Journal of Science and Engineering Technology, Vol. 1, No. 1, pp. 47-59 (2005)

以算數平均數應用在門檻值制定之研究

李德治　鄧安生
大葉大學資訊管理學系
彰化縣大村鄉山腳路 112 號

摘　要

關聯法則的產生過程大致可分爲二個步驟：第一個步驟是產生大項目集合。第二個步驟是依據第一個步驟所產生的大項目集合來產生規則。本研究主要是針對第二個步驟產生的規則所制定的門檻值問題來探討。有關門檻值的訂定有絕對門檻值與相對門檻值兩種，並無一定的標準可供依循。相較於相對門檻值而言，絕對門檻值較不具意義，但相對門檻值的產生卻需耗費較多的電腦運算時間。有鑑於此，本研究嘗試利用算數平均數的原理並配合線性內插公式訂定門檻值，這個方法不但具有相對門檻值的意義且節省電腦運算時間，同時也具備絕對門檻值之意義。

關鍵詞：資料採礦，關聯法則，大項目集合，支持度，信度

A Study on Obtaining an Association Rule Threshold by the Mean Value

DE-CHIH LEE and AN-SHENG DENG
Department of Information Management, Da-Yeh University
112 Shan-Jiau Rd., Da-Tsuen, Changhua, Taiwan

ABSTRACT

An association rule process usually has two steps: first, producing a large itemset, and second, according to the large itemset producing a rule from the first step. The first step is the bottleneck of an algorithm. Many researchers have studied this problem. This report focuses on the probability of the rule from the second step. In order to make the rule more meaningful, the rule must be greater than the given threshold of support and confidence. There are two ways to obtain the threshold, namely absolute and relative; however, there is no standard rule to be followed. Therefore, in this research we attempt to use the mean value and a linear interplant to modify the threshold. By using this method, many meaningful results as well as the absolute threshold are conserved.

***Key Word*:** data mining, association rule, large itemset, support, confidence

一、緒論

近年來由於網際網路蓬勃發展，硬體設備與資料庫之技術亦不斷地進步，推陳出新，這使得儲存媒體所儲存資料的數量也愈來愈龐大。對於資料的擷取，不僅可輕易的從企業內部的資料庫中取得，亦可透過網際網路即時獲取外部的資料。隨著時間的累積，各企業組織的電腦系統中所儲存的資料量不斷地快速增加。對企業組織而言，這些資料庫系統除了包含大量資料外，同時蘊藏著豐富且有用的資訊。針對這些龐大的資料，我們可以透過特殊的技術深入加以處理與分析，找出隱藏的有用資訊或知識，提供企業組織做爲制定決策之參考。

現今企業所面臨問題已由資料不足轉變成資料過多，然而我們可透過篩選（select）、資料豐富化（enrich）等技術整合資料，建立企業資料倉儲（data warehouse），解決所面臨的資料龐大問題。並運用『資料探勘（data mining）』技術重新賦予這些資料生命讓企業擁有更豐富的商業智慧。

在資料倉儲內的資料，往往隱藏著某些特徵（patterns）以及關係（relations）。如果使用傳統資料查詢和統計功能，並不容易找出它們的關聯性，因此便引發出資料探勘的相關技術。資料探勘技術是經由自動或半自動的方法挖掘及分析大量的資料，尋找出有效的模型及規則 [11]，並利用這些模型與規則從龐大的資料庫中萃取出有用的資訊。

近幾年來有許多學者們從事資料探勘等相關技術的研究，其中有一項被廣泛討論的議題就是從交易資料庫中挖掘關聯法則。關聯法則主要是在協助尋找資料庫中資料與資料間的相互關係。其最初的應用是在市場購物籃分析（market-basket analysis）上 [5, 9, 17]；例如：10% 的顧客買麵包後接著會買牛奶，利用關聯法則的概念記成“麵包⇒牛奶”（support: 10%，confidence: 60%）的形式。也就是說在資料庫中有 10% 的交易包含麵包和牛奶，而包含麵包的交易中，有 60% 亦包含牛奶。其參數分別爲支持度（support）及信度（confidence），此兩個參數可避免找到沒有意義性與有用性的規則。支持度用來限制所找到的規則必須高於一定的比例，也就是有足夠的代表性與意義性；而信度則是用來表示這兩個項目集合間交互關係的相關程度。支持度與信度都需要由資料探勘的專家來給定，或是由管理者依其需要給定之，也就是支持度與信度必須大於最小支持度（minimum support）與最小信度（minimum confidence）的門檻值（threshold），找出的規則才具有它的意義。

由於關聯法則門檻值的設定是人爲所制定的，依研究者與研究方向的不同，會有不同的門檻值，所以可知門檻值的設定不易，太高太低都有相關的問題產生。門檻值過低會找出過多規則，門檻值太高會錯失許多具有價值規則的缺失。本研究將針對此問題提出新的方法做修正，以便能以較快的速度與較具備意義的找出具有價值的重要規則。

此外，倘若使用的交易資料庫產品項目過多，會增加相關項目組的計算，而且大部份均是不重要之規則，所以我們可以先針對資料做層級的規劃 [4-6]，至於如何去定義層級的類別及產品之間的關係，亦是值得研究的重點之一，本研究暫不探討。

有關聯法則之演算法已有眾多的學者做深入的探討與研究，使得演算法的執行效率越來越好，故本研究暫時不將重心放在演算法之改良，僅針對門檻值的問題提出一個較佳且迅速的制定法，我們以找出候選項目集合（candidate itemset）之次數爲基礎，再利用算數平均數逐次刪減不重要之候選項目集合，即平均項目集合分割法（mean itemset divide method, Midm），本研究將利用此法以求得重要之規則。

以下是本研究預期所要達到的具體目標：

1. 以找出候選項目集合之次數爲基礎，利用算數平均數逐次刪減不重要之候選項目集合，即平均項目集合分割法，吾人將利用此法以求得重要之規則。
2. 本研究利用微軟產品中的範例資料庫，首先對產品資料做分類分層的規劃，求得最重要之類別後，再繼續往下一層做探勘的動作，以求得更細項之重要規則。
3. 與傳統門檻值設定方法比較資料探勘的優缺點。

二、文獻探討

隨著科技的進步，人類的生活型態愈加複雜，在商業交易的過程中經過日積月累後，企業所需處理的資料量也變得非常龐大，使得資料探勘技術在近幾年來愈受重視。資料探勘可結合演算法及各種統計方法分析資料，從大型資料庫中挖掘出有用的資訊與知識，以提供未來決策支援與預測。

對於資料探勘的定義，以下是三個常被引用的定義：

1. 探勘是一個確定資料中有效的、新的與可能有用的，並且最終能被理解的模式的重要過程 [8]。
2. 資料探勘是爲要發現出有意義的樣型或規則，而必須從大量資料之中以自動或是半自動的方式來探索和分析資

料 [11]。

資料探勘是一種新的且不斷循環的決策支援分析過程，它能夠從組合在一起的資料中，發現出隱藏價值的知識，以提供企業專業人員參考 [10]。

資料探勘所要處理的問題，就是在龐大的資料庫中尋找出有價值的隱藏事件，並且加以分析。而其主要的貢獻在於它能從資料庫中獲取有意義的資訊以及對資料歸納出有結構的模式，以做爲企業在進行決策時之參考依據。Fayyad et al. [7] 將資料探勘納入了資料庫知識發現（knowledge discovery in database, KDD）[12, 23] 的流程之中，是屬於資料庫知識發現中最重要的一環 [7]。從大量資料中歸納出資訊與知識的方法，其中可能包含了關聯式法則、時間序列、人工智慧、統計、資料庫等方式 [1]。

