Occupancy Detection Using Machine Learning

Yuxiang Xu 12715948
Yi Lin 12444129

Ⅰ. Introduction

Occupancy detection means to predict room occupancy by machine learning models and room data, such as light, temperature, humidity and CO2 measurements. When the related indexes change, the room occupancy status also changes. This application is useful in the office, market, and other public room centers. It helps us to know accurate room occupancy with no human intervention.

Tree-based learning algorithms are considered to be one of the best-supervised learning methods. And it is mostly used. In this paper, I chose the decision tree as my learning method. The aim of this paper is detecting room occupancy by the decision tree.

Ⅱ. Data Preparation and Analysis(Exploration)

The data related to room occupancy comes from Luis Candanedo's experiments(Candanedo, 2016). There are 7 columns in total, including 1 timely column, 5 attribute columns, and 1 target column. The target value called occupancy is binary, 0 for not occupied and 1 for occupied status. The attributes consist of light, temperature, relative humidity, CO2, and humidity ratio, which is derived from temperature and relative humidity.

Before modeling, we had to prepare our data and analyze it. Original data is always raw. We need to clean and check it carefully. The first step is a missing value. According to our observation, there is no missing value in our data. We analyzed 5 attributes one by one.

Another thing we had to consider is model evaluation. After modeling, we have to check our model to make sure it is valid in new observations. We had split the original data into two parts, 70% for training and 30% for testing. The training data is selected randomly from total data. Left data is used to test model validity.

Ⅲ. Model(Methodology) and Evaluation

The decision tree is a tree-like structure, in which every part has its role to play. The internal node represents a test on attributes. For example, whether a coin flip comes up heads or tails, and each branch represents the outcome of the test. And each leaf node represents a class label, which means the decision taken after computing all attributes. The paths from the root to leaf nodes are classification rules.

ID3(Quinlan 1986), C4.5(Quinlan, 1986), and CART(Breiman et al.1984) are commonly used algorithms in the decision tree. The difference between them is using different criteria as internal nodes. When determining how to split the original data set, ID3 and

C4.5 are depended on entropy. However, CART is based on the Gini index. The nature of entropy and the nature of Gini are the same. Both of them are used to describe the impurity of the dataset. The best choice of the split can improve the purity of the data set. In our project, we compared the different criteria in the decision tree. The comparison method is variable-controlling. We just changed the criteria parameter while keeping others unchanged. The same training data set is used twice to keep data unchanged. The results are shown below. The accuracy rate of C4.5 is about 99.18%. Meanwhile, the accuracy rate of the CART algorithm is a little higher, and it is close to 99.22%.

Algorithm	C4.5	CART
Accuracy Rate	99.18%	99.22%

Table 1: Accuracy rate of different algorithms in decision tree

Overfitting is a hard question in the decision tree. In theory, if the class labels of records that have identical attributes are the same, the decision tree can reach a 100% accuracy rate in the training data set. In this situation, the decision tree will be very complex. However, there is no point to do this because it is invalid in the test data set. In decision trees, overfitting is very common. Since the decision tree algorithm continues splitting into attributes until either it classifies all the data points or there are no more attributes to splits on.

Pruning is one of the most used methods to avoid overfitting. Pruning means restricting the max depth of the decision tree or the min number of every leaf node. Under the restriction, the decision tree can't be too complex. According to former results, the accuracy rate of the test data set is bigger than 99%. And this means that our model performs very well. Overfitting seems non-existent in our problem. But to check it, we still considered pruning. The pruning method used by us is restricting the max depth of the tree, the former one.

The experiment results are shown here. The default parameter of max depth in Python's decision tree module is none, which means infinitive. According to the figure's data, we can see that the default result is not the best. The performance in the test data set is not the best one. The accuracy rate of the default parameter is 99.18%. This value is lower than in most other situations. We can even conclude that it is the worst performance. From the figure, there is a peak point. The accuracy rate is close to 99.43% when max_depth equals to 5. So we can know that the decision tree is mort complex, the result maybe not the best. Restricting the max depth of the decision tree may be helpful to improve the accuracy rate.

