

科技部補助
大專學生研究計畫研究成果報告

計 畫 名 稱	: AI犯罪獵人：結合城市大數據與人工智慧，預測犯罪熱點及推薦巡邏路線
------------	-------------------------------------

執行計畫學生：呂姿儀

學生計畫編號：MOST 108-2813-C-006-004-E

研究期間：108年07月01日至109年02月28日止，計8個月

指導教授：解巽評

處理方式：本計畫涉及專利或其他智慧財產權，2年後可公開查詢

執行單位：國立成功大學電機工程學系（所）

中華民國 109年03月30日

(一) 摘要

在本項計畫中，我們的目標是當政府機構或警察機構需要規劃巡邏的路線時，如何從偌大的都市空間中，利用都市中不同地點的異質性巨量資料，如地理特色、人口結構、人群移動資料、動態氣象...等資訊，來有效預測不同地區在不同時段下發生犯罪案件的機率，並對即時的巡邏路線提出最佳化的推薦，期望提供一個友善且明瞭的平台來有效遏阻犯罪。

換言之，我們將開發一個案件預測與路線推薦系統，適用於城市中多種類型的巡視應用，包含違停糾舉或停車收費路徑的安排等。再者，現今針對犯罪預測的研究，多半在時間或空間上不夠細微，因此難以對巡邏路線的規劃有所幫助。此外，在巡邏路線的推薦上，過去也鮮少有研究針對多組人力遏阻犯罪路線的情境作探討。

我們透過與 CNN 相關的進階深度學習方法，將可得知任意地點在不同時間區間下的犯罪預測，用以助於進行巡邏路線的規劃。而在設計巡邏路線上，我們提出一個 A-star like 的演算法，可視人力情況有效地提供多條巡邏路線。最後，藉由本計畫特徵工程的異質性大數據分析平台可使政府機構受益，瞭解影響犯罪的可能因素，將有助於城市建設規劃與政策改善。

(二) 研究動機與研究問題

1. 研究背景

犯罪是自古以來就一直存在的社會問題，與人們的生活安危息息相關。隨著社會的發展與變遷，在現代社會，犯罪的型態也愈來愈多樣化，且在許多繁華的大都市，居高不下的犯罪率也已是常態。例如為美國金融中心城市之一的芝加哥，暴力犯罪與財產犯罪等不勝其數，甚至在 2016 年度共計有 4331 名槍擊受害者及 762 件謀殺案，創下近 20 年來最高的犯罪率，案件數甚至超過紐約與洛杉磯案件數的總和。

欲降低犯罪案件的數量，不僅需投注大量警力處理槍枝與幫派問題，警方的巡邏工作亦是不可或缺的一環，尤其需要針對犯罪熱點(Hot spot of crime)加強巡邏與臨檢，藉由提高見警率以遏止犯罪之發生，進而達到預防犯罪之功效。

此外，預測出各時段的犯罪熱點(Hot spot of crime)，除了可以使警方加強巡邏該地點以遏止犯罪，也能提醒民眾盡量避開該地點以降低受害機率，更可以藉由這些犯罪熱點(Hot spot of crime)的資訊，來規劃出警方巡邏的路線，使其巡邏工作得以更具效率。

2. 研究目標

本研究共分兩項目標：(1)給定過去歷史犯罪案件發生地點資料，以及目標欲預測的經緯度(在這篇研究中，我們將使用城市路網的節點)，如何利用 AI 從偌大的都市空間中自動學習並預測不同時段下高度可能發生犯罪的地點。(2)當高犯罪機率的節點確定以後，如何讓多組警察人馬在此路網上巡邏，盡可能地阻止犯罪案件的發生，以確保路線設立後的最大效益，亦避免資源、經費與人力的浪費。

本研究中的方法與其應用對象亦可為城市中違停糾舉或停車收費之路徑，換言之，我們希望在此計畫中開發出一個事件評估系統，讓使用者得以在都市地圖中先了解任一指定地點在不同時間區間下之事件預測，並進一步推薦巡邏路線，且支持多組人馬的路線規劃。

Definition 1: 路網.

我們可將城市中的路網看作一個有向圖： $G=(V, E)$ ，其中 V 代表所有節點的集合， E 代表邊的集合，任意 $v_i \in V$ ，代表兩條路段的交叉或一條道路的終點，任意 $e_k=(v_i, v_j) \in E$ ，代表一條從 v_i 連往 v_j 的路段，對所有的 v_i 來說，我們在資料庫中將存有其經緯度(lat_i, lng_i)。

Problem Definition1: 路網中任一指定地點在不同時間區間下之犯罪預測.

若 $V=\{v_1, v_2, \dots, v_m\}$ 為城市中的 m 個路網節點，我們的目標是針對路線 v_i 在某些特定時間 $T=\{t_1, t_2, \dots, t_n\}$ ，產生犯罪是否發生之預測：label $P_t=\{pv_{t1}, pv_{t2}, \dots, pv_{tm}\}$ ，其中 pv_{th} 為 0 或 1 的布林值，1 代表預測會有犯罪發生，0 則代表沒有。

我們將利用該城市過往的犯罪數據，訓練並建立一套預測式的模型。於實務上，時間區間 T 可代表平日或假日的不同時段，且可以是每半小時、一個小時或任何一個時段的預測；而犯罪是否發生的預測 label pv_{ti} 便是代表 t_i 時間區間內，該節點 v_i 預期是否發生犯罪的布林值。本計劃的第一目標就是希望預測正確的節點越多越好，而選擇路網節點作為預測之原因在於，警察的巡邏路線是根據城市道路結構所移動的，因此我們希望透過各節點的犯罪發生預測，來讓警方巡邏路徑的規劃更加便利。

Problem Definition2: 多組巡邏路徑規劃.

若有 k 個巡邏小組 $Q=\{q_1, q_2, \dots, q_k\}$ ，我們的目標為當使用者給定某些特定時間 $T'=\{t_1, t_2, \dots, t_l\}$ (以台灣為例， T' 通常設定為兩小時，代表一次巡邏為兩小時的路徑)，我們系統將為每一個巡邏小組 q_m 找出一組巡邏路徑 $R_{q_m}=(vt_1, vt_2, \dots, vt_l)$ ，其中 vt_s 為在時間點 t_s 需拜訪的路網節點。而我們的最佳化目標為在 R_{q_m} 包含的所有節點中，所包含的犯罪節點數量盡可能越大越好。

3. 研究動機

提出本研究的動機大致可劃分為兩個面向，一是在社會上的重要性，二則是於研究上的新穎性。

首先，誠如研究背景中所提及，預測出各時段的犯罪熱點(Hot spot of crime)，可以讓警方加強巡邏、提醒民眾避開該地點、以及幫助警方規劃巡邏路線，本研究便是以預測出犯罪熱點(Hot spot of crime)進而設計出任意巡邏路線作為出發點，期許能提供決策者一個友善便利的平台，作為規劃巡邏路線的參考，以達到降低犯罪、維護社會治安的目的，此為社會上的重要性。

