Predict the Category of Crimes That Occurred in the City by the Bay

韓文彬
National Central University, Tauyuan, Taiwan
b84414@gmail.com

郭皓磊
National Central University, Tauyuan Taiwan
ym910747@gmail.com

王銘陽
National Central University, Tauyuan Taiwan
wsp151515@gmail.com.tw

ABSTRACT

這篇報告是進行 Kaggle 網站上的其中一項競賽，該競賽為藉由以往犯罪資料來預測舊金山下一個犯罪可能發生的類型。官方提供了往年的資料，並以單、複數周作為區分，單數周為訓練資料，複數周為測試資料。官方給予的資料包含十幾種的犯罪類型，其中訓練資料還給予案件的詳細資訊及處理結果，但這兩欄並不會出現在測試資料。此外還有給予犯罪的經緯度，藉由經緯度可以在地圖中精準的畫出位置。

Categories and Subject Descriptors

Java, Weka(tool).

General Terms

Experimentation.

Keywords

Machine Learning, Data Mining.

1. INTRODUCTION

From 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz.

Today, the city is known more for its tech scene than its criminal past. But, with rising wealth inequality, housing shortages, and a proliferation of expensive digital toys riding BART to work, there is no scarcity of crime in the city by the bay.

From Sunset to SOMA, and Marina to Excelsior, this competition's dataset provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods. Given time and location, you must predict the category of crime that occurred.

2. RELATED WORK

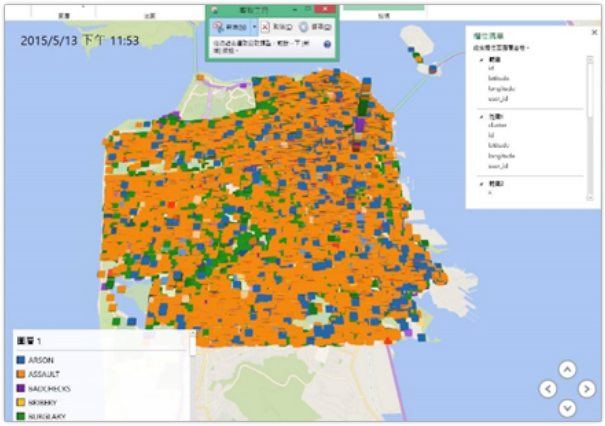
2.1 Data Information

- Dates - timestamp of the crime incident
- Category - category of the crime incident (only in train.csv)
- Descript - detailed description of the crime incident (only in train.csv)
- DayOfWeek - the day of the week
- PdDistrict - name of the Police Department District
- Resolution - how the crime incident was resolved (only in train.csv)
- Address - the approximate street address of the crime incident
- X - longitude
- Y - latitude

2.2 Data Analysis - 3D MAP

根據數據給予犯罪的經度(Longitude)以及緯度(Latitude)，我們使用 Excel 將數據轉換成 3D-Map 的形式，藉由視覺化的地圖，助於我們針對特定地方分析各個類型的犯罪。除此之外，Excel 還可以加入時間作為特徵，切割成各個時段來觀察，分析是否有某些犯罪於特定時間較容易發生。將各個犯罪案件顯示在地圖上會呈現如圖一的結果。

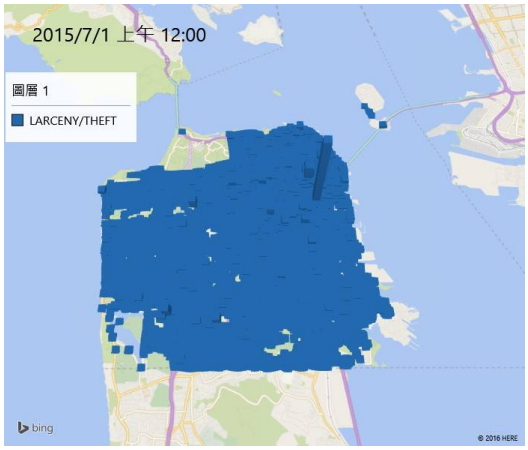
因為犯罪種類相當的多樣而且複雜，會致使我們難以分析地圖，無法得到詳細的資訊，因此我們擷取特定幾樣犯罪類型做分析，其中我們選擇竊盜、賭博和自殺三種犯罪類型。



