Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 32 (2007) 254–264
Expert Systems with Applications
www.elsevier.com/locate/eswa

A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan

Kuan-Yu Chen [a,*], Cheng-Hua Wang [b]

[a] *Department of Business Administration, Far East College and Graduate School of Business and Operations Management, Chang-Jung Christian University, 49 Chung Hua Road, Hsin-Shih, Tainan County 744, Taiwan*
[b] *Department of Business Administration, Chang-Jung Christian University, Taiwan*

Abstract

This paper proposes a hybrid methodology that exploits the unique strength of the seasonal autoregressive integrated moving average (SARIMA) model and the support vector machines (SVM) model in forecasting seasonal time series. The seasonal time series data of Taiwan's machinery industry production values were used to examine the forecasting accuracy of the proposed hybrid model. The forecasting performance was compared among three models, i.e., the hybrid model, SARIMA models and the SVM models, respectively. Among these methods, the normalized mean square error (NMSE) and the mean absolute percentage error (MAPE) of the hybrid model were the lowest. The hybrid model was also able to forecast certain significant turning points of the test time series.

Keywords: Support vector machines; Neural networks; SARIMA

1. Introduction

The Taiwanese machinery industry has steadily progressed over the recent decade, forming a critical foundation for the overall manufacturing industry in Taiwan. Furthermore, it is a major exporting industry in Taiwan. In addition to the traditional precision machinery, wafer slicing, semiconductor manufacturing equipment, high-tech anti-pollution equipment, and crucial machinery parts are being actively developed with government support. Entrepreneurial activity in the industry is mainly focused on overseas markets. However, of those manufacturers dedicated to the machinery industry, 95% of them are of medium and small enterprises. To take benefit from the globally competitive markets, these companies must respond rapidly to the change of market requirements. As a result, forecasting production value is important for the Taiwanese machinery industry (Pai & Lin, 2005).

Generally, production values in the machinery industry change over time. The changes thus can be treated as a time series process. Time series forecasting is an important area of forecasting in which past observations of the same variable are gathered and analyzed to develop a model describing the underlying relationship. The model is then used to extrapolate the time series into the unseen future. This modelling approach is especially useful when little knowledge is available on the underlying data generating process, or when no satisfactory explanatory model exists relating the prediction variable to other explanatory variables. Considerable effort has been devoted during recent decades to developing and improving time series forecasting models.

Several different approaches are available for time series modelling. One of the most popular and extensively used seasonal time series forecasting models is the seasonal auto-regressive integrated moving average (SARIMA) model. Widespread use of the SARIMA model is owing to its statistical properties, as well as the well-known Box-Jenkins methodology (1976) used for constructing the model. The SARIMA model has been successfully

* Corresponding author. Tel.: +886 6 5977611; fax: +886 6 5977610.
E-mail address: t530903@ms7.hinet.net (K.-Y. Chen).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.11.027

adopted in numerous fields (Goh & Law, 2002; Huang & Min, 2002; Li, Campbell, Haswell, Sneeuwjagt, & Venables, 2003; Navarro-Esbri, Diamadopoulos, & Ginestar, 2002). Although the SARIMA model has been highly successful in both academic research and industrial applications during the past three decades, it suffers from a major limitation owing to its pre-assumed linear form of the model. Restated, a linear correlation structure is assumed among the time series values, and thus the SARIMA model cannot capture any nonlinear patterns. It is not always suitable to apply linear models to complex real-world problems (Zhang, 2003).