在資料探勘的研究領域中，已有許多技術或方法被提出，這些技術或方法各應用在不同的需求上，而每種方法又有不同的演算法。資料探勘時需依照所應用的領域，選用適合的方法，才能有效的找出有用的資訊。常見的資料探勘型態有下列幾項：關聯法則（association rule）[5, 9, 17]、分類（classification）、推估（estimation）、預測（prediction）、群集偵測（cluster detection）等等 [6, 11]。

（一）關聯法則

關聯規則是資料探勘中最早被提出的理論技術，所以也是目前在資料探勘中最爲成熟和利用最多的一個領域，也最被許多學者廣爲討論的技術，並且已經有相當多的研究被提出。這個方法最早由 IBM Almaden Research Center 的 Agrawal et al. 提出的 [7-8]。關聯規則主要是針對交易資料庫加以分析以便找出隱藏其中的資訊與顧客的購買行爲模式，例如 10% 的顧客買麵包後接著會買牛奶的情形，利用關聯法則的概念記成“麵包⇒牛奶”（support: 10%，confidence: 60%）的形式。

每一條規則都有支持度及信度這兩個參數，用來判斷所找出的關聯法則是否有意義：支持度爲資料庫中 X∪Y 的交易記錄所佔百分比，記作 support（X∪Y）；而信度則是定義此關聯法則可信的程度，也就是 X 出現的條件下，Y 也會跟著出現的條件機率，記作 support(X∪Y)/ support(X)。依照條件機率，若某關聯法則的信度超過一定限度時，其意義爲若此交易包含 X，有很高的機率會包含 Y。而一個有效的關聯法則，其支持度及信度必須要大於或等於研究者所訂定之最小支持度及最小信度，只有滿足條件之關聯法則才具有意義性與代表性。

也就是說，關聯法則必須要找出所有 X⇒Y 形式的關聯法則，並且滿足下列條件：

1. support (X⇒Y) = support (X∪Y) ≥ min_support
2. confidence (X⇒Y) =support (X∪Y) / support (X) ≥ min_confidence

相關項目集合產生時，因爲仍未計算其支持度與信度，所以並不知道此項目集合是否大於或等於使用者所定之最小限制條件，此時的相關項目集合稱之爲候選項目集合，之後再計算其支持度與信度，假如滿足使用者所定之最小支持度與信度的限制，稱此候選項目集合爲大項目集合（large itemset）。再藉此大項目集合，推導出關聯法則。

就目前的大項目集合產生方式，計有 SETM、AIS、Apriori、AprioriTid、AprioriHybrid [7-8]、特性項導向法 [6]、雜湊法（hash-based）[9-11] 等。不管是哪一種演算法，均需設定門檻值，才能找出大項目集合。對於演算法的改良與新演算法的提出，已有許多學者從事相關研究，在此便不再針對演算法進行相關研討。本研究採用其中之一種演算法－Apriori，以新的方法制定門檻值，並利用吾人所提出的方法進行挖掘關聯法則並與相對門檻值作一比較。有關 Apriori 演算法主要包含兩個步驟：

1. 在資料庫中尋找出所有可能的大項目集合，並且要大於所設定的最小支持度。
2. 分析所產生的大項目集合，產生適當的規則。

其詳細執行步驟如下：

（1）掃描資料庫 D，找出大於等於最小支持度之項目集合，稱此項目集合爲 L_1，其中 1 代表長度爲 1 之大項目集合。

（2）利用長度爲 k-1 的大項目集合（Lk-1）來產生候選項目集合（Ck）。

（3）計算所有候選項目集合的支持度，判斷是否大於等於最小支持度，將符合條件之候選項目集合挖掘出來，成爲長度爲 k 之大項目集合（Lk）。

（4）一直重覆以上步驟，直到無法產生新的候選項目集合，宣告停止。

在圖 1 的交易資料庫中，可以知道所存在的物品項目有五種，以及四筆的交易筆數。在圖 2 中，可以很清楚的看到 Apriori 演算法的運作過程，它先以所有的物品項目當成第一階段的候選項目集合（C_1），若是大於使用者自行訂定的

Database D

TID	Items
100	ACD
200	BCE
300	ABCE
400	BE

圖 1. Apriori 的交易資料庫範例

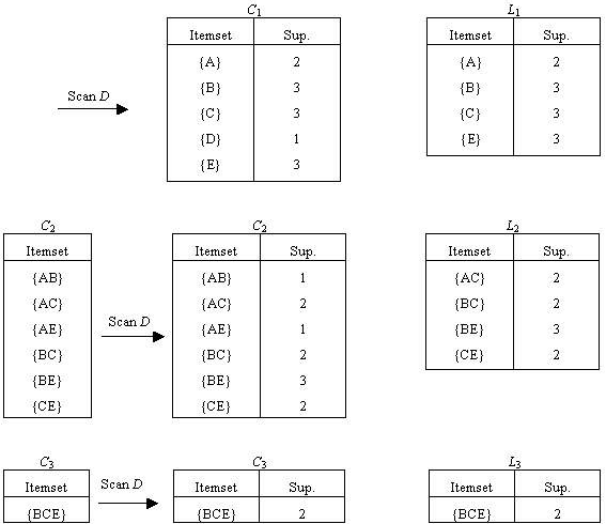

圖 2. 產生候選項目集合及大項目集合

最小門檻值（即 support），在此設定爲兩次，即可成爲第一階段的大項目集合（L_1），由圖中可知有 {A}、{B}、{C}、{E} 四個物品項目符合最小門檻值的限制。

接下來繼續進行第二階段，長度爲 k 的候選項目集合的產生（C_k）是由上一階段的大項目集合（L_{k-1}）做排列組合，也就是說 C_2 是由 L_1 所排列組合而成，在最小門檻值的訂定中，同樣地設定爲兩次，一共產生了 {AB}、{AC}、{AE}、{BC}、{BE}、{CE} 六個候選項目集合。

第三階段則是要產生 C_3 進而得到 L_3，而在 L_2 中，有兩個項目集合具有相同的第一個物品項目，即 {BC}、{BE}，而 {CE} 也存在於 L_2 中，這些過程則是經過不斷的排列組合及資料庫存取，產生 {BCE} 的大項目集合 L_3。因爲 L_3 只有一組候選項目集合，無法再繼續產生 C_4，Apriori 演算法則即告終止。

（二）廣義關聯法則

以上所探討之關聯法則，其所代表的意義都是各交易項目之間彼此的關聯性，這些項目是資料庫中原始的資料，是最底層且詳細的資料，藉由關聯法則的方法，可挖掘出一個很精確的結果。但是有時過於精確的結果，不一定是有用的資料，所以有時會採取趨向於比較明顯或大範圍的方向進行資料分析。

爲了找出比較明顯與較大趨勢之關聯法則，可採用廣義關聯法則（generalized association rules）的概念，將項目分類（taxonomy）的資訊加入關聯法則探勘，其好處是可將分類有關的資訊反映在關聯法則探勘上，充分的表示階層間的關係，並可得知不同階層間項目的關聯性。

廣義關聯法則指的就是較大階層項目間彼此的關聯性。以圖 3 爲例，“買 Jackets 也會買 Shoes”，“買 Ski Pants 也會買 Shoes”，這種是最底層且最詳細的關聯法則結果。若是以類別型態的較大階層項目間彼此的關聯法則，則可改成“買 Outerwear 也會買 Shoes” 這種較大趨勢與概括性的關聯法則。

使用廣義關聯法則，各項目間必須加以整理，每個項目屬於哪一個分類項目必須事先定義清楚，才能推導廣義關聯法則，它是利用 Taxonomy（is-a hierarchy）架構來定義，如圖 3 的範例。

