http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 13 (18): 2710-2719, 2014
ISSN 1812-5638

ansinet
Asian Network for Scientific Information

RESEARCH ARTICLE OPEN ACCESS

DOI: 10.3923/itj.2014.2710.2719

Study on Residential Hedonic Price Classification Model Based on MDLP Binning and Support Vector Machine

Heyong Wang and Ming Hong
Department of E-Business, South China University of Technology, Guangzhou, 510006, China

Corresponding Author:
Heyong Wang,
Department of E-Business,
South China University of Technology,
Guangzhou, 510006, China

ABSTRACT

Currently, studies about Support Vector Machine (SVM) and residential price focus on proving that the improved SVM models they proposed are better and forecasting numeric values of residential prices using Support Vector Regression (SVR). Most current studies improves SVM models by improving algorithms or develops forecasting models to predict numeric values of residential prices directly. It is also meaningful to give reasonable intervals of residential prices to tell the ranges of numeric values. This study improves SVM models by classifying residential price variable as well as choosing the best kernel with the best parameters. In this study, residential hedonic price classification model based on Minimal Description Length Principle (MDLP) binning and SVM are proposed. Residential price values are classified using MDLP binning. Furthermore, SVM models are built using four kernel functions with different parameters in SVM to study the accuracy of residential price groups forecasting in different models. By experiment, it proves that linear kernel function is best, RBF and polynomial kernel better and the worst is sigmoid kernel.

Key words: Residential hedonic price, MDLP binning, support vector machine

INTRODUCTION

It is a major issue faced by housing developers to determine reasonable real estate sales prices and accurately grasp the market changes to make timely modifications or adjustments, on the one hand to win customers as well as the market and to maximize profits on the other hand. The traditional methods to determine estate sales prices include cost-oriented method, market-oriented method and demand-oriented method. The widely used method currently is to determine the estate sales price based on the market price using hedonic price model (Lu, 2012).

Hedonic price model is commonly used in residential price valuation. The model can evaluate residential prices from the implicit prices of real estate features, to overcome various defects of traditional methods and can get good results. Various researches have studied residential price valuation using hedonic price model. Wang (2006) analyzes urban residential hedonic price model theoretically and Ma and Li (2003), Wen and Jia (2006), Hao and Chen (2007), Zhang and Chen (2008) and Guo *et al.* (2006) analyze the factors affecting residential prices in particular regions using hedonic price model and construct their particular hedonic price models. Butler (1982), Ozanne and Malpezzi (1985) and Crocker *et al.* (1987) study the selection of factors that affect residential prices.

In practical applications, residential price values may be processed so that they satisfy assumptions of a particular model or can be studied in other aspects. The processions can be functional transform or data discretization, etc. Some of the current studies have been involved in residential price processing. Liu and Zhao (2013) have chosen logarithmic model to show the relationship between residential prices and residential features based on the enjoy price model. They chose natural logarithm of residential price as dependent variable in the model. Li *et al.* (2011) have classified the residential prices in 12 towns (streets) in Baoshan District in Shanghai. Gu *et al.* (2011) have processed residential prices so that they meet normal distribution. Gu and Xia (2013) have applied a method to forecast real estate prices of some buildings in Nanjing and they have used hierarchical clustering method to define the state of real estate price in particular time.

SVM has been applied on the study of residential price. Gu *et al.* (2011) have predicted residential numeric values based on G-SVM and genetic algorithm. Stitson *et al.* (1997) have proposed ANOVA kernel and have applied it to Boston Housing Price dataset. Their result has shown that ANOVA kernel performs better. Hao and Chiang (2007, 2008) have proposed a model based on fuzzy analysis and SVM. Lin and Chen (2011) from Mingchi University of Technology have compared SVR and neural network in their effectiveness in forecasting residential prices. Agarwal (2002) from AT&T Labs has studied the proximal support vector machine model.

Most current studies improves SVM models by improving algorithms or developes forecasting models to predict numeric values of residential prices directly. It is also meaningful to give reasonable intervals of residential prices to tell the ranges of numeric values bacause intervals can be used to test the accuracy of other models forecasting numeric values of residential prices directly. This study improves SVM models by classifying residential price variable as well as choosing the best kernel with the best parameters. In this study, residential price classification model based on Minimal Description Length Principle (MDLP) binning and SVM (M-SVM is used for short in the context in this study) are proposed. Residential price values are classified using MDLP binning. This research studies the accuracy on residential price groups forecasting other than residential price numeric values directly because it is also meaningful to give reasonable intervals of residential prices to tell the ranges of numeric values. M-SVM model can also be used with other residential price forecasting models to improve the accuracy. Moreover, this study builds models using four kernel functions with different parameters in SVM to study the accuracy of residential price groups forecasting in different models.

By experiment, it proves that linear kernel function is best, RBF and polynomial kernel better and the worst is sigmoid kernel.

MATERIALS AND METHODS

Selection and quantification of residential characteristics: The following three types of residential features are commonly considered important in the domestic and foreign studies (Can and Megbolugbe, 1997; Ma and Li, 2003):

- Building Structure Features (BSF), such as the housing area, orientation
- Regional Characteristics (RC), such as the distance from public transportation to the city center
- Neighborhood Environmental Characteristics (NEC), such as the residential infrastructure, culture and entertainment

This study has selected 20 residential characteristic variables including 12 building structure features, 5 regional characteristics and 3 neighborhood environmental characteristics. Variables chosen are shown in Table 1.

