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

人工智慧、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](#_Toc120735400)

[Abstract ii](#_Toc120735401)

[目錄 iii](#_Toc120735402)

[圖目錄 v](#_Toc120735403)

[表目錄 viii](#_Toc120735404)

[1 第一章 緒論 1](#_Toc120735405)

[1.1 研究背景 1](#_Toc120735406)

[1.2 研究動機與目的 3](#_Toc120735407)

[1.3 研究架構 5](#_Toc120735408)

[2 第二章 文獻探討 6](#_Toc120735409)

[2.1 變數編碼 6](#_Toc120735410)

[2.1.1 順序編碼 9](#_Toc120735411)

[2.1.2 獨熱編碼 10](#_Toc120735412)

[2.1.3 二進制編碼 11](#_Toc120735413)

[2.1.4 頻率編碼 12](#_Toc120735414)

[2.1.5 目標編碼 13](#_Toc120735415)

[2.2 維度災難 15](#_Toc120735416)

[2.3 降維處理 19](#_Toc120735417)

[2.3.1 特徵選取 19](#_Toc120735418)

[2.3.2 特徵萃取 22](#_Toc120735419)

[2.4 LightGBM分類模型 25](#_Toc120735420)

[2.5 驗證指標 26](#_Toc120735421)

[3 第三章 高維二元特徵之聚合編碼技術 27](#_Toc120735422)

[3.1 二元特徵分群 29](#_Toc120735423)

[3.1.1 資料原始特徵群 29](#_Toc120735424)

[3.1.2 主成分分析群集 30](#_Toc120735425)

[3.1.3 相關係數群集 31](#_Toc120735426)

[3.2 .群內二元特徵排序 34](#_Toc120735427)

[3.2.1 二元特徵總和排序 36](#_Toc120735428)

[3.2.2 特徵純粹度排序 36](#_Toc120735429)

[3.2.3 特徵重要度排序 37](#_Toc120735430)

[3.2.4 基因演算排序法 38](#_Toc120735431)

[3.3 群組二進碼十進數編碼 40](#_Toc120735432)

[3.3.1 二進位十位數編碼數值 40](#_Toc120735433)

[3.3.2 二進位十位數編碼數值排名 41](#_Toc120735434)

[3.4 二元特徵降維技術 43](#_Toc120735435)

[4 第四章 案例研討 46](#_Toc120735436)

[4.1 連續二元分類資料測試 46](#_Toc120735437)

[4.1.1 資料集簡介與實驗架構 46](#_Toc120735438)

[4.1.2 不同連續資料集之下的測試與實驗 49](#_Toc120735439)

[4.1.3 分類結果評比與歸納 53](#_Toc120735440)

[4.2 UCI資料集 55](#_Toc120735441)

[4.3 Kaggle資料集 56](#_Toc120735442)

[4.3.1 資料集簡介與實驗架構 56](#_Toc120735443)

[4.3.2 分類結果評比與歸納 57](#_Toc120735444)

[5 第五章 結論與建議 58](#_Toc120735445)

[參考文獻列表 59](#_Toc120735446)

[附錄 A (如果有) 61](#_Toc120735447)

# 圖目錄

[圖 1.1 研究架構。 5](#_Toc120735448)

[圖 2.1 資料預處理常見步驟 (García et al., 2015)。 6](#_Toc120735449)

[圖 2.2 依據探索式資料分析進行資料視覺化 (Behrens, 1997)。 7](#_Toc120735450)

[圖 2.3 獨熱、二進位編碼後的特徵數量比較。 12](#_Toc120735451)

[圖 2.4 維度個數變化對於分類模型表現的影響 (Spruyt, 2014)。 15](#_Toc120735452)

[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](#_Toc120735453)

[圖 2.6 超球體體積對應維度變化 (Köppen, 2000)。 16](#_Toc120735454)

[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](#_Toc120735455)

[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](#_Toc120735456)

[圖 2.9 降維處理的階層化架構 (Tang et al., 2014)。 19](#_Toc120735457)

[圖 2.10 特徵選取與整體資料分析流程 (Tang et al., 2014)。 20](#_Toc120735458)

[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](#_Toc120735459)

[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](#_Toc120735460)

[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](#_Toc120735461)

[圖 2.14 以PCA、LLE視覺化Leukaemia資料集 (Hira & Gillies, 2015)。 23](#_Toc120735462)

[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](#_Toc120735463)

[圖 2.16經由PCA將資料投影至主成分座標軸 (Abdi & Williams, 2010)。 24](#_Toc120735464)

[圖 2.17 LightGBM於Flight Delay（左）與LETOR（右）兩資料集中的收斂表現 25](#_Toc120735465)

[圖 3.1 方法流程圖。 28](#_Toc120735466)

[圖 3.2 原始二元特徵資料，以動物園資料為例。 29](#_Toc120735467)

[圖 3.3 原始資料二元特徵間的相關性矩陣。 32](#_Toc120735468)

[圖 3.4 以塊模型進行置換後的二元特徵間的相關性矩陣。 32](#_Toc120735469)

[圖 3.5 群集過後的特徵組。 34](#_Toc120735470)

[圖 3.6 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。 35](#_Toc120735471)

[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，依據類別區分。 35](#_Toc120735472)

[圖 3.8 以二元特徵總和，排序各個群組中的二元特徵。 36](#_Toc120735473)

[圖 3.9 以目標特徵純粹度，排序各個群組中的二元特徵。 37](#_Toc120735474)

[圖 3.10 各項二元特徵於預訓練分類模型中的特徵重要性。 38](#_Toc120735475)

[圖 3.11 以特徵重要度，排序各個群組中的二元特徵。 38](#_Toc120735476)

[圖 3.12 經過特徵純粹度排序的第三特徵組。 40](#_Toc120735477)

[圖 3.13 BCD與Rank BCD對於新編碼後的特徵分布比較。 42](#_Toc120735478)

[圖 3.14 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。 43](#_Toc120735479)

[圖 3.15 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。 44](#_Toc120735480)

[圖 3.16 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。 44](#_Toc120735481)

[圖 3.17 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。 45](#_Toc120735482)

[圖 4.1 模擬的連續二元分類資料。 46](#_Toc120735483)

[圖 4.2 連續資料集下的實驗架構。 47](#_Toc120735484)

[圖 4.3 原始連續資料於X、Y、Z三維度上的分布情形。 48](#_Toc120735485)

[圖 4.4 二元化後的連續資料，共劃分為30個二元特徵。 48](#_Toc120735486)

[圖 4.5 連續資料集一的資料分布。 50](#_Toc120735487)

[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。 50](#_Toc120735488)

[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。 50](#_Toc120735489)

[圖 4.8 連續資料集二的資料分布。 51](#_Toc120735490)

[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。 51](#_Toc120735491)

[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。 51](#_Toc120735492)

[圖 4.11 連續資料集三的資料分布。 52](#_Toc120735493)

[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。 52](#_Toc120735494)

[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。 52](#_Toc120735495)

# 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](#_Toc120735496)

[表 1.2 常見於製造業中的在製品製程紀錄。 2](#_Toc120735497)

[表 2.1 不同變數類別的定義與描述 (Stevens, 1946)。 8](#_Toc120735498)

[表 2.2 變數類別接受運算子與範例。 9](#_Toc120735499)

[表 2.3 不同編碼方式所對應的模型準確度 (Potdar et al., 2017)。 9](#_Toc120735500)

[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](#_Toc120735501)

[表 2.5 目標編碼後的特徵欄位，以水果價格為例。 14](#_Toc120735502)

[表 3.1 變數與符號定義。 27](#_Toc120735503)

[表 3.2 具有群組資訊的二元特徵資料。 30](#_Toc120735504)

[表 3.3 缺乏群組資訊的二元特徵資料。 30](#_Toc120735505)

[表 3.4 不同主成分之下的二元特徵權重絕對值。 31](#_Toc120735506)

[表 3.5 依據主成分分析群集二元特徵。 31](#_Toc120735507)

[表 3.6 依據相關係數群集二元特徵。 33](#_Toc120735508)

[表 3.7排序前各群組中的二元特徵，由表示。 39](#_Toc120735509)

[表 3.8 染色體範例，以基因演算法排序組內特徵。 39](#_Toc120735510)

[表 3.9 常用的BCD編碼方式，與對應的十位數值。 40](#_Toc120735511)

[表 3.10 各個樣本轉換後的新數值。 41](#_Toc120735512)

[表 4.1 二元化後的二元特徵資料。 49](#_Toc120735513)

[表 4.2 類別化後的類別特徵資料。 49](#_Toc120735514)

# 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。

## 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 |

## 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。

1. 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。

1. 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。

1. 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。

## 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，並歸納出未來本研究的後續發展方向。



圖 1.1 研究架構。

# 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。

## 變數編碼

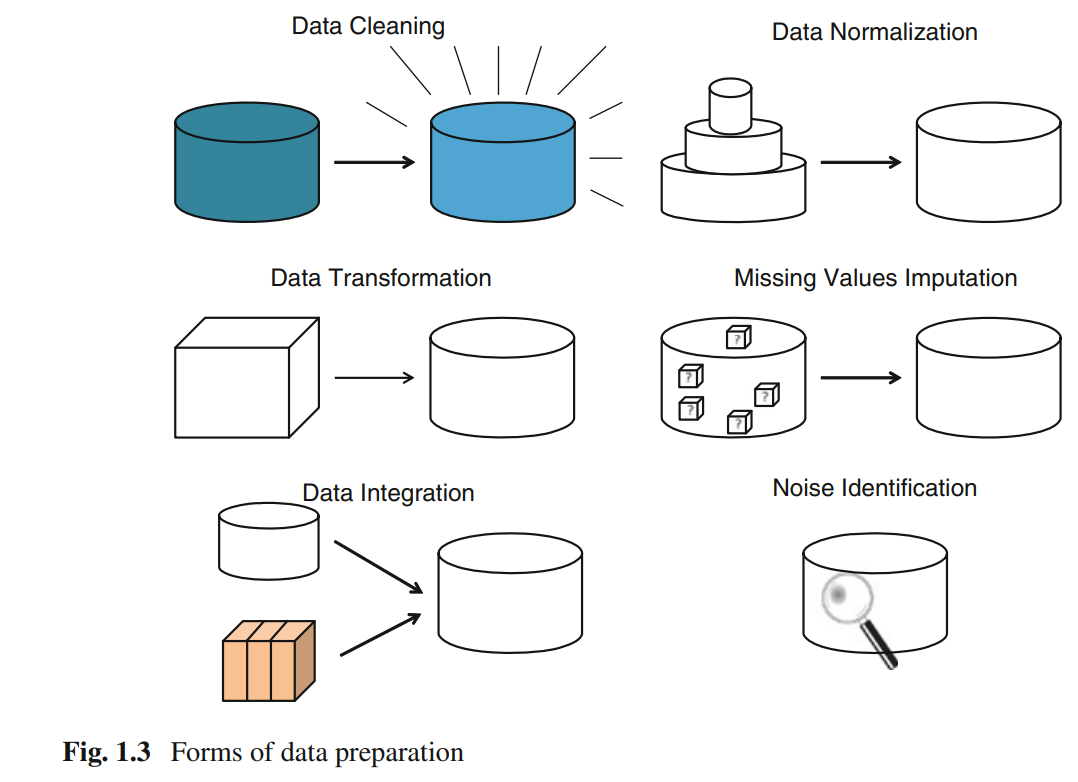


圖 2.1 資料預處理常見步驟 (García et al., 2015)。

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。

1. 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。

1. 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。

1. 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。

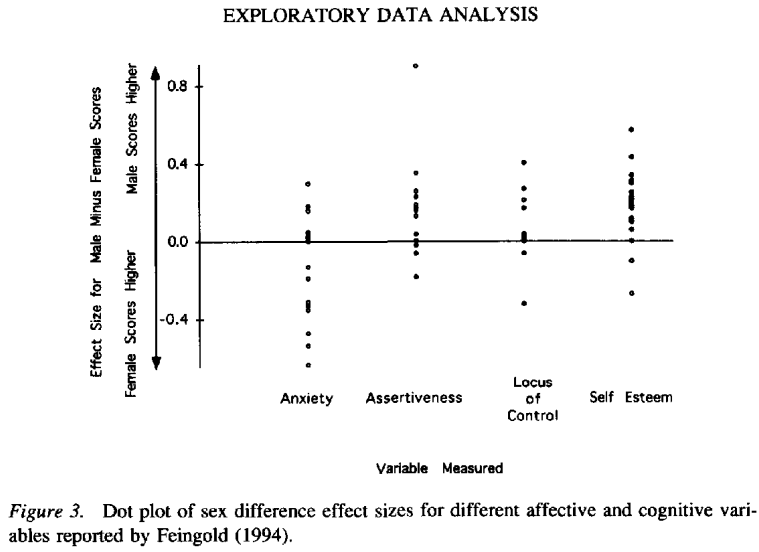


圖 2.2 依據探索式資料分析進行資料視覺化 (Behrens, 1997)。

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 |

### 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具有個類別特徵，每一類別特徵由表示，且；經由頻率編碼後的新數值資料具有*m*個數值特徵，每一數值特徵以表示，且；由0到之間的正整數組成，用以描述原先的中的種類別：

( 2.1 )

( 2.2 )

( 2.3 )

( 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。

### 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具有個類別特徵，每一類別特徵由表示，且；經由獨熱編碼後的新數值資料具有*m*個數值特徵，由複數個向量所組成；其中各個向量內僅包含零與一、且總和等於一。

( 2.5 )

( 2.6 )

( 2.7 )

( 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。

### 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，具有個種類的情形下，獨熱編碼需要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具有個類別特徵，每一類別特徵由表示，且；經由獨熱編碼後的新數值資料具有*m*個數值特徵，由複數個向量所組成；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。

( 2.9 )

( 2.10 )

( 2.11 )

( 2.12 )

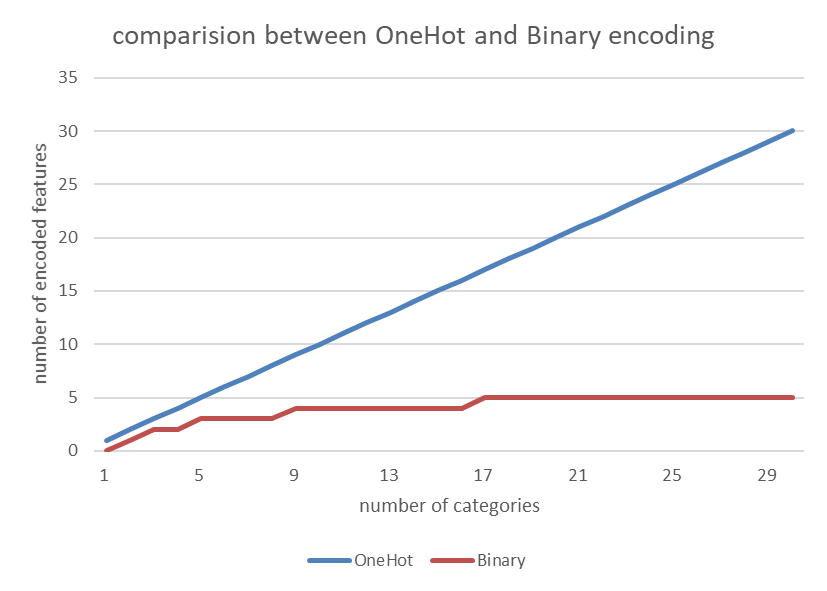


圖 2.3 獨熱、二進位編碼後的特徵數量比較。

### 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具有個類別特徵，每一類別特徵由表示，且；經由頻率編碼後的新數值資料具有*m*個數值特徵，每一數值特徵以表示，且。

( 2.13 )

