Tree based classification methods for occupancy detection

M Koklu*, K Tutuncu
Selcuk University, Konya, TURKEY

* Corresponding author: mkoklu@selcuk.edu.tr

Abstract. Latest smart buildings are not only be intelligent to allow occupant to control the light, heating, cooling, gas and other systems but also focuses on occupancy detection since accurate occupancy detection can result in saving energy up to 42% as can be seen in literature. For this aim, different autonomous systems including sensors, actuators, microcontroller and etc. are at the development phase for smart buildings. At this point, determination of classification methods to detect the occupancy together with hardware plays crucial role. Having done in this study 3 different classification methods that is based on Machine Learning Methods were applied on benchmark dataset named Occupancy by UCI Machine Learning Repository, 2016. The classifiers are Random Forest, Decision Tree and Bagging. They were chosen by following two principals. First one is to have classifier methods that were not use in literature for benchmark dataset and the second one almost never usage of tree based classifiers in the literature. The Occupancy dataset consists of light, temperature, humidity, CO2 and occupancy. It has been seen that the highest accuracy or prediction ratio was obtained as 99,368% by Decision Tree method namely; Random Forest. This result was compared with the results of the studies on the same benchmark dataset. It has been seen that it is the second best accuracy ratio after Fuzzy Granulation (Fgf) method among 16 different Machine Learning Based classification methods. Additionally, Decision Tree and Bagging had the accuracy ratio of 99.222% and 99.207, respectively. These ratios are also higher than other methods used in literature but Fgf. Thus this study showed how decision tree can be promising for occupancy detection.

Keywords: Classification methods, occupancy detection, random forest, decision tree, bagging

1. Introduction
The growth in population of the world, the climate change, unconscious usage of energy resources, global warming and other factors are threatening current energy resources of the world. As all other sectors the building sector responses to this problem and tries to build SMART buildings. The basic aim of SMART building is to provide an environment that is comfortable and healthy with less energy consumption. At this point, it is very important to control light, air-conditioning and other ambient related conditions with respect to number of people or occupant and ambient parameters. Mostly sensing, actuating and control technologies are used with Information Technologies to get environmental information and provide desired comfort and health [1]. Since motion sensors are not sensitive enough to detect the occupant who is hold still, video and some other sensors such as humidity, CO2, temperature, light and etc. are employed for detection of occupant. Since privacy and also requirement of large data storage are essential factors for everyone the video is not preferred for these applications. Thus, determining whether or not the occupant is detected needs different classification methods to evaluate the information coming from sensors. Once the occupant is detected related actuators will be triggered and then desired ambient conditions will be provided. So, the goal of occupancy detection will be provided as reducing inefficient energy consumption. The studies [2-4] showed that when appropriate and accurate occupancy detection system is employed energy saving ratio can be achieved between 30% and 42%. The authors of the study [5] claimed that they reduced energy consumption by saving 37% ratio of energy in commercial building by occupancy detection. It
is claimed in the study [6] that between 29% and 80% energy consumption reduction was achieved when the occupancy related insights were used as an input for Heating Ventilation Air Conditioning (HVAC) control system. These amounts of energy saving show how important to study in the field of occupancy detection is. The contribution of this paper to this field is to investigate the successes of the different classification methods for occupancy detection on specific benchmark dataset that is obtained from UCI Machine Learning Repository, 2016 [7]. The results of this study were only compared with the previous studies whose methods were applied on the same benchmark dataset to be consisted.

The rest of the paper is organized as follows. In the next (Section 2), a literature survey that only includes the previous studies that applied on the same benchmark Occupancy dataset is presented in the table to be more readable and comparable. In Section 3, the Occupancy dataset is introduced and then information about 3 Machine Learning Based Classification Methods, namely Random Forest (RF), Decision Tree (DT) and Begging is presented. In section 4, the results of the applications of applied methods are presented, and also suggestions for future study are presented.

2. Theoretical background
In order to be consistent and comply with rationality only the studies whose methods were applied on the same benchmark dataset is listed in this section as table. Since Occupancy by UCI Machine Learning Repository was delivered in 2016 [6,13], the studies belongs to year 2016 and later were listed in Table 1 to be more clear and compact. As can be seen from the studies in table 1, the highest accuracy ratio is obtained by Yanyong Huang et al. [10] as 99.74% using Fuzzy Granulation method.