Table 1: Selection of residential characteristics

Type	Residential characteristics variables
BSF	Total floors, floor, residential age, residential area, decoration level, orientation, number of rooms, number of living rooms, number of bathrooms, number of balconies, property type, packing spaces
RC	Floor area rate, greening rate, public facilities, education facilities, sports and commercial facilities
NEC	Area, transport facilities, CBD distance

Table 2: Quantification of residential characteristics variables

Type and variables	Method used
BSF	
Floor	Likert scale
Fecoration level	Likert scale
Orientation	Likert scale
Property type	Likert scale
Packing spaces	Dummy variables
RC	
Area	Likert scale
Transport facilities	Comprehensive index
NEC	
Public facilities	Likert scale
Education facilities	Comprehensive index
Sports and commercial facilities	Comprehensive index

In the process of establishing residential hedonic price models, non-numerical data need to be quantified. Methods commonly used for quantification include comprehensive index method, numerical method, dummy variables method, Likert scale method and fuzzy mathematics method. The non-numerical residential characteristics variables chosen in this study are quantified using one of the methods mentioned in Table 2 and the numerical variables retain their original values.

MDLP binning: Minimal Description Length Principle (MDLP) is essential to determine an output variable for MDLP binning to classfify target variable (or input variable). MDLP binning succeed only if the classified input variable can explain output variable better (Xue and Chen, 2010).

Sample set is defined as S, output varaible as C. C can be classified by k categories. $P(C_i, S)$ is the probability when value of C is C_i(C_i is the ith category, one of the k categories). The entropy of S Ent(S) can be defined using the following equation:

$$\text{Ent}(S) = \sum_{i=1}^{k} P(C_i, S)\log_2(P(C_i, S)) \quad (1)$$

Obviously, Ent(S) minimizes when there exists an i to make $P(C_i, S)$ equals to 1 with other $P(C_t, S)$ ($t = 1,2,3,\ldots, k$ and $t \neq i$) equals to 0. The larger Ent(S), values of C are more uncertain, witch means larger amount of information.

The input variable is defined as A, T as the separate point. The whole sample set S can be separated into two groups, S_1 and S_2. The entropy of S with input variable A and separate point T Ent(A, T, S) can be defined using the following equation:

$$\text{Ent}(A, T; S) = \frac{|S_1|}{|S|}\text{Ent}(S_1) + \frac{|S_2|}{|S|}\text{Ent}(S_2) \quad (2)$$

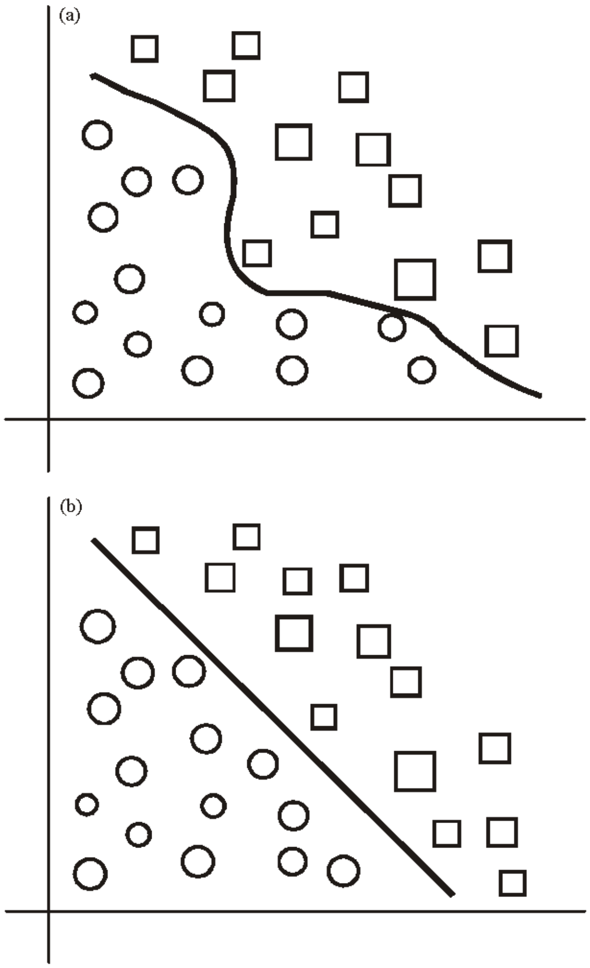

Fig. 1(a-b): Principle of SVM, (a) Original data and (b) Linearly separable mapped data

where, $|S_1|$ and $|S_2|$ are defined as the amount of records in S_1 and S_2. Furthermore, information gains of S with input variable A and separate point T Gains (A, T; S) can be defined using Ent (A, T; S) and Ent (S):

$$\text{Gains (A, T; S)} = \text{Ent (S)} - \text{Ent (A, T; S)} \tag{3}$$

Gains (A, T; S) is the criteria to determine whether separate point T is effective. In order to show more details about Gains (A, T; S), it is essential to define another two equations:

$$G = \frac{\log_2(N-1)}{N} + \frac{\Delta(A,T;S)}{N} \tag{4}$$

$$\Delta(A;T;S) = \log_2(3^k-2)-[k\text{Ent}(S)-k_1\text{Ent}(S_1)-k_2(S_2)] \tag{5}$$

In Eq. 3-5, k_1 and k_2 are the number of categories in S_1 and S_2.

