Machine learning and deep learning approaches for multivariate time series prediction and anomaly detection

Thill, M.

Chapter 1

Introduction

1.1 Background & Motivation

We often find ourselves in situations that seem unusual to us, and we get surprised by events or observations that do not seem to correspond to our expectations. But what is the reason for us reacting surprised? One possible explanation could be that we get used to continually recurring events and observations and memorize (although often not consciously) the relevant patterns that characterize them. Due to our cognitive abilities, we can process new information, match it with already existing patterns in our memory, and formulate predictions or expectations based on it (usually, also subconsciously). And as long as the new information matches our expectations, everything is fine. But as soon as we observe a new piece of information containing patterns that deviate from the norm, we run into a moment of surprise because we cannot connect these new patterns to the information stored in our long-term memory and our expectations (built up in repeated experiences) are not met. Oddly enough, we can – in many cases – naturally learn to identify what is novel or unusual without requiring any form of external assistance. As we will see later in the context of learning algorithms, this characteristic is often linked to a concept called "unsupervised learning". Our ability to identify the unusual helps us to cope with our daily life. Often, an unusual situation is also associated with a certain degree of concern for us, and we understand the situation as a warning of an impending problem.

Ignoring unusual or unexpected observations may have fatal consequences. In 1998, a high-speed train derailed and crashed into a bridge near the town of Eschede in Germany, causing the death of 101 people [42]. The accident's leading cause (followed by a sequence of unfortunate events) was a fatigue crack in one of the train's wheels, which was not discovered during inspections of the train in the weeks prior to the accident. Remarkably, already two months before the crash, several train staff members noticed and reported unusual noises and vibrations in the train compartment with the defective wheel – however, the complaints were not considered a safety issue. And although train wheels had been previously inspected with advanced testing devices, by 1998, the equipment was no longer in use due to its high false alarm rates (we will call these alarms false-positive errors and discuss the issue of false alarms in more detail later on). Instead, maintenance staff only performed visual inspections of the wheels with simple flashlights, which would not allow

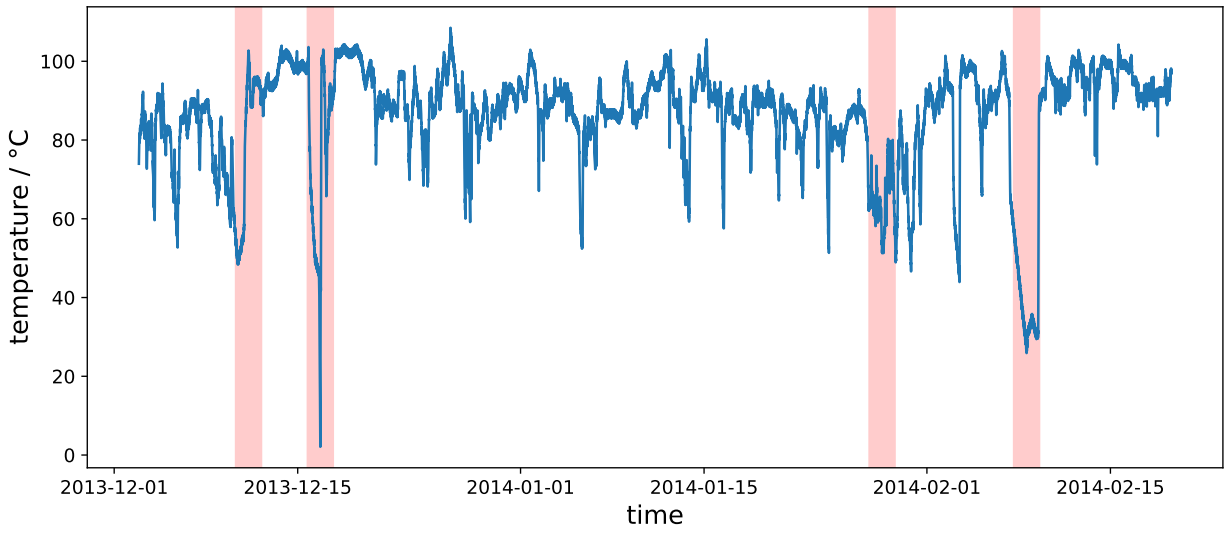

Figure 1.1: An example time series taken from the Numenta Anomaly Benchmark (NAB) [89].The graph shows the temperature sensor data of an internal component of a large industrial machine over its last few months of operation. The second anomaly (mid of December) is a planned shutdown of the machine. The catastrophic failure occurs end of February when the recordings end.

detecting fatigue cracks reliably. In the aftermath, it was discussed how microphones usage, in particular, could have possibly prevented the crash [10]. Today, so-called wayside monitoring points have been installed across the German rail network, which monitor passing trains with acoustic, optic, and other sensors and, for example, check the noise of passing trains for irregular patterns [38].