Table 1. Literature survey on the benchmark dataset

| Method                                  | Accuracy Ratio (%) | References                          |
|-----------------------------------------|--------------------|-------------------------------------|
| Fuzzy Granulation (Fgf)                 | 99.740             | Yanyong Huang et al., 2018 [10]     |
| Artificial Neural Network (ANN)         | 99.061             | Ozcan et al., 2017 [12]             |
| SVM with Standard Particle Swarm Optimisation (SVM with SPSO) | 98.980             | Shinichi Yamada et al., 2017 [9]   |
| FS-KC                                   | 98.840             | Shinichi Yamada et al., 2017 [11]  |
| Ultra-fast Optimization Multiple Kernel Learning (MKL 2) | 98.670             | Shinichi Yamada et al., 2017 [11]  |
| Kernel Construction (KC)                | 98.580             | Shinichi Yamada et al., 2017 [11]  |
| Lp-Norm Multiple Kernel Learning (MKL 1) | 98.080             | Shinichi Yamada et al., 2017 [11]  |
| SVM                                     | 98.050             | Shinichi Yamada et al., 2017 [11]  |
| Feature Selection (FS)                  | 97.960             | Shinichi Yamada et al., 2017 [11]  |
| K-Nearest Neighbour (KNN)               | 97.655             | Sachin Kumar et al., 2018 [8]      |
| Logistic Regression (LR)                | 97.029             | Sachin Kumar et al., 2018 [8]      |
| Gradient boosting machine               | 96.807             | Sachin Kumar et al., 2018 [8]      |
| Naive Bayes (NB)                        | 96.402             | Sachin Kumar et al., 2018 [8]      |
| Linear Discriminate Analysis (LDA)      | 96.156             | Sachin Kumar et al., 2018 [8]      |
| Classification and regression tree      | 92.495             | Sachin Kumar et al., 2018 [8]      |
| Support vector machine (SVM)            | 78.770             | Sachin Kumar et al., 2018 [8]      |

3. Materials and methods
3.1. Dataset description
Occupancy data set is received from UCI machine learning repository [7, 13]. The dataset contains input feature space with attributes such as temperature, humidity, light and CO2, and it consists of 20,560 samples.

3.2. Classification methods
The classifier methods used in this study were chosen such a way that they haven’t been used in literature for Occupancy benchmark dataset. Followings are definition and information about the classifier techniques of this study.

**Random Forest (RF):** Breiman (2001) A random forest is a classifier consisting of a collection of tree structured classifiers \( \{ h(x, \Theta_k), k=1, \ldots \} \) where the \( \{ \Theta_k \} \) are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input \( x \) [14].

**Decision Tree (DT):** The decision tree classifier is one of the possible approaches to multistage decision making; table look-up rules, decision table conversion to optimal decision trees and sequential approaches are others. The basic idea involved in any multistage approach is to break up a complex decision into a union of several simpler decisions, hoping the final solution obtained this way would resemble the intended desired solution [15].

**Bagging:** Bagging predictors is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets [16].

### 3.3. Commonly-accepted performance evaluation measures

The correctness of a classification can be evaluated by computing the number of correctly recognized class examples (true positives), the number of correctly recognized examples that do not belong to the class (true negatives), and examples that either were incorrectly assigned to the class (false positives) or that were not recognized as class examples (false negatives). These four counts constitute a confusion matrix shown in figure 1 for the case of the binary classification. [17].

TP, FP, TN and FN are first calculated depending on the obtained result. Then, Classification Accuracy, Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value and ROC Area are found by using TP, FP, TN and FN values. Followings are the definitions of related terms and Table 2 shows how to calculate these terms.

**Figure 1.** Confusion matrix for binary classification.

**Classification Accuracy:** It refers to the total number of records that are correctly classified by the classifier. Accuracy of a classifier is defined as the percentage of test set tuples that are correctly classified by the model.

**Sensitivity:** Refers to the true positive rate that means the proportion of positive tuples that were correctly identified.

**Specificity:** Refers the rate at which a test or diagnostic method sets a correct (ie negative) diagnosis for a patient who is not ill.

**Positive Predictive Value:** The fraction of retrieved instances that are relevant.
**Negative Predictive Value**: The result that the modeling makes a negative prediction.

