Joshua Oswari

Wen-Xin Zhou

Math189

06/06/19

Room Occupancy Prediction

Introduction

After reading the research paper written by Luis and Veronique titled, "Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models." I was intrigued. Never did I think all of the statistical method that I learned throughout the quarter can be applied in this kind of experiment. A little bit about my thoughts, I often think that Mathematics major is a little bit useless because the application is very little in the real world considering the number of classes that I usually took is pure Math. However, after looking at this project, I'm challenged to looked at the big picture and apply these Mathematical skills that I learned, especially from this class, Math189. Maybe this kind of topic written by Luis and Veronique can be explored much deeper that maybe in the future can be useful for things such as investigating a criminal case and room reservation that I think can be implemented in Geisel Library knowing that often times many library rooms are open yet there is no tracker saying that the room is available to use. Back to the original topic, about this final project, I will try the same statistical classification models using the program R. I will explore a little bit deeper why the best accuracies are obtained from training Linear Discriminant Analysis (LDA) and why Random Forest (RF) is not as accurate as LDA.

Now before we go any deeper into the summarization, we want to analyze the data by combining all of the three data and see the summary of the whole data.

```
> str(wholeData)
'data.frame':   20560 obs. of  7 variables:
 $ date         : Factor w/ 20560 levels "2015-02-04 17:51:00",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Temperature  : num  23.2 23.1 23.1 23.1 23.1 ...
 $ Humidity     : num  27.3 27.3 27.2 27.2 27.2 ...
 $ Light        : num  426 430 426 426 426 ...
 $ CO2          : num  721 714 714 708 704 ...
 $ HumidityRatio: num  0.00479 0.00478 0.00478 0.00477 0.00476 ...
 $ Occupancy    : int  1 1 1 1 1 1 1 1 1 1 ...
```

Fig. A. Structure Data

Figure A will represent the structure of the data that will be used throughout this paper.

```
Not Occupied     Occupied
        76.9         23.1
```

And then we can see from the data combined that 76.9% of the time the room is not occupied while 23.1% of the time the room is occupied.

```
  Temperature       Humidity        Light             CO2         HumidityRatio            Occupancy
 Min.   :19.00   Min.   :16.75   Min.   :   0.0   Min.   : 412.8   Min.   :0.002674   Not Occupied:15810
 1st Qu.:20.20   1st Qu.:24.50   1st Qu.:   0.0   1st Qu.: 460.0   1st Qu.:0.003719   Occupied     : 4750
 Median :20.70   Median :27.29   Median :   0.0   Median : 565.4   Median :0.004292
 Mean   :20.91   Mean   :27.66   Mean   : 130.8   Mean   : 690.6   Mean   :0.004228
 3rd Qu.:21.52   3rd Qu.:31.29   3rd Qu.: 301.0   3rd Qu.: 804.7   3rd Qu.:0.004832
 Max.   :24.41   Max.   :39.50   Max.   :1697.2   Max.   :2076.5   Max.   :0.006476
```

Fig. B. Summary of the Data

Figure B above is the summary of all of the variables that will be used throughout the paper and we see that the ranges from each variable is differ from each other. In this observation there are total 6 variables that will be a huge important factor in determining the room. The variables contained in the datasets are:[1]

(a) Date: in the form: year-month-day hour: minute: second

(b) Temperature: in Celsius

(c) Relative Humidity: in

(d) Light: in Lux

[1] The variables explanation are taken from Final_Project.pdf from Piazza

(e) CO2: in ppm

(f) Humidity Ratio: a derived quantity from temperature and relative humidity; in kg (water-vapor)/kg (air)

(g) Occupancy: 0 or 1; 0 for not occupied, 1 for occupied status. Ground-truth occupancy was obtained from time stamped pictures of the rooms that were taken every minute.

There are actually 7 variables in the dataset, however, as it is said earlier, there is only 6 variables that will be the key factor in determining whether the room is occupied or not. The variable date time year-month-day hour:minute:second is only the index of the data or the timeframe of the data, which does not have any relationship in determining the final result or will affecting the data in some ways.

Methodology

1. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis or LDA is such a powerful method that can be applied in this data because the method itself that will classify more than two variables. Moreover, LDA is a supervised method, meaning that to LDA tries to build model and predict its each dependent variable. This goes along with our problem in determining whether the room is occupied or not based on multiple variables which are temperature, humidity, and many others. In this test, I will be using 17560 observations in the train set and the rest, which is 3000 observations to predict and check the accuracy of the model.

