San Francisco Crime Classification and Prediction

Kai Wang
Department of Computer Science, UCLA
kwang42@cs.ucla.edu

***Abstract*—In this paper, we apply machine learning and time series analysis to the problem of classifying and forecasting crime incidents in San Francisco. Our dataset originates from a Kaggle competition [1]. Based on existing researches on these problems, we employ Logistic Regression and VAR(p) models respectively. The classification problem result demonstrates high challenge as our results across all the 39 crime categories achieved the accuracy of 32.67%. For the time series analysis, we revealed correlations of occurrences of difference crime categories. And the VAR(p) model can forecast the trend of number of certain crimes. We also involved demographic data for classification problem, which contributed to a slight improvement to accuracy. With richer and more up-to-date data, our results might be further improved.**

Keywords: Crime Prediction; Feature Engineering; Classification; Time Series Analysis

I. Introduction

Crime activities have always been threat to public safety, and researches on criminals have long been limited within the realm social science. With the huge volume of existing crime datasets and the advent of robust machine learning and statistical technologies, we seek to predictively analyze crime incidents. Our vision is to help police, public policy makers as well as individuals understand crime patterns and take more effective measures to prevent crimes.

In our project, we combined spatio-temporal and demographic data to predict which category of crime has the largest possibility to occur, given a timestamp, location and demographics of the surrounding area. The inputs to our classification algorithm include time (hour, day, year), location(street, latitude, longitude and police district) and demographic data(population density, median house value, poverty rate and median per capita income). The output is the type of crime that is most likely to have occurred. We performed complicated feature engineering to enrich predictors. We also tried various classification algorithms, among which Logistic Regression achieved highest accuracy.

For time series analysis, we applied vector autogression model to capture the interdependencies among multiple categories of crime time series. We tested for the optimal lag p for VAR(p) model. The input to VAR(p) is spatio-temporal data; the output is the time series model. For some crime categories, this model can reveal the trends of its occurrence.

II. Datasets

Our primary dataset is a training set obtained from Kaggle, which includes 878,049 crimes in San francisco that occurred from 2003 to 2015. Every crime record was marked a category, and there are in total 39 categories. The other dataset is the demographic data of different neighborhoods in San Francisco, including population, income, race constitution, education level etc. The demographic data is provided by US Census conducted in 2010. We also searched Wikipedia for population density.

A. Features

Each record in our dataset correspond to a particular crime incident, which includes following fields:

- Date and timestamp of the crime
- Day of the week of the incident
- The third etc ...

B. Preprocessing

III. Methods

1) Depth map computation: example of table end

TABLE I
Logfile Format

frameId		TimeStamp	IsKeyframe	
(Joint0) JointType	X	Y	Z	TrackingState
...				
(Joint19) JointType	X	Y	Z	TrackingState

example

```
float calc_offset_with_angle(Skeleton s1,
    Skeleton s2){
  float TotalOffset = 0;
  foreach(Angle A1 in s1){
```

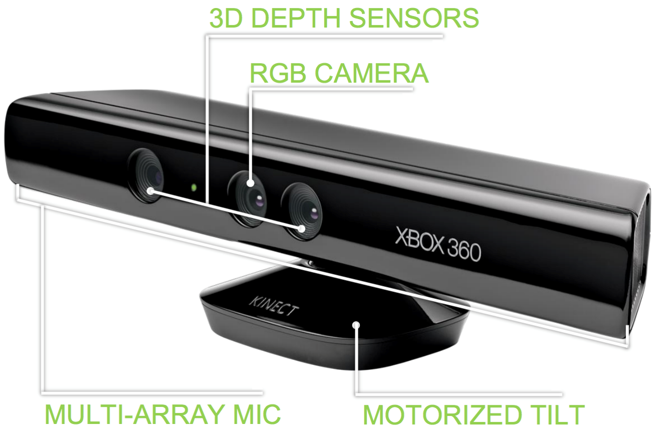

Fig. 1. Depth Sensor in Xbox 360

```
    Let A2 be the cooresponding Angle in
       s2.
  TotalOffset += Weight_of_A1
 * abs(A1 - A2)
}
  return TotalOffset/NumberOffset
}
```

$$\mathbf{w} * x - b = 0$$

example of inline formula: $(\mathbf{w} * x' - b)$ end example

IV. Experiment

V. Related Work

VI. Conclusion

VII. Acknowledgment

I would like to deeply thank Professor D. Stott Parker for his advice on feature selection and preprocessing, which facilitated my preprocessing with demographic data.

References

[1] https://www.kaggle.com/c/sf-crime