廣義關聯法則便是將交易中的項目加上分類項目，包括分類中的葉節點（leaf node，即一般項目）及內部節點（internal node，即廣義項目）。利用定義好的階層關係，各個不同階層之間的關聯法則便可被推導。

（三）關聯法則之缺點

關聯法則最初是由 IBM Almaden Research Center 的 Agrawal et al. [2] 於 1993 年所提出的，最常依循的方法是 Apriori 演算法。Apriori 演算法在 1994 年時，由 Agrawal et al. [4] 所提出。它利用簡單且循序的方式，來找出資料庫中物品項目間彼此的關係與關聯性，以形成規則。後續提出有關關聯法則的演算法大都以 Apriori 演算法爲基礎加以延伸改

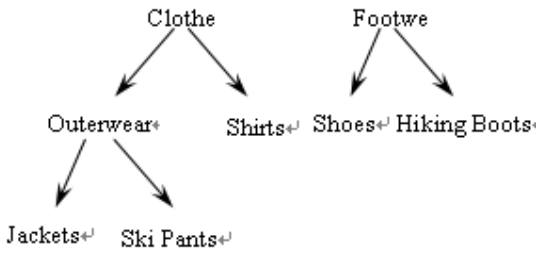

圖 3. Taxonomy 範例

善，Apriori 演算法主要包含兩個步驟：

1. 在資料庫中尋找出所有可能的大項目集合，並且要大於所設定的最小支持度。
2. 分析所產生的大項目集合，產生適當的規則。

由於第一個步驟是演算法的瓶頸所在，爲了有效率及快速的找出大項目集合，已有許多學者針對此問題進行相關的研究，也提出了許多不同的演算法。其中不僅有增進執行效率之演算法，還有針對不同領域與不同資料而提出各個適合的演算法，不同的演算法的特性均不同，但其目的卻是相同，即找出大項目集合，以利用找出的大項目集合，進而產生關聯法則。不管是何種演算法，其都有不變的最小支持度與最小信度，也就是門檻值是固定不變的。門檻值的訂定，一般來說都是均由研究者自行給訂，但是不同的研究領域與不同的研究者，所訂定的門檻值必然會不同。以數學的角度來看，門檻值的訂定，應透過專業的數學基礎加以推導比較客觀。在過去研究者在進行研究實驗的時候，常常因爲資料不足而必須強迫降低門檻值。如此一來，會因爲門檻值的訂定有一定程度上的偏差，而有相關的問題產生。

有關最小支持度與最小信度的訂定大多任意給定或使用相對值，無一定的標準，可隨研究者視研究情形彈性的自行訂定，有的甚至隨意調整門檻值的大小，以找出不同的關聯法則。許多研究人員在進行研究實驗時，在門檻值訂定的步驟，大部份都會遭遇到許多的瓶頸，除了來源資料不足，另外在設定門檻值之後仍必須經過反覆的檢視所產生的規則，檢查其是否有值得注意的規則出現。由此可知門檻值設定的不易，太高太低都有相關的問題產生。有關傳統關聯法則門檻值設定的缺點與問題，可歸納出以下四點：

1. 門檻值太高，可能遺漏重要之規則。
2. 門檻值太低，可能找出不重要之規則。
3. 若找不出規則時，則須不斷地調整門檻值，重覆進行挖掘試驗。
4. 若採用相對門檻值，則資料必須排序浪費電腦運算時間。

爲改善上述之缺點，故本研究提出利用平均數篩減的方法來制訂門檻值，透過此方法所獲得之門檻值將較絕對門檻值更具意義，同時運算速率較相對門檻值制訂法快速。

三、研究方法

（一）廣義關聯法則概念於資料分類上

本研究引用了廣義關聯法則的概念，依照相同性質屬性加以分類，即在交易中的項目加上分類項目，包括分類中的葉節點及內部節點。利用定義好的階層關係，便可推導各個不同階層之間的關聯法則。因資料庫取得不易，故本研究利用微軟 access 所附的交易資料進行分析。在本研究中不探討討廣義關聯法則之分類方法，其分類之階層按此一資料庫之內定分類爲依據。同時本研究重點在於提出新的門檻值制訂方法，故有關聯法則演算法的改進本研究亦不予討論。有關交易資料庫的分類（taxonomy）如圖 4 所示。

（二）研究基本假設

有關關聯法則的挖掘技術，仍然存在許多的問題與瓶頸，例如門檻值的訂定便讓許多研究人員不知如給定，甚至在訂定門檻值時無任何的理論根據，以致於發生挖掘出的關聯法則不是太多就是太少的情形，本研究即針對此問題加以深入的研究探討，基本假設如下：

1. 不考慮數值屬性關聯法則。
2. 不考慮限制性關聯法則。
3. 不考慮負面性關聯法則。
4. 資料量夠大，且分配均勻。

本研究之演算法架構在 Apriori 之上，並引進廣義性關連法則概念將物品項目分類。

（三）平均項目集合分割法

爲了能夠兼顧相對門檻值的意義與絕對門檻值的迅速，本研究嘗試提出一個新的方法－平均項目集合分割法，利用算數平均數配合線性內插的概念來挖掘出大項目集合，減少掃描資料庫的次數以及計算支持度與信度的煩瑣過程。

在相關文獻中，Apriori 演算法部份並沒有明確的告訴研究者如何訂定最小門檻值才是最理想的，也沒有理論依據來訂定。相關文獻僅說明視原始資料的情況與需要何種資訊等等，再由 Data Mining 的專家來給定，或是由管理者依其

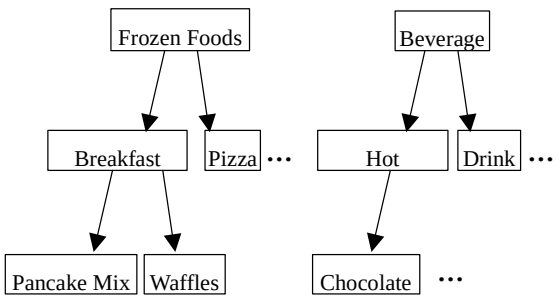

圖 4. 交易資料庫之 Taxonomy

需要所給定之，無一定的規定與標準。

本研究所提出的方法，即有感於最小門檻值的訂定非常地困難，而且也無一定的標準可依循，在同樣的資料、同樣的研究領域、同樣的資訊需求與其他狀態均相同的情況下，可因研究人員的不同，導致門檻值的訂定也有所不同。爲了解決此問題，本研究提出了平均項目集合分割法，其步驟如下：

1. 掃描資料庫 *D*，計算出各個項目集合間的出現次數，並計算出長度爲 k 的項目集合之平均次數。
2. 以項目集合之平均次數爲準，找出大於等於平均次數之候選項目集合。
3. 將上一步驟所得的候選項目集合，計算出平均次數，再以此平均次數爲準，找出大於等於平均次數之候選項目集合，且只需掃描上一步驟所得的候選項目集合。
4. 一直重覆 2、3 步驟，當資料量夠大且分配均勻時，在各階段所求出之候選項目集合的比例呈現將趨近於：前 50%、25%、12.5%、6.25%...等等。
5. 最後依據決策者需要多少百分比以上的重要規則，利用線性內插的方式，和相差最近的前後兩個百分比之候選項目集合，計算其比例差距，即可得出長度爲 k 之大項目集合。此百分比即爲最小門檻值。

由於本研究所提出的方法，並不需一開始即訂定門檻值的限制，只需決定所需多少的重要規則，再依其制定最小門檻值。當資料庫非常龐大時，花費在找出關聯法則的時間會很浪費時間。若一開始制定的門檻值所找出的規則不如預期，則必須重定門檻值再找一次規則，且下次所找出的規則又不一定是所要之重要規則，無形中浪費相當多的時間。

