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

Smart Buildings (Predictive & Neuro-Fuzzy Control)

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

1876-6102
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:

$$\Omega: \quad (x_b,\ q_b) \xrightarrow{(x_a, p_a, c_a)} P(c)$$

In other words: we aim to create a probabilistic classifier Ω, being a function implemented by our classification algorithm, which for each unclassified time series x_b characterized by the related vector of their statistical features q_b is able to predict a probability distribution P over the known set of classes $c \in < 1; r >$ as accurately as possible, given the input training time series x_a, each characterized by the related vector of their statistical features p_a and the related value of their class c_a.

4. Data set

Training data sets consisted of multivariate time series from 70% of 5,142 data points located in E.ON Energy Research Centre building. Data was collected between 15-12-2015 12:00:00 until 16-12-2015 11:59:59, with 60s sampling time. The time series of values were further characterised by the following eight features, selected upon recommendations of [10]:

- Value
- Mean
- Variance
- Skewness
- Kurtosis
- Minimum
- Maximum
- Class (1 to 22).

The values obtained from the remaining 30% of data points were used for testing the classifiers' accuracy. Assigning data points to both categories (training and testing) was performed randomly. As a result, the initial training matrix included 5,464,800 rows, while the initial testing matrix consisted of 2,337,120 rows.

Table 1. Classes of data points in E.ON Energy Research Centre building

Class number	Class name	Class description
1	AL	Malfunctioning message/Alarm/Maintenance message
2	C	Counter
3	CO2	CO_2 concentration
4	HF	Heat flow
5	OM	Operating message (On, Off, Opened, Closed, Active, Not active)
6	P	Power
7	pressure	Pressure
8	revs	Revolutions / frequency
9	rh_BAS	Relative humidity
10	SP_O	Set point (operation/operation request/release/schedule)
11	SP_Percent	Set point in percent
12	SP_T	Set point (temperature)
13	SP_T_Pot	Set point (temperature) via potential meter
14	T_g	Temperature of gaseous fluid
15	T_l	Temperature of liquid fluid
16	V_dot_g	Volume flow for gaseous fluid
17	V_dot_l	Volume flow for liquid fluid
18	VOC	Volatile organic compounds measured by BAS
19	VP	Device status in percent: valve position/power indicator
20	w	Electric work
21	WSP_percent	Working set point (0-100) and positioning/position feedback
22	WSP_T	Working set point (temperature)

5. Design of experiment

The training data was analysed using MATLAB and Salford Predictive Modeler® software suite. Thirteen types of classifiers were trained: 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. In each case 5-fold cross-validation was applied. Data was divided into five folds and then each fold was left out of the model learning process and used as a test set, repeatedly five times. For optimising the computational effort, 10,000 observations from training set and 4,000 observations from testing set were randomly selected for assessing the best classifier's performance.

6. Results

The estimated performance (the average of five trained classifiers of each type from cross-validation) is summarised in Figure 1. The models accuracy ranged from 29.7% for linear SVM to 73.2% for bagged trees.

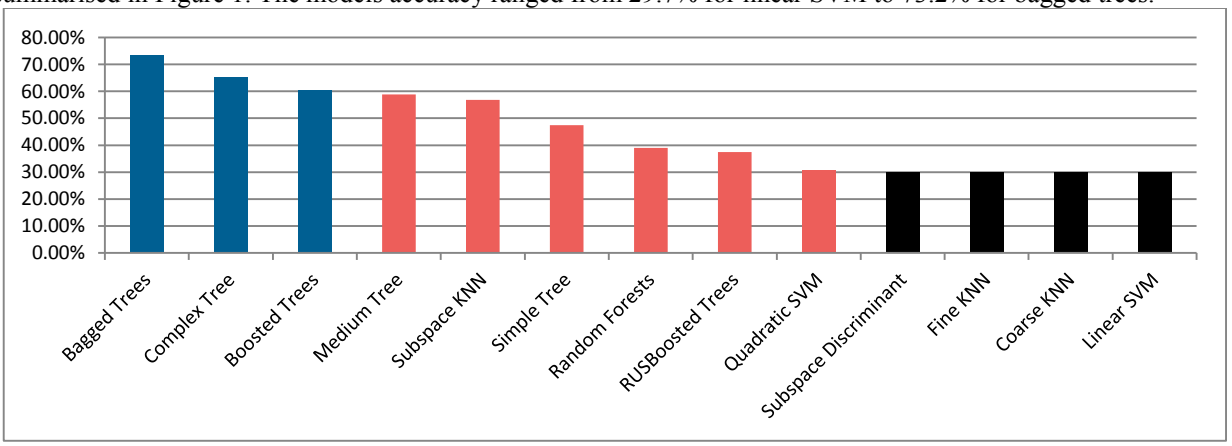

Figure 1. Accuracy of the tested types of classifiers after cross-validation

Subsequently, the most accurate classifier was analysed using Receiver Operating Characteristic (ROC) curves and confusion matrix, which are presented in

