Big Data Research ••• (••••) •••–•••

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

How to Quantify the Impact of Lossy Transformations on Event Detection ☆

Pavel Efros, Erik Buchmann, Adrian Englhardt, Klemens Böhm

Karlsruhe Institute of Technology, Karlsruhe, Germany

ARTICLE INFO

Article history:
Received 30 June 2016
Received in revised form 1 December 2016
Accepted 20 February 2017
Available online xxxx

Keywords:
Time series
Lossy transformations
Event detection
Change detection

ABSTRACT

To ease the proliferation of big data, it frequently is transformed, be it by compression, be it by anonymization. Such transformations however modify characteristics of the data. In the case of time series, important characteristics are the occurrence of certain changes or patterns in the data, also referred to as *events*. Clearly, the less transformations modify events, the better for subsequent analyses. More specifically, the severity of those modifications depends on the application scenario, and quantifying it is far from trivial. In this paper, we propose MILTON, *a flexible and robust Measure for quantifying the Impact of Lossy Transformations on subsequent event detectiON*. MILTON is applicable to any lossy transformation technique on time-series data and to any general-purpose event-detection approach. We have evaluated it with several real-world use cases. Our evaluation shows that MILTON allows to quantify the impact of lossy transformations and to choose the best one from a class of transformation techniques for a given application scenario.

1. Introduction

Event detection on time-series data is an important building block of many real-world applications [1,2]. It perceives a time series of measurements as one of events. Our notion of event encompasses changes [2,3] and frequent time-series patterns (motifs) [4–6]. Changes are points of time when properties (e.g., mean, probability distribution) of a time series change. Frequent patterns are contiguous subsequences of a time series that occur frequently, indicating a structure or information with some regularity [4]. To illustrate the notion of event detection, think of energy consumption data from a smart meter, which serves as our running example. Event detection on such data allows to detect interesting patterns (turning a device on/off, abnormal device activity). Detecting such events is necessary for demand side management, peak shifting, peak shaping, etc. – all elementary techniques to integrate renewable energy sources into the Smart Grid. However, data transformation, e.g., lossy compression or anonymization, can modify the data considerably. This in turn can aggravate the subsequent detection of those events significantly.

Example 1. An energy provider uses a lossy compression technique for time series from smart meters in order to reduce the data volume, before running an event-detection algorithm. Due to the compression loss, (a) some events might be detected at different points in time, or (b) their significance might be altered, compared to the original time series, (c) events also might go undetected at all, or (d) the compression might result in new events. Using domain knowledge, the provider can assess the importance of these impacts. Based on his assessment, he wants to select a concrete compression technique, with a good parameterization.

Approaches like lossy compression [7], estimation [8] or perturbation/anonymization [9] lossily transform the time series before event detection takes place: A lossy transformation can reduce the data volume, generate approximate versions of the data, or remove personal information from a dataset. However, existing similarity measures for time series, applied to the original time series and the one after lossy compression and decompression, do not quantify the impact of a lossy transformation on event-detection quality in a way that is conclusive in general [10–13]. Such a quantification however is needed to identify and parameterize a compression algorithm or anonymization approach, given a certain dataset and quality requirements on the event-detection result. This quantification sought is difficult, for several reasons: First, as shown in Example 1, the impact is manifold. One therefore needs to determine possible effects of a lossy transformation on events. Second,

☆ This article belongs to Online Forecasting.

E-mail addresses: pavel.efros@kit.edu (P. Efros), erik.buchmann@kit.edu (E. Buchmann), adrian.englhardt@student.kit.edu (A. Englhardt), klemens.boehm@kit.edu (K. Böhm).

http://dx.doi.org/10.1016/j.bdr.2017.02.001
2214-5796/

the definition of a measure for this impact is not obvious. It is necessary to investigate application scenarios where one is working on the transformed data, in order to come up with respective requirements. Third, the measure envisioned should be customizable to the concrete application scenario. Think of the energy provider once again. For him, it will most likely be more detrimental if compression eliminates certain events from the data, as opposed to the insertion of new ones. In other contexts, the picture is different. Fourth, identifying the specific effect of a transformation on an event (e.g., shift in time vs. disappearance) is an application-dependent procedure, which must take all events into account. This is because assigning an effect to a certain event may cascade and influence the assignment of effects to other events. Put differently, even if the measure is defined, algorithms for its efficient computation remain to be designed.

