International Journal of Forecasting 34 (2018) 89–104

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Benchmarking robustness of load forecasting models under data integrity attacks

Jian Luo [a], Tao Hong [a,b,*], Shu-Cherng Fang [c]

[a] School of Management Science and Engineering, Dongbei University of Finance and Economics, Dalian, China
[b] Department of Systems Engineering and Engineering Management, University of North Carolina at Charlotte, NC, USA
[c] Department of Industrial & Systems Engineering, North Carolina State University, NC, USA

ARTICLE INFO

Keywords:
Cybersecurity
Data integrity attack
Electric load forecasting
Linear regression
Neural network
Support vector regression
Fuzzy regression

ABSTRACT

As the internet's footprint continues to expand, cybersecurity is becoming a major concern for both governments and the private sector. One such cybersecurity issue relates to data integrity attacks. This paper focuses on the power industry, where the forecasting processes rely heavily on the quality of the data. Data integrity attacks are expected to harm the performances of forecasting systems, which will have a major impact on both the financial bottom line of power companies and the resilience of power grids. This paper reveals the effect of data integrity attacks on the accuracy of four representative load forecasting models (multiple linear regression, support vector regression, artificial neural networks, and fuzzy interaction regression). We begin by simulating some data integrity attacks through the random injection of some multipliers that follow a normal or uniform distribution into the load series. Then, the four aforementioned load forecasting models are used to generate one-year-ahead ex post point forecasts in order to provide a comparison of their forecast errors. The results show that the support vector regression model is most robust, followed closely by the multiple linear regression model, while the fuzzy interaction regression model is the least robust of the four. Nevertheless, all four models fail to provide satisfying forecasts when the scale of the data integrity attacks becomes large. This presents a serious challenge to both load forecasters and the broader forecasting community: *the generation of accurate forecasts under data integrity attacks*. We construct our case study using the publicly-available data from Global Energy Forecasting Competition 2012. At the end, we also offer an overview of potential research topics for future studies.

1. Introduction

In the era of the internet of things, cybersecurity is of growing concern to governments, financial institutions, and many other business entities (Singer & Friedman, 2014). Among various cybersecurity issues, data integrity attacks, where hackers access supposedly protected data and inject false information, are of great importance to the forecasting community, because the quality of input data affects the forecast accuracy directly. Although the topic of "outlier detection" has been studied extensively in the literature (Akouemo & Povinelli, 2016; Hodge & Austin, 2004; Rousseeuw & Leroy, 2005; Xie & Hong, 2016), the problem of forecasting under data integrity attacks is still relatively new to the forecasting community. This paper focuses on the power industry, conducting an empirical study to test and benchmark the robustness of four representative load forecasting models under various simulated data integrity attacks using publicly available data.

The power industry operates the electric grid, one of the most complicated man-made systems in the world, to

* Corresponding author at: Department of Systems Engineering and Engineering Management, University of North Carolina at Charlotte, NC, USA.
E-mail address: hong@uncc.edu (T. Hong).

http://dx.doi.org/10.1016/j.ijforecast.2017.08.004
0169-2070/

produce and deliver electricity to over 5.6 billion people worldwide. Like many other industries, the power industry requires good forecasts of the electricity supply, demand and price in order to plan and operate the grid. Electric load forecasting has been an integral part of its business operations for over a century. Applications of load forecasting are spread across virtually every segment of the power industry (Hong, 2014). Both power companies' financial bottom lines and the resilience of power grids rely on accurate load forecasts.

Over the past several decades, researchers and practitioners have used a variety of techniques in their attempts to tackle the load forecasting problem (Hippert, Pedreira, & Souza, 2001; Hong & Fan, 2016; Weron, 2006). Some of these, such as artificial neural networks (Khotanzad & Afkhami-Rohani, 1998), semi-parametric models (Fan & Hyndman, 2012; Hyndman & Fan, 2010) and multiple linear regression models (Hong, 2010; Hong, Wilson, & Xie, 2014; Papalexopoulos & Hesterberg, 1990), have been used by power companies in the production environment, while others have excelled in notable load forecasting competitions. For instance, the support vector regression model won the EUNITE load forecasting competition in 2001 (Chen, Chang, & Lin, 2004); while regression models and gradient boosting machines took the top places in the load forecasting track of the Global Energy Forecasting Competition 2012, (GEFCom2012; see Hong, Pinson, & Fan, 2014).

