)-y_i|-\varepsilon & \text{el se}\end{cases} \tag{4}$$

This function means that construct a regional of which thickness is $2\varepsilon$, centered as hyper-plane, if the samples fall into the region, ignoring the difference between predictive value and the actual value, when the samples fall outside the region, give the forecast difference a punishment [6].

The experience risk function is as follows:

$$R_{emp}=\sum_{i=1}^{l}|y_i-f(x_i)| \tag{5}$$

where $l$ is the number of samples.

According to the structural risk minimization criteria of statistical theory, Support Vector Machine determines the regression function by minimizing the risk function. So SVM regression algorithm can be expressed as the following conditions constrained optimization problems.

$$\min\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}|y_i-f(x_i)| \tag{6}$$
$$s.t.\begin{cases}y_i-\omega\varphi(x)-b\le\varepsilon+\xi\\ \omega\varphi(x)+b-y_i\le\varepsilon+\xi^*\end{cases}$$

where $\xi,\xi^*$ is the upper and lower limits of training error, and $\xi,\xi^*\ge 0$

Due to the high-dimensional attribute space, it's almost impossible to solve (6) directly, so we use Wolfe dual skills and introduce the Kernel function $k(x_i,x_j)$. Then we transfer to solve the following dual problem of (4)

$$\begin{aligned}\max(\alpha,\alpha^*)\quad & -\frac{1}{2}\sum_{i=1,j=1}^{l}(\alpha_i-\alpha_i^*)^{\mathrm{T}}(\alpha_j-\alpha_j^*)\langle x_i,x_j\rangle\\ & +\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)y_i-\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\varepsilon\\ s.t.\quad & \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0\\ & 0\le\alpha,\alpha^*\le C\end{aligned} \tag{7}$$

we can get the regression estimation function

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)k(x_i,x)+b \tag{8}$$

## IV SVM based on rough set and algorithm for real estate price forecasting

### A Sample selecting and data preprocessing

SVM has a strong advantage in prediction, through selecting a appropriate parameter $C$, $\varepsilon$, constructing a greatest and optimal hyper-plane to contains as much support vector as possible, thus avoiding the local optimal solution problem of neural network and ensuring the generalization ability of SVM. But when the training set is large-scale, especially when there are numbers of support vectors, the SVM can not distinguish the redundant information and the importance of

each indicator, while the learning process will also take up much of memory, and reduce the speed of optimization, which bring to the practical application of SVM a lot of trouble.

The shortcomings of SVM can be overcome with RS. Because its classification is based on the indiscernible relations, and it only analysis the data and can obtain a minimal attribute set on the premise of preserving key information while doesn't require any prior information or assumption. RS also has shortcomings such as it can only deal with issues such as quantitative data. According to above weaknesses combined with the advantages of both RS theory and SVM algorithm, this paper presents a method which is a combination of RS theory and SVM. Taking advantages of RS in processing of large-scale data, eliminating redundant information and other areas to reduce training data of SVM, overcoming the shortage of SVM algorithm which has low processing speed resulted from large-scale data. Take RS as a pre-system, then under the information structure pretreated using RS method, construct a SVM data prediction system [7].

The data in the decision-making table of rough set used to be expressed with discrete data, so it's necessary to make the indicators value normalized before reducing the attribute. Paper[1] introduced the commonly used equidistance division algorithm, frequency allocation algorithm, and Naïve Scalcer algorithm. This paper uses equidistance division algorithm to reduce the discrete data for the purpose to reduce the input space dimension.

