Urban Crime Rate Prediction Analysis

Ibraheem Saleh
Project URL: https://github.com/salehibr

ABSTRACT

The goal of this project is to develop a framework for crime prediction using diverse types of data available (crime date, crime time, type of crime, day of week the crime occurred, and the neighborhood in which the crime occurred). Specifically, given a location, time of day, and day of week, and auxiliary data about the location, and predict the likelihood that a crime will occur. If the model developed is accurate it can then be applied to all locations in a region for a future time period.

1. INTRODUCTION

For introduction, you need to include the following information:

1. Crime prediction is an important application of big data especially for law enforcement in large cities. This problem is interesting because of how volatile the data sets can be especially switching between regions and cities (depending on population, population density, poverty levels, etc). This project specifically uses crime date, crime time, type of crime, day of week the crime occurred, and the neighborhood in which the crime occurred to find the probability of a crime occurring at that location the following year at that exact date.
2. The goal of this project is to form a prediction model that is accurate enough to estimate and offer an educated and calculated prediction of when and where a crime will occur and what probability is has of actually occurring. By using a very large data set, the prediction model has a higher chance of being more accurate and providing usable prediction data to support law enforcement.
3. I have collected over 184000 crime occurrences with their related data from the Detroit area and used it to form the prediction model through different analysis and models.
4. Finding relevant data to use for the prediction model was a challenge as different models require different pieces of data to truly be accurate enough and output usable data. I used data that could be processed by different models in order to obtain a model in the event any specific organizational or analytical model was used.
5. The prediction model I have formed reported a ~71% accuracy rate when predicting if a crime would occur at a specific time and place. With a larger data size this would be more accurate a there would be less room for error with a higher number of crime occurrences.

CSE881-2015, Month 1–2, 2004, City, State, Country.

2. DATA

I specifically used a data set provided by the Detroit Police Department, it contains crime data dating back to 2016 in ever neighborhood in the Detroit area for every major crime reported. (Source cited below in references) I downloaded the .csv file and sifted through the columns to rewrite a file that contains the columns of data that were going to be used for the predictive algorithm.

Crime date, crime time, type of crime, day of week the crime occurred, and the neighborhood in which the crime occurred are the data points in which I was interested in. They were the only things necessary for this predictive algorithm. Some of the data issues were that certain crimes were never reported in certain time periods but then reported in other time periods in a different year, causing a reduction in accuracy for predicting that specific crime at that specific time. The data consisted of about 184000 lines of data dating back to 2016.

For preprocessing, the csv file was sifted through and the unnecessary columns removed to save from the data size and reduce the run time for the algorithm. No missing values were needed to be filled in the file. I tested a smaller sample of the data (5000 rows) and the algorithm reported an accuracy of 69%, and after using the entire data set, the accuracy level increased to ~72%.

A predictive system was created from a recommendation system, and it was able to use the provided data to calculate and estimate predictions of crimes in locations. The crime data consisted of every neighborhood in which a crime occurred in Detroit and the types of crimes ranged from theft, assault, larceny, burglary, and damage.

The attributes in the table were Incident Date, Incident Time (based on 2400 clock), Offense Category, Day of Week (weekday or weekend), and Neighborhood in which the crime took place. There were 5 columns and slightly above 184000 rows of data to process. This data set .csv is 8.7mb in size.

3. METHODOLOGY

I used decision tree classifiers to perform my classification with the assistance of the Python Surprise toolkit to handle recommendations into the prediction model using the non-negative matrix factorization approach to predict.

The test set was created from the data set being used of all the crime data provided of Detroit crime over the past 3 years. First the data was tested in small increments (5000 rows at a time) then the entire data set was processed and a higher accuracy rate resulted. KNeighbors was used for clustering (knn set to the value of 100).

- ClassProject.ipynb: this is the Jupyter notebook file that I wrote to process the entire data set and predict the crime data.
- EntireDataSetDetroit.csv: this is the .csv file containing all the crime data from Detroit all the way back to 2016. Certain pieces of data were manually changed (day of week in number form changed to weekday/weekend, date and time column changed to just date)

4. EXPERIMENTAL EVALUATION

For the experimental setup, sample sizes of 5000 rows were used to ensure that accuracy remained high and did not lower to the point where the data could not be looked at for prediction analysis. The accuracy ratings stayed above 66% for these smaller data sizes but never exceeded 70%.

4.1 Experimental Setup

This section should include:

1. Worked on Mac OSX, using Jupyter notebook to run python commands and scripts to process the data file.
2. Based off other research studies and algorithms present online, the accuracies of most project were very high (usually exceeding 80% and sometimes 90%) but this specific case (due to keeping the tested time period at 24 hours) the accuracy remained below 80%
3. Accuracy was used as a measurement for this.

4.2 Experimental Results

The experiments began with sample sizes of 5000 rows of data and then the sizes were increased for every run afterwards to ensure accuracy remained at a satisfactory level. The results were an output of every crime and location pair and the probability they had of occurring again within that time frame again the following year.

The project was not as successful as wanted. Accuracy did not reach a point where I would recommend police departments to use this as a prediction model for future crimes. With more time and further analysis and utilizing other prediction techniques a stronger and more accurate algorithm can be developed and used to assist in these predictions.

5. CONCLUSIONS

The projects findings were sub-par at best. The accuracy was above 70% in the end but that still only provides an “educated analysis” and with further work done a more beneficial algorithm can be developed. Future work on this should use other prediction techniques and factoring in auxiliary data that could influence the possibility of that crime occurring.

6. REFERENCES (at least 3 references)

[1] Vikram-Bhati. (2018, May 30). Vikram-bhati/PAASBAAN-crime-prediction. Retrieved from https://github.com/vikram-bhati/PAASBAAN-crime-prediction

[2] 7cb15. (2019, January 06). 7cb15/Predicting-Crime-in-Toronto. Retrieved from https://github.com/7cb15/Predicting-Crime-in-Toronto

[3] Nishi1612. (2018, November 02). Nishi1612/Crime-Type-Prediction. Retrieved from https://github.com/nishi1612/Crime-Type-Prediction

[4] Data for Detroit: https://bit.ly/2WTJm7P

[5] Professor Tan – CSE 482, Exercise 6 and 8

Grading criteria

Note that the project accounts for 10% of your final grade. The project will be graded based on the following criteria:

1. Presentation - structure/organization and clarity of writing (including tables and figures).
2. Technical - Correctness and thoroughness of the analysis performed. What are the challenges faced and how well did you address them? How do you evaluate the performance of the method you'd applied to the data? How much detailed discussion you provide to explain the results you'd obtained (e.g., discussion about why the method works or didn't work on the data)?
3. Difficulty level - How large is the dataset used? How much effort you had to spend to collect, integrate, preprocess, and analyze the data? Are you implementing the project on a cluster or a single machine? What tools did you use (do you have to implement them or are you simply using existing libraries)?