半導體製程資料特徵萃取與資料挖礦之研究

簡禎富、李培瑞、彭誠湧
清華大學工業工程與工程管理研究所

摘要

在半導體製造程序中，許多資料會以自動或半自動方式記錄下來。包括產品的基本資料、過站時間與機台紀錄、機台設定參數、測試資料等。由於資料維度與數量龐大且混雜的雜訊等問題，傳統統計分析方法有其限制；而工程師亦往往無法從收集的龐大資料中，迅速有效地察覺可能導致製程異常的原因。本研究目的係分析半導體多維度資料，並以具體實證研究說明，包含製程監控與事故診斷兩大部分：第一部份針對半導體製程晶圓允收測試參數資料的多維度資料，透過人工類神經網路之自我組織映射成圖網路演算法先將資料分群，以發現隱藏於資料中的樣型與良率間的關連性，再以決策樹將類別之特徵以樹狀結構呈現，透過參數表現特徵提供給工程師監控製程變化的決策依據，以改善製程提昇良率。第二部分則針對半導體製程製造測試中的電性功能針測的多維度資料，以良率爲目標變數，透過人工類神經網路之自我組織映射成圖網路演算法先將資料分群，以發現隱藏於資料中的樣型，將相同特徵者歸爲一類，再由定義之分類作爲目標以決策樹將其各分群之特徵以樹狀結構呈現其分類規則。透過綜合資訊的比較縮小診斷範圍，提供給工程師作爲事故診斷的決策依據，以快速排除事故提昇良率。

關鍵字：資料挖礦、決策樹、自我組織映射成圖網路、半導體製程、決策分析

Semiconductor Manufacturing Data Mining for Clustering and Feature Extraction

C.-F. Chien, P.-R. Lee, C.-Y. Peng
Department of Industrial Engineering & Engineering Management,
National Tsing Hua University

ABSTRACT

Owing the rise of e-commerce and information technology, a large amount of data has been automatically or semi- automatically collected in modern industry. Decision makers may potentially use the information buried in the raw data to assist their decisions through data mining for possibly identifying the specific patterns of the data. This study proposes data mining procedures for analyzing semiconductor manufacturing data for manufacturing process monitoring and defect diagnosis. In particular, SOM is applied for clustering and decision tree is applied for feature extraction to analyze multi-dimensional semiconductor manufacturing data. We used real data from a fab to conduct two case studies for validation and found that this approach can effectively limit the scope for defect diagnosis and summarize the findings in specific decision rules. We conclude this study with discussions on the results and future research.

Key words: data mining, decision tree, SOM, semiconductor manufacturing, decision analysis

壹、前言

在企業製商整合與電子化的經營環境下，大量被儲存的資料所扮演的角色屬於資產還是負債，取決於其資料價值的發揮與否。資料挖礦(data mining)是從大量資料中以自動或是半自動的方式來探索（explore）和分析資料以發掘出潛在有用的資訊，例如，有意義的樣型（Pattern）或規則（Rule）等（Berry & Linoff 1997）；並可進一步將分析後的資訊整理歸納以作爲決策之依據，而累積的資訊與決策經驗均成爲企業知識管理的具體實踐。

半導體的事故診斷通常需要依靠工程師的領域知識，儘管有些公司已引進統計與工業工程手法進行製程監控與分析，對於事故診斷大都仰賴檢視基本的統計量或者是採無母數檢定比較其參數或機台表現差異。另一方面由於資料維度與數量龐大，變數之間複雜的交互作用，且加上資料收集過程混雜的雜訊問題，傳統統計分析方法有其限制。目前，相關研究大都以製程站別或機台的差異來找出可能發生變異來源的機台；僅有少數研究透過多變量的群聚分析技術，將多維度資料的不同群聚區隔開來，並選擇合適的規則以歸納描述對應的特徵，提供給工程師作爲事故診斷的參考(Gandner & Bieker 2000；簡禎富等 2001)。

本研究目的係針對半導體製程事故診斷的資料進行特徵萃取與描述，透過人工類神經網路之自我組織映射成圖網路演算法(self-organizing map; SOM)先將多維度資料分群，以探索性資料挖礦方法發現隱藏於資料中的樣型與良率間之關連性與分佈狀況，再以決策樹（decision tree）將良率異常類別之特徵以樹狀結構呈現，並轉換爲分類規則，提供給工程師作爲監控製程變化與事故診斷的決策依據，藉以提昇良率；本研究並與某半導體廠合作以實證研究具體說明研究方法與步驟。本研究包含製程監控與事故診斷兩大部分：第一部份針對半導體晶圓允收測試的多維度資料，透過人工類神經網路之自我組織映射成圖網路演算法先將資料分群，以發現隱藏於資料中的樣型與良率間的關連性，再以決策樹分析各類別之特徵而以樹狀結構呈現，透過參數表現特徵提供給工程師監控製程變化的決策依據。第二部分則針對半導體製造測試的多維度資料，以良率爲目標變數，透過人工類神經網路之自我組織映射成圖網路演算法先將資料分群，以發現隱藏於資料中的樣型，將相同特徵者歸爲一類，再由定義之分類作爲目標以決策樹分析各類別之特徵而以樹狀結構呈現，並轉換爲具體分類規則，以改善製程提昇良率。

貳、資料挖礦與資料挖礦工具

一、資料挖礦

在過去，要將大量的資料完整的保留下來，需耗費大量的成本，許多的限制阻礙資料的有效儲存、分析與資訊管理。隨著資訊科技日益進步，各種消費、製造、服務甚至診斷等記錄可以被大量、輕易的儲存下來。在電腦的輔助之下，從大量資料中探索、挖掘出隱藏於其中的訊息。發揮資料的價值，將其轉換成對於企業有益的資訊或知識，做爲正確且快速決策支援的參考，已成爲現代經營決策的重要方向。Fayyad(1997)定義「資料庫知識發現」(Knowledge Discovery in Database; KDD)的過程：「一種非顯而易見，但卻是有效、有趣及可能有用且容易解讀的資料樣型建立過程」。整個知識發現的步驟如表 1：

表 1：KDD 主要步驟及其內容（Fayyad et al., 1996）

KDD 主要步驟：	步驟內容大要：
1.學習應用領域 (learning the application domain)	學習領域相關知識及應用目標的探索學習。
2.建立目標資料集合 (creating the target dataset)	選擇適合分析的資料集合或者定義目標的變數集合。
3.資料淨化與前置處理 (data cleaning and preprocessing)	包括空白值(empty)或遺漏值(missing)資料的處理與補值的基本整理。
4.資料投影與簡化 (data projection and reduction)	依據分析目標將資料作適當的轉換。
5.選擇資料挖礦的功能 (choosing the function of data mining)	決定演算法最終要推導出的屬於哪種類型，例如：摘要(summarization)、分類(classification)、分群(clustering)等。
6.選擇資料挖礦演算法 (choosing the data mining algorithms)	針對問題類型尋找有用的工具或技巧。
7.資料挖礦 (data mining)	從感興趣的資料中以分類規則(classification rules)、決策樹、回歸(regression)、群集分析(clustering analysis)、順序性模式(sequence modeling)等方式呈現有意義的樣型。
8.解釋、描述 (interpretation)	透過一種可以被理解、確認、觀察和再利用的表達方式來呈現所得到的資訊。
9.使用所探索出的知識 (using discovered knowledge)	將所獲得的有價值資訊應用於實際系統上，加以驗證。

其中資料挖礦爲資料庫知識發現的核心步驟。Berry & Linoff (1997)定義資料挖礦爲：「資料挖礦爲了要發現有意義的樣型或規則，必須從大量資料中以自動或半自動方式探索(exploration)與分析資料」。Kleissner（1998）亦提出知識發現的循環週期可分爲資料選擇(data selection)、資料清理(data cleaning)、資料豐富化與編碼(data enrichment and coding)及資料挖礦四個步驟。其中對資料挖礦解釋爲：「資料挖礦是新式地、反覆循環的決策支援分析過程，透過此過程從組合的資料中發現具價值且隱藏其中的知識，以提供給企業專家參考」，此種決策過程屬於「探索導向」(Discovery-Driven)，而非「假設導向」(Assumption-Driven)。