其次，我們發現以往針對犯罪的研究分析，主要為以下三點：(1)在時間上以較大的時間區段作為單位；(2)在空間上以較大的行政區域作為單位；(3)只著重於各時段犯罪熱點(Hot spot of crime)的預測。

相較之下，本研究的目標為：在更加細微的路網節點及時間區段中做出犯罪預測，進而利用該預測資訊來設計出即時的巡邏路線規劃，並針對任意設計路線作出效益評估。由此可見，本研究與過往的研究方向有明顯相異之處，於犯罪預防上，應會更具應用性及完整性，此則為研究上的新穎性。

基於以上兩點，故我們認為本研究計畫是具有社會意義且具有研究價值的。

4. 研究挑戰

要開發出一個好的犯罪評估系統並不是那麼容易的，本研究所面臨的挑戰可分為三個部分，分別是影響犯罪的外在因素、細部時空間預測的難度以及規劃路線的相互影響力。

第一，由於本研究是對任意路網中的節點做犯罪預測，而影響一個路網節點的犯罪潛在因素非常多元，其中又可為動態與靜態兩種類型。動態而言，包含了歷史犯罪熱點、即時動態抱怨資料、人群流動、天氣、時間等；靜態而言，則包含 POI(Point of interest)、路網結構、人口組成、地理特性等。要如何整合上述這些因素，進而訓練並建立一套預測式的模型，便是本研究所致力於解決的目標。

第二，過去的研究也有針對犯罪率做預測，然而他們多半在空間上以較大的行政區域作單位(例如台北市大安區)，但我們的目標是在細微的路網節點中做出預測，請該預測易

於後續的路線規劃，要如何追出與犯罪相關的細微特徵值得探討，而在時間上的挑戰也是相同，過去的研究大多分出較大的時間區段，但在本研究中，為了利於巡邏規劃，我們必須找出對於細部時間犯罪的重要特徵，至少以每半小時為基礎去做推薦，實現近乎即時性的預測。

最後，當我們面臨多組人馬要同時在某個時段做巡邏規劃時，如何安排各組人馬的路徑，規劃出涵蓋犯罪範圍愈大且彼此不重複的多條巡邏路徑，也是值得本計畫挑戰之處。

(三) 文獻回顧與探討

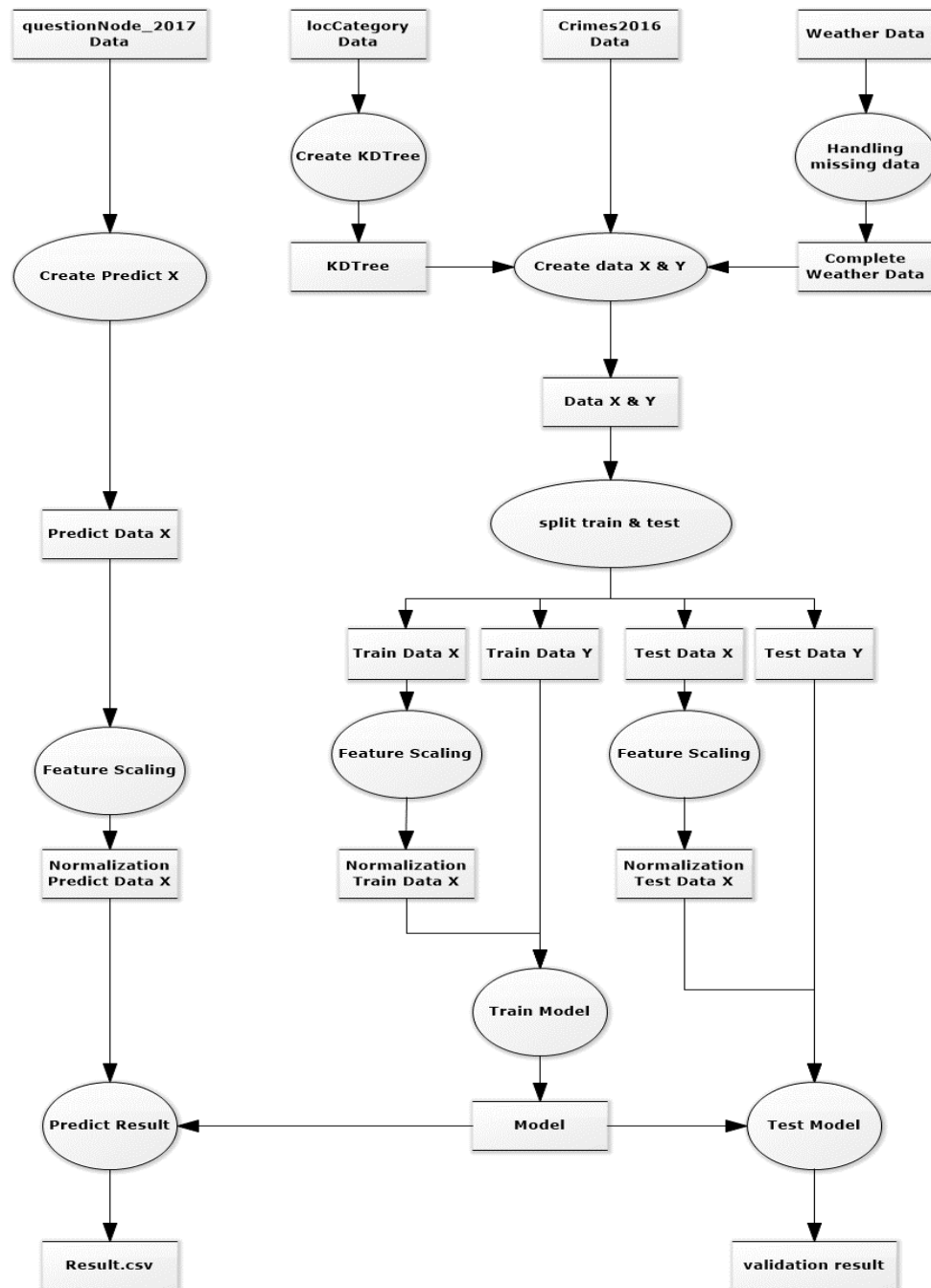
Yi et al.[8]分別在時間相關性利用 auto regression model 建模，而在空間特性上考量地區特徵的相似度來預測犯罪，最後提出 Clustered Continuous Conditional Random Field (Clustered-CCRF)，結合時間與空間相關性來對犯罪做出有效的預測。Zhao et al. [9] 利用 Tensor 拆解的方式結合氣象資料、POI、計程車資料以及 311 抱怨資料來對犯罪作預測。Huang et al. [1] 利用深度模型 Hierarchical Recurrent Networks 對於犯罪預測作建模，這樣的方式可以抓到不同段間犯罪的相關性，並在實驗結果中展現良好的預測能力，不過此模型僅適用於 POI 和 311 抱怨資料，無法表現出對於大量異質資料的結合能力。Jefferson et al.[4] 運用質化方法混合對參與芝加哥警察局預測分析軟體的電腦科學家的開放式訪談之分析，做出犯罪地圖推估。Fitterer [1] 針對 Break and entries 此種犯罪類型利用傳統 GIS 分析方法來實作犯罪預測地圖。Bowers[1]運用質性方法分析犯罪發生熱點的成因，並預測未來可能有哪些地區會開始發生大量犯罪行為，是屬於比較長期的犯罪預測研究。Mukhopadhyay et al. [6] 運用 Linear programming 根據犯罪發生來對於巡邏路線作動態規劃，但他們並未考慮犯罪發生地點的機率在時空間上的分布。