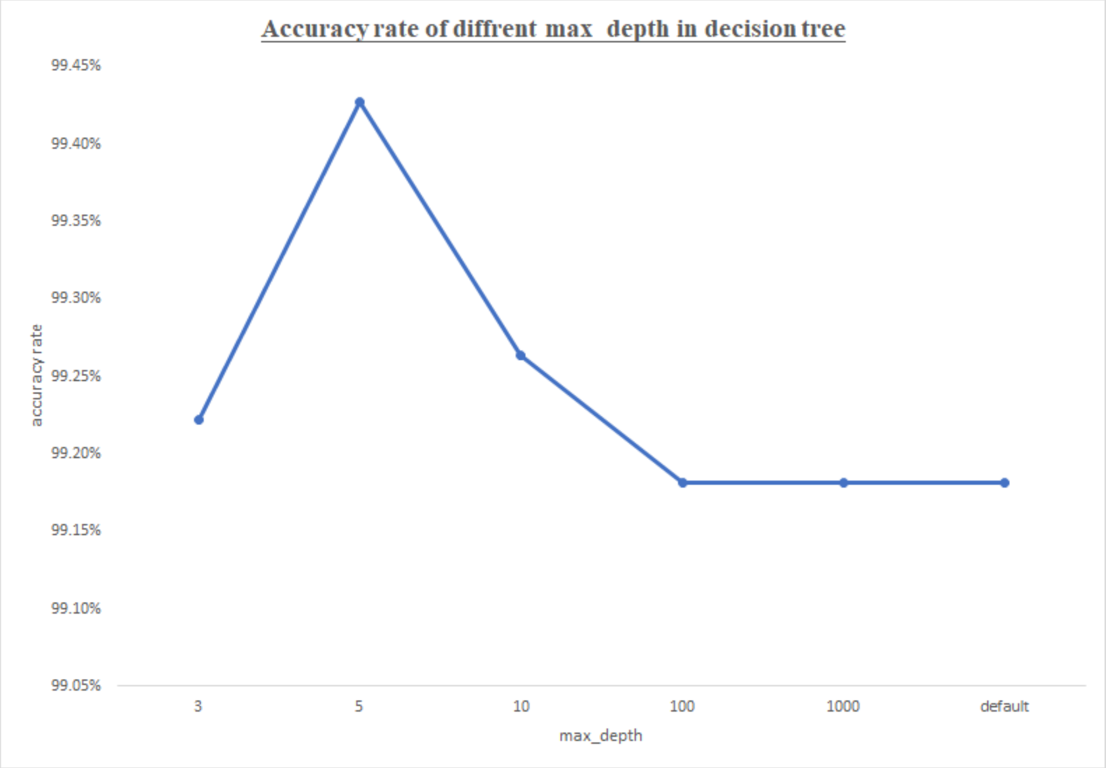

Figure 1: Accuracy rate of different max depth in decision trees.

Ⅳ. Conclusion

From our project, we can conclude some important points from two aspects. One aspect is about occupancy detection. As you can see, we can establish a valid model to predict occupancy status. And the accuracy rate if very high in any situation. The other aspect is about machine learning techniques. Different choices of criterion can bring different results. And pruning is very important to avoid overfitting. It is a useful method to improve model performance.

Ⅴ. Ethical discussion

I don't think our project has any ethical problems. It's just a simple project for occupancy detection. This can only help people predict the situation of vacant rooms and draw conclusions from the data collected in some rooms. And in the process of collecting these data, I don't know whether it infringes some personal privacy. I don't know, because these data are obtained from the public dataset website. I know the Real-world data is always raw and hard to use. I used the public dataset on a website called UCI Machine Learning Repository. There should be no infringement of other people's copyright and so on.

Reference

Breiman, L., Friedman, J., Olshen, R., Stone, C. 1984, ‘Classification and regression trees (cart) wadsworth international group’, CA, USA, Belmont.

Candanedo, L. 2016,‘ Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models’, *Energy and Buildings,* vol. 112, no. 1, pg. 28-39.

Quinlan,J,R. 1986,‘ Induction of Decision Trees’, *Machine Learning*. 1(1), vol. 1, no. 4, pg. 81–106.

Link to the GitHub file: https://github.com/12444129/Yi-Lin
Link to the YouTube: https://youtu.be/8lL6Dwc6Gzs