Support Vector Machine Prediction Modeling for Automobile Ownership

Ruidong Zhang, Xinguang Zhang

School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai, China
Email: zxg416@126.com

Abstract

To address some inherent defects of artificial neural networks, such as insufficient generalization performance, local extremum problem, and dimensional catastrophe problem, a support vector machine was proposed and applied to the modeling of automobile ownership prediction. By analyzing the data on automobile ownership and its influencing factors, the learning sample couples for automobile ownership prediction modeling were constructed, and support vector machine (SVM) was used to regression the nonlinear function relation of automobile ownership prediction model, and the established automobile ownership prediction model was used to predict the automobile ownership in different years. To reduce the impact on the accuracy of automobile ownership prediction caused by the large order of magnitude difference between the data of automobile ownership and its influence factors, the normalization method was used to pre-process the automobile ownership and its influence factors, and the inverse normalization was used to process the automobile ownership prediction results. The comparison between the automobile ownership prediction results and the statistical results shows that the automobile ownership prediction model has good generalization performance, and the support vector machine is an effective method to model the automobile ownership prediction.

Keywords

Automobile Ownership, Prediction, Support Vector Machine

1. Introduction

The accurate prediction modeling of China's automobile ownership can provide a reference for decision-making to deal with a series of urgent social problems such as traffic, environmental pollution, and energy caused by the growth of au-

tomobile ownership in China.

In predictive modeling, the mathematical model combined with computer simulation is mostly used nowadays, and the key to applying this method is to accurately determine the model parameters or functional relationship in the mathematical model.

Classical system identification methods, including the least-squares, Kalman filtering, etc., were applied in various fields. To decrease the modeling errors and enhance the control behavior in transients and steady-state, Anian Brosch used the recursive least squares forecasting approach [1] to conduct the current control. Shao *et al.* proposed the direct least-squares method [2] to address collinear data and big data. Fabio Tinazzi *et al.* described the historical current measurements using the recursive least-square algorithm [3] to predict the currency's future behavior corresponding to a limited set of voltage vectors. A least-squares [4] predictive model using discrete cosine transform was suggested to perform the prediction of hourly air quality index. The partial least squares regression model [5] was presented to develop the sea ice concentration prediction model. Xing *et al.* used the Kalman filter [6] based on the fractional-order model to perform the lithium-ion battery state of charge estimation. Mathieu Brunot presented the recursive estimator [7] that was of Kalman type for discrete-time systems characterized with measurement noise modeled from Gaussian uniform mixtures. Martincevic, A. and M. Vasak used the constrained Kalman filter [8] to identify the semi-physical building thermal models. Xiao *et al.* utilized the improved extended Kalman filter [9] to do the intelligent vehicle localization. The neuron-based Kalman filter [10] was developed to address the problems of identification of actual systems and conventional Kalman filter for empirical parameter estimation.

However, these classical methods have some inherent defects, such as the dependence on the mathematical model of the research object; the convergence and accuracy of the algorithm depending on the initial value of the parameter estimation, etc., which hinder the further development of traditional system identification methods and make their applications somewhat limited. Artificial neural network [11] makes up for the inherent defects of traditional system identification methods, does not depend on the mathematical model of the research object, has the characteristics of algorithm stability and accuracy with little dependence on the initial value of parameter estimation, and has strong nonlinear mapping ability, which can effectively describe the input-output response characteristic for the nonlinear dynamic system of automobile ownership prediction model. However, the artificial neural network also has insufficient generalization performance, local extremum problems, dimensional catastrophe problems, etc. As a new artificial intelligence method developed based on statistical theory and the principle of structural risk minimization, a support vector machine (SVM) [12] solves the local extremum problem which cannot be avoided in neural network methods and has a better generalization ability and solves the dimensional ca-

tastrophe problem skillfully. However, there is little literature on the application of support vector machines to vehicle ownership prediction modeling. The modeling used in this paper is novel, but not innovative.

