Takeshi: Application of unsupervised machine learning techniques for topology detection in building energy systems

To cite this article: Florian Stinner et al 2019 J. Phys.: Conf. Ser. 1343 012041

View the article online for updates and enhancements.
Takeshi: Application of unsupervised machine learning techniques for topology detection in building energy systems

Florian Stinner, Lukas Raßpe-Lange, Marc Baranski, Dirk Müller

1 RWTH Aachen University, E.ON Energy Research Center, Institute for Energy Efficient Buildings and Indoor Climate, Mathieustraße 10, 52074 Aachen, Germany
E-mail: fstinner@eonerc.rwth-aachen.de

Abstract. Buildings rarely achieve their energy targets defined in the planning phase. Due to a lack of time and money, operators frequently fail to identify and implement energy efficiency measures in control. Knowledge about the topology of the building energy system (BES) is important for the automatic identification of energy efficiency measures, but usually does not exist in a machine-readable form. In this paper, we present an approach to detect the topology of a BES using an unsupervised learning algorithm, called Takeshi. We apply the algorithm to real and simulated time series data of a multifunctional building. This algorithm relies on a data mining approach, in which four steps are conducted, preprocessing, partitioning, sequencing and rule detection. The results obtained using the real-life data were only partly satisfactory. The best F1 score was 43.3 %, whereby the used seasons of the year had a high influence. In order to demonstrate a broader range of applications, we applied the algorithm to the simulation data. In that case, the algorithm shows significantly better results and the F1 score reached 79.6 %. We evaluate reasons for the poor performance in the case of the real BACS data set and derive possible improvements to the methodology.

1. Introduction

According to a study, 90 % of the total floor area is controlled by faulty automation systems [1]. This shortcoming offers a high potential for energy savings. Building automation and control systems (BACS) provide a large amount of data about states of the building energy system (BES). The condition monitoring of the BES offers a high potential to convert faulty BACS into functioning and energy-saving BACS. For the detection of potential measures, a very high effort with a high level of system knowledge and a poor cost-benefit ratio is currently required.

Due to lower personnel expenses, algorithms on the basis of artificial intelligence (AI) could contribute to a rapid reduction of costs to implement energy saving measures. An important basis for the development of the measures is knowledge of the BES topology, i.e. how components of the BES are linked with each other. However, this knowledge rarely exists in a digital form. Therefore, it is necessary to use an algorithm for the preparation of the data.

Previous works have segmented BACS data according to find reoccurring patterns in the energy consumption of an entire building[2][3]. [4] and [5] shows that active control supports the detection of the connection between variable air volume (VAV) boxes and the supplied space.
To the authors’ knowledge, no previous approach has topology recognition in hydronic systems as its goal.

We present our algorithm Takeshi, based on unsupervised learning to detect the states of the BES in BACS data. Based on a general approach for BACS data [3], a 4-phase procedure is used to identify the relationships between data points in these states. We have applied suitable approaches for the respective phase that are suitable for topology detection.

Our algorithm detects the direct correlations between data points. For example, it detects which data point shows a change if a control command switches from off to on. These correlations can then be used in various advanced applications. We apply our algorithm to the real data and simulated data of the energy conversion of a multifunctional building [6].

We structure the paper as follows: first, we describe all four phases of our algorithm Takeshi and the used data sets. Afterwards, we show the results and possible future work.

2. Method
In this paper, we present our algorithm “Topology detection of Artificial intelligence based Knowledge discovery of Energy Systems and Hvac systems for energy measures Identification” (short: Takeshi) that is capable of detecting correlated sensors of a physical system in a data set, based only on the multivariate time series data of a BACS. We apply Takeshi to time series data of the control signals and sensors of thermal water systems.

As described before, the method is divided into four phases. In phase 1, the data is prepared in a preprocessing step. In phase 2, the time series are divided into clusters that constitute the states of the system and represented as a sequence database (see Figure 2). In phase 3, the sequences are then examined for association rules. Finally, phase 4 identifies which data points are contained in the individual rules.

