JOINT INSTITUTE
交大密西根学院

VE414 Project
Group 4

Shengyuan Xu, Fan Chen, Qiansiqi Hu

University of Michigan Joint Institute

Contact Information:
Shanghai Jiao Tong University, Dongchuan Road 800
Email: xushengyuan@sjtu.edu.cn
chenfred02@sjtu.edu.cn
hqsq0905_kon@sjtu.edu.cn

Abstract

Bayesian analysis, as one of the most essential topics in statistical studies, has achieved great success in many applications such as classification, multi-classification real-time prediction and recommender systems. Our work in this project aims to predict the total number of the Jiuling in the Forbidden Forest according to the record of the spell. However, there exists some challenges in solving this problem if we apply traditional classification methods because there're many variables unknown, and the exact location of targets are also undetermined. Our work provides a feasible method to classify the source of Tayes, determine their locations, as well as predict the total number of Jiuling in the Forbidden Forest through area estimation. Since the effect of direct clustering is not so good, we first design a mining algorithm to predict the number and location of Tayes based on the given data. Then we work backward to predict the number and distribution of Jiuling trees by applying the method of Expectation-Maximization based on Gaussian Mixture Model. Moreover, we also designed several algorithms to estimate the area covered by existing trips. The methods we propose greatly improves the accuracy of cluster analysis, and we also raised some thoughts for circumstances where Jiuling can actually move.

1. Introduction

Jiuling, a kind of magic tree in the Forbidden Forrest of Hogwarts, is invisible and therefore extremely difficult to detect. Only by detecting the distribution and number of its fruit, Tayes, can we estimate the footprint of this magic tree. In order to help Hermione Granger to finish her master's project of data science, our project aims at determining the number of Jiuling trees in the Forbidden Forest.

In our project, we first make some assumptions to regularize the data. Also, to strengthen the effectiveness and credibility of the prediction, we reject the traditional method of direct clustering. On the contrary, we backtrack the information of Tayes based on data collected by each spell. After that, we perform a cluster analysis based the on number and location of Tayes, with the help of the Expectation-Maximization (EM) algorithm and Gaussian Mixture Model to getter a better result.

The remainder of this poster is organized as follows, our main objectives are listed in section 2. In section 3, we summarize the assumptions we made in this project. In section 4 and 5, how our algorithms are implemented, as well as the corresponding results will be reported. In section 6, we draw some conclusions and make an expectation for future improvement.

2. Objectives

1. Visualize the locations and information that are available.
2. Propose efficient algorithms to estimate the number and location of Tayes in the Forbidden Forest.
3. Figure out the number and location of Jiulings in the Forbidden Forest.
4. Propose what we'll need to address the task if Jiuling can actually move in discussion part.

