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
- Category of the crime. Which is the label to be predicted in test data set.
- Detailed description of the crime (only provided in train data)
- Day of the week of the incident
- Name of the Police Department District the took charge of the crime.
- Resolution, which means how the crime was resolved(booked, arrested or resolved, etc)
- Approximate street address of the crime
- Longitude
- Latitude

Among these fields, resolution is discarded as it only exist in train data.

B. Preprocessing

While http://datasf.org also includes similar data, our dataset from Kaggle is more selected as there is no empty cell in any fields. However, some fields contain

unreasonable values due to probably wrong registration. Some numeric fields need to be scaled while some fields are not yet numeric. To transform the original data to suitable input of classification algorithms, we performed feature engineering described below.

1) Outlier removal: The following pictures are scatter plots of scaled longitude(X) and latitude(Y). There are in total 67 outlier points, all of their latitudes are beyond the border of San Francisco. After removal of these outliers, the scatter plot fits the territory of San Francisco.

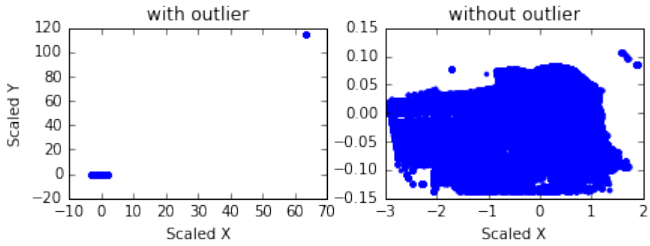

Fig. 1. Scatter plot of crime locations before and after outlier removal

2) Demographic data: Our demographic data is based on US Census from 2006 to 2010. It includes demographic data in each neighborhood of San Francisco, covering race, education, income and housing characteristics. Following heat displayed the correlation among the frequency of some typical crime categories and demographic statistics. For example, poverty rate is positively correlated to most crime categories, especially embezzlement; while high population density tend to relate to drug use and disorderly conduct.

One challenge of employing demographic data is correctly matching it the the crime location as the crime dataset does not include neighborhood. Our solution was to download the map data of San Francisco, where each neighborhood is represented as a polygon. Following pseudo-code shows how we located each neighborhood.

```
bool getNeighborhood(Lang, Lat, Polygon){
  //For acceleration
  if (Lang, Lat) in Polygon-Bounding-Box{
    if (Lang, Lat) in Polygon{
      return True
    }
  }
  return False
}
```

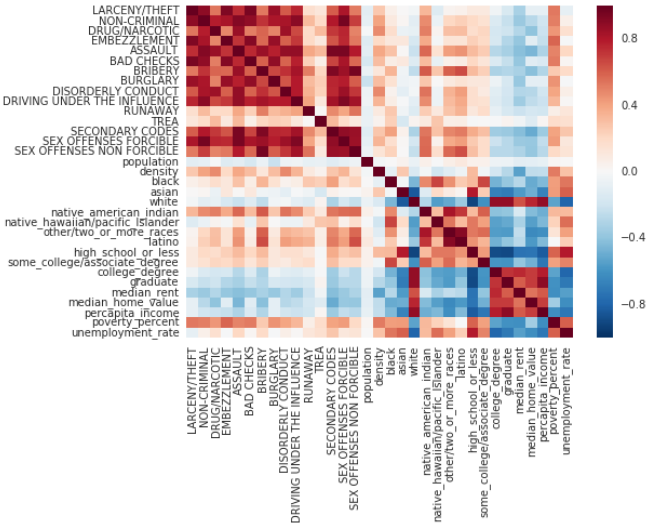

Fig. 2. heat map of correlation among crime amount and demographic data

3) Feature enrichment: For timestamps in dataset, we parsed the monthdate-of-year, hour-of-day from them. We also added season, day/night as new fields.

Another feature enrichment approach is computing the log-odds defined below, where c is crime category.

$$logodds\_c = log(\frac{c\_counts}{len(train\_data)}) - log(1 - \frac{c\_counts}{len(train\_data)})$$

There are around 23,000 addresses in the train dataset, which is small compared with number of crime entries.

4) Scaling and Transformation: Some fields are skewed, with log transformation their distribution are more close to Gaussian Distribution. Following figure shows the log transformation of percapita-income. Similar transformations are performed on median-rent and population density.

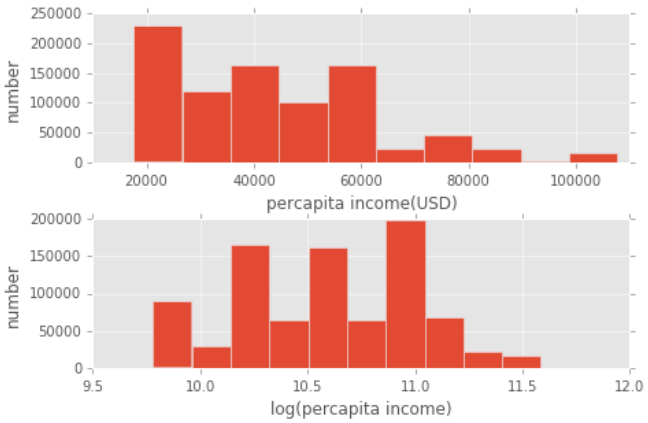

Fig. 3.