The learning samples required for the automobile ownership prediction modeling are obtained from the statistical data on automobile ownership and its impact factors. The nonlinear functional relationship in the automobile ownership prediction model is regressed using a support vector machine, and the established automobile ownership prediction model is used to predict automobile ownership in different years. To reduce the impact on the accuracy of automobile ownership prediction due to the large difference in the order of magnitude between the input and output of learning samples, the statistical data on automobile ownership and its influence factors are normalized and the results of automobile ownership prediction are processed by the inverse normalization method. The comparison between the prediction results and the statistical data shows that the automobile ownership prediction model has high performance in generalization, and SVM is a useful method to construct the automobile ownership prediction model.

2. Research Method

SVM is mainly used for pattern recognition and function regression. When used for pattern recognition, it is called support vector classification (SVC) and when used for function regression, it is called support vector regression (SVR).

The algorithm of SVR [12] can be described as follows.

Step 1: Construct the training set

$$T = \left\{ (x_1, y_1), \cdots, (x_l, y_l) \right\} \in \left(\chi \times y \right)^l, \, x_i \in R^n, \, y_i \in R, \, i = 1, \cdots, l.$$

Step 2: Select the appropriate accuracy parameter ε, penalty parameter C, and kernel function $K(x, x')$.

Step 3: Solve the following optimization problem.

$$\min_{\alpha^{(*)} \in R^{2l}} \frac{1}{2} \sum_{i,j=1}^{l} \left(\alpha_i^* - \alpha_i \right)\left(\alpha_j^* - \alpha_j \right) K\left(x_i \cdot x_j \right) + \varepsilon \sum_{i=1}^{l} \left(\alpha_i^* + \alpha_i \right) - \sum_{i=1}^{l} y_i \left(\alpha_i^* - \alpha_i \right)$$

$$\text{s.t. } \sum_{i=1}^{l} \left(\alpha_i - \alpha_i^* \right) = 0 \tag{1}$$

$$0 \le \alpha_i^{(*)} \le \frac{C}{l}, \; i = 1, 2, \cdots, l$$

The optimal solution $\tilde{\alpha}^{(*)} = \left(\tilde{\alpha}_1, \tilde{\alpha}_1^*, \cdots, \tilde{\alpha}_l, \tilde{\alpha}_l^* \right)^{\mathrm{T}}$ is obtained from Eq. (1).

Step 4: Construct the decision function.

$$f(x) = \sum_{i=1}^{l} \left(\tilde{\alpha}_i^* - \tilde{\alpha}_i \right) K\left(x_i, x \right) + b \tag{2}$$

Select the $\tilde{\alpha}_j$ or $\tilde{\alpha}_k^*$ located in the open interval $\left(0, \frac{C}{l} \right)$. If $\tilde{\alpha}_j$ is chosen, then

$$b = y_j - \sum_{i=1}^{l} \left(\tilde{\alpha}_i^* - \tilde{\alpha}_i \right) K\left(x_i, x \right) \tag{3}$$

If $\tilde{\alpha}_k^*$ is chosen, then

$$b = y_j - \sum_{i=1}^{l} \left(\tilde{\alpha}_i^* - \tilde{\alpha}_i \right) K \left(x_i, x \right) \tag{4}$$

In the structural parameters for SVR, the accuracy parameter ε is selected with a certain rule, *i.e.*, as ε increases, the sparsity of the solution becomes larger and the fitting accuracy becomes smaller; instead, the smaller ε is, the higher the fitting accuracy is. When ε decreases to 0, the number of support vectors required for SVR reaches the maximum value, which is almost the same as the number of training samples, affecting the learning efficiency. With a larger value of the penalty parameter C, the more favorable the sample fitting accuracy is but reduces the confidence and generalization performance of the learning machine. Meanwhile, the larger the C value is, the greater the penalty is for the noisy sample, but eliminating too many samples will lead to distortion of the data regression, conversely, a smaller C value may not fit the samples well enough, which likewise causes the generalization performance of the learning machine to be reduced.

3. Results and Discussion

In this paper, the training sample pairs required for SVR for automobile ownership prediction modeling were constructed by using automobile ownership and its impact factor, and the statistical data on automobile ownership and its impact factor was obtained from the National Bureau of Statistics.

The difference in the order of magnitude between automobile ownership and its impact factors can seriously affect the accuracy of automobile ownership prediction modeling. Therefore, in the paper, the automobile ownership and its impact factor statistics were first normalized and pre-processed.

