4.實例研究

在這個章節中將使用兩種不同的方法尋找出適當的關聯法則。第一種方法是利用布林值關聯法則的概念，也就是單一層級的關聯法則，單純的將各特徵因子與房產均價的關聯利用 Apriori 演算法導出適合的關聯法則。第二種方法則是複合維度的關聯法則，考慮不同維度的數量屬性轉換成絕對屬性，再利用 inter dimensional 關聯法則的概念尋找出適當的 inter dimensional 關聯法則。在這裡需要強調的是，Apriori 演算法與 inter dimensional 關聯法則概念的不同。Apriori 演算法在一開始尋找候選項目集時，已經先將房產均價區間與特徵因子區間合併尋找出大於最小支持度的關聯法則，但在 inter dimensional 關聯法則中則事先將各個特徵因子的數量屬性轉換成絕對屬性，再將滿足所有要求的絕對屬性對應至房產均價的絕對屬性中來找出複合維度的關聯法則。

2004 年台北市政府所提供的台北市土地房屋買賣實例調查估價表 1000 筆房價資料中隨機取 99 筆資料來建立我想要的關聯法則模型。在資料分料時，由於利用模糊聚類分析中的 *k-means* 分類法的效果不好，在此將不再探討。因此我們利用了自然分類法先將這 99 筆資料中的房產均價分成三個區塊，分別 P1、P2、P3。之後再把每一個特徵因子也分成三個區間。作法如下：

建築物均價：低、中、高 P1、P2、P3。
離捷運站距離：近、中、遠 M1、M2、M3。
離公園距離：近、中、遠 PK1、PK2、PK3。
離學校距離：近、中、遠 S1、S2、S3。
離醫院距離：近、中、遠 H1、H2、H3。

房產均價，捷運站距離，公園距離，學校距離及醫院距離的區塊分別如下
房產均價低、中、高 P1：[34195,56607]、P2：[56608,72171]、P3：[72172,186267]。
離捷運站距離近、中、遠 M1：[20,540]、M2：(540,1128)、M3：[1128,4915]。
離公園距離近、中、遠 PK1：[30,141]、PK2：(141,210)、PK3：[210,429]。
離學校距離近、中、遠 S1：[70,204.487]、S2：(204.487,285)、S3：[285,2427]。
離醫院距離近、中、遠 H1：[111,583]、H2：[584,1074]、H3：[1075,4012]。

但由於人工分組過於費時，接下來的資料我將使用 MATLAB 軟體，加快分組的速率，更有效的找出適當的模型。

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

接下來層級分類的工作由於過於費時即在不影響關聯法則的原則下，層級分組分類的工作，交給 MATLAB 程式處理。而此一 MATLAB 程式的結果則在附錄部分。且由於在 Apriori 演算法部分同時 (交集) 擁有各種不同特徵屬性因子的個數過於少，因此無法訂出合理的最小支持度，所以在 Apriori 演算法的部分，我們將利用擁有某一個特徵屬性因子就收集(聯集)，所以經由 MATLAB 得知，分別找出滿足 $P_i \to M_i, P_i \to PK_i, P_i \to S_i, P_i \to H_i, \forall i = 1,2,3$ 的資料。minimum support 設定爲 75％，minimum confidence 設定爲 30％。由 MATLAB 程式計算出 2 物項目集合其個數分別如下表所示。

表 4. 1a 為符合 $P_i \to M_i, \forall i = 1,2,3$ 的總個數。

	個數			Support			Confidence		
	M1	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
P1	5	13	15	5/99	13/99	15/99	5/33	13/33	15/33
P2	13	10	10	13/99	10/99	10/99	13/33	10/33	10/33
P3	15	10	8	15/99	15/99	8/99	15/33	10/33	8/33

表 4. 1b 為符合 $P_i \to PK_i, \forall i = 1,2,3$ 的總個數。

	個數			Support			Confidence		
	PK1	**PK2**	**PK3**	**PK1**	**PK2**	**PK3**	**PK1**	**PK2**	**PK3**
P1	15	8	10	15/99	8/99	10/99	15/33	8/33	10/33
P2	12	14	7	12/99	14/99	7/99	12/33	14/33	7/33
P3	6	11	16	6/99	11/99	16/99	6/33	11/33	16/33

表 4. 1c 為符合 $P_i \to S_i, \forall i = 1,2,3$ 的總個數。

	個數			Support			Confidence		
	S1	**S2**	**S3**	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
P1	7	9	13	7/99	9/99	17/99	7/33	9/33	17/33
P2	13	13	7	13/99	13/99	7/99	13/33	13/33	7/33
P3	11	12	10	11/99	12/99	10/99	11/33	12/33	10/33