( 2.14 )

( 2.15 )

( 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 |

### 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（Label）；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具有個類別特徵，每一類別特徵由表示，且；經由目標編碼後的新數值資料具有*m*個數值特徵，每一數值特徵以表示，且。

( 2.17 )

( 2.18 )

( 2.19 )

( 2.20 )

在中，則以各個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 |

## 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。

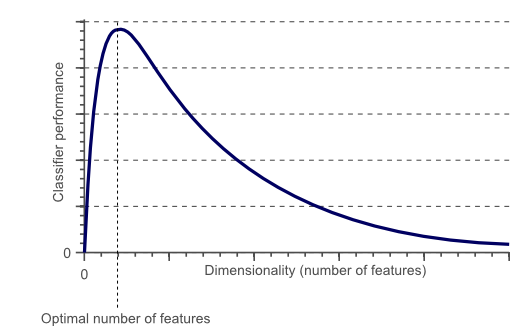


圖 2.4 維度個數變化對於分類模型表現的影響 (Spruyt, 2014)。

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和訓練資料。



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。

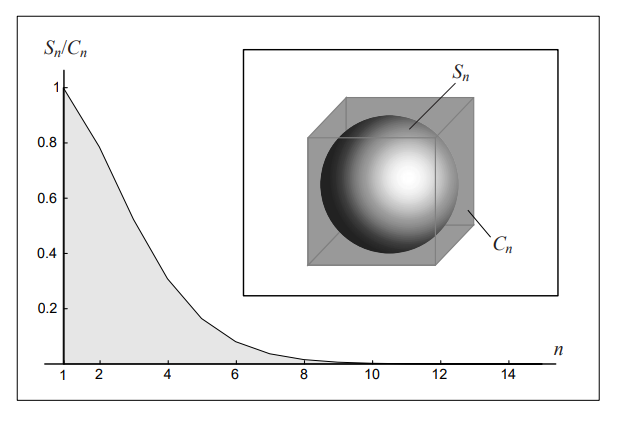


圖 2.6 超球體體積對應維度變化 (Köppen, 2000)。

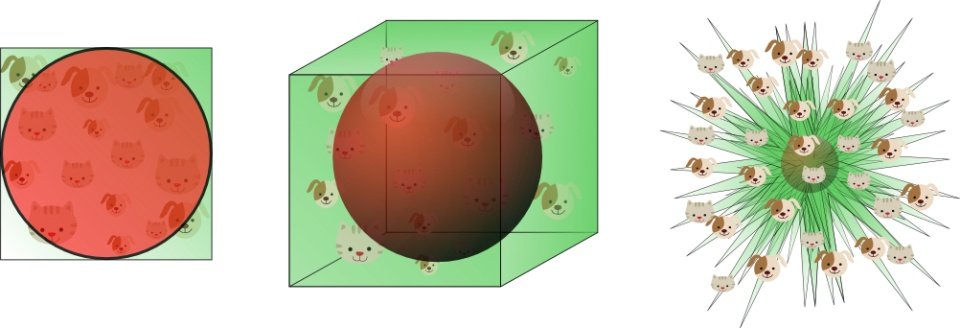


圖 2.7 資料分佈情形對應維度變化，以貓狗分類為例 (Spruyt, 2014)。

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

( 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。

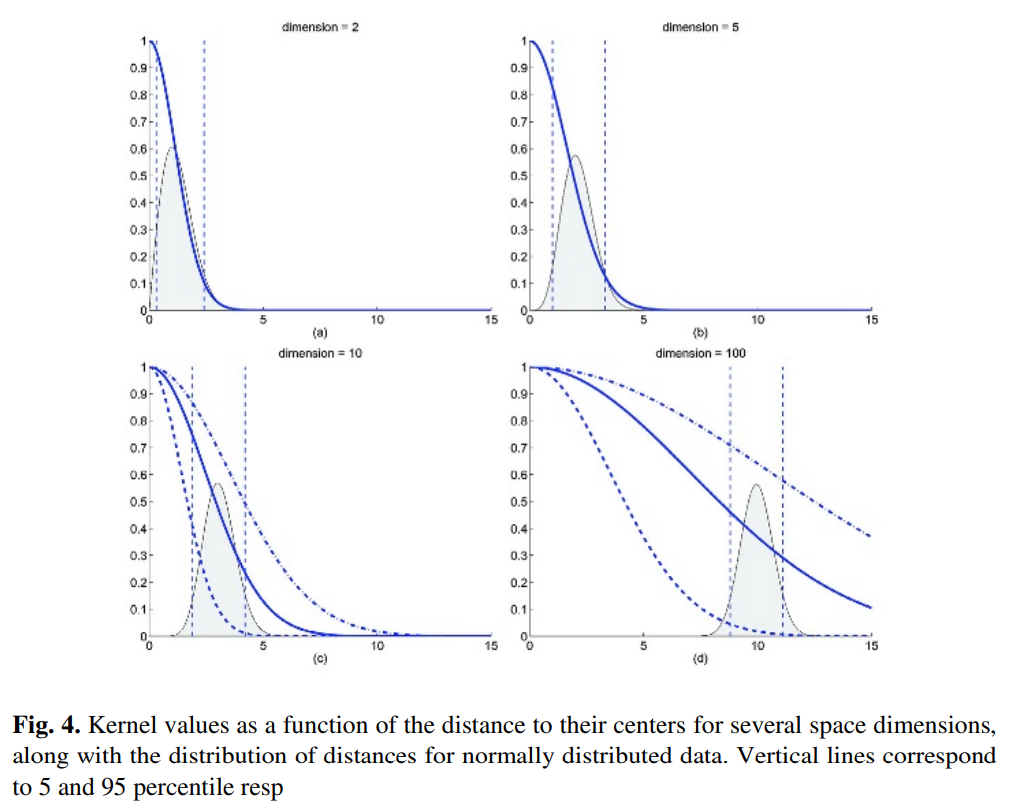


圖 2.8 高斯核函數值對應距離分布於高維度空間的變化 (Verleysen & François, 2005)。

然而，隨著近年來資料分析技術與接器學習的盛行，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。

## 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。

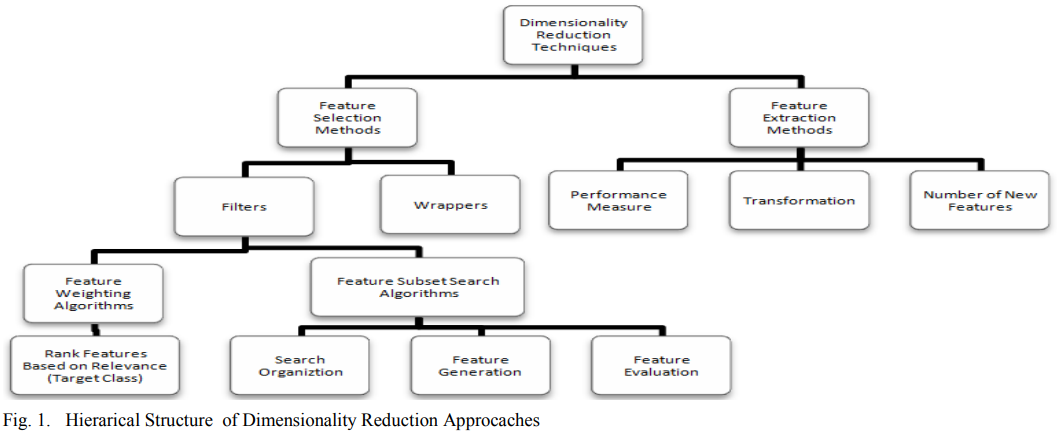


圖 2.9 降維處理的階層化架構 (Tang et al., 2014)。

### 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。

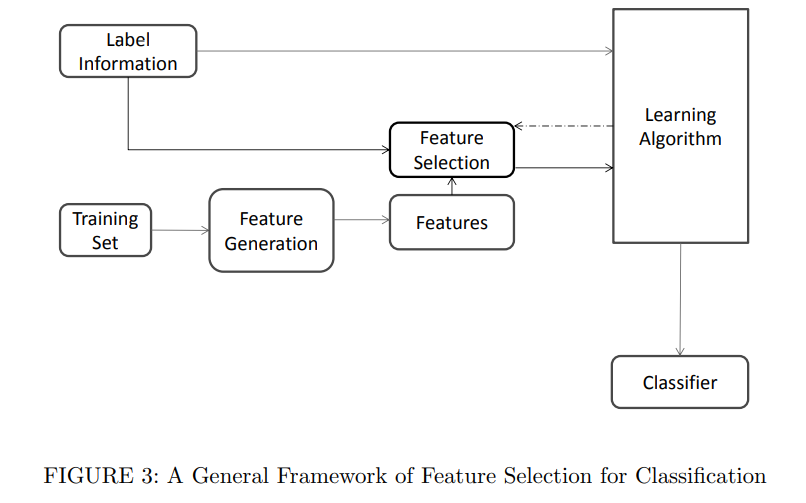


圖 2.10 特徵選取與整體資料分析流程 (Tang et al., 2014)。

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。

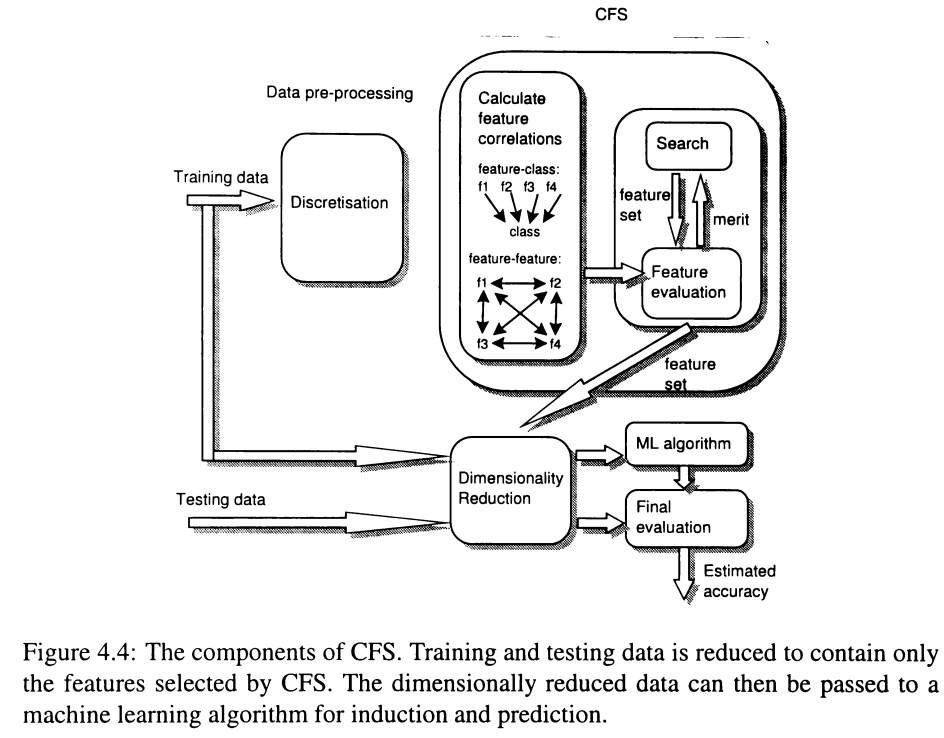


圖 2.11 基於相關係數的過濾型特徵選擇，結合機器學習流程 (Hall, 1999)。

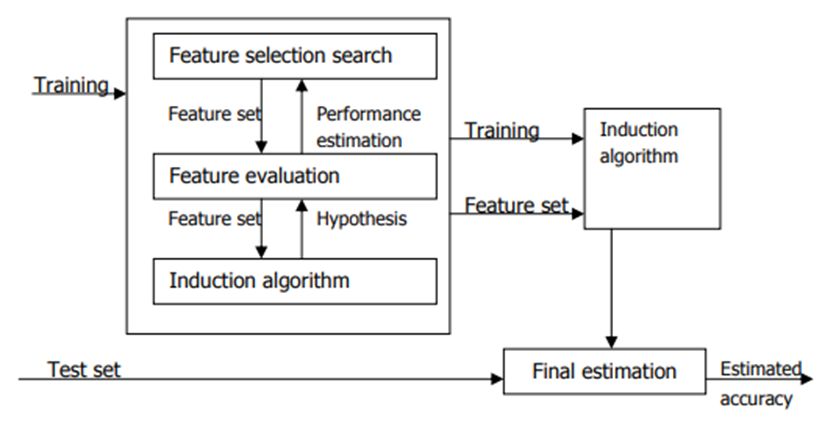


圖 2.12 基於相關係數的包裝型特徵選擇，結合機器學習流程 (Hall & Smith, 1999)。

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。

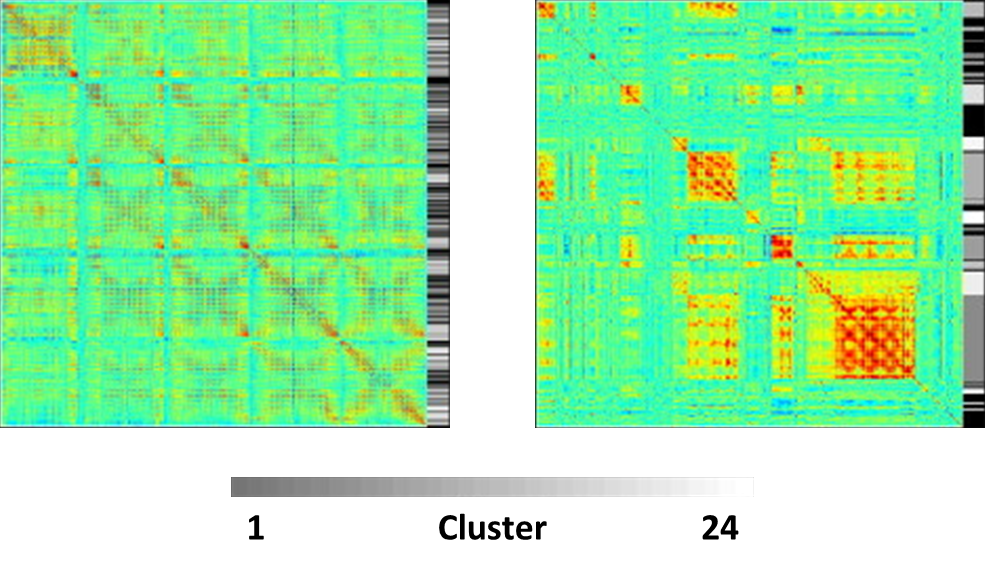


圖 2.13 以階層群集重新排序特徵之相關係數矩陣比較 (Liu et al., 2012)。

### 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、避免過度擬和等目的；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