The previous example highlights the necessity of monitoring critical systems and detecting and responding appropriately to unexpected or abnormal behavior at an early stage in order to prevent failure and consequential damage. Similar to the wayside monitoring points mentioned before, many critical systems such as IT infrastructures, industrial machines, or power plants can be equipped with specialized sensors and software to record relevant health indicators and detect abnormal behavior. If the monitored data is visualized/presented appropriately, one could manually inspect the recorded data. An example is given in Figure 1.1, where one health indicator (temperature) of a large industrial machine is plotted over time. Although not all readers will have sufficient background knowledge about the monitored machine, most will spot at least two points in time that appear to deviate from the norm significantly.

However, while in the past, one could rely on humans' aforementioned extraordinary capabilities to analyze data and to recognize unusual and possibly problematic patterns, this has become increasingly challenging in recent years.

For example, one major challenge in many domains is the enormous amounts of data that have to be processed continuously. In the past years, we could observe the increasing degree of digitalization in companies and their processes, households, and many other public institutions and an extensive interconnection of systems. Today, companies are equipping even the smallest devices with sensors that consistently supply various measurements and other indicators. New technologies for the internet of things, cyber-physical systems (CPS), and other related domains enable manufacturers to equip such devices with considerable computing power, networking capabilities, and numerous sensors. These devices can consistently record and distribute various measurements and other information. The sheer endless amount of data and its inherent complexity can even overwhelm experienced human experts when analyzing and interpreting it manually. For this reason, among others, the necessity for automated, intelligent, and adaptable analytical approaches arose, and researchers invested much effort into developing new methods and algorithms. Especially in the field of machine learning (ML), the research community has made notable progress in recent years. In particular, deep learning (DL) – a subfield of ML – has recently received much interest. Essentially, machine learning is the generic term for computer algorithms designed to learn solely through data. There is no need for programming hard-coded rules into a model. Instead, an ML model is trained with a set of examples (the so-called training data set) and autonomously learns to identify the intricate underlying patterns within the data, which later allows the model to generalize to examples it has not seen before.

While ML/DL models can be and are used in a wide range of applications, such as computer vision[174], natural language processing (NLP) [125], or general game playing (GGP) [155], in this thesis, we will focus on one particular task in ML, which is concerned with the detection of unusual events in time series. In the scientific literature, this task is commonly referred to as *time series anomaly detection*.

1.1.1 Introduction to Time Series Anomaly Detection

Generally, there is no clear definition of the term "anomaly" since the categorization into anomalous or normal is heavily dependent on the application domain and the problem context. The Oxford Language dictionary defines an anomaly as

"something that deviates from what is standard, normal, or expected" [126].

Other (similar) definitions can be found in [26, 64, 23]. In the literature, also the terms *outlier detection*, *novelty detection* or sometimes *abnormality detection* can be found. Especially the term "outlier detection" is often used synonymously with "anomaly detection". While this understanding might be reasonable in many domains, especially in the time series analysis domain, which is the main focus of this work, the notion of anomaly and outlier is mostly different: outliers are single points that deviate significantly from all other points in the time series. Outliers are typically considered as single-point anomalies. However, not all anomalies in time series are outliers. Most anomalies in time series are temporal patterns

– containing a range of points – that are unusual given their temporal context. A possible definition that considers the temporal aspect of time series and which might be useful for this work is:

> *An anomaly is an observed pattern that, when viewed in its temporal context, does not conform to the normal or expected behavior due to its characteristics.*

There is a wide range of possible applications for time series anomaly detection: As already mentioned, an important application is system health monitoring, and fault detection, which is, for example, often embedded in predictive maintenance (PdM) approaches [119]. Other relevant applications are intrusion detection systems, patient monitoring systems in medical domains, or event detection in sensor networks [26]. This thesis will focus less on the details of different application domains of anomaly detection algorithms. Instead, as discussed in more detail below, this thesis aims to investigate more general aspects of time series anomaly detection. Nevertheless, we will investigate a few examples coming from some of the domains mentioned above.

