3. 理論與方法

3.1 關聯規則

在資料相關性之研究中挖掘關聯規則（association rule），是相當重要的資料探勘問題。舉例來說，一個銷售的交易資料庫中，我們有興趣的是發現項目（item）的關聯，若在許多交易中，我們發現一個項目的出現會引發另一個項目的出現，這就是所謂的關聯規則。例如：有一關聯規則爲若購屋者購買低均價的房子，則他同時低均價的房子離學校也近，即低均價⇒近學校。關聯規則的挖掘是在所給定的資料中，找出每個資料項目之間有趣關聯的方法，例如在銷售資料中，我們無法直接從資料庫中了解消費者購買的行爲模式，但是透過資料探勘的方法，我們便可以獲得這方面的資訊，進而作爲顧客動線規劃及商品擺設位置調整的依據。因此，關聯規則在決定交易資料屬性間的相互關連是相當有價值的，而且它已經應用於行銷、財務、和零售等各方面。而關聯規則有許多種類，不過我們大體上可以將它分成以下三類：

(1) 以屬性值的型態爲基礎：如果我們所關注的焦點是在物件有沒有出現之上，這種稱爲布林值的關聯規則（Boolean association rule），例如『低均價⇒近學校(support=20%,confidence=40%)』即屬於這類關聯規則。如果我們所要描述的規則是項目或屬性在某種數量範圍下會產生某種結果的話，這種就稱爲定量的關聯規則（quantitative association rule），如下面的例子，X 是代表消費者的一個變數。離捷運站(X, 1100 公尺…5000 公尺)^離學校(X, 285 公尺…2500 公尺)⇒購得房子(X, 房產均價低)。

(2) 以規則中所涵蓋的資料維度爲基礎：如果在關聯規則中的項目或屬性僅參照單一的維度時，我們稱之爲單一維度關聯規則（single dimensional association rule），例如我們將「低均價⇒近學校」的關聯規則寫成「購買房屋(X,"低均價")⇒購買房屋(X," 近學校")」，則其著眼點的則是「購買房屋」的這個維度。但關聯規則中的項目或屬性參照兩個以上的維度時，便稱爲複合維度關聯規則（multidimensional association rule），例如上述定量的關聯規則中的例子，便包含了「購買房屋」、「低均價」以及「近學校」等三個維度。

(3) 以規則集合中所涵蓋的抽象層級爲基礎：如果挖掘關聯規則的方法會找出「購買房屋的均價(X, 35000, …, 200000)⇒購屋地點 (X, 離捷運站距離)」以及「購買房屋的均價(X, 35000, …, 200000)⇒購屋地點(X, 鄰近捷運站距離)」這些不同層級的關聯規則來（離捷運站距離較之鄰近捷運站距離屬於較高層級，也就是較爲一般化的層級），則稱這類規則爲複合層級關聯規則（multilevel association rule）。反之，如果沒有參照到項目或屬性不同層級的規則，則稱爲單一層級關聯規則（single-level

association rule)。

關聯法則中的資料包括一般交易資料或其他維度有關的資料，而知識往往存在資料所隱含與呈現的關係間。對於目前現有的資料探勘方法和資料探勘系統有四種不同的資料關係類別爲資料關聯性、週期性、有趣性、合用性。

3.2 資料分類

在傳統聚類分析中，這些聚落中心被要求形成 X 的分割，並且使得在相同分割區塊內的資料相關性較強，不同分割區塊的資料之間相關性較弱。而聚類分析的目的是找出數個聚類中心，將 X 分成數個適合的聚落。

聚類中心依排序情形可分爲三種：(1)有序聚類中心 (ordered cluster centroid): 聚類中心的散佈呈遞增或遞減之排序狀態. (2)部份排序聚類中心 (partial ordered cluster centroid):聚類中心的散佈只基於某些因素而呈排序情形. (3)加權排序聚類中心 (weighted ordered cluster centroid):聚類中心加入權重的因素，使其散佈情形依因素與權重排序的狀態。