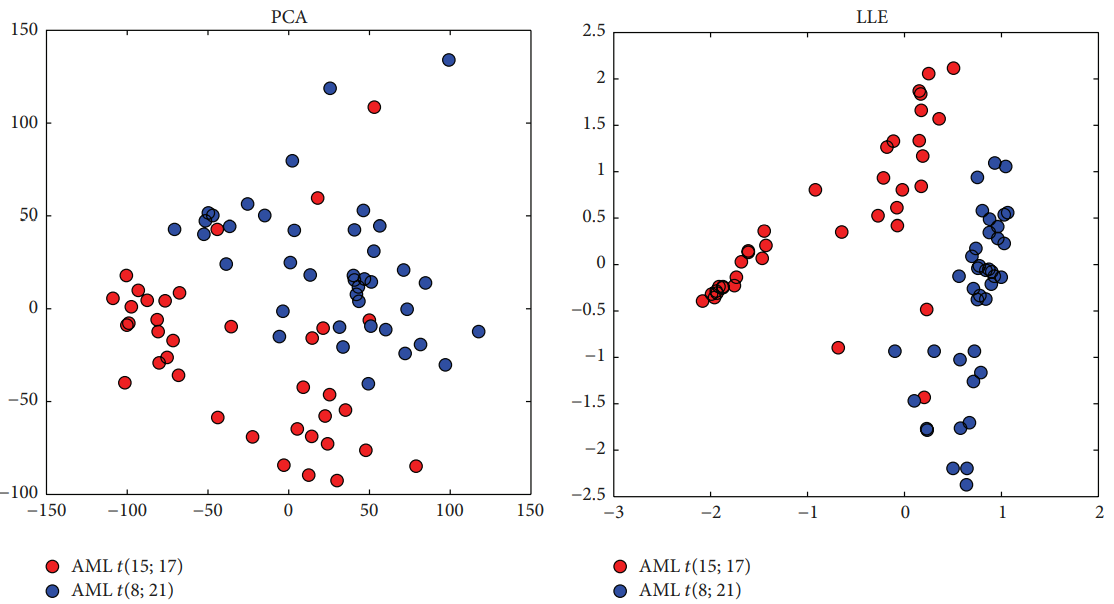


圖 2.14 以PCA、LLE視覺化Leukaemia資料集 (Hira & Gillies, 2015)。

主成分分析（PCA）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)。

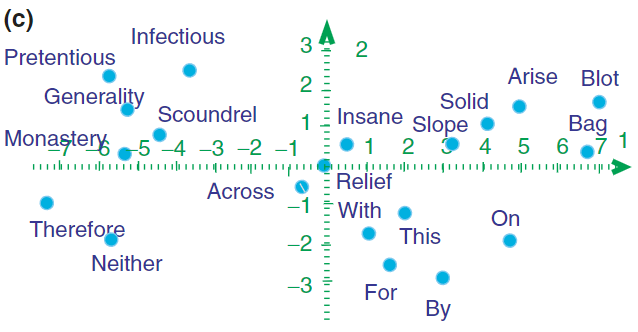


圖 2.16經由PCA將資料投影至主成分座標軸 (Abdi & Williams, 2010)。

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

## LightGBM分類模型

LightGBM為一種輕量化的梯度提升機（GBM, Gradient Boosting Machine），由Ke et al. (2017)與其微軟團隊所提出，支援梯度提升決策樹（GBDT, Gradient Boosting Decision Tree）演算法的一種框架，能以梯度提升的方式調整、訓練決策樹分類模型。梯度提升決策樹演算法乃透過決策樹做為弱分類器基底，經由不斷訓練並優化弱分類器後產生最佳的分類模型；目前已被普遍使用於各式分類模型與分類任務之中，像是XGBoost (Chen & Guestrin, 2016)、CatBoost (Dorogush et al., 2018) 皆是支援GBDT的梯度提升機。

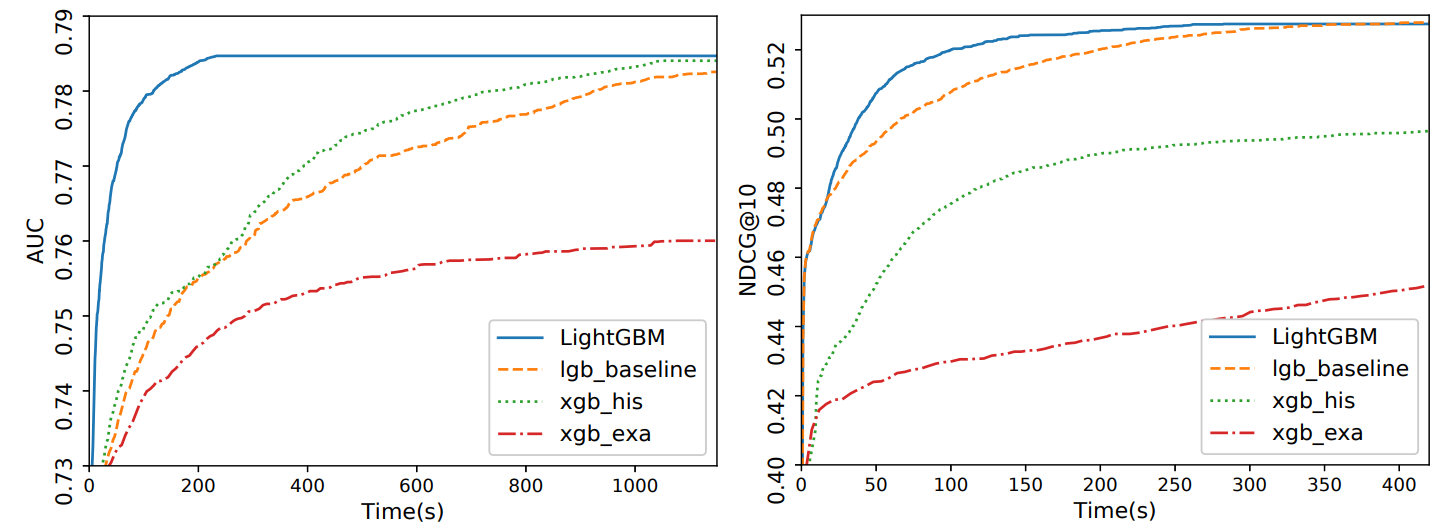


圖 2.17 LightGBM於Flight Delay（左）與LETOR（右）兩資料集中的收斂表現

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；如上圖，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。

## 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。傳統的評估標準

# 第三章 高維二元特徵之聚合編碼技術

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。

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 3.1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、以及群組內二元特徵的排序方式，來使得新產生出的數值資料有利於機器學習模型的分類任務。

表 3.1 變數與符號定義。

|  |  |
| --- | --- |
| 符號 | 定義 |
|  | 原始的二元資料，共含有個二元特徵， |
|  | 對二元特徵進行分群，產生個 |
|  | 二元特徵群組個數 |
|  | 第*j*個二元特徵群組， |
|  | 群組內二元特徵的排序方式，即調整內部的排列 |
|  | 第*j*個二元特徵群組中的特徵排列方式， |
|  | 第*j*個二元特徵群組的BCD數值 |
|  | 編碼後的數值資料，共含有個數值特徵， |
|  | 還原後的類別資料，共含有個類別特徵， |

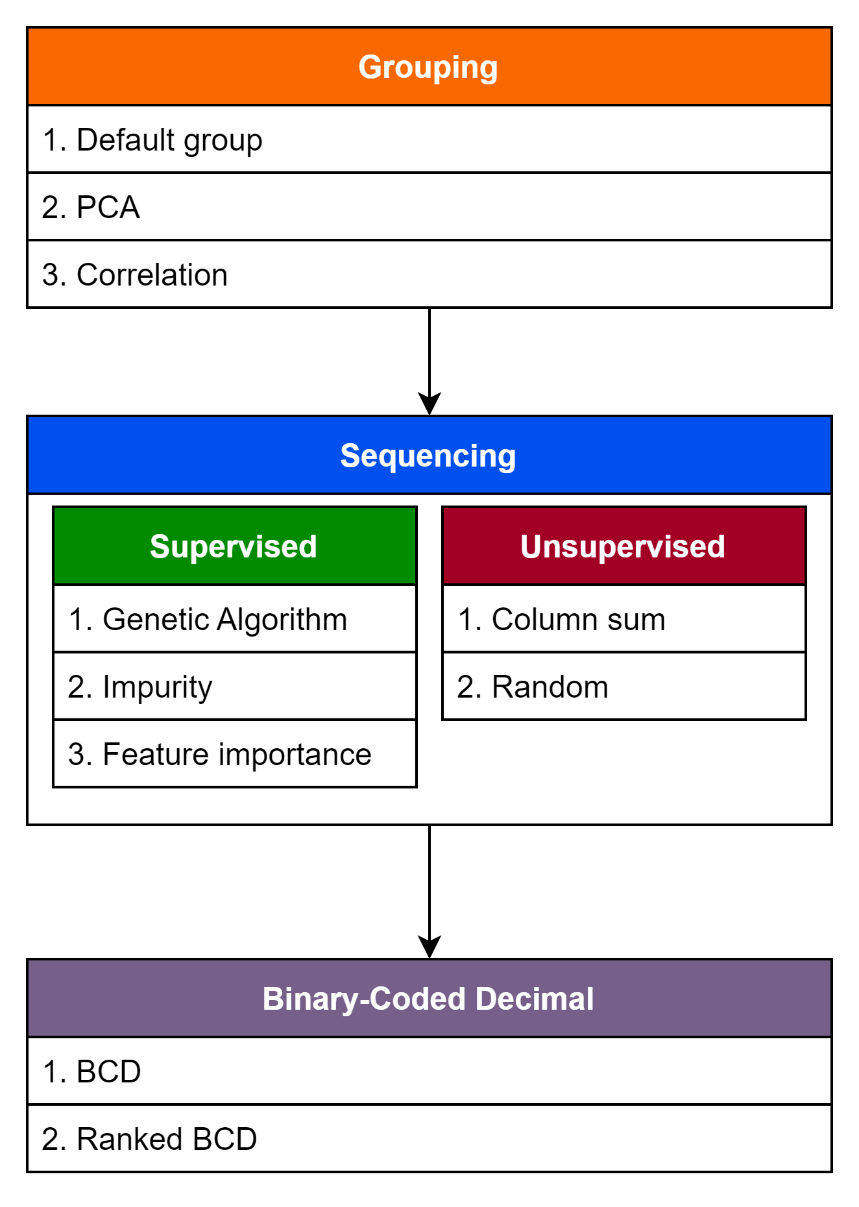


圖 3.1 方法流程圖。

## 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 3.2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。

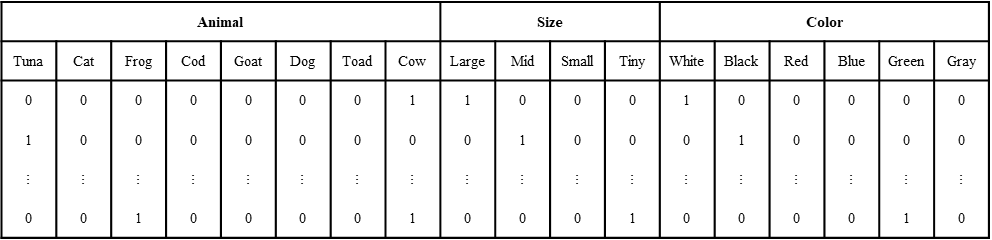


圖 3.2 原始二元特徵資料，以動物園資料為例。

### 資料原始特徵群

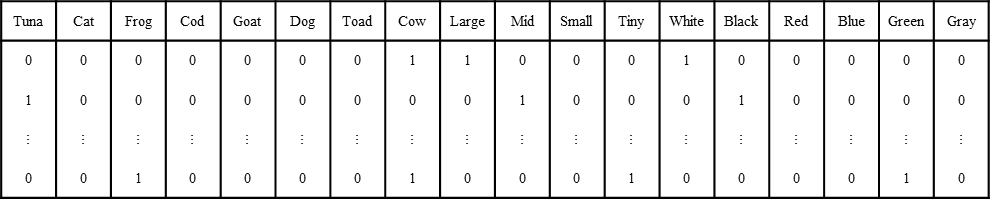
若是原始的二元特徵資料是由獨熱編碼所產生，且具有特徵之間的分群的資訊，屬於具有二元特徵群組資訊的資料時，則可以使用該群組資訊將二元特徵進行群組。如下表，表示了經由原始二元特徵的群組知識（特徵區分成「動物」、「尺寸」、「顏色」三個群組），群組後的二元特徵。

表 3.2 具有群組資訊的二元特徵資料。



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。

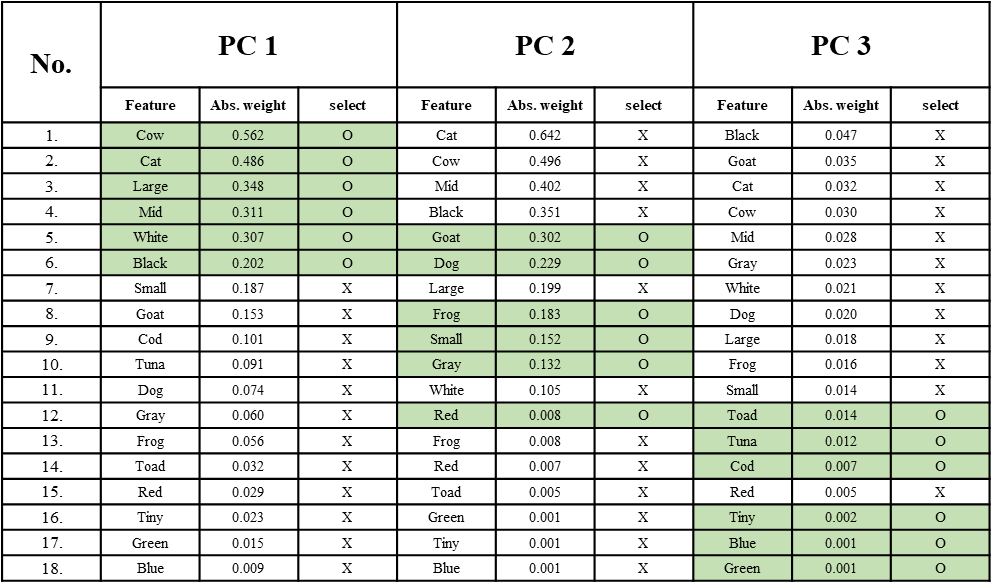
表 3.3 缺乏群組資訊的二元特徵資料。



### 主成分分析群集

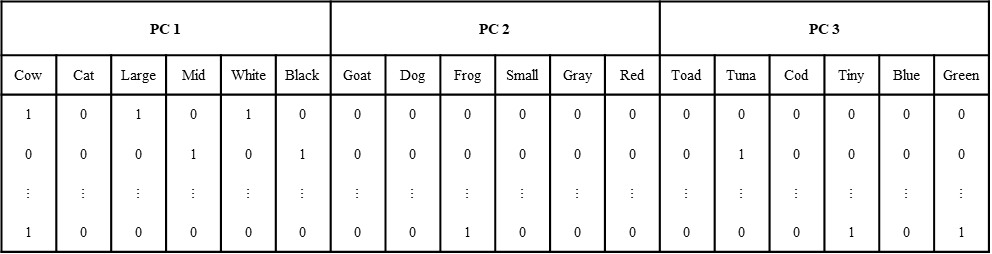
透過主成分分析（Principle Component Analysis），可以將最能解釋整體資料變異、最具資訊的二元特徵分於相同群組之內，如此使新產生的數值特徵融合了最多具有變異資訊的二元特徵。透過決定群組個數，再根據主成分的順序，依據權重的絕對值依序進行選取，將解釋資料變異度相近的二元特徵聚集於同一群組之中。