The data sets consist of a set of data points. We use the following definition for data points: “a data point is an information carrier that continuously provides information about a state “ [6]. Each data point can represent either a control signal that operates the system or a sensor that monitors it. Each data point features a time series of observations. The resulting input data set is a matrix in which each row represents an observation at a single point in time and each column is the time series of a single data point.

In the creation of the method, we expected that the following three assumptions apply to a data set: First, the system alternates between different operation modes, that are determined by the configuration of control signals. Remaining in an operation mode causes the system to develop into a stationary state. Second, the changes and patterns of temperature and volume flow sensors are strongly correlated to the configuration of the control signals. Third, different control signals change their state independently and often enough, so that the caused effects can be observed in isolation.

The goal is to isolate the effect that is caused by the change of a control signal, and analyze it to detect the data points that correspond to the changing control signal. Figure 1 shows the whole process for topology detection. At the end, Takeshi gives the probability that a certain data point is correlated with another data point (see Figure 4). This means to what extent a ”switch on” of one data point affects another data point. A value higher than 50% is interpreted as correlated data point. We compared these correlated data points with connections between systems in the real construction plans.

2.1. Data preprocessing
In the preprocessing step, we identify the control signals of the data set and clean the data of outliers. BACS data usually contains a substantial amount of outliers which can impede the results of data mining techniques [3].
Data cleaning methods

Hampel filter for outlier detection

Data transformation

Min/Max Normalization

Input:
Sequence database

Method:
TRuleGrowth algorithm, creation of subsets, analysis of variance

Output:
Correlating data points in rules

Applications

Topology detection

Figure 1. Topology recognition process: four phases (preprocessing, partitioning, knowledge discovery and post mining) to transfer raw data to applications

We use the Hampel filter to remove outliers. The Hampel filter is a decision filter that analyzes a window of a given time series and replaces the central value with the median of the window if it is identified as an outlier [7]. We use the Hampel filter with a window size of 7 and a threshold of 3.

The second task is the normalization of the data. This is necessary to achieve comparability between different data points (e.g. volume flows and temperatures). Therefore, we use the min-max normalization method. Min-max normalization puts every data point in a margin between zero and one respective to their minimum and maximum values. [8]

2.2. Data partitioning

The main purpose of data partitioning is to divide the data set into smaller subsets, to improve the results and reduce the overall computational load[9]. Therefore, in a preliminary step, the data set is split into subsets according to its control signal configurations. Each subset represents the system in one of the possible operation points (e.g. a heat pump is switched on and a boiler is switched off). Figure 2 shows an example of data partitioning.

However, the subsets still comprise observations from the stationary state and each transition into it from another operation point. For the success of this method it is desirable to analyze the transition from one stationary state into another most isolated. To achieve that, we partition each of the previously obtained subsets with help of the k-medoids clustering algorithm in order to find the stationary and transient states for all operation points.

The k-medoids algorithm is a close relative of the more popular k-means clustering algorithm, but is different in that each center of a cluster is also an object of the data set itself. [8] It was implemented in the example by Bauckhage. [10] K-medoids does not necessarily find the best solution and can converge to local minima. Consequently, we implemented the Silhouette Cluster Validity Index (SCVI) to ensure finding a good result. [11] We use for the application of the
k-medoids algorithm a k-value of 2 for stationary and transition states and as distance the value of the SCVI.

After the application of the k-medoids algorithm the results are used to build a sequence database, which describes the temporal succession of operation points and states the system reaches. Each entry in the sequence database represents one of the clusters obtained by the previous step.

2.3. Knowledge discovery

In the knowledge discovery step, we intend to find typical transitions from one operation point to another. We analyze these transitions to determine which data points react to the change of a control signal in order to identify subsystems and how they relate to one another.

