Tourism Demand Forecasting by Improved SVR Model

Li Mei
Department of Social Services, Zhengzhou Tourism College, Henan, P.R. China
hnlm2002@163.com

Abstract
The inboard tourism demand forecasting is very important to the development of tourism industry. In this paper, the SVR model is adopted to forecast monthly inbound tourism demand of China. And the elitist Non-dominated Sorting Genetic Algorithm (NSGAII) is used to parameter optimization. The NSGAII algorithm can reduce complexity of the algorithm, keeps the diversity of population and increasing the forecasting accuracy. At last, the proposed NSGAII-SVR model is used to forecast monthly inbound tourism demand of China from July 2011 to December 2011. And the experimental results show that the NSGAII-SVR has the best performance on forecasting compared with other models.

Keywords: Forecast, tourism demand, SVR, NSGA

1. Introduction
With the globalization of the economy and international exchanges continuous to deepen, the international tourism industry has been seen rapid development. Lots of countries have formulated policies to support and encourage the development of tourism industry. The tourism industry has irreplaceable function in the balance of international payments and trade structure improvement. It also plays an important role on expanding the opening to the outside world and promoting exchanges and cooperation with other countries. Therefore, in the past 20 years, the study of tourism industry got an unprecedented development. While tourism demands forecasting is the focus of the study.

The tourism demand forecasting have been studied by many researches [1-2]. Early tourism demand forecasting methods mainly use the econometrics models [3-4]. However, the econometrics models are time-consuming and difficult to determine the defect influence factor. The time series model does not have to seek inner relationship between tourism demand and influence factors compared with econometrics models. It seeks inner relations by the data of tourism demand itself. So in some ways, time series model is more suitable for tourism demand forecasting. The time series models include exponential smoothing model, Autoregressive model(AR), Autoregressive and Moving Average Model(ARMA), Autoregressive Integrated Moving Average Model(ARIMA) etc. These models have been successfully used in tourism demand forecasting[5-10]. In recent years, the artificial neural network model (ANN) is widely used in many research fields because of its good ability to handle complex nonlinear functions[11-15]. And lots of researches also adopted neural network in tourism demand forecasting[16-18].

However, the neural network model has their own shortcomings, such as slow training speed, low generalization ability, easy to fall into local optimum etc. Support vector machine(SVM) was proposed by Vapnik in 1995[19]. Support vector machine is a new algorithm which is based on the principle of structural risk minimization as the foundation. It is better than the other algorithms which are based on the principle of empirical risk minimization. At the same time, it can guarantee the extremal solution is global optimal solution because it is a convex secondary optimization problem. SVM can solve the problems which characteristics are small sample, non-linear, high dimension and local
minima. Qiao et al.[20] proposed a Bayesian evidence framework to infer the LS-SVR model parameters. The results show that the Bayesian framework of an LS-SVR significantly improves the speed and accuracy of the forecast. Hong et al. [21] employs GA-SA algorithm to choose the suitable parameter combination for a SVR model and used the SVRGA–SA model to forecast Taiwanese 3G mobile phone demand. He et al.[22] used random sampling method to improve SVM model and discussed commercial bank customer churn prediction based on SVM mode. Ch. et al [23] investigated the accuracy of the hybrid SVM-QPSO model (support vector machine-quantum behaved particle swarm optimization) in predicting monthly streamflows. And it is employed in forecasting the streamflow values of Vijayawada station and Polavaram station of Andhra Pradesh in India.

In this paper, the overall structure of the study is as follows: In section 2, we first introduce the standard SVR model. In section 3, Non-dominated Sorting Genetic Algorithm(NSGA) is presented and the NSGAII-SVR model is proposed. In section 4, we establish the optimized model and use it to predict inboard tourism demand. At last, the conclusion is given in section 5.

2. The Principle of SVR

Support vector machine was originally used for pattern recognition problems and it is mainly applied to classification. With the introduction of insensitive loss function by Vapnik, Support vector regression (SVR) is extended to solve the nonlinear regression problem. The principle of SVR model is introduced as follows.

