ACTA SCIENTIFIC COMPUTER SCIENCES

Volume 4 Issue 8 August 2022

Research Article

A Machine Learning Approach for Occupancy Detection in Smart Building

Shashi Shekhar Kumar[1] and Upendra Pratap Singh[2]*

[1]Department of Information Technology, IIIT Allahabad, Prayagraj, India

[2]Department of Computer Science and Engineering, SOA University, Bhubaneshwar, India

***Corresponding Author:** Upendra Pratap Singh, Department of Computer Science and Engineering, SOA University, Bhubaneshwar, India.

Received: May 16, 2022

Published: July 18, 2022

Abstract

The IoT has emerged as one of the fastest-growing areas of research. Community of IoT is exploring this area of study to enhance innovative building features for reliable comfort in every prospect. One of the intelligent building features is occupancy detection, which can detect the Occupancy of a particular room or premises in real-time. Different IoT sensors implemented in an intelligent building can generate enormous data. For simulation purposes, we have used python along with apache spark framework, which will analyze the data generated and consequently, we can realize the other features like humidity, Co_2 level of the room. The key contribution of this paper is to enhance the accuracy of the machine learning model for better occupancy prediction. With machine learning, the approach building feature can adjust automatically.

Keywords: Smart Building; Big Data; Occupancy Detection; Cyber-Physical System; Internet of Things

Introduction

According to Cisco's official data in 2020, the no of IoT connected devices ratio is 6.58 per person, and it is increasing day by day. Keeping tracking of the data generated by each device is still an uphill task in itself. Human life can be considerably improved using IoT devices. One of the domains of IoT is smart building [23,26]. Various sensors have been attached to building premises to continuously monitor the data generated by sensors. The collected data is further stored on the server for analysis and transformation [22,24,25].

Andrea and Fontana [1] introduced an approach using iBeacon for smart android devices to detect Occupancy in a smart building. Candanedo, Luis M and Feldheim, Ve´ronique [2] used the statistical model to detect the Occupancy using the environment measures that include lights, temperature, carbon dioxide, and humidity. Q. Hua and H. Chen [3] used support vector regression to detect the occupancy detection. Jeon, Yunwan [4] used a point extraction algorithm to predict the Occupancy of a building. They used the pm pattern of occupants' activities to find the high and low to demonstrate the algorithm. Elkhoukhi, H and NaitMalek [5] combined the IoT sensor and big data platform to detect the Occupancy in real-time. They proposed the approach by comparing the dynamic and static machine learning techniques. Zhao, Hengyang and Hua [6] presented a thermal-based sensor that used support vector regression to predict the occupancy detection in a smart building. Hailemariam, Ebenezer and Goldstein [7] proposed an approach using a decision tree and multiple environmental sensors to predict the Occupancy of building premises. Saha, Homagni [8] proposed a data analytics approach and mathematical tools to predict the Occupancy of a smart building. they used the performance matrics and benchmarking system for accurate prediction.

K. Akkaya and I. Guvenc [9] proposed a method using counting the no of occupants present inside a building or premises. They used different actuators and smart sensors to predict the Occupancy. Ji, Youngmin and Ok [10] proposed a method named occupancy detection algorithm to detect the Occupancy. They used Carbon Dioxide concentration (CO_2) level along with a passive infrared sensor to increase the accuracy level of Occupancy. Weiming Shen and Guy Newsham [11] proposed an approach to detect the Occupancy using Bluetooth technology and smartphone with an approach of detecting the Occupancy in a specified region.

M. Azam and M. Blayo [12] introduced a framework of wifi motion detection sensing along with a Random Forest and Decision tree machine learning approach to detect the most stable accuracy. Y. Gao and A. Schay and D. Hou [13] introduced an approach of usage pattern analysis along with appliances correlation analysis to trace the power consumption and predict accurate Occupancy. C. Feng and A. Mehmani and J. Zhang [14] used a deep learning model to estimate the accurate Occupancy. The model used by them was a convolution neural network and an extended short-term network. T. Ekwevugbe and N.Brown and V. Pakka and D. Fan [15] proposed a methodology of two models, filter and wrapper model, respectively. The filter model works on general characteristics without any learning algorithm. On the contrary, the wrapper model work for predictive accuracy of a predetermined learning algorithm for feature selection. A. Parise and A. Manso Callejo [16] used a support vector prediction model and IoT based cloud infrastructure to predict the accuracy of occupancy estimation. M.R. Bashir and A.Q. Gill [17] implemented a smart building prototype using a big data platform and IoT based data.

