Analyze on the Police Department Incident Reports in San Francisco, the United States of America

VO HOANG TRONG
Department of Electronics and Computer Engineering
Chonnam National University
Gwangju, Republic of Korea
187151@jnu.ac.kr

I. INTRODUCTION

In this project, I analyze the Police depart incident reports in San Francisco (SF), the US from 2003 to May 2018. First, I use K-means to cluster the specific of crimes in SF using 5 attributes: Crime type, district, a day of the week, hour and month. Then, given a day, month, hour, minute, and a name of a district, I use Decision tree and comparison with Multi-layer perceptron (MLP) to classify crimes that may happen at that time.

II. DATASET

The Police depart incident reports [1] released by the SF police department, filled by officers and by individuals through self-service online reporting for non-emergency cases. This data has 2.215.024 reports from January 1st, 2003 to May 5th, 2018 and it contains 13 attributes. The details information on each attribute is shown in TABLE 1.

TABLE 1. Attributes description of the dataset

Attribute	Description [2]
IncidntNum	The number issued on the report
Category	Contain 39 crimes
Descript	Detail description of the crime in Category, it has 895 unique description
DayOfWeek	The day the incident occurred, 7 days from Monday to Sunday.
Date	The day the incident occurred, written in format month/day/year, from 1/1/2003 to 10/17/2017.
Time	The day the incident occurred, written in format hour:minute, from 0:01 to 23:59.
PdDistrict	Name of district the incident occurred, with 10 districts
Resolution	Solution of the police for the crime, contain 17 different solution
Address	Address of the incident occurred
X	Longitude
Y	Latitude
Location	Coordinate written in format (Y, X)
PdId	An identifier unique to the dataset

I use this data and experiment on 2 problems

- Cluster: From 5 attributes: Category, PdDistrict, DayOfWeek, hour in Time, and month in Date, cluster into k groups and analyze.
- Classification: Given a day and month (in Date), hour and minute (in Time) and PdDistrict, classify which crimes (in Category) that may happen.

III. EXPERIMENT

A. Cluster

Before cluster, I preprocess the data as follows:

- Category: Since this attribute has 39 unique crimes, I encoded by using a one-hot vector has the length equal to 39. This vector has the value 1 in the corresponding category position, and 0 in the other position.
- PdDistrict: Similar to Category, I use a one-hot vector has length is 10 since this attribute has 10 unique value.
- DayOfWeek: I encoded the string from Monday to Sunday is 1 to 7 since this attribute has the order.
- Hour: I only use the hour value in Time attribute, in total, I have 24 values for this attribute, from 0 to 23.
- Month: From Date, I used the month part, I have 12 specific values from the month, from 1 to 12.

In total, I represent each incident by an attribute vector has the length is 52. Before using K-mean, I apply the elbow method to choose optimal k. I run k from 1 to 10, use formula (1) to calculate the sum-of-squares error on each k, then plot the graph to determine suitable k. In practice, I use Euclidean distance to calculate d.

$$\sum_{i=1}^{n} d(v_k - v_k^{\text{center}}) \quad (1)$$

where

v_k: Attribute vector belong to k^{th} cluster.

v_k^{center}: Center coordinate of k^{th} cluster.

n: Number of attribute vectors.