Set a sample set \( S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \). Through a nonlinear mapping \( \phi \), the data from sample set \( S \) are mapped into a high dimensional space \( F \). And the following function is used for linear regression in \( F \) space.

\[
f(x) = w \cdot \phi(x) + b, \phi: R^n \rightarrow F, w \in F \quad (1)
\]

Where \( b \) threshold value is \( w \) is regression coefficient vector. The influence factors of \( w \) are the sum of empirical risk and the flatness of the high-dimensional space \( \|w\| \), that is

\[
R(w) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} \varepsilon(f(x_i) - y_i) \quad (2)
\]

\[
\varepsilon(f(x_i) - y_i) = \begin{cases} 0, & |f(x_i) - y_i| < \varepsilon \\ |f(x_i) - y_i|, & |f(x_i) - y_i| \geq \varepsilon \end{cases} \quad (3)
\]

Where \( m \) the number of samples is \( \varepsilon(f(x_i) - y_i) \) is loss function. \( C \) is penalty factor.

In order to control the complexity of function, linear regression function should be brought to flat as far as possible. The relaxation factors \( \xi \) and \( \xi^* \) are also introduced considering the regression error that beyond the precision. The relaxation factors are used to deal with the points which do not satisfy the Eq. (3). Based on structural risk minimization criterion of statistical learning theory, the variable of \( w \) and \( b \) are determined through support vector regression method by minimizing the objective function \( R(w) \):

\[
R(w, \xi, \xi^*) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*)
\]

\[
\begin{cases} y_i - w \cdot x_i - b \leq \varepsilon + \xi_i \\ y_i - w \cdot x_i - b \geq -\varepsilon - \xi_i^* \\ w \cdot x_i + b - y_i \leq \varepsilon + \xi_i \\ w \cdot x_i + b - y_i \geq -\varepsilon - \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \quad (4)
\]
In the Eq. (4), \( \frac{1}{2} \| w \|^2 \) makes the regression function more flat and have better generalization ability. The second part of Eq.(4) is used to reduce the error. The penalty factor \( C \) is a constant number and \( C > 0 \). It is used to control the degree of punishment of samples beyond the error \( \varepsilon \). Then establish Lagrange’s equation:

\[
L(w, \xi, \xi^\ast) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^\ast) - \sum_{i=1}^{m} \alpha_i [ (\varepsilon + \xi_i) + y_i + w \cdot \phi(x_i) + b] - \sum_{i=1}^{m} \alpha_i^\ast [ (\varepsilon + \xi_i^\ast) + y_i + w \cdot \phi(x_i) - b] - \sum_{i=1}^{m} (\lambda_i \xi_i + \lambda_i^\ast \xi_i^\ast)
\]

(5)

In order to make the Eq. (5) be minimum, function \( L(w, \xi, \xi^\ast) \) of all parameters of partial derivative is equal to zero. Then the following dual optimization problem can be get.

\[
\min \left\{ \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\alpha_i - \alpha_i^\ast)(\alpha_j - \alpha_j^\ast) [\phi(x_i) \cdot \phi(x_j)] + \sum_{i=1}^{m} \alpha_i (\varepsilon - y_i) + \sum_{i=1}^{m} \alpha_i^\ast (\varepsilon + y_i) \right\}
\]

(6)

\[
s.t. \sum_{i=1}^{m} (\alpha_i - \alpha_i^\ast) = 0
\]

\[
\alpha_i, \alpha_i^\ast \in [0, C]
\]

Then, the support vector regression problems can be summed up in quadratic programming problem. By solving the quadratic programming problem Eq. (6), we can get the function \( w \) which is described by the training sample points.

\[
w = \sum_{i=1}^{m} (\alpha_i - \alpha_i^\ast) \phi(x_i)
\]

(7)

Where \( \alpha_i \) and \( \alpha_i^\ast \) is the solution of \( R(w, \xi, \xi^\ast) \). Then the regression function can be got. That is

\[
f(x) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^\ast) k(x, x_i) + b
\]

(8)

Where \( k(x, x_i) = \phi(x) \cdot \phi(x_i) \) is the kernel function Different SVR models can be got by selecting different kernel functions. In this paper, the radial basis function (RBF) is chosen in SVR model. The RBF is shown in bellow

\[
K(x, x_i) = \exp(-\gamma |x - x_i|^2)
\]

(9)

**3. Improved SVR Model**

The three parameters \( \varepsilon, C, \delta \) will have a major impact on the forecasting results of SVR model through the principle of SVR algorithm. Thus, the improved Non-dominated Sorting Genetic Algorithm (NSGA) will be used to parameters optimization.