以上研究皆針對犯罪預測有所著墨，而本研究與以上研究存有相異之處。首先，我們針對的是時間區段較小以及空間較細部的犯罪即時預測，在時空間較細微的情況下，我們需要結合更多的動靜態資料，以及更精細的、可以抓到地區局部特徵的模型來作預測。其二，在過去，鮮少有研究使用計算機科學或數學的方式去解決犯罪路徑的推薦，大多數仍採用地知知識，且每天的路徑都會相同，而本研究將會在做出犯罪預測以後，結合路網與交通距離，設計出多組人力巡邏路徑，以求能遏阻多數的犯罪發生。

(四) 研究方法及步驟

1. 犯罪預測研究架構圖

本計畫將採用 Machine Learning 的方式。圖一為犯罪預測流程圖，包含 Data Preprocessing、Training、Prediction 等步驟，並以 2016 的犯罪資料做 Training，2017 的犯罪資料做預測：



圖一：犯罪預測流程圖

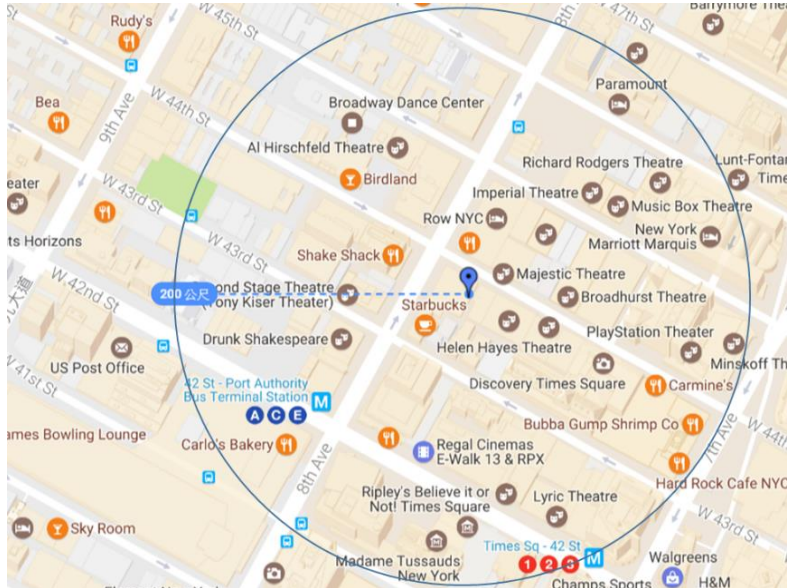
2. 特徵擷取與資料分析

(1) 犯罪 Label 前處理

芝加哥有詳實的犯罪資料內容，每筆犯罪案件都將有其經緯度、犯罪類型、及時間。雖然在犯罪資料中，有標記犯罪的類型，但由於本研究要預測的項目是有無犯罪的 Binary Classification，因此我們會將捨棄這些犯罪資料的犯罪類型，一律將 Y Label 標記為 True。

(2) 節點的周邊影響犯罪範圍定義

對於每筆預測，我們的目標為取出其周邊的環境特色與土地利用來輔助我們預測犯罪，例如若該節點周邊夜生活場所較多，或周邊店家較少(偏僻)的時候，也許就是容易發生犯罪的地點。在作法上，我們將考慮以 v_i 為圓心、 r 為半徑畫出圓面積再取出圓內地理特徵作為與犯罪發生相關的特徵。以圖二為例，假設藍色氣球點為我們想要預測犯罪率的節點，在半徑 $r=200$ 公尺，我們可以看到範圍內多為熱鬧或人群較多的地標，因此可以推斷這個地方在白天可能不是一個犯罪容易發生的地方。



圖二：地點與對應周邊範圍內的 POI 地點(本圖以 google 地圖輔助畫出)

(3) 預計採用特徵

A. 地點周邊 POI 分布

上節提到，我們定義了 r 作為地點的圓型周邊影響範圍，接下來我們便要找出所謂範圍內的 POI 分布特徵，探討 POI 分布的原因在於它代表的是在圓型內土地使用情形或不同面向的生活機能。因此，我們從圓範圍中取出的周邊異質性地點集合 $P=\{P_1, P_2...P_x\}$ ，以 P 集合內的地點來形容地路段 v_i 的周邊環境。如圖二中路線 v_i ，半徑 $r=200$ 公尺的例子中，若我們取出範圍的地點作為探討 POI 之分布，而這些地點便是代表在 v_i 周邊的基礎地理特徵。首先我們將在範圍內取出總 POI 數量，接下來，針對不同類型的 POI 再取出其對應數目，作為不同類型的特徵值。我們於表一列舉常用的 15 種 POI 類型，我們將可對 15 種 POI 類別取出重要的統計特徵值，例如該類型 POI 的數量，以作為 v_i 周邊環境的描述。

表一：常用的 15 種 POI 類型

T1: 汽車服務站 (加油站或修理站)	T9: 文化或教育設施
T2: 交通運輸站	T10: 娛樂場所
T3: 工廠	T11: 公司或辦公大樓
T4: 家具	T12: 旅館相關服務
T5: 飲食與餐廳	T13: 醫療機構
T6: 賣場、百貨或市場	T14: 夜生活場所
T7: 運動設施	T15: 警局或治安相關單位
T8: 公園	

誠如上一段落所述，一個區域中的 POI 類別象徵著該區域的土地利用及功能，因此它們可以有助於機器判斷哪些地點具有較多潛在乘客。一些 POI 類別甚至和犯罪率有直接的因果關係。例如，都市若有某區域包含許多夜生活場所(T14)，對於犯罪來說

可能在夜晚較容易發生鬥毆相關的事件，這即是 POI 類型與犯罪出線機率的正相關。相反地，在一些 POI 很少的地方也有可能造成犯罪的高度機率發生，特別是強盜或殺人之類的犯罪問題，若某個地區不同種類的 POI 類型與數量皆很大，代表可能是生活機能非常豐富多樣，將可能減低犯罪，另外 T15 的警局分布可能也會影響犯罪的發生地點分布。

對於每個節點 v_i ，我們預計取出下列對應 POI 特徵值作為地點分布特徵：

- 地點周遭 POI 密度：透過計算以 v_i 為軸心向外擴張 r 半徑內的 POI 數量，作為該路線的 POI 密度。所有類型的總 POI 密度某種程度而言代表周遭區域的人潮，所以我們預期應會對犯罪有一定的相關；除此之外，為了瞭解不同地點與 POI 類型的相關性，我們也會針對前一段落所提到的 15 種 POI 類型各別算出其密度，並都做為 v_i 節點周遭多元密度的特徵值。
- 周遭熵值(Entropy)：為了評估該地區的空間異質性對一個地方的犯罪影響，我們應用資訊理論中的 Entropy[Cover et al., 1991]計算方式來衡量一個地區的多元性。假如將 Γ 定為所有地理社交媒體服務地點類型的集合， $\gamma \in \Gamma$ 為某種地點類型，而 $N(v_i, r)$ 為 v_i 路段 r 半徑內所有地點個數， $N_\gamma(v_i, r)$ 為 v_i 路段 r 半徑內 γ 類型地點的所有個數，則

$$\text{Entropy}(v_i) = - \sum_{\gamma \in \Gamma} \left(\frac{N_\gamma(v_i, r)}{N(v_i, r)} \times \log \frac{N_\gamma(v_i, r)}{N(v_i, r)} \right)$$

