

國立臺灣大學工學院工業工程學研究所

碩士論文

Institute of Industrial Engineering

College of Engineering

National Taiwan University

Master Thesis

發展二元變數之監督與非監督式編碼架構以提升模型

預測表現

Supervised and Unsupervised Encoding Schemes of
Binary Variables for Prediction Performance Enhancement

楊雲皓

Yun-Hao Yang

指導教授：藍俊宏 博士

Advisor: Jakey Blue, Ph.D.

中華民國 112 年 01 月

January, 2023

摘要

人工智能、機器學習與深度學習在近年來被廣泛的應用於各行各業，不論是影像辨識或自然語言處理的發展，遍及製造業、金融業、市場銷售、與影像醫學辨識等領域，實作者前仆後繼地設法將機器與深度學習應用於實際的問題上，以提升日常工作的效率與準確度。然而，機器學習的模型表現並不端看模型建置的技巧與超參數的調教與設置，資料的前處理與編碼方式對於模型表現也有著極為深遠的影響。例如，在處理含有字串的類別變數時，我們往往將單一類別變數編碼成多個數值特徵，例如以獨熱編碼來將類別變數中的字串特徵轉換成二元特徵，以作為輸入資料供模型讀取。但若是類別變數中的類別繁多，進行獨熱編碼後將產生許多的二元變數特徵，如此將稀釋原變數的資訊、並造成維數災難的困境；此外編碼出的二元特徵也不全然與分類器、迴歸機有直接關係；更甚者，二元特徵本身的資料分布往往與諸多機器學習演算法的假設相左。

鑑於以上的挑戰，本研究提出了一創新的監督式與非監督式的編碼方式將二元特徵聚合成少數個整數型別的變數，編碼出來的整數型別變數，可以輕量化地餵入模型，且由於其聚合編碼是透過原二元特徵間的關聯度，因此模型訓練更有效率且最終表現亦有所提升。此創新的聚合編碼方法乃藉由探究二元特徵間的相關係數、主成份的權重等方式將二元特徵分組，再根據各組內特徵的屬性進行排序後，編碼成整數數值。整體而言，該方法力求在縮減維度、提升處理速度的同時，維持模型的整確性與變數的可解釋性。

關鍵字：類別變數、獨熱編碼、監督式／非監督式編碼、二元特徵排序

Abstract

AI techniques have recently been widely applied to the tasks of image recognition and natural language processing. Practitioners from fields such as manufacturing, finance, marketing, and radiology are eager to implement AI methods to enhance daily efficiency and effectiveness. However, AI method performance depends on not only the modeling skills and hyperparameters tuning but also the data preprocessing and encoding. While handling categorical variables, one-hot encoding is commonly used to convert strings into binary features, which can then serve as the input for model training/testing. If the number of categorical levels is large, it consequently creates a large number of features, and the curse of dimensionality would be an essential concern. Furthermore, the one-hot encoding features are created based on the levels of categorical variables and do not guarantee to be related to the classification/regression tasks. Not to mention that the binary feature values often violate the assumptions in machine learning algorithms.

In this research, we develop unsupervised and supervised encoding methods to tackle the aforementioned issues. In unsupervised encoding, we compare the feature properties, such as the column sparsity, PCA-weight, and feature importance, for consolidating related features into a semi-continuous one via binary encoding. In supervised encoding, an optimization scheme is proposed to incorporate the performance improvement of the classifier/regressor and the consolidating orders of the binary features. It is expected to reduce the number of binary features significantly as well as to enhance the classification/regression accuracy through inputting the consolidated features.

Keywords: categorical variable, one-hot encoding, supervised/unsupervised encoding, binary feature sorting

目錄

摘要	i
Abstract.....	ii
目錄	iii
圖目錄	vi
表目錄	x
1 第一章 緒論	1
1.1 研究背景	1
1.2 研究動機與目的	3
1.3 研究架構	5
2 第二章 文獻探討	6
2.1 變數編碼	6
2.1.1 順序編碼.....	9
2.1.2 獨熱編碼.....	10
2.1.3 二進制編碼.....	11
2.1.4 頻率編碼.....	12
2.1.5 目標編碼.....	13
2.2 維度災難	15
2.3 降維處理	19
2.3.1 特徵選取	19
2.3.2 特徵萃取	22
2.4 決策樹相關模型	25
2.4.1 決策樹	25
2.4.2 隨機森林	26
2.4.3 梯度提升決策樹	27
2.5 驗證指標	30
2.5.1 回歸指標	30

2.5.2 分類指標	30
3 第三章 高維二元特徵之聚合編碼技術及分析框架	33
3.1 二元特徵分群	36
3.1.1 資料原始特徵群	36
3.1.2 主成分分析群集	37
3.1.3 相關係數群集	38
3.2 群內二元特徵排序	41
3.2.1 二元特徵總和排序	43
3.2.2 特徵純粹度排序	43
3.2.3 特徵重要度排序	44
3.2.4 基因演算排序法	45
3.3 群組二進碼十進數編碼	47
3.3.1 二進位十位數編碼數值	47
3.3.2 二進位十位數編碼數值排名	48
3.4 二元特徵降維技術	50
3.5 分類與評估指標	53
4 第四章 案例研討	54
4.1 連續二元分類資料測試	55
4.1.1 資料集簡介與實驗架構	56
4.1.2 不同連續資料集之下的測試與實驗	59
4.1.3 分類結果評比與歸納	63
4.2 UCI 資料集	65
4.2.1 資料集簡介與實驗架構	65
4.2.2 分類結果評比與歸納	67
4.3 Kaggle 資料集	68
4.3.1 資料集簡介與實驗架構	68
4.3.2 分類結果評比與歸納	70
5 第五章 結論與建議	72

5.1 研究成果	72
5.2 未來研究方向	74
參考文獻列表	76
附錄 A (如果有).....	78

圖 目 錄

圖 1.1 論文架構。	5
圖 2.1 資料預處理常見步驟 (García et al., 2015)。	6
圖 2.2 依據探索式資料分析進行資料視覺化 (Behrens, 1997)。	7
圖 2.3 獨熱、二進位編碼後的特徵數量比較。	12
圖 2.4 維度個數變化對於分類模型表現的影響 (Spruyt, 2014)。	15
圖 2.5 訓練模型所需樣本個數對應維度變化，以貓狗分類為例 (Spruyt, 2014)。	16
圖 2.6 超球體體積對應維度變化 (Köppen, 2000)。	16
圖 2.7 資料分佈情形對應維度變化，以貓狗分類為例 (Spruyt, 2014)。	16
圖 2.8 高斯核函數值對應距離分布於高維度空間的變化 (Verleysen & François, 2005)。	17
圖 2.9 降維處理的階層化架構 (Tang et al., 2014)。	19
圖 2.10 特徵選取與整體資料分析流程 (Tang et al., 2014)。	20
圖 2.11 基於相關係數的過濾型特徵選擇，結合機器學習流程 (Hall, 1999)。	21
圖 2.12 基於相關係數的包裝型特徵選擇，結合機器學習流程 (Hall & Smith, 1999)。	21
圖 2.13 以階層群集重新排序特徵之相關係數矩陣比較 (Liu et al., 2012)。	22
圖 2.14 以 PCA、LLE 視覺化 Leukaemia 資料集 (Hira & Gillies, 2015)。	23
圖 2.15 (a)表示原始資料分布，(b)中綠線為 PCA 產生的兩主成分 (Abdi & Williams, 2010)。	23
圖 2.16 經由 PCA 將資料投影至主成分座標軸 (Abdi & Williams, 2010)。	24
圖 2.17 決策樹範例，以二元分類問題為例 (Song & Ying, 2015)。	25
圖 2.18 引導聚集算法示意圖 (Efron & Tibshirani, 1994)。	27
圖 2.19 梯度提升決策樹示意圖。	27

圖 2.20 引導聚集算法與提升方法的比較。	28
圖 2.21 LightGBM 於 Flight Delay (左) 與 LETOR (右) 兩資料集中的收斂表現。	29
圖 2.22 混淆矩陣於常見的衡量指標計算。	31
圖 3.1 資料處理、評估流程圖。	33
圖 3.2 研究方法流程圖。	34
圖 3.3 原始二元特徵資料，以動物園資料為例。	36
圖 3.4 原始資料二元特徵間的相關性矩陣。	39
圖 3.5 以塊模型進行置換後的二元特徵間的相關性矩陣。	39
圖 3.6 群集過後的特徵組。	41
圖 3.7 以不同方式排序二元特徵，產生的新數值資料分佈比較，依據新數值特徵分佈。	42
圖 3.8 以不同方式排序二元特徵，產生的新數值資料分佈比較，依據類別區分。	42
圖 3.9 以二元特徵總和，排序各個群組中的二元特徵。	43
圖 3.10 以目標特徵純粹度，排序各個群組中的二元特徵。	44
圖 3.11 各項二元特徵於預訓練分類模型中的特徵重要性。	45
圖 3.12 以特徵重要度，排序各個群組中的二元特徵。	45
圖 3.13 經過特徵純粹度排序的第三特徵組。	47
圖 3.14 BCD 與 Rank BCD 對於新編碼後的特徵分布比較。	49
圖 3.15 呈現整體二元資料降維至三維後的資料分佈，依據特徵純粹度、隨機與特徵和排序方式。	50
圖 3.16 呈現整體二元資料降維至二維後的資料分佈，依照特徵純粹度排序。	51
圖 3.17 呈現整體二元資料降維至二維後的資料分佈，依照隨機排序。	51
圖 3.18 呈現整體二元資料降維至二維後的資料分佈，依照特徵和排序。	52
圖 4.1 模擬的連續二元分類資料。	55

圖 4.2 連續資料集下的實驗架構。	56
圖 4.3 原始連續資料於 X、Y、Z 三維度上的分布情形。	57
圖 4.4 二元化後的連續資料，共劃分為 30 個二元特徵。	57
圖 4.5 連續資料集一的資料分布。	60
圖 4.6 連續資料集一中，不同編碼方式所得數值資料的分類成績，對應切分二元特徵數量變化。	60
圖 4.7 連續資料集一中，不同編碼方式所得數值資料的平均分類成績。	60
圖 4.8 連續資料集二的資料分布。	61
圖 4.9 連續資料集二中，不同編碼方式所得數值資料的分類成績，對應切分二元特徵數量變化。	61
圖 4.10 連續資料集二中，不同編碼方式所得數值資料的平均分類成績。	61
圖 4.11 連續資料集三的資料分布。	62
圖 4.12 連續資料集三中，不同編碼方式所得數值資料的分類成績，對應切分二元特徵數量變化。	62
圖 4.13 連續資料集三中，不同編碼方式所得數值資料的平均分類成績。	62
圖 4.14 UCI 網站上的二手車輛車況評估資料集。	65
圖 4.15 UCI 資料集下的實驗架構，圓圈內為該類特徵個數。	66
圖 4.16 UCI 資料集中，不同編碼方式所得數值資料的平均分類成績。	67
圖 4.17 Kaggle 網站上的類別特徵編碼挑戰資料集。	68
圖 4.18 Kaggle 資料集下的實驗架構，圓圈內為該類特徵個數。	69
圖 4.19 Kaggle 資料集中，不同編碼方式所得數值資料的平均分類成績。	70
圖 4.20 面對無法類別化的二元特徵資料時，所能使用的數值編碼方式。	71
圖 5.1 LightGBM 分類模型訓練後的特徵重要度。 (左) 原始二元資料 (右) 主成分分析群組	72
圖 5.2 Kaggle 資料集中，依據不同的特徵個數群組二元特徵下的分類成績。(左) 主成分分析群組 (右) 相關係數群組	74

圖 5.3 二元資料特徵相關係數圖，其中 GB 表示目標欄位。.....	75
圖 5.4 不同群組方式產生之特徵相關係數圖，其中 GB 表示目標欄位。（左）相 關係數群組（右）二元特徵群組資訊	75

表目錄

表 1.1 獨熱編碼後產生的二元特徵，以居住城市為例。	1
表 1.2 常見於製造業中的在製品製程紀錄。	2
表 2.1 不同變數類別的定義與描述 (Stevens, 1946)。	8
表 2.2 變數類別接受運算子與範例。	9
表 2.3 不同編碼方式所對應的模型準確度 (Potdar et al., 2017)。	9
表 2.4 順序、二進制、獨熱與頻率編碼的比較，以居住城市為例。	13
表 2.5 目標編碼後的特徵欄位，以水果價格為例。	14
表 3.1 變數與符號定義。	35
表 3.2 具有群組資訊的二元特徵資料。	37
表 3.3 缺乏群組資訊的二元特徵資料。	37
表 3.4 不同主成分之下的二元特徵權重絕對值。	38
表 3.5 依據主成分分析群集二元特徵。	38
表 3.6 依據相關係數群集二元特徵。	40
表 3.7 排序前各群組中的二元特徵，由 C_{ij} 表示。	46
表 3.8 染色體範例，以基因演算法排序組內特徵。	46
表 3.9 常用的 BCD 編碼方式，與對應的十位數值。	47
表 3.10 各個樣本轉換後的新數值。	48
表 4.1 二元化後的特徵資料。	58
表 4.2 類別化後的特徵資料。	58

第一章 緒論

本章節將描述將高維度二元特徵作為機器學習模型的輸入時，所面臨到的挑戰，後提及本研究之目的與架構。

1.1 研究背景

在機器學習的過程之中，處理輸入資料即時，時常會遭遇到字串型別特徵，像是對於受測者血型、居住城市的描述皆以字串形式呈現。為了將類別特徵輸入模型之中，則必須透過各種變數編碼方式來對無法作為模型輸入的類別變數進行編碼，已將其轉換為數值型別；例如，當面臨以字串描述體積的類別特徵：「大、中、小」時，我們可以依照相對體積的順序關係，將其編碼為「大：3、中：2、小：1」如此數值型態的特徵便能作為模型的輸入，此方法稱為順序編碼。然而，當今天面臨的是描述城市種類的類別變數：「紐約、倫敦、東京」時，由於城市之間並不存在著明顯的順序關係，若是編碼成「紐約：1、倫敦：2、東京：3」會使得訓練的機器學習模型誤解城市與城市之間的關聯性，而導致錯誤判讀的情況發生。當面臨類別變數內的各個種類不存在順序關係、種類繁多且出現頻率相近時，通常使用獨熱編碼（one-hot encoding）來為各個種類產生新的虛擬特徵來表示，如表 1.1 所示。

表 1.1 獨熱編碼後產生的二元特徵，以居住城市為例。

種類特徵	居住城市				
	紐約	倫敦	東京	台北	上海
樣本 1	0	1	0	0	0
樣本 2	0	1	0	0	0
樣本 3	1	0	0	0	0

這些虛擬二元特徵之間存在強烈的互斥關係，因而有高度的關聯性，並稀釋了原先單一特徵的資訊。而類別特徵中的種類越多，獨熱編碼後也將產生越多的虛擬特徵，使得模型訓練時也將消耗更多的記憶體與運算時間、最終導致模型難以有效的收斂和進行訓練。為減緩獨熱編碼造成的維度膨脹問題，許多不同的編碼方式也

被提出，像是透過較少量二元特徵來描述的二進制編碼、改以類別於特徵中出現頻率取代的頻率編碼、以及用相對於目標值的平均值取代的目標編碼等。

在製造業當中，具有眾多二元特徵的資料卻相當常見；如表 1.2，常見於描述在製品於製造過程中通過的特定機台與工序、以及品質管制當中表示合格與否的檢測項目，而且再加上由於不同製程在所使用機台、工序上的要求也大不相同，也使此類二元特徵資料並不一定滿足獨熱編碼的假設與特性；又如表 1.1 中二元特徵帶有群組資訊、或是樣本間存在特徵互斥。該如何前處理這些眾多且又相互關聯的二元特徵也成了一大難題。

表 1.2 常見於製造業中的在製品製程紀錄。

在製品編號、種類	機台 A	機台 B	工序 1	工序 2	檢驗 α	檢驗 β
1 (一般製品)	1	0	1	0	1	1
2 (重工製品)	1	1	0	1	1	0
3 (機台測試)	0	1	0	0	0	1
4 (成品檢驗)	0	0	0	0	0	1

1.2 研究動機與目的

本研究試圖透過非監督式與監督式的方法，對眾多相互關聯的二元特徵進行群組、排序與編碼後，以求大幅度縮減資料維度、壓縮資料資訊、縮減模型讀取時間、並在一定程度上維持或提升機器學習模型對於資料的分類結果。相較於傳統變數編碼僅處理類別型別的資料，本研究設法規劃出針對眾多二元特徵資料的編碼方式，透過聚合編碼二元特徵產生少量新的數值變數，以求在提升儲存效率與表述原二元特徵資料的同時，還能提升資料的可預測性，綜合而言，主要研究目標可分為以下四點：

1. 縮減資料特徵個數：

維度的增加對於機器學習模型的成果有著深遠的負面影響，過多的維度將使得模型難以收斂、延長訓練資源與計算時間。透過對二元特徵群組後進行群組後編碼，將能大幅度的縮減特徵總數；在縮減資料總維度以利表示與描述的同時，也減低模型花費在資料讀取的時間。

2. 壓縮特徵資訊：

獨熱編碼過後的二元特徵資料，總體資訊不改變，但特徵總個數的上升；意味著眾多二元特徵瓜分了原先單一類別特徵所包含的資訊。透過群組後編碼二元特徵，回復特徵平均的資訊含量與重要性。

3. 產生相互獨立的特徵：

符合獨熱編碼特性的二元特徵與彼此間具有著強烈的互斥關係，而特徵與特徵之間的不獨立導致的共線性問題，也代表著特徵能用以預測彼此。透過群組相關性、相似程度過高的特徵並編碼，產生故不相關的新特徵，來解決特徵與特徵嚴重共線性的問題。

4. 維持或提升編碼過後資料的分類表現：

只包含二元特徵的資料，即只存在大量 0、1 的資料對於機器模型的訓練與預測也造成困難。當模型再進行最佳化與求解時大多仰賴梯度計算，而眾多的 0 將

使模型無法計算梯度，導致模型的訓練緩慢、劣化預測成效等負面影響。編碼後的新變數將變為整數型別，避免二元型別影響模型的梯度計算。

1.3 研究架構

本研究目標為透過對於二元特徵的資料進行群組後的排序、與編碼，來壓縮資料維度並維持一定程度的分類成果。在本章中已簡介研究動機、目標與整體架構；由第二章文獻回顧闡明各個不同種類變數編碼的用意與目的、高維度資料所導致的維度災難、在機器學習模型分類時所遭遇的難題，及處理維度災難時常採用的方法。第三章研究方法中，描述本研究如何對於多維度二元資料特徵進行群組、組間特徵排序、以及編碼各個特徵組以生成新的數值資料。第四章案例分析將透過模擬出的測試資料、加利福尼亞大學爾灣分校提供的開源資料集（UCI Dataset）、與 Kaggle 數據建模和數據分析競賽平台上的資料集，結合本研究所提出之編碼方式，與原始二元資料、不同的編碼方式產生的數值資料一同給予機器學習模型做分類成果比較。第五章結論建議部分將對於根據案例實施的成果做出總結，並歸納出來本研究的後續發展方向。本論文的整體架構可見於圖 1.1。

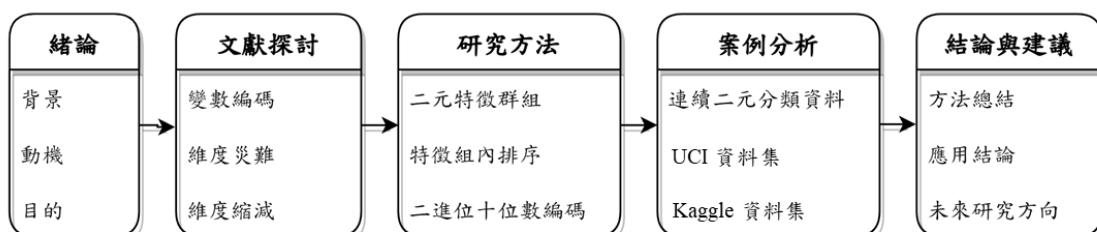


圖 1.1 論文架構。

第二章 文獻探討

本章節探討研究欲解決之問題，與研究之相關文獻。囊括了面臨類別變數時常使用的變數編碼方式；以及在處理多維度資料下所面臨的難題，而後提及透過特徵選取、特徵萃取等方式來減緩維度災難。

2.1 變數編碼

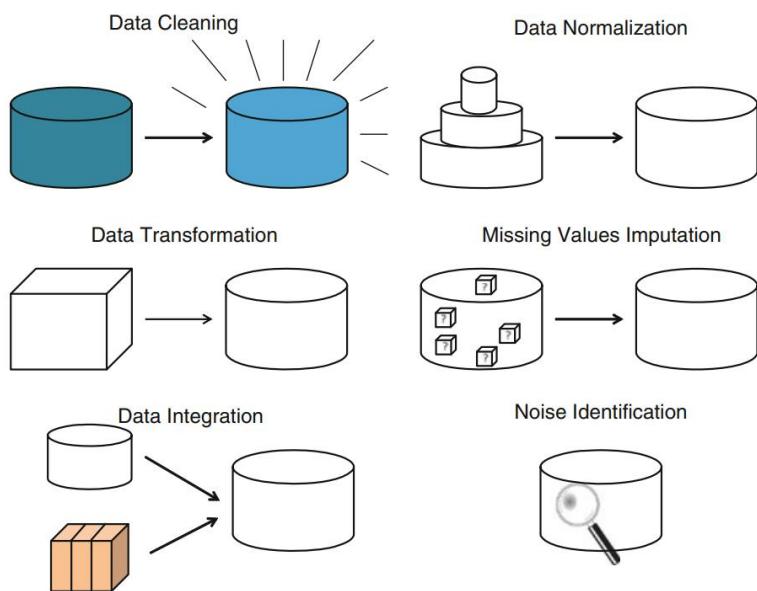


圖 2.1 資料預處理常見步驟 (García et al., 2015)。

收集完資料與定義問題後，在進入機器模型建模之前，若是資料具有如字串等無法作為模型輸入的類別特徵，則必須先經過變數編碼 (variable encoding) 步驟，以將資料調整為適合模型輸入的形式。變數編碼則屬於資料前處理 (data preprocessing) 中的階段，也屬於為 ETL (Extract-Transform-Load)，資料的前處理通常包括了以下幾點，如圖 2.1 中所描述：

1. 資料清理 (data cleaning)：

主要目標為將資料中的缺失、不完整或錯誤的數值進行刪除或填補。錯誤的數值指的是含有亂碼、無法閱讀的符號的數據；而填補的數值通常為眾數或為平均值。次要的目標則常見為透過回歸、群集的方式來對離群值 (outlier) 與噪點 (noise) 進行平滑化的處理。

2. 資料整合 (data integration) :

資料整合為將不同的資料集合併的過程。於收集資料時，時常需將來源不同的資料作合併，作為同一資料集供模型學習；為此即須規範資料中不統一的量測尺度、合併相關且過於冗餘的欄位、同時避免、型別不一致、或資料重複與衝突等問題。

3. 資料型別轉換 (data transformation) :

又可稱作特徵轉換；資料經由轉換、濃縮以滿足模型或演算法對於輸入的要求，以供模型輸入、或提升整體執行效果效率。變數編碼便常見於此過程，經由特定的編碼模式，將原先屬於字串、布林型別的變數編碼成數值型別；與此同時產生良好的訓練資料，以便模型能更好地進行最佳化。

4. 探索式資料分析 (Exploratory Data Analysis, EDA) :

由 Tukey (1977) 提出，主要以統計、視覺化等快捷的方式呈現資料，以利於分析者從各方面快速理解資料及本身與其特性；包括但不限於如觀察資料中各個特徵的分布狀態、資料點之間的距離分布、特徵之間的關聯關係、與降維之後的資料分布情形等等。由於探索式資料分析並非為必須執行的前處理步驟，故在資料預處理時常會容易被忽略，但是恰當的探索式資料分析能有效地對於資料提出洞見，並避免以盲人摸象的方式進行資料分析。

