BS 2019 2-4 SEPT ROME 16th IBPSA INTERNATIONAL CONFERENCE AND EXHIBITION

INTERNATIONAL BUILDING PERFORMANCE SIMULATION ASSOCIATION

Insight Into Predictive models: On The Joint Use Of Clustering And Classification By Association (CBA) On Building Time Series

Paul Westermann, Joel Grieco, Johanna Braun, Eamon Murphy, Ralph Evins[1]
[1]Energy Systems and Sustainable Cities group,
Department of Civil Engineering, University of Victoria, Canada

Abstract

Data-driven, black box machine learning models have received a lot of attention in the field of building control. They have been used successfully to predict building behaviour given information like weather forecasts and real time sensor information. In these models, the occupant behaviour is considered to act exogenously on the building.
We consider the users as active elements of the building operation control loop. To make educated control decisions they have to be informed about how the building will behave. Therefore, we propose a prediction model which explains to occupants the day-ahead building behaviour using a clustering and classification by association model. We benchmark this approach to a neural network regression model and only observed a small loss of accuracy.
Knowing the upcoming building behaviour, occupants can adjust their behaviour (e.g. putting on clothes) or the building systems settings (e.g. set points) accordingly. The proposed method is a promising way to decode complex regression models into readable rules, which in future may be useful in conjunction with for example voice-based virtual assistants.

Introduction

Buildings are a major energy consumer accounting for 36% of final energy and 55% of final electricity consumption worldwide (IEA, 2017). 80 to 90% of that energy is attributed to building operation (Ramesh et al., 2010). Therefore, optimizing building operation through effective energy management is a strong element of current research on sustainable buildings (Shaikh et al., 2014).
Building occupants have a major impact and partially explain why high performing building technologies (e.g. efficient HVAC systems) do not guarantee low energy use (Andersen et al., 2009). In a simulation-based study on office buildings, a difference in energy use of up to 50% is found if the worker is proactive in energy savings or not (Lin and Hong, 2013). Behavioural differences are found in their adaptive actions (e.g. opening/closing of windows, adjusting set-points) or non-adaptive actions (operation of office equipment, movement through space, etc.) (Hong et al., 2017). This shows that engaging occupants in the energy efficient control of the building will be crucial to achieving energy use targets.
Researchers have developed tools which incorporate occupancy data as input into *supervisory* building control algorithms. Supervisory control logic is implemented at a higher level than the individual controllers of the building systems. Two approaches are prevailing in research: rule-based, and model-predictive control. While rule-based control uses rules defined by HVAC specialists, MPC conducts an operational optimisation over a specified prediction horizon. In both approaches temperature set-points for the whole building are adjusted, or HVAC systems activated taking occupant actions (adapative or non-adaptive) into account. The occupant behaviour is either hard-coded in schedules or detected based on data (Lu et al., 2010). Detection of occupancy patterns (e.g. sleeping, or absent) is a key element of smart thermostat technologies which already exist.[1]
A characteristic of rule-based and model-predictive control is that they *monitor* human behaviour instead of involving occupants as sensing and active element in the control loop (*direct* human-in-the-loop control, HIL). Recent publications envision an interplay of occupants and automated controls where comfort conditions are traded-off with minimizing energy use (D'Oca et al., 2018). This negotiation of comfort conditions demands not only machines to learn occupancy patterns, but also occupants to understand the computer controlling the building.
This study contributes by providing a forecasting method which features a human-readable set of information to explain the expected building behaviour given the computer-based controls already existing in the building. We use a combination of clustering and associate rule mining. Cluster analysis enables to find typical 24-hour temperature profiles and

[1]See for example: `https://nest.com/thermostats/nest-learning-thermostat/overview/`

