Contents

Bibliography

[MDRD12] Max Mustermann, John Doe, Mario Rossi, and Jean Dupont. An Example Title with Double Braces that is Useful for Weird Capitalization and IEEE Acronyms. In *Proceedings of the 7th International Symposium on WeIrd SHit (ISoWISH), Aachen, Germany*, April 2012.

2

Problem Statement

[ ]Give outline of the subsections and their focus

2.1 Challenges

[x]For data modelling, challenges of unstructured datapoint names, no standarised data format for capturing building context information

[x]For TSC, challenge of not knowing building datapoint metadata. Which tasks does this affect?

[x]Why does faulty data need to be handled? (sensor error, missing data)

A key challenge in the building sector is the lack of a standarised naming convention for buiding datapoints. The use of numeric identifiers and acronyms in datapoint names causes them to not be human-readable. This leads to challenges in several analysis tasks such as development of Digital Twins, comparison of similar datapoints across buildings, efficiency monitoring. A related challenge is the absence of a sector-wide data model for capturing context and objective information within buildings. While the adoption of data models is influenced by multiple factors such as database provider, downstream applications, scalability, the use of different data models leads to difficulties in cross-platform data exchange and overhead of conversion into compatible data formats. Sensors and actuators comprising a Building Management System (BMS) collect and transmit data constantly, leading to a large corpus of tracked data. A data model may define the structure of datapoints monitored by a BMS, and represent logical relationships amongst them. Another challenge to be dealt with is data prone to errors caused by faulty sensor measurements, signal noise, missing data, incorrect timestamps, among other causes. Appropriate handling of faulty data is crucial to ensure data quality for downstream tasks such as anomaly detection and event forecasting.

2.2 Related Work

2.2.1 Data Modelling in Building Management Systems

[ ]BRICK schema, its key features (ontology, based on semantic web)

[ ]Smart Data Models, its key features

[ ]How do these schemas address challenges?

2.2.2 Time Series Classification

[ ]Bode et al. (EBC paper). Unsupervised time series clustering. Why poor results?

[ ]Chen at al. (A Metadata Inference Method for Building Automation Systems With Limited Semantic Information)

[ ]Final paragraph summarising research gaps over mentioned existing work

Automated detection of BMS datapoint metadata has been addressed in recent work using time series data and datapoint labels for detection.

El mokhtari et al. (2021) propose classification of sensor time series using a CNN that receives datapoint time series and its derivative as inputs. They further propose a grey-box model to calculate correlation scores between estimated control time-series and unknown control time-series datapoints. This study however focuses only on sensor time series for the CNN and does not consider boolean, control command, and setpoint time series. They also consider only 4 classes of sensor time series.

Mertens et al. (2022) evaluate several traditional machine learning algorithms to classify ventilation system time series and achieve best results *(mention accuracy value)* with XGBoost, ExtraTrees. For the classification of a test datapoint, they classify each 24-hour window of its time series individually, and then assign the majority predicted class over all windows as the final prediction of a datapoint. However, this 2-step method is impractical for real-time scenarios where the predicted classification is consumed immediately by a subsequent process. Secondly, this study does not address sparsely annotated and/or human-unreadable datapoint labels and only considers ventilation system datapoints.

2.3 Problem Statement

[x]How does a good data model look like? What goals should it achieve?

[ ]What novelty / research gap does our data model plug (compared to BRICK, Smart Data Models)?

[ ]How do you plan to use the data model for time series classification? (mention anecdotes, conceptual sketch)

We divide our solution into three parts: developing a data model for BMS datapoints, unsupervised representation learning for time series data, exploiting the data model for time series classification of BMS datapoints.

Since there is no single correct choice for a data model, we consider a set of objectives that our data model must fulfill. Datapoints from the EBC industrial test hall are used as the working set around which we construct our data model. Since datapoint labels chosen by building operators are cryptic and reveal little or no semantic information about their metadata, our data model aims to encode context information for datapoints (eg: human-readable label, location, highest measurement resolution). Datapoints often have constraints on the underlying values they track, owing to the physical nature of the measured metric. Value range (minimum, maximum), datatype, unit of measure are examples of such constraints that our data model must define. Logical relationships between entities are key to understand the complete building topology and functioning of building energy systems. We build a data model based on RDF schema to utilize RDF triples for encoding entity relationships. This enables the data model to be visualized as a graph. In order to address the challenge of cross-platform data exchange between different BMS, our data model must be serializable into a machine-readable format. In this thesis, we (mention chosen serialization (NGSI-LD or NGSI-v2?)).

Although existing research has investigated approaches for supervised classification of BMS time series using semantic labels with success, unsupervised approaches using only time series data have been largely unaddressed to our knowledge. Addressing the unsupervised setting is an important research goal because datapoint labels are often unavailable, poorly annotated, especially for legacy buildings. In these cases, a classification system that can be trained solely on timeseries data is beneficial. Further, learning representations of time series in an unsupervised setting for BMS has not been explored. Learning discriminative representations for time series fulfills two main goals: (1) learned representations are application agnostic and can be applied across different BMS thereby being scalable, (2) representations can be readily used for several time series analysis tasks (classification, forecasting, fault detection). To elaborate on the idea, we find an embedding space where feature embeddings of time series datapoints can be accurately differentiated.

[Paragraph on using data model for classification]