Crime prediction with machine learning

Sida Zhang
Electrical and Computer Engineering Department, Marquette University
Course: EECE5850 Introduction to Intelligent Systems

Abstract— **Crime is a big issue of society. A lot percentage of crimes are unpredictable. Therefore, if we can use some data to forecast some crime happen in the future, it may do a big favor to social security. On the other side, whether the criminal would be caught is also very important to victims. In this project, according to analysis all the crime happened in Chicago during 2011-2016, I can do a prediction whether a crime just happened can be caught immediately. I will use four machine learning algorithms which are Random Forest, Logistic Regression, Adaboost, and Majority Voting. In these four algorithms, Random Forest has the best performance which the success rate is 0.8951. Logistic Regression is the least one which the success rate is about 0.823.**

I. Introduction

In early 2010, the police department in Los Angeles has already predicted the crime by analyzing a big amount of crime data. Police found that a crime occurs somewhere in the city, the probability of secondary crime happened here may have a lot of factors in common such as time of a day, location, and types of crime. Therefore, according to this rule, they developed a software which can predict the happened of a crime. By using data analysis, the software provides a crime hot map. The police may clearly see which place are more likely to have crimes. Then, the police department can send more police to patrol those blocks. This method may obviously decrease the crime rate. At present, in Los Angeles, theft has decreased 0.33, and violent crime has decreased 0.21 after using this software. The software is being tested in more than 150 cities in the United States and will be gradually used in the whole country. This can save a lot of expenditure for police department. However, if some crimes just happened unstoppable, finding criminal may also spend a lot of expenditure. Therefore, a prediction which can forecast whether the criminal will be caught is needed. In my project, by analyzing the crime datasets in Chicago, the program can predict whether the criminal would be caught. The inputs would be block, IUCR, primary description, secondary description, location description, domestic, ward, time, and mouth. The output would be arrested or not.

II. Related Work

I have read two papers about crime prediction. The first one is Using Machine Learning Algorithms to Analyze Crime Data. This paper can be divided into two parts which are data mining and machine learning. My project is also can be divided into data mining and machine learning. In data mining, the authors do the following things: (1) association (2) classification (3) clustering (4) forecasting (5) visualization. In my project, I also do some same things. In my data mining, I also classified my data, and deleted some features which is obviously useless. In addition, I used several images to show the description of some feature which can help me to do a preliminary forecast. In machine learning part, the authors used Linear Regression, Additive Regression and Decision Stump. The methods used by authors have fast modeling and handling some irrelevant features. However, they also have disadvantages such as linear regression must be used in linear relationship between features. Another paper I read used two machine learning algorithms which are KNN and boosted decision tree. In my opinion, KNN can easily ignore some abnormal value but it can not used in a large amount of dataset because of its computational complexity. As for the algorithms I used in the project, the Random Forest has the best precision and it is suitable for handling high dimensional data. However, Random Forests have been shown to overfit on classification or regression problems with big noise. The Adaboost

would not overfitting but data training may cost a lot of time.

III. Dataset and Features

The dataset I used in the project is all the crime happened in Chicago during 2011 to 2016 which including more than 250,000 crime information. Each crime was recorded their time, data, block, primary and secondary description, arrested or not domestic, ward, longitude and latitude.

A. Classification

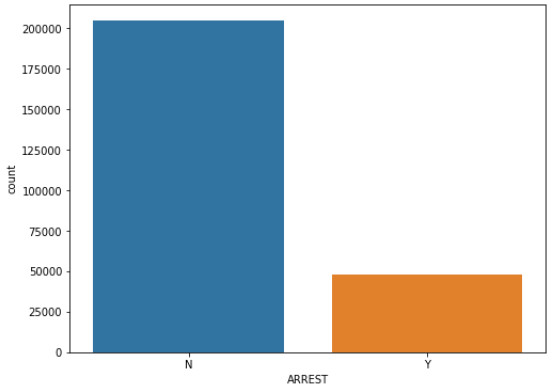
In my opinion, some of them is useless such as the latitude and longitude. The reason is that we have already had the information about block which is more comprehensible than latitude and longitude. Therefore, I deleted five features which are not necessary for crime forecasting.

