國立交通大學

電控工程研究所

碩士論文

基於互信息的變數分群和變數選取

Variable Clustering and Variable Selection Based on Mutual Information

研 究 生：王景文

指導教授：周志成　博士

中華民國一百年六月

基於互信息的變數分群和變數選取

Variable Clustering and Variable Selection Based on Mutual Information

研 究 生：王景文　　Student：Chin-Wen Wang

指導教授：周志成　　Advisor：Chi-Cheng Jou

國 立 交 通 大 學

電控工程研究所

碩 士 論 文

A Thesis

Submitted to Department of Electrical and Control Engineering

College of Electrical Engineering

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master

in

Electrical and Control Engineering

June　2011

Hsinchu, Taiwan, Republic of China

中華民國一百年六月

基於互信息的變數分群和變數選取

學生：王景文　　指導教授：周志成

國立交通大學電控工程研究所

摘　　要

資訊爆炸時代各領域處理的資料量不斷倍增，變數選取——如何從龐大的資料中挑選出最有價值的變數——已成爲一個至關重要的問題。變數選取的目的有二：藉由挑選代表變數達到簡化系統的效果，以及針對給定的目標變數挑選有效的解釋變數以建立高準確度的預測模型。變數分群是變數選取的一種實現過程，其功用在將相似度高的變數聚在一群，再從每一群中挑出具代表性的變數。傳統變數分群和變數選取的方法受到變數間必須呈線性關係、資料型態必須是連續及呈多變量常態分布這些條件的限制，本論文提出一種基於互信息理論的變數分群、變數選取方法，使用互信息來衡量變數的相似度可以克服傳統方法的限制。變數分群有兩種模式，一是以互信息當作變數間的“距離”使用 k-中心分群，二是先推論出互信息網路，在其上使用譜分群。變數選取則依兩種不同的目的分別以互信息和中心度來挑選每一群的代表變數。最後以晶圓製程的資料來驗證我們的方法，結果顯示 k-中心分群所選出來的變數在兩種變數選取的目的上均有較佳的表現。

Variable Clustering and Variable Selection Based on Mutual Information

Student：Chin-Wen Wang Advisor：Dr. Chi-Cheng Jou

Department of Electrical and Control Engineering

National Chiao Tung University

ABSTRACT

In the era of information explosion, data in each field is increasing rapidly. Variable selection -- how to select valuable variables from numerous dataset -- has been an important issue. There are two purposes of variable selection: to simplify the system by selecting representative variables and to build up high – accuracy predictive models for a given target variable by selecting effective explanatory variables. Variable clustering is a realization of variable selection. Cluster variables with high similarity into groups then choose representatives for each group. There are three constraints of traditional methods of variable clustering and variable selection, namely the relation between variables must be linear, the dataset must be continuous and multivariate normal distributed. We propose a new approach based on mutual information theory. By taking mutual information as the similarity among variables, we can circumvent these constraints. There're two approaches for variable clustering, one is to treat mutual information as the distances between variables, which can be an input for k-medoids clustering, the other is

to inference the mutual information network first then apply spectral clustering on it. Finally we verify our method with a dataset of wafer process, the result shows that variable selection based on k-medoids clustering has better performance on both two purposes.

誌謝

兩年的研究生涯，轉眼間就過去了，其中的酸甜苦辣滋味，相信唯有親自走過才能體會。在此由衷地感謝恩師　周志成博士的指導與教誨，使學生得以窺探資料分析領域的奧妙。每次與您談話完總是會有新的啓發，不論是研究或待人處事上均受益良多，雖千言萬語無法表達對恩師的感激。

口試承蒙　楊谷洋教授與蔡雅蓉博士於百忙之中親臨指導，對論文細心地斧正，並提供寶貴的意見，使論文更爲嚴謹完善，僅致以深摯謝意。

感謝學長仲翔、慶陽及同學偉勳、恕緣、承綱及學弟智勇、駿程，謝謝你們的陪伴與鼓勵。

最後特別感謝我的家人，感謝你們完全的支持與關愛。願與你們分享這份成果與喜悅。

目錄

圖目錄

表目錄

第一章　　緒論

1.1　研究動機

隨著科技的進步，我們進入一個資訊爆炸的時代，各領域處理的資料量不斷倍增，如何從龐大的資料中挖掘出最有價值的資訊已成爲一個重要的問題。考慮一個廣義的系統，系統中存在許多變數，眾多的變數固然提供了豐富的資訊，同時也造成處理上的負擔，爲了掌握系統的特性而同時監控這些龐大數量的變數是沒有效率且不切實際的。以半導體晶圓製程爲例，一片晶圓經過數百或數千個製程步驟才得以完成，每一步驟又有數個機台參數或測量項目，可能是某個步驟中某金屬層的厚度，或是某程序中受控制的溫度或壓力等。若將每個項目都視爲一個系統變數，那麼每一片晶圓經過完整製程後我們就會得到一包含有數百甚至數千個變數的資料。這麼大量的資料不論是儲存、分析或運算都會是相當龐大的負擔，因此，在分析資料和建立系統模型前我們首要考慮的就是如何減少變數量。如何從大量的系統變數中挑選出少數、帶有重要資訊的變數，以簡化系統並降低計算量。這正是本論文所要探討的問題，稱爲變數選取 (variable selection)。

變數選取的目的有二：

簡化系統：爲了掌控系統狀態，我們可能會設置一些感測器 (sensor) 或利用估計的方法來獲得某些系統參數 (變數)。當變數數量龐大時，我們不可能監控到每一個變數，此時變數選取可以決定哪些變數是比較重要的，我們只監控那些少數的重要變數。通常會先將變數分群 (variable clustering)，相似度 (similarity) 高的變

數爲一群，再從每一群中選出代表變數 (representative variable)，理想情況下可保留住大部分原本多數變數所攜帶的資訊。

解釋目標變數：目標變數通常就是我們關心的系統輸出，如晶圓製程中的良率 (yield)。爲了預測未知的目標變數，我們會選出一些系統變數作爲解釋變數 (explanatory variable) 以建立預測模型。一般而言解釋變數越多，預測就越準確，但過多的解釋變數有可能造成它們之間的共線性* (colinearity) 或訊號雜訊比 (signal-to-noise ratio, SNR) 降低，同時基於計算時間等因素考量下，我們希望選取較少的解釋變數但預測準確度又不至於太差。

進行變數選取前，有必要考慮以下面向：

資料型態：資料是連續的 (continuous)，例如數值，或是離散的 (discrete)，例如類別 (category)。表 1.1 水果資料中的重量變數爲連續而顏色變數爲離散。