Motivation

As India is moving towards the digital age, the smart building is one of the prominent areas to be explored. In the coming years, most of the buildings will be fully automated in terms of every application of the smart building. As we all know, Smart building generates a lot of IoT based data, and It becomes really challenging to analyze the data. All the literature cited above has been focused on a methodology based on static data and a very small set of data. The motivation behind this paper is to detect the Occupancy of smart buildings with a machine learning approach. We have used linear regression to implement a huge size of the dataset.

Organization

The rest of the paper has been organized follows as. Section II shows the motivation to write up the paper. Section III represents the experimental setup and specification of the hardware and software. In section IV, we have explained the proposed methodology. Section V represents the performance evaluation. Section VI discusses the results obtained and Section VII concludes the work.

Research contribution of the work

Based on the information explained above, the following contribution has been presented in this paper.

- A Spark based framework is used for predicting the room occupancy using a Logistic regression classifier.
- Data Acquisition proceeds using IoT sensors deployed at different locations in the building; the data streams collected from these sensors are fed to the underlying machine learning classifier for its training and subsequent inference.

IOT sensor and data collection

Sensor node

The sensor can collect the data it depends on the user's requirement and how we can use it. We have used two different types of sensors to collect the data in real-time. The sensors are followed as below.

- **Passive infrared sensor:** This sensor is used to detect indoor motion and activity. It works on infrared radiation, which varies from the temperature and surface characteristics object in front of the sensor. It senses the movement of anything which comes in front of the sensor. The PIR sensor works on the infrared radiation generated when a movement occurs in the sensor zone.
- **Air quality sensor:** The sensor is used to detect CO_2 level of the room. The Sensor measures gaseous carbon dioxide based on the concentration of CO_2 in the air, by detecting the quantity of infrared radiation absorbed by carbon dioxide molecules.

Data collection

Data collection is an important aspect of any proposed experimental implementation. Without useful data, we can't propose any research. We need to be cautious while collecting the data from sensors. For collecting data through sensors, we have used raspberry Pi and Arduino. The captured data is streamed

Figure 1: Different sensors.

through a TCP connection and injected into the cloud server using the apache flume agent. The collected data need to be preprocessed before it is used because the captured data may contain noise or vague or empty data which can cause concern while implementing the model.

Prototype architecture

In this section, we have shown a structure for the collection of data. The sensors can capture movement. With the help of raspberry, pi data is stored in database storage in the form of excel format. After storage, the data preprocessing has been done to combat the null values in the datasets.

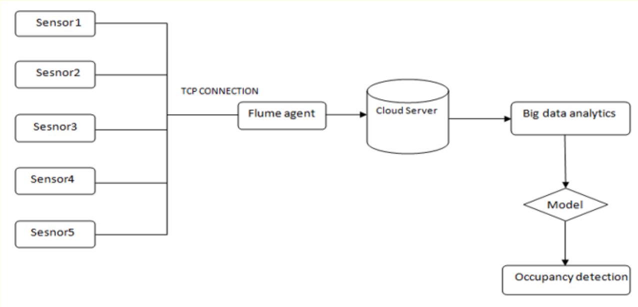

Figure 2: Architecture of the proposed work.

Experimental set up

In this section, we have presented a detailed description of the proposed method to detect Occupancy. Before proceeding with the method proposed, we will explain some of the points that will help us understand it better.

Machine learning

Machine learning is an approach to designing a statistical model to solve a specific problem without being explicitly programmed. Although machine learning may apply different techniques, it can be categorized into three types.

- **Supervised learning:** Supervised learning algorithm is an approach to learning by example. The term supervised emphasizes training the model based on the given class. While some of the training is matched with the correct output during the training, the algorithm will search for the correct pattern and correlate it with the desired output. After training a supervised learning algorithm, we will feed the remaining input data to classify the data. Supervised learning is where we have an input variable x and output variable y, and we use a mapping function to map both values. The goal is to approximate the mapping function when we have a new input data (x) that can predict output variable (y).
- **Unsupervised learning:** The algorithm in which only data is given, and there is no corresponding value is available. Unlike supervised learning, here, there is no predefined labelled data. The algorithm is left to its own devices to discover and present its own structure in data. Unsupervised learning can be further classified into clustering and Association.

