國立交通大學

管理學院資訊管理學程碩士班

碩　士　論　文

整合正向與負向關聯規則探勘於
發掘 DRAM 製造之問題機台

Integrating Positive and Negative Association Rule Mining for Identifying Root Cause Machines in DRAM Manufacturing

研 究 生：林 居 逸

指導教授：劉 敦 仁 博士

中 華 民 國 九 十 九 年 六 月

整合正向與負向關聯規則探勘於發掘 DRAM 製造之問題機台

Integrating Positive and Negative Association Rule Mining for Identifying Root Cause Machines in DRAM Manufacturing

研 究 生： 林 居 逸　　Student: Jiu-Yii Lin

指導教授： 劉 敦 仁 博士　　Advisor: Dr. Duen-Ren Liu

國 立 交 通 大 學

管理學院資訊管理學程碩士班

碩 士 論 文

A Thesis

Submitted to Institute of Information Management

College of Management

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Information Management

June 2010

Hsinchu, Taiwan, the Republic of China

中 華 民 國 九 十 九 年 六 月

整合正向與負向關聯規則探勘於發掘 DRAM 製造之問題機台

研究生：林居逸　　指導教授：劉敦仁 博士

國立交通大學管理學院資訊管理學程碩士班

摘　要

半導體產業競爭日益激烈，製程愈趨複雜，若能快速的找出製程中的根本問題站點或機台，便能儘早的解決問題以提升良率，比對手更具競爭力。本研究提出整合正向與負向關聯規則的資料探勘方法，以產品的良率與站點機台間的關聯規則，推估造成缺陷產品的根本問題站點機台。正向關聯規則為造成缺陷產品的問題站點機台，而負向關聯規則為產出正常產品中非該站點機台與其它站點機台的關聯。一站點機台 X 若在正向關聯規則 $X \Rightarrow Y$ 及負向關聯規則 $\neg X \Rightarrow Y$ 皆出現，則 X 更能確認為問題站點機台。因此整合正向與負向關聯規則探勘，分析站點機台間關聯規則，能更有效的找出問題機台。實驗評估以某 DRAM 半導體廠商資料為例，由結果得知，將負向關聯規則納入考量的整合關聯規則方法較單以正向關聯規則分析的方法表現佳，佐證本研究提出方法之實際效益。

關鍵字：資料探勘、關聯規則、Apriori 演算法、問題分析

Integrating Positive and Negative Association Rule Mining for Identifying Root Cause Machines in DRAM Manufacturing

Student：Jiu-Yii Lin　　Advisor：Dr. Duen-Ren Liu

Institute of Information Management

National Chiao Tung University

Abstract

Manufacturing processes have become more and more complex in semiconductor industry. Improving yield by identifying root cause machines is the key factor of improving competitiveness. In this study, we propose to identify root cause machines by integrating positive and negative association rule mining of correlations between machines from the manufacturing processes of defective and normal products, respectively. If a machine X exists both in positive association rules (eg. $X \Rightarrow Y$) and negative association rules (eg. $\neg X \Rightarrow Y$), then X might be the most doubtful machine. The experimental results of real DRAM manufacturing datasets show that the proposed integrated approach is more effective than the approach with considering only positive association rules.

Keyword : Data mining, Association Rule, Apriori algorithm, Root cause analysis

誌謝

首先，感謝我的指導教授劉敦仁博士，在劉老師的諄諄教誨及悉心指導下，讓我在求學的過程中，不斷的進步及自我突破。

然要能順利完成此項任務，背後也需要有強大的力量支持著，而家人就是我繼續前進的最大能量，所以特別要感謝我的家人，無論是我的父母在生活上的照顧或是毅玨在心理層面給我的支持，你們都是我能全心全力投注在課業上，安心完成學業的推手，感謝你們。

而此時，在我人生旅途中會是個特別值得紀念的時刻，因為在此論文撰寫完成後，我的阿公並未能與我一同分享這份喜悅，但我相信他在另一個世界，也同樣能感受到我對他感恩的心，謝謝他不求回報的付出及教導，最後，謹以此誌來感念阿公的恩情。

目錄

圖目錄

表目錄

第一章 緒論

1.1 研究背景

半導體產業競爭日益激烈，尤其在 2008 年爆發了全球性的金融海嘯，更使得企業生存割喉戰白熱化。以 DRAM(動態存取記憶體)產業為例，過去十幾年來幾乎是以每三年一次景氣循環，2001 年到 2003 年連續三年虧損，2004 年到 2006 年轉為連續三年獲利，2007 年到 2009 年又落入了連續三年嚴重虧損，如圖 1.1 所示，然在 2010 年因全球景氣復甦，帶動消費者及企業換機潮，以及智慧型手機帶動行動記憶體需求大幅成長下產業動力強勁，又將進入獲利循環；在這需要大量產出以搶佔市佔率的時機，卻因 2008 年的金融海嘯各家 DRAM 廠商大幅縮減支出，且短時間內轉進下世代製程所需的浸潤式機台供給有限，致製程皆未能及時轉至 40 奈米以下增加產出，值此局面各家廠商惟有將自身的良率提高以達增加產出之效，故良率除了在平時製造業廠商能力指標佔有相當重要的代表性，在危機時更是企業存續與否的關鍵，而如何提升良率則是各家廠商提高競爭力不可或缺的重要議題。

Figure-1 DRAM Makers'OP Margin from 1999- 2009

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Winbond	15.8%	21.6%	-49.4%	-5.6%	-5.5%	10.9%	-7.9%	11.2%	-11.5%	-27.6%	-30.7%
ProMOS	42.0%	41.2%	-44.9%	-8.5%	3.4%	27.0%	0.2%	28.7%	-13.0%	-78.1%	
PSC	8.1%	24.8%	-53.2%	-10.5%	0.8%	40.0%	10.7%	30.2%	-14.2%	-74.8%	-58%
Inotera						5.5%	24.9%	39.0%	6.9%	-48.1%	-27%
Nanya	5.2%	7.1%	-86.4%	11.6%	-0.2%	20.5%	1.2%	19.1%	-17.0%	-66.4%	-38%
Micron	-0.8%	37.6%	-24.8%	-39.6%	-38.4%	5.7%	4.4%	6.6%	-4.9%	-27.3%	-34.9%
Elpida				-37.7%	-26.3%	7.3%	0.1%	14%	-6.2%	-44.5%	-17.3%
SEC(Semi)	33.4%	47.1%	7.9%	30.8%	28.7%	41%	29.8%	26.4%	11.9%	0.7%	9.8%
Hynix	10.7%	16.9%	-32.4%	-31.3%	-7.2%	31.5%	24.9%	24.7%	5.9%	-28.2%	2.5%

Source: Companies data, compiled by DRAMeXchange

圖 1.1 DRAM 各家廠商歷年盈虧列表(資料來源: DRAMeXchange)

1.2 研究動機

影響良率的則為製造過程中加工的程序及處理的機台，然精密的半導體產品往往需要經過數百道加工的程序，及上百台的機台處理，過程中通常會有系統紀錄著幾乎所有實際生產的程序及機台資訊，但龐大且複雜的資訊，在實務上卻有可能讓決策者難以閱覽，而本研究的動機則是提出一資料探勘手法應用在分析製程資料，找出根本造成良率不佳的站點機台，以系統化的方式提供決策者輔以更有效及有效率的解決問題，以提升良率。

1.3 研究目的

以往在半導體業界或許既有系統可能以統計手法，例如迴歸分析，或提供大量的圖表給工程師人工判讀，然而這樣的方式卻不夠有效率，本研究企圖透過分析產出良率及製程站點機台的資料，藉其產出良率與其製程機台歷程的關聯，及站點機台間的關聯，發掘有效且是使用者感興趣的樣式(pattern)，其中站點機台間的關聯，因在半導體業界普遍機台組合有其生產計劃為依據，故機台間的關聯規則樣式發掘較為困難，然本研究認為發掘的樣式雖然較稀少，卻是有重要意義的資訊，因而提出不只是在缺陷產品中萃取正向關聯規則，更在正常產品中萃取其相對應的負向關聯規則，以期資訊的完整性。

本研究提出的程序大致分為四個階段，資料前置處理、產生候選值階段、萃取關聯規則，及可能根本問題機台排名列表，在資料前置處理階段將所有產出良率以使用者訂定的良率規格分別標示為缺陷產出或是正常產出，再與製程歷史紀錄表彙整成缺陷產品歷程紀錄表及正常產品歷程紀錄表。

產生候選值階段分為兩步驟，(1)篩選顯著站點、(2)產生有趣的候選值。首先以 Kruskal-Wallis 統計檢定法篩選拒絕虛無假設 H_0:S 站點中

的各機台，良率表現皆無差異的顯著站點，選出顯著的前 50 名作為初步的候選站點；接著再以支持度最小門檻值及有趣度最小門檻值產生既發生次數較頻繁且使用者較感興趣的候選值，此處的有趣度為有效率的修剪策略，使機台間發生的關聯性較低者能提前被修剪剔除，以期讓後續的關聯運算能更符合使用者期望。

萃取關聯規則階段，則是企圖透過良率與製程站點機台的關聯回推各機台間互相影響的樣式，找出間接影響良率的機台組合關聯，其中以負向關聯規則輔正向關聯規則，因為一般在思考問題時，較常以正向的思維尋找解答，然許多時候，若能逆向思考反而會有不同於平常的更佳解，例如，有 2 個類別 C_1 和 C_2，3 個屬性值 X、Y 和 Z，在支持度-可信度的架構下找出屬性值與類別的正向關聯規則，$X \rightarrow C_1$，$Y \rightarrow C_1$，$X \rightarrow C_2$，$Y \rightarrow C_2$，若有一新資料列同時擁有 X 及 Y 的屬性值，則此資料列較不確切會被分類為 C_1 或 C_2；但若同時考慮負向關聯規則時，如 $X \rightarrow C_1$，$Y \rightarrow C_1$，$\neg Z \rightarrow C_1$，$X \rightarrow C_2$，$Y \rightarrow C_2$，$Z \rightarrow C_2$，以此情況看來，這筆同時擁有 X 及 Y 但沒有 Z 屬性值的新資料列就會被歸類在 C_1，足以見負向關聯規則在分類程序中的可應用性[16]，所以本研究在此階段提出整合 Wu 等於 2004 年[25]提出正向及負向關聯規則的萃取流程，應用於 DRAM 製造問題機台的發掘，過程不只在缺陷產品資料中萃取正向關聯規則，更在正常產品資料中萃取出負向關聯規則，藉整合兩者以期達到更整體全面性的分析。

最後在可能根本問題排名列表階段，透過三種計算方式比較各有趣候選項目經由萃取關聯規則階段萃取出的各機台間關聯的間接關聯可信度，得排序積分，以作為最後根本問題機台排序的依據，藉排序列表找出可能為根本問題的機台，提供決策者及時找出問題，以便有效率的切中要點，解決問題。

1.4 論文架構

本論文參考相關歷史文獻所提出的方法與架構，提出一個更有效率及有效的半導體製程根本問題分析方法。本研究分為五個章節，其架構如圖 1.2 所示，說明如下：

第一章緒論：主要是說明研究背景及動機、研究目的及論文架構。

第二章文獻探討：以本研究相關之文獻知識作為探討之主題，其內容包括 Kruskal-Wallis 無母數統計法、資料探勘、負向關聯規則及其相關應用等。

第三章研究方法與架構：本章以本研究所採用的研究方法為討論重點，包含問題定義、研究方法與架構。

第四章實驗與評估：本章實際以半導體廠 DRAM 製造業之製程資料進行實驗，後以實驗結果進行評估。

第五章結論與未來研究方向：針對此篇研究做一結論及說明未來的研究方向。

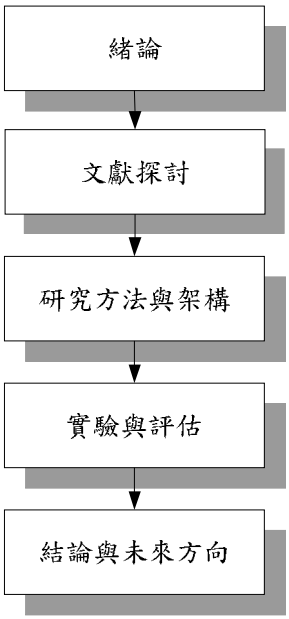


圖 1.2 論文架構

第二章 文獻探討

2.1 Kruskal-Wallis 無母數統計檢定法

Kruskal-Wallis 統計檢定法是由 Daniel 於 1978 年提出的無母數統計檢定法，所謂無母數意謂不考慮參數的特性，所以較無假設限制，且計算簡單，不太受極端值的影響，但也因此這類檢定法的缺點是檢定力較弱，然本研究採用此檢定法做初步篩選，故評估未受其檢定力弱而有所影響，其基本假設如下[32]。

