ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Data learning from big data

José L. Torrecilla [a,*], Juan Romo [b]

[a] Institute UC3M-BS of Financial Big Data, Universidad Carlos III de Madrid, Spain
[b] Department of Statistics. Universidad Carlos III de Madrid, Spain

ARTICLE INFO

Article history:
Available online xxxx

MSC:
00-01
99-00

Keywords:
Big data
Data learning
Statistics

ABSTRACT

Technology is generating a huge and growing availability of observations of diverse nature. This big data is placing *data learning* as a central scientific discipline. It includes collection, storage, preprocessing, visualization and, essentially, statistical analysis of enormous batches of data. In this paper, we discuss the role of statistics regarding some of the issues raised by big data in this new paradigm and also propose the name of *data learning* to describe all the activities that allow to obtain relevant knowledge from this new source of information.

1. Introduction

Big data is one of the most fashionable concepts nowadays: everybody talks about it, is permanently in the media, and companies and governments try to exploit the new amount of available information (Lohr, 2012; John Walker, 2014; James, 2018). The ideas behind this interest are mainly two. First, the fact that at present, most activities generate data (with very low cost) that contains (potentially valuable) information. The second one is well summarized in John Walker (2014): "Data-driven decisions are better decisions - it is as simple as that. Using big data enables managers to decide on the basis of evidence rather than intuition". The opportunities offered by big data are undeniable, but there is still a debate about the scope and usefulness of this (Secchi, 2018; Bühlmann and van de Geer, 2018). The opinions of the most fervent followers speak of the end of the theory and the models and, in articles like the controversial "The end of theory" (Anderson, 2008) they argue that "with enough data, the numbers speak for themselves". On the other hand, there have been more critical voices that question whether the optimism and the faith that is being put into the big data are really justified. In this line, Tim Harford wonders if "we are making a mistake" in another provocative article (Harford, 2014). In this paper we will review some of the big data aspects that can generate doubts from the point of view of a statistician trying to scrutinize if the data are sufficient by themselves or it is necessary to give them a sense.

First, it is convenient to be more specific. Although there is no single definition, there seems to be a certain consensus that big data encompasses the study of problems so "Big" that conventional tools and models cannot handle them, either because they are not adequate or because they require too much time. In any case, whatever the definition we choose or where we put the emphasis, what is clear is that current technology generates huge amounts of data, so we have to be able to extract the best information from them and use it to make the best decisions. How to get it and the challenges associated to this new framework have become common discussion topic in the last years (Lynch, 2008; Fan et al., 2014; Gandomi and Haider, 2015) and the best way to tackle the problem has also been subject of debate. As an example, we can cite the former paper by Breiman et al. (2001) about the two cultures of statistical modeling: stochastic models and algorithms (see Dunson, 2018 for recent discussion in the context of big data). In what follows, we discuss the role of statistics regarding some of the

* Corresponding author.
E-mail address: joseluis.torrecilla@uc3m.es (J.L. Torrecilla).

https://doi.org/10.1016/j.spl.2018.02.038
0167-7152/

Please cite this article in press as: Torrecilla J.L., Romo J., Data learning from big data. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.038.

issues raised by big data in this new paradigm and also propose the name of *data learning* to describe all the activities that allow us to obtain relevant knowledge from this new source of information.