https://doi.org/10.26868/25222708.2019.211236

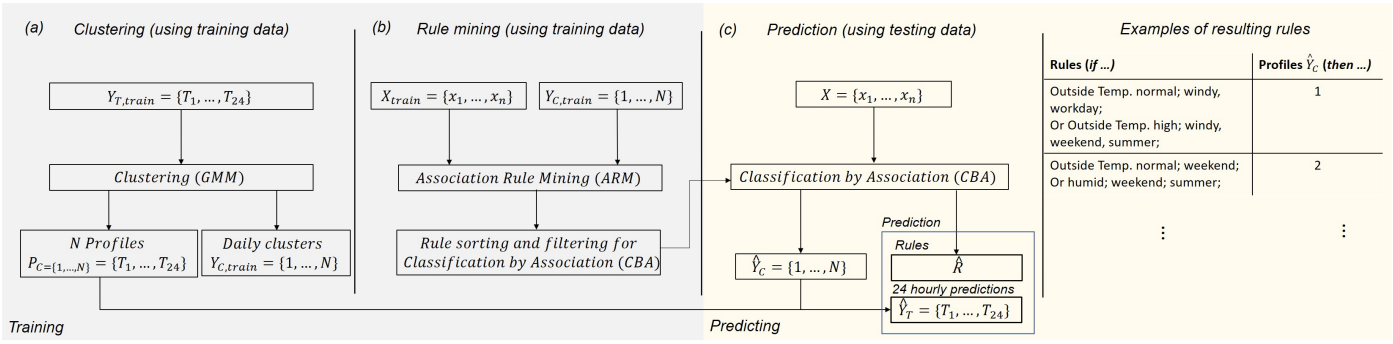

Figure 1: Overview of the proposed approach.

associate rule mining allows to assign a set of rules connecting each of the profiles to weather conditions and occupancy. Based on those rules we select a 24-h profile for the upcoming day with the classification by association (CBA) algorithm. As a result, the occupant has access to numerical building behaviour predictions which are explained by human-readable association rules in the form of "*the predicted profile is x because y*" .

The combination of clustering and association rule mining has been leveraged on building time series data before. Mirebrahim et al. (2017) and Xiao and Fan (2014) used it to receive insight on the control of heating, ventilation and air conditioning (HVAC) systems. Both cases exemplify the strength of the approach for analytical purposes, however it has never been used for forecasting of building time series.

We showcase the use of the method in a study where we derive 24-h indoor temperature forecasts for the upcoming day. Indoor temperature was chosen as it inherently captures the trade-off between occupant comfort and energy demand.

The use of a set of temperature profiles and of categorical features (e.g. binned outdoor air temperature) instead of continuous ones for rule-based prediction limits model complexity. We benchmarked our approach against a 24h prediction of a deep multiple output feed-forward neural network.

In this paper we familiarize the reader with the applied method and provide details on the clustering algorithm used (Gaussian Mixture Modelling), associate rule mining and the classification by association algorithm. Then, the performance and limits of the approach are shown in a case study on indoor temperature prediction in an office building.

Methodology

The proposed approach combines clustering (Fig. 1, a) and rule-mining (Fig. 1, b) to give insightful time series predictions which provide numerical forecasts as well as explanatory rules causing that forecast (Fig. 1, c). The method can be applied to any time series data which is formatted as daily sets of 24 hourly values. In the case study below we focussed on indoor temperature forecasting only, hence the model outputs ($\hat{Y}_T$) are labelled T.

The methodology consists of two steps to train the model:

1. Derive N typical daily profiles using a Gaussian Mixture Model (GMM). The number of profiles has to be chosen by the modeller and is treated as a hyperparameter to be optimized in a grid search (see Table 2).
   Clustering converts hourly output values $Y_{T,train}$ to daily ones $Y_{C,train}$ which contain the derived cluster numbers for each day of the training data.

2. The prediction model, a CBA model, uses association rules for the cluster number $Y_{C,train}$ given features X_{train}. In our case study, the n number of features include daily mean weather forecast data, date-time information (incl. holidays), occupancy data and the cluster of the previous day.
   To derive the CBA model, we first generate association rules between X_{train} and $Y_{C,train}$ using the Apriori algorithm (Agrawal et al., 1994). Then the number of rules is reduced to a small set which only includes those rules with the highest confidence. The high confidence rules form the CBA model.

After this model training process is terminated, the CBA model can be used to predict hourly indoor temperatures for the upcoming day given a new set of unseen features X (Fig. 1, c). It uses the profiles (cluster centroids) and rules determined on the training data. Note that, in the following prediction performance is quantified solely by comparing predicted hourly values, $\hat{Y}_{T,test}$, to observed hourly values, $Y_{T,test}$. We fully neglect whether clusters are predicted correctly.

In the sections below we provide more details on the two steps to derive the prediction model.