If Gains (A, T:S)>G, T is effective, witch means the classified output variable can explain input variable better. If not, T is not effective and it t to choose a new one. If there are more than one T, the T which leads to the largest Gains (A, T:S) is chosen. According to the effective T and a chosen output variable, the input variable can be separated into two groups.

Support vector machine: SVM is a both linear and non-linear classification method. It can be used for not only regression but also classification. SVM serves small and non-distributed datasets well.

SVM classifies data by searching the best separating hyperplane. It is the specific approach to map the original training data to a higher demension on witch the algorithm searches the best separating hyperplane. The principle of SVM is shown in Fig. 1. The original data can not be separate linearly but after they are mapped to a higher demension, they become linearly separable so they can be separated by the best separating hyperplane.

Considering a linearly separable dataset, vector X is defined as one of the records in the dataset, vector W is defined as weight vector and b is defined as bias. Separating hyperplane can be defined as:

$$W \times X + b = 0 \tag{6}$$

If $X = (x_1, x_2)$, then Eq. 6 can be rewritten as:

$$w_0+w_1x_1+w_2x_2 = 0 \tag{7}$$

Two planes can be defined as following by adjusting weight values w_1 and w_2:

$$H_1\text{: } w_0+w_1x_1+w_2x_2 \geq 1 \tag{8}$$

$$H_2\text{: } w_0+w_1x_1+w_2x_2 \leq -1 \tag{9}$$

Points above H_1 are classified as one class and points below H_2 are classified as second class. Points on H1 and H2 are called support vectors witch are the hardest tuple to classify and owns most classification information. They are more than one separating hyperplanes can separate points (Fig. 2). But distance between two classes (that is between H_1 and H_2) can be maximized to make the best classification by only one hyperplane witch is called Maximum Marginal Hyperplane (MMH). MMH is the best separating hyperplane shown in Fig. 3 (Han and Kamber, 2007).

For the non-linear case, data are mapped to a higher dimension non-linearly. Best separating hyperplane is search in the higher dimension. Kernel functions are applied to map data. Kernel functions used in this study are shown in Table 3.

There are two parameters in SVM: Regularization parameter (represented by C) and regression precision (represented by ε). C controls the trade-off between maximizing the margin and minimizing the training error term. Value of C should normally be between 1 and 10 inclusive. Increasing the value improves the classification accuracy

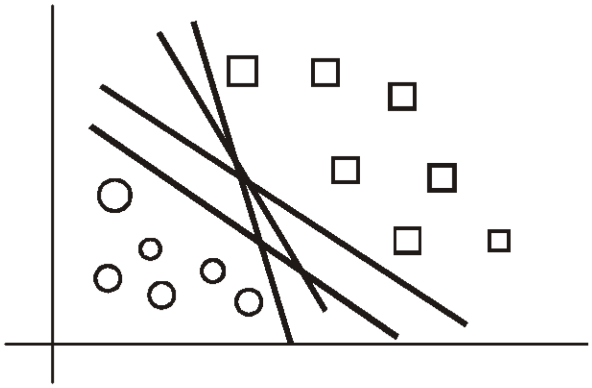

Fig. 2: Separating hyperplanes

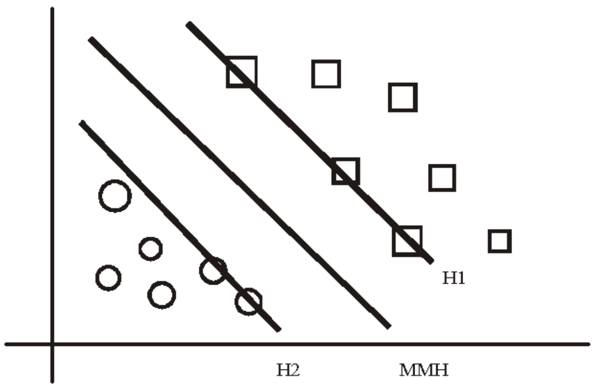

Fig. 3: H1, H2 and MMH

Table 3: Kernel functions

Kernel function	Equation	Parameters
Linear	-	-
Polynomial	$(\gamma\times\mu'\times v+\text{coef }0)^D$	γ, coef 0, D
RBF	$\exp(-\gamma\times\lvert\mu-v\rvert^2)$	γ
Sigmoid	$\tanh(\gamma\times\mu'\times v+\text{coef }0)$	γ, coef 0

(or reduces the regression error) for the training data, but this can also lead to overfitting. ε used only if the data type of the target field is range. Increasing the value may result in faster modeling but at the expense of accuracy for more details about SVM (Nello and John, 2004; Wu and Kumar, 2013; Tan *et al*., 2011).

Experiments

Experimental preparation: The experimental data used in this study are collected by two ways. Most of them such as building structure features variable are collected from professional real estate information websites such as gz.soufun.com. Data of some regional characteristics and some neighborhood environmental characteristics can not be collected directly and these are collected from Baidu Map and Guangzhou Electric Map. Totally, 407 complete records are collected.

Analysis software used in the experiment is SPSS Clementine.

