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.

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
adopted in numerous fields (Goh & Law, 2002; Huang & Min, 2002; Li, Campbell, Haswell, Sneeuwjag, & 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 $\epsilon$-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 (Lawrence, 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. Clemens (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 \((p,d,q) (P,D,Q)\), process of Box and Jenkins time series model if

$$\phi_p(B)\Phi_P(B^s)(1-B)^dZ_t = \Theta_q(B)\Theta_Q(B^s)e_t,$$

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

$$\phi_p(B) = 1 - \phi_1B - \phi_2B^2 - \cdots - \phi_pB^p,$$
$$\Phi_P(B^s) = 1 - \Phi_1B^s - \Phi_2B^{2s} - \cdots - \Phi_PB^{ps},$$
$$\Theta_0(B) = 1 - \theta_1B - \theta_2B^2 - \cdots - \theta_qB^q$$
and
$$\Theta_Q(B^s) = 1 - \Theta_1B^s - \Theta_2B^{2s} - \cdots - \Theta_QB^{qs}$$
are polynomials in \(B\) of degree \(p, q, P, \) and \(Q\). \(B\) is the backward shift operator, and \(e_t\) 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 \((p,d,q) (P,D,Q)\) 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 \(e_t\) should be independently and identically distributed as normal random variables with mean \(= 0\) and constant
variance $\sigma^2$. The roots of $\phi(Z) = 0$ and $\theta(Z) = 0$ should all lie outside the unit circle. In addition, it is suggested by Box and Jenkins (1976) that a minimum of 50 (preferably 100) observations should be used for the SARIMA model.

2.2. Support vector machines for regression

SVM can be applied to regression problems by introducing an alternative loss function, with very encouraging results (Vapnik, 1995; Vapnik et al., 1997). SVM maps the input data $x$ into a higher-dimensional feature space $F$ by nonlinear mapping, to yield and solve a linear regression problem in this feature space. Therefore, the regression approximation addresses the problem of estimating a function based on a given data set $G = \{(x_i, d_i)\}^n_{i=1}$, where $x_i$ denotes the input vector, $d_i$ denotes the desired value, and $n$ denotes the total number of data patterns. In SVM, the regression function is approximated by the following function:

$$f(x) = \omega \phi(x) + b,$$

$$\phi : \mathbb{R}^n \rightarrow F, \omega \in F,$$  \hspace{1cm} (2)

where $b$ is a scalar threshold; $\phi(x)$ is the high dimensional feature space which is nonlinearly mapped from the input space $x$. Thus, the linear regression in the high-dimensional feature space responds to nonlinear regression in low dimension input space, which disregards the inner product computation between $\omega$ and $\phi(x)$ in the high-dimension feature space. The coefficients $\omega$ and $b$ are estimated by minimizing

$$R_{\text{SVM}}(\omega) = R_{\text{emp}} + \frac{1}{2} \lVert \omega \rVert^2 = C \times \frac{1}{n} \sum_{i=1}^{n} L_\epsilon(d_i, y_i) + \frac{1}{2} \lVert \omega \rVert^2,$$  \hspace{1cm} (3)

$$L_\epsilon(d, y) = \begin{cases} |d - y| - \epsilon, & |d - y| \geq \epsilon, \\ 0 & \text{otherwise}. \end{cases}$$  \hspace{1cm} (4)

In the regularized risk function given by Eq. (3), the first term $C \times \frac{1}{n} \sum_{i=1}^{n} L_\epsilon(d_i, y_i)$ represents the empirical error (risk), which is estimated using the $\epsilon$-insensitive loss function in Eq. (4). The loss function is introduced to obtain enough samples of the decision function in Eq. (2) by using fewer data points. The second item $\frac{1}{2} \lVert \omega \rVert^2$ represents the regularization term. The regularized constant $C$ calculates the penalty when an error occurs, by determining the trade-off between the empirical risk and the regularization term, which represents the ability of prediction for regression. Raising the value of $C$ increases the importance of the empirical risk relative to the regularization term. The penalty is acceptable only if fitting error is greater than $\epsilon$. The $\epsilon$-insensitive loss function is used to stabilize estimation. In other words, the $\epsilon$-insensitive loss function can decrease noise. Thus, $\epsilon$ can be considered as a tube size equivalent to the approximation accuracy in training data. In empirical analysis, $C$ and $\epsilon$ are the parameters selected by users.