在平均項目探勘法中必須計算其相關項目在資料庫中出現的次數和平均次數，大於等於平均次數即成爲候選項目集合，逐次比對之掃描資料庫的次數分別爲：n、$\frac{n}{2}$、$\frac{n}{4}$、$\frac{n}{8}$、…。掃描資料庫的次數相較於傳統方法減少了許多，執行的效率也大大的提高，而且也可求得在候選項目集合中的任一比例做爲大項目集合。

當交易資料非常龐大時，經過計算之後所得的候選項目集合數量也會非常地多，然而欲在如此龐大的候選項目集合之中，訂定一個最小門檻值，並非是一件容易的事。不管計算後的候選項目集合是以次數或是百分比的方式表達，都將因候選項目集合的數量太多，而必須先經過排序的動作，才能較容易的訂定門檻值。如果不先經過排序，是比較無法一次就能訂定出一個最佳的門檻值，所以只能靠專家的專業知識與經驗來決定。但此情況下，通常是經過多次改變門檻值與不斷地重複篩選，才能訂出一個較合理的門檻值，無形中浪費很多時間。

有關排序演算法已有許多不同的方法可以選擇使用，在此列舉幾項於表 1 中，以平均時間來看，最快的排序演算法的時間複雜度爲 O($n$log$n$)，最慢爲 O(n^2)。但平均項目集合分割法所處理的步驟，第一次是處理 n 筆資料，第二次以後的資料量便逐次減少，因此在計算時間複雜度時，平均花費時間複雜度爲 O(n)。利用平均項目集合分割法訂定門檻值不須經過排序，使處理的時間增快許多。

圖 5 即是平均項目集合分割法之示意圖，每回合均需找尋其平均次數，大於等於平均次數之項目集合將做爲下一回合之候選項目集合，如此反覆，以漸近式的概念找出出現次數最多的大項目集合，不同比例的大項目集合，可做爲不同的資訊需求，提供不同的解決方案。在第 4 部分實驗與結果評估將針對此方法做一實驗，將本研究所提出的方法實作出來，以驗證所提之理論。例如要找出超過 90% 的規則，則可在 $\overline{x}_3$ 與 $\overline{x}_4$ 之間，利用線性內插的方式計算出比例差距，經過運算後而得到結果，即我們想要的絕對門檻值，大於此門檻值的項目集合，便是我們要的關聯法則。

（四）演算法實作

首先由使用者輸入欲找出多少規則，x 介在 0~1 之間，即表示找出的規則是 x 百分比以上之規則。

$$1-\left(\frac{1}{2}\right)^{n-1} < x < 1-\left(\frac{1}{2}\right)^{n} \qquad (1)$$

表 1. 時間複雜度之比較

		時間		
		平均	最差	最佳
排序演算法	Insertion	O(n^2)	O(n^2)	O(n)
	Bubble	O(n^2)	O(n^2)	O(n)
	Quick	O(nlogn)	O(n^2)	O(nlogn)
	Merge	O(nlogn)	O(nlogn)	O(nlogn)
	Heap	O(nlogn)	O(nlogn)	O(nlogn)
平均項目集合分割法	Midm	O(n)	O(n)	O(n)

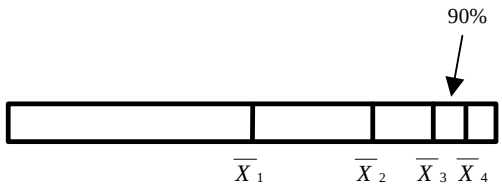

圖 5. 平均項目集合分割法概念圖

其中 n 爲計算平均數的次數

解方程式（1）可得

令 $t = \log_{\frac{1}{2}}(1-x)$

$$\begin{cases} n = [t]+1 & ,t \notin Z \\ n = t & ,t \in Z \end{cases} \qquad (2)$$

其中：[ ]爲高斯符號。

求出比對次數 n 之後，再利用線性內插，即可求出給定的門檻值 d。

$$\begin{cases} d = \overline{d}_{n-1} + \dfrac{x + \left(\frac{1}{2}\right)^{n-1} - 1}{\left(\frac{1}{2}\right)^{n-1} - \left(\frac{1}{2}\right)^{n}} \times \left(\overline{d}_n - \overline{d}_{n-1}\right) & ,t \notin Z \\ d = \overline{d}_n & ,t \in Z \end{cases} \qquad (3)$$

四、實驗與結果評估

（一）關聯法則之探勘

爲能順利進行研究與方便起見，本研究採用微軟公司在 Access 產品中所附的交易範例資料（FoodMart 2000.mdb）。此外，利用隨機亂數來產生項目集合資料，以便進一步探討平均項目分割法所得的結果與實際在資料所佔百分比做一比較，以驗證其適用性。

在 Access 範例資料庫中包含了兩年的銷貨資料，產品項目共有 1,560 種，資料筆數約有 25 萬筆資料。爲了能順利進行資料探勘，按照原有之資料的屬性及性質，將產品區分成 110 種的產品類別，如此可以減少執行時間，同時可探討產品的廣義項目之概括性的規則。在此範例資料庫中的 110 種產品類別，是已經被預先定義。

從實驗得知，進行第一次的篩選後大於平均數的項目集合有 1887 條規則。接著再以 1887 條規則爲處理來源，繼續進行重覆的動作。我們所找到的項目集合依序有 615、201、66 條。若不斷地重復分割的步驟，將使得規則數目逐步地減少。當規則數目愈少時，表示此規則爲一重要之規則。因爲當門檻值愈高，規則數目就愈少，而規則數目愈少時，通常代表著其規則愈重要。

我們除了利用平均項目集合分割法挖掘出每次分割後所得到的規則與其在資料庫中實際所占全部項目集合之百分比外，另外計算出資料庫的資料分佈情形的一些相關統計參數（parameter）；包括標準差（standard deviation）、變異數（variance）、偏態（skewness）及峰態（kurtosis）等。這些統計參數能讓我們進一步觀察出整個資料庫的分配情形，同時可提供後續的研究人員做進一步的研究與分析。以統計的觀點來看，當資料量足夠大且趨近對稱分配時，平均數（meam）會趨近於中位數（median）。但在實際上平均數並不一等於中位數。我們大膽預測其誤差可能與偏態值或峰態有某種關聯，故將其相關資料與百分比列於表 2 中。在表 2 中可看出本研究使用範例資料庫的分析結果，其標準差爲 12.5858、偏態爲 4.333、峰態則是 37.923。由此可知，此資料庫的項目集合次數是呈現右偏和高峽峰的情形，即大部份的項目集合會集中在某一範圍內，出現極不平均的現象，因此經平均項目集合分割法處理後，每次分割所得到的項目集合數目實際上占全部項目集合之百分比依序爲：0.3148、0.1026、0.0335 及 0.0110，即表 2 中的 $\overline{X}_1$、$\overline{X}_2$、$\overline{X}_3$ 及 $\overline{X}_4$，與理想值的 0.5、0.25、0.125 及 0.0625 仍有些差距。因此我們猜測其差距可能跟偏態與峰態有某種關係，其相關的資料列於表 2。

在計算出 $\overline{X}_1$、$\overline{X}_2$、$\overline{X}_3$、$\overline{X}_4$ …之後，依照決策者輸入的百分比數值，再利用線性內插的方式求出門檻值。例如要找出超過 90% 的規則，則可在 $\overline{X}_3$ 與 $\overline{X}_4$ 之間，利用線性內插的方式計算出比例差距，經過運算後而得到的結果，即我們想要的絕對門檻值。大於此門檻值的項目集合，便是我們所要的關聯法則。由於篇幅的關係，無法將超過 90% 的