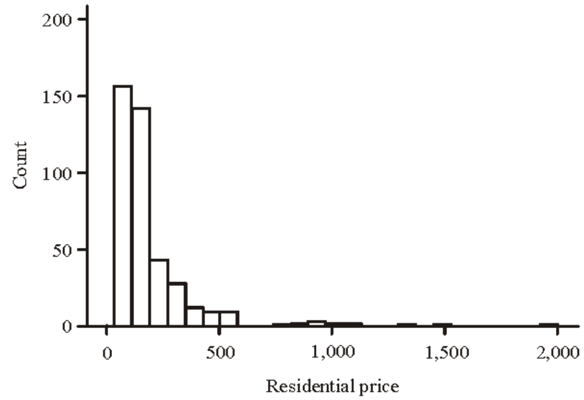

Fig. 4: Distribution of residential price variable of the whole dataset

Table 4: Separating training sets

Ratio of training set (%)	Number of records in training set
50	209
60	229
70	285
80	320

Experimental step:

- To separate training sets and testing sets according to particular ratios
- To check the distributions of residential price variables of training sets and testing sets and then group values of residential price variables
- To build M-SVM models using different kernel functions with different parameters using training sets
- To compare the accuracy of M-SVM models with the same kernel functions whose values of parameters differ using testing sets. Values of parameters witch make the best accuracy are chosen for its kernel function to build M-SVM model
- To change the ratio of training set and repeat the above steps and finally get the conclusion

Separating training sets and testing sets: In this study, different ratios of training sets are shown in Table 4 (totally 407 records).

Residential price variables classification: Before classification, it is necessary to observe the distribution of residential price variable. The distribution of residential price variable of the whole data set is shown in Fig. 4.

It shows that the distribution of residential price variable is skewed to the right in Fig. 5b. So, the distribution of residential price variables of training sets and testing sets are required to be the same as that in the whole data set, that is, also to be skewed to the right so that different experiments are comparable. Figure 5a-d show the distributions of residential price variables of training sets and testing sets. Figures on the left show the distributions of testing sets and the right show the distributions of training sets.

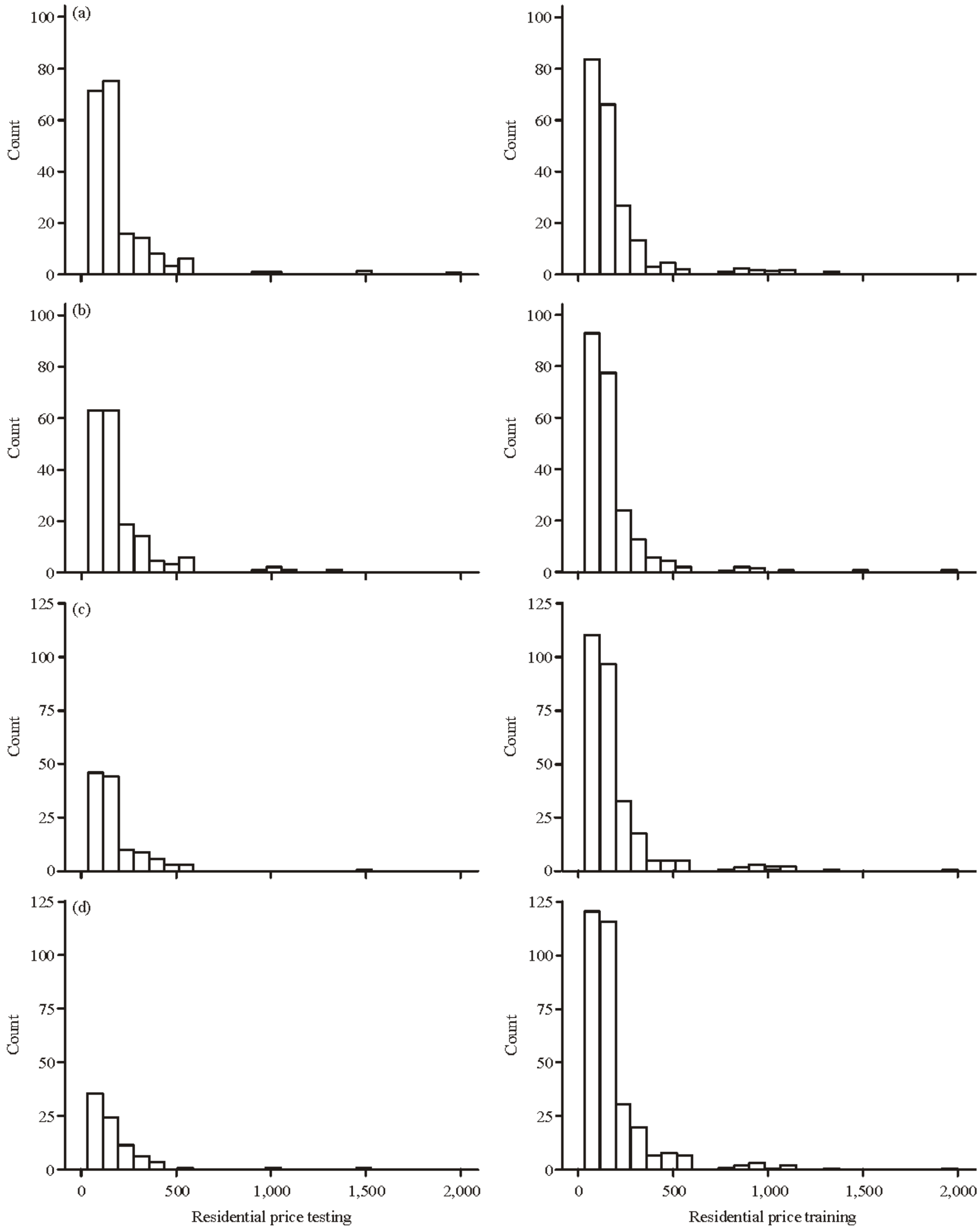

Fig. 5(a-d): Training set of (a) 50%, (b) 60%, (c) 70% and (d) 80%

Form the distributions shown, it is obvious that most of the values are located in the interval [0, 500], very few of them are located in the interval [500, 2000] sparsely. For MDLP, data can only separate into two groups. Area variable is chosen as the output variable because residential prices in different areas are obviously different. Areas of house are ranking in 3 levels (Fig. 6) and it is clear that distributions of residential price in 3 levels are differ from each other.

