Information Systems 65 (2017) 65–77

Contents lists available at ScienceDirect

Information Systems

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

An evaluation of combinations of lossy compression and change-detection approaches for time-series data

Gregor Hollmig, Matthias Horne, Simon Leimkühler, Frederik Schöll, Carsten Strunk, Adrian Englhardt, Pavel Efros*, Erik Buchmann, Klemens Böhm

Karlsruhe Institute of Technology, Karlsruhe, Germany

ARTICLE INFO

Keywords:
Time series
Compression
Change detection

ABSTRACT

Today, time series of numerical data are ubiquitous, for instance in the Internet of Things. In such scenarios, it is often necessary to compress the data to, say, reduce data-transmission costs, and to detect changes on it. More specifically, both methods are used *in combination*, i.e., data is lossily compressed and later decompressed, and then change detection takes place. There exists a broad variety of compression as well as of change-detection techniques. This calls for a systematic comparison of different combinations of compression and change-detection techniques, for different data sets, together with recommendations on how the values of the various (typically non-linear) parameters should be chosen. This article is such an evaluation. Its design is not trivial, necessitating a number of decisions. We work out the details and the rationale behind our design choices. Next to other results, our study shows that the choice of combinations of change detection and compression algorithm and their parameterization does affect result quality significantly. Our evaluation also indicates that results are highly contingent on the nature of the data.

1. Introduction

Nowadays, time-series data is ubiquitous. More and more applications like the Smart Grid or the Internet of Things that produce and/or process time-series data are proliferating. Such data is often used to detect certain events and to react to them as soon as possible. In other words, change-detection methods are indispensable. On the other hand, because of the many devices generating data, the huge amount of data and the high data-transfer rates, an efficient compression is essential. Lossless compression reduces the statistical redundancy of the data. However, compression rates are relatively low; as an example, rates of 4 have been reported for smart-meter readings from individual buildings using the bzip2 algorithm [1]. Lossy compression in turn yields significantly higher compression ratios than the lossless one. At the same time, data compressed in this manner is still useful for many applications. In this study, we focus on lossy compression. Putting things together, it often is necessary to combine lossy compression[1] and change-detection techniques.

Example 1. *Smart meters may deliver data to a central analysis system via a wireless network. To save bandwidth and to reduce costs, the data is compressed directly on the device. The central data-analysis system can then do change detection to react to events such as a sudden increase in overall power consumption.*

When combining lossy compression and change detection, several issues arise. First, lossy compression introduces errors. In particular, changes can be lost, or new false changes can occur. Therefore a lossy compression method must be chosen which preserves the change information as much as possible. Furthermore, different use cases generate different kinds of time-series data, as we will explain. Thus, it is necessary to choose a good combination of compression and change-detection technique *per use case*. This is difficult due to the large number of possible combinations. Next, compression as well as change-detection algorithms usually have several parameters, which often have non-obvious effects on the outcome. The expectation typically is that domain experts select the parameter values. This means that these experts must have a deep insight into the algorithms used. But even if they have selected the values, it is hard to determine whether their selection is a good one. To investigate how combinations of compression and change-detection algorithms perform on different datasets, a systematic comparison is necessary. This article is such a

* Corresponding author.
E-mail addresses: gregor.hollmig@student.kit.edu (G. Hollmig), matthias.horne@student.kit.edu (M. Horne), simon.leimkuehler@student.kit.edu (S. Leimkühler), schoell@ira.uka.de (F. Schöll), carsten.strunk@student.kit.edu (C. Strunk), adrian.englhardt@student.kit.edu (A. Englhardt), pavel.efros@kit.edu (P. Efros), erik.buchmann@kit.edu (E. Buchmann), klemens.boehm@kit.edu (K. Böhm).

[1] For improved readability we usually refer to compression and later decompression simply as compression.

http://dx.doi.org/10.1016/j.is.2016.11.001
Received 28 January 2016; Received in revised form 12 November 2016; Accepted 15 November 2016
Available online 26 November 2016

study.

Designing our study has been challenging, partly due to the issues just mentioned. To illustrate, one of the various design decisions is as follows: It is difficult to choose the parameterization of the compression and the change-detection algorithms such that the comparison is fair. Reusing the parameter values suggested in the original publications may not be the best option. This is because proper choices of parameter values depend on the data the algorithms are applied to. Thus, we have decided to perform an optimization on each dataset that yields the parameter values that give way to change-detection results after compression that are closest to some carefully chosen reference point. This article lists the design questions encountered in the context of our comparison, together with explanations behind our choices.

In line with these design decisions, we have implemented a framework that can be used for the evaluation of virtually any combination of compression and change-detection methods. In our specific study, we examine five compression algorithms like APCA [2] and five change-detection algorithms like Online-Kernel Change Detection [3] on five datasets, resulting in 125 possible combinations. We focus on result quality and leave aside criteria such as runtime performance or total cost of ownership, which highly depend on specifics of the implementations and the runtime environment as well as on characteristics of the underlying optimization framework.

