**Literature review**

Crime behavior is a major concern of our society. Data mining is a powerful tool that enables criminal investigators who may lack extensive training as data analysts to explore large databases quickly and efficiently[1]. The primary goal of crime data analysis is to find the hidden pattern of crime information and then make meaningful summary in a practical way. Some of the most widely used techniques are classification, clustering and so on.

There are some related researches on crime data mining, but not that much. Chen Hsinchun from University of Arizona[2], studies several different data mining techniques including entity-extraction techniques to analyze the behavioral patterns, association, clustering, neutral networks and pattern visualization, then matches them with most suitable crime types for analysis, such as traffic violations, sex crime, theft, fraud, arson, gang/drug offense, violent crime, and cybercrime which can facilitate police work and enable investigators to save their time to other valuable tasks.There is some related research on crime data mining, but not that much. Iqbal Rizwan from Universiti Putra Malaysia[3], covers investigation of data pre-precessing and classification methods on crime dataset acquired from UCI machine learning communities and discusses the results of the classification al- gorithms for predicting the Crime Category attribute. Basically, this paper applies two classification algorithms Naive Bayes and Decision Trees, to put U.S states into different categories such as low, medium and high crime rates, mainly based on median household income, population density, unemployment rate and population that is under poverty threshold. The results are evaluated by accuracy (correctly classified instances), pre- cision, recall values and F-measure for both of the algorithms. For decision tree, the accuracy and precision are 83.9519% and 83.5%, which performs better than Naive Bayes with accuracy and precision of 70.8124% and 66.4% respectively. Shyam Varan Nath from Florida Atlantic University[4], uses clustering algorithm to detect crime pat- terns. Along aside clustering method, the research also uses k- means clustering with some enhancements to assist the process of pattern identification, semi-supervised learning technique to help increase predictive accuracy and weighting scheme for attributes to deal with limitations of clustering tools and techniques. The proposed combined method mentioned above is applied with geo-spatial plot to display the result graphically. As a result, significant attributes have been identified and crime patterns have been formulated.

Data Set -- Data visualization

Further exploration for the dataset, we found some other information from more than 878,000 records of different crimes.

This data set covers a wide variety of crime, from Figure 1, we see that there are 39 categories in total. Among these categories, larceny/theft, other offenses, non-criminal, assault, and drug/narcotic rank top five in frequency. Statistically, top five crime categories make up about 66% of all recode, which means that several mostly occurred crimes make up the majority. We conclude that it is necessary for the police to put more force on dealing with major top five crimes.

In addition, we plot to analyze number of total crimes occurred in each of ten police department districts. We can see, from Figure 2, that Police department of Southern district deals with most amount of crime. Follow with Mission district, Northern district, Bayview district and other six districts. Richmond district has the least number of crimes, which is around one forth of Southern district.

Apart from above exploratory analysis, we also attempt to have a geometrical view of the dataset so we plot scatter of crimes as well as density line on San Francisco map. From Figure 3, it could be inference that crimes occurred mostly in northern part of the city. After some online searches, we found that the northern part is the downtown of San Francisco. It is consistent with what we expected from the former two plots.

Moreover, we are interested in the amount of total crimes occurred hours of day. As shown in Figure 4, crime rate is relatively low between 3 am and 6 am, and reach to their peak at 12 pm, and second peak around 5 pm to 6 pm.

Then, we want to know whether there is correlation between number of crimes of each of ten police department districts and the hour of the day. As indicated in Figure 5, there is a valley through wave of all ten districts at around 5 am and most of crimes break out in afternoon. At about 12 pm there is a peak in crime number, which may be caused by chaos of afternoon break. We notice that all ten districts have the same pattern, which is consistent with Figure 4. Therefore, we suggest that police can concentrate their force in some crime peak times.

[1]  .U.M. Fayyad and R. Uthurusamy, *Evolving Data Mining into Solutions*  *for Insights*. Comm. ACM, Aug. 2002, pp. 28-31.

[2]  .Chen, Hsinchun, et al., *Crime data mining: a general framework and some examples*. Computer 37.4 (2004): 50-56.

[3]  .Iqbal, Rizwan, et al., *An experimental study of classification algorithms for crime prediction*. Indian Journal of Science and Technology 6.3 (2013): 4219-4225.

[4]  .Nath, Shyam Varan, *Crime pattern detection using data mining*. Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI- IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on. IEEE, 2006.