Time Series Data Forecasting

Vincent Cho
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

Businesses are recognizing the value of data as a strategic asset. This is reflected by the high degree of interest in new technologies such as data mining. Corporations in banking, insurance, retail, and healthcare are harnessing aggregated operational data to help understand and run their businesses (Brockett et al., 1997; Delmater & Hamcock, 2001). Analysts use data-mining techniques to extract business information that enables better decision making (Cho et al., 1998; Cho & Wüthrich, 2002). In particular, time series forecasting is one of the major focuses in data mining. Time series forecasting is used in a variety of fields, such as agriculture, business, economics, engineering, geophysics, medical studies, meteorology, and social sciences. A time series is a sequence of data ordered in time, such as hourly temperature, daily stock prices, monthly sales, quarterly employment rates, yearly population changes, and so forth.

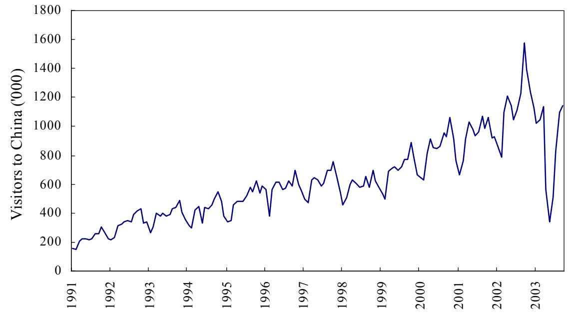
BACKGROUND

The objective of studying time series is to identify the pattern of how a sequence of data changes over time, and, thus, future forecasting can be made to help in scientific decision making.

The typical time series forecasting applications are related to economics, finance, and business operations. Data on economic time trends like GDP and tourist arrivals (Cho, 2001, 2003); financial time trends such as stock indices (Cho et al., 1999; Cho & Wuthrich, 2002; Wuthrich, et al., 1998), exchange rates; and business operations on inventory management, yield management (Choi & Cho, 2000), staff planning (Cho & Ngai, 2003), customer demands and spending patterns (Cho & Leung, 2002), telecommunication traffic (Layton et al., 1986), and marketing (Nijs et al., 2001; Dekimpe & Hanssens, 2000) are common forecasting domains. Figure 1 shows a typical time series that has obvious periodical pattern with some disturbances. The pattern can be captured by time series analysis techniques.

In order to have a more reliable forecasting of a time series, usually the time series need to be under a stable environment, and extensive underlying factors determining the time series should be included in the analysis.

Figure 1. Visitors to China 1991-2004



Moreover, an adequate training data set should be captured for the model building, and the model should be retrained with a moving window, which covers most of the recent cases.

MAIN THRUST

The common techniques for time series forecasting are exponential smoothing, ARIMA, transfer functions, Vector Auto-Regression (VAR), and Artificial Neural Network (ANN). The interrelationship among time series is usually described by the cross-correlation. In this article, ARIMA and ANN are presented for time series studies. These two techniques are selected because they are quite different in their natures. ARIMA was developed based on theories of mathematics and statistics, whereas ANN was developed based on the inspiration of nerve structure in human brains. Details are described as follows.

ARIMA

ARIMA models are flexible and widely used in time-series analysis. ARIMA (AutoRegressive Integrated Moving Average) combines three types of processes: Auto Regression (AR), differencing to strip off the integration (I) of the series and moving averages (MA). Each of the three types of processes has its own characteristic way of responding to a random disturbance.

Identification is a critical step in building an $ARIMA(p, d, q)(sp, sd, sq)_L$ model, where p is the AR order that indicates the number of coefficients of AR process, *d* is the number of times the data series must be differenced to induce a stationary series *Z*, *q* is the MA order that indicates the number of coefficients of the MA process, *sp* is the seasonal AR order that indicates the number of coefficients of seasonal AR process, and *sq* is the seasonal MA order that indicates the number of coefficients of seasonal MR process, *sd* is the number of times the data series needs to be seasonally differenced to induce a seasonally stationary series, and *L* indicates the seasonal periodicity.