B. Encoding Features

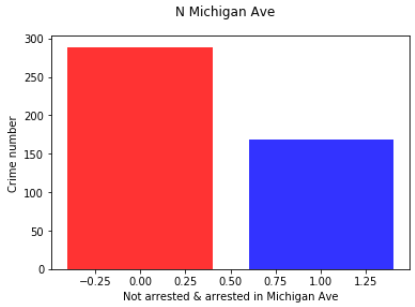
Firstly, I count that how many different types in each feature. For example, the number of different blocks which dad happened at least one crime is 27,377. If I just use 27,377 names of blocks, it is very difficult for agent to train and predict. Therefore, I used 1 to 27,377 to replace 27,377 different blocks. At the same time, I also used numbers to replace different types of other 11 features.

C. Data Mining

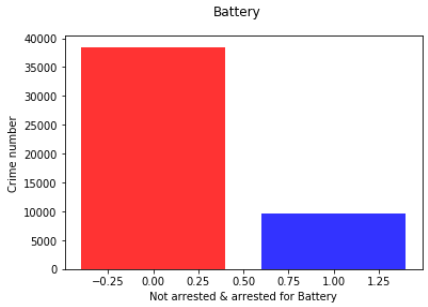
The following graph shows the numbers of arrested crimes and not arrested crimes.



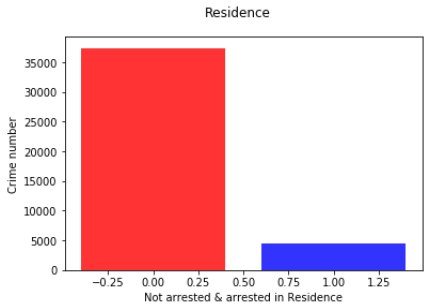
Then, I found the block which happened most crimes in Chicago around 2011 to 2016 which the result is 008XX N MICHIGAN AVE. The following graph shows he the numbers of arrested crime and not arrested crimes.



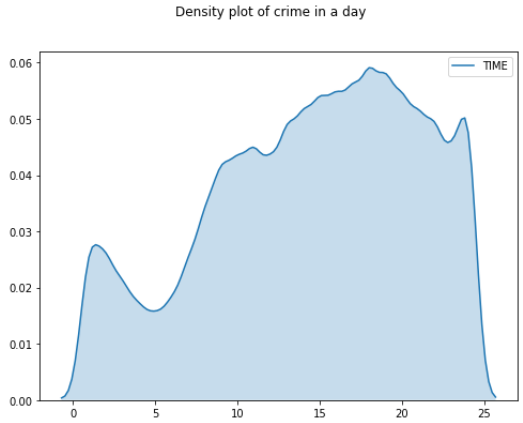
In addition, I also find the most crimes happened in Chicago is “battery”. The following graph shows the numbers of arrested crimes and not arrested crimes about “battery”.



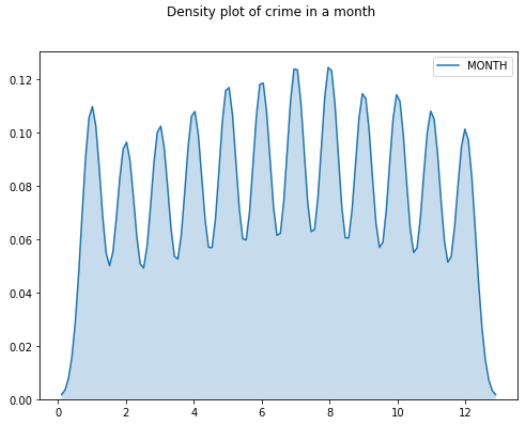
As for location, the most crime happened is Resident. The following graph shows the numbers of arrested crimes and not arrested in resident.



I also analyzed the distribution of the time when crime happened and made the following graph. It is easily to find that the most crimes happened at nightfall and the number reaches a peak of a day at 6 o’clock PM. The number reaches an off-peak before sunrise.



The following graph show the distribution of the number of crimes in different month. There is not a really big difference among 12 months. However, summer and January are more than other months. In my opinion, people's activities in summer are more than winter, so that it makes the curve has a peak in summer.



IV. Machine Learning

I used four different kinds of machine learning algorithms which are Random Forest, Logistic Regression, Adaboost, and Majority Voting. I will introduce them and give my statements.

A. Random Forest

Random forest randomizes some behaviors (such as feature selection and sample selection) in the process of constructing a decision tree to generate many decision trees. Then uses the results of these decision trees as votes. The most voted prediction will be the final solution. The effect of randomization is to reduce the bias of a single decision tree.

When building each decision tree, first of all, there are put back n times of sampling, n is the size of training set. It can be proved that two-thirds of the samples will be selected and the remaining one-third will be training data. In each node splitting, m features are randomly removed from m features (M is far less than m), and the best segmentation point is selected from these sub features. It is worth to point out that, because only the segmentation points are selected from the M features, the calculation amount is greatly reduced. Each decision tree is calculated to the end and is not cropped.

