Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 122 (2017) 337–342

Energy Procedia

www.elsevier.com/locate/procedia

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

Indoor Environment Quality (User Comfort, Health and Behaviour)

Comparison of Different Classification Algorithms for the Detection of User's Interaction with Windows in Office Buildings

Romana Markovic[a,*], Sebastian Wolf[b], Jun Cao[a], Eric Spinnräker[a], Daniel Wölki[a], Jérôme Frisch[a], Christoph van Treeck[a]

[a] *RWTH Aachen University, Mathieustr. 30, 52074 Aachen, Germany*
[b] *Technical University of Denmark, Asmussens All, Building 303B, 2800 Lyngby, Denmark*

Abstract

Occupant behavior in terms of interactions with windows and heating systems is seen as one of the main sources of discrepancy between predicted and measured heating, ventilation and air conditioning (HVAC) building energy consumption. Thus, this work analyzes the performance of several classification algorithms for detecting occupant's interactions with windows, while taking the imbalanced properties of the available data set into account. The tested methods include support vector machines (SVM), random forests, and their combination with dynamic Bayesian networks (DBN). The results will show that random forests outperform all alternative approaches for identifying the window status in office buildings.

Keywords: occupant behavior; window opening; office buildings, SVMs; Random forest;

1. Introduction

Occupant behavior is seen as the main source of discrepancy between predicted and measured energy consumption in buildings. Hence, understanding occupant behavior is crucial for achieving high performance and low-energy use, both in the commercial and the residential field [15].

Users' interactions with windows, in terms of window opening and closing are needed for modelling air exchange through ventilation. Nonetheless, they should be taken into account for controlling strategies of buildings' mechanical ventilation. Modelling window opening behavior is an important part of building performance simulation, in order to make reliable predictions of the buildings' energy consumption [24]. However, the conventional building simulation approaches still rely on synthetic window opening predictions, that do not lead to a realistic occupant's influence on the energy performance.

Occupant behavior and perceived thermal comfort in office buildings have been investigated in numerous studies [13], [8], [7], [11], [23], [9]. In addition, there are multiple studies that investigated window opening behavior in offices

[*] Corresponding author. Tel.: +49-241-80-25541 ; fax: +49-241-80-22030.
E-mail address: markovic@e3d.rwth-aachen.de

10.1016/j.egypro.2017.07.333

[14], [10] and residential buildings [2], [20], [22]. Haldi and Robinson [14] showed that a Markov model provides higher accuracy compared to the logistic regression and agent based method. However, even though the model provided over 80 % of correct predictions in case of closed windows, the ability to predict an open window remained low. D'Oca and Hong [9] applied a data-mining approach to discover the patterns of window opening and closing in office buildings. They identified several behavioral patterns, including motivational and opening duration patterns. Similarly to Haldi and Robinson, they showed that indoor and outdoor air-temperature together with the time of a day and presence durations were the strongest factors leading to window opening and closing actions.

Machine learning (ML) and artificial intelligence (AI) techniques are widely used for predicting and evaluating occupant's actions in buildings [5], [16] as well as buildings' energy consumption [12], [1], [3]. However, there is little work that uses smart algorithms for modelling the occupants' interactions with windows in case of office buildings. Furthermore, human behavior, including window opening and closing actions, cannot be modelled using analytical physical approaches. As a result, occupant actions have to be modelled using data-driven methods. For this purpose, machine learning methods offer a comprehensive alternative to modelling the occupant behavior in buildings and its' influence on the energy consumption.

This paper models occupant's actions and the resulting window status in office buildings by applying support vector machines (SVMs) and random forest. Based on monitoring data, the window status is defined as classification problem, where the status, open or closed, is identified. In addition, the temporal dependence of window actions is investigated by implementing a dynamic Bayesian network (DBN), with the aim of smoothing the classification results.