Clustering (Gaussian Mixture Model)

The GMM is suitable for clustering problems. It has been applied to time series data before (Eirola and Lendasse, 2013) and specifically on building re-

lated time series data (Melzi et al., 2017) (Mirebrahim et al., 2017). It is a classification algorithm which describes a cluster by its mean and covariance. Both are composed of a mixture of Gaussian distributions. This allows it to identify inhomogeneous, multimodal clusters as required for time series profile clustering of temperature data. In comparison to the k-means or hierarchical clustering, GMM is a soft clustering algorithm, i.e. individual samples influence the centroids of all clusters and not only the one they belong to (a comparison of both approaches is found in Park et al., 2019). Soft clustering may be suitable for the given problem as indoor temperatures are inherently continuous and cannot be sorted into discrete, separable bands. Comparing and picking the best performing clustering algorithm is not within the scope of this study but would be valuable future work.
The output of the GMM is a probability density function $P_k(x)$ for each of the clusters $k \in K$ given a set of features X. The density functions consist of a linear combination of multiple Gaussian distributions $N(x; \mu_{kr}, \Sigma)$ (Hastie et al., 2009).

$$P_k(X) = \sum_r \pi_{kr} N(X; \mu_{kr}, \Sigma) \quad (1)$$

Here all clusters share the same covariance matrix Σ. The optimum value of all parameters, i.e. the mean of each Gaussian distribution, the mixing proportion $\pi_{k,r}$ for each of the R Gaussian distributions and covariance matrix Σ are chosen by maximising the log-likelihood

$$\sum_k^K \sum_{g_i=k} log \left[\sum_{r=1}^{R_k} \pi_{kr} N(x_i; \mu_{kr}, \Sigma) \prod_k \right] \quad (2)$$

of all clusters $k \in K$ simultaneously, where $\prod_k$ represents the clusters prior probability. The cluster with the highest probability given a set of parameters x is the one proposed by the GMM. Fitting the GMM is done using the expectation-maximisation (EM) algorithm (Dempster et al., 1977).

Before the GMM is fitted to the data, the number of clusters is picked manually. The common way is to use information criteria like BIC or AIC which enable to qualitatively compare accuracy of models with different number of clusters. In our case, we optimized the number of clusters to maximize predictive accuracy of the whole approach in Figure 1.

Model derivation

Association rule mining

Like GMM, association rule mining (ARM) is an unsupervised learning technique that identifies interesting relationships between features and targets (Jirı and Kliegr, 2012). It was initially applied to market basket analysis for the identification of simple rules to understand consumer behaviour.
First, the discretized features X and targets Y_C are stored in a transactional database. The transactional database is scanned for association rules using one of the existing ARM algorithms (here: Apriori algorithm, Agrawal et al., 1994). The quality of a rule is quantified by calculating support and confidence of each rule described in the following equations.

$$supp(A) = |t \in T; A \subseteq t| \,/\, |T| \quad (3)$$

$$conf(A \Rightarrow B) = supp(A \cup B)/supp(A) \quad (4)$$

Let A be a feature set, A ⇒ B an association rule and T a set of transactions of a given database. Support captures how likely it is that A and B occur jointly ($P(A, B)$) while the confidence provides a value for how likely the occurrence of B is if A is given. A minimum value for support is used to place a limit on the number of rules.
For classification purposes, the rule mining algorithm is adjusted to restrict the consequent B to only contain the target variable Y_C. Ma and Liu (1998) formulated the framework for creating association rules in this manner, naming them class association rules (CARs). A predictive classification model is created by a subset of CARs which are picked using Classification by Association (CBA).

Classification by Association (CBA)

CBA is a supervised machine learning algorithm which stands out due to its simplicity. It takes CARs as inputs, sorts them and outputs a subset of useful rules that can classify sets of features. As outlined by Ma and Liu (1998), to derive the CBA model, CARs are sorted by the confidence, then support, and then the order the rules are generated in. Each entry of the training data is covered by at least one rule.
The CARs derivation, sorting and deleting of rules is conducted based on training data and therefore may be regarded as model training. Afterwards, the remaining rules can be applied to unlabelled data picking the first rule within the list of sorted rules that is satisfied by a given set of new features.
The rules picked by CBA are a useful output in themselves, because they provide a human readable list of the most predictive features for target selection. Ma and Liu (1998) describe this as the discovery of understandable rules. The CBA framework can provide more understandable and more predictive rules than association rule mining alone. In addition to the prediction of targets on unseen data the outputted rule set can assist in achieving the human readable functionality desired in many applications.
In this study we used the pyFIM and PyARC libraries for ARM and CBA implementation (Borgelt, 2012)(Jirı and Kliegr, 2012).

Case Study

The methodology is applied to predict indoor temperatures in a small room ($\approx 10m^2$, one worker, one window) of an office building in British Columbia.

Table 1: Overview on the dataset split into target and features.

Type	Sensor name	properties
Target Y_T	Indoor air temperature [$^\circ C$]	hourly mean (15 min. data)
Features X:	outdoor air temperature forecast* [$^\circ C$]	daily mean on-site measured data
	wind chill* [$^\circ C$]	daily mean, on-site measured data
	heat index* [$^\circ C$]	daily mean, on-site measured data
	relative humidity* [%]	daily mean, on-site measured data
	occupancy [%]	daily mean, on-site measured data
	lagged profile number []	profile number from previous day (predicted by GMM)
	date - time []	day of the week, month, season

*discretized by equal frequency binning.