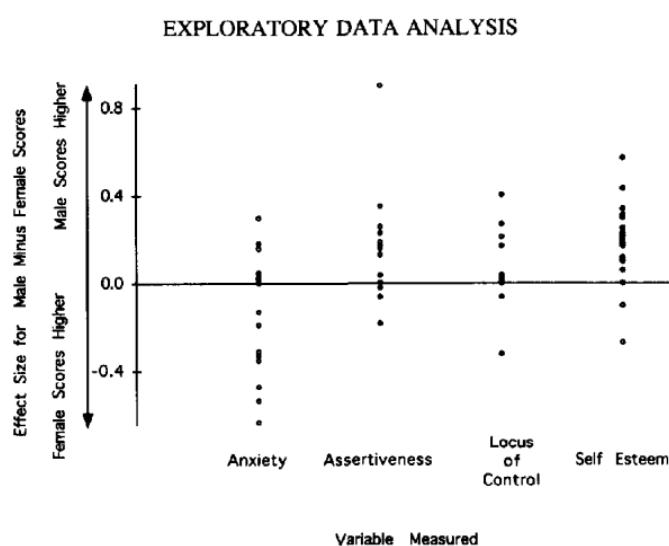


圖 2.2 依據探索式資料分析進行資料視覺化 (Behrens, 1997)。

在進行變數編碼之前，還應理解類別特徵的屬性和類別間的關聯關係，如此才

能挑選適當的編碼方式。Stevens (1946) 將變數的刻畫區分的相當詳盡，能依據不同經驗法則、辨別標準與數值結構進行區分，分為名義尺度 (nominal scale)、順序尺度 (ordinal scale)、等距尺度 (interval scale)、以及比例尺度 (ratio scale) 等四種不同的變數型態，其中等距尺度與比例尺度以數值的形式表示，可以直接作為輸入供模型使用；然而名義尺度與順序尺度則為類別型式紀錄，以該二種型態出現的變數稱之為類別變數 (categorical variable)，若欲將其做為模型輸入還需經由變數編碼做型別轉換，詳細的變數屬性如表 2.1 所示。

表 2.1 不同變數類別的定義與描述 (Stevens, 1946)。

Scale	Basic empirical operations	Mathematical group structure	Permissible statistics (invariantive)
Nominal	Determination of equality	Permutation group	Number of cases, Mode, Contingency correlation
Ordinal	Determination of greater of less	Isotonic group	Median, percentiles, Percentiles
Interval	Determination of equality of intervals of difference	General linear group	Mean, Standard deviation, Rank-order correlation Product-moment correlation
Ratio	Determination of equality of ratios	Similarity group	Coefficient of Variation

在這四種變數型態中，等距尺度、比例尺度皆是以數值方式呈現，可直接進行加減的數值運算，能直接交由機器學習模型作為輸入；而順序尺度與名義尺度以字串、或是布林的型別出現，為此需要進行變數編碼來將其轉換為數值，以處理機器學習模型無法接受類別變數型態。針對類別間具有關聯關係的順序尺度，一般常以順序編碼 (ordinal encoding) 處理，依照關聯性給予連續的正整數值取代原先類別；面對類別之間互不關聯的名義尺度時，常見且有效的方法有獨熱編碼 (one-hot encoding)、目標編碼 (target encoding)。

表 2.2 變數類別接受運算子與範例。

尺度	類型	接受運算子	範例
名義尺度	非計量（類別）	=、≠	性別（男性、女性）
順序尺度		=、≠、>、<	體積（大、中、小）
等距尺度	計量（數值）	=、≠、>、<、+、-	滿意度（1, 2, 3）
比例尺度		=、≠、>、<、+、-、×、÷	體重、身高等

不同的編碼方式將對機器學習模型造成一定程度的影響，Potdar et al. (2017) 使用的處理類別變數時常見的不同編碼方式對 UCI 的車輛評估資料集（Car Evaluation Data Set, 1990）進行預處理後交由 ANN 訓練分類，探討了不同的編碼方式將對於 ANN 分類模型訓練完後的預測準確度有著明顯影響，如表 2.3 所示。可見在選擇特定的變數編碼前，必須先明白各個特徵本身的刻度關係，再選擇相對應的編碼方式，才能使機器學習模型能正確的識別類別特徵關係，使模型導出更優異的分類成果。

表 2.3 不同編碼方式所對應的模型準確度 (Potdar et al., 2017)。

Encoding Technique	Accuracy (Percentage)
One Hot Coding	90
Ordinal Coding	80
Sum Coding	95
Helmert Coding	89
Polynomial Coding	91
Backward Difference Coding	95
Binary Coding	90

2.1.1 順序編碼

對於以字串來描述具有相關順序的類別特徵，例如身高：「高、中、低」；體重：「重、中、輕」等具有明確物理上的意義順序、且可以通過順序排列類別的特徵，便適合使用順序編碼（Ordinal Encoding）。概念便是根據特定物理意義、或是類別含意，以類別數量個整數來描述原先的類別特徵，範例可見表 2.4。

假設有一類別特徵資料 Z 具有 l 個類別特徵，每一類別特徵由 Z_k 表示，且 $0 \leq k < l$ 。

$k \leq l$ ；經由頻率編碼後的新數值資料 Y 具有 m 個數值特徵，每一數值特徵以 Y_j 表示，且 $0 \leq j \leq m$ ； Y_j 由 0 到 $|Z_k| - 1$ 之間的正整數組成，用以描述原先的 Z_k 中的 $|Z_k|$ 種類別：

$$Z = \{Z_1, Z_2, \dots, Z_l\} \quad (2.1)$$

$$m = l \quad (2.2)$$

$$Y = \{Y_1, Y_2, \dots, Y_m\} \quad (2.3)$$

$$Y_j = \{0, 1, \dots, |Z_k| - 1\} \quad (2.4)$$

順序編碼有著編碼簡單、點位密集的優勢、且維持了原先的特徵數量；但是等分位類別的處理方式也限制了描述類別間距的彈性；倘若類別與類別間的差距不一致、甚至或差甚遠時，順序編碼將無法反映出此一關係。

2.1.2 獨熱編碼

獨熱 (one-hot) 為在數位電路與機器學習領域之中，描述一種位元組或是向量的表現形態。在同一樣本之中，互相關連的獨熱欄位群組中只允許存在一個 1，其餘相關欄位必須為零；而在統計、經濟學中，這些相關的獨熱欄位則被稱呼為虛擬變數。經過獨熱編碼 (One-hot Encoding) 後的資料請見表 2.4。

當面對的資料特徵並非數值、且種類之間沒有物理與特性上的順序時，便可使用獨熱編碼進行變數的轉換，來避免模型誤解種類之間存在特定的關聯關係。獨熱編碼為透過虛擬變數 (dummy variable) 來描述原先的類別特徵。

假設有一類別特徵資料 Z 具有 l 個類別特徵，每一類別特徵由 Z_k 表示，且 $0 \leq k \leq l$ ；經由獨熱編碼後的新數值資料 Y 具有 m 個數值特徵，由複數個向量 v_k 所組成；其中各個向量 v_k 內僅包含零與一、且 v_k 總和等於一。

$$Z = \{Z_1, Z_2, \dots, Z_l\} \quad (2.5)$$

$$m = \sum_{k=0}^l |Z_k| = \sum_{k=0}^l |v_k| \quad (2.6)$$

$$Y = \{v_1, v_2, \dots, v_l\} \quad (2.7)$$

$$v_k \in \{0, 1\}^{|Z_k|}; \sum_{h=0}^{|Z_k|} v_{kh} = 1 \quad (2.8)$$

獨熱編碼在一定程度上協助了機器學習模型遭遇屬性變數的時的處理能力，雖然會使得資料總體維度上升，但其仍是個有效且直觀的變數編碼方法。假如今天的種類特徵是在描述台灣的 21 個縣市，則獨熱編碼過後便會產生 21 個虛擬變數欄位；雖然提升了特徵總數，但總體的資訊卻沒有增加，代表獨熱欄位的 1 零散地被 0 所包圍，並散落在這些虛擬變數的之中，導致資料趨為稀疏。而相同群組內的特徵間存在於完全的互斥關係，只要掌握樣本獨熱（數值為一）的特徵位置便能預測出其餘為冷（數值為零）的特徵數值，使獨熱編碼後的高維資料存在著嚴重的共線性問題。

雖然有著造成總特徵個數膨脹、使資料失去梯度的問題存在，但獨熱編碼本身的便利性使得其仍然被廣泛地採用。也有許多的研究在針對獨熱編碼進行改良。像是二進制編碼，即是希望以更為精簡、少量的二元虛擬特徵欄位來描述原始類別資料。

2.1.3 二進制編碼

相較獨熱編碼，二進制編碼（Binary Encoding）以更少的二元虛擬特徵描述了相同數量的特徵種類，在某一類別特徵，具有 n 個種類的情形下，獨熱編碼需要 n 個虛擬特徵描述原始特徵中的各個類別；然而二進制編碼只使用 $\lceil \log_2(n) \rceil$ 個虛擬特徵，在一定程度上減緩了維度的膨脹，如圖 2.3、與表 2.4 所示。缺點則在於面對無序特徵時，二進制編碼後的虛擬變數無法如獨熱編碼一樣有效的解釋變數所包含的意義。

假設有一類別特徵資料 Z 具有 l 個類別特徵，每一類別特徵由 Z_k 表示，且 $0 \leq k \leq l$ ；經由獨熱編碼後的新數值資料 Y 具有 m 個數值特徵，由複數個向量 v_k 所組

成；二進制編碼可視為將順序編碼後產生的新特徵，做了一次十進位到二進位的轉換。

$$Z = \{Z_1, Z_2, \dots, Z_l\} \quad (2.9)$$

$$m = \sum_{k=0}^l (|Z_k|)_2 = \sum_{k=0}^l |v_k| \quad (2.10)$$

$$Y = \{v_1, v_2, \dots, v_l\} \quad (2.11)$$

$$v_k \in \{0, 1\}^{(|Z_k|)_2} \quad (2.12)$$

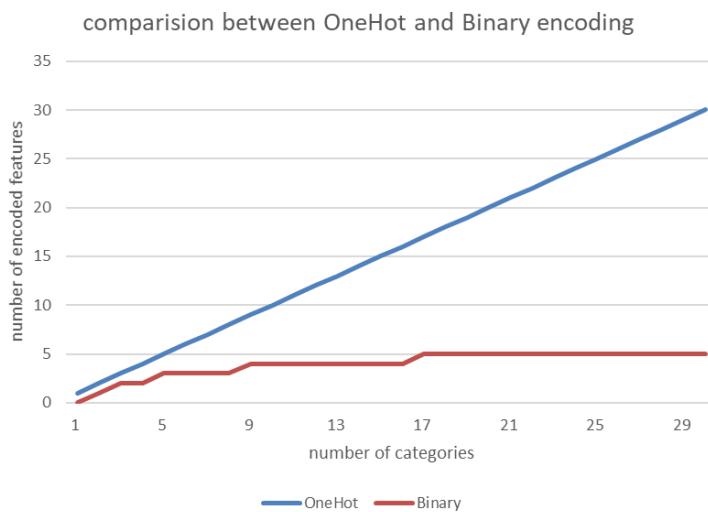


圖 2.3 獨熱、二進位編碼後的特徵數量比較。

2.1.4 頻率編碼

頻率編碼 (Frequency Encoding) 的想法相當簡單直接，便是以某一特定種類於類別特徵中出現的頻率來取代該類別字串；頻率編碼假設在收集資料時，資訊的重複比例即是富有價值的資訊，並進一步設法利用該資訊，來進行變數編碼。

假設有一類別特徵資料 Z 具有 l 個類別特徵，每一類別特徵由 Z_k 表示，且 $0 \leq k \leq l$ ；經由頻率編碼後的新數值資料 Y 具有 m 個數值特徵，每一數值特徵以 Y_j 表示，且 $0 \leq j \leq m$ 。

$$Z = \{Z_1, Z_2, \dots, Z_l\} \quad (2.13)$$

$$m = l \quad (2.14)$$

$$Y = \{Y_1, Y_2, \dots, Y_m\} \quad (2.15)$$

$$Y_j \leftarrow f(Z_k) \quad (2.16)$$

然而當有特徵中有類別的出現頻率相同時，便會造成混淆，如表 2.4 所示。因此當資料本身類別多、且有重複出現頻率的類別時要格外留意此一狀況的發生。有時若遭遇相同頻率的類別時，也會以各別增減特定數值以利區分，例如：「台北：0.2, 桃園：0.2, 新竹：0.2」改以「台北：0.15, 桃園：0.2, 新竹：0.25」。

表 2.4 順序、二進制、獨熱與頻率編碼的比較，以居住城市為例。

居住城市	順序編碼	二進制編碼	獨熱編碼					頻率編碼
台北	0	0 0 0	0	0	0	0	1	0.2
桃園	1	0 0 1	0	0	0	1	0	0.2
新竹	2	0 1 0	0	0	1	0	0	0.2
台中	3	0 1 1	0	1	0	0	0	0.2
台南	4	1 0 0	1	0	0	0	0	0.2

2.1.5 目標編碼

目標編碼（Target Encoding）又可以稱為平均值編碼（Mean Encoding），不同於前面所提及許多的編碼方法，目標編碼為一種監督式的變數編碼方式，意即在編碼的過程之中，有參照了目標欄位；編碼的方式為把某一類別特徵中，同樣種類的資料對應的目標欄位數值加總後，除以類別個數取得該類別對應目標的平均值，並且將這平均值做為新的數值特徵。相較於獨熱、二進制編碼，順序、頻率與目標編碼轉換後的特徵欄位個數維持在單一數值特徵當中，避免了獨熱編碼後造成特徵膨脹的問題。

假設有一類別特徵資料 Z 具有 l 個類別特徵，每一類別特徵由 Z_k 表示，且 $0 \leq k \leq l$ ；經由目標編碼後的新數值資料 Y 具有 m 個數值特徵，每一數值特徵以 Y_j 表

示，且 $0 \leq j \leq m$ 。

$$Z = \{Z_1, Z_2, \dots, Z_l\} \quad (2.17)$$

$$m = l \quad (2.18)$$

$$Y = \{Y_1, Y_2, \dots, Y_m\} \quad (2.19)$$

$$Y_j = t(Z_k) \quad (2.20)$$

在 F_j^{num} 中，則以各個 F_k^{cate} 中元素的目標值平均做取代；如表 2.5 所示，原先的類別特徵有著數值不一的目標數值，但是再經由目標編碼之後，原先的類別特徵將由同一類別對應目標欄位的平均值所取代。

表 2.5 目標編碼後的特徵欄位，以水果價格為例。

類別特徵	目標欄位	目標編碼
香蕉	5	10
香蕉	15	10
鳳梨	40	35
鳳梨	30	35
蘋果	30	30

2.2 維度災難

維度災難（Curse of dimensionality）又可稱之為 Hughes 現象（Hughes Phenomenon），是描述在樣本總數不改變的情形下，當特徵（同為樣本的空間維度）增加時，將面臨到的難題；包括資料分布範圍增大，而導致樣本之間距離增大、數據變為稀疏；因而造成高維空間中所需訓練樣本的數量不足，低維度的空間特性、統計性質無法推廣至高維空間，使歐式距離的計算與對資料的常態假設失去效用，因而使得在維度持續提升的情形之下，機器學習模型的成效不增反降的結果。

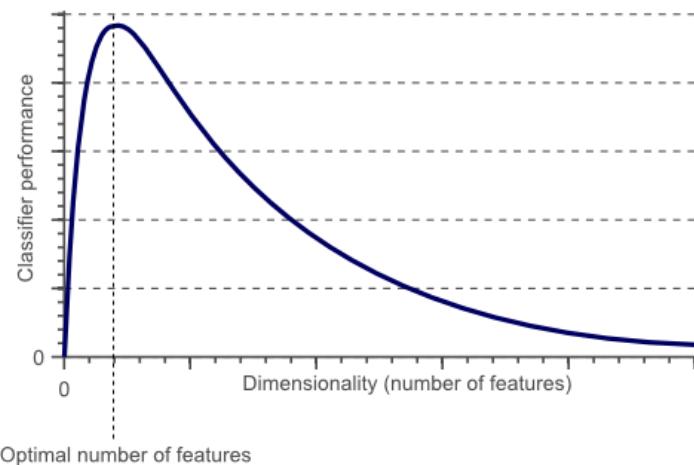


圖 2.4 維度個數變化對於分類模型表現的影響 (Spruyt, 2014)。

為了使模型達到更佳分類成果，分析者往往藉由收取更多資訊、提供更多特徵給模型分析的方式；然而，當特徵個數超過一定的水平後，模型的成效將會不增反減，主因為過多的維度將使得模型訓練、調整的參數增加、機器學習模型擬合了訓練資料中的噪音誤差，導致無法對測試資料做出適當的泛化而導致過擬合的情形，分類成效也將隨著維度上升而下降；而除了難以收斂與有效的訓練模型之外，同時也將導致訓練所需資料、與訓練時間的增加。如圖 2.5 中所描述，假設欲以全部樣本的百分之二十做為訓練資料，隨著維度提升，每一特徵所需的樣本比例也隨之提高以應付資料分佈趨於稀疏，才使得在樣本個數固定的情形之下，提升維度卻造成了模型過度擬合訓練資料。

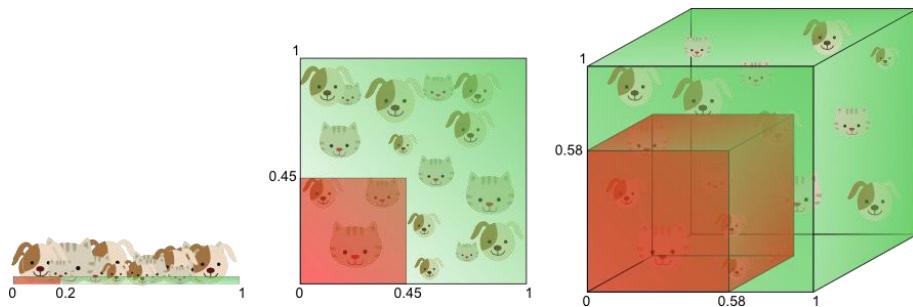


圖 2.5 訓練模型所需樣本個數對應維度變化，以貓狗分類為例 (Spruyt, 2014)。

更糟糕的是，面對高維度資料時，必須重新審視一些對於資料的假設；像是在低維度時可以假設資料為常態分佈，並使用統計手法推斷資料本身特性、以及透過歐式距離、馬式距離來描述樣本距離。然而維度的提升導致的資料稀疏性將使這些常用的方法難以再被使用。如圖 2.6 中，隨著維度升高，中心超球體體積將不斷減小，分布於超球體內的資料個數也隨之減少；反之表示，維度若是持續增加，隨著中心超球體體積減小，多數的資料將開始集中於超球體外的角點之上，見圖 2.7。

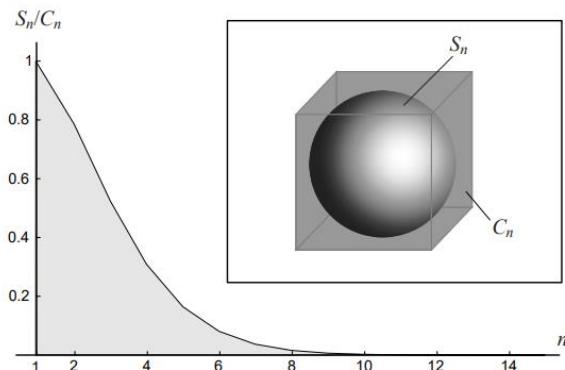


圖 2.6 超球體體積對應維度變化 (Köppen, 2000)。

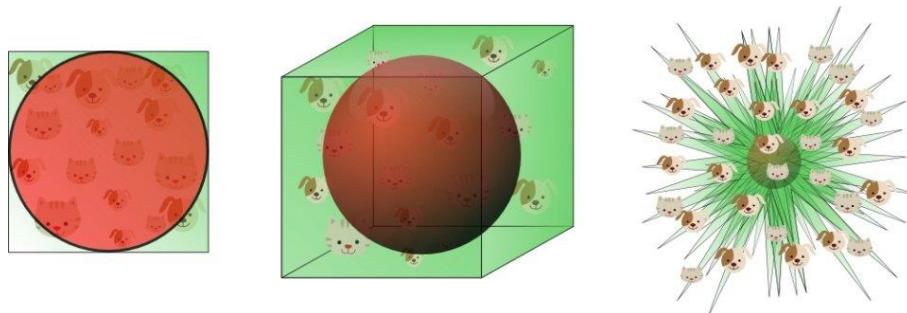


圖 2.7 資料分佈情形對應維度變化，以貓狗分類為例 (Spruyt, 2014)。

因此，當特徵總維度個數趨向無窮大時，從各個樣本點到質心的最小和最大歐

幾里得距離之差與最小距離本身之比趨於零，造成了距離計算在高維度的空間中失去作用，無法為仰賴距離分類的分類器量測有意義的距離，如下方程式所示

$$\lim_{d \rightarrow \infty} \frac{dist_{\max} - dist_{\min}}{dist_{\min}} \rightarrow 0 \quad (2.21)$$

距離的增加導致了資料之間的稀疏性，在多數機器學習的模型之中，如 RBFN (Radial-Basis Function Networks)、SVM (Support Vector Machines) 與 LS-SVM (Least-Squares Support Vector Machines) 等，皆是以高斯核函數來做為計算距離的方式，來斷定樣本之間的遠近關係。如圖 2.8，隨著維度升高，資料點間的距離也開始增加，使得任兩點間的高斯距離分佈（鐘形曲線）逐漸右移，5% 和 95%（垂直虛線）距離對應的核函數數值（實心遞減曲線）也愈來愈相似，表示在高維度空間之中，以高斯核函數則無法有效地區分出樣本間的遠近關係，相較於處於低維度空間時的顯著成效將有著明顯落差。

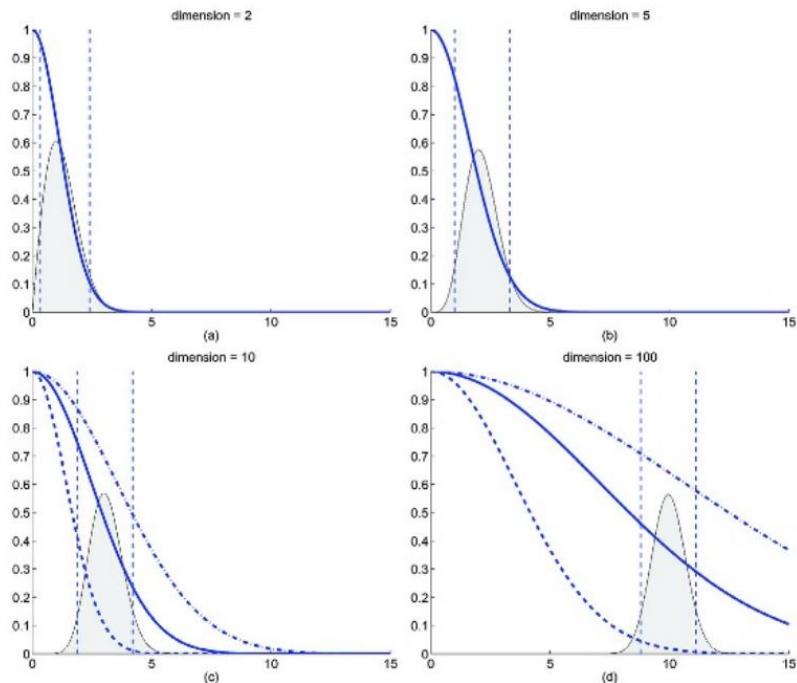


圖 2.8 高斯核函數值對應距離分布於高維度空間的變化 (Verleysen & François, 2005)。

然而，隨著近年來資料分析技術與機器學習的盛行，高維度資料集也變得相當的廣泛、常見。而該如何處理高維度資料也成了富有研究價值的主題。目前較為常見的做法包括降低維度、或是以特徵萃取做為高維資料的前處理方式。其目的皆在

以降低變數、維度的個數來描述原先的高維度資料，同時保留原先資料樣本之間的特性與關聯性，可使用於資料視覺化與模型訓練。

2.3 降維處理

當所欲處理資料之維度過高，無法進行視覺化或是分析無效率時，降維 (Dimension Reduction) 是資料前處理的一種手段，亦是當前機器學習中所謂的特徵工程的主流方法。其目的乃是希望以壓縮原資料原有特徵、同時維持原資料本身特性、和資料點之間的關聯關係；改將原資料投影到低維度的座標上，來做為原資料的代表，便於進行資料的分析、讀取、甚至作為新的模型輸入。如下圖，可以分為特徵選取與特徵萃取兩種方法。