一般而言，熵值大的地區將被預期為包含有較多類型的地點，熵值小的地區則意味著此區域的地點偏向某一特定類別，例如住宅區或商業區。

B. 打卡動態資料

單純分析 POI 類型的密度或熵值是不夠的，我們可以更進一步透過分析人類足跡資料如 Foursquare 或 Facebook 打卡資料來了解人口流動的情形。知道過去有多少人曾經拜訪過這些 POI 可能對於我們評估人潮有所幫助：

- 該地區 POI 熱門程度：為了評估 v_i 半徑 r 範圍內的整體人潮數目對於 v_i 的影響，我們測量在該地點區域內觀察到的總打卡數作為特徵值，這個特徵值也代表了人潮的往來的數量，同樣地，對於前節所述之 15 個 POI 類別我們也都可以分別取出該類別的熱門程度做為特徵值。
- 轉移密度(Transition Density)：假設該地點 v_i 半徑 r 範圍內移動性越多，就代表越多人在這個範圍內做一次以上的拜訪不同地點，而我們將計算此密度。這個特徵值也代表了該範圍內的人潮往來頻率，我們預期此特徵與犯罪率會有相關。
- 進入流動(Incoming Flow)：用於解釋該地點 v_i 半徑 r 範圍外的用戶流向該地區的流量特徵值。我們考慮由 $\text{pair}(m, n) \in T$ 表示連續兩個地點打卡的地點轉換，當第一個位置 m 位於範圍外，而在 n 位於範圍內時，就是所謂的 Incoming Flow 發生，我們計算此次數的多寡作為特徵值以代表外面帶進來的人潮流量，我們預期此特徵會與犯罪率有所相關。

C. 311 抱怨資料

目前許多城市的 Open data 有包含了動態的民眾反映資料，例如美國的 311 Request data 或台灣的 1999 市民電話，裡面包含了豐富的各種類型民眾抱怨內容，以及其發生的地點和時間，這方面的即時資料可以代表一種民眾動態的抱怨累積，當抱怨累積較多的時候有可能對於犯罪有所助長的影響，因此我們預計將範圍內的不同抱怨特徵(例如：空氣品質、動物、電力、火災、熱、無家可歸、停車、噪音、交通和水系統)，以類似動態打卡特徵的方式處理，觀察是否與犯罪相關。

D. 人口組成

除了考量動態方面的人群流動情形，靜態而言，我們透過分析政府開放資料，進而了解該地區的人口結構組成。一地區的人口結構組成雖與 POI 具有部分的相關性；然而，透過人口組成的分析可能對於我們評估潛在犯罪率有所幫助。我們將考量地區的教育程度或年齡人口組成以及年收入…等，以作為犯罪的特徵考量。

E. 時間

即使是同樣的地點，在不同時間的犯罪發生機率也應有所不同，因此時間資訊也是很重要的。

- a. Month，犯罪發生的月份，1 月—12 月
- b. Day，犯罪發生的日期，1 日—31 日
- c. Hour，犯罪發生的時間，0 點—23 點
- d. Week，犯罪發生的星期，1(星期一)—7(星期日)

F. 地點（距離）

每個候選地點應與和歷史上的犯罪熱點距離有著高度相關，而經緯度也是可以參考的特徵，我們會透過實驗去分析兩者是否相關。

- a. Latitude，犯罪發生的緯度
- b. Longitude，犯罪發生的經度
- c. 取現有 k 個最近犯罪熱點中心的距離作為特徵(KNN)

G. 天氣

即時的天氣概況也應與犯罪有相關，例如大雪或暴雨時的犯罪可能應較低，我們會以離預測目標最接近的整點，去找尋天氣資料。例如 01:16:00，會被算為 01:00:00。

- a. Humidity，犯罪發生時的溼度
- b. Pressure，犯罪發生時的氣壓
- c. Temperature，犯罪發生時的溫度
- d. wind Direction，犯罪發生時的風向
- e. Wind Speed，犯罪發生時的風速
- f. Weather Description，犯罪發生時的天氣描述

由於天氣描述的種類太多，如果直接進行 One-Hot Encoding，會造成 Features 數量太多，可能造成 Overfitting，因此我們將數種天氣的描述，進行摘要與提取。經過分析後，將數種天氣歸納成四個種類，最後使用 One-Hot Encoding 的方式，將天氣描述的種類編碼成以下四個特徵：

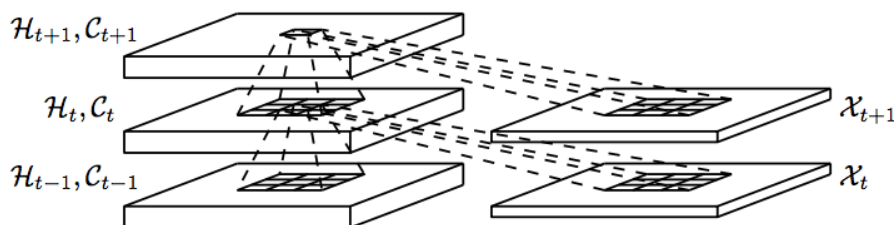
- (a) Snow，只要有下雪，都被歸為此類
- (b) Rain & Thunderstorm，只要有下雨或打雷，都被歸為此類
- (c) Smoke & Foggy，造成視線不佳，例如起大霧或是霧霾
- (d) Sunny & Overcast，除了上述三種以外，其餘的天氣描述，例如晴天、多雲和陰天都算此類

3. 犯罪預測方法

在本計畫中的方法，我們預計採用以卷積神經網路(Convolutional Neural Networks, CNN)為基礎的模型作預測，雖然 CNN 方法目前來說被驗證最有效的領域是影像辨識方面的應用，然而其實可以把城市中地圖視為二維的圖像，藉由 CNN 抓到城市中不同區域重要資料特性，我們預期應該會不錯預測效果。本研究方法預計採用兩種與 CNN 相關之研究方法進行預測，分別為 ConvLSTM 和 Graph Convolutional Neural Networks。

(1) ConvLSTM

Shi et al. [7]提出了如何將 Convolution 應用至 LSTM 的想法，顧名思義，其不僅具有 LSTM 的時序建模能力，而且還能像 CNN 一樣刻畫局部特徵，可以說是時空特性具備的預測模型，傳統的 LSTM 內部 gate 之間是依賴於類似前饋式神經網路來計算的，對於時序資料可以很好地處理，但對於空間資料卻無法抓到地區局部特性，ConvLSTM 嘗試解決此問題，做法是將 LSTM 中 Input-to-state 和 State-to-state 部分由前饋式計算替換成卷積的形式。ConvLSTM 內部結構如下圖三所示：



