Complex Adaptive Systems Modeling

RESEARCH Open Access

Complex building's energy system operation patterns analysis using bag of words representation with hierarchical clustering

Usman Habib[1,3*], Khizar Hayat[2] and Gerhard Zucker[1]

*Correspondence:
usmanhabib@ciit.net.pk
[1] Energy Department, AIT Austrian Institute of Technology, Giefinggasse 2, 1210 Vienna, Austria
Full list of author information is available at the end of the article

Abstract

Purpose: Due to the large quantity of data that are recorded in energy efficient buildings, understanding the behavior of various underlying operations has become a complex and challenging task. This paper proposes a method to support analysis of energy systems and validates it using operational data from a cold water chiller. The method automatically detects various operation patterns in the energy system.

Methods: The use of k-means clustering is being proposed to automatically identify the On (operational) cycles of a system operating with a duty cycle. The latter's data is subsequently transformed to symbolic representations by using the symbolic aggregate approximation method. Afterward, the symbols are converted to bag of words representation (BoWR) for hierarchical clustering. A gap statistics method is used to find the best number of clusters in the data. Finally, operation patterns of the energy system are grouped together in each cluster. An adsorption chiller, operating under real life conditions, supplies the reference data for validation.

Results: The proposed method has been compared with dynamic time warping (DTW) method using cophenetic coefficients and it has been shown that the BoWR has produced better results as compared to DTW. The results of BoWR are further investigated and for finding the optimal number of clusters, gap statistics have been used. At the end, interesting patterns of each cluster are discussed in detail.

Conclusion: The main goal of this research work is to provide analysis algorithms that automatically find the various patterns in the energy system of a building using as little configuration or field knowledge as possible. A bag of word representation method with hierarchical clustering has been proposed to assess the performance of a building energy system.

Keywords: Building energy performance, Fault detection and diagnosis (FDD), Clustering, Symbolic aggregate approximation (SAX), Bag of words representation (BoWR), Hierarchical clustering, Coefficient of performance (CoP), Dynamic time warping (DTW), Heating ventilation and air conditioning (HVAC), Gap statistics analysis

Background

This paper is an extension of work originally presented in proceedings of Frontiers of information technology (FIT'15) Conference 2015 (Habib and Zucker 2015). The energy systems of a typical contemporary building are usually complex and may contain several subsystems

Springer Open

deployed independently of each other. In order to analyze various energy performance aspects of a given building, a lot of raw data is recorded during its monitoring Khan et al. (2011). The recorded data is studied at later stages in order to find interesting features, using a variety of visualization tools (Mourad and Bertrand-Krajewski 2002). The massive amount of recorded data makes any detailed performance analysis a formidable task. Moreover, there is a high chance of overlooking some important patterns in the data which, if noticed properly, may help identify faults that can compromise energy efficiency.

Patterns are regular, usually repetitive, sequences in a given data and may owe their existence to a specific event. A pattern is thus dependent on the characteristics of a system and may represent the underlying processes and structure of the system. Methods that can automatically identify interesting patterns from buildings' data, help to get useful insights into the various parameters of energy usage as well as the source of faults in different components. In this context, data mining techniques like clustering are feasible tools to address these issues. The process of automatically finding the various patterns in the data can make the subsequent analysis easier, more feasible and lesser laborious (Miller et al. 2015; Iglesias and Kastner 2013; Narayanaswamy et al. 2014; Lin and Li 2009).

We aim to exploit machine learning for finding various patterns in energy related building data . The idea is to realize all this with minimum possible configuration changes and knowledge of the relevant field. More specifically, in order to automatically find different patterns in the adsorption chiller's operation, in this article, we propose to use a bag of words representation (BoWR) with subsequent hierarchical clustering. The suggested method has been applied to the operation data of a water chiller and compared to another approach called dynamic time warping (DTW) using cophenetic correlation. The dynamic time warping (DTW) method uses a dynamic programming technique for defining the best alignment between the two time series data Keogh and Ratanamahatana (2004). Furthermore, the cophenetic correlation demonstrates that the cluster tree has a strong correlation with the distances between objects in the distance vector Lin and Li (2009). The On/Off state information required for the suggested technique is detected by using the k-means clustering algorithm. As we are taking the sensor readings that are placed outside the chiller, the sensors reading will reflect the behavior of the chiller during its operational cycle. Thus, the On (operational) states are of greater importance for assessing the performance of chillers and faults detection and diagnosis (FDD). Moreover, avoiding the Off cycle for finding different patterns will reduce the amount of data as well. The On (operational) cycles are discretized by using the symbolic aggregate approximation (SAX) method. These discretized values are called symbols or words. After transformation of the On cycles to words, a normalized histogram for each On cycle is created; called bag of words representation (BoWR). The normalized BoWR is used because the On (operational) cycle's vary in there duration. The hierarchical clustering uses the normalized BoWR of the On cycles for finding the various operational patterns of the chiller. The details of the different clusters created by hierarchical clustering are also explained in detail.

The rest of the paper is arranged as follows. The next section discusses the state of the art methods available in the literature. In the subsequent section, the design of the demonstrated system is elaborated. This is followed by a section describing the methodology of the proposed solution for finding the patterns in the data. The penultimate section explains the different experiments and results, followed by a "Conclusion" section.

State of the art

This section discusses state of the art methods, from literature, proposed for finding operation patterns in different energy systems, in the context of buildings. The energy systems can be modeled using simulation tools.

Complex systems (CA)

"Complex systems consist of many interacting components and many hierarchical layers and, in general, it is impossible to reduce the overall behavior of the system to a set of properties characterizing the individual components. The interaction between components is able to produce properties at the collective level that are simply not present at the component considered individually" Avram and Rizescu (2014); Moffat (2010). The focus of complex systems is to study the system modules (subsystems), their interaction with each other, and how each module is contributing to the overall behavior of the system. Examples of complex systems are:

- people creating social systems,
- our nervous system, with brain spinal cord and neurons being the subsystems, and
- a weather forecast system with factors like wind flow, pressure and temperature contributing in predictions.
- cities can be considered as system and different aspect such as social physics, urban economics, transportation theory, regional science, and urban geography can be considered as subsystem (agents) for designing the cities Batty (2007).

The knowledge of complex systems can be used in all traditional disciplines of science, along with engineering and management. The main focus in the complex systems is on questions regarding parts (independent), overall behavior and relationships. The information provided with the complex systems, using sophisticated tools, is helpful to think analytically about these systems in detail and contributes in the modeling or simulation analysis of these systems.

Complex adaptive systems (CAS)

A Complex adaptive system (CAS) can be defined as an open system "with large variability and diversity of elements or agents, with dynamic interactions among them that create non-linear feedback systems" (Faucher 2010). Such systems are usually linked to the learning activities, in order to provide various features of CAS, like self-organization and unpredictability. They are also described as "special cases of complex systems, which can be called as 'complex macroscopic collection' of relatively similar micro-structures that are partially connected. These macro-structures are formed to adapt the changes in the environment, and increase its survivability" (Kayman 2014).

The subsystems of a complex system are generally modeled as agents. The agents are usually goal-oriented, variable in number and, the condition of the environment can be affected by other agents. The three basic properties of the CAS are (Andrii 2014):

1. Adaptation: This characteristic of CAS is relevant to the adaptability of the system to changes in the environment.

2. Self-organizing: This characteristic is dependent on the structure of the system as well as its internal processes; the underlying question being how the larger dynamic system organizes itself in critical situations.
3. Emergence: This characteristic defines the qualitative change in the behavior of the system during a change in its observation scale. It is one of the common characteristic of CAS where the behavior of the system is more complex than the sum of the behaviors of the components of the system. The emergent property is lost when the system is decomposed into its component parts or when an elimination of some component occurs.