假設：

1. 簡單隨機樣本。
2. 獨立樣本。
3. 母群的形態相同。
4. 所有樣本大於等於 5。

整個檢定程序如圖 2.1 所示。

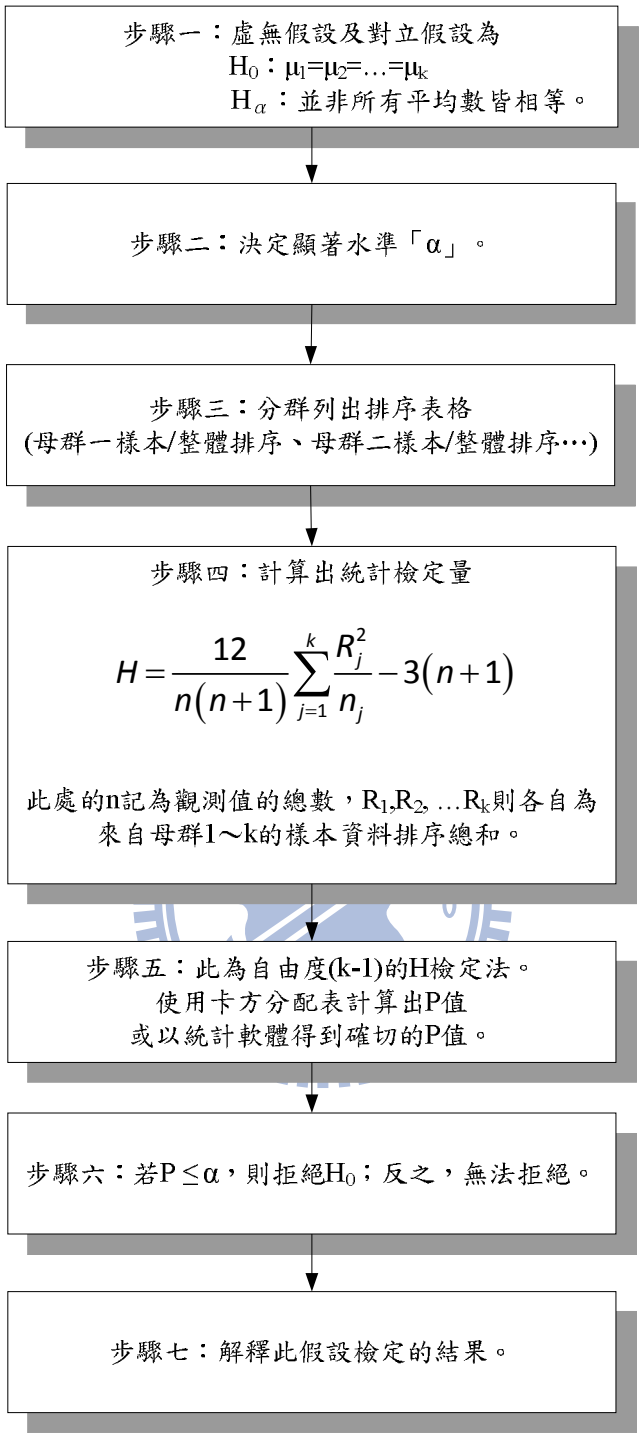


圖 2.1 Kruskal-Wallis 檢定流程

以表 2.1 資料為例，三組簡單隨機、獨立且型態相同的母群樣本資料，A、B、C，每一組有 6 個數值，分析各組間是否表現相同，依照上述檢定程序。

表 2.1 三組母群樣本

A	B	C
86	91	92
79	83	84
82	85	88
81	78	89
90	75	77
87	80	76

步驟一虛無假設為 H_0:u_A=u_B=u_C，對立假設為並非所有平均數相等；步驟二:決定顯著水準 α=0.05；步驟三:分群列出排序表格，如表 2.2，其中排序是依據整體資料的順序給予序號。

表 2.2 分群排序表格

A		B		C	
Value	Order	Value	Order	Value	Order
86	12	91	17	92	18
79	5	83	9	84	10
82	8	85	11	88	14
81	7	78	4	89	15
90	16	75	1	77	3
87	13	80	6	76	2

步驟四:計算其統計檢定量，

$$K\text{-}W=\frac{12}{18(18+1)}\left(\frac{61^2}{6}+\frac{48^2}{6}+\frac{62^2}{6}\right)\text{-}3(18+1)=0.7135$$；步驟五至七:使用卡方分配表計算出 P 值，k=3，故查卡方分配表 P-value>0.05，所以 H_0 不顯著，即沒有證據說三群組間有顯著差異。

在半導體業界，常以統計上變異數分析(ANOVA)或 Kruskal-Wallis 檢定法，透過產品的良率不佳的資料，回推其製造過程中，可能是哪一道製程或是製造機台造成的良率上與其它製程或機台的差異，藉以找出可能的根本問題。本研究使用 Kruskal-Wallis 檢定法是因為 ANOVA 的假設其母體必須是常態分布的資料，但在實務上往往不是如此，所以本研究

採用 Kruskal-Wallis 檢定法而非 ANOVA，在歷史文獻中，Kruskal-Wallis 檢定法在半導體製程資料分析的應用，如表 2.3 所示。

表 2.3 Kruskal-Wallis 檢定法在半導體資料分析的應用

論文	作者	應用描述
Data mining for yield enhancement in semiconductor manufacturing and an empirical study	Chien, C.F.，2007[7]	以半導體為實證對象，利用無母數 Kruskal-Wallis 統計檢定與決策樹找出造成良率不佳的晶圓過製程站點，其中採用 Kruskal-Wallis 檢定降低製程站點決策分支的維度，以增加其分析效率。
整合決策樹與關聯規則之資料挖礦架構及其實證研究	楊景晴，2003[33]	以台電配電事故表與某半導體實際資料進行實證研究，整合決策樹與關聯規則之方法，而初步篩選採 Kruskal-Wallis 找出發生故障性較高的原因，降低資料維度，縮小診斷範圍。
建構半導體晶圓允收測試資料挖礦架構及其實證研究	林鼎浩，2000[31]	以半導體允收測試資料為目標變數，利用無母數 Kruskal-Wallis 統計檢定法與決策樹，找出允收測試資料不佳的機台與時間。

2.2 資料探勘

「探勘」，直覺應該是類似礦工在漆黑的礦坑裡，使用十字鎬、圓鍬等原始的工具，僅憑團隊經驗及奮鬥不懈的毅力，一吋一吋的往可能藏有珍貴礦石、或是稀有資源的洞穴深處挖掘，如果將「資料」看做是

挖掘過程中處理的各種原石，「知識」是挖掘萃取出珍貴礦石，而「資料探勘」就是要將這些龐大雜亂的「資料」，經過資料選擇、處理、轉換、資料挖掘、評估等幾個步驟，評估分析後得到的未知新資訊或知識，則為「資料庫的知識發現」(Knowledge Discovery in Database)[11]。

在「資料庫的知識發現」過程中分為七個步驟，資料清理(Data cleaning)、資料整合(Data integration)、資料選擇(Data selection)、資料轉換(Data transformation)、資料探勘(Data mining)、評估模型(Pattern evaluation)、知識具現化(Knowledge visualization)。

資料探勘的模型普遍可歸納為四種類型，例如:關聯規則(association rule)、叢集(clustering)、分類(classification)、預測(prediction)[14],[10], [28]，而以下將針對本研究主要使用的關聯規則及其延伸應用加以探討。

2.2.1 關聯規則

關聯規則是學者 Rakesh Agrawal [1]在 1993 年提出，探討如何探勘市場購物籃型態的大量項目資料集合，以最小可信度(minimum specified confidence)找出項目集(itemsets)之間的關聯規則。例如，大部分顧客購買麵包時，也會購買牛奶，所以市場商店可以降價促銷麵包，同時也提高麵包與牛奶的銷售量，或是透過此關聯規則，可以研擬商品擺放位置的策略，以增加各項商品的銷售量。

Agrawal 於 1994 年[2]定義關聯規則如下: $I = \{I_1, I_2, \ldots, I_m\}$為包含 $I_1, I_2, \ldots, I_m$ 項目集，D 則為所有項目的交易資料集。T 則為交易資料集中的其中一筆交易資料，標示為 $T \subseteq I$，每一筆交易資料皆有其唯一識別編號 TID。若此交易資料 T 中包含 X，則標示為 $X \subseteq T$。而又若同一筆資料中包含 X，且又包含 Y，且兩者交集為空集合，標示為 $X \subseteq T$，$Y \subseteq T$，$X \cap Y = \varnothing$，則可萃取其在資料集 D 中 X 發生同時 Y 發生，標示為 $X \Rightarrow Y$ 的關聯規則。為了使其關聯規則有足夠的證據，以支持度 s(support)的最小門檻值篩選，支持度的計算則為在資料集 D 中 $X \cup Y$ 同時發生的機率，

標示為 support($X \Rightarrow Y$)=P($X \cup Y$)。而 $X \Rightarrow Y$ 的可信度 c (confidence)，計算 X 發生的情況下 Y 也發生的條件機率，標示為 confidence($X \Rightarrow Y$)=P($Y|X$)=P($X \cup Y$)/P(X)。故惟有同時滿足大於最小支持度門檻值且可信度在最小可信度門檻值以上的關聯規則方足以稱為強關聯規則(strong association rules)。

Apriori 演算法為探勘關聯規則中最為廣泛被應用於找出候選值的方法之一，以反覆程序的層級運算找出頻繁的項目集組合，其中反覆程序、層級意謂的是由符合最小支持度門檻值的 k 層項目集，排列組合產生 k+1 層的項目集，直到整個項目集組合皆被完全探索，無法產生新的為止。以 Apriori 找出頻繁項目集後，再計算其可信度，則可找出強關聯規則，而關聯規則在半導體製程的相關應用介紹如下。

2.2.2 關聯規則在半導體製程的相關應用

知識在製造產業中扮演著承先(保留舊有的傳統價值)啟後(開創新的學識領域)的強而有力角色，以解決錯綜複雜的問題，進而創造核心競爭力，開創對於個人與組織，現在與未來的新局面[8]。然環顧現代的製造產業中，其自動化及數位化的製程環境，於製造過程中產生前所未有的資料量。這些資料或許涵蓋了設計、產品、加工機台、材料、存貨、維護、計劃與控制、裝配、運輸物流、性能等等面向，其中隱藏著資料的樣式、趨勢、關聯與相依性。然而這些被積聚的資料在實際使用上卻相當有限，造成"資料充足卻資訊貧乏"的問題[23]。如何透過 KDD 的資料探勘手法，將隱藏在複雜資料裡面的樣式、趨勢、關聯與相依性，以系統化的方式探勘有意義的資訊，進而萃取有價值的知識，則是一重要的議題，故資料探勘在製造產業的研究上已然是一門不可或缺的顯學。

資料探勘模型的類型主要分為特徵描述、關聯規則、叢集、分類、預測，各類型在半導體製造的領域中皆有眾多研究與貢獻。Choudhary et al.於 2009 年[8]使用文字探勘，找出資料探勘類型與製造產業各領域的關係與連結性，如圖 2.4。製造產業各領域涵蓋 SCM(Supply Chain

Management)、產品設計、生產管理、故障診斷(fault diagnosis)、製造流程、良率提升、生產排程、日常維護、品質控制、及製造系統等。

特徵描述類型的資料探勘與生產排程具有較深的連結關係，其中不乏以基因演算法(GA)或模糊理論應用在生產排程上，Liu, Huang, 及 Lin 於 2005 年[27]提出的 SAMA(supervised attribute mining algorithm)，以解決排程規則的屬性選擇問題；分類類型的資料探勘較常應用在品質控管與故障診斷領域，這一類的應用通常都能有很好的成效，例如 Rojas 及 Nandi 於 2006 年[20] 應用 SVM(support vector method)的故障分類，以發掘故障樣式提供工程師診斷，提供更有效率及更有效的分類程序。叢集類型的資料探勘在產品設計也是應用廣泛，然值得一提的是叢集在良率提升的應用，Gardner 及 Bieker 於 2000 年[13]提出自我映射組織(SOM)的類神經方法與規則引證結合，分析晶圓製造的資料，以找出影響良率的關鍵因子，藉以改善且提升良率，其對產業的貢獻良多。預測類型的資料探勘在製造產業扮演著重要的角色，例如廖泰祥於 2007 年[34]應用通用迴歸類神經(general regression neural network)分析生產機台之參數資料，以達成即時檢測及節省實際量測成本的效益。Li 等於 2006 年[15]應用基因演算法為基礎的資料探勘手法，預測良率且提供自動探索造成良率低下的因素，這方面的應用在實務上也廣泛的被採用。

關聯規則研究文獻中最常以購物籃原理分析說明，然在製造產業中其應用不只在市場行銷、存貨及倉儲分析，也常應用在故障分析、製程能力等，萃取製程與良率關聯的重要因素，使製程能更有效率的調整參數或機台的健康狀況，以提升良率、生產力，進而降低生產成本，以下將依領域分類列表說明，如表 2.4。

表 2.4 關聯規則在半導體的應用

領域	論文	作者	應用描述
製造程序	Configuration of cellular manufacturing systems using association rule induction	Chen，(2003)[4]	透過關聯規則找出機台或機台群組間的組合關係，藉以提高單元成型(cell-formation)的效率及彈性。
	Integration of variable precision rough set and fuzzy clustering: An application to knowledge acquisition for manufacturing process planning	Wang 等，(2005)[24]	整合模糊叢集與 VPRS(variable precision rough set)找出有效的最小門檻值訂定策略，使萃取關聯規則能更有效的訂定製造程序計劃。
	A novel manufacturing defect detection method using association rule mining techniques	Chen 等，(2005)[6]	透過萃取製程中的機台組合造成缺陷產品的關聯規則，進而找到根本問題機台組合，提高製程效率。
	Data mining for improvement of product quality	Chuha 等，(2006)[9]	應用關聯規則找出最佳的組裝程序，將製造過程的風險因子降低，提升產品品質。

支援決策	An expert system to generate associativity data for layout design	Chao 等，(1997)[3]	提供關聯資料使佈局合成(layout generation)自動化之智慧型決策支援系統。
	Integrating data mining and rough set for customer group-based discovery of product configuration rules	Shao 等，(2006)[22]	以關聯規則找出顧客群與產品規格群組間的關聯，進而有效率的設計管理產品組態。
顧客關係管理	An association based clustering approach to order batching considering customer demand pattern	Chen and Wu，(2005)[5]	以關聯規則分析顧客訂單產品組合的關聯性，進而引導出顧客在下訂單時的關聯產品項目，增加顧客下訂單時的便利性。

由以上資料探勘的各種模型類型在半導體的領域應用足可見，不論是資料探勘又或是其關聯規則在製造產業的應用廣泛及其深耕的程度，更在在顯示其重要性，其中又以 Chen 提出的 Root-cause machine identifier 方法與關聯規則在根本問題機台的半導體探勘領域最為相關，以下大概說明其方法。

2.2.3 根本問題機台判定方法

根本問題機台判定方法(Root-cause machine identifier method)是 Wei-Chou Chen 在 2005 年提出的一透過關聯規則的資料探勘手法，以偵測製程上的缺陷，進而提升製程的良率，降低成本[6]。

