# 摘要

隨著航空運輸的發展日益興盛，提升了人類探索周圍與世界的能力，也拓展了經濟發展與各種商業行為。任何飛航服務都基於由國際民航組織所頒布的安全規範，以及民航法規來運作的。當嚴重事故發生，最先影響的是民眾的搭機意願、減少商務活動、甚至可能造成經濟衰退的問題。

過去有很多學者也投入飛航安全的研究，大多數都著重在特徵的獲取和篩選，而對各種分類演算法之間的準確度差異較少去，所以本篇將會把各個分類演算法的結果進行比較，嘗試找出最適合應用在飛航安全領域的分類演算法。本實驗在資料預處裡的階段為了解決資料不平衡的問題(傷亡與非傷亡的事故筆數差異懸殊)會使用Synthetic Minority Oversampling Technique (SMOTE)技術使資料量趨於平衡，接著使用九種分類演算法去各自進行分類(採用10折交叉驗證的方式)進行訓練以及測試，最後取各自的評估指標(準確度、F-measure、AUC、Recall)去進行比較，希望在互相比較後，能從中發現各方法的不同特點，並針對篩選出的特徵屬性進行更進一步的探討。

關鍵詞：飛航安全、分類演算法、資料預處理

**第一章、介紹**

　　根據IATA(International Air Transport Association) 2018的飛航安全報告 (Safety Report) 顯示航空運輸業客人數每百萬航班損失率平均達0.19，相當於每五百四十萬就會有一次嚴重的事故發生。如果能透過飛安事故的歷史資料去進行探勘，或許可以從中發現一些導致傷亡的關鍵原因並進而防範，以降低發生事故的風險。正因為飛航安全非常重要，如何降低飛航風險的方法一直被廣泛研究。

本研究所採用的資料來源微美國聯邦航空聯署(Federal Aviation Administration. FAA)提供的一般民航飛航事故事件庫(Accident/Incident Data System, AIDS)。在這個資料集中都是以發生飛航事故意外的事件，但大多數的意外都能夠被解決，例如機體輕微的損傷，並不會導致急降。不過，少部分的事故意外還是會有嚴重的損死傷，因此，本研究的目標是在飛航事故發生後，找出導致傷亡產生的關鍵因素。

本研究所採用的資料集是選擇較為常見的波音(Boeing)機型為例子。此資料集有不平衡資料集(Imbalanced Dataset)的問題，即發生傷亡與非傷亡資料筆數懸殊。為彌平資料不平衡對於少類別中分類的結果影響，本研究採用SMOTE增樣技術(Synthetic Minority Oversampling Technique, SMOTE)方法以增加少類別樣本。特徵屬性篩選採用資訊獲利(Information Gain, IG)的篩選方式進行。分類建模部分，本研究採用多種分類方法進行比較，分別是簡易貝氏(Naive Bayes)、貝氏網路(Bayes Net)、支持向量機(Support Vector Machine, SVM)、決策樹(C4.5)、徑向基底函數網路(Radial Basis Function Network, RBFN)等十種分類器去分析。最後，本研究使用混亂矩陣作為主要效能指標。

1. 尋找飛航事故發生後，哪些特徵易導致傷亡產生。

2. 應用SMOTE增樣技術以解決資料類別不平衡問題。

3. 應用資訊獲利進行屬性排序並利用門檻值進行篩選特徵屬性。

# 第二章、方法

此章節將提及本研究進行資料前處理的步驟和所使用的演算法及其概念，簡易貝氏、貝氏網路、支持向量機(SVM)、決策樹(C4.5)、徑向基底函數網路(RBFN)。

## 一、實驗過程

* 步驟一：資料收集

本研究資料取自美國聯邦航空總署(2016)所提供的飛安事件資料庫(AIDS)。僅考慮常見的波音系列飛航事故事件，期間為西元2000年1月至2019年12月，共1209筆。