In order to study the complex systems, one has to take into account all; be it the components, their interaction or the overall behavior of the system. Still the emergent behavior cannot be discounted. One of the common methods used is to ignore some system details that mean to find a higher abstraction level of the system. In multidimensional scenarios, the space can be reduced by using mapping (generating a few equivalence classes). There are different factors usually ignored by CAS designers that can constrain the system and influence their long term performance, called as energy. The authors in Hadzikadic (2010) discusses the changes and influence due to the variation in available energy. Furthermore, adding the concepts of efficiency and resilience in complex adaptive systems can be beneficial in modeling (Korhonen and Snäkin 2015).

Buildings as CAS

Complex systems scale from large systems like ecosystem (Levin 1998; Grimm et al. 2005) or social ecological systems (Olsson et al. 2004) to smaller systems such as secure authentication systems (Habib et al. 2011) or buildings Oosterhuis (2012) and their energy system (Azar and Menassa 2010, 2011; Jensen et al. 2016). Limited area notwithstanding, the analysis of a building's energy system is a complex task as it consists of several subsystems. In order to make a detailed analysis of the energy systems, the buildings are monitored using sensors. Nowadays, it is feasible to maintain a record of the historic operation data in the building. While there exist other domains that have considerably higher amounts of data, the operation data in buildings are specifically challenging, since there is commonly no appropriate underlying data model that can be generally applied to operation data; data is very specific to one building or component. Thus, data analytics methods have to supply a high degree of unsupervised automation in order to treat different types of data. Thus, today the main approach for data analysis is a simple visualization of the process parameters using time graphs as visualization tools. Such visualizations may further require manual followup performance analysis. Methods of analysis like these can be time intensive and there is always a chance to miss out some areas of interest that may eventually be of greater importance (Mourad and Bertrand-Krajewski 2002).

In order to make the buildings energy efficient, their prototype models are simulated for energy performance. For better designing and the ability to handle the dynamic nature of the building's characteristics, each component of the building can be modeled as an active part; thus different components of the building will constitute a complex network (Oosterhuis 2012). There are many energy modeling methods that are generally used for predicting the buildings performance during the design phase. The actual energy consumption reading usually deviates from the predicted value during the modeling phase (Azar and Menassa 2010, 2011). Some of the reasons for this deviation are the dynamic parameters like occupant's behavior, climate, and buildings properties (Azar and Menassa 2011). The agent based modeling can be used to handle such dynamic parameters. For example, the dynamic nature of occupants' behavior can be correlated with the impact on energy consumption in commercial buildings (Azar and Menassa 2010, 2011) or in managing ventilation system in residential buildings (Jensen et al. 2016). There are several bottom up approaches put forward for the agent based modeling. The authors in Grimm et al. (2005) have proposed a framework using a pattern oriented approach for agent based modeling to handle the complexity and uncertainty problems.

Other methods for analysis of energy systems in buildings

The reasons for analyzing data from the energy subsystems are manifold and include such objectives like assessment of the overall system performance, comparison with other systems, calculation of operating costs, and prediction of energy consumption and faults etc. The International energy agency (IEA) has launched an implementing agreement (i.e., a technology initiative) called "IEA Solar Heating and Cooling Programme (SHC)". Within this implementing agreement IEA SHC Task 38 "Solar Air-Conditioning and Refrigeration" was one of the research topics. The IEA SHC Task 38 (subtask A3a-B3b: "Monitoring procedure for solar cooling system") defines a generic monitoring policy that provides information on sensor locations and naming for the evaluation of systems, evaluation of the system performance, and comparison of different energy systems (Napolitano et al. 2011). In the literature, one can find many methods for faults detection and diagnosis (FDD) in building components. One important area is concerned with the Heating, Ventilation and Air-Conditioning (HVAC) (Pietruschka et al. 2015; Isermann 2005; Fan and Qiao 2011; Katipamula and Brambley 2005; Capozzoli et al. 2015; Katipamula and Brambley 2005; Lee and Eun 2015; Narayanaswamy et al. 2014). Prior knowledge about the system can be useful in finding some of the simple undetected faults using first principles (i.e. energy balance, mass balance and other physical principles), but still there is a requirement for more sophisticated techniques to judge various aspects of a building's energy performance. One known class of techniques that makes use of historic operation data describes the behavior of the system, characterized as black box models, which are fitted using the historical data (Katipamula and Brambley 2005, 2005). Faults can also be detected in buildings with machine learning algorithms using the information from the installed electricity consumption meters as shown in (Figueiredo et al. 2005), Domínguez et al. (2013). There are different parameters available that can be useful for the prediction of electricity consumption for each HVAC component; multivariate analysis can be used to calculate these parameters (Djuric and Novakovic 2012).

In order to detect various patterns in any energy system using data driven techniques, the focus is on extracting information from the recorded data using little to none domain expertise. There are several machine learning techniques that can be used for extracting information from the data, e.g. clustering can be used for finding similar daily performance patterns in the buildings (Miller et al. 2015; Seem 2005), detecting the abnormal performance from electricity consumption Seem (2007), and further enhancing the performance optimization algorithms (Kusiak and Song 2008). Moreover, at a larger scale, wavelet transformations and clustering can be used for the classification of electrical demand profiles of buildings (Florita et al. 2013).

The data is usually stored as a time series for later analysis. The time series data can be represented with different available techniques that can further help in finding the similarity between the data having same behavior. An example is the symbolic aggregate approximation (SAX), a category of Piecewise Aggregate Approximation (PAA), that can be used to improve the speed and usability of several analysis techniques Lin et al. (2007). The similarities between different time series data can even be calculated by simply using the Euclidean distance parameter, but the problem in this method is that even a slighter shift of data can lead to erroneous results (Lin and Li 2009). A comparison of time series data similarity algorithms (Euclidean, DTW, wavelets) is carried out in Lin and Li (2009) par rapport the method of bag of patterns using hierarchical clustering. The authors have concluded that the bag of patterns representation (BoPR) approach performed better for finding similarities in the time series data as compared to other methods. The use of bag of words model can be seen in various fields with classification (Anwar et al. 2015).

One of the well-known methods used for finding similar groups via data mining is clustering (Armano and Javarone 2013; Shah et al. (2015). The decision of the optimal number of clusters is an important issue in unsupervised methods, in general, and in hierarchical clustering, in particular. A clustering algorithm can give better results if the inter-cluster variations are minimum and intra-cluster variations are maximum (Tibshirani et al. 2001). Clustering algorithms can also be used for finding various energy states in the building, e.g., k-means clustering can be used to detect the state (On/Off) of machine, as data toggle between these two states (Habib et al. 2015; Zucker et al. 2015a, b). Another example of using clustering for finding system states can be found in Zucker et al. (2014), where the X-Means clustering algorithm is used for automatically detecting the system states (On/Off), in order to examine the operational data of adsorption.

Cluster evaluation methods

Cluster evaluation is usually carried out using graphical methods. One such way is to plot error measurement against the number of clusters. In this method the position, where the plot creates an "elbow" in the graph, can be taken as the number of clusters, since the "elbow" occurs at the point of sudden decrease in the error measurement (Ketchen and Shook 1996). There are other methods that can be used to find optimal number of clusters in the data, e.g. the Silhouettes criterion method Rousseeuw (1987), Davies-Bouldin's criterion method Davies and Bouldin (1979) and Calinski-Harabasz criterion method Caliński and Harabasz (1974). Other than these techniques, a method from the literature is based on gap statistics analysis wherein the gap criterion finds the optimal number of clusters by estimating the "elbow" location as the number of clusters against the largest gap value (Tibshirani et al. 2001). The gap value can be defined as (Tibshirani et al. 2001)

$$Gap_n(k) = E_n\{log(W_k)\} - log(W_k), \tag{1}$$

where E_n is the expected value, n is the size of the sample, k is the number of clusters that are being evaluated, and W_k is the dispersion measurement within the cluster and can be find as

$$W_K = \sum_{r=1}^{k} \frac{1}{2_{n_r}} D_r, \tag{2}$$

where n_r represents the count of data points in the cluster r, and D_r denotes the sum of the pairwise distances for all data points in the cluster r.