B. Logistic Regression

When facing a regression or classification problem, we can use the cost function. Then, the optimal model parameters are solved iteratively by the optimization method, and then we test and verify the quality of our solution model. Although logistic regression has "regression" in its name, it is actually a classification method. It is mainly used for two objectives classification problems (there are only two types of output, which represent two categories). In the regression model, y is a qualitative variable, such as y = 0 or 1. Logistic regression is mainly used to find out the probability of some events.

C. Adaboost

AdaBoost is used to build a strong classifier from a weak classifier. The general operation process of AdaBoost is to train every sample in the data and give it a weight to form the corresponding weight vector D. First, all training samples have the same weight. Then, use the weak classifier to classify and calculate the error rate, and then find the weak classifiers and train them. In the second training, the weight of each sample will be adjusted. In the first training, the sample weight of the first team will be reduced, and the sample weight of the first error will be increased. Finally, we can get a set of classifiers, and give each classifier a weight value alpha according to the error rate of each classifier. The value of alpha is as follows:

$$\alpha = 0.5ln((1-c)/c) \tag{1}$$

After we get the alpha, we update the weight vector D to reduce the weight of correctly classified samples and increase the weight of misclassified

samples. The calculation method is as follows: if a sample is correctly classified, the weight value is changed to:

$$D_i^{(t+1)} = D_i^{(t)} e^{-\alpha} / Sum(D) \tag{2}$$

D. Majority Voting

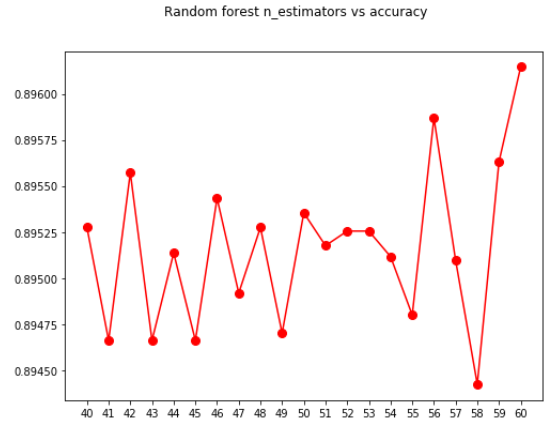
Each time you find a pair of different elements from the array, remove them from the array until you traverse the entire array. Because this problem has shown that there must be an element that appears more than half of the times, there must be at least one element in the array after traversing the array. The algorithm defines a sequence element (m) and a counter (i) in the local variable. The counter is 0 in the case of initialization; The algorithm scans the elements in the sequence. When processing element x, if the counter is 0, then assign x to m, and then set the counter (i) to 1; If the counter is not 0, the sequence elements m and X will be compared. If they are equal, the counter adds 1. If they are not equal, the counter minuses 1; After processing, the sequence element (m) is the most element in this sequence.

V. RESULTS AND DISCUSSION

First, I divided all the crime data into two parts. Four fifth of the data is training set and last one fifth is test set.

A. Random Forest

The number of false negative, false positive, true negative, true positive are 40047, 1047, 4256, and 5192. The success rate of predicting not arrested is 0.904. The success rate of predicting arrested is 0.833. The totally success rate is 0.8951. I also used 21 different value of K (from 40 to 61). The following graph shows the distribution of each success rate.The highest score of Random Forest is 0.900 when estimator equals to 53.



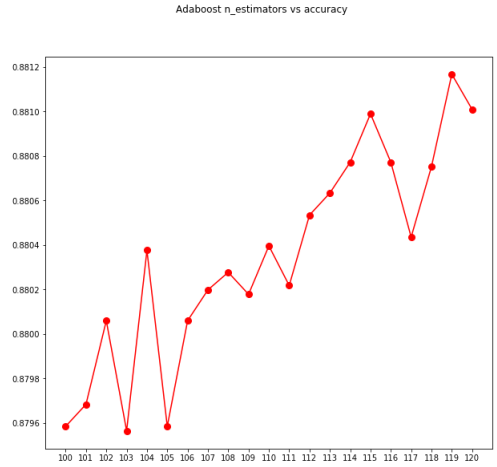
B. Logistic Regression

The number of false negative, false positive, true negative, true positive are 40192, 902, 8046, and 1402. The success rate of predicting not arrested is 0.833. The success rate of predicting arrested is 0.609. The totally success rate is 0.823.

