KNOWLEDGE MANAGEMENT AND DATA MINING: IN DEMAND ESTIMATION APPROACH FOR PERFECT REVENUE MANAGEMENT

Masaud Yaghini
Department of Rail Transportation Engineering, Iran University of Science & Technology, Iran
Neda Sakhaee
Department of Rail Transportation Engineering, Iran University of Science & Technology, Iran
sakhaei@alborzniroo.ir
Hamidreza Shahbaznezhad
Department of Information Technology Management, Tehran University, Iran

ABSTRACT

Knowledge management and data mining techniques have been widely used in many important applications in both scientific and business domains in recent years. The importance of forecasting has also been exposed in many practical zones. Considering that accurate forecasts are the base for correct decisions in revenue management, the purpose of this paper is to study the role of knowledge management and data mining techniques for a demand estimating process. For this reason this paper deals with neural network structures, which is a data mining technique, for railway passenger demand forecasting. It is advised that information should be made available in specific positions in neural network operations rather than combining all the accessible information at the beginning. The results can be useful for all knowledge based transportation businesses.

KEYWORDS

Knowledge Management, Data Mining, Data Mining techniques, Forecasting, Neural Networks, Multi Layer Perception, Railway passenger demand.

1. INTRODUCTION

Outputs of forecasting models can be inputs for operational planning. In railway industries, forecasting is the key to the success of revenue management (RM) and is one of four important systems that can be used. Forecasting renders demand information, which can be used as inputs for the other three systems: seat allocation, overbooking and pricing. In order to make correct decisions, accurate forecasts are essential (Tsai, Lee & Wei, 2009).

The most common methodology to cope with short-term forecasting problems in the literature is extrapolation. Vlahogianni, Golias, and Karlaftis (2004) thoroughly reviewed papers in short-term traffic forecasting and indicated that neural networks are the most potential techniques.

Extrapolation is also the most general methodology to deal with demand forecasting problems. The literature identify a number of extrapolation methodologies, such as exponential smoothing (Faraway & Chatfield, 1998; Kyungdoo & Thomas, 1995), ARIMA-type models (Tavana & Mahmassani, 2000; Williams & Hoel, 2003), non-parametric regression (Clark, 2003; Robin & Polak, 2005) and neural networks (Ishak, Kotha, & Alecsandru, 2003; McFadden, Yang, & Durrans, 2001). Although proportional studies

(Makridakis & Hibon, 2000; Ord et al., 2001) and selection models (Prudêncio, Ludermir, & de Carvalho, 2004; Venkatachalam & Sohl, 1999) have discussed the issue of models selection, it is still controversial to say which model can achieve the best predictive performance. The purpose of this research is to adapt neural networks for the development of railway passenger demand forecasting models in the Iranian Railway. According to various sources a neural network structure was usefully applied in multilayer perceptions in many instances (Dougherty, 1995; Hippert, Pedreira, & Souza, 2001).

2. FACTOR MINING

In this section data and factors that will be used in this research will be introduced.

2.1. Data

The daily sales data of all trains from March 2006 to March 2009 was collected from the Islamic Republic of Iran Railways. To illustrate how to design network structures based on railway data factors, this study focuses on one particular train service. Sales data from Tehran to Mashhad of all types of trains over three years will be observed. The reason of choosing Tehran and Mashhad is because they are the biggest two cities in Iran and the rail path between them has prosperous demand. Two categories of data factors are observed. The first category of data factors are chronological factor. Chronological factor is widely used in scrutinizing time series data in the literature. Typically, trend, periodic pattern and cycle are three major components of a chronological factor. The second category of factors is level swing. Level swing is made because of special happenings. In addition, this study studies the daily and monthly amounts of the applied data to extort data factors from different levels of combined data. In the following section, factors are firstly defined. These factors are then used to monitor the railway data and present potential variables for model construction.

