Determine Occupancy of an Office Room using Statistical Techniques

Paras Amitkumar Lad
1170211
Lakehead University
Thunder Bay, ON, Canada
plad2@lakeheadu.ca

Priyank Patel
1150518
Lakehead University
Thunder Bay, ON, Canada
ppatel92@lakeheadu.ca

***Abstract*—Accurate information related to occupancy can help us in various application like seat allocation and energy conservation. This article will showcase different statistical techniques like Logistic Regression, Decision Tree and SVM used to classify this binary classification problem which takes Light, Temperature, Humidity and CO2 of an office room as parameters to classify whether room is occupied or not.**

I. Introduction

Energy Conservation is a hot topic in today's era. If we can accurately determine the occupancy in a particular area then we can move a step forward in conserving it. Cameras and other visual devices can easily be used to determine it but some areas may have privacy concerns and we can't use them. Nowadays, different sensors are available at an affordable price which can be used to determine temperature, light, humidity and other factors of a particular area. Our method tries to determine the occupancy based on data captured by different sensors in an office room.

Our approach takes temperature, light, CO2 and Humidity captured from different sensors and try to detect the occupancy based on it through statistical learning methods for classification. Dataset is available in UCI library named as 'Occupancy Detection Data Set' https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+ It consists of one training set containing 8143 samples and two testing set. First testing set consist of features when office door is open and another when door is closed.Our approach will Logistic Regression,Decision Tree and Support Vector Machine(SVM) to classify our features and will find which approach will yield optimal results.We will also look at the impact of each feature on our accuracy. Let's say we will consider only Temperature from our feature set and capture our results. Same we can do for other three features (Light, Humidity and CO2) or we can take two or three feature combination and determine which component of our dataset will contribute more and which has least affect on our outcome.

II. Literatue Review

When occupancy detection was used in experiments, 37 times more energy was conserved in [1] and when used with HVAC control algorithm as input it ranged 29-80%[2].Real time occupancy detection was presented by [4] using decision tree. CO2, sound, motion sensors, power use and Light were used as parameters and accuracy reached around 97.5%. He found that implementing too many reading from sensors may decrease our accuracy due to over-fitting.

CO2, motions,Relative humidity, computer temperature and air temperature were captured from sensors and neural network model was implemented on that but that reach accuracy around 75 to 85%[6]. Sources like illumination, CO2, noise, motion, humidity and temperature was taken into consideration and was trained under SVM and Neural Networks to get an accuracy around 80%[7]. PIR Sensors and Thermal Camera were used to track occupants in determine occupancy[8].KNN, Linear Regression and ANN were used for classification.

This dataset was implemented with statistical methods like LDA, Classification and Regression Trees(CART) and Random Forest(RF) models and accuracy was received accuracy around 95-99% [3].It is interesting to know that with only one feature (Temperature), LDA model outputs an accuracy of 85 to 83 %. Our approach will try to address this data to different statistical models compare the result with each approach and fetch the one having best results. We will also focus on which feature from our dataset will be most impactful and which will impact the least to our outcome.

III. Methodology and Justification

There are many factors which determine the accuracy of our results. Calibration of sensors plays a huge role in determining the accuracy of our target. If sensors are not appropriately located, they may not trace the accurate data required to generate our datasets. We have used dataset from below UCI Library.

https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+

The data was fetched from different sensors used to monitor humidity, temperature, CO2 and Light within an office room. Our data also include humidity ratio which is determined by the temperature and relative humidity. The overall process of data collection is mentioned in "Accurate occupancy detection of an office room from light, temperature, humidity and CO2measurements using statistical learning models" [3].

We have performed normalization of our data so that every component of our data are in common scales without affecting

differences in their range of values. We performed Min-Max Normalization described as below:

X[i] = (X[i]-min(X)/max(X)-min(X)) * (new_max – new_min) + new_min

Here X[i] is a sample from list X, min(X) and max(X) is the minimum and maximum element of the list X respectively. New_min and new_max are values to which we want our data to be ranged, so we have set new_min as 0 and new_max as 1 to make all our data between 0 and 1.