資料挖礦所處理的問題類型 (Berry & Linoff 1997; Pyle 1999; Two Crows Corporation 1998)，可分爲下列四種：

㈠分類(*classification*)：

將資料中各屬性（Attribute）分門別類地加以定義，透過訓練大量資料後，所得到的規則來建立類別(Class)模式。例如，鳶尾花分類問題，利用輸入花瓣及花萼的長度、寬度，訓練模型的分類穩定性以建立區分三種不同花種的規則。

㈡預測(*prediction*)：

利用過去的歷史資料來預測未來可能發生的行為或結果。例如，股票價位的預測、銷售量的推估等。預測分析的資料大多是時間序列型的資料，此種類型的資料會隨著時間增加而大量累積。

㈢分群(*clustering*)：

透過相似程度的定義將資料分別不同的群集。其中相似程度可以利用不同的距離或相似度(similarity)來定義。而群集結果的意義要靠事後的闡釋才能衡量。因此，找出群集本身並非目的，瞭解集群的意義才是重要的。譬如，透過群集分析瞭解信用卡顧客的特殊消費樣型或者市場區隔。

㈣關聯規則分析(*association rule*)或購物籃分析(*market basket analysis*)：

透過資料尋找同時發生的事件（event）或記錄（record）並加以分析且以規則的形式來表達搜尋結果。例如，超市顧客的交易記錄可能會發現：「若」顧客 A 在星期五晚上買了啤酒，「則」他也會買尿布。這樣的關聯規則可以幫助超市決策者擬定銷售策略及賣場擺設方式。

在確認問題類型後，必須選擇適合的資料挖礦工具，其中包括傳統的統計分析，以及人工類神經網路(Artificial Neural Network)、決策樹、關聯規則、基因演算法(genetic algorithm)等。

資料挖礦已經應用到許多領域，如：市場行銷(Brachman et al. 1996)、財務投資(Deboeck & Kohonen 1998)、製造生產(Milne & Renoux 1998)等。資料挖礦過程中，其實是不斷地重複四個步驟（Berry & Linoff 1997）：（1）確認問題：定義問題，了解問題的本質和分析的目標。（2）資料分析：使用合適的工具從龐大的資料中挖掘有用的資訊。（3）採取行動：根據所得到的有用資訊採取行動，做出決策。（4）評估結果：根據執行成果來評估這次資料挖礦的成效，並有效地運用這次結果與經驗反覆修正模式，作為下一循環之改善行為，以建立決策支援的機制。

二、資料挖礦工具

㈠人工類神經網路：自我組織映射網路

人工類神經網路涵蓋軟體與硬體，是指利用大量簡單的相連人工神經元來模仿生物神經網路能力的網路架構。整個網路架構是從外界環境或其他人工神經元取得資訊，經過運算後輸出結果到外界環境或其他人工神經元（Gurney 1997）。當有大量資料要被模型化而其物理意義尚未被瞭解，以致無法使用統計方法時，人工類神經網路就變的非常有用，但由於訓練的過程是屬於黑箱作業（Black Box）的方式，所以在使用之後會很難對模型參數作具物理意義的解釋（Kittler & Wang 1999）。

人工類神經網路依照其學習方式的不同亦可分成兩大類：監督式(supervised)及非監督式(unsupervised)神經網路。監督式神經網路是一種萃取輸入與輸出之間關係的技術(Deboeck and Kohonen, 1998)，透過適應地、反覆地學習程序偵測輸入與輸出之間的關係。由於有明確的輸出或者可稱作學習的目標(target)可供學習修正之用，故稱作監督式。其輸入與輸出間的關係可以轉換為數學式表達，因此可以用來預測或支援決策的訂定。監督式學習網路通常會以誤差函數(error function)來衡量學習的品質。非監督式學習網路則是一種

用來區別(classifying)、組織(organizing)及視覺化(visualizing)大量資料的技術(Deboeck and Kohonen, 1998)，利用降低網路優勝單元的連結加權值所構成的向量與輸入向量間的距離，以達到每一個輸出單元的連結加權值向量代表一群訓練範例樣本在樣本空間中的聚類中心。

自我組織映射網路是非監督式學習網路的一種，屬於資料導向的演算法則，其輸出層的類神經元是以矩陣方式排列於一維或二維的空間中。根據目前的輸入向量，以競爭方式取得調整鍵結值向量的機會，而最後輸出層的類神經元會根據輸入向量的「特徵」以有意義的「拓樸結構」(topological structure)展現於輸出的空間(蘇木春、張孝德 1999)，其網路架構如圖 1。自我組織映射網路的基本學習法則如下（林昇甫、洪成安 1993）：

(1)設定初始值，包括鄰域集合的初始大小及初始鍵結值。
(2)依時間，輸入圖樣向量。
(3)尋找獲勝神經元並計算其輸出。
(4)調整優勝神經元及其鄰域的鍵結值。
(5)回到步驟二直至學習速率降至零。

學習後穩定的網路架構，便可藉由輸出神經元的相對位置展現。人工類神經網路的自我組織映射成圖演算法可用於多變量群聚區隔的模型選擇上。這種演算法的好處在於能夠處理大量且高維度的多變量資料，且能保留資料所隱含的資訊。透過向量量化與向量投影，將多維度的資料映射到二維的拓樸座標上，透過視覺化的方式呈現。對於群聚結果的表達，在人類訊息處理的接受度方面也較高(Kohonen 1995)。目前相關研究與應用領域包含製程監控與分析(Kasslin et al. 1992; Kohonen 1996)、材料科學應用(Kessler et al. 1993; Cai 1994)、財務預測(Deboeck & Kohonen 1998; Back et al. 1997; Kiviluoto 1998)與樣型識別(Lampinen & Oja 1995; 薛如珊 2001)等。

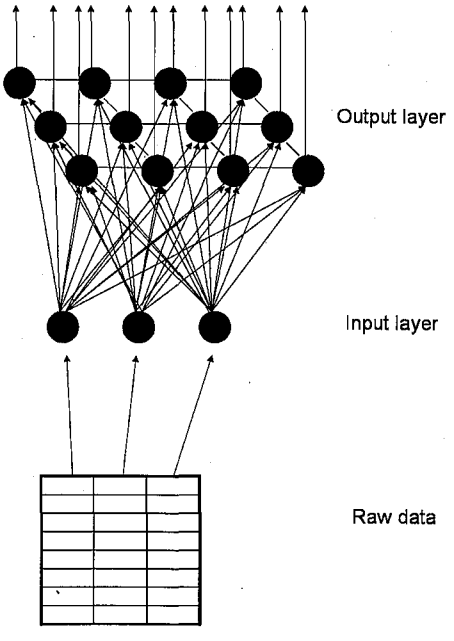

圖 1：SOM 網路基本架構

㈡決策樹

特徵萃取的演算法亦可分為監督與非監督式兩大類(Bursteinas & Long 2000)。非監督式的特徵萃取，如主成分分析法(PCA)，不會使用到分類器，通常用在訊號再現與處理。監督式的特徵萃取則，例如決策樹(decision tree)，會選擇最有差異的方式將資料類別作區隔。決策樹能夠將變數依據某種規則或資料特性對資料進行分類，通常扮演監督式特徵萃取與描述的角色，解決分類(classification)型態的問題。透過變數的選擇與目標(target)的指定對資料進行分類，並以樹枝狀方式表現類別之間由輸入變數所造成的區別，可對資料進行層級架構的分類。

由於決策樹的容易解釋、規則化分枝結果，因此利用訓練完成的決策樹架構，亦可對資料進行比較或預測分析。例如半導體製程的應用便可經由決策樹的分類規則提供事故診斷的決策參考與進一步製程設定(Irani et al. 1993 ；Gandner & Bieker 2000；簡禎富等2001)。

