Integrating Heterogeneous Contextual Data for Enhanced Time Series Analysis

Saifullah Burero

Supervised by Anton Dignös and Johann Gamper
Free University of Bozen-Bolzano, Italy

Abstract

In the rapidly evolving industrial landscape, sensors are integral to automation applications. Capturing and analyzing the vast amount of time series data is crucial for optimizing processes. However, analyzing this sensor data in isolation presents challenges, particularly in time series analysis, due to the influence of various external contextual factors that are not always apparent. Integrating these contextual factors with time series data is essential for time series analysis. However, these contextual factors are often heterogeneous in the time dimension due to the diverse nature of the data, makes integration challenging. Therefor, as a part of this PhD research that is currently at the beginning of the second year, we aim to introduce a systematic approach for integrating contextual factors with heterogeneous time dimensions. This integration enables the transformation of data with heterogeneous time dimensions into a format that can be effectively processed by machine learning and deep learning models for time series analysis. We use Water Distribution Systems (WDSs) as a representative use case and aim to demonstrate how this integration enhances the accuracy and reliability of time series analysis.

Keywords

time series analysis, heterogeneous time dimensions, time series forecasting, anomaly detection

1. Introduction

In this work, we use Water Distribution Systems (WDSs) as a representative application use case, where sensors are employed for monitoring the consumption of water. These observations serve as crucial inputs for, e.g., detecting water losses or estimating the water demand. WDSs are equipped with networks of sensors and control units, such as flow and pressure sensors, to ensure efficient resource management [1]. These patterns can be analyzed and used to forecast the water consumption, helping to fulfill water demand, and to detect possible losses. However analyzing these patterns solely on sensor measurements can be complex, as various contextual factors influence consumption patterns and/or measurements, such as maintenance, sensor calibration, weather, yearly seasons, tourist trends, temperature etc. These factors pose challenges for machine learning and deep learning models for time series analysis, as they may not be apparent in the data. Figure 1 highlights the importance of contextual information when analyzing time series data, where water consumption data from water distribution system based in Trentino, Italy is reported. The data is segmented and color coded to highlight different consumption patterns. Green segments illustrate "regular" consumption, i.e., low consumption during night with peaks in the morning, afternoon, and evening. Red segments indicate the impact of weekends, pattern that are also observed during bank holidays, reflecting altered consumption behavior during these periods as compared to regular days. Yellow segments highlight deviations likely caused by scheduled maintenance, which occurs monthly from the 7th to the 10th. Additionally, in the lower part of the figure, highlighted in orange, an extended period where the consumption behavior diverges from the normal behavior is highlighted, potentially due to external factors like school vacations, tourist trends, and/or seasonal temperature changes. Although these external factors are not directly visible, their influences can be inferred, making it essential to consider such factors to accurately interpret patterns in time series data.

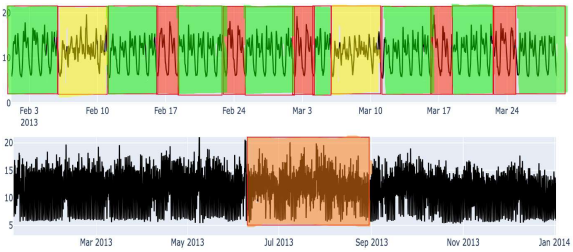

Figure 1: Importance of Context

In the domains of WDSs, many methods have been proposed that rely on historical consumption patterns for analysis. The work by Zanfei et al. [2] emphasizes the importance of integrating external factors such as temperature, humidity, radiation, and rainfall, since these factors strongly correlate with consumption patterns. While such studies focus on contextual data that are time series, such as meteorological factors, which can be integrated with sensory information by aligning them to the same sampling frequency, our approach addresses a broader range of contextual factors. These factors are heterogeneous in the time dimension and cannot be simply treated as time series data. For instance, consider a period of drought that spans several months, where the water consumption pattern at the beginning of this period might still be regular, but water consumption tends to increase the longer the drought persists. In WDSs and in many industrial domains such contextual factors are characterized by heterogeneity in the time dimension. We have identified four distinct types: *static* contextual data that is not changing with time, such as sensor location, sensor sensitivity and other specifications etc., *interval* contextual data that occurs over a period of time, such as drought, periods of hight tourism, vacations etc., *event* contextual data that happens at possibly irregular time points, such as bank holidays, sensor replacements, calibration, cleaning etc., and *secondary time series* contextual data, that are recorded at a regular frequency, such as temperature, humidity, pressure

Published in the Proceedings of the Workshops of the EDBT/ICDT 2025 Joint Conference (March 25-28, 2025), Barcelona, Spain
saifullah.burero@student.unibz.it (S. Burero)

etc. These external contextual factors play a crucial role in shaping how resources are utilized and can significantly impact the accuracy of forecasting and anomaly detection in systems like WDSs. An effective integration of this diverse mixture of data enhances the decision making process by providing a comprehensive understanding of processes. The main objective of this PhD will be on the systematic integration of contextual information with heterogeneous time dimensions (static, interval, event, and secondary time series) for the effective analysis of time series, i.e., missing values imputation, anomaly detection and forecasting.