Logistic regression

Among numerous machine learning algorithms available, we have chosen to use this statistical model to implement on data because it suited most according to our requirement and available sample and contains the highest accuracy among other models. This machine-learning algorithm is based on a classification-based problem's predictive analysis technique. The model is widely used in statistics to estimate the probability value based on some given input data and binary data. In logistic regression, the dependent variable is in the definite form of either 0 or 1 and depending upon the attribute, we get a probability of either ”yes” or ”No”. We can assess the logistic regression model using a confusion matrix which shows the no of a correct and incorrect prediction made by comparing the classification model and the actual outcome from sample data. The correct and incorrect predictions are summarized with count values and broken by each class. The matrix does not only give errors being made by the classifier but types of errors that are made. It allows visualization of the performance of the algorithm. The logistic regression can be defined by the following given equation. Where e is the numerical Euler's constant, and x is the input given to the logistic function. The accuracy of the model can be defined as the following expression.

$$Ac = \frac{correct_{prediction}}{\times \; number\ of_{correct\ of\ prediction}} \quad 100 \qquad (1)$$

Streaming

Big data ingest the data that is added to the queue for seamless integration and transformation of data in real-time. Big data streaming is ideally a process that processes a stream of continuous data. Real-time streaming is a single pass analysis because once the data stream is passed, it cant be reanalyzed again. The various platforms leverage many features to stream the data in real-time. Apache spark is a big open-source data processing framework used nowadays because of its various unmatched features and in-memory computing facility. Spark runs as a standalone or on the top of yarn, where data can be directly read by the Hadoop distributed file system. Apart from these features, it has various machine learning libraries and graph processing.

Proposed work

In this paper, we have proposed a method to detect Occupancy in a smart building. For this purpose, we have five different sensors to collect the data from the different sensors. The sensor is deployed on the building premises to monitor and fetch the data collected from the sensor.

Data ingestion

We are collecting data from five different sensors that is CO_2 sensor, temperature, electricity consumption, lightning sensor. These sensors are physically attached inside the closed premises of a building. We assume that data captured from different sensors are in a structured form. For data ingestion, we have used apache flume, which can efficiently collect from external sources and move a large amount of log data to the server.

Buffer

We have used a buffer during data collection from sensors because there might be the chances of any interruption from the detector while streaming data from the sensor. Buffer is programmed and has a functionality of boolean to check the status. Suppose the status of the error log is true. There might be an error in streaming data. The error log is false, saying that data is streaming without any mistake. The buffer's functionality also ensures that streaming data will only be stored in the cloud in case the error log is false for any particular sensor.

$$\gamma = \begin{cases} 1, & \text{if error in streaming.} \\ 0, & \text{No error} \end{cases} \qquad (2)$$

Cloud storage

The cloud storage stores the data streaming from a sensor that is already in a structured form. The data will be moved to cloud storage only after checking the error log status. The big data can directly access the data stored in the cloud and further processed. We are not using any preprocessing of data assuming that data is in a structured format. With cloud storage, the spark framework can directly access the streaming data to process the data in real-time.

Error log

An error log has been defined to report the error in the streaming data from sensors. The overall error rate is calculated using no sample coming from the sensor to no sample data received to cloud storage, excluding the error log.

$$\rho = \frac{\alpha - \beta}{Sample_{received}} \times 100 \qquad (3)$$

With the above equation, we can predict the error rate in the streaming data.

Dataset

The total datasets collectively from various sensors will be processed for further evaluation to get the occupancy detection. The aim is to detect the Occupancy of an office room or building premises based on given attributes such as CO_2, temperature, lights, humidity and occupancy variable to detect the status of the room, either occupied or not. The various attribute has been illustrated below, at which we are going to predict the occupancy status.

- The CO_2 attribute inside an office shows the status and their rising level due to the presence of a human.
- Temperature attribute shows the presence of humans inside the room or office, and it varies from time to time.

- Lights attribute of dataset shows that electricity consumption due to humans inside the office.
- humidity level of the room can be significantly increased or decreased based on human presence.
- Occupancy variable 0 and 1 defines whether the room is occupied based on the given attribute.
- Training set is a part of sample data excluded from total sample data to train the model.
- Testing set is a part of a dataset other than the training dataset, which is used to test the data based on the trained model.
- Validation set is also a part of sample data which will be used to validate the model.

The following table shows the no of samples collected from the sensor.

S.NO	Name	No of sample
1	Training	8146
2	Testing	9580
3	Validation	2664

Table 1: Sample table.

Simulation set up

For simulation purpose we have used different tools which is illustrated below.