* 步驟二：資料預處理

本研究對多種演算法使用Weka執行並互相進行比較。Weka輸入格式有提供幾種，例如arff或csv，而本研究所使用的格式為csv。確定輸入格式後，需刪減無法匯入Weka的符號 (如：%, ‘ )。若欄位內缺失值過於嚴重則刪除此欄位，其它有缺失值的筆數也斟酌刪減，少部分缺失值因欄位有UNKOWN屬性值而取代為此。接著執行SMOTE，使傷亡比數比非傷亡筆數相同，進而解決資料不平衡的問題，此時資料總筆數為1463筆。在資料平衡之後，開始做欄位刪減的動作，首先，將像是單筆飛航事故事件的識別編號或者飛行載具註冊編號，這類直覺上對飛航安全事故不會有直接或間接影響的欄位，我們也會把它做刪除的動作。接著，利用資訊獲利(Information Gain)來評定屬性重要度，對資料進行降維，將重要度較低的欄位也刪除。原始資料屬性達27種，合併傷亡2屬性後為26種，於此階段將去除無效屬性(資訊獲利為0者)與非重要屬性(如：單筆飛航事故事件的識別代號)。

* 步驟三：使用分類演算法進行建模動作

貝氏網路、簡易貝氏、支持向量機(SVM)、決策樹(C4.5)、徑向基底函數網路(RBFN)和多層感知器(Multilayer Perceptron)等七種演算法建立分類模型，佐以評估指標(總體分類準確度、Recall、F-measure、AUC)來評估此模型的效能。採用10-fold cross validation方式進行樣本訓練與測試。

* 步驟四：結果討論

最後階段將對於實驗結果作闡述，本研究期盼於互相比較之下，能夠發現各方法中所得出的不同特點並針對篩選出的特徵屬性進行更進一步的探討。

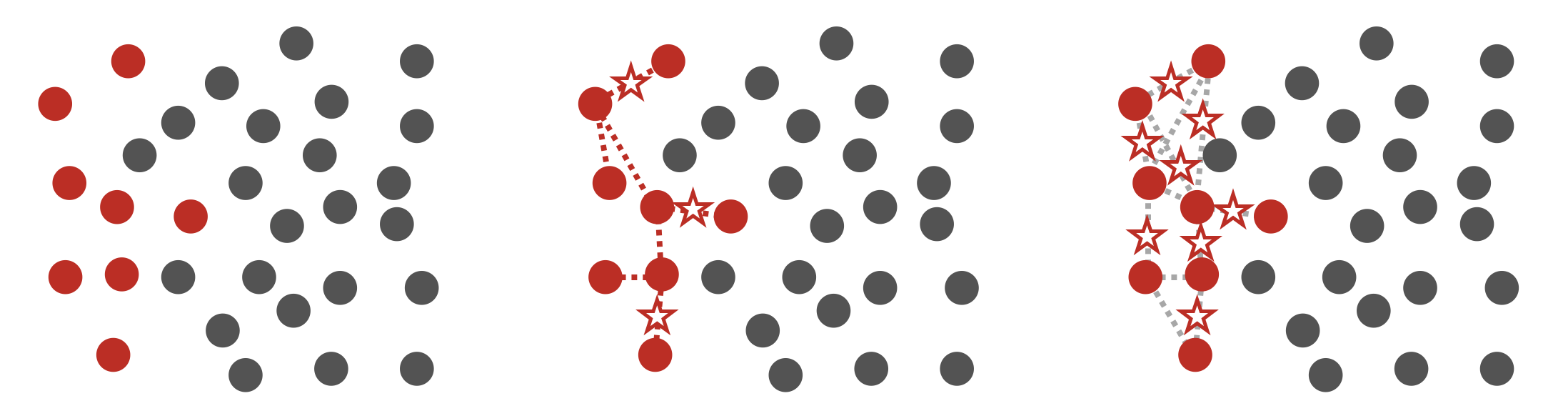
## 二、Synthetic Minority Oversampling Technique (SMOTE)

類別不平衡指的是假設資料分為兩個類別，其原始資料呈現出兩類別之間少數與多數的比例差距大，導致分類結果呈現出多數類別準確率高，但少數類別的分類效果差的情況。

一般傳統的分類方式總是將所有類別的資料樣本平等看待，導致如上述所說的結果發生，但對於大多情形少數類別較為重要，例如本研究對於產生傷亡數的少數資料較為重視一樣，所以必須解決類別不平衡的問題。為解決類別不平衡的問題，可以從兩方向著手，一是尋找針對資料類別不平衡做分類的演算法，二是通過樣本取樣技術，以符合一定平衡比例再去進行分類。前者多通過對於少數類別樣本容易被分類錯誤的特性去增加其權重，再經過多次修正以實現最小損失為收斂目的。後者則是對於資料本身做取樣的動作，這種方式較常被人所用，因為其成本較低。對於解決類別不平衡的問題，本研究所採取的方式是樣本取樣。樣本取樣技術的主要作法是增加少數類別樣本或減少多數類別樣本，以平衡資料類別間的數量差異。

