Urban Crime Rate Prediction in Chicago

Chenshu Xu
517-402-9299
1236 Woodcrest Lane Apt 103 East Lansing, MI 48823, United States
xuchensh@msu.edu

Hongyu Yan
517-580-2526
300 Pin Oak Ln Apt 104 East Lansing, MI 48823, United States
yanhong1@msu.edu

ABSTRACT

The goal of the whole project is to develop a effective predictive modeling framework with the historical crimes data, weather data, traffic data and demographic data. The predictive modeling will be used to predict the possibility that a crime will occur with a specific location in Chicago.

The potential contributions of this project is that offering a model for user to predict the possibility of crime in specific location with different time period, weather, traffic condition and day of week. It will let user know the at when and where the crime will happen in a highly chance and they can try to avoid it. Also, it can be used by police and whit the predicted result, the police can prevent the crime effective.

KeyWords

Crime prediction, Weather, Income, Traffic, Chicago, Crime Rate

1. INTRODUCTION

The problem we're investigating are the crime rate. Specifically, we are building a model to predict the crime rate using variables such as time period, day of week, weather conditions etc. The challenge is to handle the missing data. For example, the traffic data for several months are missing. The weather data is huge, has so many types of weather, we have to classify some weather type to a group of weather type. Also there is not weather code for sunny day in the documentation. The traffic data is huge, it takes a long time to merge with its segment location. Also the traffic data only has a last update time, which is all the same. So we cannot get traffic information in each time period. In addition, the location is hard to determine, because it's hard to find a database that have longitude and latitude information of a specific district or street.

2. PRELIMINARIES

We collected traffic data, weather data, census data and also crime data to build the prediction model. The main problem is to pre-process the data. By first looking at the data, we found there are many data in traffic is missing. Also the rest of traffic data is huge, and needs to take a lot time to process. So the pre-processing data efficiency needs to take in consideration.

3. METHODOLOGY

For traffic data, there are mainly two type of files. One is specific traffic flow speed in each segment ID in each day. Another is the table that gives relation on information of segment ID and range of longitude and latitude information. Base on the traffic data, we can get a table with columns: Average Traffic Condition, Segment ID, Start Longitude, End Longitude, Start Longitude, End Latitude. In crime data, it has column "Latitude" and "Longitude" so we can find corresponding traffic flow speed when that crime record happens. However, the are some crime records that don't have match with traffic records. We decede to ignore them, since null value in traffic condition cannot help us building relations in the prediction model.

For census data, we mainly focus on the average income in that community area. Since we have a column "Community Area" in crime data. we can merge then together. It's hard to decide how to classify the average income into "High", "Medium" and "Low" since we cannot find a specific definition of that is "High", "Medium" and "Low" on income. So we decide to classify them in same width. First calculate the max difference of average income between communities, then divide into three. This classification may not be the best way, and can be improved with other ways.

For weather data, it's not regular csv file with specific weather type in that day. The weather type is based on its

weather code and is has 21 types weather in records. So we need to group some of them into a general weather. We group Also there are so many weather stations in the records. We just picked one station in O'Hare International Airport to represent the general weather for Chicago.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Setup

The initial crime data set we have collected in two months ago has a bunch of attributes like crime type, community area, date, location and etc.

We spend a lot of time to decide which attributes should be keeped in the sub-data and are going to be used in combine with other sub-datas.

The initial weather data set we have collected are from 2001 to 2018 which is less than 1 MB. When we were going to preprocess it, we met a risk that the data only contain the station name, date and weather type(WT from 01 to 22) . Also, the different year has the different weather type columns. So, we shrink our original datasets into 2 years (2013 to 2014) and convert those weather type code into the general weather type like “Sunny”, “Cloudy”. “Snow” and “Rain”. The final dataset of weather only contain the weather type and date.

The initial income data set have a lot of attributes which we were not planning to use like “Percent Aged Under 18 or Over 64”. We filter out those attributes and only keep the “Community Area Number”, “Community Area Name” and “Per Capita Income”. Then we use equal frequency approaches to produce “Income” into 3 bins called “High”, “Medium” and “Low”. The final dataset of income only have 3 columns.

4.2. Experimental Results

We didn’t figure out how to do the train model with regression technique. We use classification technique to train the data and calculate the accuracy of decision tree between crime rate and one of traffic condition, income, weather, time of period and day of week. Then we find that the accuracy of traffic condition and day of week have a higher accuracy than other attribute which is 0.755500354862 and 0.734563520227. The lowest is the income which is 0.408209131772.

GitHub link:
https://github.com/Newton222/cse482-project.git

4.3. Discussion

If we can find more data, it will help a lot on accuracy. For example, if we can find every day traffic data with more detailed time information such as traffic condition in each hour in a day, then we can have prediction more accurate. Right now, we are using average traffic flow data in that segment each day, it’s relatively a long time span. Because morning and evening may have different traffic conditions. That can affect the accuracy of prediction. Also for the income data, if we can find more detailed income information of that area, it would be better. Right now, the only data we can find is from 2008 - 2012 census data. It may outdated. Also the area of a community is too big for each crime record.

5. CONCLUSIONS

Collecting data is also a very important part of big data analysis. Detailed data can help us improve the results.

6. REFERENCES

[1] Data.cityofchicago.org. (2018). *Census Data - Selected socioeconomic indicators in Chicago, 2008 – 2012 | City of Chicago | Data Portal*. [online] Available at: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2 [Accessed 30 Apr. 2018].

[2] Data.cityofchicago.org. (2018). *Chicago Traffic Tracker - Congestion Estimates by Segments | City of Chicago | Data Portal*. [online] Available at: https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Congestion-Estimates-by-Se/n4j6-wkkf [Accessed 30 Apr. 2018].

[3] Data.cityofchicago.org. (2018). *Chicago Traffic Tracker - Historical Congestion Estimates by Segment | City of Chicago | Data Portal*. [online] Available at: https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Historical-Congestion-Esti/77hq-huss [Accessed 30 Apr. 2018].

[4] Data.cityofchicago.org. (2018). *Crimes - 2001 to present | City of Chicago | Data Portal*. [online] Available at: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2 [Accessed 30 Apr. 2018].

[5] DOC/NOAA/NESDIS/NCEI > National Centers for Environmental Information, U. (2018). *U.S. Hourly Precipitation Data - Data.gov*. [online] Catalog.data.gov. Available at: https://catalog.data.gov/dataset/u-s-hourly-precipitation-data [Accessed 30 Apr. 2018].

[6] Kaggle.com. (2018). *Crimes in Chicago* | *Kaggle*. [online] Available at: https://www.kaggle.com/currie32/crimes-in-chicago [Accessed 30 Apr. 2018].