Establishing of Passengers Forecasting Model between the Cross-Strait Using Support Vector Regression, Genetic Programming, and Exponential Smoothing

Shih-Huang Chen, Meng-Lung Chung, Wei-Yu Chen, Yi-Ting Chao, and Shyuan Lin
Dept. of Transportation Technology and Management, Feng Chia University, Taichung, 407, Taiwan,
Email:shc@mail.fcu.edu.tw, irene78130@gmail.com, {toro-1228, jpcj8, a309532004}@hotmail.com

Abstract—Demand Forecasting is a very important index to make the management policy in the future. This study focused on the accuracy of forecasting on the number of passengers between the cross-strait in the future. It utilized both a traditional method and an artificial intelligence method, and compared various forecasting performance. This study adopted methods of support vector regression, genetic programming, and exponential smoothing model to establish the forecasting model of passengers between the cross-strait. The MAPE value would be utilized to be the assessing index on the model. By comparing SVR, GP, and Holt model; the MAPE value in SVR forecasting passengers between the cross-strait shown the lowest value and it was under 10%. It represented ideal ability to forecast the future passengers’ capacity. For the verification result from January to July, 2010, the average deviation was only 9%. Therefore, this study has adopted the method of SVR to construct an appropriate forecast model on passengers between the cross-strait. The model could accurately forecast the number of passengers on the required market in the future.

Index Terms—forecast, SVR and GP

I. INTRODUCTION

In recent years, the development of mainland China expanded rapidly. Interactions, such as economic activities, work and academic affairs, as well as family relative visits, between the cross-strait have increased by days. Hence the direct cross-strait flight has a vital role in the transportation development in Taiwan. This study collected the data on total number of passengers that entered and exited the border between the cross-strait in these few years, and discovered a gradual increasing trend on passengers. In future, the study may integrate and reform with air and marine transportation, and develop a diverse and innovative operating model that could be used to reduce time and cost to accommodate the increasing passengers between the cross–strait. The one-day living circle between the cross-strait was beginning to form. The model may only determine an accurate forecast for the future demand; it could also be adjusted in accordance with major infrastructures through holiday period and peak seasons. Therefore, promoting the continual development and increasing the operative efficiency of air and marine transportation between the cross-strait, this study focused on the accuracy of forecasting on the number of passengers between the cross-strait in the future. It utilized both a traditional method and an artificial intelligence method, and compared various forecasting performance. The research scope was on the monthly number of passengers that entered and exited the border between the cross-strait from 2000 to 2009. By conducting comparison on various forecasting performance through various model of practice and verification, the obtained results were analyzed to determine the accuracy level of each model. It would then adopt the Mean Absolute Percent Error (MAPE) to be the assessment model standard of practical value and forecast value, and obtain the best performance model in order to conduct estimation on the future number of passengers between the cross-strait. The final result would be the trend of future variation of passengers; it would then be compared with previous data and conduct analysis, which could verify more on the accuracy of the number of passengers forecasted between the cross-strait in the future.

Tripathi et al. (2006) proposed that by adopting the support vector machine (SVM) to establish the scale model of monthly rainfall, the result have shown that the SVM model under the statistical base, it could provide weather impact researches on hydrology aspects. Cao et al. (2003) adopted the SVM and back-propagation network as forecasting tool to research five types of commodity future contract in the Chicago Board of Trade. The difference comparatives were the Normalized Mean Square Error (NMSE), Weighted Directional Symmetry (WDS), and Mean Absolute Error (MAE). It was discovered that the forecast performance of SVM was better that the back-propagation network. Chen (2006) proposed a new model of Genetic Algorithms combined Support Vector Regression (GA-SVR). It utilized the practical value genetic algorithms to find the most optimal parameter combination in the SVR. The model have collected the closing prices of the Taiwan Stock
Exchange Capitalization Weighted Stock Index from the period of 30th of October, 2001, to 9th of August, 2002; which totaling of 190 entries. Lastly, the result of model data and traditional model of Random Walk, as well as the forecast result of back-propagation network model, had all been compared. The outcome shown that forecasting accuracy of SVR was better than other two forecasting models, and the MAPE value reached 1.501125% ; it also verified the accuracy of the model forecasting. Chen et al. (2010) have adopted the data of marine container freight handling capacity from main ports in Taiwan for the past 30 years to seek for the most optimal forecasting model. The utilized methods were the genetic programming (GP), X-11(Decomposition Approach), and SARIMA. The MAPE value on result of these three methods was 2.28%, 3.6%, and 3.41%; which were all within 4% and had ideal forecasting ability. By adopting ANOVA table to conduct the analysis, the forecasting result of GP and other two methods have significant difference. Among the three methods, the MAPE value of GP was 2.28%, which was the lowest. Therefore, such study has selected GP to establish the forecast of marine container freight handling capacity in Taiwan.