Various Statistical Models were used over the dataset which we will discussing in the following sections.

A. Logistic Regression

Logistic Regression is one of the basic classification models used for binary classification of data. It is used to analyze the relationship between different dependent and independent variables in our data by evaluating probabilities. Logistic Regression uses linear kernel given by given by e(WX)/ (1 + e(WX)) which assures that our output resides in the range 0 and 1. We can use a threshold let's say 0.5, and any value between 0.5 and 1 belongs to one class and the value between 0 and 0.5 belongs to another. Figure 1 shows a curve of a logistic function. So this classifier is simple to implement with linear kernel and thus help us to categorize our data samples.

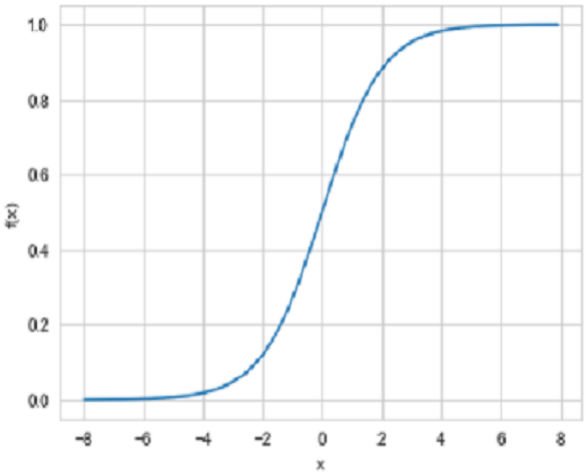

Fig. 1. Logistic Regression

B. Decision Tree

Decision Tree follows a graphical representation of all possible solutions to a decision. It follows a tree structure. Root node is the first node of our tree from where our classification starts. At any node of a tree, a feature of the data is compared with a threshold value and it's result is used to select one of the two child nodes attached to it. When there is no child node of a particular node or it's a leaf node then the class predicted by the node is our prediction of our model. All child nodes follow the feature from their parent node. Decision Tree is depicted in Figure 2 . Within a given dataset, to determine the root node, it is selected as the attribute having the most information gain and the threshold is selected based on the gain ratio. Gini Index is also being calculated which depicts the probability of an attribute that is being wrongly classified when it is randomly chosen. While traversing through a tree, each path exhibits a combination of attributes that contributes towards distinguishing of classes thus provides useful information of inner working of the classifier which can't be measured in Neural Networks, SVM or any other classifiers. This feature of Decision Tree fascinated us to select it for classifying our dataset.

Fig. 2. Decision Tree

C. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression problems. SVM is unique than other classifiers as it takes the decision boundary that maximizes the distance between the closest sample point of all the classes. We can say SVM tries to find an optimal boundary that separates all classes in given dataset. The nearest points taken into consideration to determine the decision boundary between classes are called Support Vectors. We can use linear or non-linear kernel functions to map our data from low dimension space to high dimension space which makes SVM more powerful than other classifier models. Figure 3 shows a basic example of SVM. Figure 4 depicts overview of our entire process.

