Neural Comput & Applic (2017) 28:2633–2647
DOI 10.1007/s00521-016-2215-x

ORIGINAL ARTICLE

Sales forecasting by combining clustering and machine-learning techniques for computer retailing

I-Fei Chen[1] · **Chi-Jie Lu**[2]

Received: 22 May 2015 / Accepted: 19 January 2016 / Published online: 3 February 2016

Abstract Sales forecasting is a critical task for computer retailers endeavoring to maintain favorable sales performance and manage inventories. In this study, a clustering-based forecasting model by combining clustering and machine-learning methods is proposed for computer retailing sales forecasting. The proposed method first used the clustering technique to divide training data into groups, clustering data with similar features or patterns into a group. Subsequently, machine-learning techniques are used to train the forecasting model of each group. After the cluster with data patterns most similar to the test data was determined, the trained forecasting model of the cluster was adopted for sales forecasting. Since the sales data of computer retailers show similar data patterns or features at different time periods, the accuracy of the forecast can be enhanced by using the proposed clustering-based forecasting model. Three clustering techniques including self-organizing map (SOM), growing hierarchical self-organizing map (GHSOM), and K-means and two machine-learning techniques including support vector regression (SVR) and extreme learning machine (ELM) are used in this study. A total of six clustering-based forecasting models were proposed. Real-life sales data for the personal computers, notebook computers, and liquid crystal displays are used as the empirical examples. The experimental results showed that the model combining the GHSOM and ELM provided superior forecasting performance for all three products compared with the other five forecasting models, as well as the single SVR and single ELM models. It can be effectively used as a clustering-based sales forecasting model for computer retailing.

Keywords Sales forecasting · Computer retailing · Clustering algorithm · Machine learning

1 Introduction

Sales forecasting is the basis of each stage of firm management planning. Effective sales forecasting can boost firm performance regarding inventory management, merchandise procurement, and sales management, thereby increasing firm profits and decreasing costs. Thus, to improve business management performance, firms must have appropriate sales forecasting models to effectively estimate sales of all products within a specific future period [1–5].

For computer retailers, the accuracy of product sales forecasts is potentially more critical than for other industries. Rapid technological development and accelerating rate of product innovation have intensified competition in the computer market, leading to shortening of product life cycle. Poor sales forecasting may lead firms to maintain insufficient product inventories or overstock inventories. Moreover, these firms may fail to satisfy customer needs and subsequently profit less, decreasing their competitiveness. Consequently, how to construct an effective sales forecasting model for computer products is a critical problem for computer retailers [5, 6].

Numerous studies have investigated sales forecasting in diverse industries such as fashion [7, 8], clothing [3, 9],

✉ Chi-Jie Lu
jerrylu@uch.edu.tw; chijie.lu@gmail.com

[1] Department of Management Sciences, Tamkang University, Tamsui District, New Taipei City, Taiwan, ROC

[2] Department of Industrial Management, Chien Hsin University of Science and Technology, Zhongli District, Taoyuan City 32097, Taiwan, ROC

food products [1], electronics [10, 11], and automobiles [12]. However, few studies have investigated sales forecasting for the information technology or computer products. Lu et al. [4] used multivariate adaptive regression splines to construct sales forecasting models for computer wholesalers. Lu [5] combined the variable selection method and SVR to construct a hybrid sales forecasting model for computer products. Lu and Shao [6] integrated ensemble empirical model decomposition and an extreme learning machine (ELM) for forecasting computer product sales. Dai et al. [13] utilized three different independent component analysis algorithms and support vector regression (SVR) for forecasting sales of an information technology chain store.

Most of the existing studies that have focused on modeling the sales forecasting of computer products have directly used all training data to construct the forecasting model without considering the extent of the relevance between the training data and the data to be forecasted (test data). In such cases, forecasting accuracy may be reduced because the training data possibly contain excessive data irrelevant to the test data, thereby increasing training errors. To reduce computational time and obtain promising forecasting performance, several recent studies have proposed using clustering algorithms to divide the whole forecasting data into multiple clusters having consistent data characteristics before constructing forecasting models [14–20]. However, they have often been used to predict stock prices.

For example, Tay and Cao [14] integrated self-organizing map (SOM) and support vector machine (SVM) to construct a forecasting model, which comprised a two-stage network architecture. In the first stage of their study, the input variable space was divided into multiple irrelevant clusters through the SOM. In the second stage, SVM was used to construct forecasting model for each cluster. They employed a stock market index and five actual futures as empirical data and showed that using the integrated forecasting system exhibited superior performance to that of using an SVR model alone. Similarly, Cao [15] integrated an SOM with support vector regression (SVR) to construct an expert forecasting system for predicting stock price indices. The results showed that the expert system presented favorable forecasting performance and a high convergence speed. Lai et al. [17] adopted K-means clustering to cluster stock price indices, subsequently analyzing the data in each cluster using a fuzzy decision tree to forecast stock prices. Their empirical results indicated that a hybrid method can yield better forecasting results. By combining SOM and SVR to construct forecasting models for predicting Taiwan stock price indices, Huang and Tsai [18] showed that the integrated model exhibited a higher forecasting performance. In a study constructing a stock price forecasting model for the India Nifty index, Badge and Srivastava [19] used K-means clustering to cluster historical stock data into different groups and applied autoregressive integrated moving average (ARIMA) model to the selected suitable group to construct a forecasting model. Their results indicated that the combined model using clustering data in model training stage can improve the forecasting accuracy.

Few studies have applied clustering-based forecasting models to sales forecasting [11, 16, 20–24]. Moreover, no study has investigated the sales forecasting of computer products. Hadavandi et al. [11] integrated genetic fuzzy systems and K-means algorithm to construct a sales forecasting expert system for forecasting monthly printed circuit board sales. Their empirical results exhibited that the clustering-based forecasting method can generate better prediction results. Thomassey and Fiordaliso [16] proposed a clustering-based sales forecasting system by combining K-means algorithm and C4.5 decision tree algorithm for new items of textile distributors. They utilized 285 real sales items from a French textile distributor as empirical data and showed that the proposed forecasting system outperformed the five competing models. Kumar and Patel [20] used Fisher's clustering method to present a hybrid sales forecasting method for retail merchandising. The results showed that their model produced significantly better sales forecasting results than the individual forecasting model without clustering. Chang and Lai [21] combined an SOM with case-based reasoning (CBR) to forecast the sales amounts of new books, showing that using this integrated model yielded more accurate forecasting results compared with using CBR alone. Chang et al. [22] constructed a monthly sales forecasting model for printed circuit boards in Taiwan. By integrating K-means clustering with a fuzzy neural network (FNN), they developed a KFNN hybrid forecasting model, which exhibits higher forecasting accuracy compared with the four other forecasting models. Lu and Wang [23] integrated independent component analysis, growing hierarchical self-organizing map (GHSOM), and SVR to construct a sales forecasting model for computer wholesalers. Their experimental results indicated that the integrated model accurately forecasted the sales of computer wholesalers. Murlidha et al. [24] utilized standard hierarchical agglomerative clustering algorithm with a new sales pattern distance between two sales series to propose a new clustering-based forecasting model to forecast product demand sales of retailers. They demonstrated that the proposed clustering-based sales forecasting model can generate the best prediction performance.

In the present study, a clustering-based forecasting model integrating clustering and machine-learning techniques is proposed for predicting computer product sales.