The result is shown in the following:

- **Group 1:** (0, 102)
- **Group 2:** (102, 1999)

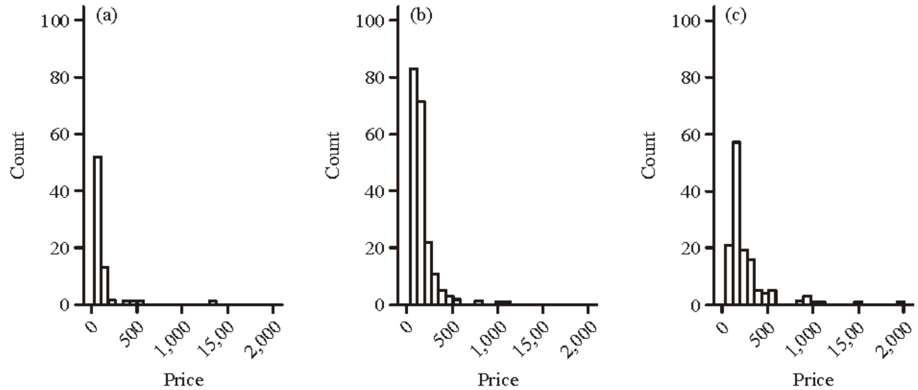

Fig. 6(a-c): Residential prices in areas of different levels, (a) 1.0, (b) 2.0 and (c) 3.0

RESULTS

M-SVM models are built using kernel functions mentioned in previous section. In this study, SVM parameter ε has no affects in building M-SVM because the output variable is flag type. SVM parameter C is set to 10. Parameter coef0's of polynomial kernel and sigmoid kernel are set to 0 bacause "coef0 = 0" can satisfy most cases. There are more details of the four kernel functions in previous section.

M-SVM based on RBF: Generally, parameter γ of RBF can be set from 3/k to 6/k. k is the number of selected input variables. In this study, 20 variables are selected (previous section), so k = 20. The larger γ leads to the higher accuracy of training sets but it may also lead to overfitting. The accuracies of testing sets with different ratios testing by M-SVM based on RBF with different γ's are shown in Fig. 7. It shows in Fig. 7 that accuracies are always high when ratio of training set gets to 80% and accuracies of training sets with different ratios are relatively higher when γ is set to 0.20. So, γ is set to 0.20 for M-SVM based on RBF.

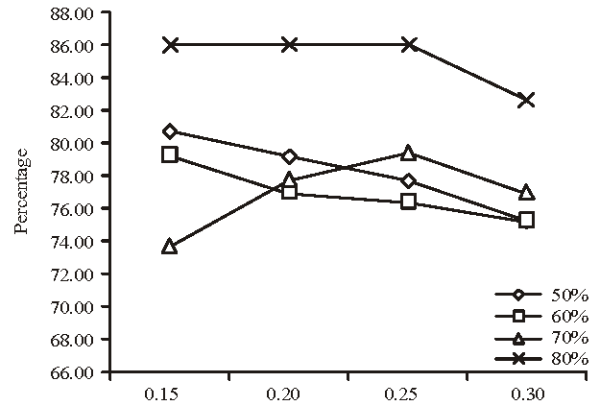

Fig. 7: Accuracy of M-SVM based on RBF

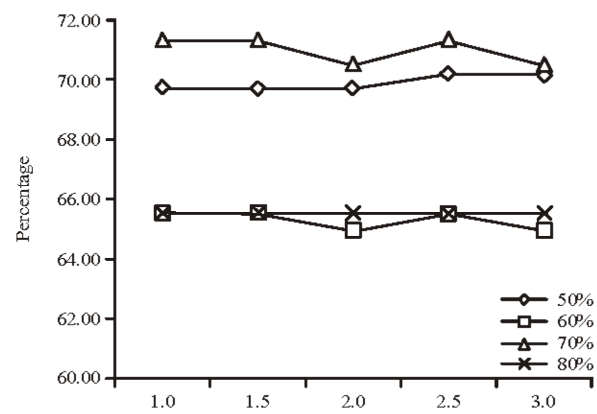

Fig. 8: Accuracy of M-SVM based on sigmoid kernel

M-SVM based on sigmoid kernel: Parameter γ of sigmoid kernel is set to 1.00 or 1.50 or 2.00 or 2.50 or 3.00. The larger γ leads to the higher accuracy of training sets but it may also lead to overfitting. The accuarcies of testing sets with different ratios testing by M-SVM based on sigmoid with different γ's are shown in Fig. 8. It shows in Fig. 8 that accuracies of trianing sets with different ratios are relatively higher when γ is set to 2.50. So, γ is set to 2.50 for M-SVM based on sigmoid kernel.