We use the TRuleGrowth algorithm [12] to examine the previously constructed sequence database for temporal association rules. These rules indicate which transitions from one operational point to another occur frequently in the system. We construct subsets that contain occurrences of a certain rule. In these subsets only those data points that correspond to the changing control signal change significantly and can be identified by their high variance.

For each rule, we iterate through all data points of the subset that occur in one rule. Then we can calculate the variance of the data point in this subset. If the variance does not exceed a predefined threshold it is discarded. The remaining data points are those that react to the change of the control signal described by the rule and are saved as a rule of correlating data points. The outcome is a rule of correlating data points for each of the temporal association rules found by the TRuleGrowth algorithm. We use the TRuleGrowth algorithm for all rules with a window size of 2, a minimum support of 0.2, a minimum confidence of 0.5, a maximum antecedent size of 1, a maximum consequent size of 1 and a variance threshold of 0.01.

2.4. Post mining

In the final step, we analyze and interpret the results. The rules of correlating data points are evaluated to determine the probability of connections between data points. This results in a co-occurrence matrix which can be used as a basis for a heat map or hierarchical clustering algorithm to visualize the topology of the system (see Figure 4). In case of a known Ground Truth, the results can also be evaluated by a F1 score, accuracy, recall and precision.

2.5. Data set

2.5.1. Real data set

The data set we use originates from a multifunctional building in which various types of energy conversion plants, distribution systems and room supply systems are installed. [6] gives an overview of the used data set and further information on the usage of the data set. We used the temperature and volume flow (discretized as control signal) of the energy conversion (boiler, combined heat and power (CHP), heat pump and chiller).

We sampled the data sets from five representative months in 2015 and 2017 and split them into their 4 weeks. This results into 19 data sets (as 1 has to be dropped due to too many missing values). Each data set consist of 20 data points with 10080 observations (1 week - 1 min resolution).

2.5.2. Simulated data set

We developed a simplified model of energy conversion that simulates the pipelines of the real building together with the technical equipment. It contains a source and a sink on the heat and cold side, which represent the energy usage in the building. A boiler, a CHP and a heat pump are implemented as technical components.

The motivation for using simulated data is to check whether topology detection achieves a better result in a controlled and undisturbed environment. We expect that the disturbance variables in a real BES have a high influence on the detection during normal operation.
We use control signals that are distributed alternately over the entire runtime. These switch
the technical systems on and off in repetitive patterns. Each data set consists of 16 data points
and 10080 observations (1 week - 1 min resolution).

3. Results

Figure 3. Results of real and simulated BACS data with average accuracy, precision,
recall and F1 score over all data sets.

Figure 4. Results of synthetic BACS data as confusion matrix with hierarchical clustering.
Each square visualizes the probability that a “switch-on” signal of one data point
influences another data point (black=100% and white=0%).

Figure 3 shows that our algorithm Takeshi achieves an accuracy [13] of up to 64.2% in real
data, while the F1 score only reaches a maximum of 43.3%. The average F1 score is 32.5%.
These results differ greatly from the simulation data.

Due to the external conditions, only one technical system was used for the supply in some
weeks. As a result, this system was not switched off. Therefore, the TRuleGrowth algorithm
is not able to generate a rule describing the change of the system (e.g. by the boiler switching
on and off). As a consequence the method can not find any sensor corresponding to the control
signal of the boiler because it does not appear in any rule as its variance is zero.

When a boiler is on continuously, temperatures of 75-80 °C occur. With min-max
normalization, these values are normalized as 0-1. Due to the incorrect found switch event,
rules incorrectly include the temperature sensors, because the variance filter is unable identify
them as irrelevant, since their normalized variances exceed the threshold.

In addition it occurs that components only switch on for a short interval or are constantly
switching on and off and the system does not reach stationary states. In some example weeks,
too few stationary states were found in the data. However, this is decisive for the method.
Apparently, the implemented control leads to technical systems either running for a very long
time (boiler) or changing their state very quickly (heat pump).