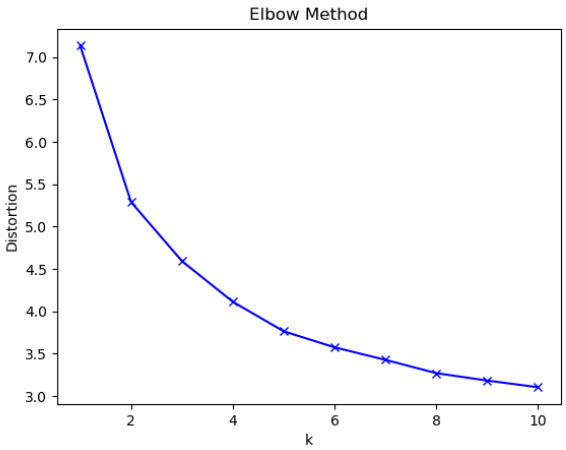


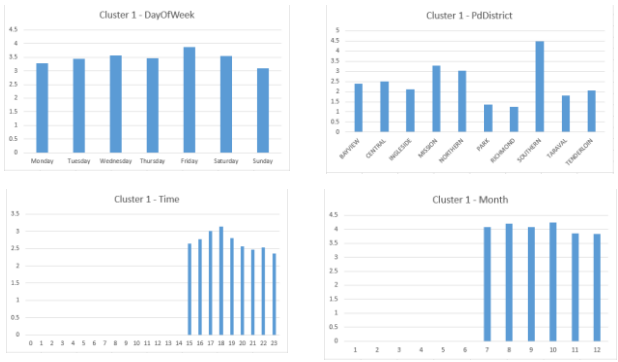
Figure 1. Elbow method to determine k

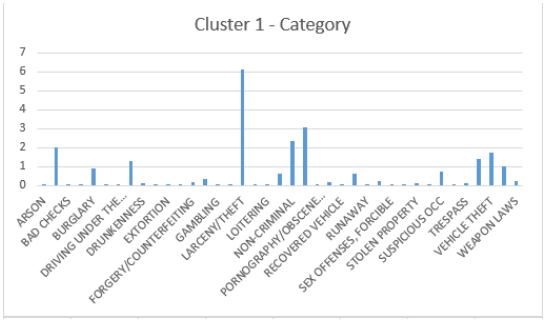
From Figure 1, I choose $k = 5$. After clustering, I have the result on each group below.

Group 1.

This group has 537929 incidents. On top 5 Category that appears most in this group. nearly 6.16% of larceny/theft will happen on this group, followed by other offenses (3.095%), non-criminal (2.35%), assault (2%) and vehicle theft (1.745%). It happens most in the Southern district (4.48% in total belong to this group), Mission (3.27%), Northern (3.03%) and Central (2.52%). Those crimes may happen nearly on all days, but most on Friday (3.88% in total of data) and Wednesday (3.57%), start from 3 p.m (2.64%) and getting higher to 6 p.m (3.14%), then it starts getting down to 11 p.m (2.358%). In a year, this group will happen from July (4.07%), get to the highest peak at October (4.24%) and end at December (3.84%). TABLE 2 shows 5 figures of 5 attributes.

TABLE 2. Figures of 5 attributes on group 1

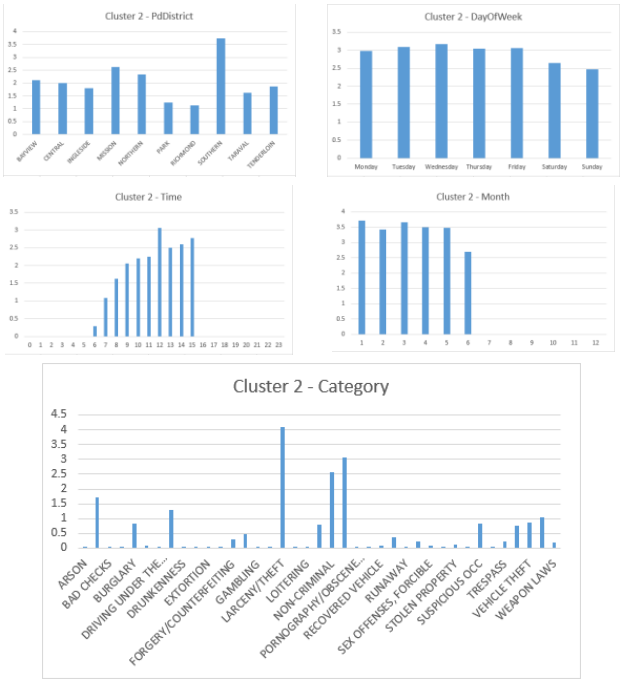




Group 2.

This group has 453473 incidents. TABLE 3 shows 5 figures of 5 attributes. On top 5 crimes, larceny/theft has 4.09% in total of the original data, followed by other offenses (3.077%), non-criminal (2.58%), assault (1.71%) and drug/narcotic (1.28%). Most of this crimes happen on the Southern district (3.73%), Mission (2.62%), and Northern district (2.33%) at the middle of the week, which the highest day is Wednesday (3.17%). However, Saturday and Sunday are the 2 lowest days (2.65% and 2.47%). Time occurs those crimes start from 6 a.m (0.288%), reach the highest point at 12 p.m (3.068%) and decrease to an end at 3 p.m (2.78%). It happens only on the first half on the year, start from January (3.71%) and end at June (2.7%).

