Mining sensor data from complex systems

Vespier, U.

Chapter 1

Introduction

Over the last decades, the advances in computational power, storage technology and sensor networks have made data an abundant resource [79]. Today, virtually everything, from natural phenomena to complex artificial and physical systems, can be measured and the resulting information collected, stored and analyzed in order to gain new insight, optimize existing processes or both.

The term *Big Data* has gained popularity, in academia, industry and the public opinion, to describe the opportunities and the challenges connected to this huge explosion in data availability [64]. In a report by IDC [33], the authors estimate that the total amount of data in the digital world will amount to 40.000 exabytes[1] by the end of 2020, almost doubling its size every year. The data sources are diverse, ranging from user-contributed material on social networks (i.e. posts, tweets and status updates) to consumer behavioral data collected by online retailers such as product views and purchases.

In particular, advances in measuring technology and sensors networks [3, 16] greatly contributed to the explosion of data. The adoption and deployment of measurement systems for all sorts of industrial, commercial and consumer applications, is paving the way to important opportunities for monitoring and analyzing all kinds of systems over time at a level of detail never experienced

[1] This equates to $4 \cdot 10^{22}$ bytes.

before.

In fact, sensing technology and the ability to manage big data represent a fundamental improvement in our ability to measure complex systems. Multiple types of sensors, high sampling rates, advances in noise reduction techniques, to cite a few, are all improvements that are contributing to making progressively better representations of systems in data.

As a result, new challenges in the analysis and visualization of this large amount of sensor data have emerged and, over the last decade, the efforts of the research community to provide solutions to these problems have soared [1]. Methods and algorithms, in fact, will have to advance in order to cope with the increased complexity of the time series datasets available and to improve the ability to learn from the greater level of detail present in them.

A side effect of the exponential explosion of data collection is that *labeled* information will be an increasingly scarce resource in the future, as it is extremely costly to produce labeled datasets in relation to the current rate of data growth. Because of this, the task of extracting structured information from unlabeled data is of paramount importance when dealing with the challenges posed by big data. The algorithms and methods presented in this thesis are designed to work in this scenario where novel insights have to be extracted in an unsupervised way.

In particular, the focus of this thesis is the analysis of complex sensor data in the form of *time series*. Time series are sequences of observations sampled periodically over time. We approach the analysis of such data from a *data mining* perspective, with the end goal of extracting previously unknown knowledge and insight in the data. Data mining [37, 104] (DM) is a discipline aimed at discovering useful and structured patterns in large collections of data. Data mining methods lie at the intersection between computer science, machine learning, database systems and statistics, and are a fundamental part of any KDD (Knowledge Discovery in Databases) process [27]. Time series, on the other hand, are a ubiquitous type of data, and mining time series data represents an important branch of DM.

In this thesis, we introduce data mining and visualization methods for large time series data collected from complex physical systems by means of sensors.

Our work is motivated by InfraWatch [55], a Structural Health Monitoring project centered around the management and analysis of data collected by a large sensor network deployed on a highway bridge. The sensor network comprises strain, temperature and vibration sensors, sampling continuously at 100 Hz. A highway bridge is a complex system and so is the data collected. The behaviour of the bridge, and consequently the properties of the data, is affected by external factors such as the temperature, weather conditions, traffic activity and deterioration of the concrete. Moreover, InfraWatch data contains repeated patterns at different resolutions due to the bridge's response to recurring events such as passing vehicles or traffic jams. Because of these characteristics, InfraWatch is an ideal testbed for evaluating the methods we introduce.

In this thesis, we will develop and discuss solutions to the following fundamental questions when dealing with large and complex time-series data collected from sensors like the one provided by InfraWatch:

- What are the relevant temporal scales of analysis for a given time series?
- Which are the recurring multi-scale patterns present in a given time series?
- How can we effectively model and recognize events in time series data?
- How can we support efficient and interactive visualization of massive time series data?

Complex systems are often affected by several phenomena at multiple temporal scales and this effect is reflected in the collected data. Consider, for example, the time series produced by one of the strain sensors of the InfraWatch bridge. This data is the result of the superimposed effects of the passing vehicles, traffic jams and more long-term effects such as the day-night cycle in temperature, which in turn affect the response of the structure. Throughout the thesis, we will introduce a method to discover which are the relevant

temporal scales of analysis and introduce a decomposition of the original time series such that every component represents a single phenomenon at its characteristic scale. The goal of the method is to find the underlying factors that explain the input data.

These multiple phenomena, moreover, are often characterized by the presence of patterns that repeat over time and reflect their effect in the data. Consider again the strain sensor example. The effect of traffic jams will produce similar recurring patterns in the data, for example every morning during rush hour. The same would happen with the effect of passing vehicles, although they will appear at a shorter time scale (in the order of seconds) and potentially superimposed on the traffic jam patterns. We will introduce a method to mine recurring patterns, so-called *motifs* in the literature, from time series data at multiple temporal scales.