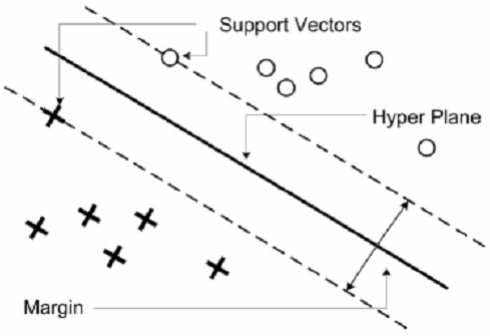

Fig. 3. Support Vector Machine

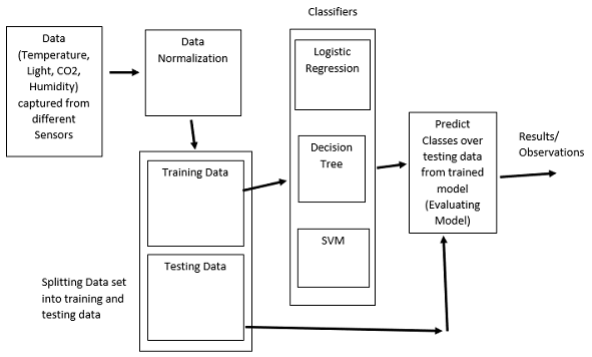

Fig. 4. System Overview Diagram

IV. Experiments

As approaching to our goal to find occupancy in a room, our first target was to find the relationships between the parameters which determine it. To fetch the relationship between different parameters, we plotted a correlation matrix as shown in Fig. 5. Correlation matrix shows how each random variable X has a relationship with other parameters in the dataset. It is depicted as a table where diagonal elements are always set to 1 as the correlation with the variable itself is always 1. We found that Temperature and Humidity is inversely proportional to each other which means if temperature increases, humidity decreases. All other variables are proportional with each other.

	Temperature	Humidity	Light	CO2	Humidity Ratio
Temperature	1.000000	-0.141759	0.649942	0.559894	0.151762
Humidity	-0.141759	1.000000	0.037828	0.439023	0.955198
Light	0.649942	0.037828	1.000000	0.664022	0.230420
CO2	0.559894	0.439023	0.664022	1.000000	0.626556
Humidity Ratio	0.151762	0.955198	0.230420	0.626556	1.000000

Fig. 5. Correlation Matrix

We performed Logistic Regression with Newton Raphson approach over our dataset to fetch our result. To train our network it took around 20 epochs to converge the parameters and achieve the best accuracy to be reached over the model.

We analyze Decision Tree over the dataset by considering maximum depth of tree as 3 to yields the features having the highest Information Gain as the root node and others act as child node of that root node. Same procedure follows for all node until we get a leaf node which determine our class whether the room is occupied or not. Due to high complexity of Decision tree, time taken to train the model is larger compared to other models.

We accomplish Support Vector Machine (SVM) with linear kernel considering lambda as 0.001, learning rate as 0.001 and loop over 500 iterations to optimize our parameters weight and bias after it was confirmed that no significant decrease was noticed to achieve the best decision boundary that separate our classes.

We have two test datasets, one when office room is open and another when office room is closed each having 2665 and 9752 samples respectively. To get insights of our dataset and its features, we performed analysis over different combination of features to visualize the impact on our outcome. For example, while training our model, we took Temperature, Light, CO2 as feature set and produce our results and test our model accuracies.

V. Results

Training and Testing Accuracies over different combination of features of datasets over Logistic Regression, Decision Tree and Support Vector Machine is depicted in figure 6, 7 and 8. In figure T, C, L, H, HR corresponds to Temperature, CO2, Light, Humidity and Humidity ratio respectively.

We examined that the best training accuracies were noticed in Decision Tree although Testing accuracies were not quite appreciative. This occurs as it tries to memorize the training samples and do perform well over it but when new sample arrives it is under performed. Logistic Regression too was not good in determining testing accuracies especially for testing dataset 2, which contains record when office door was open. This may be due to high correlation between different features in the dataset which leads to wrong training of parameters and optimizing cost function. SVM performed well in both training and testing accuracies and proves to be a better option then Decision Tree or Logistic Regression. It gives as optimal boundary to separate our two classes which can't be achieved in any other models.

Combination of all the features results in high accuracies as compared to individual features. So, each parameters plays a role to achieve better accuracies. Although combination of CO2, Light and Humidity results in best accuracy around 99.32 for test dataset 2 and Light alone produce the best accuracy of 97.86 for test dataset 1 for SVM model.

Combination of Temperature, CO2 and Humidity ratio yields the least accuracy for testing dataset 2. As door is open, so all these factors doesn't play an important role in determining accuracy as all these parameters would be changed and will not accurately determine the data of a room. Similarly, Humidity plays the least role in testing dataset1 when room is closed.