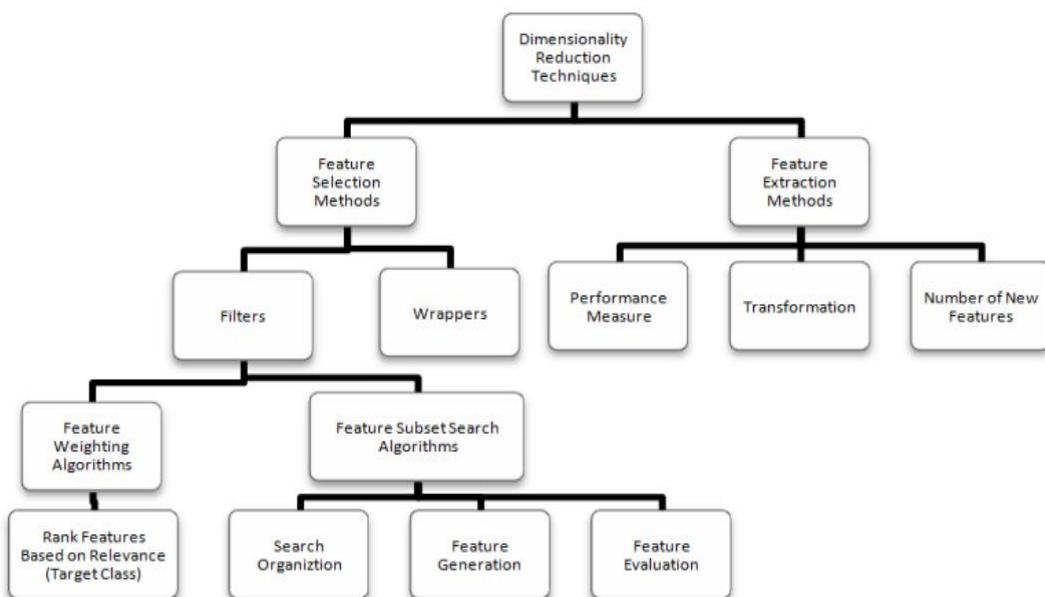


圖 2.9 降維處理的階層化架構 (Tang et al., 2014)。

2.3.1 特徵選取

特徵選取 (Feature Selection) 旨在假設數據之中含有許多冗餘或無關的特徵，並透過從資料集中移除部分不具備夠多信息的特徵，從原有資料集的特徵之中挑選出最具代表性、富有資訊的重要特徵子集合。若特徵選取得當，包含了極具鑑別能力的最優特徵子集，便能達到簡化機器學習模型的訓練時長、避免過度擬合、提升模型準確度，以及便於理解特徵於模型輸出之間的關聯關係等目的。

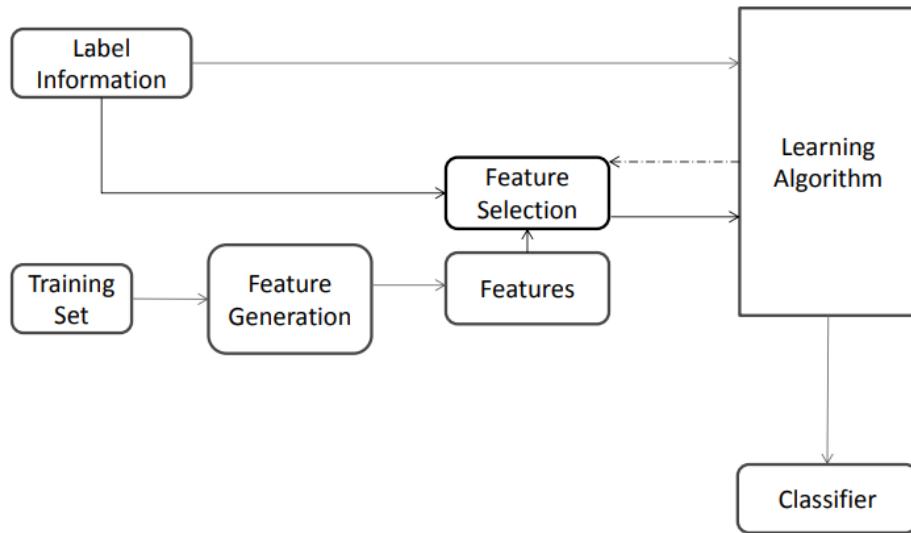


圖 2.10 特徵選取與整體資料分析流程 (Tang et al., 2014)。

依據不同特徵選取的方式，可以區分為三類：

1. 過濾法 (Filter)：透過指標評比每一特徵，並依據閾值或者預選定特徵個數案大小排名選取；這些指標可能為皮爾森相關係數、解釋變異等。
2. 包裝法 (Wrapper)：用模型測試、評比若干不同的特徵子集，並依據分數排除或選取特徵。
3. 嵌入法 (Embedded)：與過濾法相似；即在模型與演算法訓練的同時計算各特徵權重與指標分數，依此進行選取。

Hall (1999)則透過結合相關係數作為選取、或是群組特徵的依據，並結合過濾法與包裝法便發展出了一種監督式結合概念的特徵選取方法；其主要依據各個特徵對應目標欄位的相關性給予特徵排名，並篩選、保留與目標欄位有強相關性的特徵，以減少資料總維度，其篩選流程如圖 2.11、圖 2.12 所呈現。

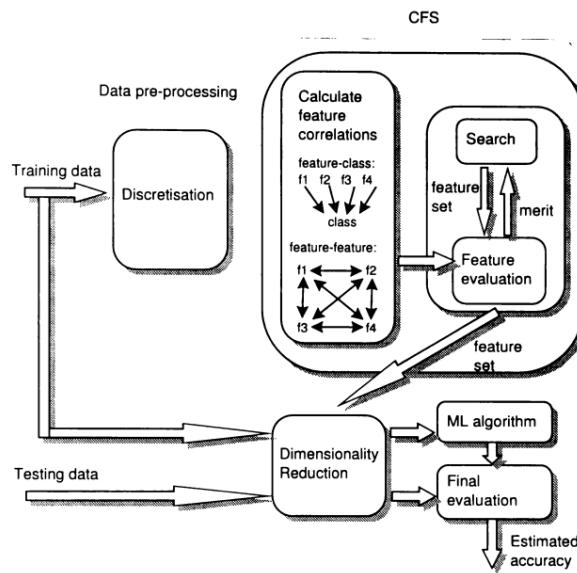


圖 2.11 基於相關係數的過濾型特徵選擇，結合機器學習流程 (Hall, 1999)。

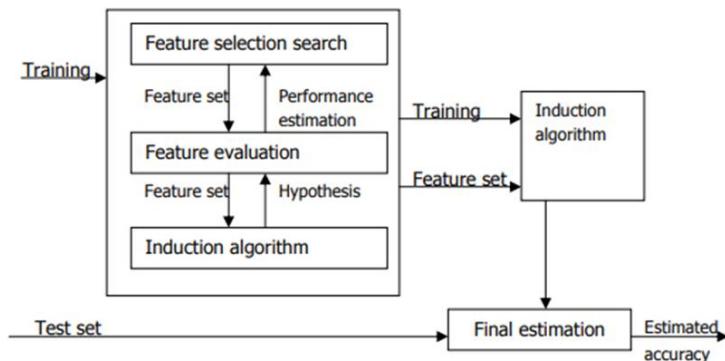


圖 2.12 基於相關係數的包裝型特徵選擇，結合機器學習流程 (Hall & Smith, 1999)。

而非監度式的相關係數特徵群組方式則有如塊模型、階層群集等方式；其方法皆在觀察特徵間的相關係數矩陣，並依此調整特徵順序，以產生特徵群組；且在同一群組內的特徵皆與彼此有高度相關性、但對於群組外的特徵卻不具有太多的相關。Liu et al. (2012)便在研究中使用了以相關係數矩陣為基礎的階層群集方法，來萃取影像特徵間的相關性規律，如圖 2.13，左方為原始相關係數矩陣、右方則為經由階層分析重新排序特徵後，的新相關係數矩陣。

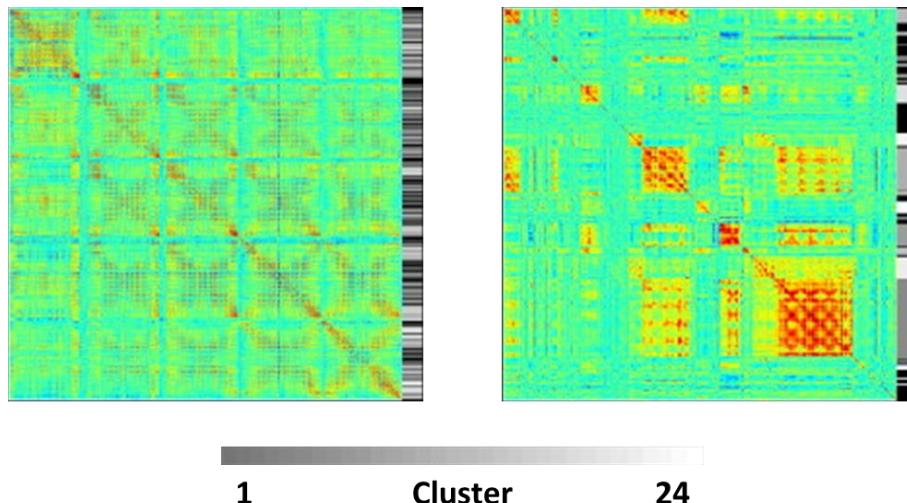


圖 2.13 以階層群集重新排序特徵之相關係數矩陣比較 (Liu et al., 2012)。

2.3.2 特徵萃取

比起一個具有大量特徵、龐大且冗餘的資料集，一個相對精簡且切中要點、融合了大多數特徵的資料集更能使分析者理解與詮釋這筆資料。特徵萃取 (Feature Extraction) 目的在於此，藉由變數之間的組合與處理同時維持資料本身的準確性，來降低整體資料的維度，以達到有效運用運算資源、避免過度擬合等目的；而相較於特徵選取，特徵萃取不論原始資料為何，皆能將原始資料解析、融合後轉換出能作為機器學習模型讀入的新特徵。

依據縮減維度的方法不同可分為以下兩類：

1. 線性降維：主成分分析 (PCA)、線性判別分析 (LDA)、MDS
2. 非線性降維：局部線性嵌入 (LLE)、T-SNE

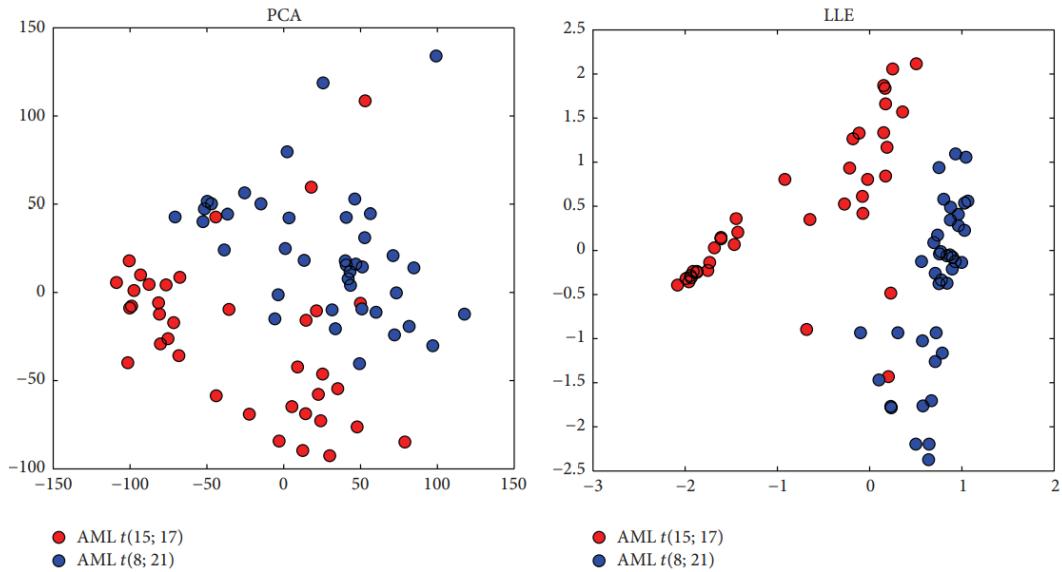


圖 2.14 以 PCA、LLE 視覺化 Leukaemia 資料集 (Hira & Gillies, 2015)。

主成分分析 (PCA) 最早由 Karl Pearson (1901) 發明，而後由 Harold Hotelling (1930) 重新定義與命名，其原為多變量統計中的一項分析手法，在機器學習領域之中則常作為縮減資料維度的工具使用。透過主成分分析可以在特徵空間中依次找到原資料中最大變異量的投影軸，對全體資料進行投影，最大限度的將資料與資料之間的差異顯示在新產生的座標軸之上，以便更輕鬆的對資料進行區分與評比。

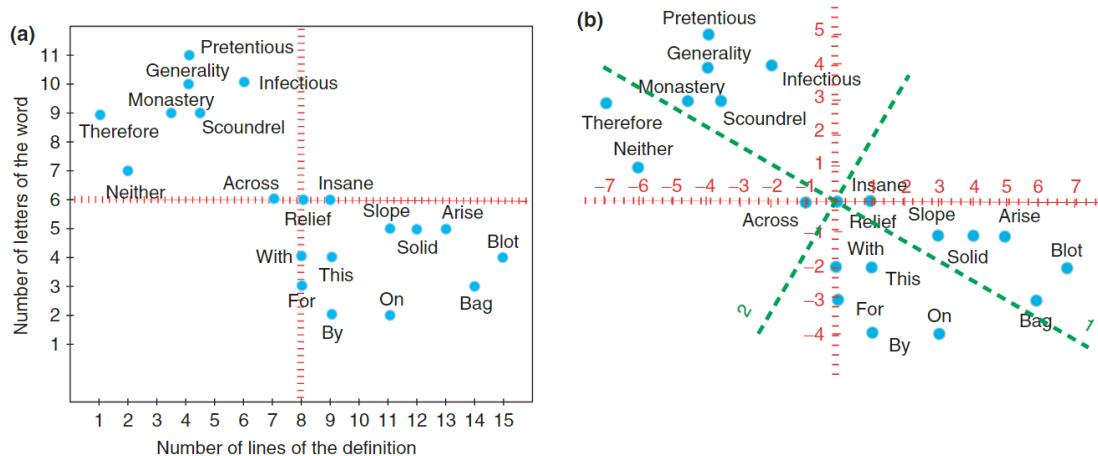


圖 2.15 (a)表示原始資料分布，(b)中綠線為 PCA 產生的兩主成分

(Abdi & Williams, 2010)。



圖 2.16 經由 PCA 將資料投影至主成分座標軸 (Abdi & Williams, 2010)。

使用主成分分析的主要目的有以下幾點：

1. 從原資料中萃取資訊
2. 簡化資料描述方式
3. 壓縮資料資訊

2.4 決策樹相關模型

決策樹做為有效、簡單的早期機器學習模型，早已被廣泛應用於分類、甚至是回歸的任務之中；而歷經近年對於決策樹相關方法的發展，使愈加彈性與萬能，並且得以跟上如今資料趨向大樣本、複雜化的趨勢；因決策樹能在短時間之內有效的分類、歸納複雜的訓練資料，適用於本研究發展方法所針對的資料特性，故在分類的任務之中將採用決策樹相關模型。以下將介紹近年來對於決策樹相關機器學習模型的發展過程。

2.4.1 決策樹

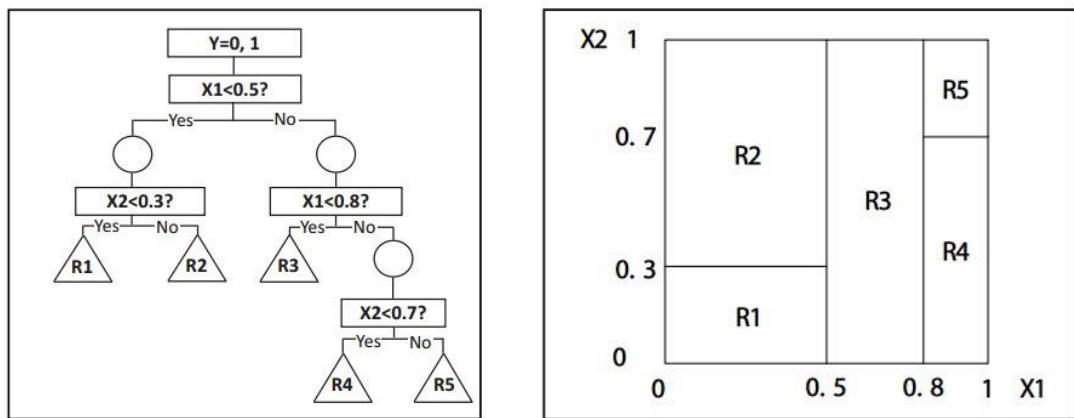


圖 2.17 決策樹範例，以二元分類問題為例 (Song & Ying, 2015)。

決策樹最早由 Morgan and Sonquist (1963)共同提出，透過建立各節點、與節點間的連線來產生樹狀結構，以做為區分資料的依據；在各個節點上對樣本特徵進行條件運算式判斷樣本走向，於最後一層來決定最終樣本歸屬類別。如上左圖，樣本具有 X_1 、 X_2 二特徵、且整體資料具有兩種類別 Y ；第一步先由頂點，也稱為根節點的位置出發，以 X_1 是否小於 0.5 作為評判，若 X_1 小於 0.5 則將樣本歸屬為左方節點；接著確認 X_2 是否小於 0.3；若 X_2 小於 0.3，則將樣本歸屬於 R_1 類別，此代表分類結果的最終節點則稱之為葉節點。根據該決策樹，將全部樣本歸類於 $R_1 \sim R_5$ 類別，而 $R_1 \sim R_5$ 則擇一代表著 $Y = 0$ 與 $Y = 1$ 某類特定類別；如上右圖，則將決策樹的決策方式改以樣本空間呈現。

決策樹通常依據最終切分後各節點中的資料純粹度來評判該決策樹的好壞，而常見於計算資料純粹度的方法有資訊增益（Information gain）、基尼係數（Gini index）；然而為了避免決策樹分支過於細碎，造成分類結果過於擬合分類資料，則以比例增益（Ratio gain）來規範整體決策樹的分支個數。

決策樹作為分類模型具有簡單、且高度可解釋性與便於視覺化的優點；同時得益於本身為條件判別式的特性，相較於其他種分類模型，決策樹所使用的運算資源與花費時間也相對微小許多。然而缺點在於當資料類別過於複雜、難以區分時將大幅增加決策樹的深度，導致過擬合的情形產生。

2.4.2 隨機森林

隨機森林由 Breiman (2001)提出，主要概念是以多棵較為低矮的決策樹作為分類模型。如上文所提及，當資料趨向多維度、複雜且難以簡單方式區分時，建構出的決策樹時常過度的複雜與深入，導致冗長的運算時間及過擬合的可能性。隨機森林則透過結合決策樹與引導聚集算法(Bootstrap aggregating, 又可簡稱為 Bagging)，嘗試解決結合此問題。Efron and Tibshirani (1994)所發展的引導聚集算法，乃透過將整體資料切分出若干個小型子資料集，並分別依據這些小型子資料訓練若干回歸／分類模型，而後根據這些訓練出的眾多模型以取平均值／多數票的方式，來進行資料的評判。其概念類似比起傳統由單一領導者進行決策，改為由各個不相同領域的專家討論、表決的方式進行，有更高機率產出較為適當的決策。

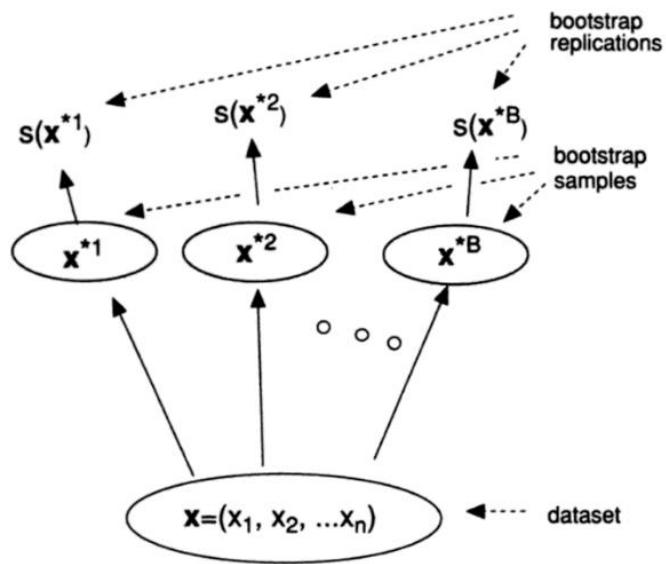


圖 2.18 引導聚集算法示意圖 (Efron & Tibshirani, 1994)。

隨機森林即是改用眾多較為簡單、低矮的決策樹做為分類模型。隨機森林內的決策樹之間相互獨立，且會使用到何種資料、何種特徵皆是隨機給定，如此的設計可以同時建構並平行訓練多棵決策樹，同時解決單一決策樹容易過度擬合的問題。

2.4.3 梯度提升決策樹

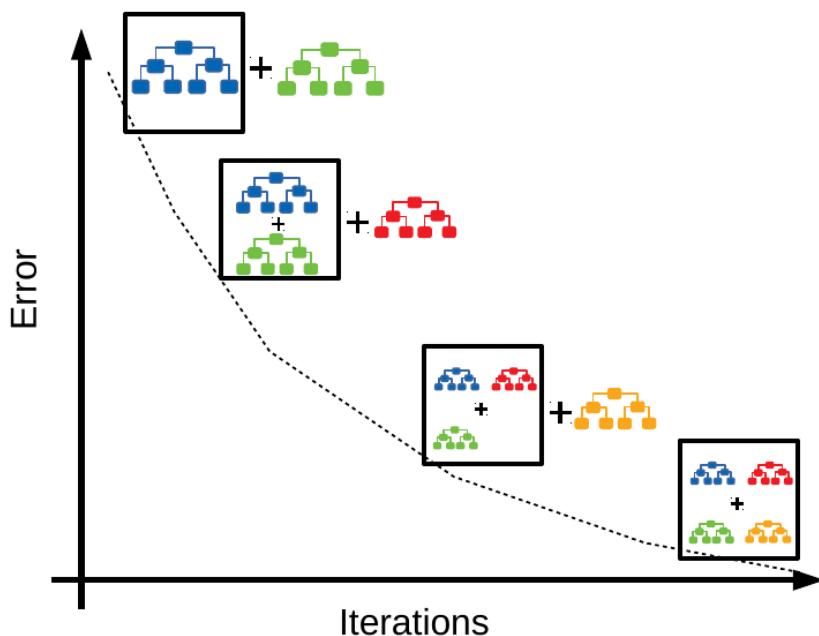


圖 2.19 梯度提升決策樹示意圖。

梯度提升決策樹 (GBDT, Gradient Boosting Decision Tree) 意指採用了以梯度

提升方法（Gradient Boosting）迭代訓練出新的決策樹，以彌補先前決策樹群的殘差，並以最終決策樹群做為目標模型的模型建構流程。

提升方法（Boosting）屬於一種集成學習演算法。主要概念為透過訓練眾多的弱學習器（weak learner）的組合來達到單一強學習器（strong learner）所具有的學習效果；其中弱學習器意指一個單純模型，其分類／回歸結果皆不夠強勢、而強學習器則意指在分類／回歸任務中有出色表現的模型。提升方法在建構的過程之中，使用迭代的方式添加弱學習器以逐步減少誤差，並持續訓練、優化機器學習模型直到迭代次數上限，如圖 2.20 所表示，相比於引導聚集算法切分資料後平行訓練弱學習器、提升方法則是迭代添加弱學習器以優化整體模型。常見的提升方法有如 Freund and Schapire (1996)提出的自適應增強（AdaBoost, Adaptive Boosting）。

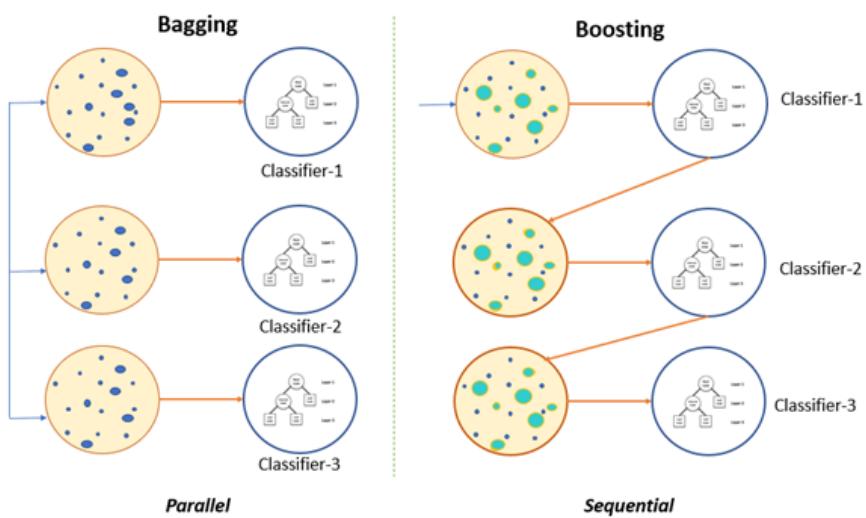


圖 2.20 引導聚集算法與提升方法的比較。

梯度提升方法則為提升方法的一種分支，由 Leo Breiman(1997)提出以觀測目標函數梯度優化模型的概念後，而後 Friedman (2001)便提出梯度提升機（GBM, Gradient Boosting Machine）作為梯度提升方法的實際應用。其方法應用了梯度計算於函數空間中，並根據負梯度作為整體模型的建構方向，在補足前一代次弱學習器的殘差同時，嘗試最佳化函數空間中的目標函數。代表性的梯度提升方式有如 Chen and Guestrin (2016)提出的極限梯度提升(XGBoost, eXtreme Gradient Boosting)、Dorogush et al. (2018) 提出的種類梯度提升(CatBoost, Categorical Boosting)、與 Ke

et al. (2017) 提出的輕量化梯度提升機 (LightGBM)。目前各式梯度提升機皆已被普遍使用以支援、訓練梯度提升決策樹模型，並應用於各式分類／回歸任務之中。