決策樹依演算法不同使用不同的分枝規則與判斷方式以選擇分枝變數、分枝方式與層次架構而將資料做一系列的分類。主要的決策樹演算法包括分類與迴歸樹(Classification and Regression Tree)(Breiman et al. 1984)、卡方自動交互檢測(Chi-squared Automatic Interaction detection ; CHAID)(引自 Berry & Linoff 1997)，以及 ID3(Interactive Dichotomiser 3)、C4.5、C4.0 等一系列方法(Quinlan 1993 1996)。決策樹演算法的目標皆是在分枝的時候最大化群類間的「距離」，因「距離」的衡量方式不同亦可視為區別不同決策樹演算法的一向重要因素。

參、以資料挖礦方法進行特徵萃取與描述

本研究提出半導體製程資料之資料挖礦方法包括兩大階段，第一部份是利用自我組織映射網路（Kohonen 1981）進行資料的多變量分群，第二部分則是利用決策樹以分析萃取與表達前述分群結果的特徵，作為製程監控與事故診斷之參考依據。本研究方法之重點為針對半導體龐大且複雜的工程資料而言，透過良率等關聯變數與多變量分群結果以發覺特殊樣型，然後利用決策樹進行後續特徵萃取與分析。本研究有別於傳統的統計分析方法，利用探索式的多變量分群，可以在已知變異因子下進行多變量樣型分析，亦可在未知變異因子狀況下藉由所發掘的樣型與萃取特徵規則以解決問題。本研究根據半導體製程資料特性與資料挖礦的四個循環過程，主要步驟包括：問題定義與架構、資料準備、自我組織映射網路分群、決策樹特徵萃取、結果解釋與討論(如圖 2)。以下先具體說明各研究步驟，第肆章將以個案實例進一步說明。

一、問題定義

對問題的釐清與定義非常重要的，可以預設資料挖礦之方向。資料挖礦工作者(data miner)便可以根據此一目標取得相關所需的資料，以進行後續相關的挖掘工作。對於半導體製造廠來說，其目標在於產品狀況的監控(monitor)或者縮短事故診斷的時間範圍，以提昇產品的良率。然而因為製造程序複雜影響變數眾多，工程師往往無法從龐大資料中，迅速有效地察覺製程異常的原因或因素，更遑論從資料中發現先前隱藏不知的重要訊息(簡禎富等2001)。因此必須瞭解半導體領域相關知識，再根據問題的目的，收集或回溯相關的製程資

料，選擇適當的方法或模式進行挖掘，以找出可能解釋事故發生的原因。

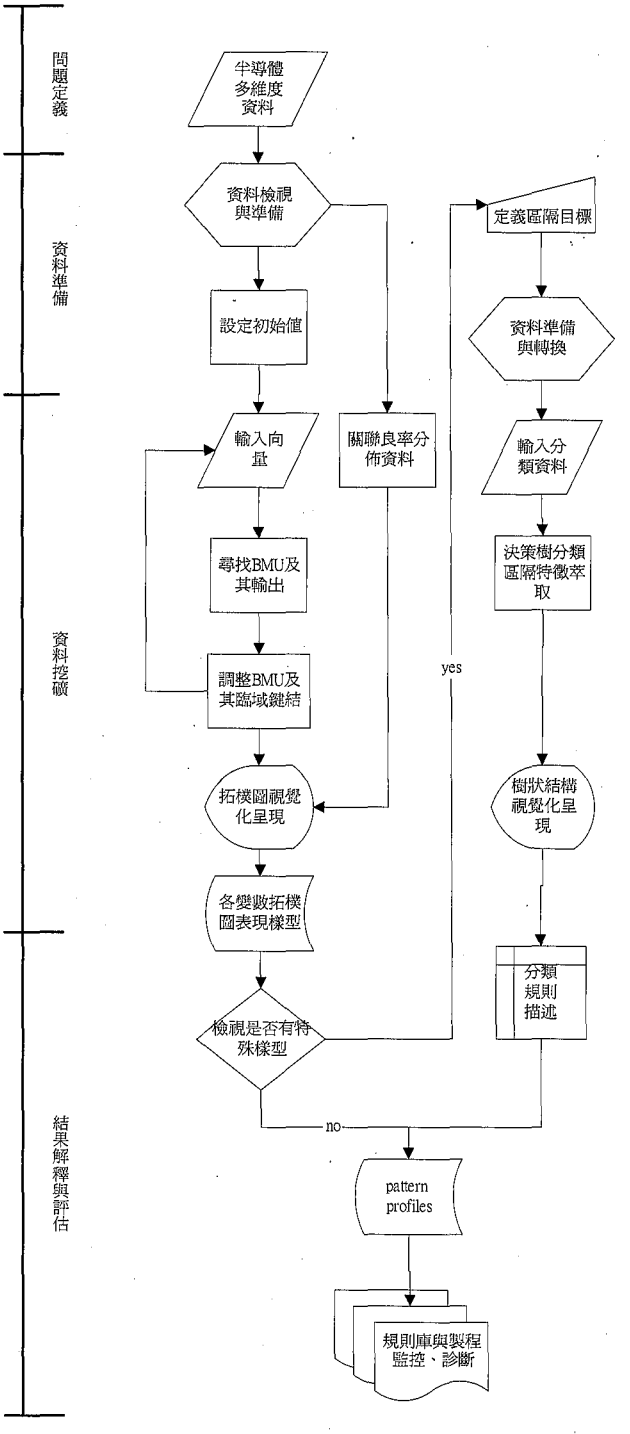

圖 2：半導體資料特徵萃取與描述之資料挖礦步驟

二、資料準備

對於資料挖礦來說，對半導體領域知識及問題有基本認識之後，必須透過與領域專家的訪談、討論，進一步認識資料的型態與特性。透過線上之工程資料分析系統 (Engineering Data Analysis System；EDAS) (Peng and Chien, 2002)，取得要分析的資料並將其潔淨化整理至可分析處理的型態，其中包括資料檢視、遺漏值或空白值的處理、以及不同處理平台間的資料轉換等。但對半導體資料來說，將越大量的資料放入分析，不一定保證就能挖掘到有價值的資訊，越多的資料雖然可以增加發現樣型的機會，但相對地更多無用的(redundant)資料或雜訊也跟著增加，因而對於資料處理的效能與結果的正確程度造成相對的影響。因此在資料取得後必須先作一連串基本的資料準備階段後，再進行後續的分析。主要步驟如下：

㈠資料的基本分佈與檢視：

對於資料的基本數理特性作一全面性的檢視(exploration)，瞭解資料的長相與分佈以便作後續處理。

㈡偏離值(*outlier*)的檢測與處理：

對於散佈在正常分佈外的資料點做出處理。比如直接刪除還是保留其資訊。

㈢空白值與遺漏值的處理：

空白值與遺漏值所代表的意義並不相同，要刪除該筆資料或者以特殊方式補值對於挖礦結果的解釋會有不同的影響。

㈣資料維度轉換：

資料的維度對於挖礦模型的建立有一定的影響程度，一般而言高維度的資料計算複雜，所花費的時間較多，因此維度的降低亦是一門重要的課題，統計多變量方法例如主成分分析等可以降低維度。但是有些資料格式的轉換 (例如加工時間)，反而會增加資料的維度 (如，日期、班次、工序等)。資訊的保存與資料的處理效率間的權衡，需要挖掘者的判斷與決定。

㈤資料格式或分佈轉換：

有些資料挖礦的模型，只能針對特定的資料格式作分析，像是數字或者是文字、加工時間等。因此在考慮資料處理時，資料的格式轉換與其所代表的相對意義都需瞭解。

三、自我組織映射網路的多變量分群

本研究採用 SOM 神經網路來進行多變量集群分析，其目的在於透過兩層順向連結的神經網路，將高維度的圖樣特徵，映射到一維或二維的輸出神經元陣列。當圖樣特徵之間具某種順序關係時，透過鍵結值的學習，在輸出神經元之間亦能保持拓樸空間上的關係。藉由對大量多變量資料的集群分析，利用視覺化結果的呈現觀察資料集群的樣型。並將未放入運算但欲建立關聯的變數建立關聯，檢視是否有特殊樣型產生。透過對特殊樣型的觀

