IChemE
www.icheme.org/journals
doi: 10.1205/cherd.04246

0263–8762/05/$30.00+0.00

Trans IChemE, Part A, August 2005
Chemical Engineering Research and Design, 83(A8): 1009–1018

CUSTOMER DEMAND FORECASTING VIA SUPPORT VECTOR REGRESSION ANALYSIS

A. A. LEVIS and L. G. PAPAGEORGIOU*

Centre for Process Systems Engineering, Department of Chemical Engineering, UCL (University College London), London, UK

This paper presents a systematic optimization-based approach for customer demand forecasting through support vector regression (SVR) analysis. The proposed methodology is based on the recently developed statistical learning theory (Vapnik, 1998) and its applications on SVR. The proposed three-step algorithm comprises both nonlinear programming (NLP) and linear programming (LP) mathematical model formulations to determine the regression function while the final step employs a recursive methodology to perform customer demand forecasting. Based on historical sales data, the algorithm features an adaptive and flexible regression function able to identify the underlying customer demand patterns from the available training points so as to capture customer behaviour and derive an accurate forecast. The applicability of our proposed methodology is demonstrated by a number of illustrative examples.

Keywords: customer demand forecasting; support vector regression; statistical learning theory; nonlinear optimization.

INTRODUCTION

Recent studies have clearly identified customer demand as the ultimate driver of business management in process industries ranging from the traditional oil and gas industry (Lasschuit and Thijssen, 2004) to the high-risk agrochemical and pharmaceutical industries (Maravelias and Grossmann, 2001; Levis and Papageorgiou, 2004). Given the importance of customer demand, one can easily realize the potential benefits of an accurate customer demand forecasting tool in process industries.

Forecasting has gained widespread acceptance as an integral part of business planning and decision-making in areas such as sales planning, marketing research, pricing, production planning and scheduling (Makridakis and Wheelwright, 1978). From a historical perspective, exponential smoothing methods and decomposition methods were the first forecasting approaches to be developed back in the mid-1950s. During the 1960s, as computer power became more available and cheaper, more sophisticated forecasting methods appeared. Box–Jenckins (Box and Jenckins, 1969) methodology gave rise to the autoregressive integrated moving average (ARIMA) models. Later on during the 1970s and 1980s, even more sophisticated forecasting approaches were developed including econometric methods and Bayesian methods (Makridakis and Wheelwright, 1982). The consolidation and improvement of the aforementioned approaches provided forecasting tools of ever increasing complexity, before artificial neural networks (ANN) emerged as a novel and promising forecasting approach in the 1990s taking full advantage of the number-crunching capabilities of super-computers (Foster *et al.*, 1992).

However, it is very interesting to notice that the increasing complexity of forecasting approaches is not always accompanied by the desired increased predictive accuracy as pointed out by Makridakis and Hibon (1979). Their remark is consistent with recent criticism towards the excessive use of parameters in unnecessarily complex ANN applications in chemical engineering problems (Bhat and McAvoy, 1992). There exists a clearly identified need for a new generation of forecasting tools that share all the benefits of ANN while at the same time maintain the underlying formulation as simple as possible.

Support vector machines (SVM) constitute such a novel learning paradigm that provides an inherently simple formulation and yet offers the promise of increased flexibility. The growing popularity of the SVM is mainly attributed to the solid theoretical foundations and the practical applications in a broad range of the scientific spectrum. Based on the statistical learning theory recently developed by Vapnik (1998), SVM applications have been proposed for a number of classification and regression problems ranging from discrete manufacturing (Prakasvudhisarn *et al.*, 2003) to bioinformatics

*Correspondence to: Dr L. G. Papageorgiou, Centre for Process Systems Engineering, Department of Chemical Engineering, UCL (University College London), London WC1E 7JE, UK.
E-mail: l.papageorgiou@ucl.ac.uk

(Myasnikova *et al.*, 2002). Agrawal *et al.* (2003) portray SVM as a useful tool for process engineering applications while Kulkarni *et al.* (2004) and Chiang *et al.* (2004) provide support vector classification applications in process engineering problems.

However, in order to present a balanced perspective, we must also mention that so far the applicability of support vector regression (SVR) is hindered by the notorious problem of parameter selection. Although, the number of parameters to be tuned is not prohibitively large, parameter values affect significantly the predictive capabilities. Exhaustive grid-search (Chang and Lin, 2001) and heuristic-based rules for parameter selection (Cherkassky and Ma, 2004) are currently used for SVR while further research in this area is in progress. As a typical case of any emerging forecasting research field, those heuristic rules can be regarded an initial step towards the identification of a more formal way for parameter selection in the near future.