The indoor climate of the room is controlled by a trickle vent and slab heating or cooling. The trickle vent preheats or cools fresh air using a coil. Both systems are connected to a central heat pump.

Data and Feature selection

In the proposed approach the selection of input features is crucial as they form the rules shown to occupants to understand temperature predictions. For now, we limit the set to only a small selection of features, constrained by data availability and quality.
The considered data set spans three years (2014-2017). It consists of measured values on the building systems and the internal and external climate conditions. The data is not public but information on the building are publicly available.[2] The data from multiple sensors was cleaned and aligned to a frequency of one hour (Y_T) or one day (X).[3] All continuous features are discretized into bins with equal numbers of samples. Besides the listed features, we also had access to temperature set point (occupant input) data of the room which was constant over the whole period and therefore ignored.
Among the features in Table 1, we selected a subset based on an exhaustive grid search (see next section). In future applications, more occupant inputs (adaptive actions, see Section 1), building system data and sensor data of adjacent rooms might be important to ensure useful explanations for forecasts.

Model derivation

The model was trained on two years of data (Nov. 2014 to Nov. 2016) and tested on the following year. As the CBA algorithm ranks rules based on support and confidence values derived on the training data, it is crucial that the training data consists of the same number of samples from each season. Otherwise, the support (Eq. 3) for rules of an underrepresented season will be relatively low in comparison to rules of other seasons. Similarly, the confidence of rules (Eq. 4) would be skewed.
In Figure 2, the resulting seven temperature profiles generated with Gaussian Mixture modelling are shown. All results in this section were derived using the optimized number of clusters, bin size and set of features (see Table 2). The profiles may be sorted from hot to cold and by differences in shape. Two profiles are rather flat with low average temperature. The other five profiles fluctuate strongly between day and night and the temperature is warmer on average.
Next we apply associate rule mining and extract the classification by association rules (CARs). We receive distinct explanations for each cluster (see Fig. 3). The rules in Fig. 3 show the three rules with the highest support value for each cluster. Some clusters have less than three rules in which case all of the associated rules are shown.
The most days (highest support) in the training data

[2] https://www.reliablecontrols.com/corporate/facility/

[3] Instances of sensor outages were found at various points in the data set. Days with one or more missing indoor temperature values are ignored leading to a loss of 12.9% of samples. In future, measurement gaps could be filled with rolling mean values.

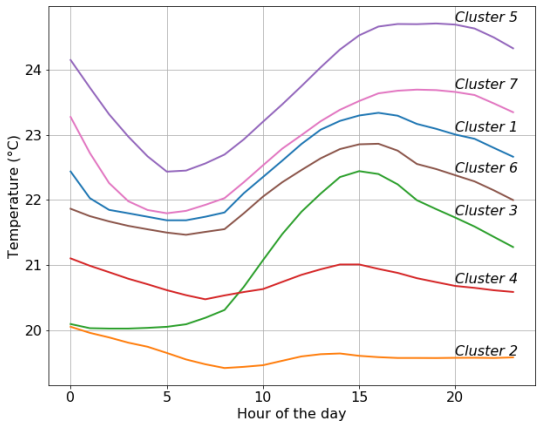

Figure 2: Temperature profiles for each of the seven clusters.

Consequent (then...)	Antecedent (if...)	confidence	support
Cluster 1	day-of-week=Tuesday,Previous_day=0.0	0.556	0.009
	season=Fall,day-of-week=Tuesday,Wind_Chill=high	0.556	0.009
Cluster 2	day-of-week=Sunday,month=12	1.000	0.013
	day-of-week=Sunday,Previous_day=Cluster 4,Heat_Index=low	0.600	0.011
	oat_mean=very low,Occupancy=False,month=12	0.600	0.011
Cluster 3	Occupancy=True,day-of-week=Monday,season=Winter	0.400	0.015
	Occupancy=True,day-of-week=Monday,quarter=4,Previous_day=Cluster 2	0.750	0.011
Cluster 4	Previous_day=Cluster 6,Occupancy=False,Heat_Index=low	0.750	0.022
	Heat_Index=low,day-of-week=Saturday	0.857	0.022
	day-of-week=Sunday,Occupancy=False,quarter=1	0.588	0.018
Cluster 5	Wind_Chill=very high,Previous_day=Cluster 5,Occupancy=False,season=Summer	1.000	0.015
	Previous_day=Cluster 5,OA_RH=low	0.667	0.011
	day-of-week=Monday,month=6	0.556	0.009
Cluster 6	quarter=1,Occupancy=True	0.785	0.133
	Wind_Chill=medium,oat_mean=medium	0.737	0.102
	Occupancy=True,season=Spring,Wind_Chill=medium	0.800	0.051
Cluster 7	quarter=3	0.664	0.165
	Heat_Index=very high,quarter=3	0.795	0.113
	Heat_Index=very high,Occupancy=True,season=Summer	0.753	0.100