On our feature set, we found Light to be a major factor which determines our results. That's proves to be true as whenever any occupant is in the room, light sensor would receive high value and when no person is present and lights are turned off, it's value sharply decreases. So any feature combination with Light as one of it's feature, we usually notice high accuracies for our data.

VI. Discussions and Conclusion

We found that there is high correlation between some parameters in our dataset. Logistic Regression to functional more efficiently it requires less or no co linearity between independent parameters as repetition of information can lead to wrong training of parameters which minimizing cost function. So we found in our results that combination of some variables fetched low accuracy during training and testing.

Model	Parameters	Training Accuracy	Testing Accuracy	
			Door Closed	Door Open
Logistic Regression	T,C,L,H,HR	98.62	96.47	74.34
	T	78.7	63.03	78.65
	C	88.92	64.11	70.05
	L	94.3	68.5	75.31
	H	78.75	63.18	78.97
	HR	78.75	63.48	78.97
	T,C	87.05	64.05	57.19
	T,L	96.8	96.51	97.6
	T,H	74.75	58.01	76.99
	T,HR	78.76	63.52	78.98
	C,L	85.43	97.82	85.4
	C,H	90.96	73.8	71.84
	C,HR	81.53	71.44	66.69
	L,H	91.37	90.54	95.26
	L,HR	90.55	86.26	93.96
	H,HR	85.04	71.81	24.92
	T,C,L	97.5	92.42	84.39
	T,C,H	93.15	79.77	73.67
	T,L,H	95.66	97.29	98.6
	T,L,HR	95.71	96.06	98.72
	T,H,HR	84.18	76.92	23.23
	T,C,HR	93.34	77.82	69.12
	C,L,H	91.94	85.02	78.51
	C,L,HR	83.78	86.56	73.53
	C,H,HR	89.75	73.17	71.6
	L,H,HR	91.77	89.15	64.87
	T,C,L,H	98.256	95.984	88
	T,C,L,HR	98.26	94.82	85.05
	T,L,H,HR	98.79	97.11	45.11
	T,H,C,HR	92.59	77.22	66.19
	C,L,H,HR	91.94	81.91	76.88

Fig. 6. Logistic Regression Accuracy

Model	Parameters	Training Accuracy	Testing Accuracy	
			Door Closed	Door Open
Decision Tree	T,C,L,H,HR	99.17	97.82	95.1
	T	86.45	74.89	79.64
	C	92.27	84.65	76.39
	L	98.84	97.52	97.93
	H	81.7	63.45	63.65
	HR	83.86	59.58	65.42
	T,C	94.19	77.26	59.84
	T,L	99.05	95.04	98.2
	T,H	90.87	83.15	81.38
	T,HR	91.14	83.82	69.53
	C,L	98.98	97.82	98.53
	C,H	94.81	85.77	62.36
	C,HR	94.44	77.41	46.82
	L,H	98.91	97.78	97.16
	L,HR	98.93	97.63	98.36
	H,HR	87.21	73.17	53.35
	T,C,L	99.17	97.82	95.87
	T,C,H	95.27	67.87	44.98
	T,L,H	99.36	95.23	94.96
	T,L,HR	99.32	95.05	97.09
	T,H,HR	91.34	83	58.65
	T,C,HR	94.73	77.67	44.06
	C,L,H	98.98	97.82	97.76
	C,L,HR	98.98	97.82	98.53
	C,H,HR	94.81	85.77	62.36
	L,H,HR	98.91	97.78	97.16
	T,C,L,H	99.17	97.82	95.1
	T,C,L,HR	99.17	97.82	98.87
	T,L,H,HR	99.36	95.23	94.96
	T,H,C,HR	95.05	79.06	49.93
	C,L,H,HR	98.98	97.82	97.76

Fig. 7. Decision Tree Accuracy

We choose Decision tree to classify our data as it provides the information which tells us how model classify the data but huge disadvantage of it is it increases the time and the complexity of the model. We also found that decision tree gives high training accuracy but don't provide accurate testing accuracy. Over-fitting of data is the main reason behind it. It tries to memorize the data and thus new samples does not produce results as expected. Pruning can be a step to reduce this over-fitting problem.

