Expert Systems with Applications 38 (2011) 3383–3386

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Housing price forecasting based on genetic algorithm and support vector machine

Gu Jirong [a,*], Zhu Mingcang [b], Jiang Liuguangyan [a]

[a] Key Laboratory of Land Resources Evaluation and Monitoring in Southwest, Sichuan Normal University, Chengdu 610068, China
[b] Land and Resources Department of Sichuan Province, Chengdu 610072, China

ARTICLE INFO

Keywords:
Housing price
G-SVM
Forecasting model
Forecasting accuracy

ABSTRACT

Accurate forecasting for future housing price is very significant for socioeconomic development and national lives. In this study, a hybrid of genetic algorithm and support vector machines (G-SVM) approach is presented in housing price forecasting. Support vector machine (SVM) has been proven to be a robust and competent algorithm for both classification and regression in many applications. However, how to select the most appropriate the training parameter value is the important problem in the using of SVM. Compared to Grid algorithm, genetic algorithm (GA) method consumes less time and performs well. Thus, GA is applied to optimize the parameters of SVM simultaneously. The cases in China are applied to testify the housing price forecasting ability of G-SVM method. The experimental results indicate that forecasting accuracy of this G-SVM approach is more superior than GM.

1. Introduction

Housing price involves multifarious economic interest, including government, the developers, ordinary people, etc. Accurate forecasting for future housing price is very significant for socioeconomic development and national lives (Liu, 2004). In the past years, grey model (GM) is widely applied to forecast housing price. Tough grey model can be constructed by only a few samples, it only depicts a monotonously increasing or decreasing process (Du & Cross, 2007; Hsu and Chen, 2003; Wu & Chen, 2005). Generally, the housing price data can take on fluctuation to a certain extent. Thus, the forecasting results of grey model is unsatisfactory for the housing price forecasting.

Support vector machine (SVM) is based on structural risk minimization principle, which minimizes the expected error of a learning machine and reduces the problem of overfitting (Juhos, Makra, & Tóth, 2008; Varol, Oztop, Koca, & Avci, 2009). This learning machine has been proven to be a robust and competent algorithm for both classification and regression in many applications (Kim, 2003; Pai & Hong, 2006; Tay & Cao, 2002). However, how to select the most appropriate training parameter value is the important problem in the using of SVM. The solution of the problem is very important because the selection of the training parameters of SVM can influence the forecasting accuracy of the SVM. At present, many techniques have been used for the parameter optimization. The most common optimization techniques are Grid algorithm and genetic algorithm (Berti, 2006; Maulik & Bandyopadhyay, 2000; Rubenstein-Montano, Anandalingam, & Zandi, 2000). Grid algorithm method is time consuming and does not perform well. Genetic algorithm can consume less time to perform the optimization of SVM parameters simultaneously. Therefore, the G-SVM method presented in this study is adopted to forecast housing price. Genetic algorithm is used to perform the optimization of SVM parameters. The cases in China are applied to testify the housing price forecasting ability of G-SVM method.

2. The regression theory of SVM

Consider a set of data $G = \{(x_i, y_i)\}_i^n$, where *xi* is a vector of the model inputs, y_i represents the corresponding scalar output, and n is total number of data patterns. To solve a nonlinear regression problem, the inputs are first nonlinearly mapped into a high dimensional feature space wherein they are correlated linearly with the outputs. The SVM formalism considers the following linear estimation function

$$f(x) = (w \cdot \phi(x)) + b, \tag{1}$$

where w is weight vector, b is a constant, $\phi(x)$ denotes a mapping function in the feature space.

ε-insensitive can be used in the SVR formulation. The robust ε-insensitive loss function (L_ε), given below is the most commonly adopted as

$$L_\varepsilon(f(x), y) = \begin{cases} |f(x) - y| - \varepsilon, & |f(x) - y| \geqslant \varepsilon, \\ 0, & otherwise, \end{cases} \tag{2}$$

where ε is a precision parameter representing the radius of the tube located around the regression function $f(\mathbf{x})$.