2.2. Factor definition

With regard to chronological factor there are three important factors: trend, periodic pattern and cycle. Trend is used to study whether there is an upward increase or a downward decrease over several following periods. The formula of $y_t = a+b\times t$ is applied to test whether a trend exists, where t is the time index. The trend exists if b is significantly different from zero. Periodic pattern is used to observe whether there is a frequent behavior in the data. Day-of-week pattern in the daily data and month-of-year pattern in the monthly data are considered. The formula $yt = a+b\times y_{t-k}$ is applied, where t = k + 1~n, and k is the length of the periodic pattern. If b is significantly different from zero, then the periodic pattern exists. It should be noted that k = 7 for the daily data and k = 12 for the monthly data in the study. In addition, cycle describes a frequent pattern. The major difference between periodic pattern and cycle is that the duration of a cycle is not constant. Cycle usually exists in economic data or data series related to economic fluctuations. Railway passenger demand is a function of economic forces. For example, income may increase during an economic phenomena and provoke passengers to choose more qualified transportation modes. However, it was not possible to consider the influence of cycle in short-term data because at least two complete cycles of data are needed and it needs much more than three-year data to observe two economic cycles. Cyclic influences are better disregarded if there is no firm evidence (Armstrong, 2001). In this study, a cycle is not present.

In the class of level swing, the vacation is the only significant factor used to check whether the mean of data changes during NOROZ celebration and summer vacations. In Iran,

NOROZ celebration vacations are usually from the last third of March to the first week of April, and summer vacations are usually from the end of June to the middle of September. It is estimated that passenger demand swings to a higher level during vacations and swings back to the previous level at the end of vacations.

is estimated that passenger demand swings to a higher level during vacations and swings back to the previous level at the end of vacations.

2.3. Analyzing and exploring factors

Three-year daily data was surveyed and Figure 1 illustrates the applied data, and Table 1 summarizes the observed factors. As can be seen, the coefficient of the trend is significant. The positive mark of the coefficient means that there is an upward trend, as shown by a straight line in Figure 1.