III. Methods

A. Naive Bayes

B. Logistic Regression

C. Random Forest

D. Adaboost

E. Time Series

A time series is a sequence of data points with continuous time interval and successive measurements made over a time interval. Time series analysis comprises approaches for extracting meaningful statistics and other characteristics of the data. And time series forecasting is using a model to predict future values based on previously recorded data. It has been applied to crime analysis for decades [2].

VAR models (vector autoregressive models) are desiged for multivariate time series. It view each variable as a linear combination of past lags of itself and past lags of other variables. The variables(number of crimes in our case) are collected in a 39*1 vector. At time t, define the vector as y_t, a p-th order VAR, namely VAR(p), is

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + ...A_p y_{t-p} + e_t$$

where y_{t-j} is called the $j-th$ lag of y, c is constant intercept and e_t is an error term. The VAR model assumes the data is **weakly stationary** and above attributes of e_t must hold

1) $E(e_t) = 0$, zero mean error
2) $E(e_t e_t') = \Omega$, the contemporaneous covariance matrix of e_t is positive semidefinite.
3) $E(e_t e_{t-k}') = 0$, for any positive k. There is no series correlation in error term.

An Augmented Dickey-Fuller unit root test [3] is often employ to test stationarity. Using the library from statsmodel [4] , we performed the ADF test on each category with constant, linear and quadratic trend. We chose BIC(Bayesian Information Criterion) for penalty strictness. (results in Fig 1). There are 23 categories with scores below 5% level, and we selected these categories for time series model construction.

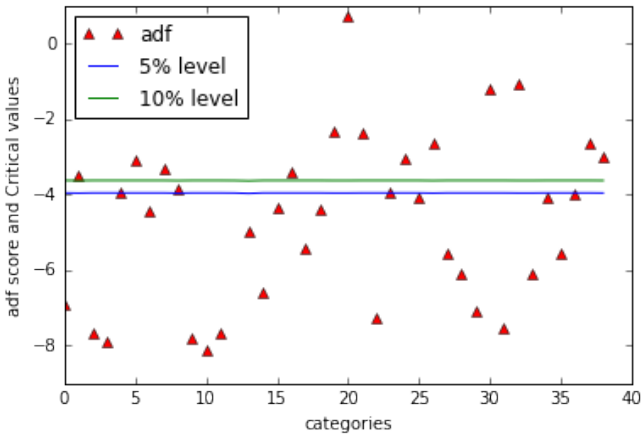

Fig. 4. ADF test score against 5% and 10% threshold

IV. Experiments

A. Time Series

1) Motivations: Based on the stationarity test in previous chapter, the VARp model is suitable for finding trends in the crime dataset. We selected 28 categories with low ADF score, and computed the pair-wise correlation among them. Following picture shows the heat-map of correlation matrix which computes the Pearson's r. In the heat-map, we can find large positively correlation coefficients, for example, BURGLARY & ASSAULT, as well as negative coefficients like DRIVING UNDER THE INFLUENCE & VEHICLE THEFT.

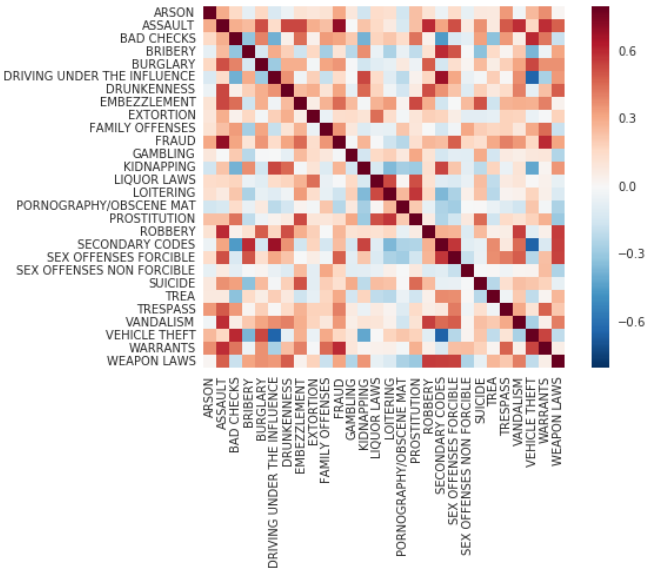

Fig. 5. Heatmap of time series correlation matrix

Auto-correlation function(ACF) and Cross-correlation

function(CCF) are also effective approaches to reveal statistical connections. Following picture demonstrate the CCF between category 'DRIVING UNDER THE INFLUENCE' and 'VEHICLE THEFT'. The increase of the first category precedes the decrease of the other. These kinds of correlations are further exploited in the correlation matrix.