Figure 3: Top 3 rules for classification of each cluster (sorted by support).

are members of *Cluster 6* and *7*. *Cluster 6* represents occupied days during winter (quarter 1) and shoulder season (spring) with medium outside air temperature and wind chill, and *Cluster 7* is the typical profile for occupied days in summer (quarter 3, Season = Summer). *Cluster 4* and *2* show the profiles for unoccupied days. During unoccupied days in winter the temperature typically drops to below $20°C$. *Cluster 3* has a very distinct shape. It captures the reheating process after unoccupied days in winter which typically occurs on Mondays. *Cluster 5* describes overheating inside the room. The rules show that this happens on days when wind chill is high meaning high ambient temperature and low wind speeds. The strong impact of wind speed is due to the fact that the room features trickle vents which rely on natural ventilation for cooling. Lastly, *Cluster 1* has very low support values. This is surprising as it lies between the two most common clusters. The reason may be that the control routine of the heating and cooling system leads to indoor profiles very close to *Cluster 7* OR *Cluster 6* and nothing in between.
Finally, we apply the derived model to unseen data and compare the results to the observed indoor temperature profiles. Model derivation and testing was conducted iteratively in an exhaustive grid search with the number of clusters, the number of bins for variable discretization and the selection of features as hyperparameters. To speed up the process the features were grouped into four sets (Table 1). The optimal parameter settings are shown in Table 2. Especially, the use of clusters of the previous day increased the accuracy significantly. They were derived using the mixture model trained on the training data.

Table 2: Results of grid search.

Hyper-parameter	Range	Final choice
No. of clusters	[1,15]	7
Bin size	[2,10]	5
Feature subsets	[Date time],[Lagged Clusters],[Weather], [Workday], [Occupancy]	[Date time],[Lagged Clusters], [Weather], [Occupancy]

Model validation

Testing the method on unseen data gives a Mean Absolute Error (MAE) of $0.558°C$ and 62.5% of the variation in the indoor temperature is explained ($R^2 = 0.625$). Figure 4 shows the characteristics of cluster based prediction with a cap at high temperatures and floor at low temperatures. Furthermore, due to the discrete classification of profiles the predictions exhibit a gap between $19.7°C$ and $20.3°C$. To better understand the performance and the causes of inaccuracies, we decomposed the inaccuracies and benchmarked our algorithm to two different applications of neural networks.
In a first step, the loss of variance caused by using

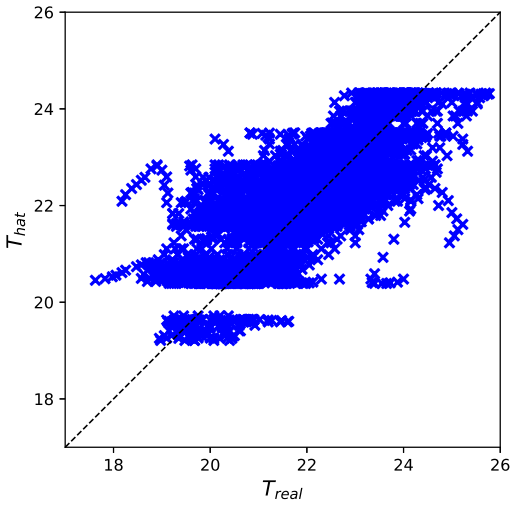

Figure 4: Observations vs. predictions on test data.