察，定義欲區別的「群別」，以作爲決策樹分類所需的目標。本研究方法不同於傳統的統計分析，只針對某些特定變因檢定其顯著性或差異與否，透過 SOM 神經網路此種探索性的方法，可以同時考慮多變量的因子，甚至察覺出先前未知的訊息，而不需事先侷限住可能變因的範圍。

在初始階段，必須設定輸出節點(node)個數、學習速率、及初始鍵結值的大小。透過軟體，我們只需設定資料來源並定義需要放入分析的變數。接著設定輸出的節點個數、及其輸出的排列方式。由於在最後視覺化的拓樸圖上，越相鄰的節點代表其對應的輸入向量越相似。根據硬體運算及最終集群結果的視覺化考量，本研究中以輸入向量的筆數來決定輸出神經元的多寡。在運算可能的情況下，輸出神經元的個數盡可能與輸入筆數相同。

輸入向量後網路會計算優勝神經元(BMU)的輸出，並依照臨域函數(neighborhood function)調整其臨域神經元的鍵結。透過此學習過程，最後收斂至穩定狀態，並以拓樸圖來呈現輸入向量之間的關係。拓樸圖除了可以展現資料之間的近似關係外，各變數對於特定群集的貢獻程度亦可透過檢視個別變數的分群狀況以顏色加以區分。檢視拓樸圖的集群分佈後，引入欲建立關聯的相關變數，例如良率。探討其各群集對於相關變數的分佈後，將要區隔的群集新增一類別欄位，以作爲後段決策樹分類作業的目標(target)變數。

四、決策樹以萃取與表達各分群之特徵

由於決策樹容易解釋結果的特性，在前一階段建立欲區別的群別作爲目標(target)分類變數後，接著將以決策樹進行分類而以樹枝狀架構呈現分類結果，並轉換爲容易理解的規則。因資料型態的差異對於決策樹演算法的選擇與設定也會不同。

決策樹以樹枝狀架構呈現其分類結果，其中指向同一分群的規則可視爲其樣型表徵(profile)。決策樹各節點分類結果可以透過計算其準確率與可信度，以作爲衡量分類規則的指標。可信度代表此分類節點的純度，以準確率代表此節點相對於原有類別個數被正確區隔的比例。換言之，決策樹分類之目的是找到準確率與可信度高的規則來代表特定分群（如，低良率貨批）之特徵。

五、結果解釋與評估

在分析過程乃至於最後挖掘的結果，不論是數據、視覺化圖形或者規則化敘述，應不斷與領域專家討論，以獲得其經驗與進一步改良的意見。挖掘的結果對於工程師而言是否有提供幫助，整個挖掘的過程是否達到預期效果，皆需透過結果解釋與討論重複循環，釐清樣型特徵所代表的意義與價值，才可使得研究模式與結果更加完備，之後並可進一步將相同屬性的規則類型儲存至規則庫，結合領域專家的經驗與質化說明，以建立製程監控或事故診斷的決策支援機制與知識管理系統。

肆、實證研究

本章根據第參章之研究方法與步驟，以某半導體廠之實際資料進行實證研究。此公司是積體電路研發、製造、測試及銷售之整合元件製造廠(Integrated Device Manufacturer；IDM)，專注於非揮發性記憶體(Non-Volatile Memory) 及系統整合晶片 IC 產品，爲全球非揮發性記憶體主要供應商。

在半導體製造程序中，許多資料會以自動或半自動方式記錄下來。包括產品的基本資料、過站時間與機台紀錄、機台設定參數等。在晶圓完成所有加工步驟後，都會在製程結束前進行晶圓允收測試(Wafer Acceptance Test；WAT)，或稱電子特性測試（E-Test）。測試的目的在於測試半導體元件上的電性特性，針對不同需要，量測不同的電子特性，如電阻、電壓、電流等，所以電性測試的參數往往會超過上百項。測試的每一個參數皆用來監控元件的特性，通常也會與特定一層或多層的製程特性有關連。工程師便可藉由電性測試結果配合領域知識診斷晶圓發生異常之原因。在產品最後出廠前也會實行電性功能針測(Circuit Probe Test；CP）的功能測試，以確保產品的功能性符合客戶要求。由於半導體製造程序複雜、影響變數眾多，工程師往往無法從收集的龐大資料中，迅速有效地察覺可能導致製程異常的原因，或是歸納產品品質不良的特性，甚至是先前根本隱藏不知的重要訊息。

一、半導體製程監控分析實例

㈠問題定義與架構

因此針對某半導體製程監控的問題，本研究以多變量的角度針對半導體晶圓允收測試資料進行群聚分析，將良率分佈的特殊樣型進行特徵萃取，瞭解參數表現的特性(profiles)與良率間的關係，以協助工程師進行製程監控與後續之快速事故診斷。

㈡資料準備

面對收集的多維度資料，首要工作是資料探索與準備。由於要挖掘的是 WAT 資料與良率間具特定分佈的樣型，透過其工程資料分析系統擷取測試時間為 2001 年一月份的資料。資料欄位包含每批晶圓的批號、測試時間及各測試參數的量測記錄。由於量測變數眾多且皆為連續型變數，在與工程師討論之後選擇此項產品的 41 項主要量測參數(key parameter)作進一步分析，共取得 264 筆資料。在進行運算之前的前置處理，我們對每個變數實行以全距為標準的尺度(scale)調整。

㈢群聚分析

瞭解資料特性與分佈後，設定 SOM 網路架構以進行群聚分析。在考慮硬體運算的效率下，設定 SOM 的節點(node)數為 1000，並將資料輸入 SOM 網路。透過向量量化與向量投影，其群聚結果如圖 3，在瞭解其資料點在拓樸圖上的分佈後以顏色來區分群聚，共可分為四群。根據上述分群方式進一步針對這些資料點的良率值進行分析(如圖 4)，其中良率值的相對表現越高，顏色越接近紅色，反之則接近藍色。由良率分佈可以發現在右下角的群聚其良率相較於另外三群是較低的，大都在 0.75 以下。針對這樣特殊的樣型，藉由與領域工程師的討論後，我們將這些資料曾被工程師所下過的診斷紀錄進行整理，並引入拓樸圖中。

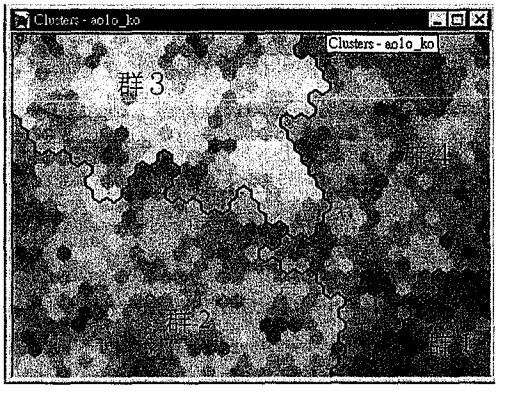

圖 3：WAT 資料群聚現象拓樸圖

由整理結果發現，群聚位在拓樸圖右下角的資料點皆被工程師下過相同的診斷記錄(代號皆爲"#")且對應圖 4 皆影響到良率的表現。由於群聚樣型是根據 WAT 參數的表現而分群，因此群 1 的特徵與低良率現象有某種關連。透過檢視各變數對群聚現象的貢獻程度拓樸圖亦可找出哪些參數對於群聚及良率有其較大的貢獻。但由於 SOM 群聚分析著重以視覺化方式表現群聚，因此接著以決策樹進行特徵萃取與分類規則的描述。