表 3.4 不同主成分之下的二元特徵權重絕對值。



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。



### 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）、或是塊建模（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。

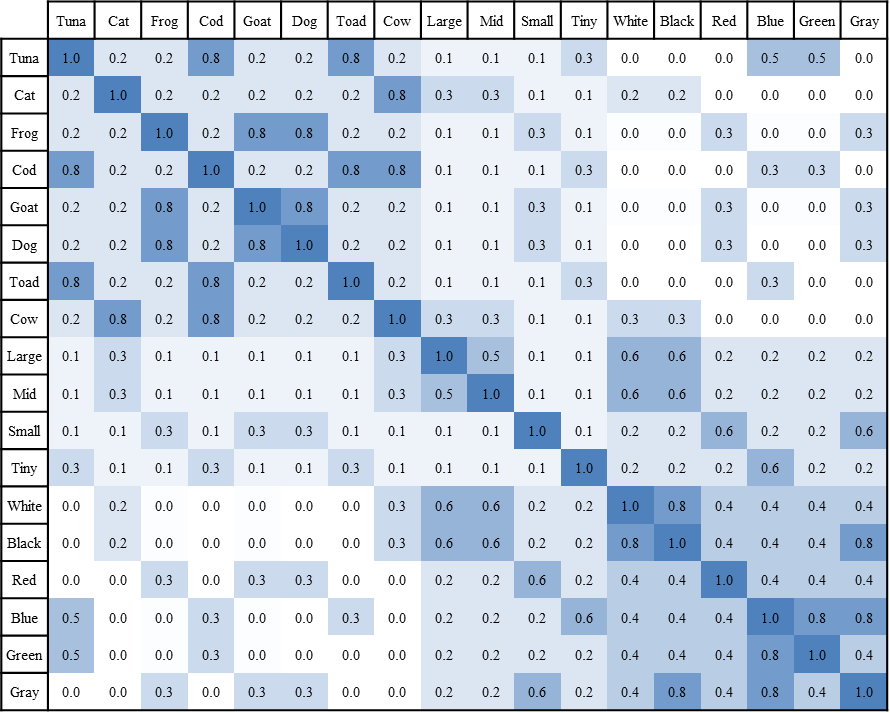


圖 3.3 原始資料二元特徵間的相關性矩陣。

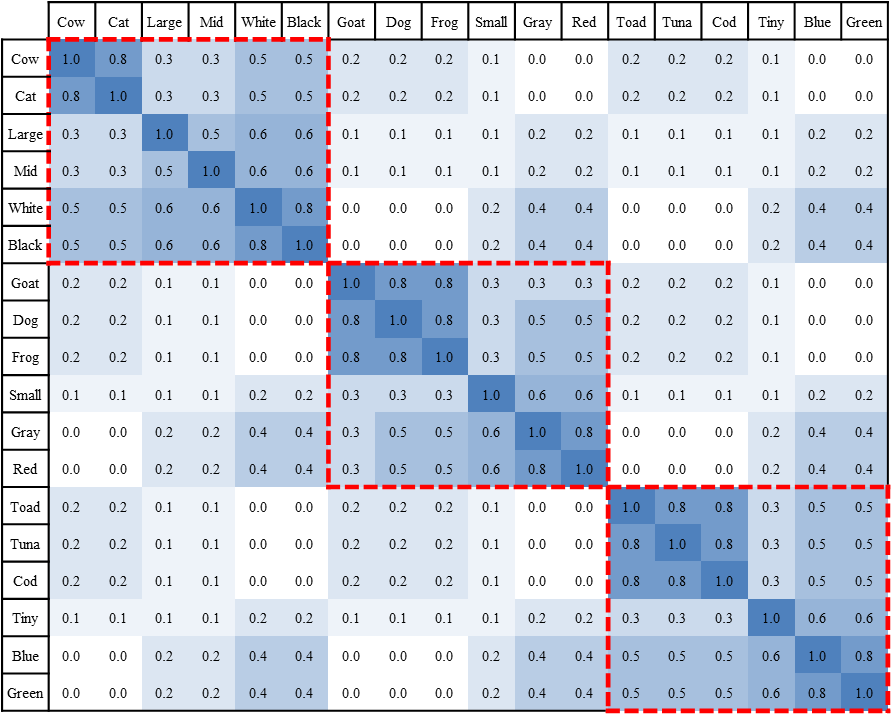


圖 3.4 以塊模型進行置換後的二元特徵間的相關性矩陣。

排序前的原始資料圖 3.3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 3.4的新相關係性矩陣。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。



## .群內二元特徵排序



圖 3.5 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。

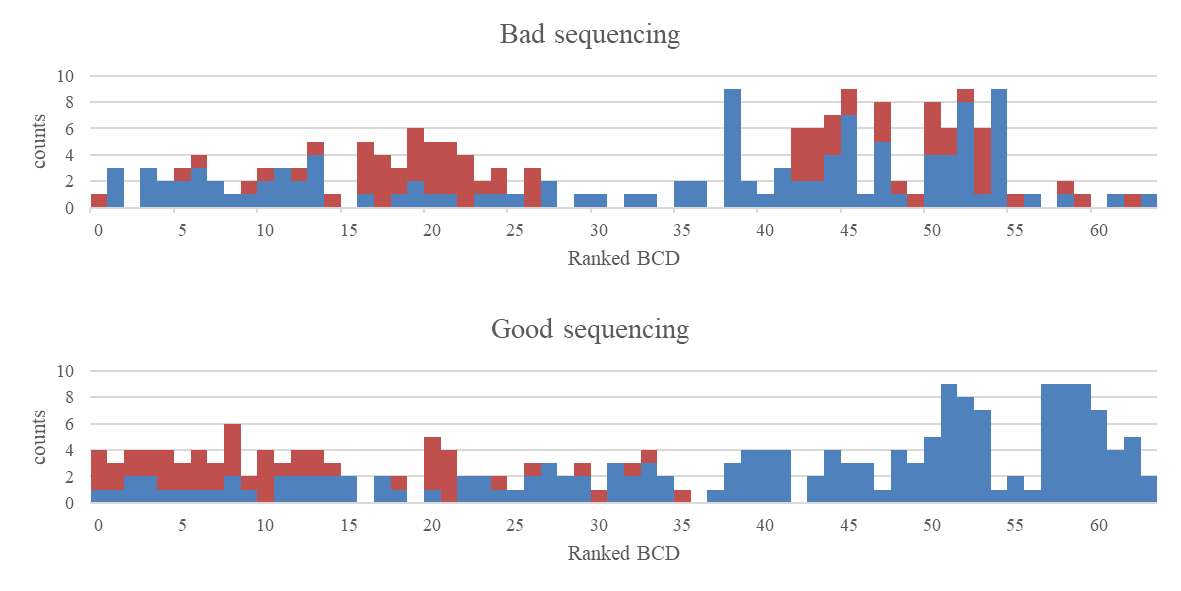


圖 3.6 以不同方式排序二元特徵，產生的新數值資料分佈比較，依據新數值特徵分佈。

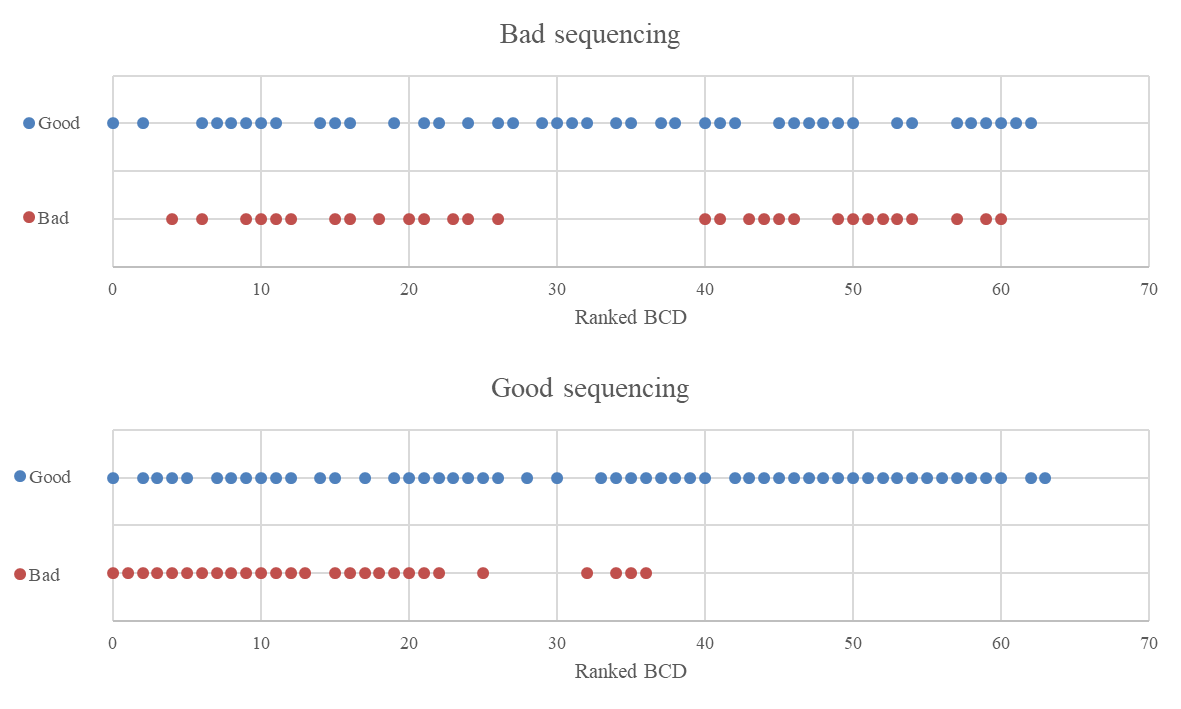


圖 3.7 以不同方式排序二元特徵，產生的新數值資料分佈比較，依據類別區分。

圖 3.6、圖 3.7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。

### 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的該類別的出現頻率，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。