Design of the demonstration system

This section discusses the architecture of the system that has been under observation for applying the proposed method. For this research, the data from selected solar adsorption chillers is used for the period of 16 months from January 2014 until April 2015. The monitoring policy with naming convention of IEA SHC Task 38 for solar and cooling had been followed (Napolitano et al. 2011). The design of the system is shown in Fig. 1, showing three different cycles in the system along with the installed sensors.

The process involves the three main parts which can be summarized as follows:

- The low temperature (LT)cycle is representing the part of the system that is handling the low temperature water produced by the chiller.
- The medium temperature (MT) cycle represents the system portion where the unwanted heat of the system is transferred to the environment using cooling tower.
- The high temperature (HT) cycle is showing the section of the system where heat is provided to produce cold water by the chiller.

The 18 different parameters of interest along with their description are given in Table 1.

In order to find patterns in the operational data (On cycles), different tests were performed in consultation with the experts in the field. There are additional features added

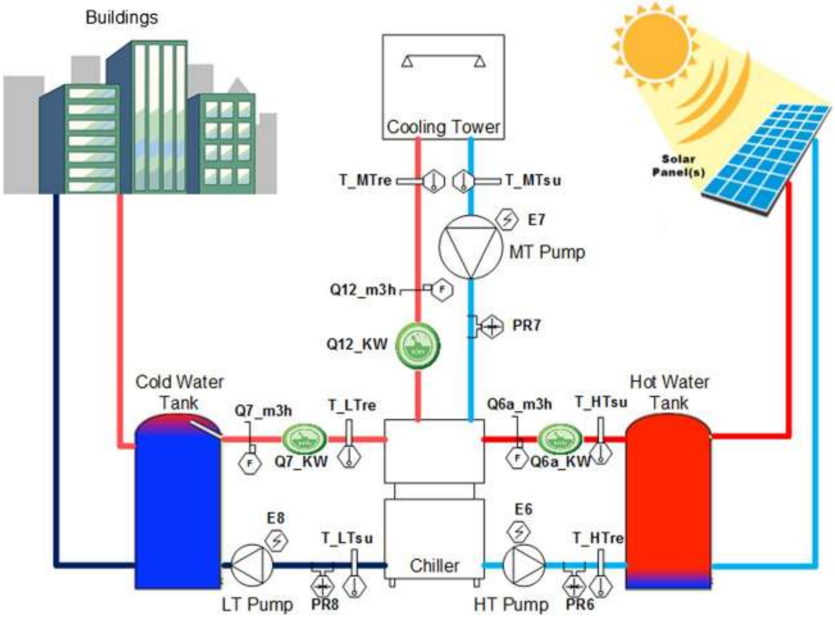

Fig. 1 System architecture with the the required sensors

Table 1 Parameters description

Sensors	Description
E6	High temperature (HT) electricity consumption meter
E7	Medium temperature (MT) electricity consumption meter
E8	Low temperature (LT) electricity consumption meter
Q6a_m3h	HT cycle Flow (water) reading
Q12_m3h	MT cycle Flow (water) reading
Q7_m3h	LT cycle Flow (water) reading
T_HTre	HT cycle temperature on return side
T_HTsu	HT cycle temperature on supply side
T_MTre	MT cycle temperature on return side
T_MTsu	MT cycle temperature on supply side
T_LTre	LT cycle temperature on return side
T_LTsu	LT cycle temperature on supply side
Q6a_KW	HT cycle Energy consumption reading
Q12_KW	MT cycle Energy consumption reading
Q7_KW	LT cycle Energy consumption reading
PR6	Pressure in HT cycle
PR7	Pressure in LT cycle
PR8	Pressure in MT cycle

for better results. The temperature difference between the return and supply temperature sensors of each of the cycle had been used as a feature that are given as,

$$\Delta Temp\_LT = |T\_LTre - T\_LTsu| \tag{3a}$$

$$\Delta Temp\_HT = |T\_HTre - T\_HTsu| \tag{3b}$$

$$\Delta Temp\_MT = |T\_MTre - T\_MTsu| \tag{3c}$$

Therefore, the new set of features are added with other selected parameters for hierarchical clustering in next step. The following Table 2 shows the different features that have been used for the hierarchical clustering.

Table 2 Selected features for hierarchical clustering

Features	Descrition
$\Delta Temp\_LT$	Temperature difference of low temperature cycle
$\Delta Temp\_HT$	Temperature difference of high temperature cycle
$\Delta Temp\_MT$	Temperature difference of medium temperature cycle
Q6a_m3h	Flow in high temperature cycle
Q7_m3h	Flow in low temperature cycle
Q12_m3h	Flow in medium temperature cycle
Q6a_KW	Energy reading in high temperature cycle
Q7_KW	Energy reading in low temperature cycle
Q12_KW	Energy reading in medium temperature cycle

Methods

This section describes the methodology proposed in this paper. The first step followed in the analysis of data is always the preprocessing and finding outliers. The data used has already been processed; therefore it can be used without the preprocessing step.

The three methods used in this work were selected keeping in view their independence from two main factors, viz. configuration information and domain knowledge. The algorithms used in this research paper do not any require domain knowledge or configuration information, as illustrated in the Table 3.

On state (operational) detection using K-means clustering

The distribution of the states for chiller vary a lot in the two states (On/Off), therefore the data can be classified in two clusters. It can be readily observed from Fig. 2 that the mean, minimum and maximum of the two clusters can be observed based on the difference between temperatures, flows, energy readings and pressures. For this purpose, the K-means clustering with two clusters and Euclidean distance setting is used to detect the On and Off state. After the detection of On/Off state, at each point of the time, the consecutive On states are marked as one On cycle. The same procedure is adopted for all consecutive Off states. The sensors are placed at the outer points of the solar cooling system, which means that during the On cycle, the data will be representing the system behavior; otherwise, during the non-operational period it will represent the behavior of the environment. Our interest lies in finding various patterns in the chiller's data, therefore, only On cycles were considered for clustering.