傳統上我們常用 k 平均法 (*k-means*) 進行資料分類，其分類步驟爲：(1) 設聚類個數 c，並計算各群之聚類中心。(2) 計算每一樣本到各中心之距離，並將各樣本分配到與其最近的聚類。(3) 重新計算新聚類中心 v_i 。(4) 重複步驟(2)和(3)，直到各群沒有重新分配樣本的情形爲止。在模糊聚落分析中有兩個基本方法，其中之一依據模糊 c 分類，被稱爲模糊 c 平均分類法。另一方法，乃根據模糊等價關係計算，稱爲模糊等價關係聚落分類法 (Clustering Method Based Upon Fuzzy Equivalence Relation)。 底下將介紹這兩種方法，並用例子作說明。

模糊 c 平均分類演算法

模糊分類法中之模糊 c 平均法，除了須事先指定聚類數 c，另外訂定一個實數 m 和一個代表停止準則的微小正數 ε 。與傳統分類法最大的不同在於隸屬度函數與特徵函數的差別。很明顯地，隸屬度函數的值域爲介於 0 到 1 之間的所有實數。模糊聚類的意義爲:給一組資料 $X=\{x_1,x_2,\cdots,x_n\}$ ，將其分類成一組模糊子集 $P=\{P_1,P_2,\cdots,P_c\}$ ，且滿足下列條件：

$$\sum_{i=1}^{c} P_i(x_j)=1,\ \text{對所有}\ j\in N,\ \text{且}\ 0<P_i(x_j)<1\ 。$$

例如：令一組資料集 $X=\{x_1,x_2,x_3,x_4\}$ ，若 $P=\{P_1,P_2\}$ 爲 X 的一分割，其隸屬情形如下表:

X的元素	x_1	x_2	x_3	x_4

屬於P_1的隸屬度	0.2	0.9	0.6	0
屬於P_2的隸屬度	0.8	0.1	0.4	1

即 $P_1 = 0.2I_{x_1} + 0.9I_{x_2} + 0.6I_{x_3} + 0I_{x_4}$ ， $P_2 = 0.8I_{x_1} + 0.1I_{x_2} + 0.4I_{x_3} + 1I_{x_4}$ 。這裡的 $P = \{P_1, P_2\}$ 就是一個模糊 2 分類。

給一組資料的集合 $X = \{x_1, x_2, \cdots, x_n\}$ ，一般來說 x_k 是一向量。模糊聚類分析的目標是找出一組有 c 個聚落中心 $v_1, v_2, \cdots, v_c$ 的模糊分類 $P = \{P_1, P_2, \cdots, P_c\}$ ，這些聚落中心要儘可能清楚表示樣本分布情況。我們需要一些規則來表達此一概念，也就是結果能使同一聚落中的元素有較強的關聯，且不同的聚落中的元素關聯性較弱的規則需考慮。因此定義一個模糊 c 分割矩陣 $M_{fc} = \{P \mid \mu_{ij}\}$ ， μ_{ij} 表示樣本 j 屬於 i 聚類的隸屬度，且滿足式(3.1)，(3.2)及(3.3)

$$\mu_{ij} \in [0,1];\ i=1,\cdots,c;\ j=1,\cdots,n; \tag{3.1}$$

$$\sum_{i=1}^{c} \mu_{ij} = 1;\ j=1,\cdots,n; \tag{3.2}$$

$$0 < \sum_{j=1}^{n} \mu_{ij} < n; \tag{3.3}$$

則模糊聚類之目標函數爲

$$J_{fc}(P,\mathbf{v}) = \sum_{j=1}^{n}\sum_{i=1}^{c} (\mu_{ij})^m \left\| x_j - v_i \right\|^2; \tag{3.4}$$

其中 $1 \le m < \infty$ 爲模糊指數，此加權參數控制著分類過程的模糊性. $\| \cdot \|$ 是在空間中的內積， $\left\| x_j - v_i \right\|^2$ 是 x_j 與 v_i 的歐基里德距離.