3. Data Visualization

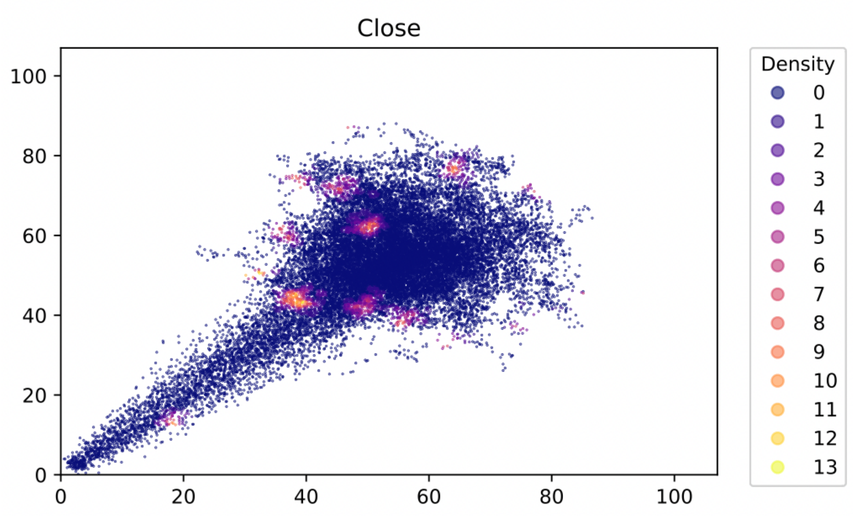

Figure 1: Data Visualization

In Figure 1, we visualize all the footprints. The color the each point is based on density, or the number of Tayes within radius of 1 meter. The lighter color indicates more Tayes in that area. As we can see, almost every cluster of Tayes is round or oval, but they do not share a common density. Also, the explored area only takes up a small portion. Thereby, in section 4, we make several assumptions according to this rudimentary visualization.

4. Assumptions

1. Position of Tayes falling from a certain Jiuling follows a 2D Gaussian distribution
2. Number of Tayes falling from a certain Jiuling follows a 1D Gaussian distribution
3. Jiuling are uniformly randomly distributed in the Forbidden Forrest.

5. Methodology & Implementation

In this project, we choose to perform an EM cluster analysis on the data points. We apply the `mclust` package in R to realize EM clustering. The optimal number of clusters, which can represent the number of Jiuling, is selected during this process. However, since the exact location of Tayes is undetermined, performing clustering directly can lead to inaccurate results. Therefore, we should first estimate the number and position of Tayes based on the data with term "Close" larger than zero. To solve this problem, we design a "mine" algorithm, which is shown on the right. The implementation of function `getLoss` is introduced in attached slides. Based on this algorithm, we can estimate the number and location of Tayes with high accuracy. With these messages known, we can perform a cluster analysis based on EM algorithm to determine the number of Jiuling.

Moreover, in order to estimate the covered area, we apply two kinds of methods. One is the Graham-Scan Algorithm, which can be used to find the convex hull of a finite set of data points. It can determine the boundary of the travelled regions so that the covered area can be found. Besides, we modifies the algorithm by setting the starting point at concave regions where large positive deviations can easily occur. The other method is to divide the forest into 107×107 grids. We propose three ways to perform grid estimation. The first one is to treat each trip point as a circle with radius 1.

Algorithm 1: Generate the prediction on fruit distribution

Input: Nearby Fruit Distribution *Target*
Output: Prediction of fruit number and location *Prediction*

function predict(Target)
Initialize *Prediction*, *loss*, tot_{loss}, $\Delta_{tot_{loss}}$,
while $\Delta_{tot_{loss}} > 3$ *or max term in loss* > 2 *or min term in loss* < -2 **do**
  newLoss, MSE, *loss* = getLoss(target, prediction)
  update $\Delta_{tot_{loss}}$, tot_{Loss}, $loss_{max}, loss_{min}$
  if $loss_{max} + loss_{min} > 0$ *or* $loss_{max} + loss_{min} == 0$ *and* $tot_{loss} > 0$ **then**
    while $loss_{max}$ *doesn't change* **do**
      x, y is the position of $loss_{max}$
      increase $Prediction[x, y]$
      update $\Delta_{tot_{loss}}$, $\text{tot}_{Loss}, loss_{max}, loss_{min}$
  else if $loss_{max} + loss_{min} < 0$ *or* $loss_{max} + loss_{min} == 0$ *and* $tot_{loss} < 0$ **then**
    while $loss_{min}$ *doesn't change* **do**
      x, y is the position of $loss_{min}$
      decrease $Prediction[x, y]$
      update $\Delta_{tot_{loss}}$, $\text{tot}_{Loss}, loss_{max}, loss_{min}$
  else
    return *Prediction*

The other method is to divide the forest into 107×107 grids. We propose three ways to perform grid estimation. The first one is to treat each trip point as a circle with radius 1. Grids that are covered or reached by those circles are marked as detected. The second one is to mark all the "loose" grids that contain trip points as detected, while the third one is to mark "strict" grids that contain at least two trip points as detected. Then the area of detected grids are summed together to estimate the total covered area. According to our assumption, we can estimate the number of Jiuling in the whole forest through the ratio of forest area to covered area.

