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

針對二元變數提升預測表現的種類變數編碼方式

Encoding Methods for Binary to Categorical Variable 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111年06月

June, 2022

# 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個領域當中；不論是影像辨識、自然語言處理或是機台數值預測。遍及了製造業、金融業、市場銷售、與影像醫學辨識的各領域實作者，皆前仆後繼地設法將機器學習應用於遭遇的問題上，來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特徵時，我們便需使用獨熱編碼來將類別變數中的字串特徵轉換成二元變數，以利於作為輸入資料供模型讀取。但若是類別變數的種類繁多，進行獨熱編碼後將產生眾多的二元變數特徵，如此將稀釋該特徵的資訊、並面臨維數災難這一窘境。此外編碼出的二元特徵也不全然與分類、回歸有關聯關係；更甚者，二元變數特徵本身在一定程度上便違反了機器學習演算法的假設。

本研究提出了監督式與非監督式的編碼方式協助二元資料轉換為整數型別的類別變數，以求解決以上提及的問題。透過二元特徵之間的關聯關係、PCA權重等方式群組特徵，再根據各組間特徵自身的屬性進行排序後編碼成整數的類別變數。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／非監督式學習、二元變數編碼

# 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 encoding

# 目錄

[摘要 i](#_Toc117678890)

[Abstract ii](#_Toc117678891)

[目錄 iii](#_Toc117678892)

[圖目錄 v](#_Toc117678893)

[表目錄 vi](#_Toc117678894)

[1 第一章 緒論 1](#_Toc117678895)

[1.1 研究背景 1](#_Toc117678896)

[1.2 研究動機與目的 2](#_Toc117678897)

[1.3 研究架構 3](#_Toc117678898)

[2 第二章 文獻探討 4](#_Toc117678899)

[2.1 變數編碼（Variable Encoding） 4](#_Toc117678900)

[2.1.1 順序編碼（Ordinal encoding） 7](#_Toc117678901)

[2.1.2 獨熱編碼（One-hot encoding） 8](#_Toc117678902)

[2.1.3 二進制編碼（Binary encoding） 9](#_Toc117678903)

[2.1.4 頻率編碼（Frequency encoding） 9](#_Toc117678904)

[2.1.5 目標編碼（Target encoding） 10](#_Toc117678905)

[2.2 維度災難 11](#_Toc117678906)

[2.3 降維處理（Dimension reduction） 14](#_Toc117678907)

[2.3.1 特徵選取（Feature selection） 15](#_Toc117678908)

[2.3.2 特徵萃取（Feature extraction） 16](#_Toc117678909)

[2.4 資料不平衡 17](#_Toc117678910)

[3 第三章 針對多維度二元特徵資料的變數編碼 18](#_Toc117678911)

[3.1 依據特徵工程群組二元特徵 18](#_Toc117678912)

[3.1.1 資料原始特徵群 18](#_Toc117678913)

[3.1.2 主成分分析群集 18](#_Toc117678914)

[3.1.3 相關係數群集 18](#_Toc117678915)

[3.2 排序組間特徵 18](#_Toc117678916)

[3.2.1 特徵和（Column sum） 19](#_Toc117678917)

[3.2.2 吉尼不純度（Gini impurity） 19](#_Toc117678918)

[3.2.3 特徵重要性（Feature importance） 19](#_Toc117678919)

[3.2.4 以基因演算最佳化 19](#_Toc117678920)

[3.3 對各特徵組進行二進碼十進數編碼 20](#_Toc117678921)

[3.4 以Light GBM分類 20](#_Toc117678922)

[4 第四章 案例研討 21](#_Toc117678923)

[4.1 依據連續二元分類資料測試 21](#_Toc117678924)

[4.2 UCI資料集 21](#_Toc117678925)

[4.3 面板廠製程資料 21](#_Toc117678926)

[5 第五章 結論與建議 22](#_Toc117678927)

[參考文獻列表 23](#_Toc117678928)

[附錄 A (如果有) 24](#_Toc117678929)