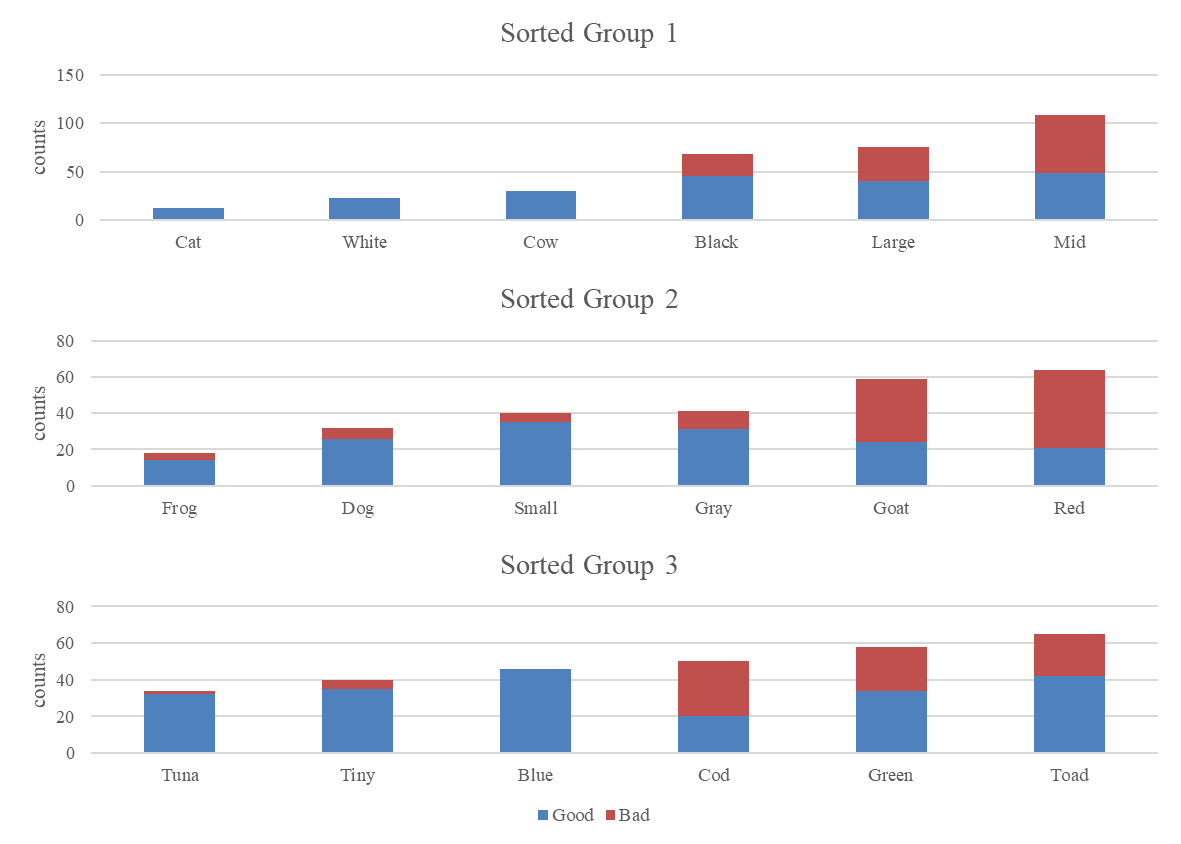


圖 3.8 以二元特徵總和，排序各個群組中的二元特徵。

### 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。

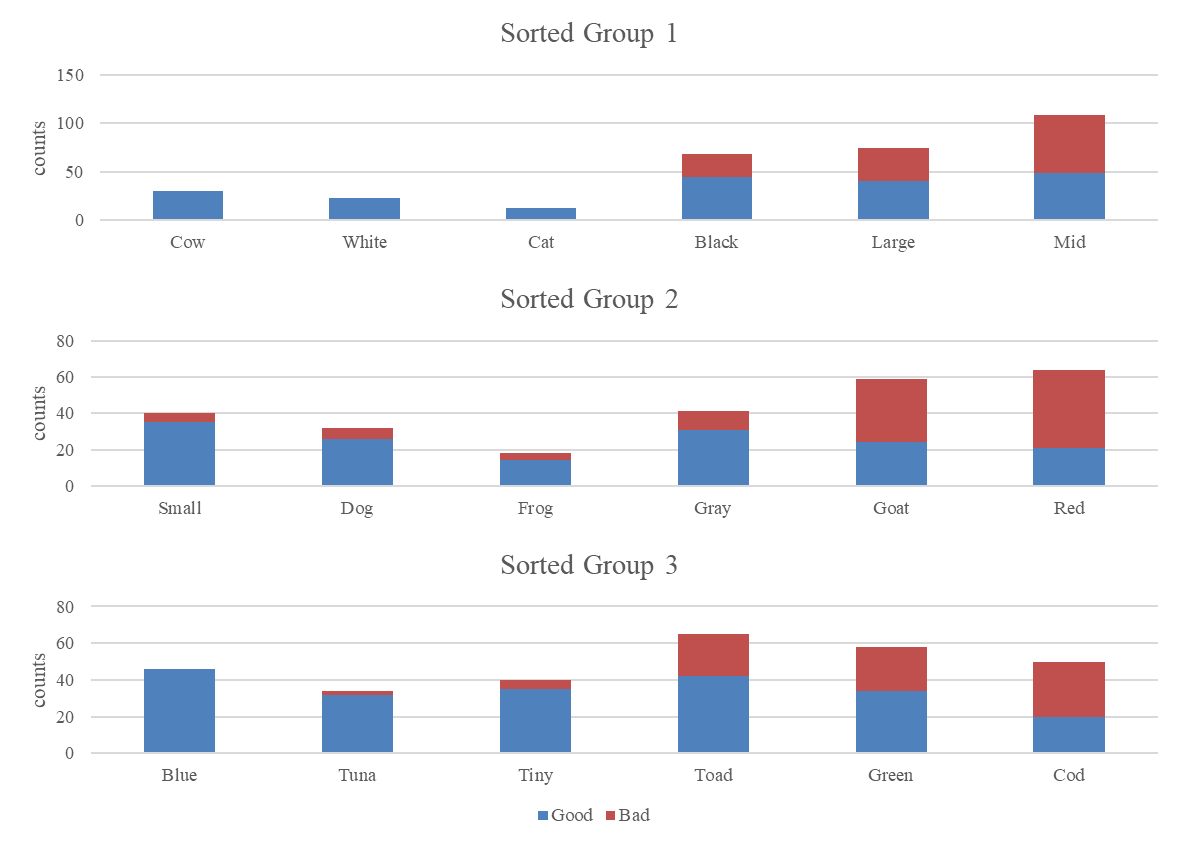


圖 3.9 以目標特徵純粹度，排序各個群組中的二元特徵。

### 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 3.10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 3.11為根據特徵重要度排序各組二元特徵的成果。

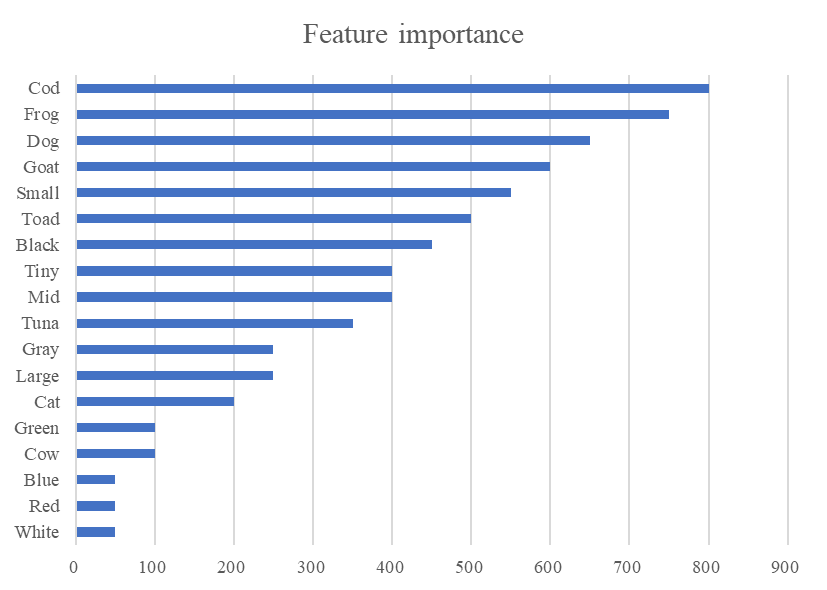


圖 3.10 各項二元特徵於預訓練分類模型中的特徵重要性。



圖 3.11 以特徵重要度，排序各個群組中的二元特徵。

### 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，表示了群組後、排序前的二元特徵，其中表示特徵所歸屬的群組、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，由表示。

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group 1 | | | | | | Group 2 | | | | | | Group 3 | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

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 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Chromosome 2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Chromosome 3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ⁝ | ⁝ | | | | | | ⁝ | | | | | | ⁝ | | | | | |

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。

( 3.1 )

( 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。

## 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。



### 二進位十位數編碼數值

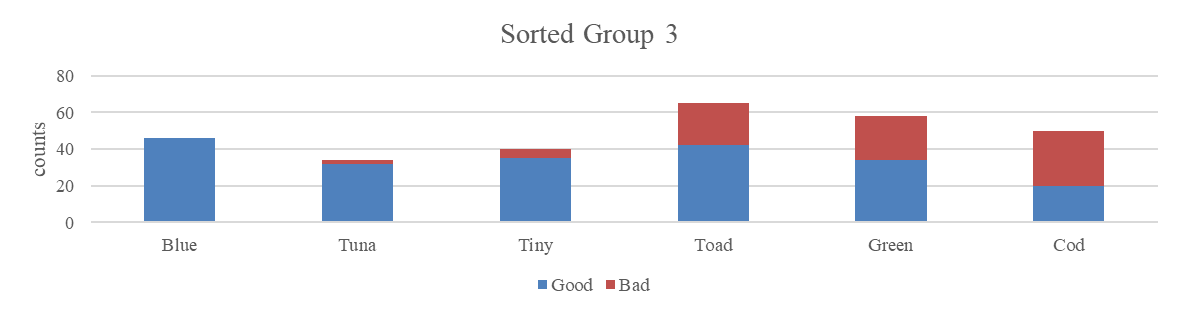
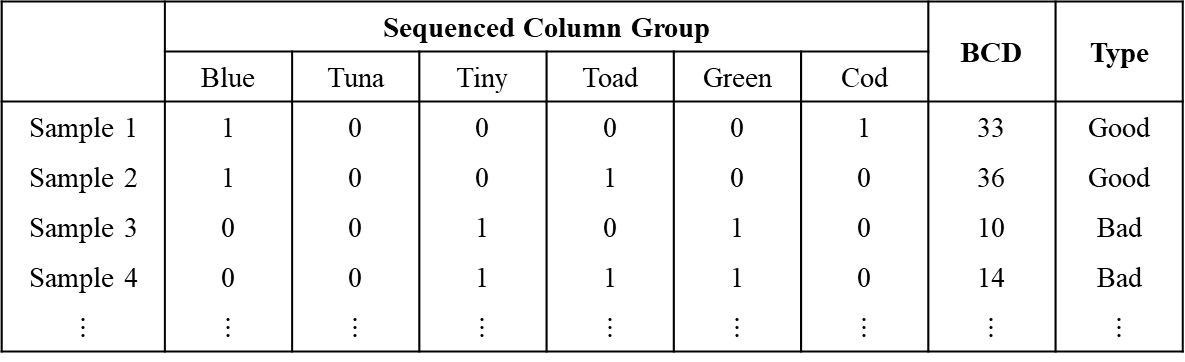


圖 3.12 經過特徵純粹度排序的第三特徵組。

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 3.12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，如如下表所表示。

表 3.10 各個樣本轉換後的新數值。



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。

### 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 3.13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。

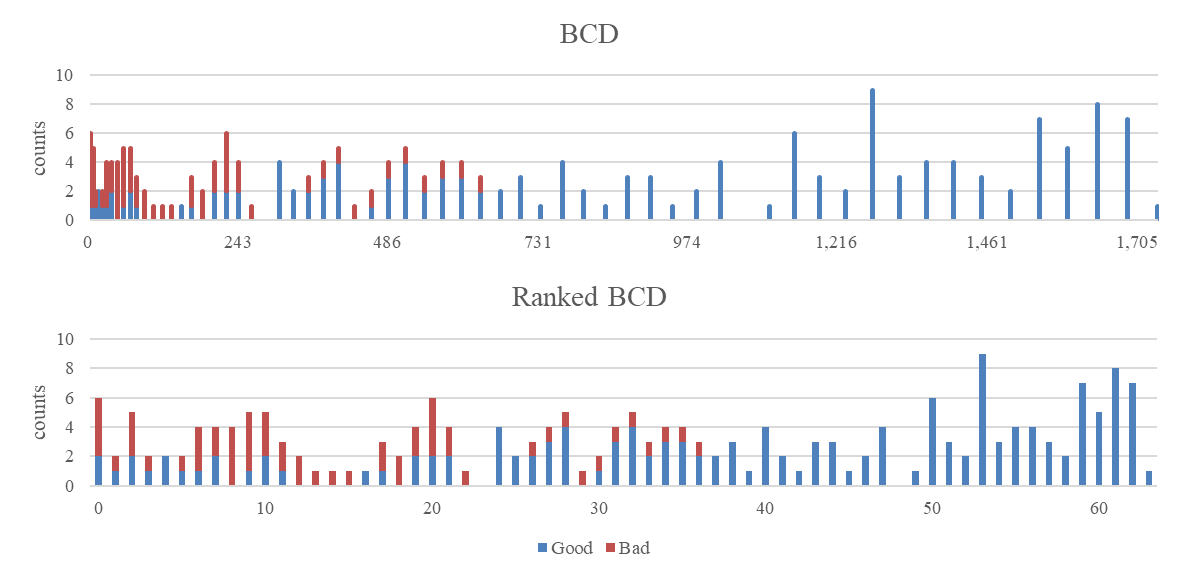


圖 3.13 BCD與Rank BCD對於新編碼後的特徵分布比較。

## 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，便能將原先總共為維度的二元資料，轉變成僅具有維度的數值資料，。依據動物園資料為例，圖 3.5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。

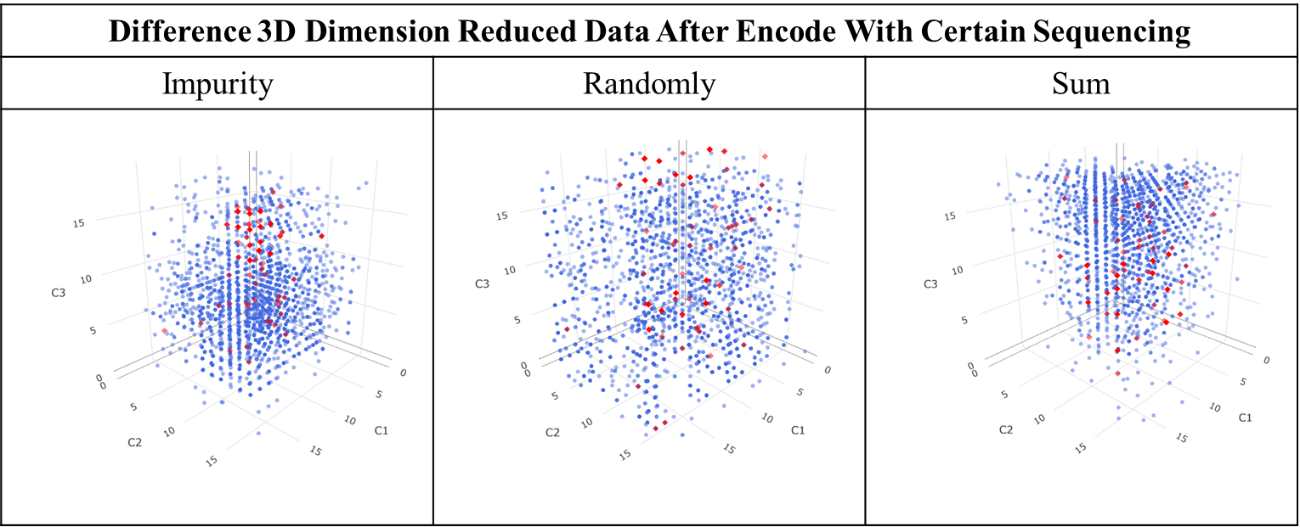


圖 3.14 呈現整體二元資料降維至三維後的資料分佈，依據特徵純粹度、隨機與特徵和排序方式。

如圖 3.14，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 3.15、圖 3.16、圖 3.17中，則呈現降至二維度後的資料分布，同時將如圖 3.6、圖 3.7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。