（圖一、所有犯罪案件）

2.2.1 Larceny/Theft

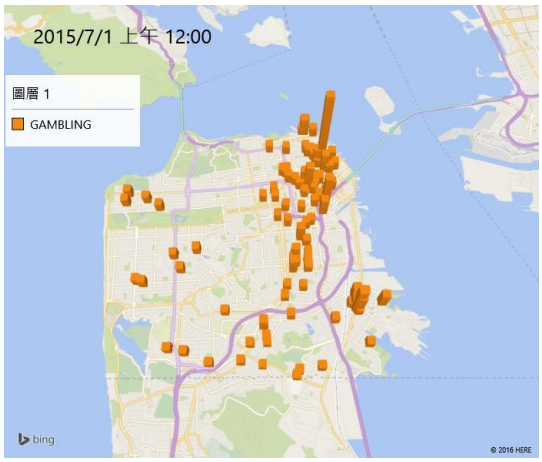
竊盜是最廣泛發生的犯罪類型，也是犯罪的種類中出現頻率最高的犯罪，從地圖（圖二）中觀察可以發現，竊盜分布的地點相當平均，其中右上角有一個非常高的柱狀體，代表該點發生的數量非常多，令人匪夷所思，因此我們拿該點的經緯度去搜尋 Google 地圖，發現該點是一家大型警察局，也就是說，如果警察局收到報案時，該案件的位置會標註在警察局的位置，所以那個地點才會有相當多的數據呈現在地圖上，此外，該點的犯罪理應被排除。



（圖二、竊盜案發生地點）

2.2.2 Gambling

第二個分析的犯罪類型是賭博，從地圖（圖三）中呈現該右上角較為密集，因為右上角為舊金山的市中心，當人的數量越高時，發生案件的機率越高，比較有趣的是，中間偏右的港口的數量也算密集，我們從 Google Map 的街景去查看，發現該位置大多都是平房，數量不多，但是往南方走一段距離會有很多家銀行，或許和賭博也有關聯，另外該地黑人的比例也比較高，或許跟郊區因素有關。

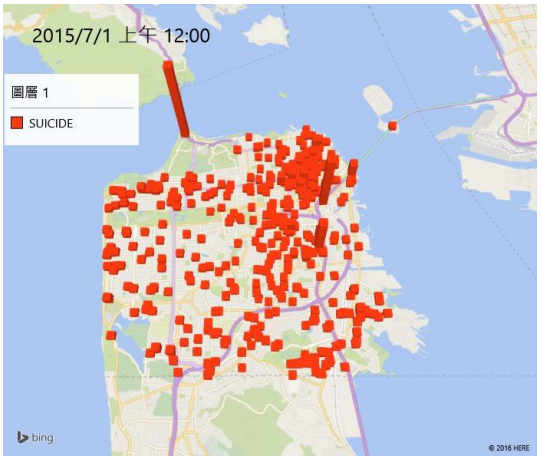


（圖三、賭博發生地點）

2.2.3 Suicide

第三個犯罪類型為自殺，從地圖（圖四）中可以觀察到，自殺分布也是滿平均的，右上角因為是市中心，當人的數量越多的時候，自殺的數量也會比較多，但是和其他地區相比時，比例是相同的。

不過有趣的是左上角出現的次數遠遠超過其他地方，上網搜尋資料後，得知該區是「舊金山自殺聖地」，據說在這裡自殺就有了普遍認同的理由：為什麼在這兒自殺？因為你無法在別的地方得到如此絢麗的死亡，這兒是死亡聖殿，只要跳了，你就是獻給死亡的神聖幡祭。因此在燈光迷霧下的金門大橋優雅地死去，是一件很寂寥卻美麗的結尾， 也造就了該地自殺率如此高的原因。