- Python V3.6 has been used to simulate the big data along with machine learning model.
- Spark V2.1.3 has been identified as a big data framework for analysis and transformation of the data received from the cloud server.
- Machine learning model has been used for prediction purpose and to achieve the standard accuracy. logistic regression has been used as it has highest accuracy among others.

Performance evaluation

The algorithm given below collects the data from different sensor and then passed to the machine learning model to find the prediction. The algorithm proposed below has been used for enhancement of accuracy for occupancy detection.

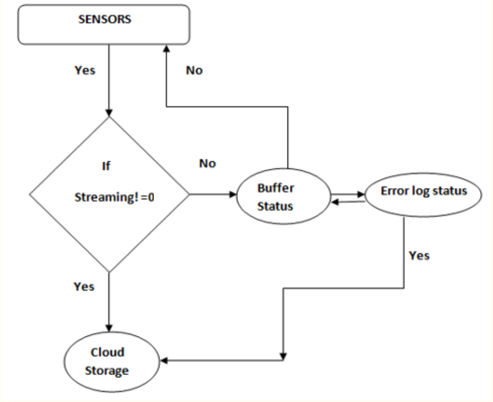

Figure 3: Flow Diagram.

Input: Sensor data from cloud server.

Output: Enhanced accuracy measure

- Preprocess the data
- While Data is Structured go to step 3
- Read the data through pandas library of python
- Apply the logistic regression model for accuracy enhancement
- Train the model with the training dataset
- Test the model with the testing sample
- Validate the dataset to find the accuracy of the model
- End of While
- Exit.

Algorithm 1: Pseudo code for the enhancement algorithm used in our work.

Result and Future Works

Smart building is one of the fastest-growing areas of research. As we are moving towards the next generation, our lifestyle has also been adapted to the extent of comfort level. The occupancy detection of the smart building is one of the functions which detects the Occupancy of the building premises or any particular area in real-time based on the environmental data collected through different sensors. We have used the spark framework and machine learning for realtime occupancy detection for the prediction. We have trained the model from the training data (part of sample data) and used 3 fold cross-validation of the logistic regression model. We have an accuracy rate of 99.29% and 99.56% for validation and testing samples.

Testing

Based on the trained model the testing sample has an accuracy of 99.56%.The following shows the ROC curve of testing sample.

Validation

The model has been validated with validation sample which has an accuracy of 99.29%. The results obtained above correspond to the experimental settings as detailed in Section IV; moreover, as some computations are also performed on the cloud including storage and consequent retrieval, the proposed system might not be scalable for larger networks. The scalability of the proposed methodology needs to be evaluated rigorously prior to its deployment in the real world setting.

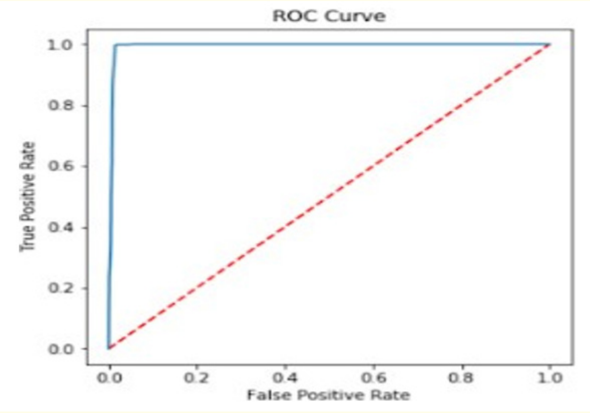

Figure 4: Testing.

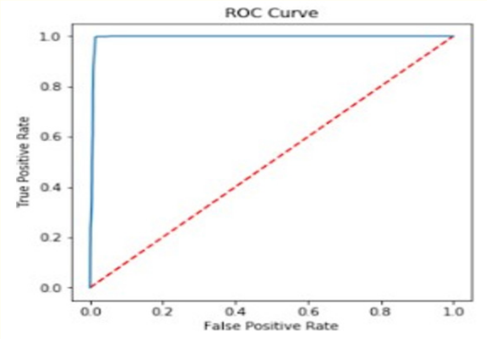

Figure 5: Validation curve obtained in the experiments.

Conclusion

In this work, different smart building solutions are proposed using Spark framework for predicting the room occupancy using a logistic regressor. The IoT sensors are deployed at different locations for data acquisition; the stream data collected using these sensors is fed to the classifiers for predicting room occupancy. We experimented with different hyperparameter settings; empirically, the classifier is able to achieve a recognition accuracy of 99.29% and 99.56% on training and testing datasets, respectively.