	重量(g)	顏色
取樣1	205	橘
取樣2	3370	綠
取樣3	350	黃

表 1.1　水果資料

變數間的關係：變數間的關係是否爲線性，或是非線性。

資料分布：資料是否呈多變量常態分布 (multivariate normal distribution)。

* 共線性專指線性相關，但也有可能爲非線性相關，如之後會提到的“冗餘 (redundancy)”概念。

以上所提是在選擇分析方法前所必須要考慮的。傳統變數分群、變數選取的方法很多，常用的經典方法如因素分析 (factor analysis)、階層式分群 (hierarchical clustering)、逐步選取法 (stepwise selection)、MRMR (maximum relevance minimum redundancy) 等。因素分析可用於變數分群和變數選取，因爲使用相關係數 (correlation coefficient)，故僅適用於連續資料，同時變數間必須大致上呈線性關係，此外還要求資料必須是多變量常態分布[1]。階層式分群可用於變數分群，雖然可以置換相似度的定義方式以克服離散資料和非線性關係，但它只能做變數分群，無法從群中挑選代表變數。逐步選取法 (stepwise selection) 是針對解釋目標變數的變數選取方法，內嵌於線性迴歸 (linear regression) 中，故也要求變數間是線性關係。MRMR (maximum relevance minimum redundancy) 亦是針對解釋目標變數的變數選取方法，宣稱可以克服非線性但效果有待驗證。以上提到的方法在以下章節中均會介紹，先以表 1.2 列出其特性。

方法	變數分群	變數選取	限制
因素分析	V	簡化系統	連續、線性、常態分布
分類樹	V		
逐步選取法		解釋目標變數	連續、線性
MRMR		解釋目標變數	

表 1.2　傳統變數分群、變數選取方法的特性

1.2　研究方法

本論文提出一套基於互信息 (mutual information) 的方法，可以達到變數分群

並同時滿足兩種不同目的的變數選取。我們宣稱這套方法可以克服非線性、離散資料和非常態分布的限制，其關鍵就在於互信息，兩變數 X 、Y 之間的互信息定義如下：

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{1.1}$$

$p(x)$ 和 $p(y)$ 分別爲 X 和 Y 的機率密度函數 (probability density function)。互信息是一種比相關係數更廣義的相關性測量，爲一個恆正的值，越大代表兩變數相關性越強，等於零時兩變數互相獨立 (independent)。互信息的本質是一個變數已知後，另一變數的熵 (entropy) 的減少量，因此使用互信息可以掌握變數間的線性與非線性關係。(1.1)式是互信息的定義也是計算公式，其中機率密度函數的估計並不受限於連續資料，在離散資料下也可計算；同時也不在乎資料是否呈常態分布。我們利用以上互信息的這些特性來克服傳統變數分群、變數選取方法所遇到的問題。有關熵與互信息的詳細介紹留待第三章。

我們整體的構想如圖 1.1

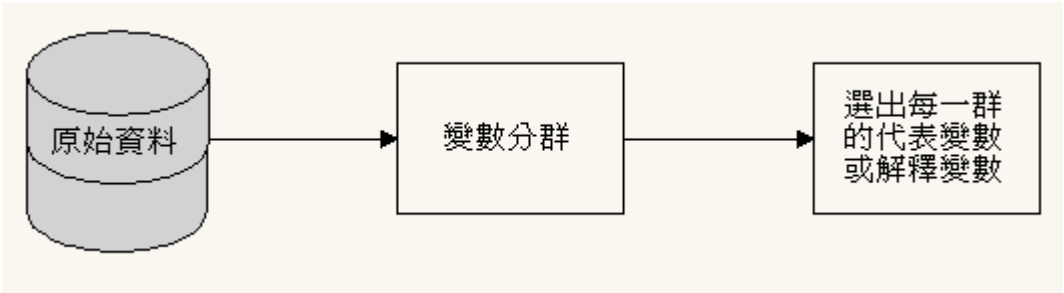


圖 1.1　構想流程圖

首先由原始資料估計出兩兩變數間的互信息，以互信息當作變數的相似度並根據

相似度將變數分群。分群的作法有二：一是直接用 k-中心 (k-medoids) 分群；二是先推論出變數的互信息網路 (mutual information network)，接著用譜分群 (spectral clustering)，以譜分群中圖分割 (graph partitioning) 的觀點來分割互信息網路以達到變數分群的目的。兩種方式的差別在於互信息網路提供一個可視覺化的變數關係網路圖，我們可以從圖上判讀出哪些變數之間具有真正的直接關係*。

變數選取接在變數分群後進行，若目的是簡化系統，則我們從每一群中選出中心度 (centrality)† 最大的變數作爲代表變數；若目的是解釋目標變數，則從每一群中選與目標變數互信息最大的變數作爲解釋變數。值得一提的是，不論選變數的目的是哪一個，我們提出的方法差別只在最後選變數的方式，而這步驟佔總計算量的比例相當小，此特性亦算是我們的優點之一。

1.3 論文結構

本論文共分爲六章。第二章簡介逐步選取法、因素分析和階層式分群這三種傳統的變數分群、變數選取方法。第三章介紹互信息的意義、互信息的估計方法以及互信息網路的推論。第四章介紹 k-中心、譜分群這兩種分群方法以及中心度的概念並說明如何利用中心度和互信息來選取變數。第五章是實驗，先對實驗的資料作簡介，接著討論變數分群和變數選取的結果，對於以簡化系統爲目的的變數選取以我們自定義的一個指標來評分；對於以解釋目標變數爲目的的變數選取

* 直接關係是相對於虛假相關而說的，虛假相關在第三章中會介紹。

† 度量一個節點在網路中的重要程度，第四章中會介紹。

分別各以五種分類和迴歸的方法做驗證，並和逐步選取法、MRMR 作比較。最後一小節爲離散資料的變數分群、變數選取結果。第六章結論。

第二章　　文獻回顧

2.1　逐步選取法

迴歸分析中，逐步選取法針對某個目標變數來挑選解釋變數，可分爲三種：

順向選取 (forward selection)：迴歸模型內的變數從空集合開始，在每一次的選擇步驟中，選出對目標變數貢獻最大的解釋變數進入迴歸模型，並對尚未進入迴歸模型的解釋變數進行比較，以決定下一次的選取可能被納入的解釋變數。選取的標準依各解釋變數 F 檢定 (F-test) 中的 p 值 (p-value)來決定，選擇對應 p 值最小的解釋變數且該 p 值必須小於某個門檻值 (p-enter)。重複此步驟直到所有變數的 p 值均大於門檻值爲止。

反向選取 (backward selection)：首先將所有的解釋變數都納入迴歸模型中，然後逐一剔除對目標變數貢獻最小的解釋變數，並對仍保留在迴歸模型中的各解釋變數進行比較，以決定某個解釋變數是否要被剃除或保留。剔除對應 p 值最大的解釋變數且該 p 值必須大於某個門檻值 (p-remove)。重複此步驟直到所有變數的 p 值均小於門檻值爲止。

綜合法：綜合順向選取和反向選取，從空集合開始，每一步驟中先採順向選取，接著對被選取的變數執行反向選取，意即從外部納入有用的變數進入迴歸模型，接著剃除模型中沒用的變數，反覆交叉執行順向選取和反向選取，納入與剔除分別根據兩個門檻值 (p-enter 和 p-remove)。