To estimate $\omega$ and $b$, Eq. (3) is transformed to the primal function given by Eq. (5) by introducing the positive slack variables $\xi_i, \xi_i^*$ as follows:

$$\text{Minimize} \quad R_{\text{SVM}}(\omega, \xi) = \frac{1}{2} \lVert \omega \rVert^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$  \hspace{1cm} (5)

Subjected to

$$d_i - \omega \phi(x_i) - b_i \leq \epsilon + \xi_i, \quad \omega \phi(x_i) + b_i - d_i \leq \epsilon + \xi_i^*,$$

$$\xi_i, \xi_i^* \geq 0.$$  \hspace{1cm} (6)

Finally, by introducing Lagrange multipliers and exploiting the optimality constraints, the decision function given by Eq. (2) has the following explicit form Vapnik (1995):

$$f(x, x_i, x_i^*) = \sum_{i=1}^{n} (x_i - x_i^*) K(x_i, x_i) + b.$$  \hspace{1cm} (7)

2.2.1. Lagrange multipliers

In Eq. (6), $x_i$ and $x_i^*$ are the so-called Lagrange multipliers. They satisfy the equalities $x_i \times x_i^* = 0, x_i \geq 0$ and $x_i^* \geq 0$ where $i = 1, 2, \ldots, n$, and are obtained by maximizing the dual function of Eq. (5), and the maximal dual function in Eq. (5) has the following form:

$$\text{Max} \quad R(x, x_i') = \sum_{i=1}^{n} d_i (x_i - x_i') - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_i') (x_j - x_j') K(x_i, x_j)$$  \hspace{1cm} (7)

with the constraints

$$\sum_{i=1}^{n} (x_i - x_i') = 0, \quad 0 \leq x_i \leq C, \quad i = 1, 2, \ldots, n,$$

$$0 \leq x_i^* \leq C, \quad i = 1, 2, \ldots, n.$$  \hspace{1cm} (7)

Based on the Karush-Kuhn-Tucker’s (KKT) conditions of solving quadratic programming problem, $(x_i - x_i')$ in Eq. (7), only some of them will be held as non-zero values. These approximation errors of data point on non-zero coefficient will equal to or larger than $\epsilon$, and are referred to as the support vector. That is, these data points lie on or outside the $\epsilon$-bound of decision function. According to Eq. (7), the support vectors are clearly the only elements of the data points employed in determining the decision function as the coefficient $(x_i - x_i')$ of other data points are all equal to zero. Generally, the larger the $\epsilon$ value, the fewer the number of support vectors, and thus the sparser the representation of the solution. Nevertheless, increasing $\epsilon$ decreases the approximation accuracy of training data. In this sense, $\epsilon$ determines the trade-off between the sparseness of representation and closeness to the data (Tay & Cao, 2001).

2.2.2. Kernel function

The term $K(x_i, x_j)$ in Eq. (6) is defined as kernel function, where the value of kernel function equals the inner product of two vectors $x_i$ and $x_j$ in the feature space $\phi(x_i)$ and $\phi(x_j)$, meaning that $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$. The kernel function is intended to handle any dimension feature space without
the need to calculate \( \phi(x) \) accurately (Tay & Cao, 2001). If any function can satisfy Mercer’s condition, it can be employed as a kernel function (Vapnik, 1995). The typical examples of kernel function are the polynominal kernel \( K(x, y) = (x \cdot y + 1)^d \) and the Gaussian kernel \( K(x, y) = \exp(-((x - y)^2 / 2\sigma^2)) \). In these equations, \( d \) represents the degree of the polynominal kernel, and \( \sigma^2 \) represents the bandwidth of Gaussian kernel. These parameters must be selected accurately, since they determine the structure of high dimensional feature space and govern the complexity of the final solution.