The third research question addresses the problem of clustering time series subsequences in order to model and recognize fixed-length events in the data. Time series clustering has proven to be a difficult task, as it is hard to model the subsequence space properly without introducing artifacts in the results [49]. We will introduce a novel distance measure to cope with this problem.

Finally, we will address the problem of massive time series visualization. Although visualization is not directly related to data mining, it is a fundamental task in every data science project, especially to support the exploratory phases and build an idea of the data at hand. When exploring and visualizing a dataset, interactivity is important as it permits testing ideas and assumptions quickly without having to wait excessive periods. We will see how we made the interactive visualization of terabyte-sized datasets possible by introducing an ad-hoc storage scheme for time series, which effectiveness has been proven by a real world software package called VizTool.

Although these research questions find a natural application in the InfraWatch project for the analysis of bridge sensor data, we stress that they are instrumental to the understanding of many complex systems. In fact, the presence in the data of multiple temporal scales and recurring phenomena, as well as

the need for effective visualization, are general challenges shared among all complex physical and artificial systems measured by sensor networks.

The methods and algorithms introduced in this thesis combine concepts from data mining, signal processing, and information theory. In particular, in order to formally characterize the concept of temporal scales, we will make use of concepts from the field of signal processing, such as the theory of scale-space [103, 62].

As we are interested in extracting new insights from the data, such as the relevant temporal scales and the recurring events, we approach the problem from a *compression* standpoint. The idea of using concepts from the theory of compression in order to learn new facts about the structure of a dataset has been widely considered and explored in the literature [82, 105]. Data compression techniques geared towards learning have been employed for categorizing text [29], clustering data [17], devising similarity measures [99, 52], in genomic analysis [36], data discretization [57], pattern mining [100, 68, 84], stream mining [89], and as the base of parameter-free data mining methods [51].

We will see how we can define parametrized compression schemes for time series in order to find the one that best compresses the data and, at the same time, results as simple as possible. We employ the Minimum Description Length framework [34], a formalization of the Occam's Razor principle [91], in order to select the best compression model among the many possible and conceptually discern what is notable, and what can be ignored, in the data. We will see how this approach deals with many of the challenges present when analyzing real-world time series data, such as the presence of noisy measurements, the occurrence of spurious and anomalous events and, ultimately, the risk of over-fitting the data with models that would be hardly general.

1.1 Thesis Outline

Below, we give a brief outline of the dissertation, summarizing the contents of the following chapters. As most chapters are based on previous publications by the author, we also give the appropriate references to them when this is the case.

In **Chapter 2: Sensor Data and Applications**, we give the main motivations behind this work and introduce the InfraWatch project.

In **Chapter 3: Preliminaries and Background**, we introduce fundamental material and concepts that will be used throughout the rest of the thesis, especially in Chapter 4 and Chapter 5.

In **Chapter 4: Identifying the Relevant Temporal Scales**, we discuss a method for discovering the most relevant scales of analysis, and their corresponding scale components, in time series data. This work was published in the following paper [96]:

> Vespier U., Knobbe A., Nijssen S., and Vanschoren J., *MDL-based Analysis of Time Series at Multiple Time-Scales*, in Proceedings ECML-PKDD 2012, Bristol, UK.

This work is also part of the following book chapter [95]:

> Vanschoren, J., Vespier, U., Miao, S., Meeng, M., Cachucho, R., Knobbe, A., *Large-scale sensor network analysis — Applications in structural health monitoring*, in Big Data Management, Technologies, and Applications, IGI Global, 2013

In **Chapter 5: Mining Variable-Length Motifs at Multiple Scales**, we introduce a method for mining variable-length, and potentially overlapping, motifs at multiple temporal scales in sensor-based time series data. This work was published in the following paper [98]:

> Vespier, U., Knobbe, A., Nijssen, S., *Mining Characteristic Multi-Scale Motifs in Sensor-Based Time Series*, in Proceedings CIKM 2013, San Francisco, USA.

In **Chapter 6: Subsequences Clustering for Events Modeling**, we discuss a distance measure and an associated method for the effective clustering time-series subsequences for events modeling. This work was published in the following paper [97]:

> Vespier, U., Knobbe, A., Vanschoren, J., Miao, S., Koopman, A., Obladen, B., Bosma, C., *Traffic Events Modeling for Structural Health Monitoring*, in Proceedings IDA 2011, Porto, Portugal.

This work is also part of a book chapter [95]:

> Vanschoren, J., Vespier, U., Miao, S., Meeng, M., Cachucho, R., Knobbe, A., *Large-scale sensor network analysis — Applications in structural health monitoring*, in Big Data Management, Technologies, and Applications, IGI Global, 2013

In **Chapter 7: Interactive Time-Series Visualization**, we introduce a method and a software platform for visualizing terabyte sized time-series dataset. This work was published in the following paper [6]:

> Baggio, A., Vespier, U., Knobbe, A., *Automated Selection of Data-Adaptive Approximations for Large Time-Series Visualization*, in Proceedings Benelearn 2013, Nijmegen, the Netherlands

In **Chapter 8: Conclusions**, we draw the overall conclusions regarding this work and highlight some final considerations about its impact and potential future work.