Despite the parameter selection problem, SVR still enjoys numerous advantages when compared with other forecasting methodologies. Similar to ANN, SVR employs an adaptive basis regression function without postulating any pre-determined family of basis functions (e.g., high-order polynomial parametric regression). Support vectors provide a completely new way of parameterization of the adaptive function (Cherkassky and Mulier, 1998) leading to increased flexibility while avoiding the trap of over-complexity. Unlike ANN, SVR employs only a handful of parameters while its unique mathematical formulation guarantees a computationally tractable global optimal solution. This is a very attractive feature for applications in the process industries where repeatability and consistency are of paramount importance. Furthermore, SVR requires no *a priori* fundamental understanding of the process being studied since it is a training data-driven methodology and therefore is very well-suited for process industry forecasting applications where historical data is abundant.

Overall, support vector regression is identified as a novel emerging forecasting technique and the aim of this paper is to validate the applicability of SVR analysis for forecasting customer demand in process industries. The rest of the paper is organized as follows. In the next section, the main characteristics of the customer demand forecasting problem are described. The SVR section provides a brief mathematical description of support vector regression analysis. A three-step algorithm is then described in the Proposed Forecasting Algorithm section before it is validated through a number of illustrative examples in the following section. Finally, some concluding remarks are drawn in the last section of the paper.

PROBLEM DESCRIPTION

We assume that customer demand at time period t (y_t) depends on a number of z independent variables (x_{1t}, x_{2t}, $\ldots$, x_{zt}) that are called attributes and form the associated input vector x_t.[1] Therefore, the dependant variable y_t is a function of the input vector $\boldsymbol{x}_t$ which in turns contains the multiple independent variables as shown in the following equations.

$$\boldsymbol{x}_t = \begin{pmatrix} x_{1t} \\ x_{2t} \\ . \\ . \\ . \\ . \\ . \\ x_{zt} \end{pmatrix} \tag{1}$$

$$y_t = F(\boldsymbol{x}_t) \tag{2}$$

The problem of customer demand forecasting via SVR analysis can be formally stated as follows. Given a set of training data (time series) in the form of N training points $(\boldsymbol{x}_1, y_1)$, $(\boldsymbol{x}_2, y_2)$, $\ldots$, $(\boldsymbol{x}_N, y_N)$ where $\boldsymbol{x}_t$ is the input vector and y_t is the associated customer demand for every $\boldsymbol{x}_t$, as well as a forecasting horizon of size M, we would like to *determine* the output values $\hat{y}_{N+1}$, $\hat{y}_{N=2}$, $\ldots$, $\hat{y}_{N+M}$. The mean absolute percent error (MAPE) is a commonly used forecasting error metric for quantifying and assessing the accuracy of the predicted output values. Mathematically, it is given by the following formula (Makridakis and Wheelwright, 1978):

$$\mathrm{MAPE} = 100 \cdot \frac{\sum_{t=N+1}^{N+M} |(y_t - \hat{y}_t)/y_t|}{M} \tag{3}$$

where y_t is the actual customer demand and $\hat{y}_i$ is the predicted demand at time period t. Clearly, accurate predictions would result in low MAPE values, which implies small absolute deviations between the actual and predicted output values.

Based on the available training points $(\boldsymbol{x}_t, y_t)$, the ultimate goal of support vector regression analysis is to extract as much information as possible from the historical data so as comprehend the complicating relationships between customer demand and all the different attributes before identifying an appropriate regression function F able to accurately predict future unknown output values from a given input vector of attributes.

In our time series forecasting problem, customer demand attributes can be classified into a number of different main categories such as:

- *Past demand attributes*: those attributes represent customer demand for a predetermined number of previous time periods. Employing past demand attributes can be extremely helpful to relate present customer demand with historical customer demand values. According to our experience, past demand attributes prove to be very efficient when dealing with periodical customer demand patterns.
- *Calendar attributes*: those attributes illustrate a specific characteristic of the time period under investigation and are usually treated as binary parameters representing true or false statements with one and zero values respectively. For example, calendar attributes could be

[1]Symbols in bold fonts represents vectors.