Big Data Research 5 (2016) 9–15

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Machine Learning with Big Data
An Efficient Electricity Generation Forecasting System ☆

Mohammad Naimur Rahman, Amir Esmailpour *, Junhui Zhao

Department of Electrical & Computer Engineering and Computer Science, University of New Haven, West Haven, CT, 06516, United States

ARTICLE INFO

Article history:
Received 1 June 2015
Received in revised form 26 January 2016
Accepted 19 February 2016
Available online 24 March 2016

Keywords:
Artificial neural network
Backpropagation
Big Data
Electricity generation forecast
Hadoop
MapReduce

ABSTRACT

Machine Learning (ML) is a powerful tool that can be used to make predictions on the future nature of data based on the past history. ML algorithms operate by building a model from input examples to make data-driven predictions or decisions for the future. The growing concept "Big Data" has brought much success in the field of data science; it provides data scalability in a variety of ways that empower data science. ML can also be used in conjunction with Big Data to build effective predictive systems or to solve complex data analytic problems. In this work, we propose an electricity generation forecasting system that could predict the amount of power required at a rate close to the electricity consumption for the United States. The proposed scheme uses Big Data analytics to process the data collected on power management in the past 20 years. Then, it applies a ML model to train the system for the prediction stage. The model can forecast future power generation based on the collected data, and our test results show that the proposed system can predict the required power generation close to 99% of the actual usage. Our results indicate that the ML with Big Data can be integrated in forecasting techniques to improve the efficiency and solve complex data analytic problems existing in the power management systems.

1. Introduction

The United States (U.S.) is currently the second largest electricity producer and consumer in the world [1]. The U.S. enjoys a magnificent geographical diversity among states with a high amount of power consumption. This makes it challenging to deploy a centralized power management system that can control the power generation and regulate the consumption. The electricity is mostly generated from natural resources, such as coal, gas, nuclear, petroleum, oil, and renewable energy. The consumption sectors can be detailed in terms of commercial, industrial, residential and other user communities.

Due to lack of centralized control, there is a large disparity in the ratio of power consumption/power generation from one state to the next. This imbalance results in wasting large quantities of power generated in states where generation significantly exceeds consumption, while other states are suffering from insufficient amount of power generation. Due to the size and the geographical diversity of different states in the U.S., it is farfetched to prescribe centralized control over the power system. Merely, the interstate segments are regulated by the federal government [2,3], and the majority of the rest of the nation is delimited by individual states. Fig. 1 shows the electricity generation and consumption in the U.S. during 1980–2014. In this figure, the green line at the bottom shows the consumption, the red line in the middle represents the actual generation, and the blue line on top indicates total generation including net import (i.e. from neighboring countries). The difference between the generation (red line) and consumption (green line) is attributed to system losses, uncounted loads, and the lack of centralized control.

Fig. 2 shows electricity generation in the U.S., by state. States shown in lighter brown color are not producing enough electricity to meet their demand. Other states (shown in darker orange color) produce excess electricity, which could be used to compensate for the brown states lacking sufficient power generation. Further deficiencies are fulfilled by importing electricity from neighboring countries.

Power generation is in direct correlation with the amount of resources used to generate the electricity such as coal, gas, nuclear, petroleum, oil, and renewable energy. In Fig. 1, the red line in the middle (representing the power generation in the U.S.) provides two types of information: the amount of energy consumed and the quantity to be imported. Therefore, predicting power generation might provide vague information about power demand; hence

☆ This article belongs to Analytics & Applications.

* Corresponding author.
E-mail addresses: MRahm1@unh.newhaven.edu (M. Naimur Rahman), aesmailpour@newhaven.edu (A. Esmailpour), JZhao@newhaven.edu (J. Zhao).

http://dx.doi.org/10.1016/j.bdr.2016.02.002

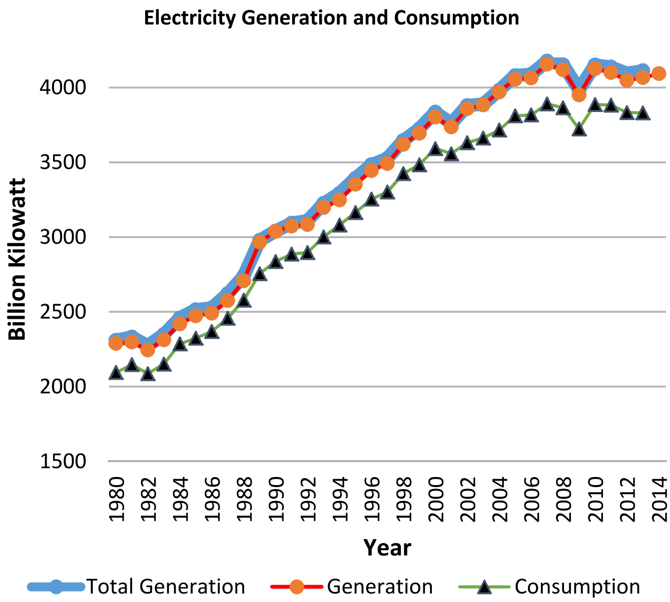

Fig. 1. Electricity generation and consumption graph. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 2. Electricity generation in the U.S., by state. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