半導體製程中，產品材料經過許多站點，而每個站點又由可能不只一台機台負責。產品材料則因生產計劃、產品類別、參數不同，會經過不同的站點及機台，而後產出最終產品。

此一透過關聯規則的資料探勘手法，以偵測製程上的缺陷，其主要架構流程如圖 2.5，說明如下。

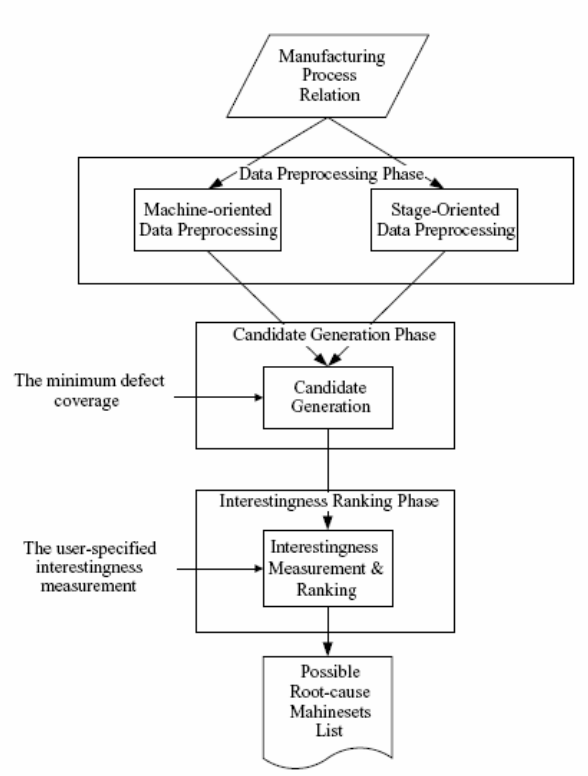


圖 2.2 根本問題機台判定方法流程[6]

(1) Data Preprocessing Phase:

前置資料處理階段，分為機台導向與站點導向兩個面象分析，機台導向程序專注在產品經過的機台，而略過站點。儘管一台機台可能提供多個功能，但只要經過一站點上的某機台，就看作為一筆關聯資料。

(2)Candidate Generation Phase:

此階段是根據缺陷產出的比例，以符合最小缺陷產出比例的候選鍵組合，一層一層產生後挑選。此候選鍵的挑選乃根據 Apriori 演算法，只要是符合最小缺陷產出比例，由第一層級候選機台組合，至第二層級候選機台組合，直至適當的候選機台組合被挑出，或是沒有更高一層級的候選組合。

(3)Interestingness Ranking Phase:

有趣度量測值，以用來評估候選機台組合與缺陷產品的關連性指標。此階段提出兩有趣度量測值，Confidence 與 Continuity。

可信度 Confidence，是最廣為人知的有趣度量測值，由 Piatetsky-Shapiro 等在 1991 年[18]提出，用以評量當 A 發生時 B 也發生的機率。連續性函數 Continuity，則為 Wei-Chou Chen 在該篇論文中提出，代表當機台候選鍵之間連續性越高，則提高其可信度，而其計算方式則根據鄰近的缺陷產品平均距離計算而得。評估指標則根據以上兩種組合，排序後作為輔助分析工程師決策哪一種機台組合是根本問題機台之依據。

2.2.4 負向關聯規則及其相關應用

非預期樣式(unexpected patterns)與例外樣式(exceptional patterns)意指一些例外狀況的規則，也稱之為意外週期樣式(surprising patterns)。所謂例外狀況就是跟眾所皆知的事實相左的偏差樣式(deviation patterns)，表現的令人出乎意料之外。通常內含有否定的用詞，因此可認定其為特殊之負向規則案例[25]。這類的例外狀況在研究上是有趣的，當現象與平時所認知的事實有所異時，其原因更是值得研究發掘的。

而所謂負向關聯規則(negative association rules)就是指在關聯規則裡項目與項目之間有著相反的關係，例如，某事件的發生與某事件的不發生。一般關聯規則裡所標示的"$X \Rightarrow Y$"，表示前項 X 發生次數增加會使得後項 Y 的發生次數跟著增加，延伸其負向關聯規則則有"$X \Rightarrow \neg Y$"、"$\neg X \Rightarrow Y$"、"$\neg X \Rightarrow \neg Y$"，分別表示前項 X 發生次數增加會使得後項 Y 不發生的次數增加、前項 X 不發生的次數增加會使得後項 Y 發生的次數增加，及前項 X 不發生的次數增加會使的後項 Y 不發生的次數也跟著增加，這些都可歸類為相對"$X \Rightarrow Y$"正向關聯規則的負向關聯規則。

負向關聯規則有其不可忽視的重要性，雖在關聯規則的相關文獻中仍屬著墨不多，但已有愈來愈廣泛深入研究的趨勢。而最早的負向關聯規則挖掘演算法是在 1998 年由 Savasere 學者提出[21]，當顧客通常買了某些商品時，卻不買哪些商品，這一類的規則定義為負向關聯規則。若在一般交易資料庫中探勘負向關聯規則將會挖掘出大量極端不有趣的關聯規則，因此 Savasere 學者提出以領域知識為依據，利用「商品分類架構」(taxonomy) 將商品分類，其假設各類商品分別有該類的商品與商品項目集的期望支持度(expected support)，設計一有趣度指標 RI(rule interest measure)，如式 2.1，計算該類商品分類中，商品與商品項目集的實際支持度與期望支持度差距，另設定 RI 的最小門檻值，若

差距高於最小門檻值，則代表該規則隱含的資訊較有意義，即為愈有趣，藉以有效率的篩選有趣的負向關聯規則。

$$RI = \frac{\varepsilon[supp(X \cup Y)] - supp(X \cup Y)}{supp(X)} \qquad \text{式 2.1}$$

此評估關聯規則有趣度的指標函式在 2002 年 Yuan 等學者[29]將其另行定義，以實際可信度與估計可信度差距作為評量的標準；然以商品分類架構為基礎的方法，其關鍵在於分類，必須依產業有其不同的領域知識作分類，而分類的品質又將大大的影響挖掘的關聯規則品質。

Wu 等學者於 2004 年[25]提出兩階段的萃取正向與負向關聯規則流程，使得探勘更有效率，首先找到有趣的正向與負向候選項目集，主要是依據 Agrawal 等學者於 1994 年[2]提出的支持度-可信度(support-confidence)架構，正向候選項目集即定義為以支持度最小門檻值為條件，發生次數高於門檻值的項目集，即為「頻繁項目集」；相對地，負向候選項目集則為發生次數低於最小支持度門檻值的項目集，也可稱為「非頻繁項目集」。其中，「有趣」的定義是以該篇研究提出的修剪(prune)策略評估，當前項(antecedent)與後項(consequent)兩事件的發生機率接近獨立事件機率的定義，即計算所得之有趣度小於有趣度最小門檻值，其提出的論點闡述由此兩項目集產生的關聯規則對研究來說將屬於不有趣的，故予以提前修剪移除，藉以限制搜尋空間，以達更有效率的運算，找到真正有用的候選項目集。第二階段則根據第一階段產生有趣的正向與負向候選項目集，根據 Zhang 與 Zhang 在 2002 年[30]提出的 PR 模式(probability ratio model)，分別以條件機率遞增比例函式(conditional-probability increment ratio)計算評估，若大於訂定的可信度最小門檻值，則可分別依正向候選項目集與負向候選項目集萃取正向關聯規則”$X \Rightarrow Y$”，及更明確的萃取出各類負向關聯規則” $X \Rightarrow \neg Y$”、” $\neg X \Rightarrow Y$”、” $\neg X \Rightarrow \neg Y$”。

Maria 與 Osmar 學者在 2004 年[16]提到，雖然愈來愈多的研究顯示

負向關聯規則的重要性，但同時萃取正向與負向關聯規則，也只有 Wu 等學者這一組團隊提出的演算法能有效的探勘，有鑑於此，Maria 與 Osmar 提出另一方法同時挖掘潛在的正向與負向關聯規則，以實作一關聯式分類(associative classifier)系統。其演算法不同於之前單純只以支持度-可信度架構為基礎的萃取流程，另提出了應用相關係數(correlation coefficient)的正向或負向關聯的判定，使該項目集與該類別的相關係數為正時，則將該項目集歸類為正向關聯分類規則，若為負的，則反之，將該項目集歸類為負向關聯分類規則。

第三章 研究方法與架構

3.1 問題定義

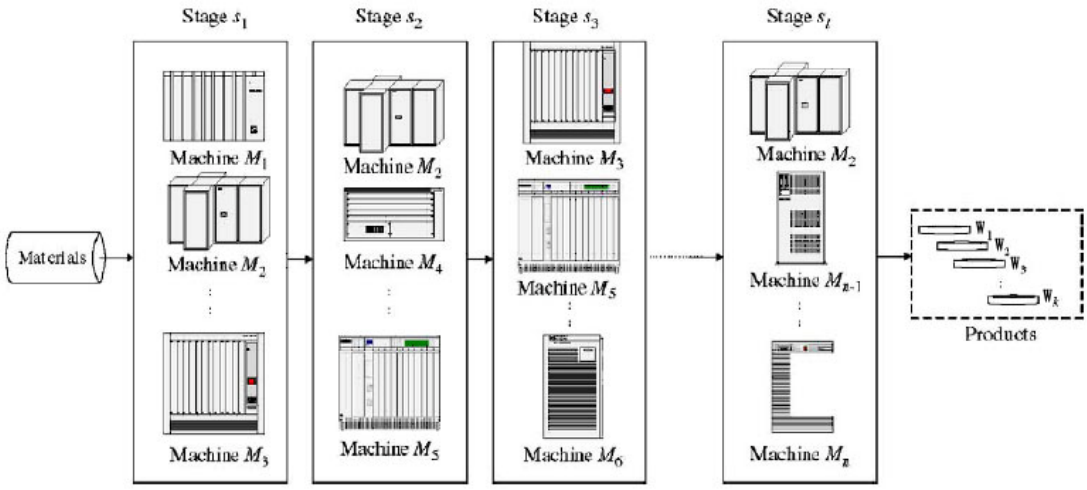


圖 3.1 製造站點與機台程序圖[6]

半導體製程中，產品材料經過許多站點，而每個站點又由可能不只一台機台負責。產品材料則因生產計劃、產品類別、參數不同，會經過不同的站點及機台，而後產出最終產品。如圖 3.1 所示，*k* 個產品批號{W_1, W_2, W_3,…, W_k}，每個產品經過 *l* 個站點{S_1,S_2,S_3,…,S_l}，總共有 *n* 台機台{M_1, M_2, M_3,…, M_n}。

此研究目的則是要透過資料探勘方法，找出真正造成缺陷批號的根本問題機台組合。在此之前，同樣的目的已有數篇研究，利用不同的方法論，如學者 Mieno 於 1999 年[17]提出迴歸樹分析對製造過程中的失誤進行研究、Raghavan 於 2002 年[19] 提出透過決策樹探討造成良率缺失的根本原因、及 Wei-Chou Chen 等人於 2005 年[6]提出透過關聯規則的根本問題機台判定方法。

根本問題機台判定方法利用製程站點、機台與有缺陷產出批號之間的關聯規則，進而找出有興趣的樣式(pattern)，即定義為可能的根本問

題機台組合，但該方法僅以正向關聯規則分析，而近年來，關聯規則探勘方法也由原來較單純的頻繁次數分析，演化出眾多相關的議題探討，例如，Freitas 於 1999 年[12]提出的樣式有趣度設計，針對如何才能確定找出的項目集是使用者感興趣的，闡述設計及評估有趣度的重要性；Yao 於 2006 年[26]提出利潤式的關聯規則，以權重方式計算項目集不只頻繁次數，還有實際產生的利潤，進而找到對使用者更有價值的樣式。

另外，Wu 於 2004 年[25]針對正向及負向關聯規則提出更有效率的探討，因為不只正向的關聯規則對使用者是有意義的，負向關聯規則也在近年來更被廣泛的應用在例外規則的探勘程序裡，例如，最常被拿來舉例關聯規則的顧客購物籃原理，當多數顧客購買果汁飲料時，伴隨著會購買洋芋片，我們以此正向關聯得知，可將果汁飲料及洋芋片做組合促銷，然多數顧客購買果汁飲料時，多半不會再額外添購瓶裝水，則我們也可以此負向關聯輔以作決策，將瓶裝水使用在現打果汁或其他飲料的調配上，再搭配洋芋片做促銷，提高瓶裝水的使用率及整體銷售額，但若僅以正向關聯規則決策時，無法得知可以如此提升整體銷售額的策略，因此，負向關聯規則在決策的應用上有其明顯及不可忽略的效益。

3.2 研究架構

故本研究企圖整合正向及負向關聯規則，探討製程站點、機台與有缺陷產出批號間的負向關聯規則，針對原正向關聯規則增添額外資訊的樣式，進而更能鎖定最可能為根本問題的機台，使工程師能更迅速精準的解決問題，此研究解析根本問題機台的流程如圖 3.2。

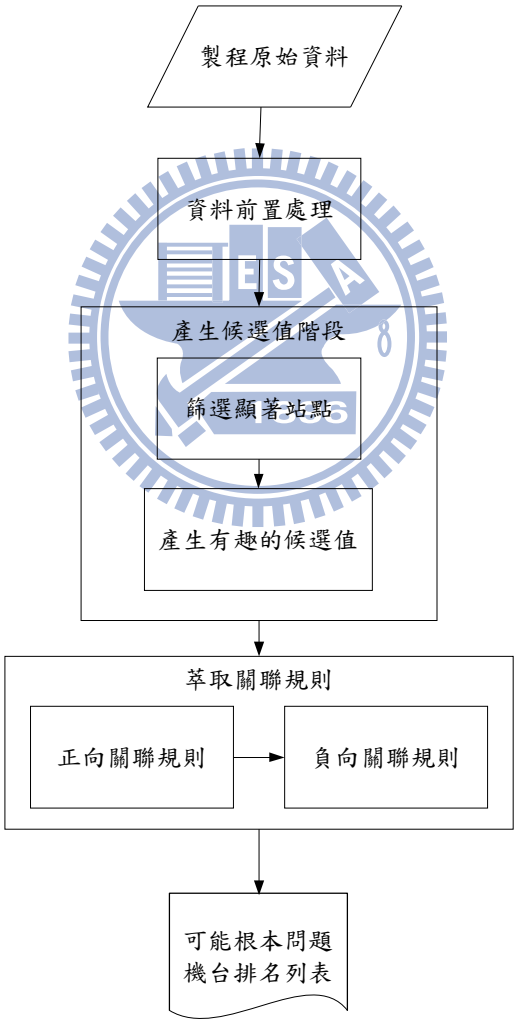


圖 3.2 根本問題機台判定流程圖

主要步驟為(1)資料前置處理、(2)產生候選值階段、(3)萃取關聯規則及(4)可能根本問題機台排名列表，以下將分別說明其詳細程序。

3.2.1 資料前置處理

製造生產過程中，資料無所不在且非單一方向的資料流，產品有產品的資料，機台端有機台的各類狀態資料，或製造流程中各項設定及實際量測的資料，而各類資料有可能以不同格式存在，如二位元的訊息格式、廠商自訂的檔案格式、或統一的 XML 格式，資料可謂是複雜又龐大，所以業界較具規模的廠商都不免需建置一套屬於製造生產過程中各類資料儲存的資料倉儲系統。而本研究從個案公司資料倉儲擷取出相關資料記錄的資料表「製程歷史記錄表」、「產品良率資料表」。