In 1995, Vapnik developed a neural network algorithm called Support Vector Machine (SVM), which is a novel learning machine based on statistical learning theory, and which adheres to the principle of structural risk minimization seeking to minimize an upper bound of the generalization error rather than minimize the training error (the principle followed by neural networks). This induction principle is based on the bounding of the generalization error by the sum of the training error and a confidence interval term depending on the Vapnik–Chervonenkis (VC) dimension. Based on this principle, SVM achieves an optimum network structure by striking a right balance between the empirical error and the VC-confidence interval. This balance eventually leads to better generalization performance than other neural network models (Tay & Cao, 2001). Additionally, the SVM training process is equivalent to solving linearly constrained quadratic programming problems, and the SVM embedded solution meaning is unique, optimal and unlikely to generate local minima. Originally, SVM has been developed to solve pattern recognition problems. However, with the introduction of Vapnik's ε-insensitive loss function, SVM has been extended to solve nonlinear regression estimation problems, such as new techniques known as support vector machines for regression, which have been shown to exhibit excellent performance (Vapnik, Golowich, & Smola, 1997).

Different forecasting models can achieve success each other in capturing patterns of data sets, and numerous authors have demonstrated that combining the predictions of several models frequently results in higher prediction accuracy than that of the individual models (Lawerence, Edmundson, & O'Connor, 1986; Makridakis, 1989; Makridakis & Winkler, 1983). Since the early work of Reid (1968) and Bates and Granger (1969), the literature on this topic has expanded significantly. Clemen (1989) provided a comprehensive review and annotated bibliography on this area. Wedding and Cios (1996) described a combining methodology using radial basis function networks and the Box–Jenkins models. Luxhoj, Riis, and Stensballe (1996) developed a hybrid econometric and ANN approach for sales forecasting. Pelikan, de Groot, and Wurtz (1992) and Ginzburg and Horn (1994) proposed combining several feed forward neural networks to enhance the accuracy of time series forecasting. Tseng, Yu, and Tzeng (2002) proposed a hybrid forecasting model, which combines the seasonal time series ARIMA (SARIMA) and the neural network back propagation models. Furthermore, Zhang (2003) combined the ARIMA and feed-forward neural network models for forecasting. In this study, we combine the SARIMA and SVM models to forecast time series involving seasonality.

The remainder of this study is organized as follows. In Section 2, the SARIMA, the SVM models, and the hybrid models are described. Section 3 elaborates on the GA-SVM model. Section 4 describes the data source. Section 5 discusses the evaluation methods used for comparing the forecasting techniques. Section 6 analyzes the results of real-code genetic algorithms and optimizes SVM's parameters, and also explains the determining parameters process of the SARIMA models. Section 7 compares the results obtained from the hybrid model against the SARIMA model and the SVM model. Section 8 provides concluding remarks.

2. Methodology

Both the SARIMA and SVM models are summarized in the following as foundation to describe the hybrid model.

2.1. SARIMA model

SARIMA is the most popular linear model for forecasting seasonal time series. It has achieved great success in both academic research and industrial applications during the last three decades. A time series $\{Z_t|t=1,2,\ldots,k\}$ is generated by SARIMA $(p,d,q)\ (P,D,Q)_s$ process of Box and Jenkins time series model if

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D Z_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t, \tag{1}$$

where p, d, q, P, D, Q are integers, s is the season length;

$$\phi_p(B) = 1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p,$$
$$\Phi_P(B^s) = 1-\Phi_s B^s-\Phi_{2s}B^{2s}-\cdots-\Phi_{Ps}B^{Ps},$$
$$\theta_q(B) = 1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q \text{ and}$$
$$\Theta_Q(B^s) = 1-\Theta_s B^s-\Theta_{2s}B^{2s}-\cdots-\Theta_{Qs}B^{Qs}$$

are polynomials in B of degree p, q, P, and Q. B is the backward shift operator, and ε_t is the estimated residual at time t. d is the number of regular differences, D is the number of seasonal differences; Z_t denotes the observed value at time t, $t=1,2,\ldots,k$.

Fitting a SARIMA model to data involves the following four-step iterative cycles:

(a) Identify the SARIMA $(p,d,q)\ (P,D,Q)$s structure;
(b) Estimate unknown parameters;
(c) Perform goodness-of-fit tests on the estimated residuals;
(d) Forecast future outcomes based on the known data.

The ε_t should be independently and identically distributed as normal random variables with mean $=0$ and constant