# 圖目錄

[圖 1.1 研究架構 3](#_Toc117105202)

[圖 2.1 資料預處理常見步驟 (García et al., 2015) 4](#_Toc117105203)

[圖 2.2 獨熱、二進位編碼後的特徵數量比較 9](#_Toc117105204)

[圖 2.3 維度個數變化對於分類模型表現的影響 (Spruyt, 2014) 11](#_Toc117105205)

[圖 2.4 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) 12](#_Toc117105206)

[圖 2.5 超球體體積對應維度變化 (Köppen, 2000) 12](#_Toc117105207)

[圖 2.7 資料分佈情形對應維度變化，以貓狗分類為例 (Spruyt, 2014) 13](#_Toc117105208)

[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) 14](#_Toc117105209)

[圖 2.9 主成分分析步驟示意圖（John, 2010） 17](#_Toc117105210)

# 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](#_Toc117678930)

[表 2.1 不同變數類別的定義與描述 (Stevens, 1946) 6](#_Toc117678931)

[表 2.2 變數類別接受運算子與範例 6](#_Toc117678932)

[表 2.3 不同編碼方式所對應的模型準確度 (Potdar et al., 2017) 7](#_Toc117678933)

[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 10](#_Toc117678934)

[表 2.5 目標編碼後的特徵欄位，以水果價格為例 10](#_Toc117678935)

[表 3.1 染色體範例，以基因演算法排序組內特徵 19](#_Toc117678936)

# 第一章 緒論

本章節將描述將多維度二元特徵作為機器學習模型的輸入時，所面臨到的窘境與難題，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，則必須透過各種編碼方式來對無法作為模型輸入的類別變數做可變編碼；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」如此便能作為模型的輸入，此為序號編碼。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」會使得訓練模型誤解城市之間的關聯性。當面臨類別變數不存在順序關係、種類多且出現比例相近時，通常透過獨熱編碼或是來為各個種類產生新的虛擬變數來表示，如下表。

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，並且稀釋了原先單一特徵的資訊。而類別特徵中的種類越多，也將產生越多的虛擬變數，使得模型訓練時也將消耗更多的記憶體與運算時間，導致模型難以收斂和訓練。但是在製造業當中，經過獨熱編碼、同時虛擬欄位眾多的二元特徵資料卻相當常見，常用於描述製品於製造過程中通過的機台或是工序、以及品質管制當中表示合格與否的檢測項目；該如何前處理這些眾多且相互關聯的二元特徵也成了一大難題。

為減緩維度膨脹的問題，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。

## 研究動機與目的

本研究試圖透過監督以及非監督式的方法，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。透過重新編碼獨熱編碼過後的資料，設法規劃出新的預處理編碼方式，以提升資料的可預測性，主要目標可分為以下幾點。

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編碼，將能大幅度的縮減特徵總數；在縮減資料大小以利儲存的同時，也減低模型花費在資料讀取的時間。

1. 壓縮特徵資訊：

經由獨熱編碼過後，總體資訊不改變，但特徵總數的上升；意味著眾多二元特徵瓜分了原先單一特徵所包含的資訊。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：

多數的二元特徵與彼此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. 維持編碼過後資料的分類成果：

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。當模型再進行優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。編碼後的特徵改為整數型別，避免二原資料型別影響模型的梯度計算。

## 研究架構

本研究程序如圖 1.1所示，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。在緒論中簡介研究動機、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、多維度二元特徵資料所導致的維度災難，且其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特徵資料群組特徵、排序組間特徵、以及編碼各個特徵組。第四章案例分析將透過產生的測試資料、加利福尼亞大學爾灣分校提供的開源資料集（UCI Dataset）、與面板製造廠的製程資料，結合本研究所提出之編碼方式，與原始二元資料、與不同的編碼方式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編碼類型；以及在處理多維度資料時的難題，而後提及減緩維度災難時所採用的方法。