圖三：ConvLSTM 原理與架構

(2) Graph Convolutional Networks(GCN)

Kipf et al. 2017[1]年提出 Graph Convolutional Networks，由於過於雜亂的數據並不易於 CNN 中進行池化，因此便須採用針對網絡(Graph)進行處理的卷積神經網路(GCN)。於本次研究中，我們利用 Convolutional layers 所得 Rectified Linear Unit(ReLU)之 Activation functions 以確保其非線性的狀態。此外，由於在類神經網路中透過 Chebyshev polynomials 與自由參數進行學習，並在空間向量中得到一近似於二維 CNN 所得之平滑濾波，因此利用 GCN 進行學習將可大幅減少模型訓練時間並提高對於網絡處理的精確度。

首先，所有神經網路層將可被定義為一非線性方程式 $H^{l+1} = f(H^l, A)$ ，其中 $H^0 = X$ ， $H^L = Z$ ；而 L 代表神經網路的總層數、A 代表網絡結構的 Adjacency matrix、Z 則為以節點(Node)作為區分依據的輸出結果，至於目標在於求得表述網絡 $G = (V, E)$ 的方程式。另一方面，中介層的傳遞則是透過 $f(H^l, A) = \sigma\left(\hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}} H^l W^l\right)$ ，其中 W^l 即為第 l 層的 Weight matrix；而 $\hat{A} = A + I$ ， \hat{D} 則是代表 \hat{A} 的 Diagonal node degree matrix。

由於異質性的 Features 很多，而過多的 Features 不一定是好事，它可能導致 Model 容易 Overfitting，又或者特徵之間有 Dependency，導致特徵組合會降低預測準確度。因此除了上述兩個 Model 的參數需要做測試調整以外，也需挑選正確組合的 Features，才能得到更好的結果。除了上述的特徵之外，也有其他參數可以修改，例如訓練資料的數量、半徑 r 的大小以及 KNN 的 k 值等。

4. 路線規劃方法

經過前節的 ConvLSTM 或 GCN 預測方法以後，我們就可以得知每個節點預期是否發生的犯罪的機率，接下來，我們預期使用並修改 A*搜尋演算法來作巡邏路線規劃。

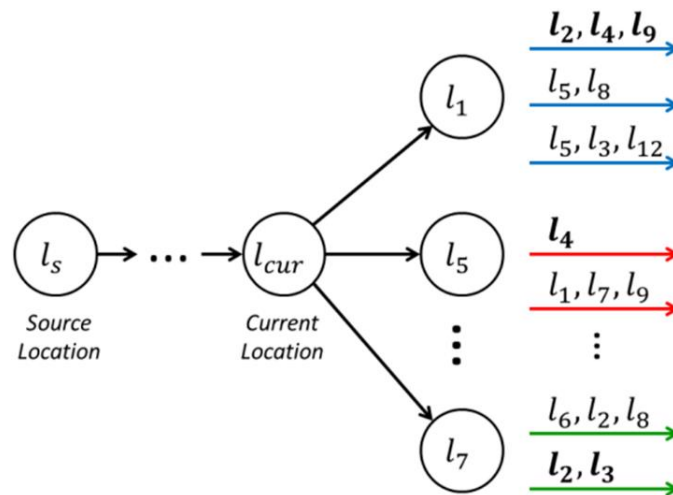
(1) A* algorithm 的 g(n)和 h(n)

這是一種在圖形平面上，有多個節點的路徑，求出最低通過成本的演算法。該演算法綜合了最良優先搜尋和 Dijkstra 演算法的優點：在進行啟發式搜尋提高演算法效率的同時，可以保證找到一條最佳路徑（基於目標函式）。在此演算法中，如果以 g(n) 表示從起點到任意頂點 n 的實際距離，h(n) 表示任意頂點 n 到目標頂點的估算距離（根據所採用的評估函式的不同而變化），那麼 A*演算法的估算函式為：

$$f(n) = g(n) + h(n)$$

選用 A* 搜尋演算法之原因如下：(1) 雖然 A* 演算法原本被提出是處理距離的相關應用，但是否發生犯罪這個機率我們也可以看作是犯罪可能性的分數，也就是距離的相反數，因此使用 A* 演算法是合理的。(2) 本問題無法使用暴力法將所有可能巡邏路線列出以後再取通過最多犯罪地點的路線，以芝加哥為例，一個警局轄區可能有 3 萬個節點以上，同時又要考慮節點之間的距離，以及多組人馬之間的巡邏路線不得重複來看，要對於一段時間內的巡邏作出一個有效率的規劃是個 NP-hard 的問題。因此，我們目標是將 A* 演算法中的 $g(n)$ 作為得到犯罪的分數，而在 $h(n)$ 中，我們將同時考慮目前節點到巡邏終點的預計可以得到的犯罪分數和其他人重複巡邏的機率分數。圖四為我們方法的示意圖，而以下的 f^* 為 A* 演算法計算是否加入巡邏地點的分數，若 l_{cur} 為目前加入的巡邏地點，則考慮下一個巡邏地點時，我們將會在 $g(n)$ 考慮可以得到的犯罪分數，在 $h(n)$ 考慮到下個巡邏地點之後可能得到犯罪分數以及給予一個懲罰權重去考慮，不斷地加入節點，並最終長成一條巡邏路徑，而當有多組巡邏人力的時候，我們將使各組分別長出一條路線並最終產生出多條路線。 β 在這裡為控制重要性的參數，我們將透過 data 去學習設定出一個好的 β 值。

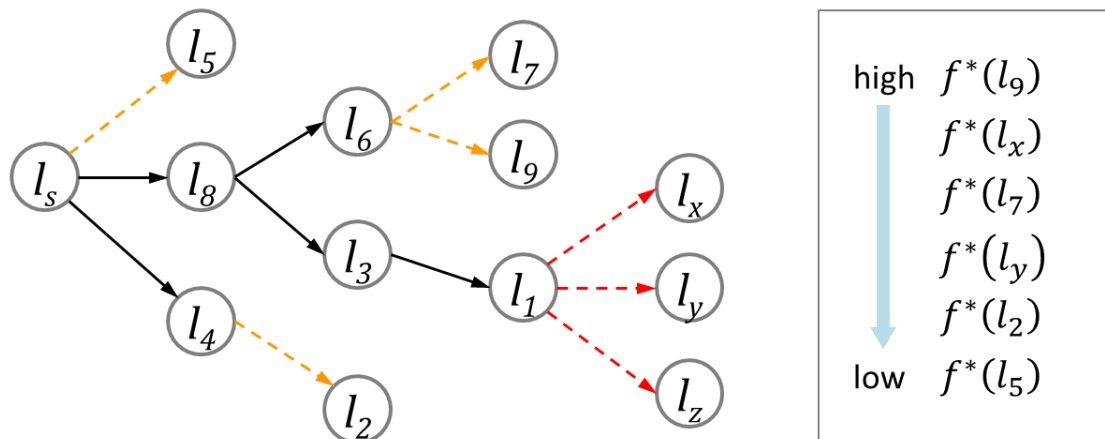
$$f^*(l_1) = (1 - \beta) \times g(l_1) + \beta \times h(l_1)$$