Since computer manufacturers periodically launch new products or remodel the existing merchandise on the markets to keep pace with the technological advancement, computer retailers accordingly have to orchestrate their marketing campaigns timely for implementation of annual sales plans. As a result, the sales data of computer products exhibit similar data patterns or features at different time periods, and it is believed that a clustering-based forecasting model for sales can be effectively applied to computer products. Three clustering techniques including SOM, GHSOM, and K-means algorithm and two machine-learning techniques including SVR and extreme learning machine (ELM) are used in this study. The SOM, GHSOM, and K-means are commonly adopted clustering methods in previous studies [25]. The SOM has demonstrated its compelling performance of analyzing highly dimensional input data and visualizing data in a comprehensive manner, while GHSOM has lent itself to investigate the hierarchical relationships of input data via its dynamic topology to further elicit the insights of clusters underlying in high-dimensional large datasets. On the other hand, K-means is a very efficient and common clustering algorithm in a variety of data mining applications. Nevertheless, it is prone to terminate its processing iterations so rapidly to obtain a local optimal solution.

Referring to forecasting techniques, SVR is an effective machine-learning algorithm [26, 27]. It is derived from the structural risk minimization principle for estimating a function by minimizing an upper bound of the generalization error and has been receiving increasing attention for solving nonlinear regression estimation problems. The ELM is a novel learning algorithm for single-hidden-layer feed-forward networks. It provides enhanced generalization performance with faster learning speeds and avoids many problems faced using traditional neural network algorithms such as the stopping criterion, learning rate, number of epochs, local minima, and over-tuning [28]. These two prediction methods have been widely applied to various forecasting problems [29–35] and sales forecasting problems [4–6, 36–38].

In the proposed scheme, first, the clustering technique is employed to divide the training data into multiple small training data sets (i.e., clusters) possessing similar data features or patterns before machine-learning technique is used to train the forecasting models. After the cluster containing data patterns most similar to those of the test data is identified by using the average linkage method, the forecasting model trained using this cluster is applied for sales forecasting. We combined three clustering techniques (i.e., SOM, GHSOM, and K-means) and two machine-learning techniques (i.e., SVR and ELM) to construct six clustering-based forecasting models, which are called SOM-SVR, SOM-ELM, GHSOM-SVR, GHSOM-ELM, K-SVR, and K-ELM. The empirical retail data for notebook computers (NBs), personal computers (PCs), and liquid crystal displays (LCDs) from three computer retailers are collected and used as the numeric examples in the present study due to their dominance in computer product retailing market. They are generally the three highest priced products and the most crucial stock keeping units of computer retailers. The datasets of NBs, PCs, and LCDs are employed for evaluating the forecasting performance of the six clustering-based forecasting models and two single machine-learning techniques (i.e., single SVR and single ELM) without using clustering algorithm to partition training data. The forecasting accuracy of the six clustering-based forecasting schemes, single SVR, and single ELM is compared to identify whether the clustering-based forecasting models outperform the single machine-learning techniques and which of the six clustering-based forecasting models is the most appropriate scheme for computer retailing sales forecasting.

The rest of this paper is organized as follows. Section 2 gives a brief introduction about SOM, GHSOM, K-means, SVR, and ELM algorithms. The proposed clustering-based sales forecasting model is thoroughly described in Sect. 3. Section 4 presents the experimental results from three computer products sales data. The paper is concluded in Sect. 5.

2 Research methodology

2.1 SOM

The SOM algorithm proposed is a kind of artificial neural networks with unsupervised learning and referred to as a nonlinear, ordered, smooth mapping method for high-dimensional input data onto one- or two-dimensional display [39]. The fundamental principle of an SOM is to identify certain similar features, rules, or relations between unlabeled sample groups and group samples with similar patterns into the same category. SOM functions with competitive learning that earns activation opportunities through competition between neurons of the output layer. Different from the general competitive learning neurons, SOM rather relies on the principle of "reciprocity" competition.

The SOM network comprises a set of i units deployed in 2D grid with weight vector m_i, normally randomly initialized. A typical SOM architecture composed of input and output layers allows lateral interaction between the neurons to activate and inhibit one another. In each training session, when a neuron has a minimum Euclidean distance from the input vector x, the neuron represents the winning neuron expressed as W, as shown in the following [39]:

$$W(t) = \arg\min_i \{\|x(t) - m_i(t)\|\} \quad (1)$$

The weight vector of the winning neuron is incrementally adapted to the input signal vector of nearby winning neurons by a certain fraction of Euclidean distance, a time-decreasing leaning rate (α). Adaptation herein means a gradual reduction in relative element difference between input patterns and the vector model, as shown in Eq. (2).

$$m_i(t+1) = m_i(t) + \alpha(t) \cdot h_{Wi}(t) \cdot [x(t) - m_i(t)], \quad (2)$$

where t represents the current training iteration and x denotes input vector.

The amount of movement is controlled by the learning rate α. The principle for adjusting α is to make substantial adjustments in the initial learning stage of the network. When the learning time lengthens, α decreases gradually. The neighbor neurons near the winning neuron are expressed using neighbor kernel h_{Wi} to represent the distance between neuron i in the output space and winning neuron W of the cycle. The neighbor kernel is limited to a scalar quantity between one and zero to ensure the distance intensity adjusted by the nearby unit is larger than that of the remote units. A Gaussian function is commonly used, as shown in Eq. (3). Generally, when the distance to the winning neuron increases, the neighborhood function is a simple decreasing function surrounding the winning neuron [39].

$$h_{Wi} = \exp\left(-\frac{\|r_W - r_i\|^2}{2 \cdot \delta(t)^2}\right) \quad (3)$$

where $\|r_W - r_i\|^2$ represents the distance between W and i in the output space and r_i represents the two-dimensional vector unit in the grids. Time variant δ is neighborhood range. This learning procedure leads similar patterns to mapping into neighboring regions while dissimilar patterns are apart.

2.2 GHSOM

A GHSOM is hierarchical deployment of SOMs of various sizes which allows the size and dimensionality of its map to incrementally grow during the training process to adapt the training dataset based on the defined parameters. A GHSOM comprises multiple layers in hierarchical architecture. Instead of adding rows or columns to a SOM structure, each layer of GHSOM inserts a new independent SOM which maps the detailed patterns represented by a specific neuron. A GHSOM grows in two orientations and also is controlled by two parameters, τ_1 and τ_2, respectively. The former determines the growth of a map, whereas the latter dominates the hierarchical growth of the GHSOM [40]. The training and growing process mainly depends on the quantization error (QE) of a neuron which is an index of the error occurred in the mappings of the data onto a neuron. It is noted that the larger QE, the higher heterogeneity of the data cluster.

As to the basic GHSOM algorithm, the upmost layer of GHSOM (layer 0) contains a sole neuron which represents the mean of all input samples [41, 42]. The mean quantization error (MQE), referred as to a measurement of deviation of samples in the input space, can be obtained by Eq. (4)

$$\mathrm{MQE}_0 = \frac{1}{\Omega(X)} \cdot \sum_{x_j \in X} \|m_0 - x_j\| \quad (4)$$

where X is the set of all input samples, m_0 is the sole model vector of layer 0, and $\Omega(X)$ indicates the number of samples. Conforming to SOM learning algorithm, the offspring layers are hierarchically created below the ancestor layer after a predetermined iterations, and then the mean quantization errors for all units can be defined by Eq. (5)

$$\mathrm{MQE}_i = \frac{1}{\Omega(S_i)} \cdot \sum_{x_j \in W_i} \|m_i - x_j\| \quad (5)$$

where S_i is a subset of samples for unit i.