6. Results & Discussion

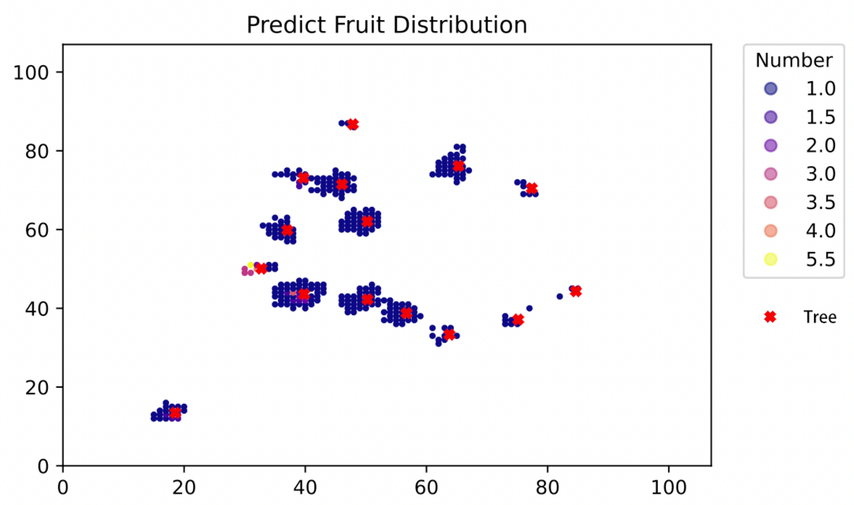

Figure 2: Distribution of Jiuling & Tayes

Results derived through Algorithm 1 and EM cluster analysis are shown in Figure 2. There're 15 clusters in total, each with source represented by a red cross. Therefore, every red cross represents the location of a Jiuling, and it's estimated that there're 15 Jiuling trees in the explored region.

The next step is to estimate the number of Jiulings in the whole Forbidden Forest. We need to figure out the area of explored region in addition to Assumption 3. Results of the estimated area of the covered region through five different methods are listed in Table 1. Regular Graham scan algorithm construct a polygon containing all points, while in modified version we manually select a starting point to better depict the region. The true value of area is roughly the average of results of previous two methods, or roughly 23.6%. Loose Grid method divides the forest into 107×107 grids, and counts the number of visited grids, resulting in 23.96%. Thereby, it's reasonable to argue that the ratio of explored area to forest area ranges is 23.96%, hence the total number of Jiuling is approximately 62.60, and they're uniformly randomly distributed in the uncovered area.

Method	Value	Proportion	Tree Number
Regular Graham scan	2927.89	25.57%	58.66
Modified Graham scan	2480.89	21.66%	69.25
Circular Coverage	3228	28.19%	53.21
Loose Grid	2743	23.96%	62.60
Strict Grid	2283	19.94%	75.23

Table 1: Total Tree Number Estimation

For cases where Jiuling can actually move, two key questions about Tayes can be raised: do Tayes exist permanently? On what basis do Tayes fall from Jiuling? For example, if Tayes exist permanently, fall from Jiuling continuously and frequently, then we can determine the moving range of each Jiuling since gaps exist between different clusters. If tayes exist permanently and fall from Jiuling periodically, then we cannot easily deny that there's only one Jiuling. Different assumptions can lead to very different results. On the other hand, the existing data is far not enough. One most important thing is to record the starting time of each trip, since the position of each Jiuling may vary over time. Moreover, we need more people travelling at the same time. Simultaneous record of data & observations at different locations can be extremely helpful. Besides, more groups of data covering the whole forest are needed. Because one Jiuling can move towards anywhere inside the forest, estimating from local to whole lacks accuracy and becomes less convincing.

References

[1] A. Gelman and J. B. Carlin. *Bayesian Data Analysis*. CRC Press, 3rd edition, 2014.