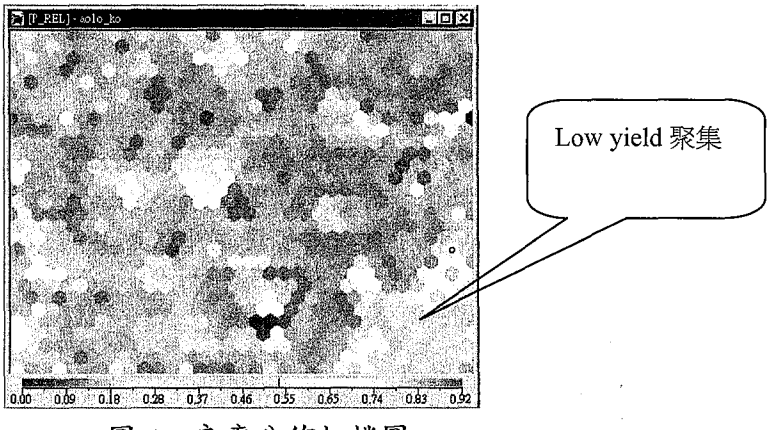

圖 4：良率分佈拓樸圖

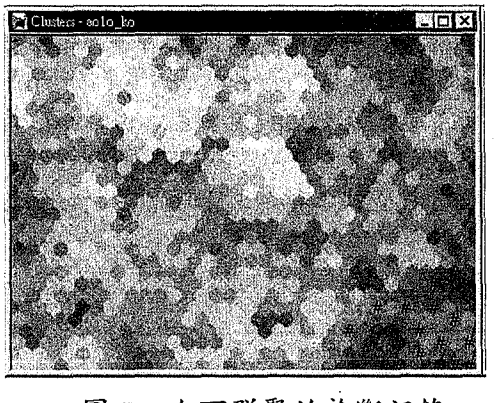

圖 5：右下群聚的診斷記錄

（四）決策樹萃取特徵規則

由前述 SOM 分群中發現位於拓樸圖右下方的群 1 有相同的事故紀錄，因此我們以群 1 爲目標將其類別訂爲 Bad(共 21 筆)，其餘三群的標籤則訂爲 Other(共 243 筆)。利用決策樹進行分類，而 WAT 參數表現的差異將群 1(Bad)與其他(Other)區隔開來，本研究採取卡方自動交互檢測(Chi-squared Automatic Interaction detection ; CHAID)演算法，利用資料挖礦軟體 SAS-EM（SAS 軟體股份有限公司，1999），並將顯著水準設爲 0.05。因此在決策樹的每一個分枝節點，決策樹會選擇一個對於分類目標最有貢獻的變數進行卡方檢定。在決策樹的深度上訂爲 6，一個節點最多可向下分兩枝。

在衡量決策樹分類規則的時候，我們以可信度代表此分類節點的純度，以準確率代表此節點相對於原有類別個數被正確區隔的比例。換言之，我們期望找到準確率與可信度高的規則來代表聚類特徵，決策樹的分類結果如圖 6。由圖 6 可以發現藉由變數 9 與變數 18 便可將 Bad 的與 Other 區隔開來且其正確率可達 90.4%(19/21)。若以分類規則的純度衡量其可信度，當變數 9 < 8.59 且變數 18 > -7.8 時，其規則的可信度可達 95%(分類節點純度爲 95%)。當然規則的可信度越高越好，但由於我們關心的是大部分的群 1 所顯現出的特徵差異，因此其他的規則雖然純度很高，所能代表群 1 特徵的資料點數卻很稀少，故在此只作爲部分參考。

根據上述決策樹的分類規則，代表在群聚拓樸圖中群 1 與其他群之間依照 WAT 量測表現作爲區隔的標準。換句話說，這樣的分類規則代表的是群 1 與其他群聚間差異的特徵，而這樣的特徵對應至良率的分佈可能都會造成低良率的情況。亦即以變數 9 < 8.59 且變數 18 > -7.8 代表群 1 的 WAT 特定表徵，且這種表徵與特定的事故現象是高度相關的。另一方面，透過特定表徵與事故診斷連結的規則庫，可以提供工程師於製程的監控、分析與良率預測的參考。

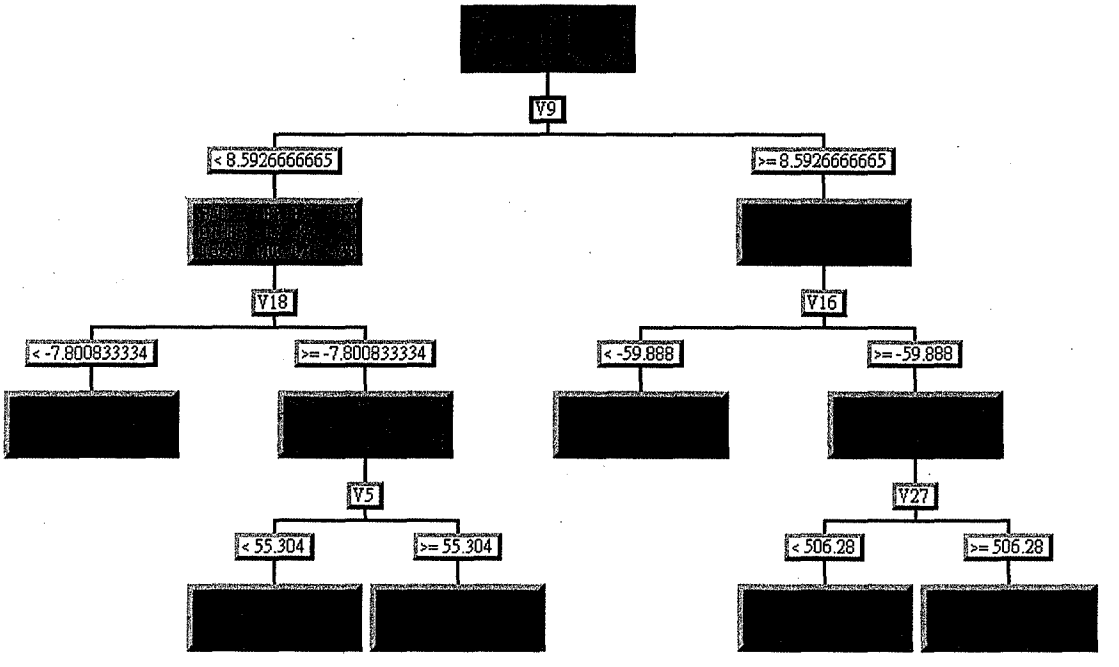

圖 6：決策樹分類結果

㈤結果討論

使用決策樹分類時最常碰到的便是資料變數間具共線性或高度相關的情況。本研究在分枝節點檢視 logworth(其中，logworth = -log(p-value))值可以瞭解每一個變數在此分枝點對於分類目標變數的貢獻程度，logworth 值越大表示其貢獻程度越高。在初次分割的階段其各變數的 logworth 值如表 2：

表 2：各變數貢獻程度(因篇幅限制故只列出前幾項)

貢獻排序	變數	logworth	貢獻排序	變數	logworth
1	v9	41.74	11	v15	21.033
2	v22	40.829	12	v23	20.934
3	v11	38.142	13	v16	13.243
4	v19	37.604	14	v36	11.135
5	v18	37.57	15	v26	10.506
6	v14	34.69	16	v34	10.274
7	v4	34.658	17	v12	9.438
8	v10	30.958	18	v32	7.479
9	v20	25.119	19	v33	7.331
10	v35	23.214	...	...	...

表 3：貢獻程度前九名分類結果

	description	bad(/21)	other(/243)
rule1	**V22** >= -7.994	18	3
rule2	**V11** < -8.2658	17	3
rule3	**V19** < -1.8284 & **v35** >= 7.866	20	1
rule4	**V18** >= -7.764	16	2
rule5	**V14** < 8.686 & **v10** >= 3.457	18	0
rule6	**V4** < 8.4653 & **v10** >= 3.4571	18	0
rule7	**V10** >= 3.58 & **v4** < 8.48	17	0
rule8	**V20** >= 58.716 & **v4** < 8.48	18	2
rule9	**V35** >= 8.058 & **v4** < 8.476	19	1

檢視各變數的貢獻程度，可以發現在第一次分割的時候雖然變數 9 仍較變數 22 貢獻大，但兩者貢獻是相當接近的。有鑑於此，除了第一次以變數 9 作切割萃取特徵規則以代表群 1profile 的情況外，我們選取貢獻程度另外前九名(v22~v35)的變數，各自進行分割以討論其分類結果，各項選取分類的結果規則如表 3。