The impact factors affecting the automobile ownership are used as the learning sample inputs, including seven impact indicators such as gross national income, GDP per capita, total import and export, disposable income per urban resident, steel production, road passenger traffic, and total social consumer goods, and the automobile ownership was used as the learning sample output.
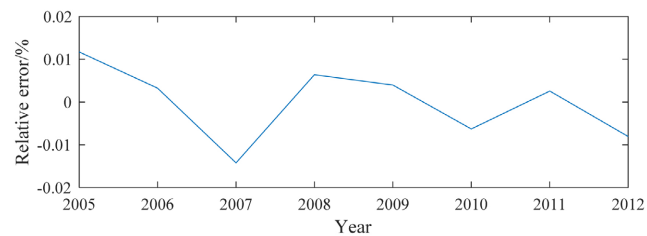
The statistical data on automobile ownership and its impact factors before and after normalization are shown in Table 1 and Table 2.

In the SVR structural parameters, the penalty parameter C is chosen as 10^5 and the precision parameter ε is chosen as 0. To confirm the generalization performance for the automobile ownership regression function constructed by SVR, a learning sample set is constructed from the automobile ownership and its impact factor data for the period from 1994 to 2004, while a verification sample set is constructed from the automobile ownership and its impact factor data for the period from 2005 to 2012.

After the inverse normalization of the automobile ownership prediction results, the predicted values of automobile ownership for the period from 2005 to 2012 are obtained, and the comparison of the predicted values and the statistical results is shown in Figure 1.
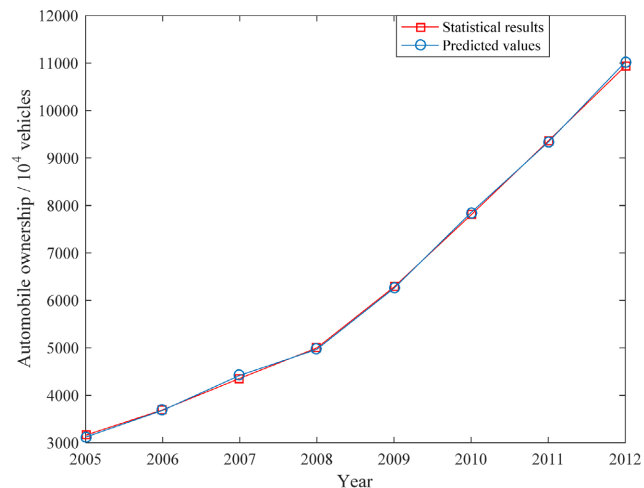
(a)

(b)

Figure 1. Automobile ownership prediction based on SVR. (a) Comparison between automobile ownership estimated values and statistical results; (b) Automobile ownership prediction error.

Table 1. Automobiles ownership and its impact factors on statistical data.

year	1994	…	2010	2011	2012
Car ownership/10^4 liang	942	…	7802	9356	10,933
Gross national income 10^8 yuan	48,108	…	399,760	468,562	516,810
Per capita GDP/yuan	4044	…	30,015	35,198	38,459
Total import and export/10^8 yuan	20,382	…	201,722	236,402	244,160
Per capita disposable income of urban residents/yuan	3496	…	19,109	21,810	24,565
Steel output/10^4 tons	8428	…	80,277	88,620	95,578
Highway passenger volume/10^4	953,940	…	3,052,738	3,286,220	3,557,010
Total retail sales of consumer goods/10^8 yuan	18,623	…	156,998	183,917	210,307

Table 2. Automobiles ownership and its impact factors on normalized statistical data.

year	1994	...	2010	2011	2012
Car ownership/10^4	−1.00	...	−0.99	−0.99	−0.99
Gross national income/10^8 yuan	−0.90	...	−0.74	−0.71	−0.71
Per capita GDP/yuan	−0.99	...	−0.98	−0.98	−0.98
Total import and export/10^8 yuan	−0.96	...	−0.87	−0.86	−0.86
Per capita disposable income of urban residents/yuan	−0.99	...	−0.99	−0.99	−0.99
Steel output/10^4 tons	−0.98	...	−0.95	−0.95	−0.95
Highway passenger volume/10^4	1.00	...	1.00	1.00	1.00
Total retail sales of consumer goods/10^8 yuan	−0.96	...	−0.90	−0.89	−0.88

As seen in Figure 1, the constructed automobile ownership prediction model is very well generalized and SVM presents an attractive tool for the automobile ownership prediction modeling.

