Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 122 (2017) 943–948

Energy Procedia

www.elsevier.com/locate/procedia

CISBAT 2017 International Conference – Future Buildings & Districts – Energy Efficiency from Nano to Urban Scale, CISBAT 2017 6-8 September 2017, Lausanne, Switzerland

Application of selected supervised learning methods for time series classification in Building Automation and Control Systems

Johannes Fütterer[a]*, Maksymilian Kochanski[b], Dirk Müller[a]

[a]RWTH Aachen University, E.ON ERC, Institute for Energy Efficient Buildings and Indoor Climate, Mathieustraße 10, 52074 Aachen, Germany
[b]Research and Innovation Centre Pro-Akademia, Poland

Abstract

Acquiring knowledge from the growing amount of Building Automation and Control Systems (BACS) data is becoming a more and more challenging and complex engineering task. However, it is also a prerequisite for smart and sustainable energy management as well as improving energy efficiency and comfort of building users. This report analyses the prospects of applying selected supervised learning methods for time series classification in BACS. Our training and testing data covered multivariate time series from 5,142 data points located in E.ON Energy Research Center building, describing observations from 22 classes, such as temperatures of gaseous fluid, CO2 concentrations, heat flows, and operating messages. We trained thirteen types of classifiers: complex tree, medium tree, simple tree, linear Support Vector Machines, quadratic Support Vector Machines, boosted trees, bagged trees, subspace discriminant, subspace KNN, RUSBoosted Trees, Fine KNN, Coarse KNN and random forests. The highest demonstrated average classification accuracy concerned bagged trees (56.76%), with the maximum accuracy level of 76.54%. However, the maximum accuracy achieved by random forests was even higher, reaching 78.95%. Finally, we identified factors that may have a substantial influence on performance of particular methods.

Peer-review under responsibility of the scientific committee of the CISBAT 2017 International Conference – Future Buildings & Districts – Energy Efficiency from Nano to Urban Scale

Keywords: big data, supervised learning, time series classification, operating data, building, building automation, building energy system, building performance optimization

* Corresponding author. Johannes Fütterer
E-mail address: jfuetterer@eonerc.rwth-aachen.de

Peer-review under responsibility of the scientific committee of the CISBAT 2017 International Conference – Future Buildings & Districts – Energy Efficiency from Nano to Urban Scale
10.1016/j.egypro.2017.07.428

1. Introduction

Energy consumption in buildings in developed countries is responsible for approximately 40% of total energy use and is above industry and transport figures in the EU and USA [1]. Buildings' indirect and direct impact on human health, environment, and economy is tremendous, though it is often not in the spotlight of sustainability considerations. Improvements of buildings energy efficiency are of enormous potential, however, any decision-making processes in this field need to be based on reliable evidence. In view of this fact, Building Automation and Control System (BACS) are not only getting more and more popular, but they are also collecting more and more data. By 2020, the 980 million meters installed in buildings worldwide will generate 431,000 petabytes of data a year [2]. Acquiring knowledge from this unprecedented amount of data is becoming a more and more complex engineering task. Beside meters, the same accounts for BACS operational data points. Gathering knowledge is a prerequisite for smart energy management, improving energy efficiency and comfort of building users.

The E.ON Energy Research Center's main building may serve as a representative example of information-intensive source of BACS data. With over 9000 data points, a significant amount of human work has been done to assign the data points to correct classes of data (e.g. liquid temperature, gas temperature, operational messages, alarms), which is usually one of first steps for data analysis. Application of selected supervised learning methods for time series classification aims to present possibilities of automating this process so that as much data as possible is categorized automatically. The basic motivation of this research lies in increasing engineering productivity and efficiency while setting up monitoring projects aimed at improving sustainability of buildings.

2. Related work

Supervised learning methods have already gained widespread popularity in various areas, such as movement recognition (e.g. Kinect for Xbox) [3], text recognition [4], 3D brain scans [5], insects monitoring [6], DNA sequences identification [7] and isolation of household devices based on electricity usage profiles from smart meters data [8]. Within [9], the authors present a comprehensive review of research towards unsupervised statistical learning and visual analytics techniques applied to building performance analysis. Still, to the best of our knowledge, we are the first to address the possibilities of applying supervised learning methods in improving BACS efficiency.

3. Problem statement

First, we define a set of data points X with n elements: $X = \{x_1, x_2, \dots, x_n\}$, with two subsets:

- Subset of training data points X_{tr}, $X_{tr} \subseteq X$; with a elements: $X_{tr} = \{x_{tr_1}, x_{tr_2}, \dots, x_{tr_i}, \dots, x_{tr_a}\}$
- Subset of testing data points X_{te}, $X_{te} \subseteq X$; with b elements: $X_{te} = \{x_{te_1}, x_{te_2}, \dots, x_{te_j}, \dots, x_{te_b}\}$

where: $X = X_{tr} \cup X_{te}$. All elements of $X\{x | x \in X\}$ are coincidentally distributed to X_{tr} or X_{te}, with the following condition: $a = \|0.7\, n\|$ and $b = n - a$.

Second, we define exactly one time series (column) x_a for each training data point x_{tr_i}:

$$\forall\, x_{tr_i} \;\; i = 1, 2, \dots, a \qquad \exists!\; x_a = \{x_{a1}, x_{a2}, \dots, x_{am_a}\}$$

The number of observations m_a is time-series-specific for each time series: $m = \{m_1, m_2, \dots, m_a\}$. By analogy, we define exactly one time series (column) x_b for each testing data point x_{te_j}:

$$\forall\, x_{te_j} \;\; j = 1, 2, \dots, b \qquad \exists!\; x_b = \{x_{b1}, x_{b2}, \dots, x_{bk_b}\}$$

The number of observations k_b is time-series-specific for each time series: $k = \{k_1, k_2, \dots, k_b\}$.

Third, we derive vectors p_a and q_b including s characteristic statistical features for each time series x_a and x_b respectively:

$$\forall\, x_a \;\; \exists!\;\; p_a = \{p_1, p_2, \dots, p_s\}$$
$$\forall\, x_b \;\; \exists!\;\; q_b = \{q_1, q_2, \dots, q_s\}$$

Fourth, we assign a class c_a to each data point included in the subset of training data points X_{tr}:

$$\forall\, x_{tr_i} \;\; i = 1, 2, \dots, a \qquad \exists!\; c_a$$

where $c_a = 1 \vee c_a = 2 \vee \dots \vee c_a = r$ and $r \in \mathbb{Z}$ is the number of data types (classes) in X.

Finally, we pose the following research question: