Midterm Report: Predictive Analysis of City Crime Hotspots

Sahil Deshpande, Vinit Kumar
{*sahilsd, vinitn*}@*bu.edu*

***Abstract*—With increased digitization of data around the world, various type of data is now easily available. This includes criminal reports for most of the major cities in the US. In order to constructively use this data and the available data mining tools in Python, we are trying to analyze crime hotspots and predict next most likely crime location.**

I. Introduction

Crimes have become a common part of city life that seriously affect quality of life and economic growth of a society. Universal organizations are spending a lot of resources trying to identify safest and most dangerous cities to help local authorities manage their workforce. As more and more people shift to the city, this concentrated population makes it important to find safer places. Using the open-source data from official websites and some of the basic data mining approaches, we are attempting to help this process [1].

From past analysis of various cities and their neighbourhoods, it has been shown that certain parts of the city are more prone to criminal activities than others making them a criminal hotspot. Even though there doesn't seem to be an intuitive pattern for criminals, they tend to favor certain areas that makes it possible to predict such malicious activities. Using the information about past activities, law enforcements can effectively serve their duties.

To achieve this predictive analysis, we have looked at various datasets available for free on the Internet[2-3]. After some research on which cities have most relevant and most updated data, we have decided to work with Boston, Los Angeles and Raleigh. The first part of this project is to get the data in a specified format (in this case json) and parse it to extract meaningful information. Using these parsed data structures, we are interested in finding the most relevant crime types such as assault, robbery, hit and run, to name a few. We select the types that affect a particular area the most using regression techniques[4]. After trimming the dataset, we have applied k-means clustering to plot these various types of activities using GMplot[5]. We now plan to use classification methods in order to predict most vulnerable parts of the city. Later sections describe the results we've seen so far.

II. Technique

First we searched on the Internet for reliable data sources for criminal activities in major cities in US. After going through large number of datafiles, we decided to work with data from [3]. After getting the data in json format, we parsed it to get attributes such as crime location, latitude and longitude, time of crime, day of the week, type of the crime. Since these attributes help us distinguish between most of the important crime activities, we plan to use these attributes for further analysis.

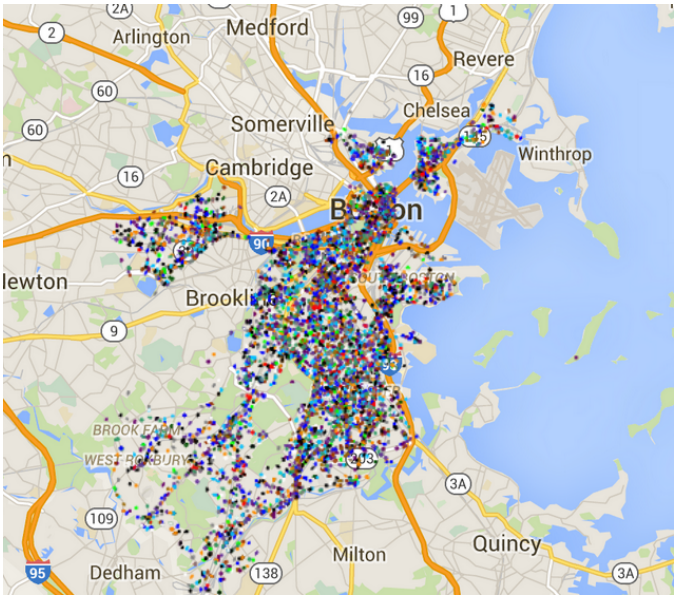

Fig. 1. Overall crime distribution across Boston shows concentrated crime data.

While parsing the data, we have also divided the timestamps into six categories and day of the week as weekday or weekend. This further enables more meaningful clustering because coarse datasets are clustered more effectively.

In order to reduce this large data, we are now planning to use linear and logistic regression to find which crime types affect an area the most. Here we can calculate an average score for one area based on crime types (as we did in homework) and get top five coefficients that will denote the most common crime types. As we know, more data helps in better analysis and prediction we wanted to minimize this data reduction and thought this technique would be most effective.

For now, we have skipped the regression step and directly went on to categorize the data in types like assault, robbery, disturbance, white collar crime and others. But this can be easily replaced with the types we get from regression results so we are first focusing on analysis part. Using these categories, we have applied K-means clustering from Python sklearn. The clustering results are explained in the next section. We have

```
------------------------------
Assault : 12228
DrugAlcohol : 0
Othercrimes 0
PublicDisorder 0
Theft 0
WhiteCollarCrime 0
T1 926
T2 845
T3 2399
T4 3107
T5 2937
T6 2014
Weekday 12228
Weekend 0
------------------------------
```