*B The selection of kernel function of SVM*

After the pre-treatment to the data in decision-making table, we use the attribute set in the reduced decision-making table as input space of SVM. There are many types of kernel function, and the commonly used types mainly are polynomial function, RBF function, Sigmoid function, etc. In this paper we chose RBF function. Firstly, RBF kernel can insinuate the input space to feature space nonlinearly to deal with the nonlinear problem in reality facilitatelly. Secondly, compared with the polynomials kernel and Sigmoid kernel, RBF kernel will be less numerical difficulties encounted [8]. Its form is as follows:

$$k(x_i, x) = \exp(-\frac{\left\| x_i - x_j \right\|^2}{2\sigma^2}) \tag{9}$$

*C Example*

In this paper, we use the actual data in literature [4] to analysis. $a_1, a_2, \cdots a_{12}$ represents the residential style and facilities, type of construction, architectural exterior style, traffic, the surrounding environment and education, life, whether suit for investment, operation evaluation, the time when it be trade, whether decoration twelve indicators respectively. Take them as a condition attribute set and sales price as a decision-making attribute set. Paper[4] had processed the indicators data, so this paper only normalize the original data, using the following formula:

$$a_{ij} = \frac{m_{ij} - \min(m_j)}{\max(m_j) - \min(m_j)} \tag{10}$$

Here equidistance division algorithm is used to normalize the discrete data, interval breakpoint is set to five, 0～0.2，0；0.2～0.4，1；0.4～0.6，2；0.6～0.8，3；0.8～1.0，4. Then we got a new decision-making table as Table 1

Table 1 Price indicators

| Sample | Value of Condition Indicators | | | | | | | | | | | | Decision Attribute |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ | $a_9$ | $a_{10}$ | $a_{11}$ | $a_{12}$ | |
| $X_1$ | 3 | 3 | 4 | 1 | 0 | 3 | 3 | 2 | 1 | 1 | 4 | 0 | 2 |
| $X_2$ | 4 | 4 | 4 | 2 | 1 | 3 | 1 | 2 | 2 | 4 | 4 | 0 | 2 |
| $X_3$ | 4 | 4 | 3 | 4 | 3 | 3 | 3 | 4 | 4 | 2 | 3 | 4 | 4 |
| $X_4$ | 3 | 4 | 3 | 2 | 0 | 1 | 1 | 1 | 1 | 0 | 3 | 0 | 2 |
| $X_5$ | 1 | 4 | 1 | 3 | 2 | 1 | 2 | 2 | 4 | 0 | 4 | 0 | 2 |
| $X_6$ | 3 | 2 | 4 | 1 | 1 | 4 | 0 | 2 | 0 | 2 | 4 | 0 | 2 |
| $X_7$ | 3 | 4 | 4 | 3 | 4 | 3 | 2 | 3 | 2 | 1 | 4 | 0 | 2 |
| $X_8$ | 1 | 3 | 1 | 3 | 2 | 1 | 1 | 2 | 3 | 1 | 3 | 0 | 2 |
| $X_9$ | 1 | 4 | 1 | 1 | 3 | 1 | 1 | 1 | 3 | 0 | 3 | 0 | 1 |
| $X_{10}$ | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 4 | 0 | 0 |
| $X_{11}$ | 4 | 4 | 3 | 2 | 0 | 1 | 3 | 2 | 4 | 0 | 3 | 0 | 3 |
| $X_{12}$ | 0 | 3 | 0 | 2 | 2 | 0 | 2 | 1 | 2 | 0 | 3 | 0 | 1 |
| $X_{13}$ | 1 | 4 | 1 | 3 | 1 | 3 | 1 | 2 | 0 | 1 | 4 | 0 | 2 |
| $X_{14}$ | 1 | 4 | 4 | 4 | 2 | 3 | 4 | 4 | 4 | 0 | 3 | 0 | 3 |
| $X_{15}$ | 1 | 4 | 1 | 2 | 2 | 3 | 1 | 1 | 3 | 0 | 0 | 0 | 1 |
| $X_{16}$ | 0 | 2 | 3 | 1 | 2 | 1 | 0 | 1 | 3 | 0 | 2 | 0 | 1 |
| $X_{17}$ | 3 | 4 | 4 | 2 | 0 | 3 | 4 | 2 | 4 | 2 | 2 | 0 | 2 |
| $X_{18}$ | 1 | 4 | 3 | 2 | 2 | 1 | 3 | 1 | 3 | 0 | 0 | 0 | 1 |
| $X_{19}$ | 3 | 2 | 3 | 3 | 1 | 3 | 2 | 2 | 1 | 1 | 4 | 0 | 2 |
| $X_{20}$ | 0 | 0 | 0 | 1 | 2 | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 |
| $X_{21}$ | 1 | 2 | 4 | 3 | 4 | 3 | 3 | 3 | 3 | 2 | 4 | 0 | 3 |
| $X_{22}$ | 3 | 3 | 4 | 2 | 1 | 3 | 3 | 2 | 2 | 2 | 4 | 0 | 2 |
| $X_{23}$ | 1 | 4 | 1 | 2 | 2 | 3 | 3 | 1 | 3 | 0 | 4 | 0 | 2 |