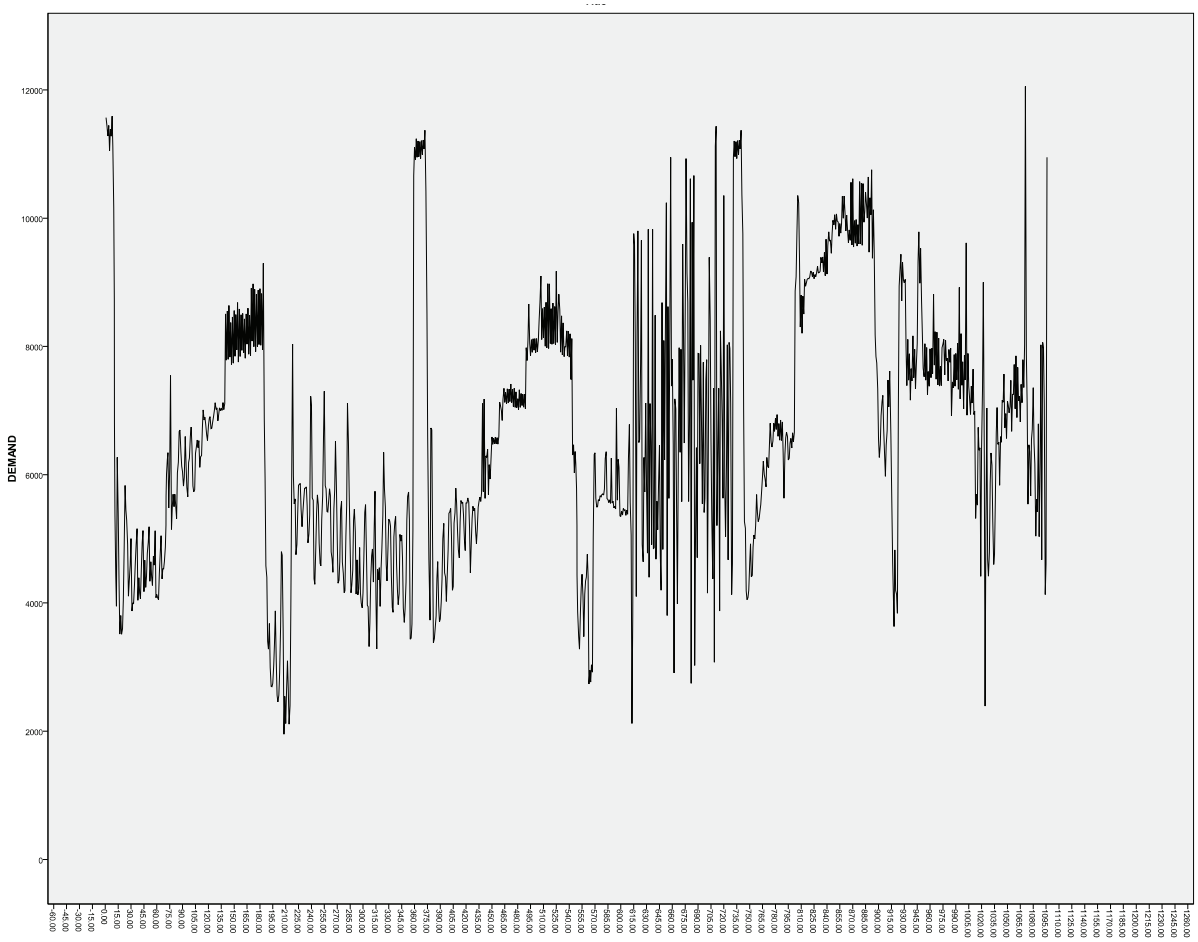

Figure 1: Daily trend

The coefficient of the day-of-week pattern is significant (see Figure 1). Moreover, the influences of vacations, as demonstrated by rectangles (shows the periods of summer vacations and NOROZ celebration vacations) in Figure 1, are prominent. It shows that passenger demand swings to a higher level through vacations and shifts back to the previous level when vacations are over.

Table 1: Factors significance

factor	daily trend	monthly trend	weekly pattern	monthly pattern	NOROZ celebration	summer vacation
significance of correlation(p-value)	0.000	0.000	0.000	0.000	0.000	0.000
correlation	0.339	0.343	0.394	0.42	0.302	0.335
existence situation in model	exist	exist	exist	exist	exist	exist

The daily data is combined to attain the monthly data. In Table 1 the significance of all factors is surveyed by a correlation test and each factor which has a p-value less than 0.05, is significant. It can be seen that the trend is a significant factor. Also, the month-of-year pattern is significant. Furthermore, cycle is not expected to be observed in the three-year data.

In short, the study found that daily and monthly trend, day-of-week, month of year and vacation is four prominent variables in the daily data and monthly data. In the next section, these observed factors are represented via suitable forms for constructing daily passenger demand forecasting models. Table 3 shows the input set and Figure 2 shows the idea of model construction, where attention is put on network designs.

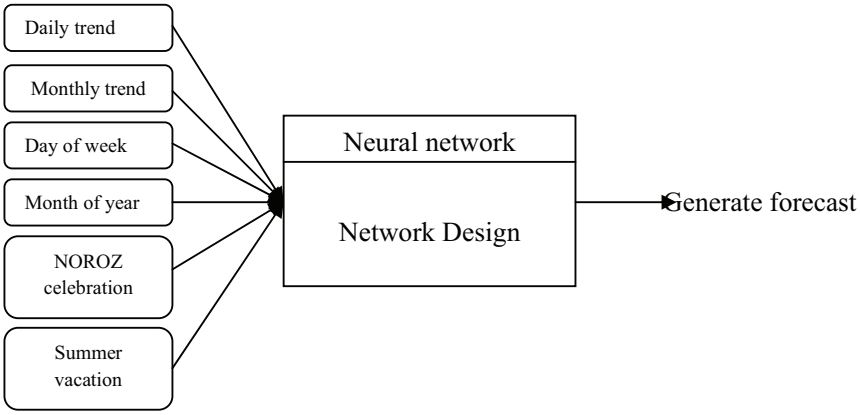

Figure 2: Idea of model construction

3. MODEL STRUCTURE

MLP is a forward connected network and it can be seen in Figure 3 that it has three layers namely an input, hidden and output layer. The input layer is used for getting data from external inputs. The number of neurons in the input layer is related to the number of input factors. The decision of the number of neurons in a hidden layer is different and is still contentious. The output layer generates forecasts and propagates errors for parameter estimation. The number of neurons in the output layer depends on how many lead-time forecasts are requested. In this study the forecast variable is demand which is a clustered variable. The clustering process for making a demand variable is based on the k-means algorithm which means that three major clusters are made. This means that the number of

neurons in the output layer is three. Table 2 illustrates the layer characteristics for applied network construction.

Table 2: Characteristics of layers

Input	number of variable	6
hidden	number of hidden layer	1
	Number of Units	6
	Rescaling Method for Covariates	standard
	Activation Function	Hyperbolic tangent
output	Number of Units	Softmax
	Activation Function	3

As can be seen in Table 3, there are six types of inputs, and consequently twelve neurons are applied in the input layer. As a result, the number of neurons in the hidden layer is six. Each line between the layers has one weight on it and it represents a parameter.