圖四：A-star 演算法搜尋下一個巡邏地點之示意圖

(2) Backward Checking Mechanism

在 A* 演算法搜尋中，為了以防演算法不小心走入 Bad local optimal solution，我們將保留先前搜尋樹在展開的各種可能性，以下圖五為例，若一開始選擇 l_4 卻發現從 l_4 發展會加入 l_2 這個極差的點則我們將返回 l_s 考慮由 l_8 重新往下長。



圖五：Backward Checking Mechanism

(五) 預期結果

本研究將進行實驗來驗證本計畫所提出的犯罪預測和巡邏路徑是否具有相當的效益。

首先，我們會將預期和犯罪有關之特徵值進行結合，再透過 GCN 和 ConvLSTM 等實作來驗證犯罪預測的結果。我們將帶入美國洛杉磯、紐約、芝加哥、舊金山...等知名城市的犯罪歷史數據和相關 Open data，並使用是否犯罪做為目標值的取樣，進而利用 Training 和 Testing 的方式，將實際發生犯罪和推估會發生犯罪的地點作比較，看兩個推出的集合是否相近。因此我們將使用 F1 score 當作我們的驗證指標，它能夠有效且客觀地去推測這些節點的犯罪可能性，是否能被我們的方法有效地預測出來。此外，我們將使用至少為年度以上的犯罪資料來作驗證，例如利用 2017 年的犯罪資料來對 2018 年的不同時段作預測，預測的時間基準以小時或半小時為單位。

其次，在巡邏路線的驗證中，假設有 k 組巡邏人力，本研究將驗證我們所設計的 A* 演算法是否能夠盡可能地涵蓋到愈多的真實犯罪節點，也將測試在不同數值 k 的情形之下，此方法的效益為何。因此我們將以 Hit rate 當作我們的驗證指標，其算法為走過且真實發生犯罪的節點數除以總犯罪發生節點數。

(六) 實作成果

1. 犯罪預測

(1) Feature

A. 時間

- a. Month，犯罪發生的月份，1 月—12 月
- b. Day，犯罪發生的日期，1 日—31 日
- c. Hour，犯罪發生的時間，0 點—23 點
- d. Week，犯罪發生的星期，1(星期一)—7(星期日)

B. 地點（距離）

- a. Latitude，犯罪發生的緯度
- b. Longitude，犯罪發生的經度
- c. Location Category，犯罪發生地點附近的建築物

本研究總共實作四種方法，其架構和流程皆大同小異，只是 Feature 的表示有略微不同，可使用 CONFIG 檔輕鬆調整。

首先我們會使用 KDTree 找尋離犯罪地點最近的 K 個建築物，接著在以犯罪地點為圓心，從 K 個建築物中保留半徑 R 範圍以內的建築物，其餘距離太遠的則會被剔除。由於建築物總共有兩萬多個，如果直接將這些建築物進行 One-Hot Encoding，Feature 數會變得太大，容易造成 Overfitting，因此我們使用 Foursquare 提供的 Venue Categories，將建築物摘取成 10 個種類。最後使用 One-Hot Encoding 的方式，將 Location Category 編碼成十個 Features。

One-Hot Encoding 主要是針對 Nominal Data，因為這種資料的種類並沒有順序或大小之分，因此無法直接轉換成有大小的數值，只能將其拆成許多種類特徵，例如本研究的 Location Category 總共有 10 類，因此會變成 10 個特徵，每個特徵都是 Binary，1 代表有該種類，0 代表沒有該種類，而通常每個資料都只有一個種類，造成只有一個值為 1，其餘都為 0，因此才會得到 One-Hot encoding 的命名。由於從 K 個建築物中保留半徑 R 範圍以內的建築物數量可能是複數，將會同時有多個特徵的數值都是 1，因此我們的作法算是 One-Hot encoding 的變

形，例如一個犯罪地點的附近的建築物總共有 3 種，則會有三個特徵值為 1，其餘為 0。

以下是四種調整的作法：

	R	Encoding
(a)	$R < 1$	One
(b)	$R > 100$	One
(c)	$R < 1$	Multiple
(d)	$R > 100$	Multiple

- (a) 半徑 R 很小，很可能從 K 個建築物中剔除部分，剩下在半徑內的建築物。接著利用 Encoding 產生特徵向量，若有兩個建築物的種類是相同的，則該種類的特徵值依然用 1 表示。
- (b) 半徑 R 很大，因此所有 K 個建築物都會被保留。接著利用 Encoding 產生特徵向量，若有兩個建築物的種類是相同的，則該種類的特徵值依然用 1 表示。
- (c) 半徑 R 很小，很可能從 K 個建築物中剔除部分，剩下在半徑內的建築物。接著利用 Encoding 產生特徵向量，若有兩個建築物的種類是相同的，則該種類的特徵值會用 2 表示。
- (d) 半徑 R 很大，因此所有 K 個建築物都會被保留。接著利用 Encoding 產生特徵向量，若有兩個建築物的種類是相同的，則該種類的特徵值會用 2 表示。

C. 天氣

我們以犯罪發生時間最接近的整點，去找尋天氣資料。例如 01:16:00，會被歸類為 01:00:00；而 01:46:00，會被歸類為 02:00:00。

- a. Humidity，犯罪發生時的溼度
- b. Pressure，犯罪發生時的氣壓
- c. Temperature，犯罪發生時的溫度
- d. wind Direction，犯罪發生時的風向
- e. Wind Speed，犯罪發生時的風速
- f. Weather Description，犯罪發生時的天氣描述

由於天氣描述的種類太多，如果直接進行 One-Hot Encoding，會造成 Features 數量過多，造成 Overfitting，因此我們將數種天氣的描述，進行 Abstraction 摘要與提取。經過分析後，將數種天氣歸納成四個種類，最後使用 One-Hot Encoding 的方式，將 Weather Description 編碼成四個特徵。以下為四種天氣描述的種類：

- (a) Snow，只要有下雪，都歸為此類
- (b) Rain & Thunderstorm，只要有下雨或打雷，都歸為此類
- (c) Smoke & Foggy，造成視線不佳，例如起大霧或是霧霾，都歸為此類
- (d) Sunny & Overcast，除了上述三種以外，其餘的天氣描述，例如晴天、多雲和陰天，都歸為此類

由於 Features 眾多，而過多的 Features 不一定是好事，可能導致 Model 容易 Overfitting，又或者特徵之間有 Dependency，導致特徵組合會降低預測準確度。因此我們得經過許多嘗試或思考後，挑選正確組合的 Features，才能得到較好的結果。為了滿足上述的目標，我們將可挑選的特定 Feature 寫成設定檔，這樣就可以較為簡單地測試特徵的組合。

```
[FEATURE]
month = yes
day = no
hour = yes
week = yes
latitude = no
longitude = no
location_category_arts_entertainment = no
location_category_college_university = no
location_category_event = no
location_category_food = no
location_category_nightlife_spot = yes
location_category_outdoors_recreation = yes
location_category_professional_other_places = no
location_category_residence = yes
location_category_shop_service = no
location_category_travel_transport = yes
humidity = no
pressure = no
temperature = yes
wind_direction = no
wind_speed = no
weather_description = no
```