SMOTE全名是Synthetic Minority Oversampling Technique，即合成少數類別過度取樣技術(Chawla el at.,2002)。此演算法是基於改善隨機過度取樣演算法(Random Oversampling)的一種方案，由於隨機過度取樣只會簡單複製以增加少數類別樣本，這樣容易產生過度配適(Over Fitting)的問題。SMOTE演算法的基礎想法是利用所有少數類別樣本中擁有的特性進行重新編排合成，進而得出新的樣本並添加回資料集中。SMOTE(T,N,k)表輸入此演算法的變數含少數類別樣本T，增加N%，與合成樣本時，需參考的k個最近鄰居樣本。

舉個例子，如下圖所示，最左邊為原始資料；中間圖深色的個體是被用來找 k-nearest neighbors 的陽性個體，此處假設我們選擇 k = 3，則對於這兩個點 SMOTE 演算法會先辨認出最鄰近的 3 個鄰近點，接下來會隨機挑選其中 1 個鄰點用來產生新樣本，最後會在被挑到的個體與對應鄰點的連線上隨機產生一個新的個體，並當作這個個體是陽性的；當我們選擇了很多不同的個體去找出 k-nearest neighbors 以及合成新樣本後，最後會的結果會如下圖所示。



El-Sayed et al.(2015) 在研究中使用SMOTE進行增樣以用於處理自閉症不平衡數據，並提高準確性、可信度，結果表明使用SMOTE是可靠的。Dong(2015)利用DNA特徵來預測前列腺癌(prostate cancer)，其原始資料為高維度且類別不平衡。Dong的研究主要使用兩種抽樣方法，一個過採樣(oversampling)和一個欠採樣(undersampling)，以及五個特徵選擇方法，三個分類器訓練，然後評估哪種組合產生最佳結果，實驗結果表明SMOTE顯著有用。Sarakit et al.(2015)進行Youtube的文本情緒分類，因資料集有不平衡問題，所以採用SMOTE進行平衡。並使用三項機器學習算法，即多項式樸素貝葉斯（MNB），決策樹（DT）和支持向量機（SVM）進行測試，結果顯示SMOTE技術可以解決不平衡數據問題並獲得有效改進的分類效果。基於上述，因此本研究採用SMOTE進行類別平衡。

## 三、簡易貝氏及貝氏網路

此分類演算法核心基於貝氏定理而成。它需假設已知目標類別的事前機率，此機率經常從訓練樣本中得知。在給定任一目標類別下，其參與的屬性皆假設為相互獨立，此點與貝氏定理假設相呼應。此方法因其建構速度快，所以適合進行高維度資料分類。於公式面P(C | X) = P(X | C)\*P(C)/P(X)來看，假設訓練資料中有X屬性集合={X1,X2,...,Xk}，且X非包含目標類別屬性，而C為目標類別之屬性值集合，C={C1,C2,...,Cn}，以本研究來說，C包含有傷亡與沒傷亡兩種。P(C | X)為某X屬性集合下，目標類別C出現的機率。於預測資料時，會經由C集合內各目標類別之屬性{C1,C2,...,Cn}進行評比後，選擇機率最多的目標類別屬性值以作為預測分類結果。P(X | C)為某目標類別C下，某X屬性集合所出現的案件機率。P(C)為目標類別的機率。P(X)為某X屬性集合下所出現的案件機率。

貝氏網路，是藉由有向非循環的關係而成的機率圖形。貝氏網路與簡易貝氏不同的地方在於屬性間存在互相影響的關係且非循環，此關係必須先事前得知。舉例而言，貝氏網路可用來表示疾病和其相關症狀間的機率關係；倘若已知某種症狀下，貝氏網路就可用來計算各種可能罹患疾病之發生機率。由於屬性間存在依賴關係，因此比起簡易貝氏來說，更能夠去凸顯出決策者有興趣的目標假設。

