CIS 635 Term Project

By Zhen Lu

Topic: Crime Forecasting

Introduction

Crime prevention is consistently a paramount concern for communities and law enforcement agencies worldwide. Accurate crime forecasting plays a crucial role in ensuring public safety, and extensive research has been conducted to enhance the precision of crime predictions. Many studies have compared various models for accuracy. However, a significant question remains: how much data should be used for the best result?

It is widely acknowledged that crime hotspots are more likely to experience recurring criminal activities. This project aims to investigate whether the length of data used impacts the accuracy of crime forecasting. To address this issue, I will adopt a data-driven approach that encompasses the collection and analysis of historical crime data, thorough preprocessing, model development, and rigorous validation.

Our primary goal is to compare the final accuracy of models that use different time durations for training and the same duration but predict different time periods. By doing so, I aim to uncover the potential improvements in accuracy brought about by considering how much reference data should be used for prediction. As a result, I found that accuracy increased with the prediction length, and with more historical data, the accuracy improved as well. However, the significance of the length of historical data usage is low.

Related Work

Numerous research papers have addressed the task of crime prediction. For instance, Zhuang (2017) leveraged historical data to construct a deep recurrent predictive model and then compared its performance with that of conventional algorithms. Similarly, Zhang (2020) adopted a similar approach by evaluating various algorithms on historical data to identify the most accurate machine learning algorithms for crime prediction. Wilpen (2003) studied a short-term, univariate method to proceed with crime prediction. He highlighted the limitation of small-scaled data and emphasized the importance of increasing scale for improved forecast accuracy, which will support my project.

In this project, I build upon the findings and methodologies of previous research. My focus is on whether the number of historical data will affect the output accuracy. However, my investigation deviates from previous studies, as I aim to explore the impact of the length of data to use on the accuracy of crime prediction and how the length of prediction will affect accuracy. This unique perspective seeks to contribute to a deeper understanding of the intricacies of crime forecasting.

Methods

I used data from the National Institute of Justice, specifically calls-for-service data from Portland, Oregon. The primary reason for selecting this dataset is its inclusion of Spatio-Temporal Data and historical data since 2012, making it well-suited for my analysis of crime hotspots and the impact of the length of training data on prediction accuracy. In preparing the dataset for analysis, I anticipated the need for the following preprocessing steps: data cleaning, data transformation, and data reduction.

First, I imported all the data, focusing on three attributes: x-coordinate, y-coordinate, and occ-date. I used Python libraries Pandas and NumPy; for the libraries to work, I transformed all the dates to datetime format, filtered the desired time frame, and transformed all dates to numerical format. I also dropped all empty or invalid dates to avoid errors during training. Next, I used the NumPy library to add all data into a grid-like 3D array, with the grid dimensions shown below:

Table 1. Grid dimension

	Min	Max
x	7603950	7717500
y	651190	733990

I used a resolution of 600x600 for the grid, resulting in a grid with 189x138 cells. In each cell, it contains the number of calls for service cases. I then ranked the cells to get the top 100 cells with the highest case count. I used these cell coordinates to predict the next top 100 crime hotspots and compared them with the actual top 100 hotspots. Prediction Accuracy Index (PAI) and Prediction Efficiency Index* (PEI*) were calculated for numerical comparison.

There are two measures, PAI and PEI, as shown below:

A) Prediction Accuracy Index (PAI)

- The *PAI* will measure the effectiveness of the forecasts with the following equation:

$$PAI = \frac{\frac{n}{N}}{\frac{a}{A}}$$

- Where *n* equals the number of crimes that occur in the forecasted area, *N* equals the total number of crimes, *a* equals the forecasted area, and *A* equals the area of the entire study area.

(B) Prediction Efficiency Index* (PEI*)

- The *PEI** will measure the efficiency of the forecast with the following equation:

$$PEI^* = \frac{PAI}{PAI^*}$$

- Where *PEI** equals the maximum obtainable PAI value for the amount of area forecasted, *a*. As such:

$$PEI^* = \frac{n}{n^*}$$

- Where *n** equals the maximum obtainable *n* for the amount of area forecasted, *a*.