4. Conclusion

In this paper, the SVR based modeling method for automobile ownership prediction was suggested, the learning sample set was constructed using the statistical results of automobile ownership and its influence factors, and the decision function for automobile ownership prediction was developed. The comparison of the automobile ownership prediction results with the statistical results confirms the SVR based automobile ownership prediction modeling method. However, there was no theoretical basis for the selection of penalty parameter C and accuracy parameter ε in SVR, and the arbitrariness in the selection of penalty parameter C and accuracy parameter ε will directly affect the accuracy of automobile ownership prediction modeling. Therefore, the next research work was to use the optimization algorithms including genetic algorithms, etc. for the optimization of penalty parameter C and accuracy parameter ε to improve the accuracy of SVR based automobile ownership prediction modeling.

Acknowledgements

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Brosch, A., Hanke, S. and Wallscheid, O. (2021) Data-Driven Recursive Least Squares

Estimation for Model Predictive Current Control of Permanent Magnet Synchronous Motors . *IEEE Transactions on Power Electronics*, **36**, 2179-2190. https://doi.org/10.1109/TPEL.2020.3006779

[2] Shao, Z.T., Zhai, Q.Z. and Wu, J. (2021) Data Based Linear Power Flow Model: Investigation of a Least-Squares Based Approximation. *IEEE Transactions on Power Systems*, **36**, 4246-4258. https://doi.org/10.1109/TPWRS.2021.3062359

[3] Tinazzi, F., Carlet, P.G. and Bolognani, S. (2020) Motor Parameter-Free Predictive Current Control of Synchronous Motors by Recursive Least-Square Self-Commissioning Model. *IEEE Transactions on Industrial Electronics*, **67**, 9093-9100. https://doi.org/10.1109/TIE.2019.2956407

[4] Yang, Z.C. (2020) DCT-Based Least-Squares Predictive Model for Hourly AQI Fluctuation Forecasting. *Journal of Environmental Informatics*, **36**, 58-69. https://doi.org/10.3808/jei.201800402

[5] Ye, X.C. and Wu, Z.W. (2021) Seasonal Prediction of Arctic Summer Sea Ice Concentration from a Partial Least Squares Regression Model. *Atmosphere*, **12**, 230-249. https://doi.org/10.3390/atmos12020230

[6] Xing, L.K., Ling, L.Y. and Gong, B. (2022) State-of-Charge Estimation for Lithium-Ion Batteries Using Kalman Filters Based on Fractional-Order Models. *Connection Science*, **34**, 162-184. https://doi.org/10.1080/09540091.2021.1978930

[7] Brunot, M. (2020) A Gaussian Uniform Mixture Model for Robust Kalman Filtering. *IEEE Transactions on Aerospace and Electronic Systems*, **56**, 2656-2665. https://doi.org/10.1109/TAES.2019.2953414

[8] Martincevic, A. and Vasak, M. (2020) Constrained Kalman Filter for Identification of Semiphysical Building Thermal Models. *IEEE Transactions on Control Systems Technology*, **28**, 2697-2704. https://doi.org/10.1109/TCST.2019.2942808

[9] Xiao, G.C., Song, X.L. and Cao, H.T. (2020) Augmented Extended Kalman Filter with Cooperative Bayesian Filtering and Multi-Models Fusion for Precise Vehicle Localisations. *IET Radar, Sonar & Navigation*, **14**, 1815-1826. https://doi.org/10.1049/iet-rsn.2020.0155

[10] Bai, Y.T., Wang, X.Y. and Jin, X.B. (2020) A Neuron-Based Kalman Filter with Nonlinear Autoregressive Model . *Sensors*, **20**, 299. https://doi.org/10.3390/s20010299

[11] Zhang, C., Guo, Y. and Li, M. (2021) Review of Development and Application of Artificial Neural Network Models. *Computer Engineering and Applications*, **57**, 57-69.

[12] Zhang, X.-G. and Zou, Z.-J. (2013) Estimation of the Hydrodynamic Coefficients from Captive Model Test Results by Using Support Vector Machines. *Ocean Engineering*, **73**, 25-31. https://doi.org/10.1016/j.oceaneng.2013.07.007