Table 3 Selected algorithms analysis

Methods	Algorithms	Knowledge of the field required	Configuration required
Duty cycle detection	k-means	No	No
Duty cycle representation	BoWR	No	No
Clustering	Hierarchical clustering	No	No

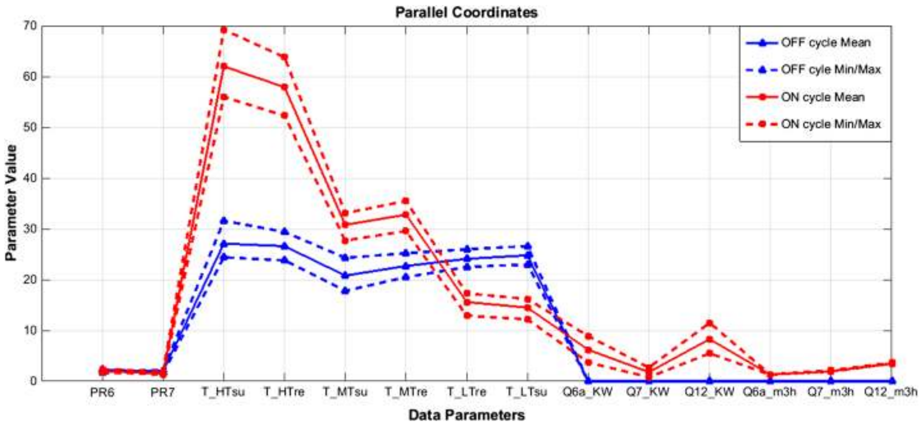

Fig. 2 Parallel coordinates showing the On/Off cycle mean,minimum and maximum patterns

Symbolic aggregate approximation (SAX) transformation

After the detection of On/Off cycles, the data will be in the form defined by Eq. 4:

$$C_i = \{S_1, S_2, S_3, \ldots\ldots., S_N\}, \tag{4}$$

where C_i is the ith cycle and S_t is the sensor value at time tick t. The data is normalized using Z-Scores which is given as:

$$Z(Data) = \frac{S_t - \mu}{\sigma}, \tag{5}$$

where $Z(Data)$ is the Z-score normalized form of the data, S_t is the sensor data at tth time tick, μ represents the mean and σ is the standard deviation. After applying the Z-Score normalization, the data will be now in the form as defined in Eq. 6 while using the On/Off information from the k-means clustering algorithm,

$$Cycle_i = \{Z_1, Z_2, Z_3, \ldots\ldots Z_{N_i}\}, \tag{6}$$

where $Cycle_i$ is the ith cycle of the data and in case if odd count of i is representing Off cycle then even count of i will be presenting the On cycle in data. Z_t is the normalized sensor value while N_i is the event count of $Cycle_i$.

Each cycle data is first broken down into M non overlapping sub-sequences, in a uniform manner, just like the example illustrated in Fig. 3, wherein the partitions are represented by alphabets a, b, c and d. This process is called as chunking, and the period (x-axis) can be of different time length (P) depending on the application where it is used (Miller et al. 2015). The value of P is taken as 5 min in this research. The symbol of each data point is assigned according the breakpoints. The number of break points (M) taken for this research is 60. This transforms the data for each cycle to symbols. The SAX representation is specific for each a length of each cycle. In order to generalize the symbolic representation for each cycle with different lengths, the BoWR is used.

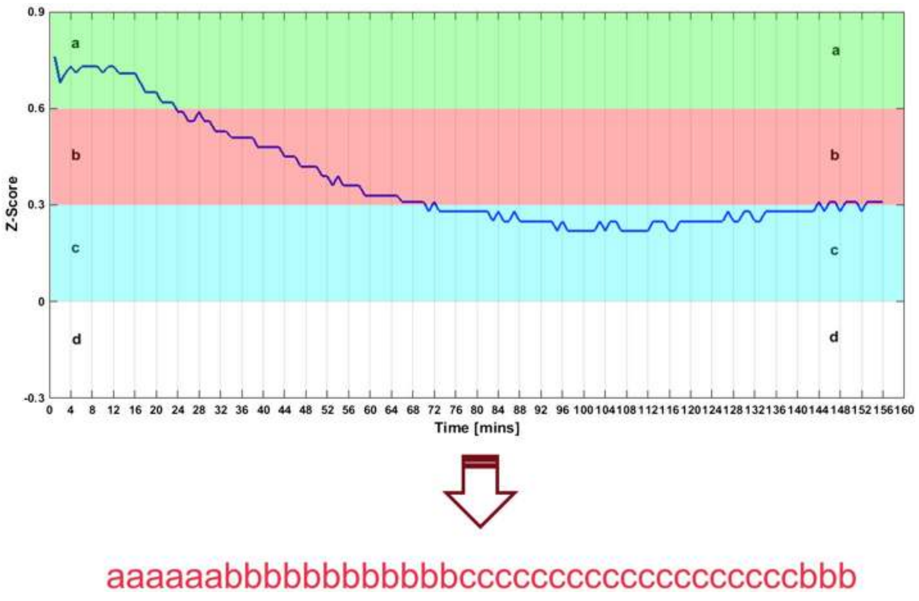

Fig. 3 Example of SAX transformation of On duty with $M = 4$ and $P = 5$

Bag of words representation (BoWR)

In order to represent the complete behavior of a cycle, with different parameters taken under consideration, each sensor's data is converted to a BoWR of 60 characters and put together in a 540 character representation, as can be seen in Fig. 4.

All the required parameters will be converted to z-score before transformation to SAX symbols. The value of M is taken as 60 for this research. $BoWR_i$ pertains to BoWR of the ith On cycle containing all features (shown in Table 2). A pattern can be defined as:

$$\mathrm{BoWR_i} = \{\mathrm{BoWR\_Sensor}^i_1, \mathrm{BoWR\_Sensor}^i_2, ..., \mathrm{BoWR\_Sensor}^i_P\}, \tag{7}$$

The $BoWR\_Sensor^i_P$ consists of a vocabulary set $\{w_1, w_2, w_3, ..., w_M\}$ of sensor P. The associated histogram vector $BoWR\_Sensor^i_P$ for ith On cycle will be like the following:

$$\mathrm{BoWR\_Sensor^i_P} = \left(\mathrm{V}^i_1\ \mathrm{V}^i_2\ \mathrm{V}^i_3 \ldots \mathrm{V}^i_M\right), \tag{8}$$

where P is representing the features (see Table 2) selected for finding out different patterns in the chiller's data. V^i_j is the number of occurrences of w_j in the ith cycle, i.e.

$$\mathrm{V^i_j} = \mathrm{Count}_i(\mathrm{w_j}), \tag{9}$$

where the subscript i in $Count_i$ refers to the ith cycle.

In order to handle cycles of variable time lengths, a better idea is to normalize, i.e. use relative frequencies. With this in view, Eq. 9 can be modified as Eq. 10 below:

$$\mathrm{V}^i_j = \frac{\mathrm{Count}_i(\mathrm{w_j})}{\mathrm{N_i}}, \tag{10}$$

where N_i is representing the number of time ticks in the ith cycle.

Hierarchical clustering

The hierarchical clustering technique groups the data over different scales by creating a cluster tree called dendrogram (Vesanto and Alhoniemi 2000). A dendogram shows a multilevel hierarchy of clusters, where the clusters (groups) at one level are joined together to constitute a cluster for the next level. This property of hierarchical clustering allows to decide the level of clustering that is the most appropriate for the task it is used for. The BoWR for each cycle is clustered using the hierarchical clustering technique. Figure 5 shows the dendogram of the $BoWR_i$ given as input to the hierarchical clustering.

There are different techniques available to decide the best level or number of clusters for hierarchical clustering. One such technique is the gap method Tibshirani et al.