The modern power grid relies heavily on communication networks and information technologies. Unfortunately, while such integration is essential for the evolution of the power grid, it also makes the grid more vulnerable to cyber-attacks by hackers around the world. For instance, a cyberattack to the supervisory control and data acquisition (SCADA) system of a Ukrainian power company disconnected seven substations for three hours (Perez, 2016). Thus, cybersecurity is an emerging field in power system research. While some researchers have studied data integrity attacks against "state estimation" (Hu & Vasilakos, 2016; Liu, Ning, & Reiter, 2011), little work has been done on data integrity attacks against "load forecasting". This paper attempts to tackle the problem of *load forecasting under data integrity attacks* by first benchmarking some of the existing models.

Addressing the issue fully requires extensive effort to be put into (1) detecting attacks, (2) identifying maliciously manipulated data, and (3) cleansing and recovering attacked data, before formally using a load forecasting system to generate forecasts. However, there could be many possible forms of malicious data integrity attacks, some of which may even be beyond our current understanding. An intelligent hacker might be able to inject false information without being detected by the state estimator (Liu et al., 2011) or load forecaster, which makes it extremely complex and technically difficult to assure that the data are attack-free all the time. The most relevant literature may be that on outlier detection and data cleansing in electric load forecasting (Xie & Hong, 2016). However, detecting and cleansing the attacked data points in load forecasting (Xie & Hong, 2016) and state estimation (Liu et al., 2011) is very difficult and expensive. Hence, it is imperative to first benchmark the performances of representative load forecasting models under possible data integrity attacks.

To establish a basic framework for benchmarking, we consider the case where hackers get hold of the historical load data and select a random set of data points (a given percentage of the whole data set) to multiply by a set of maliciously injected multipliers. The range of the multipliers is assumed to be either normally or uniformly distributed. These two types of data integrity attacks are called normally-distributed and uniformly-distributed data integrity attacks, respectively. Essentially, this simulation method can be viewed as injecting noise into the input data. It is worth noting that the "noise" in time series forecasting is natural and is usually small, while these simulated multipliers can reach large magnitudes, such as many times the original load values.

This study begins by simulating the corresponding data integrity attacks, then benchmarks the point forecast accuracy under various simulated data integrity attacks for the following four load forecasting models: multiple linear regression (MLR), artificial neural network (ANN), support vector regression (SVR), and fuzzy interaction regression (FIR). We select these four models for comparison because they are representative, in the sense that their characteristics range from black-box to non-black-box, statistical to fuzzy, and classical to emerging.

This paper contributes to the field in at least three ways: (1) it introduces an important emerging problem to the forecasting community, namely forecasting under data integrity attacks; (2) it proposes a systematic data integrity attack simulation framework for load forecasting; (3) it benchmarks and analyzes the robustness of four representative load forecasting models under data integrity attacks at various levels. Since the data used in this paper are accessible publicly, the results of this paper can be reproduced freely or used by readers directly for other benchmarking purposes.

The rest of the paper is arranged as follows. Section 2 provides an overview of the four representative load forecasting models. Section 3 introduces the settings of the benchmark study, including the GEFCom2012 data, the accuracy of the four load forecasting models without data integrity attacks, and the data integrity attack simulation framework. Section 4 presents the computational results of the four load forecasting models under two data attack scenarios and discusses the robustness of these models. Section 5 discusses several future research directions. Finally, the paper concludes in Section 6.

2. Four load forecasting models

This section introduces the four representative load forecasting models, namely MLR, ANN, SVR and FIR. Note that all of the models implemented in this study use temperature information for constructing explanatory variables. In other words, we did not use any of the well-known naïve (e.g., random walk and seasonal naïve) or time series (e.g., exponential smoothing and autoregressive integrated moving average) models for our comparisons, because load forecasting models that do not take weather inputs are of limited use in practice (Wang, Liu, & Hong, 2016).