「製程歷史記錄表」為製程程序中每站製程站點的各項資料，包含製程站點、機台、及製程完成時間，一批產品批號將有多筆資料存於此記錄表內，主索引鍵欄位訂為 Lot_ID、OPE 欄位，而外索引鍵欄位則以 Lot_ID 與「產品良率資料表」關聯，資料綱要如表 3.1 所示。

表 3.1 製程歷史記錄表綱要

欄位名稱	欄位說明	變數型態	說明
Lot_ID	產品批號	文字	
OPE	製程站點	文字	ex:$Stage_1$
EQP	機台	文字	ex:$Machine_1$
主索引鍵欄位: Lot_ID+OPE；外索引鍵欄位: Lot_ID			

「製程歷史記錄表」關聯資料表示為 $r=\{w_1w_2,w_3,...,w_k,\}$，範例資料如表 3.2，以 W_1 為例，經過第一道製程站點 S_1 時，由機台 M_2 負責該道製造的程序；經過第二道製程站點 S_2 時，由機台 M_5 負責該道製造的程序；經過第三道製程站點 S_3 時，由機台 M_8 負責該道製造的程序；然在第四道製程時，並沒有機台資訊，是因為各式產品所需的製造步驟不同，所經的製造站點也可能不同，所以並無 S_4 的資料；而第五道製程 S_5 又記錄著 M_{11} 的資料，以此類推。

表 3.2 製程歷史記錄範例

Lot_ID	OPE	EQP	Lot_ID	OPE	EQP	Lot_ID	OPE	EQP	Lot_ID	OPE	EQP
W_1	S_1	M_2	W_4	S_2	M_4	W_7	S_5	M_{12}	W_{12}	S_1	M_1
W_1	S_2	M_5	W_4	S_3	M_8	W_8	S_1	M_2	W_{12}	S_2	M_4
W_1	S_3	M_8	W_4	S_5	M_{12}	W_8	S_2	M_6	W_{12}	S_5	M_{12}
W_1	S_5	M_{11}	W_5	S_1	M_1	W_8	S_4	M_9	W_{13}	S_1	M_1
W_2	S_1	M_2	W_5	S_2	M_4	W_8	S_5	M_{11}	W_{13}	S_2	M_4
W_2	S_2	M_4	W_5	S_3	M_8	W_9	S_1	M_1	W_{13}	S_3	M_8
W_2	S_3	M_7	W_5	S_5	M_{12}	W_9	S_2	M_4	W_{13}	S_5	M_{11}
W_2	S_4	M_9	W_6	S_1	M_1	W_9	S_3	M_7	W_{14}	S_1	M_3
W_2	S_5	M_{11}	W_6	S_2	M_4	W_9	S_5	M_{12}	W_{14}	S_2	M_4
W_3	S_1	M_1	W_6	S_3	M_8	W_{10}	S_1	M_1	W_{14}	S_5	M_{12}
W_3	S_2	M_4	W_6	S_4	M_9	W_{10}	S_2	M_4	W_{15}	S_1	M_3
W_3	S_3	M_7	W_6	S_5	M_{12}	W_{10}	S_5	M_{12}	W_{15}	S_2	M_4
W_3	S_4	M_{10}	W_7	S_1	M_1	W_{11}	S_1	M_1	W_{15}	S_5	M_{12}
W_3	S_5	M_{12}	W_7	S_2	M_5	W_{11}	S_4	M_{10}			
W_4	S_1	M_2	W_7	S_3	M_8	W_{11}	S_5	M_{13}			

「產品良率資料表」則為產品批號完成所需製程後，送至晶圓測試單位量測之良率記錄，每一批產品批號將只有一良率資料存於此資料表內，故此資料表的主索引鍵欄位為 Lot_ID，Yield 良率是產出產品經過晶圓測試後扣除不良品後的良品比例，Defective 則是衍生欄位，判定良率小於使用者自訂的規格的為 1，代表有缺陷的產品批號，反之則為 0，代表正常的產品批號，資料綱要如表 3.3 所示。

表 3.3 產品良率資料表綱要

欄位名稱	欄位說明	變數型態	說明
Lot_ID	產品批號	Key	
Yield	良率	數值	ex:0.88、0.92
Defective	是否為缺陷產出	衍生欄位	1:判定良率小於使用者訂定之規格
			0:判定良率大於或等於使用者訂定之規格
主索引鍵欄位：Lot_ID			

「產品良率資料表」關聯資料表示為 r={w_1w_2,w_3,...,w_k,}，範例資料如表 3.4，在 W_8 完成以上的製造程序後，送交晶圓測試單位測試良率為 0.8，若工程師將產出良率規格訂為 0.85，則產品良率小於 0.85 的將列為不良品，故 W_8 的 Defective 標示為 1，以示其為有缺陷的產出批號，相對地，W_{12} 的良率為 0.89 大於或等於 0.85，所以 W_{12} 的 Defective 標示為 0，表示為正常的產出批號。

表 3.4 產品良率資料表範例

Lot_ID	Yield	Defective	Lot_ID	Yield	Defective
W_1	0.84	1	W_9	0.78	1
W_2	0.82	1	W_{10}	0.79	1
W_3	0.83	1	W_{11}	0.86	0
W_4	0.79	1	W_{12}	0.89	0
W_5	0.75	1	W_{13}	0.9	0
W_6	0.81	1	W_{14}	0.89	0
W_7	0.83	1	W_{15}	0.88	0
W_8	0.8	1			

雖然在資料倉儲中，已經過一定程度的資料處理，但要使其能更方便及有效的應用在本研究，仍必須經過客製化的資料前置處理，資料前置處理主要分為四種形式，資料清理、資料整合、資料轉換、資料化約，[14]，在此研究主要的處理說明如下：

(1) **缺值(Missing value)的處理**:資料在儲存或處理的過程中，可能會因通訊或其他因素，造成資料遺失(Data loss)，而有缺值的產生，但在本研究都視其為無效之資料，為使其不影響研究結果，除了刪除該筆資料外，其以外索引鍵關聯之資料也將一併刪除，以確保資料完整性。

(2) **無意義資料的處理**:空白值或是非製程站點資訊對於本研究為無意義的資料，也會予以篩選排除。

(3) **資料分佈檢視**:在本研究的程序中會以 Kruskal-Wallis 無母數統計檢定法，進行初步的可疑站點篩選，所以必須進行資料分佈檢視，

為常態、非常態、偏右、偏左等，了解其特性，再視資料的分佈狀態做後續處理。

(4) **資料轉換**:資料倉儲中資料存放方式，為各類資料擷取及分析所用，以較正規化的綱要設計存放，如前面表 3.2、表 3.4 所示，然為了特定分析目的，需轉換資料格式，以便有效率及更正確的資料分析，而本研究範例資料轉換後，如表 3.5 所示，分為(a)缺陷產品歷程記錄表與(b)正常產品歷程記錄表，本研究企圖找出的是經過哪些站點機台，會造成缺陷產品批號的樣式，所以將缺陷產品批號的記錄當作是正向的交易記錄，相反的，正常產出的產品批號則為負向交易記錄。而此兩歷程記錄表在萃取正向關聯規則及負向關聯規則時，分別為其運算的基礎。

表 3.5 資料轉換後資料範例

(a) 缺陷產品歷程記錄表

Lot_ID	Items	Defective
W_1	M_2S_1，M_5S_2，M_8S_3，$M_{11}S_5$	1
W_2	M_2S_1，M_4S_2，M_7S_3，M_9S_4，$M_{11}S_5$	1
W_3	M_1S_1，M_4S_2，M_7S_3，$M_{10}S_4$，$M_{12}S_5$	1
W_4	M_2S_1，M_4S_2，M_8S_3，$M_{12}S_5$	1
W_5	M_1S_1，M_4S_2，M_8S_3，$M_{12}S_5$	1
W_6	M_1S_1，M_4S_2，M_8S_3，M_9S_4，$M_{12}S_5$	1
W_7	M_1S_1，M_5S_2，M_8S_3，$M_{12}S_5$	1
W_8	M_2S_1，M_6S_2，M_9S_4，$M_{11}S_5$	1
W_9	M_1S_1，M_4S_2，M_7S_3，$M_{12}S_5$	1
W_{10}	M_1S_1，M_4S_2，$M_{12}S_5$	1

(b) 正常產品歷程記錄表

Lot_ID	Items	Defective
W_{11}	M_1S_1，$M_{10}S_4$，$M_{13}S_5$	0
W_{12}	M_1S_1，M_4S_2，$M_{12}S_5$	0
W_{13}	M_1S_1，M_4S_2，M_8S_3，$M_{11}S_5$	0
W_{14}	M_3S_1，M_4S_2，$M_{12}S_5$	0
W_{15}	M_3S_1，M_4S_2，$M_{12}S_5$	0

3.2.2 產生候選值階段

本階段主要分為兩步驟，(1)篩選顯著站點、(2)產生有趣的候選值。對於半導體製造業來說，其製造過程耗時又繁雜， 往往一批貨，從原料進入製程程序，以至最後完成送交晶圓測試單位測試，需要約 8、90 天的週期，方能完成一批貨，其中又必須經過約 300 道程序，每道程序可能由不只一台機器負責加工，因此可能會有 3000 種以上組合搭配生產，如果將這 3000 多種製程站點、機台組合以關聯規則運算，其排列組合將呈幾何級數成長。

3.2.2.1 篩選顯著站點

故在此階段第一步驟提出業界普遍使用的 Kruskal-Wallis 無母數統計檢定法，用來檢定各製程站點之間，機台良率表現差異的情況，進而初步篩選可疑站點，降低資料維度，縮小範圍，以達更有效率的分析。因為在半導體製程的資料分佈通常不是常態分佈，所以採用此檢定法，其假設檢定 H_0:S 站點中的各機台，良率表現皆無差異；H_1:S 站點的機台中，至少有一台的良率表現有差異；且在顯著水準 α=0.05 以下的前提進行檢定。

再將有差異的各站點，依其 P-value 由小到大排序，其 P-value 越小則其站點的機台間表現差異相對顯著，依需求取前幾名站點，本研究預設將取 500 道程序的 1/10，前 50 名站點作後續的探討。

3.2.2.2 產生有趣的候選值

在關聯規則的應用中，最被廣泛應用的即為學者 Agrawl 於 1994 年[2]提出的 Apriori 演算法，首先要找到在所有交易資料裡，發生次數較頻繁的項目集，作為候選值，而頻繁與否則以支持度(support)，即其項目集的發生機率來評量，符合支持度門檻值(minimum support)，ms

的則列為第一層級的候選值，再依第一層級的候選值項目集排列組合，產生第二層級的候選值項目集，以此重複的運算，直到沒有新的或更高層級的候選值項目集產生。

然應用於問題機台的正向候選值運算中，對應的則是先將表 3.5 的 Defective=1 的資料作為正向候選值運算的所有交易資料，製程站點、機台資料則為項目集；在候選的過程中，如 Wu.等[25]在標示有趣的候選值階段所提，修剪策略在此步驟也該被考慮進來，以確保候選值的項目集，在關聯規則被萃取時，對使用者而言是有趣的；在此前提下，修剪策略的設計就關係到是否為有效率搜尋方法的一關鍵因素。

Piatetsky-Shapiro 於 1991 年[18]提出當 $supp(X \cup Y) \approx supp(X) \times supp(Y)$ 時，規則 $X \Rightarrow Y$ 並不會讓使用者覺得感興趣，因為當前因(antecedent)X 與後果(consequent)Y 兩事件相當接近獨立事件，則該規則的因果關係將變得不令人感興趣。因此 Wu.等[25]設計了一有趣度函數，如式 3.1。

$$interest(X,Y) = \left| supp(X \cup Y) - supp(X) \times supp(Y) \right| \qquad \text{式 3.1}$$

對應其有趣度函數的門檻值，則為最小有趣度 mi (minimum interestingness)，故當 $interest(X,Y) \geq mi$，則 $X \cup Y$ 為有趣的候選值，若 $interest(X,Y) < mi$，則反之，所以將其整合到此階段的有趣候選值產生階段，則為 Wu.等[25]設計的有趣的候選值函數，如式 3.2。

$$fipis(X,Y) = \left\{ \begin{array}{l} true\ ,\ then\ X \Rightarrow Y\ is\ a\ frequent\ itemset\ of\ potential\ interest \\ false,\ then\ X \Rightarrow Y\ is\ not \end{array} \right\}$$

where 式3.2

$$fipis(X,Y) : X \cap Y = \varnothing \wedge$$
$$f(X,Y,ms,mi) = 1$$
$$f(X,Y,ms,mi) = \frac{supp(X \cup Y) + interest(X,Y) - (ms+mi) + 1}{\left| supp(X \cup Y) - ms \right| + \left| interest(X,Y) - mi \right| + 1}$$

其中，X 及 Y 為兩獨立的事件，且 f(X,Y,ms,mi)=1，意味 X 及 Y 同時發生的支持度大於設定的最小支持度門檻值，及 X⇒Y 的有趣度也大於設定的最小有趣度門檻值，則 fipis(X,Y)為 true，即本研究定義之感興趣的候選值，當 fipis(X,Y)為 false 時則反之。

表 3.5(a)中 Defective=1 的資料視為正向候選值運算的所有記錄資料，在此範例中將 ms 設為 0.5，mi 設為 0.05，M_1S_1 在 10 筆記錄資料裡，有缺陷的產出批號，就包含了 W_3、W_5、W_6、W_7、W_9、W_{10} 等 6 筆記錄，其 $Supp = 6/10 = 0.6$；本研究有興趣的是站點機台間的關聯，而第一層因只有一項目，尚無法計算其有趣度，故僅以 Supp 評估是否為第二層的候選值，則第一層計算結果如表 3.6，因此，在第一層的篩選(f 函數運算結果為 1)，M_1S_1、M_4S_2、M_8S_3、$M_{12}S_5$ 選為第二層的候選值。