## 四、支持向量機

支持向量機(SVM)是一種將高特徵維度進行空間轉化，並對應到一超平面以方便進行分類動作的監督式學習方法。其主要運行方式是將這超平面兩側不同目標類別點進行區隔，其核函式有線性、多項式、半徑型與切線型等。以線性方法來切割為例，假設一超平面為f(x) : W．X＋b=0，X為高維度座標，W為一向量，b為一常數。超平面f(x)能夠將不同的高維度類別集合分開，此切割選擇需以最易切割出兩方類別為優先，簡單來說就是找到一個決策邊界(decision boundary)讓兩類之間的邊界(margins)最大化，使其可以完美區隔開來。於此公式中W與b的最小值常以拉格朗日方法求解。於 (1) 則為利用拉格朗日方法L(a)並配合分類屬性、求解後而得的判斷公式。此公式內為拉格朗日變數，此變數需假設已經求得，、為高維度座標，n為對應於此超平面的點總數，為分類屬性值(如：正類別設為1，負類別設為-1)，則為SVM之核函數公式。其中的變數與r為核心參數，T為轉置矩陣，而d為屬性維度數。

|  |  |
| --- | --- |
|  | (1) |

## 五、決策樹(C4.5&LMT&Random Decision Tree)

C4.5是一種建立決策樹的演算法(Quinlan, 1993)，其前身為ID3演算法。本研究使用C4.5作為分類演算方法的原因，除了以ID3為基礎的分類法常被使用之外，於國內飛航事故因素的研究中，C4.5所建立的分類模型較少被討論。

首先，由於C4.5演算法會涉及到ID3演算法之基礎，因此，先簡述ID3演算法之基礎。ID3演算法所採用的選取屬性節點策略為資訊獲利，資訊獲利的公式於下列 中定義，當資訊獲利代入某一測試屬性A後，此公式將利用屬性A分割前的資訊熵與分割後的資訊熵相減後以得出Gain值。屬性A分割前的資訊熵 中，為第i分類值下的次數，假設為{,…,}之次數總和， =/。屬性A分割後的資訊熵 (2) 中，則與分割前的資訊熵的變數相似，但需以屬性A作為切割主役，並將屬性A之n種值域分開計算。其中j為n種值域下某屬性值之序號，i為某屬性A目標類別的序號，m為分類類別總數。即為某屬性A於第j屬性值所對應之第i種目標類別出現之次數。

|  |  |
| --- | --- |
|  | (2) |
|  |

C4.5演算法基於ID3演算法改良而成，選取屬性的方法是採用資訊獲利比率() = Gain(A) / (A)，此比率需要用到ID3演算法選取節點時所使用的資訊獲利，且配合分裂資訊() (3) 以求得此值。分裂資訊公式中，A屬性下的值域個數有n種，S為資料總筆數，為A屬性下值域子集第j種之出現個數。

|  |  |
| --- | --- |
|  | (3) |

邏輯模型樹(LMT)在計算機科學中是一種分類模型，本身具有關聯的監督式訓練算法，該算法結合了邏輯斯迴歸和決策樹學習。邏輯模型樹是基於模型樹的早期思想，決策數的葉子上具有線性回歸模型，以提供分段線性回歸模型。基本的邏輯模型樹歸納算法使用交叉驗證來找到許多不會過度擬合訓練數據的LogitBoost迭代。

在機器學習中，隨機森林(Random Decision Tree)是一個包含多個決策樹的分類器，其輸出的類別結果是由各個樹輸出類別的眾數而定。隨機森林的優點有以下幾種：

* 可以處理大量的輸入變數。
* 可以在決定類別時，評估變數的重要性。
* 它在遇到有大部分資料遺失的情況時，仍可以維持準確度。
* 對於不平衡的分類集來說，它可以平衡誤差。
* 它對資料探勘、偵測離群值和將資料視覺化非常有用。
* 學習過程快速。

## 六、徑向基底函數網路

類神經網路是一種模擬生物神經細胞的一種分類方式，其主要構成包含輸入層、隱藏層與輸出層。於此三層之間相互以節點連接且單向構成，而每個節點間由連接線連接著，這些連接線為權重。輸入的資料在隱藏層內處理後，才得以於輸出層進行分類。於隱藏層的所使用的函式以 來表示，為分類類別總數，為輸入值總數，為輸出之目標分類屬性值，函式將函式內計算之數值轉換為分類類別代表符號，為輸入值權重，為輸入之屬性值，為偏移量。