表 2. 範例資料庫分析結果

標準差 (SD)	變異數 (Var)	偏態 (SK)	峰態 (K)	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_3$	$\overline{X}_4$
12.5858	158.4024	4.3330	37.9230	0.3148	0.1026	0.0335	0.0110

全部規則列出，僅列出 15 個被購買次數最多的關聯法則。其由高到低依序列舉如下：

- Fresh Vegetables ⇒ Fresh Fruit
- Fresh Vegetables ⇒ Dried Fruit
- Cheese ⇒ Fresh Vegetables
- Soup ⇒ Fresh Vegetables
- Cookies ⇒ Fresh Vegetable
- Wine ⇒ Fresh Vegetables
- Fresh Fruit ⇒ Dried Fruit
- Soup ⇒ Fresh Fruit
- Fresh Vegetables ⇒ Canned Vegetables
- Cheese ⇒ Fresh Fruit
- Nuts ⇒ Fresh Vegetables
- Cookies ⇒ Fresh Fruit
- Wine ⇒ Fresh Fruit
- Frozen Vegetables ⇒ Fresh Vegetables
- Paper Wipes ⇒ Fresh Vegetables

以上是將產品項目進行分類之後，利用 Apriori 演算法以及平均項目集合分割法所找到最重要的 10 個規則。這個方法不但能在眾多的產品項目做一有效率的分類，且能看出較大趨勢之關聯法則，提供給決策人員做決策之參考依據。

（二）關聯法則之探勘

在實驗的過程中，由於只有取得一個範例交易資料庫，故我們進一步利用隨機亂數產生摸擬資料，以便對本研究所提的方法進一步做詳細的分析。由於本研究是探討關聯法則門檻值之制定，因此在產生摸擬資料時，並不是產生如同範例交易資料庫的資料一樣，而是必須符合本研究方法所需要的資料。一般在交易資料庫中，須先經過計算出候選項目集合的次數，方能依照所訂定之最低門檻值而篩選出大項目集合，然而本研究是著重在門檻值制定之研究。因此在產生摸擬資料時，即直接產生項目集合資料庫，使得系統能立即的計算出最佳門檻值，減少電腦運算時間。

本研究利用程式產生模擬資料，使用隨機亂數產生一萬至五萬筆的項目集合資料庫各 10 個，每個資料庫只存有項目集合之次數，根據這些次數資料，做爲提供本研究之資料來源。產生的資料庫結構及資料如表 3。由於本研究是在研究門檻值之制定，故不考慮關聯法則之長度爲何，均適合用本研究方法找出門檻值。在表 3 中，即實際資料庫存放資料的範例，存放有項目集合編號（itemset.no）及支持度（support）。

表 3. 項目集合資料庫範例

Itemset.no	Support
1	10
2	20
3	15
4	22
5	36

當資料庫產生完畢之後，再將支持度做必要的計算，求出必要的數值，以了解每個資料庫的大概情形，表 4 即是由表 3 所計算的敘述統計量結果。

依照表 2 產生模擬資料之後，再計算其相關的統計量，結果顯示在表 4，本研究即根據此步驟一一計算出所產生的模擬資料庫之統計量，以便能夠清楚的知道每個資料庫的特徵，更可利用這些相關的數據資料，提供爲後續之研究。表 5 即是計算模擬資料庫之後，所得到的敘述統計量結果。

（三）實驗分析

本研究用程式產生模擬資料，使用隨機亂數產生一萬至十萬筆的項目集合資料庫，再利用本研究方法，分別計算出不同資料庫之相關統計參數數與該筆資料在資料庫中之實際百分位數。我們企圖在後續研究中，進一步找出規則性，以便能發展出更快速的門檻值制定法。

利用亂數產生模擬資料於 100 個資料庫中，並以本研究所提出的平均項目集合分割法爲依據，對這 100 個資料庫進行運算處理，從經過處理完畢的資料庫中，分析其產生的結果。本研究爲了驗證所提之方法，必須有大量的資料做爲資料來源以供測試，受限於時間的關係僅以 100 個模擬資料庫，做爲本研究的資料來源，實驗結果如表 6 所示。在表 6 與表 5 中，兩表格中每個編號即代表一個資料庫，兩表的編號若相同即代表同一個資料庫。兩個表格不同的地方，在於表 5 記錄著每個資料庫的統計量，表 6 則是利用平均項目集合分割法針對每個資料庫做運算，得到相關的數據資料。其數據資料所代表的是每次分割後所得到的項目集合佔全部項目集合之百分比。

表 4. 項目集合資料庫範例之統計量

項目集合數	總和	平均數	標準差	變異數	偏態	峰態
5	103	20.6000	9.7877	95.8000	1.0170	1.5150

表 5. 模擬資料庫之統計量

編號	平均數	標準差	變異數	偏態	峰態
1	4196.6460	2044.9509	4181824.2	-0.259	2.046
2	3582.3707	2183.6949	4768523.6	0.111	1.779
3	4284.2746	1996.0442	3984192.4	-0.311	2.139
4	3461.5209	2221.2397	4933905.7	0.189	1.758
5	2351.6343	1974.4302	3898374.6	1.055	3.056
6	3185.5548	1963.1325	3853889.0	0.294	2.173
7	3160.2209	1949.4809	3800475.9	0.314	2.210
8	3301.6426	2031.5103	4127034.0	0.248	2.045
9	4144.3069	2193.8938	4813170.0	-0.248	1.809
10	3363.8475	2054.3806	4220479.5	0.212	2.008
11	3559.5299	2181.8258	4760364.0	0.119	1.780
12	4191.0186	2024.4184	4098269.9	-0.253	2.073
13	2064.1941	1808.7047	3271412.8	1.380	4.118
14	2286.2159	1926.1540	3710069.4	1.129	3.285
15	4276.5626	1983.5723	3934559.0	-0.292	2.137
16	3989.9257	2205.7107	4865159.7	-0.153	1.763
17	5319.7511	1885.4870	3555061.2	-1.228	3.579
18	3952.2002	2201.8933	4848334.1	-0.125	1.764
19	4100.4563	2203.9834	4857542.6	-0.222	1.791
20	5587.0286	1685.7264	2841673.6	-1.581	5.008
21	3161.2586	2192.5518	4807283.6	0.391	1.889
22	2717.6323	2127.9660	4528239.3	0.731	2.297
23	4253.5930	2000.1754	4000701.6	-0.272	2.114
24	2524.3290	2053.0343	4214949.8	0.898	2.653
25	2997.5102	2180.5095	4754621.8	0.514	1.994
26	4302.6242	2207.2817	4872092.6	-0.365	1.849
27	3284.5769	2028.0829	4113120.3	0.259	2.070
28	4917.5107	2073.8576	4300885.2	-0.843	2.532
29	3393.0272	2070.8931	4288598.3	0.206	1.974
30	4451.1387	2189.3474	4793242.1	-0.469	1.946
31	2809.3360	2143.9397	4596477.6	0.654	2.176
32	4011.5080	2098.4319	4403416.4	-0.152	1.912
33	2819.5941	2146.8105	4608795.4	0.642	2.161
34	3166.4978	2200.2037	4840896.4	0.391	1.871
35	2306.4203	1943.8402	3778514.9	1.096	3.180
36	5094.3170	1992.5266	3970162.4	-1.010	2.935
37	5010.7156	2032.8505	4132481.3	-0.924	2.723
38	3199.4593	1989.0062	3956145.7	0.303	2.144
39	4211.7081	2198.4989	4833397.3	-0.301	1.819
40	4178.0659	2205.3250	4863458.5	-0.277	1.800
41	3576.6826	2193.5701	4811749.9	0.113	1.766
42	3424.8259	2211.1316	4889102.8	0.211	1.777
43	2400.6326	2004.1774	4016727.1	1.019	2.943
44	1973.3454	1741.6268	3033263.9	1.485	4.574
45	3752.8756	2165.6399	4689996.2	-0.005	1.807
46	5390.0682	1838.3392	3379491.0	-1.314	3.889
47	4090.8361	2202.7317	4852025.1	-0.222	1.788
48	3473.8411	2084.7577	4346214.9	0.163	1.945
49	3702.7624	2150.6487	4625289.8	0.028	1.824
50	5336.1594	1867.8774	3488966.0	-1.271	3.737