**3.1 NSGA Algorithm and its Improvements**

Non-dominated Sorting Genetic Algorithm (NSGA) is proposed by Srinivas and Deb [24] in 1995. It is a genetic algorithm based on pare to optimal concept. It can embody Goldberg thought best from these multi objective genetic algorithm optimization methods. The NSGA improves the choice of regeneration based on the basic genetic algorithm. Each
individual is stratified in accordance with their dominated and non-dominated relations. Before the selection operation, the population has been classified and sorted according to the dominated and non-dominated relations. And all the individuals of the population are assigned a virtual fitness value. The individuals at the same level have same virtual fitness values. This will ensure that the individuals of same level have the same probability of replication. Then doing a selection operation, so that the algorithm are obtained very satisfactory results in the multi-objective optimization. The characteristics of NSGA are turning multiple objective function calculation into virtual fitness calculation [25]. The flow chart of NSGA is shown in Fig. 1.

NSGA retains the excellent individuals by the non-dominated sorting algorithm and uses fitness sharing function to keep the diversity of the population. But NSGA also has obvious shortcomings, mainly be reflected in two aspects: 1. High computational complexity; 2. The sharing radius needs to be specified in advance.

In order to overcome the shortcomings of NSGA, Deb proposed the Elitist Non-dominated Sorting Genetic Algorithm (NSGAII). The algorithm mainly improved on NSGA in the following three aspects:

(1) The complexity of the algorithm is reduced by a proposed a fast non dominated sorting algorithm based on the classification.

(2) Put forward congestion and the congestion comparison operator. The fitness sharing strategies do not need to specify the sharing radius. As the winning standards after sorting between same classes, the pareto domain of the elements can be extended to the entire pareto domain. The distribution is uniform and keeps the diversity of population.

(3) Introduce the elitist strategy and increase the sampling space. The population of the next generation is got by the competition of parent population and its progeny population in order to get better next generation.

The detailed process of NSGA II algorithm is shown in bellow.

(i) Fast non-dominated sorting algorithm. \( n_p \) is the number of dominant individual \( p \) in population. \( S_p \) is the set that controlled by individual \( p \).

Step 1: Find the individual which is \( n_p = 0 \) in the population. And they are stored in the current set \( E_1 \).

Step 2: For each individual \( i \) in the current set \( E_1 \), the dominating set for individual is \( S_i \). Travel each individual \( l \) from \( S_i \). Let \( n_l = n_i - 1 \), if \( n_l = 0 \), the individual \( l \) is stored in set \( H \).

Step 3: Let the individual got from \( E_1 \) be the first individual of non-dominated layer. Let \( H \) be the current set, repeat the above steps until the entire population is classified.

(ii) The determination of the degree of congestion.

Step 1: Let \( n_d = 0 \), \( n = 1, 2, \ldots, N \) \( n_d \) is the degree of congestion.

Step 2: For each objective function

a. The population is sorted based on the objective function.

b. Let the degree of congestion of boundaries of two individuals crowded be infinite. That is \( l_d = N_d = \infty \).

c. Calculate the updating value \( n_d \).

Step 3: The comparison of the degree of congestion. After a fast non-dominated sorting algorithm and congestion degree calculation, each individual in the population has two attributes: the non-dominated order \( n_{\text{rank}} \) and congestion degree \( n_d \). The degree of congestion comparison operator is defined as \( \geq_{\text{con}} \). If \( i \geq_{\text{con}} j \), let individual \( i \) is better than individual \( j \), if and only if \( i_{\text{rank}} \leq j_{\text{rank}} \) any \( i_d > j_d \).
Finally, we give the calculation steps of NSGA II algorithm.