表 4.1d 為符合 $P_i \rightarrow H_i, \forall i = 1,2,3$ 的總個數。

	個數			Support			Confidence		
	H1	**H2**	**H3**	**H1**	**H2**	**H3**	**H1**	**H2**	**H3**
P1	9	6	18	9/99	6/99	18/99	9/33	6/33	18/33
P2	8	16	9	8/99	16/99	9/99	8/33	16/33	9/33
P3	16	10	7	16/99	10/99	7/99	16/33	10/33	7/33

由上表 4.1a、4.1b、4.1c、4.1d 所示可以由 Apriori 演算法得知 2 物項目集合爲下表所示但由於有 28 個 2 物項目集下表只以 P1 爲例。又由於本研究基於房產均價與各個特徵因子間的關聯法則，所以在 2 物項目集合中只取與房產均價有相關聯的大於所設定之最小支持度的 2 物項目集。且接下來的 3 物候選項目集也不再與相同特徵因子區間做合併。其步驟圖如下圖 4.1a。

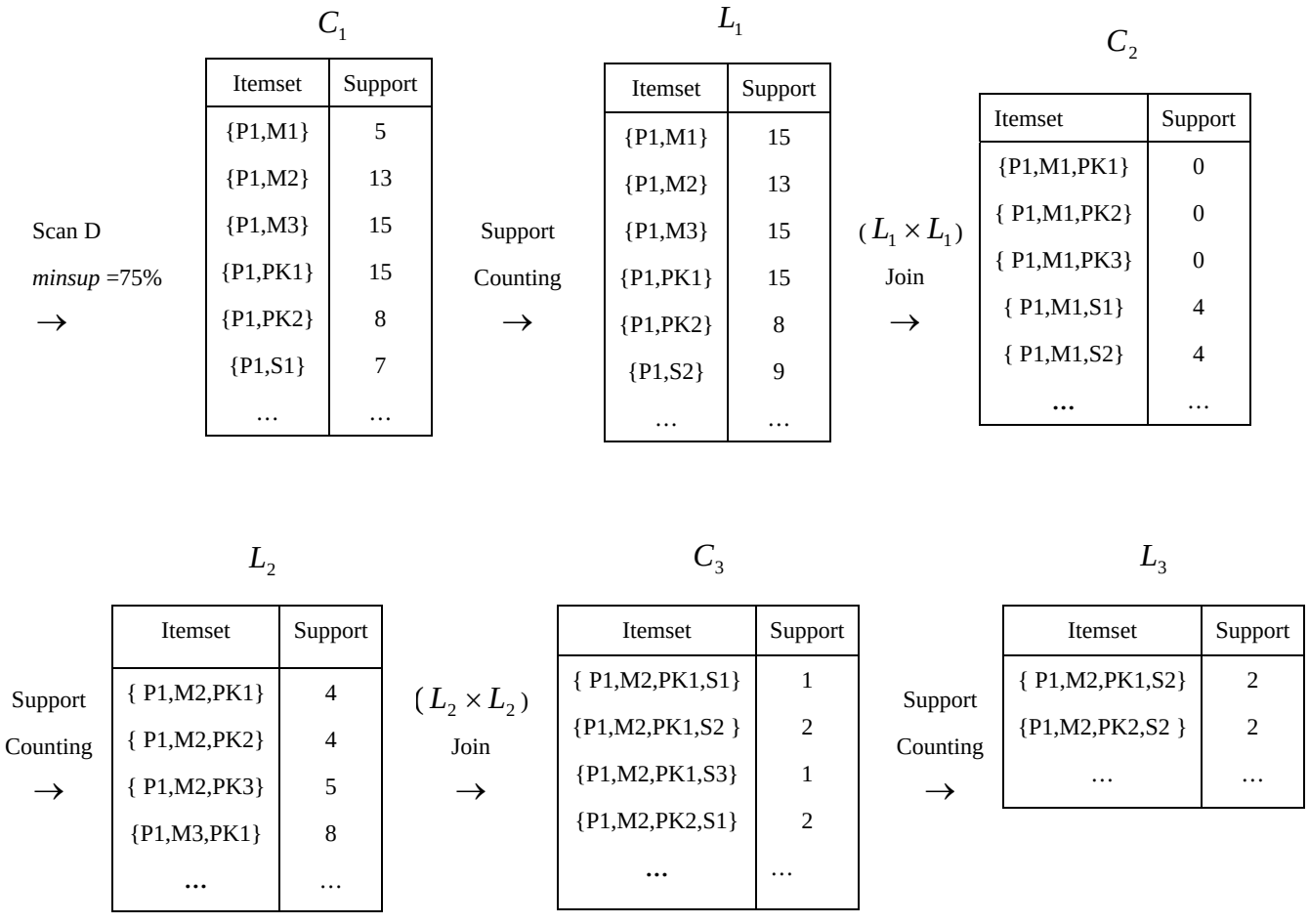

圖 4.1 實例的 Apriori 演算法

上述 Apriori 演算法流程爲 MATLAB 程式結果。在做至 4 物候選項目集後，決定高頻 4 物項目集，在掃瞄完整各資料庫後發現，候選 5 物項目集的支持度均小於所設定的

最小支持度 60%，因此停止作業，決定布林值關聯規則。決定出整理出來房產均價與三個特徵因子區間的布林值關聯規則共有 153 條規則。但若把最小支持度訂爲 55%時就可以得到 5 物候選集，進而得到 17 條滿足高頻 5 物集的關聯法則，分別爲：

P1 的關聯法則：{ P1, M2, PK1, S2, H1 }，{ P1, M2, PK2, S1, H1 }，{ P1, M2, PK2, S1, H2 }，{ P1, M3, PK2, S2, H2 }，{ P1, M3, PK3, S1, H2 }。

P2 的關聯法則：{ P2, M1, PK2, S3, H3}，{ P2, M1, PK3, S3, H1}，{ P2, M2, PK3, S3, H1}，{ P2, M3, PK3, S1, H1 }，{ P2, M3, PK3, S2,H2 }，{ P2, M3, PK3, S3, H1 }。