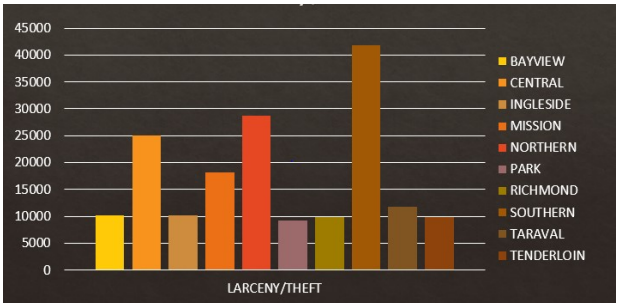
（圖四、自殺發生地點）

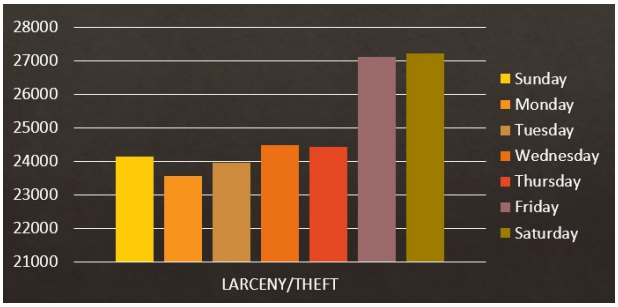
2.3 Data Analysis - Table

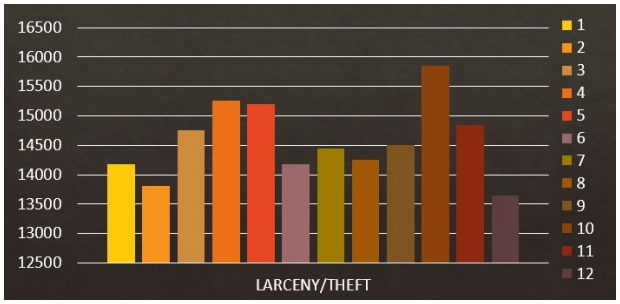
完成地圖地分析後，我們針對 PdDistrict、DayOfWeek 和 Month 三種犯罪類席分析各個特徵下的狀況，利用 Excel 製作了圖表，每個犯罪類型會有三種圖表，共九張圖。

2.3.1 Larceny/Theft

下方圖表為根據 PdDistrict、DayOfWeek 和 Month 做分類的樞紐分析表。第一張圖表是分析 PdDistrict 和竊盜犯罪的分布狀況，可以看到 Southern、Northern 和 Central 發生的次數是最高的。接著再以星期做分類，很明顯地在星期五和星期六的犯罪次數超過了其他天許多，推測是星期五和星期六的觀光人潮較多，故發生竊盜的機會就會比較大。最後是以月份做的分布結果，並無太明顯的特徵，沒有太多的資訊可以獲得。因此若要預測竊盜案件，以地區和星期作為特徵效果會好許多。