2. Related Work

According to a recent survey on anomaly detection in the IoT and IIoT domain by Rodríguez et al. [3] that reviewed 99 articles only 8% consider context aware information. Recent studies [4, 5, 6, 7] in contextual anomaly detection application explore combining contextual and behavioral features for improved contextual anomaly detection. These approaches categorize features into contextual and behavioral depending upon domain knowledge, with clustering used to establish context and separate models built for each group.

Daniel et al. [8] developed a three-stage model incorporating temporal features and a sliding window technique for feature representation. Rozhin Yasaei et al. [9] proposed an RNN-based model for clustering sensor behaviors, using a consensus algorithm for anomaly localization. Kosek et al. [10] focused on detecting malicious voltage control actions in the power grid with a deep neural network approach.

Most of the existing works consider internal contexts while analyzing sensory data, such as day, time, yearly seasons, months, and years etc., that are typically uniform and consistent and easily handled, as it is derived from the time dimension of the time series data. On the other hand, integrating external context is more challenging and requires a transformation because the information might be represented using a different time dimension and often lacks uniformity.

3. Objective and Research Questions

The main objective of this thesis is to enhance the performance of machine learning and deep learning models for time series analysis by integrating contextual information with heterogeneous time dimensions together with sensory readings. By acknowledging the crucial role of context in time series analysis, this research endeavors to develop novel methodologies that leverage contextual information to significantly enhance the accuracy and robustness of these models. Through rigorous experimentation and analysis, this study aims to establish a deeper understanding of how contextual information can be effectively integrated into existing frameworks, thereby advancing the state-of-the-art in time series analysis within diverse domains. The main research questions for the PhD are as follows.

RQ1 How to combine relevant contextual information with heterogeneous time dimensions and time series data to build data driven models for time series analysis, such as anomaly detection and forecasting?

RQ2 How to detect patterns in time series data that deviate in specific contexts using data driven approaches?

RQ3 How to build a data driven model that effectively handles heterogeneity and adapts to time series dynamics?

4. Methodology

Figure 2 provides an overview of the proposed approach in combination with an anomaly detection task. Each stage is responsible for performing different task. Initially, our time series data and contextual data with heterogeneous time dimensions from various sources is provided as input to the representation stage. The data undergoes a first transformation step into a homogeneous format understandable to machine learning and deep learning models. After data representation, the data undergoes a feature extraction stage. In this phase, time series features are extracted from the homogeneous format, along with other relevant features depending on the time dimension of the context. The feature extraction process yields many features, necessitating a subsequent feature selection stage. Once important features are obtained, a model based on machine learning or deep learning is constructed to perform anomaly detection.

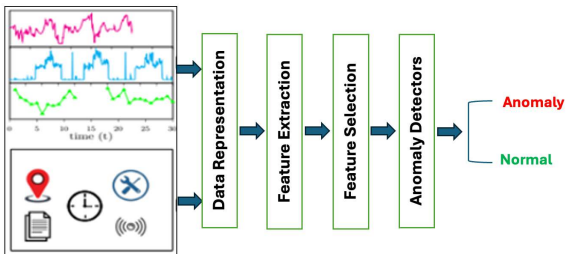

Figure 2: Analyzing time series with contextual data

4.1. Data with Heterogeneous Time Dimensions

Machine learning and deep learning techniques for time series analysis, such as Random Forest, Support Vector Machine (SVM), Long Short Term Memory (LSTM) etc., require input data with a uniform or homogeneous format, i.e, numerical values (vectors) recorded over regular time steps. However, contextual information with heterogeneous time dimensions lacks this uniformity. The effective integration of data with heterogeneous time dimensions involves distinct strategies for each type, and we illustrate it on the example of WDSs.

4.1.1. Time Series Data

WDSs utilize networks of sensors such as flow sensor to monitoring the consumption of water, that generate time series data recorded either regularly or irregularly, with varying sampling frequencies. Along with consumption patterns (primary time series) to be analyzed, secondary time series may also be considered, such as temperature or rainfall. For example, water consumption is highly correlated with temperature, which is also time series data. Tables 1 and 2 exemplify time series data collected from a

water flow sensor and a temperature sensor. This data can be aligned with specific timestamps to ensure consistency, typically achieved through interpolation or resampling to establish regular time steps.

