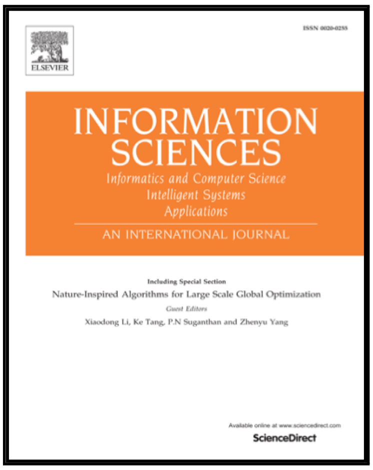

A novel Spark-based multi-step forecasting algorithm for big data time series

A. Galicia, J.F. Torres, F. Martínez-Álvarez, A. Troncoso

PII:	S0020-0255(18)30449-3
DOI:	10.1016/j.ins.2018.06.010
Reference:	INS 13703
To appear in:	*Information Sciences*
Received date:	10 August 2017
Revised date:	30 May 2018
Accepted date:	3 June 2018

Please cite this article as: A. Galicia, J.F. Torres, F. Martínez-Álvarez, A. Troncoso, A novel Spark-based multi-step forecasting algorithm for big data time series, *Information Sciences* (2018), doi: 10.1016/j.ins.2018.06.010

A novel Spark-based multi-step forecasting algorithm for big data time series

A. Galicia and J. F. Torres and F. Martínez-Álvarez and A. Troncoso

Division of Computer Science, Universidad Pablo de Olavide, ES-41013 Seville, Spain.
{agalde, jftormal}@alu.upo.es, {fmaralv, ali}@upo.es

Abstract

This paper presents different scalable methods for predicting big time series, namely time series with a high frequency measurement. Methods are also developed to deal with arbitrary prediction horizons. The Apache Spark framework is proposed for distributed computing in order to achieve the scalability of the methods. Prediction methods have been developed using Spark's MLlib library for machine learning. Since the library does not support multivariate regression, the prediction problem is formulated as h prediction sub-problems, where h is the number of future values to predict, that is, the prediction horizon. Furthermore, different kinds of representative methods have been chosen, such as decision trees, two tree-based ensemble techniques (Gradient-Boosted and Random Forest) and a linear regression method as a reference method for comparisons. Finally, the methodology has been tested in a real time series of electrical demand in Spain, with a time interval of ten minutes between measurements.

Keywords: Big data, scalable, electricity time series, forecasting

1. Introduction

It is well known that advances in technology have led, in recent years, to the increasing amount of data generated and stored, to the extent that 90% of the data that exist in the world has been generated during the last two years. The need to process this huge amount of information has made it essential in recent years to develop and evolve tools that have been included under the heading of Data Mining. This evolution has given rise to the term Big Data. An essential component in the nature of the data is that information is normally indexed over time, a process that is known in the literature as time series. This case is very common in the field of Big Data, giving rise to the term Big Data Time Series. For example, two of Big Data's main sources are open data repositories, which are proposed by management for transparency policies, such as smart cities, where multiple sensors provide information on consumption, traffic, pollution, etc. These two types of data make sense if their analysis is performed with respect to their evolution over time: data that measure electrical demand or pollution can be analysed for various purposes: to predict their evolution; to predict anomalous values; to obtain patterns that allow us to compare their evolution with other data; to establish relations between certain variables with respect to others, and so forth.

Nowadays, the main existing frameworks for the massive data processing have been developed thanks to leading technology companies such as Google and Yahoo!. MapReduce technology was developed by Google [6], which for processing purposes divides the input data into blocks and then integrates the output information of each block into a single solution. Later, Yahoo! developed Hadoop [37], an open-source implementation based on the MapReduce paradigm, now part of the Apache Foundation. The limitations of MapReduce when implementing algorithms that need to iterate over data have required the creation of new tools, such as Spark [15], developed by the University of Berkeley in California, also within the Apache Foundation.

Spark's deployment on the Hadoop Distributed File System (HDFS) allows the parallelization of data processing in-memory, achieving much faster processing speeds than with Hadoop. Apache Spark is also