TABLE 3. Figures of 5 attributes on group 2

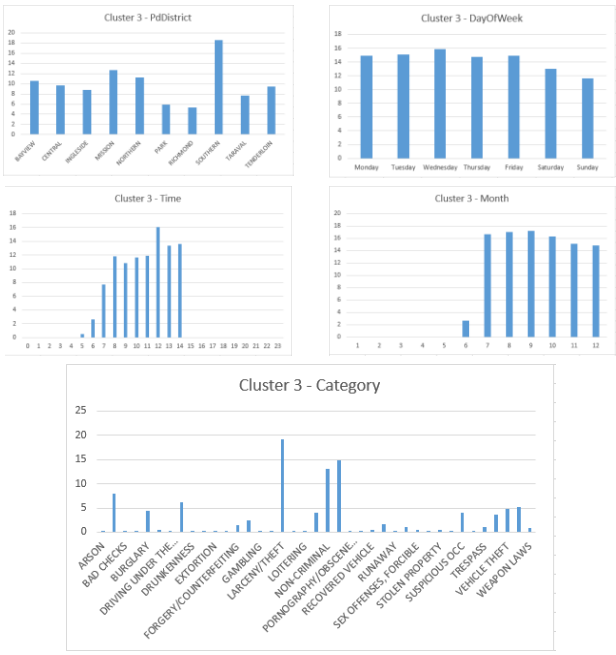


Group 3.

This group has 187723 incidents, top 5 crimes are: Larceny/theft (19.22%), other offenses (14.76%), non-criminal (13.12%), assault (8.02%) and drug/narcotic (6.15%). Most of the crimes happen on the Southern district (18.6%), followed by Mission (12.65%), and Northern (11.22%). It happens most on Tuesday (15.07%), Wednesday (15.83%) and less happen on Sunday (11.59%). For the crime in group 3, it starts

happening at 5 a.m (0.47%), then accelerates to 16.03% on 12 p.m and start decreasing and stop at 2 p.m (13.6%). The crimes happen on the last half of the year, start from June (2.68%), then quickly accelerate to September (17.23%), and decrease to 14.9% in December. TABLE 4 show figures of 5 attributes on group 3.

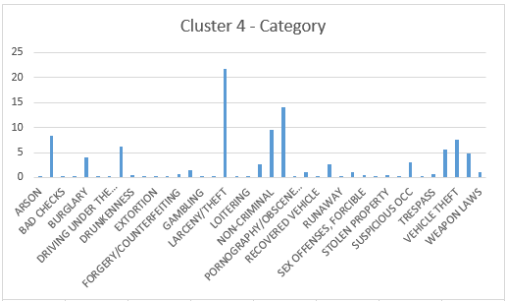
TABLE 4. Figures of 5 attributes on group 3



Group 4.

This group has 234772 incidents and figures of 5 attributes in this group are shown in TABLE 5, and top 5 crimes are larceny/theft (21.69%), other offenses (14.06%), non-criminal (9.46%), assault (8.4%), and vehicle theft (7.62%). Those crimes happen most on Southern (17.46%), Mission (14.07%), and Northern district (12%). It increasing from Monday (13.54%) to Friday (15.97%), and start decreasing to Sunday (12.84%). On the time, it begins at 4 p.m (13.09%), increasing to 7 p.m (12.97%), then it decreases to 9 p.m (11.12%) and again increase to 11.9% at 10 p.m, and decrease at 11 p.m (10.88%). The crimes on this group appear from January (16.99%) to June (16.3%).

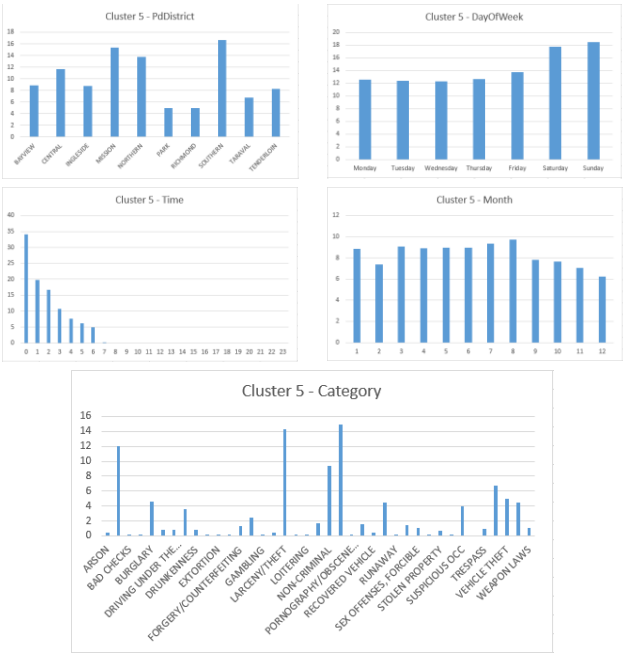
TABLE 5. Figures of 5 attributes on group 4



Group 5.