Since the MQE measures the dissimilarity between the input vector and a specific unit, high MQE values represent that the input space is not correctly clustered. The unit possessing the highest MQE is selected as an error unit υ, as shown in Eq. (6). Between the error unit υ and its most dissimilar neighbor d, a new column or row is added, resetting the learning rate and neighborhood ranges.

$$\upsilon = \arg\max_i \left(\sum_{x_j \in W_i} \|m_i - x_j\|\right) \quad (6)$$

The growing process continues until the MQE_m (i.e., the mean of all MQE_i values) reaches the fraction τ_1 of MQE_u (i.e., the MQE of the corresponding unit u in the upper layer), as shown in Eq. (7).

$$\mathrm{MQE}_m < \tau_1 \cdot \mathrm{MQE}_u \quad (7)$$

Please note that the smaller the τ_1 is and the longer the training time is, the larger the resulting map is. If the units of a completely trained map exhibit low similarity, the next layer of the map is continuously created. The threshold parameter of similarity between the units is τ_2. Equation (8) serves as the termination criterion to halt the growing process. If unit i satisfies the condition of Eq. (8), the next layer of expansion is not required; otherwise, a new map grows in the next layer.

$$\mathrm{MQE}_i < \tau_2 \cdot \mathrm{MQE}_o \quad (8)$$

It appears that the smaller the τ_2 is, the more easily the units expend to the next layer, the deeper hierarchical architecture a GHSOM has.

2.3 K-means

The K-means is one of the simplest and most efficient clustering algorithms [25]. The main idea of K-means clustering is to divide a set of data into mutually exclusive k clusters and assign each sample to the cluster whose center is nearest to the assigned sample, based on minimization of the squared error criterion function [43].

Initially, the k cluster centers are randomly designated among all the input samples. Then, a serial of local search is conducted to minimize the squared error between sample points and cluster centers and to obtain the optimum of Eq. (9)

$$E = \arg\min \sum_{i=1}^{n} \sum_{j=1}^{k} \omega_{ij} \left\| x_i - \theta_j \right\|^2 \tag{9}$$

where n is the size of data samples, k is the predetermined number of clusters, x_i is the ith sample point, θ_j is the center of cluster j, and ω_{ij} is the affiliation element which specifies the x_i cluster membership, given w_{ij} is

$$\omega_{ij} = \begin{cases} 1, & \text{if } \left\| x_i - \theta_j \right\| \le \left\| x_i - \theta_m \right\|, \forall m \ne j \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

subject to $\sum_{j=i}^{k} \omega_{ij} = 1, \quad i = 1, 2, \ldots, n \quad \text{and} \quad \sum_{i=1}^{n} \sum_{j=1}^{k} \omega_{ij} = n$.

2.4 SVR

Support vector regression (SVR) based on the principle of structural risk minimization is a machine-learning algorithm. The concept of SVR involves converting low-dimensional nonlinear regression problems into high-dimensional linear regression problems. The basic function of SVR can be expressed as the following equation:

$$f(x) = (w \cdot \phi(x)) + b \tag{11}$$

where w is weight vector, b is bias, and $\phi(x)$ is a kernel function which use a nonlinear function to transform the nonlinear input to be linear mode in a high dimension feature space. Traditional regression gets the coefficients through minimizing the square error which can be considered as empirical risk based on loss function. Vapnik [27] introduced so-called ε-insensitivity loss function to SVR. It can be expressed as:

$$L_\varepsilon(f(x) - y) = \begin{cases} |f(x) - y| - \varepsilon & \text{if } |f(x) - y| \ge \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where y is the target output, ε is the region of ε-insensitivity, and when the predicted value falls into the band area, the loss is zero. Contrarily, if the predicted value falls out the band area, the loss is equal to the difference between the predicted value and the margin.

Considering empirical risk and structure risk synchronously, the SVR model can be constructed to minimize the following programming:

$$\text{Min} : \frac{1}{2} w^T w + C \sum_i \left(\xi_i + \xi_i^* \right)$$
$$\text{Subject to} \begin{cases} y_i - w^T x_i - b \le \varepsilon + \xi_i \\ w^T x_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \quad \xi_i^* \ge 0 \end{cases} \tag{13}$$

where $i = 1, 2, \ldots, n$ is the number of training data; $(\xi_i + \xi_i^*)$ is the empirical risk; $\frac{1}{2} w^T w$ is the structure risk preventing over-learning and lack of applied universality; C is modifying coefficient representing the trade-off between empirical risk and structure risk. Equation (13) is a quadratic programming problem. After selecting proper modifying coefficient (C), width of band area (ε), and kernel function (K), the optimum of each parameter can be resolved though Lagrange function. The general form of the SVR-based regression function can be written as [27]

$$f(x, w) = f(x, \alpha, \alpha^*) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x, x_i) + b, \tag{14}$$

where α_j and α_j^* are Lagrangian multipliers and satisfy the equality $\alpha_j \alpha_j^* = 0$; $K(x_i, x_i')$ is the kernel function. Any function that meets Mercer's condition can be used as the kernel function.

Although several choices for the kernel function are available, the most widely used kernel unction is the radial basis function (RBF) defined as [44] $K(x_i, x_j) = \exp\left(\frac{-\left\| x_i - x_j \right\|^2}{2\sigma^2} \right)$, where σ denotes the width of the RBF. Thus, the RBF is applied in this study as kernel function.

2.5 ELM

Extreme learning machine (ELM) proposed by Huang et al. [28] is a new learning method for single-hidden-layer feed-forward neural networks (SLFNs). An ELM is a simple, rapid, and efficient SLFN, which focuses on the input weight values of SLFNs being random. In other words, the parameters of the hidden layer nodes are selected randomly. After the hidden nodes parameters are chosen randomly, SLFN becomes a linear system where the output weights of the network can be analytically determined using simple generalized inverse operation of the hidden layer output matrices.

Consider N arbitrary distinct samples (x_i, t_i) where $x_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^T \in R^n$, and $t_i = [t_{i1}, t_{i2}, \ldots, t_{im}]^T \in R^m$. SLFNs with $\tilde{N}$ hidden neurons and activation function $g(x)$ can approximate N samples with zero error. This means that

$$\mathbf{H}\beta = T \tag{15}$$

where

$$H(w_1, \ldots, w_{\tilde{N}}, b_1, \ldots, b_{\tilde{N}}, x_1, \ldots, x_N) = \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_1 + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_N + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_N + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}};$$

$$\beta_{\tilde{N} \times m} = (\beta_1^T, \ldots, \beta_{\tilde{N}}^T)^t; \; T_{N \times m} = (T_1^T, \ldots, T_N^T)^t$$

where $w_i = [w_{i1}, w_{i2}, \ldots, w_{in}]^T$, $i = 1, 2, \ldots, \tilde{N}$, is the weight vector connecting the ith hidden node and the input nodes, $\beta_i = [\beta_{i1}, \beta_{i2}, \ldots, \beta_{im}]^T$ is the weight vector connecting the ith hidden node and the output nodes, b_i is the threshold of the ith hidden node, and $w_i \cdot x_j$ denotes the inner product of w_i and x_j. $\mathbf{H}$ is called the hidden layer output matrix of the neural network; the ith column of $\mathbf{H}$ is the ith hidden node output with respect to inputs $x_1, x_2, \ldots, x_N$.

Thus, the determination of the output weights (linking the hidden layer to the output layer) is as simple as finding the least-square solution to the given linear system. The minimum norm least-square (LS) solution to the linear system (i.e., Eq. 15) is [28]

$$\hat{\beta} = H^{\Psi} T \tag{16}$$

where $\mathbf{H}^{\Psi}$ is the Moore–Penrose generalized inverse of matrix $\mathbf{H}$. The minimum norm LS solution is unique and has the smallest norm among all the LS solutions.

The first step of ELM algorithm is randomly assign input weight w_i and bias b_i; Then, the hidden layer output matrix $\mathbf{H}$ is calculated; finally, one can calculate the output weight β, $\hat{\beta} = \mathbf{H}^{\Psi} T$, where $T = (t_1, \ldots, t_N)^t$. For the details of the ELM algorithm, see Huang et al. [28].