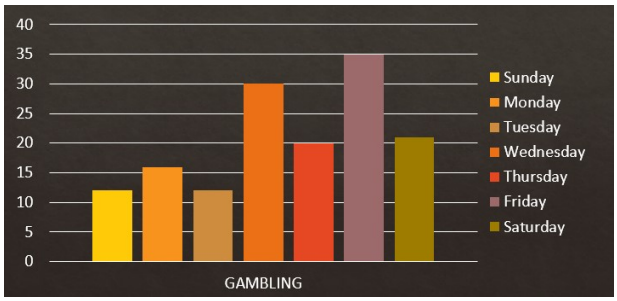






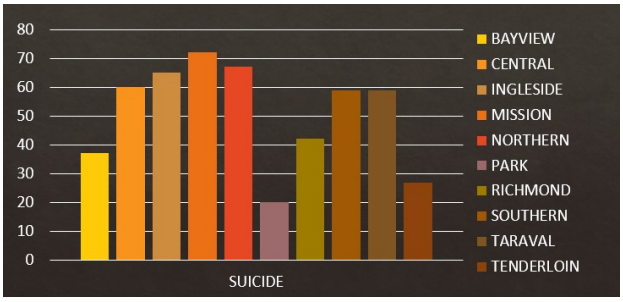
2.3.2 Gambling

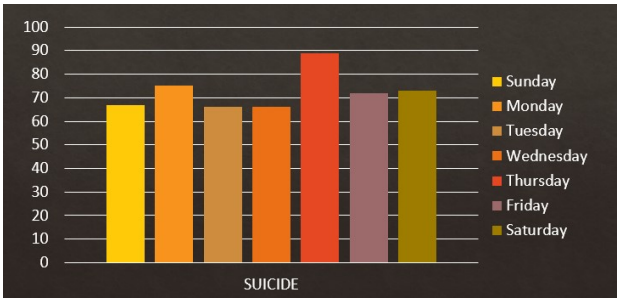
接著分析賭博的犯罪類型，第一張圖表是分析 PdDistrict 和賭博犯罪的分布，在 Bayview 和 Central 這兩個地區的犯罪次數超出其他地方滿多的，就和前面地圖分析的結果一致。接著分析星期和賭博的分布結果，其中星期三和星期五發生的次數明顯高於其它天，推測是因為星期三為小周末夜晚及星期五是周末夜晚，是上班族放鬆的日子，因此賭博的機率也比較高。最後一張圖表是月份和賭博的分布結果，一樣地，這張圖並無太過明顯的差別，因此無法獲得很多資訊。

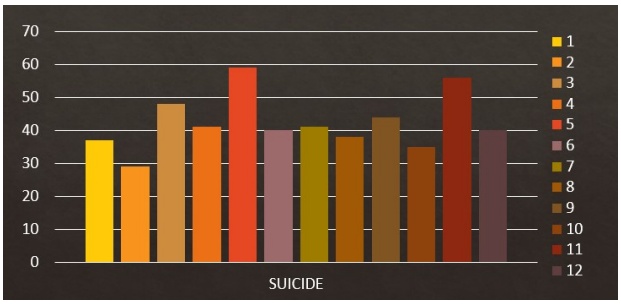


2.3.3 Suicide

最後針對自殺類型最分析，第一張圖表是分析 PdDistrict 和自殺的分布，不過地區並無太過明顯的差別，所以無法獲得足夠多的資訊。接著分析星期和自殺的分布結果，從星期一至星期日的分布挺平均的，因此也沒有明顯的資訊。最後是月份和自殺的分布結果，同樣地也算是平均的分布，因此這三種特徵都無法拿來訓練。







3. METHODS

藉由以上分析後，我們從得到的資訊著手安排我們的預測方法，經過多方嘗試後，正確率一直很低，或許是對於初心者而言，一次預測多數類別並且要很準確是一件很困難的事，因此我們先簡化分類，挑出其中一種犯罪類型，並將剩下的類型分成 Others，也就是總共會有兩種類別，再從 Weka 進行訓練和預測，使用的方法有 ZeroR、OneR 和 Naïve Bayes。最後藉由正確率和 F1-measure 來了解我們的模組。

4. EXPERIMENTS

根據前面的觀察，我們選出前面做分析的三種類型：竊盜、賭博和自殺。由於再起初我們針對時間做分類，發現時間因素並不會特別影響這三種犯罪，所以我們不再對時間做特別處理。下方表格是執行完的各項結果。

4.1 Larceny/Theft

可以從圖表發現，正確率高達 80%，藉由簡化分類能夠得到較好的數據。圖中的 F1-measure 是以竊盜對竊盜的正確率做運算的，所以只有 0.2 左右，但是加上 Others 對 Others 的正確率後會提升許多。

Classified Instances	Correctly	Incorrectly
ZeroR	703149 (80.0808 %)	174900 (19.9192 %)
OneR	706677 (80.4826 %)	171372 (19.5174 %)
NaiveBayes	703978 (80.1753 %)	174071 (19.8247 %)