## 變數編碼（Variable Encoding）

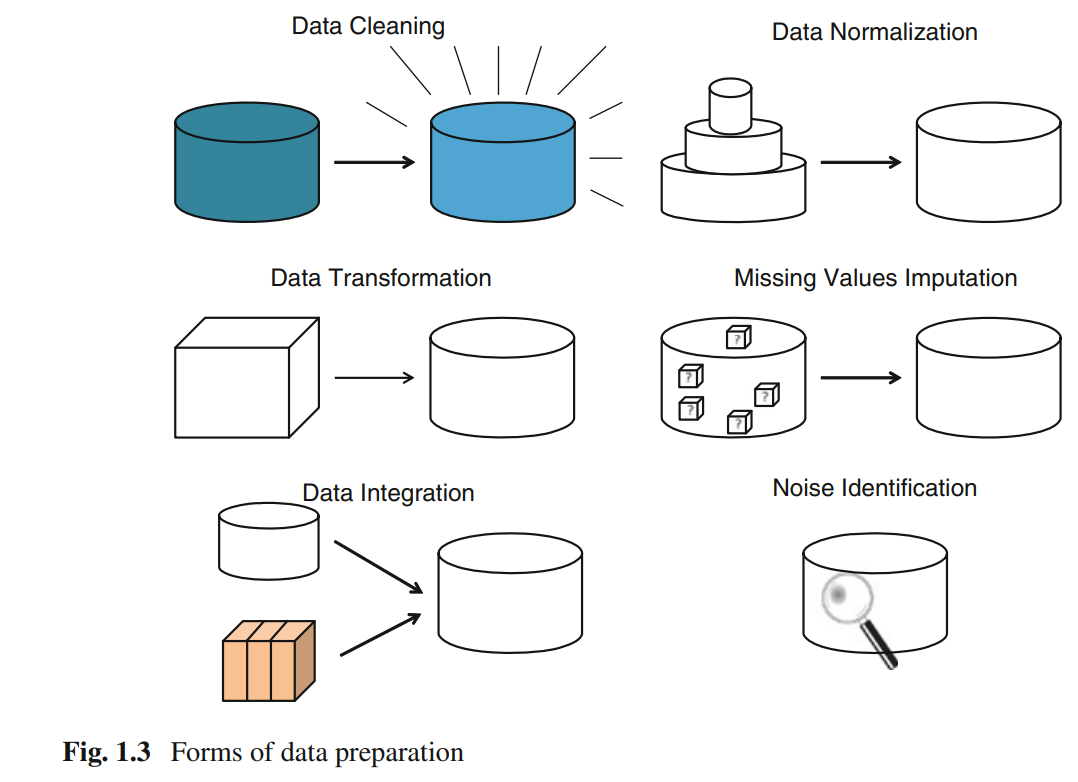


圖 2.1 資料預處理常見步驟 (García et al., 2015)

收集完資料與定義問題後，在進入機器模型建模之前，必須先經過變數編碼步驟，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所示：

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子（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是必須的步驟，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。

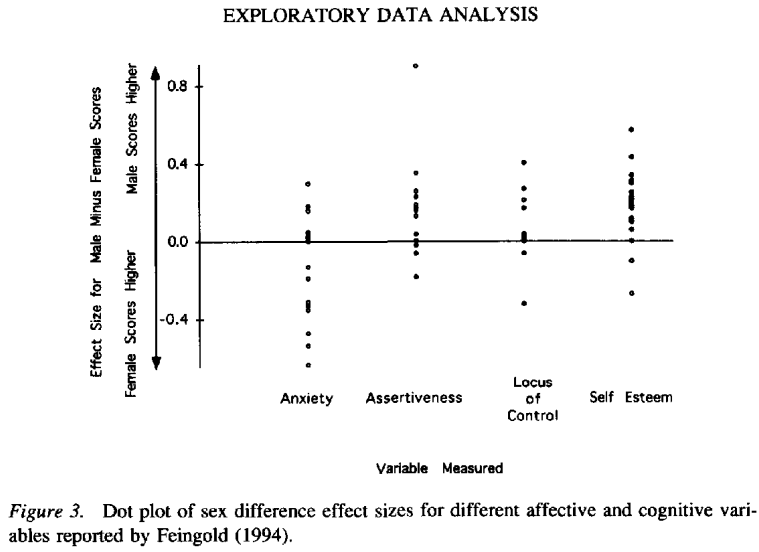


圖 2.2 依據探索式資料分析進行資料視覺化 (Behrens, 1997)

