Occupancy Detection using Machine Learning

Sara Ranjit
UG Student, Division of Computer Science, School of Engineering,
Cochin University of Science and Technology, Kerala, India

Abstract—Occupancy detection is used to detect the presence of a person in a room. The aim is to successfully and correctly predict whether the room is occupied or not. This can be used to control lighting and can lead to energy and cost savings. The analysis has been performed using machine learning in Python. Occupancy Detection Data Set from the UCI Machine Learning Repository has been referred for analysis. Seven classifiers are considered here. The models for these are built and evaluated and predictions are made using the best one with largest accuracy. Here, Random Forest model has the largest accuracy. Its confusion matrix and classification report has also been generated.

Keywords—*occupancy detection, machine learning classifiers, classification.*

I INTRODUCTION

Motion sensors are used to control electric lighting in rooms. The room need not be lit if no motion is detected because the space will be empty. Under such circumstances, turning off the lights can save considerable amounts of energy and the electricity bill also comes down. Significant energy savings can also be achieved by operating ventilation, heating and air conditioning controllers after obtaining a feedback from sensor-based occupancy detection techniques. Hence, energy management in commercial buildings is made possible by accurate occupancy detection. Numerous sensors are usually deployed to gather occupancy details.

II. LIERATURE SURVEY

Occupancy modeling is evaluated using twelve ambient sensor variables [1]. Evaluation is done in both single-occupancy and multi-occupancy offices using six machine learning algorithms and the decision-tree technique yielded the best overall accuracy (i.e. 96.0% to 98.2%). Symmetrical uncertainty analysis was used for feature selection [2]. Selected multi-sensory features were combined using a neural network. Estimation accuracy reaching up to 75% was obtained for occupied periods. It was shown that 42% annual energy savings could be obtained using strategies based on sensor network occupancy model predictions [3]. In [4], symmetrical uncertainty analysis was used to carry out feature selection and back-propagation neural network was used for fusion of sensor features and with occupant count accuracy exceeded 74%. Occupancy was estimated with accuracies of 85% and 83% in the two testing sets using LDA model [5]. In [6], when the results were compared after applying various Artificial Neural Network algorithms to the dataset, it was observed that the highest accuracy rate of 99.061% was obtained with Limited

Memory Quasi-Newton algorithm. Occupancy detection accuracy of 97.16 % was obtained using the sensors of volatile organic compounds (VOCs) [7]. In [8], models were presented to conduct a situation-centric profiling using context sources that are commonly available in commercial buildings. The potential for detecting occupancies with accuracy as high as 90% was exhibited.

III. DATASET

Occupancy Detection Data Set from the UCI Machine Learning Repository has been referred for analysis [9]. 20560 instances and 6 attributes have been taken to perform this analysis. The attributes are
Temperature in Celsius
Relative Humidity in %
Light in Lux
CO2 in ppm
Humidity Ratio, Derived quantity from temperature and relative humidity, in kg water-vapor per kg-air
Occupancy, 0 or 1
The final attribute Occupancy takes two values which are 0 if the room is not occupied and 1 if the room is occupied. This is a classification problem and all attributes used here are numeric.

IV. CLASSIFIERS

The seven classifiers considered in this paper are [10]-
K-Nearest Neighbours (KNN)
Classification and Regression Trees (CART)
Random Forest (RF)
Gaussian Naive Bayes (NB)
Support Vector Machines (SVM)
Logistic Regression (LR)
Linear Discriminant Analysis (LDA)

K-Nearest Neighbors (KNN)

KNN is a non-parametric technique that falls in the supervised learning family of algorithms. It has been successful in a large number of classification and regression problems.

Classification and Regression Trees (CART)

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Random Forest(RF)

Random forest or is an ensemble learning method for classification, regression and other tasks that create a set of decision trees at training time and decides the final class after obtaining the results from decision trees.

Gaussian Naive Bayes (NB)

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes theorem which assume that all the features are independent of each other. GaussianNB is specifically used when the features have continuous values. It is also assumed that all the features follow a gaussian distribution.

Support Vector Machines (SVM)

Support vector machines (SVMs) are a set of supervised learning methods used for multi-classification, regression and outliers' detection. It can be used as a discriminative classifier formally defined by a separating hyperplane.

Logistic Regression (LR)

Logistic Regression is a statistical technique to analyze a dataset in which there are one or more independent variables that determine an outcome. It is used for binary classification where there are only two possible outcomes.

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is a linear classification technique that uses Bayes Theorem. It can be used to perform supervised dimensionality reduction before later classification.