* Corresponding author.
E-mail address: guJirong@163.com (J. Gu).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.08.123

The coefficients w and b can thus be estimated by minimizing the following regularized risk function.

$$R(C) = C\frac{1}{n}\sum_{i=1}^{n} L_{\varepsilon}(f(x), y) + \frac{1}{2}\|w\|^2, \tag{3}$$

where C is the regularization constant. $\frac{1}{2}\|w\|^2$ is the regularization term which controls the trade-off between the complexity and the approximation accuracy of the regression model.

Here, the positive slack variables ξ_i and ξ_i^* are introduced, which make Eq. (3) transform into the following constrained form:

$$\text{Min } R(w, \xi, \xi^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{4}$$

$$\text{s.t. } \begin{cases} y_i - [w \cdot \varphi(x)] - b \leqslant \varepsilon + \xi_i, & \xi_i \geqslant 0, \\ [w \cdot \varphi(x)] + b - y_i \leqslant \varepsilon + \xi_i^*, & \xi_i^* \geqslant 0. \end{cases}$$

a_i and a_i^* are the Lagrange multipliers. They satisfy the equalities $a_i \cdot a_i^* = 0,\ a_i \geqslant 0,\ a_i^* \geqslant 0$.

The maximal dual function has the following form:

$$\text{Max } R(a_i, a_i^*) = \sum_{i=1}^{n} y_i(a_i - a_i^*) - \varepsilon\sum_{i=1}^{n}(a_i + a_i^*) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i - a_i^*)(a_j - a_j^*)K(x_i, x_j) \tag{5}$$

$$\text{s.t. } \sum_{i=1}^{n}(a_i - a_i^*) = 0, \quad a_i, a_i^* \in [0, C].$$

Finally, by introducing Lagrange multipliers and exploiting the optimality constraints, the decision function has the following explicit form:

$$f(x, a_i, a_i^*) = \sum_{i=1}^{n}(a_i - a_i^*)K(x_i, x) + b. \tag{6}$$

Based on the Karush–Kuhn–Tucker's (KKT) conditions of solving quadratic programming problem, only some of $(a_i - a_i^*)$ will be held as non-zero values. $K(x_i, x) = \phi(x_i) \cdot \phi(x)$ is called the kernel function. In the kernel functions of SVM, radial basis function (RBF) is the most widely used, which is defined as $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|x_i - x_j\|/2\sigma^2)$, where σ denotes the width of the RBF. Thus, C, σ and ε are the parameters which need determination by user.

3. Housing price forecasting model based on G-SVM

3.1. The preprocessing of housing price data

After the normalization of housing price data, the training sample sets are constructed, which is expressed as followings:

$$X = \begin{bmatrix} d_1 & d_2 & \cdots & d_m \\ d_2 & d_3 & \cdots & d_{m+1} \\ \vdots & \vdots & & \vdots \\ d_{n-m} & d_{n-m+1} & \cdots & d_{n-1} \end{bmatrix}, \quad Y = \begin{bmatrix} d_{m+1} \\ d_{m+2} \\ \vdots \\ d_n \end{bmatrix}, \tag{7}$$

where m is the dimension of the input vector.

3.2. The structure of G-SVM

The structure of G-SVM for forecasting housing price is shown in Fig. 1. In the model, kernel function parameter σ, insensitive loss parameter ε, soft margin constant C penalty parameter of support vector machine (SVM) is selected by GA. The structural and operational process of the G-SVM is described below.

