

全矩陣式資料視覺化與資訊探索

Dimension Free Data Visualization and Information Mining

陳君厚

中央研究院 統計科學研究所

摘要

全矩陣式資料視覺化技術,是一個非降維的視覺化方法,為探索式資料分析(EDA)之一環。其可分析資料維度大小僅受限於電腦顯示器或印表機等輸出尺寸。基本上全矩陣式資料視覺化並不是一項新的技術,只是新一代電腦的計算、記憶與顯示能力賦予它一個新的舞台。

全矩陣式視覺化的基本概念是很簡單的,就是將完整資料矩陣 (data matrix) 或關係矩陣 (proximity matrix) 有效的呈現於電腦顯示器或報表紙上。我們將全矩陣視覺化推廣為四個主要部分:

1. 原始資料之呈現與關係矩陣之選擇
2. 關係矩陣與資料矩陣之排序
3. 關係矩陣與資料矩陣之分割 (partition)
4. 充分統計圖 (sufficient statistical graph)

全矩陣視覺化具高度之變通性 (flexibility),針對不同的資料結構與應用需求可以輕易進行改造。全矩陣式視覺化雖然不完全是一個新的研究領域,卻仍是一個尚待開發且頗具潛力的礦場,存在許多的研究課題與應用技術。以下是數個全矩陣式視覺化發展中及可能發展課題:

1. 類別型 (categorical) 資料之全矩陣視覺化
2. 多時點 (相同變項) 資料之全矩陣視覺化
3. 多條件 (不同變項) 資料之全矩陣視覺化
4. 條件式 (變項校正) 全矩陣視覺化
5. 相依 (dependent) 或群集 (clustered) 資料之全矩陣視覺化
6. 巨量資料之全矩陣視覺化
7. 全矩陣視覺化之遺漏值 (missing value) 處理

我們將針對類別型 (categorical) 資料之全矩陣視覺化及相關應用,作較詳盡之介紹,並以數個實例說明。