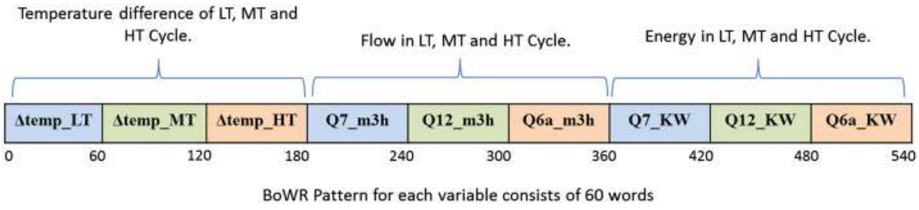
Temperature difference of LT, MT and HT Cycle.
Flow in LT, MT and HT Cycle.
Energy in LT, MT and HT Cycle.
Δtemp_LT
Δtemp_MT
Δtemp_HT
Q7_m3h
Q12_m3h
Q6a_m3h
Q7_KW
Q12_KW
Q6a_KW
0
60
120
180
240
300
360
420
480
540
BoWR Pattern for each variable consists of 60 words

Fig. 4 540 character long representation of Cycle

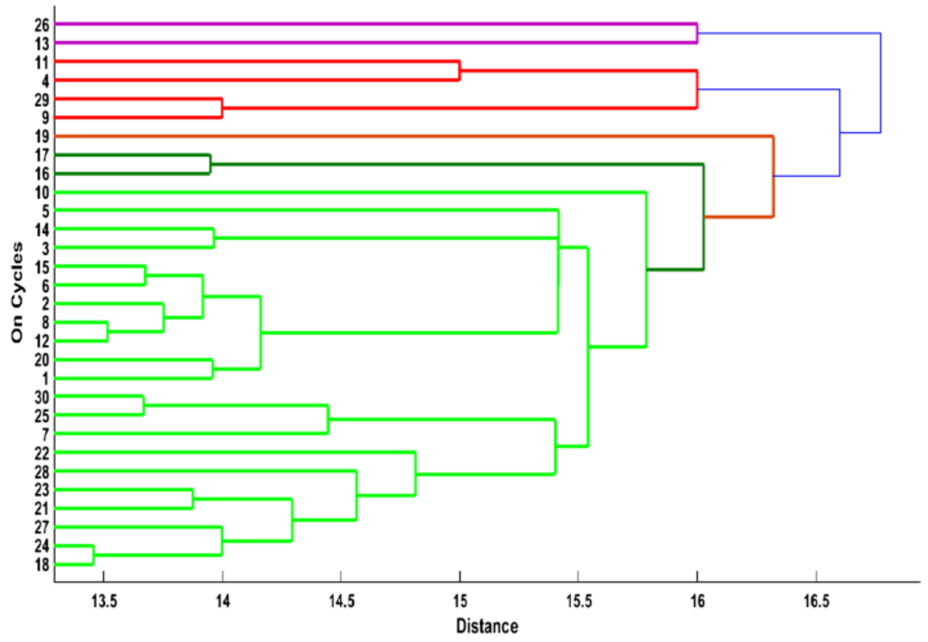

Fig. 5 Dendogram of BoWR

(2001). A clustering algorithm gives better results when the intra-cluster difference is as small as possible while the inter-cluster difference is as high as possible.

Methodology overview

The steps involved, in the proposed method, are illustrated in Fig. 6. Below is a brief stepwise description of the method:

- The first step is to find the On (operational) cycles in the data by using the k-means algorithm. The latter can be applied to any energy system because the two states are their in any energy dependent system and On duty cycle can be readily detectable.
- The On cycles data are transformed to symbolic data with the SAX transformation method. This step also does not need any field knowledge and is applicable to almost all energy systems.

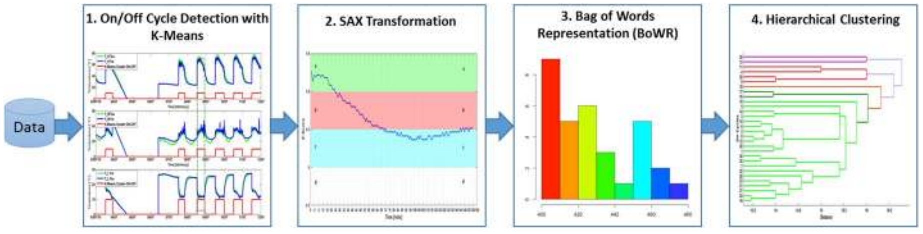
Data
1. On/Off Cycle Detection with K-Means
2. SAX Transformation
3. Bag of Words Representation (BoWR)
4. Hierarchical Clustering

Fig. 6 Methodology

- A BoWR was created for the symbols of each On cycle. This procedure does not need any field knowledge.
- The BoWR are clustered by using the hierarchical clustering for finding various operation patterns of the chiller. This process does not need any field knowledge.
- The gap statistics is used to find the optimal number of clusters in the data. This procedure does not need any field knowledge.
- The cluster patterns can be further investigated using the average performance indicators of each cluster.

Experiments and results

The experiments had been performed on a data from water based chillers. Only On cycles were considered for clustering, as these are more appropriate for finding faults in the chiller. The hierarchical clustering was applied with dynamic time warping (DTW) and the proposed BoWR method. The comparison of the hierarchical clustering performance of the two methods was carried out with the help of cophenetic coefficients (Saraçli et al. 2013). The cophenetic correlation is technique that demonstrates the cluster tree strong correlation with the distances between objects in the distance vector. Table 4 lists the cophenetic coefficients with different hierarchical clustering methods Levin (2007) using the BoWR and DTW techniques. The BoWR has strong correlation with distance with all other objects in all the clustering methods. The best results for BoWR are attained with the *Average* method for hierarchical clustering.

The first step performed for the BoWR, was to find the On cycles automatically. The results of the k-means clustering algorithm can be seen in Fig. 7. The last graph in Fig. 7 shows the On/Off cycles of the chiller. It can be observed from the behavior of the temperatures at the low temperature (LT), medium temperature (MT) and high temperature (HT) cycle are responding according to the detected On/Off state. It is clear from Fig. 7 that during the detected On cycle, the LT temperature decreases showing the cooling operation. At the same time, increase in the temperatures at HT and MT cycle of the chiller can be noticed. These simultaneous variation in temperature gives a clear signal that the chiller is in operational mode, which has also been detected by the proposed method of k-means clustering.

The selected features of the detected On cycles were converted to BoWR. The hierarchical clustering makes a clustering tree (dendogram) that gives the option to select the

Table 4 Cophenetic coefficients of dynamic time warping (DTW) and BoWR

No.	Clustering methods	Bag of word representation (BoWR)	Dynamic time warping (DTW)
1	Average	0.9897	0.0375
2	Centroid	0.9851	0.037
3	Complete	0.9753	0.035
4	Median	0.9803	0.0363
5	Single	0.9848	0.0414
6	Ward	0.9835	0.0363
7	Weighted	0.9888	0.0368

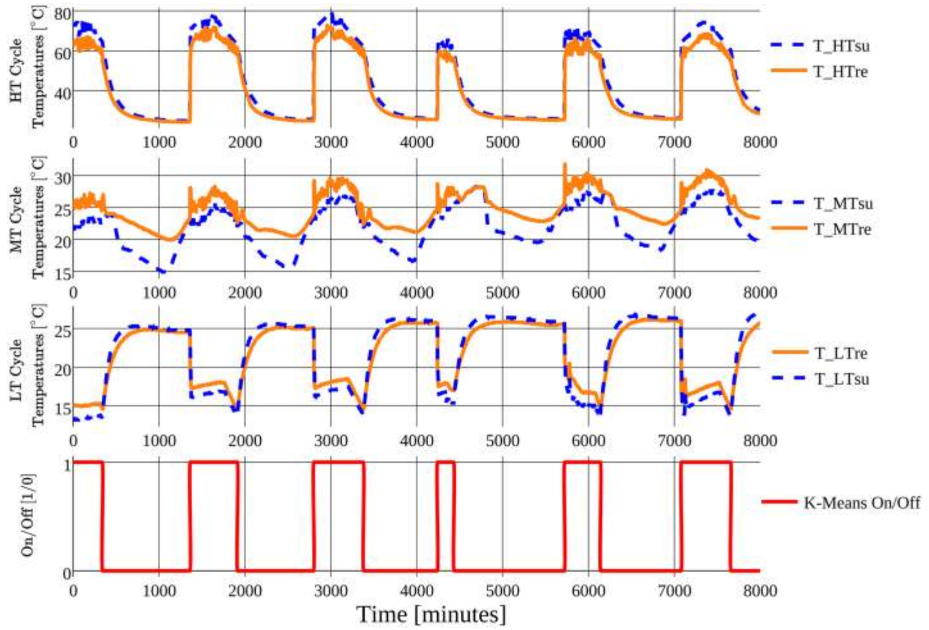

Fig. 7 Demonstration of On/Off cycle detection with different temperature sensors using k-means clustering

level (cutoff) for clustering. The gap statistics Tibshirani et al. (2001) had been used to find the optimal number of clusters, depending on the gap between different clusters. As it can be observed from Fig. 8, the gap statistics analysis gives the best gap distance with five clusters.