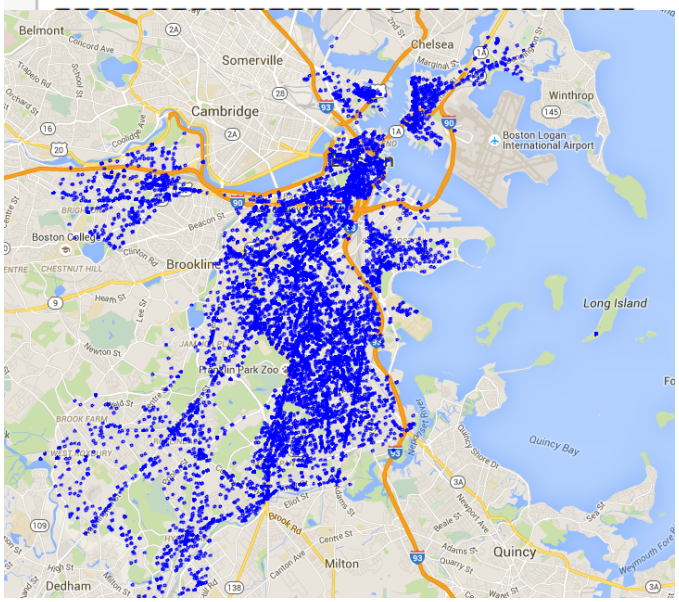

Fig. 2. Cluster showing only Assault crimes and only on weekdays.

```
Assault : 0
DrugAlcohol : 593
Othercrimes 7510
PublicDisorder 51
Theft 174
WhiteCollarCrime 271
T1 0
T2 8599
T3 0
T4 0
T5 0
T6 0
Weekday 0
Weekend 8599
------------------------------
```

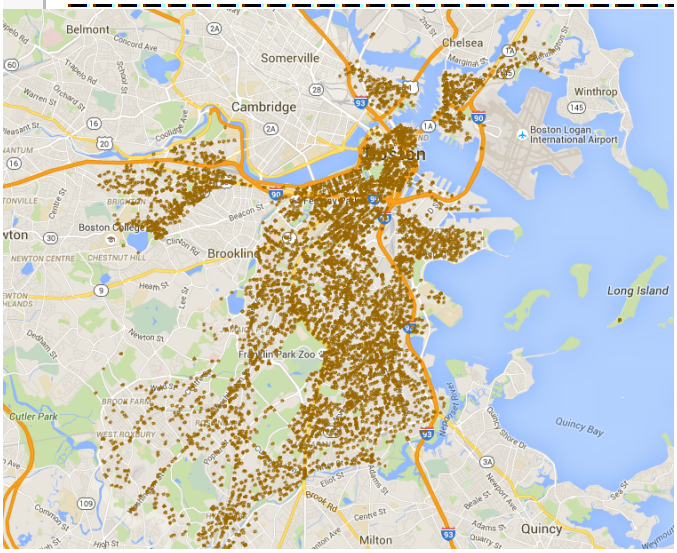

Fig. 3. Cluster showing only crimes only on weekends.

normalized the latitude and longitude during clustering so that they don't add any extra weight. While plotting the clusters on the map, we have used the original latitudes and longitudes to give correct crime hotspots.

So this has given us a basic idea of malicious activities and their patterns in Boston. We can easily identify which areas show more assaults or weekend crime. This example already starts to unfold the advantages after seeing which areas are relatively safer and which would need more attention of local law enforcement. Using this clustered data, we are now exploring classifier methods in python which were discussed in the class. Right now we have come up with two approaches on how we can predict a crime but this part needs more work. First approach is to directly use latitude/longitude for predicting most likely crime activity but this has very fine granularity which makes it less favorable. On the other hand, we are planning to utilize area codes in the data set. This divide the city in certain code like A4, G1, etc. We can use these and one of the six timestamps and try to get a tuple similar to, ¡type of crime, area code, time category, day category¿. We think this will be more clear after the Amazon prediction assignment. We plan to revisit this once more.

III. Experiments

For initial phase of analysis, we tried to focus on dataset for Boston. The initial code is available in a private repository at [6]. This stage consists of parsing data, reducing data and

clustering. After reviewing the dataset, we managed to select few crime types for initial experiments. Using these types along with time of day, day of the week and latitude/longitude pair we executed K-means clustering and plotted it using GMplot.

IV. Results

As discussed before, the elementary results give a basic idea about different types of crime in the city and their distribution. For example, after plotting the clusters, we could point out which area has which type of crime as prominent activity. We can also see which areas show troubles only on weekends or only late at night. This analysis itself gives residents a slightly better picture of which areas to avoid at what times and hopefully will help law enforcers to look after these hotspots.

In Figure 1, where there are less dots we can most probably say that these areas are the safest areas in Boston. Figure 2 shows one of the clusters and crime statistics along with timestamps. Here, we can see the areas that experienced only $Assault$ crimes and only on weekdays. Similarly, we plotted a few more clusters and Figure 3 shows the areas that experienced crimes only on a weekend and during a particular time (5 a.m to 9 a.m) .

V. Conclusion

So far the we have completed collecting required data and parsed it to extract required information for Boston. The primary analysis and results are encouraging enough to pursue classification techniques to get more insight on most likely next crime. We would also like to believe our future work will reach the levels of $Minority\ Report(2002)$[7] which is where we got this idea from.

References

[1] CRIME PREDICTION BASED ON CRIME TYPES AND USING SPATIAL AND TEMPORAL CRIMINAL HOTSPOTS Tahani Almanie, Rsha Mirza and Elizabeth Lor Department of Computer Science, University of Colorado, Boulder, USA.

[2] Crimereports.com, 2015. [Online]. Available: https://www.crimereports.com. [Accessed: 20- May-2015].

[3] 'Crime Datasets - US City Open Data Census', Us-city.census.okfn.org, 2015. [Online]. Available: http://us-city.census.okfn.org/dataset/crime-stats. [Accessed: 20- May- 2015].

[4] Regression techniques reference. http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html

[5] GMPlot reference. https://pypi.python.org/pypi/gmplot/1.0.5

[6] GitHub code repository. https://github.com/sahilsd/591-crime-project

[7] Minority Report (2002) http://www.imdb.com/title/tt0181689/