C. Adaboost

The number of false negative, false positive, true negative, true positive are 40374, 720, 5366, and 4082. The success rate of predicting not arrested is 0.883. The success rate of predicting arrested is 0.851. The totally success rate is 0.8796.

I also used 21 different value of i (from 100 to 121). The following graph shows the distribution of each success rate.The highest score of Adaboostis 0.880 when estimator equals to 119.

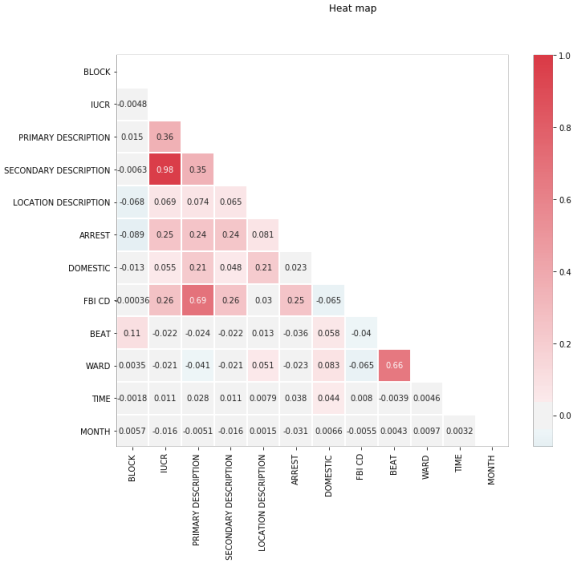


D. Majority Voting

The number of false negative, false positive, true negative, true positive are 40578, 516, 5437, and 4011. The success rate of predicting not arrested is 0.882. The success rate of predicting arrested is 0.887. The totally success rate is 0.882.

E. Heatmap of each features

Finally, the following heatmap shows the correlation between each feature. It is obviously that not too many features have strong connection. There is no doubt that some features should have strong connection. In my opinion, the reason why the heatmap does not show the strong connection is that the accuracy is very low and independence between each two features are high but between all features are not. It means, for example, the correlation between arrest and domestic is 0.023, which cannot exactly show the correlation. However, if we put all features together, the correlation will be strong.



F. Shortage in the project

Finally, I found a problem with my results. As we have seen in Data Mining part, the number of arrested crimes is lease than not arrested crime. Especially when crimes happened in residence, the number of not arrested crimes is over ten times of arrested crimes. It may cause the success rate of arrested crime. Another shortage is that the number of arrested crimes is no enough, so that the success rate of predicting arrested is low.

I used KNN, but it's unsuitable for a large amount of dataset. The results would be affected by the value of K. The success rate will increase when K is increasing. However, when the value of K beyond 12, the success rate may no longer affected by K. In my opinion, the main reason of this is overfitting and that is why I did not use KNN in my project.

VI. Conclusion and Future Work

In conclusion, Random Forest has the best performance which the success rate is 0.8951. Logistic Regression is the least one which the success rate is about 0.823. Random Forest has the best precision and it is suitable for handling high dimensional data. However, Random Forests have been shown to overfit on classification or regression problems with big noise. The Adaboost would not overfitting but data training may cost a lot of time. Logistic Regression is fast and suitable for binary classification. In my project, arrested or not is a typically binary variate. As for the last one Majority Voting, it is easy to understand and fast.

As for future works, I think I can use CNN to do machine learning. CNN is more accurate than the algorithms i used in the project. However, the things i worry about is that CNN is like a black box process. It may seems not reliable enough.

References

[1] S. Kim, P. Joshi, P. S. Kalsi, P. Taheri, "Crime Analysis Through Machine Learning" Fraser International College, Simon Fraser University, DOI: 10.1109/IEMCON.2018.8614828

[2] Y. Lin, T. Chen and L. Yu, "Using Machine Learning to Assist Crime Prevention," 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), Hamamatsu, 2017, pp. 1029-1030. doi: 10.1109/IIAI-AAI.2017.46

[3] Z. M. Wawrzyniak et al., "Data-driven models in machine learning for crime prediction," 2018 26th International Conference on Systems Engineering (ICSEng), Sydney, Australia, 2018, pp. 1-8. doi: 10.1109/ICSENG.2018.8638230 through deep reinforcement learning. Nature, 518(7540):529–533.

[4] City of Chicago, cityofchicago/crimes-one-year-prior-to-present, Data.world, Dec, 1, 2017. Available:https://data.world/cityofchicago/crimes-one-year-prior-to-present