決策者可依時間、成本和預測準確度來調整門檻值以決定選取變數的數目。

2.2 因素分析

因素分析是一種分析變數變異 (variability) 的統計方法，做法是找出數目較少的潛在變數 (latent variable) 或稱因素 (factor) ，使得原有變數近似於潛在變數的線性組合。利用原有變數和潛在變數間的關係可將原有變數分群。萃取因素的方法很多，最著名也最常用的就是主成分分析 (principal component analysis, PCA)，主成分分析的目的是找出互相正交 (orthogonal) 的軸*使得原有變數在這些軸上有最大的變異量 (variance)。令 $\mathbf{R}^{p\times p}$ 爲 p 個變數之間的相關係數矩陣，其第 (i, j) 元爲第 i 個變數和第 j 個變數的相關係數。計算 $\mathbf{R}$ 特徵值 (eigenvalue) 與特徵向量 (eigenvector)，因爲 $\mathbf{R}$ 是一個實對稱矩陣，可正交對角化，故

$$\mathbf{R} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1} = \mathbf{V}\mathbf{D}\mathbf{V}^{T} \tag{2.1}$$

其中 $\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_p \end{bmatrix}$，$\mathbf{D} = diag\left(\lambda_1,\ldots,\lambda_p\right)$，$\lambda_1 \geq \ldots \geq \lambda_p \geq 0$，$\lambda_i$ 和 $\mathbf{v}_i$ 爲一組互相對應的特徵值和特徵向量，$\mathbf{v}_i$ 之間互爲正交。PCA 找出的軸及變數在軸上的變異量分別爲 $v_1,...,v_p$ 與 $\lambda_1,\ldots,\lambda_p$。因素負荷 (factor loadings) 定義爲原有變數和因素的相關係數。例子如表 2.1 (絕對值大於 0.7 以粗體標示)，可以看出變數 3、變數 5 和因素 2 高相關，其餘變數均和因素 1 高相關，沒有變數與因素 3 高相關。

* 此處的"軸"即爲因素

	因素1	因素2	因素3
變數1	**0.87**	0.15	0.48
變數2	**0.97**	-0.15	-0.17
變數3	0.58	**-0.79**	-0.06
變數4	**-0.91**	-0.42	0
變數5	0.11	**-0.97**	0.13
變數6	**0.96**	0.07	-0.28
變數7	**0.99**	0.12	0.04

表 2.1　因素負荷的例子

太多變數與同一個因素高相關或者同一個因素可以同時代表多個變數，這對於變數分群來說是不利的，通常我們希望變數能分得散一點，而不是都擠在同一個因素上 (同一群)。解決方法就是將因素 (軸) 做旋轉，旋轉的方式有很多，我們介紹最常用的最大變異 (varimax) 旋轉[2]。最大變異旋轉的概念和 PCA 非常類似——旋轉軸使得因素負荷的平方在旋轉後的軸上變異量最大

$$\arg\max \sum_{j,l}\left(f^2_{j,l} - (\bar{f}^2_l)\right) \tag{2.2}$$

$f_{j,l}$ 爲第 j 個變數在第 l 個因素上的因素負荷，$\bar{f}^2_l$ 爲第 l 個因素的因素負荷平方的平均。表 2.2 爲使用最大變異旋轉後的因素負荷，可看出變數較旋轉前分得散，因素 3 也得到一個代表變數。此時變數大致上可分爲四群：{ 變數 2、變數 6、 變數 7 }、{ 變數 4 }、{ 變數 3、變數 5 }、{ 變數 1 }。可挑選每群中與因素負荷最大者爲代表變數如變數 6、變數 4、變數 5 和變數 1。因素分析使用相關係數做計算，故受到變數之間關係必須爲線性的限制。

	因素1	因素2	因素3
變數1	0.52	-0.07	**0.85**
變數2	**0.89**	-0.34	0.29
變數3	0.37	**-0.88**	0.11
變數4	**-0.83**	-0.21	-0.51
變數5	-0.1	**-0.98**	-0.02
變數6	**0.96**	-0.11	0.23
變數7	**0.84**	-0.1	0.53

表 2.2　最大變異旋轉後的因素負荷

2.3　階層式分群

階層式分群依演算法可分爲聚合 (agglomerative) 和分裂 (divisive) 兩種，所謂聚合是指初始時每個變數都各自爲一群，接著將距離*最近的兩個變數合爲一群，距離最近的兩群合爲一群，一次次的聚合使得群數越來越少，最後全部的變數爲一群。兩群之間距離的定義方式有

單連接 (single linkage)

$$d_{A,B} = \min_{i \in A,\, j \in B} d_{i,j} \tag{2.3}$$

全連接 (complete linkage)

$$d_{A,B} = \max_{i \in A,\, j \in B} d_{i,j} \tag{2.4}$$

* 一般對取樣分群可使用歐幾里得距離 (Euclidean distance)、馬哈拉諾畢斯距離 (Mahalanobis distance)、城市區塊距離 (city block distance) 等，但對於變數間的距離，較常使用的是相關係數的平方。

平均連接 (average linkage)

$$d_{A,B} = \sum_{i \in A} \sum_{j \in B} d_{i,j} / n \tag{2.5}$$

華德法 (Ward's method)

$$d_{A,B} = \sum_{i \in A \cup B} (d_{i,u})^2 \tag{2.6}$$

$d_{A,B}$ 爲 A 群與 B 群的距離，i, j 分別爲 A 群內與 B 群內的變數，n 是加總距離的總數，u 爲 A 群與 B 群聚合後的群中心。單連接、全連接和平均連接分別以兩群內變數之間的最短距離、最長距離和平均距離做爲兩群間的距離。華德法則是以兩群合併後的群中心到兩群內所有變數的距離平方做爲兩群間的距離，也就是試圖尋找聚合後變異最小的兩群。

分裂和聚合是相互對應的，分裂先從全部的變數爲一群開始，逐漸分裂成多群直到所有的變數都各自爲一群，此法較不常用故不多做介紹。圖 2.1 是使用華德法對 30 個變數分群的例子，橫軸是變數編號，縱軸是邊數間的距離，定義成 1 減去相關係數的絕對值 (注意刻度已經過對數處理)。我們很難從圖上判定變數要分成幾群以及如何挑選每一群的代表變數。