daily temperature profile clusters instead of predicting each hourly temperature value individually is shown in Figure 5. Using seven clusters, which was determined to be optimal by the grid search, a maximum R^2 of 0.79 is theoretically achievable if all clusters are predicted correctly. Hence, there is a 21% loss in theoretically explainable variance by the process of converting continuous hourly target values to seven discrete daily clusters.
Another simplification of the prediction process is the use of association rules instead of a complex statistical regression model. To quantify the loss of accuracy induced by rule based prediction, we conducted the cluster prediction with a parameterized black-box classifier. Here, we use a feed-forward neural network classifier whose parameters were again optimized in a grid search. It outperformed the CBA algorithm only by a little ($R^2 = 0.665$).
After having decomposed the loss of accuracy, we benchmarked the algorithm against a state-of-the-art deep neural network regressor which predicts 24 temperature values individually for each day. The regressor is fed with the same set of inputs as before while having 24 temperature outputs. The network is composed of three layers with 200 neurons each and was pruned by increasing the regulation term α (Hastie et al., 2009) step-by-step until optimum performance was achieved.[4] The accuracy is much higher (R^2=0.815) than the proposed cluster- and rule-based approach but with loss of explainability. Also it shows that given the current set of features the neural network fails to explain 19.5% of the variance. Probably, more features on occupants and other unknowns may be helpful to further increase accuracy.

[4]The process of *pruning* refers to gradually increasing the regularization term until variance and bias of the neural network are balanced.

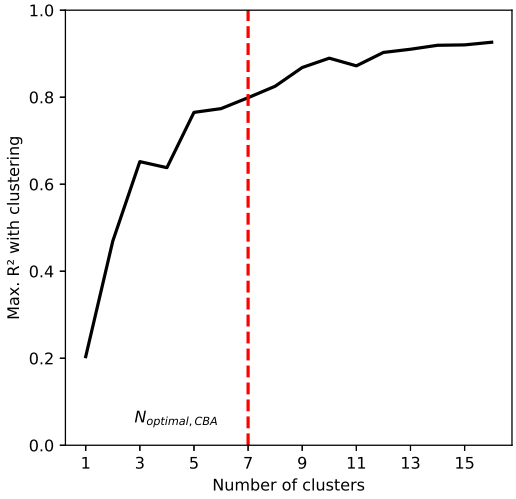

Figure 5: Maximum achievable R^2 score for a given number of clusters.

Model application

The functionality of the proposed algorithm is shown in Fig. 6. One week of indoor temperature predictions for each season is shown. The weeks were selected randomly among the weeks without missing data. The rules which caused the CBA algorithm to predict one of the seven profiles for the upcoming day are shown below each of the predicted 24h temperature profiles (split by black lines). For example, on 8th April 2017 the model predicts *Cluster 2* because the previous day was *Cluster 6*, the heat index is medium, and the building was unoccupied.

Generally, we find that the cluster- and rule-based prediction is capable of capturing the indoor temperature behaviour well. Weekends are identified and depending on weather conditions different profiles are picked during the week (see winter week). However, we also see that *Cluster 6* in winter and *Cluster 7* in summer are classified on most days. Rarely, a significant misclassification of a day can be observed as for example found on 22nd October 2017.

The dominance of two clusters is explainable due to the impact of the heating and cooling system, and due to the fact that the temperature set point was never changed by the worker in the training and testing data. As a consequence, our classifier mainly distinguishes between the seasons and between days where the HVAC system is switched on and those when it is switched off.

Misclassification may be caused by ambiguous information provided by the features. On 22nd October 2017, the classifier predicts the building to be heated but instead the heating system was switched off as it is Sunday. On that day the occupancy sensor recorded some activity in the room. This triggers the CBA algorithm to predict the wrong cluster, because in this specific case occupancy-based rules have higher confidence than rules which consider that it is a Sunday and the room should be unheated. A similar misclassification is observed on the 19th March which was also a Sunday.

Discussion

The case study showed that the proposed method is convenient to apply. Once a pipeline of clustering and rule-mining is established, it generates forecasts alongside of comprehensible sets of rules. In Fig. 6 a maximum of four variables per rule were generated which seems suitable for rapid forecast analysis.

The data available for the case study lacks information on occupant action. The rules like *it will be hot (Cluster 5) because wind chill is high and the building is occupied* (see Fig. 3), do not recommend any occupant action.[5] In further applications, the data should be complemented with behavioural features. For example, if an occupant knows that *it will be hot because wind chill is high, the building is occupied and windows are closed*, he or she will open the window to increase comfort.

Model parameters and model performance considerations

The number of clusters is the only model parameter of the GMM which was optimized. Its covariance matrix, another parameter of the GMM, was set to be full, i.e. each cluster has a different, full covariance matrix. A brief study showed that this is better than all other choices of covariance matrix type (all clusters sharing the same matrix or the matrices may only have diagonal elements).

The rule mining process has four modelling parameters, i.e. minimum support, minimum confidence, the bin size of the variable discretization and the involved features. We included the latter two into the hyper-parameter optimization process. Minimum confidence was removed (set to zero) and minimum support set to five days. This ensures that any derived rule is found at least five times in the data.