除了上述的特徵之外，也有其他參數可修改，例如訓練資料的數量、半徑 R 的大小、KNN 的 K 值等，同樣使用設定檔調整與測試。

```
train_data_size_subset = 4
k_nearest_neighbor = 1
radius = 100
multiple_location_type = no
```

(2) Label

雖然在蒐集來的犯罪資料中，有標記犯罪的類型，但由於本研究要預測的項目，是有無犯罪的 Binary Classification，因此將這些有犯罪的資料，捨棄掉其犯罪類型，一律將 Y Label 標記為 True。

(3) Model

以下會介紹本研究使用的 Model，以及 Tune Model 時調整的相關參數。

A. SVM

a. Kernel Function

由於 SVM 的數學公式略為複雜，難以評斷對 Model 造成的影響，因此我們把常見的 Kernel Function 都稍微試過，包含 linear、polynomial、rbf 和 sigmoid。

B. Decision Tree

a. Maximum Depth of Tree

由於 Decision Tree 很容易發生 Overfitting，因此我們會控制樹的最高層數，限制樹的高度，防止其延伸太多層，以達到遏止 Overfitting。

b. Minimum Samples for A Node Split

由於 Decision Tree 很容易發生 Overfitting，因此我們會控制此參數，規定每次的分裂都必須有特定數量以上的 Samples，否則不能進行分裂，以達到遏止 Overfitting。

c. Minimum Samples for A Leaf

由於 Decision Tree 很容易發生 Overfitting，因此我們會控制此參數，規定樹的每個 Leaf 都必須有特定數量以上的 Samples，以達

到遏止 Overfitting。

d. Measure Function

嘗試了 Entropy 和 Gini Index 這兩種計算方式，發現最終的效果相差不多。

C. Random Forest

a. The Number of trees

由於 Random Forest 是由許多 Decision Tree 組成的，因此要調整樹的數量，太少的話表現可能會不好，太多的話則可能會讓訓練時間太久。

b. Attributes of Decision Tree

上述 Decision Tree 所調整的參數，在 Random Forest 這裡也都會調整，目的也是為了防止 Overfitting。

D. Multi-layer Perceptron classifier

a. Hidden Layer

調整 Hidden Layer 的層數，以及每層 Layer 擁有的 Neuron 數量以及使用的 Activation Function，我們使用的 Activation Function 是常見的 Sigmoid 和 Relu。

b. Optimizer

使用主要常見的兩個 Optimizer，Adam 和 SGD。

(4) Data Set

A. 產生假資料

由於我們只有犯罪的資料，也就是 Positive Data，完全沒有 Negative Data，這種 Imbalanced Data 很容易產生問題，如此訓練出來的 Model 就會傾向於資料較多數的那一方，造成預測結果永遠都是會有犯罪發生，這將會是一個不好的結果。

解決 Imbalanced Data 的方法有很多種，我們採用的是手動產生 Negative Data，讓 Positive 和 Negative 雙方的資料量差不多。我們從 2016 年曾經發生犯罪的所有地點中，隨機挑選一個地點，接著將該地點曾經發生犯罪的時間點剔除，再從剩餘的時間點中隨機挑選一個取代，這樣就可以產生一筆沒有犯罪的假資料，也就是 Negative Data。重複以上動作，直到 Negative Data 數量和 Positive Data 數量相同為止。

(5) Experimental results

A. Feature Selection

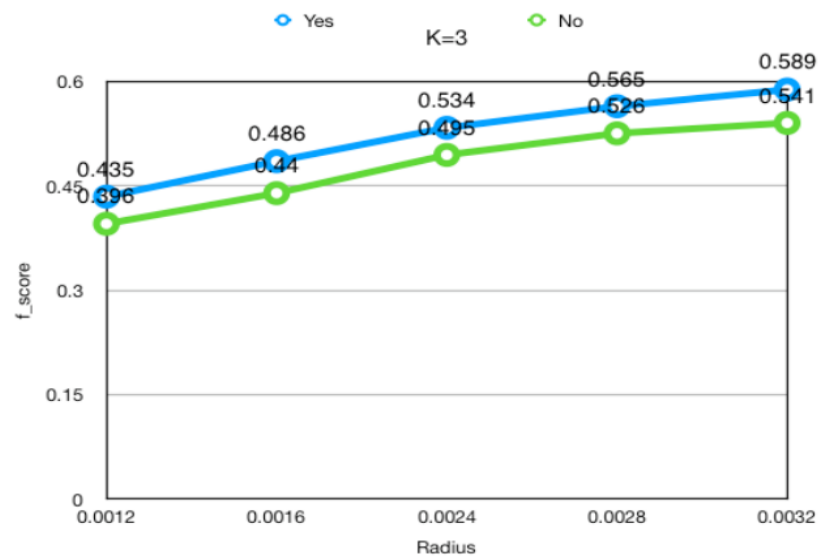
由於 SVM 是 Binary Classification 最常使用的其中一個 Model，因此我們一開始先選擇實驗 SVM，調整各個 Feature 以及參數，並記錄測試結果。雖然不同的 Model，適合的 K 和 R 數值可能會有所變動，但目前的實驗可以當作一個參考值，在測驗其他 Model 時可先用此數值，若表現得不夠好可再做調整。

a. 地點

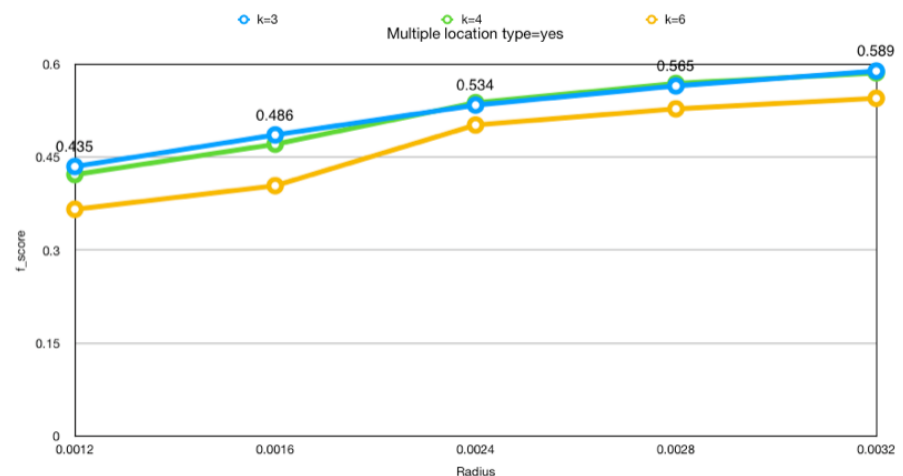
對於時間和天氣相關的特徵，基本上都只需考慮是否納入，把我們認為重要的特徵加進 Model 進行訓練。但對於地點相關的特徵而言，由於 KNN 的 K 值，以及半徑 R 都是未知數，對於準確度的影響是未知的，因此我們首先進行地點相關的特徵調整。