3 Proposed clustering-based sales forecasting scheme

This study uses clustering algorithm and machine-learning technique to propose a clustering-based forecasting model for computer retailing sales forecasting. The research scheme of the proposed methodology is presented in Fig. 1. As shown in Fig. 1, the proposed methodology consists of two phases: training and testing.

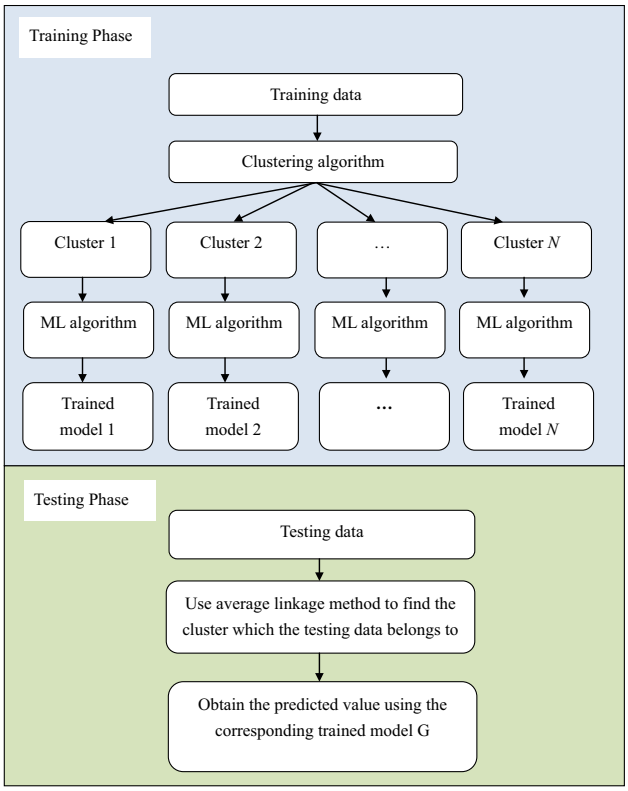

Fig. 1 Proposed clustering-based sales forecasting scheme

In the training phase, the purpose is to divide the overall training data and its complex data characteristics into multiple small training data sets having consistent data characteristics, as well as to train individual forecasting models for the clusters. The detailed procedure of the training phase can be summarized in the following steps:

1. First, historical sales data with a time length t are collected as the training data $X = [x_i]$, $i = 1, 2, \ldots, t$. Because historical sales data are favorable forecasting variables for sales data [4, 5], we use an appropriate historical data (i.e., window size) as the forecasting variables. Subsequently, we apply a moving window method to construct a forecasting variable matrix with the dimensions of $L \times q$, $q = t - L$ on data X under a predetermined window length (L; i.e., L forecasting variables) as follows:

$$\mathbf{X}_F = [f_1, f_2, \ldots, f_q] = \begin{bmatrix} x_L & x_{L+1} & \cdots & x_{t-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_2 & x_3 & \cdots & x_{t-L+1} \\ x_1 & x_2 & \cdots & x_{t-L} \end{bmatrix}$$

The corresponding target variable $Y = [y_1, y_2, \ldots, y_q] = [x_{L+1}, x_{L+2}, \ldots, x_t]$ features a dimension of $1 \times q$. For

example, when $t = 100$ and $L = 3$, then $q = 100–3 = 97$, $\mathbf{X}_F = [f_1, f_2, \ldots, f_{97}] = \begin{bmatrix} x_3 & x_4 & \cdots & x_{99} \\ x_2 & x_3 & \cdots & x_{98} \\ x_1 & x_2 & \cdots & x_{97} \end{bmatrix}$, and the corresponding target variable $Y = [y_1, y_2, \ldots, y_{97}] = [x_4, x_5, \ldots, x_{100}]$.

2. Then, in order to divide the whole forecasting data $\mathbf{X}_F$ into N clusters which possess consistent data characteristics, the clustering technique is used to partition f_i into N clusters, $i = 1, 2, \ldots, q$. Three types of clustering including K-means, SOM, and GHSOM algorithms are considered in this study.
3. Finally, the machine-learning technique is employed to train the forecasting model with the training data of each cluster. With N clusters, N forecasting models are trained. In this phase, the machine-learning techniques considered in this study are SVR and ELM.

The estimation accuracy of SVR and ELM may highly depend on the choice of parameters. However, there are no general rules for setting the parameters of SVR and ELM. For modeling SVR, the grid search proposed by Lin et al. [45] is a common and straightforward method using exponentially growing sequences of C and ε to identify good parameters. The parameter set of (C, ε, σ) which generate the minimum forecasting root mean square error (RMSE) is considered as the best parameter set. In this study, the grid search is used in each cluster to determine the best parameter set for training SVR forecasting model.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - e_i)^2}{n}} \tag{17}$$

where y_i and e_i represent the actual and predicted value at week i, respectively, and n is the total number of data points.

As discussed in Sect. 2.5, it is known that the most important and critical ELM parameter is the number of hidden nodes and that ELM tends to be unstable in single run forecasting [28]. Therefore, the ELM models with different numbers of hidden nodes varying from 1 to 30 are constructed. For each number of nodes, an ELM model is repeated 10 times and the average RMSE of each node is calculated. The number of hidden nodes that gives the smallest average RMSE value is selected as the best parameter of ELM model.

After the forecasting models trained using the clusters of training data, in the testing phase, the cluster with data patterns most similar to those of the test data is identified. And the trained forecasting model of the cluster is adopted to yield sales forecasting result. The detailed steps of the testing phase are described as follows:

1. If the sales data in time t (y_t') are the forecast target, the sales data from time $t - 1$ to $t - L$ are used as corresponding forecasting variable data $P = [y_i']$, $i = 1, 2, \ldots, L$. Note that L is number of forecasting variables.
2. Then, the average linkage method based on Euclidean distance is applied to measure the similarity between the test data and every cluster. That is, the Euclidean distances (d_i, $i = 1, 2, \ldots, L$) between the center of forecasting variable data P and the center of each cluster are computed, where d_i represents the Euclidean distance between the test data and the cluster i.
3. The cluster with minimal Euclidean distance (d_i) is the cluster which has the most similar data features or patterns to those of the test data. It is called cluster B, $B = arc\min(d_i)$. The trained forecasting model of cluster B is the most suitable model for predicting test data.
4. The predicted value of the test data is obtained using the trained forecasting model corresponding to the cluster B. The best parameter set of the forecasting model is determined in the training phase.

4 Empirical study

4.1 Empirical data and performance evaluation criteria

This study constructs six clustering-based forecasting models, namely SOM-SVR, SOM-ELM, GHSOM-SVR, GHSOM-ELM, K-SVR, and K-ELM. As the biweekly sales amount is more practical than daily and weekly sales amount for the sales and inventory management of computer retailers, the biweekly sales data for PC, NB, and LCD products of three computer retailers were collected and used as illustrative examples. The research data comprised 124 points of biweekly sales data from January 2005 to September 2009.