SVM classify our dataset quite well with linear kernel and the decision boundary obtained is optimal. So SVM would be

Model	Parameters	Training Accuracy	Testing Accuracy	
			Door Closed	Door Open
SVM	T,C,L,H,HR	98.24	97.48	99.21
	T	84.2	80.75	78.25
	C	90.1	86.45	79.31
	L	95.66	97.86	96.3
	H	78.76	63.52	78.98
	HR	78.76	63.52	78.98
	T,C	89.2	86.71	76.41
	T,L	95.67	97.86	96.42
	T,H	82.27	78.23	78.7
	T,HR	83.75	80.18	80.4
	C,L	98.55	96.54	99.31
	C,H	91.09	87.01	78.6
	C,HR	90.26	86.6	78.36
	L,H	98.73	97.67	98.85
	L,HR	98.75	97.56	98.02
	H,HR	79.25	72.83	80.92
	T,C,L	98.39	97.78	98.68
	T,C,H	89.2	86.97	76.05
	T,L,H	98.62	95.94	98.85
	T,L,HR	98.74	97.44	98.15
	T,H,HR	83.87	80.03	80.63
	T,C,HR	89.19	86.94	75.98
	C,L,H	98.6	96.47	99.32
	C,L,HR	98.57	96.24	99.26
	C,H,HR	90.23	86.79	79.73
	L,H,HR	98.75	97.52	98.24
	T,C,L,H	98.4	97.56	98.97
	T,C,L,HR	98.51	97.63	99.12
	T,L,H,HR	98.64	95.98	98.32
	T,H,C,HR	89.08	86.9	76.05
	C,L,H,HR	98.52	96.39	99.32

Fig. 8. SVM Accuracy

the best model to classify our data. Although results obtained are appreciable, we are still not aware of the best kernel to choose from, which can further work upon to acquire better outcome than this linear kernel.

Future work can be done on implementing some new/high functional sensors and collaborating it which can provide more accurate results.We can also move a step forward towards finding the number of occupants present in the room based on different information presented by different sensors.

VII. Acknowledgement

Paras Amitkumar Lad : Literature Review, Methodology and Justification (Logistic Regression and Decision Tree), Experiments, Results of model implemented, Discussion and Conclusion

Priyank Patel : Abstract and Introduction, Methodology and Justification (Logistic Regression and SVM), Results of model implemented, Discussion and Conclusion

References

[1] An experimental investigation of occupancy-based energy-efficient control of commercial building indoor climate — IEEE Conference Publication — IEEE Xplore

[2] Energy-efficient control of under-actuated HVAC zones in commercial buildings

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

[4] E. Hailemariam, R. Goldstein, R. Attar, A. Khan, Real-time occupancy detection using decision trees with multiple sensor types, in: Proceedings of the 2011Symposium on Simulation for Architecture and Urban Design, Society forComputer Simulation International, San Diego, CA, 2011, pp. 141–148

[5] Ekwevugbe T., Brown N., Pakka V. (2013). Real-time building occupancysensing for supporting demand driven HVAC operations. 13th InternationalConference for Enhanced Building Operations, Montreal, Quebec

[6] T. Ekwevugbe, N. Brown, V. Pakka, D. Fan, Real-time building occupancysensing using neural-network based sensor network, in: 7th IEEEInternational Conference on IEEE, Digital Ecosystems and Technologies(DEST), Menlo Park, California, 2013, pp. 114–119.

[7] K.P. Lam, M. Höynck, B. Dong, B. Andrews, Y.-S. Chiou, R. Zhang, D. Benitez, J.Choi, Occupancy detection through an extensive environmental sensornetwork in an open-plan office building, IBPSA Build. Simulat. 145 (2009)1452–1459.

[8] A. Beltran, V.L. Erickson, A.E. Cerpa, Thermosense: occupancy thermal basedsensing for hvac control, in: Proceedings of the 5th ACM Workshop onEmbedded Systems For Energy-Efficient Buildings, ACM, Rome, Italy, 2013,pp. 11:11–11:18.