編號	平均數	標準差	變異數	偏態	峰態
51	2637.8570	2088.7658	4362942.6	0.799	2.454
52	4215.5595	2009.7603	4039136.6	-0.262	2.090
53	2681.6474	2109.4025	4449578.8	0.757	2.363
54	3992.3711	2107.6323	4442114.0	-0.144	1.911
55	2586.5139	2076.9436	4313694.7	0.839	2.524
56	4023.3525	2198.6007	4833844.9	-0.169	1.772
57	3355.4588	2048.4193	4196021.7	0.228	2.019
58	5336.3126	1873.4556	3509835.9	-1.264	3.693
59	4174.2735	2209.7735	4883098.9	-0.270	1.797
60	4856.1549	2098.9442	4405566.6	-0.792	2.420
61	2417.6794	2004.4249	4017719.2	0.996	2.894
62	2823.0982	2144.2026	4597604.8	0.642	2.171
63	2373.3110	1982.9523	3932099.7	1.042	3.021
64	3410.3743	2206.5433	4868833.4	0.218	1.778
65	2902.6489	2163.3660	4680152.4	0.580	2.075
66	4737.3738	2125.4087	4517362.2	-0.690	2.248
67	5506.8622	1753.0030	3073019.6	-1.464	4.494
68	3827.0332	2176.9621	4739164.1	-0.051	1.783
69	4985.1509	2049.1006	4198813.4	-0.909	2.676
70	3656.9661	2139.5431	4577644.8	0.058	1.842
71	3456.9330	2202.3940	4850539.1	0.187	1.779
72	3463.8092	2196.2571	4823545.0	0.185	1.777
73	4106.6670	2063.6704	4258735.5	-0.206	1.983
74	2544.5956	2054.5812	4221303.7	0.878	2.613
75	4350.9358	1961.8417	3848823.0	-0.324	2.195
76	3599.2687	2129.1590	4533318.0	0.092	1.857
77	4485.7369	2183.3378	4766963.9	-0.493	1.967
78	4409.2018	2192.6544	4807733.2	-0.442	1.918
79	3620.9910	2131.2041	4542031.0	0.077	1.858
80	4549.6801	2168.5916	4702789.5	-0.549	2.038
81	2786.5399	2138.0447	4571235.1	0.677	2.218
82	2623.1497	2081.8937	4334281.4	0.802	2.457
83	1877.3676	1653.9862	2735670.4	1.616	5.214
84	2405.8522	2000.0635	4000254.0	1.007	2.920
85	4121.0267	2060.1178	4244085.2	-0.214	1.996
86	5539.8237	1722.6479	2967515.8	-1.504	4.676
87	4598.7532	2165.2407	4688267.5	-0.581	2.078
88	4835.4717	2101.0569	4414440.0	-0.778	2.397
89	3627.1427	2133.4553	4551631.5	0.070	1.850
90	4088.1798	2209.0090	4879720.8	-0.220	1.779
91	2501.0701	2040.2371	4162567.3	0.917	2.696
92	3918.0498	2121.8318	4502170.0	-0.104	1.875
93	2444.4053	2018.8870	4075904.8	0.972	2.833
94	3522.1833	2195.7799	4821449.6	0.148	1.773
95	2197.2594	1887.6326	3563156.8	1.220	3.558
96	5622.9455	1654.7750	2738280.2	-1.615	5.211
97	4914.6387	2080.4219	4328155.1	-0.841	2.522
98	4313.6543	2195.0927	4818431.8	-0.378	1.869
99	3192.4568	1978.4442	3914241.5	0.304	2.159
100	4338.2111	2193.1708	4809998.1	-0.390	1.875

表 6. 實驗分析結果

編號	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_3$	$\overline{X}_4$
1	0.5214	0.2618	0.1308	0.0658
2	0.4880	0.2449	0.1206	0.0616
3	0.5256	0.2625	0.1310	0.0652
4	0.4789	0.2389	0.1208	0.0608
5	0.3380	0.1666	0.0830	0.0406
6	0.4826	0.1965	0.0971	0.0474
7	0.4857	0.1964	0.0997	0.0496
8	0.4774	0.2080	0.1018	0.0510
9	0.5314	0.2696	0.1342	0.0681
10	0.4861	0.2136	0.1074	0.0544
11	0.4875	0.2407	0.1204	0.0608
12	0.5181	0.2572	0.2483	0.2483
13	0.3504	0.1536	0.0764	0.0389
14	0.3404	0.1628	0.0814	0.0401
15	0.5201	0.2578	0.1279	0.0634
16	0.5208	0.2655	0.1347	0.0670
17	0.6545	0.3277	0.1656	0.0838
18	0.5147	0.2630	0.1314	0.0651
19	0.5287	0.2719	0.1365	0.0692
20	0.6310	0.3153	0.1571	0.0787
21	0.4442	0.2210	0.1108	0.0554
22	0.3722	0.1866	0.0929	0.0473
23	0.5159	0.2571	0.1294	0.0642
24	0.3378	0.1682	0.0844	0.0421
25	0.4177	0.2087	0.1045	0.0522
26	0.5548	0.2808	0.1402	0.0704
27	0.4820	0.2052	0.1017	0.0516
28	0.6505	0.3256	0.1621	0.0807
29	0.4847	0.2194	0.1090	0.0538
30	0.5721	0.2887	0.1441	0.0723
31	0.3889	0.1955	0.0969	0.0483
32	0.5123	0.2582	0.1297	0.0641
33	0.3927	0.1960	0.0969	0.0479
34	0.4430	0.2227	0.1109	0.0556
35	0.3378	0.1644	0.0815	0.0407
36	0.6660	0.3339	0.1667	0.0832
37	0.6643	0.3324	0.1669	0.0841
38	0.4787	0.1970	0.0974	0.0499
39	0.5420	0.2750	0.1385	0.0679
40	0.5374	0.2750	0.1368	0.0688
41	0.4859	0.2430	0.1210	0.0604
42	0.4719	0.2359	0.1193	0.0602
43	0.3331	0.1668	0.0841	0.0419
44	0.3612	0.1476	0.0726	0.0367
45	0.5004	0.2506	0.1249	0.0625
46	0.6489	0.3243	0.1638	0.0814
47	0.5305	0.2696	0.1351	0.0679
48	0.4868	0.2259	0.1134	0.0563
49	0.4973	0.2459	0.1222	0.0612
50	0.6525	0.3256	0.1643	0.0813