表 3.6 範例資料第一層正向候選值

Item	Supp	Interest	f value
M_1S_1	**0.6**	**-**	**1**
M_2S_1	0.4	-	0.82
M_3S_1	0	-	0.33
M_4S_2	**0.7**	**-**	**1**
M_5S_2	0.2	-	0.54
M_6S_2	0.1	-	0.43
M_7S_3	0.3	-	0.67
M_8S_3	**0.5**	**-**	**1**
M_9S_4	0.3	-	0.67
$M_{10}S_4$	0.1	-	0.43
$M_{11}S_5$	0.3	-	0.67
$M_{12}S_5$	**0.7**	**-**	**1**
$M_{13}S_5$	0.1	-	0.43

接著，將第一層候選值 M_1S_1、M_4S_2、M_8S_3、$M_{12}S_5$ 排列組合，產生第二層{M_1S_1、M_4S_2}，{M_1S_1、M_8S_3}，{M_1S_1、$M_{12}S_5$}，{M_4S_2、M_8S_3}，{M_4S_2、$M_{12}S_5$}，{M_8S_3、$M_{12}S_5$}，其中以{M_1S_1、M_4S_2}為例，$Supp(M_1S_1 \cup M_4S_2) = 0.5$，$Supp(M_1S_1) = 0.6$，$Supp(M_4S_2) = 0.7$，所以得

$Interest = 0.5-(0.6*0.7)=0.08$，而

$f(M_1S_1,M_4S_2,ms,mi)=\frac{(0.5+0.08)-(0.5+0.05)+1}{|0.5-0.5|+|0.08-0.05|+1}=1$，其餘以此類推，如表 3.7 所示。

表 3.7 範例資料第二層正向候選值

Itemset	Supp	Interest	f value
M_1S_1、M_4S_2	**0.5**	**0.08**	**1**
M_1S_1、M_8S_3	0.3	0	0.6
M_1S_1、$M_{12}S_5$	**0.5**	**0.14**	**1**
M_4S_2、M_8S_3	0.3	0.05	0.49
M_4S_2、$M_{12}S_5$	**0.6**	**0.18**	**1**
M_8S_3、$M_{12}S_5$	0.3	0	0.6

第三層，只有一組候選值產生出來，如表 3.8 所示，無法有新的項目集產生，所以重複運算到此告一段落，{M_1S_1、M_4S_2、$M_{12}S_5$}則為第三層的正向候選值。

表 3.8 範例資料第三層正向候選值

Itemset	Supp	Interest	f value
M_1S_1、M_4S_2、$M_{12}S_5$	**0.5**	**0.2**	**1**

在此階段，有趣的候選值第一至三層，然我們接下來要探討的是站點機台間的關聯，所以只將第二層以後的所有正向候選值整理列表，以供後續關聯規則的計算運用，如表 3.9 所示。

表 3.9 範例資料有趣的正向候選值

Itemset	Supp	Interest	f value
M_1S_1、M_4S_2	**0.5**	**0.08**	**1**
M_1S_1、$M_{12}S_5$	**0.5**	**0.14**	**1**
M_4S_2、$M_{12}S_5$	**0.6**	**0.18**	**1**
M_1S_1、M_4S_2、$M_{12}S_5$	**0.5**	**0.2**	**1**

3.2.3 萃取關聯規則

截至目前，已將有趣的候選值產生，但要如何從中找到關聯規則呢?本篇研究將採取 Wu.等[25]設計之條件機率增加比例函數(conditional-probability increment ratio)，CPIR，如式 3.3。

$$CPIR(Y|X)=\begin{cases}\dfrac{supp(X\cup Y)-supp(X)supp(Y)}{supp(X)(1-supp(Y))}, if\ p(Y|X)\geq p(Y),\ p(Y)\neq 1\\ \dfrac{supp(X\cup Y)-supp(X)supp(Y)}{supp(X)supp(Y)}, if\ p(Y)>p(Y|X),\ p(Y)\neq 0\end{cases} \quad \text{式 3.3}$$

CPIR，是由統計的相依性(Dependence)延伸而來的函數，用以判斷在 X 事件發生的情況下 Y 事件發生的相對機率是增加還是降低，若增加則代表 X 事件的發生也會提高 Y 事件的發生機率，降低則相反，以界定兩者的關聯性。Wu 也提到為了同時可發掘正向與負向關聯規則，在各種情況驗證後，使用 CPIR 作為萃取正向與負向關聯規則時的可信度，所以本研究後續提到之可信度皆為 CPIR 計算而得。

3.2.3.1 產生正向關聯規則

以表 3.8 的資料為例，若把經過 M_1S_1 的發生當作 X 事件，經過{M_4S_2、$M_{12}S_5$}的發生當作 Y 事件，藉以嘗試找出在缺陷產品記錄中的有趣候選值其中各站點機台的發生是否有其關聯性，即為經過 M_1S_1 的發生的是否會使得經過{M_4S_2、$M_{12}S_5$}的發生機率增加，則

$$Confidence(M_1S_1\Rightarrow\{M_4S_2、M_{12}S_5\})=CPIR(\{M_4S_2、M_{12}S_5\}|M_1S_1)$$

$$=\frac{supp(M_1S_1\cup\{M_4S_2、M_{12}S_5\})-supp(M_1S_1)supp(\{M_4S_2、M_{12}S_5\})}{supp(M_1S_1)(1-supp(\{M_4S_2、M_{12}S_5\}))}=\frac{0.5-0.6*0.6}{0.6*(1-0.6)}=0.583$$

，解釋為在缺陷產品批號紀錄中，經過 M_1S_1 的發生會使得經過{M_4S_2、$M_{12}S_5$}發生機率增加的可信度為 0.583，若超過使用者訂定的最小可信度門檻值 mc(minimum confidence)，在此訂為 0.4，即可定義此規則為

本篇研究中欲萃取的正向關聯規則，故將在此階段產生的正向關聯規則整理列表，如表 3.10 所示。

表 3.10 範例資料正向關聯規則

Itemset	Level	CPIR
$M_1S_1 \Rightarrow M_4S_2$	2	0.44
$M_4S_2 \Rightarrow M_1S_1$	2	0.29
$M_{12}S_5 \Rightarrow M_4S_2$	2	1
$M_4S_2 \Rightarrow M_{12}S_5$	2	0.64
$M_{12}S_5 \Rightarrow M_1S_1$	2	0.58
$M_1S_1 \Rightarrow M_{12}S_5$	2	0.58
$M_{12}S_5 \Rightarrow M_1S_1$、$M_4S_2$	3	0.67
$M_4S_2 \Rightarrow M_1S_1$、$M_{12}S_5$	3	0.43
$M_1S_1 \Rightarrow M_4S_2$、$M_{12}S_5$	3	0.58

本研究在此階段想探討的是各站點的機台間的關聯，以推估哪一站點的機台會使得其他站點機台組合發生的機率增加，以找出真正的根本問題機台，與之前幾篇研究主要是以機台組合與有缺陷的產品之間的關聯為主要探討議題不同的是，大部份探討機台組合與缺陷產品之間的關聯都只由資料集萃取正向關聯規則，因為只要證明該機台組合與缺陷產品間的正相關聯程度高過一定信心程度即可論述該機台組合造成缺陷產品的可疑成分高，然本研究因為探討的是各站點機台間的關聯，若只萃取出在缺陷產品資料中的關聯規則，可能會犯了統計學上定義的型二錯誤(Type 2 error)，如果該站點機台佔了生產流程的大部分步驟或為重要節點的機台，單以缺陷產品資料則會將其納入可疑站點機台，錯將非根本原因的站點機台當做根本原因，所以本研究提出此階段的第二步驟，產生負向關聯規則，以使萃取的關聯規則不但是我們感興趣的，更是精準正確的可疑機台。

3.2.3.2 產生負向關聯規則

當一項目在資料集中發生的次數高於最小支持度門檻值，則可能是正向的候選值，再依其各項目間的可信度找出關聯規則，則為正向關聯

規則，表示為 X⇒Y；反之，一項目不發生的次數高於最小支持度門檻值，則定義為不頻繁發生的項目，而同樣依該項目與其他項目間的可信度找出的關聯規則，則為負向關聯規則，表示為¬X⇒Y，Wu.等[25]定義的負向關聯規則須符合式 3.4 的函數運算定義。

$$iipis(X,Y)=\left\{\begin{array}{l} true\ ,then\ X\Rightarrow Y\ is\ a\ infrequent\ itemset\ of\ \ potential\ interest \\ false,then\ X\Rightarrow Y\ is\ not \end{array}\right\}$$

where 式3.4

$$iipis(X,Y):X\cap Y=\varnothing\wedge$$

$$g(\neg X,Y,ms,mi)=2$$

$$g(\neg X,Y,ms,mc,mi)=$$

$$\frac{supp(\neg X\cup Y)+confidence(\neg X,Y)+interest(\neg X,Y)-(ms+mc+mi)+1}{\left|supp(\neg X\cup Y)-ms\right|+\left|confidence(\neg X,Y)-mc\right|+\left|interest(\neg X,Y)-mi\right|+1}$$

$$+\frac{supp(X)+supp(Y)-2ms+1}{\left|supp(X)-ms\right|+\left|supp(Y)-ms\right|+1}$$

X 與 Y 為獨立之兩事件，且 g(¬X,Y,ms,mc,mi)=2，$\neg X\cup Y$ 意味 X 的不發生次數及 Y 發生次數的支持度大於設定的最小支持度門檻值，$\neg X\Rightarrow Y$ 意味 X 不發生次數增加會使得 Y 發生次數增加的可信度也大於設定的最小可信度門檻值，及 X 的不發生與 Y 的發生的有趣度也大於設定的最小有趣度門檻值，則 iipis(X,Y)為 true，及本研究定義之負向關聯規則，當 iipis(X,Y)為 false 時則不是本研究定義的負向關聯規則；其中在 g(¬X,Y,ms,mc,mi)函數中的後項，判斷 X 及 Y 事件發生的支持度是否大於最小支持度門檻值，因為本研究關注的負向關聯規則必須依據其相對應的正向候選值也需發生一定頻繁次數以上為本研究所感興趣的，所以藉由此項計算函數，產生有意義的負向關聯規則。

表 3.5(b)中 Defective=0 的資料視為負向關聯規則運算的所有記錄資料，在此範例中將 ms 設為 0.2，mc 設為 0.15，mi 設為 0.03，本研究感興趣的是藉由缺陷產品歷程紀錄表產生的正向關聯規則，其前因即 X 在正常產品歷程紀錄表中相對應的負向關聯規則，所以例如表 3.10 中正向關聯規則 $M_1S_1\Rightarrow M_4S_2$、$M_{12}S_5$，經過 M_1S_1 看作 X 事件而經過 M_4S_2、

$M_{12}S_5$ 則看作 Y 事件，所以相對應的負向關聯規則 $\neg X \Rightarrow Y$ 則表示為 $\neg M_1S_1 \Rightarrow M_4S_2$、$M_{12}S_5$，計算其

$$g(\neg M_1S_1,\{M_4S_2、M_{12}S_5\},ms,mc,mi) =$$

$$\frac{supp(\neg M_1S_1 \cup \{M_4S_2、M_{12}S_5\})+confidence(\neg M_1S_1,\{M_4S_2、M_{12}S_5\})+interest(\neg M_1S_1,\{M_4S_2、M_{12}S_5\})-(ms+mc+mi)+1}{|supp(\neg M_1S_1 \cup \{M_4S_2、M_{12}S_5\})-ms|+|confidence(\neg M_1S_1,\{M_4S_2、M_{12}S_5\})-mc|+|interest(\neg M_1S_1,\{M_4S_2、M_{12}S_5\})-mi|+1}$$

$$+\frac{supp(M_1S_1)+supp(\{M_4S_2、M_{12}S_5\})-2ms+1}{|supp(M_1S_1)-ms|+|supp(\{M_4S_2、M_{12}S_5\})-ms|+1}$$

$$=\frac{0.4+1+0.16-(0.2+0.15+0.03)+1}{|0.4-0.2|+|1-0.15|+|0.16-0.03|+1}$$

$$+\frac{0.6+0.6-2*0.2+1}{|0.6-0.2|+|0.6-0.2|+1}=2$$

其餘以此類推，如表 3.11 所示。

表 3.11 範例資料負向關聯規則

Itemset	Level	Supp	Interest	CPIR	g value
$\neg M_1S_1 \Rightarrow M_4S_2$	2	0.4	0.08	0.08	1.89
$\neg M_{12}S_5 \Rightarrow M_4S_2$	2	0.2	0.12	-0.375	1.35
$\neg M_4S_2 \Rightarrow M_{12}S_5$	2	0	0.12	-1	0.89
$\neg M_{12}S_5 \Rightarrow M_1S_1$	**2**	**0.4**	**0.16**	**1**	**2**
$\neg M_1S_1 \Rightarrow M_{12}S_5$	**2**	**0.4**	**0.16**	**1**	**2**
$\neg M_1S_1 \Rightarrow M_4S_2$、$M_{12}S_5$	**3**	**0.4**	**0.16**	**1**	**2**
$\neg M_4S_2 \Rightarrow M_1S_1$、$M_{12}S_5$	3	0	0.16	-1	-0.1
$\neg M_{12}S_5 \Rightarrow M_1S_1$、$M_4S_2$	**3**	**0.2**	**0.04**	**0.16**	**2**

其中，$\neg M_1S_1 \Rightarrow M_4S_2$、$M_{12}S_5$ 計算之 CPIR 為 1，代表在正常產品歷程紀錄裡，未經過 M_1S_1 發生次數增加會使得同時經過 M_4S_2、$M_{12}S_5$ 的發生次數增加的可信度為 100%；而 $\neg M_{12}S_5 \Rightarrow M_1S_1$、$M_4S_2$ 之 CPIR 為 0.16，則表示未經過 $M_{12}S_5$ 發生次數會使得同時經過 M_1S_1、M_4S_2 發生次數增加的可信度為百分之十六。