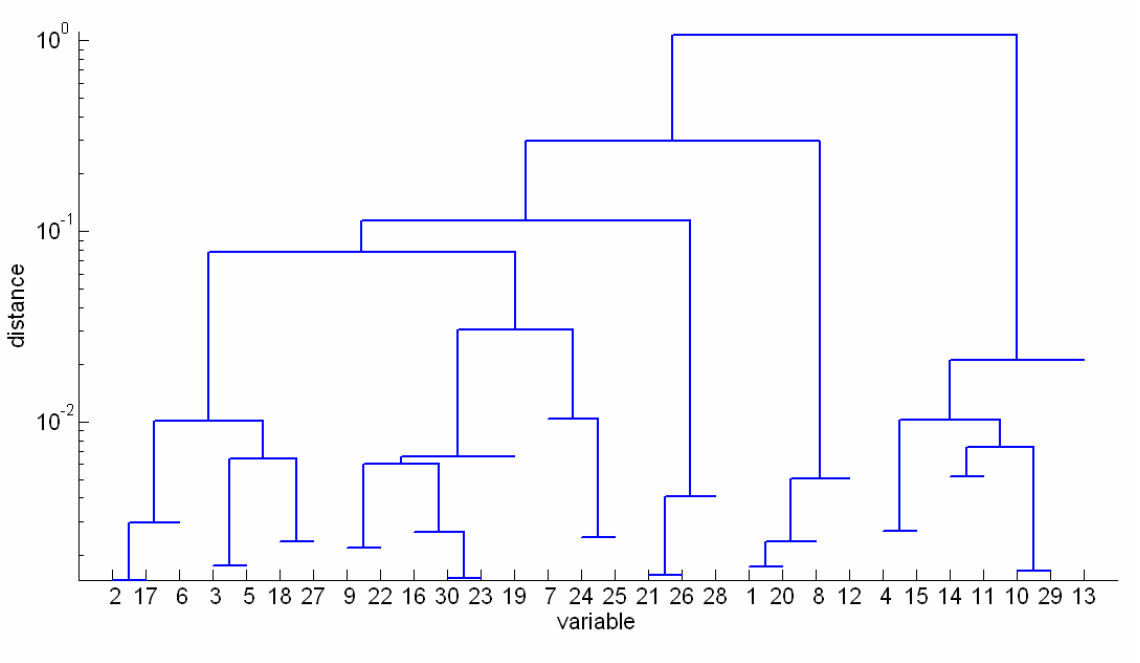


圖 2.1　使用華德法做 30 個變數分群的例子

第三章　　互信息和互信息網路

3.1　互信息與熵

互信息是由熵衍生而來，因此在介紹互信息之前有必要先介紹熵。熵用來測量一個隨機變數 (random variable) 的不確定性 (uncertainty)，或稱混亂度。對於一個離散隨機變數 X ，其熵 $H(X)$ 的定義如下

$$H(X) = -\sum_{x \in X} p(x) \log p(x) dx \tag{3.1}$$

$p(x)$ 爲 X 的離散機率密度函數。對於一個連續隨機變數 X ，其熵 $h(X)$ 的定義如下：

$$h(X) = -\int_{x \in X} f(x) \log f(x) dx \tag{3.2}$$

$f(x)$ 爲 X 的連續機率密度函數。聯合熵 (joint entropy) $H(X,Y)$ 和條件熵 (conditional entropy) $H(Y \mid X)$

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y) \tag{3.3}$$

$$H(X \mid Y) = H(X,Y) - H(Y) \tag{3.4}$$

$p(x,y)$ 爲 X 和 Y 的聯合機率密度函數 (joint probability density function)。聯合熵 $H(X,Y)$ 用來測量 X 和 Y 整體的不確定性，而條件熵 $H(X \mid Y)$ 測量在 Y 爲已知的情況下， X 的不確定性。

互信息用來測量兩個隨機變數 X 和 Y 之間的相關性 (dependence)，定義如下：

$$\begin{aligned} I(X;Y) &= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \\ &= H(X) - H(X \mid Y) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{3.5}$$

$$I(X;Y) = I(Y;X) \tag{3.6}$$

互信息 $I(X;Y)$ 可解讀成在 Y 已知的情況下， X 不確定性的減少量 ((3.5)式第二個等式)；換句話說，在 Y 爲已知的情況下，若 X 的不確定性較原本的 (Y 未知) 降低很多，則 X 和 Y 高度相關， $I(X;Y)$ 的值大。(3.6)式表示互信息具有交換律。以上聯合熵、條件熵和互信息只列出隨機變數爲離散時的式子，連續情況可依此類推，將總和 ($\sum$) 替換成積分 ($\int$) 即可。

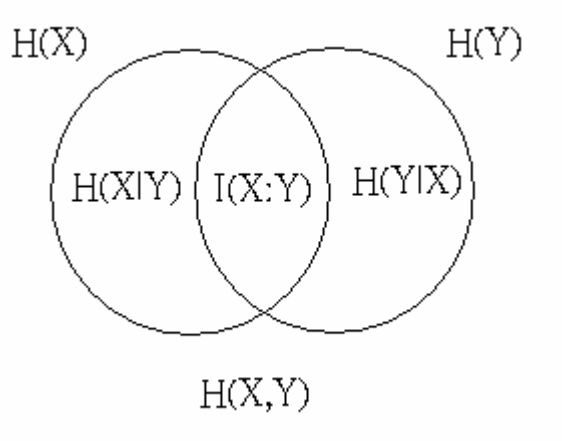


圖 3.1 熵與互信息之關係圖

3.2 互信息的估計

由 $I(X;Y)$ 的定義(3.5)可知，估計 $I(X;Y)$ 等同於估計 $p(x)$ 、 $p(y)$ 和 $p(x,y)$ 。機率密度函數的估計主要可分爲兩種方法：離散化 (discretization) 法[3]和核密度估計法 (kernel density estimation)[4]。考慮 $(x_1, x_2, ..., x_n)$ n 個從某個機率密度函數產生的取樣 (sample)，假設彼此之間獨立。離散化法顧名思義就是將取樣做離散化處理 (binning)，把取樣分布的範圍切成若干個相等大小的區間 (bin)，接著計算落

在每個區間內的取樣數目，將區間內的取樣數目除以取樣總數和區間寬度即爲機率密度函數在這個區間內的值。核密度估計法是一種無母數 (non-parametric) 的方法，結果寫成

$$\begin{aligned}\hat{p}_k(x) &= \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \\ &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)\end{aligned} \tag{3.7}$$

$K(\cdot)$ 稱作核，爲一對稱函數，常用的核函數如高斯 (Gaussian)、葉帕涅奇尼科夫 (Epanechnikov)、三角 (triangular) 等。$h > 0$ 是一個決定核寬度的參數，稱作頻寬 (bandwidth)。

舉一個簡單的例子：已知 6 個一維的取樣 $(-2.1,\ -1.3,\ -0.4,\ 1.9,\ 5.1,\ 6.2)$，分別使用離散化法和核密度估計法去估產生這組取樣的機率密度函數。對於離散化法，我們將 $[-4\ 8]$ (自定，但必須包含所有取樣) 分成 6 個區間，每個區間寬度爲 2。觀察這些取樣落在區間內的情形，從左到右的區間內分別有 1,2,1,0,1,1 個取樣，將這些數除以取樣總數 6 和區間寬度 2 得到 $\frac{1}{12}, \frac{1}{6}, \frac{1}{12}, 0, \frac{1}{12}, \frac{1}{12}$，此即爲取樣落在各個區間內的機率估計值，如圖 3.2 左。對於核密度估計法，核函數我們選擇用高斯，令 $h = 1.8$，結果如圖 3.2 右，虛線爲 6 個個別的高斯核函數，加總後得到機率密度函數的估計結果 (實線)。

