CHAPTER 11

關聯法則 – Association rules

・・學・習・目・標・・

- 瞭解何謂關聯規則
- 瞭解何謂支持度(support)及可靠度(confidence)
- 瞭解關聯性法則的步驟
- 瞭解何謂Apriori演算法
- 瞭解Apriori演算法處理程序
- 瞭解何謂候選項目集合和高頻項目集合
- 瞭解候選項目集合和高頻項目集合的計算流程
- 瞭解IBM SPSS Modeler資料格式與設定
- 瞭解IBM SPSS Modeler個案實作的步驟
- 瞭解IBM SPSS Modeler實際個案分析實作

關聯規則 (Association rules) 是一種機率關係的應用，它是憑藉著過去的經驗或紀錄而在大型資料庫中尋找資料的屬性 (Agrawal, Imilienski & Swami, 1993)。關聯規則演算法起初是由不同型態的資料而發展，例如購物籃分析就是使用零售商的交易資料，而演算法的核心理念則是由 Agrawal, Imilienski & Swami 等人在 1993 年提出 (Shahbaz, Srinivas, Harding and Turner, 2006)。

11-1 關聯法則Apriori基本概念

在資料探勘的領域之中，**關聯性法則 (association rule)** 是最常被使用的方法。關聯性法則在於找出資料庫中的資料間彼此的相關聯性，這種方法現已經普遍運用於各領域。此外，在關聯性法則之使用中，Apriori 是最爲著名且廣泛運用的演算法。

假設在資料庫中，$L = \{l_1, l_2, \cdots, l_n\}$ 是所有顧客的知識與需求之集合，其中 X 及 Y 均爲決策變數且是 L 的子集合 (subset) 並互相獨立，因此關聯性法則的表示形式爲：$X \rightarrow Y$，$X \subset L$，$Y \subset L$ 且 $X \cap Y = \varnothing$。關聯性法則的產生由兩個參數來決定：**支持度 (support)** 及**可靠度 (confidence)**(Wang, Chuang, Hsu & Keh, 2004)。

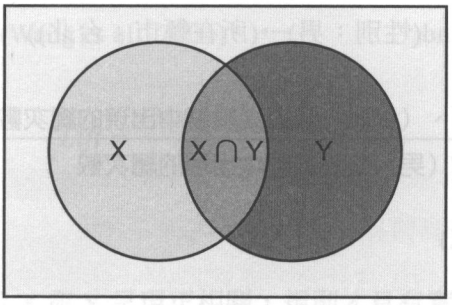
支持度的定義爲決策變數在資料庫中所出現的比例，表現形式爲 Sup(X)，也就是在整個資料庫 L 中出現的比例，支持度越高，越值得重視。支持度代表事件的發生機率。$Sup(X \rightarrow Y)$ 代表同時發生 X 和 Y 兩個交易事項的機率，支持度介於 0% 和 100% 之間。

$$Sup(X) = \frac{\text{項目集合X在資料庫中出現的總次數}}{\text{資料庫中的總交易筆數}}$$

可靠度的定義此關聯性法則可信的程度，也就是某決策變數 X 已確知或成立時，另一決策變數 Y 發生或成立的機率，與統計中的條件機率相同，表現形式爲 Conf (X → Y)。Conf(X → Y) 代表發生 X 的交易事項下，發生 Y 交易事項的機率，可靠度介於 0% 和 100% 之間。

$$Conf(X \rightarrow Y) = \frac{Sup(X \cap Y)}{Sup(X)}$$

在圖 11-1 中，我們以機率中典型的文氏圖來說明這個機率的簡單概念。假設藍色方框的面積是整個事件出現的集合 (set)，因此其機率表示是 100%，也可以 P(S)=1 來表示。X 在整個事件中所佔的面積就是 X 在整個事件中出現的機率，可用 P(X) 來表示。Y 在整個事件中所佔的面積就是 Y 在整個事件中出現的機率，可用 P(Y) 來表示。中間深色區塊表示 X 與 Y 同時出現的機率，即爲 X 與 Y 的交集，所以可以用 P(X ∩ Y) 來表示。則事件 X 的支持度即爲 X 在事件中出現的機率，故 Sup(X) = P(X)。當已知 X 已經成立時，那麼 X 與 Y 同時出現的交集機率爲 $P(X \cap Y)/P(X)$，也就是 $Conf(X \rightarrow Y) = Sup(X \cap Y)/Sup(x)$。