編號	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_3$	$\overline{X}_4$
51	0.3612	0.1789	0.0897	0.0450
52	0.5194	0.2578	0.1291	0.0642
53	0.3692	0.1829	0.0918	0.0462
54	0.5126	0.2558	0.1284	0.0644
55	0.3506	0.1749	0.0868	0.0434
56	0.5215	0.2653	0.1330	0.0674
57	0.4831	0.2131	0.1068	0.0527
58	0.6547	0.3275	0.1639	0.0814
59	0.5378	0.2735	0.1368	0.0693
60	0.6405	0.3204	0.1603	0.0795
61	0.3347	0.1665	0.0830	0.0419
62	0.3949	0.1969	0.0989	0.0503
63	0.3345	0.1663	0.0834	0.0419
64	0.4699	0.2349	0.1177	0.0587
65	0.4058	0.2026	0.1012	0.0507
66	0.6171	0.3087	0.1545	0.0771
67	0.6383	0.3180	0.1587	0.0801
68	0.5056	0.2545	0.1273	0.0634
69	0.6657	0.3315	0.1671	0.0831
70	0.4949	0.2426	0.1219	0.0611
71	0.4764	0.2369	0.1194	0.0597
72	0.4752	0.2379	0.1184	0.0590
73	0.5169	0.2580	0.1293	0.0645
74	0.3424	0.1705	0.0856	0.0430
75	0.5208	0.2603	0.1301	0.0652
76	0.4908	0.2392	0.1188	0.0593
77	0.5767	0.2908	0.1457	0.0727
78	0.5672	0.2871	0.1431	0.0718
79	0.4932	0.2392	0.1193	0.0602
80	0.5876	0.2949	0.1471	0.0736
81	0.3853	0.1929	0.0967	0.0481
82	0.3601	0.1791	0.0889	0.0444
83	0.3745	0.1361	0.0680	0.0338
84	0.3341	0.1669	0.0830	0.0411
85	0.5169	0.2588	0.1295	0.0643
86	0.6342	0.3172	0.1583	0.0795
87	0.5952	0.2999	0.1500	0.0750
88	0.6364	0.3186	0.1591	0.0795
89	0.4937	0.2413	0.1207	0.0601
90	0.5292	0.2692	0.1356	0.0677
91	0.3333	0.1662	0.0833	0.0416
92	0.5098	0.2539	0.1271	0.0639
93	0.3346	0.1664	0.0831	0.0418
94	0.4813	0.2394	0.1199	0.0605
95	0.3438	0.1621	0.0807	0.0405
96	0.6255	0.3138	0.1567	0.0785
97	0.6503	0.3255	0.1627	0.0811
98	0.5557	0.2811	0.1403	0.0708
99	0.4795	0.1974	0.0988	0.0497
100	0.5589	0.2823	0.1408	0.0699

利用隨機產生項目集合資料庫和 ACCESS 所提供的範例資料庫做本研究的實驗，計算出各個項目集合資料庫之標準差、變異數、偏態及峰態。我們預估利用本研究方法在資料量極大的資料庫中找出的平均次數會趨近於呈現 50%、75%、87.5%、93.75% 的百分位置，這是在最理想的狀態之下所得到的結果。但平均數所在位置並不一定恰等於中位數。我們大膽預測其誤差可能與偏態值或峰態值有某種關聯，爲了找出其相關性，我們可以針對各種不同的資料庫加以實驗。

經由實驗發現，若項目集合資料庫若趨近於常態分配，其偏態值會趨近於 0，峰態值則會趨近於 3 時。利用平均項目集合分割法處理後，所得結果則會趨近於理想值。在表 5 與 6 中，編號第 45 的項目集合資料庫，項目集合筆數有 50000 筆、標準差爲 2165.6399、變異數爲 4689996.2、偏態值爲 -0.005、峰態值爲 1.807，非常接近理想值。在資料數量非常龐大的情形下，中位數會趨近於平均數，因此在計算第一次平均項目集合時的理想值是位於 50% 的百分位置，實際上超過此平均項目集合的項目集合個數則占全部項目集合總數的 50.04% 的百分位置。之後，再以超過第一次所得到的平均項目集合之項目集合爲依據，計算其平均項目集合個數，依此類推，所佔的百分比依序爲 25.06%、12.49%、6.25%。

表 7 與圖 6 則是資料筆數爲 10000 筆時的執行結果。

在其他情況不變之下，當資料筆數增加至 50000 筆時，執行結果如表 8 與圖 7 所示。

在其他情況不變之下，當資料筆數增加至 100000 筆時，執行結果如表 9 與圖 8 所示。

表 7. 一萬筆資料之實驗結果

偏態	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_3$	$\overline{X}_4$
-0.311	0.5256	0.2625	0.1310	0.0652
-0.259	0.5214	0.2618	0.1308	0.0658
-0.248	0.5314	0.2696	0.1342	0.0681
0.111	0.4880	0.2449	0.1206	0.0616
0.189	0.4789	0.2389	0.1208	0.0608
0.212	0.4861	0.2136	0.1074	0.0544
0.248	0.4774	0.2080	0.1018	0.0510
0.294	0.4826	0.1965	0.0971	0.0474
0.314	0.4857	0.1964	0.0997	0.0496
1.055	0.3380	0.1666	0.0830	0.0406

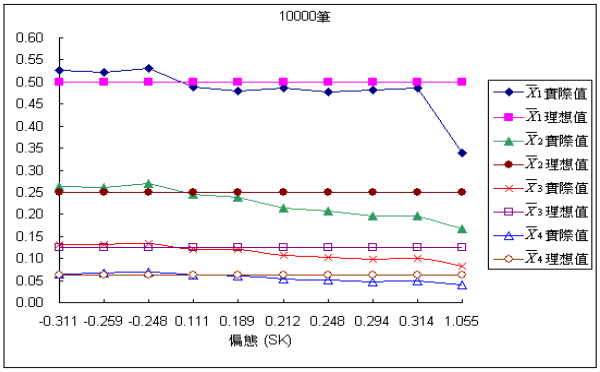

圖 6. 一萬筆資料之結果分析

表 8. 五萬筆資料之實驗結果

偏態	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_3$	$\overline{X}_4$
-1.314	0.6489	0.3243	0.1638	0.0814
-1.271	0.6525	0.3256	0.1643	0.0813
-0.222	0.5305	0.2696	0.1351	0.0679
-0.005	0.5004	0.2506	0.1249	0.0625
0.028	0.4973	0.2459	0.1222	0.0612
0.113	0.4859	0.2430	0.1210	0.0604
0.163	0.4868	0.2259	0.1134	0.0563
0.211	0.4719	0.2359	0.1193	0.0602
1.019	0.3331	0.1668	0.0841	0.0419
1.485	0.3612	0.1476	0.0726	0.0367

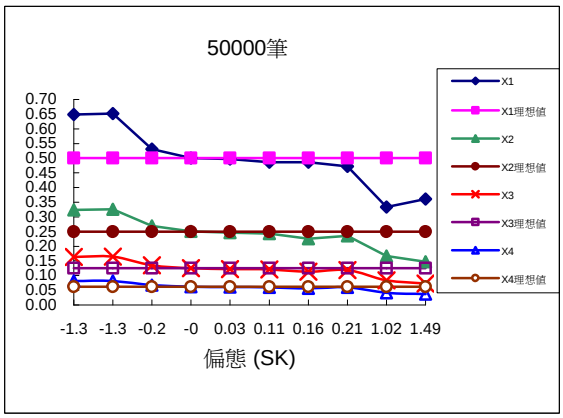

圖 7. 五萬筆資料之結果分析

由實驗中，可看出在利用平均項目集合分割法處理隨機資料時，偏態值愈接近 0 的資料庫，其實際的項目集合數目將愈接近理想值。若偏態值愈偏離 0 的位置，則項目集合數目將愈偏離理想值。

表 9. 十萬筆資料之實驗結果

偏態	$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_3$	$\overline{X}_4$
-1.615	0.6255	0.3138	0.1567	0.0785
-0.841	0.6503	0.3255	0.1627	0.0811
-0.390	0.5589	0.2823	0.1408	0.0699
-0.378	0.5557	0.2811	0.1403	0.0708
-0.104	0.5098	0.2539	0.1271	0.0639
0.148	0.4813	0.2394	0.1199	0.0605
0.304	0.4795	0.1974	0.0988	0.0497
0.917	0.3333	0.1662	0.0833	0.0416
0.972	0.3346	0.1664	0.0831	0.0418
1.220	0.3438	0.1621	0.0807	0.0405