From the classical statistics point of view this massive amount of data could be a blessing as we should be nearer to the real populations and the asymptotic convergence of the models. However, in practice, it looks more like a curse. Big sample sizes entail storage and processing problems and a huge effort must be done in terms of improving computational performance and developing new tools (both hardware and software) in order to handle such volume of data. But the real challenge for statisticians is to deal with the heterogeneity inherent to big data which appears in different forms. Populations are no longer homogeneous or normal, consisting of different groups and communities, what makes invalid (at least partially) most classical statistical approaches based on convergence and central limit theorems. But heterogeneity also appears in the variety and complexity of the data managed in the areas where this kind of problems appears, such us medicine (Howe et al., 2008; Raghupathi and Raghupathi, 2014) or business (McAfee et al., 2012). Therefore, the data at hand are far from standard random vectors in $\mathbb{R}^n$ stored in simple matrices. On the contrary, data can be completely unstructured coming from surveys, call records, web activity, social networks, an so on. All of this opens the door to a wide variety of problems including a proper information encoding and the combination of different types of data structures (categorical and continuous variables, functional data, time series, images, trees, text, networks, video...), some of them so recent and complex that they are a new field of study in statistics by themselves (Marron and Alonso, 2014). Moreover, big data tends to include (and often increase) the usual problems of high dimensionality. So we are in front of a double curse in both the number of features and observations.

Hence, the "Big" in big data refers, at least, to three different directions: velocity, complexity (number of variables and variety of data structures) and sample size. Usually these three problems appear together, but they entail different issues and the role of statisticians is quite different at each of them.

2. Velocity

Nowadays, we are able to generate such huge amounts of data (structured, unstructured and semi-structured) that would be very costly and take too much time to analyze them with conventional methodologies. Therefore, it is necessary to develop new tools and architectures in order to process the increasing and unstructured volume of data in a proper way. This represents a paradigm shift that affects all aspects of data processing that must be reconsidered (Jagadish et al., 2014).

Maybe, the best known indicators of these changes are the transition from sequential to parallel/cloud computing and the emergence of non-relational data bases. Substantial efforts have already been dedicated to parallelization at all levels. Platforms like Amazon Web Services, Azure or Google offer big data services and provide the infrastructure "in the cloud" on-demand and with elastic load balance, what solves many of the problems related to scalability and the maintenance of a physical cluster (Hashem et al., 2015). Beyond the cloud computing, softwares like Cloudera or Hortonworks provide a wide catalog of parallelized services (storage, pipelines, data bases, data processing) and control the communication among them. Most of this software is based on Apache Hadoop (2008), a software framework that allows distributed storage and data processing using MapReduce. Finally, the cluster-computing framework Apache Spark (2014) is gaining popularity and substituting Hadoop for certain tasks. One of the reasons in favor of Spark is the programming interfaces for languages such as Python or R, largely used in machine learning and statistics (Singh and Reddy, 2015).

Furthermore, non-relational or NoSQL databases arise from the need to both treat and search the data quickly and organizing the unstructured information. These models go a step further the parallelization of conventional relational databases (which is possible using Hive or HBase) by proposing new approaches that allow for greater dynamism and facilitate maintenance and scalability.

We could go deeper into these points, which are of great interest, and talk about other technical difficulties related to velocity (for example, in visualization or with the consistency of data), but the central task of statisticians here is, firstly, the development of computationally efficient algorithms. This aspect, which is usually forgotten or overlooked in the statistical literature, is critical in this context because even an optimal algorithm is completely useless if it cannot be applied.

3. Complexity

Mostly, the complexity inherent to big data comes from the high dimensionality of the observations and the unstructured nature of the data we generate with smart phones, sensors, social networks, internet searches, GPS devices, emails, and so on.

Problems associated with the dimensionality are well known by statisticians and other researchers. They have been extensively studied and many proposals have been done. Beyond the particular properties of each technique, the dimension reduction methods can be grouped in two big families that we could call projection data and variable selection. Both approaches (variable selection and projection methods) have been extensively studied and compared at different contexts. The difference between these two approaches is the role of the original variables in the reduced space. While variable selection is restricted to the original variables, projection methods allow certain combinations of them in order to obtain the new components. This makes variable selection techniques provide meaningful reductions (highly appreciated in areas such as biology or medicine).

Both methodologies have proven to be effective in a wide range of problems and are already being applied to big data questions where the issues typically associated with the high dimension take on special relevance. For example, Singh et al.

Please cite this article in press as: Torrecilla J.L., Romo J., Data learning from big data. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.038.