The cluster information of the five clusters are given in Table 5. The interesting pattern group is $Cluster_1$ and $Cluster_2$, as the average operation time in these clusters is greater than around 1 hour. For finding faults, $Cluster_1$ patterns are more suitable since the average Coefficient of Performance (CoP) of cycles in this cluster is 0.16, in comparison to the average operational time of cycles that is around 68 hours, thus showing that the chiller's performance is bad. A majority (98.75 %) of the On cycles lies in $Cluster_1$. The

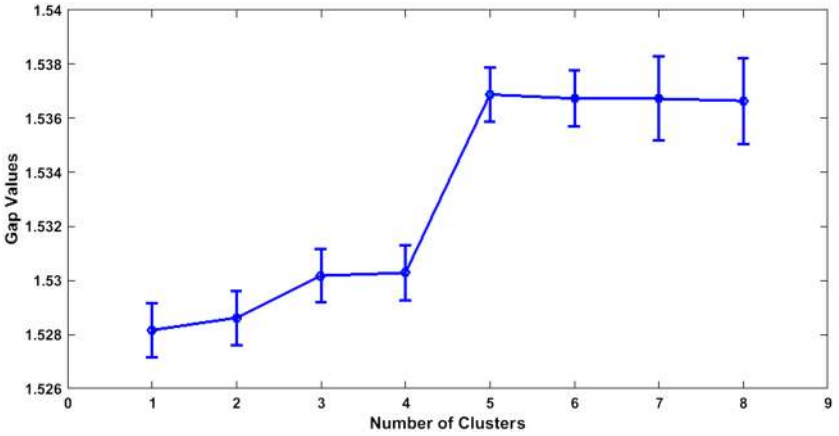

Fig. 8 Gap distance of the BoWR data

Table 5 Cluster information of the five clusters with hierarchical clustering

Cluster_no	Percentage of cycles in cluster (%)	Average CoP of on cycles in cluster	Average time of on cyclesin cluster (hours)
$Cluster_1$	0.73	0.16	67.65
$Cluster_2$	98.75	0.54	0.95
$Cluster_3$	0.06	0.62	0.09
$Cluster_4$	0.34	0.87	0.07
$Cluster_5$	0.12	0.7	0.06

latter represents the cycles with normal operational behavior, since its average CoP is 0.54 while the average operational time of the cycles in this cluster is around one hour. $Cluster_3$, $Cluster_4$ and $Cluster_5$ are representing the cycles with shorter operational time, as the machine is in transient phase; thus the patterns in these clusters are different from a normal operational behavior of the chiller and are thus not plausible.

For further investigation of the cycle behaviors in $Cluster_1$, the graph in Fig. 9 had been drawn in order to show the behavior pattern of one of the On cycles in $cluster_1$. The three graphs in Fig. 9 display the temperature difference ($\Delta Temp\_LT$, $\Delta Temp\_HT$, $\Delta Temp\_MT$),

flows ($Q7\_m3h$, $Q6a\_m3h$, $Q12\_m3h$) and energy meter readings ($Q7\_KW$, $Q6a\_KW$, $Q12\_KW$) in the low, high and medium temperature cycles, respectively. The x-axis displays the time (in minutes) for the On cycle. In each graph, the values are represented with the intensity of the color given in the form of a vertical color code bar at the right hand side of each plot. It can be observed form Fig. 9 that at 30 min, the $\Delta Temp\_MT$ becomes zero, showing that the cooling tower is not operating normally, thus causing no change in the temperature of MT cycle. It is also important to note that the flow variable ($Q12\_m3h$) for MT cycle is showing flow throughout the cycle. Due to this effect, the $\Delta Temp\_LT$ has also started decreasing and at around 80 minute, the cooling has been stopped by the chiller. At the same time, the deriving heat ($\Delta Temp\_HT$) has been provided to the chiller but the chiller is not able to match the cooling load, thus showing

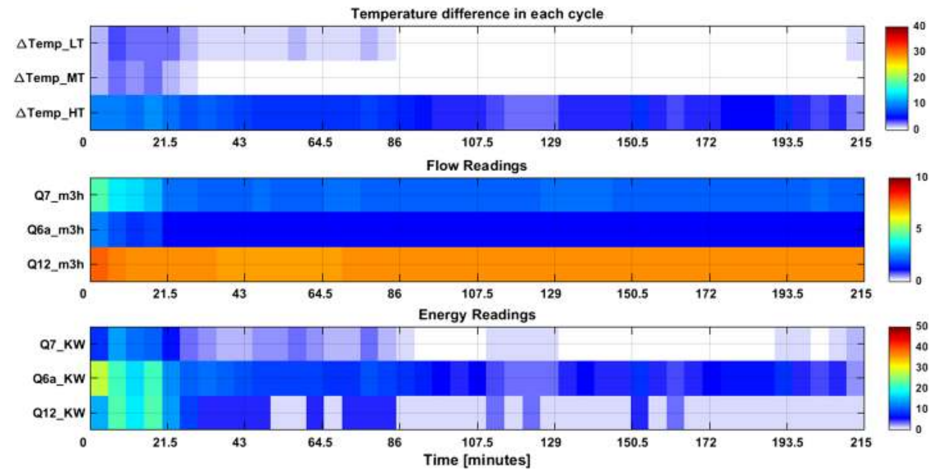

Fig. 9 On cycle pattern showing low performance (faulty) operation of chiller

lower coefficient of performance. This pattern of the chiller is giving a clue about the faults in the chiller that need to be diagnosed. Furthermore, to support the argument that $Cluster_1$ is representing group of On cycles having bad performance, the histogram of CoP of On cycles grouped in $Cluster_1$ are shown in Fig. 10. In order to find the CoP, the following equation is used (Napolitano et al. 2011).

$$\text{CoP} = \frac{\text{Q7\_KW}}{\text{Q6a\_KW}}. \tag{11}$$

The histogram shows that 80 % of the On cycles have CoP less than 0.2, thus representing low performance of the chiller.

The same procedure had been adopted to see the behavior pattern of one of the On cycles in $cluster_2$, as can be seen in Fig. 11. The x-axis displays the time, in minutes, for the On cycle.

