Available online at www.sciencedirect.com
SCIENCE DIRECT
Int. J. Production Economics 84 (2003) 231–253
international journal of production economics
www.elsevier.com/locate/dsw

Efficacy of end-user neural network and data mining software for predicting complex system performance

Paul F. Schikora[a,*], Michael R. Godfrey[b]

[a] Analytical Department, Indiana State University, Terre Haute, IN 47809-0001, USA
[b] College of Business Administration, University of Wisconsin Oshkosh, Oshkosh, WI 54901-8675, USA

Received 3 March 2001; accepted 25 October 2002

Abstract

The performance of a university's dial-up modem pool under various time limit policies and customer behavior patterns was studied. Because the system is very complex, simulation offered the only method to obtain a limited set of steady-state performance measure estimates. A more generalized predictive model must be built from the simulated output. Traditional methods available to practitioners for predicting system performance across a range of environmental and decision variables have typically been limited to linear regression models. However, when the system being studied is highly complex and its performance is nonlinear in nature, the effectiveness of linear models can be limited. While more advanced nonlinear methods, such as neural networks, have been shown to perform better than traditional regression analysis in these situations, the knowledge needed to implement them "from scratch" is beyond most practitioners. Fortunately, these advanced methods are now available in ready-to-use desktop software programs, making them more attainable for practitioner use. The efficacy of these end-user programs compared to more traditional methods in practice is of interest.

Multiple variable linear regression models were developed for predicting six output measures in a simulation study and were compared to nonlinear regression models developed using a data mining software package (PolyAnalyst 4.3 Evaluation Software from Megaputer Intelligence) and two commercial neural network software packages (Statistica Neural Networks from Statsoft, and Predict from NeuralWorks). Comparisons of the models' predictive ability were made on both the data used to design the models and on a test set of data. Statistical analysis shows that predictive performance on the test data was usually best with one of the neural network models, but relative performance of the different models varied widely.

Keywords: Neural networks; Data mining; Regression analysis; Dial-up modem pools

1. Introduction

The current study stems from ongoing research into the performance of a major university's dial-up modem pool (DMP) system under various configurations and resource allocation policies. Simulation was

*Corresponding author: Tel.: +812-237-2092.
E-mail address: sdpaul@isugw.indstate.edu (P.F. Schikora).

0925-5273/03/$ - see front matter
doi:10.1016/S0925-5273(03)00003-3

used to examine system performance as time limits were imposed on customers in two alternative types of DMP systems. Of interest in the ongoing research is how the systems will perform across a wide range of configuration options and customer behavior patterns. To answer that question, different predictive models were built using the simulation output data. The performance measures in most cases proved to be nonlinear functions of the decision variables. While neural networks have been shown to generally perform better than regression on nonlinear data, the specialized skills needed to create and work with neural networks are generally beyond information systems managers. Recently, commercial desktop software programs that implement neural networks and other nonlinear modeling techniques, and which are easy to use, have become more widely available. These packages are purposefully generic in nature, and it is unknown whether these "one size fits all" end-user programs will perform as well as specialized neural networks have proven to. The predictive models were therefore built using traditional multiple linear regression, and commercially available data mining and neural network software packages. In this current study, we compare the predictive capability of the more traditional regression models to those generated by the end-user data mining and neural network software packages.

1.1. Physical model and problem description

Universities, corporations, and on-line network service providers (e.g., CompuServe) have long-provided external access to large central computer networks via modems. Modems allow data to be transmitted over standard analog telephone connections, providing long-distance access to the central computer(s) from any place that has a telephone connection and a modem-equipped computer. Although the data transfer rates through a modem-to-modem connection are much slower than a direct network connection, modems are essential for providing flexible external access to a central network for most users.

Typically, the operator of a central computer network will establish a DMP, a collection of modems that can all be accessed via a single telephone number. Each modem can provide a connection to a single user at any one time. Therefore, the maximum number of users that can be connected via a DMP at any given time is equal to the number of modems in the pool. When all of the modems are busy, the system is full, and a user attempting to dial into the system will be unable to immediately connect to the system.

Our ongoing research examines various configurations for an existing DMP system at a large university. This DMP system is characterized by full system usage for most of the 15 hour time period between 9:00 A.M. and 12:00 A.M. This usage is driven by 12% of the customers who consume approximately half of all available capacity (Schikora, 1999). In many queuing systems where service times are assumed to be exponentially distributed, some skewing of the service time is expected, but in this DMP system this characteristic is grossly exaggerated. It is reasonable to assume that these high consumption customers are wasting a great deal of capacity on nonproductive connect time.

DMP systems fit a general queuing model known as a retrial queue. Generally, a retrial queue (also called a queue with repeated calls, returning customers, repeated orders, etc.) describes a system that operates in the following manner: a customer arriving at the system when all servers are busy (i.e., the customer is blocked) leaves the service area but returns after some random time to repeat his demand (Falin, 1990). Previously blocked customers waiting to retry for service are said to be in orbit (or an orbit queue or a retrial queue). Obviously, there is no queue per se—the orbit is an artificial construct to account for the blocked customers who will be returning for service. As such, there is no queuing discipline in orbit, and determination of the next customer to be served follows a random process. Customers in orbit are in a sense competing with other orbiting customers and new arrivals for the next available server (modem).

A DMP system can further be defined by its availability, i.e., by the number of separate modem pools available. When there is a single pool (by definition accessed through a single telephone number), any free server (modem) in the system can be accessed by any customer. This type of system is called a full-available system. On the other hand, when the modems serving a local area are broken into m different pools, with