Hang Zhao

CSE 482

Pang-Ning Tan

03/24/2019

Crime Rate Prediction of Chicago

Chicago is one of the largest cities in the United States, and it is also one of the cities that has a large crime rate. Big data can sufficiently predict the crime rate. First, this project is to analyze the crime rate of Chicago at different times, like a weekday, weekend, morning of the day, and afternoon of the day. Second, this project utilizes the same way of analyzing the crime rate of the City of Chicago as analyzing the crime rate of each community in Chicago. It also adds the data of population density to analyze the correlation of population density and the crime rate within the community. Finally, this project will analyze individual crime situation of each community, such as the percentage of different type of crime, the rate of crime in the morning, afternoon and evening, and the rate of weekday and weekend.

The three data sets used in this project are the Chicago Crime data from 2001 to Present, Chicago Community Area Census data, and the Chicago Community Areas data. The Chicago Crime data is the main data in this project, this data set contains all Chicago crime information, like occur data, location, type of the crime, and so on. This project will use those data to analysis the crime rate of Chicago, the crime rate of each community in Chicago, and the crime rate for different time frames. The census data includes all the population of the community, but only two data are used in the project, one is the “total population for each community” and the other is the “community code”. The last data set is community areas data, which contains all community area information, such as community code, community name, geometric information, and so on. The reason of using this data is because there is only community code in crime data sets. This project will analyze the crime rate of each community, and the codes need to be transformed into community names so that it would be intuitive to see which community this is.

The primary method to use in the preprocessing data is Pandas. It is necessary to remove some data since the data is large in the crime data set. For instance, CaseId and FBI code are not useful data sets for this project. The DataFrame data structure was used to save data when extracting data of each community, the data will group by clustering the time and then calculate the rate of population and crime rate for the given time period. Using the total number of crimes of the given time to divided by the total number of crime in the community to get the crime rate for the given time period. The below table is the example of the data of the community.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Community Name | Time | Day | Crime rate | Population rate |
| Loop | Morning | Weekend | 0.2 | 0.2 |
| Loop | Morning | Weekday | 0.1 | 0.2 |
| Loop | Afternoon | Weekend | 0.3 | 0.2 |

Table 1: Example of the data of the community

The classification technique will be used to analyze the date in this project because all attribute is based on the crime rate in this project. In the preprocessing data part, the data will be assigned to different predefined categories, like crime rate with each community, crime rate with different time or day, and crime rate with population density. The results will be easily identified through Bar plot method to show the result of classification.

There is one significant challenge in analyzing data. There is 77 community in Chicago, so there may be huge work to do this for analyzing each community.

The project timeline is shown below. Collect Date, Preprocess Date, and Project intermediate report are complete. Begin with Analyze data and redo some preprocessing steps if needed in the analysis, and the final report will be done when all analysis are done.图片包含 屏幕截图

描述已自动生成