P3 的關聯法則：{ P3, M1, PK1, S2, H3 }，{ P3, M1, PK2, S3, H3 }，{ P3, M2, PK1, S2, H2 }，{ P3, M2, PK2, S3, H1 }，{ P3, M2, PK2, S3, H2 }，{ P3, M3, PK2, S2, H3 }。

在這邊所建立的關聯法則均滿足步驟 1 所設定的最小支持度及最小信心度，這一節利用 Aporior 演算法所建立的布林值關聯法則將在本章的第三節中選取另 45 筆資料進行比對，看所建立的關聯法則是否適當。而在下一節將利用複合層級概念找出符合 99 筆資料的複合層級關聯法則。

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

在傳統關聯法則挖掘中幾乎都在分析單一層次間的關聯，這些方法並沒有考慮層次內項目的階層關係。隨著層次內的項目多樣化，有時候會很難去找出具有強烈關聯性 (strong association rules) 的關聯法則，可能導致很多項目組無法滿足我們所定的最小支持度門檻值，使得找出來的關聯法則數目變少，甚至產生許多隱藏的之事無法被挖掘。若資料分析的層次提昇，可以容易找出較多且更多樣性的關聯法則以滿足不同使用者的需求，此即複合維度關聯法則。其各屬性因子之關係立體圖形如下圖所示(以捷運距離、公園距離、學校距離爲三個維度的屬性因子)。而在這邊使用的複合維度關聯法則因爲距離述語重複出現(捷運站距離，遠、中、近)，(公園距離，遠、中、近)，(學校距離，遠、中、近)，(醫院距離，遠、中、近)→(房產均價，高、中、低)，所以在這邊使用的複合維度關聯法則爲 Hybrid 複合維度關聯法則的概念，概念立體圖及步驟圖爲圖 4.2。

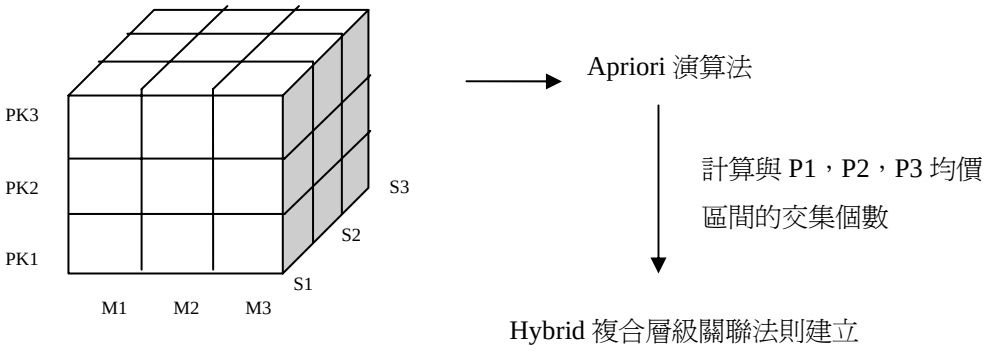

圖 4.2　Hybrid 複合層級關聯法步驟圖

接下來依照 Apriori 演算法前四個步驟，首先訂定最小支持度爲 60%，最小支持度爲 30%，找出候選項目集及高頻項目集，最後再將符合的高頻項目集與房產均價區間對應，觀察取出的高頻項目集中和者與房產均價區間最有關聯性。