在(3.1)，(3.2)及(3.3)的條件下，對(3.2.4)式求取最小，可得 μ_{ij} 及 v_i 爲

$$v_i = \frac{\sum_{j=1}^{n} (\mu_{ij})^m x_j}{\sum_{j=1}^{n} (\mu_{ij})^m},\ i=1,2,\cdots,c. \tag{3.5}$$

$$\mu_{ij} = \frac{(1/\left\| x_j - v_i \right\|^2)^{1/(m-1)}}{\sum_{i=1}^{c} (1/\left\| x_j - v_i \right\|^2)^{1/(m-1)}} \quad i=1,2,\cdots,c;\ j=1,\cdots,n. \tag{3.6}$$

觀察由(3.5)式計算出的 v_i ，它可被視爲模糊分類 P_i 的聚落中心，因爲它是 P_i 中資料的加權平均，而權數是 x_j 在模糊集合 P_i 中隸屬度的 m 次方。

目標函數用來測度模糊聚落中，聚落中心與元素距離加權後的和。所以，較小的 $J_m(P)$ 值對應較佳的模糊分類。模糊 c 平均分類法的目標是找出一組模糊分類 P 擁有最

小的 $J_m(P)$；聚落問題轉成了最佳化的問題。

模糊等價關係聚落分類法

在使用模糊 c 平均分類法時，必須給定聚落中心的數目。當聚落問題並沒有給定這一數目時，這是一個很不利的條件。在這種情況下，自然的，聚落中心的數目應由資料的結構反應出。依據模糊關係的分類法正是以這樣的方法運作。

每個模糊關係對每一個分割都可以導出明確的分類。模糊聚類問題可視爲辨識適當地資料模糊關係。雖然這通常無法直接做出，但我們應用合適的距離函數，就很容易決定模糊適合關係。

如同前面所做，給一組資料集 $X=\{x_j \mid x_j \in R^p,\ j=1,...,n\}$ 。令在 X 上的模糊適合性關係 R，以 Minkowski 分類上適當的距離函數如下的公式：

$$R(x_i, x_k) = 1 - \delta(\sum_{l=1}^{p} |x_{il} - x_{kl}|^q)^{\frac{1}{q}} \quad ; \tag{3.7}$$

其中 δ 爲一常數使 $R(x_i, x_k)$ 介於 0 與 1 之間。明顯地，δ 是 X 中最大距離的倒數。

一般說來，(3.7)式所定義的模糊適合性關係，不一定是模糊等價關係。因此我們通常必須決定 R 的遞移封閉關係，這可用較簡單的演算法完成；然而 R 是適合性關係，我們可利用下面的定理去發展更有效率的演算法。

定理 3.2：令 R 為有限宇集 X($|x| = n$)上的模糊適合性關係。則 R 的最大-最小的遞移閉集合是關係 $R^{(n-1)}$.

這定理提供了計算遞移封閉的方法，$R_T = R^{(n-1)}$，也就是經由計算關係矩陣的數列：

$$R^{(2)} = R \cdot R$$

$$R^{(4)} = R^{(2)} \cdot R^{(2)}$$

$$\vdots$$

$$R^{(2^k)} = R^{(2^{k-1})} \cdot R^{(2^{k-1})}.$$

直到沒有新的關係矩陣被算出，或 $2^k \geq n-1$ 。這演算法在計算上比一般的演算法更有效率，但只適用模糊反身關係而已。

例如：給一組資料集合 X 如下：

J	1	2	3	4	5
x_{j1}	0	1	2	3	4

x_{j2}　0　1　3　1　0

在此例中，爲了看出式(3.7)中參數 q 的影響力，在分析資料時取 q=1, 2.

(1) q=1 時，這時的距離函數爲 Hamming 距離。最大的距離是 5 (x_1 與 x_3 的 Hamming 距離)，取 $\delta = 1/5 = 0.2$ 。由式(3.7)，得到

$$R = \begin{bmatrix} 1 & .6 & 0 & .2 & .2 \\ .6 & 1 & .4 & .6 & .2 \\ 0 & .4 & 1 & .4 & 0 \\ .2 & .6 & .4 & 1 & .6 \\ .2 & .2 & 0 & .6 & 1 \end{bmatrix}.$$