2. Method

2.1. Data Set

Data set includes monitoring data collected over two years in an office building in Frankfurt, Germany [17], [21]. The available data are collected on ten monitored offices in ten minutes time-steps. Due to the very low occupancy rate, one of the ten offices is excluded from the further evaluation. Measured data include indoor climate features (indoor air temperature) and outdoor climate features (outdoor air temperature, precipitation, wind velocity, wind direction, CO_2 concentration and relative humidity) as well as occupant's presence and actions (position sun protection, occupancy, time presence, occupancy state).

The window status is defined as a binary problem, where 0 and 1 refers to a closed and an open window, respectively. In addition, it is not distinguished between both windows in each office. As a result, all data points where at least one of the two windows is opened are labelled as class 1. In case of SVMs, features are scaled in range between 0 and 1 prior to data splitting into training and evaluation set. Since random forests does not require feature scaling, the random forest data remained in the original monitored range.

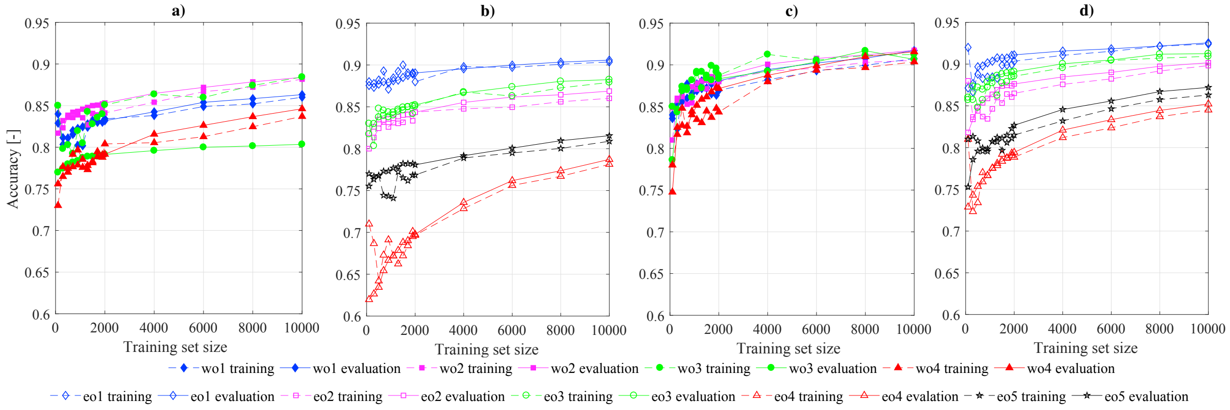

Fig. 1. Training and evaluation accuracy for varied training set size for SVMs ((a) and (b)) and random forest ((c) and (d)).

In order to avoid over-fitting, an optimal training set size had to be determinated for each implemented classification algorithm. For that purpose, it is iterated over the number of training data points until the convergence criteria of training accuracy and evaluation accuracy is fulfilled. The results of iterations over training set size are presented in Figure 1. Eventually, it is opted for the training set size of 4000 data points, which corresponds to approximately 4 weeks of monitoring data. With the aim to perform training on data collected during all seasons, the training set consists of measured values over one week in January, April, July and October, respectively.

2.2. SVMs

SVMs are set of machine learning methods that extract models or patterns from data [19], which tries to find a hyperplane that maximizes the margin between different classes. For a detailed theoretical background, the reader is referred to [4].

An overview of the developed model is presented in Figure 2. Training is preformed using a radial based function (RBF Kernel), and an optimal combination of the penalty factor C and the inverse of radis σ is searched with the aim of achieving the highest prediction accuracy. C is varied in range between 2^0 and 2^{10}, while possible σ values are iterated from 0.05 to 1.0. Tested cases include the cases where all labels are weighted with factor 1, as well as the case of higher weights for the penalty coefficient for misclassified labels for the under-represented class. A detailed result for each case are presented in Section 5. For each training iteration, the model is validated using a five−fold cross validation. The trained models are exported and eventually used for model evaluation. For the model evaluation purpose, data that was not used in the training procedure is randomly shuffled and fed into the trained model.