Results and Discussion

In general, the accuracy increased as the predicting for a longer time period. For training set, I used 2012 whole year data as input, and for testing set I used as shown in table 2.

Table 2. 1 year training data and prediction

Time	**Result**
14 days	Total count using indices from the last layer: 80.0 Total count for the actual top 100 cells: 14337.0 Sum of Points: 146927.0 points PEI: 0.005579967915184488
30 days	Total count using indices from the last layer: 2165.0 Total count for the actual top 100 cells: 16432.0 Sum of Points: 146927.0 points PEI: 0.13175511197663098
90 days	Total count using indices from the last layer: 6370.0 Total count for the actual top 100 cells: 20596.0 Sum of Points: 146927.0 points PEI: 0.3092833559914547

150 days	Total count using indices from the last layer: 10912.0 Total count for the actual top 100 cells: 25115.0 Sum of Points: 146927.0 points PEI: 0.43448138562611976

as we see from the table, the PEI value increase as the predicting time increase, which means the prediction gets more accurate. Then I changed the input set, I changed the training set to half year only instead of whole year and the result are shown in table 3.

Table 3. half year training data and prediction

Time	**Result**
14 days	Total count using indices from the last layer: 1052.0 Total count for the actual top 100 cells: 15523.0 Sum of Points: 89922.0 points PEI: 0.06777040520517941
30 days	Total count using indices from the last layer: 2183.0 Total count for the actual top 100 cells: 16432.0 Sum of Points: 89922.0 points PEI: 0.13285053554040896
90 days	Total count using indices from the last layer: 6346.0 Total count for the actual top 100 cells: 20596.0 Sum of Points: 89922.0 points PEI: 0.3081180811808118
150 days	Total count using indices from the last layer: 10872.0 Total count for the actual top 100 cells: 25115.0 Sum of Points: 89922.0 points PEI: 0.4328887119251443

Interestingly, I found that the result end up very similar, and for 14 days prediction, half year data set have a better performance than 1 year data set. I think the reason for that is what Wilpen (2003) said, small scaled data are full of randomness, so it could be that time frame it fits better with half year data prediction.

Then I slice the data into layers, I used the previous layer to predict for next one, which means my training set also start from 14 to 150. And table 4 shows what I found: Table 4. 2012 Street Crime Prediction with different time frame

Table 4. **2012 Street Crime Prediction with different time frame**

Time Frame = 14	Time Frame = 30	Time Frame = 90	Time Frame = 150
Layer 1 - Time Interval: 2012-03-01 to 2012-03-15 Total count for the actual top 100 cells: 299.0 Sum of Points: 910 points Layer 2 - Time Interval: 2012-03-15 to 2012-03-29 Total count using indices from the last layer: 144.0 Total count for the actual top 100 cells: 285.0 Sum of Points: 920 points PEI: 0.5052631578947369 Layer 3 - Time Interval: 2012-03-29 to 2012-04-12 Total count using indices from the last layer: 153.0 Total count for the actual top 100 cells: 298.0 Sum of Points: 904 points PEI: 0.5134228187919463 … … Layer 21 - Time Interval: 2012-12-06 to 2012-12-20 Total count using indices from the last layer: 250.0	Layer 1 - Time Interval: 2012-03-01 to 2012-03-31 Total count for the actual top 100 cells: 524.0 Sum of Points: 1941 points Layer 2 - Time Interval: 2012-03-31 to 2012-04-30 Total count using indices from the last layer: 398.0 Total count for the actual top 100 cells: 586.0 Sum of Points: 2215 points PEI: 0.6791808873720137 Layer 3 - Time Interval: 2012-04-30 to 2012-05-30 Total count using indices from the last layer: 411.0 Total count for the actual top 100 cells: 583.0 Sum of Points: 2273 points PEI: 0.7049742710120069 … … Layer 10 - Time Interval: 2012-11-26 to 2012-12-26 Total count using indices from the last layer: 451.0	Layer 1 - Time Interval: 2012-03-01 to 2012-05-30 Total count for the actual top 100 cells: 1498.0 Sum of Points: 6429 points Layer 2 - Time Interval: 2012-05-30 to 2012-08-28 Total count using indices from the last layer: 1454.0 Total count for the actual top 100 cells: 1673.0 Sum of Points: 7581 points PEI: 0.8690974297668859 Layer 3 - Time Interval: 2012-08-28 to 2012-11-26 Total count using indices from the last layer: 1821.0 Total count for the actual top 100 cells: 2102.0 Sum of Points: 9051 points PEI: 0.8663177925784968	Layer 1 - Time Interval: 2012-03-01 to 2012-07-29 Total count for the actual top 100 cells: 2537.0 Sum of Points: 11440 points Layer 2 - Time Interval: 2012-07-29 to 2012-12-26 Total count using indices from the last layer: 2390.0 Total count for the actual top 100 cells: 2633.0 Sum of Points: 11621 points PEI: 0.9077098366881884