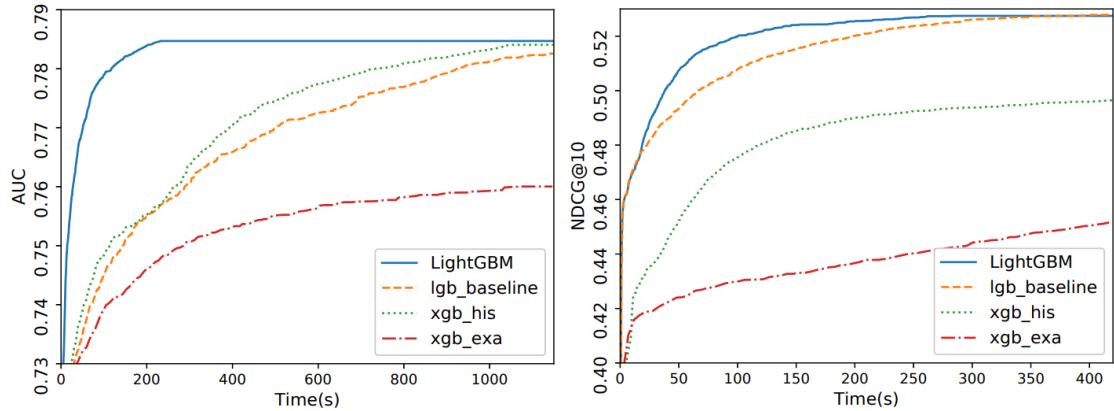


圖 2.21 LightGBM 於 Flight Delay (左) 與 LETOR (右) 兩資料集中的收斂表現。

相較於採用原始版本的演算方式，LightGBM 整體而言具有更多的廣泛優點，如快速的訓練速度與模型收斂、支援類別變數作為輸入、較低的記憶體用量與運算資源、較快速的訓練效率、產出更佳的分類準確度、支援平行運算，得以處理大規模數據等等；如圖 2.21，展示了相比於傳統梯度提升機，LightGBM 有著更快速的收斂速度，可以在相同的迭代次數之中找出更具優勢的解。

2.5 驗證指標

驗證指標（validation index）協助針對不同資料集、分類模型之間進行統一、有基準性的評估。針對處理問題的不同，可以將指標進行細分為針對回歸問題的回歸指標、與針對分類問題的分類指標。

2.5.1 回歸指標

在處理與數值預測相關的回歸問題時，針對預測值 (\hat{y}) 與目標值 (y) 之間的誤差 ($y - \hat{y}$) 便時常做為評判回歸模型與公式的一大方式。常見的回歸指標包括但不限於以下誤差計算方式：

1. 平均均方誤差 (MSE, Mean Squared Error)

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.22)$$

2. 平均絕對誤差 (MAE, Mean Absolute Error)

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.23)$$

3. 均方根誤差 (RMSE, Root Mean Square Error)

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2.24)$$

4. 平均絕對百分差 (MAPE, Mean Absolute Percentage Error)

$$MAPE(y, \hat{y}) = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (2.25)$$

2.5.2 分類指標

然而，當面對分類問題時缺乏數值運算，便無法再根據誤差進行評判。如此便需要以混淆矩陣(Confusion matrix)來總結整體的分類成果，歸納各個樣本的判斷、與其真實類別結果於矩陣當中，再依此矩陣計算衡量指標。如圖 2.22 所呈現，當 TP、FN (圖中綠色區域) 的數值越大，表示模型的判斷成功的將大多數樣本歸納成其的真實類別，便是個優秀的分類結果；反之，若落於 FP、FN (圖中紅色區域) 中的樣本愈多，則表示模型無法有效的預測樣本類別。

		True Condition		
		Positive	Negative	
Prediction	Positive	True Positive (TP)	False Positive (FP)	Positive Predictive Value (PPV), Precision $\frac{TP}{TP + FP}$
	Negative	False Negative (FN)	True Negative (TN)	False Omission Rate (FOR) $\frac{FN}{FN + TN}$
		True Positive Rate (TPR), Sensitivity, Recall $\frac{TP}{TP + FN}$	True Negative Rate (TNR), Fall-out $\frac{FP}{FP + TN}$	

圖 2.22 混淆矩陣於常見的衡量指標計算。

根據混淆矩陣中的各類數值，也得以計算出各類衡量指標：

1. 準確度 (Accuracy)

$$Accuracy = \frac{TP+TN}{T} \quad (2.26)$$

2. 召回率 (Recall)

$$Recall = \frac{TP}{TP+FN} \quad (2.27)$$

3. 精確率 (Precision)

$$Precision = \frac{TP}{TP+FP} \quad (2.28)$$

4. F1-score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.29)$$

5. G-measure

$$G = \sqrt{Precision \times Recall} \quad (2.30)$$

該使用以上的何種指標，應當根據問題的特性做調整。假設有個每次皆將樣本歸類於陰性的分類模型，嘗試分類一不平衡的資料集，而在這資料集中，具有九十九個陰性、與一個陽性樣本；分類後，此模型將能獲得百分之九十九的準確度，雖

然其根本不具有分類價值。因此在面對不平衡的資料時，為了關注那正確分類的稀少陽性樣本，則會依據欲確保的重點部分，考慮使用召回率與精確率做為衡量標準；而 F1 score 與 G measure 則綜合了兩者，使其可以被同時評估。

第三章 高維二元特徵之聚合編碼技術及分析框架

經由第二章文獻探討得以發現，以多維度二元特徵資料作為機器學習模型的輸入時所遭遇到的難題。傳統的變數編碼方式僅針對、並轉換類別變數為數值型別；為此，此研究規劃了對於二元特徵資料的編碼方法，得以透過群組、排序與二進位十進數表示的方式，轉換二元特徵資料成數值資料。

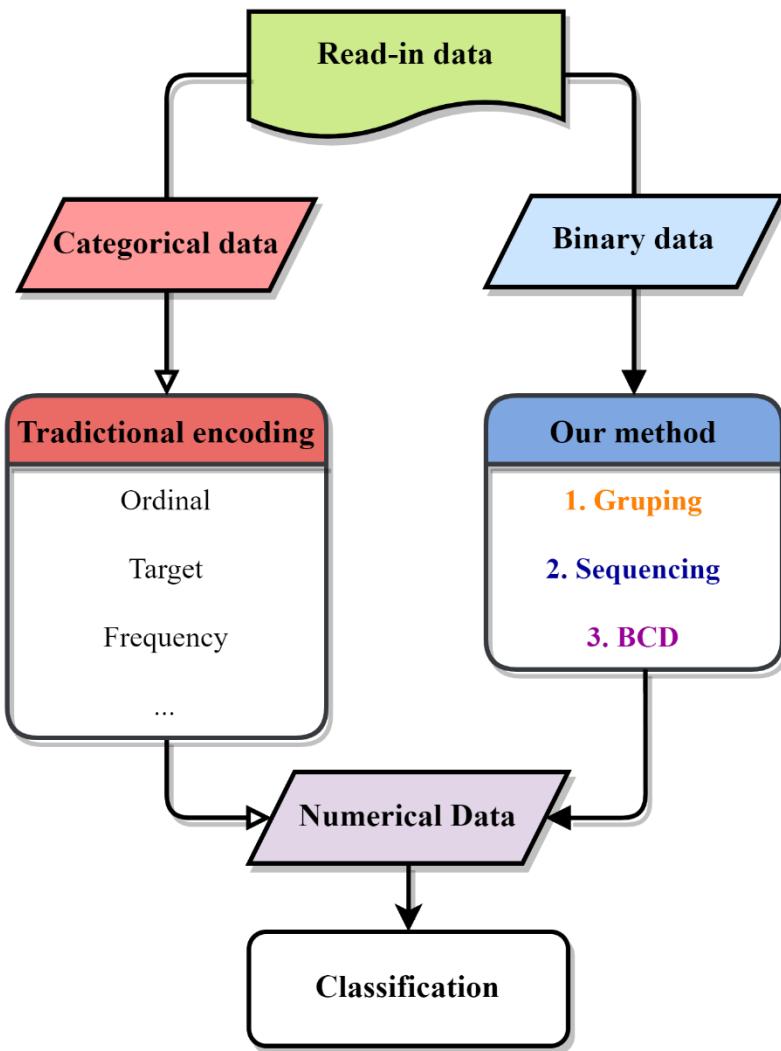


圖 3.1 資料處理、評估流程圖。

圖 3.1 展示了本研究的評估流程。實驗資料將分別透過傳統的變數編碼方式、與本研究的研究方法進行編碼，產生編碼過後的新數值資料，並由 LightGBM 建置分類模型，以評估各數值資料的分類成果。而於此的實驗資料採用可於類別與二元兩種形式間轉換的資料，以確保兩種資料只在形式的表現上有所不同，但在本質上

具有相同的意義，如原始的類別資料經由獨熱編碼，轉換出二元資料，像是表 2.4；或是第四章第一小節中的連續資料切分出的二元資料，如表 4.1、以及由此二元資料所類別化後的類別資料，如表 4.2。

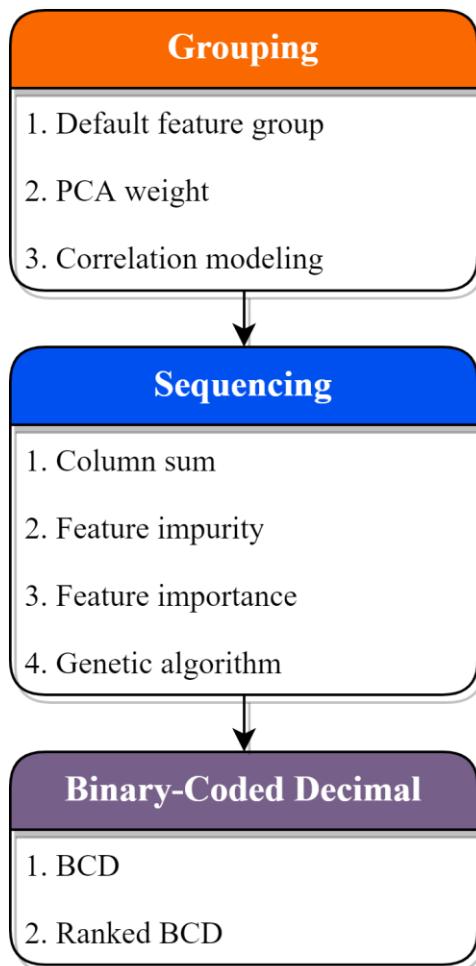


圖 3.2 研究方法流程圖。

本研究所發展針對二元特徵的編碼方式流程如上圖，透過群組、排序與 BCD (Binary-Coded Decimal) 此三個步驟將原先的二元資料轉換為數值資料，再與傳統編碼產生的數值資料做分類成果比較，如圖 3.1。而這框架之下的三個步驟中也能透過各種不同的方式來達成，像是群組二元特徵時能依據 PCA、相關係數或是原始群組資訊；而特徵純粹度、特徵重要性或是特徵和皆能作為組內的排序法則等。

表 3.1 變數與符號定義。

符號	定義
X_i	原始的二元資料，共含有 n 個二元特徵， $0 \leq i \leq n$
$g(X)$	對二元特徵進行分群，產生 m 個 G_j
m	二元特徵群組個數
G_j	第 j 個二元特徵群組， $0 \leq j \leq m$
$s(G_j)$	群組內二元特徵的排序方式，即調整 G_j 內部二元特徵的排列
S_j	第 j 個二元特徵群組中的特徵排列方式， $0 \leq j \leq m$
$BCD(G_j)$	第 j 個二元特徵群組的 BCD 數值
Y_j	編碼後的數值資料，共含有 m 個數值特徵， $0 \leq j \leq m$
Z_k	類別化後的類別資料，共含有 l 個類別特徵， $0 \leq k \leq l$

在表 3.1 中，定義了本研究內的各個變數符號於其代表的意涵；而整體的流程如圖 3.2 所示，先是透過群組相關的二元特徵、而後進行特徵組內的特徵排序、再透過二進位十位數編碼的方式將排序後的特徵組轉換為數值資料。本研究著重於找尋合理、適當的二元特徵的群組方式 $g(X_n)$ 、以及群組內二元特徵的排序方式 $s(G_j)$ ，來使新產生出的數值資料有利於機器學習模型的分類任務。 Z 則表示當原始資料 X 為符合獨熱編碼法則的二元資料時，得以轉換回的編碼前類別資料，如表 2.4 中以「獨熱編碼」欄位，轉換回「居住城市」欄位。

3.1 二元特徵分群

在分群 (Grouping) 階段針對具備有相同物理意義、具有相關性、或是重要度相近的二元特徵進行群組，以便將數量較多、資訊較弱的相互關連的二元特徵整合進入同一群組內，以便後續將群組內的特徵。在本研究之中，提出了根據特徵選取與特徵萃取的方式、或是依據相關性分析，將二元特徵區分至各個群組之中。

為了便於理解，假設某一動物園針對園區內所有三百隻動物進行健康狀態普查，針對各個動物的物種、尺寸與顏色進行檢測，並將其身體健康狀態以藍色與紅色兩色作為區分成「好」或「壞」，最後收錄健康普查結果，如圖 3.3 所示。可見到在「尺寸：中等」的欄位中，園中有接近一百隻動物的體型屬於此尺寸，且其中「健康狀況：好」、與「健康狀況：壞」的動物數量各占一半。

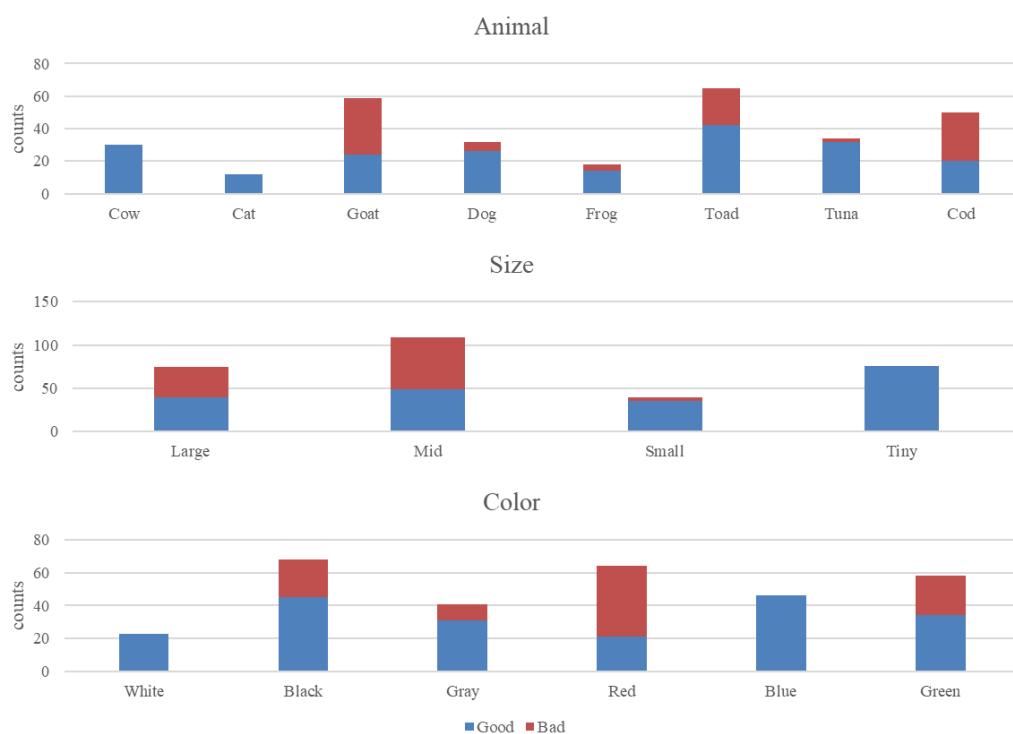


圖 3.3 原始二元特徵資料，以動物園資料為例。

3.1.1 資料原始特徵群

若是原始的二元特徵資料是由獨熱編碼所產生，且具有特徵之間的分群的資訊，屬於具有二元特徵群組資訊的資料時，則可以使用該群組資訊將二元特徵進行

群組。如下表，表示了經由原始二元特徵的群組知識(特徵區分成「動物」、「尺寸」、「顏色」三個群組)，群組後的二元特徵。

表 3.2 具有群組資訊的二元特徵資料。

Animal								Size				Color					
Tuna	Cat	Frog	Cod	Goat	Dog	Toad	Cow	Large	Mid	Small	Tiny	White	Black	Red	Blue	Green	Gray
0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0

此一群組方式確保了群組內二元特徵的高度相關性，也作為其餘群組方式嘗試所分群的最佳結果；然而，當面對的二元資料並不具有群組資訊時，如表 3.3，此種方式便不再適用，僅能以別種方式將二元特徵進行分群。而如何將缺乏群組資訊的二元特徵資料進行編碼，轉換成數值資料，也同為本研究重點。

表 3.3 缺乏群組資訊的二元特徵資料。

Tuna	Cat	Frog	Cod	Goat	Dog	Toad	Cow	Large	Mid	Small	Tiny	White	Black	Red	Blue	Green	Gray
0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0

3.1.2 主成分分析群集

透過主成分分析 (Principle Component Analysis)，可以將最能解釋整體資料變異、最具資訊的二元特徵分於相同群組之內，如此使新產生的數值特徵融合了最多具有變異資訊的二元特徵。透過決定群組個數，再根據主成分的順序，依據權重的絕對值依序進行選取，將解釋資料變異度相近的二元特徵聚集於同一群組之中。

表 3.4 不同主成分之下的二元特徵權重絕對值。

No.	PC 1			PC 2			PC 3		
	Feature	Abs. weight	select	Feature	Abs. weight	select	Feature	Abs. weight	select
1.	Cow	0.562	O	Cat	0.642	X	Black	0.047	X
2.	Cat	0.486	O	Cow	0.496	X	Goat	0.035	X
3.	Large	0.348	O	Mid	0.402	X	Cat	0.032	X
4.	Mid	0.311	O	Black	0.351	X	Cow	0.030	X
5.	White	0.307	O	Goat	0.302	O	Mid	0.028	X
6.	Black	0.202	O	Dog	0.229	O	Gray	0.023	X
7.	Small	0.187	X	Large	0.199	X	White	0.021	X
8.	Goat	0.153	X	Frog	0.183	O	Dog	0.020	X
9.	Cod	0.101	X	Small	0.152	O	Large	0.018	X
10.	Tuna	0.091	X	Gray	0.132	O	Frog	0.016	X
11.	Dog	0.074	X	White	0.105	X	Small	0.014	X
12.	Gray	0.060	X	Red	0.008	O	Toad	0.014	O
13.	Frog	0.056	X	Frog	0.008	X	Tuna	0.012	O
14.	Toad	0.032	X	Red	0.007	X	Cod	0.007	O
15.	Red	0.029	X	Toad	0.005	X	Red	0.005	X
16.	Tiny	0.023	X	Green	0.001	X	Tiny	0.002	O
17.	Green	0.015	X	Tiny	0.001	X	Blue	0.001	O
18.	Blue	0.009	X	Blue	0.001	X	Green	0.001	O

選取流程如表 3.4 所示，首先由最能解釋全體變異的第一主成分根據各特徵的權重絕對值做選取；而後依序交由之後的主成分選擇，且同時須避免選取到已被先前主成分所選取的二元特徵。群組後的二元特徵可見表 3.5。

表 3.5 依據主成分分析群集二元特徵。

PC 1						PC 2						PC 3					
Cow	Cat	Large	Mid	White	Black	Goat	Dog	Frog	Small	Gray	Red	Toad	Tuna	Cod	Tiny	Blue	Green
1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1

3.1.3 相關係數群集

在群組二元特徵時，也可以透過審視二元特徵間的相關性，以階層群集 (Hierarchical clustering)、或是集區建模 (Block modeling) 的方式群組相互高度相關的特徵；其主要的目的便在於，使群組過後的特徵在同一特徵群組內彼此有高度的相關性，同時特徵組與特徵組之間卻不過度相關。

	Tuna	Cat	Frog	Cod	Goat	Dog	Toad	Cow	Large	Mid	Small	Tiny	White	Black	Red	Blue	Green	Gray
Tuna	1.0	0.2	0.2	0.8	0.2	0.2	0.8	0.2	0.1	0.1	0.1	0.3	0.0	0.0	0.0	0.5	0.5	0.0
Cat	0.2	1.0	0.2	0.2	0.2	0.2	0.2	0.8	0.3	0.3	0.1	0.1	0.2	0.2	0.0	0.0	0.0	0.0
Frog	0.2	0.2	1.0	0.2	0.8	0.8	0.2	0.2	0.1	0.1	0.3	0.1	0.0	0.0	0.3	0.0	0.0	0.3
Cod	0.8	0.2	0.2	1.0	0.2	0.2	0.8	0.8	0.1	0.1	0.1	0.3	0.0	0.0	0.0	0.3	0.3	0.0
Goat	0.2	0.2	0.8	0.2	1.0	0.8	0.2	0.2	0.1	0.1	0.3	0.1	0.0	0.0	0.3	0.0	0.0	0.3
Dog	0.2	0.2	0.8	0.2	0.8	1.0	0.2	0.2	0.1	0.1	0.3	0.1	0.0	0.0	0.3	0.0	0.0	0.3
Toad	0.8	0.2	0.2	0.8	0.2	0.2	1.0	0.2	0.1	0.1	0.1	0.3	0.0	0.0	0.0	0.3	0.0	0.0
Cow	0.2	0.8	0.2	0.8	0.2	0.2	0.2	1.0	0.3	0.3	0.1	0.1	0.3	0.3	0.0	0.0	0.0	0.0
Large	0.1	0.3	0.1	0.1	0.1	0.1	0.1	0.3	1.0	0.5	0.1	0.1	0.6	0.6	0.2	0.2	0.2	0.2
Mid	0.1	0.3	0.1	0.1	0.1	0.1	0.1	0.3	0.5	1.0	0.1	0.1	0.6	0.6	0.2	0.2	0.2	0.2
Small	0.1	0.1	0.3	0.1	0.3	0.3	0.1	0.1	0.1	0.1	1.0	0.1	0.2	0.2	0.6	0.2	0.2	0.6
Tiny	0.3	0.1	0.1	0.3	0.1	0.1	0.3	0.1	0.1	0.1	0.1	1.0	0.2	0.2	0.2	0.6	0.2	0.2
White	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.3	0.6	0.6	0.2	0.2	1.0	0.8	0.4	0.4	0.4	0.4
Black	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.3	0.6	0.6	0.2	0.2	0.8	1.0	0.4	0.4	0.4	0.8
Red	0.0	0.0	0.3	0.0	0.3	0.3	0.0	0.0	0.2	0.2	0.6	0.2	0.4	0.4	1.0	0.4	0.4	0.4
Blue	0.5	0.0	0.0	0.3	0.0	0.0	0.3	0.0	0.2	0.2	0.2	0.6	0.4	0.4	0.4	1.0	0.8	0.8
Green	0.5	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.2	0.2	0.2	0.2	0.4	0.4	0.4	0.8	1.0	0.4
Gray	0.0	0.0	0.3	0.0	0.3	0.3	0.0	0.0	0.2	0.2	0.6	0.2	0.4	0.8	0.4	0.8	0.4	1.0