increase the quantity to be imported from neighboring countries. It is critical to explore possibility of centralized power management and to determine the allocation of natural resources.

The prediction is challenging due to the accuracy requirement, and it becomes even more cumbersome when datasets are enormous in volume and have excessive noise and high volatility. Several forecasting methods using different species of Machine Learning (ML) algorithms, such as fuzzy neural network [15,22], gray algorithm [16], gray Markov model [17], and support vector regression [18] have been proposed to deal with electricity forecasting problems. Those models were showing impressive results in terms of forecasting accuracy. However, they might not be as effective dealing with Big Data, where more efficient schemes must be employed to deal with large volumes and complexity of datasets. On the other hand, large penetration of renewable energy sources, such as wind and solar systems, increases the uncertainty in generation [29,30].

It is important to forecast the power generation in order to allocate resources that produce the power and to calculate the demand and the quantity to be imported from neighboring countries. To reach this goal, ML methods based on Artificial Neural Network (ANN) algorithms have been developed. However, there still remains the problem of how to deal with large data size and complex mining process, and how to make the algorithms scalable and intact in their performance. In this study, a prediction method is developed based on a three step framework that incorporates Big Data analytics. First, raw data were processed and converted to suitable format; then, the data were normalized to get better performance from the ML algorithm; and finally, the data were fed into an ANN model for training purposes. The deployment begins by collecting past power generation data from all the states in the U.S., and storing it in a distributed database. Then Big Data tools are used to deal with the processing of the data. Data are first distributed to a group of computing nodes inside Hadoop cluster, and distributed algorithms are implemented in form of MapReduce to take advantage of distributed high performance computing paradigm in the laboratory environment. Afterwards, data are fed into the ANN algorithm to train the network. Finally, forecasted results from ANN are compared to the actual generation.

Fig. 3 depicts different steps in the framework for the proposed strategy. In the first step, the framework collects past power generation data from all U.S. states and stores them in a distributed database. This is the raw data with redundant information, some of which are in a completely unstructured format such as text files; others are not in any desired structured format such as csv formatted file. In the next step, Big Data tools are applied, MapReduce is implemented on top of Hadoop cluster to deal with such large datasets. Data are stored in multiple computing nodes, and distributed algorithms are implemented in the form of MapReduce. MapReduce is used to allocate assignment and to handle large datasets. Manipulated data is extracted from each computing node in the desired format. Then, data are normalized to increase the effectiveness of the ML algorithm. Finally, data from each node are used on ANN for training to predict the future power generation.

Forecasting electricity generation will eventually yield information on the demand, since there is a linear relationship between the two. Also, it is easier to deploy centralized control if we have enough information about generation and consumption for individual states as well as for the entire nation. Therefore, knowing the total generation eventually determines the amount of electricity to be imported from neighboring countries.

The remainder of this paper is organized as follows. Section 2 briefly introduces ML and ANN methods. Section 3 describes some related works, and Section 4 presents the detailed strategy and the design of the framework, followed by the results in Section 5. Finally, Section 6 concludes the paper with a discussion.

2. Background

Machine Learning (ML) and Artificial Neural Networks (ANN) are parts of cognitive science, initially evolved from two important concepts, pattern recognition and computational learning, both parts of Artificial Intelligence (AI) [4,7,8]. ML deals with analyzing algorithms that can be trained to make predictions for the future based on the past information. ANN is a learning process based on statistical models and human biological neural networks. ANN is used to estimate values based on a large number of inputs. ANN interconnects neurons with numeric values, adjustable based on experience, allowing them to use the inputs in the learning process. In this study we employ these concepts to build a framework for the electricity generation predictions with large volume of data.

ML and data mining processes have strong ties with mathematical optimization to build complex models, where designing and programming explicit and rule-based algorithms are infeasible. There are several ML algorithms, where the learning process can be supervised or unsupervised. ANN is one of the popular supervised learning process methods [7,26,27].