再轉換變數之前，還應理解變數的屬性和變數之間的關係，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的辨別標準、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的型式紀錄，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，為此勢必需要進行變數編碼，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，常見的手法有獨熱編碼、目標編碼來轉換該類別特徵。

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中所示。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特徵關係，使模型產出更佳的分類成果。

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（Ordinal encoding）

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」等明確物理可以通過順序排列類別的特徵，便適合使用順序編碼。概念便是根據特定物理意義、或是類別含意，以零到類別數量之間的整數描述原先的類別特徵，如表 2.4。

假設有一類別特徵具有個獨特類別，每一類別由表示；經由順序編碼後的新特徵欄位，則以0到之間的正整數來描述原先的個不同類別：

(2.1)

(2.2)

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類別間的彈性；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（One-hot encoding）

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，如表 1.1所示。

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。獨熱編碼為透過個虛擬變數（Dummy variable）來描述原先的個種類的類別特徵。假設有一類別特徵具有個獨特類別，每一類別由表示；經由獨熱編碼後的新特徵欄位為，則向量表示原先的個不同類別：

(2.3)

(2.4)

(2.5)

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，雖然在同時間使得資料總體維度上升。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。而稀疏的資料列間存在於完全的互斥關係，只要掌握樣本獨熱的特徵位置便能預測出其餘特徵數值，使獨熱後的高維資料存在著嚴重的並行性與共線性等問題。

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，但獨熱編碼本身的易行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作改良、甚至取代。像是如下所提及的二進制編碼，即是希望以更為精簡、少量的虛擬特徵欄位來描述原始類別變數。

### 二進制編碼（Binary encoding）

相較獨熱編碼，二進制編碼以更少的虛擬變數描述了相同數量的特徵種類，在特徵種類個的情形下，獨熱編碼需要個虛擬特徵描述原始特徵的各個類別，即位每一類別；但二進制編碼只需要以個虛擬特徵，在一定程度上減緩了維度的膨脹，如表 2.4、與圖 2.3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有效的解釋變數所包含的意義。假設有一類別特徵具有個獨特類別，每一類別由表示；經由獨熱編碼後的新特徵欄位，則向量表示原先的個不同類別：

(2.6)

(2.7)

(2.8)

換言之，二進制編碼即是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。

圖 2.3 獨熱、二進位編碼後的特徵數量比較

### 頻率編碼（Frequency encoding）

頻率編碼的想法相當簡單直接，便是以該類別在特徵中出現的頻率做為數值來取代該類別字串；頻率編碼預設在收集資料時，資訊的重複比例即是富有價值的資訊。假設有一類別特徵具有個獨特類別，每一類別由表示；經由順序編碼後的新特徵欄位，則以類別在該特徵中的出現頻率取代：

(2.9)

(2.10)

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小心與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（Target encoding）

目標編碼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）；編碼的方式為把同樣類別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特徵。相較於獨熱、二進制編碼，目標編碼轉換後的特徵欄位個數維持在一個欄位當中，避免了獨熱編碼後造成特徵膨脹的問題。

待補方程式。

表 2.5 目標編碼後的特徵欄位，以水果價格為例

|  |  |  |
| --- | --- | --- |
| **目標欄位** | **原始特徵** | **目標編碼** |
| 5 | 香蕉 | 10 |
| 30 | 蘋果 | 30 |
| 40 | 鳳梨 | 35 |
| 15 | 香蕉 | 10 |
| 30 | 鳳梨 | 35 |