Step 1: Randomly generate initial population $P_0$. Then non-dominated sorting is used to the population. Each individual is assigned with pareto rank. Then do selection, crossover and mutation operation of initial population and get a new population $Q_0$, let $t = 0$.

Step 2: In the generation iteration of $t$, form new groups $R_t = Q_t \cup U_t$. The non-dominated sorting is used to the population $R_t$. Get the non-inferior front $E_1, E_2, \ldots$.

Step 3: Calculate all the degree of congestion of non-inferior front $E_t$.

Step 4: Do partial order selection, choose the best $N$ individuals to form population $P_{t+1}$.

Step 5: Do selection, crossover and mutation operation to population $P_{t+1}$ and get population $Q_{t+1}$.

Step 6: Determine whether the termination condition reaches. If doesn’t, turn to Step 2. Then output results.

![Figure1. The Flow Chart of NSGA](image-url)
3.2 NSGAII-SVR Model

The parameters ε, C, δ play an important role in the SVR model. How to get the parameters accurately is very important. And in this paper, NSGAII algorithm is used in parameters optimization and the NSGAII-SVR model is established in order to forecast tourism demand of China.

The flow chart of NSGAII-SVR is shown in Fig 2 and the basic steps of NSGAII-SVR algorithm is as follow:

Step 1: Experimental design. Select the important influential variables as the design variables. Choose the experimental design method and design training samples.

Step 2: Optimize the parameters of SVR model.

Step 3: Establish SVR model. The SVR model is established through training samples and optimized parameters. If approximation accuracy of the SVR model can not meet the requirements, each iteration optimization results are chosen as a new training sample. Then the new training samples are used to update model and increase the accuracy.

Step 4: Multi-objective optimization of NSGAII. The NSGAII algorithm is used to optimize SVR model and design space exploration. Using the cross validation method to optimize NSGA II-SVR model. The quality of pareto solution set is evaluated by uniformity and diversity of distribution of pareto solution set. If it satisfies the requirement, then output the pareto solution set. If it doesn’t, the population and the maximum number of iterations of evolution are increased. Then turn to step 3.

Step 5: According to the preference information, select the optimal solution.

![Figure 2. The Flow Chart of NSGAII-SVR](image-url)
4. Model Construction and Prediction

In this paper, the NSGAII-SVR model is established to forecast the monthly inbound tourism demand of China. The data is chosen from January 2000 to December 2011. The unit of inbound tourism demand of China is ten thousand person-times. The forecasting data is from July 2011 to December 2011. The impact factor is shown in Table 1.

<table>
<thead>
<tr>
<th>Impact factors</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDP per capita of passenger source</td>
<td>Yuan</td>
</tr>
<tr>
<td>The disposable income of residents of passenger source</td>
<td>Yuan</td>
</tr>
<tr>
<td>Exchange rate(Dollar against the RMB)</td>
<td>Dollar/RMB</td>
</tr>
<tr>
<td>The ratio of CPI</td>
<td>%</td>
</tr>
<tr>
<td>The conditions of tourism resources</td>
<td>Score</td>
</tr>
</tbody>
</table>

(1) GDP per capita of main passenger source.

The main passenger source refers to the countries which their inbound tourists account for more than 80% of the total inbound tourists. Calculating formula of GDP per capita of main passenger source is

\[
X_{t_i} = \frac{\sum M_i}{\sum N_i}
\]  

(10)

\(M_i\) is the GDP of main passenger source. \(N_i\) is the population of main passenger source.

(2) The disposable income of residents of main passenger source.

The calculating formula is

\[
X_{z_{2i}} = \frac{\sum C_i N_i}{\sum N_i}
\]  

(11)

\(C_i\) is the income per capita of main passenger source. \(N_i\) is the population of main passenger source.

(3) Exchange rate (Dollar against the RMB).

The exchange rate is \(X_{x_{20}}\). It means the number of RMB that one dollar can be exchanged.

(4) The ratio of CPI.

\[
X_{a_{4i}} = \frac{CPI_i}{CPI_{2000}}
\]  

(12)

\(CPI_i\) is consumer price index at time \(t\).