The accuracy of the model is significantly lower than 24h predictions of a deep neural network as shown in Table 3. However, one could argue that a loss of $0.19°C$ in MAE may be acceptable if the method helps occupants to improve energy efficiency of the building by adjusting their behaviour. This trade-off in loss of accuracy and improved occupant behaviour has yet to be studied in a field test.

The prediction accuracy of the model can be improved by deriving better rules to predict more

[5] High wind chill index refers to high ambient temperatures and low wind speeds.

Table 3: Model validation and benchmarking for 24h predictions.

Error Type	GMM + CBA	GMM + ANN_C	ANN_{Reg}
MAE $[°C]$	0.558	0.548	0.37
R^2	0.625	0.665	0.815

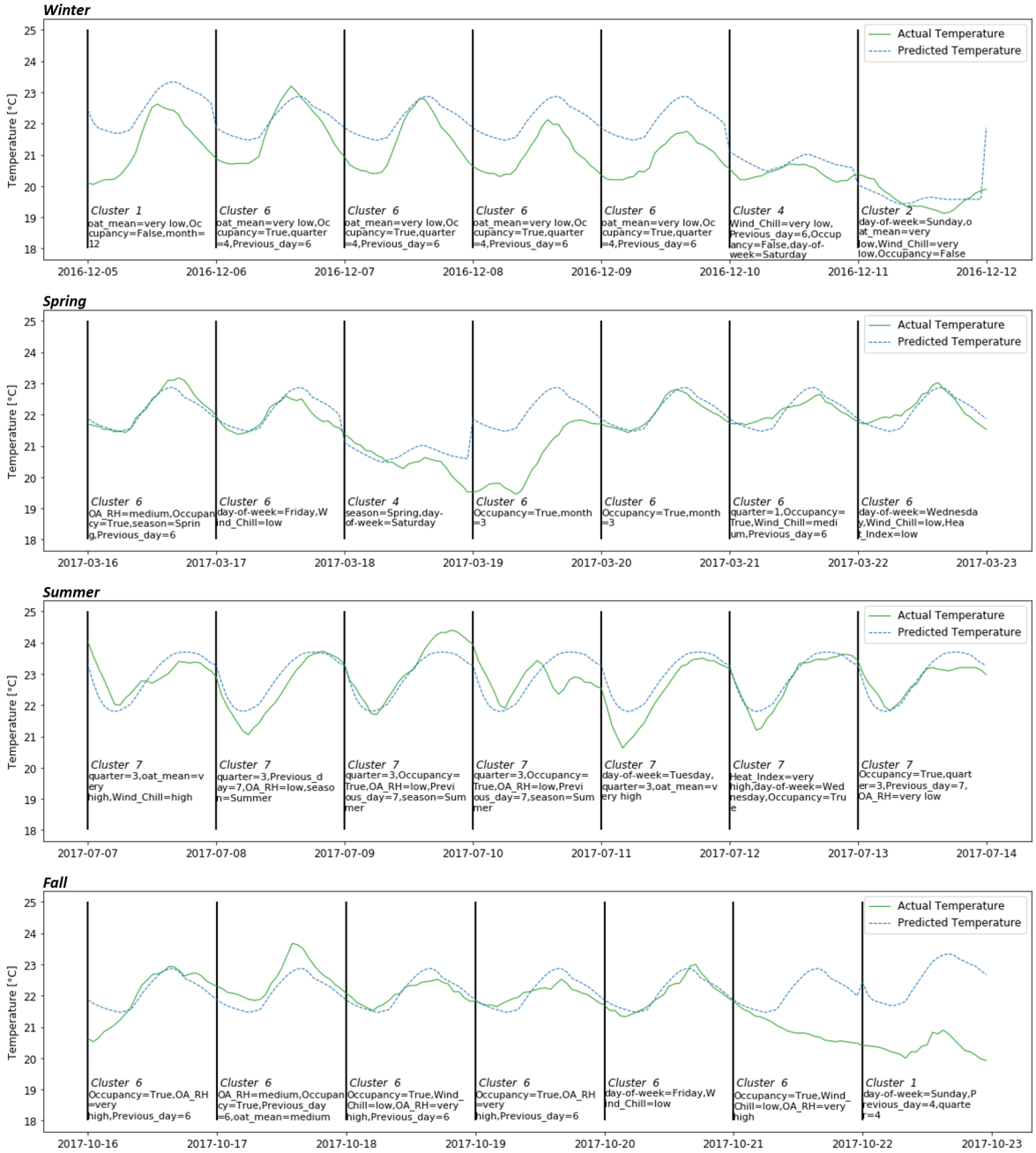

Figure 6: Predictions and associated rules for one week of each season in the test data.

clusters accurately (see Figure 5). For example, by using ten clusters the maximum achievable R^2 score would increase from $R^2_7 = 0.79$ to $R^2_{10} = 0.88$. With the current set of features and the resulting rules, we determined seven clusters to optimal. Our rule set is not explanatory enough to accurately predict more clusters. If more or better features are found, more clusters could be accurately predicted.