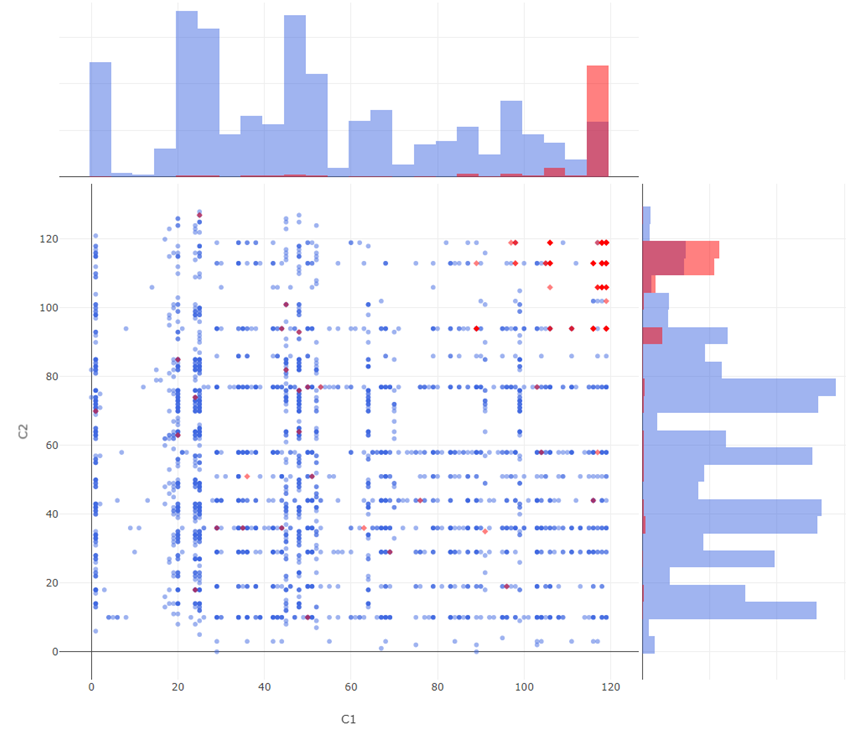


圖 3.15 呈現整體二元資料降維至二維後的資料分佈，依照特徵純粹度排序。

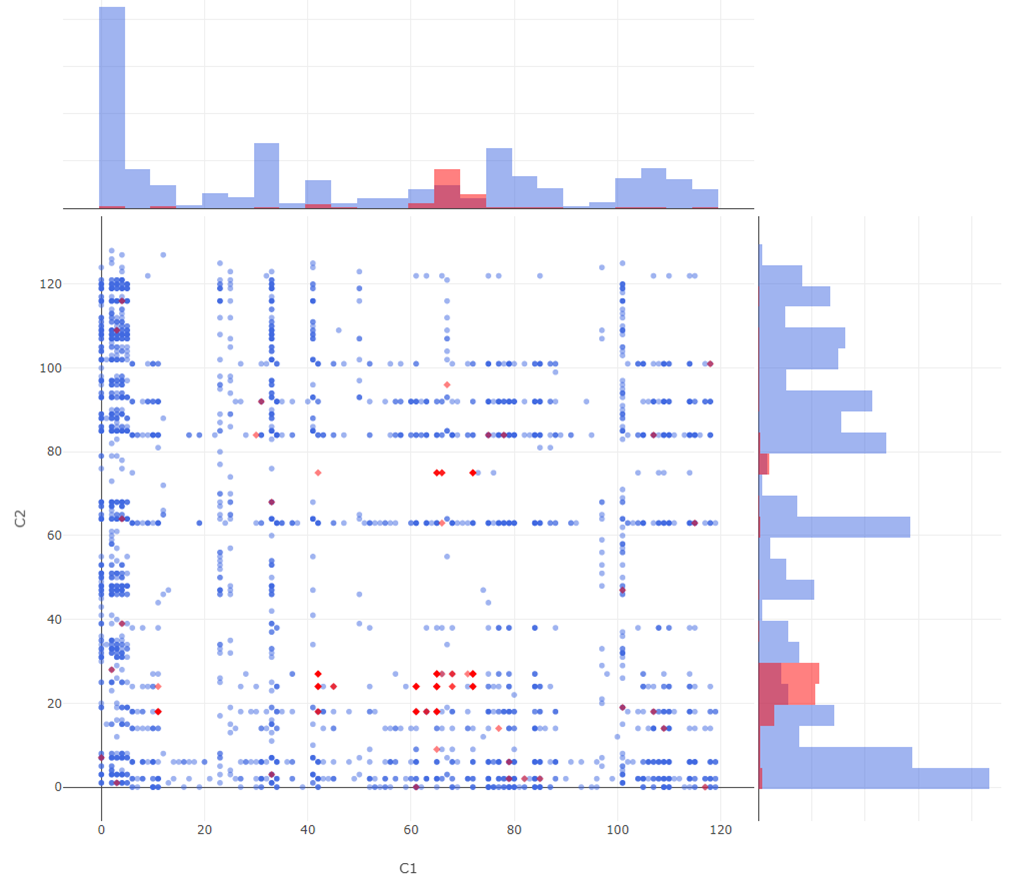


圖 3.16 呈現整體二元資料降維至二維後的資料分佈，依照隨機排序。

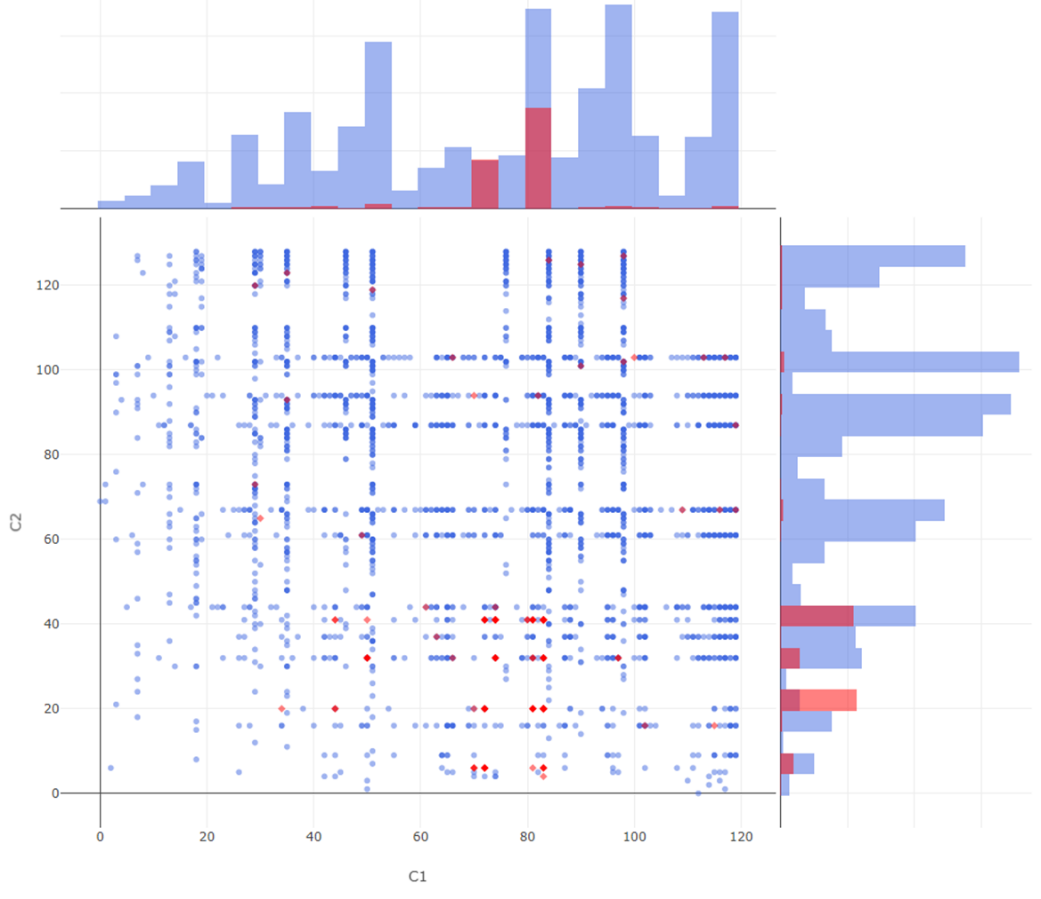


圖 3.17 呈現整體二元資料降維至二維後的資料分佈，依照特徵和排序。

# 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本研究所發展方法所產生的新數值資料，對於機器學習模型分類成效之影響。

## 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300個紅色（少數類別、型二樣本）與3000個藍色（多數類別、型一樣本）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，目的在於增加分類的複雜度。

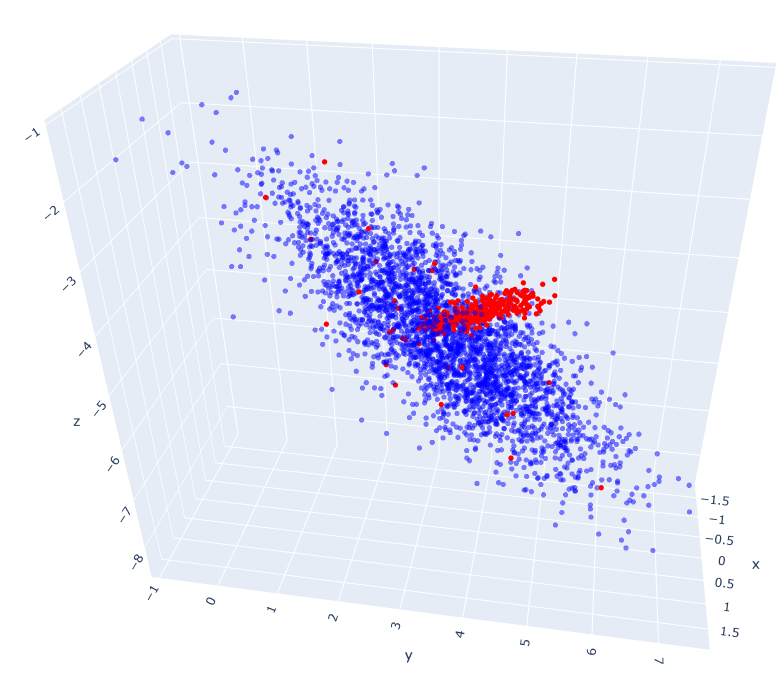


圖 4.1 模擬的連續二元分類資料。

### 資料集簡介與實驗架構

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，以產生作為輸入的二元特徵資料。同時也將二元資料類別化，以作為傳統變數編碼方式的輸入，後比較各種編碼後的資料集的分類成果。