在高頻 3 物項目集中與 P1 房產均價區間具有較大關聯性的關聯法則有 9 種，與 P2 房產均價區間具有較大關聯性的關聯法則有 1 種，而與 P3 房產均價區間具有較大關聯性的關聯法則有 3 種。在高頻 4 物項目集中與 P1 房產均價區間具有較大關聯性的關聯法則有 23 種，與 P2 房產均價區間具有較大關聯性的關聯法則有 17 種，而與 P3 房產均價區間具有較大關聯性的關聯法則有 17 種。所建立起來的布林值關聯法則與複合層級關聯法則將在接下來的章節利用 45 筆不同於建立關聯法則的資料比對，驗證所建立起的關聯法則是否準確。

4.3 關聯法則結果分析

爲了確定上一章節的關聯法則是否能準確的預測到房產均價與特徵因子之間的關連，將利用上一章節所建立起來的布林值關聯法則及複合式關聯法則拿以此 99 筆資料外的 45 筆資料做討論。此 45 筆資料將列於附錄之中。而表 4.3a 爲未落在所建立的布林值關聯法則的資料。

布林值的關聯規則（Boolean association rule）

45 筆資料中滿足布林值的關聯規則的資料有 40 筆，下面爲不符合的 5 筆資料的數據。

表 4.3a　5 筆未落在關聯法則的資料

	交易總價	建物總面積	捷運站距離	公園距離	學校距離	醫院距離	房產均價
A1	8,700,000.00	161.31	**1,028.28**	272.51	300.93	138.93	53,933.42
A2	9,000,000.00	166.64	352.40	**272.84**	71.30	403.58	54,008.64
A3	9,850,000.00	144.07	245.30	179.54	**220.01**	358.14	68,369.54
A4	16,350,000.00	219.49	1,999.61	253.30	194.91	**2,115.19**	74,490.87

	交易總價	建物總面積	捷運站距離	公園距離	學校距離	醫院距離	房產均價
A5	10,800,000.00	76.96	987.05	**245.45**	278.64	691.53	140,332.64

在這 5 筆資料中我們可以發現，雖然其中均有一個屬性因子沒落在所建立的關聯法則中，但非關聯法則建立錯誤，而是說明了當選擇了此類建物均間區間內的房屋時，可能會有一類屬性因子並未滿足需求，但並非全部選擇此房產均價區間時擁有的屬性因子不落在所建立的關聯法則中，且在布林值關聯法則建立的定義之下，並非所有特徵因子均會落在其中，而是說明了，倘若選擇了這條布林值關聯法則，購屋消費者可以選擇某些到特徵(屬性)因子。所以，建立起的布林值關聯法則具有一定的可靠性。

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

在此所建立的複合維度關聯法則，45 筆比對資料均落在其中表示所建立的複合維度關聯法則有相當的可信度，而且複合維度的關聯法則不同於 Apriori 演算法建立的布林值關聯法則，原因在布林值的關聯法則中，是將房產均價固定於 Apriori 演算法中，不高於最小支持度與最小信賴度的關聯法則刪除，這樣可能會造成一些潛在關聯法則遺失，且建立出來的關聯法則也較爲單調。所以相對於傳統布林值的關聯法則只討論單一維度的方法與複合維度的關聯法則比照起來，所建立起來的複合層維度關聯法則也較爲廣泛且多樣化。而且 Apriori 演算法在尋找關聯法則時，較複合維度關聯法則來的耗時，不但更增長了尋找時間，還得每一尋找出候選項目集後掃瞄整個資料庫，實爲浪費時間與人力物力。而複合維度的關聯法則在建立過程中，不但可以觀察到更多特徵因子間的關聯，且建立起來的複合層級關聯法則也較爲人性化，不但提供了購屋消費者更多的選擇外，也讓房地產業者有更多企畫方案呈現，不但可以刺激房地產業者的發展，更說明了現代人對房產的需求。但是，Apriori 演算法結合布林值並非全無好處，此方法雖較爲費時，但卻是發展關聯法則的根本，許多在 Apriori 演算法之後延伸尋找關聯法則的方法都由此開始，所以雖較爲費時，但仍有其必要存在的意義。