下圖的參數調整，是將 KNN 的 K 定為 3，找出離犯罪地點最近的 3 個建築物，接著觀察隨著 R 的變動調整，相對應的 F-Score 的數值變化。以下所測試的是 Location Category 四個做法中的(a)和(c)，綠線代表(a)，若有兩個建築物的種類是相同的，則表示該種類的特徵值依然

用 1 表示；藍線代表(c)，若有兩個建築物的種類是相同的，則表示該種類的特徵值會用 2 表示。



以上的實驗數據顯示，藍線平均比綠線高，也就是作法(c)平均比作法(a)好，因此後續實驗將會以作法(c)為主。下圖的參數調整，是固定使用作法(c)，接著觀察隨著 R 和 K 的變動調整，相對應的 F-Score 的數值變化。從圖中可發現，不論 K 值為多少，當 R 愈大時，則 F-Score 的表現愈好。另外，當 K 為 3 或 4 時，表現都較為良好；當 K 為 5 時，表現則會下降些許。



b. 時間

時間的特徵主要選擇 Month、Hour 和 Week。

- (a) Month，月份會影響季節，可能夏天和冬天等季節對犯罪有所影響，因此將 Month 納入特徵
- (b) Hour，不同時間可能會影響犯罪，半夜人較少時可能較容易發生犯罪，因此將 Hour 納入特徵
- (c) Week，平日是上班/上學的時間，周末是假日，這些可能都會影響犯罪，因此將 Week 納入特徵

實際上，利用上述時間特徵訓練出來的 Model，與利用上述地點特徵的成果差異不大，Accuracy 大約都在 54~58% 左右。

c. 天氣

天氣的特徵主要選擇天氣描述。

- (a) 天氣描述，下雪或下雨等天氣，都可能會影響犯罪機率，因此將天氣描述納入特徵。

實際上，利用天氣特徵描述進行 Model 的訓練，Accuracy 大約都在 54~58% 左右。即使加入了其他天氣特徵（溫度、濕度等），也沒有提升的跡象。

d. 總和

將地點、時間以及天氣特徵進行結合，Accuracy 和 Precision 都沒有明顯的變化，都維持到 54~60% 左右，但 Recall 有明顯的進步，大約都可以到 70% 以上，因此 F-Score 都維持在 60~70% 左右。

B. SVM

不論 Kernel 是哪個，預測結果都沒有突破性的進步，都維持在 50~60% 左右。

C. Decision Tree

Decision Tree 很容易 Overfitting，因此有很多可以調整的參數都是為了避免其發生。Overfitting 會讓 Model 過度 Fit Training Data，雖然在 Training Data 的預測表現特好，但在其他資料上，可能表現就差強人意。但我們目前連 Training Data 都表現不夠良好，預測的 Accuracy 都在 58% 左右，Entropy 都 0.9 多，因此調整這些參數沒有多大的意義。

D. Random Forest

和上述的 Decision Tree 的原因差不多，因此不需要去調整避免 Overfitting 的參數。我們需要調整的參數是樹的數量，測試的數量從 10、30、50 到 100 左右，當樹的數量上升時，預測表現會愈來愈好，但只有些微的上升，Accuracy 只到達 58% 左右。

E. Multi-layer Perceptron classifier

Hidden Layer 的層數，我們嘗試了 1~4 層。每層的 Neuron 數量，嘗試了 20~100。常見的 Activation Function，試了 Sigmoid 和 Relu。Optimizer 的選擇，試了 Adam 和 SGD。上述這些參數組合的測試，無論是哪種組合，Loss Function 所算出來的數值，都大約在 0.68 左右，不太會再往下降了。不論我們怎麼調整參數，收斂的效果都沒有變得更好。

2. 巡邏路線規劃

(1) 高斯分布

依照預測出的犯罪案件，以每件案件為中心做出二維高斯分布，離中心愈近機率密度愈大，離中心愈遠機率愈小。

(2) 高斯分布疊加

將網格圖中的二維高斯分布疊加，每個網格上不同高斯分布計算出的機率密度相加，得出整個網格圖的犯罪機率密度。

(3) 路徑規劃

巡邏路線從起點出發，將犯罪機率密度最大的鄰接點加進路徑，重複此步驟完成路徑規劃。

(七) 總結與討論

1. 犯罪預測

從上述的實驗可以發現，不論是特徵的選擇或是 Model 的參數調整，預測結果都沒有表現很良好，Accuracy 和 Precision 都在 55~60%左右，而 Recall 則在 70%左右，最後 F-Score 都在 65%左右。

在本研究原先的 Model 設計中，主要是預測犯罪有無發生，因此當初在 Preprocessing Data 時，不論是發生的犯罪是屬於搶劫還是殺人，我們都把這些資料標記為有犯罪的 Y Label，而這可能就是導致成果表現不夠良好的原因，因為不同的犯罪類型，可能會有不同的犯罪特性，例如 THEFT 容易在半夜發生、OFFENSE INVOLVING CHILDREN 發生的時間可能接近學生上課或下課的時間。上述這兩種犯罪類型的特性，所需要、所注重的特徵就可能就會有所不同。因此，當我們忽略了犯罪的類型，可能就會導致最後訓練出來的 Model 表現得不夠良好。

2. 巡邏路線規劃

從上述的實驗可以發現，使用高斯分布疊加可以考慮到各個犯罪案件之間的距離和密度，從而影響路徑的規劃，使規劃出的路徑不僅考慮到附近案件最多處，還考慮到了接下來前進的方向，最終規劃出可以接觸到較多犯罪案件的路徑。

(八) 參考文獻

- [1] KJ. Bowers, SD. Johnson, K. Pease (2004). Prospective Hot-Spotting: The Future of Crime Mapping? British Journal of Criminology.
- [2] J. Fitterer, T.A. Nelson and F. Nathoo(2014). Predictive crime mapping. Police Practice and Research. 2014.
- [3] C. Huang, J. Zhang, Y. Zheng, and NV Chawla(2018). DeepCrime: Attentive Hierarchical Recurrent Networks for Crime Prediction. CIKM 2017.
- [4] BJ. Jefferson(2018). Predictable Policing: Predictive Crime Mapping and Geographies of Policing and Race. Annals of the American Association of Geographers.
- [5] T. N. Kipf, and M. Welling(2017). Semi-Supervised Classification with Graph Convolutional Networks. ICLR 2017.
- [6] A. Mukhopadhyay, C. Zhang, Y. Vorobeychik, M. Tambe, K. Pence, and P. Speer. Optimal Allocation of Police Patrol Resources Using a Continuous-Time Crime Model
- [7] X. Shi, Z. Chen, H. Wang, D.Y. Yeung, W. Wong, W. -C. Woo(2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. NIPS 2015.
- [8] F. Yi, Z. Yu, F. Zhuang, X. Zhang, H. Xiong (2018). An Integrated Model for Crime Prediction Using Temporal and Spatial Factors. IEEE ICDM 2018.
- [9] X. Zhao, and J. Tang (2017). Modeling Temporal-Spatial Correlations for Crime Prediction. CIKM 2017.