Figures 2, 3, and 4, respectively, show the trend of the sales for the PCs, NBs, and LCDs. From the figures, it can be observed that the sales data of each computer products exhibit similar data features in its different time periods. From the figures, it can be observed that the sales data of each computer products exhibit similar data features in its different time periods. However, the structure of the sales data for the three examined products differed. First, regarding PCs, the sales data revealed a stable sales performance from 2005 to 2006 (the prior 52 sample points). Thereafter, because of competition with other substitute products (e.g., Tablets and NBs) and changes to retailers' sales strategies, PC sales fluctuated drastically, generating a sales trend distinct from that prior to 2006. Consequently, a low level of similarity for the data structure of PCs at

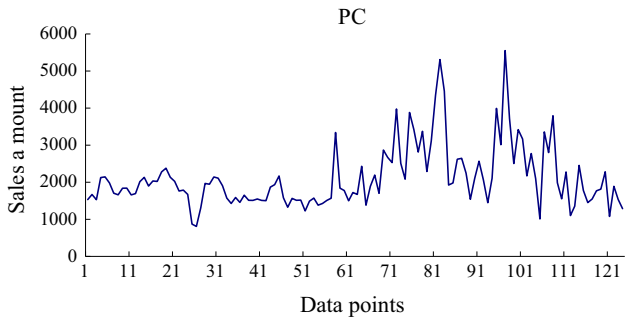

Fig. 2 PC sales amounts

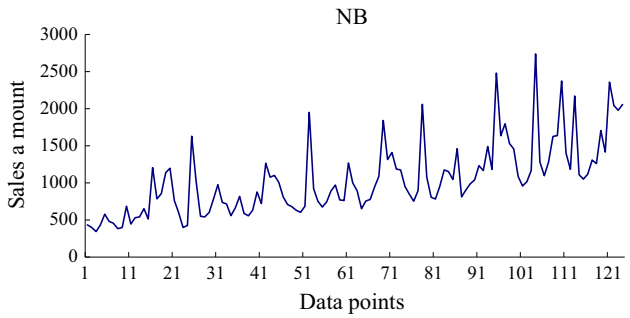

Fig. 3 NB sales amounts

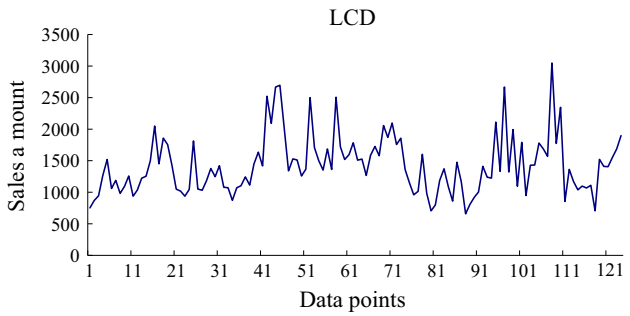

Fig. 4 LCD sales amounts

different periods was observed. Unlike the PC product, the NB and LCD products exhibited an obvious periodic sales trend and similar data patterns at different time points. Furthermore, compared with LCDs, the NB products were associated with a more apparent and stable data structure because changes in their product specification as well as their demand and sales characteristics are relatively constant.

The sales amounts of previous six periods (i.e., $t-1$, $t-2$, …, $t-6$) are used as six forecasting variables. Moreover, the first 88 data points (71 % of the total sample points) are used as the training sample, while the remaining 36 data points (29 % of the total sample points) are holdout and used as the testing sample for out of sample forecasting. The moving (or rolling) window technique is used to forecasting the training and testing data. All of the eight forecasting schemes are used for one-step-ahead forecasting of biweekly sales data.

Regarding the criteria for the forecasting performance evaluations, we use mean absolute percentage error (MAPE) and root mean square percentage error (RMSPE) to evaluate forecasting accuracy. A smaller value or small error indicated that the forecasting value and actual value were approximate. The definitions of these criteria are as follows:

$$\text{MAPE} = \frac{\sum_{i=1}^{n} \left| \frac{y_i - e_i}{y_i} \right|}{n}$$

$$\text{RMSPE} = \sqrt{\frac{\sum_{i=1}^{n} \left(\frac{y_i - e_i}{y_i} \right)^2}{n}}$$

where y_i and e_i represent the actual and predicted value at week i, respectively, and n is the total number of data points.

The SVR, ELM, GHSOM, and SOM analyses are conducted using MATLAB version 7.8.0 (R2009a) toolbox (MathWorks, Natick, MA, USA), and the K-means is performed using SPSS version 12.0 software (SPSS, Inc., Chicago, IL, USA).

4.2 Results of single SVR and single ELM models

In modeling the single SVR model, the whole training data are used and the grid search is applied for determining the best parameter set of (C, ε, σ). The parameter-searching scope of the three variables ranged from 2^{-15} to 2^{15}. The parameter set with minimal testing errors is the optimal parameter set. A list of the SVR testing errors of different parameter sets for the PCs is given in Table 1. As given in Table 1, the parameter set ($C = 2^{11}$, $\varepsilon = 2^{-13}$, $\sigma = 2^{9}$) provides a minimum testing RMSE and is considered the best parameter set for the single SVR model in forecasting sales for the PCs. When the grid search is also used for the NBs and LCDs, the best SVR parameter sets for the NBs and LCDs are ($C = 2^{-13}$, $\varepsilon = 2^{-15}$, $\sigma = 2^{9}$) and ($C = 2^{11}$, $\varepsilon = 2^{-15}$, $\sigma = 2^{-15}$), respectively.

Regarding the single ELM model, as mentioned in Sect. 3, we test the numbers of hidden nodes from 1 to 30 and repeat the test 10 times in each node for calculating average RMSE. Figure 5 shows the average RMSE values of the single ELM model with different numbers of hidden nodes. As shown in Fig. 5, the single ELM model with seven hidden nodes has the lowest average RMSE values and is therefore the optimal ELM model for forecasting sales of LCD. By following the same procedure, the appropriate number of hidden nodes of the single ELM for the NBs and LCDs is 9 and 5, respectively.

Table 1 Model selection results of the single SVR model for PCs

C	ε	σ	Testing RMSE
2^7	2^{-11}	2^7	875.64
	2^{-13}	2^9	875.64
	2^{-15}	2^{11}	875.64
2^9	2^{-11}	2^7	881.76
	2^{-13}	2^9	812.83
	2^{-15}	2^{11}	812.83
	2^{-11}	2^7	768.34
2^{11}	**2^{-13}**	**2^9**	**768.15**
	2^{-15}	2^{11}	768.76
2^{13}	2^{-11}	2^7	912.23
	2^{-13}	2^9	897. 21
	2^{-15}	2^{11}	887.47
2^{15}	2^{-11}	2^7	902.93
	2^{-13}	2^9	997.74
	2^{-15}	2^{11}	928.84

Bold values indicate the best parameter sets

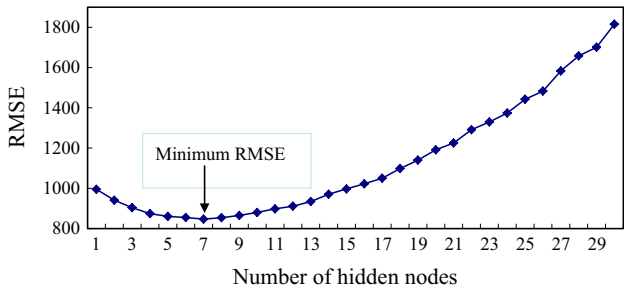

Fig. 5 Average RMSE values of the single ELM model for PCs with different numbers of hidden nodes

Table 2 Sales forecasting results for PCs, NBs, and LCDs using the single SVR and single ELM models

Products	Models	Metrics	
		MAPE (%)	RMSPE (%)
PC	Single SVR	36.67	48.31
	Single ELM	**25.50**	**34.12**
NB	**Single SVR**	**21.24**	**26.42**
	Single ELM	33.71	37.96
LCD	Single SVR	27.04	35.32
	Single ELM	**19.10**	**24.78**

Bold values indicate the best parameter sets

Table 2 shows a list of the forecasting results of the single SVR and single ELM models for the PC, NB, and LCD products. The table shows that the forecasting results of the single SVR model were superior regarding NB sales, whereas the single ELM model generated low forecasting errors for PC and LCD sales. In summary, the single ELM method had more satisfactory forecasting performance than that of single SVR model.