M-SVM based on polynomial kernel: Parameter γ of polynomial kernel is set to 1.00 or 1.50 or 2.00 or 2.50 or 3.00. The larger γ leads to the higher accuracy of training sets but it may also lead to overfitting. Parameter D of polynomial kernel controls the complexity of mapping space (that is, dimension) and generally D is not greater than 10. D is set to an integer not smaller than 2 and not greater than 10. The accuarcies of testing sets with different ratios testing by M-SVM based on polynomial kernel with different γ's are shown in Fig. 9.

It shows in Fig. 9 that for different D, values of γ has no affects on M-SVM based on polynomial kernel. γ is set to 3.00 in this study.

It shows in Fig. 10 that in the case where D equals to 5, accuracy of testing set increases as the ratio of training set getting greater compared with other values of D and "D = 5" lead to greater accuracies. So, D is set to 5.

M-SVM comparison: In this section, best parameters of each of the four kernel functions (proved above) are chosen for M-SVM models. The accuracies of testing sets with different ratios testing by M-SVM models with different kernel functions are shown in Fig. 11 (R represents ratios of training sets). It shows in Fig. 11 that linear kernel is the best for dataset in this study, polynomial kernel and RBF kernel better and the worst is sigmoid kernel.

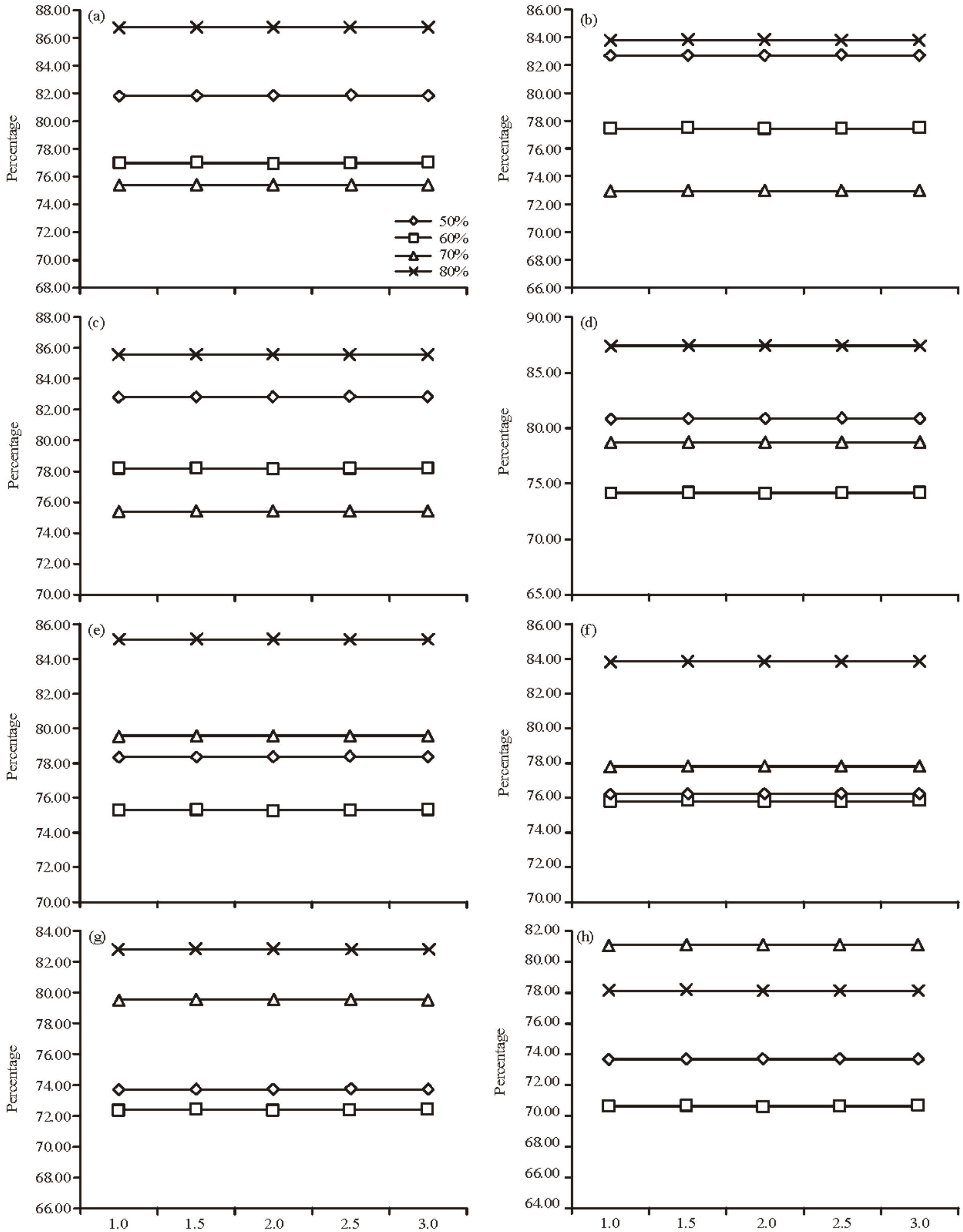

Fig. 9(a-i): Continue

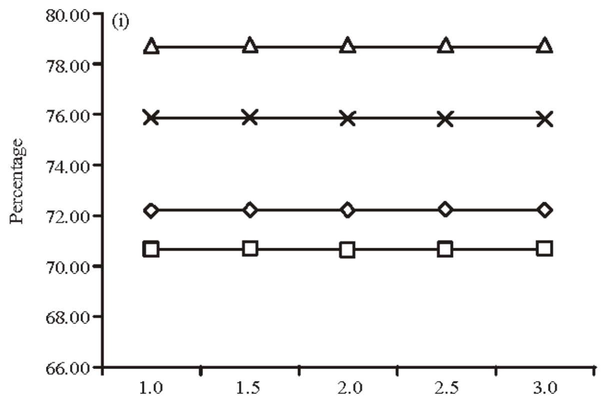

Fig. 9(a-i): Accuracy of M-SVM based on polynomial kernel for different D, (a) D = 2, (b) D = 3, (c) D = 4, (d) D = 5, (e) D = 6, (f) D = 7, (g) D = 8, (h) D = 9 and (i) D = 10