◀ 圖11-1 文氏圖 ▶

另外由下表及計算過程來說明支持度與可靠度的計算方式：

表11-1 原始資料表

ID	年齡	性別	所在縣市
001	27	男	台北
002	28	女	高雄
003	33	女	台北
004	34	女	高雄
005	29	男	台北

表11-2 關聯法則關係表

Rule	Support	Confidence
(年齡：25~30)and(性別：男)→(所在縣市：台北)	40%	100%
(所在縣市：台北)→(性別：女)	60%	33.3%

Sup ((年齡：25~30)and(性別：男))

$$=\frac{\text{(年齡：20 ~ 25) } and \text{ (性別：男) 在資料庫中出現的總次數}}{\text{資料庫中的總交易筆數}}=\frac{2}{5}=40\%$$

Conf((年齡：25~30)and(性別：男)→(所在縣市：台北))

$$=\frac{\text{(25 ~ 30) } and \text{ (男) 、 (台北) 同於資料庫中出現的總次數}}{\text{(25 ~ 30) } and \text{ (男) 在資料庫中出現的總次數}}=2/2=100\%$$

Sup (所在縣市：台北)

$$=\frac{\text{(所在縣市：台北) 在資料庫中出現的總次數}}{\text{資料庫中的總交易筆數}}=\frac{3}{5}=60\%$$

Conf((所在縣市：台北)→(性別：女))

$$= \frac{\text{（所在縣市：台北）、（性別：女）同時再資料庫中出現的總次數}}{\text{（所在縣市：台北）在資料庫中出現的總次數}} = 1/3 = 33.33\%$$

一般而言，關聯性法則的支持度及可靠度皆必須分別大於使用者訂定的最低限制，才能據此判定其爲有意義的關聯性法則 (Padmanabhan & Tuzhilin, 2002; Coenen, Goulbourne & Leng, 2004; Kouris, Makris & Tsakalidis, 2005; Wang et al., 2004)。

關聯性法則的建立，按照 Agrawal & Srikant(1994) 兩位學者所設計的流程，有以下二個步驟：

1. 從資料庫中找出**高頻的項目集合(large itemsets)**，亦即此集合之各個決策變數的組合，同時要大於所設定之**最低支持度(minimum support)**。
2. 接著，用前述步驟所產生的高頻項目集合產生關聯性法則，並計算其可靠度，若高於所設定的**最低可靠度(minimum confidence)**，則此法則確定成立。

此外，爲減少僅憑藉此兩項指標可能造成之偏誤，因此，應該要再考量相關性 (correlation)，進行**相關分析 (correlation analysis)**，此處所提到相關分析，即爲**增益值 (lift)**(Wang et al., 2004)。

$$Lift = \frac{Confidence(X \rightarrow Y)}{Support(Y)}$$

若：

增益值＞ 1，表示 X 與 Y 呈現正相關，規則才具有實用性。

增益值＝ 1，表示 X 與 Y 爲獨立事件。

增益值＜ 1，表示 X 與 Y 呈現負相關，比亂數取得之結果更差。

11-2 Apriori演算法簡介

在關聯性法則之使用中，Apriori 是最爲著名且廣泛運用的演算法。最早是由 Agrawal & Srikant 等兩位學者於 1994 年首先提出，而在這之後許多應用的相關演算法，僅是修正 Apriori 中的部分概念而來，例如 DHP 演算法、DLG 演算法、DIC 演算法與 FP-Tree 演算法等，其處理程序說明如下：

1. 定義**最低支持度(minimum support)**及**最低可靠度 (minimum confidence)**。
2. Apriori演算法使用了**候選項目集合(candidate itemsets)**的觀念，若候選項目集合的支持度大於或等於最低支持度(minimum support)，則該候選項目集合爲**高頻項目集合(large itemsets)**。
3. 首先由資料庫讀入所有的交易，得到第一候選項目集合(candidate 1-itemset) 的支持度，再找出第一高頻項目的集合(large 1-itemset)，並利用這些高頻單項目集合的結合，產生第二候選項目集 (candidate 2-itemset)。
4. 再掃描資料庫，得出第二候選項目集合的支持度以後，再找出第二高頻項目集合，並利用這些第二高頻項目集合的結合，產生第三候選項目集合。
5. 反覆掃描整個資料庫，再與最低支持度相比較，產生高頻的項目集合，再結合產生下一層候選項目集合，直到不再結合產生出新的候選項目集合爲止。