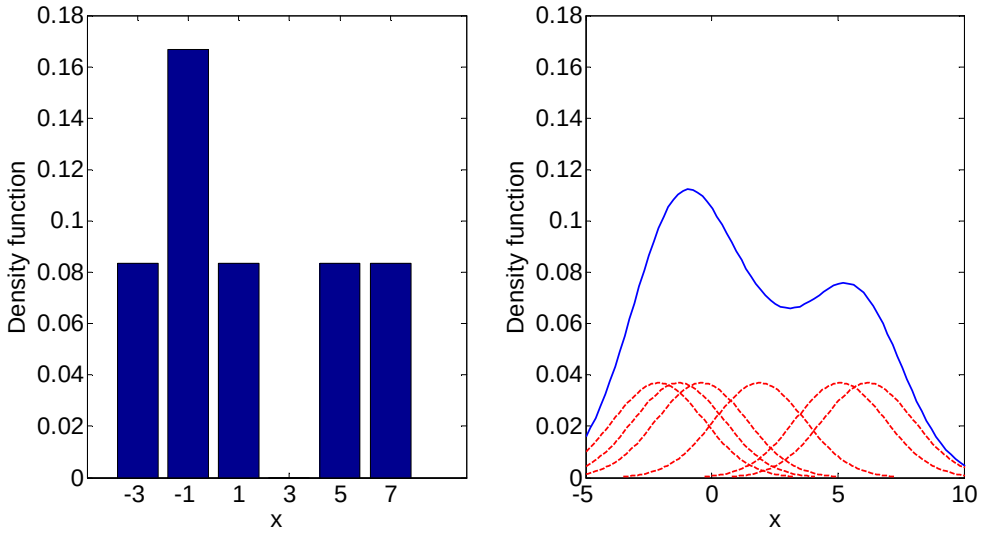


圖 3.2　離散化法 (左) 和核密度估計法 (右) 之結果比較

3.3　互信息網路

互信息網路原先的用途是分析基因與基因之間的關係，用於疾病檢測或藥物製作[5]。推論出的網路以有權重的無向圖 (weighted, undirected graph) 呈現，圖上的節點 (vertex) 代表基因，邊 (edge) 的權重 (weight) 代表邊連接的兩個基因之間的關係，通常是一個非負的值，越大表示這兩個基因越相關。有許多不同的推論演算法，但目標都是一致的，舉例敘述如下：有甲乙丙三個基因，甲基因和乙基因有關係 (可能是某種生物上的互動機制)，乙基因和丙基因有關係。甲基因和丙基因就某種程度上來說是獨立的 (條件獨立)，他們的關係是建立在乙基因上。這是真實的情況，但是我們得到的資料 (互信息) 卻會顯示甲乙丙三基因兩兩之間都有關係，我們可能因此而被誤導。互信息網路就是在解決這樣的問題，如圖 3.3 (a)，推論演算法會試著推論出一個沒有虛線邊 (或虛線邊被賦予很低的權重) 的網路。

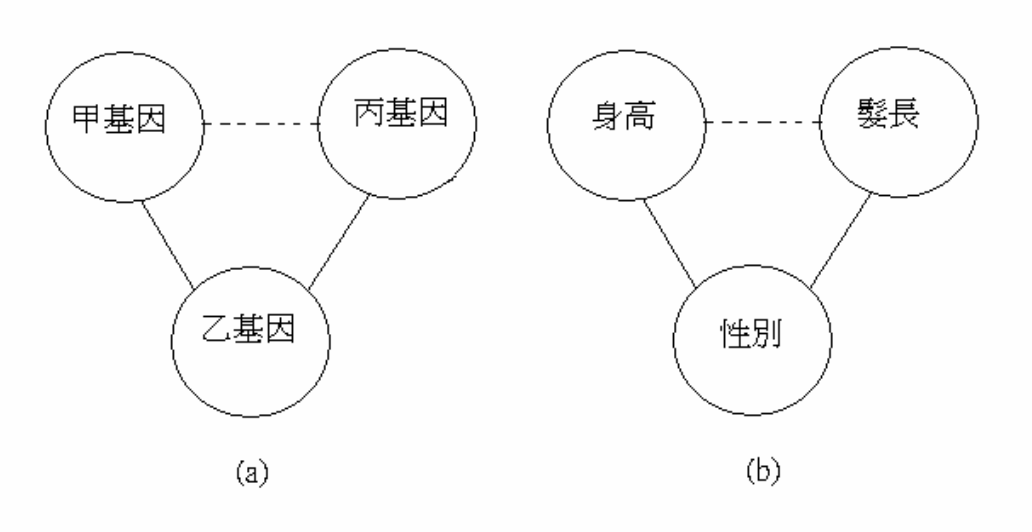


圖 3.3　虛假相關

現在把問題轉向變數，變數間也存在著類似以上敘述的情形，稱爲變數間的虛假相關 (spurious correlation)，例如：有一包資料內含性別、身高和頭髮長度三個變數，分析這包資料我們可能會得到這樣的推論：身高和頭髮長度呈負相關，如圖 3.4 (a)。

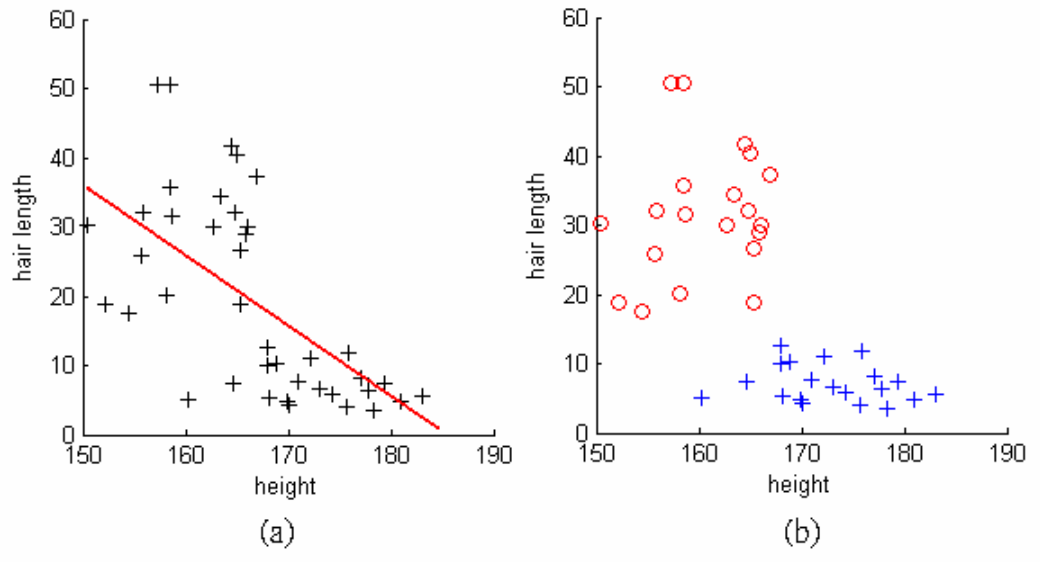


圖 3.4　虛假相關的例子

事實真的是如此嗎？身高越高的人頭髮越短？答案應該是否定的，因爲我們沒有考慮到性別這個變數，身高和髮長的關係是建立在性別上的 (一般而言，男性身高較高，頭髮較短，女性則剛好相反)。若已知性別，那麼身高與髮長應該是無關的，如圖 3.4(b) (圈圈代表女性，十字代表男性)。我們想藉由互信息網路來消除變數間的虛假相關，如圖 3.3 (b)。