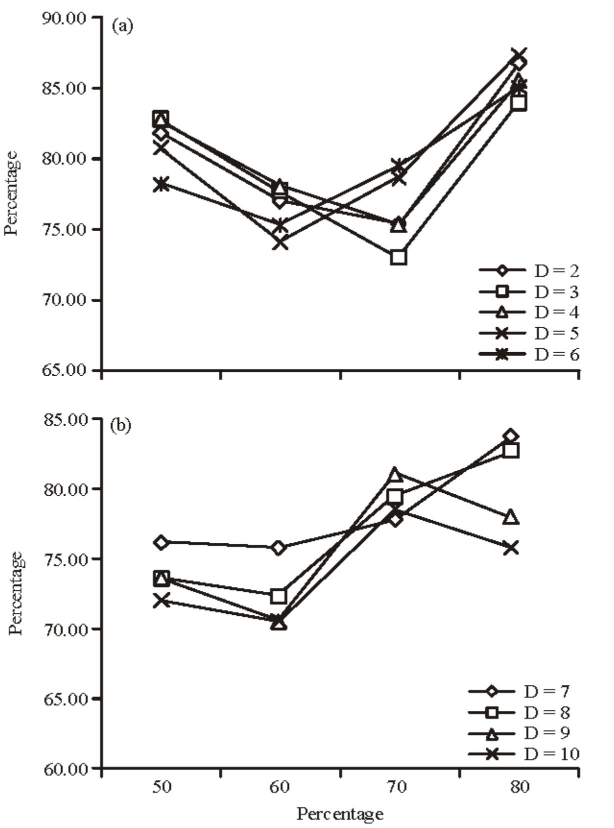

Fig. 10(a-b): Accuracy of M-SVM based on polynomial kernel

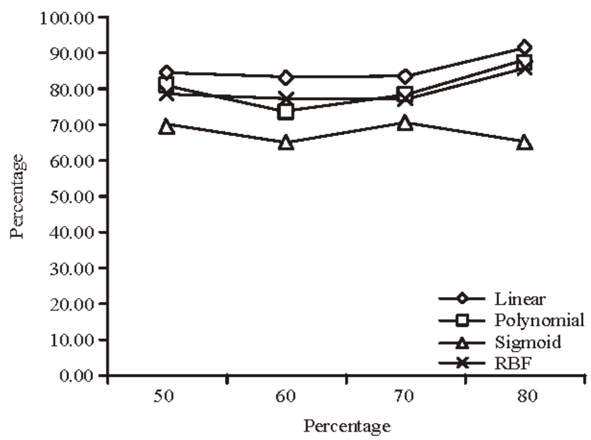

Fig. 11: M-SVM models comparison

DISCUSSION

Some previously published studies have discuss about kernel functions of SVM and optimal parameters of these kernel functions. Ohn *et al.* (2004) compare three kernel functions (Inverse Multi-Quadric, Radial and Neural). In their experiments, Neural kernel performs the worst and the other two show the same efficiency. Howley and Madden (2005) shows percentage classification accuracy of SVM using RBF kernel, polynomial kernel and sigmoid kernel. It is indicated in experiments that RBF kernel performs worse that polynomial kernel of low degree of d and sigmoid kernel. Boolchandani *et al.* (2011) explore performance of several kernel functions (RBF, Polynomial, Power, Bias, Multipled and Log) in serveral datasets. In their results, RBF performs more poorly than other kernel functions. Sahak *et al.* (2012) mention in their study that from previous experiments exploring optimal parameters of kernel functions of SVM, the optimal regularization parameter and gamma for SVM were found to be 1 and 0.009. Chiroma *et al.* (2014) analyze crude oil price data using SVM. They compare performance of several kernel functions (RBF, Polynomial, Wave, Sigmoid and Exponential). Experimental results shows that performance wave kernel is the best.

This study compares percentage classification of SVM using four kernels (Linear, RBF, Polynomial and Sigmoid). Different to studies methioned above, linear case is considered in this study because a given dataset may be linear separatable and if this assumption is correct, computational complexity is reduced which results in less time to construct SVM models. Experimental results in this study show that data set used is actually linear separatable.

CONCLUSION

In this study, MDLP binning and SVM are applied to study residential hedonic price and proposes M-SVM model. First, residential characteristics and quantification technique are introduced and then this study gives brief introductions of MDLP binning and SVM. M-SVM models are built according to four different kernel functions with different parameters. By comparing constructed M-SVM models, it is the conclusion that linear kernel is the best for dataset in this study, polynomial kernel and RBF kernel better and the worst is sigmoid kernel.

ACKNOWLEDGMENTS

This research was supported by project of National Science Foundation, Grant No. 13BTJ005.

REFERENCES

Agarwal, D.K., 2002. Shrinkage estimator generalizations of proximal support vector machines. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada, pp: 173-182.