Table 1
Time series data from sensors

Sensor ID	DateTime	Water Flow
1	01-01-2023 01:30	9.01
1	01-01-2023 02:30	8.85
2	01-01-2023 01:30	6
2	01-06-2023 12:30	7.5
1	01-01-2023 03:30	7.90

Table 2
Time series data from locations

Location	DateTime	Temp
Bolzano	01-01-2023 01:00	4
Bolzano	01-01-2023 02:00	5
Bolzano	01-01-2023 02:30	5
Bronzolo	01-01-2023 02:30	6
Bronzolo	01-01-2023 03:30	7

4.1.2. Static Data

Static data, as illustrated in Table 3, remains constant over time and may be associated with every time point, maintaining its validity throughout the period of interest. In WDSs, static data, such as sensor location and characteristics play a crucial role for cross sensor analysis. For instance the location of sensors can provide valuable insights into the population distribution of different regions, which can directly influence factors such as water consumption. Areas with higher population density are likely to exhibit different consumption patterns compared to sparsely populated regions. By considering sensor locations, models can better capture and interpret variations in water consumption data, leading to more accurate and informed analyses in decision making.

Table 3
Static data

Sensor ID	Location	min flow	max flow
1	Bozen	5000	10000
2	Bronzolo	1000	5000

4.1.3. Event Data

Event data as illustrated in Table 4, captures specific time points of an event occurrence. Such data may be integrated by marking relevant time points with event indicators, offering a comprehensive view of discrete events over time. In WDSs, event data captures irregular occurrences such as sensor calibration, cleaning, replacement or bank holidays. Some of this events can directly influence sensor readings, potentially leading to inaccuracies in interpreting water consumption data. Therefore, it is crucial to include such type of information for accurate analysis of water consumption.

4.1.4. Interval Data

Interval data as illustrated in Table 5, are characterized by start and end time points for which a particular contextual information is valid. In the domain of WDSs, interval data represents periods, such as tourism seasons that may or may not occurring annually over some period of time. It is crucial to consider tourism related information as it significantly influences water consumption. During the peak of a tourism season, water consumption increases substantially compared to off-season periods. and therefore, incorporating tourism seasonality into the analysis is essential for accurately understanding the water consumption patterns.

Table 4
Process events

Sensor ID	DateTime	Maintenance
1	01-08-2023 02:00:00	calibration
2	01-09-2023 01:30:00	replacement

Table 5
Interval data

Location	Period	Tourist Trend
Bolzano	01-06-2023—30-10-2023	high
Bronzolo	01-07-2023—30-10-2023	high

4.2. Homogeneous Data

During the data representation stage, data with heterogeneous time dimensions is transformed into a uniform or homogeneous format suitable for machine learning models that require numerical vectors over regular time steps. This stage following the sampling rate of the main time series or user parametrization extracts different values from the data with heterogeneous time dimensions using simple value extraction, different aggregation functions, or one-hot encoding (similar to resampling). For instance at each time step, for constant data it may just extract a value or a one-hot encoding, for event data it may extract a one-hot encoding and/or count of events, for interval data it may extract a count and/or an indicator if a period just started or ended, and for secondary time series it may apply a resampling based on the new frequency. This initial transformation ensures that data has the same homogeneous representation, such as for instance in Table 6 (do not consider the last two columns for now). Each row in the table indicates to a specific time step, while each column represents values extracted from the heterogeneous data. Figure 3 illustrates this transformation process to homogeneous data graphically, making it compatible with machine learning models. This step can be achieved through a combination of densification and aggregation, and some of these values extracted at each time step will be used later to extract features that are related to previous (cross) time steps.

Table 6
Uniform data representation

Sensor ID	Location	flow	Temp	DateTime	Tourism	Maintenance	Last Cal.	Dur. high tourism
1	Bolzano	9.01	4	01-01-2023 01:00	low	calibration	30	0
1	Bolzano	8.05	5	01-01-2023 02:00	low	no	0	0
1	Bolzano	7.90	5	01-01-2023 03:00	high	cleaning	0	2
1	Bolzano	7	5	01-01-2023 04:00	high	calibration	3	2
2	Bronzolo	6.01	6	01-01-2023 01:00	low	no	0	0
2	Bronzolo	9.01	7	01-01-2023 02:00	high	replacement	0	1

4.3. Feature Extraction

After data representation, the subsequent step involves feature extraction, where features from the generated values