## 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時，當特徵（即樣本的空間維度）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的樣本數量不足，低維度的空間特性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。

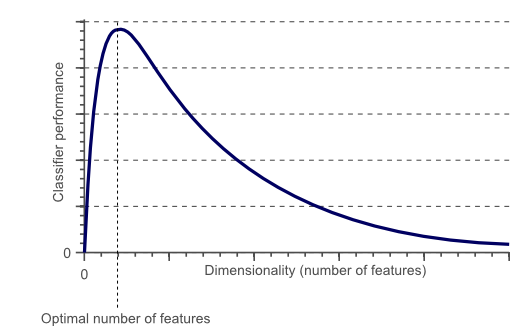


圖 2.4 維度個數變化對於分類模型表現的影響 (Spruyt, 2014)

為了使模型達到更佳的準確度或是分類成果，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數機器學習模型擬合訓練資料中的噪音誤差、模型訓練使用、調整的參數增加，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所示，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，在樣本個數固定的情形之下，導致模型過度擬和訓練資料。

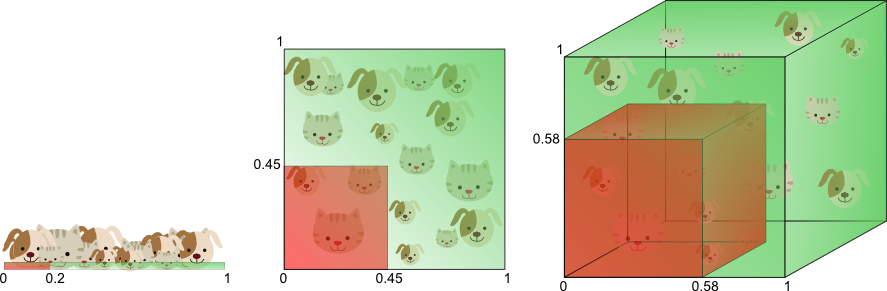


圖 2.5 訓練模型所需樣本個數對應維度變化，以貓狗分類為例 (Spruyt, 2014)

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，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，如圖 2.7。

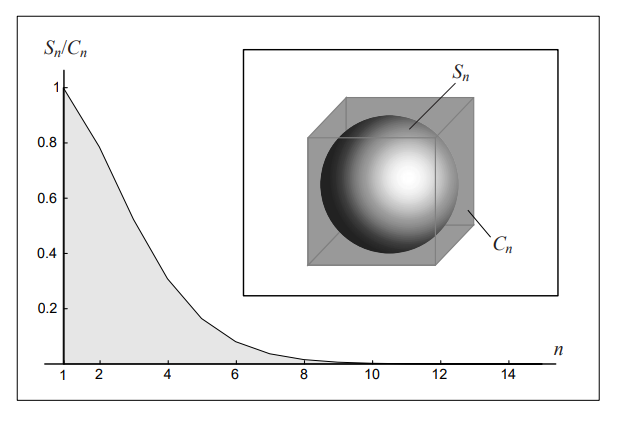


圖 2.6 超球體體積對應維度變化 (Köppen, 2000)

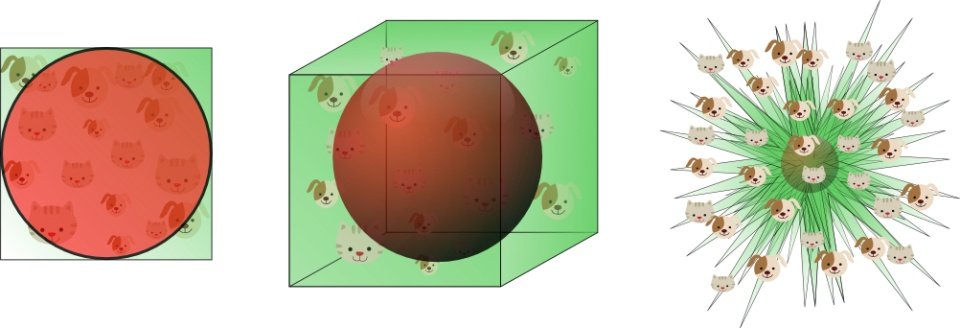


圖 2.7 資料分佈情形對應維度變化，以貓狗分類為例 (Spruyt, 2014)