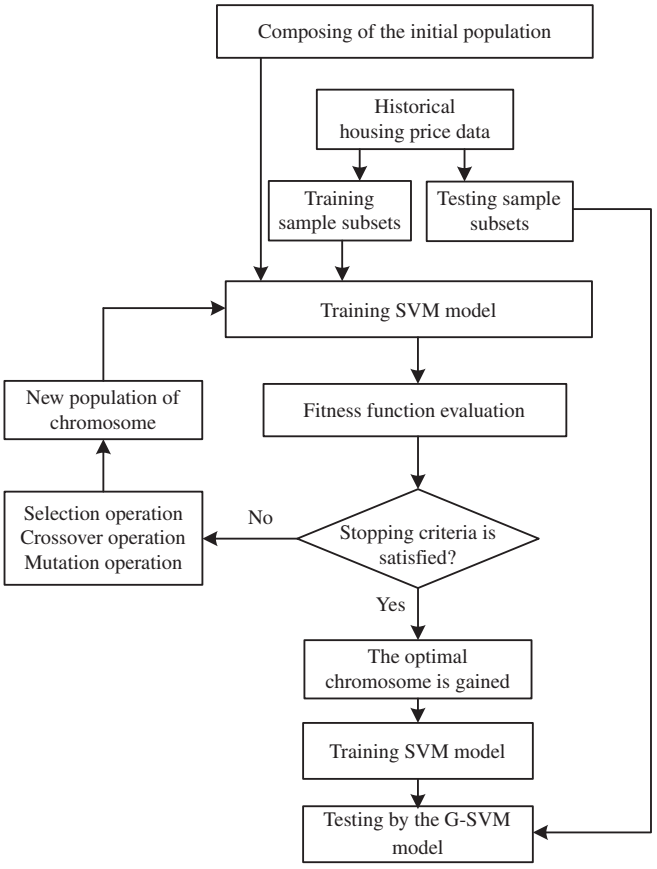

Fig. 1. The operational process of the G-SVM.

3.2.1. Composing of the initial population

The chromosome of G-SVM consists of kernel function parameter σ, insensitive loss parameter ε, penalty parameter C. The initial population of the genetic algorithm is consisted of 20 chromosomes. Each chromosome is consisted of three segments. The first segment of a chromosome represents kernel function parameter σ. The second segment of a chromosome represents penalty parameter C. The third segment of a chromosome represents value of insensitive loss parameter ε. Randomly generate an initial population of the chromosomes.

3.2.2. Calculation of the fitness function

The training subsets are used for calculating the fitness value of a chromosome of G-SVM. In the training subsets, one of them is taken as validation set in turn, others are taken as training set. The fitness function is defined as $\frac{1}{l}\sum_{i=1}^{l}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$, Where y_i and $\hat{y}_i$ represent the actual and validation values, respectively.

3.2.3. Selection operation

In the study, excellent chromosomes are selected to reproduce by means of the roulette wheel.

3.2.4. Crossover operation

Randomly genes between two chromosomes are exchanged by means of the single-point, the probability of creating new chromosomes in each pair is set to 0.8.

3.2.5. Mutation operation

Mutation is performed to alter binary code. That is, if a bit is equal to 1, it is changed to 0; if it is equal to 0, it is changed to 1. The rate of mutation is set to 0.01.

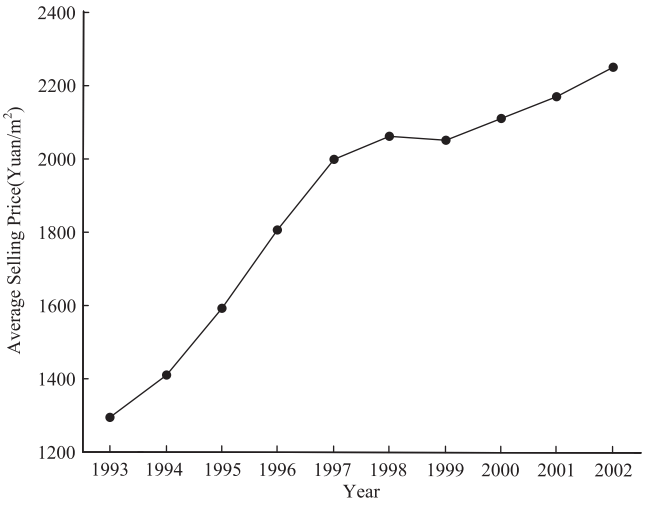

Fig. 2. National average selling price in China from 1993 to 2002.