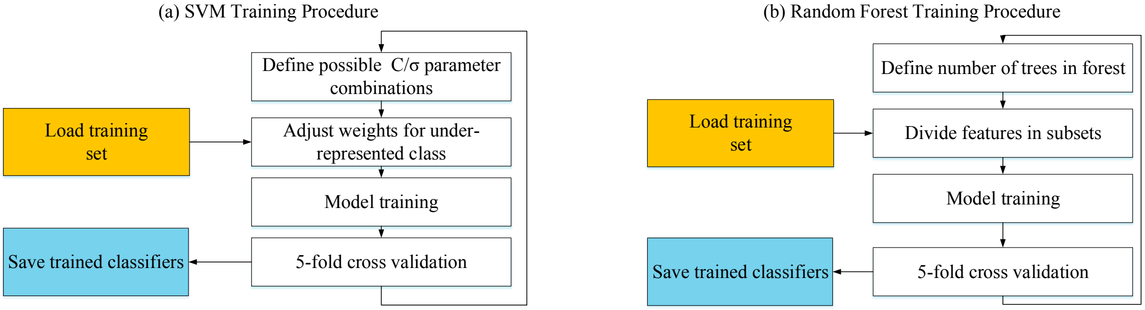
(a) SVM Training Procedure
Define possible C/σ parameter combinations
Load training set
Adjust weights for under-represented class
Model training
Save trained classifiers
5-fold cross validation
(b) Random Forest Training Procedure
Define number of trees in forest
Load training set
Divide features in subsets
Model training
Save trained classifiers
5-fold cross validation

Fig. 2. (a) SVM training procedure; (b) Random forest training procedure.

2.3. Random Forest

The idea of random forests is to use a set of decision trees as weak classifiers, in order to provide a strong classifier. For detailed theoretical background about random forests, the reader is referred to [4] and to the original random forest application [6]. A predictive model is developed using bagged trees on the varied data subsets. Data subsets are created by randomly choosing a predefined number of input features. In case of the random forest classification models, the empirical rule is to train trees on subsets consisting of a feature number equal to squared root of the overall feature amount. In this case, the models were trained using between 3 and 11 features in each subset. In addition, an optimal number of decision trees for each model is investigated in the range between 10 and 200 decision trees.

2.4. Dynamic Bayesian Network

The classification of the identified window status is performed without taking temporal dimension and duration of each opening event into account. The role of a dynamic Bayesian network is to predict how long a window stays opened without changing the status to closed. This is achieved by incorporating likelihood for windows to remain opened to each following time stamp.

An overview of the implemented network is presented in Figure 3. The transitions (blue nodes) correspond to corrected window status, while the measurements (white nodes) represent the classification results by implemented classification models (SVMs and random forest). The likelihoods for the nodes and edges to have status 1 (open) or 0 (closed) are

computed for each office separately.
Eventually, the results of the case where dynamic Bayesian network are used to build a time-serie from the classifier results. They are evaluated and compared to the case where only classifiers are used for window status identification.

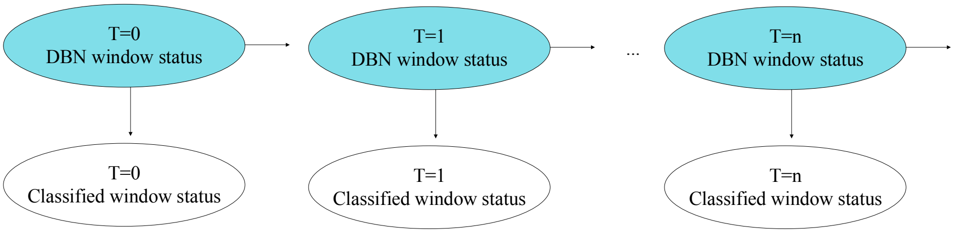

Fig. 3. Dynamic Bayesian network used in combination with trained classifiers.