Table 3: Characteristics of layers

Data unit	Denote	# of neurons	Value
Daily trend	DT	1	1, 2, 3, . . ., 365
Monthly trend	MT	1	1, 2, 3, . . .,12
NOROZ celebration	NOROZ	2	0,1
Summer vacation	TABESTAN	2	0,1
Day of week pattern	DP	1	0,1
Month of year pattern	MP	1	0,1

The number of training samples is determined by three factors. Firstly, the number of samples for model construction must be ample to show the attributes of the input set. In this study, at least two-year data seems reasonable to realize the pattern. Secondly, passenger demand is dynamic. Thirdly, it is also vital to keep observing predictive ability over several periods so that believable results can be achieved. In this study, two year data is used as training samples. One month data which is dated exactly after the end of the training samples is used as testing samples. The update frequency is one month.

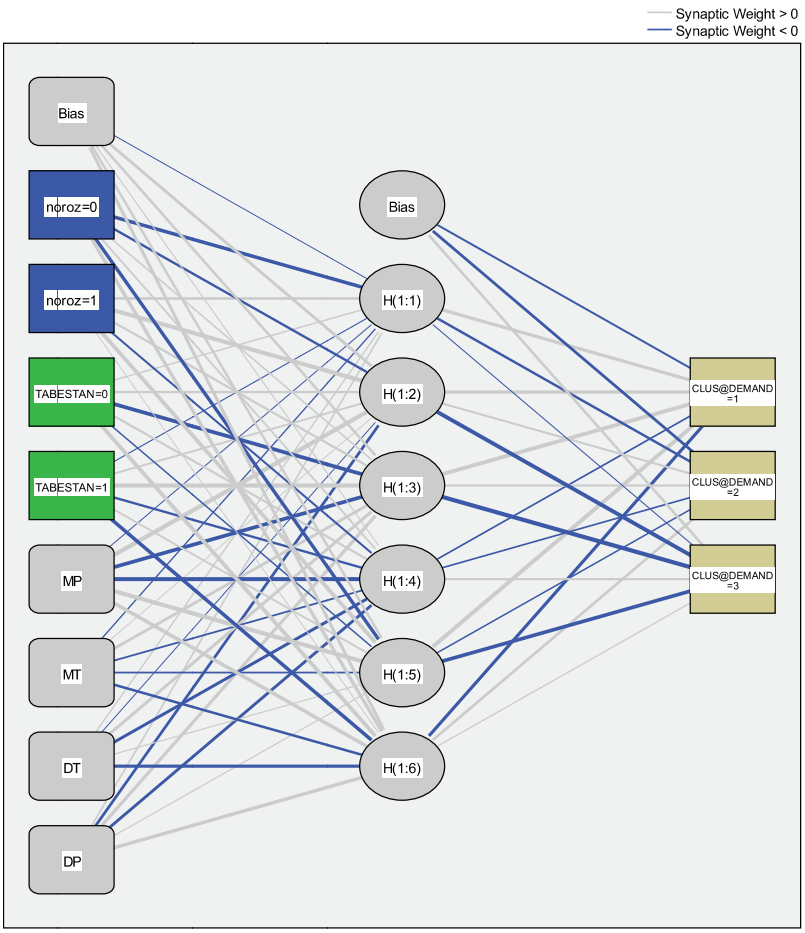

Figure 3: MLP structure for short-term railway passenger demand forecasting

4. EMPIRICAL RESULTS

This section evaluates the performance of used network structures and regression models from MSE perspectives. Firstly, the study observes analytical performance, which is based on an out-of-sample evaluation. The mean square error (MSE) is approved for performance evaluation. Eq (1) shows the formulas. Secondly, paired sample t test is then applied to test whether the two methods have a significant difference.

$$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n}(y(t) - \hat{y}(t))^2 \qquad (1)$$

The result MLP can attain 6.3% upgrading of MSE in comparison with regression. From the above comparisons, the study finds that the proposed MLP network is preferred to a usual regression model for railway passenger demand forecasting in the study.

5. CONCLUSION

In this research, one new neural network model is anticipated. The results show that MLP is favored to regression in terms of forecasting error. It is recommended that further research should be conducted to expand the proposed models.

The paper tried to examine knowledge management techniques and methodologies that can be used for knowledge extraction in a business to estimate demand. In particular, the results of this research are useful for revenue management based on knowledge management. This indicates a new linkage between knowledge management and revenue management.