圖 3.4 原始資料二元特徵間的相關性矩陣。

	Cow	Cat	Large	Mid	White	Black	Goat	Dog	Frog	Small	Gray	Red	Toad	Tuna	Cod	Tiny	Blue	Green
Cow	1.0	0.8	0.3	0.3	0.5	0.5	0.2	0.2	0.2	0.1	0.0	0.0	0.2	0.2	0.2	0.1	0.0	0.0
Cat	0.8	1.0	0.3	0.3	0.5	0.5	0.2	0.2	0.2	0.1	0.0	0.0	0.2	0.2	0.2	0.1	0.0	0.0
Large	0.3	0.3	1.0	0.5	0.6	0.6	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.2	0.2
Mid	0.3	0.3	0.5	1.0	0.6	0.6	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.2	0.2
White	0.5	0.5	0.6	0.6	1.0	0.8	0.0	0.0	0.0	0.2	0.4	0.4	0.0	0.0	0.0	0.2	0.4	0.4
Black	0.5	0.5	0.6	0.6	0.8	1.0	0.0	0.0	0.0	0.2	0.4	0.4	0.0	0.0	0.0	0.2	0.4	0.4
Goat	0.2	0.2	0.1	0.1	0.0	0.0	1.0	0.8	0.8	0.3	0.3	0.3	0.2	0.2	0.2	0.1	0.0	0.0
Dog	0.2	0.2	0.1	0.1	0.0	0.0	0.8	1.0	0.8	0.3	0.5	0.5	0.2	0.2	0.2	0.1	0.0	0.0
Frog	0.2	0.2	0.1	0.1	0.0	0.0	0.8	0.8	1.0	0.3	0.5	0.5	0.2	0.2	0.2	0.1	0.0	0.0
Small	0.1	0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.3	1.0	0.6	0.6	0.1	0.1	0.1	0.1	0.2	0.2
Gray	0.0	0.0	0.2	0.2	0.4	0.4	0.3	0.5	0.5	0.6	1.0	0.8	0.0	0.0	0.0	0.2	0.4	0.4
Red	0.0	0.0	0.2	0.2	0.4	0.4	0.3	0.5	0.5	0.6	0.8	1.0	0.0	0.0	0.0	0.2	0.4	0.4
Toad	0.2	0.2	0.1	0.1	0.0	0.0	0.2	0.2	0.2	0.1	0.0	0.0	1.0	0.8	0.8	0.3	0.5	0.5
Tuna	0.2	0.2	0.1	0.1	0.0	0.0	0.2	0.2	0.2	0.1	0.0	0.0	0.8	1.0	0.8	0.3	0.5	0.5
Cod	0.2	0.2	0.1	0.1	0.0	0.0	0.2	0.2	0.2	0.1	0.0	0.0	0.8	0.8	1.0	0.3	0.5	0.5
Tiny	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.3	1.0	0.6	0.6
Blue	0.0	0.0	0.2	0.2	0.4	0.4	0.0	0.0	0.0	0.2	0.4	0.4	0.5	0.5	0.5	0.6	1.0	0.8
Green	0.0	0.0	0.2	0.2	0.4	0.4	0.0	0.0	0.0	0.2	0.4	0.4	0.5	0.5	0.5	0.6	0.8	1.0

圖 3.5 以塊模型進行置換後的二元特徵間的相關性矩陣。

排序前的原始資料圖 3.4 所示，而經由塊模型、或是階層群集等相關係數群集手法、調換特徵的順序之後，可以得到如圖 3.5 的新相關係數矩陣。可以由圖中看

出相較於紅線外的二元特徵，在紅線內部的二元特徵彼此更為相關，因此將紅線內部的二元特徵歸類於同一群組之中，群集結果如下表所示。

表 3.6 依據相關係數群集二元特徵。

Corr 1						Corr 2						Corr 3					
Cow	Cat	Large	Mid	White	Black	Goat	Dog	Frog	Small	Gray	Red	Toad	Tuna	Cod	Tiny	Blue	Green
1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1

3.2 群內二元特徵排序

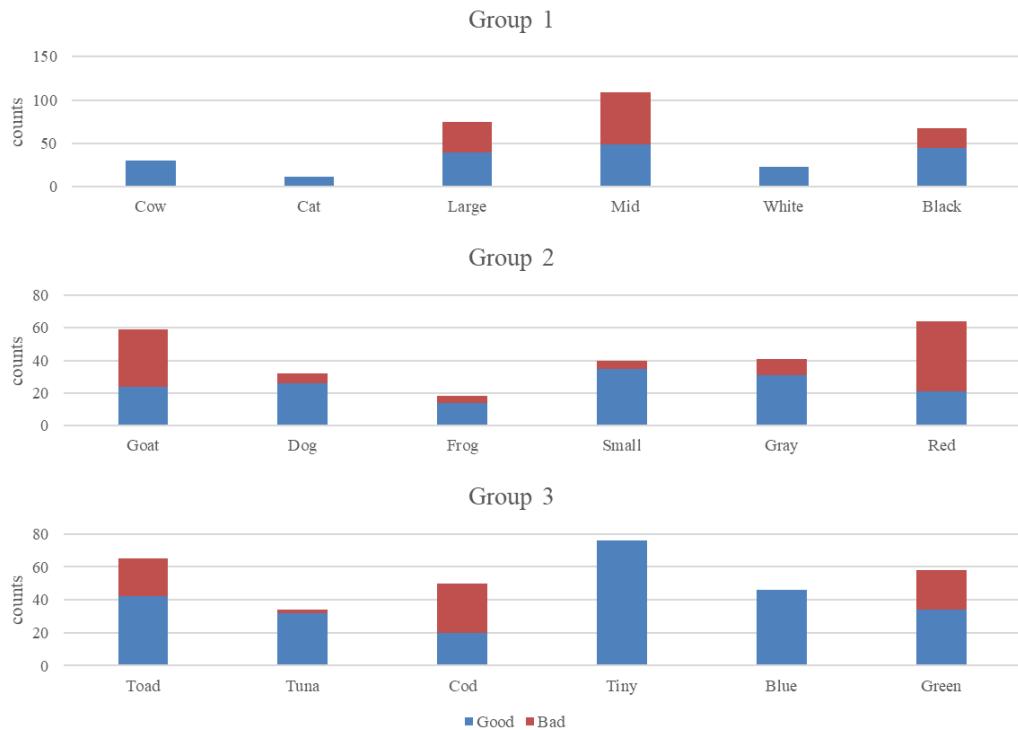


圖 3.6 群集過後的特徵組。

在群組了原始資料的二元特徵後，為了使產出後的新數值特徵更具有目標資訊、與分類價值，將依據各群組內各個二元特徵本身屬性，再對各組內二元特徵做排序（Sequencing），來調動編碼過後的數值。這些屬性包括但不限於二元特徵總值、特徵純粹度、預訓練模型的特徵重要性、或甚至以隨機指派的方式作為排序依據，而後對比不同的排序方式對於分類結果的影響。除了以特徵屬性作排列依據之外，也可以將此描述為一最佳化問題，嘗試以不同最佳化方法進行求解，例如基因演算、捷思法等。

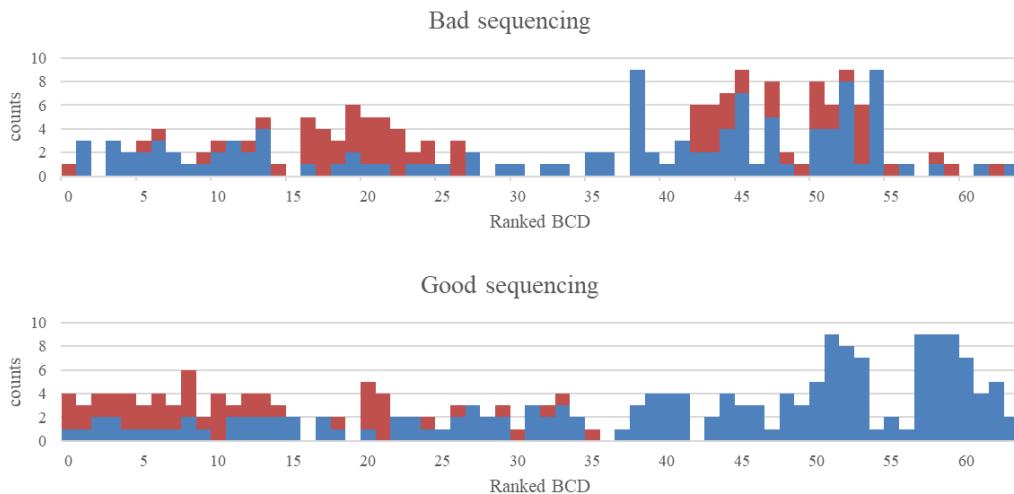


圖 3.7 以不同方式排序二元特徵，產生的新數值資料分佈比較，依據新數值特徵分佈。

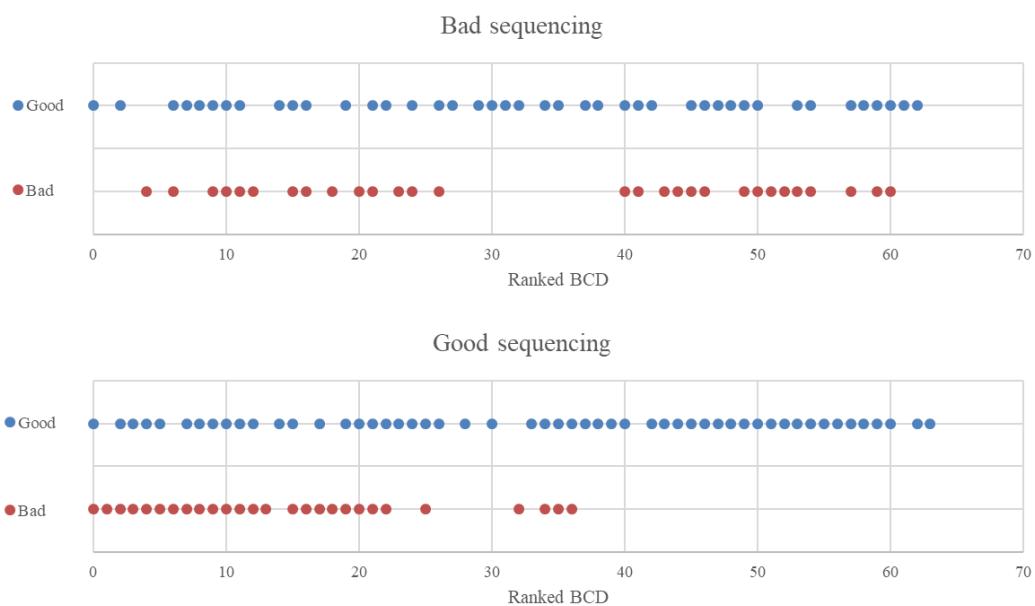


圖 3.8 以不同方式排序二元特徵，產生的新數值資料分佈比較，依據類別區分。

圖 3.7、圖 3.8 描述了在不同的排序方式之下，由二元特徵組編碼後，所產生的新數值特徵的資料分布。對於常規的分類模型而言，依據優勢的排序方式將使產生的數值特徵可以在數線上找到一個更佳的數值切分點來區分出紅色與藍色兩種資料，因此能達到較好的分類成果。如果能設法找尋出好的特徵排序方式，對於新數值資料的分類任務將有莫大的幫助。

3.2.1 二元特徵總和排序

依據特徵的各樣本數值總和作為排序；以表 3.2 的資料為例，若是原始資料符合獨熱編碼原則，且帶有二元特徵的群組資訊，則數值一出現頻率表示了該類別的出現頻率，同時代表了該類別具有相當多的資料個數。該排序目的在於將總和較少的二元特徵置於群組前方，如此那些含有此稀有二元特徵的少量樣本便會在編碼時被投影到離數線原點較遠的位置，以利區分出這些與眾不同、較偏離整體資料分布的離群樣本。

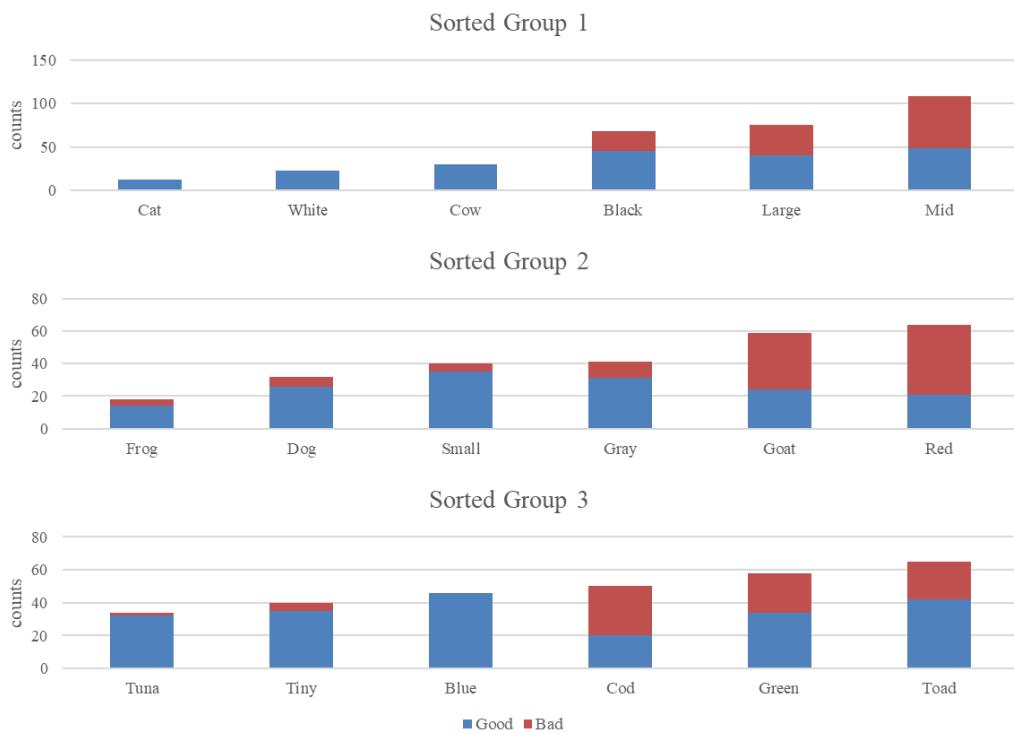


圖 3.9 以二元特徵總和，排序各個群組中的二元特徵。

3.2.2 特徵純粹度排序

計算特徵的特徵純粹度，以進行特徵排序；以二元分類為例，將最有分類價值、純度越高的資料分布於特徵群組的前端、與後端，而中間段的特徵則是最不純粹的，如圖所示；因此當群組中的二元特徵經由 BCD 轉換為數值資料時，兩種類別會因為這些被置於前方、含有資訊與分類價值的二元特徵的緣故，編碼出兩種數值差異大的兩種數值。因為在排序的過程之中參考了目標欄位，因此為監督式的排序方式。

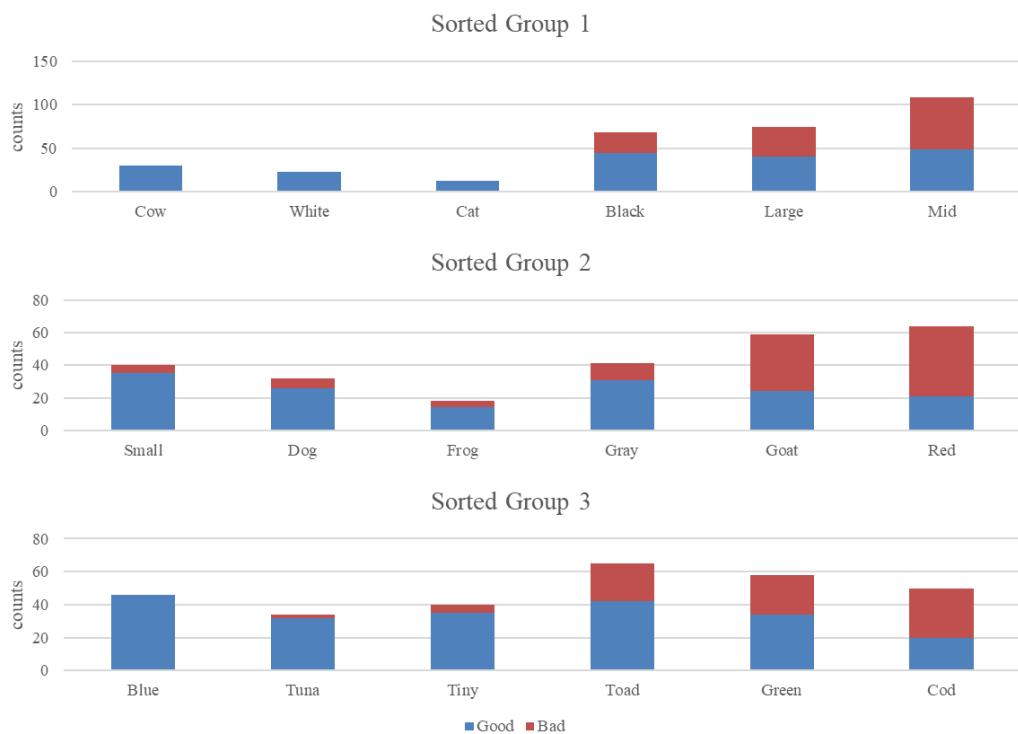


圖 3.10 以目標特徵純粹度，排序各個群組中的二元特徵。

3.2.3 特徵重要度排序

根據預先訓練的分類模型的特徵重要性來進行組內的特徵排序，以期透過機器學習模型分辨出最具有分類資訊的特徵，並調整至於群組的最前方，以求新編碼後的數值特徵能因此具備更優秀的分辨能力。圖 3.11 表示各個二元特徵於預訓練模型中的特徵重要度，可依此作為排序依據；圖 3.12 為根據特徵重要度排序各組二元特徵的成果。

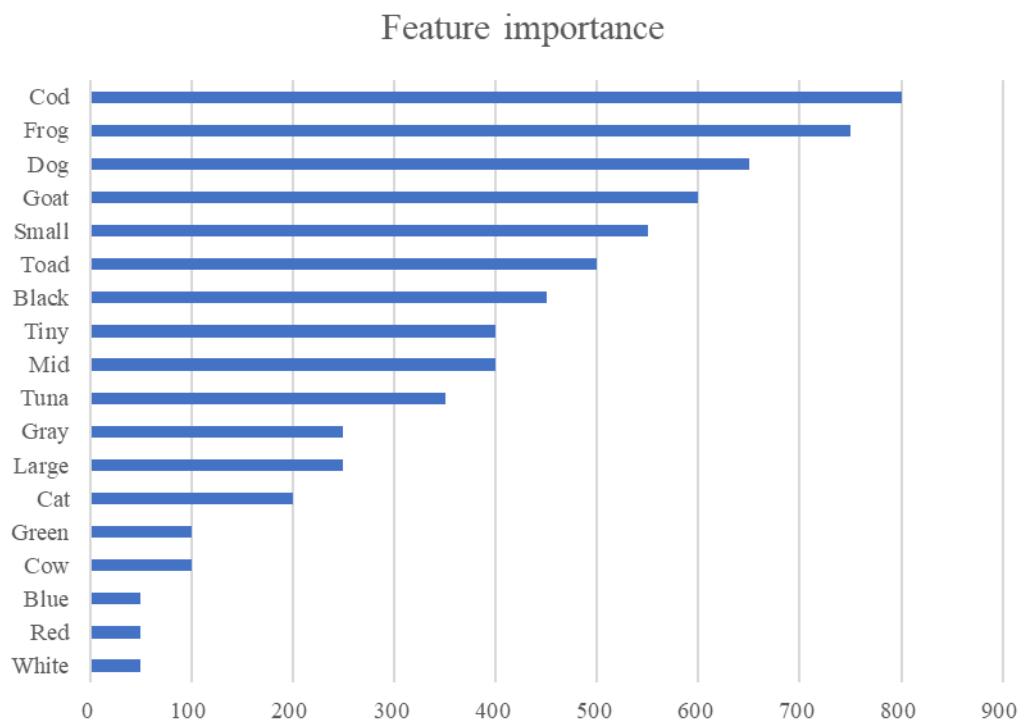


圖 3.11 各項二元特徵於預訓練分類模型中的特徵重要性。

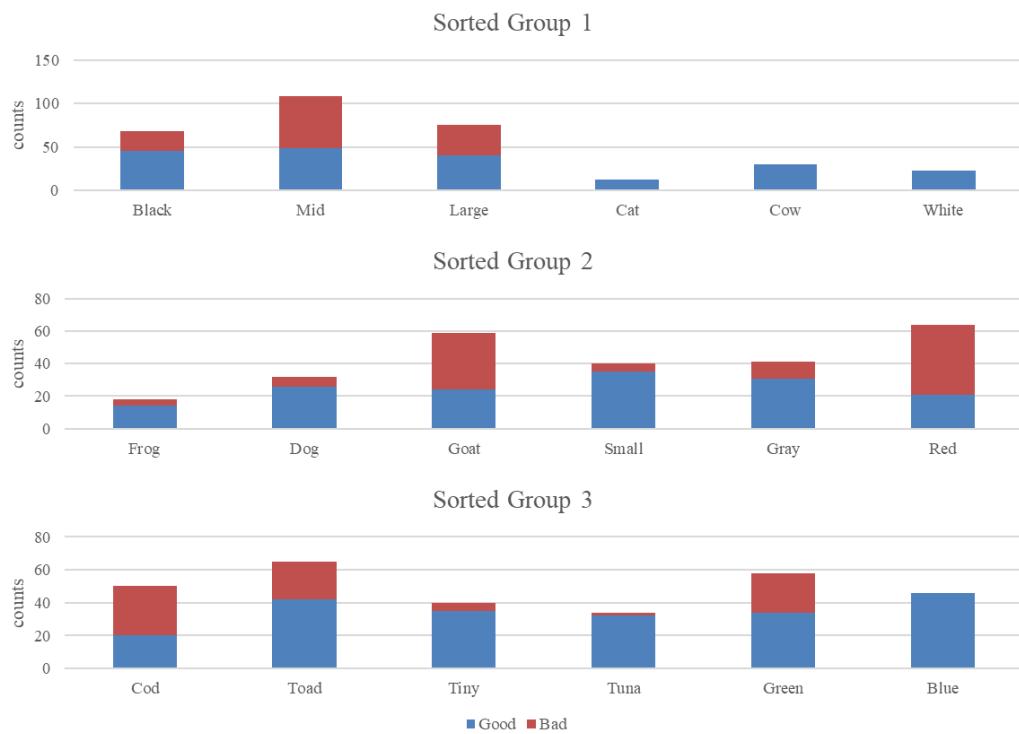


圖 3.12 以特徵重要度，排序各個群組中的二元特徵。

3.2.4 基因演算排序法

本研究也嘗試以基因演算法排序特徵；只要將問題設定成多組的銷售員路徑

排序問題（permutation GA）、以及計算適應度（fitness）後，即可以 GA 作為排序方法。在表 3.7 中， C_{ij} 表示了群組後、排序前的二元特徵，其中 i 表示特徵所歸屬的群組、 j 表示特徵於組內的排序。

表 3.7 排序前各群組中的二元特徵，由 C_{ij} 表示。

Group 1						Group 2						Group 3					
C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{21}	C_{22}	C_{23}	C_{24}	C_{25}	C_{26}	C_{31}	C_{32}	C_{33}	C_{34}	C_{35}	C_{36}

如表 3.8 所示，透過基因演算，能循序漸進、方向性地搜尋樣本空間中，各式特徵排序的可能性，經由多次迭代後取得最佳適應值的排序組合，並得出在有限的世代中找尋出最高適應值的染色體，由此作為該群組特徵的排序方式。

表 3.8 染色體範例，以基因演算法排序組內特徵。

	Group 1						Group 2						Group 3					
Chromosome 1	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{21}	C_{22}	C_{23}	C_{24}	C_{25}	C_{26}	C_{31}	C_{32}	C_{33}	C_{34}	C_{35}	C_{36}
Chromosome 2	C_{16}	C_{15}	C_{14}	C_{13}	C_{12}	C_{11}	C_{26}	C_{25}	C_{24}	C_{23}	C_{22}	C_{21}	C_{36}	C_{35}	C_{34}	C_{33}	C_{32}	C_{31}
Chromosome 3	C_{11}	C_{13}	C_{15}	C_{12}	C_{14}	C_{16}	C_{22}	C_{24}	C_{26}	C_{21}	C_{23}	C_{25}	C_{31}	C_{33}	C_{35}	C_{32}	C_{34}	C_{36}
:	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

本研究考量了編碼後的新數值資料中，兩種或單種類別的統計資訊，作為染色體適應值的計算方式。像是以編碼過後的組間變異除以組內變異或是最小化某一特定數類別的變異，參見以下兩方程式。皆可做為適應值的計算方式，供基因演算法在不同世代間篩選染色體。

$$\text{Max}\left(\frac{SS_B}{SS_E}\right) \quad (3.1)$$

$$\text{Max}(1/SS_E) \quad (3.2)$$

相較於其餘排序方式，基因演算將花費更多運算資源與時間成本，且相當受制於適應值的計算，倘若使用了對於提升新資料分類效用不甚明顯的適應值，則將導致基因演算法往錯誤的樣本空間方向進行探索與演化，使得最終所費成本收效甚微。