In this paper, we propose and evaluate MILTON, *a practical and flexible Measure which quantifies the Impact of various Lossy Transformation methods for time series on subsequent event detectiON.* MILTON is applicable whenever one wants to know how much a certain transformation approach for time series reduces the result quality of an event-detection technique, as compared to event detection on the original data. This lets an operator, say, decide how much he can compress or perturb data without affecting event detection considerably. Thus, MILTON is useful when choosing from several lossy transformation techniques, by quantifying their impacts on events. To ensure flexibility, we do not impose any restriction on the event detection or the transformation approach used, and we allow to flexibly weight effects on events. We have also investigated several cases that we deem recurrent and propose corresponding parameterizations of MILTON. In contrast to metrics of the quality of event detection methods (e.g., recall, precision, F-score), MILTON's purpose is to quantify the impact of lossy transformations on events. It is applicable to any event-detection algorithm that takes place subsequently.

At first sight, a complement of MILTON or even an alternative to it might be a model of the loss of data quality due to a transformation. However, such a model would have to be generally applicable. But it is difficult to impossible to integrate each of the many existing lossy transformation techniques and event-detection approaches into one model.

In this article, we now make the following contributions:

- We study characteristics of application scenarios that do event detection on lossily transformed time-series data.
- We propose a measure of the impact of time-series transformation methods on subsequent event detection.
- We carry out an evaluation of our measure using five different use cases, namely compression, estimation, anonymization, assisted living and activity hiding.

We have carried out extensive experiments, which have revealed interesting insights on the relationship between the transformation technique in use and event-detection quality. For instance, different anonymization techniques may have a very different impact on event detection, although they protect against noise filtering equally well. We also have found MILTON suitable with any combination of lossy transformation technique and event-detection approach we have encountered. In addition, the design of MILTON enables a flexible customization of the different effects a lossy transformation may have on events. Finally, it is applicable in many application areas in a straightforward manner.

Paper structure: Section 2 describes five application scenarios for our measure. Section 3 introduces and explains MILTON, which Section 4 evaluates. Section 5 reviews related work, and Section 6 concludes. – This article is an extended version of [14]. The extensions are the following: Our study of application scenarios that consider time-series patterns is broader, as is our respective evaluation. Next, we have extended all of our approach so that it now also subsumes time-series patterns, as opposed to only changes.

2. Application scenarios

In this section, we describe five scenarios – this will serve as motivation behind our measure. We then derive the requirements on it. We have consciously decided to describe these scenarios in much detail, in order to reveal the subtle differences between them, which then give way to the requirements.

2.1. Compression scenario

2.1.1. Description

The growing number of smart meters as well as the increasing frequency at which data is collected make storing and transferring the data much more expensive. To illustrate, while smart meter readings often take place every 15 minutes, meters that collect and send data every second are now proliferating. Moreover, such meters now collect and send several values instead of just one, e.g., voltage, current, frequency, active power, etc. The collected data is useful for analyses such as energy-consumption forecasts [15] or energy disaggregation [16]. To store and communicate this data, recent research has produced numerous model-based lossy compression techniques [7,17–19]. In contrast to lossless ones, they can obtain significantly higher compression ratios. Lossy methods typically produce a piecewise approximation of the original data within an error threshold ϵ. Thus, they do not only modify the original data, but also the changes present in it. An energy provider intending to use the compressed data for analytics needs to take these effects into account.

2.1.2. Problem domain

The result of lossy compression methods depends on the models they use (e.g., constants, straight lines, polynomials), and how they use them. To evaluate their impact on changes in the data, one thus needs to consider different classes of models. Another important parameter here is the error threshold ϵ. We expect compression results, and consequently their impact on changes, to strongly depend on this parameter.

2.1.3. Setting

The energy provider employs a forecasting application that uses the compressed time series to predict the energy consumption. By detecting changes in data streams and integrating them in the learning model, he can improve forecasting [20] or enhance stream mining [21]. Thus, here, detecting a change in the time series triggers an update of the model behind the forecasting algorithm, to improve predictions. In such a case, it makes sense to penalize changes which disappear ("missed", also referred to as "false negative" in the literature) more than those which emerge ("false positives") as an effect of the transformation. This is because a missed change prevents the forecasting algorithm from updating its model when necessary. This may impact its accuracy significantly. A false-positive change in turn will trigger an unnecessary update of the model, which may cause additional effort, but should not affect forecasting accuracy considerably. Regarding shifts of changes in time, the provider deems them important for forecasting, as they will delay or vice-versa advance the update of the underlying model. On the other hand, modifications of the importance of changes are not crucial in this case, so he chooses to ignore them altogether. This makes sense here because, once a change is detected, the model is updated regardless of that importance.

Please cite this article in press as: P. Efros et al., How to Quantify the Impact of Lossy Transformations on Event Detection, Big Data Res. (2017), http://dx.doi.org/10.1016/j.bdr.2017.02.001