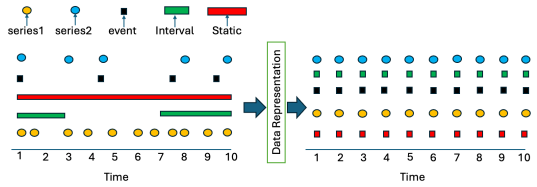

Figure 3: Homogeneous data representation

and primary time series are extracted. Similarly to data representation this stage is sensitive to the type of homogeneous time dimension from which the values were extracted. For instance, for event data, the number of events in the past or elapsed time since a specific event last occurred may be generated, e.g., see "Last Cal." in Table 6 that records the number of hours since the last calibration event. For interval data the time since a period last ended or the duration so far may be generated, e.g., see "Dur. high tourism" in Table 6 that records the duration of high tourism in days. These additional features provide information across time to the model and help setting a given time step into context, ultimately improving model performance. For instance, sensor readings with very high values for the last calibration may be subject to higher variability or outliers, or water consumptions deep in the tourism season may be generally very high. While the previous stage was achieved through a combination of densification and aggregation, this step can generated through window functions aggregating over previously generated values. The key challenge is to identify the required sequence of dependencies.

4.4. Feature Selection

Feature extraction may yield hundreds of features and at this stage it becomes imperative to select features with high importance. Feature selection is indeed crucial for several reasons, including reducing computational complexity and mitigating the risk of overfitting models. Features are chosen based on the model's performance, and depending on the model's output, features are iteratively selected, continuing this process recursively during feature selection. This iterative approach enhance the algorithm's explainability, identifying which features are indeed valuable for decision making.

4.5. Anomaly Detection Model

After feature extraction and selection, the final step involves the development of machine learning and deep learning models for multivariate time series analysis using the example of anomaly detection. To validate the impact of the integrated contextual information in the previous steps, the performance evaluation is twofold. Firstly, model training without incorporating contextual data to establish a baseline performance solely based on available features or signals. Secondly, contextual features are integrated to observe their impact on the model's predictive capabilities. Comparing the model's performance in both settings facilitates understanding of how contextual information enhances predictive accuracy and robustness.

5. Conclusion

Sensor readings often display complex properties influenced by numerous contextual factors. Integrating contextual factors in time series analysis is crucial for accurately interpreting data pattern and identifying factors behind variations. By incorporating contextual factors, we aim to improve performance of machine learning and deep learning models for time series analysis. In this PhD research, currently at the beginning of the second year, we have identified four categories of contextual information with heterogeneous time dimensions. Based on these, we will develop a systematic data management approach for the effective and efficient integration of contextual features for time series analysis.

References

[1] M. Mutchek, E. Williams, Moving towards sustainable and resilient smart water grids, Challenges 5 (2014) 123–137. doi:10.3390/challe5010123.

[2] A. Zanfei, B. M. Brentan, A. Menapace, M. Righetti, A short-term water demand forecasting model using multivariate long short-term memory with meteorological data, Journal of Hydroinformatics 24 (2022) 1053–1065. doi:10.2166/hydro.2022.055.

[3] M. Rodríguez, D. P. Tobón, D. Múnera, Anomaly classification in industrial internet of things: A review, Intell. Syst. Appl. 18 (2023) 200232. doi:10.1016/J.ISWA.2023.200232.

[4] E. Calikus, S. Nowaczyk, M. Bouguelia, O. Dikmen, Wisdom of the contexts: active ensemble learning for contextual anomaly detection, Data Min. Knowl. Discov. 36 (2022) 2410–2458. doi:10.1007/S10618-022-00868-7.

[5] M. A. Hayes, M. A. M. Capretz, Contextual anomaly detection framework for big sensor data, J. Big Data 2 (2015) 2. doi:10.1186/S40537-014-0011-Y.

[6] Z. Li, M. van Leeuwen, Explainable contextual anomaly detection using quantile regression forests, Data Min. Knowl. Discov. 37 (2023) 2517–2563. doi:10.1007/S10618-023-00967-Z.

[7] Y. Shulman, Unsupervised contextual anomaly detection using joint deep variational generative models, CoRR abs/1904.00548 (2019). arXiv:1904.00548.

[8] D. B. Araya, K. Grolinger, H. F. ElYamany, M. A. M. Capretz, G. T. Bitsuamlak, Collective contextual anomaly detection framework for smart buildings, in: IJCNN, IEEE, 2016, pp. 511–518. doi:10.1109/IJCNN.2016.7727242.

[9] R. Yasaei, F. Hernandez, M. A. A. Faruque, Iot-cad: Context-aware adaptive anomaly detection in iot systems through sensor association, in: IEEE/ACM ICCAD, IEEE, 2020, pp. 9:1–9:9. doi:10.1145/3400302.3415672.

[10] A. M. Kosek, Contextual anomaly detection for cyber-physical security in smart grids based on an artificial neural network model, in: CPSR-SG, 2016, pp. 1–6. doi:10.1109/CPSRSG.2016.7684103.