探勘多期資料下樣式變化之研究

戴群達

成功大學資訊管理研究所

資料探勘可以找出隱藏在資料下的知識，其中包括樣式發現；一般而言，樣式被認為是一類很重要的知識，因為它代表在大量資料中某些有意義的事件，同時也可以用來產生關聯法則。因此，瞭解樣式的變化是一個重要的研究議題。變化探勘即是為了瞭解資料的變化情況，期望更進一步地協助探勘者制訂決策；然而過去關於樣式變化的研究，皆是採用探勘頻繁樣式的處理程序，也就是在第一階段先刪除不夠頻繁的樣式，只保留符合使用者所設定門檻值下的樣式。如此做法不但會遺失部份資訊，亦不甚合理，因為變化探勘的目標是尋找有「變化」的樣式，而不只是變化的「頻繁」樣式；另外，大部份的研究皆只是比對兩期的資料，但若能觀察愈多期的資料，則所得到的結論也將會愈可靠；若能處理多期的資料，也就可以進行時間性資料探勘，譬如分析每日、每月或每季資料的變化情況等以提供管理者有用的資料變化資訊，方便其掌握市場的變動趨勢或開拓新的客源。由於相關議題在文獻中並無太多著墨，因此我們將以探勘多期時間下樣式的變化為主要的研究議題，期能提供不同於探勘頻繁樣式之外的另一個研究方向。

本研究提出一個新的演算法來探勘所有樣式；並直接採用以樣式的成長幅度作為篩選的標準，以篩選出所有符合變化趨勢的樣式。為了改善探勘效率，本研究發展一個候選樣式森林的特殊資料結構及演算法以探勘多期資料的變化；並為該特殊的資料結構設計一套彈性的的機制以增加發現樣式變化的可能性，以及可以避免太多不具參考價值的樣式。為了測試本研究所提出之演算法效率，我們另外設計並實作了一套以探勘頻繁樣式為基礎之演算法來探勘多期時間下樣式的變化。進行一連串的測試結果之後，我們發現本研究所提出之資料結構與演算法的確可以非常地有效率探勘多期時間下樣式的變化，而且探勘愈多期的資料集合愈能突顯本研究所提出的演算法之優越性。此外，我們亦針對各期資料集合間的相似度，以及樣式的變化率門檻值對不同的變化探勘演算法所造成的影響加以探討。

***關鍵字*：** 樣式發現；變化探勘；候選樣式森林；時間性資料探勘；演算法；資料結構

論文指導教授: *王逸琳* ilinwang@mail.ncku.edu.tw
http://ilin.iim.ncku.edu.tw
***戴群達* 於 2005 年就讀成大資管所碩士班，於 2007 年獲得碩士學位**