2.3. The hybrid methodology

Both the SARIMA and SVM models have been successfully applied to their own linear or nonlinear problems. However, neither is a universal model suitable for all situations. Since it is difficult to completely know the characteristics of the seasonal time series data in a real problem, a hybrid strategy that involves both linear and nonlinear modelling abilities provides a good alternative for forecasting seasonal time series. Both the SARIMA and the SVM models have the unique strength to capture characteristics of the seasonal time series data in a real problem. However, neither is a universal model suitable for all situations (Duan, Keerthi, & Poo, 2001; Lin, 2001). These parameters include:

1. Kernel function: The kernel function is used to construct a nonlinear decision hyper-surface on the SVM input space. Generally, using Gaussian function will yield better prediction performance (Smola, 1998). Therefore, the Gaussian function is used as the SVM kernel function in this study.
2. Regularization parameter \( C \): \( C \) determines the trade-off cost between minimizing the training error and minimizing the model’s complexity.
3. Bandwidth of the kernel function \( (\sigma^2) \): \( \sigma^2 \) represents the variance of the Gaussian kernel function.
4. The tube size of \( \varepsilon \)-insensitive loss function \( (\varepsilon) \): It is equivalent to the approximation accuracy placed on the training data points.

In this study, the author proposes a new method known as GA-SVM, which optimizes all SVM’s parameters simultaneously. This model adopts real value genetic algorithms (RGA) to seek the optimal values of SVM’s parameters and improve the prediction efficiency. The proposed GA-SVM model, dynamically optimizing the values of SVM’s parameters through RGA evolutionary process, and using acquired parameters to construct optimized SVM model in order to proceeded prediction. Fig. 1 illustrates the algorithm process of the GA-SVM model.

Initial populations consisting of chromosomes were randomly generated from RGA to search for optimal values of SVM’s parameters. The values of the three parameters, i.e., \( C \), \( \sigma^2 \) and \( \varepsilon \), were directly coded in the chromosomes with real value data. Details of our proposed GA-SVM are described as follows:

1. Representation.

When GA-SVM solves optimal problems, the relative real valued parameters or variables can be directly used form a chromosome, unlike traditional binary genetic algorithms which must be translated into binary codes. Thus, in RGA, the chromosome representation is straightforward. The three SVM parameters, \( C \), \( \sigma^2 \) and \( \varepsilon \), were directly coded to generate
(2) Fitness definition.

The fitness of the training data set is easy to calculate, but is prone to over-fitting. This problem can be handled by using a cross validation technique. In this context, the $k$-fold cross validation offers the best compromise between computational cost and reliable parameter estimates, and was successfully adopted by Duan et al. (2001). In $k$-fold cross validation, the training data set is randomly divided into $k$ mutually exclusive subsets (folds) of approximately equal size. The regression function is built with a given set of parameters $\{C, \sigma^2, \varepsilon\}$, using $k-1$ subsets as the training set. The performance of the parameter set is measured by the MAPE on the last subset. The above procedure is repeated $k$ times, so that each subset is used once for testing. Averaging the MAPE over the $k$ trials (MAPE cross validation) gives an estimate of the expected generalization error for training on sets of size $(k-1/k) \times l$, where $l$ is the number of training data sets. Finally, the best performing parameter set is specified. Conventionally, the training error of $k$-fold cross validation is applied to estimate the generalization error ($k=5$ suggested by Duan et al., 2001). Therefore, the fitness function is defined as the MAPE cross validation of the 5-fold cross validation method on the training data set, as follows:

$$\text{Min} f = \text{MAPE}_{\text{cross validation}} = \frac{1}{l} \sum_{i=1}^{l} \left| a_i - p_i \right| / a_i \times 100\%,$$  \hspace{1cm} (14)  

where $l$ is the number of training data samples; $a_i$ is the actual value, and $p_i$ is the predicted value. The solution with a smaller MAPE cross validation of the training data set has a smaller fitness value, and thus has a better chance of surviving in the successive generations.

(3) Population initialization.

In this study, the initial population was composed of 8 randomly created chromosomes. The population size of 8 was selected as a trade-off between the convergence time and the population diversity.

(4) Fitness evaluation.

The fitness value for each chromosome was calculated according to Eqs. (14) and (15).

(5) Selection.

A standard roulette wheel was employed to select 8 chromosomes from the current population.

(6) Crossover.

The simulated binary crossover (Deb & Agrawal, 1995; Deb & Kumar, 1995) was applied to randomly paired chromosomes. The probability of creating new chromosomes in each pair was set to 0.8. The newly created chromosomes constituted a new population.

(7) Mutation.

