完整關聯法則於生物資訊之應用

孫光天、鄭吉宏、李耀全
國立臺南大學 數位學習科技研究所

E-mail: ktsun@mail.nutn.edu.tw

摘要

資料探勘技術在生物與醫學資訊系統中常被用來探勘隱藏的訊息或作自動化決策處理等相關應用。在資料探勘中，關連法則探勘是一種從大量的資料理面，探勘每一個資料項目之間是否具有特性規律的技術，常用於生物序列分析，可以尋找出不同核苷酸或是胺基酸之間是否具有關聯性。但關連法則演算法的設計，並非是針對生物序列資料的特性設計。除此之外，在有些狀況下會因為最小支持度的關係而刪除含有不頻繁項目的序列，但該序列可能隱含重要生物意義與訊息。有些情況則又面臨太多規則產生，並沒有效精簡，導致生物實驗者無法直接根據探勘結果做生物實驗。因此，必須有另一種關連法則演算法，具有綜整全部規則的特性。本研究將生物序列資料導入布林代數及相關運算，以簡化規則之間邏輯關係的方式，發展適用於探勘生物序列資料的關聯法則演算法，有效發掘生物序列隱藏的訊息。

關鍵字：生物資訊、資料探勘、關連法則、Apriori 演算法

Abstract

Data mining technologies often used in finding hidden information and automatic decision of biological and medical information systems. In data mining, association rule is a technology of mining the items if including special regular patterns in large database. Association rule applied in bioinformatics was analyzed biological sequences to search some relation between different nucleotide or protein sequences. But the design of association rule was not based on biological sequence data properties. In some case, the algorithm would delete the sequences that include important biological message by the minimal support value. In other case, it would produce a large number of rules that are beyond experiments. In this paper, the Boolean algebra operations were used to deal with biological sequences. It greatly reduces the number of rules and avoid the inconsistency developes between them. The proposed method is an efficient technologe to process the bio-information.

Keywords：*Bioinformatics*、*Data mining*、*Association rule*、*Apriori*、*Logic*