The benchmarking analysis showed that our method with the current way of feature engineering does not fully leverage the information hidden in the data. A neural network achieved much higher accuracy given the same set of information. More work on feature engineering could be done, but also it may be concluded that an increase of explainability leads to a loss in accuracy.

Conclusions and Future Work

This study introduced and benchmarked a novel approach to provide hourly forecasts on building behaviour for the upcoming day. It combines the analytical power of unsupervised machine learning (clustering, associate-rule mining) with the prediction ability of supervised machine learning methods given by the CBA algorithm. As a result each forecast is complemented with rule-based explanations why a certain forecast was given. This would enable occupants to adapt and adjust their actions. In future, we imagine the method could help to involve occupants in the building control loop which may lead to an increase in building energy efficiency.

After having benchmarked the accuracy of the

method against *black-box* models, the next step is to conduct a second case study where rules are provided to actual occupants of a building. This could be done by implementing the forecasting method on an intelligent personal assistant device to communicate the explanations and recommendations associated with temperature or energy consumption forecasts. This will clarify if influencing occupant actions can increase overall building efficiency.

Acknowledgements

We thank Reliable Controls Corporation for providing the data and NSERC for funding the research work.

References

Agrawal, R., R. Srikant, et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Volume 1215, pp. 487–499.

Andersen, R. V., J. Toftum, K. K. Andersen, and B. W. Olesen (2009). Survey of occupant behaviour and control of indoor environment in danish dwellings. *Energy and Buildings 41*(1), 11–16.

Borgelt, C. (2012). Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2*(6), 437–456.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

D'Oca, S., T. Hong, and J. Langevin (2018). The human dimensions of energy use in buildings: A review. *Renewable and Sustainable Energy Reviews 81*, 731–742.

Eirola, E. and A. Lendasse (2013). Gaussian mixture models for time series modelling, forecasting, and interpolation. In A. Tucker, F. Hppner, A. Siebes, and S. Swift (Eds.), *Advances in Intelligent Data Analysis XII*, Lecture Notes in Computer Science, pp. 162–173. Springer Berlin Heidelberg.

Hastie, T., R. Tibshirani, and J. Friedman (2009). The elements of statistical learning: data mining, inference, and prediction.

Hong, T., D. Yan, S. D'Oca, and C.-f. Chen (2017). Ten questions concerning occupant behavior in buildings: The big picture. *Building and Environment 114*, 518–530.

IEA (2017). Energy technology perspectives. Technical report, International Energy Agency.

Jirı, F. and T. Kliegr (2012). Classification based on associations (cba)-a performance analysis.

Lin, H.-W. and T. Hong (2013). On variations of space-heating energy use in office buildings. *Applied Energy 111*, 515–528.

Lu, J., T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field, and K. Whitehouse (2010). The smart thermostat: using occupancy sensors to save energy in homes. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pp. 211–224. ACM.

Ma, B. L. W. H. Y. and B. Liu (1998). Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*.

Melzi, F. N., A. Same, M. H. Zayani, and L. Oukhellou (2017). A dedicated mixture model for clustering smart meter data: Identification and analysis of electricity consumption behaviors. *10*(10), 1446.

Mirebrahim, S. H., M. Shokoohi-Yekta, U. Kurup, T. Welfonder, and M. Shah (2017). A clustering-based rule-mining approach for monitoring long-term energy use and understanding system behavior. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*, pp. 5. ACM.

Park, J. Y., X. Yang, C. Miller, P. Arjunan, and Z. Nagy (2019). Apples or oranges? identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. *Applied Energy 236*, 1280–1295.

Ramesh, T., R. Prakash, and K. Shukla (2010). Life cycle energy analysis of buildings: An overview. *Energy and buildings 42*(10), 1592–1600.

Shaikh, P. H., N. B. M. Nor, P. Nallagownden, I. Elamvazuthi, and T. Ibrahim (2014). A review on optimized control systems for building energy and comfort management of smart sustainable buildings. *Renewable and Sustainable Energy Reviews 34*, 409–429.

Xiao, F. and C. Fan (2014). Data mining in building automation system for improving building operational performance. *Energy and buildings 75*, 109–118.