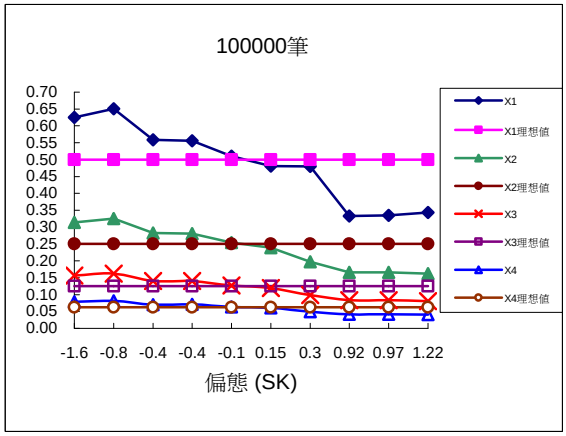

圖 8. 十萬筆資料之結果分析

當本研究在進行時，或許會認爲當資料量愈大時，所實驗的結果是否愈接近理想值，因爲在統計的觀念裡，資料量愈大時，則會趨近於常態，當然其也會愈接近理想值。所以本研究則針對十種不同資料筆數的資料庫，且在每種資料庫中，各選擇出偏態值較接近於 0 的資料庫做實驗分析，執行結果如表 10 與圖 9。

由表 10 與圖 9 之分析結果來看，可清楚了解並不是資料量愈大時，則愈接近理想值，可能和偏態值有關聯。雖然這十種資料庫之偏態值並不相同，但均非常接近 0，未來將會朝向這方面做更深入的研究與探討。

表 10. 不同資料庫之偏態值

萬筆	1	2	3	4	5	6	7	8	9	10
偏態	0.111	0.119	0.206	-0.152	-0.005	-0.144	-0.051	0.077	0.070	-0.104

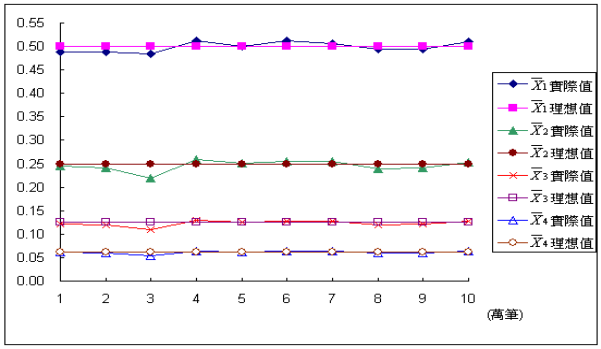

圖 9. 不同資料筆數之比較

五、結論

在本篇論文中，我們提出了一個新的門檻值制訂方法，用來改善傳統相對門檻值方法的缺點，並同時具備絕對門檻值的意義。本研究所提出的方法主要能找出決策者所需的規則數量，利用平均項目集合分割法，逐次的找出決策者所需的規則數量，再利用線性內插方式，計算其比例，即可得之此百分比的門檻值，且可彈性的供決策者輸入此百分比，系統即可自動的找出絕對門檻值。

每一種資料庫所儲存資料的分配情形均不相同，且大都是雜亂無章的，無一定規則循序可循。本研究在使用平均項目集合分割法時，平均數與中位數不盡然相等。在此大膽預測其誤差可能與偏態值或鋒態甚至其他的統計參數有某種關連。故在未來的研究可針對資料的分佈情形，將偏態與鋒態值或其他統計參數加以探討，期望能找出依經驗法則。利用此經驗法則，使往後在訂定門檻值時，能馬上訂出最佳門檻值，無須再做調整，而得到的關聯法則，也是決策者最需要的規則。

參考文獻

1. Adriaans, P. and D. Zantinge (1996) *Data Mining*, Addison Wesley Longman, Edinbrugh Gate, England.
2. Agrawal, R., T. Imilienski and A. Swami (1993) Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD Intel Conference on Management of Data, 207-216.
3. Agrawal, R. and R. Srikant (1994) Mining sequential patterns. Proceedings of the Int'l Conference on Data Engineering (ICDE), Taipei, Taiwan, March 1995. Expanded version available as IBM Research Report

RJ9910.

4. Agrawal, R. and R. Srikant (1994) Fast algorithm for mining association rules in Large Databases. In: Proceedings Int'l Conference VLDB, 478-499. Santiago, Chile.
5. Brin, S., R. Motwani and C. Silverstein (1997) Beyond market baskets: Generalizing association rules to correlations. In: Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD '97), 265-276. J. M. Peckman, Ed. Tucson, AZ.
6. Chen, M. S., J. Han and P. S. Yu (1996) Data mining: an overview from database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 866-883.
7. Fayyad, U., G. P. Shapiro and P. Smyth (1996) The KDD process for extracting useful knowledge from volumes of data. *Communications of The ACM*, 39(11), 27-34.
8. Fayyad, U. M. (1996) Data mining and knowledge discovery: making sense out of data. *IEEE Expert* [see also IEEE Intelligent Systems], 11(5), 20-25.
9. Han, J. and M. Kamber (2000) *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, California.
10. Kleissner, C. (1998) Data mining for the enterprise. System Sciences, Proceedings of the Thirty-First Hawaii International Conference.
11. Michael J., A. Berry and G. Linoff (1997) Data mining technique: For marketing, sales and customer support. Wiley Computer Publishing, New York, NY.
12. Olaru, C. and L. Wehenkel (1999) Data mining. *IEEE Computer Applications in Power*, 12(3), 19-25.
13. Park, J. S., M. S. Chen and P. S. Yu (1995) An effective hash-based algorithm for mining association rules. Proceedings ACM SIGMOD, 175-186.
14. Park, J. S., M. S. Chen and P. S. Yu (1997) Using a hash-based method with transaction trimming and database scan reduction for mining association rules. *IEEE Trans. on Knowledge and Data Engineering*, 9(5), 813-825.
15. Savasere, A., E. Omiecinski and S. Navathe (1995) An efficient algorithm for mining association rules in large databases. In Proceedings of the 21st Int. Conference on Very Large Data Bases.
16. Savasere, A., E. Omiecinski and S. Navathe (1998) Mining for strong negative associations in a large database of customer transactions. In Proceedings of the IEEE 14th Int. Conference on Data Engineering, Orlando, FL.
17. Simoudis, E. (1996) Reality check for data mining. *IEEE Expert* [see also IEEE Intelligent Systems], 11(5), 26-33.
18. Srikant, R., Q. Vu and R. Agrawal (1997) Mining association rules with item constraints. Proceedings of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California.
19. Srikant, R. and R. Agrawal (1995) Mining generalized association rules. In Proceedings of the 21st Int'l. Conference on Very Large Data Bases.
20. Srikant, R. and R. Agrawal (1996) Mining quantitative association rules in large relational tables. Proceedings of the ACM SIGMOD Conference on Management of Data, Montreal, Canada.
21. Srikant, R. and R. Agrawal (1997) Mining generalized association rules. *Future Generation Computer Systems*, 13(December), 2-3.
22. Yongjian, Fu (1996) Discovery of multiple–level rules from large databases. Ph. D. Dissertation, Simon Fraser University, Burnaby, British Columbia, Canada.
23. Yongjian, Fu (1997) Data mining tasks, techniques and applications. *IEEE Potentials*, 16(4), 18-20.
24. Zaki, M. J., S. Parthasarathy, M. Ogihara and W. Li (1997) New algorithms for fast discovery of association rules. American Association for Artificial Intelligence.

收件：92.11.10　修正：93.02.10　接受：93.03.05