REFERENCES

Clark, S. (2003). Traffic prediction using multivariate nonparametric regression. Journal of Transportation Engineering, 129(2), 161–168.
Dougherty, M. (1995). A review of neural networks applied to transport. Transportation Research Part C, 3(4), 247–260.
Faraway, J., & Chatfield, C. (1998). Time series forecasting with neural networks: A comparative study using the airline data. Applied Statistics, 47(2), 231–250.
Frasconi, P., Gori, M., & Soda, G. (1992). Local feedback multilayered networks. Neural Computation, 4, 120–130.
Hippert, H. S., Pedreira, C. E., & Souza, R. S. (2001). Neural networks for short-term load forecasting: A review and evaluation. IEEE Transactions on Power Systems, 16(1), 445.
Ishak, S., Kotha, P., & Alecsandru, C. (2003). Optimization of dynamic neural networks performance for short-term traffic prediction. Transportation Research Record, 1836, 45–56.
Lee, A. O. (1990). Airline reservations forecasting: Probabilistic and statistical models of the booking process. MIT dissertation, USA.
Makridakis, S., & Hibon, M. (2000). The M3-competition: Results, conclusions and implications. International Journal of Forecasting, 16, 451–476.
McFadden, J., Yang, W.-T., & Durrans, S. R. (2001). Application of artificial neural networks to predict speeds on two-lane rural highways. Transportation Research Record, 1751, 9–17.
Morantz, B. H., Whalen, T., & Zhang, G. P. (2004). A weighted window approach to neural network time series forecasting. In G. P. Zhang (Ed.), Neural networks in business forecasting. USA: IRM Press.
Nelles, O. (2001). Nonlinear system identification. Germany: Springer. Ord, K. et al. (2001). Commentaries on M3-competition. International Journal of Forecasting, 17, 537–584.
Parlos, A. G., & Patton, A. D. (1993). Long-term electric load forecasting using dynamic neural network architecture. In IEEE/NTUA Athens power tech conference, Athens. Pham, D. T., & Liu, X. (1993). Identification of linear and nonlinear ynamic systems using recurrent neural networks. Artificial Intelligence in Engineering, 8, 67–75.
Prudêncio, R. B. C., Ludermir, T. B., & de Carvalho, F. A. T. (2004). A model symbolic classifier for selecting time series models. Pattern Recognition Letters, 25, 911–921.
Robin, S., & Polak, J. W. (2005). Modelling urban link travel time with inductive loop detector data using the k-NN method. In 2005 TRB annual meeting, Washington, DC.
Smith, B. L., & Demetsky, M. J. (1994). Short-term traffic flow prediction: Neural network approach. Transportation Research Record, 1453, 98–104. Tavana, H., & Mahmassani, H. S. (2000). Estimation and application of dynamic speed-density relations by using transfer function models. Transportation Research Record, 1710, 47–57.
Tsung-Hsien Tsai, Chi-Kang Lee & Chien-Hung Wei,(2009) Neural network based temporal feature models for short-term railway passenger demand forecasting, Expert Systems with Applications (36) 3728–3736
Venkatachalam, A. R., & Sohl, J. E. (1999). An intelligent model selection and forecasting system. Journal of Forecasting, 18, 167–180. Vlahogianni, E. I., Golias, J. C., & Karlaftis, M.

G. (2004). Short-term traffic forecasting: Overview of objectives and methods. Transport Reviews, 24(5), 533–557.
Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2005). Optimized and metaoptimized neural networks for short-term traffic flow prediction: A genetic approach. Transportation Research Part C, 13, 211–234.
Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech, and Signal Processing, 37(3), 328–339.
West, D., & Dellana, S. (2004). Predicting wastewater BOD levels with neural network time series models. In G. P. Zhang (Ed.), Neural networks in business forecasting. USA: IRM Press.
Williams, B. N., & Hoel, L. A. (2003). Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. Journal of Transportation Engineering, 129, 664–672.
Yasdi, R. (1999). Prediction of road traffic using a neural network approach. Neural Computing and Applications, 8, 135–142.
Yun, S. Y., Namkoong, S., Rho, J. H., Shin, S. W., & Choi, J. U. (1998). A performance evaluation of neural network models in traffic volume forecasting. Mathematic Computing Modelling, 27(9–11), 293–310.