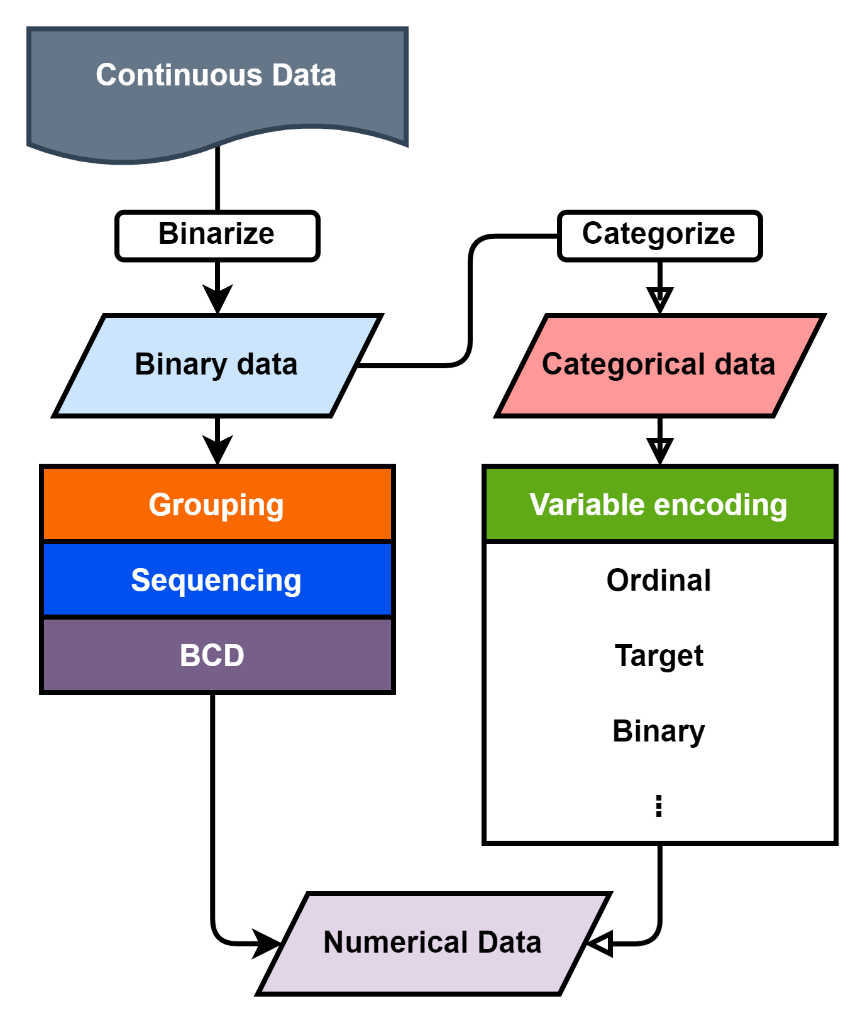


圖 4.2 連續資料集下的實驗架構。

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。

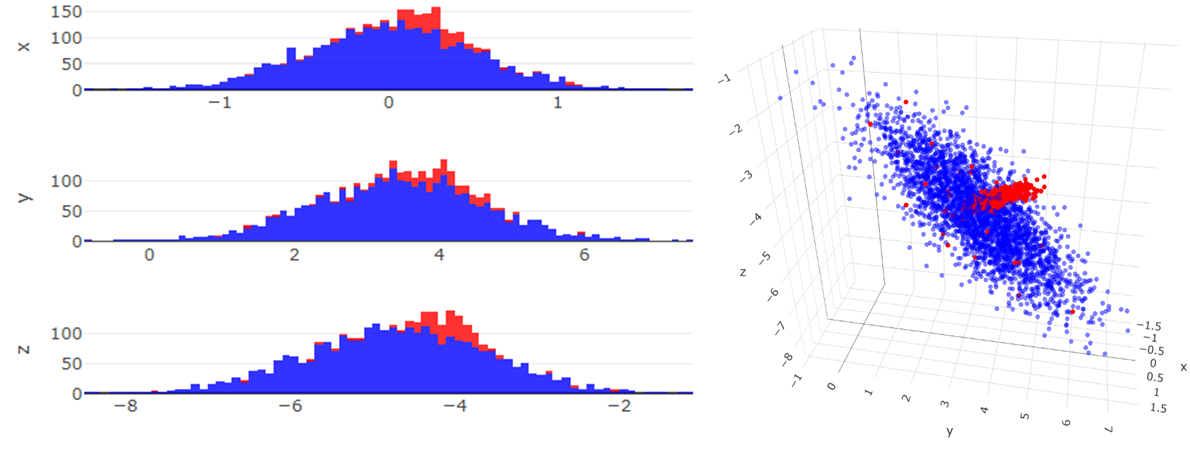


圖 4.3 原始連續資料於X、Y、Z三維度上的分布情形。

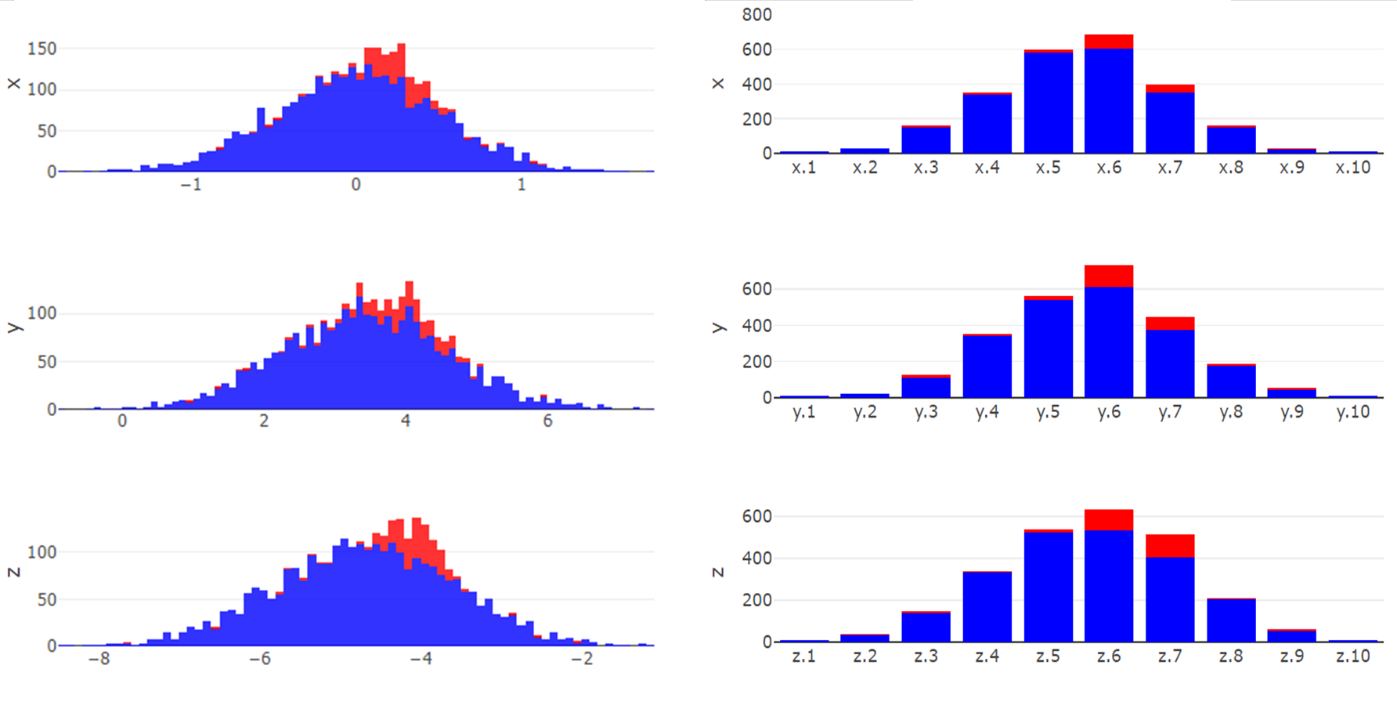


圖 4.4 二元化後的連續資料，共劃分為30個二元特徵。

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。

表 4.1 二元化後的特徵資料。

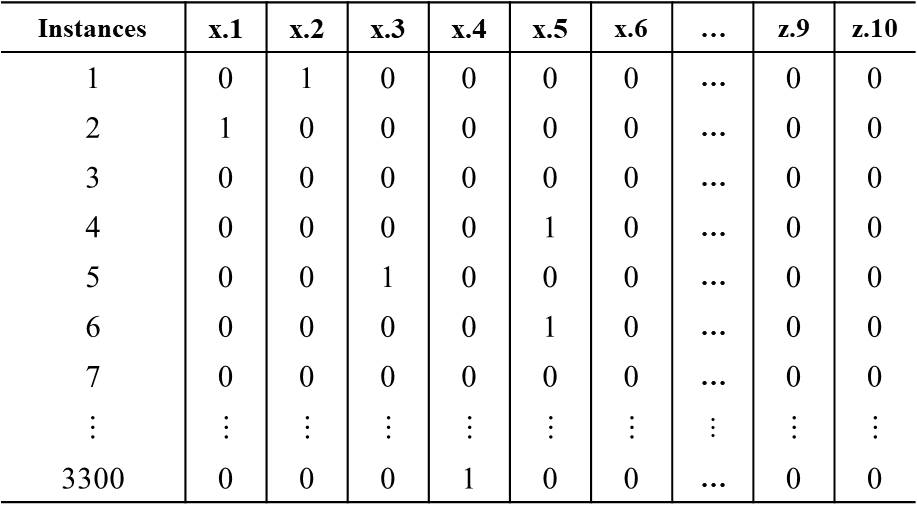
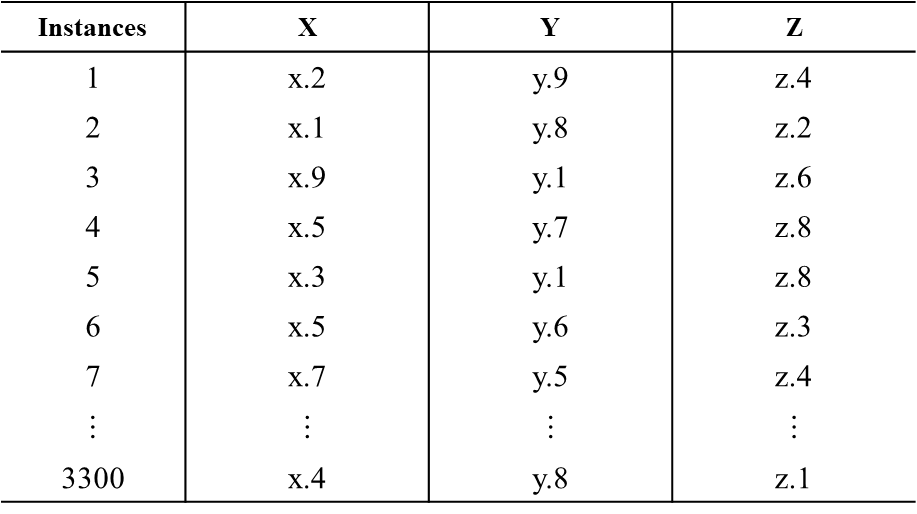


表 4.2 類別化後的特徵資料。



### 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，比較了本研究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，以F1分數作為評比資料及分類難易度的指標。

在群組二元特徵時，本研究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、Correlation、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，依序記為Sum、Purity、FI、RND。