3.2.4 根本問題機台排序列表

以上經過有趣的候選階段產生候選值，接著分別萃取出在缺陷產品歷程紀錄裡的正向關聯規則及正常產品歷程紀錄裡的負向關聯規則，然有可能產生大量的關聯規則，但哪些關聯規則對於解決找出根本問題機台研究是有意義的，又該如何判定各關聯規則的重要性排序，本研究在

排序列表的程序裡以三種評估方法來實驗比較，其中包含(1)整合關聯規則最大可信度排序、(2)整合關聯規則可信度加總排序及(3)正向與負向關聯規則正規化可信度加總排序。

上述的(1)及(2)方法，分別利用整合關聯規則的最大可信度或可信度加總來做排序，而所謂整合關聯規則乃意指將前項與後項相對等的正向及負向關聯規則進行整合的程序，其整合流程如下。

***Algorithm**: Integration of Positive & Negative Association Rules*
***Input**: PAR: positive association rules; NAR: negative association rules;*
Wp: weight of positive impact; Wn: weight of negative impact;
(Wp+Wn=1)
***Output**: IAR: integration of positive & negative association rules*
***Method**:*
(1) IAR ← ∅; / initial of IAR*/*
*(2) **foreach** positive association rule A in PAR **do***
*(3) **foreach** expression A = {X ⇒ Y} **do***
*(4) **begin***
(5) B = ∅;
*(6) **if** {¬X ⇒ Y} exists in NAR **then***
(7) B = {¬X ⇒ Y};
(8) IAR ← IAR ∪ {X ⇒ Y }
(9) with CPIR′ = Wp A.CPIR + Wn*B.CPIR;*
*(10) **if** B=∅ **then***
(11) IAR ← IAR ∪ {X ⇒ Y}
(12) with CPIR′ = Wp A.CPIR;*
*(13) **end;***
*(14) **return.***

此流程的輸入分別有正向關聯規則集合 PAR、負向關聯規則集合 NAR，即為前幾階段萃取出來的關聯規則集合；然正向及負向關聯規則可能會依使用者的重視程度不同而有不同的權重，為輸入值的另兩項 Wp、Wn，分別代表正向關聯規則的影響權重，及負向關聯規則的影響權重，惟此兩項權重係數介於 0 到 1 之間，相加等於 1，此權重的設定

後續會以設計實驗測試及評估；整合流程的輸出即為整合關聯規則集合及各規則由正向與負向關聯規則的 CPIR 加權計算後的 *CPIR'*。

本研究提出的整合流程，主要是視正向關聯規則相對應的負向關聯規則是否存在(6)，若存在則將其正向關聯規則的可信度與負向關聯規則的可信度分別乘上正向及負向的影響權重後相加(9)而得整合關聯規則，此處的正向關聯規則是由缺陷歷程紀錄表中萃取，負向關聯規則是由正常產品歷程紀錄表中萃取，所以在意義上來說，負向關聯規則的可信度是可增加正向關聯規則的可信度，故以兩項相加的方式計算整合關聯規則的可信度；反之，若在負向關聯規則集合中並未找到正向關聯規則相對應的負向關聯規則(10)，則其最後的可信度則為正向關聯規則的可信度乘上正向關聯規則的影響權重(12)。

以範例資料表 3.10 及表 3.11 合併產生則可得表 3.12 的正向及負向整合關聯規則。

表 3.12 範例資料正向及負向整合關聯規則

Itemset	Level	CPIR	Integrated Itemset	*CPIR'* (Wn=0.5)
$M_1S_1 \Rightarrow M_4S_2$	2	0.44	$M_1S_1 \Rightarrow M_4S_2$	0.22
$M_4S_2 \Rightarrow M_{12}S_5$	2	0.64	$M_4S_2 \Rightarrow M_{12}S_5$	0.32
$M_{12}S_5 \Rightarrow M_4S_2$	2	1	$M_{12}S_5 \Rightarrow M_4S_2$	0.5
$M_1S_1 \Rightarrow M_{12}S_5$	2	0.58	$M_1S_1 \Rightarrow M_{12}S_5$	0.79
$\neg M_1S_1 \Rightarrow M_{12}S_5$	2	1		
$M_{12}S_5 \Rightarrow M_1S_1$	2	0.58	$M_{12}S_5 \Rightarrow M_1S_1$	0.79
$\neg M_{12}S_5 \Rightarrow M_1S_1$	2	1		
$M_1S_1 \Rightarrow M_4S_2$、$M_{12}S_5$	3	0.58	$M_1S_1 \Rightarrow M_4S_2$、$M_{12}S_5$	0.79
$\neg M_1S_1 \Rightarrow M_4S_2$、$M_{12}S_5$	3	1		
$M_4S_2 \Rightarrow M_1S_1$、$M_{12}S_5$	3	0.43	$M_4S_2 \Rightarrow M_1S_1$、$M_{12}S_5$	0.22
$M_{12}S_5 \Rightarrow M_1S_1$、$M_4S_2$	3	0.67	$M_{12}S_5 \Rightarrow M_1S_1$、$M_4S_2$	0.42
$\neg M_{12}S_5 \Rightarrow M_1S_1$、$M_4S_2$	3	0.16		

例如，在負向關聯規則的影響權重為 0.5 情況下，正向關聯規則 $M_1S_1 \Rightarrow M_4S_2$、$M_{12}S_5$ 與其相對應負向關聯規則 $\neg M_1S_1 \Rightarrow M_4S_2$、$M_{12}S_5$ 的 CPIR

分別為 0.58 及 1，經由乘上權重後加總，可得其一整合關聯規則 $M_1S_1 \Rightarrow M_4S_2$、$M_{12}S_5$ 及其 *CPIR′*=(1-0.5)*0.58+0.5*1=0.79，其餘項目以此類推。而完成此步驟後將以其最大可信度及可信度加總後排序評估，分別在 3.2.4.1 及 3.2.4.2 小節中介紹。

3.2.4.1 整合關聯規則最大可信度排序

以整合關聯規則最大可信度排序，即直接以整合關聯規則集合依照其可信度排序，惟此方法只考慮有相對的負向關聯規則的整合關聯規則，若未有相對的負向關聯規則不納入排序，以濾掉過多前項重複且相對無意義的關聯規則，如表 3.13，濾掉了 $M_1S_1 \Rightarrow M_4S_2$、$M_4S_2 \Rightarrow M_{12}S_5$、$M_{12}S_5 \Rightarrow M_4S_2$、$M_4S_2 \Rightarrow M_1S_1$、$M_{12}S_5$，其中以整合後的 $M_1S_1 \Rightarrow M_4S_2$、$M_{12}S_5$、$M_{12}S_5 \Rightarrow M_1S_1$ 及 $M_1S_1 \Rightarrow M_{12}S_5$ 整合後可信度 *CPIR′* 排名在第一順位，故可判定 M_1S_1、$M_{12}S_5$ 這兩機台皆有可能為最大可信度排序方法找出的根本問題機台。

表 3.13 範例資料整合關聯規則最大可信度排序

Integrated Itemset	*CPIR′* (Wn=0.5)
$M_1S_1 \Rightarrow M_4S_2$、$M_{12}S_5$	**0.79**
$M_{12}S_5 \Rightarrow M_1S_1$	**0.79**
$M_1S_1 \Rightarrow M_{12}S_5$	**0.79**
$M_{12}S_5 \Rightarrow M_1S_1$、$M_4S_2$	0.42

3.2.4.2 整合關聯規則可信度加總排序

此研究提出的可信度為站點機台與站點機台間的關聯規則可信度，意指經過 M_1S_1 站點機台的發生促使經過 M_4S_2、$M_{12}S_5$ 站點機台的發生頻率增加的可信度，所以在此評估方法欲將與 M_1S_1 站點機台所有整合關聯規則的可信度加總，視為 M_1S_1 此站點機台的排序積分，故在產生有趣候選值的第一層正向候選值項目集中所有項目(1)取出計算其排序積分，加總所有整合關聯規則裡前項與該項目相同的(7)(8)關聯規則可信度(9)，則為該項目的排序積分，如下所示。

Algorithm: *calculate sorting score of candidate item with IAR*
Input: L_1 : *frequent* 1-*itemset*; *IAR*: int*egrated association rules*;
Output: *itemset with sorting score*
Method:
(1) ***foreach*** *item I in* L_1 ***do***
(2) ***begin***
(3) ***let*** $I.Score = 0$
(4) ***for*** $n = 2; n <= IAR.\max Level; n++$
(5) ***begin***
(6) ***let*** $LAR = \varnothing$
(7) ***foreach*** *association rule* $R \leftarrow \{X \Rightarrow Y\}$ *in IAR.n-rules* ***do***
(8) ***if*** $I = X$ ***then*** $LAR \leftarrow LAR \cup R$
(9) ***if*** $LAR \neq \varnothing$ ***then*** $I.Score \mathrel{+}= \mathbf{Sum}(LAR.rules.CPIR')$
(10) ***end;***
(11) ***end;***
(12) ***return.***

所以範例資料依此方法計算，M_1S_1 的可信度加總為 0.22+0.79+0.79=1.8，依此類推，結果如表 3.14 所示，則可判定 M_1S_1 為此方法所找出的根本問題機台。

表 3.14 範例資料整合關聯規則可信度加總排序

Item	Score (Wn=0.5)
M_1S_1	**1.8**
$M_{12}S_5$	1.71
M_4S_2	0.54
M_8S_3	0

3.2.4.3 正向與負向關聯規則正規化可信度加總排序

前一評估方法根據整合後的關聯規則加總其可信度，然而在萃取階段負向關聯規則因篩選條件比正向關聯規則更為嚴苛，所以通常負向關聯規則的數量較正向關聯規則少，比例上也有蠻大的差距，又考量到本研究提出的正向與負向各最小門檻值不盡相同，所以以正規化的程序將兩者取得單位上的一致性與可比較性，其程序如下。

Algorithm: *calculate sorting score of candidate item with PAR & NAR*

Input: L_1 : *frequent* 1 - *itemset*;

PAR : *positive association rules*; *NAR* : *negative association rules*;

Output: *itemset with sorting score*

Method:

(1) ***foreach*** *item I in* L_1 ***do***

(2) ***begin***

(3) ***let*** $I.Score = 0;\ I.PosScore = 0;\ I.NegScore = 0;$

(4) ***for*** $n = 2; n <= PAR.\max Level; n++$

(5) ***begin***

(6) ***let*** $LPAR = \varnothing$

(7) ***foreach*** *association rule* $R \leftarrow \{X \Rightarrow Y\}$ *in PAR.n - rules* ***do***

(8) ***if*** $I = X$ ***then*** $LPAR \leftarrow LPAR \cup R$

(9) ***if*** $LPAR \neq \varnothing$ ***then*** $I.PosScore \mathrel{+}= \textbf{Sum}(LPAR.rules.CPIR)$

(10) ***end;***

(11) ***for*** $n = 2; n <= NAR.\max Level; n++$

(12) ***begin***

(13) ***let*** $LNAR = \varnothing$

(14) ***foreach*** *association rule* $R \leftarrow \{\neg X \Rightarrow Y\}$ *in NAR.n - rules* ***do***

(15) ***if*** $I = X$ ***then*** $LNAR \leftarrow LNAR \cup R$

(16) ***if*** $LNAR \neq \varnothing$ ***then*** $I.NegScore \mathrel{+}= \textbf{Sum}(LNAR.rules.CPIR)$

(17) ***end;***

(18) $$I.Score = Wp * \frac{I.PosScore}{\textbf{Sum}(PAR.rules.CPIR)} + Wn * \frac{I.NegScore}{\textbf{Sum}(NAR.rules.CPIR)}$$

(19) ***end;***

(20) ***return.***

將第一層正向候選值項目集中所有項目(1)取出，分別在正向關聯規則集(7)與負向關聯規則集(14)中計算其正向與負向的排序積分，首先分別找出所有與正向關聯規則前項相同的項目(8)，及所有與負向關聯規則前項相同的項目(15)，加總其正向與負向關聯規則的可信度則可得該項目的正向(9)與負向(16)的排序積分，最後分別除以正向關聯規則可信度總合與負向關聯規則可信度總合以正規化兩項數值，再分別乘以正向與負向影響權重(18)，則得此項目的最終排序積分。

以 M_1S_1 為例，

$$M_1S_1.Score=0.5*\frac{0.44+0.58+0.58}{(0.44+0.64+1+0.58+0.58+0.58+0.43+0.67)}+0.5*\frac{1+1}{(1+1+1+0.16)}=0.48$$

，以此類推，則可得其排序結果，如表 3.15 所示，則可判定 M_1S_1 為此方法所找出的根本問題機台。

表 3.15 範例資料正向及負向關聯規則正規化可信度加總排序

Item	Score (Wn=0.5)
M_1S_1	**0.48**
$M_{12}S_5$	0.41
M_4S_2	0.11
M_8S_3	0

第四章 實驗與評估

4.1 系統環境與設定

本研究提出整合正向與負向關聯規則的資料探勘應用於找出製程中根本問題機台的方法，在 AMD 3.01G Hz 四核中央處理器搭載 2G 記憶體的桌上型電腦上，以 C#.net 開發應用程式，資料庫則是 SQL Server 2005。資料以台灣某 DRAM 製造業 P 公司的製程資料為例，P 公司廠固定以每週為單位，分析該週產出的良率及製程機台歷程，探討造成該週良率低落的根本問題機台，藉此作為評估本研究提出方法找出根本問題機台的準確度。以八週良率資料為樣本，如表 4.1 所示。

表 4.1 八週樣本資料的相關製程資訊

Dataset	Data size (Bad/Good by user-defined spec)	Number of machines	Number of stages	Number of machines*stages
Case 1	323 (61/262)	769	833	3300
Case 2	304(49/255)	646	857	3383
Case 3	386(199/187)	640	853	3346
Case 4	289(114/175)	616	802	3037
Case 5	369(71/298)	674	835	3403
Case 6	315(56/259)	661	838	3344
Case 7	347(65/282)	657	839	3347
Case 8	319(80/239)	591	786	2725

舉例來說，標示 Case 1 的資料集中，以批為單位的產品批號 323 批，其中以使用者自訂良率規格為 0.86 的情況下，61 片標示為缺陷產品，262 片則為正常產品批號，而這段時間產出的產品批號製造歷程中經過了 833 道製程站點，769 台機台負責加工，每道製程站點約有 6~7 台機台不等，所以站點機台組合實際上則有 3300 組。