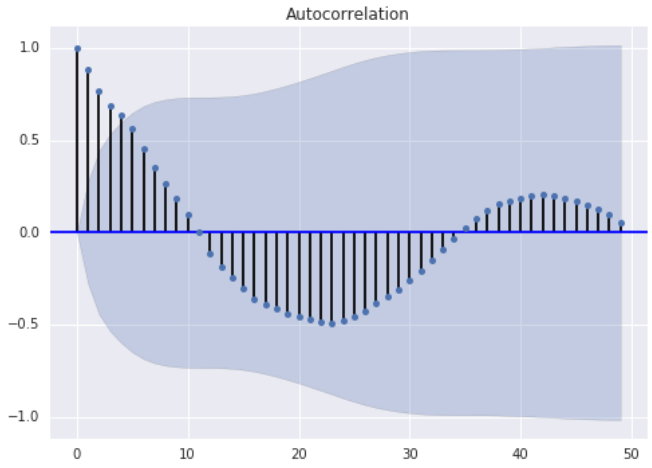

Fig. 6. CCF between 'DRIVING UNDER THE INFLUENCE' and 'VEHICLE THEFT'

2) Model Selection: Choosing the order p for VAR(p) model is a feature selection problem. We tried different values of p under two penalty criterions AIC(Akaike Information Criterion) and BIC. As is shown in following picture, with AIC, the penalty score is minimal when $p = 2$. For BIC, the penalty is 30.88 at $p = 1$ and 30.89 at $p = 2$.

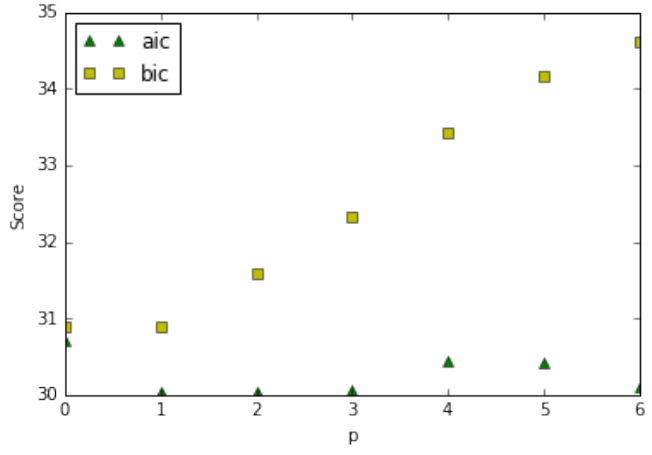

Fig. 7. aic/bic

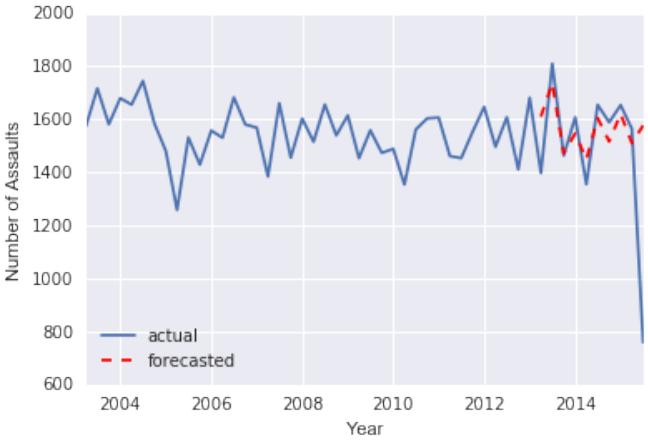

Fig. 8. Forecast of 'ASSAULT'

3) Fitting Model: We choose $p = 2$ and AIC penalty criterion for VAR(p) model fitting. And tried the model on "ASSAULT" time series. The first 40 time units were used for fitting, and the last 10 were forecasted. As the following picture shows, the model is not able to capture the exact quarterly variation, but it predicts the trend of increase or decrease in crime number.

V. Related Work

VI. Conclusion

VII. Ackknowlegement

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.

TABLE I
LOGFILE FORMAT

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

References

[1] https://www.kaggle.com/c/sf-crime
[2] Greenberg, David F. "Time series analysis of crime rates." Journal of Quantitative Criminology 17.4 (2001): 291-327.
[3] Said, Said E., and David A. Dickey. "Testing for unit roots in autoregressive-moving average models of unknown order." Biometrika 71.3 (1984): 599-607.
[4] Seabold, Skipper, and Josef Perktold. "Statsmodels: Econometric and statistical modeling with python." Proceedings of the 9th Python in Science Conference. 2010.