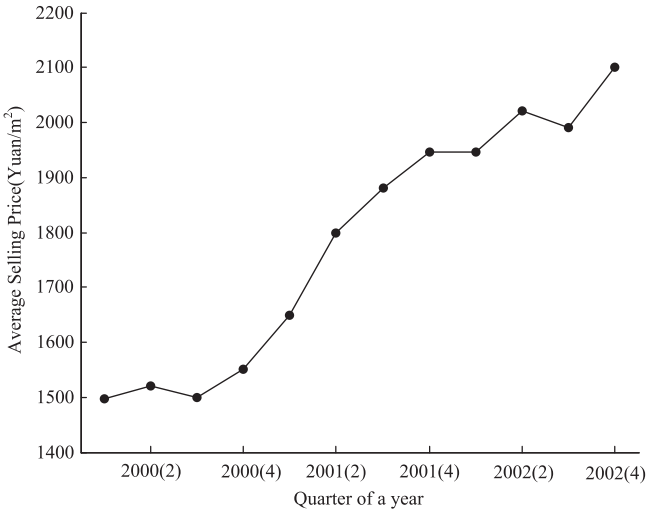

Fig. 3. Average selling price of a certain district in Tangshan city in quarter of a year from 2000 to 2002.

Table 1
The suitable parameters for the G-SVM model in the housing price forecasting cases.

No.	C	σ	ε
1	6.69	1.38	0.00002
2	6.84	1.47	0.00002

These operational processes of the G-SVM approach are realized, respectively for each of the 20 chromosomes in population for a generation. If the evolutionary process proceeds until stopping criteria is satisfied, otherwise goes to step 2.

4. Experimental studies for forecasting housing price based on G-SVM

We employ national average selling price in China from 1993 to 2002 and average selling price of a certain district in Tangshan city in quarter of a year from 2000 to 2002 to study the housing price forecasting performance of the G-SVM method compared with GM. The cases data are shown in Figs. 2 and 3, respectively. In the case of national average selling price, the data from 1993 to

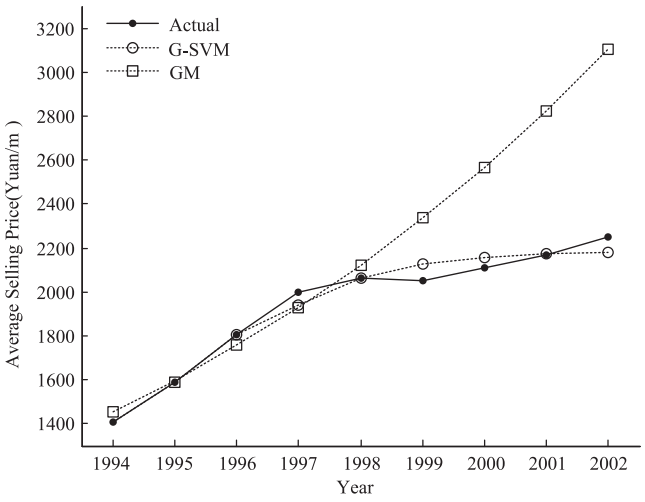

Fig. 4. Forecasting values by G-SVM and GM in case 1.

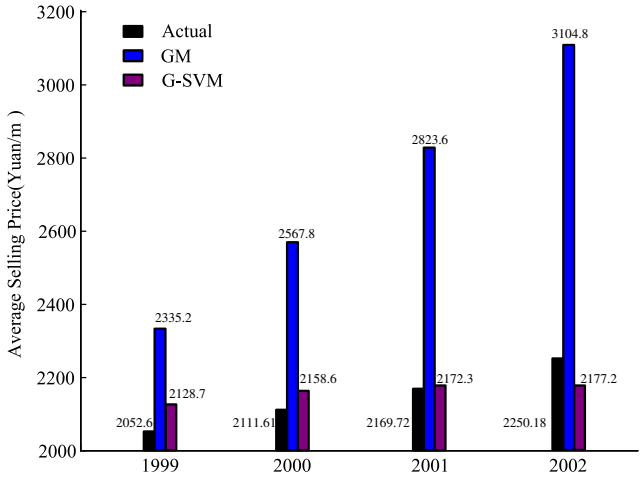

Fig. 5. Comparison of forecasting values between G-SVM and GM in case 1.