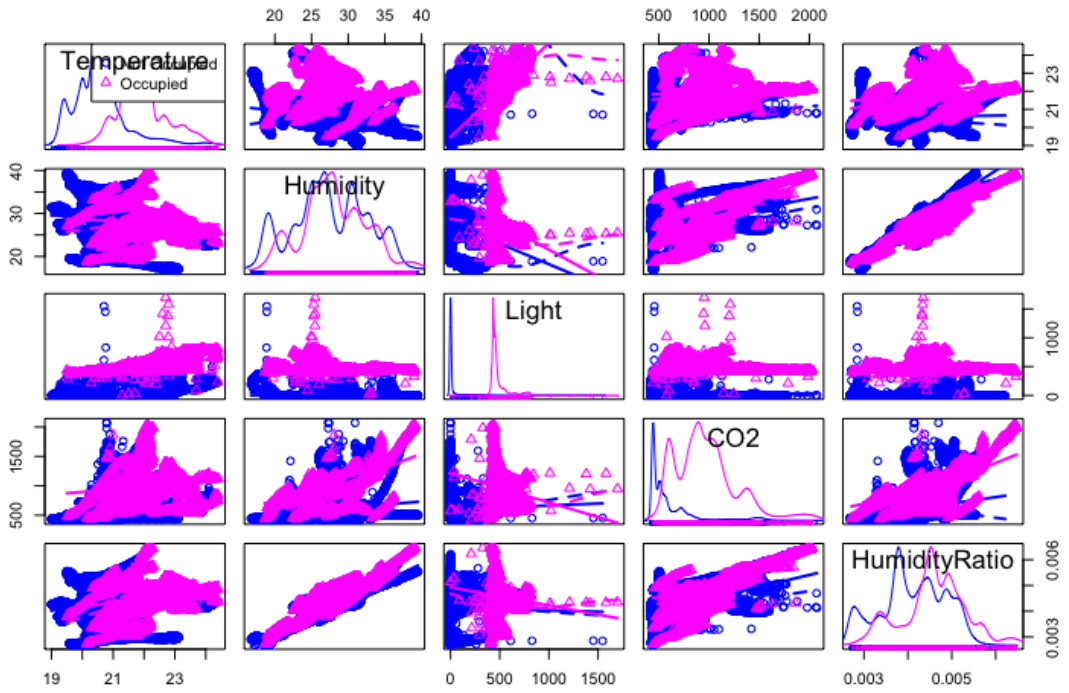

Fig. C. LDA classification scatter matrix plot

From the scatter matrix plot we can see, the pink triangle indicating that the room is occupied while blue indicating that the room is not occupied. From the plot I can see that there are high correlation between temperature and humidity, Co2 and temperature, which somehow indicate there is a link between these two classes. And the highest accuracies that I obtained using LDA models is 98.3223% for the first test and 99.38% for the second test. And I think this is already a good start because I got such high accuracy in predicting whether the room is occupied or not.

2. Random Forest (RF)

Random Forest or RF is a tree-based method. With Random Forest, the RF algorithm in R will construct a multi decision tress and then will create a bunch of small regions recursively with assigned class label, and in this case, this will end with label 0 and 1, with 0 as the room is not occupied and 1 meaning that the room is occupied. In this

classification I will be using the same amount of training observations and testing observations in LDA classification. I'm repeating the number of trees ranging from ntree 100 to 1000 to find out which result will get the highest ratio. And after waiting such long time knowing that the data observed is more than ten thousand data, I got the highest rate of accuracies with ~99.4% which is very high considering Luis and Veronique claim that the highest accuracies are obtained through training Linear Discriminant Analysis (LDA) (Luis and Veronique 2).

Conclusion

In conclusion, between Linear Discriminant Analysis and Random Forest, Random Forests will generate better performance in regard of predicting the accuracies from the data with accuracies of 99.4% while LDA 99.38% resulting only difference of 0.02% which somewhat is not fair to say Random Forest perform better in this case. However, when I tried to choose ntree between 100 to 1000, I can see that the accuracies that RF generate exponentially increasing as the number of trees being input are higher. Therefore, I can still say that the Random Forest is the best bet for getting the highest accuracies in this test. However, there are some drawbacks that I can conclude from each classification. The tree-based method is straightforward and somewhat easy to understand and interpret, yet, in this case, we have such a large number of variables. Therefore it is a little bit hard, and the decision-tree can create such complex trees that even can't be observed. And the amount of time the code calculating is very long, which is more than 2 minutes to generate.

This conclusion, however, contradicting to what Luis and Veronique claim, Luis and Veronique affirm that the LDA will perform higher than the other classification. I suspected that this has happened because the code that I and Luis and Veronique differs, or the amount of training and test observations split are different. In the future, I would suggest that the time frame from the data is more widely collected as well as other variables that need to be taken into account. Because in this data, I noticed that the data collected for only two weeks. This might be caused in different conclusion from what I concluded, and Luis and Veronique concluded.

Works Cited

Candanedo, Luis M., and Véronique 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. Crossref. Web.