This group has 157616 incidents. Top 5 crimes are other offenses (14.93%), larceny/theft (14.27%), assault (12%), non-criminal (9.44%) and vandalism (6.7%). Most of them happen in Southern (16.68%), Mission (15.35%), and Northern district (13.74%). At the beginning of the week, it slightly increasing from Monday (12.6%) to Thursday (12.67), then start accelerating to Sunday (18.53%). Most of the crime happens at midnight (34.17%), then deeply decrease to 4.88% at 6 a.m, and 0.14% at 7 a.m. On first 8 months it happens quite a stability (approximate 9.75% on August), then decrease on the last 4 months (approximate 6.22% on December). TABLE 6 show figures of 5 attributes on group 5.

TABLE 6. Figures of 5 attributes on group 5



B. Classification

For the classification, I do the following problem: "Given a specific day, month, hour, minute, and district, classify which crime will happen". To solve this problem, from the raw dataset, I use the Category for the class, which mean I have 39 classes. For the feature, I use day and month on the Date attribute, hour and minute on the Time attribute, and PdDistrict to get the name of the district. Because PdDistrict is the category

feature, I encode it by using one-hot vector has 10 in length. In total, the length of a feature vector is 14.

From the raw dataset, I notice that 2 or more crimes can happen at the same time, date, and location, but in the experiment, I only use a one-hot vector to describe classes of the data and see how much noise it will affect on the training process.

To solve this problem, I use Decision tree and MLP. 70% of data are used for training, and the remain 30% use for testing. The train set has 1.550.499 data and the test set has 664.525 data. After training, I calculate the accuracy. Micro and macro precision, recall, and F1 score.

1. Decision tree

I use function `DecisionTreeClassifier()` on Python 3.5. After creating the tree, it has 44 in depth, and the Gini index shows that the Decision tree uses all 14 attributes. When applying to the test set, the result is shown in TABLE 7.

TABLE 7. Classification result when using Decision tree

Accuracy	59.72%
Micro-precision	59.72%
Micro-recall	59.72%
Micro-F1 score	59.72%
Macro-precision	64.88%
Macro-recall	44.86%
Macro-F1 score	49.28%

2. MLP

I train MLP using TensorFlow. For the MLP model, it contains 2 hidden layers, the first hidden layer has 512 nodes and the second hidden layers have 256 nodes. I use Sigmoid as the activation function after each hidden layer, and Stochastic Gradient Descent for backpropagation. The learning rate is set to 0.0001 and batch size is 16. I train on 100 epochs. For convenient, I use the test set as the validate set and choose weights that has the highest accuracy on the test set. shows the train accuracy process. After 100 epochs, the average accuracy on the train set is 29.45%, which is underfitting when using the MLP model, plus the noise of the data, where a feature vector may belong to many classes, affect worse on the training process. TABLE 8 shows the accuracy on the test set, which is lower than using a decision tree. In fact, there are some classes are not predicted.

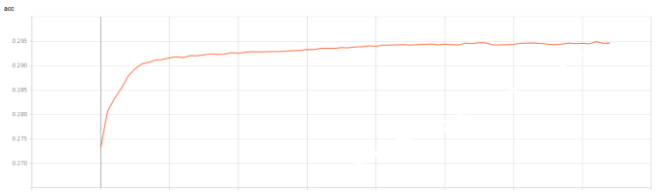


Figure 2. Train accuracy

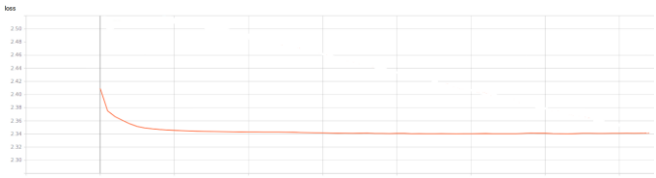


Figure 3. Train loss

TABLE 8. Classification result when using MLP

Accuracy	25.19%
Micro-precision	25.19%
Micro-recall	25.19%
Micro-F1 score	25.19%
Macro-precision	3.333%
Macro-recall	3.971%
Macro-F1 score	2.722%

IV. References

[1] DataSF, "Police Department Incident Reports: Historical 2003 to May 2018," 13 September 2018. [Online]. Available: https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-Historical-2003/tmnf-yvry.

[2] DataSF, "Police Department Incident Reports 2018 to Present," 2018. [Online]. Available: https://bit.ly/2x7Ta2P.