3. Results

In case of SVMs, the highest performance for all tested offices is achieved for the penalty factor C equal to 2.0 and σ values between 0.95 and 1.0, while an optimal weight coefficient for the under-represented class is 3. The σ values implies that an application of a linear SVMs would lead to the identical accuracy. In case of random forest, the highest accuracy is achieved for 3 features per each random feature subset. Supplementary, a larger number of trees resulted in higher accuracy and higher computational costs. In order to find a meaningful trade-off between model complexity and predication accuracy, it is opted for 200 decision trees as an optimal number of weak classifiers.
The performance of the investigated methods is evaluated using an overall accuracy (ACC) and confusion matrix, where ACC corresponds to the proportion of overall correctly classified evaluation data points. The confusion matrix consists of true positive rate (TPR), true negative rate (TNR), false positive rate (FPR) and false negative rate (FPR). TPR is defined as coefficient of true positive classified points to the sum of true positive and false negative points, while TNR is the coefficient of true negative points to the sum of true negative and false positive points.
Mean prediction accuracy for the tested models is in the range between 0.82 % in case of SVMs and 0.89 % in case of random forest. In addition, the evaluation accuracy remained higher for the data points where windows were closed compared to the case of open windows. This may be interpreted as the consequence of the data set imbalance. However, the random forest showed satisfying performance for identifying the opened windows. A detailed overview of the models' performance can be found in Table 1.

Table 1. Performance comparison of SVMs and random forest for investigated offices.

	SVMs			**RF**			**DBN+ SVMs**			**DBN+ RF**		
office ID	**ACC**	TPR	TNR	**ACC**	TPR	TNR	**ACC**	TPR	TNR	**ACC**	TPR	TNR
wo1	**0.84**	0.53	0.91	**0.89**	0.51	0.97	**0.83**	0.44	0.93	**0,87**	0,45	0,97
wo2	**0.87**	0.69	0.91	**0.90**	0.67	0.95	**0.86**	0.55	0.93	**0,89**	0,56	0,96
wo3	**0.82**	0.28	0.94	**0.90**	0.71	0.97	**0.82**	0.25	0.95	**0,90**	0,55	0,98
wo4	**0.82**	0.82	0.84	**0.89**	0.72	0.94	**0.81**	0.50	0.89	**0,88**	0,55	0,96
eo1	**0.87**	0.66	0.93	**0.92**	0.58	0.97	**0.87**	0.56	0.94	**0,89**	0,51	0,98
eo2	**0.84**	0.66	0.89	**0.89**	0.57	0.96	**0.84**	0.52	0.91	**0,87**	0,48	0,97
eo3	**0.88**	0.73	0.91	**0.90**	0.67	0.95	**0.86**	0.57	0.93	**0,89**	0,55	0,96
eo4	**0.69**	0.86	0.65	**0.83**	0.80	0.85	**0.76**	0.36	0.85	**0,83**	0,45	0,91
eo5	**0.80**	0.61	0.84	**0.85**	0.61	0.93	**0.80**	0.44	0.88	**0,86**	0,48	0,95
mean value	**0.83**	0.64	0.87	**0.89**	0.65	0.94	**0.83**	0.47	0.91	**0.88**	0.51	0.96

In case of the imbalanced data set, in which one class is represented in significantly higher proportion in comparison to other, the receiver operating characteristic (ROC) is the reliable mean of evaluation. The ROC curve is graphically presented as TPR against FPR. ROC diagramms for the evaluated cases are shown in Figure 4.