II. RESEARCH PLAN

This study adopted methods of support vector regression, genetic programming, and exponential smoothing model to establish the forecasting model of passengers between the cross-strait. MAPE value would be utilized to be the assessing index on the model. The three research methodologies adopted in this study would be introduced separately in the following:

A. Support Vector Machine used for Regression

The Support Vector Machines (SVM) was proposed by Vapnik and AT&T laboratory team. It was a machine learning approach that based on the statistical learning theory (Vapnik, 1995:1998). In recent years, the SVM have been widely adopted on problems of regression analysis and movement of forecasting trends. In 1997, Vapnik had lead-in the SVM into the ε-Insensitive Loss Function, allowing the SVM to extend its solving ability on problems of non-linear regression (Vapnik, 1997). Such new forecasting approach was been referred as the Support Vector Regression (SVR). SVR was the composition of a training system and testing system. Initially, the data was lead-in into the training system to conduct system learning. Then, the input value and output value would substitute into the training system to be the Supervised Learning. After the training learning have completed, a set of training target would be obtained. From such target set, it could establish the predictive function and testing system. As the forecasting was required, it could obtain the predict value by only requiring to lead-in another set of unseen data into the testing system. As the SVM been used to estimate the regression equation, there were three unique characteristics: 1, as the SVM was used to estimate regression, the definition of linear equation system in a higher dimensional space was used. 2, the structure of SVM followed the minimization principle to achieve the ability of regression estimation. Such structural risk has adopted the ε-Insensitive Loss Function by Vapnik to conduct the estimation. 3, SVM used the penalty factor from experience error deviation and structural risk minimization principle to composite the risk function. When data were non-linear, the solution was to map the data to a higher dimensional feature space through kernel function, and seek for an optimal classification of hyperplane. Actual calculation of ψ(x) was not required in the regression process; if learning in feature space was required, it only required training the sample through inner product training method and obtaining the inner product value from the calculation. The direct mapping of data to the feature space was not required. Such decision function became: as for the kernel function type selection, there were no relative literature reviews to conclude on which type would perform better results on the SVM. Therefore, by establish an appropriate network model and seek for the optimal forecast and classification result, the selected kernel function would not be very important.

B. Exponential Smoothing Model

The exponential smoothing model was a form of weighted moving average, which reviewed the contribution of every phase of historical data as time changes. It had a tapering exponential relationship. The model was combined under the concept of data decomposition; the advantage was that the weighted-form could easily transform to incorporate with specific requirements. By conducting value forecasting from previous periodic data, the value would adjust in accordance with the deviated error of the previous forecasted value. The size of smoothing constant (α) would determine the forecasting response on deviation error level of the previous forecast, which α has to be in between 0–1. In general, 0.2–0.5 was commonly applied smoothing constant. These values had shown that the current forecast should accord with previous forecasted deviation error by adjust the value from 20% to 50%. As the constant value became larger, the generated response would be quicker. However, it might also generate abnormal norms. As for the smaller constants, the generated forecast value might show significant lagged results. Exponential Smoothing Model would analyze the characteristics and research development of the data, and determine whether the data possessed tendency and periodic cycles. It further divides the model into three categories: the first category assumed the data does not have influence of linear tendency and seasonality. It was referred as the Single Exponential Smoothing Model. The second category includes horizontal seasonal influence, but not the tendency influence. It was referred as the Horizontal Seasonal Exponential Smoothing Model. The third category includes both the seasonal and liner influence, such as the Holt-Winter Model. For most of requirement, it required type generally affected by tendency and seasonal influence. Such tendency often rise or drop under stable rate. As for Holt model, the characteristic did not include seasonality, but it had
tendency requirement. The equation was established to be (Chen, 2006)

\[ D_t = at + b \]

Future estimation value:

\[ F_{t+1} = L_t + T_t \]

\[ F_{t+n} = L_t + nT_t \]

After the requirement of observation period \( t \), the adjusted level and estimation value of tendency:

\[ L_{t+1} = aD_{t+1} + (1 - \alpha)(L_t + T_t) \]

\[ T_{t+1} = \beta(L_{t+1} - L_t) + (1 - \beta)T_t \]

\( D_t \) : the observed actual requirement in period \( t \)

\( F_t \) : the required estimation value in period \( t \) (in period \( t-1 \) or earlier estimation)

\( L_t \) : the estimate value of level as period \( t \) ends.