The mutation operation follows the crossover operation and determines whether a chromosome should be mutated in the next generation. This study applied polynomial mutation methods (Deb & Goyal, 1996) to the proposed model. Each chromosome in the new population was subject to mutation with a probability of 0.05.

(8) Elitist strategy.

The fitness value was calculated for the chromosomes of the new population. If the minimum fitness value of the new population is smaller than that of the old population, then the old chromosome can be replaced with the new chromosome of the minimum fitness value.

(9) Stopping criteria.

The process was repeated from 4 to 8 until the number of generations was equal to 100.

According to Chtioui, Bertrand, and Barba (1998), the converged solution is mostly affected by the setting probability of parameter. In this study, the crossover probability is recommended from Holland (1975). The choices of other parameters such as the mutation probability, population size are based on numerous experiments as those values provide the smallest MAPE cross validation on the training data set.
4. Data set

The machinery industry is one of the main exporting industries in Taiwan. The production values of the Taiwanese machinery industry have increased continuously during past years. In this study, SVM, SARIMA and the hybrid model are used to forecast the seasonal time series data of production values for the Taiwanese machinery industry. Production values of the Taiwanese machinery industry during the period January 1991 to December 1996 was employed as experimental data. Fig. 2 illustrates the production values of the Taiwanese machinery industry during the study period. In Fig. 2, we can observe the strong seasonality and growth trends of the data. The sharp drop that generally occurs in January or February each year is due to plants closing for the Chinese New Year holiday. The datasets contained 72 data points in time series, and are the same as the datasets cited in Tseng et al. (2002) and Pai and Lin (2005). The collected data were divided into two sets, training data (in-sample data) and testing data (out-of-sample data), in order to testify the performance of the three proposed forecasting methods. To achieve a more reliable and accurate result, a long period served as the training period. Based on these considerations, the period from January 1991 to February 1996 was chosen as the training period, and that from March 1996 to December 1996 served as the testing period. Each data point will be scaled by Eq. (16) within the range of (0,1). This scaling for original data points will be helpful for improving the forecasting accuracy (Chang & Lin, 2001).

\[
\frac{X_t - X_{\min}}{X_{\max} - X_{\min}} \times 0.7 + 0.15.
\]  

(16)

\(X_t\): The production values at time-\(t\).
\(X_{\max}\): The maximum of production values during the period of data source.
\(X_{\min}\): The minimum of production values during the period of data source.

5. Performance criteria

5.1. Quantitative evaluations

Some statistical metrics such as NMSE (normalized mean square error), MAPE (mean absolute percentage error) and \(R\) (correlation coefficient) were used to evaluate the forecasting performance of the three models. Table 1 shows these performance metrics and their calculations. NMSE and MAPE were used to measure the deviation between the actual and predicted values. The smaller the values of NMSE and MAPE, the closer were the predicted values to the actual values. The metric \(R\) was adopted to measure the correlation of the actual and the predicted values.

5.2. Turning point evaluations

While the above criteria provide good measures of the deviations of the predicted values from the actual values, they cannot reflect model ability to predict turning points (directional change). Therefore, the turning points are as important as the forecast value itself. To assess competing models, model abilities were evaluated using two directional tests: the directional change accuracy (DCA) and the regression test.

DCA is a non-parametric test for the directional accuracy of forecasting that focuses on accurately predicting the directional change for the variable under consideration (Pesaran & Timmermann, 1992). Let \(Z_t\) denote the predicted value at time \(t\), while \(Z_t\) represents the actual value at time \(t\). This test does not require any quantitative information and only uses the signs of \(Z_t\) and \(Z_t\). The following indicator variables can be defined:

\[
A_t = \begin{cases} 
1, & \text{if } Z_t \text{ change from a low to high value,} \\
0, & \text{otherwise}, 
\end{cases}
\]

\[
F_t = \begin{cases} 
1, & \text{if } \tilde{Z}_t \text{ change from a low to high value,} \\
0, & \text{otherwise}, 
\end{cases}
\]

\[
D_t = \begin{cases} 
1, & \text{if value of } (A_t, F_t) \text{ are } (1, 1) \text{ or } (0, 0), \\
0, & \text{otherwise}. 
\end{cases}
\]