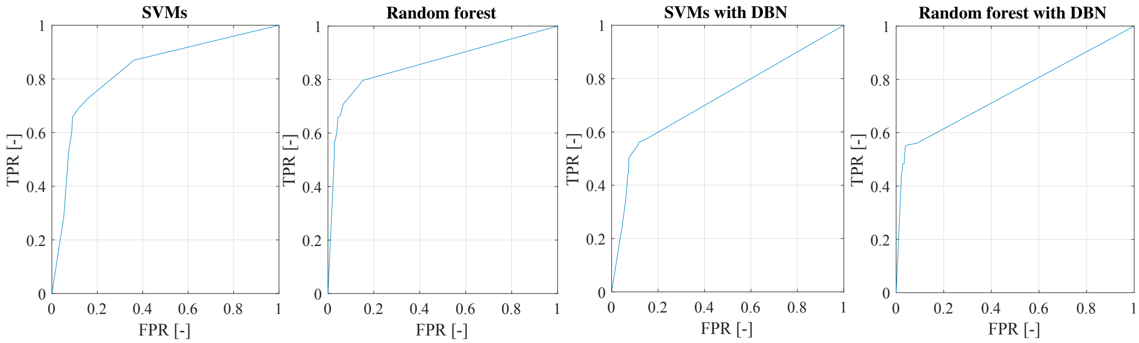

Fig. 4. ROC curves for the investigated cases.

4. Discussion

Realistic and accurate modeling of occupant behavior is necessary for building automation and controlling in order to perceive a satisfying standard of indoor climate and comfort. Thus, a reliable model of occupants behavior in terms of window opening is needed to complete HVAC controlling strategies and develop automatized window controlling. An application of SVM algorithms for identifying window status in office building showed satisfying performance. The highest performance is achieved in case of taking the unbalanced properties into account by weighting the penalty factors for the under-represented class. However, a drawback of SVMs is a computationally expensive model training procedure.

The random forest for modeling window status significantly outperformed all alternative methods for identifying the window status, and the presented results are state-of-the-art. Due to it characteristics of combining a large number of weak classifiers, in this case 200 decision trees, it showed a satisfying accuracy for the under-represented label. As a result, the random forest algorithm can be used on a wide range of offices with little prior knowledge about the frequency and time duration of the window opening actions.

An implementation of dynamic Bayesian network for modeling the window position in the temporal domain based on prior knowledge about the durations of the window openings did not improve the classification results. This may be caused by low probabilities for windows to remain open over longer time intervals (over 30 minutes), which resulted in a misclassification of longer periods where windows remained opened. However, a more complex graphical model which includes further temporal features could lead to a higher prediction accuracy.

Models are trained separately for each office in order to find the best fit for each user, which resulted in computationally expensive training procedures. Although in case of random forests all offices achieved an optimal performance in case of the identical number of bagged trees, the model trained on one office would not show a sufficient performance in case of different offices. This is caused by the individualized entropies and tree structures for each trained model. Therefore, training procedures may be optimized by finding the generic model parameters that may provide a sufficient accuracy to a group or cluster of behavioral models. As a part of this work, models are trained of a data set consisting of nine single- or two person offices. However, training should be conducted on a large number of offices and independent data sets.

5. Conclusion

Based on the evaluation results, the proposed approaches outperformed alternative methods for window opening models, and the implementation of a random forest with 200 trees scored state-of-the-art results.

However, a very intense training procedure which must be applied for each individual occupant cannot be covered by computational resources of conventional software used in building automation. In addition, its' implementation in thermal building simulation for window opening controlling may lead to slower and less computationally efficient simulation process.

As a result, a generic window opening model should be developed based on the proposed methods. A generic model should consist of a window opening model that can allow a slightly lower accuracy, but it would be applicable for a cluster of users without model re-training. In addition, it may be supported by a model that may learn the occupant's behavioral patterns leading to the highly individualized window opening behavior.

6. Acknowledgements

The authors appreciate the financial support of this work by the German Federal Ministry of Economics and Energy (BMWi) as per resolution of the German Parliament under the funding code 03ET1289D. As well, we thank Marcel Schweiker and Andreas Wagner of Karlsruhe Institute of Technology, Germany, for providing the dataset of the office building.

References

[1] Ahmad A.S., Hassan M.Y., Abdullah M.P., Rahman H.A. A review on applications of ANN and SVM for building electrical energy consumption forecasting. In: Renewable and Sustainable Energy Reviews 2014; 33:102−109.

[2] Andersen R., Fabi V., Toftum J., Corgnati S.P., Olesen B.W. Window opening behaviour modelled from measurements in Danish dwellings. In: Building and Environment 2013. 69:101−113.

