應用序列樣式探勘技術建構成對序列相似核方法於支援向量機分類器

研究生：姚佑俞　　指導教授：蔡介元 博士

元智大學 工業工程與管理研究所

摘要

在現實生活中，資料探勘中的序列分類經常被研究和討論，例如文本分類和蛋白質功能預測等。近年來支援向量機備受關注和廣泛的使用在序列分類領域裡，因為支援向量機可以處理分類中的非線性資料，且擁有較高的效率。然而，核方法的設計是支援向量機最重要的一部份，因為適合的核方法對支援向量機的分類有決定性的影響。因此，本研究提出一成對序列相似核(pairwise sequence similarity kernel)，利用挖掘出來的序列樣式當作參考序列，而非用 k-mers，然後再利用映射函數（map function）去計算參考序列和資料庫序列的相似分數。為了得到所需要的序列樣式，三種不同挖掘序列樣式的方法分別從序列資料庫中萃取序列樣式(frequent sequential pattern)、封閉序列樣式(frequent closed sequential pattern)和最大序列樣式(frequent maximal sequential pattern)，本研究評估這三種序列樣式何者能擁有較高的準確率。核方法的映射函數為編輯距離演算法，是用來計算所提出來核之參考序列和資料庫序列的相似分數。接著，依據所提出的成對序列相似核來建立序列支援向量機分類器，並透過所提出的序列分類方法，可以精準預測新序列的類別。本研究應用一組人工資料和一組蛋白質序列資料測試所提出支援向量分類器的成對序列相似核方法，其實驗結果證明本研究所提出支援向量分類器的成對序列相似核方法是有效率且可行的。

關鍵詞：序列樣式探勘、序列分類、核方法、支援向量機分類器