遞移閉集爲

$$R_T = \begin{bmatrix} 1 & .6 & .4 & .6 & .6 \\ .6 & 1 & .4 & .6 & .6 \\ .4 & .4 & 1 & .4 & .4 \\ .6 & .6 & .4 & 1 & .6 \\ .6 & .6 & .4 & .6 & 1 \end{bmatrix}.$$

由此關係式，可找出不同 α 分割的分類如下：

$$\alpha \in [0, 0.4] : \{\{x_1, x_2, x_3, x_4, x_5\}\}$$

$$\alpha \in (0.4, 0.6] : \{\{x_1, x_2, x_4, x_5\}, \{x_3\}\}$$

$$\alpha \in (0.6, 1] : \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}\},$$

(2) q=2 時，這時的距離函數爲歐基里德距離。我們的第一步是要決定。在 X 中，最大的歐基里德距離是 4(x_1 與 x_5 兩點的距離)，故取 $\delta = 1/4 = 0.25$ 。接著由式(3.7)來計算 R 的隸屬度， $R(x_1, x_3) = 1 - 0.25(2^2 + 3^2)^{0.5} = 0.1$ ，故得到矩陣

$$R = \begin{bmatrix} 1 & .65 & .1 & .21 & 0 \\ .65 & 1 & .44 & .5 & .21 \\ .1 & .44 & 1 & .44 & .1 \\ .21 & .5 & .44 & 1 & .65 \\ 0 & .21 & .1 & .65 & 1 \end{bmatrix}.$$

這關係矩陣不是最大最小遞移；它的遞移封閉爲

$$R_T = \begin{bmatrix} 1 & .65 & .44 & .5 & .5 \\ .65 & 1 & .44 & .5 & .5 \\ .44 & .44 & 1 & .44 & .44 \\ .5 & .5 & .44 & 1 & .65 \\ .5 & .5 & .44 & .65 & 1 \end{bmatrix}.$$

當對應不同的α分割時，這關係矩陣導出 4 種不的分類如下：

$$\alpha \in [0, .44] : \{\{x_1, x_2, x_3, x_4, x_5\}\},$$

$$\alpha \in (0.44, 0.5] : \{\{x_1, x_2, x_4, x_5\}, \{x_3\}\},$$

$$\alpha \in (.5, .65] : \{\{x_1, x_2\}, \{x_3\}, \{x_4, x_5\}\},$$

$$\alpha \in (0.65, 1] : \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}\}.$$

而在接下來這些地價$x_1, x_2, ..., x_n$的資料中，我會先將其按照房產均價及其他特性指標因子的最高與最低分出三個區間將所有資料分成三區間涵蓋起來，如：在地價最低的這個區塊中房產均價的最高值爲A_1，最低值爲B_1，即用$[A_1, B_1]$表示最低均價的範圍。在按照各個均價區間內有關的特性指標，如：此區內有捷運通過，或有著名的商圈，若或有醫院…等。將這些特性指標作成一關聯法則。當作完適當的資料分析之後，必須解決其最小支持度（minimum support）與最小信賴度（minimum confidence）的問題，讓所取出的候選項目集滿足在尋找關聯法則過程中所訂定最小支持度及最小信賴度，則可取出高頻項目集按照這個步驟直到找不到高頻項目集爲止，即可建立關聯法則，讓所建立的關聯法則的支持度及信賴度達到所訂定的標準，而後找出滿足所建立之關聯法則的特徵區間，以讓購屋消費者從有限的購買金額內找出適合的區間選取符合需求區域（ 如：較注重學區或交通 ），而另一方面，讓房地產業者在進行企畫時有更多的方案及訂定建築物總價的標準。

所以在這關聯法則中給於每一個地區的成交因素一合適的模型，再加以整理分析，讓買方在選擇購屋地點時有更多方面的考量，使得買屋時的選擇性加以集中及縮小範圍，讓購屋方能得到最好及最有效的資訊，而不是胡亂投資而造成投資失當，也避免花了大錢買了一棟對自己各方面需要都不方便的房子。而房地產業者方面則是提供更多樣化的企畫來吸引更多消費族群。