互信息網路的推論演算法都以兩兩變數的互信息當作輸入，以下介紹兩種：ARACNE[6]與 MRNET[7]。ARACNE 的概念是根據信息理論中的一個不等式——資料處理不等式 (data processing inequality)：若變數 X_1 經由 X_2 而影響 X_3 (如上述例子)，則

$$I(X_1; X_3) \le \min\left(I(X_1; X_2),\, I(X_2; X_3)\right) \tag{3.8}$$

ARACNE 首先建立一個完全圖 (complete graph)*，邊上的權重爲邊連接的兩變數的互信息 $I(X_i; X_j)$。首先將權重小於某個門檻值 I_0 的邊刪除，接著檢視所有的三三一組 (triplet) 的變數組合，對於每一組，權重最小的那個邊將被標上刪除標記。檢視完所有的變數組合後，刪除被標上刪除標記的邊。若網路架構是樹 (tree) 且變數只有兩兩之間的關係，則 ARACNE 可推論出真正的網路。

MRNET 的概念是根據一個變數選取的方法名爲 MRMR (maximum relevance minimum redundancy)，以下先介紹 MRMR。令 Y 爲目標變數，V 爲所有輸入變數 X_i 的集合。MRMR 的目標是選擇 V 中的變數做爲 Y 的解釋變數，選的先後順序根據以下準則

---

* 任意兩節點均有連接的圖。

$$X_j = \arg \max_{X_j \in V \setminus S} s_j$$
$$s_j = u_j - r_j \tag{3.9}$$

$$u_j = I(X_j; Y) \tag{3.10}$$

$$r_j = \frac{1}{|S|} \sum_{X_k \in S} I(X_j; X_k) \tag{3.11}$$

S 爲已被選取的變數集合。u_j 爲 X_j 與 Y 的互信息，稱作相關項 (relevance term)。r_j 爲 X_j 與 S 中所有變數的互信息平均，稱作冗餘項 (redundancy term)。s_j 爲第 j 個變數的得分 (相關減掉冗餘)，從得分最高的變數開始選起，在每一次的選擇過程中都會重新計算 s_j。MRMR 權衡相關和冗餘，將要被選的變數和目標 Y 越相關越好，和已經被選擇的變數越不相關越好。

MRNET 在網路的推論過程中會重複地執行 MRMR 來完成邊的權重賦予，當要決定連接任意兩變數 X_i 與 X_j 的邊的權重 w_{ij} 時，令 X_i 爲目標 Y ，V 爲 $\{X_1, X_2, \ldots, X_p\} \setminus \{X_i\}$，根據(3.9)式 MRMR 會回傳 s_j (選到 X_j 時的分數)；同理，令 X_j 爲目標 Y ，V 爲 $\{X_1, X_2, \ldots, X_p\} \setminus \{X_j\}$ 也會得到 s_i ，此時

$$w_{ij} = \max(s_i, s_j) \tag{3.12}$$

重複以上直到所有邊的權重都被計算出，最後將權重小於某個門檻值 w_0 的邊刪除即完成 MRNET 互信息網路推論。MRNET 的精神是：若變數 X_i 與 X_j 相連，則 X_i 是 X_j 的良好預測變數 (well-ranked predictor) ($s_i > w_0$) 或 X_j 是 X_i 的良好預測變數 ($s_j > w_0$)。

第四章　　分群和變數選取

分群 (類) 的方法非常多，著名的如 k-平均 (k-means)、k-NN (k-nearest neighbor)、貝氏分類器 (Bayes classifier)等。考慮非監督式學習 (unsupervised learning) 及分群後挑選代表變數這兩個條件，我們選擇 k-中心和譜分群。

4.1　k-中心和影值

k-中心是和 k-平均極爲類似的分群演算法，差別在於 k-平均用群內成員的平均做爲群中心而 k-中心的群中心是群內的某一成員。事實上，我們選擇 k-中心做分群的最大原因在於我們使用變數間的互信息當作分群的依據，變數沒有計算平均所需的“座標”。k-中心演算法如下：

k-中心演算法

輸入：兩兩變數的互信息、群數 k

1. 隨機選擇 k 個變數做爲群中心
2. 群中心以外的變數被分到和群中心互信息最大的那群
3. 產生新的群中心，與群內變數互信息總合最大者當之
4. 重複 2.和 3.直到群中心不再變動

輸出：分群結果

k-中心和 k-平均有相同的問題，由於初始群中心是隨機產生的，有可能會收斂

到不好的分群結果，因此我們重複執行 k-中心多次然後根據影值選擇結果最好的那一次做爲最終的分群結果。影值是一個評估分群結果好壞的指標，每一個變數 i 都可以計算一個影值 $s(i)$ 定義如下

$$s(i)=\begin{cases}1-b(i)/a(i) & \text{if } a(i)>b(i)\\ 0 & \text{if } a(i)=b(i)\\ a(i)/b(i)-1 & \text{if } a(i)<b(i)\end{cases} \tag{4.1}$$

其中 $a(i)$ 是和自己群內變數的平均相似度 (互信息)， $d(i,C)$ 是和第 C 群內變數的平均相似度， C 是除了自己這一群外的其他群， $b(i)=\max_{C} d(i,C)$ 。 $s(i)$ 是一個介於 1 和-1 之間的值，越接近 1 表示分群結果越好，意即與自己群內的變數相似度 ($a(i)$) 大，與他群內的變數相似度 ($b(i)$) 小。