因此，當特徵維數趨向無窮大時，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

(2.11)

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。見圖 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。

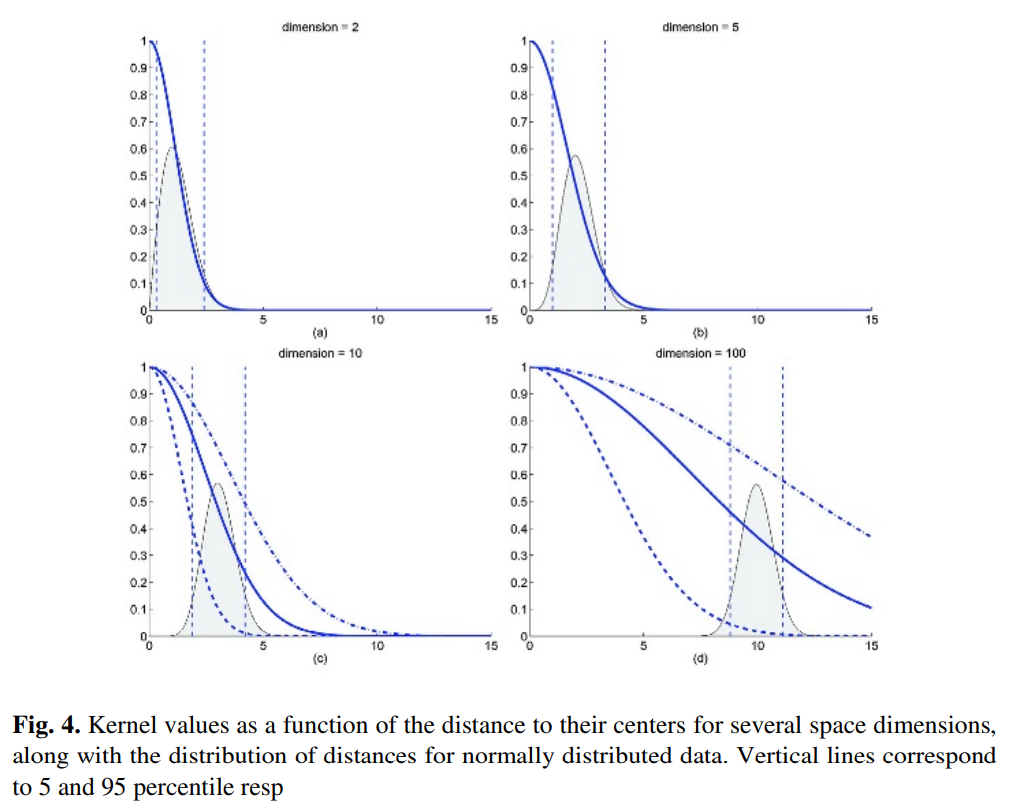


圖 2.8 高斯核函數值對應距離分布於高維度空間的變化 (Verleysen & François, 2005)

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及富有研究價值的主題。目前通常透過降低維度或是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，便於資料視覺化與模型訓練。

## 降維處理（Dimension reduction）

處理資料維度過高，無法進行視覺化或是分析時，降維則是個不錯的方法。降維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；以低維度的座標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。