Table 2 and Figure 2 respectively. The biggest area under ROC curve (0.9999) was obtained for class 2 (C), class 3 (CO2), and class 18 (VOC), while the smallest area concerned class 10 (SP_O) – 0.8424, class 21 (WSP_percent) – 0.8479 and class 1 (AL) – 0.8659.

Table 2. Receiver operating characteristics for each class predicted by Bagged Trees after cross-validation

Class	Area under ROC curve	Class	Area under ROC curve
1	0.8659	12	0.9599
2	0.9999	13	0.9907
3	0.9999	14	0.9945
4	0.9998	15	0.9977
5	0.8991	16	0.9073
6	0.9610	17	0.9823
7	0.9662	18	0.9999
8	0.9114	19	0.9633
9	0.9873	20	0.9996
10	0.8424	21	0.8479
11	0.9028	22	0.9887

As shown in s class 5 (operating messages)., most of misclassifications in the cross-validation procedure concerned classes characterised by binary values. Classes 1, 10, 11, 12 and 21 (AL, SP_O, SP_Percent, SP_T, and WSP_percent respectively) were most often falsely classified as class 5 (operating messages).

As a next step, the model was tested with untrained time series that were obtained from the data points not included in the training set. The overall accuracy reached 75.75%, which was even higher than the cross-validation results (73.20%). Classification accuracy results are presented in Table 3. The following classes were classified with 100% accuracy: AL, C, CO2, HF, P, pressure, revs, rh_BAS, T_l, V_dot_g, V_dot_l, VOC, WSP_percent. Accuracies calculated for classes related with small testing sets (e.g. w, C, revs) have to be treated with limited confidence.

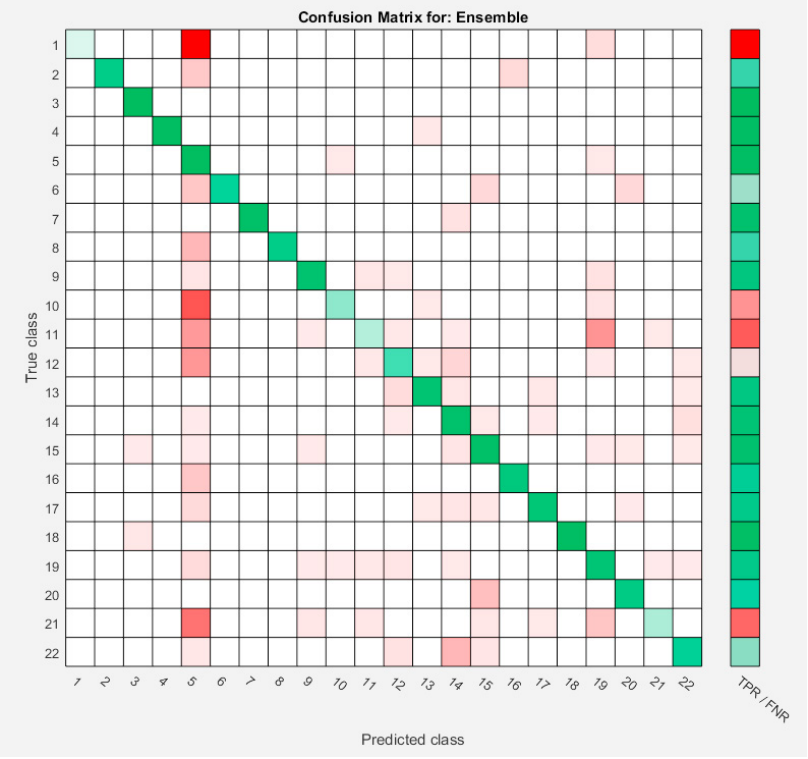

Figure 2. Confusion matrix for Bagged Trees after cross-validation

Table 3. Classification accuracies for testing set (Bagged Trees)