\( T_t \) : the estimate value of tendency as period \( t \) ends.

\( \alpha \) : \( 0 < \alpha \leq 1 \) level of smoothing constant

\( \beta \) : \( 0 < \beta \leq 1 \) tendency of smoothing constant

C. Genetic Programming

The Genetic programming (GP) was proposed by Koza (1992). The basic concept was followed from the Genetic Algorithm. The calculation structure includes selection, duplication, mating, mutation, and adaptation assessment. Initially, the computer program would select a better proposal from the mother trend. The duplication would select the most adaptable individual, and duplicated into a most optimal computer program. The aim was to expect an excellent gene in the mother chromosome could prolong to the next generation. The mutation was to prevent chromosome fall into partial optimal solution, or refrain from unrestricted expansion of chromosome layers. After reproduction and mating, the daughter gene would replace the mother gene. For the new individual among every generation, it should conduct adaptation assessment in order to assure the problem-solving ability of the individual. The aim of Genetic Programming was to hope that, under the condition of not providing detailed programming instructions, the computer could learn to solve problem autonomously. In the genetic programming, every individual was referred as chromosome, and the genetic value of every chromosome was generated by randomization. The chromosome cluster formed in every generation was referred as population. Among the generation, every individual competed with each other. A more suitable environment would have better adaptability, and the chromosome with better adaptability could choose to evaluate to a next generation individual. It could then choose to mate in order to produce next generation, and expect that the next generation would have higher adaptability.

III. RESEARCH FINDING

A. SVR Result

This study collected the monthly data of passengers between the cross-strait. The research period was from Jan 2000 to Dec 2009, totaling of 120 entries to conduct training (Fig. 1). The SVR adopted Matlab SVM and Visual Studio C++ software to build an appropriate forecasting model on passengers between the cross-strait, and utilize MAPE value assessment standard to determine the effectiveness of the model. For MAPE value that was smaller than 10% of the forecasting result, it represented that the model has high accurate forecasting ability. It was acknowledged that by utilizing SVR to train on number of entry passengers, the MAPE value reached 3.19%, whereas the MAPE value for number of exit passengers reached 2.45%; both had shown high accurate forecasting result. The Fig. 2 and Fig. 3 of this study were the comparison of forecasted values and actual number of passengers entering and exiting the border.

Figure. 1. Passengers no. between the Cross-Strait

Figure. 2. Forecasting of SVR and actual from China to Taiwan

Figure. 3. Forecasting of SVR and actual from Taiwan to China
B. Holt Mode Result

A lot of time series analyses were established on stable series, however, the data collected for this study had the tendency to increase over time. It also has shown norm of instability. Therefore, this study made adjustment to the source data, allowing the data to be stable series and increased the future accuracy of the forecasting movement. Holt-Winter was the demand forecasting model of season and tendency. This study adopted the SPSS13.0 Trends software to analyze and forecast the number of required passengers between the cross-strait in the future. For forecasting the passengers entering and exiting the border, results of Holt model shown that the MAPE value were 38% and 10%. It possessed rational ability to forecast the number of required passengers between the cross-strait in the future.

C. GP Result

By utilizing 120 entries of monthly data of passengers between the cross-strait to be the object of the empirical analysis, through multiple verification and adjustments on the data, finally, ¼ of the data entries were selected as samples of training data. The rest ¾ of data samples would be used as verification data. Before conducting of GP, settings of system parameters had to be considered, such as the group size, adaptability function, mutation ratio, mating ratio, and parsing tree height. However, the system would automatically search the most optimal evaluate parameters. When data became convergent or generates oscillation, it could terminate evolution. By knowing that the R2 training and data verification were 0.7, the model possessed rational explanation ability and had not shown norm of over-training. In the testing aspect, for the number of passengers between the cross-strait over the years, the MAPE of forecast value and actual value have reached 13%; which possessed ideal forecasting ability.