在關聯規則的萃取過程中，使用者訂定最小門檻值，以期更有效率篩選有價值的知識規則，如表 4.2 所示，本研究分別微調產生候選值階段的支持度(ms)、有趣度最小門檻值(mi)，萃取正向關聯程序的可信度最小門檻值(mc)及負向關聯程序的支持度(ms)、有趣度(mi)、可信度(mc)的最小門檻值，設定三種不同的參數設定，進而對本研究提出的方法架構作進一步的評估。

表 4.2 最小門檻值設定列表

Scenario no.		一	二	三
Positive threshold	ms	0.2	0.2	0.2
	mc	0.05	0.05	0.05
	mi	0.05	0.05	0.05
Negative threshold	ms	0.3	0.2	0.2
	mc	0.05	0.05	0.03
	mi	0.03	0.02	0.01

4.2 實驗與評估

以此三種參數設定透過整合正向及負向關聯規則程序進行，萃取出的正向關聯規則及負向關聯規則數量統計結果如圖 4.1 顯示，灰色斜線柱狀圖表示只有正向關聯規則的整合關聯規則數量，灰色填滿柱狀圖則表示有萃取得負向關聯規則的整合關聯規則數量，明顯可見，負向關聯規則的數量比正向關聯規則少，且數量比例差距也頗大。

4.2.1 參數設定評估

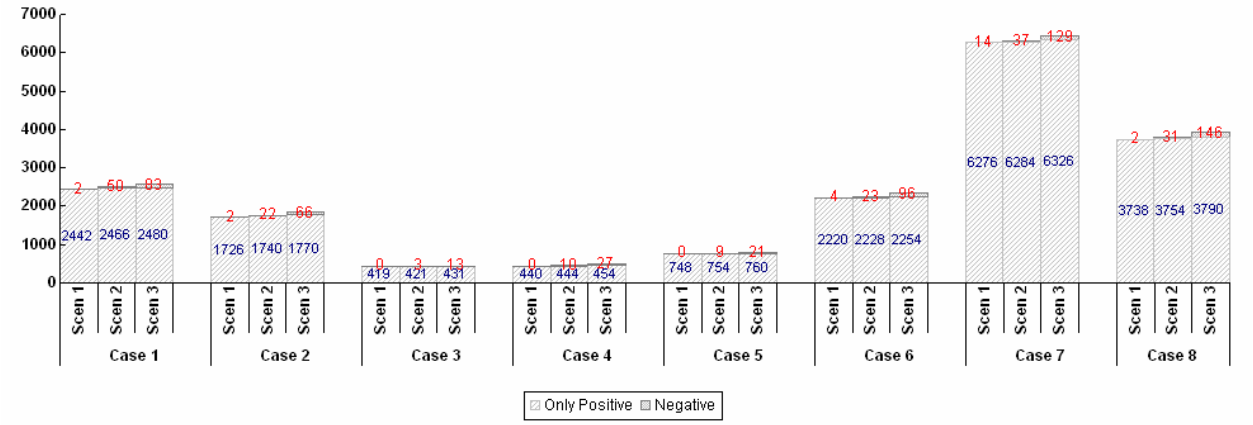


圖 4.1 正向及負向關聯規則統計圖

參數設定各有不同的負向關聯最小門檻值，如參數設定一所示，若將負向關聯程序的支持度最小門檻值設為 0.3，在本研究的實驗結果看來，限制負向關聯規則數量的效果非常顯著，主要是因為負向關聯規則在萃取的過程中，除了驗證該項目在資料集中的不發生的支持度大於支持度最小門檻值，也驗證該項目的支持度大於最小門檻值，如式 3.4 中的加法後項所示，其涵義是為了確保萃取所得的負向關聯規則是有意義的例外狀況，而非該項目發生與不發生支持度都小於最小門檻值的偏離值(outlier)，然在本研究的實驗中或許 0.3 在條件上是相對嚴苛的，所以參數設定二、三將負向關聯程序的支持度最小門檻值設為 0.2，在後續的實驗結果看來，既能避免沒意義的負向關聯規則，又不至於萃取的數量過少，不足以分析評估，或許是較適宜的。

可信度的最小門檻值設定目的是為了確保萃取所得的是強關聯規則，負向關聯規則裡代表的是前項的不發生使得後項發生的可信度，在半導體產業，站點機台組合大部分是照生產計劃安排的流程執行製造加工，此意味著某批產品執行過某特定站點機台與某特定站點機台的組合有可能是很固定的，所以要萃取其相反的負向關聯規則是較困難的，若可信度最小門檻值設定過高，還有可能會完全濾掉本研究定義的負向關

聯規則，所以可由圖 4.1 各負向關聯規則程序的可信度最小門檻值結果而得知，參數設定三的 0.03 比參數設定一、二的 0.05 萃取的數量相對的多。

由上而知，無論是支持度、可信度或是有趣度的最小門檻值設定將影響萃取關聯規則的數量及其有效性，但哪種設定組合適宜，將在後續的實驗結果來探討。

4.2.2 實驗結果

實驗分為(1)整合關聯規則最大可信度排序、(2)整合關聯規則可信度加總排序及(3)正向與負向關聯規則正規化可信度加總排序三種方法進行評估，實驗結果分別如圖 4.2、4.3 及 4.4，X 軸為負向關聯的影響權重設定值，Y 軸為其排序，每個案例在實際資料上都有其相對應，由工程師監測分析該週最可疑的問題站點機台，依各實驗結果的排序列表，該實際問題站點機台在排序列表中的排名，如圖 4.2(a)案例 7，其數列資料皆為 11，意指該週實際的問題站點機台在整合關聯規則最大可信度排序結果裡，排名 11，與案例 1 實際問題站點機台與排序結果相符的 1 相比，案例 1 的表現比案例 7 理想，而有些未有數值的代表實際問題站點機台並未在排序結果中，如圖 4.2(a)的案例 2、3、4、5、6、8 是因為整合關聯規則可信度最大排序法裡，只考慮有負向關聯規則的整合關聯規則，所以當最小門檻值較嚴苛時，大部分項目都無法排序，故在此圖未有數值可比較，其餘的以下將分別舉證說明實驗結果。

(1) 由圖 4.2、4.3 及 4.4 可得知，或許排序的方法不同，但大部分的實驗結果皆以將負向關聯規則納入考量的方法表現較佳，更甚者有些案例在整合關聯規則可信度加總排序，圖 4.3(b)、(c)除了案例 7、8 以外，只考慮負向關聯規則即 Wn=1 的情況下，反而排名表現會是最好的。

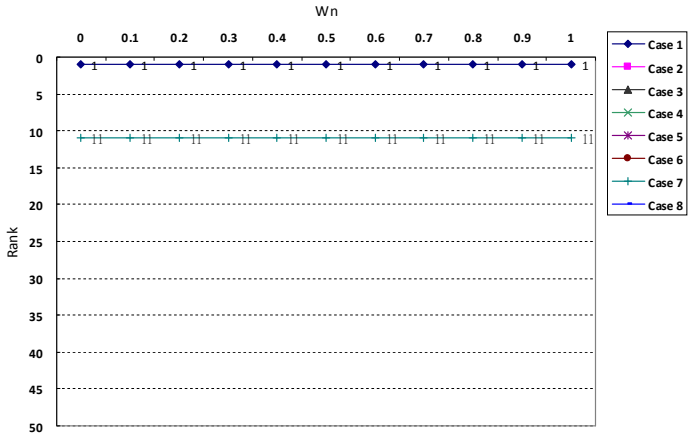
(2) 惟圖 4.3(b)、圖 4.4 (b)的案例 7、8，及圖 4.3(c)、4.4(c)的案例 7 趨勢與實驗結果(1)較不同，推論與案例 7、8 所找到的正向關聯規則與負向關聯規則數量比例懸殊有關，因找到的正向關聯規則數量龐大，資料充足，而若以加總的方式，則會相對的比負向關聯表現佳，而案例 8 到了參數設定三的情況下，則又因負向關聯規則數量有所提升，而與其他結果趨勢相同，可見負向關聯規則與正向關聯規則的數量比例，在模型評估上不可忽視其重要性。

(3) 參數設定在此程序中扮演相當重要的角色，在正向與負向關聯規則正規化加總方法結果看來，若在適當的參數設定下，排序結果有顯著的增長。

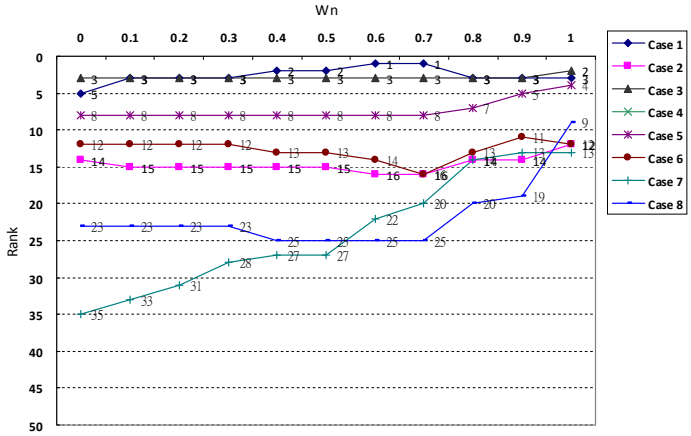
(4) 在整合關聯規則可信度加總與正規化加總排序兩種方法，由正規化加總排序在表現上變異較小的比較結果看來，正規化的程序的確能降低正向關聯規則與負向關聯規則在數量上的不對等性，使得結果較合理。

圖 4.2 整合關聯規則最大可信度排序結果

(a) 參數設定一



(b) 參數設定二



(c) 參數設定三

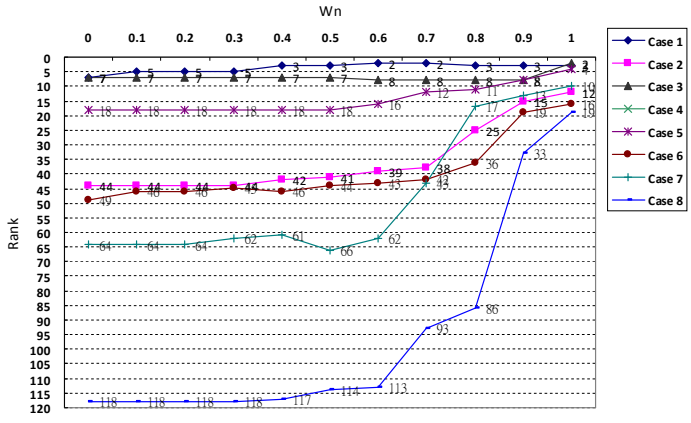
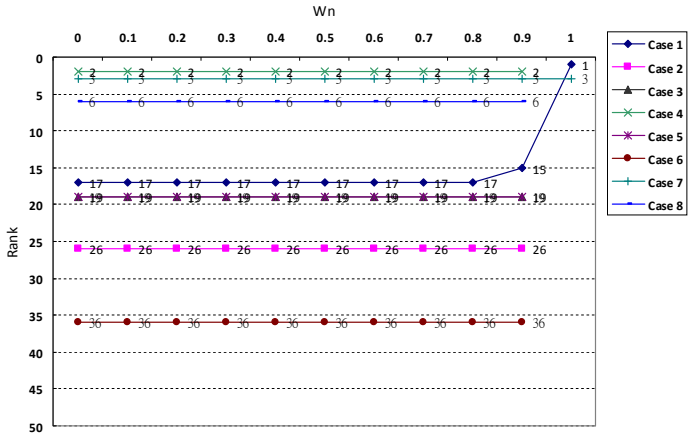
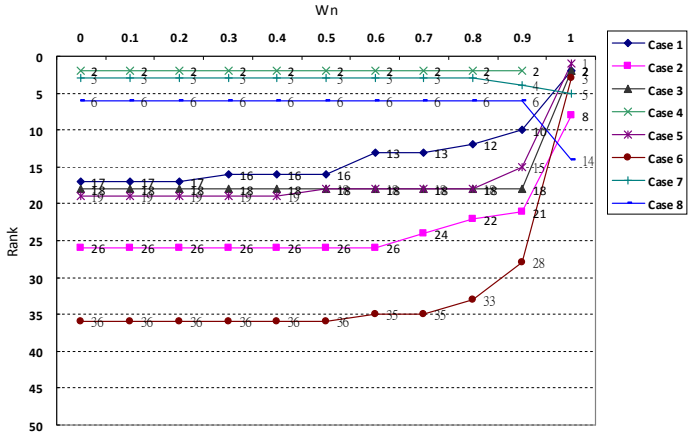


圖 4.3 整合關聯規則可信度加總排序結果

(a) 參數設定一



(b) 參數設定二



(c) 參數設定三

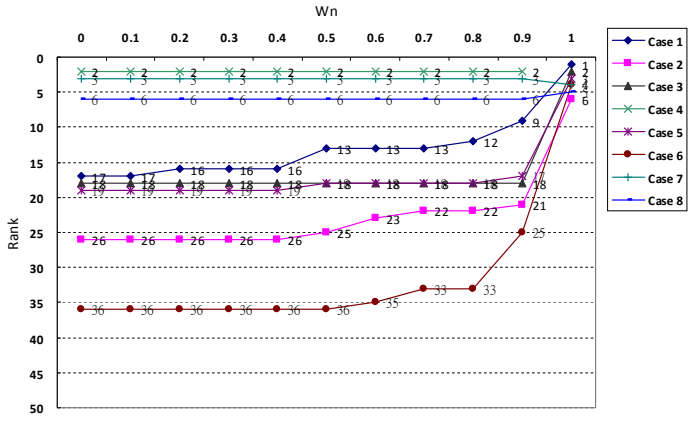
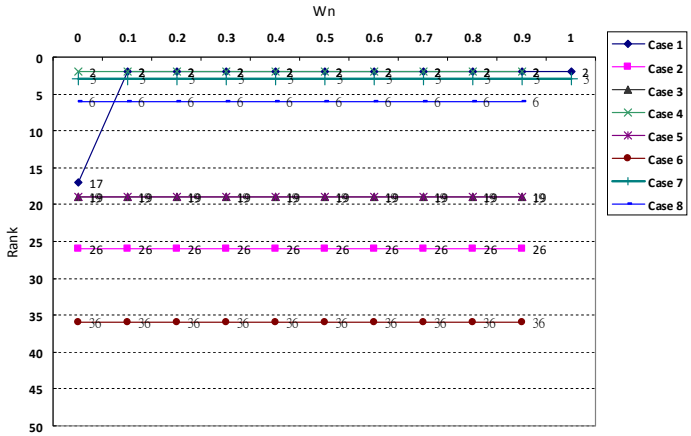
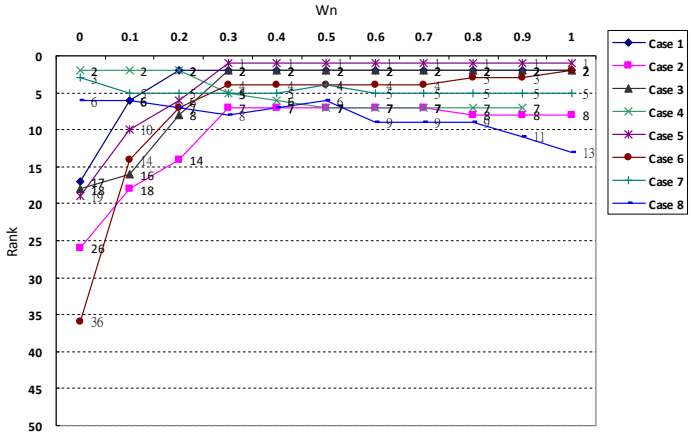