近年來，類神經網路已被視為非常有效的非線性模型建構工具。因此，本研究採用徑向基底函數網路，公式如下所示 ，為分類類別總數，為隱藏層節點總數，函式為轉換函式，為權重值，是對於隱藏層至輸出層的連結進行權重加成，為徑向基底函數，為輸入值，為隱藏層第節點之中心點，此中心點功用為搭配群集方式求解所使用，為偏移量。

## 七、多層感知器

多層感知器(Multilayer Perceptron, 縮寫MLP)是一種人工神經網路。多層感知器可以被當作是一個有向圖，由多個節點層所組成，每一層都完全連接到下一層。除了輸入節點之外，每個節點都是一個帶有非線性激活函數的神經元(又稱處裡單元)。多層感知器本身可以使用任何形式的激活函數，像是階梯函數或者S函數，但為了使用反向傳導算法進行有效學習，激活函數必須限制為可微函數。

MLP在過去曾是相當流行的機器學習方法，因為它擁有廣泛的應用領域，如語音辨識、圖像識別、機器翻譯等等，不過在後來90年代，支援向量機的出現讓多層感知器遇上了強勁的對手。到近期，由於深度學習的發展成功，多層感知器就又重新得到了關注。也因為如此，本次實驗也將多層感知器納入實驗的方法當中。

## 八、評估績效指標

本研究指標F-measure = 2\* Precision\* Recall/(Precision + Recall)，此指標結合了精確率(Precision) =TP/(TP+FP)與召回率(Recall)=TP/(TP+FN)此兩種方法結合而成。F-measure值域為0到1，得分越高者越佳。AUC為ROC線下面積，得分越高分就表示ROC越趨向左上方，即分類效能越好。在最後我們會將已發生的飛安事故中，做Apriori關聯法則去了解較常發生的關聯組合，或許從中可以發現一些造成飛安的共通特徵，並進一步推測組合中各自特徵的潛在問題。

# 第三章、結果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Precision | Recall | F-measure | AUC | Accuracy |
| Bayes Network | 0.952 | 0.951 | 0.951 | 0.987 | 95% |
| Hidden Naive Bayes | 0.963 | 0.962 | 0.962 | 0.990 | 96.24% |
| Naive Bayes | 0.949 | 0.947 | 0.947 | 0.987 | 94.73% |
| RBF Network | 0.953 | 0.952 | 0.952 | 0.975 | 95.21% |
| SMO | 0.963 | 0.963 | 0.963 | 0.963 | 96.31% |
| J48 | 0.965 | 0.965 | 0.965 | 0.977 | 96.51% |
| LMT | 0.961 | 0.961 | 0.961 | 0.980 | 96.1% |
| Random Forest | 0.964 | 0.964 | 0.964 | 0.988 | 96.45% |
| Random Tree | 0.951 | 0.950 | 0.950 | 0.976 | 95% |
| Multilayer Perceptron | 0.954 | 0.954 | 0.954 | 0.985 | 95.35% |

|  |  |  |
| --- | --- | --- |
| Input attributes | | Output |
| Primary Flight Type=SCHEDULED AIR CARRIER 30 | | Is\_Fatal = 1 |
| Primary Flight Type=SCHEDULED AIR CARRIER | Nbr of Engines=2 29 |
| Aircraft Engine Make=RROYCE 24 |
| Aircraft Engine Make=RROYCE 24 | |
| Aircraft Engine Model=UNKNOW 24 | |

# 第四章、結論與討論

經過Information Gain及部分人為的篩選後，我們發現”事件發生的機場”、”事件發生的城市”和”事件發生的狀態”，這些欄位特徵對飛安影響較大。其中事件發生的城市、州別和機場這類地域相關的屬性可能跟該地區航線的密度有關。

另外一部分屬性間接的人為關係，像是這邊的Operator(飛行載具所隸屬的航空公司)就可能與營運方的組織方針有關。Aircraft Model(載具型號)可能跟該型號的結構設計有關。

在還未使用SMOTE之前，由於傷亡筆數整體資料集中佔極少數，導致分類結果分類很差，找不到適合特徵。但在使用了SMOTE之後，分類結果就明顯變好許多。這次實驗的分類器比較結果中我們發現J48的分類結果最好。飛航類型、引擎個數和引擎製造商在關聯規則中是導致飛安的重要因子。

這次的實驗資料來源是AIDS資料庫，因為內容有限無法提供更多屬性來做探勘，因此部分導致飛安傷亡的潛在屬性可能還存在，關於這部分希望往後會提供更多的屬性來做資料探勘的研究。