In each graph, the values are represented with the intensity of the color given in the attached color bar. It can be observed from Fig. 11 that the duration of the On cycle is 195 minutes. The temperature difference parameters show that there is cooling in the LT

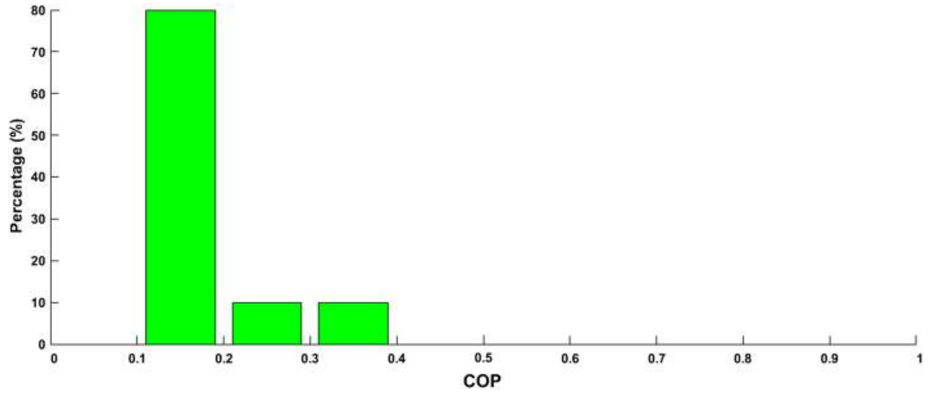

Fig. 10 Histogram of On cycle's associated with cluster of faulty operation of chiller

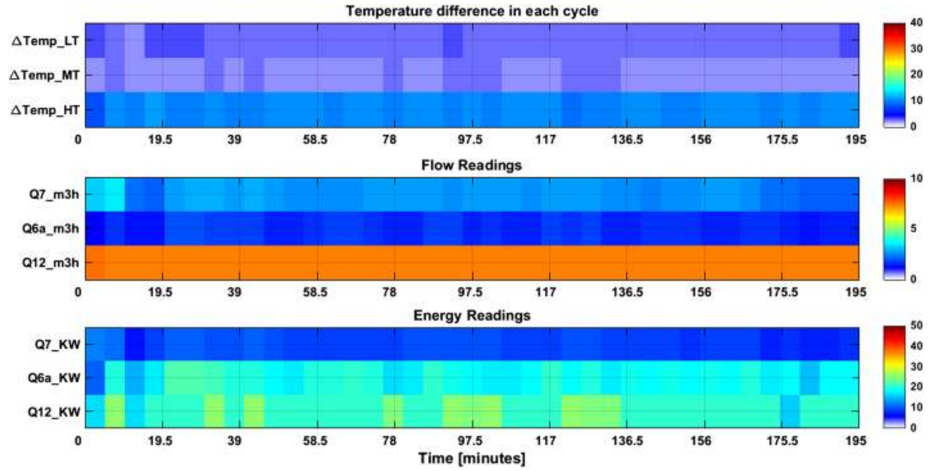

Fig. 11 On cycle pattern showing normal operation of chiller

cycle, as $\Delta Temp\_LT$ is representing it with time. The effect can be seen in the MT cycle as well. At the same time, HT cycle shows that the constant driving heat was provided to the chiller. It is also important to mention that the flow variable displays the flow in all the three cycle, whereas, same had been observed for the energy parameters. This behavior pattern shows the normal operation of the chiller. The histogram of CoP of On cycles, grouped in $Cluster_2$, is shown in Fig. 12 in support of the argument that $Cycle_2$ is representing the group of On cycles corresponding to the normal performance of the chiller. The histogram shows that the chiller is performing with CoP between 0.4 to 0.8, thus representing the normal behavior of the chiller.

Comparison of proposed method with CAS modeling

The main point in this research is to find various patterns in the operation of the energy system in buildings using minimum possible input from the engineers. For the analysis of the energy system, the data has been selected using IEA SHC Task 38. The issues that may surface, while modeling a current system using CAS, can be traced back to the complete knowledge of the system, its behaviors or states and the interaction of the subsystems; a problem of scale dealing with a very large state-space representation. Secondly, complex dynamic systems will require transitions between completely different behaviors in the form of what is called phase transitions. Hence a critical transition detection will require a detailed state-space model.

Conclusions

The main goal of this research work is to provide analysis algorithms that automatically find the various patterns in the energy system of a building using as little configuration or field knowledge as possible. A bag of word representation method with hierarchical clustering has been proposed to assess the performance of a building energy system. In the first phase, a k-means clustering algorithm is used to find the On (operational) cycles of the chiller. These On cycles are represented with symbols by using symbolic aggregate approximation (SAX) method. Furthermore, the symbolic representation is transformed to BoWR, which is provided to hierarchical clustering. The proposed method has been compared with dynamic time warping (DTW) method using cophenetic coefficients and it has been shown that the BoWR has produced better results as compared to DTW. The

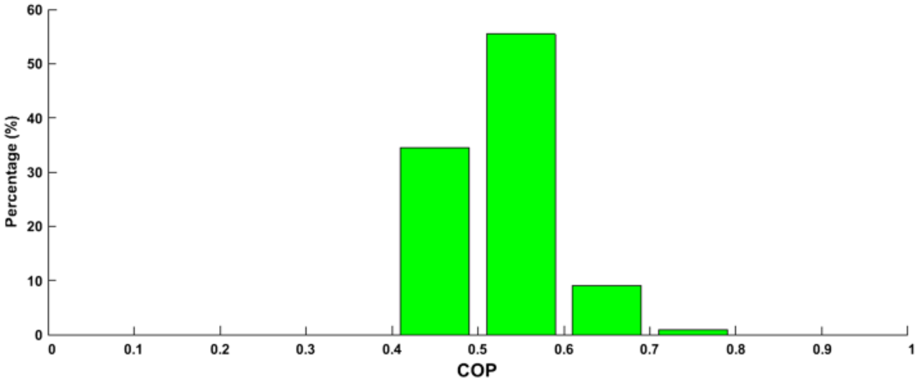

Fig. 12 Histogram of On cycle's associated with cluster of normal operation of chiller

results of BoWR are further investigated and for finding the optimal number of clusters, gap statistics have been used. At the end, interesting patterns of each cluster are discussed in detail.

In future, the current research can be used in the field of automatic faults detection and diagnostics (FDD) in buildings, as the current research helps in finding the different performance patterns. This would help the experts in the field to look only for those areas where the performance is bad. Further research is needed in order to find intelligent ways of diagnosing the faults

Authors' contributions

UH, KH and GZ conceived and designed the experiments. The experiments are performed by UH. The data has been analyzed by UH, KH and GZ. The paper is written by UH, KH and GZ. All authors read and approved the final manuscript.

Author details

[1] Energy Department, AIT Austrian Institute of Technology, Giefinggasse 2, 1210 Vienna, Austria. [2] College of Arts and Sciences, University of Nizwa, Nizwa, Sultanate of Oman. [3] Computer Science Department, COMSATS Institute of Information Technology, Abbottabad, Pakistan.

Acknowledgements

This work was partly funded by the Austrian Funding Agency in the funding programme e!MISSION within the project "extrACT", Project Number 838688.

Competing interests

The authors declare that they have no competing interests.

Received: 1 February 2016 Accepted: 13 May 2016
Published online: 13 June 2016

References

Andrii C (2014) Exploring behavioral patterns in complex adaptive systems. PhD thesis, University of Pittsburgh, Pennsylvani

Anwar H, Zambanini S, Kampel M (2015) Efficient scale and rotation invariant encoding of visual words for image classification. IEEE Signal Process Lett 22(10):1762–1765

Armano G, Javarone MA (2013) Clustering datasets by complex networks analysis. Complex Adapt Syst Model 1(1):5

Avram V, Rizescu D (2014) Measuring external complexity of complex adaptive systems using onicescu's informational energy. Mediterr J Soc Sci 5(22):407

Azar E, Menassa CC (2011) Agent-based modeling of occupants and their impact on energy use in commercial buildings. J Comp Civ Eng 26(4):506–518

Azar E, Menassa C (2010) A conceptual framework to energy estimation in buildings using agent based modeling. In: Proceedings of the 2010 winter simulation conference (WSC), pp 3145–3156

Batty M (2007) Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals. The MIT press, Massachusetts

Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat 3(1):1–27

Capozzoli A, Lauro F, Khan I (2015) Fault detection analysis using data mining techniques for a cluster of smart office buildings. Expert Syst Appl 42(9):4324–4338

Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1(2):224–227

Djuric N, Novakovic V (2012) Identifying important variables of energy use in low energy office building by using multivariate analysis. Energy Build 45:91–98

Domínguez M, Fuertes JJ, Alonso S, Prada MA, Morán A, Barrientos P (2013) Power monitoring system for university buildings: architecture and advanced analysis tools. Energy Build 59:152–160

Fan W, Qiao P (2011) Vibration-based damage identification methods: a review and comparative study. Struct Health Monit 10(1):83–111

Faucher JB (2010) Reconceptualizing knowledge management: knowledge, social energy, and emergent leadership in social complex adaptive systems. PhD thesis, University of Otago, Dunedin

Figueiredo V, Rodrigues F, Vale Z, Gouveia JB (2005) An electric energy consumer characterization framework based on data mining techniques. IEEE Trans Power Syst 20(2):596–602

Florita AR, Brackney LJ, Otanicar TP, Robertson J (2013) Classification of commercial building electrical demand profiles for energy storage applications. J Solar Energy Eng 135(3):031020–031020

Grimm V, Revilla E, Berger U, Jeltsch F, Mooij WM, Railsback SF, Thulke H-H, Weiner J, Wiegand T, DeAngelis DL (2005) Pattern-oriented modeling of agent-based complex systems: lessons from ecology. Science 310(5750):987–991

Habib U, Jørstad I, Thanh DV, Khan IA (2011) A framework for secure linux based authentication in enterprises via mobile phone. J Basic Appl Sci Res 1(12):3058–3066

Habib U, Zucker G (2015) Finding the different patterns in buildings data using bag of words representation with clustering. In: 2015 13th International conference on Frontiers of information technology, pp 303–308

Habib U, Zucker G, Blochle M, Judex F, Haase J (2015) Outliers detection method using clustering in buildings data. In: Industrial electronics society, IECON 2015—41st Annual Conference of the IEEE, pp 000694–000700

Hadzikadic M (2010) Energy in the context of complex adaptive systems: Predator-prey dynamics. In: LAWDN-Latin-American workshop on dynamic networks, p 1

Iglesias F, Kastner W (2013) Analysis of similarity measures in times series clustering for the discovery of building energy patterns. Energies 6(2):579–597

Isermann R (2005) Model-based fault-detection and diagnosis—status and applications. Ann Rev Control 29(1):71–85

Jensen T, Holtz G, Baedeker C, Chappin ÉJ (2016) Energy-efficiency impacts of an air-quality feedback device in residential buildings: an agent-based modeling assessment. Energ Build 19(1):4

Katipamula S, Brambley MR (2005) Review article: methods for fault detection, diagnostics, and prognostics for building systems—a review. HVAC&R Res 11(1):3–25

Katipamula S, Brambley MR (2005) Review article: methods for fault detection, diagnostics, and prognostics for building systems—a review. HVAC&R Res 11(2):169–187

Kayman EA (2014) Chaos in education as an intelligent complex adaptive system. Chaos and complexity theory in world politics 280

Keogh E, Ratanamahatana CA (2004) Exact indexing of dynamic time warping. Knowl Inf Syst 7(3):358–386

Ketchen DJ, Shook CL (1996) The application of cluster analysis in strategic management research: an analysis and critique. Strateg Manag J 17(6):441–458

Khan A, Hornbæk K (2011) Big data from the built environment. Proceedings of the 2Nd International Workshop on Research in The Large, LARGE '11ACM, New York, pp 29–32

Korhonen J, Snäkin J-P (2015) Quantifying the relationship of resilience and eco-efficiency in complex adaptive energy systems. Ecol Econom 120:83–92

Kusiak A, Song Z (2008) Clustering-based performance optimization of the boiler-turbine system. IEEE Trans Energ Convers 23(2):651–658

Lee ET, Eun HC (2015) Damage identification through the comparison with pseudo-baseline data at damaged state. Eng Comp 40:1–8

Levin SA (1998) Ecosystems and the biosphere as complex adaptive systems. Ecosystems 1(5):431–436

Levin MS (2007) Towards hierarchical clustering (Extended Abstract). In: Diekert V, Volkov MV, Voronkov A (ed) Computer Science—theory and applications: proceedings of second international symposium on computer science in Russia, CSR 2007, Ekaterinburg, pp 205–215

Lin J, Keogh E, Wei L, Lonardi S (2007) Experiencing SAX: a novel symbolic representation of time series. Data Min Knowl Discov 15(2):107–144

Lin J, Li Y (2009) Finding structural similarity in time series data using bag-of-patterns representation. In: Winslett M (ed) Scientific and statistical database management, vol 5566, Lecture notes in computer science. Springer, Berlin, pp 461–477

Miller C, Nagy Z, Schlueter A (2015) Automated daily pattern filtering of measured building performance data. Autom Constr 49:1–17

Moffat J (2010) Complexity theory and network centric warfare. DIANE Publishing, Pennsylvania

Mourad M, Bertrand-Krajewski JL (2002) A method for automatic validation of long time series of data in urban hydrology. Water Sci Technol 45(4–5):263–270

Napolitano A, Sparber W, Thür A, Finocchiaro P, Nocke B (2011) Monitoring procedure for solar cooling systems. Technical Report IEA Task 38, international energy agency

Narayanaswamy B, Balaji B, Gupta R, Agarwal Y (2014) Data driven investigation of faults in HVAC systems with model, cluster and compare (MCC). In: Proceedings of the 1st ACM conference on embedded systems for energy-efficient buildings. ACM, New York, pp 50–59

Narayanaswamy B, Balaji B, Gupta R, Agarwal Y (2014) Data driven investigation of faults in HVAC systems with model, cluster and compare (MCC). Proceedings of the 1st ACM conference on embedded systems for energy-efficient buildings, BuildSys '14ACM, New York, pp 50–59

Olsson P, Folke C, Berkes F (2004) Adaptive comanagement for building resilience in social-ecological systems. Environ Manag 34(1):75–90

Oosterhuis K (2012) Simply complex, toward a new kind of building. Front Arch Res 1(4):411–420

Pietruschka D, Dalibard A, Ben I, Focke H, Judex F, Preisler Helm M, Ohnewein P, Frein A, Muscherá M (2015) Report for self-detection on monitoring procedure. Technical Report IEA Task 48/B6, international energy agency

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comp Appl Math 20:53–65

Saraçli S, Doǧan N, Doǧan İ (2013) Comparison of hierarchical cluster analysis methods by cophenetic correlation. J Inequal Appl 2013(1):1–8

Seem JE (2005) Pattern recognition algorithm for determining days of the week with similar energy consumption profiles. Energy Build 37(2):127–139

Seem JE (2007) Using intelligent data analysis to detect abnormal energy consumption in buildings. Energy Build 39(1):52–58

Shah MA, Abbas G, Dogar AB, Halim Z (2015) Scaling hierarchical clustering and energy aware routing for sensor networks. Complex Adapt Syst Model 3(1):5

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc 63(2):411–423

Vesanto J, Alhoniemi E (2000) Clustering of the self-organizing map. IEEE Trans Neural Netw 11(3):586–600

Zucker G, Habib U, Blöchle M, Judex F, Leber T (2015) Sanitation and analysis of operation data in energy systems. Energies 8(11):12776–12794

Zucker G, Habib U, Blöchle M, Wendt A, Schaat S, Siafara LC (2015) Building energy management and data analytics. In: 2015 international symposium on smart electric distribution systems and technologies (EDST), pp 462–467

Zucker G, Malinao J, Habib U, Leber T, Preisler A, Judex F (2014) Improving energy efficiency of buildings using data mining technologies. In: 2014 IEEE 23rd international symposium on industrial electronics (ISIE), pp 2664–2669