由表 2 與表 3 中可以發現，所選出的前幾名變數，皆可以將原本的資料做出區隔，但是區隔的規則正確率及可信度皆不相同。舉 rule 1 爲例，其分類的正確率可達 85.71%(18/21)，但是其可信度卻也下降至 85.71%。以 rule3 來說，第一次以變數 19 作爲分枝時，其正確率可高達 95.23%(20/21)，但規則可信度卻只有 71.43%(20/28)，在引入變數 35 繼續分枝的情況下，才能在正確率不變的情況下將可信度提高至 95.23%。由於在第一個節點進行分枝時變數 19 對於目標分類的顯著程度不如變數 9，因此在預設以貢獻程度高的變數進行分枝的條件下，會以變數 9 進行分枝萃取規則，而得到"當變數 9 < 8.59 且變數 18 > -7.8 時，屬於某項會造成低良率的事故原因"這樣的 profile。

另一方面，在與工程師討論之後發現，群 1 這種 WAT profile 的特殊樣型是由於蝕刻過程造成殘餘物質，影響部分元件的漏電電壓而導致良率下降，針對各變數對於群聚現象的貢獻，檢視各變數的拓樸圖也能發現與量測漏電電壓有關的參數表現與良率間的相關性(如圖 7)，其中群 1 在這些變數的表現上相對於其他群皆是迥異的。在決策樹所萃取出的參數表現特徵當中，亦可發現變數 9 與變數 18 量測的皆是元件的漏電電壓。由此可以驗證在群聚現象中所發現的樣型，藉由特徵的萃取與描述，對於建立 WAT 參數表現與低良率之間的特徵規則是有幫助的。

二、半導體製程事故診斷分析實例

㈠問題定義與架構

由於半導體製造所收集的資料往往相當龐大，而工程師要迅速發掘問題並將其解決也

變得相當困難。現行此公司的作法，大都採用統計檢定的方式以比較產品在測試資料表現或者機台的差異來提供事故診斷的參考。由於當前方法只能由單變量提供資訊，因此對於事故診斷的幫助有限，且需大量藉助工程師的領域知識。因此本研究以多變量方法，針對半導體資料中低良率的部分進行特徵萃取，提供工程師綜合性的資訊以提高事故診斷時效，並做爲決策支援雛形系統的參考。由圖 8 的良率趨勢可以發現製程發生良率下降的情況，因此工程師必須開始對於良率下降的原因進行分析與瞭解，以快速發掘問題原因修正問題，以提升良率。

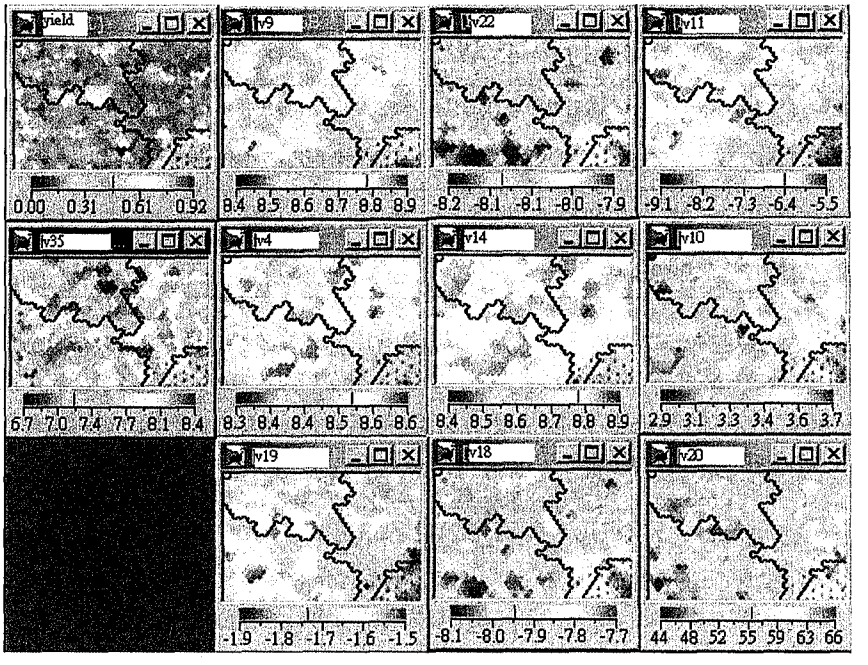

圖 7：各群聚在量測漏電電壓相關變數的相對表現(限於篇幅只列舉其中幾項)

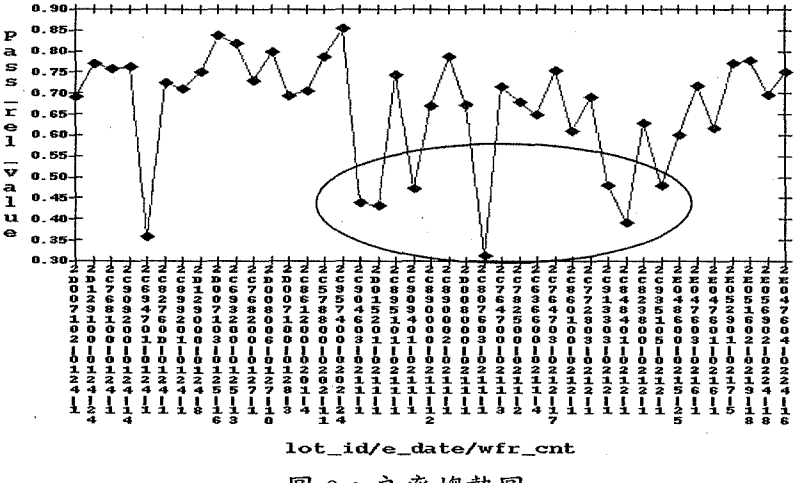

圖 8：良率趨勢圖

㈡資料準備

在發現良率下降的情形下，我們選定時間範圍爲某年 1 月 1 號到 3 月 31 號的資料。在這段期間之內的資料包含在趨勢圖上所看到的低良率的貨批。透過工程資料分析系統選定這些批的電性功能針測參數測試資料、電性測試資料及過站記錄將資料擷取下來。因此進行挖礦的資料裡包含每一批晶圓的編號、電性功能針測測試時間、電性功能針測所有測試參數的值、通過的站別及所用的機台。

本研究先採用 CP 的測試資料及過站記錄組成分析資料的主體。在 CP 測試資料方面屬於連續型的資料，由於許多參數的測試記錄有發生遺漏值或空白值的狀況，在與工程師討論之後選擇其中的 18 個主要參數(bin)作爲代表。另一方面，過站記錄資料亦相當龐大，故我們只選取其中的四個程序。過站記錄的機台資料屬於類別資料，且以字串的形式記錄下來，因此轉換的方式爲增加資料維度，使資料變成零一變數(binary variable)，以滿足群聚分析軟體的計算需求(Pyle 1999)，整理後的資料包含 76 個批(lots)及 42 個變數。

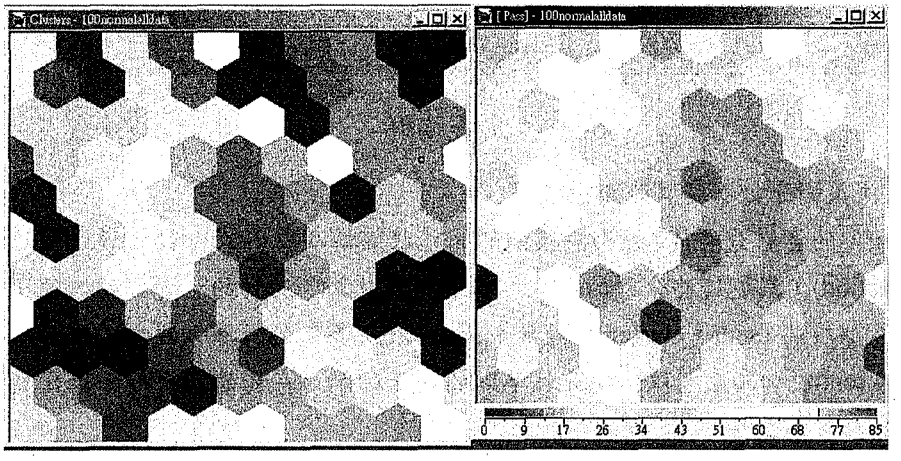

圖 9：綜合表現(左)及良率拓樸圖(右)