ZeroR	Others	Larceny/Theft
Others	703149	0
Larceny/Theft	174900	0

OneR	Others	Larceny/Theft
Others	682340	20809
Larceny/Theft	150563	24337

Precision = 0.539
Recall = 0.139
F1 = 0.221

NaiveBayes	Others	Larceny/Theft
Others	671290	31859
Larceny/Theft	142212	32688

Precision = 0.506
Recall = 0.186
F1 = 0.272

4.2 Gambling

賭博和自殺有些欄位為零，因此沒有特別去計算 Recall 和 Precision，因為兩者的樣本數量相較於整體是非常少的，因此預測的準確率也會低很多。未來能夠考量樣本數量問題進行平衡，將兩者的數量達到相近時，預測準確率可能會上升。

Classified Instances	Correctly	Incorrectly
ZeroR	877903 (99.9834 %)	146 (0.0166 %)
OneR	877903 (99.9834 %)	146 (0.0166 %)
NaiveBayes	877903 (99.9834 %)	146 (0.0166 %)

ZeroR	Others	Gambling
Others	877903	0
Gambling	146	0

OneR	Others	Gambling
Others	877903	0
Gambling	146	0

NaiveBayes	Others	Gambling
Others	877903	0
Gambling	146	0

4.3 Suicide

Classified Instances	Correctly	Incorrectly
ZeroR	877541 (99.9421 %)	508 (0.0579 %)
OneR	877540 (99.942 %)	509 (0.058 %)
NaiveBayes	877389 (99.9248 %)	660 (0.0752 %)

ZeroR	Others	Suicide
Others	877541	0
Suicide	508	0

OneR	Others	Suicide
Others	877540	1
Suicide	508	0

NaiveBayes	Others	Suicide
Others	877389	152
Suicide	508	0

5. CONCLUSIONS

將類別簡化後，我們能夠預測出高達 80%的正確率，此百分比包含 True Positive 與 True Negative，當類別比較少時，特徵的界線比較明顯，對 Weka 而言也比較好做預測分類，不需要建造太複雜的分類樹。未來能夠加入更多的特徵提升正確率，以及以兩種或兩種以上不同的類別來進行比較，讓預測的結果更能夠被應用在生活中。

6. HOW TO RUN THE CODE

我們是自己寫一隻 java 程式處理資料，包含過濾某些欄位等等，所以在 package 會有多的.class，但實際上只會使用到一個，改該檔案裡的程式碼後就可以生成各種我們需要的欄位，包括一開始我們分成季節、月份和早晚的時間分類。得到處理完後的資料，再用 EXCEL 去做 3D Map 和樞紐分析表，藉以分析資料，有許多功能可以在 EXCEL 內完成，像是篩選、排序等等，因此前處理完成後就不會再使用到其他 code 了。最後預測的部分直接使用 Weka 做機器學習，在 Weka 中也可以逕自刪除欄位，省下了許多麻煩。

7. REFERENCES

[1] Predict the category of crimes that occurred in the city by the bay. *Kaggle.com.* DOI= https://www.kaggle.com/c/sf-crime

[2] Mir Henglin. *Mapping and Visualizing Violent Crime in San Francisco*. DOI= https://www.kaggle.com/mircat/sf-crime/violent-crime-mapping

[3] Microsoft Office. *Get started with 3D Maps.* Technical Report. DOI= https://support.office.com/en-us/article/Get-started-with-3D-Maps-6b56a50d-3c3e-4a9e-a527-eea62a387030

[4] Weka 3: Data Mining Software in Java. DOI= http://www.cs.waikato.ac.nz/ml/weka/

Division Table

韓文彬：程式 **Code**、分析資料、投影片製作、整理報告。

郭皓磊：分析資料、資料預測、投影片製作、報告分工。

王銘陽：**Excel 3D Map**、分析資料、資料預測、報告分工。