4.3 Results of the clustering-based schemes

In modeling six clustering-based forecasting models, the number of clusters is a critical parameter. An excess number of clusters lead to an overly low number of training data; thus, satisfactory forecasting models cannot be produced. By contrast, an excessively low number of clusters cause samples in the training data to contain features or patterns dissimilar to those of the testing data, which also leads to poor forecasting models. To obtain satisfactory forecasting results, each model tests two to six clusters

Table 3 Forecasting results of the six clustering-based forecasting models for PC sales

Product	Number of clusters	Models	MAPE	RMSPE
PC	None	Single SVR	36.67	48.31
		Single ELM	25.50	34.12
	2	GHSOM-SVR	24.89	30.75
		K-SVR	29.31	42.52
		SOM-SVR	26.03	36.42
		GHSOM-ELM	15.90	23.39
		K-ELM	23.83	33.91
		SOM-ELM	19.63	25.73
	3	GHSOM-SVR	24.14	37.76
		K-SVR	28.49	38.30
		SOM-SVR	27.87	41.05
		GHSOM-ELM	16.25	21.78
		K-ELM	20.83	27.71
		SOM-ELM	20.92	31.34
	4	GHSOM-SVR	20.58	29.17
		K-SVR	28.17	38.43
		SOM-SVR	26.74	39.49
		GHSOM-ELM	13.99	19.48
		K-ELM	18.18	25.80
		SOM-ELM	16.65	26.93
	5	GHSOM-SVR	18.04	22.16
		K-SVR	26.43	36.66
		SOM-SVR	26.71	38.79
		GHSOM-ELM	**7.42**	**9.21**
		K-ELM	17.83	24.26
		SOM-ELM	18.64	26.66
	6	GHSOM-SVR	14.72	19.42
		K-SVR	25.44	31.86
		SOM-SVR	24.56	36.34
		GHSOM-ELM	7.60	9.79
		K-ELM	20.99	29.23
		SOM-ELM	15.55	22.17

Bold values indicate the best parameter sets

when the clustering-based forecasting models are constructed. The cluster number with the minimal forecasting error is the optimal number of clusters. Moreover, during the construction of the six clustering-based forecasting models, the procedure used for selecting the optimal parameter of the SVR and ELM is adapted from the procedure used for the single SVR and single ELM models mentioned previously.

Table 3 shows the forecasting results of the six clustering-based forecasting models when different numbers of clusters were used. Regardless of the number of clusters used, the GHSOM-ELM model generates the best forecasting results and yields the lowest forecasting errors (MAPE, 7.42 %; RMSPE, 9.21 %) when five clusters are employed. Thus, based on the results given in Table 3, the GHSOM-ELM demonstrates the highest forecasting performance for PC product sales of all of the clustering-based forecasting models, including the single SVR and single ELM models.

Regarding NB products, the results of the six clustering-based forecasting models when using different numbers of clusters are given in Table 4. The GHSOM-ELM model yields promising forecasting results when using different numbers of clusters, except for with three clusters. In addition, the MAPE (11.43 %) and RMSPE values (15.24 %) of the GHSOM-ELM model using four clusters are the lowest. Thus, when using four clusters, the GHSOM-ELM model exhibits the most optimal forecasting performance superior to the other five clustering-based forecasting models, single SVR, and single ELM.

Table 4 Forecasting results of the six clustering-based forecasting models for NB sales

Product	Number of clusters	Models	MAPE (%)	RMSPE (%)
NB	None	Single SVR	21.24	26.42
		Single ELM	33.71	37.96
	2	GHSOM-SVR	17.46	22.23
		K-SVR	18.79	24.57
		SOM-SVR	24.38	30.74
		GHSOM-ELM	17.28	21.26
		K-ELM	16.49	20.36
		SOM-ELM	19.78	26.23
	3	GHSOM-SVR	20.24	24.37
		K-SVR	23.25	29.41
		SOM-SVR	21.36	27.35
		GHSOM-ELM	12.49	17.21
		K-ELM	18.08	23.48
		SOM-ELM	20.94	24.90
	4	GHSOM-SVR	15.45	20.14
		K-SVR	20.87	25.59
		SOM-SVR	33.33	39.63
		GHSOM-ELM	10.43	14.24
		K-ELM	15.59	19.06
		SOM-ELM	19.33	22.76
	5	GHSOM-SVR	13.97	19.55
		K-SVR	20.87	25.59
		SOM-SVR	23.09	29.33
		GHSOM-ELM	**12.30**	**16.72**
		K-ELM	15.59	19.06
		SOM-ELM	18.75	24.50
	6	GHSOM-SVR	17.57	22.09
		K-SVR	17.78	21.88
		SOM-SVR	23.42	27.47
		GHSOM-ELM	17.23	20.49
		K-ELM	18.47	22.31
		SOM-ELM	23.75	30.41

Bold values indicate the best parameter sets

Table 5 Forecasting results of the six clustering-based forecasting models for LCD sales

Product	Number of clusters	Models	MAPE (%)	RMSPE (%)
LCD	None	Single SVR	27.04	35.32
		Single ELM	19.10	24.78
	2	GHSOM-SVR	15.58	20.42
		K-SVR	25.11	35.35
		SOM-SVR	25.82	33.44
		GHSOM-ELM	11.63	16.29
		K-ELM	16.84	22.73
		SOM-ELM	19.16	25.82
	3	GHSOM-SVR	17.78	21.33
		K-SVR	20.52	26.76
		SOM-SVR	25.30	35.33
		GHSOM-ELM	11.46	15.20
		K-ELM	14.08	19.12
		SOM-ELM	17.34	21.35
	4	GHSOM-SVR	19.10	24.22
		K-SVR	20.32	25.53
		SOM-SVR	22.98	30.18
		GHSOM-ELM	11.15	13.74
		K-ELM	15.12	20.53
		SOM-ELM	15.95	19.84
	5	GHSOM-SVR	10.65	14.59
		K-SVR	21.22	28.28
		SOM-SVR	19.88	26.36
		GHSOM-ELM	**8.31**	**10.14**
		K-ELM	14.14	19.05
		SOM-ELM	12.52	17.00
	6	GHSOM-SVR	13.20	16.89
		K-SVR	17.09	22.99
		SOM-SVR	19.87	25.22
		GHSOM-ELM	11.44	14.63
		K-ELM	11.67	16.45
		SOM-ELM	11.48	14.95

Bold values indicate the best parameter sets

Therefore, this sales forecasting model is the most suitable scheme for NB products.

Table 5 shows the sales forecasting results of the six clustering-based forecasting models for LCD products when using different numbers of clusters. The GHSOM-ELM model also demonstrates the best forecasting performance when using any number of clusters. Moreover, the lowest MAPE (9.31 %) and RMSPE (11.41 %) values are observed when the cluster number is five, thereby creating a forecasting result superior to that of the other seven models. Therefore, as given in Table 5, the GHSOM-ELM model is suitable for forecasting LCD sales by using six clusters.

Overall, as given in Tables 3, 4, and 5, the forecasting errors of the GHSOM-ELM model are lower than those of the GHSOM-SVR, K-SVR, SOM-SVR, K-ELM, and SOM-ELM models, as well as those of the single SVR and single ELM models, for sales data of all three computer products. Using different numbers of clusters, the GHSOM-ELM model generated promising forecasting results for all three products, except for the NB product when using three clusters. These results demonstrate that the GHSOM-ELM model is a robust forecasting model.

Besides, in order to demonstrate the effective of the GHSOM-ELM model, the best forecasting results of each clustering-based model for PC, NB, and LCD products are summarized and compared in Table 6. Note that the number in the parentheses means the most suitable numbers of clusters for each clustering-based forecasting model. For example, for forecasting PC sales, GHSOM-SVR(6) indicates that six clusters can generate the best forecasting results when using GHSOM-SVR

model. From Table 6, it can be found that the GHSOM-ELM model yields the best forecasting results for forecasting the sales of the three computer products. Based on the findings discussed above, it can be inferred that the GHSOM-ELM model is suitable for computer retailing sales forecasting.