4.2 譜分群

譜分群[8-9]有不同的觀點可以切入，它既是是一種圖分割的方法，也是廣義的 k-平均，或是一種將座標轉換到另一個空間進而做分群的技巧。這裡我們引用它圖分割的觀點，延續第二章介紹的互信息網路，利用譜分群對互信息網路做分割，達到變數分群的目的。令 $G=(V,E)$ 爲代表互信息網路的圖， $V=\{v_1,...,v_n\}$ 爲圖上 n 個節點 (變數)的集合， **W** 爲接鄰矩陣 (adjacency matrix)，其第 (i, j) 元 w_{ij} 爲連接節點 i 和節點 j 的邊的權重，定義

$$W(A_1,A_2)=\sum_{i\in A_1,\, j\in A_2} w_{ij} \tag{4.2}$$

$$cut(A_1,\ldots,A_k)=\frac{1}{2}\sum_{i=1}^{k} W(A_i,\bar{A}_i) \tag{4.3}$$

$A_1,\ldots,A_k$ 爲分割後的節點子集合，$\bar{A}_i$ 是 A_i 的補集，$A_i \cap A_j = \phi$ 且 $A_1 \cup \ldots \cup A_k = V$ 。基本上，圖分割時所切斷的邊應該要越少越好，或者說切斷的邊的權重越小越好。因此我們希望 *cut* 越小越好，其次還有一個條件必須滿足，那就是對於每一群的節點數目不能太少，否則我們只要不斷分離其中一個邊最少或邊權重最小的節點即可讓 *cut* 最小，但這顯然不是我們所要的結果，加上第二個條件後定義

$$RatioCut(A_1,\ldots,A_k) = \sum_{i=1}^{k} \frac{cut(A_i,\bar{A}_i)}{|A_i|} \tag{4.4}$$

$|A_i|$ 爲 A_i 分割子圖內的節點個數。爲簡單起見，我們從分成兩群 $k=2$ 開始，圖分割可以轉成這個最佳化的問題

$$\min_{A \subset V} RatioCut(A,\bar{A}) \tag{4.5}$$

定義圖拉普拉斯 (graph Laplacian) 矩陣 $\mathbf{L}$

$$\mathbf{L} = \mathbf{D} - \mathbf{W} \tag{4.6}$$

$\mathbf{D} = diag\{d_1,...,d_n\}$ ， $d_i = \sum_{j=1}^{n} w_{ij}$ ， $\mathbf{L}$ 有一個性質是

$$\mathbf{v}^T\mathbf{L}\mathbf{v} = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(v_i - v_j)^2 \tag{4.7}$$

對於任意 $\mathbf{v} = [v_1 \quad \cdots \quad v_n]^T$ 皆成立。令 $\mathbf{f} = [f_1 \quad \cdots \quad f_n]^T$

$$f_i = \begin{cases} \sqrt{|\bar{A}|/|A|} & \text{if } v_i \in A \\ -\sqrt{|\bar{A}|/|A|} & \text{if } v_i \in \bar{A} \end{cases} \tag{4.8}$$

把 $\mathbf{f}$ 帶入(4.7)式

$$
\begin{aligned}
\mathbf{f}^T\mathbf{L}\mathbf{f} &= \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 \\
&= \frac{1}{2}\sum_{i\in A,\, j\in \bar{A}} w_{ij}\left(\sqrt{\frac{|\bar{A}|}{|A|}} + \sqrt{\frac{|A|}{|\bar{A}|}}\right)^2 + \frac{1}{2}\sum_{i\in \bar{A},\, j\in A} w_{ij}\left(-\sqrt{\frac{|\bar{A}|}{|A|}} - \sqrt{\frac{|A|}{|\bar{A}|}}\right)^2 \\
&= cut(A,\bar{A})\left(\frac{|\bar{A}|}{|A|} + \frac{|A|}{|\bar{A}|} + 2\right) \\
&= cut(A,\bar{A})\left(\frac{|A|+|\bar{A}|}{|A|} + \frac{|A|+|\bar{A}|}{|\bar{A}|}\right) \\
&= |V|\, RatioCut(A,\bar{A})
\end{aligned}
\tag{4.9}
$$

由(4.8)式$\mathbf{f}$的定義可知$\mathbf{f}\perp\mathbf{1} = [1 \quad \cdots \quad 1]^T$及$|\mathbf{f}|^2 = n$，結合(4.9)的結果，原本(4.5)式的問題可轉換成

$$
\begin{aligned}
&\min_{A\subset V} \mathbf{f}^T\mathbf{L}\mathbf{f} \\
&\text{subject to } \mathbf{f}\perp\mathbf{1},\ |\mathbf{f}| = \sqrt{n}
\end{aligned}
\tag{4.10}
$$

因爲$\mathbf{f}$的定義使得問題依舊困難，試著放寬$\mathbf{f}$的限制

$$
\begin{aligned}
&\min_{\mathbf{f}\in R^n} \mathbf{f}^T\mathbf{L}\mathbf{f} \\
&\text{subject to } \mathbf{f}\perp\mathbf{1},\ |\mathbf{f}| = \sqrt{n}
\end{aligned}
\tag{4.11}
$$

如此便可利用瑞利商 (Rayleigh quotient)* 得到上述問題的解$\mathbf{f}$其實就是對應$\mathbf{L}$第二小特徵值的特徵向量。對應最小特徵值的特徵向量爲$\mathbf{1}$，不滿足$\mathbf{f}\perp\mathbf{1}$，因此取對應第二小的。解出$\mathbf{f}$後如何對原本圖上的節點分群呢？最簡單的方式就是看$f_i$的正負號，對應$f_i \geq 0$的$v_i$爲一群，對應$f_i < 0$的$v_i$爲一群。或者將$f_i$視爲變數$v_i$的座

---

* 對於任意埃爾米特 (Hermitian) 矩陣$\mathbf{A}$，$\min_{\mathbf{x}\neq 0}\frac{\mathbf{x}^*\mathbf{A}\mathbf{x}}{\mathbf{x}^*\mathbf{x}} = \lambda_{\min}$，$\lambda_{\min}$爲$\mathbf{A}$的最小特徵值，*表示共軛轉置(conjugate transpose)

標，在這一維的空間中用 k-平均分群。我們省略 $k = n$ (分成 n 群) 的推導，直接給出演算法如下：

譜分群演算法

輸入：權重接鄰矩陣 $\mathbf{W}$ 、群數 k

1. 計算 $\mathbf{L} = \mathbf{D} - \mathbf{W}$

2. 計算對應 $\mathbf{L}$ 前 $k+1$ 小特徵值的特徵向量 $\mathbf{u}_1,...,\mathbf{u}_{k+1}$

3. 令 $\mathbf{U} = [\mathbf{u}_2 \quad \cdots \quad \mathbf{u}_{k+1}]$，將 $\mathbf{U}$ 的第 i 列當成第 i 個節點的座標送入 k-平均分群

輸出：分群結果

譜分群其實是一種將資料轉換到另一個空間後再進行分群的方法，這麼做也有降低維度的效果，因爲我們分的群數 k 通常不會太大，在這不大的 k 維空間中 k-平均的計算量一定會比原本的空間中的小。譜分群輸入的接鄰矩陣爲兩兩節點的相似度，爲了能讓分群更順利，通常要考慮區域的聯通性，也就是原本在圖上就有一區一區的傾向。互信息網路即具有這樣的特性，因爲消除了虛假相關 。