Class	Correctly classified	Total sample size	Class	Correctly classified	Total sample size
AL	100%	34	WSP_percent	100%	26
C	100%	3	SP_O	97%	146
CO2	100%	73	T_g	95%	417
HF	100%	42	SP_T_Pot	94%	77
P	100%	7	SP_Percent	91%	47
pressure	100%	13	WSP_T	80%	76
revs	100%	3	Average for all classes	76%	4000
rh_BAS	100%	64	VP	75%	434
T_l	100%	298	SP_T	68%	92
V_dot_g	100%	14	OM	60%	1973
V_dot_l	100%	90	w	50%	4
VOC	100%	67			

7. Discussion

We discovered that there are classes that are easier to predict and classes that are harder to predict. This depends on how unique the respective time series' set of attributes is in comparison with others. As we have shown, it is possible to identify important time series, such as temperatures, CO2-concentration and volume flows. Taking a closer look at the results reveals that binary values as well as set point values that do not change frequently are harder to allocate. Their sets of attributes are harder to allocate less due to their lower uniqueness. In order to address this result, we will reassigning classes within future research. The first step will be to summarize classes with binary values. We assume that it is possible to differentiate between binary values and set point values.

We assume that the transferability of a classifier trained with data of one building will work within another building, which is a prerequisite for achieving work reduction during the engineers' mapping activities as addressed below. We will address this issue within our future research.

The main application idea is to have a classifier that supports data analysts via pre-labeling. The results show that it is possible to allocate arbitrary time series to classes. For the building-performance-optimizing engineer, the knowledge about a class makes it far easier to map the time series with a semantic meaning within the building energy system. Further, it will be possible to apply direct and fully automated data analysis with the gathered knowledge. Application of the proposed method can thus lead to work reduction for data-driven building performance assessment and optimization.

8. Conclusion

In this paper, we demonstrated that applying supervised learning methods are not only feasible, but may also deliver meaningful, accurate and useful results in solving practical engineering tasks concerning BACS development. We identified a new area for further investigations that has a significant potential for both basic and applied research.

We confirmed that together with Bagged Trees, Random Forest classifiers are among the most successful methods currently available to handle big data originating from BACS.

Next steps of research should optimise the classification performance, i.e. maximise the classification accuracy and minimise the computational cost. At the same time, potential end-user requirements have to be clearly defined and followed. Research limitations (e.g. testing the classification accuracy with data from other buildings) have to be addressed as well.

References

[1] EIA, "Annual Energy Outlook 2012 Early Release," 2012.
[2] C. L. Stimmel, Big data analytics strategies for the smart grid. Boca Raton, FL: CRC Press/Taylor & Francis Group, 2015.
[3] R. M. Araujo, G. Graña, and V. Andersson, Towards skeleton biometric identification using the microsoft kinect sensor: ACM. Available: http://dl.acm.org/ft_gateway.cfm?id=2480369&type=pdf.
[4] M. Zahedi and Eslami S, "Improvement of Random Forest Classifier through Localization of Persian Handwritten OCR," International Journal on Information Technology, vol. 2, no. 1, 2012.
[5] W. Qiang, D. Kontos, L. Guo, and V. Megalooikonomou, "Application of time series techniques to data mining and analysis of spatial patterns in 3D images," in Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on, 2004, pp. iii-525-8.
[6] S. Kasetty, C. Stafford, G. P. Walker, X. Wang, and E. Keogh, "Real-Time Classification of Streaming Sensor Data," in 2008 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), pp. 149–156.
[7] H.-M. Hsueh, D.-W. Zhou, and C.-A. Tsai, "Random forests-based differential analysis of gene sets for gene expression data," Gene, vol. 518, no. 1, pp. 179–186, 2013.
[8] J. Lines, A. Bagnall, P. Caiger-Smith, and S. Anderson, "Classification of Household Devices by Electricity Usage Profiles," pp. 403–412.
[9] C. Millera, Z. Nagyb, A. Schluetera, "A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings", ETH Zürich, Institute of Technology in Architecture (ITA), Architecture and Building Systems (A/S), Zürich, Switzerland, and Intelligent Environments Laboratory, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, Austin, TX, USA, Working Paper under submission, DOI: 10.13140/RG.2.2.12121.93287, 2017
[10] A. Nanopoulos, R. Alcock, Y. Manolopoulos, Feature-based Classification of Time-series Data.