4.4 Significance test

For evaluating whether the proposed GHSOM-ELM model is superior to the GHSOM-SVR, K-SVR, SOM-SVR, K-ELM, and SOM-ELM in computer retailing sales forecasting, the Wilcoxon signed-rank test is employed. The test is a distribution-free, nonparametric technique that does not require any underlying distributions in the data, and deals with the signs and ranks of the values and not with their magnitude. It is one of the most commonly adopted tests in evaluating the predictive capabilities of two different models to see whether they are statistically significant different between them [4, 46, 47]. For the details of the Wilcoxon signed-rank test, please refer to Diebold and Mariano [46] and Pollock et al. [47].

Based on the forecasting results in Table 6, the test is used to evaluate the predictive performance of the six clustering-based forecasting models. Table 7 shows the Z statistic values of the two-tailed Wilcoxon signed-rank test for MAPE values between the GHSOM-ELM model and other five competing models, where the numbers in parentheses are the corresponding p values. It can be observed from Table 7 that the MAPE values of the GHSOM-ELM model are significantly different from the GHSOM-SVR, K-SVR, SOM-SVR, K-ELM, and SOM-ELM, except the GHSOM-SVR model in LCD product. It can be concluded that the GHSOM-ELM model significantly outperforms the other five clustering-based models for computer retailing sales forecasting.

Table 6 Comparison of the best forecasting results of the six clustering-based forecasting models for PC, NB, and LCD sales

Products	Models	MAPE (%)	RMSPE (%)
PC	GHSOM-SVR (6)	14.72	19.42
	K-SVR (6)	25.44	31.86
	SOM-SVR (6)	24.56	36.34
	GHSOM-ELM (5)	**7.42**	**9.21**
	K-ELM (5)	17.83	24.26
	SOM-ELM (6)	15.55	22.17
NB	GHSOM-SVR (5)	13.97	19.55
	K-SVR (6)	17.78	21.88
	SOM-SVR (3)	21.36	27.35
	GHSOM-ELM (4)	**10.43**	**14.24**
	K-ELM (4)	15.59	19.06
	SOM-ELM (5)	18.75	24.50
LCD	GHSOM-SVR (5)	10.65	14.59
	K-SVR (6)	17.09	22.99
	SOM-SVR (6)	19.87	25.22
	GHSOM-ELM (5)	**8.31**	**10.14**
	K-ELM (5)	11.67	16.45
	SOM-ELM (5)	11.48	14.95

Numbers of clusters in parentheses

Bold values indicate the best parameter sets

4.5 Robustness evaluation

To evaluate the robustness of the proposed GHSOM-ELM method, the performance of the six clustering-based forecasting models, single ELM model, and single SVR model was computed using different ratios of training and testing sample sizes. The testing plan is based on the relative ratio of the size of the training dataset size to complete dataset size. In this section, three relative ratios, 60, 70, and 80 %, are considered. Table 8 presents the prediction performance of all eight forecasting models for the three products (PC, NB, and LCD) at different relative ratios when MAPE was used as the indicator. Sections 4.2 and 4.3 describe the process of using the eight prediction models to predict the three products at a relative ratio of 70 %. We undertook the same procedure to generate prediction results for relative ratios of 60 and 80 %.

Table 7 Wilcoxon signed-rank test results between the GHSOM-ELM and the five competing clustering-based models by different computer products

Models	Products	GHSOM-SVR	K-SVR	SOM-SVR	K-ELM	SOM-ELM
GHSOM-ELM	PC	−1.999	−3.524	−3.025	−2.018	−2.005
		(0.048)*	(0.000)**	(0.000)**	(0.040)*	(0.046)*
	NB	−2.045	−2.653	−3.783	−2.057	−2.797
		(0.048)**	(0.008)**	(0.000)**	(0.048)*	(0.002)*
	LCD	−1.873	−2.877	−2.991	−1.998	−1.996
		(0.060)	(0.001)**	(0.001)**	(0.049)*	(0.049)*

p value in parentheses

** p value < 0.010

* p value < 0.050

Table 8 Robustness evaluation of the six clustering forecasting schemes, single ELM model, and single SVR model by different training and testing sample sizes

Relative ratio (%)	Models	MAPE (%)		
		PCs	NBs	LCDs
60	Single SVR	38.55	22.54	28.79
	Single ELM	35.23	35.66	20.95
	GHSOM-SVR	34.72	15.35	12.52
	K-SVR	35.24	19.74	18.82
	SOM-SVR	34.33	23.07	21.79
	GHSOM-ELM	**34.02**	**12.19**	**10.11**
	K-ELM	35.33	17.42	13.53
	SOM-ELM	34.55	20.18	12.92
70	Single SVR	36.67	21.24	27.04
	Single ELM	25.50	33.71	19.10
	GHSOM-SVR	14.72	13.97	10.65
	K-SVR	25.44	17.78	17.09
	SOM-SVR	24.56	21.36	19.87
	GHSOM-ELM	**7.42**	**10.43**	**8.31**
	K-ELM	17.83	15.59	11.67
	SOM-ELM	15.55	18.75	11.48
80	Single SVR	35.87	20.62	26.24
	Single ELM	24.51	33.07	18.59
	GHSOM-SVR	13.93	13.43	10.68
	K-SVR	24.47	16.95	16.36
	SOM-SVR	23.56	20.67	19.17
	GHSOM-ELM	**8.90**	**10.13**	**8.25**
	K-ELM	17.32	14.75	10.93
	SOM-ELM	15.61	18.14	10.94

Bold values indicate the best parameter sets

Table 8 reveals that, when predicting the three computer products at three different relative ratios by using the MAPE indicator, the GHSOM-ELM method generated the smallest prediction error compared with the other methods. This result indicates that this forecasting method outperformed the other seven methods. According to Table 8, compared with the other five clustering prediction techniques and the two single machine-learning methods, the proposed GHSOM-ELM method showed superior performance in predicting the NB and LCD products at three different relative ratios. However, for the PC product, although the GHSOM-ELM method significantly outperformed the other seven methods at relative ratios of 70 and 80 %, the prediction errors of this method and the other five clustering techniques did not differ considerably when the relative ratio was 60 %. Moreover, the prediction results of the GHSOM-ELM method were not evidently superior to those of the ELM. Figure 2 reveals a possible explanation for this result. As shown in Fig. 2, the trend chart of the PC product reveals that at a relative ratio of 60 %, the data pattern or structure of the training data (the prior 74 observations) is different from that of the testing data (the latter 50 observations). This may make the clustering results of the training phase of the proposed forecasting model cannot proper capture the pattern of the testing data. Therefore, the GHSOM-ELM method would perform similarly to the other five clustering-based forecasting models and would also obtain results similar to those of both single SVR and single ELM models. In other words, when the sales data show dissimilar data characteristics or patterns between training and testing datasets, the GHSOM-ELM method cannot outperform the other clustering prediction techniques. Subsequently, we reviewed the study conducted by Choi et al. [48] to further determine and analyze the prediction performance of the proposed clustering-based forecasting models when different sales data structures are taken into account. In the future, we will use appropriate indicators (e.g., auto correlations functions, ACF) and time-series analysis technique (e.g., wavelet transform) to extensively analyze the sales data of computer products and subsequently evaluate the applicability and validity of the proposed method.