3.3 關聯法則使用方法

布林值的關聯規則（Boolean association rule）及 Apriori 演算法介紹

Agrawal 等學者逾 1993 年提出發掘關聯法則之數學領域，用以說明發掘關聯法則之問題。首先提出之關聯法則應用於交易資料庫中，其中各項用詞定義如下(Agrawal et al. , 1993; Agrawal and Sirkant, 1995; Chen et al.,1996)。

(1) 令 $I=\{i_1,i_2,...,i_m\}$ ， I 即是欲討論的項目(items)所組成的集合。

(2) 令 D 為一交易資料集合，其中每一筆交易資料 T 為一組項目集合，表示為 $T \subseteq I$ 。

(3) 每一筆交易資料皆具有一識別之交易序號 TID 。

(4) X,Y 為包含數各項目之項目集。

(5) $T \subseteq I$ 表示一交易資料 T 包含項目集 X 。

(6) 關聯規則 $X \rightarrow Y$ ， X,Y 為項目集，且 $X \cap Y = \phi$

(7) 支持度(Support)： $X \rightarrow Y$ 之支持度 s：含 $X \cap Y$ 筆數/$|D|$ = s % 。
Support($X \rightarrow Y$) 之支持度為 $P(X \cap Y) = s\%$

(8) 信賴度(Confidence)： $X \rightarrow Y$ 之信賴度 c：含 $X \cap Y$ 筆數 /含 X 筆數 = c %
Confidence$(X \rightarrow Y)$之信賴度為 $\frac{P(X \cap Y)}{P(X)} = c$ %。

(9) 最小支持度(minimum support)：在關聯法則挖掘時，若某項目及的支持度太低，則不具顯著性，因此在關聯法則挖掘前會先訂定一個支持度的門檻值(threshold)，此門檻值稱為最小支持度及象徵具有顯著性的最小值。

(10) 最小信賴度(minimum confidence)：與最低支持度類似，使用者在挖掘關聯法則前需先訂定一最低信心度，將信心度未達門檻值的關聯法則刪除，因不具顯著性。此最低信心度稱為最小信賴度。

(11) 高頻項目集(Large k-itenset, L_k)：符合最小支持度的項目集。

(12) 候選項目集(Candidate (k+1)-itemset, C_{k+1})：由 L_k 中項目集兩兩結合(Join)組成長度為 k+1 之候選項目集，表示為 C_{k+1} 。

Apriori 演算法(R. Agrawal et al.，1994)。Apriori 演算法首先於 1994 年由 Agrawal et al.提出的。而在大型資料庫中的銷售交易資料如何發現關聯法則中，Apriori 演算法是最具代表性的演算法。此方法爲研究關聯式法則的入門演算法，其利用循序漸進的方式，

找出資料庫中項目的關係，以形成規則。執行步驟：

步驟 1：訂定最小支持度及最小信賴度。

步驟 2：Apriori 演算法使用了候選物項集合的觀念，首先產生出項目集合，稱為候選項目集合，若候選項目集合的支持度大於或等於最小支持度，則該候選項目集合為高頻項目集合(Large itemset)。

步驟 3：在 Apriori 演算法的過程中，首先由資料庫讀入所有的交易，得出候選單物項目集合(Candidate 1-itemset)的支持度，再找出高頻單物項目集合(Large 1-itemset)，並利用這些高頻單物項目集合的結合，產生候選 2 物項目集合(Candidate 2-itemset)。

步驟 4：再掃描資料庫，得出候選 2 物項目集合的支持度以後，再找出高頻 2 物項目集合，並利用這些高頻 2 物項目集合的結合，產生候選 3 物項目集合。

步驟 5：重覆掃描資料庫、與最小支持度比較，產生高頻物項目集合，再結合產生下一級候選項目集合，直到不再結合產生出新的候選項目集合為止。Apriori 演算法執行步驟如圖 3.3a：