IV. DISCUSSION

By using 120 entries of monthly data on numbers of passengers between the cross-strait from 2000 to 2009, SVR, GP and Holt model were constructed. This study aims to find the most optimal model and calculate the MAPE value from the actual number of passengers and forecasted value in 2009. For example regarding border exiting passengers, the MAPE value from SVR forecasting model reached 2.3%, the MAPE value from Holt mode reached 10.7%, and GP have indicated 5%. From the training result of Fig. 4, despite of February and September has higher deviation of SVR model due to Chinese New Year Festival and Summer holiday period, it was observed that other months have high accurate forecasting ability to forecast passengers between the cross-strait. Therefore, the selection of SVR model would only reflect the closeness towards the actual value. Other than that, this study applied the Airbus A330-300 (295 seats) and B747-400 (416 seats) to compare the deviation between the various forecasting models and actual numbers of flight passengers in 2009. In Table I, SVR model and actual numbers of flight passengers only deviated to about 4~5 planes. It indicated that not only the result would increase the efficiency of airline passengers’ carriage ratio significantly, SVR model was also able to precisely forecast the future tendency of passengers’ capacity between the cross-strait.

To further illustrate, Table I reveals a comparison of the various methods applied in forecasting the number of passengers between the cross-strait. Holt-Winter model (Holt) achieved a 2.3% MAPE, while SVR model achieved a 3.7% MAPE, and GP model achieved a 5.1% MAPE. The Holt-Winter model was selected for further analysis, as it demonstrated the lowest MAPE value.

By viewing the source data, it was apparent that the MAPE of forecast value and actual numbers of flight passengers in 2009 were 38% and 10%, respectively. It possessed rational forecasting ability to forecast the number of required passengers between the cross-strait in the future.

<table>
<thead>
<tr>
<th>Month</th>
<th>SVR</th>
<th>Holt</th>
<th>GP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>720</td>
<td>722</td>
<td>78</td>
</tr>
<tr>
<td>Feb</td>
<td>724</td>
<td>726</td>
<td>79</td>
</tr>
<tr>
<td>Mar</td>
<td>728</td>
<td>730</td>
<td>80</td>
</tr>
<tr>
<td>Apr</td>
<td>732</td>
<td>734</td>
<td>81</td>
</tr>
<tr>
<td>May</td>
<td>736</td>
<td>738</td>
<td>82</td>
</tr>
<tr>
<td>Jun</td>
<td>740</td>
<td>742</td>
<td>83</td>
</tr>
<tr>
<td>Jul</td>
<td>744</td>
<td>746</td>
<td>84</td>
</tr>
<tr>
<td>Aug</td>
<td>748</td>
<td>750</td>
<td>85</td>
</tr>
<tr>
<td>Sep</td>
<td>752</td>
<td>754</td>
<td>86</td>
</tr>
<tr>
<td>Oct</td>
<td>756</td>
<td>758</td>
<td>87</td>
</tr>
<tr>
<td>Nov</td>
<td>760</td>
<td>762</td>
<td>88</td>
</tr>
<tr>
<td>Dec</td>
<td>764</td>
<td>766</td>
<td>89</td>
</tr>
</tbody>
</table>

V. CONCLUSION

By viewing the source data, it was apparent that the Mainland China has developed rapidly in recent years. The trade, business, and academic interactions between the cross-strait increased by days. From 2000 to 2009, monthly data on passenger numbers between the cross-strait had shown a tendency of gradual increase over the time. For example, in the period between the first weekend direct flights between the cross-strait on the 4th of July, 2008, and the scheduled flights on the 31th of August, 2009, the demand for air flights capacity increased in multiplying tendency. In this year, two flights were scheduled per day between Taipei and
Shanghai. At the same time, new flight destinations between the cross-strait were established in order to reduce passengers’ travel time and to accommodate the condition of capacity insufficiency. It indicated the importance of forecasting the number of passengers between the cross-strait.

In the model construction process, the study adopted SVR, GP, and Holt mode to develop the forecast model. By comparing SVR, GP, and Holt model; the MAPE value in SVR forecasting passengers between the cross-strait shown the lowest value and it was under 10%. It represented ideal ability to forecast the future passengers’ capacity. For the verification result from January to July, 2010, the average deviation was only 9%. Therefore, this study has adopted the method of SVR to construct an appropriate forecast model on passengers between the cross-strait. The model could accurately forecast the number of passengers on the required market in the future.

As the passengers demand would differ due to seasonal and specific holiday festival influence, the monthly travel capacity would be different as well. In order to optimize the satisfaction of passengers requirement from scarce resources, the coordination on management of airlines, such as flight schedules, number of seats, allocation of supply various destination, plane deployment; should adjust in accordance to holiday and non-holiday period and high or low seasons.

ACKNOWLEDGMENT

The research team of this study was sincerely grateful on the funding and assistance by Feng-Chia University and National Science Council, Taiwan.

REFERENCES