In future, we will be extending this study to experiment with different machine learning classifiers in varying conditions of room illumination conditions.

Bibliography

1. A Corna., *et al*. “Occupancy detection via ibeacon on android devices for smart building management”. in 2015 Design, Automation and Test in Europe Conference and Exhibition (DATE). IEEE (2015): 629-632.

2. LM Candanedo and V Feldheim. “Accurate occupancy detection of an office room from light, temperature, humidity and Co_2 measurements using statistical learning models”. *Energy and Buildings* 112 (2016): 28-39.

3. Q Hua., *et al*. “Occupancy detection in smart buildings using support vector regression method”. in 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) 02 (2016): 77-80.

4. Y Jeon., *et al*. “Iot-based occupancy detection system in indoor residential environments”. *Building and Environment* 132 (2018): 181-204.

5. H Elkhoukhi., *et al*. “Towards a real-time occupancy detection approach for smart buildings”. *Procedia Computer Science* 134 (2018): 114-120.

6. H Zhao., *et al*. “Thermal-sensor-based occupancy detection for smart buildings using machine-learning methods”. *ACM Transactions on Design Automation of Electronic Systems (TODAES)* 23.4 (2018): 1-21.

7. E Hailemariam., *et al*. “Real-time occupancy detection using decision trees with multiple sensor types”. in Proceedings of the 2011 Symposium on Simulation for Architecture and Urban Design (2011): 141-148.

8. H Saha., *et al*. “Occupancy sensing in buildings: A review of data analytics approaches”. *Energy and Buildings* (2019).

9. K Akkaya., *et al*. "Iot-based occupancy monitoring techniques for energy-efficient smart buildings". in 2015 IEEE Wireless Communications and Networking Conference Workshops (WCNCW) (2015): 58-63.

10. Y Ji., *et al*. "Occupancy detection technology in the building based on iot environment sensors". in Proceedings of the 8th International Conference on the Internet of Things (2018): 1-4.

11. W Shen and G Newsham. "Smart phone based occupancy detection in office buildings". in 2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD) (2016): 632-636.

12. M Azam., *et al*. "Occupancy estimation using wifi motion detection via supervised machine learning algorithms". in 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP) (2019): 1-5.

13. Y Gao., *et al*. "Occupancy detection in smart housing using both aggregated and appliance-specific power consumption data". in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (2018): 1296-1303.

14. C Feng., *et al*. "Deep learning-based real-time building occupancy detection using ami data". *IEEE Transactions on Smart Grid* (2020): 1-1.

15. T Ekwevugbe., *et al*. "Real-time building occupancy sensing using neural-network based sensor network". in 2013 7th IEEE International Conference on Digital Ecosystems and Technologies (DEST) (2013): 114-119.

16. A Parise., *et al*. "Indoor occupancy prediction using an iot platform". in 2019 Sixth International Conference on Internet of Things: Systems, Management and Security (IOTSMS) (2019): 26-31.

17. MR Bashir and AQ Gill. "Towards an iot big data analytics framework: Smart buildings systems". in 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (2016): 1325-1332.

18. OP Qwerty. "History of the goofy layout of keyboards". Ph.D. dissertation, Podunk IN, (1996).

19. R Swearingen. "Morpholoty and syntax of british sailors' English". New York NY, Tech. Rep. (1985).

20. T Upsilon. "Obscure greek letters and their meanings in mathematics and the sciences". in Proceedings of the seventh international trivia conference, V. W. Xavier, Ed. Philadelphia PA: Last Resort Publishers, (1987): 129-158.

21. J Tetazoo. "A brief guide to recreational pyromania". (2005/06/12).

22. R Huo., *et al*. "A comprehensive survey on blockchain in industrial internet of things: Motivations, research progresses, and future challenges". *IEEE Communications Surveys and Tutorials* (2022).

23. E Khaoula., *et al*. "Machine learning and the internet of things for smart buildings: A state of the art survey". in 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). IEEE (2022): 1-10.

24. S Zhu., *et al*. "Survey of testing methods and testbed development concerning internet of things". *Wireless Personal Communications* 123.1 (2022): 165-194.

25. J J P Abad´ıa., *et al*. "A systematic survey of internet of things frameworks for smart city applications". *Sustainable Cities and Society* (2022): 103949.

26. G F Huseien and K W Shah. "A review on 5g technology for smart energy management and smart buildings in singapore". *Energy and AI* 7 (2022): 100116.