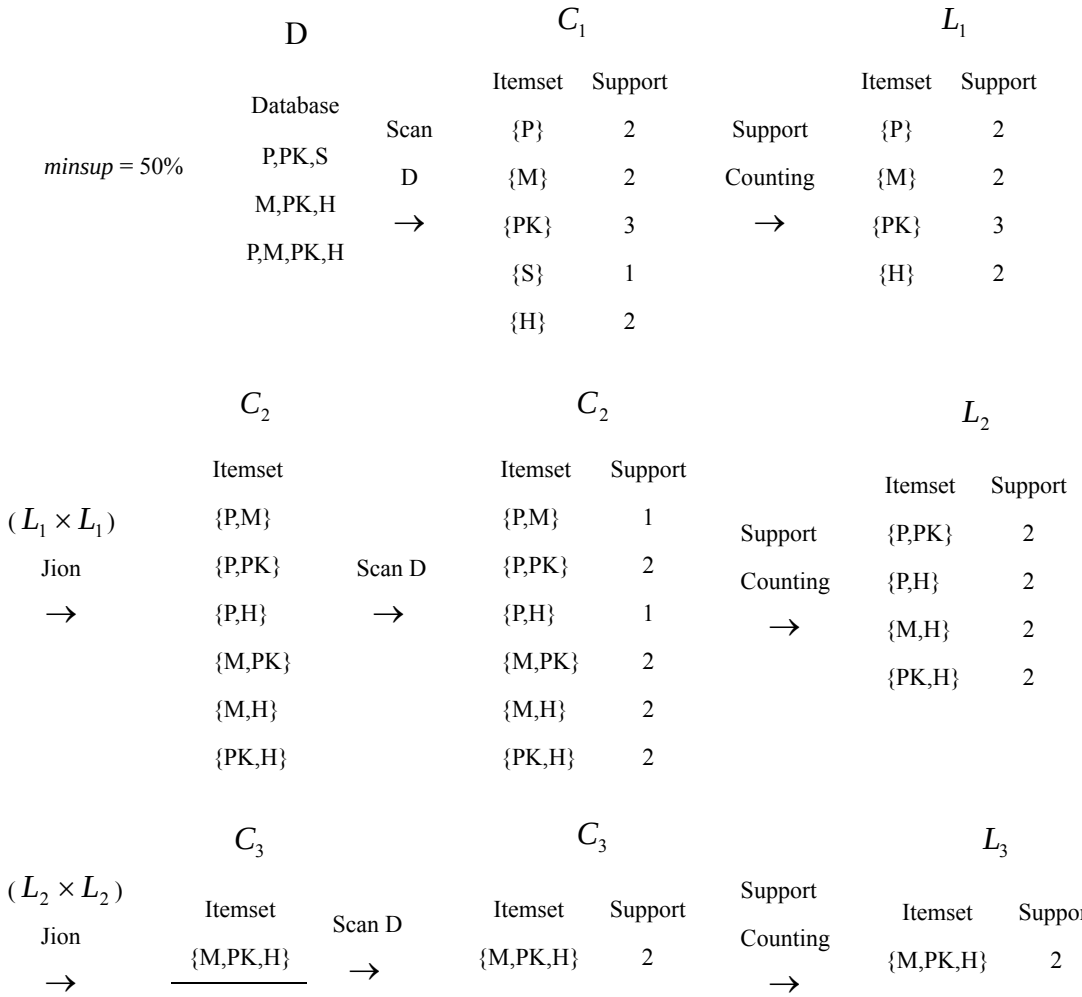

圖 3.3a　Apriori 演算法步驟圖

而布林值的關聯規則（Boolean association rule）爲基礎的處理方法，並把數值屬性分割成多個細小的區間，將這些區間視爲同一屬性裡不同的屬性值，再把各屬性值域皆當作布林值屬性利用 Apriori 演算法來產生關聯法則。屬於一種單一層級的關聯法則。所以在此關注的焦點是在資料是否在特徵因子區間有出現。利用此上述步驟計算其支持度。最後爲了驗證所建立的關聯法則是否準確，再拿不同於建立關聯法則的資料來驗證所列特徵因子與價格是否吻合。若有不吻合，則計算其單筆資料的誤差大小。

使用 Apriori 演算法的步驟分析出的布林值關聯法則，能分析出 P1 應落在哪一個特徵因子的區間內。再將 P1 與其他特徵因子區間的布林值關聯法則列出（P2、P3 亦是）。最後再將單筆或大筆的資料丟入分析的布林值關聯法則中檢驗是否準確。