Use the software ROSETTA SAVGenetic Reduce (Genetic algorithm) to reduce the attributes in table 1, the reduced results are: the residential style and facilities, type of construction, the surrounding environment and education, whether suit for investment six indicators. They constitute a new condition attributes set and take it as a input space of SVM. Sampling $X_1$ - $X_{15}$ as the training set, $X_{16}$ - $X_{23}$ test the prediction accuracy of this model and compared with conventional SVM. Using Matlab 7.1 software to do this, the results are as Table 2

**Table 2** Comparison of RS-SVM and Conventional SVM

| | $X_{16}$ | $X_{17}$ | $X_{18}$ | $X_{19}$ | $X_{20}$ | $X_{21}$ | $X_{22}$ | $X_{23}$ |
|---|---|---|---|---|---|---|---|---|
| SVM (%) | 0.93 | 0.82 | 1.12 | 1.09 | 1.21 | 0.84 | 0.97 | 0.96 |
| RS-SVM (%) | 0.8 | 0.76 | 1.03 | 0.95 | 1.18 | 0.78 | 0.81 | 0.87 |

Error is calculated as following formula:

$$e = \frac{\hat{y}_i - y_i}{y_i} \tag{11}$$

where $\hat{y}_i$ is the predicted value, $y_i$ is the real value

## CONCLUSIONS

Through the verification of our example, we conclude that SVM based on the principle of structure risk minimal has the incomparable advantages in forecasting area, avoiding the local optimal defects of the neural network. But because of the input of large numbers of samples, it will take a lot of system memory in the process of running, thus has put forward higher requirements to the machine and requires a longer running time.

After we analysis the advantages and disadvantages of RS and SVM, we try to combine Rough Set with SVM for model prediction. Use Rough Set to reduce the price feature indexes, then use the reduced indexes to train SVM, this greatly reduced the dimension of SVM input space. With illustrate it and compared with conventional SVM, this achieved a better effects.

## REFERENCES


[1]. Jin-pei Wu ,De-shan Sun Contemporary Date analysis[M] Beijing Mechanical Industrial Press 2006

[2]. Hai-zhen Wen , Sheng-hua Jia "Housing characteristics and hedonic price：Analysis based on hedonic price model" Journal of Zhejiang University(Engineering Science) Vo1．38 NO．10 0ct．2004

[3]. Cheng-dong Shi, Ju-hong Chen, Jian-Hu "Study of supply chain per formance prediction based on rough sets and BP neural network". Computer Engineering and Applications, 2007, 43( 33)

[4]. An-ming Chen "Housing Project Hedonic Price Model Based on Principal Components Analysis" Journal of Chongqing University (Natural Science Edition) Vol. 29 No. 6 Jun. 2006.

[5]. Qing-bao Zhang ,Hao-zhong Cheng ,Qing-shan Liu，Ji-wei Zheng，Dong-hai Ni "Short-Term Load Forecasting Based on Attribute Reduction Algorithm of Rough Sets and Support Vector Machine" Power System Technology ，Vo1．30 NO．8 Apr .2006

[6]. Rong-jin Zhu "The Forest of Petroleum Price Based on Support Vector Machine" Industrial Technology &Economy Vol.26,No.2 Feb.2007

[7]. Yuan-cheng Li,Ting-jian Fang "Study of Forecasting Algorithm for Support Vector Machines Based on Rough Sets" Journal of Data Acquisition&Processing Vol.18, No.2 Jun.2003

[8]. Vapnik V. The Nature of Statistical Learning Theory (2nd ed) [M] .Berlin : Springer , 1999