Matrix Visualization of Categorical Big Data

高君豪

淡江大學統計與資料科學學系

Abstract

近年來，資訊科技的迅速發展導致需要處理和分析的資料量大幅增加。涵蓋計算技術和統計分析方法的巨量資料分析已成為一個關鍵的研究趨勢。在巨量資料的深入分析中，資料視覺化和探索性資料分析(EDA)將扮演重要角色。然而，值得注意的是，當前巨量資料分析的研究主要集中在連續型資料上，對類別資料的分析則相對獲得較少關注。在象徵型資料分析(Symbolic Data Analysis, SDA)的範疇內，類別模態多值型(Categorical Modal Multi-Valued type)的概念為類別巨量資料分析提供了另一種途徑。基於已知群組或分群方法的結果，類別巨量資料轉換為類別模態多值型象徵型資料，再利用 SDA 的方法進行分析。

本研究針對使用矩陣視覺化方式對類別巨量資料進行視覺化的三個主要挑戰進行相關研究並提出解決方案：(1)計算和排序關係矩陣的計算能力限制。(2)有效利用色彩空間來表示類別矩陣資料。(3)在有限螢幕空間的限制下有效呈現類別巨量資料矩陣圖。我們提出使用廣義相關圖(GAP)結合象徵型資料分析來進行矩陣視覺化和群集分析。通過將類別巨量資料轉換為模態多值型象徵資料，並採用廣義相關圖中的關係矩陣、排序和矩陣視覺化方法，建構一個能夠有效視覺化類別巨量資料的新 EDA 工具：cBigGAP。旨在克服類別巨量資料的關係矩陣計算、排序和呈現等相關的挑戰。

Keyword: Symbolic data analysis, Modal multi-valued data, Matrix visualization, Generalized association plots, Categorical big data, Dimension reduction