所以按照 Apriori 演算法中的步驟 1，先將最小支持度及最小信賴度訂出，在步驟 2 中將所有的資料分類，再找出所有的候選項目集。在按照步驟 3 再找出高頻單物項目集，在按照接下來的步驟工作，直到不再結合產生出新的候選項目集合爲止就能建構出所想要的關聯法則。

複合維度關聯規則（multi-dimensional association rule）

在複合維度的領域中，關聯規則可能包括含了兩個或更多維度(dimension)或述語(predicate)，例如：捷運站距離(X, 1100 公尺, …, 5000 公尺)^學校距離(X, 285 公尺, …, 2500 公尺)⇒房屋(X, 房產均價低)。上述的例子就包含了「捷運站距離、學校距離、房屋」三個述語，如果每個述語在規則中均只有出現一次，我們將之稱爲 inter-dimension 關聯法則，上述的例子就是屬於 inter-dimension 關聯法則。還有一種複合維度的關聯法則稱爲 intra-dimension 關聯法則，是指關聯法則是在某一指定維度的內部。若有一些述語重複出現，我們則稱爲 hybrid-dimension 關聯法則。所以，Kamber, Han and Chiang 定義了上述三種複合維度關聯法則，定義分別爲

定義 3.3a （ Intra-dimensional 關聯法則 ）：

在一系列的交易中組成的成分維度為 k 即 $\{d_1, d_2, ..., d_k\}$，則 intra-dimensional 關聯法則表達為 $(d_i, "x_1"), (d_i, "x_2"), ..., (d_i, "x_m") \rightarrow (d_i, "y_1"), (d_i, "y_2"), ..., (d_i, "y_n")$，且 $1 \le i \le k$，$x_1, x_2, ..., x_m$ 與 $y_1, y_2, ..., y_n$ 代表在第 i 維度下的樣本。

例如：在單一維度的關聯法則中(產品，健怡可樂) → (產品，薯片)，這個關聯法則告訴我們，凡是顧客買了「產品」健怡可樂也會帶一包"產品"薯片。所以整個單一維度的關聯法則中，可以說成爲在「產品」中的 intra-dimensional 關聯法則。

定義 3.3b Inter-dimensional 關聯法則：

在一系列的交易中組成的成分維度為 k，則 inter-dimensional 關聯法則表達為

$(X_1,"x_{1m}"),(X_2,"x_{2m}"),...,(X_i,"x_{im}") \rightarrow (Y_1,"y_{1n}"),(Y_2,"y_{2n}"),...,(Y_j,"y_{jn}")$ ， 且 $1 \le i \le k, 1 \le j \le k, 1 \le m \le i, i \le n \le j$ ， $x_{im} \in X_i$ 與 $y_{jn} \in Y_j$ 代表在第 i, j 維度下的樣本。但 X_i, Y_j 的屬性不同。

例如：在單一維度的關聯法則中(地點，政治大學)，(學歷，大學) → (產品，薯片)，這個關聯法則告訴我們，大部分政治大學的學生幾乎都會買薯片。這就是屬於一種 inter-dimensional 關聯法則。在這個例子中，所有的述語即屬性因子均只出現過一次。所以將此將其區分爲 Inter-dimensional 關聯法則。

定義 3.3c Hybrid 關聯法則：

在一系列的交易中組成的成分維度為 k ，則 Hybrid 關聯法則表達為 $(X_1,"x_{1m}"),(X_2,"x_{2m}"),...,(X_i,"x_{im}") \rightarrow (Y_1,"y_{1n}"),(Y_2,"y_{2n}"),...,(Y_j,"y_{jn}")$ ， 且 $1 \le i \le k, 1 \le j \le k, 1 \le m \le i, i \le n \le j$ ， $x_{im} \in X_i$ 與 $y_{jn} \in Y_j$ 代表在第 i, j 維度下的樣本。但 X_i, Y_j 的屬性可相同。