1.1.2 Challenges

Up till today, anomaly detection in general and more specifically in time series has remained a complicated task. There are several challenges in the field of time series anomaly detection, which we believe are still not sufficiently addressed in the literature:

Availability of labeled data Many machine learning approaches require so-called *labeled (annotated) data* for their training to learn a general model that can later predict the outputs for new inputs. In these cases, the training algorithm processes examples with pairs consisting of an input and the desired output. After training (during the inference phase), the model can predict the unknown outputs for new given inputs. Since a "teacher" or "supervisor" is needed to pass the desired outputs to the training algorithm, this type of learning is called *supervised learning*. In the context of time series anomaly detection, "labeled data" means that each data point in a time series has been classified as either normal or anomalous by a domain expert. However, in a vast amount of real-world anomaly detection problems, no or insufficient labeled data exists. This increases the difficulty of the training and benchmarking of anomaly detection algorithms. One reason for this situation is that the data set's annotation process is usually very cumbersome and expensive; a human expert has to analyze and judge every piece of data. Since most anomalies in time series are not single points but longer sub-sequences, a particular problem in the annotation process is to identify the anomalous regions. Often, it is not evident at which instance of time an anomaly begins and where it ends. While large publicly available datasets for benchmarking new methods are available in other domains, such as visual object recognition (ImageNet [37], Open Images [84]), the amount of benchmark data for time series anomaly detection is somewhat limited.

Anomalies are rare events Another critical challenge for many anomaly detection problems is the simple fact that anomalies are rare events. Even if large amounts of data are available, they will mostly represent normal (nominal) instances. Due to the high class-imbalance (many examples which are normal and only very few which are anomalous), it is hard to train a supervised machine learning model.
An alternative machine learning type is so-called unsupervised learning. No annotations are needed in such setups, and the algorithms attempt to learn patterns and structure in the data without external supervision.

Finding suitable boundaries between normal and anomalous behavior In many applications, it is not trivial to define where to draw the line between normal and anomalous behavior. Especially since anomalies are only encountered very rarely, algorithms do not have many examples (sometimes not any examples at all) in order to learn a decision boundary. A common problem is that not all anomalous patterns which might occur can be known beforehand, and almost every anomaly has a unique pattern or fingerprint. Often, models are trained under the assumption that the data only contains normal behavior. The models attempt to learn a boundary that surrounds this presumably normal data. This task is often referred to as one-class classification [114] and is considered more difficult than other classification tasks. In practice, the data itself usually contains noise, and also the supposedly normal data is regularly "contaminated" with anomalies, which adds extra complexity to the learning task.

Non-stationary environments In many real-world problems, one has to deal with environments which dynamically change over time. Such environments undergo so-called concept drifts or concept changes [47, 170, 177], where certain statistical properties of variables continuously change over time. Time series with constantly changing means, variances, or trends are simple examples of such behavior. In practice, one often encounters dynamically changing time series in which also the regions of normality/abnormality permanently change over time. The challenge for anomaly detection algorithms is progressively updating their models, generalizing the data, and adapting their understanding of normality/abnormality. Especially if an algorithm has to operate on a data stream, the necessity arises to adapt quasi "on-the-fly" to new situations or concepts. Machine learning tasks involving such streaming data are considered rather challenging since many classical learning approaches are not applicable due to their offline character: offline learning algorithms are trained on a static batch of data. They require a complete repetition of the training procedure if new examples are added to the data set. Due to memory and time constraints, such approaches are typically infeasible for problems with streaming data. In such cases, it is necessary to operate in an online (an extreme case is "in real-time", which adds time constraints to the problem) setting on the data and process the data in an example-by-example manner and incrementally learn from every new example, without re-training a completely new

model each time. However, online approaches are associated with several additional new challenges, as discussed in the next chapter(s).

High false alarm rates Since the misclassification of true anomalies in the real-world usually is associated with severe consequences, most anomaly detection algorithms are configured very sensitively. In effect, the algorithms will trigger many alarms and are thus much more likely to classify normal data points as anomalous. A normal data point incorrectly classified as anomalous is called false positive. Since all alarms have to be investigated in real-world applications, false alarms cause much overhead (e.g., for machine operators). It might also happen that the responsible persons become less attentive and, as a consequence, ignore actual anomalous situations. If the false-positive rate gets too large, the expected beneficial effect of the anomaly detection algorithm has to be questioned. As described in the example mentioned earlier about the tragic derailment accident in Eschede, the use of advanced technical equipment for testing train wheels was discontinued since too many false-positive errors were reported.

High-dimensional Time Series While many might consider the detection of anomalies in univariate (1-dimensional) time series as relatively simple, the detection in their multivariate counterparts is typically of tremendous complexity since anomalies cannot be identified in individual series generally. Many anomalies only become apparent by analyzing the time series in all dimensions simultaneously. In other cases, the anomalous patterns range over very few or even only one dimension. However, they are very hard to spot since they are embedded in a high-dimensional space. While it might be possible to perform the anomaly detection in reduced lower-dimensional space for the latter case, it is much more challenging to detect high-dimensional temporal anomalies.