V. METHODOLGY

The analysis has been completed using machine learning in Python. The aim is to create seven machine learning models and to find the best one. All the modules, functions and objects to be used have to be imported. The dataset has to be loaded. Then, analyse the statistical summary of the dataset. Take a look at the number of instances that belong to Occupancy – 0 (not occupied) and Occupancy – 1 (occupied). Also try to find whether any relationship exists between the attributes. Now, some instances of the dataset have to be selected for validation. The algorithms will not be able to view the instances used for validation. These validation instances are used to test the accuracy of the models. The dataset is split such that 80% is used for training and 20% is used for testing. Use k-fold cross validation to evaluate predictive models [11]. The original dataset is randomly split into k-parts of equal sizes. Out of the k parts, 1 part is used to perform the final validation testing and the other parts are used for training. This is repeated k times. Here, seven algorithms have been chosen. Find the accuracy estimations for each. Build and evaluate the models. Evaluation is done based on accuracy [12]. Accuracy is number of correctly predicted instances divided by the total number of instances given in the dataset multiplied by 100 to get a percentage. The evaluation results can also be compared using graphs. Compare the models and find the model with the highest accuracy. Finally, make predictions using that model and obtain the confusion matrix and classification report.

VI. RESULTS AND ANALYSIS:

The following are the results and its analysis inferred:-

TABLE I shows that there are 15810 instances with Occupancy – 0 (unoccupied) and 4750 instances with Occupancy – 1 (occupied) in our dataset.

TABLE I. NUMBER OF INSTANCES

Occupancy	Instances
0	15810
1	4750

TABLE II gives the statistical information- count, mean, standard deviation, minimum and maximum of each attribute computed using Python.

TABLE II. STATISTICAL INFORMATION OF ATTRIBUTES

Measures	Temperature	Humidity	Light	CO2	Humidity Ratio	Occupancy
count	20560.000000	20560.000000	20560.000000	20560.000000	20560.000000	20560.000000
mean	20.906212	27.655925	130.756622	690.553276	0.004228	0.231031
std	1.055315	4.982154	210.430875	311.201281	0.000768	0.421503
min	19.000000	16.745000	0.000000	412.750000	0.002674	0.000000

max	24.408333	39.500000	1697.250000	2076.500000	0.006476	1.000000

Fig. I gives an idea of distribution of values of each attribute of the dataset. It is a univariate plot.

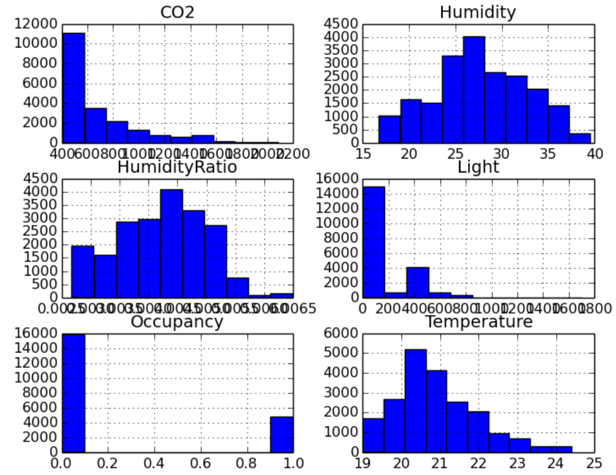

Fig. I. HISTOGRAM OF C02, HUMIDITY, HUMIDITY RATIO, LIGHT, OCCUPANCY, TEMPERATURE

Fig. II shows the relationship between the attributes. It is a multivariate plot.

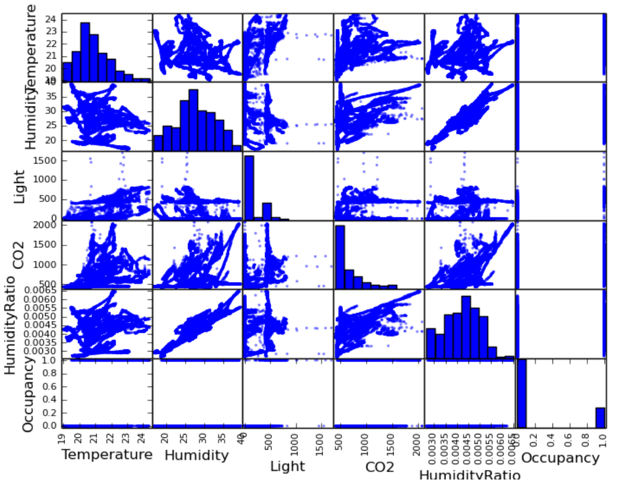

Fig. II. SCATTER PLOT OF DEPENDENCIES AMONG ATTRIBUTES

Fig. III gives the line graph comparing different classifiers after the seven models were evaluated.

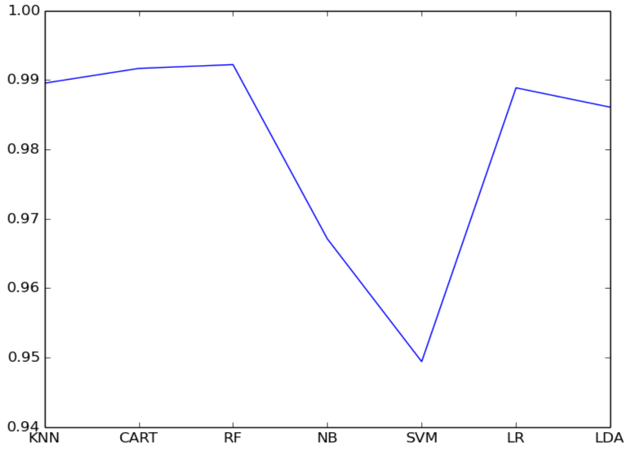

Fig. III. COMPARISON OF KNN, CART, RF, NB, SVM, LR, LDA

TABLE III gives the obtained accuracy of seven classifier models after evaluation.

