Energy and Buildings 144 (2017) 191–206

Contents lists available at ScienceDirect

Energy and Buildings

journal homepage: www.elsevier.com/locate/enbuild

An ensemble learning framework for anomaly detection in building energy consumption

Daniel B. Araya[a], Katarina Grolinger[a], Hany F. ElYamany[a,c], Miriam A.M. Capretz[a,*], Girma Bitsuamlak[b]

[a] Department of Electrical and Computer Engineering, Western University, London, Ontario, Canada N6A 5B9
[b] Department of Civil and Environmental Engineering, Western University, London, Ontario, Canada N6A 5B9
[c] Department of Computer Science, Suez Canal University, Ismailia, Egypt

ARTICLE INFO

Article history:
Received 19 July 2016
Received in revised form 3 November 2016
Accepted 26 February 2017
Available online 10 March 2017

Keywords:
Anomaly detection
Ensemble learning
Autoencoder
Support vector regression
Random forest
Building energy consumption

ABSTRACT

During building operation, a significant amount of energy is wasted due to equipment and human-related faults. To reduce waste, today's smart buildings monitor energy usage with the aim of identifying abnormal consumption behaviour and notifying the building manager to implement appropriate energy-saving procedures. To this end, this research proposes a new pattern-based anomaly classifier, the *collective contextual anomaly detection using sliding window* (CCAD-SW) framework. The CCAD-SW framework identifies anomalous consumption patterns using overlapping sliding windows. To enhance the anomaly detection capacity of the CCAD-SW, this research also proposes the *ensemble anomaly detection* (EAD) framework. The EAD is a generic framework that combines several anomaly detection classifiers using majority voting. To ensure diversity of anomaly classifiers, the EAD is implemented by combining pattern-based (e.g., CCAD-SW) and prediction-based anomaly classifiers. The research was evaluated using real-world data provided by Powersmiths, located in Brampton, Ontario, Canada. Results show that the EAD framework improved the sensitivity of the CCAD-SW by 3.6% and reduced false alarm rate by 2.7%.

1. Introduction

Commercial and residential buildings account for roughly 60% of the world's electricity consumption [1]. During building operation, a significant portion of energy consumption may be wasted due to various equipment or human-related faults. By reducing building energy waste and enhancing building consumption efficiency, facilities can minimize utilities cost and reduce the associated negative impact on the environment. This can also help address the growing energy demand that the world faces today.

One promising approach to energy efficiency goals is to monitor building energy consumption with the aim of identifying abnormal consumption patterns. Once identified, this abnormal consumption behaviour can be reported to the building manager, who can subsequently perform appropriate energy-saving procedures. In recent years, with the proliferation of sensor devices, monitoring building consumption behaviour for anomaly detection purposes has become easier.

Anomaly detection refers to the process of detecting abnormal events that do not conform to expected patterns [2]. Broadly, depending on their types, anomalies can be classified as point, contextual or collective anomalies [2]. If a data instance is anomalous compared to the rest of the data, then it is referred to as a *point anomaly*. For instance, a daily lighting energy consumption value might be anomalous compared to previously recorded daily values. If a data instance is normal in one context but anomalous in another, then it is referred to as a *contextual anomaly*. For instance, an hourly school lighting consumption value might be anomalous on weekends when there are no classes, but not on weekdays when there are classes. If a group or collection of related data instances is anomalous in comparison to the rest of the dataset, then it is referred to as a *collective anomaly*. Individually, these values might not be anomalous, but collectively they represent an anomalous occurrence. For instance, the individual values of a daily profile of heating, ventilating, and air conditioning (HVAC) consumption data recorded every hour might be normal compared to previous recorded values, but collectively, the daily profile might represent an anomalous consumption pattern.

* Corresponding author.
E-mail addresses: dberhane@uwo.ca (D.B. Araya), kgroling@uwo.ca (K. Grolinger), helyama@uwo.ca (H.F. ElYamany), mcapretz@uwo.ca (M.A.M. Capretz), gbitsuam@uwo.ca (G. Bitsuamlak).