We define \(P\), \(P_f\), and \(P_a\) as the proportion of times that the signs of \(D_t\), \(F_t\), and \(A_t\), are, respectively, predicted correctly. Therefore, \(P = \sum_{t=1}^{k} D_t/k\), and \(P_f\) and \(P_a\) have

Table 1

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMSE</td>
<td>(\text{NMSE} = \frac{1}{n} \sum_{t=1}^{n} (X_t - \hat{X}_t)^2)</td>
</tr>
<tr>
<td>MAPE</td>
<td>(\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \left</td>
</tr>
<tr>
<td>(R)</td>
<td>(R = \frac{\sum_{t=1}^{n} X_t \hat{X}<em>t}{\sqrt{\sum</em>{t=1}^{n} X_t^2} \sqrt{\sum_{t=1}^{n} \hat{X}_t^2}})</td>
</tr>
</tbody>
</table>
similar expressions with variables $F_t$ and $A_t$. Furthermore, under the assumption that $F_t$ and $A_t$ are independently distributed, we have

$$P^* = P_a_P_f + (1 - P_a)(1 - P_f).$$

(17)

In general, the standardized test statistics are given by

$$S_n = \frac{P - P^*}{\sqrt{\text{var}(P) - \text{var}(P^*))^{0.5}} \sim N(0, 1),$$

(18)

where

$$\text{var}(P) = k^{-1}P(1 - P),$$

$$\text{var}(P^*) = k^{-1}(2P_a - 1)^2(P_f)(1 - P_f),$$

$$+ k^{-1}(2P_f - 1)^2(P_a)(1 - P_a)$$

$$+ 4k^{-2}(P_f)(P_a)(1 - P_f)(1 - P_a).$$

The regression method had developed by Cumby and Modest (1987). Cumby and Modest (1987) suggest the following regression equation:

$$F_t = a_0 + a_1 A_t + \varepsilon_t,$$  

(19)

where $\varepsilon_t$ is error term; and $a_1$ is the slope of this linear equation. Here, $a_1$ should be positive and significantly different from 0 in order to demonstrate those $F_t$ and $A_t$ have a linear relationship. This reflects the ability of a forecasting model to capture the turning points of a time series.

6. Parameters determination of three models

In this study, determining the parameters in three models plays a significant role in obtaining good forecasting performance. Therefore, this section discusses the parameter determining process of three models.

6.1. SVM model

For the SVM model, no standard procedure exists to determine the free parameters, $C$, $\sigma^2$, and $\varepsilon$. Hence, this study optimized the SVM parameters by RGA. First, the 5-fold cross validation technique and RGA were applied for searching, obtaining SVM optimal parameter sets when the MAPE cross validation of 5-fold cross validation is at its minimum. The searching process of optimal parameters was operated with 100 generations in total. The whole evolutionary process was recorded. Fig. 3 illustrates the correlation curves between the optimal and average population fitness arising from the generation number. In the evolutionary process observed from Fig. 3, the MAPE cross validation of the optimal fitness and average population fitness decreased with increasing generation number. When the sample evolution to Generation 28, the MAPE cross validation of 5-fold cross validation converged, indicating that the searching of the RGA is featured with quite excellent efficiency. Within the overall searching process, the optimal fitness was reached at Generation 82 with the MAPE cross validation value rated at 5.4815%. Thus, the individual at Generation 82 produced the optimal parameters, which were $C = 26$, $\sigma^2 = 4.1632$ and $\varepsilon = 0.0198$. These optimal parameter sets were applied to construct the SVM models.

6.2. Hybrid model

In this study, the author proposed three hybrid models which are fundamentally derived from SVM. Therefore, the determining process for parameters is similar to SVM model. The searching processes of optimal parameters are shown as Figs. 4–6 and the obtained optimal parameters for each hybrid model seen as Table 2.

6.3. SARIMA model

The production values of the Taiwanese machinery industry are varied according to seasonal change, and show
increasing or decreasing tendencies annually. The SARIMA model can be applied to analyze time series with characteristics of seasonal and tendency. The SARIMA models are fitted to stationary time series, and seasonal data require regular and seasonal differencing to become stationary. The SARIMA model is denoted as SARIMA\((p,d,q)\)(\(P,D,Q\))\(_{12}\), where \(p\) and \(P\) represent the orders of the non-seasonal and seasonal autoregressive parameters, respectively; \(q\) and \(Q\) represent the orders of the non-seasonal and seasonal moving average parameters, respectively, and \(d\) and \(D\) represent the numbers of the regular and seasonal differences required, respectively.