Boolchandani, D., A. Ahmed and V. Sahula, 2011. Efficient kernel functions for support vector machine regression model for analog circuits' performance evaluation. Analog Integr. Circuits Signal Process., 66: 117-128.

Butler, R.V., 1982. The specification of hedonic indexes for urban housing. Land Econ., 58: 96-108.

Can, A.S. and I. Megbolugbe, 1997. Spatial dependence and house price index construction. J. Real Estate Finance Econ., 14: 203-222.

Chiroma, H., S. Abdulkareem, A.I. Abubakar and T. Herawan, 2014. Kernel Functions for the Support Vector Machine: Comparing Performances on Crude Oil Price Data. In: Recent Advances on Soft Computing and Data Mining, Herawan, T., R. Ghazali and M.M. Deris (Eds.). Springer International Publishing, New York, ISBN: 978-3-319-07691-1, pp: 273-281.

Crocker, J., L.L. Thompson, K.M. McGraw and C. Ingerman, 1987. Downward comparison, prejudice and evaluations of others: Effects of self-esteem and threat. J. Personality Soc. Psychol., 52: 907-916.

Gu, J., M. Zhu and L. Jiang, 2011. Housing price forecasting based on genetic algorithm and support vector machine. Expert Syst. Applic., 38: 3383-3386.

Gu, Y. and L.T. Xia, 2013. The application of weighted Markov chain on forecasting in real estate price. J. Chongqing Univ. Technol. (Nat. Sci.), 27: 125-130.

Guo, W.G., X.M. Cui and H.Z. Wen, 2006. Hedonic price analysis of urban housing: The experiential research on the Hangzhou City. Econ. Geogr., 26: 172-176.

Han, J. and M. Kamber, 2007. Data Mining Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publishers, San Francisco, CA., USA.

Hao, P.Y. and J.H. Chiang, 2007. A fuzzy model of support vector regression machine. Int. J. Fuzzy Syst., 9: 45-50.

Hao, P.Y. and J.H. Chiang, 2008. Fuzzy regression analysis by support vector learning approach. IEEE Trans. Fuzzy Syst., 16: 428-441.

Hao, Q.J. and J. Chen, 2007. Distance to CBD, transportation accessibility and the spatial differences of residential housing prices in Shanghai. World Econ. Pap., 1: 22-35.

Howley, T. and M.G. Madden, 2005. The genetic kernel support vector machine: Description and evaluation. Artif. Intell. Rev., 24: 379-395.

Li, Z.F., J.W. Li and W. Ji, 2011. Discriminant analysis and prediction of house prices based on support vector machines. J. Hubei Normal Univ. (Nat. Sci.), 31: 60-65.

Lin, H. and K. Chen, 2011. Predicting price of Taiwan real estates by neural networks and support vector regression. Proceedings of the 15th WSEAS International Conference on Systems, July 14-16, 2011, Corfu Island, Greece, pp: 220-225.

Liu, Y. and H. Zhao, 2013. Study on city residential prices based on enjoy price model. J. Yuncheng Univ., 4: 17-20.

Lu, Q.H., 2012. Neural networks in commercial residential pricing-case of Hangzhou. J. Zhejiang Univ. Technol. (Soc. Sci.), 11: 337-341.

Ma, S.X. and A. Li, 2003. House price and its determinations in Beijing based on hedonic model. China Civil Eng. J., 36: 59-64.

Nello, C. and S.T. John, 2004. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Publishing House of Electronic Industry, China.

Ohn, S.Y., H.N. Nguyen and S.D. Chi, 2004. Evolutionary parameter estimation algorithm for combined kernel function in support vector machine. Proceedings of the Advanced Workshop on Content Computing, November 15-17, 2004, ZhenJiang, JiangSu, China, pp: 481-486.

Ozanne, L. and S. Malpezzi, 1985. The efficacy of hedonic estimation with the annual housing survey: Evidence from the demand experiment. J. Econ. Soc. Meas., 13: 153-172.

Sahak, R., Y.K. Lee, W. Mansor, A.I.M. Yassin and A. Zabidi, 2012. Optimal features for classifying asphyxiated infant cry using support vector machine with RBF kernel. Proceedings of the 5th European Conference of the International Federation for Medical and Biological Engineering, September 14-18, 2011, Budapest, Hungary, pp: 98-101.

Stitson, M.O., A. Gammerman, V. Vapnik, V. Vovk, C. Watkins and J. Weston, 1997. Support vector regression with ANOVA decomposition kernels. Technical Report, CSD-TR-97-22, Royal Holloway University of London, November 27, 1997.

Tan, P.N., M. Steinbach and V. Kumar, 2011. Introduction to Data Mining. Posts and Telecom Press, Beijing, China.

Wang, X.Y., 2006. Theoretical analysis of hedonic model of urban housing. Shanghai Manage. Sci., 4: 68-69.

Wen, H.Z. and S.H. Jia, 2006. Market segment and hedonic price analysis of urban housing. J. Zhejiang Univ. Technol., 36: 155-161.

Wu, X. and V. Kumar, 2013. The Top Ten Algorithms in Data Mining. Tsinghua University Press, China.

Xue, W. and H. Chen, 2010. Clementine Data Mining Methods and Application. Electronic Industry Press, China.

Zhang, M. and S.M. Chen, 2008. Shanghai real estate prices empirical analysis based on hedonic price theory. Financial Econ., 6: 72-74.