圖 4.4 正向與負向關聯規則正規化可信度加總排序結果

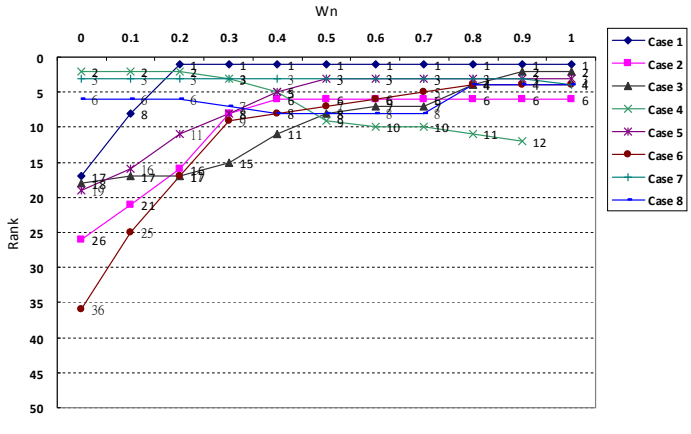
(a) 參數設定一



(b) 參數設定二



(c) 參數設定三

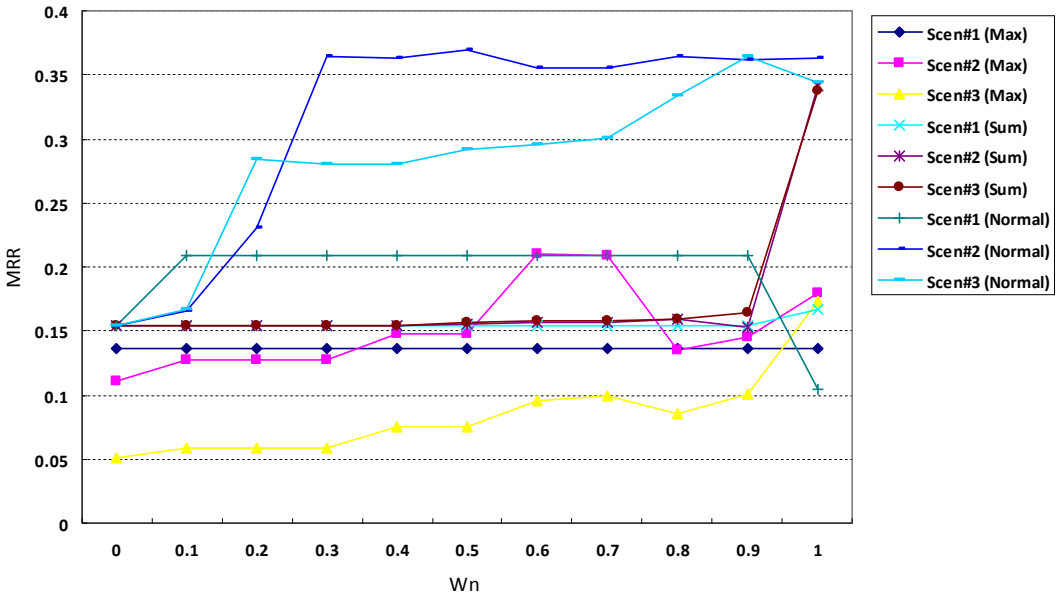


最後，本研究以 MRR(Mean Reciprocal Rank)比較評估此三種方法，MRR 是以排名的倒數平均而得一績效評估指標，分數愈高則績效表現愈好，其公式如式 4.1。

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{Q}\frac{1}{rank_i} \qquad \text{式 4.1}$$

Q 代表各方法中該參數設定的案例個數，$rank_i$ 則是第 i 個案例的排名結果，MRR 則將所有排名結果的倒數相加後平均，若未有排名結果本研究將視其倒數為 0，其實驗結果如圖 4.5 所示。

圖 4.5 實驗結果 MRR



結果顯示，以整合關聯規則可信度加總的方法是有其效益的，但若先分別將正向與負向關聯規則的可信度正規化後再加總，絕大部分案例表現有更明顯的增加，且更能正確對等的調整正向與負向影響權重。

第五章 結論與未來研究方向

5.1 結論

透過本研究提出的架構，可得到以下幾點效益:

(1) **有效率的篩選可疑站點**:因關聯規則運算需層代的反覆運算，往往產生太多又不知是否重要的規則，尤以負向關聯的萃取程序更甚，故本研究於產生候選值階段第一步驟提出 K-W 檢定法，初步篩選顯著站點，以降低資料維度，降低運算成本。

(2) **修剪剔除不有趣的候選值**:產生候選值階段第二步驟除了以支持度最小門檻值，更加入有趣度最小門檻值，使產生的候選值不只符合發生條件，更是使用者感興趣的組合，使其後續產生關聯規則時可更有效率的運算。

(3) **有效的分析**:同時考慮造成站點機台與站點機台間在缺陷產品資料中發生頻繁的正向關聯規則與在正常產品資料中該項目不發生的負向關聯規則，在解決找出半導體製程中根本問題機台有明顯的效果，因在半導體製造業中，站點機台的組合照著其生產計畫搭配加工，發掘而得的負向關聯規則反而是關鍵少數，更得以增加分析的全面性及可靠度，使關聯規則的探勘方法更具效益。

5.2 未來研究方向

本研究提出整合正向與負向關聯規則，以增加關聯規則在分析資料的全面性，然由實驗結果看來，若需要有較好的分析結果仍需建立在合適的最小門檻值設定上，因產生有趣候選值階段，及後續的關聯規則建

立，皆以使用者設定其支持度、有趣度及可信度的最小門檻值為是否該關聯規則屬強關聯規則的依據，故不難察覺支持度、有趣度及可信度的最小門檻值的設定是重要的標準，然本研究並未提供一勞永逸的方法，仍採用實驗方式來驗證本研究主軸的負向關聯規則貢獻度，故在往後的研究，該繼續努力在支持度、有趣度及可信度的最小門檻值設定適宜性的研究上。

另在整合正向與負向關聯時使用的負向關聯影響權重，若能接續本研究提出合理權重的設計，或是回饋式的訓練此模型，定能使此研究更加全面、完整。

參考文獻

一、英文部分

1. *Agrawal, R., T. Imielinski, and A. Swami, Mining association rules between sets of items in large databases. SIGMOD Record, 1993.* ***22****(2): p. 207-216.*

2. *Agrawl, R., &Srikant, R., Fast algorithm for mining association rules, in Proceedings of ACM VLDB Conference. 1994. p. 487-499.*

3. *Chao, K.M., et al., An expert system to generate associativity data for layout design. Artificial Intelligence in Engineering, 1997.* ***11****(2): p. 191-196.*

4. *Chen, M.C., Configuration of cellular manufacturing systems using association rule induction. International Journal of Production Research, 2003.* ***41****(2): p. 381-395.*

5. *Chen, M.C. and H.P. Wu, An association-based clustering approach to order batching considering customer demand patterns. Omega-International Journal of Management Science, 2005.* ***33****(4): p. 333-343.*

6. *Chen, W.C., S.S. Tseng, and C.Y. Wang, A novel manufacturing defect detection method using association rule mining techniques. Expert Systems with Applications, 2005.* ***29****(4): p. 807-815.*

7. *Chien, C.F., W.C. Wang, and J.C. Cheng, Data mining for yield enhancement in semiconductor manufacturing and an empirical study. Expert Systems with Applications, 2007.* ***33****(1): p. 192-198.*

8. *Choudhary, A.K., J.A. Harding, and M.K. Tiwari, Data mining in manufacturing: a review based on the kind of knowledge. Journal of Intelligent Manufacturing, 2009.* ***20****(5): p. 501-521.*

9. *Da Cunha, C., B. Agard, and A. Kusiak, Data mining for improvement of product quality. International Journal of Production Research, 2006.* ***44****(18-19): p. 4027-4041.*

10. *Fayyad, U., G. PiatetskyShapiro, and P. Smyth, The KDD process for extracting useful knowledge from volumes of data. Communications of the Acm, 1996.* ***39****(11): p. 27-34.*

11. *Feyyad, U.M., Data mining and knowledge discovery: making sense out of data. IEEE Expert, 1996.* ***11****(5): p. 20-25.*

12. *Freitas, A.A., On rule interestingness measures. Knowledge-Based Systems, 1999.* ***12****(5-6): p. 309-315.*

13. *Gardner, M. and J. Bieker, Data mining solves tough semiconductor manufacturing problems, in Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. 2000, ACM: Boston, Massachusetts, United States. p. 376-383.*

14. *Han, J. and M. Kamber, Data mining : concepts and techniques. 2nd ed. The Morgan Kaufmann series in data management systems. 2006, Amsterdam ; Boston San Francisco, CA: Elsevier ; Morgan Kaufmann. xxviii, 770 p.*

15. *Li, T.S., C.L. Huang, and Z.Y. Wu, Data mining using genetic programming for construction of a semiconductor manufacturing yield rate prediction system. Journal of Intelligent Manufacturing, 2006.* ***17****(3): p. 355-361.*

16. *Luiza, A.M. and R.Z. Osmar. An Associative Classifier Based on Positive And Negative Rules. in Proceedings of The 9th ACM SIGMOD Workshop on Research Issues in Data Mining And Knowledge Discovery. 2004. Paris, France.*

17. *Mieno, F., et al., Yield improvement using data mining system. 1999 IEEE International Symposium on Semiconductor Manufacturing Conference Proceedings (Cat No.99CH36314)|1999 IEEE International Symposium on Semiconductor Manufacturing Conference Proceedings (Cat No.99CH36314), 1999: p. 10.1109/ISSM.1999.808818.*

18. *Piatetsky-Shapiro, G. and W.J. Frawley, Knowledge discovery in databases. 1991, Menlo Park, Calif.: AAAI Press : MIT Press. xii, 525 p.*

19. *Raghavan, V., Application of decision trees for integrated circuit yield improvement. 13th Annual IEEE/SEMI Advanced Semiconductor Manufacturing Conference. Advancing the Science*

and Technology of Semiconductor Manufacturing. ASMC 2002 (Cat. No.02CH37259), 2002: p. 262-5|v+430.

20. *Rojas, A. and A.K. Nandi, Practical scheme for fast detection and classification of rolling-element bearing faults using support vector machines. Mechanical Systems and Signal Processing, 2006.* ***20****(7): p. 1523-1536.*

21. *Savasere, A., E. Omiecinski, and S. Navathe. Mining for strong negative associations in a large database of customer transactions. in Data Engineering, 1998. Proceedings., 14th International Conference on. 1998.*

22. *Shao, X.Y., et al., Integrating data mining and rough set for customer group-based discovery of product configuration rules. International Journal of Production Research, 2006.* ***44****(14): p. 2789-2811.*

23. *Wang, X.Z. and C. McGreavy, Automatic classification for mining process operational data. Industrial & Engineering Chemistry Research, 1998.* ***37****(6): p. 2215-2222.*

24. *Wang, Z.H., et al., Integration of variable precision rough set and fuzzy clustering: An application to knowledge acquisition for manufacturing process planning, in Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Pt 2, Proceedings, D. Slezak, et al., Editors. 2005. p. 585-593.*

25. *Wu, X.D., C.Q. Zhang, and S.C. Zhang, Efficient mining of both positive and negative association rules. Acm Transactions on Information Systems, 2004.* ***22****(3): p. 381-405.*

26. *Yao, H. and H.J. Hamilton, Mining itemset utilities from transaction databases. Data & Knowledge Engineering, 2006.* ***59****(3): p. 603-626.*

27. *Yi-Hung Liu, H.-P.H., and Yu-Sheng Lin, Attribute Selection for the Scheduling of Flexible Manufacturing Systems Based on Fuzzy Set-Theoretic Approach and Genetic Algorithm. Journal of the Chinese Institute of Industrial Engineers, 2005.* ***22****: p. 46-55.*

28. *Yongjian, F., Data mining. Potentials, IEEE, 1997.* ***16****(4): p. 18-20.*

29. *Yuan, X.H., et al., Mining negative association rules, in Iscc 2002: Seventh International Symposium on Computers and*

Communications, Proceedings, A. Corradi and M. Daneshmand, Editors. 2002, Ieee Computer Soc: Los Alamitos. p. 623-628.

30. *Zhang, C.Q. and S.C. Zhang, Association rule mining - Models and algorithms - Introduction, in Association Rule Mining: Models and Algorithms. 2002. p. 1-+.*

二、 中文部分

31. *林鼎浩, 建構半導體製程資料挖礦架及其實證研究, in 工業工程與工程管理學系. 2000, 國立清華大學. p. 102.*

32. *陳順宇、鄭碧娥, 統計學, ed. 第四版. 2004, 台北: 華泰書局.*

33. *楊景晴, 整合決策樹與關聯規則之資料挖礦架構及其實證研究, in 工業工程與工程管理學系. 2003, 國立清華大學. p. 89.*

34. *廖泰翔, 以 Grnn 為預測工具之虛擬量測, in 製造工程研究所碩博士班. 2007, 國立成功大學. p. 46.*