Data-Driven Simulation-Based Analytics

D

Durai Sundaramoorthi
Washington University in St. Louis, USA

INTRODUCTION

This chapter will introduce a novel data-driven simulation-based approach to study a dynamic system. Classification and regression trees - a data mining method – is utilized to extract non-linear dynamics of the system based on "big data" collected over a long period of time. Kernel density estimates are used at the terminal nodes of the tree to model the probability distribution of the response variable. Then the simulation model repeatedly samples responses from the kernel density estimates based on the present state of the system to mimic the evolution of the system. This way of simulation is completely data-driven and parameter free, it avoids misrepresentation of the system.

Application of this method is illustrated with two diverse applications. In the first application, nurse-activity is simulated based on a big data set collected at a northeast Texas hospital. Five tree structures are developed: (a) four classification trees from which transition probabilities for nurse movements are determined, and (b) a regression tree from which the amount of time a nurse spends in a location is predicted based on factors such as the primary diagnosis of a patient and the type of nurse.

In the second example, the dynamics of the financial market movement in the United States is estimated by simulating the Standard and Poor's 500 Index. Unlike any other indices, the Standard and Poor's 500 Index (S&P500) is the most commonly used index in the New York Stock Exchange to understand the market in the United States. It was introduced in its current form in March, 1957. The 500 publicly traded companies that make up the index are chosen by a committee to best reflect the overall market in the United States. The data set included the daily movement of financial markets in seven countries in Asia and Europe in relation to the daily movement of the S&P500. Trees also utilized data on the currency exchange rates to capture the financial dynamics between the US and other countries. The simulation model repeatedly samples from four trees to know how the opening and closing values of the S&P500 move in tandem with the other markets.

BACKGROUND

There are two major components in data-driven simulation-based analytics: data mining and simulation. In this section, a brief literature review on these topics and developments in nurse planning and S&P500 prediction are provided.

Learning from data can be broadly classified into two groups: supervised analytics and unsupervised analytics. In supervised analytics, an outcome variable is present to guide the learning process. In unsupervised analytics or clustering, one wants to observe only the features and have no measurements of the outcome. Supervised analytics is the subject of interest in this chapter as we deal with simulating the amount of time nurses spend with patients in the first application, the S&P500 open and close values in the second application, and the ground-level ozone concentration in the third application. Classification and Regression Trees (Breiman et al., 1984) - a data mining tool for prediction and classification – is used in this research for its readily usable tree structures in simulation. The readers without any background in CART are encouraged to read Shalizi, 2012 to get a quick idea about the methodology with

DOI: 10.4018/978-1-4666-5202-6.ch063

simple examples. Application of data mining tools to health care problems is quite common and has produced a significant amount of literature. For instance, recently, Ceglowski, & Churilov 2008, and Ceglowski, Churilov, & Wassertheil, 2005 used self organizing maps, a clustering technique, to determine treatment paths of emergency room patients and Ramon et al. 2007 used decision trees, first order random forests, naive Bayes, and tree augmented naive Bayes to predict patients' length of stay, patient survival, and endangering states. Similarly, application of data mining tools to financial modeling has been proven successful. For instance, in Huang, Nakamori, & Wang, 2005, Support Vector Machines (SVM) was used to model weekly movement direction of NIKKEI 225 index in tandem with S&P500. Their results show that SVM performs better than Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Neural Networks in predicting NIKKEI 225 Index. In another research, SVM was combined with self-organizing maps to model time series of currency exchange rates provided in Santa Fe Time Series Prediction Analysis Competition (Tay, & Cao, 2001). In Zhang, & Berardi, 2001, a neural network ensemble was used to predict exchange rates. This approach performed marginally better than the traditional random walk model. In a similar research (Tsai, Lin, Yen, & Chen, 2011), classifier ensembles consisting multi-layer perception neural network, classification and regression trees, and logistic regression were used to predict stock returns. This work also presents a good comparison - in terms of factors used for prediction and prediction results - from other relevant models.

Studying industrial systems using simulation was prevalent as early as the late 1950's and early 1960's. A comprehensive review of health care simulation models can be found in Klein, Dittus, Roberts, & Wilson, 1993, and Jun, Jacbson, & Swisher, 1999. In recent years, Zenios, Wein, & Chertow, 1999; Kreke, Schaefer, Angus, Bryce, & Roberts, 2002; and Shechter et al., 2005, utilized simulation models to study organ allocation systems. Similarly, simulation modeling has been widely used for different problems in finance. A comprehensive review of Monte Carlo simulation in finance can be found in Glasserman, 2004.

The simulation modeling approaches in the literature, both deterministic and stochastic, required the knowledge of experts to estimate parameters and order of events in the simulation. If the system under consideration is complex, such as nurse movement and financial market, then it is impossible even for the experts to comprehend the intricacies of the system by observation. Whereas, the data-driven simulation approach discussed in this chapter captures the system dynamics from a real data set collected from the system and requires only minimal subjective input from the experts.

MAIN FOCUS

The main focus of this chapter is to introduce a novel analytics tool for constructing efficient simulation models based on data mining. This way of simulation modeling avoids misrepresentation of system dynamics because it is entirely based on the pattern learned from a real data set collected from the actual system over a long period of time. Moreover, this approach reduces simulation states and is consequently more efficient to run. The rest of this section explains how this approach is developed by efficiently utilizing data mining and simulation.

CLASSIFICATION AND REGRESSION TREES

Classification and regression trees (cart) is a data mining technique for prediction and classification. Cart utilizes recursive binary splitting to uncover structure in a high dimensional space. Cart, on application to a data set, will partition the input space into many disjoint sets, where values within a set have a more similar response measure than values in different sets (breiman, 1984). Salford systems'