3.3 群組二進碼十進數編碼

在最後的這一階段，將各個群組內的複數二元特徵經由 Binary-Code Decimal (BCD) 編碼的方式，轉換成整數數值。其中本研究所採用 8421-BCD，其概念相當直接，即是將一連串的二元數值十位數化。如下表所示，假設某一樣本在一特徵群組內的二元特徵數值依序為 0、1、1、0，則其轉換後的十位數數值為 6，將作為該樣本於該群組轉換出新特徵的數值。

表 3.9 常用的 BCD 編碼方式，與對應的十位數值。

Decimal digit	BCD Code											
	8421				4221				5421			
0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	0	1	0	0	0	1
2	0	0	1	0	0	0	1	0	0	0	1	0
3	0	0	1	1	0	0	1	1	0	0	1	1
4	0	1	0	0	1	0	0	0	0	1	0	0
5	0	1	0	1	0	0	1	1	1	0	0	1
6	0	1	1	0	1	1	0	0	1	0	1	0
7	0	1	1	1	1	1	0	1	1	0	1	1
8	1	0	0	0	1	1	1	0	1	1	0	0

3.3.1 二進位十位數編碼數值



圖 3.13 經過特徵純粹度排序的第三特徵組。

將各群組內的二元特徵數值作為二進位數值，並轉由十位數值進行表示，如此便產生了新的數值新特徵。根據本研究的實驗結果，二元特徵的群組、排序將對該數值產生相當的影響，且對於新資料的分類結果也將有顯著的影響。如表 3.10 表

示了經由特徵純粹度排序、再 BCD 編碼後的新數值資料；可以由圖 3.13 見到，「藍色」這個二元特徵欄位被放置於群組的最前方、而且屬於「藍色」特徵的樣本的健康狀態皆為「好」的狀態，因此有著「藍色」特徵的樣本在經由 BCD 編碼轉換時將會對應到較大的數值，因而跟屬於健康狀態屬於「壞」的樣本拉開距離，如下表所表示。

表 3.10 各個樣本轉換後的新數值。

	Sequenced Column Group						BCD	Type
	Blue	Tuna	Tiny	Toad	Green	Cod		
Sample 1	1	0	0	0	0	1	33	Good
Sample 2	1	0	0	1	0	0	36	Good
Sample 3	0	0	1	0	1	0	10	Bad
Sample 4	0	0	1	1	1	0	14	Bad
:	:	:	:	:	:	:	:	:

此處也體現出了排序群組內特徵的重要性，有意義的排序將使得產生的數值資料更加的有利於機器學習模型進行分類任務。

3.3.2 二進位十位數編碼數值排名

針對各組排序過後的二元特徵組，進行二進碼十進數編碼，產生新的整數型別的類別變數，新資料的特徵個數將等於原先的二元特徵群組數，然而，新編碼過後的資料因為原先數字 1 的分布稀疏，也將導致編碼後的數值資料全距過大、且分布稀疏。例如：若組內二元特徵的個數為十個，則此群組編碼出的新數值特徵範圍將達到 0 至 1024 之間。為此，可透過排名編碼為了改善編碼後資料之間的稀疏程度，同時避免儲存過大正整數，導致整數溢位等問題，如圖 3.14 所示，使用 Rank BCD 可將原先全距大的 BCD 數值，以較小的樣本全距描述。

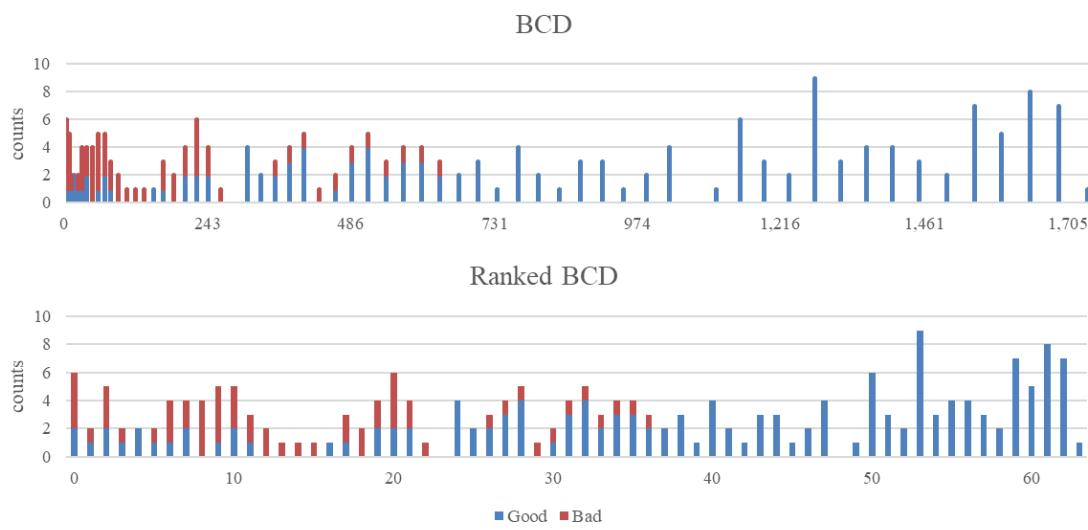


圖 3.14 BCD 與 Rank BCD 對於新編碼後的特徵分布比較。

3.4 二元特徵降維技術

本研究發展的二元變數編碼方式也可以做為二元資料的降維方法來使用。只需選定群組與排序方式，便能將原先總共為 n 維度的二元資料，轉變成僅具有 m 維度的數值資料， $0 < m \leq n$ 。依據動物園資料為例，圖 3.6 中共群組了十八個二元特徵進入三個特徵組中，而後再經由排序與 BCD 產生數值資料，即代表著資料特徵個數由十八縮減至了三個特徵之中。本研究的方法中特徵組個數可以進行調整，若將群組個數設定成二維或是三維，即可將原先多維度的二元資料投影至可視覺化的維度進行分析，亦可同時比較不同排序與群組方式的有效度。

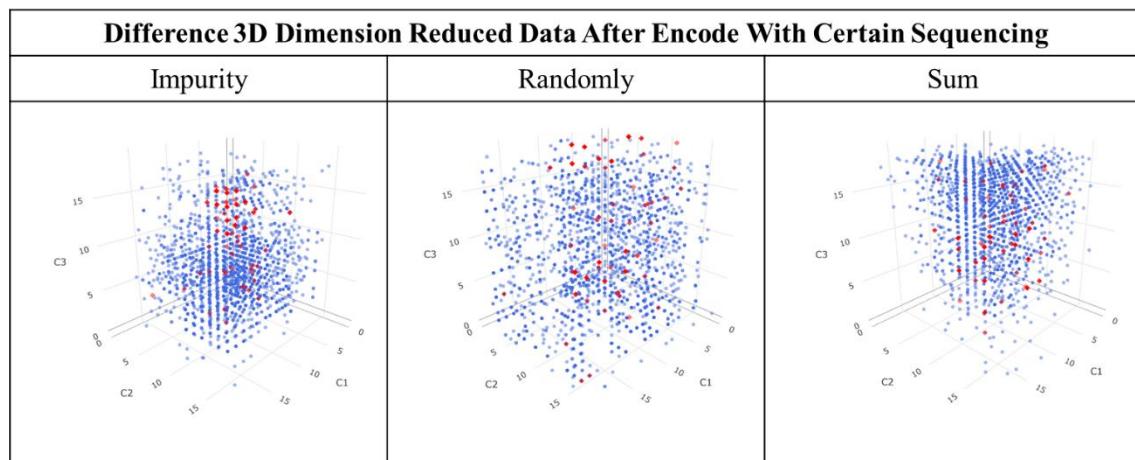


圖 3.15 呈現整體二元資料降維至三維後的資料分佈，依據特徵純粹度、隨機與特徵和排序方式。

如圖 3.15，即是根據案例分析的資料、使用相同群組方式；但依照不相同的特徵排序，將整體資料降至三維後，新的數值分佈；可以發現相較於根據特徵純粹度排列，隨機排列無法有效的聚集相同類別於同一處，且資料分佈的更為寬廣。

在圖 3.16、圖 3.17、圖 3.18 中，則呈現降至二維度後的資料分布，同時將如圖 3.7、圖 3.8 中的兩新數值資料做分佈度對比；於此更可見排序對於編碼後資料後的差異；比起隨機排序，特徵純粹度的排序更能產生適合分類的數值資料，投影了兩種類別資料至數線的兩端，得以讓編碼後的數值資料更易以類別區分；反觀隨機排列，兩類資料則與彼此相混雜，無法找出適合的切分點來區隔特徵。

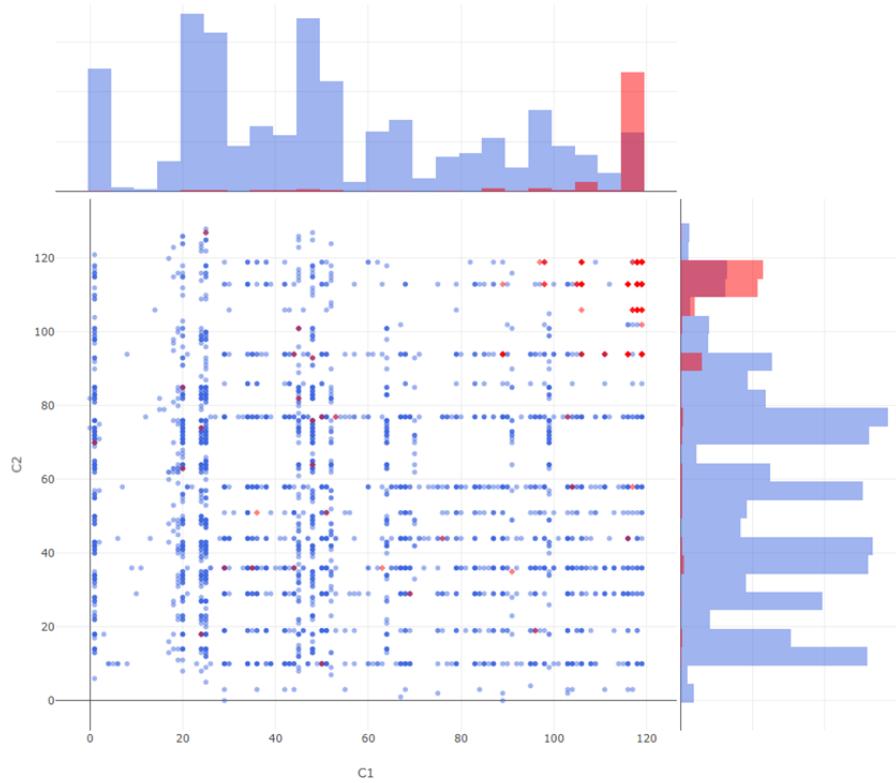


圖 3.16 呈現整體二元資料降維至二維後的資料分佈，依照特徵純粹度排序。

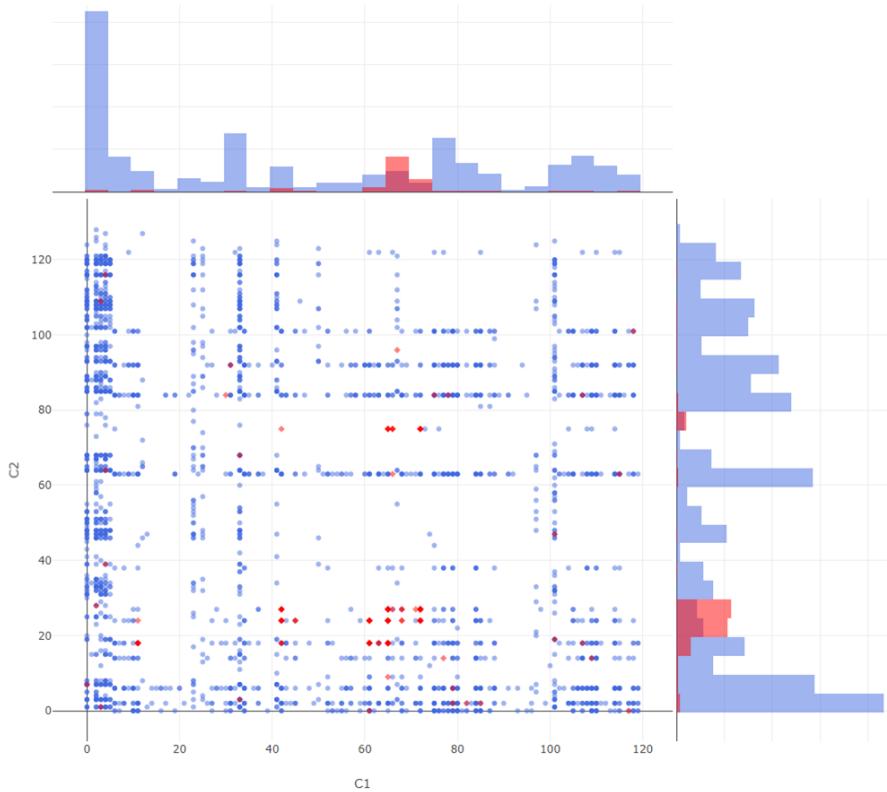


圖 3.17 呈現整體二元資料降維至二維後的資料分佈，依照隨機排序。

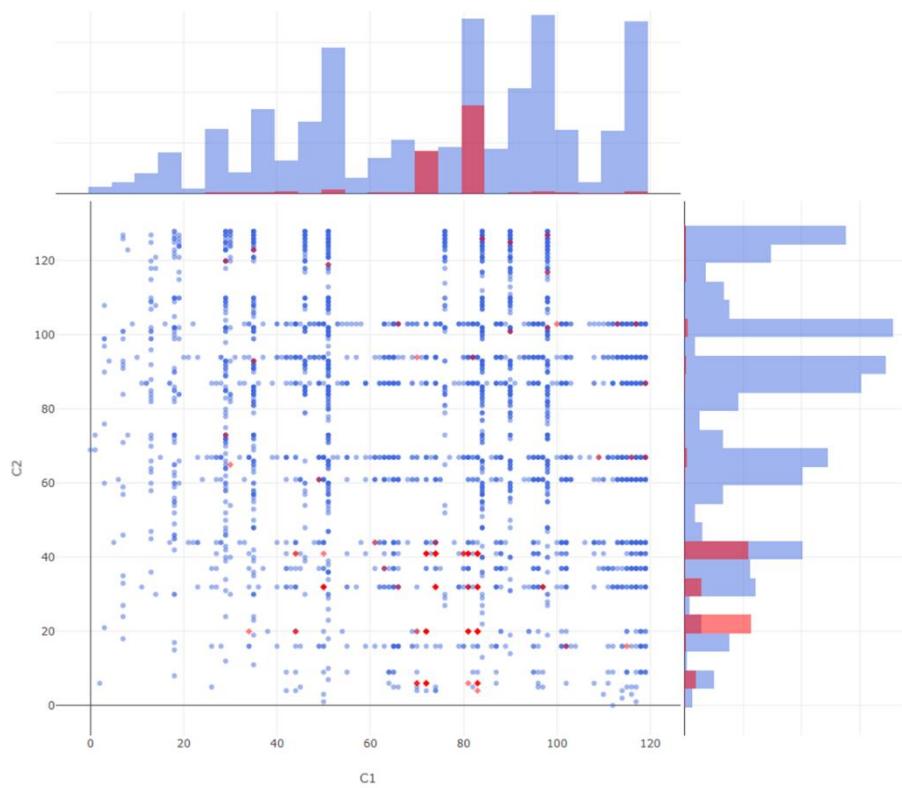


圖 3.18 呈現整體二元資料降維至二維後的資料分佈，依照特徵和排序。

3.5 分類與評估指標

本研究所採納之實驗資料具有大樣本、類別不平衡的特性，且以二元分類為目標。因此在考量了時間以及運算成本後，本研究採用 LightGBM 結合梯度提升決策樹建構分類模型，因為其具有在較短時間內並行處理、分類大量樣本資料的能力，適用於實驗資料的大樣本特性。面對類別不平衡的資料時，通常無法只根據準確度進行評估，而是須計算精確率以及召回率做為判斷依據，為此在評估指標的部分則以 F1-score 作為衡量標準，因其能在一定限度之內，同時兼顧到精確率以及召回率做為類別不平衡資料集的分類成果指標。

→也許移置第四章節序？

第四章 案例研討

本研究探討了多種資料集。在發展實驗方法與研究架構時，透過模擬出三維連續的二元分類資料，並依照比例切分為多個二元特徵作為資料集；同時也對於 UCI, Kaggle 等開源資料集平台上的類別特徵資料集進行研討，以比較傳統變數編碼與所發展之方法產生的新數值資料，對於機器學習模型分類成效之影響。

在比較的同時囊括了常見的傳統變數編碼方式，如獨熱、二進位、順序與目標編碼。本研究在群組二元特徵時，則考慮了以原先群組資訊、主成分分析、相關係數與隨機群組的方式來整合二元特徵，以下依序簡記為 Default、PCA、Corr、RND；而在排序群組內二元特徵時，則考慮的特徵和、目標類別純粹度、特徵重要性、與隨機排列，依序簡記為 Sum、Purity、FI、RND。

因所採納之實驗資料具有大樣本、類別不平衡的特性，且以二元分類為目標。因此在考量了時間以及運算成本後，本研究採用 LightGBM 結合梯度提升決策樹建構分類模型，因為其具有在較短時間內並行處理、分類大量樣本資料的能力，適用於實驗資料的大樣本特性。面對類別不平衡的資料時，通常無法只根據準確度進行評估，而是須計算精確率以及召回率做為判斷依據，為此在評估指標的部分則以 F1-score 作為衡量標準，因其能在一定限度之內，同時兼顧到精確率以及召回率做為類別不平衡資料集的分類成果指標。

4.1 連續二元分類資料測試

為了生成具有分類價值與背景知識的二元特徵下的分類資料，本研究透過模擬三維座標點分布的方式，生成了兩種類別的浮點數座標；如圖 4.1 所示，資料點具有 X、Y、Z 三維座標資訊，且區分為三百個紅色（少數類別）與三千個藍色（多數類別）兩種類別；兩種類別各自服從自身的資料分布，且在多數的藍色類別中，可見到有少量的紅色類別的躁點存在，目的在於增加分類任務的複雜度。

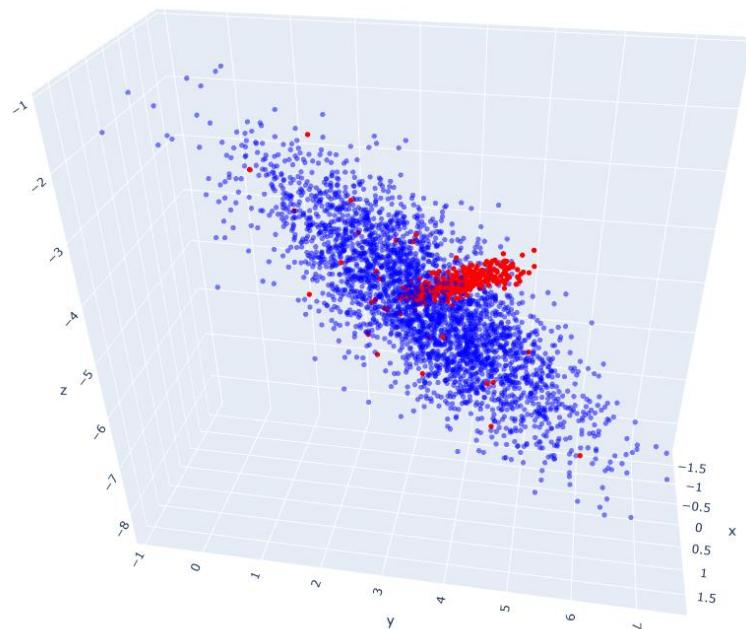


圖 4.1 模擬的連續二元分類資料。

4.1.1 資料集簡介與實驗架構

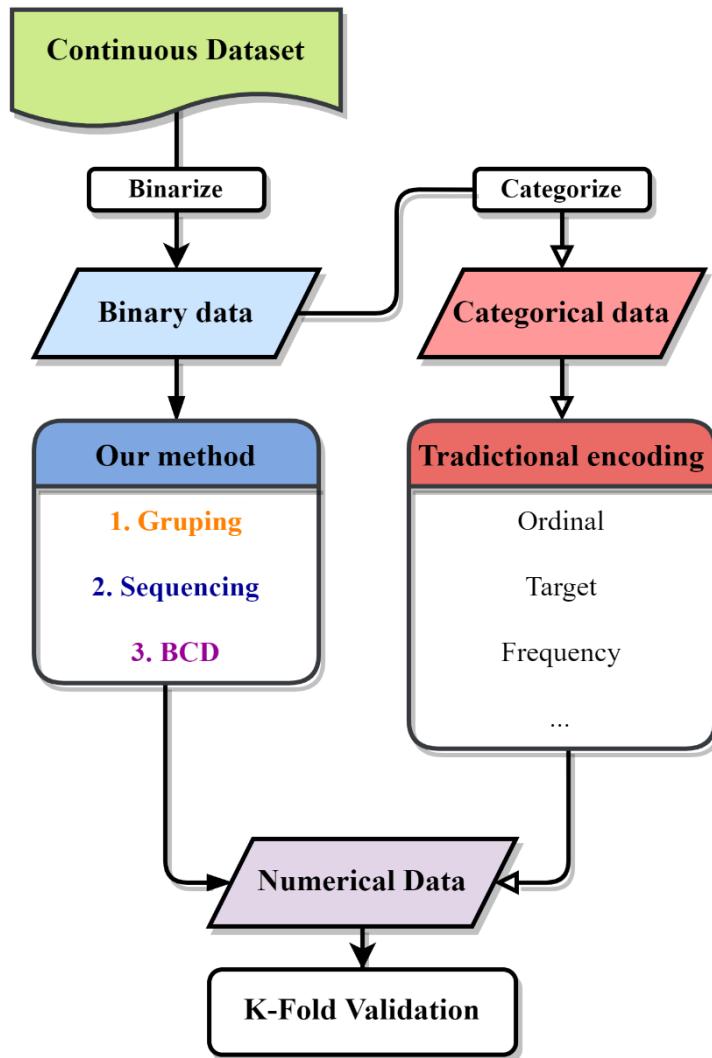


圖 4.2 連續資料集下的實驗架構。

整體實驗流程如圖 4.2 所示，模擬出二元分類的連續資料後，先擬定欲產生的二元特徵個數，便依此將三個連續座標軸依據比例切分來進行二元化，以產生二元特徵資料作為本研究方法輸入。同時也將二元資料類別化，以作為傳統變數編碼方式輸入，後交由 LightGBM 以交叉驗證的方式比較各種編碼後的數值資料集的分類成果。

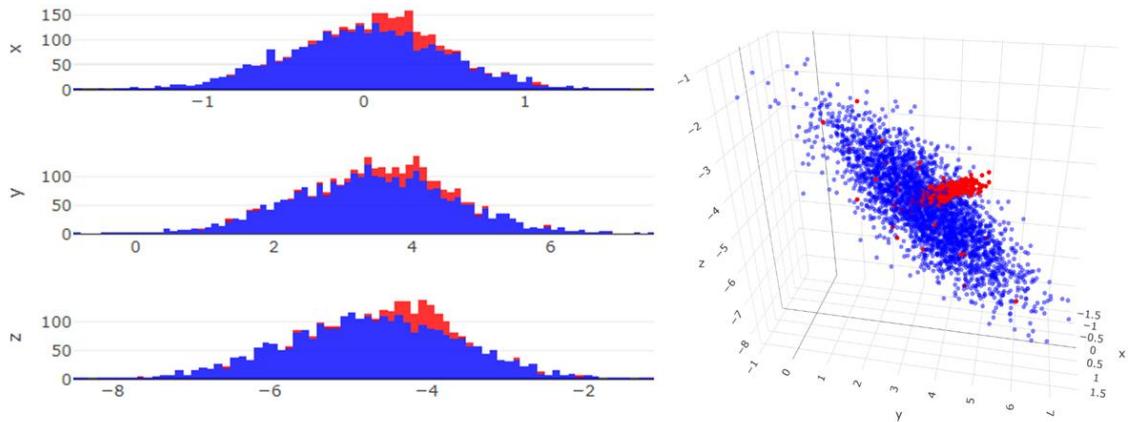


圖 4.3 原始連續資料於 X、Y、Z 三維度上的分布情形。

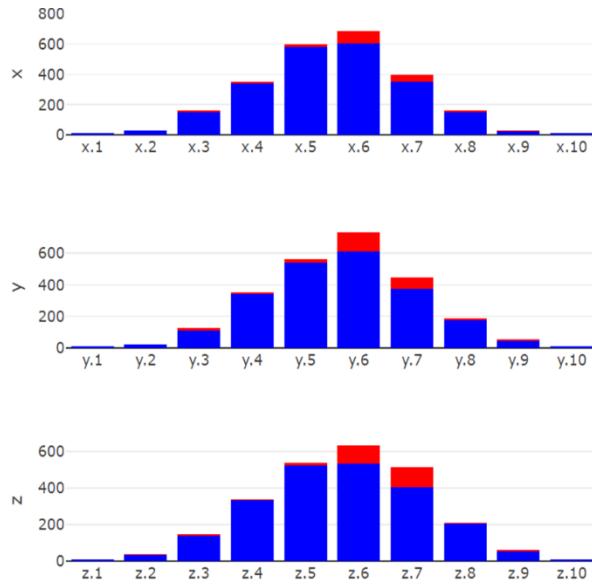


圖 4.4 二元化後的連續資料，共劃分為 30 個二元特徵。

圖 4.3 表示了原始連續資料於三個座標軸上的資料分布情形，且少數與多數類別由顏色區分；在決定了要劃分出的二元特徵總數之後，便依據比例切分各個座標軸的全距，並將資料點劃分入新產生的二元特徵之中，如此，便產生了具備分類知識的二元特徵資料；如圖 4.4 所示，各個座標軸各被切分出了十個區域，總計產生了三十個新的二元特徵，作為二元資料。