[3] Basnayake B.A.D.J.C.K., Amarasinghe Y.W.R., Attalage R.A., Udayanga T.D.I., Jayasekara A.G.B.P. Artificial intelligence based smart building automation controller for energy efficiency improvements in existing buildings. In: International Journal of Advanced Information Science and Technology 2015. 40:150−156.

[4] Bishop C. Pattern Recognition and Machine Learning. Springer 2011.

[5] Bonte M., Perles A., Lartigue B., Thellier F. An occupant behavior model based on artificial intelligence for energy building simulation. In: Proceedings of BS2013. P. 1467−1473.

[6] Breiman L.(2001). Random Forests. In: Machine Learning 2001. 45:1, p. 5−32.

[7] O'Brien W., Gunay B. The contextual factors contributing to occupants' adaptive comfort behaviors in offices − A review and proposed modeling framework. In: Building and Environment 2014. 77:77−87.

[8] Costanzo V., Donn M. Thermal and visual comfort assessment of natural ventilated office buildings in Europe and North America. In: Energy and Buildings 2017. 140:210−223.

[9] D'Oca S., Hong T. Occupancy schedules learning process through a data-mining framework. In: Energy and Buildings 2015. 88:395−408.

[10] D'Oca S., Hong T. (2014). A data−mining approach to discover patterns of window opening and closing behavior in offices. In: Building and Environment. 82:726−739.

[11] Gaetani I., Hoes P.J., Hensen J. Occupant behavior in building energy simulation: Towards a fit−for−purpose modeling strategy. In: Energy and Buildings 2016. Volume 121, p. 188−204.

[12] Gunay B., Shen W., Newsham G. Inverse black box modelling of the heating and cooling load in office buildings. In: Energy and Buildings 2017. 142:200−210.

[13] Hailemariam E., Goldstein R., Attar R., Khan A. Real−Time Occupancy Detection using Decision Trees with Multiple Sensor Types. In: Proceedings of the 2011 Symposium on Simulation for Architecture and Urban Design. P. 141-148.

[14] Haldi F., Robinson D. Interactions with window openings by office occupants. In: Building and Environment 2009. 44:12, p. 2378−2395.

[15] Hong T., Yan D., D'Oca S., Chen C. Ten questions concerning occupant behavior in buildings: The big picture. In: Building and Environment 2017. 114:518−530.

[16] Khosrowpour A., Gulbinas R., Taylor J. E. Occupant workstation level energy-use prediction in commercial buildings: Developing and assessing a new method to enable targeted energy efficiency programs. In: Energy and Buildings 2016. 127:11331145.

[17] Kleber M., Wagner A. Results of monitoring a naturally ventilated and passively cooled office building in Frankfurt am Main, Germany. Proceedings of EPIC 2006 AIVC Conference: Lyon, France 2006.

[18] Liu Y., Stouffs R., Tablada A. Coupling simulation and neural network for predicting building electricity consumption at the urban scale. In: Proceedings of BS2015. P. 633−639.

[19] Magoules F., Zhao H.-X. Data Mining and Machine Learning in Building Energy Analysis. Wiley-ISTE 2016.

[20] Psomas T., Heiselberg P., Lyme T., Duer K. Automated roof window control system to address overheating on renovated houses: Summertime assessment and intercomparison. In: Energy and Buildings 2017. 138:35−46.

[21] Schakib-Ekbatan K., Cakici F.Z., Schweiker M., Wagner A. Does the occupant behavior match the energy concept of the building? Analysis of a German naturally ventilated office building. In: Building and Environment 2015. 84:142−150.

[22] Schweiker, M., Haldi, F., Shukuya, M., Robinson, D. (2012). Verification of stochastic models of window opening behaviour for residential buildings. In: Journal of Building Performance Simulation, 5:55−74.

[23] Schweiker M., Wagner A. The effects of occupancy on perceived control, neutral temperature, and behavioral patterns. In: Energy and Buildings 2016. 117:246-259.

[24] Wolf S., Wölki D., Robinson D., van Treeck C. Evaluation and Re-training of Two Window Opening Models Using an Independent Dataset. In: Healthy Buildings Europe Conference 2017.