例如：在單一維度的關聯法則中 (地點，政治大學)，(學歷，大學)，(產品，健怡可樂)→ (產品，薯片)，這個關聯法則告訴我們，大部分在政治大學的學生幾乎在買健怡可樂時還會買薯片。這就是屬於一種 Hybrid 關聯法則。而在這個例子中，產品述語出現過兩次。所以將此區分爲 Hybrid 關聯法則。

由上述的定義 3.3a 至定義 3.3c 可清楚得知，若在單一維度討論關聯法則爲 intra-dimensional 關聯法則。而若在整個交易過程中含有多個維度或述語，且每一個維度或述語均只有出現過一次則稱爲 inter dimensional 關聯法則。反之，若有一些維度或述語重複出現，則稱爲 Hybrid 關聯法則。而 Hybrid 關聯法則其實爲 intra-dimensional 關聯法則與 inter dimensional 關聯法則的結合。而複合維度的關聯法則其實可看成一立體方塊，每一區塊代表著每一述語，其立體圖形如下圖 3.3b。

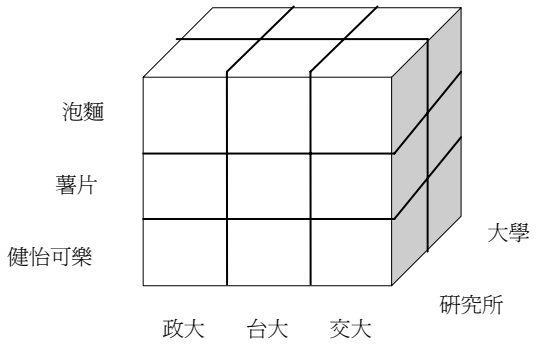

圖 3.3b 複合維度的關聯法則立體圖

在這邊特別一題的是資料庫中的屬性含有兩大類：絕對屬性(categorical attribute) 與數量屬性(quantitative attribute)。絕對屬性的特點是他的值爲物品的名稱，例如：職業、性別、顏色等等，他的值並沒有絕對的順序之分，所以又可稱爲名詞屬性(nominal attribute)：數量屬性子意味著他的值有一定的順訓之分，例如年齡、收入、價格等。

挖掘數量屬性的處理方法可分成三種：第一種作法是運用事先定義好的概念階層來區分數量屬性，例如將整塊房產均價區間分成幾個範圍，房產均價低、中、高 P1：[34195,56607]、P2：[56608,72171]、P3：[72172,186267]，切割的範圍是固定的，如此便可以將之視爲絕對屬性來處理。第二種方法是依據資料的分散狀況進行區分，存入某個儲存項目 (bins) 中，這些儲存項目在資料探勘過程中也可能再進一步組合，其分割範圍則是動態估計的，以滿足某些挖掘所設定的標準。一般在進行儲存項目區分時有三個策略：(1) 每個儲存項目的間隔相同；(2) 每個儲存項目大約都儲存相同的紀錄數目；儲存項目的大小決定後，資料庫的紀錄均勻地分散在每個儲存項目中。第三種方法依據區間資料的語意 (semantic meaning) 來區分範圍，它也是採取動態方式進行，但考慮的因素則是資料之位置的差距，也就是間隔 (interval) 中的位置之密度或數量，以及間隔中位置的接近程度，所以又稱爲 distance-based 關聯法則。而每個數量屬性的間隔則可以用屬性值的群聚 (clustering)來進行估計。

而由於論文中將要研究房產均價與各個述語及屬性因子之間的關係，且只是單純在距離方面上討論，但述語卻不同，且不重複出現，如：離捷運站距離，離公園距離…等。所以在這裡將採用數量屬性轉換成絕對屬性後，再利用 inter dimensional 關聯法則推導出適合的關聯法則。其建構步驟 1 至步驟 4 與 Apriori 演算法相似，不同的爲在候選項目集中不加入房產均價屬性因子而在步驟 5 時重覆掃描資料庫、與最小支持度比較，產生高頻物項目集合，再結合產生下一級候選項目集合，直到不再結合產生出新的候選項目集合爲止，在最後再將房產均價因子加入，找出最適合的複合維度關聯法則。