Total count for the actual top 100 cells: 471.0 Sum of Points: 1693 points PEI: 0.5307855626326965	Total count for the actual top 100 cells: 617.0 Sum of Points: 2332 points PEI: 0.7309562398703405		

As the table shows, when training data is small and prediction is also not very accuracte, as the training data size increase, it gets more accurate. Then I used more data to verify as shown below:

Table 5. 2012-2013 Call for Service perdition with 90 days' time frame

```
Layer 1 - Time Interval: 2012-03-01 to 2012-07-29
  Total count for the actual top 100 cells: 12221.0
  Sum of Points: 72140 points

Layer 2 - Time Interval: 2012-07-29 to 2012-12-26
  Total count using indices from the last layer: 11872.0
  Total count for the actual top 100 cells: 12261.0
  Sum of Points: 72361 points
  PEI: 0.968273387162548

Layer 3 - Time Interval: 2012-12-26 to 2013-05-25
  Total count using indices from the last layer: 10752.0
  Total count for the actual top 100 cells: 11274.0
  Sum of Points: 66069 points
  PEI: 0.9536987759446512

Layer 4 - Time Interval: 2013-05-25 to 2013-10-22
  Total count using indices from the last layer: 18354.0
  Total count for the actual top 100 cells: 19051.0
  Sum of Points: 110403 points
  PEI: 0.9634139940160622
```

With 90 days of time frame give the model enough data points to be accurate enough and averaging I got 96% PEI, which indicate that prediction and actual are almost the same.

Conclusion

In conclusion, the study revealed that utilizing historical data facilitates more accurate predictions for longer time frames of crime hot spots, and leveraging the seasonality pattern by using the last time frame to predict the next one yielded the best results. However, limitations arise from relying solely on historical data, emphasizing the need for a more robust algorithm. Dependence on historical data may not capture sudden changes or emerging patterns, preventing the learning of features. The model is static and does not learn patterns or features from historical data; it purely depends on locations with the highest crime count. Despite its high accuracy rate, using historical data for prediction can serve as a baseline model for reference. For future work, I attempted to add a CNN model to predict future cell counts; however, there was confusion about how to split the Spatio-Temporal Data into a training set and test set, and I was not successful in training the model. Therefore, it would be worthwhile to continue working on using CNN to train the model and assess the results.

Data and Software Availability

In this section, the link to the project Github page is provided, as shown below:

Github link

And in the Github repository there are two test files included to run, and more data can be downloaded from the following:

Call for service data

References

- Gorr, Wilpen, Andreas Olligschlaeger, and Yvonne Thompson. "Short-term forecasting of crime." *International Journal of Forecasting* 19.4 (2003): 579-594.

- X. Zhang, L. Liu, L. Xiao and J. Ji, "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots," in IEEE Access, vol. 8, pp. 181302-181310, 2020, doi: 10.1109/ACCESS.2020.3028420.

- Y. Zhuang, M. Almeida, M. Morabito and W. Ding, "Crime Hot Spot Forecasting: A Recurrent Model with Spatial and Temporal Information," 2017 IEEE International Conference on Big Knowledge (ICBK), Hefei, China, 2017, pp. 143-150, doi: 10.1109/ICBK.2017.3.