**Area Under Curve (AUC)**: Classifier’s ability to avoid false classification.

**Table 2.** Performance values measures for classifiers

| Performance Evaluation Measures          | Formula |
|------------------------------------------|---------|
| Classification Accuracy (%)              | \( \frac{TP + TN}{TP + FP + TN + FN} \times 100 \) |
| Sensitivity (%)                          | \( \frac{TP}{TP + FN} \times 100 \) |
| Specificity (%)                          | \( \frac{TN}{FP + TN} \times 100 \) |
| Positive Predictive Value (%)            | \( \frac{TP}{TP + TN} \times 100 \) |
| Negative Predictive Value (%)            | \( \frac{TP}{TN + FN} \times 100 \) |
| Area Under Curve (AUC)                   | \( \frac{1}{2} \left[ \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right] \times 100 \) |

4. Results and discussions

In order to apply RF, DT and Begging classification techniques; first Occupation dataset was normalized, then 10-layer cross folder was applied on normalized dataset and finally related classification algorithms were applied on the dataset. To evaluate the performance measures of classifiers the formula presented in table 2 were used.

The “normalization” of the data plays very important role with respect to not corrupting the relationship between the variables, the accuracy of the analysis and the network performance. Normalization is a squeezing process between the upper and lower bounds of the activation function used for each data in the dataset. The activation function used in the analysis is "sigmoid" and the data are squeezed to the range [0,1]. Before application of classifiers Occupancy dataset is normalized by using equation (1).

\[
x_i = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \quad (1)
\]

For 10-layer cross-validation method the given dataset is divided into 10 parts as shown in figure 2 [18]. Any method that needs to be trained and tested will use \( k \) different training and test sets (where \( k \)=number of data/10), and still achieves \( k \) success measures accordingly. For this reason, to determine the success of cross-validation, the arithmetic mean of the obtained \( k \) performance measures is taken [19, 20]. For occupancy dataset that consists of 20560 samples 10-fold cross-validation is applied by dividing it into the groups; where each group has 2056 elements. Since there are 10 groups, 10 different test results are formed. The average success of these tests will be the overall success, that is, the success of the method to be applied.

After applying 10-fold-cross validation and obtaining the results for RF, DT and Bagging methods the confusion matrix for three classifiers shown in Table 3 is obtained.

**Table 3.** Confusion matrix obtained from Random Forest, Decision Tree and Bagging classifiers

| Actual | RF Predicted | DT Predicted | Bagging Predicted |
|--------|--------------|--------------|-------------------|
|        | P            | N            | P                 | N                 | P                 | N                 |
| P      | 4703 (TP)    | 47 (FN)      | 4696 (TP)         | 54 (FN)           | 4709 (TP)         | 41 (FN)           |
| N      | 83 (FP)      | 15727 (TN)   | 106 (FP)          | 15704 (TN)        | 122 (FP)          | 15688 (TN)        |

4
Figure 2. 10-Fold cross-validation method.

By using confusion matrix presented in table 3 and the formulas presented in table 2 performance evaluation measures of RF, DT and Bagging were calculated and presented in table 4.

Table 4. Performance evaluation measures for RF, DT and Bagging classifiers

| Performance Evaluation Measures | Random Forest | Decision Tree | Bagging |
|---------------------------------|---------------|---------------|---------|
| Classification Accuracy (%)     | 99,368        | 99,222        | 99,207  |
| Sensitivity (%)                 | 99,011        | 98,863        | 99,137  |
| Specificity (%)                 | 99,475        | 99,330        | 99,228  |
| Positive Predictive Value (%)   | 98,266        | 97,793        | 97,475  |
| Negative Predictive Value (%)   | 99,702        | 99,657        | 99,739  |
| Area Under Curve (AUC)          | 0,992         | 0,991         | 0,992   |

5. Conclusion

As it’s mentioned before Occupancy dataset by UCI was used to be benchmark dataset. It should be noted that this dataset has allows user to have binary classification since the output variables called “occupancy” and has a value of either 1 or 0. Additionally this dataset was published in 2016 by UCI and only the studies from 2016 and then were listed in literature survey and compared with this study. When the RF, DT and Bagging were chosen for Machine Learning Based classification methods two principals were carefully followed. First one is to have classifier methods that were not use in literature for benchmark dataset and the second one almost never usage of tree based classifiers in the literature. Thus, the author of this paper would like see the success and/or failure of tree based classifiers. It has been seen that the highest accuracy or prediction ratio was obtained as 99,368% by Random Forest. So, it is the second best accuracy ratio after Fuzzy Granulation (Fgf) method among other 16 methods in the literature that we can reach. Additionally. DT and Bagging classifier showed higher accuracy than other 15 methods in the studies listed in literature section. Thus this study showed how decision tree can be promising for occupancy detection.