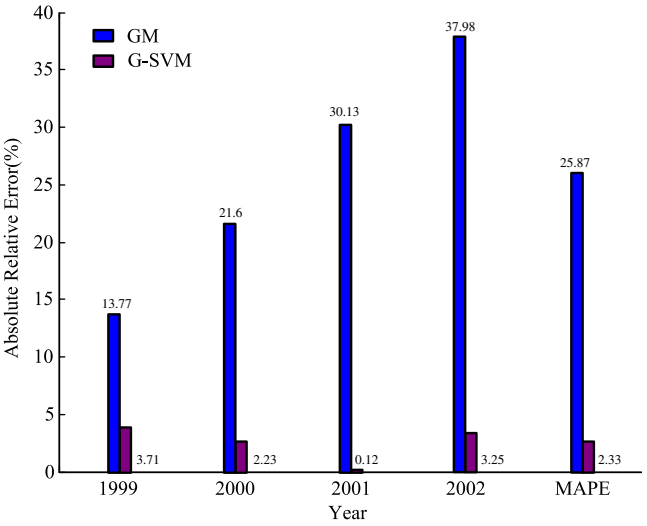

Fig. 6. Comparison of absolute relative forecasting error between G-SVM and GM in case 1.

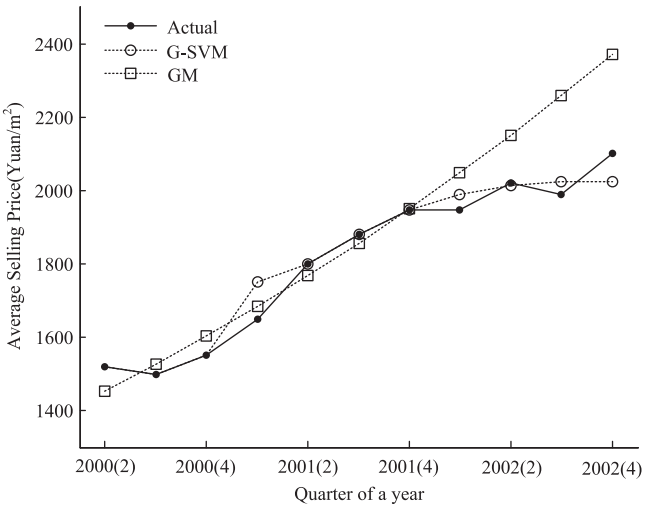

Fig. 7. Forecasting values by G-SVM and GM in case 2.

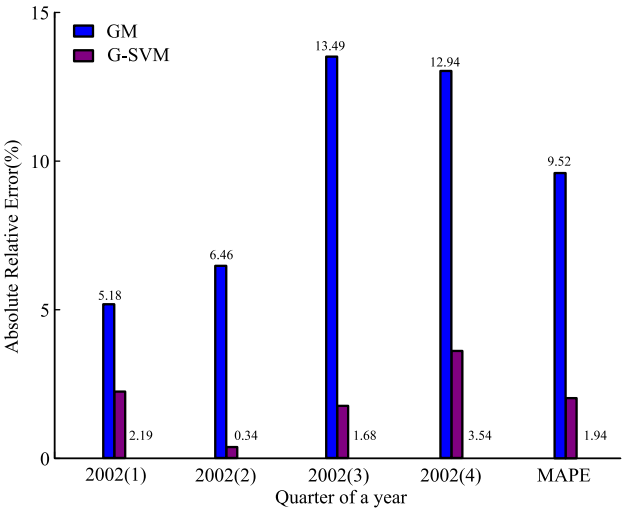

Fig. 9. Comparison of absolute relative forecasting error between G-SVM and GM in case 2.

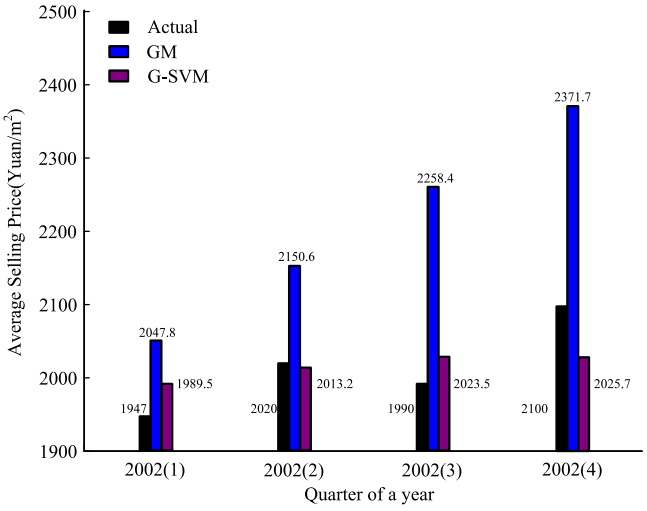

Fig. 8. Comparison of forecasting values between G-SVM and GM in case 2.