1.1.3 Underlying Research Questions

Based on the aforementioned challenges in anomaly detection, we motivate the work in this thesis with the following central research questions:

> **Q1:** Is it possible to successfully train and apply novel unsupervised machine learning models for anomaly detection and do they advance the state of the art?

This first research question is very general and will accompany us throughout the thesis. There are several sub-questions and assumptions linked to this question. For example, how can "successful" be quantified? The following chapter will describe several well-known performance measures (such as the F_1-score) which indicate an algorithm's performance on an individual time series or even a whole set of time series. By "successful", we mean achieving a performance (according to the specified measures) at least similar to other (unsupervised) state-of-the-art algorithms, ideally even surpassing them.
A central point of this work will be the investigation of *unsupervised* approaches. We will

assume that the training data is not annotated (no anomaly labels will be passed to the algorithms). Only for assessing the performance, the real anomaly labels will be used. We will make the mild assumption that the majority of the training data contains normal instances and that anomalies are relatively rare. However, if not mentioned otherwise, we will not remove any anomalies (or other noisy instances) from the training data, which increases the difficulty of learning and requires noise-resilient algorithms.
Although mentioned before, another aspect that should be highlighted again is the algorithms' requirement to operate on higher-dimensional time series. Considering the inherently temporal nature of the considered data and the multidimensionality increases the anomaly detection task significantly.

> **Q2:** Given certain characteristics of the time series data, can we advise which algorithm is most suited for detecting anomalies?

The second research question is a continuation of the first one. However, it focuses on whether it is possible to classify time series according to various properties and subsequently recommend which anomaly detection algorithm is likely to work best for the given problem. In an ideal case, *one* unique algorithm can be found that is applicable to a wide range of time series and outperforms all other algorithms on the considered problems. However, since this question is framed particularly in light of the data's complex temporal nature and great diversity, it is unlikely that a universally good method exists which performs equally well on all kinds of time series data. For example, such an algorithm would have to be able to deal with complex short- and long-term patterns, with periodic (seasonal) or quasi-periodic behavior and predictable or unpredictable time series. In order to answer this research question, it is necessary to identify distinct characteristics of different time series and find algorithmic approaches to address them.

> **Q3:** How can online learning methods be successfully used for anomaly detection in time series or data streams?

This last research question especially (but not exclusively) addresses problems with non-stationary environments. It has to be investigated how to design online algorithms that can incrementally adapt to new concepts in the data and learn the new notion of "normal". A particular challenge in online setups is the so-called stability-plasticity dilemma [21]: The learning process requires stability in order to protect the already acquired knowledge, but also plasticity, for the integration of new knowledge. A suitable balance has to be found to prevent that too much plasticity causes catastrophic forgetting on the one hand and that too much stability hinders the model from learning new concepts.

In the following chapters, we will address these research questions and propose several approaches in an attempt to answer these questions.

1.2 Thesis Outline

Chapter 2 gives an overview of the state of the art in the field of time series anomaly detection and introduces several benchmarks and algorithms that we use in this work. Furthermore, we describe the general benchmarking/scoring process for the remaining chapters. Chapter 4 is mainly concerned with research questions RQ1 & RQ3 and presents an unsupervised simple online regression anomaly detection algorithm (SORAD) and investigates the importance of its online elements on different types of time series. In chapter 5, we introduce the DWT-MLEAD algorithm. The algorithm uses the discrete wavelet transform (DWT) to analyze time series and detect anomalies at different time scales. We first propose an offline variant of the algorithm and subsequently discuss and implement necessary modifications to obtain a fully online algorithm. A deep learning (DL) approach, called LSTM-AD, based on stacked long short-term memory (LSTM) networks, is introduced in chapter 6. Like SORAD, LSTM-AD predicts a time series's normal behavior, and the prediction errors on several prediction horizons are used to detect anomalous behavior. We apply LSTM-AD to a set of electrocardiogram (ECG) recordings containing various types of arrhythmias. TCN-AE, a temporal convolutional autoencoder based on dilated convolutional neural networks, is presented in chapter 7. The model encodes time series into significantly shorter sequences and uses its decoder's reconstruction error as an indicator for anomalies. TCN-AE is evaluated on anomalous Mackey-Glass time series and, further on, on a more extensive ECG signal set. Finally, in chapter 8, we conclude this thesis and give an outlook on possible future work. In all chapters, we investigate research question RQ1 and implicitly also RQ2. Additionally, chapters 4 & 5 attempt to find an answer for research question RQ3.