表 4.1 二元化後的特徵資料。

Instances	x.1	x.2	x.3	x.4	x.5	x.6	...	z.9	z.10
1	0	1	0	0	0	0	...	0	0
2	1	0	0	0	0	0	...	0	0
3	0	0	0	0	0	0	...	0	0
4	0	0	0	0	1	0	...	0	0
5	0	0	1	0	0	0	...	0	0
6	0	0	0	0	1	0	...	0	0
7	0	0	0	0	0	0	...	0	0
:	:	:	:	:	:	:	:	:	:
3300	0	0	0	1	0	0	...	0	0

表 4.2 類別化後的特徵資料。

Instances	X	Y	Z
1	x.2	y.9	z.4
2	x.1	y.8	z.2
3	x.9	y.1	z.6
4	x.5	y.7	z.8
5	x.3	y.1	z.8
6	x.5	y.6	z.3
7	x.7	y.5	z.4
:	:	:	:
3300	x.4	y.8	z.1

經由二元化後的資料如表 4.1 所呈現，原先三個維度的連續資料經由二元化之後，便產生了具有三十個二元特徵，如此便可以做為本研究的方法輸入；而為了與傳統的變數編碼進行對比，再加上具備了此筆二元資料的二元特徵群組知識，得以經由類別化，產生出類別資料，如表 4.2 所示。此兩種資料乃為一體兩面，僅僅在表現方式上有所差異，但其本質代表的資料性質實為相同；類似於原始類別資料，之於獨熱編碼轉換出的二元資料。

4.1.2 不同連續資料集之下的測試與實驗

本研究嘗試了多個不同資料分布的連續資料集，比較所提出的特徵群組與排序方式、與傳統變數編碼產生的數值資料集，在切分了不同的二元特徵個數之下，對應 Light GBM 機器學習模型分類成效的變化；其中，選用 F1 score 作為評比資料及分類難易度的指標。

4.1.2.1 連續資料集一

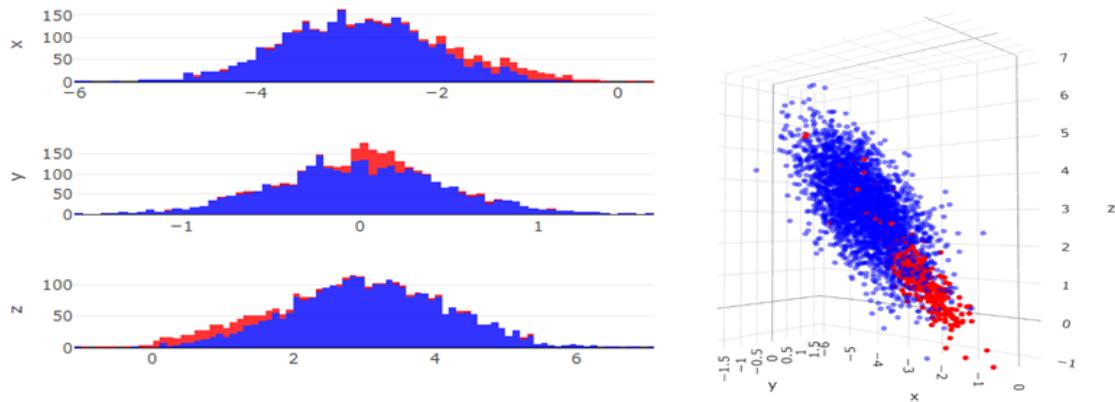


圖 4.5 連續資料集一的資料分布。

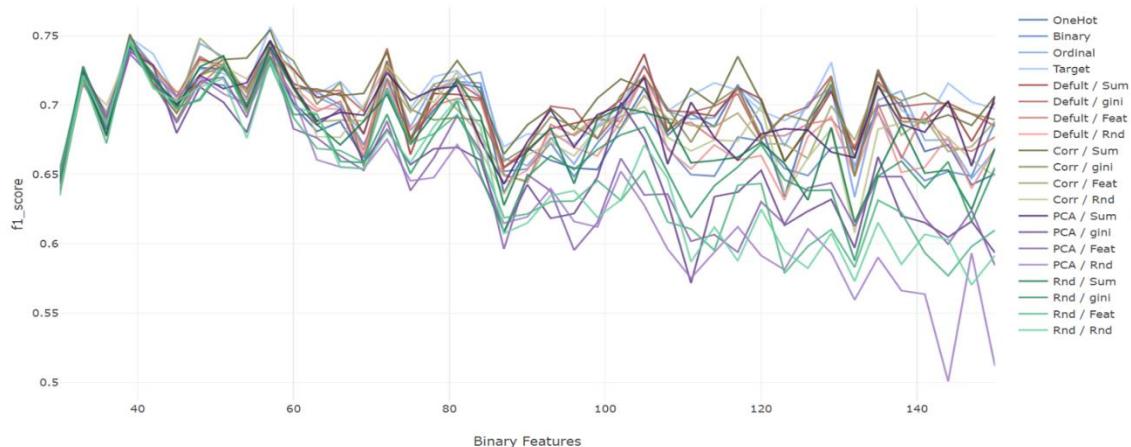


圖 4.6 連續資料集一中，不同編碼方式所得數值資料的分類成績，對應切分二元特徵數量變化。

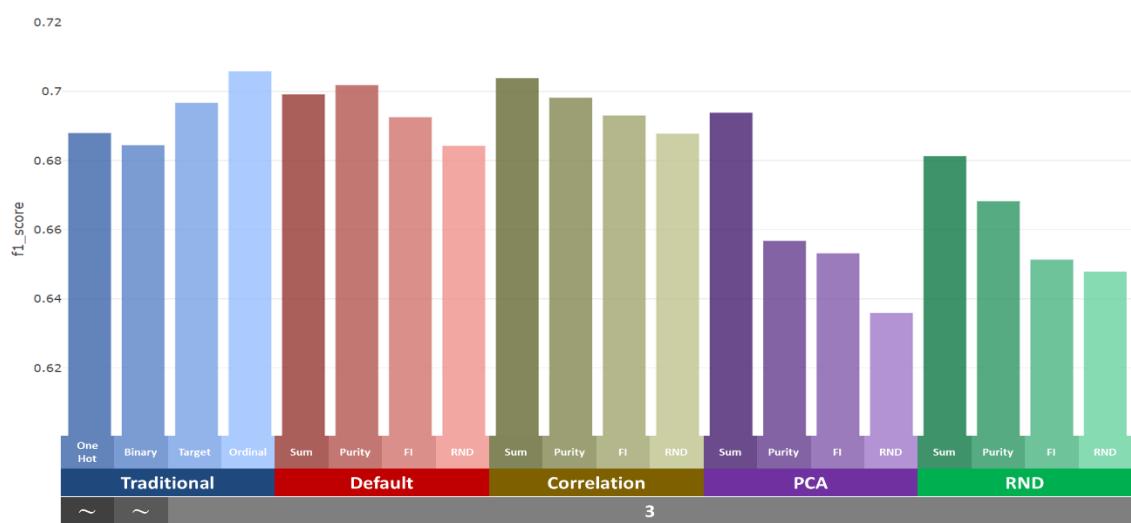


圖 4.7 連續資料集一中，不同編碼方式所得數值資料的平均分類成績。

4.1.2.2 連續資料集二

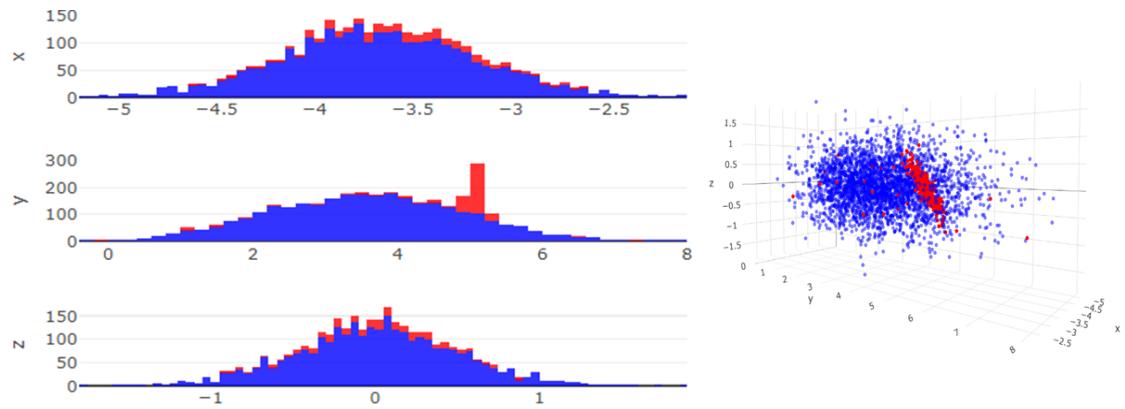


圖 4.8 連續資料集二的資料分布。

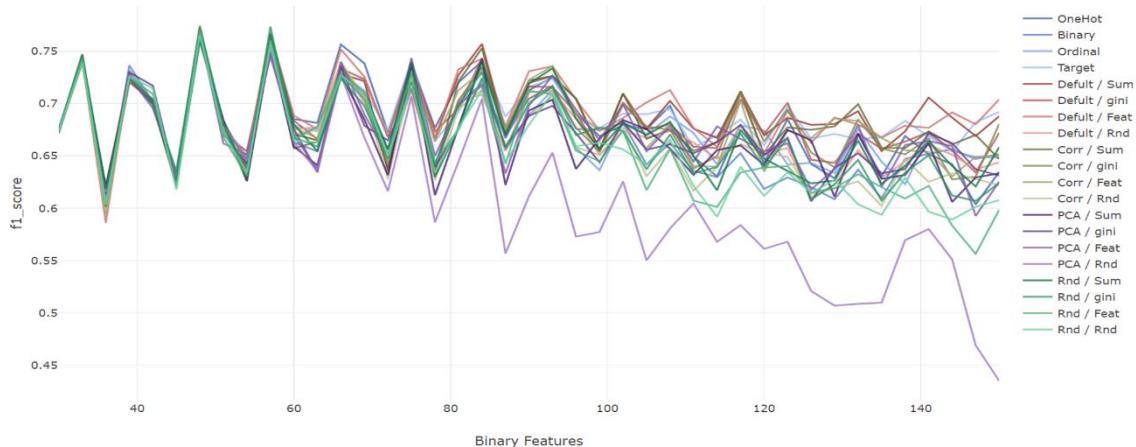


圖 4.9 連續資料集二中，不同編碼方式所得數值資料的分類成績，對應切分二元特徵數量變化。

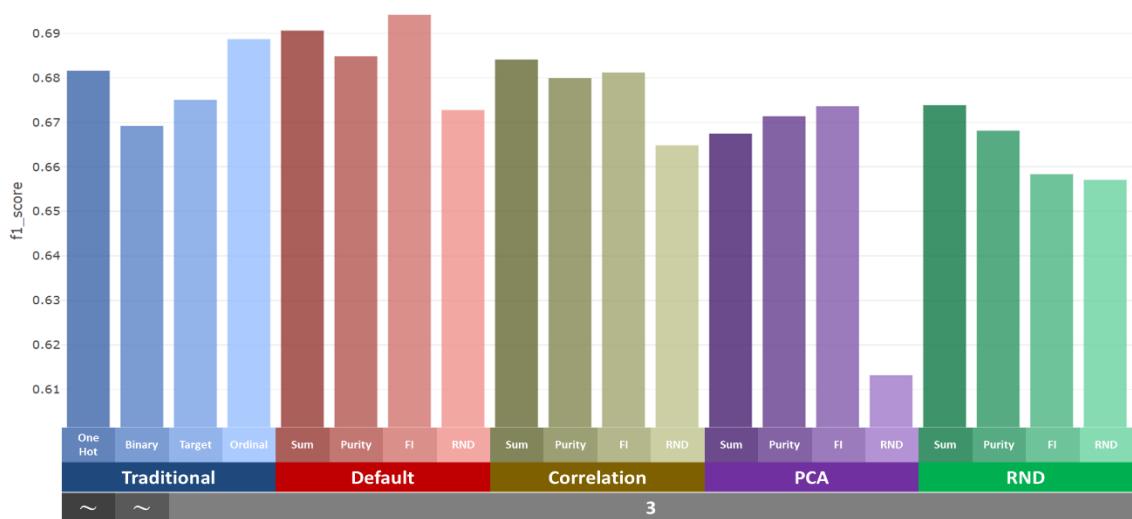


圖 4.10 連續資料集二中，不同編碼方式所得數值資料的平均分類成績。

4.1.2.3 連續資料集三

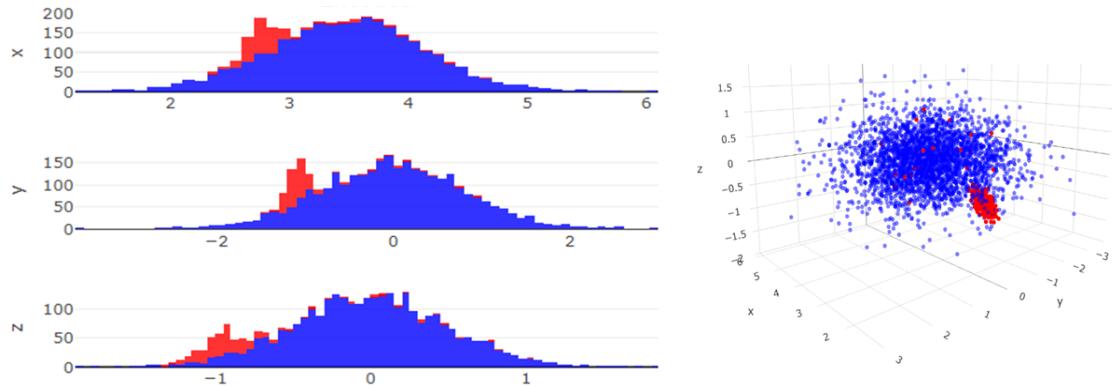


圖 4.11 連續資料集三的資料分布。

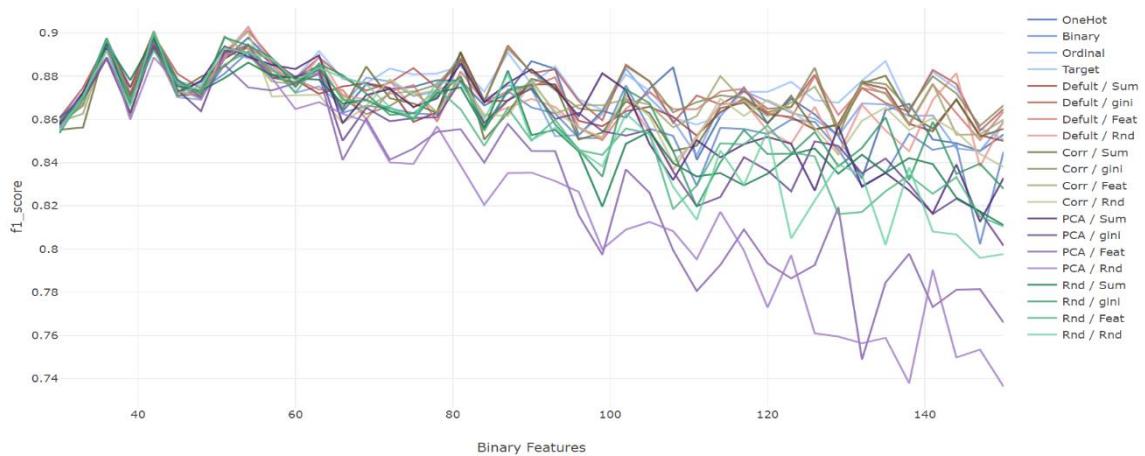


圖 4.12 連續資料集三中，不同編碼方式所得數值資料的分類成績，對應切分二元特徵
數量變化。



圖 4.13 連續資料集三中，不同編碼方式所得數值資料的平均分類成績。

4.1.3 分類結果評比與歸納

綜觀此三個連續資料集經由不同編碼方式後，所產生的新數值資料的分類成果，即圖 4.7、圖 4.10、圖 4.13 後，得以歸納出以下幾點現象：

1. 群組二元特徵對於編碼後數值資料的分類成果具有一定影響：

由以上三圖可見，經由不同群組方式產生的數值資料有著不盡相同的平均成績，不論根據何種方式排序組內特徵，分類成績似乎仍部分受到群組方式的影響。依據以上三圖可大致發現，由主成分、隨機的群組方式將使編碼後的數值資料，相較於實施本研究方法前的二元特徵資料 (OneHot) 的平均分類成效減低，並無法在縮減特徵總數時，有效的維持或提升編碼資料的分類成績；然而，根據原先群組資訊、相關係數進行群組的方式卻能在縮減整體特徵個數的同時，維持一定的分類成績，可見群組二元變數的方式將提升或限制編碼後新數值資料的分類成果。

2. 排序組內特徵對於編碼後數值資料的分類成果具有一定影響：

在以上三圖中也不難發現，不論群組方式為何，隨機排序組內群組特徵所產生數值資料的分類成績皆不甚理想，幾乎都屬於該群組方式中成績最低的排序方法。而使用特徵純粹度、特徵重要性等監督式排序時可取得在分類表現上較為優異的數值資料。可見排序特徵組內特徵也將對編碼後數值資料的分類成果造成影響。

3. 目標編碼的表現相當優異：

相較於其餘傳統變數編碼，目標編碼參看了資料的目標欄位，屬於監督式的變數編碼，使得其對於新數值資料的分類成果有著一定程度的幫助，導致編碼後的新資料的分類表現出眾，可與本研究所發展的監督式排序方法作比較；然而，目標編碼僅處理類別資料，而本研究所發展的方法著重編碼無群組資訊、且眾多二元特徵的資料之上；對於此類資料，目標編碼將無法進行處理。

4. 不同編碼方式之間，所使用特徵個數不盡相同：

如以上三圖底部灰階欄位所標示，不同的編碼方式編碼後數值資料的特徵個數與彼此相異。獨熱編碼、與二進位編碼的特徵個數隨著二元化連續資料的目標個

數改變；二元化後的連續資料即經由獨熱編碼的新資料，其特徵個數恆等於切分出的二元特徵個數；而二進位編碼的特徵個數則為 $\lceil \log_2(n) \rceil$ ，其中 n 為二元化連續資料的目標個數。而目標、順序編碼與經由原先群組資訊進行編碼的資料則依循原類別資料的群組方式，維持 X、Y、Z 三個特徵群組。經由相關係數、主成分分析與隨機群組後的新數值資料的特徵個數則未有規範，可自行調整編碼後欲使用的特徵個數，在此實驗當中設定為三，與原始群組個數相當，以便於與其餘編碼方式進行比較。

根據圖 4.6、圖 4.9、圖 4.12 三圖則可以觀察到在切分的二元變數數量較少時，所有變數編碼產生的數值資料有著相近的分類成績；然而，隨著所切分的二元變數個數增加，變數編碼產生數值資料的分類成績之間的變異也逐漸增加、且分類成績也有著緩步下降的趨勢，類似於圖 2.4 中維度災難發生時導致的模型成果下降。

4.2 UCI 資料集

為了驗證本研究發展之編碼架構，於 UCI 線上資料及網站之中搜尋適合的標竿資料集，來測驗本研究提出的不同排序、編碼方式對於新編碼資料分類成效的影響，同時對於傳統的類別變數編碼方式進行比較。

4.2.1 資料集簡介與實驗架構

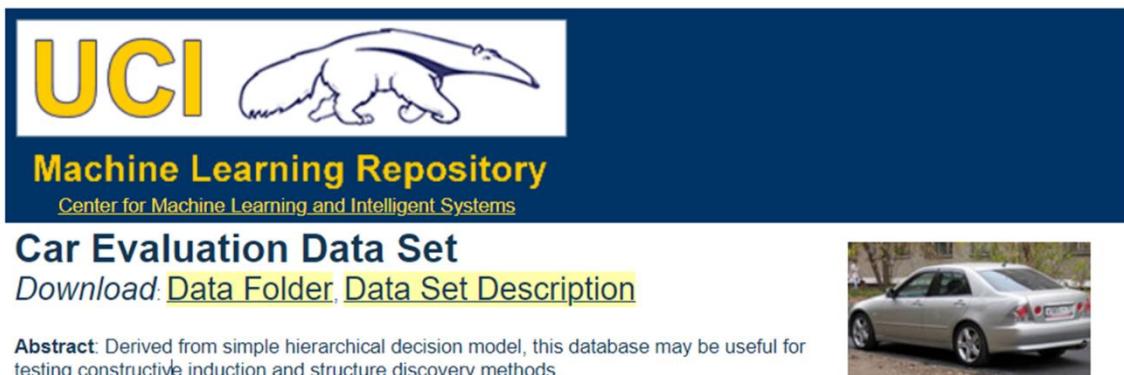


圖 4.14 UCI 網站上的二手車輛車況評估資料集。

本實驗選用與 Potdar et al. (2017)比較不同變數編碼時，所採用的二手車輛評估資料，並比較各式編碼產生數值資料於相同分類器之下的分類成果。此資料集的特徵、樣本個數皆相對較為稀少，且獨熱編碼產生後的二元特徵數量也相對較少，於本研究之中最為單純的資料集。在此資料集中，具有六個類別特徵，分別用以描述一千七百多台二手車的價格、維護費用、車門個數、乘載人數、行李箱尺寸以及安全程度，用以推論該二手車車況的好壞。

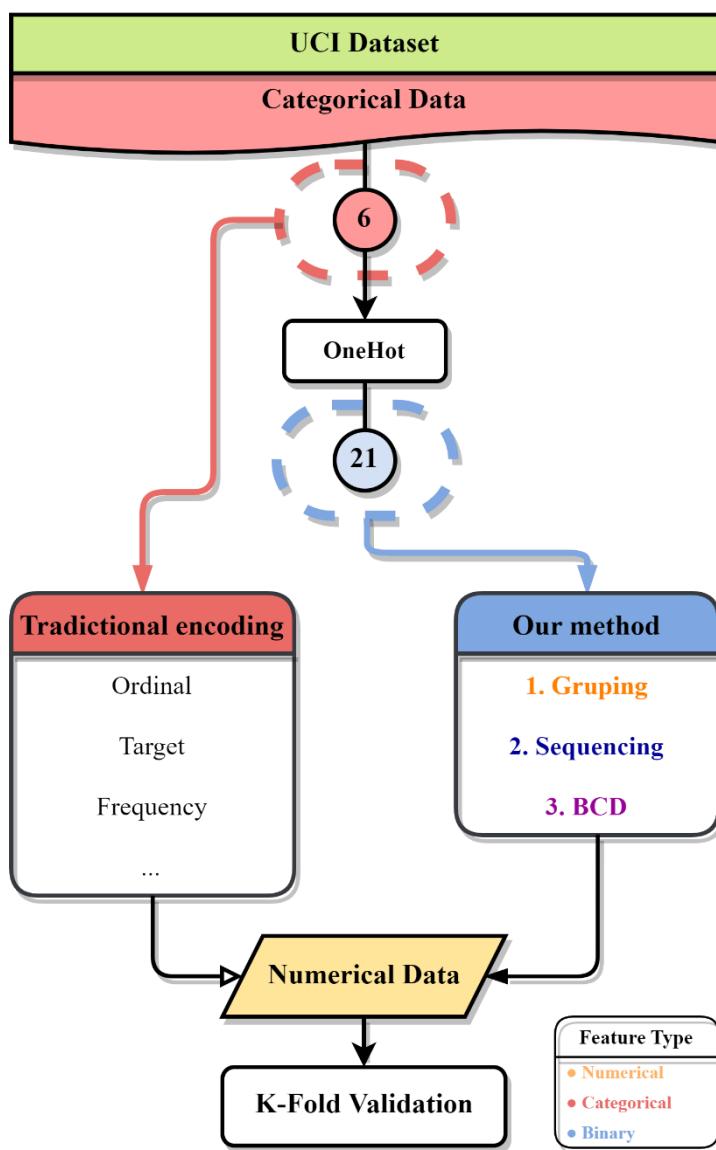


圖 4.15 UCI 資料集下的實驗架構，圓圈內為該類特徵個數。

如圖 4.15 所示，傳統變數編碼轉換類別特徵為數值資料，而經由獨熱編碼之後將產生二十一個二元特徵，則交由本實驗發展之編碼方式便針對此類二元特徵進行編碼，協助其轉換成為數值型別資料。最終，藉由 LightGBM 產生的梯度提升決策樹模型來為各數值資料集做交叉驗證，並評比各數值資料集的分類成果。