(5) The conditions of tourism resources.

\[
X_{s_i} = \sum L_i P_i 
\]  

(13)

\(L_i\) is the number of different types of tourism resources. \(P_i\) is the score of different types of tourism resources. The score of different types of tourism resources is shown in Table 2.
Table 2. The Score of Different Types of Tourism Resources

<table>
<thead>
<tr>
<th>Types of tourism resources</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>5A tourist scenic spots</td>
<td>10</td>
</tr>
<tr>
<td>4A tourist scenic spots</td>
<td>5</td>
</tr>
<tr>
<td>3A tourist scenic spots</td>
<td>1</td>
</tr>
</tbody>
</table>

In order to eliminate the influence on the prediction results of unit, the 0-1 standardized method is adopted. That is

\[
\tilde{x} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}
\]  

(14)

Fig. 1 shows the forecasting curves of NSGAI-SVR model, original SVR model and BP-NN model. From this figure, we can’t obviously find out which model has the best performance on forecasting the monthly inbound tourism demand of China since these three models have similar forecasting curves. Then we calculate the root mean squared error (RMSE) and mean absolute error (MAE) to compare the forecasting accuracy of different models. From Table 3, we can clearly see that the NSGAI-SVR model has the best performance on RMSE and MAE for forecasting the monthly inbound tourism demand of China compared with the original SVR model and BP-NN model.

Figure 3. The Forecasting Curves of Different Models

Table 3. The Performance of Different Models

<table>
<thead>
<tr>
<th></th>
<th>NSGAI-SVR</th>
<th>Original SVR</th>
<th>BP-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>0.019374932</td>
<td>0.023328432</td>
<td>0.0273087</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.006521143</td>
<td>0.009454236</td>
<td>0.0114357</td>
</tr>
</tbody>
</table>

Where \( RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \), \( MAE = \frac{1}{n} \sum_{i=1}^{n} \left| \hat{y}_i - y_i \right| \)

The Table 4 shows the absolute error of NSGAI-SVR model and the compared models from July 2011 to December 2011. In this table, we can find that these three models could give accuracy forecasting. But the NSGAI-SVR model almost performs best in each month compared with the other two models in this situation. So the proposed NSGAI-SVR
can increase the forecasting accuracy for the monthly inbound tourism demand of China.

<table>
<thead>
<tr>
<th></th>
<th>NSGAII-SVR</th>
<th>Original SVR</th>
<th>BP-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011.07</td>
<td>-0.027504845</td>
<td>-0.038273892</td>
<td>-0.044334</td>
</tr>
<tr>
<td>2011.08</td>
<td>-0.015558515</td>
<td>-0.017608192</td>
<td>-0.02131</td>
</tr>
<tr>
<td>2011.09</td>
<td>0.017392872</td>
<td>0.0142207</td>
<td>0.0208174</td>
</tr>
<tr>
<td>2011.10</td>
<td>-0.028918981</td>
<td>-0.03240422</td>
<td>-0.038338</td>
</tr>
<tr>
<td>2011.11</td>
<td>0.011320554</td>
<td>0.012705657</td>
<td>0.0102018</td>
</tr>
<tr>
<td>2011.12</td>
<td>-0.015553827</td>
<td>-0.024757933</td>
<td>-0.028851</td>
</tr>
</tbody>
</table>

5. Conclusion

In this paper, the SVR model is adopted to forecast the monthly inbound tourism demand of China. The elitist Non-dominated Sorting Genetic Algorithm (NSGAII) is used to search the optimal parameters of SVR model. Thus, the nonlinear and high volatile of monthly inbound tourism demand problem can be solve by SVR model while the parameters optimization of SVR model can be solved by NSGAII algorithm. And the experimental results show that the NSGAII-SVR model is better than the compared models on forecasting monthly inbound tourism demand of China.

References


Authors

Li Mei, 1974.10, Zhengzhou, Henan, P. R. China. She is a lecturer of Department of Social Services, Zhengzhou Tourism College, Henan, P.R. China. Her scientific interests are tourism Management, Hotel Management, Golf Club Management.