The study shows that, while the choice of the dataset does have a huge impact on which combination of compression and change-detection technique performs best, some algorithms like Chebyshev Approximation [4–6] and Bayesian Online Change Detection [7] yield good results in many settings. We also observe that a good change detection is possible even on strongly compressed data. Next, our results are particularly interesting because studying the algorithms in isolation (e.g., compression without subsequent change detection) may yield a different picture. In [8] for instance, competing algorithms have outperformed Chebyshev Approximation with regard to the compression ratio. In our context in turn, this algorithm has proven to be suitable in combination with many change-detection algorithms.

Paper outline: Section 2 describes some application scenarios. Section 3 explains our design decisions. Section 4 summarizes the algorithms evaluated. Section 5 describes the experimental setup and Section 6 presents the results. Section 7 concludes.

2. Application scenarios

In this section we describe two scenarios, with slightly different perspectives on the importance of compression and change detection. In the first scenario, compression and change-detection quality are roughly equally important. In the second one, the benefits of a high compression ratio tend to exceed those of the change detection.

2.1. Smart grid

The Smart Grid is an intelligent communication network which monitors and controls a power network. The integration into such networks of renewable energy producers alters the conventional power flow [9]. These producers are inconsistent and have performance peaks, which in turn demand intelligent power distribution systems.

Consider a company which has to manage a power- distribution network. The company collects, stores and analyzes the data delivered by the many devices (e.g., smart meters, power plants) in its network. The data needs to be analyzed in real-time, thus online change detection is indispensable. To significantly reduce communication and storage costs, the data must be lossily compressed. Now think of a sudden increase in power consumption. The company must react as soon as possible for example by powering up additional power plants. To this end, it must detect the change in the first place, which is not only the consumption measured by one single device, but an aggregate of the entire grid. As a takeaway, we observe that good compression and high-quality change detection are both very important in this scenario.

2.2. Internet of things

Internet of Things (IoT) refers to large networks of small or embedded devices, which communicate wirelessly. For many IoT entities, energy optimization is a primary constraint, as they are powered by batteries or use energy harvesting methods like micro solar panels. Thus, wireless data transmission often is the biggest factor regarding energy consumption, as the power required to transmit data increases quadratically or even with the power of 4 with the distance between sender and receiver [10]. The power consumption of data compression in turn increases only linearly with the size of the data.Thus, it is reasonable to send data that is lossily compressed over a distance. Detecting changes is often computationally heavy (e.g., overall computational complexity Bayesian Online Change Point Detection is $O(n^5)$, where n is the length of the sequence under consideration [7]) and should be performed on the central unit; it therefore has to take place on compressed and later decompressed data [11]. Now consider a home automation system, where a central control unit can adapt the heating when several temperature or humidity sensors detect a change in the weather. Online change detection is needed to react in short time. This specific scenario benefits more from a high compression ratio than from better change detection, in contrast to the previous scenario.

3. Design decisions

Designing the comparison study envisioned is challenging; in particular, there are various design decisions that one must address. Fig. 1 is an overview of our study framework. The framework takes as input a dataset and ground-truth change points. It then uses these change points to derive optimal parameters for the change-detection methods used. It subsequently uses an adequate error measure to evaluate the impact of compression on the changes in the dataset. This is in order to obtain an optimal combination of compression ratio and change-detection quality. In the following, we describe the important design alternatives and the rationale behind our choices. Even though the subsequent subsections will introduce them explicitly, Table 1 lists the basic notions we use and their definitions.

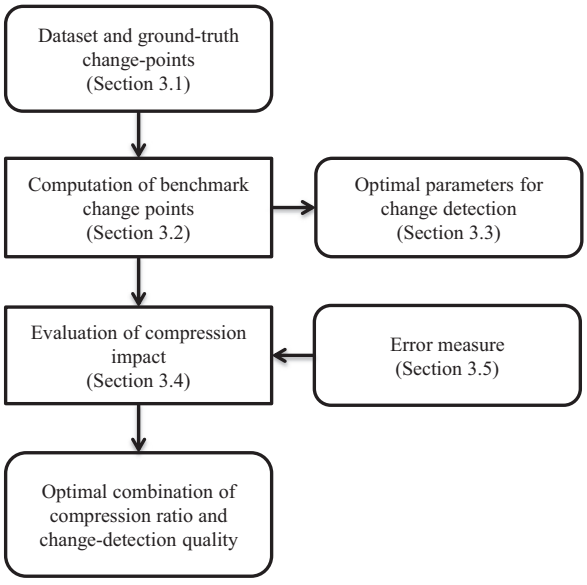

Fig. 1. Overview of evaluation framework.