These parameters usually are determined by inspecting the behavior of the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) (Box et al., 1994). The ACF and PACF of a stationary series should show either a cutoff or rapidly dying pattern. In practice, the determination of *d* and *sd* requires guessing different combinations among the possible values, until the desired patterns of ACF and PACF are achieved. Next will be the identification of the parameters, *p* and *q*, which involve the study of the behavior of the ACF and PACF. On these procedures, we can establish a tentative ARIMA model. However, all parameters are determined by observation and subjective guessing, which is rather unreliable and inaccurate.

Traditionally, identification is a rough procedure applied to a set of data to indicate the kind of representational model that is worthy of further investigation. The specific aim here is to obtain some idea of the values of *p*, *d*, and *q* needed in the general linear ARIMA model, and to obtain estimates for the parameters.

Parameter Searching Algorithm

Upon the previous drawback in estimating the parameters of the ARIMA model, an algorithm (Cho, 2003) to find the best combination of parameters is devised as follows:

This algorithm tries all combinations of parameters, which are limited to an integer lying between zero and two. The combination with the least Akaike AIC will be searched. With such range of parameter searching, the algorithm generates $3^6 = 729$ combinations. The range limitations of the parameters are set to restrict the search to a reasonable scope. Parameters greater than two make a model become too complicated, and the forecasting ability of a complicated model is seldom better than one with less coefficients. For example, for a model with p=6, q=5, sp=4, and sq=3, there would be 18 coefficients that would have to be estimated in the model, which can hardly be interpreted. Even if a complicated model is slightly better than a simple one in terms of accuracy, the simple one often is chosen because of its simplicity. Therefore, parameters greater than two are rarely used in practice.

For example, the series of visitors to China shown in Figure 1 was modeled as AR order *p=1*, MA order *q=0*, *sp=0* and *sq=1* with differencing *d=0* and seasonal differencing *sd=1*. The corresponding AIC is the lowest among all different combinations of parameters. Moreover, the solution space was restricted so that the estimated coefficients are all within a predetermined confidence limit of 95%.

Figure 2. Algorithm of finding ARIMA parameters

```
For p, d, q, sp, sd, and sq each = 0 to 2
Do
   Execute ARIMA with the set parameters.
   Record the parameters and corresponding fitting error.
Until all possible combinations are tried.
Report the parameters that produce the least AIC.
```

Artificial Neural Network (ANN)

Artificial Neural Networks are computing devices inspired by the function of nerve cells in the brain. They are composed of many parallel, interconnected computing units. Each of these performs a few simple operations and communicates results to its neighboring units. In contrast to conventional computer programs, where step-by-step instructions are provided to perform a particular task, neural networks can learn to perform tasks by a process of training on many different examples.

Typically, the nodes of a neural network are organized into layers, with each node in one layer having a connection to each node in the next layer, as shown in Figure 3. Associated with each connection is a weight, and each node has an activation value. During pattern recognition, each node operates as a simple threshold device. A node sums all the weighted inputs by multiplying the connection weights with the state of the previous layer nodes, and then the sum will be applied to a typically non-linear activation function. If the result is greater than the threshold value, the node will be activated. The result of the output nodes will be compared with the known result in the training set. The error terms will be fed backward for weighting adjustment in the hidden layers, so as to make the neural network resemble the training set more.

Neural networks provide a relatively easy way to model and forecast non-linear systems. This gives them an advantage over many current statistical methods used in business and finance, which are primarily linear. They also are very effective in learning cases that contain noisy, incomplete, or even contradictory data. The ability to learn and the capability to handle imprecise data make them very effective in handling financial and business information. A main limitation of neural networks is that they lack explanation capabilities. They do not provide users with details of how they reason with data to arrive

Related Content