㈢群聚區隔

我們將資料轉換成 Excel 的工作表格式，在將其輸入到 SOM 的模型。因爲資料有 76 筆，因此在參數的設定方面選擇 100 個節點(node)。經由向量量化與投影，可以在拓樸圖上看出不同的群聚。透過顏色的區別，在分群後可以得到綜合表現如下圖，由綜合表現成圖結果可以發現，群聚高達 52 群。

進而針對分群結果就單一變數良率的表現來探討，其中良率表現越高，顏色越偏向紅色。在良率表現的拓樸圖亦可發現左上角的部分及左下角皆有低良率的分佈。因爲本研究目標爲異常良率的分析，因此以良率表現拓樸圖爲主，以選擇(select)方式定義群聚，以作爲接下來決策樹演算法的分類目標(target)。其中將良率拓樸圖左上角同樣表現低良率的群聚選擇定義爲第一族群(c1)，左下角則定義爲第二族群(c2)，其他的部分都定義爲其他族群

(oc)。由於綜合表現才能眞正代表不同批的相似程度，因此在定義族群的同時爲避免切割到同一族群，需以良率圖爲主、綜合圖爲輔，作爲選擇之依據。將區分的三族群定義如圖10所示。

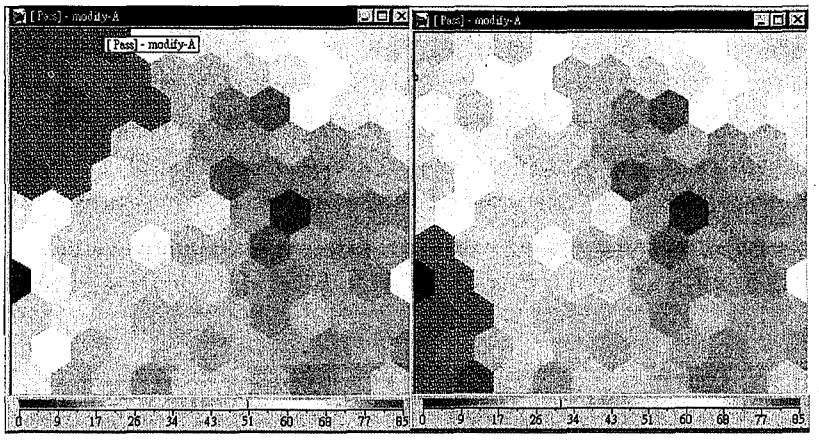

圖 10：定義族群

（四） 決策樹分析以萃取分類特徵

將表現類似的資料分群後，經由決策樹的樹枝狀架構及規則(rule)的表達方式進行資料分類(classification)，描述這些分群的特徵，以達到特徵萃取的目的。資料的處理上，我們將所有的變數都設爲 input，cluster 則設爲 target。在決策樹的計算方式方面，選擇的是計算節點的亂度(entropy)爲主，以作爲分枝準則的評估標準。在樹的深度方面設定爲 8，每一節點最大的分枝數爲 2。

將表現類似的批歸在一群，並藉由決策樹歸納出屬於此群的特徵其變數的表現會有哪幾種情況，以多種的規則代表群聚特徵嘗試發現其中有趣的樣型，是資料挖礦階段最主要的目的。整個決策樹的分類結果如圖 11，其中分類爲第一分群的規則有三條，第二分群的規則有三條，對應其他群的分類規則有四條。

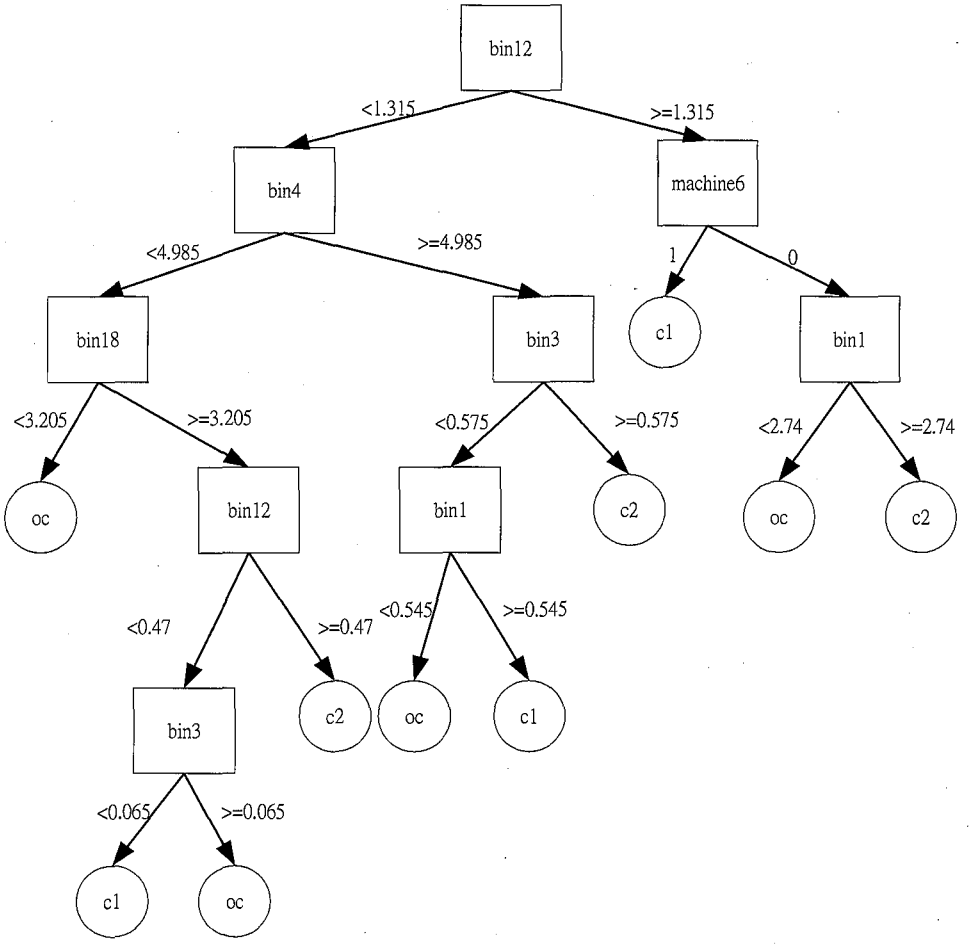

圖 11：決策樹分類結果

針對異常良率的族群一及族群二，決策樹所產生的分類規則如表 4。表 4 中所記錄的是族群一與族群二以規則的表達方式所描述的族群特徵。在每一條規則下，分別有三個指標代表此規則的重要程度。比例（proportion）代表的是此條規則對於族群中的資料能解釋的相對比例，亦即前一部份所提到之準確率；頻率（frequency）代表的是族群中滿足此規則的絕對資料筆數，與前一部份所提之可信度意義相同；錯誤率（error rate）代表決策樹各分枝分類的錯誤率，由於在這裡選擇將樹長到底，因此各底層枝葉最終節點的錯誤率皆爲零。由於我們關心的在於低良率的第一分群和第二分群，其描述規則整理如表 4。

㈤結果與討論

儘管在規則中提供電性功能針測參數的表現對於事故診斷並無直接的幫助，但是卻可以協助工程師以其領域知識找出關連性；另一方面，發現第一分群的分類規則一出現機台 6 作爲分枝準則，相較於其他的規則有其特別之處。根據規則所提供的資訊，工程師實際作事故診斷亦發現眞的是機台 6 造成良率下降，並影響到 8 個批的良率，因此可以直接縮小工程師的診斷範圍，直接由機台作事故診斷。檢驗 SOM 的拓樸圖標記這幾批的位置，

亦如預期出現在第一族群中，換句話說，在決策樹中的第一條規則診斷出這 8 個批。其他的規則雖無直接指出可能的問題機台，卻也提供綜合的參數表現，讓工程師瞭解參數的變異情況對於良率的影響。