With simulative data, i.e. if we can determine the control of the BACS ourselves, Takeshi
achieves an average accuracy of 85.5% and an F1 score of 77.6%, as Figure 4 shows. The results
of the simulation data sets are very similar to each other.
4. Future work
First of all, in a data set with a time interval of one week, it often occurs that a component does not switch its state at all. The results of simulation data shows that a controlled undisturbed environment improves results. It only makes sense to investigate systems that change their status. Here, Takeshi would have to be preceded by a selection mechanism that selects beforehand whether a system with inclusion of further time series has changed.

In addition, the fact that a stationary state is rarely reached, depending on the system, is a problem for finding rules for connections. TRuleGrowth showed very good results in a controlled and undisturbed environment, but was not suitable for non-stationary conditions to find appropriate connections. An algorithm that does not need these stationary states to find suitable rules would be helpful here.

5. Conclusion
We were able to show that our 4-phase algorithm Takeshi (preprocessing, partitioning, knowledge discovery and post mining) based on Hampel filter, k-medoid, TRuleGrowth algorithm and hierarchical algorithm is suitable for finding connections between sensors. We concentrated on the thermal energy supply, that has not been selected before for topology detection. In this field, connection detection is particularly difficult to achieve due to thermal inertia. Nevertheless, we show with simulation models that Takeshi can achieve good results with an F1 score of 77.6%.

We were also able to determine what the biggest obstacles are to topology detection based on real data: no control actions and too many control actions. For good topology recognition, a good tradeoff must be found. Especially outside normal operation, this algorithm offers potential for improving energy efficiency of building automation and control systems.

Acknowledgments
The authors would like to acknowledge the financial support of the German Federal Ministry of Economic Affairs and Energy (promotional reference 03SBE0006A).

References
[1] Waide P, Ure J, Karagianni N, Smith G and Bordass B 2014 The scope for energy and co2 savings in the eu through the use of building automation technology: Final report
[2] Miller C, Nagy Z and Schlueter A 2015 Automation in Construction 49 1–17
[3] Fan C, Xiao F, Madsen H and Wang D 2015 Energy and Buildings 109 75–89
[4] Koh J, Balaji B, Akhlaghi V, Agarwal Y and Gupta R 2016 Quiver: Using control perturbations to increase the observability of sensor data in smart buildings
[5] Pritoni M, Bhattacharya A A, Culler D and Modera M 2015 Short paper: A method for discovering functional relationships between air handling units and variable-air-volume boxes from sensor data Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments ed Culler D, Agarwal Y and Mangharam R (ACM) pp 133–136 ISBN 9781450339810
[6] Stinner F, Yang Y, Schreiber T, Bode G, Baranski M and Müller D 2019 Generating generic data sets for machine learning applications in building services using standardized time series data ISARC - 36th International Symposium on Automation and Robotics in Construction
[7] Pearson R K, Neuvo Y, Astola J and Gabbouj M 2016 EURASIP Journal on Advances in Signal Processing 2016 115
[8] Han J, Kamber M and Pei J 2012
[9] [9] Fan C, Xiao F and Yan C 2015 Automation in Construction 50 81–90
[10] Bauckhage C 2015 Numpy / scipy recipes for data science: K-medoids clustering
[11] Luna-Romera J M, Garcia-Gutiérrez J, Martinez-Ballesteros M and Riquelme Santos J C 2017 Progress in Artificial Intelligence 550 92
[12] Fournier-Viger P, Wu C W, Tseng V S, Cao L and Nkambou R 2015 IEEE Transactions on Knowledge and Data Engineering 27 2203–2216
[13] Nanopoulos A, Alcock R and Manolopoulos Y 2001 Feature-based classification of time-series data Information processing and technology ed Mastorakis N and Nikolopoulos S D (Nova Science Publishers, Inc) pp 49–61 ISBN 1-59003-116-8