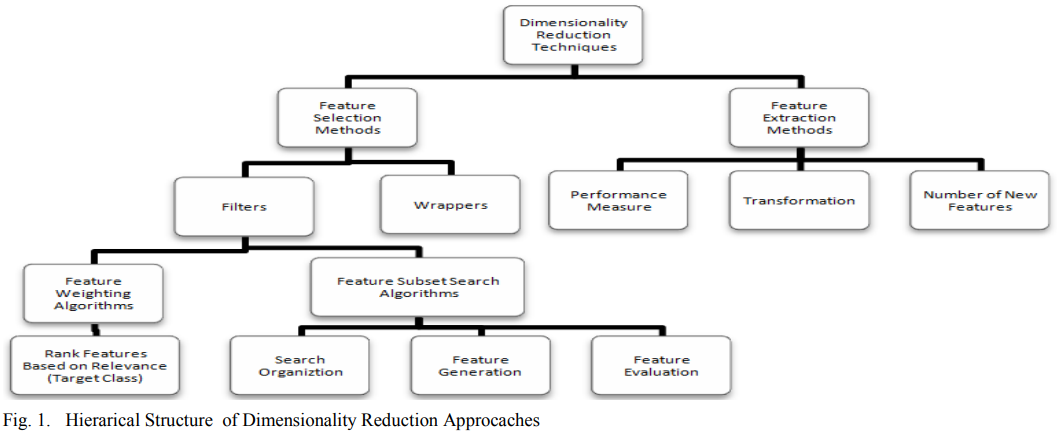


圖 2.9 降維處理的階層化架構 (Tang et al., 2014)

### 特徵選取（Feature selection）

特徵選取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。

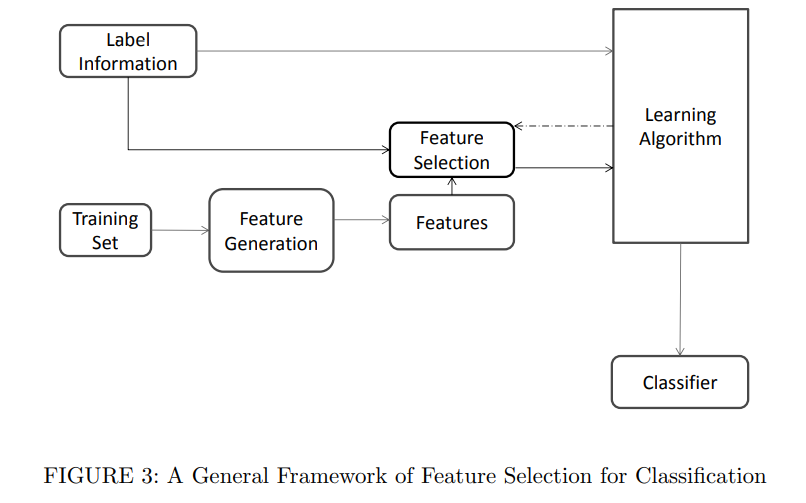


圖 2.10 特徵選取與整體資料分析流程 (Tang et al., 2014)

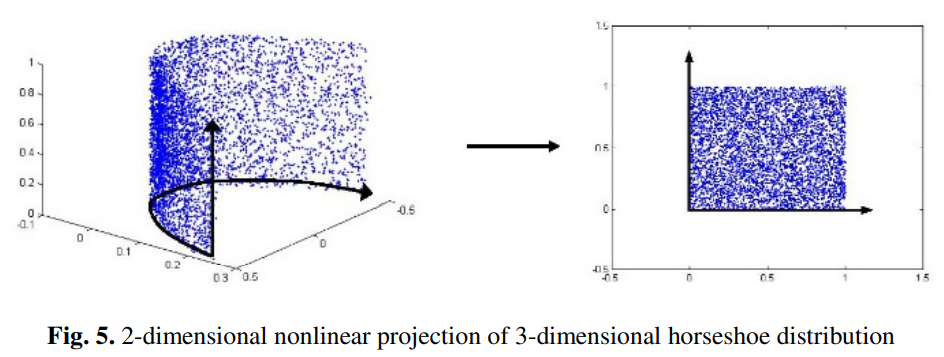
依據不同特徵選取的方式，可以區分為三類：

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。

### 特徵萃取（Feature extraction）

不論原始資料為何，特徵萃取能將原始資料解析、融合後轉換出能作為機器學習模型讀入的新特徵，而後便能針對這些新特徵進行特徵選取。依據降維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