6. References
[1.] Lam K P, et al 2009 occupancy detection through an extensive environmental sensor network in an open-plan office building Proc. 11th Intl. IBPSA Conf. (Glasgow, Scotland)
[2.] Erickson V L, et al 2011 OBSERVE: Occupancy-based system for efficient reduction of HVAC energy Proc. 10th Intl. Conf. on Information Processing in Sensor Networks (Chicago, USA) pp 258–269
[3.] Erickson V L, et al 2014 Occupancy modeling and prediction for building energy management ACM Trans. Sensor Networks 10(3)
[4.] Dong B and Andrews B 2009 Sensor-based occupancy behavioral pattern recognition for energy and comfort management in intelligent buildings Proc. 11th Intl. Building Performance Simulation Association Conference
[5.] Brooks J, et al 2014 An experimental investigation of occupancy-based energy-efficient control of commercial building indoor climate Proc. IEEE 53rd Annual Conf. on Decision and Control (Los Angeles, USA) pp 5680–5685
[6.] Brooks J, et al 2015 Energy-efficient control of under-actuated HVAC zones in commercial buildings Energy Build 93 pp 160–168
[7.] Lichman M and Bache K 1998 UCI machine learning repository (University of California, USA) URL: https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+
[8.] Kumar S et al 2018 Machine Learning-Based Method and Its Performance Analysis for Occupancy Detection in Indoor Environment Proc. Intl. Symp. on Signal Processing and Intelligent Recognition Systems (Bangalore, India) pp 240-251
[9.] Yamada S and Neshatian K 2017 Hyper-parameter Search in Support Vector Machines using PSO with Cellular Fitness Approximation Proc. 2017 IEEE Symposium Series on Computational Intelligence (Honolulu, USA) DOI: 10.1109/SSCI.2017.8285406 (978-1-5386-2726-6/17/$31.00)
[10.] Huang Y, et al 2018 Dynamic Fusion of Multi-source Interval-valued Data by Fuzzy Granulation IEEE Transactions on Fuzzy Systems DOI 10.1109/TFUZZ.2018.2832608
[11.] Yamada S and Neshatian K 2017 Kernel Construction and Feature Subset Selection in Support Vector Machines Proc. Asia-Pacific Conf. on Simulated Evolution and Learning (Shenzhen, China) pp 605–616 https://doi.org/10.1007/978-3-319-68759-9_49
[12.] Cataltas O, Tutuncu K and Koklu M 2017 Occupancy detection through light, temperature, humidity and CO2 sensors using ANN Intl. Journal of Industrial Electronics and Electrical Engineering 5(2) pp 63–67
[13.] Candanedo L M and Feldheim V 2016 Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models Energy and Buildings 112 pp 28-39
[14.] Breiman L 2001 Random Forests Machine Learning 45(1) pp 5-32
[15.] Safavian S R and Landgrebe D 1991 A survey of decision tree classifier methodology IEEE transactions on systems, man, and cybernetics 21(3) pp 660-674
[16.] Breiman L 1996 Bagging predictors Machine learning 24(2) pp 123-140
[17.] Sokolova M and Lapalme G 2009 A systematic analysis of performance measures for classification tasks Information Processing & Management 45(4) pp 427-437
[18.] Randell T 2018 Classification and Prediction (Data Mining: Concepts and Techniques) URL: http://slideplayer.com/slide/8148400/
[19.] Golub G H, Heath M and Wahba G 1979 Generalized cross-validation as a method for choosing a good ridge parameter Technometrics 21(2) pp 215-223
[20.] Shao J 1993 Linear model selection by cross-validation Journal of the American statistical Association 88(422) pp 486-494

Acknowledgments
This study has been supported by Scientific Research Project of Selcuk University.

Authors’ background

| Name            | Prefix       | Research Field                  | Email                |
|-----------------|--------------|---------------------------------|----------------------|
| Murat KOKLU     | Assist. Prof. Dr. | Artificial intelligence and applications | mkoklu@selcuk.edu.tr |
| Kemal TUTUNCU   | Assist. Prof. Dr. | Artificial intelligence and applications | ktutuncu@selcuk.edu.tr |