The test time series data were processed by taking the first-order regular difference and the first seasonal difference to remove the growth trend and seasonality characteristics. The authors used Eviews statistical software to formulate the SARIMA model. Akaike Information Criterion (AIC) was used to identify the best model. The model generated from the data set is SARIMA \((1,1,1)(1,1,1)_{12}\), as listed in Table 3. The equation used is presented in Eq. (20):

\[
Y_t = 2092.319 - 0.399Y_{t-1} + 0.990Y_{t-12} + 0.395Y_{t-13} + ε_t + 0.997ε_{t-1} + 0.933ε_{t-12} + 0.930ε_{t-13}, \tag{20}
\]

Table 2

<table>
<thead>
<tr>
<th>Model</th>
<th>C</th>
<th>(σ^2)</th>
<th>(ε)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARIMASVM1</td>
<td>79</td>
<td>2.8169</td>
<td>0.0612</td>
</tr>
<tr>
<td>SARIMASVM2</td>
<td>713</td>
<td>34.0136</td>
<td>0.0047</td>
</tr>
<tr>
<td>SARIMASVM3</td>
<td>632</td>
<td>161.2903</td>
<td>0.0051</td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>(t)-Statistics</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2092.319</td>
<td>4687.666</td>
<td>0.446346</td>
<td>0.6576</td>
</tr>
<tr>
<td>AR(1)</td>
<td>-0.398789</td>
<td>0.148220</td>
<td>-2.690517</td>
<td>0.0101</td>
</tr>
<tr>
<td>SAR(12)</td>
<td>0.989673</td>
<td>0.024128</td>
<td>41.01768</td>
<td>0.0000</td>
</tr>
<tr>
<td>MA(1)</td>
<td>-0.997404</td>
<td>0.080663</td>
<td>-12.36504</td>
<td>0.0000</td>
</tr>
<tr>
<td>SMA(12)</td>
<td>-0.933081</td>
<td>0.029215</td>
<td>-31.93866</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

R-squared         | 0.866399    | Mean dependent var | 133.6250 |
Adjusted R-squared| 0.855972    | S.D. dependent var  | 2751.079 |
S.E. of regression | 1051.288    | Akaike info criterion | 16.85175 |
Sum squared resid  | 47,523,921  | Schwarz criterion   | 17.04676 |
Log likelihood     | -399.4221   | F-Statistics        | 69.71376 |

Fig. 6. The fitness alternation during evolutionary process (SARIMASVM3 model).

Fig. 7. Comparison of prediction error of the different hybrid model. Null hypothesis of the existence of the same means of the forecasted MAPE generated by the hybrid model and other models. *To be rejected at 10% significance level.
where $Y_t$ denotes the observed value at time $t$, $t = 1, 2, \ldots, 72$, and $e_t$ is the estimated residual at time $t$. The $e_t$ should be independently and identically distributed as normal random variables with a mean of zero and a constant variance.

7. Experimental results

The hybrid model integrated the SARIMA and SVM models and tested with the raw seasonal data. Three hybrid models were built using various input layers. To evaluate the performance of the models, model performance was tested by calculating the MAPE, NMSE and $R$ of the testing data set. Besides MAPE, NMSE, and $R$ measurements, the $T$-value was also used to test the hypothesis that SARIMASVM1 and SARIMASVM2, as well as SARIMASVM3, have the same means of absolute forecast errors. If this hypothesis were statistically significant, we would have demonstrated a better model. After RGA was applied to search for the optimal parameter sets of SVM, the forecasting models were built. The optimum results are reported in Fig. 7. From Fig. 7, the SARIMASVM2 model (input nodes $Z_{t-1}$, $Z_{t-12}$ and residual value $e_t$) has the lowest out-of-sample error. $T$-tests also indicated a rejection of the hypothesis that the MAPE of the SARIMASVM2 model is the same as those of the SARIMASVM1 and SARIMASVM3 models. Therefore, this study suggests the SARIMASVM2 model as a preferred hybrid model.