1998 are adopted as training data, the data from 1999 to 2002 are adopted as testing data. In the case of average selling price of a certain district in Tangshan city, the data from the first quarter in 2000 to the last quarter in 2001 are adopted as training data, the data from the first quarter in 2002 to the last quarter in 2002 are adopted as testing data. GA is applied to search for the optimal parameter settings. The suitable parameters for the G-SVM model are shown in Table 1. The forecasting values and absolute error between G-SVM and GM are shown in Figs. 4–9. The experimental shows that G-SVM has more excellent performance than GM in forecasting housing price.

5. Conclusion

In this study, a hybrid of genetic algorithm and support vector machines (G-SVM) approach is presented to forecast housing price. Compared to Grid algorithm, genetic algorithm method consumes less time and performs well. Thus, GA is applied to the optimization of SVM parameters simultaneously in the paper. National average selling price in China from 1993 to 2002 and average selling price of a certain district in Tangshan city in quarter of a year from 2000 to 2002 are employed to study the housing price forecasting performance of the G-SVM method compared with GM. The experimental results indicate that forecasting accuracy of this G-SVM approach is more superior than GM.

References

Berti, Guntram (2006). GrAL—the grid algorithms library. *Future Generation Computer Systems, 22*(1–2), 110–122.

Du, Jia-Chong, & Cross, Stephen A. (2007). Cold in-place recycling pavement rutting prediction model using grey modeling method. *Construction and Building Materials, 21*(5), 921–927.

Hsu, Che-Chiang, & Chen, Chia-Yon (2003). Applications of improved grey prediction model for power demand forecasting. *Energy Conversion and Management, 44*(14), 2241–2249.

Juhos, István, Makra, László, & Tóth, Balázs (2008). Forecasting of traffic origin NO and NO_2 concentrations by support vector machines and neural networks using principal component analysis. *Simulation Modelling Practice and Theory, 16*(9), 1488–1502.

Kim, Kyoung-jae (2003). Financial time series forecasting using support vector machines. *Neurocomputing, 55*(1–2), 307–319.

Liu, Da-jiang (2004). The application of Gray–Markov estimation model in real estate price prediction. *Journal of Tangshan College, 17*(4), 44–46.

Maulik, Ujjwal, & Bandyopadhyay, Sanghamitra (2000). Genetic algorithm-based clustering technique. *Pattern Recognition, 33*(9), 1455–1465.

Pai, Ping-Feng, & Hong, Wei-Chiang (2006). Software reliability forecasting by support vector machines with simulated annealing algorithms. *Journal of Systems and Software, 79*(6), 747–755.

Rubenstein-Montano, Bonnie, Anandalingam, G., & Zandi, Iraj (2000). A genetic algorithm approach to policy design for consequence minimization. *European Journal of Operational Research, 124*(1), 43–54.

Tay, Francis E. H., & Cao, L. J. (2002). Modified support vector machines in financial time series forecasting. *Neurocomputing, 48*(1–4), 847–861.

Varol, Yasin, Oztop, Hakan F., Koca, Ahmet, & Avci, Engin (2009). Forecasting of entropy production due to buoyant convection using support vector machines (SVM) in a partially cooled square cross-sectional room. *Expert Systems with Applications, 36*(3), 5813–5821.

Wu, Wann-Yih, & Chen, Shuo-Pei (2005). A prediction method using the grey model GMC(1, *n*) combined with the grey relational analysis: a case study on Internet access population forecast. *Applied Mathematics and Computation, 169*(1), 198–217.