#### 主成分分析（PCA）

主成分分析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。

|  |  |
| --- | --- |
|  |  |
|  | |

圖 2.11 主成分分析步驟示意圖（John, 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

## 資料不平衡

待補

# 第三章 針對多維度二元特徵資料的變數編碼

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。

|  |  |
| --- | --- |
| 符號 | 定義 |
|  | 原始資料 |
|  |  |
|  |  |

## 依據特徵工程群組二元特徵

群組對於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。在本研究之中，根據特徵選取與特徵萃取的方式，

( 3.1)

### 資料原始特徵群

在類別變數中，原始資料所代屬的母特徵；

### 主成分分析群集

### 相關係數群集

在審視二元特徵間的相關系後，可以以階層群集（Hierarchical clustering）、或是塊建模（Block modeling）的方式群組相關特徵。

## 排序組間特徵

群組完相關特徵後，將依據特徵本身屬性，再對各組內特徵做排序，來調動編碼過後的數值。這些屬性包括但不限於特徵總值，吉尼不純度，或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，也可以將此問題描述為一最佳化問題，嘗試以不同優化手法進行求解，例如基因演算、捷思法等等。甚至是先預訓練模型，再依據模型給出的特徵重要度進行排列。

### 特徵和（Column sum）

依據特徵的各樣本數值總和作為排序；以獨熱編碼而言，數值1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。

### 吉尼不純度（Gini impurity）

計算特徵的資訊不純度，以進行特徵排序；以二元分類為例，純度越高的資料將分布於特徵群組的兩側，而中間的特徵則是最不純粹的，如圖所示。



因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。

### 特徵重要性（Feature importance）

### 以基因演算最佳化

本研究也嘗試以基因演算法排序特徵；只要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。像是以編碼過後的組間變異除以組內變異（SSB/SSE）、或是最小化其中任一類別的變異，皆可做為適應值供GA篩選。如…所示，十二個特徵分為四個特徵組，而針對各組內的特徵順序進行調換，以提升最佳的適應值。

表 3.1 染色體範例，以基因演算法排序組內特徵

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Group 1 | | | Group 2 | | | Group 3 | | | Group 4 | | |
| Chromosome 1 |  |  |  |  |  |  |  |  |  |  |  |  |
| Chromosome 2 |  |  |  |  |  |  |  |  |  |  |  |  |

( 3.2

透過基因演算，能循序漸進、方向性地搜尋特徵排序的可能性，經由多次迭代後取得最佳適應值的排序組合。相較於其餘排序方式，基因演算將花費更多的時間成本，且該如何設計適應值的計算也尤為重要。

## 對各特徵組進行二進碼十進數編碼

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，也將導致新資料全距過大、且分布稀疏；例如：若組內二元特徵的個數為十個，則此群組編碼出的整數類別變數全距範圍將達到0-1024。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。

## 以Light GBM分類

再資料編碼過後，以Light GBM作為分類器，同時比較以獨熱編碼、順序編碼、目標編碼、二進位編碼、以及本研究提出的編碼方式所產生的訓練資料，訓練出分類器的分類成效。

# 第四章 案例研討

本研究探討了多種資料集。在發展方法與架構時，透過產生三維連續資料，並切分為多個二元特徵作為資料集；同時也對於UCI，Kaggle等資料集平台上的資料集進行研討。

## 依據連續二元分類資料測試

## UCI資料集

## 面板廠製程資料

# 第五章 結論與建議

內文

# 參考文獻列表

APA格式，依照第一作者姓的字母排序，第一作者同姓則比第二作者，如果兩論文作者全部相同，則以年代較舊者先列。各式文獻範本如下：

<https://apastyle.apa.org/style-grammar-guidelines/references/examples>

<https://subjectguides.library.american.edu/c.php?g=675993&p=4847486>

# 附錄 A (如果有)

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*(2), 131.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (Vol. 72). Springer.

Köppen, M. (2000). The curse of dimensionality. 5th online world conference on soft computing in industrial applications (WSC5),

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,