#### 連續資料集一

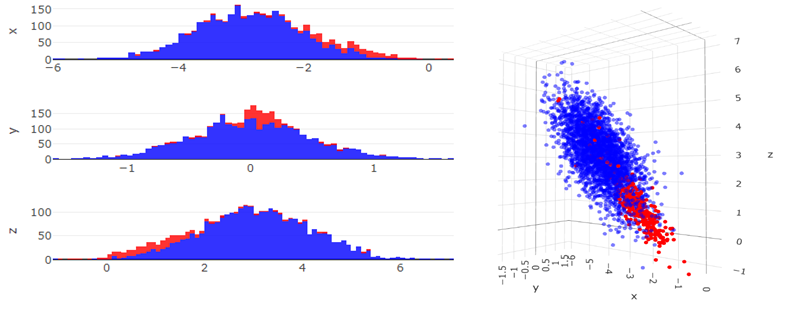


圖 4.5 連續資料集一的資料分布。

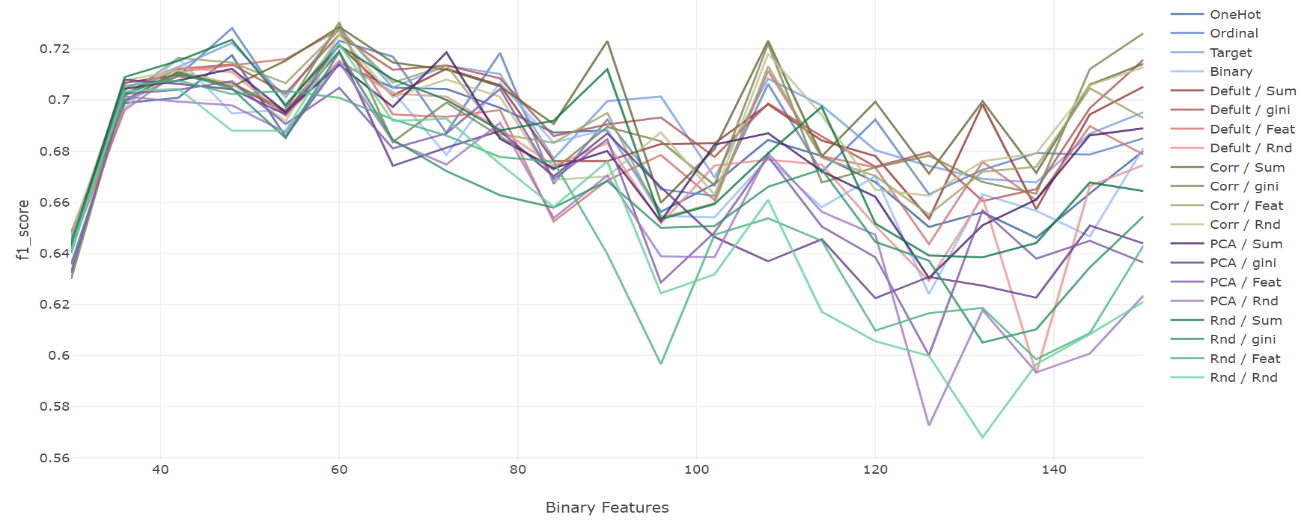


圖 4.6 連續資料集一中，不同編碼方式所得數值資料的分類成績，對應切分二元特徵數量變化。

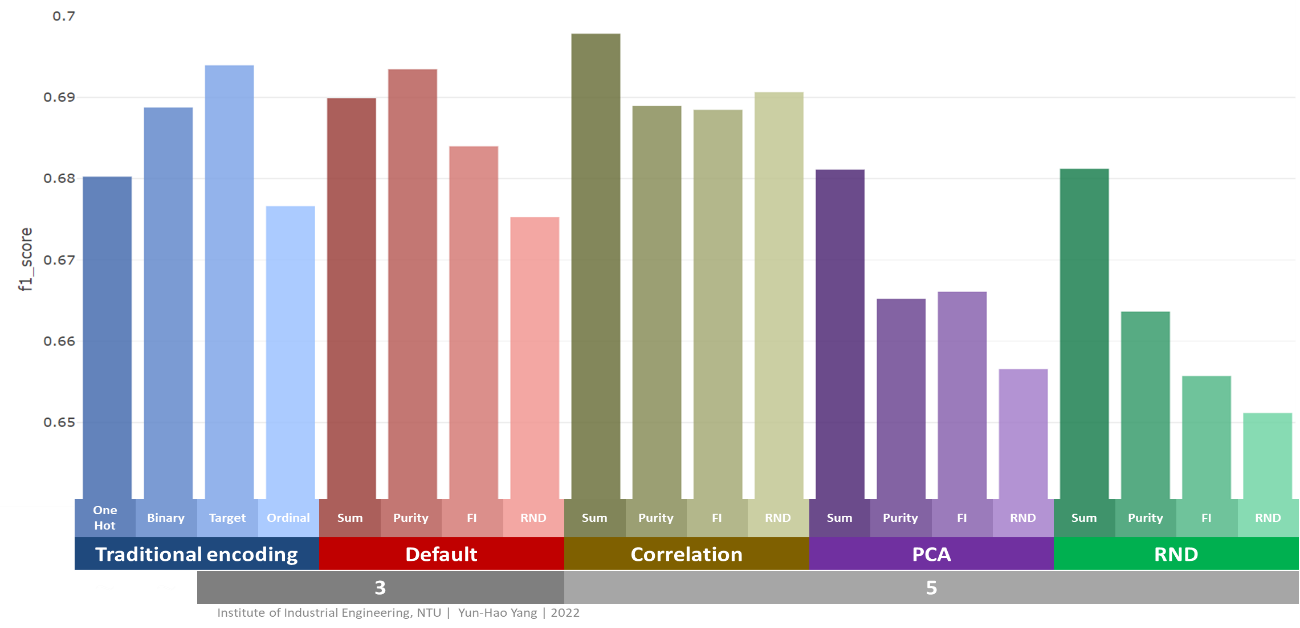


圖 4.7 連續資料集一中，不同編碼方式所得數值資料的平均分類成績。

#### 連續資料集二

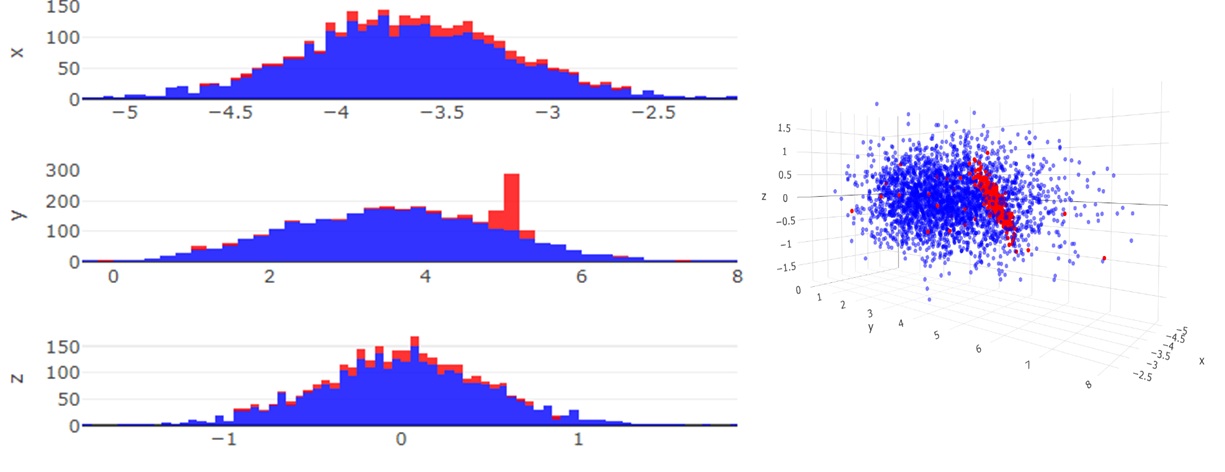


圖 4.8 連續資料集二的資料分布。

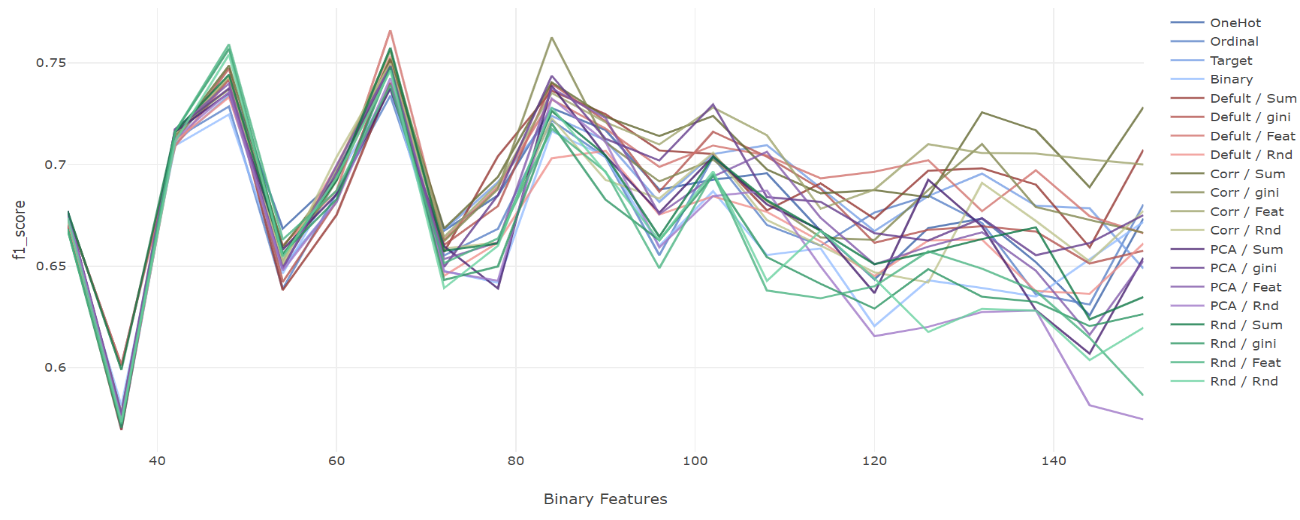


圖 4.9 連續資料集二中，不同編碼方式所得數值資料的分類成績，對應切分二元特徵數量變化。

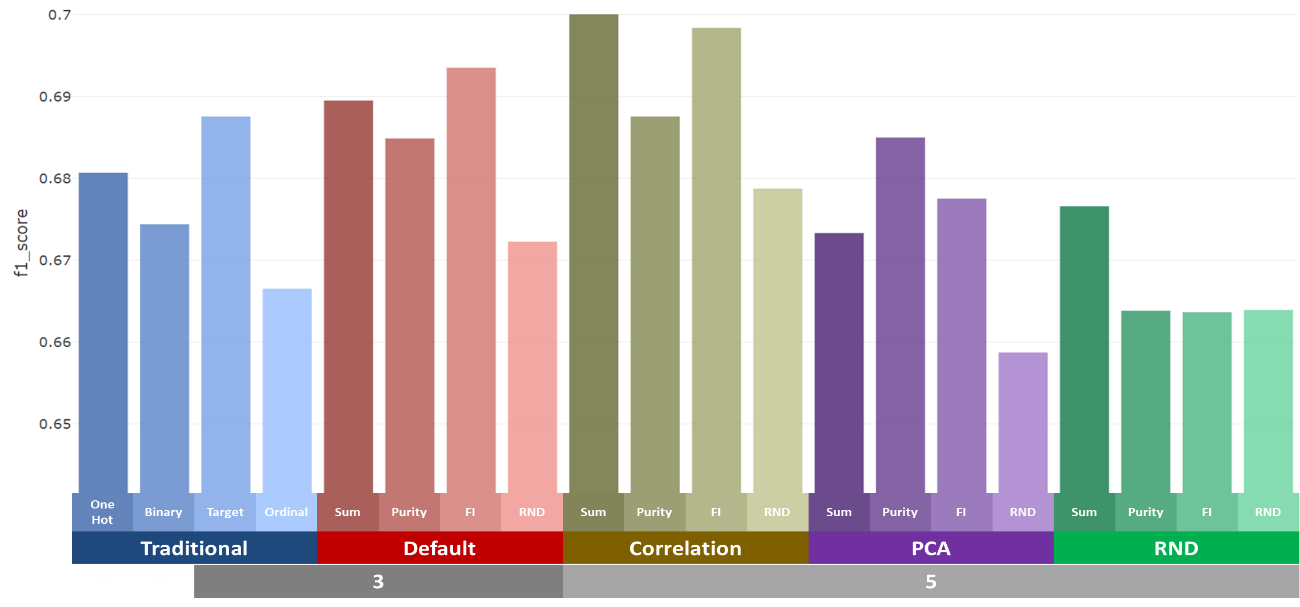


圖 4.10 連續資料集二中，不同編碼方式所得數值資料的平均分類成績。

#### 連續資料集三

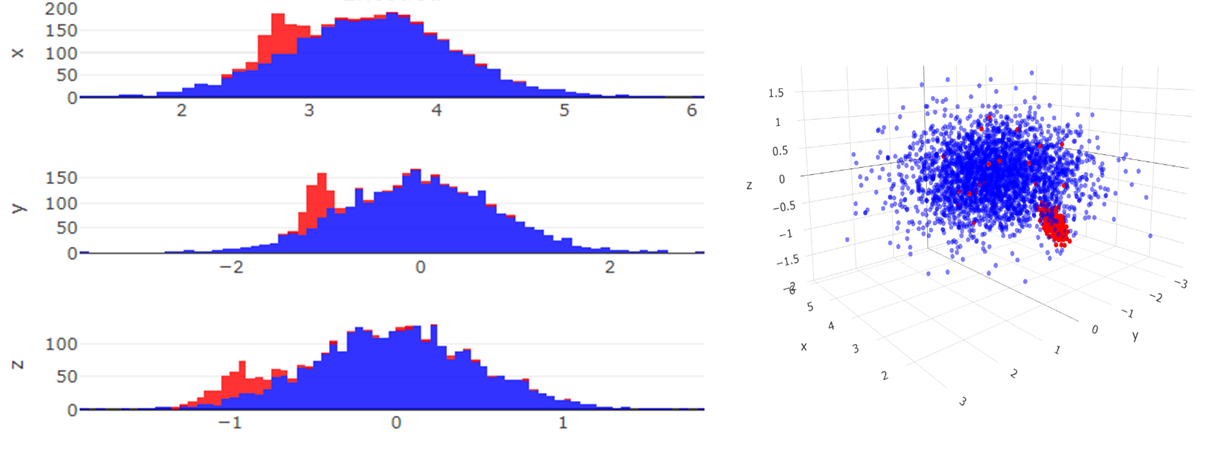


圖 4.11 連續資料集三的資料分布。

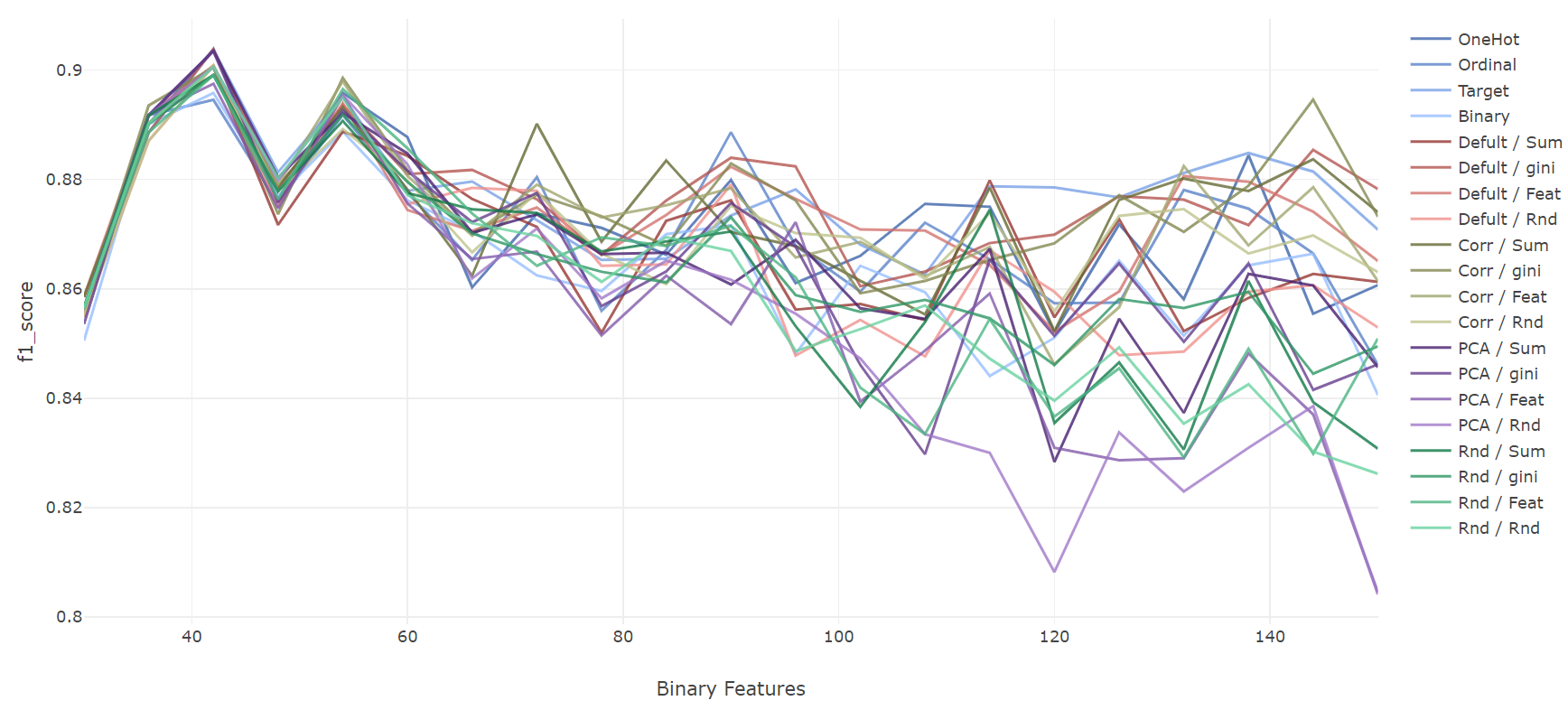


圖 4.12 連續資料集三中，不同編碼方式所得數值資料的分類成績，對應切分二元特徵數量變化。

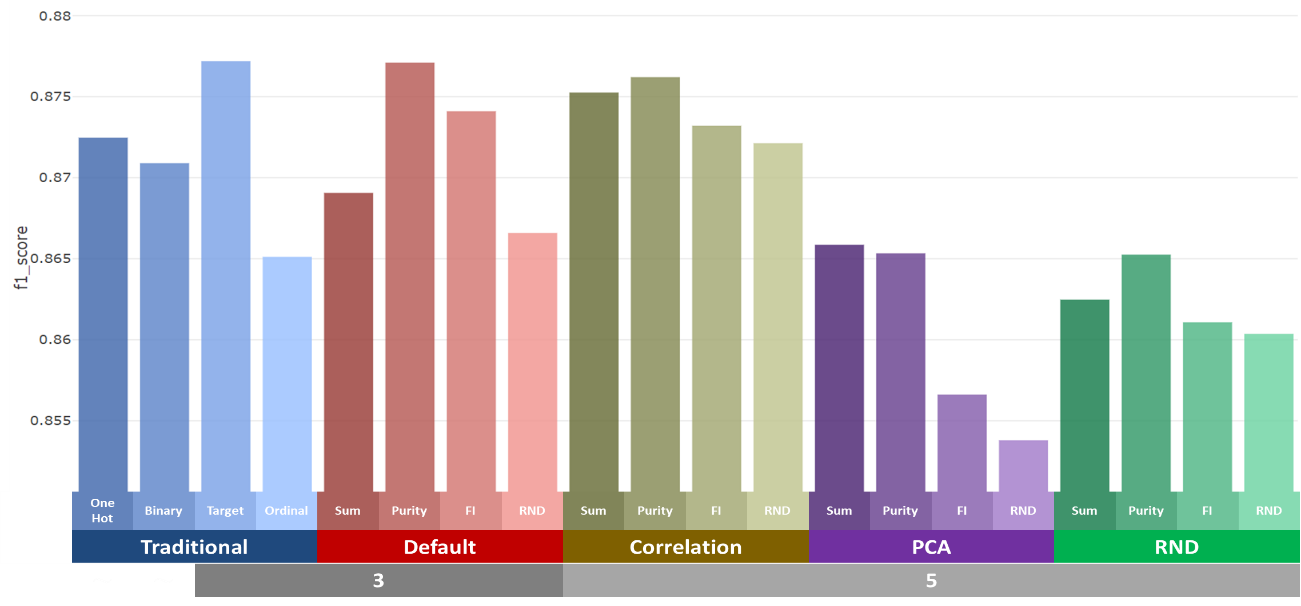


圖 4.13 連續資料集三中，不同編碼方式所得數值資料的平均分類成績。

### 分類結果評比與歸納

在綜觀此三個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，得以觀察出以下幾點現象：

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，不論根據何種方式排序組內特徵，依據主成分、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對於編碼後新數值資料的分類成果具有一定程度上的影響。

1. 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、特徵重要性等兼度式排序時可取得在分類表現上較為優異的數值資料。

1. 目標編碼的表現相當優異：

相較於其餘傳統的變數編碼，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，使其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變數，而本研究所發展的方法著重於無群組資訊的二元特徵之上；對於此類資料，目標編碼將無法進行處理。

1. 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，其中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五。

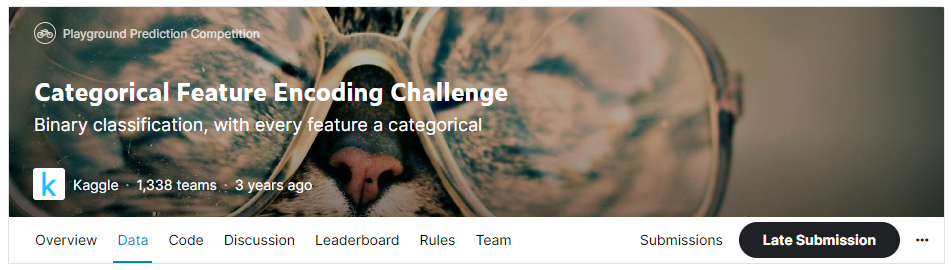
根據圖 4.6、圖 4.9、圖 4.12三圖則可以觀察到在切分的二元變數數量較少時，所有變數編碼產生的數值資料有著相近的分類成績；然而，隨著所切分的二元變數個數增加，變數編碼產生數值資料的分類成績之間的變異也逐漸增加、且大多數相較較少的二元特徵之下，分類成績有著下降的趨勢，主要乃因維度災難所導致，更多的二元特徵稀釋並瓜分了原先連續資料中的資訊。

## UCI資料集

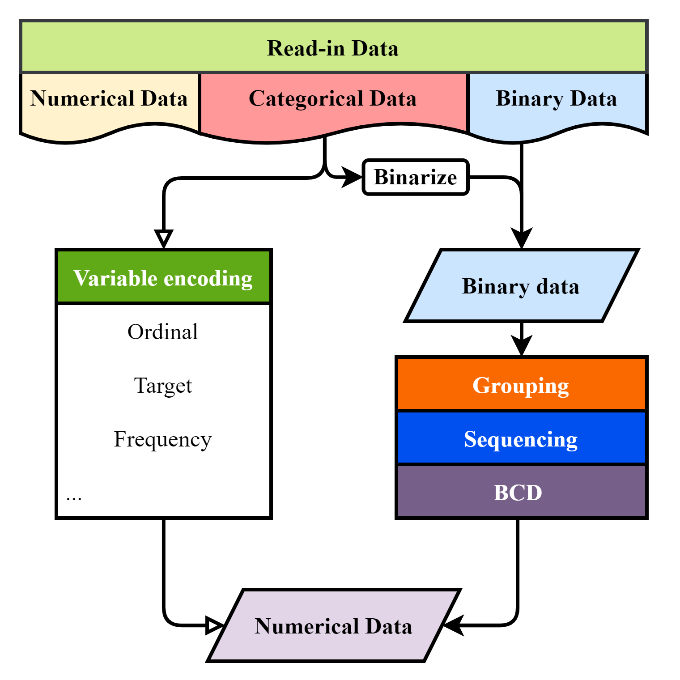
## Kaggle資料集

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。

### 資料集簡介與實驗架構



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，主要原因為該資料集同時具有數值、類別與二元變數；為二元分類問題；經由讀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；為了簡化實驗某些特徵欄位紀錄的是如序號、亂碼等資料，且該特徵欄位具有過多圖特類別



### 分類結果評比與歸納

# 第五章 結論與建議

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等步驟。不論原先的二元特徵資料是否包含特徵群組的資訊，皆可以

# 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.

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*.

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.

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., & François, 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)