Moreover, this study compares the results obtained from the hybrid models (SARIMASVM2) with the forecasting results from the individual models (SARIMA and SVM models). The results are shown in Figs. 8 and 9 and Table 4. For out-of-sample error comparisons of the machinery production values time series, Fig. 8 shows that the SARIMASVM2 model outperformed the SARIMA and the SVM models. The NMSE of the SARIMA and

![Fig. 7.](image)

![Fig. 8.](image)

![Fig. 9.](image)

Table 4

<table>
<thead>
<tr>
<th>Model</th>
<th>MAPE</th>
<th>NMSE</th>
<th>$R$</th>
<th>$T$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARIMA</td>
<td>3.5158</td>
<td>0.2898</td>
<td>0.9992</td>
<td>1.98*</td>
</tr>
<tr>
<td>SVM</td>
<td>2.9258</td>
<td>0.4625</td>
<td>0.9986</td>
<td>0.71</td>
</tr>
<tr>
<td>SARIMASVM2</td>
<td>2.0012</td>
<td>0.1142</td>
<td>0.9996</td>
<td>—</td>
</tr>
</tbody>
</table>

Fig. 8. Comparison of prediction error of the hybrid models and the individual models. Null hypothesis of the existence of the same means of the forecasted MAPE generated by the hybrid model and other models. *To be rejected at 10% significance level.

Table 4 Results of turning point forecasting capability of the four models

<table>
<thead>
<tr>
<th>Model</th>
<th>DCA</th>
<th>SVM</th>
<th>SARIMASVM1</th>
<th>SARIMASVM2</th>
<th>SARIMASVM3</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCA</td>
<td>0.83</td>
<td>1.36</td>
<td>0.83</td>
<td>3.33*</td>
<td>0.83</td>
</tr>
<tr>
<td>$a_0$ ($t$-value)</td>
<td>0.25 (0.94)</td>
<td>0.25 (0.98)</td>
<td>0.25 (0.94)</td>
<td>0 (-)</td>
<td>0.25 (0.94)</td>
</tr>
<tr>
<td>$a_1$ ($t$-value)</td>
<td>0.25 (0.73)</td>
<td>0.42 (1.26)</td>
<td>0.25 (0.73)</td>
<td>1*</td>
<td>0.25 (0.73)</td>
</tr>
</tbody>
</table>

Null hypothesis of the existence of the $a_0$ or $a_1$ is equal to zero. The $t$-value of the slope coefficient $a_1$ is statistically different from zero and the DCA test is statistically significant. This implies that for the out-of-sample data, the model had turning point forecasting power.

* To be rejected at 5% significance level.
SVM models were 0.2898 and 0.4625, respectively, while the NMSE of the SARIMASVM2 model was clearly the lowest, at 0.1142. The MAPE of the SARIMASVM2 model was just 2.0012, which was a full percentage point better than that of the other three models. T-tests also showed the rejection of the hypothesis that the MAPE of the SARIMASVM2 model is the same as that of the SVM models, while the SARIMA model is insignificant.

Finally, the turning point evaluation method using DCA and regression Eq. (19) are shown in Table 4 for the machinery production value time series. The $T$-value of the slope coefficient $a_1$ of the SARIMASVM2 model shows that it is statistically different from zero. Furthermore, the individual models show poor directional change detectability, as evidenced from the insignificance of the DCA tests. This indicates that the SARIMASVM2 model has good ability to forecast turning points.

8. Conclusions

The experimental results showed that the hybrid model (SARIMASVM2) is superior to the individual models (SARIMA and SVM models) for the test cases of the production value of the Taiwanese machinery industry. The NMSE and MAPE were all lowest for the hybrid model. The hybrid model also outperformed other models in terms of overall proposed criteria, including NMSE, MAPE, $R$, and turning point forecasts. Overall, the results obtained by the hybrid models were superior to those obtained using the individual models, in terms of both prediction errors and directional change detectability. Most of the individual models evaluated showed poor ability to detect directional change, as evident from the insignificance of the DCA tests. This problem can be overcome with the use of the hybrid model. Besides superior turning point detectability, the hybrid models could achieve superior predictive performances and showed promising results. Therefore, the experimental results suggested that the proposed hybrid model is typically a reliable forecasting tool for application within the forecasting fields of seasonal time series.

References