5 Conclusion

Because of the rapid technological development, computer products are frequently replaced. Consequently, to compete with numerous competitors, computer retailers rely on accurate sales forecasting as the basis for effective management of marketing and inventories. This study used K-means, SOM, and GHSOM as three clustering techniques and a SVR and an ELM as two machine-learning techniques to construct six clustering-based forecasting models for computer product sales forecasting. The actual sales amounts for the PC, NB, and LCD products of three computer retailers were used as the empirical data. The results showed that the GHSOM-ELM model exhibited the most promising performance for forecasting the sales of three computer products when compared with the other five clustering-based forecasting models, single SVR, and single ELM. In addition, the GHSOM-ELM model is a robust sales forecasting model that generated the lowest forecasting errors regarding the data of the three computer products when using different numbers of clusters. Thus, the proposed GHSOM-ELM model is an effective sales forecasting model that is suitable for forecasting sales in a computer retail environment.

Acknowledgments This work is partially supported by the Ministry of Science and Technology of the Republic of China, Grant no. MOST 103-2221-E-231-003-MY2. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

1. Philip D, Alex A, Panagiotis P, Haralambos S (2006) Time series sales forecasting for short shelf-life food products base on artificial neural networks and evolutionary computing. J Food Eng 75:196–204
2. Luis A, Richard W (2007) Improved supply chain management based on hybrid demand forecasts. Appl Soft Comput 7:136–144
3. Thomassey S (2010) Sales forecasts in clothing industry: the key success factor of the supply chain management. Int J Prod Econ 128:470–483
4. Lu CJ, Lee TS, Lian CM (2012) Sales forecasting for computer wholesalers: a comparison of multivariate adaptive regression splines and artificial neural networks. Decis Support Syst 54:584–596
5. Lu CJ (2014) Sales forecasting of computer products based on variable selection scheme and support vector regression. Neurocomputing 128:491–499
6. Lu C-J, Shao YE (2012) Forecasting computer products sales by integrating ensemble empirical mode decomposition and extreme learning machine. Math Prob Eng. 2012:831201. doi:10.1155/2012/831201
7. Choi TM, Hui CL, Liu N, Ng SF, Yu Y (2014) Fast fashion sales forecasting with limited data and time. Decis Support Syst 59:84–92
8. Xia M, Wong WK (2014) A seasonal discrete grey forecasting model for fashion retailing. Knowl Based Syst 57:119–126
9. Thomassey S, Happiette M (2007) A neural clustering and classification system for sales forecasting of new apparel items. Appl Soft Comput 7:1177–1187
10. Chang PC, Liu CH, Wang YW (2006) A hybrid model by clustering and evolving fuzzy rules for sales decision supports in printed circuit board industry. Decis Support Syst 42:1254–1269
11. Hadavandi E, Shavandi H, Ghanbari A (2011) An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering: case study of printed circuit board. Expert Syst Appl 38:9392–9399
12. Sa-ngasoongsong A, Bukkapatnam STS, Kim J, Iyer PS, Suresh RP (2012) Multi-step sales forecasting in automotive industry based on structural relationship identification. Int J Prod Econ 140:875–887
13. Dai W, Wu J-Y, Lu C-J (2014) Applying different independent component analysis algorithms and support vector regression for IT chain store sales forecasting. Sci World J. 2014:438132. doi:10.1155/2014/438132
14. Tay FEH, Cao LJ (2001) Improved financial time series forecasting by combining support vector machines with self-organizing feature map. Intell Data Anal 5(4):339–354
15. Cao LJ (2003) Support vector machines experts for time series forecasting. Neurocomputing 51:321–339
16. Thomassey S, Fiordaliso A (2006) A hybrid sales forecasting system based on clustering and decision trees. Decis Support Syst 42(1):408–421
17. Lai RK, Fan CY, Huang WH, Chang PC (2009) Evolving and clustering fuzzy decision tree for financial time series data forecasting. Expert Syst Appl 36:3761–3773
18. Huang CL, Tsai CY (2009) A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. Expert Syst Appl 36:1529–1539
19. Badge J, Srivastava N (2010) Selection and forecasting of stock market patterns using K-mean clustering. Int J Stat Syst 5:23–27
20. Kumar M, Patel NR (2010) Using clustering to improve sales forecasts in retail merchandising. Ann Oper Res 174:33–46
21. Chang PC, Lai CY (2005) A hybrid system combining self-organizing maps with case-based reasoning in wholesaler's new-release book forecasting. Expert Syst Appl 29:183–192
22. Chang PC, Liu CH, Fan CF (2009) Data clustering and fuzzy neural network for sales forecasting: a case study in printed circuit board industry. Knowl Based Syst 22:344–355
23. Lu CJ, Wang YW (2010) Combining independent component analysis and growing hierarchical self-organizing maps with support vector regression in product demand forecasting. Int J Prod Econ 128:603–613
24. Murlidha V, Menezes B, Sathe M, Murlidhar G (2012) A clustering based forecast engine for retail sales. J Digit Inform Manag 10:219–230
25. Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recogn Lett 31:651–666
26. Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10:988–999
27. Vapnik VN (2000) The nature of statistical learning theory. Springer, Berlin
28. Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing 70:489–501
29. Kao LJ, Chiu CC, Lu CJ, Chang CH (2013) A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting. Decis Support Syst 54:1228–1244
30. Lu CJ (2013) Hybridizing nonlinear independent component analysis and support vector regression with particle swarm optimization for stock index forecasting. Neural Comput Appl 23:2417–2427
31. Bao Y, Xiong T, Hu Z (2014) Multi-step-ahead time series prediction using multiple-output support vector regression. Neurocomputing 129:482–493

32. Xiong T, Bao Y, Hu Z (2014) Multiple-output support vector regression with a firefly algorithm for interval-valued stock price index forecasting. Knowl Based Syst 55:87–100
33. Hong WC (2012) Application of seasonal SVR with chaotic immune algorithm in traffic flow forecasting. Neural Comput Appl 21:583–593
34. Ju FY, Hong WC (2013) Application of seasonal SVR with chaotic gravitational search algorithm in electricity forecasting. Appl Math Model 37:9643–9651
35. Wu CL, Chau KW (2013) Prediction of rainfall time series using modular soft computing methods. Eng Appl Artif Intell 26:997–1007
36. Sun ZL, Choi TM, Au KF, Yu Y (2008) Sales forecasting using extreme learning machine with applications in fashion retailing. Decis Support Syst 46:411–419
37. Wong WK, Guo ZX (2010) A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm. Int J Prod Econ 128:614–624
38. Xia M, Zhang Y, Weng L, Ye X (2012) Fashion retailing forecasting based on extreme learning machine with adaptive metrics of inputs. Knowl Based Syst 36:253–259
39. Kohonen T (1989) Self-organization and associative memory, 3rd edn. Springer, Berlin
40. Palomo EJ, North J, Elizondo D, Luque RM, Watson T (2012) Application of growing hierarchical SOM for visualisation of network forensics traffic data. Neural Netw 32:275–284
41. Dittenbach M, Rauber A, Merkl D (2002) Uncovering hierarchical structure in data using growing hierarchical self-organizing map. Neurocomputing 48:199–216
42. Chattopadhyay M, Dan PK, Mazumdar S (2014) Comparison of visualization of optimal clustering using self-organizing map and growing hierarchical self-organizing map in cellular manufacturing system. Appl Soft Comput 22:528–543
43. Johnson RA, Wichern DW (1992) Applied multivariate statistical analysis. Prentice hall, New Jersey
44. Cherkassky V, Ma Y (2004) Practical selection of SVM parameters and noise estimation for SVM regression. Neural Netw 17:113–126
45. Lin CJ, Hsu CW, Chang CC (2003) A practical guide to support vector classification, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei
46. Diebold FX, Mariano RS (1995) Comparing predictive accuracy. J Bus Econ Stat 13:253–263
47. Pollock AC, Macaulay A, Thomson ME, Onkal D (2005) Performance evaluation of judgemental directional exchange rate predictions. Int J Forecast 21:473–489
48. Choi TM, Yu Y, Au KF (2011) A hybrid SARIMA wavelet transform method for sales forecasting. Decis Support Syst 51(1):130–140