以下則利用簡單的例子，來看 Apriori 演算法的處理過程。若資料庫中有四筆交易，每筆交易都具有不同的 ID 作代表，而交易中都包含了有數種物品，如右所示：

表11-3 資料庫中交易紀錄

ID	Items
001	ACD
002	BCE
003	ABCE
004	BE

表11-4 Apriori 演算法產生的候選項目集合和高頻項目集合

C1	
Itemset	Support
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

Scan Database →

L1	
Itemset	Support
{A}	2
{B}	3
{C}	3
{E}	3

→

C2
Itemset
{AB}
{AC}
{AE}
{BC}
{BE}
{CE}

Scan Database →

C2	
Itemset	Support
{AB}	1
{AC}	2
{AE}	1
{BC}	2
{BE}	3
{CE}	2

→

L2	
Itemset	Support
{AC}	2
{BC}	2
{BE}	3
{CE}	2

C3
Itemset
{BCE}

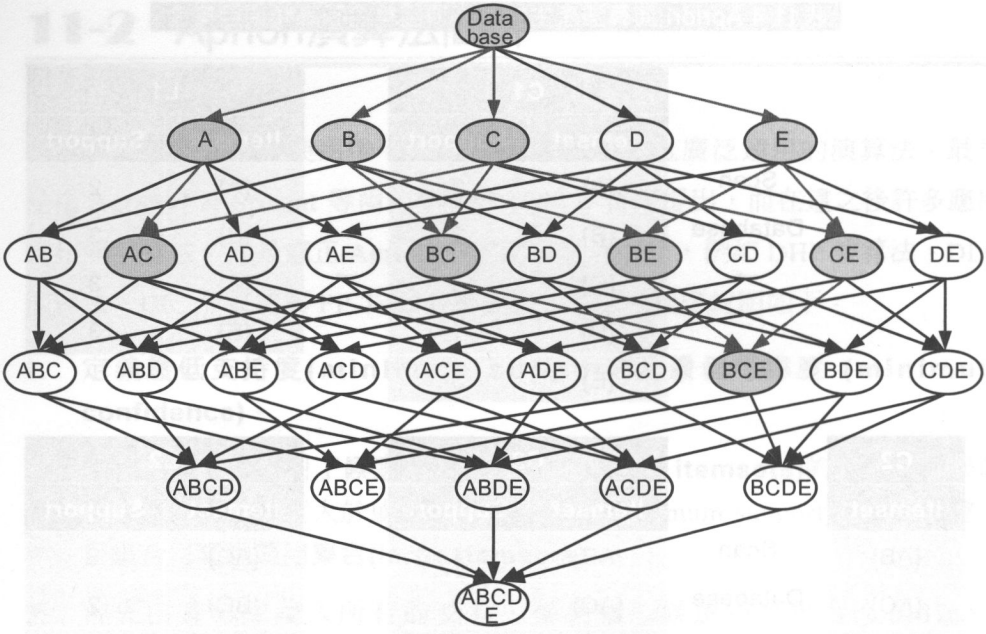
Scan Database →

C3	
Itemset	Support
{BCE}	2

→

L3	
Itemset	Support
{BCE}	2

資料來源：From "Using Information Retrieval techniques for supporting data mining," by Kouris, I. N., Makris, C. H. & Tsakalidis, A. K., 2005, ***Data & Knowledge Engineering***, 52, 362.



◀ 圖11-2 五維度的子集合示意圖 ▶

資料來源：From "Tree Structures for Mining Association Rules," by Coenen, F., Goulbourne, G. & Leng, P., 2004, ***Data Mining and Knowledge Discovery***, 8, 28.

則 Apriori 產生候選項目集合和高頻項目集合的計算流程如下：首先在掃瞄完整個資料庫後，將所有出現商品的次數予以計數，如此即得 C1 表 (第一候選項目集合)，將不符合最小支持度之項目剔除後，即得 L1 表 (第一高頻項目集合)。藉此反覆遞迴的過程，依次產生第二高頻項目集合與第三高頻項目集合 (表 11-4)。

當我們想要產生第三候選項目集合時，所產生的集合項目中，必須皆已產生於第二高頻項目集合中，由圖 11-2 可以很清楚的看到整個演算的路徑。因此第三候選項目僅剩 {BCE}，無法再產生 C4，所以演算法就此終止。