表 4：第一族群及第二族群分類規則

➤ **Cluster 1(total 12 lots):Rule1: bin12>=1.315, machine6=1**
Proportion: 83.3% (10/12)
Frequency: 10
Error rate: 0 *
Rule2: bin12<1.315, bin16>=4.985, bin3<0.575, bin1>=0.545
Proportion: 8.33% (1/12)
frequency: 1
Error rate: 0
Rule3: bin12<0.47, bin16<4.985, bin18>=3.205, bin3<0.065
Proportion: 8.33% (1/12)
Frequency: 1
Error rate: 0

➤ **Cluster 2(total 7 lots):**
Rule1: bin12<1.315, bin16>=4.985, bin3>=0.575
Proportion: 57.14% (4/7)
Frequency: 4
Error rate: 0 *
Rule2: 0.47<= bin12<1.315, bin16<4.985, bin18>=3.205
Proportion: 28.57% (2/7)
frequency: 2
Error rate: 0
Rule3: bin12<0.47, bin16<4.985, bin18>=3.205, bin3<0.065
Proportion: 14.29% (1/7)
Frequency: 1
Error rate: 0

*由於我們設定讓決策樹長到底，因此最終節點裡的資料皆是屬於單一類別，無混合的情況產生，所以 error rate 都是零。

伍、結論與未來研究

目前半導體廠在導入資料挖礦時，多偏重於資料倉儲之建置，對於資料準備、資料處理、而至資料挖礦的完整過程仍相當缺乏相關經驗與成功案例。本研究發展並用實例說明完整之分析步驟，其中包括利用自我組織映射成圖網路演算法將多變量的資料進行拓撲分群，再以決策樹分類規則的方式萃取特徵並描述低良率特殊樣型的分群特徵，提供給工程

師作爲製程監控的依據與未來事故診斷的參考。本研究以某半導體廠製程資料爲實證，包含製程監控與事故診斷兩大部分：第一部份針對半導體資料製程晶圓允收參數的多維度資料，發現隱藏於資料中的樣型與良率間的關連性，再透過參數表現特徵提供給工程師監控製程變化的決策依據，以改善製程提昇良率。第二部分則針對半導體製程製造測試的多維度資料，透過綜合資訊的比較縮小診斷範圍，提供給工程師作爲事故診斷的決策依據，以快速排除事故提昇良率。雖然每次事故問題發生的類型並非都固定一樣，然而仍可參考本研究所提出分析步驟。

能不能挖掘到埋藏於資料礦山中的寶藏，放入分析的資料對於目標的貢獻度是否足夠亦是重要的因素之一。但對半導體資料來說，資料取得不是問題，但將越大量的資料放入分析，不能保證越能挖掘到有價值的資訊，針對半導體工程資料龐大且混雜之特性，傳統統計分析方法有其限制，本研究所提出資料挖礦方法結合資訊科技與工程資料分析，經過自我組織映射成圖網路演算法進行群聚分析找出與良率分佈相關的樣型，以決策樹分類規則特徵萃取與描述表達群聚特徵的探索性方法，可以在已知變異因子下進行多變量樣型分析，亦可在未知變異因子狀況下藉由所發掘的樣型來協助工程師進行產品的監控及事故診斷的參考，實例驗證發現對半導體生產之良率提昇有具體貢獻。未來研究方向除了更可加入其他相關的資料例如製程記錄資料、量測資料等，進行多變量對多變量關連性分析後藉由特徵差異的分析與比較，協助工程師進行事故診斷與製程最佳化，加速判別產品的良率水準及故障類別。另一方面亦可結合晶圓圖的故障樣型做爲決策支援系統雛形架構的參考。

致謝

本研究承蒙旺宏電子股份有限公司委託計畫與與國科會（NSC89-2213-E-007-118）之經費補助，並特別感謝徐紹鐘、張文哲與黃佳琪等之協助。

參考文獻

1.林昇甫、洪成安，1996，類神經網路入門與圖樣辨識，台北：全華科技。
2.薛如珊，2001，使用自組織映射網路進行資料群集和資訊樣型採擷的資料探勘法，台灣大學工業工程學研究所碩士論文。
3.簡禎富、林鼎浩、徐紹鐘、彭誠湧，2001，「建構半導體晶圓允收測試資料挖礦架構及其實證研究」，工業工程學刊，第十八卷. 第四期，37-48 頁。
4.蘇木春、張孝德，1999，機器學習：類神經網路、模糊系統以及基因演算法則，台北：全華科技。
5.SAS 軟體股份有限公司，1999，Enterprise Miner V2.0 資料挖礦軟體。
6.Berry, M. and Linoff, G. Data Mining Techniques for Marketing, Sales and Customer Support, John Wiley &Sons, New York, NY, 1997.
7.Brachman, R. J., T. Khabaza, W. Kloesgen, G. Piatetsky-Shapiro and E. Simoudis "Mining business database," Communication of ACM (39:11) 1996, pp: 42- 48.
8.Breiman, L. Friedman, J.H., Olshen, R.A., and Stone, C.J. Classification and Regression Trees, International Thomson Publishing, 1984.
9.Bursteinas, B., Long, J.A. "Tools with Artificial Intelligence," Proc. IEEE 2000, pp: 274 -280.
10.Cai, Y. "The application of the artificial neural network in the grading of beer

quality," Proc. WCNN94 (1) 1994, pp: 516-520.
11.Deboeck, G. and T. Kohonen, Eds Visual Exploration in Finance with Self-Organizing Maps. Springer-Verlag, London, 1998.
12.Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. "The KDD Process for Extracting Useful Knowledge form Volumes of data," Communication of ACM, (39:11) 1996, pp: 27-34.
13.Fayyad, U. "Data Mining and Knowledge Discovery in Database: Implication for Scientific Databases", Scientific and Statistical Database Management 1997, pp: 2-11.
14.Gandner, M. and Bieker, J. "Data Mining Solves Tough Semiconductor Manufacturing Problem", Proc. KDD2000, 2000.
15.Gurney, K., An introduction to neural networks, UCL Press, London, 1998.
16.Irani, K.B., Cheng, J., Fayyad, U.M., Qian, Z. "Applying Machine Learning to Semiconductor Manufacturing," IEEE Expert, Volume: 8 Issue: 1 , Feb. 1993 pp: 41 -47.
17.Kasslin, M. Kangas, J. and Simula O. "Process state monitoring using self-organizing maps" in Artificial Neural Network, Aleksander, I. and Taylor, J. Eds., (2) 1992, pp: 1532-1534. North-Holland, Amsterdam.
18.Kessler, W. Ende, D. Kessler, R.W. and Rosenstiel, W. "Identification of car body steel by an optical on line system and Kohonen's self-organizing map," Proc. SPIE (1588) 1993, pp: 64-75
19.Kittler, R., and Wang, W. 1999,「資料分析漸露頭角」,中文半導體技術雜誌,79-85頁。
20.Kiviluoto K. "Predicting bankruptcies with the self-organizing map," Neurocompt. (21) 1998, pp: 191-201.
21.Kleissner, C. "Data Minig for the Enterprise," IEEE Proc. 31st Annual Hawaii International Conference on System Sciences (7) 1998, pp: 295-304.
22.Kohonen, T. Self-Organizing Map, Springer-Verlag, Berlin, 1995.
23.Kohonen, T., Oja, E. Simula, O., Visa, A. and Kangas, J. "Engineering applications of the self-organizing map," Proc. IEEE (84:10) 1996, pp: 1358-1384.
24.Lampinen, J. and Oja, E. "Distortion tolerant pattern recognition based on self-organizing feature extraction," IEEE Transaction on Neural Networks (6) 1995, pp: 539-547.
25.Milne, R., Drummond, M. and Renoux P. "Predicting paper making defect on-line using data mining," Knowledge-Based Systems (11) 1998, pp: 331-338.
26.Peng, C. and Chien, C. "Data value development to enhance competitive advantage: a retrospective study of EDA system for semiconductor fabrication", International Journal of Services Technology and Management, (forthcoming).
27.Pyle, D. Data Preparation for Data Mining, Morgan Kaufrnann, San Francisco, CA, 1999.
28.Quinlan, J.R. "Introduction to decision trees", Machine Learning (1) 1986, pp: 81-106.
29.Quinlan, J.R. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco, CA, 1993.