4.2.2 分類結果評比與歸納

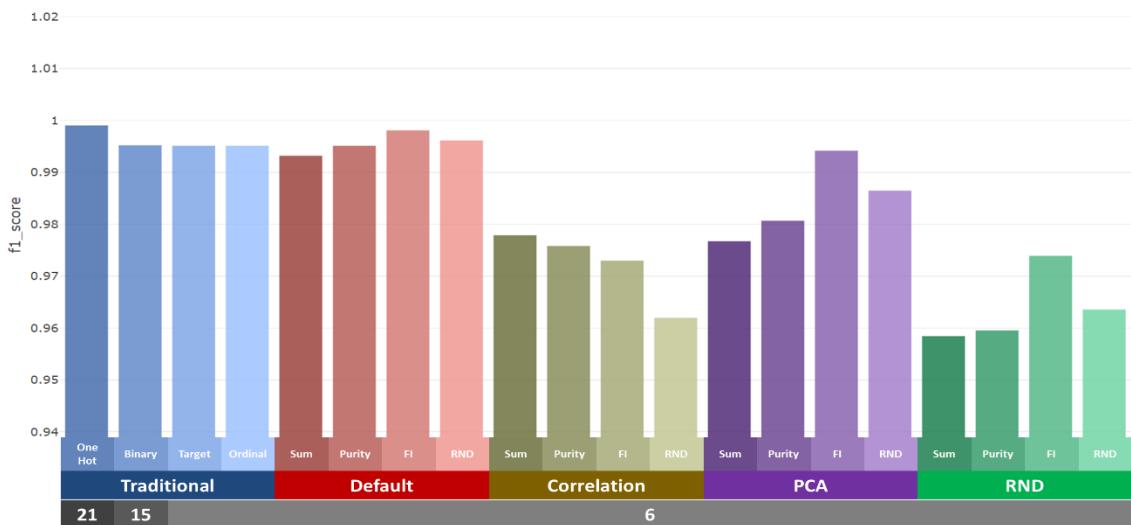


圖 4.16 UCI 資料集中，不同編碼方式所得數值資料的平均分類成績。

在圖 4.16，可以看見傳統變數編碼方式、以及依照原先群組資訊而產生的數值資料集皆表現得相當優異，甚至是單純的二元特徵資料即表現的相當不錯；反之，透過相關係數、主成分分析或隨機群組的方式皆造成分類成績的損失。

本研究推論此結果的原因有二：其一是此資料本身過於單純，任一編碼方式的 F1 score 皆高於零點九，而 LightGBM 在二十一個二元特徵之下仍能表現出不錯的分類能力，當遭遇維度較低的二元資料問題時，使用本研究所發展之二元特徵編碼方式將無法有效用的使新數值資料的分類成績有大幅度的提升；其二則是群組方式仍需改進，根據相關係數、主成分分析的群組方式，皆無法達到根據原先的二元資料群組資訊相同的組分類平均，而且也無法明顯的比隨機群組方式來的優越。

依據上述結論，本研究欲找尋具有更多二元特徵、更多樣本，且類別更為不平衡、更難以分類資料集作為實驗資料，以觀察群組、排序二元特徵後產生資料集的相關特性。

4.3 Kaggle 資料集

為了對本研究發展之方法做進一步的試驗，透過線上資料集網站—Kaggle 中搜尋具有類別變數、且須經由變數編碼的資料集作為測試資料；在實踐本研究方法的同時，也能一同與傳統變數編碼後的新資料做分類成績上的比較。

4.3.1 資料集簡介與實驗架構



圖 4.17 Kaggle 網站上的類別特徵編碼挑戰資料集。

在瀏覽眾多資料集後，選用對於類別變數編碼挑戰的資料集進行測試。採用的原因主要為該資料集同時具有數值、類別與二元變數；且同為二元分類問題；經由獨熱編碼後將產生眾多二元特徵。在該資料集中，三十萬個樣本由五個二元特徵、十六個類別特徵、天數、月份與目標欄位所構成；而在某些特定的類別特徵欄位中，紀錄的是如序號、亂碼等種類眾多的特徵資料，為了簡化輸入資料，本研究無視該類類別特徵，僅納入其中的九個種類較為均衡的類別特徵。

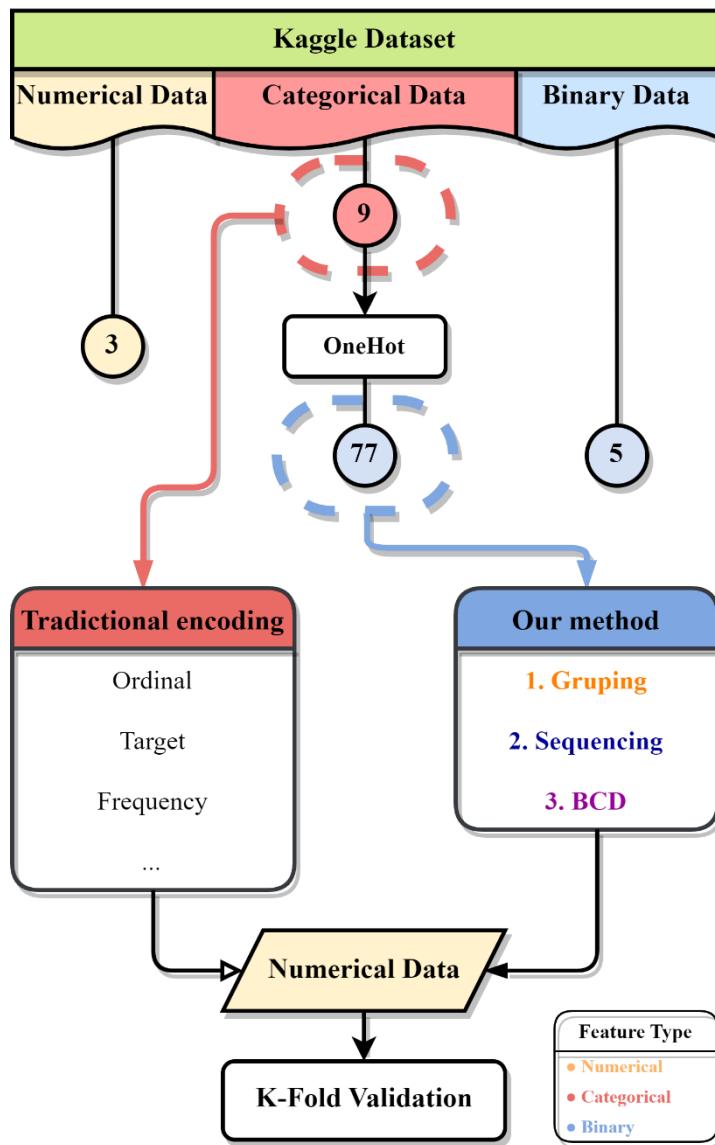


圖 4.18 Kaggle 資料集下的實驗架構，圓圈內為該類特徵個數。

如圖 4.18，篩選過後的資料集經由獨熱編碼後，將具有七十七個二元特徵，作為本研究方法的輸入，後產生的數值資料將與原始資料的數值特徵、二元特徵相結合，作為編碼完成的資料集；而傳統的變數編碼則針對原始資料中的類別特徵進行方法不同的轉換，轉換後同與原始資料的數值特徵、二元特徵相結合，作為編碼完成的資料集。而後由 LightGBM 訓練梯度提升樹作為分類模型，以交叉驗證的方式比較各式傳統變數編碼方式、與本研究提出之各種群組、排序手法產生之數值資料集的分類成績。

4.3.2 分類結果評比與歸納



圖 4.19 Kaggle 資料集中，不同編碼方式所得數值資料的平均分類成績。

如圖 4.19，可以明顯發現分群、與排序對於編碼後資料集的明顯影響，以及本研究方法與傳統編碼方式之間的差異。像是在傳統的變數編碼中，相較於其餘非監督式的編碼方式，監督式的目標編碼對於產生之數值資料的分類成績有明顯提升。而不同的群組方式也會影響分類成績，依照原先群組資訊群組特徵將有著更高的平均分類成績，再來則分別是以主成分分析、相關係數及隨機群組的方式。在各個群組方式中、不同排序方法之下，可以發現以特徵純粹度排序將產生出分類成績較佳的數值資料。

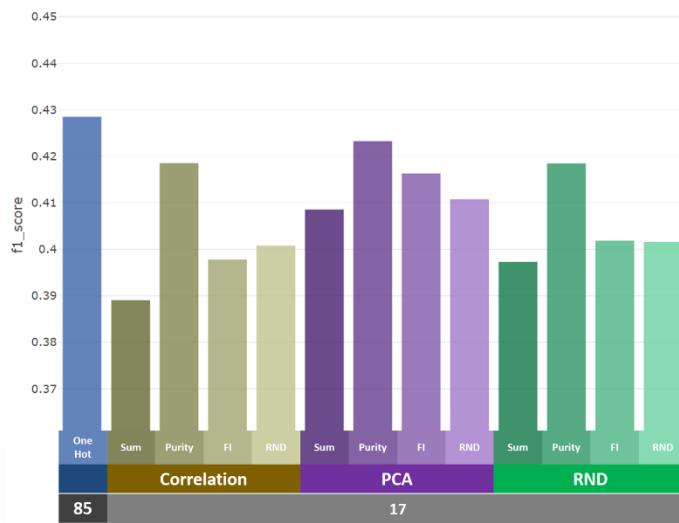


圖 4.20 面對無法類別化的二元特徵資料時，所能使用的數值編碼方式。

雖然依照主成分分析、相關係數的群組方式產生的數值資料，無法在分類成果上優於依據原先群組方式所產生的資料，但當原始資料缺乏獨熱編碼的群組資訊、或單純為多維度的二元資料時，便只能使用本研究所發展方法，協助二元資料轉換為數值資料。而在圖 4.20 可見，表示若能以適當的群組與排序方式，由十七個數值特徵構成的新資料中能與具有八十五個二元特徵原始資料有相似的分類成果；如圖 4.19，雖然依據特徵純粹度排序能有效提升分類成績，但目前缺乏能達到原先群組資訊一樣優秀分類成績的群組方式。

第五章 結論與建議

5.1 研究成果

相較於傳統變數編碼僅針對類別型態的資料，本研究提出了一種將二元特徵資料編碼成數值資料的架構設計，透過群組、排序與二進位十位數編碼等一連串步驟，不論原先的二元特徵資料是否包含二元特徵群組的資訊，皆可以將其編碼成數值資料的形式。由二元特徵群組，得以縮減高維度二元資料維度至較低維度的數值資料；而後由特徵組內的排序，得以提升編碼後數值資料集的分類成績；最終，藉由二進位十位數編碼，將同一群組內的複數二元特徵轉換成為單一數值特徵。

在第四章案例研討可見，相較於單純的資料集，當本方法應用於複雜且擁眾多二元特徵的資料時更能展現出本實驗提出編碼框架的優勢。相比於編碼之前的二元特徵資料、經由本研究編碼後產生的數值資料可具有以下特性：

1. 縮減資料特徵個數：

相比於原始二元資料具有眾多的特徵個數，編碼後的數值資料能大幅度縮減所使用到的資料維度，在一定程度上減少分類模型處理資料時所耗費的運算資源。

2. 壓縮特徵資訊：

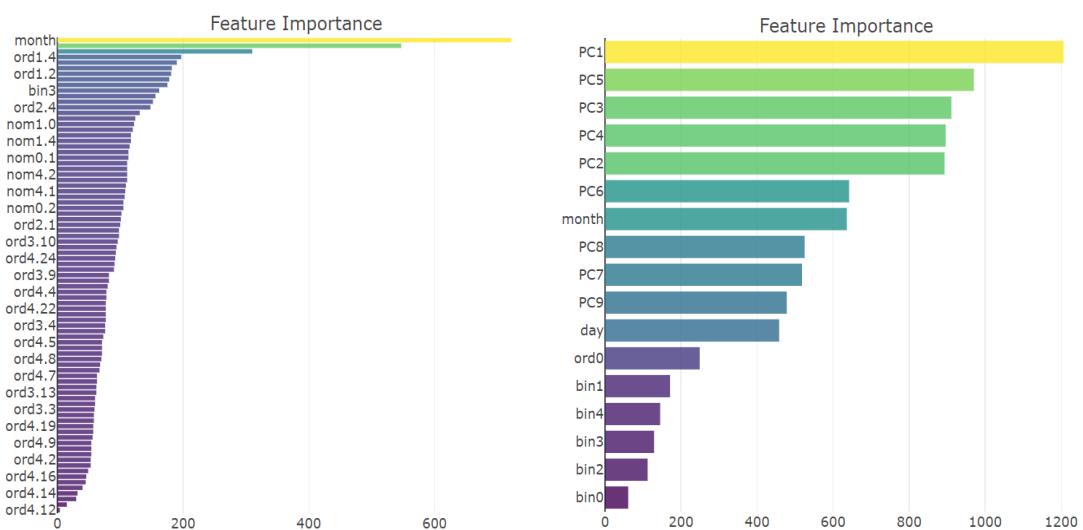


圖 5.1 LightGBM 分類模型訓練後的特徵重要度。

(左) 原始二元資料 (右) 主成分分析群組

根據特徵選取方式，將相關、相近的二元特徵劃分為同一群組之內，並重組這些二元特徵成為數值特徵。如圖 5.1，以 Kaggle 資料集為範例；由 LightGBM 訓練出的梯度提升樹分類模型中的各特徵重要性可以發現，相較於編碼前的二元資料，編碼後的數值資料的平均特徵重要度有著顯著提升，代表原先難以區分目標類別的二元特徵經由壓縮與融合，成為了更具有分類價值的數值特徵。

3. 維持或提升編碼過後資料的分類表現：

透過觀察傳統編碼以及本研究提出之各式群組、排序方式所產生數值資料，於 LightGBM 梯度提升樹分類模型的分類表現之後，可以發現群組、排序方式對於資料集分類成績存在著一定的影響，且依據如特徵純粹度、特徵重要度等監度式的排序方法結合原先的群組資訊，能於匹敵目標編碼的分類成果，並普遍優於其餘傳統變數編碼方式。

4. 作為高維度二元資料的降維方法：

本方法主要意在協助處理具有眾多二元特徵、無法經由傳統變數編碼轉換的二元資料，同時也可以做為此類資料的降維方式，不論改以三維、還是二維做為目標維度，皆可以看見排序對於編碼後的整體數值資料的影響；若採用優勢的排序方式，可以達到區分不同類別的效果，進而提升編碼後數值資料的分類成果。

5.2 未來研究方向

本研究提出的二元特徵編碼框架，可以由不同的群組、排序方式再進行擴充或改良。例如以別的特徵選取方法來群組二元特徵、或是以其他最佳化的方式進行特徵組內的排序等。也可以針對最佳的特徵群組個數進行探討，依據編碼後資料集的分類成績，尋找最佳的特徵群組個數。如圖 5.2 表示，當嘗試將二元資料以不同群組數，分類成績隨著群組個數增加，有著類似於維度災難發生之前的成績提升。

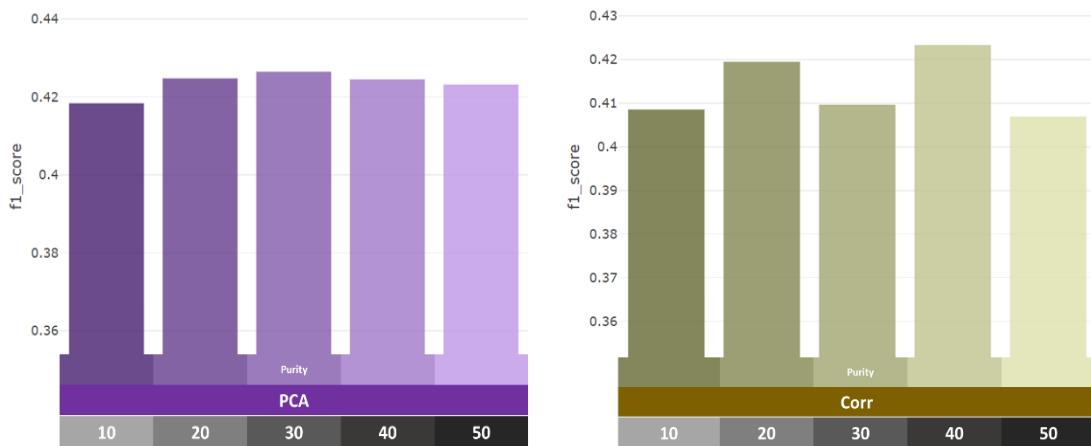


圖 5.2 Kaggle 資料集中，依據不同的特徵個數群組二元特徵下的分類成績。

(左) 主成分分析群組 (右) 相關係數群組

在第四章案例研討中，觀察許多資料集與不同編碼方法產生的新數值資料的分類成績可以發現到，若是能以原始特徵的群組資訊來群組二元特徵，並採用特徵純粹度排序之後，產生數值資料往往在分類成績上有著不俗的表現，且能與目標編碼相互比較；然而，目前採用的主成分分析、相關係數群組方式似乎無法達到了如此優良的效果。觀察圖 5.3、圖 5.4 可以發現到依照原先特徵的資訊群組能使新特徵之間的相關性更小，也許便是使其編碼後資料有著優異分類成績的原因之一；該如何改良群組方式也是本研究未來的研究方向。

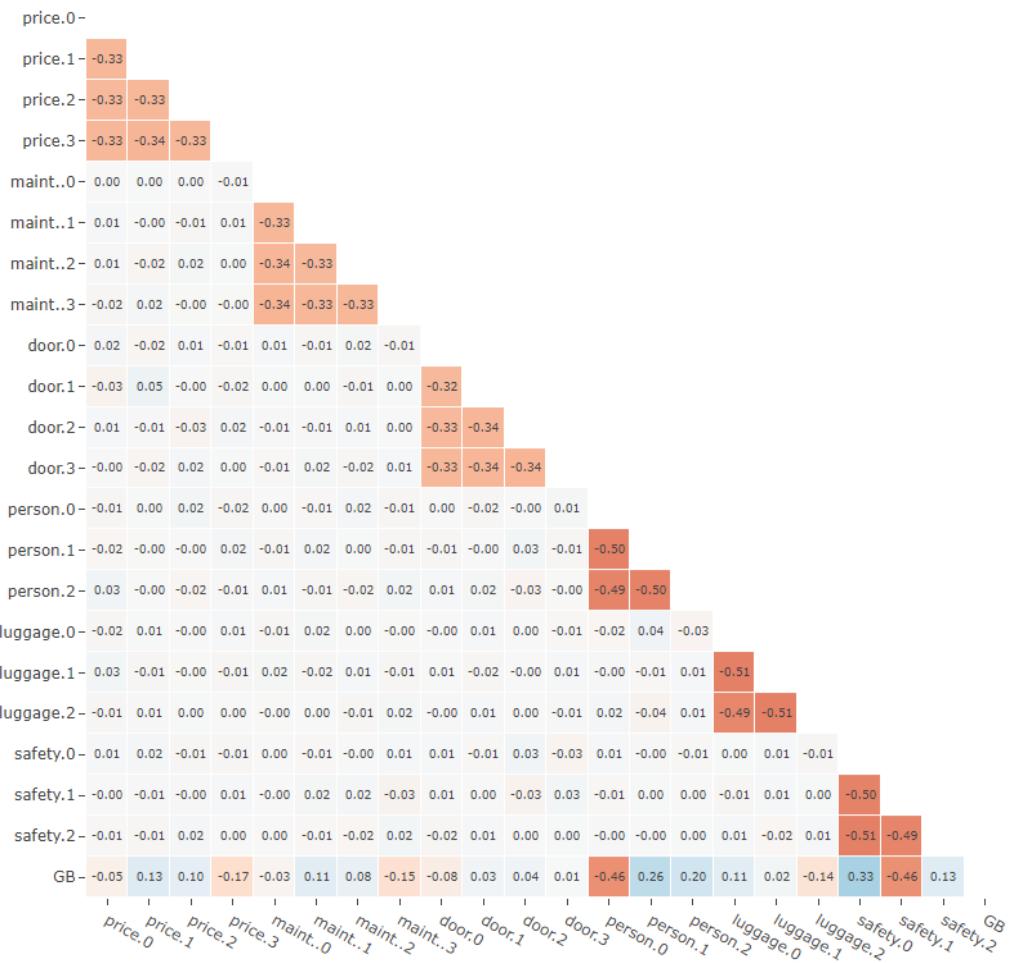


圖 5.3 二元資料特徵相關係數圖，其中 GB 表示目標欄位。

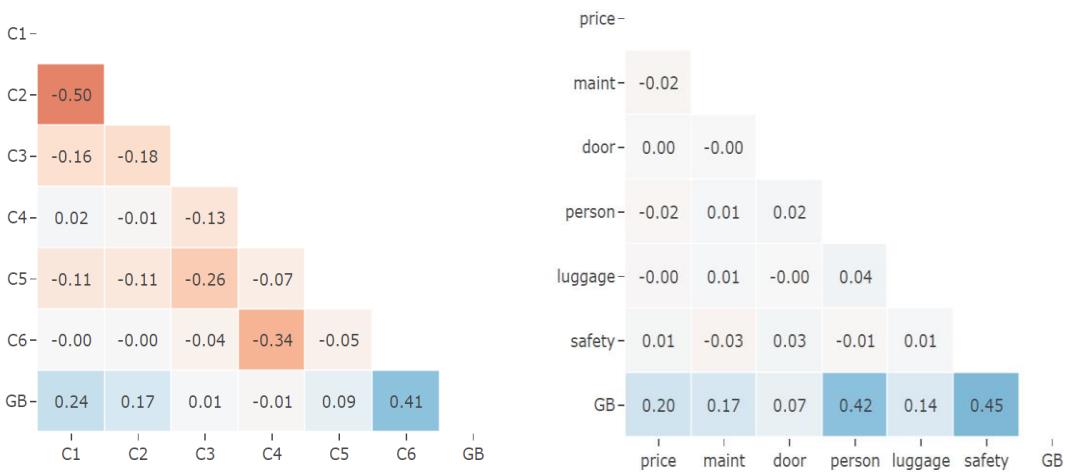


圖 5.4 不同群組方式產生之特徵相關係數圖，其中 GB 表示目標欄位。

(左) 相關係數群組 (右) 二元特徵群組資訊

參考文獻列表

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4), 433-459.
- Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, 2(2), 131.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,
- Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.
- Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. icml,
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.
- García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (Vol. 72). Springer.
- Hall, M. A. (1999). *Correlation-based feature selection for machine learning* [The University of Waikato].
- Hall, M. A., & Smith, L. A. (1999). Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. FLAIRS conference,
- Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015.
- Köppen, M. (2000). The curse of dimensionality. 5th online world conference on soft computing in industrial applications (WSC5),
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
- Liu, X., Zhu, X.-H., Qiu, P., & Chen, W. (2012). A correlation-matrix-based hierarchical clustering method for functional connectivity analysis. *Journal of neuroscience methods*, 211(1), 94-102.
- Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, 175(4), 7-9.
- Song, Y.-Y., & Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2), 130.
- Spruyt, V. (2014). *The Curse of Dimensionality in classification*.
<https://www.visiondummy.com/2014/04/curse-dimensionality-affect->

[classification/](#)

- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677-680.
- Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, 37.
- Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA.
- Verleysen, M., & Fran ois, D. (2005). The curse of dimensionality in data mining and time series prediction. International work-conference on artificial neural networks,

附錄 A (如果有)