http://dx.doi.org/10.1016/j.enbuild.2017.02.058
0378-7788/

Nomenclature

A	learner algorithm
AC	anomaly classifier
AUC	area under the curve
BAS	building automation system
C	contextual features
$CCAD$	collective contextual anomaly detection
$CCAD-SW$	collective contextual anomaly detection using sliding window
D	real dataset
D_{train}	real training dataset
EAD	ensemble anomaly detection
EPE	ensemble performance evaluator
err_neg	test output of normal dataset
err_pos	test output of anomalous dataset
err_value	test output of new sensor dataset
EVD	eigenvalue decomposition
FP	false positive
FPR	false positive rate
G	generated features
$HVAC$	heating ventilating and air conditioning
IQR	interquartile range
L	learning model
MSE	mean square error
MT	Model tester
N	real testing dataset
$OETD$	optimal ensemble threshold determinator
Opt_E_{thresh}	optimal ensemble threshold
P	artificial anomalous dataset
$pAUC$	partial area under the curve
$pAUC_c$	standardized partial area under the curve
PCA	principal component analysis
PSO	particle swarm optimization
$Q1$	first quartile
$Q2$	second quartile
$Q3$	third quartile
R	number of learning rounds
RBF	radial basis function
RF	random forest
ROC	receiver operating characteristics
S	a set of unique error values
$S_n - S_1$	difference between first and last sliding window data values
SVD	singular value decomposition
SVM	support vector machines
SVR	support vector regression
SW	sliding window
TN	true negative
TNR	true negative rate
TP	true positive
TPR	true positive rate

One of the problems of existing collective anomaly detection approaches is that there is little concern for the context of the anomaly under consideration. For example, a daily HVAC consumption pattern might be anomalous in winter, but not in summer. An important application of collective contextual anomaly detection is a *building automation system* (BAS), a built-in control system which is nowadays present in most modern buildings (smart buildings [3]). The BAS enables building managers to oversee the energy efficiency aspects of a building by providing early detection and notification of abnormal consumption behaviour. Identifying collective contextual anomalies of a facility at various granularities can be a useful tool for short-term energy saving and disaster mitigation, as well as for meeting long-term energy efficiency targets. For instance, identifying hourly collective contextual anomalies in HVAC consumption can be useful for achieving short-term energy-saving goals. Identifying anomalies in annual HVAC consumption profile is more useful for long-term energy efficiency plans such as replacing energy-wasting equipment, analyzing the cost of services over a long period of time, and planning specific energy-saving targets. For this reason, this research proposes a new pattern-based anomaly classifier, the *collective contextual anomaly detection using sliding window* (CCAD-SW) framework. The CCAD-SW framework uses overlapping sliding windows to improve significantly the anomaly detection performance of the CCAD [4] framework. In addition, it identifies anomalies earlier and substantially reduces false positives. To enhance the anomaly detection capacity of the CCAD-SW, this research also proposes the *ensemble anomaly detection* (EAD) framework. The EAD is a generic framework that combines several anomaly detection classifiers using majority voting. In this study, it is assumed that each anomaly classifier has equal weight. To ensure diversity of anomaly classifiers, the EAD framework is implemented by combining pattern- and prediction-based anomaly classifiers.

In this study, the EAD framework combines the CCAD-SW, which is implemented using autoencoder, with two prediction-based anomaly classifiers that are implemented using the machine learning algorithms support vector regression and random forest. More importantly, the EAD framework identifies an ensemble threshold that provides an anomaly classifier with optimal anomaly detection performance and minimum false positives. Results show that the EAD performs better than the individual anomaly detection classifiers.

The remaining sections of the paper are organized as follows: Section 2 provides the background information and Section 3 describes related work. Sections 4 and 5 outline the CCAD-SW and EAD frameworks proposed in this research. Section 6 presents the experimental results and discussion, and finally Section 7 concludes the paper.

2. Background information

This section first presents an overview of the learning approach used in this study, i.e., the ensemble learning approach. Moreover, it describes the machine learning algorithms used in this research: autoencoder, PCA, support vector regression, and random forest. In addition, the performance metrics used to compare anomaly detection classifiers are described.

2.1. Ensemble learning

Ensemble learning is a machine learning approach that solves a problem by training multiple learners. Unlike ordinary machine learning approaches in which a single hypothesis is learned from training data, ensemble approaches attempt to build a set of hypotheses and combine them to build a new hypothesis [5]. Previous studies show that an ensemble often performs better than the individual learners, also known as *base learners* of the ensemble [6].

Most ensemble approaches rely on a single *base learning algorithm* to produce what are referred to as *homogeneous* base learners. However, some approaches use multiple learning algorithms and are referred to as *heterogeneous* learners [5]. The primary objective of ensemble learning is to improve the performance of a model by combining multiple learners.

Normally, ensembles are constructed in two steps. Initially, several base learners are built, and then these learners are combined. Several combination techniques are used. For anomaly