TABLE III. ACCURACY OF SEVEN CLASSIFIER MODELS

Classifier	Accuracy (%)
KNN	98.9543
CART	99.1671
RF	99.2218
NB	96.7109
SVM	94.9417
LR	98.8874
LDA	98.6077

From the TABLE III and the line graph in Fig. III, it can be observed that Random Forest has the greatest accuracy of 99.2218% after k-fold cross evaluation has been applied to all algorithms. Perform validation on Random Forest model separately using the validation instances to make predictions. After this, confusion matrix, final accuracy score and a classification report were obtained.

TABLE IV gives the obtained confusion matrix for Random Forest model.

TABLE IV. CONFUSION MATRIX FOR RANDOM FOREST MODEL

n=4112	0	1
0	3153	18
1	13	928

TABLE IV indicates that 31 instances were wrongly predicted. 18 instances were wrongly predicted as 1 (occupied) and 13 instances were wrongly predicted as 0 (not occupied). 3153 instances were correctly predicted as not occupied and 928 instances were correctly predicted as occupied.

Accuracy is given by the number of correctly predicted instances divided by the total number of instances multiplied by 100. From the confusion matrix, accuracy is sum of numbers in the diagonal divided by the total sum of numbers in the matrix multiplied by 100. Hence, the final accuracy is 4081/4112*100= 99.24610894%

TABLE V gives the obtained classification report of Random Forest model.

Precision is the accuracy of positive predictions. Precision = TP/(TP + FP) where TP is True Positives and FP is False Positives.

FN – False Negatives

Recall is the fraction of correctly identified positives.

Recall = TP/(TP + FN) where TP is True Positives and FN is False Negatives.

F1 Score is a metric for comparing two classifiers. It is obtained by finding the the harmonic mean of precision and recall.

The support is the number of occurrences of each class.

TABLE V. CLASSIFICATION REPORT OF RANDOM FOREST MODEL

occupancy	precision	recall	f1-score	support
0.0	1.0	0.99	1.0	3171
1.0	0.98	0.99	0.98	941
avg/ total	0.99	0.99	0.99	4112

VII. CONCLUSION

Occupancy Detection Data Set from the UCI Machine Learning Repository has been referred for analysis. Seven classifiers were taken for analysis. The models for these were built and evaluated. Here, Random Forest model has the largest accuracy of 99.2218% after k-fold cross validation. Its confusion matrix and classification report has also been generated. Validation was performed on Random Forest model separately using the validation instances to find its independent accuracy. The final accuracy was obtained as 99.24610894%.

VIII. REFERENCES

[1] Z. Yang, N. Li, B. Becerik-Gerber, M. Orosz, A systematic approach to occupancy modeling in ambient sensor-rich buildings, Simulation vol. 90, Issue 8, pp. 960–977, August 2014.

[2] Real-time building occupancy sensing using neural-network based sensor network, Tobore Ekwevugbe, Neil Brown, Vijay Pakka, Denis Fan, published in Digital Ecosystems and Technologies (DEST), 7th IEEE International Conference on 24-26 July 2013.

[3] OBSERVE: Occupancy-based system for efficient reduction of HVAC energy, Varick L. Erickson, Miguel Á. Carreira-Perpiñán, Alberto Cerpa, published in Proceedings of the 10th ACM/IEEE International, 2011.

[4] Advanced occupancy sensing for energy efficiency in office buildings,Tobore Ekwevugbe, Neil Brown, Vijayanarasimha Pakka, Denis Fan, published in Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering.

[5] Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. Luis M. Candanedo, VÃ©ronique Feldheim. Energy and Buildings, vol. 112, pp. 28-39, 15 January 2016.

[6] Kemal Tutuncu, Ozcan Cataltas, Murat Koklu, Kemal Tutuncu, Ozcan Cataltas, Murat Koklu "Occupancy Detection Through Light, Temperature, Humidity and Co2 Sensors Using Ann" , International Journal of Industrial Electronics and Electrical Engineering , vol. 5, Issue 2, February 2017.

[7] Occupancy Detection using Gas Sensors, Andrzej Szczurek, Monika Maciejewska and Tomasz Pietrucha, In Proceedings of the 6th International Conference on Sensor Networks - vol 1: SENSORNETS, pp. 99-107, 2017, Porto, Portugal.

[8] Occupancy detection in commercial buildings using opportunistic context sources, Sunil Kumar Ghai ; Lakshmi V Thanayankizil ; Deva P. Seetharam ; Dipanjan Chakraborty, published in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on 19-23 March 2012.

[9] Source Luis Candanedo, UMONS, https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+

[10] The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) Hardcover – 19 Apr 2017 by Trevor Hastie, Robert Tibshirani, Jerome Friedman.

[11] http://scikit-learn.org/stable/modules/cross_validation.html

[12] https://developers.google.com/machine-learning/crash-course/classification/accuracy