4.3 中心度和變數選取

截至目前爲止我們介紹了互信息、互信息網路以及 k-中心、譜分群兩種分群方法，變數分群後剩下的工作就是選出群的代表。首先介紹中心度[10]這個概念，中心度用來描述一個節點在網路 (圖) 中的重要程度，或是衡量一個節點在網路中是否佔有一個核心的位置。舉例來說，捷運網路中轉運站的中心度應該就會比其

他站的中心度高。常見的中心度包括度中心度 (degree centrality)、近中心度 (closeness centrality)、介中心度 (betweenness centrality) 和特徵向量中心度 (eigenvector centrality)。

在一個邊無權重的網路中，度中心度就是一個節點的鄰居數，一個具有高度中心度的節點可能是重要的節點，因爲它可以影響很多節點或是被很多節點所影響。近中心度爲一個節點與其他節點的平均距離，一個具有高近中心度的節點也可能是重要的節點，因爲它距離其他節點都很近，可以快速地影響很多節點或者被很多節點所影響。連通網路中，任意兩節點至少都有一條相互連接的最短途徑。介中心度就是去量化一個節點出現在所有最短途徑上的次數，一個具有高介中心度的節點有可能是重要的，因爲節點間訊息的傳播經常都要靠它作爲橋樑。最後是特徵向量中心度，衡量的準則是：一個節點若連接到高特徵向量中心度的節點，那麼它本身也會得到較高的特徵向量中心度。

依分群的方法 (k-中心、譜分群) 和變數選取的目的 (簡化系統、解釋目標變數) 我們會得到四種組合。若變數選取的目的爲簡化系統，則不論是 k-中心或譜分群我們都選出度中心度最大的作爲代表變數。在邊有權重的網路中，變數 i 的度中心度的定義如下式

$$C_D(i) = \sum_j w_{ij} \qquad (4.12)]$$

其中 w_{ij} 爲接鄰矩陣 $\mathbf{W}$ 的 (i, j) 元。若變數選取的目的爲解釋目標變數則從每一群中選出與目標變數最相關的 (互信息最大) 變數作爲解釋變數。

4.4 研究方法總結

以圖 4.1 總結我們的方法流程。首先由原始資料估計出兩兩變數間的互信息，以互信息當作變數的相似度並根據相似度將變數分群。分群的方式有二：一是直接用 k-中心分群；二是先推論出變數的互信息網路，接著用譜分群，以譜分群中圖分割的觀點來分割互信息網路以達到變數分群的目的。兩種方式的差別在於互信息網路提供一個可視覺化的變數關係網路圖，我們可以從圖上判讀出哪些變數之間具有真正的直接關係。

變數選取接在變數分群後進行，若目的是簡化系統，則我們從每一群中選出中心度最大的變數作爲代表變數；若目的是解釋目標變數，則從每一群中選與目標變數互信息最大的變數作爲解釋變數。

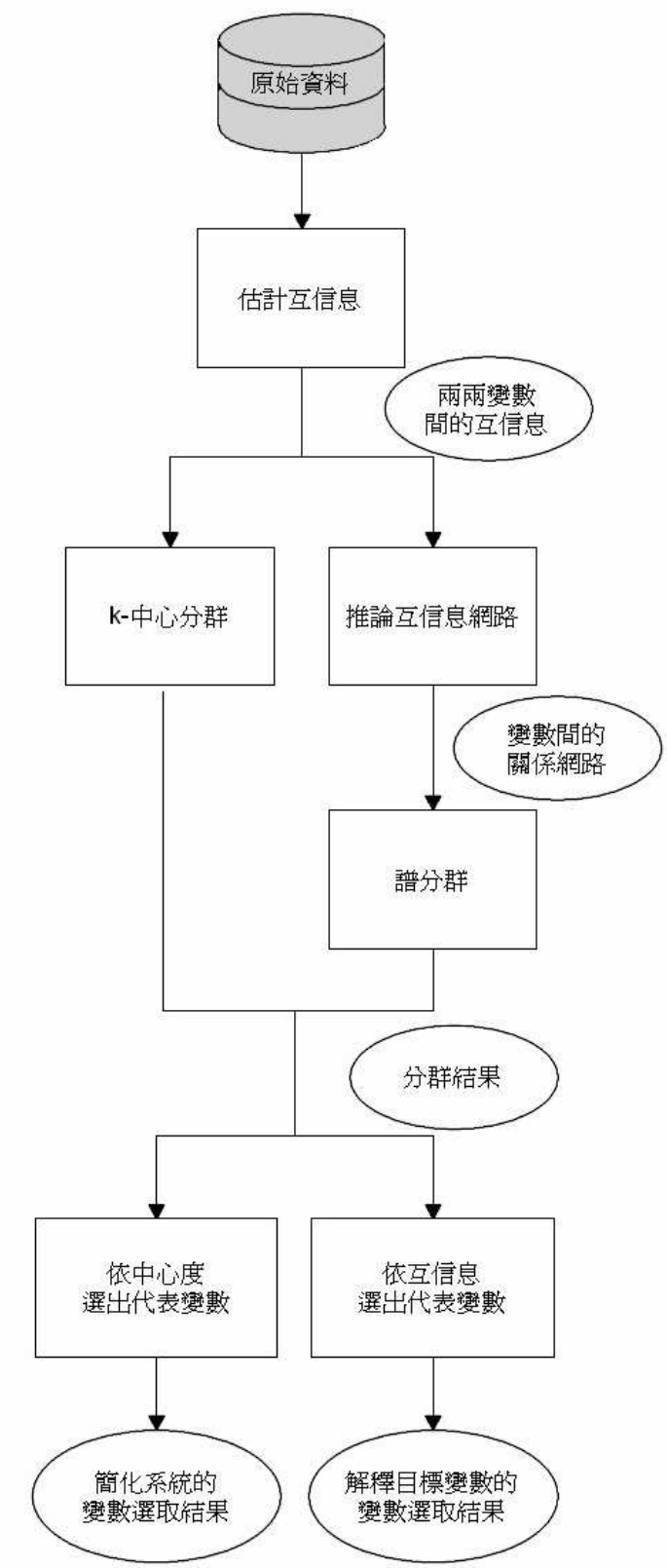


圖 4.1　研究方法詳細流程圖

第五章　　實驗

5.1　實驗簡介

實驗使用的資料取自 Statistica 的內建資料集。這包資料是真實晶圓自動化製程中所得的原始數據，共有 31 個變數，其中一個爲目標變數 Yield (良率)，其餘 30 個皆爲預測良率的變數。爲方便起見，將 30 個變數重新命名爲 var1、var2、…、var30。總共有 154 個樣本數，剔除離群值後留下 151 個樣本。原資料中有少數幾個缺值，以平均值填入。從 30 個變數中取 9 個變數加上良率畫成散布圖如圖 5.1 (最右一行爲良率)，可發現這包資料的變數大都是非線性關係。

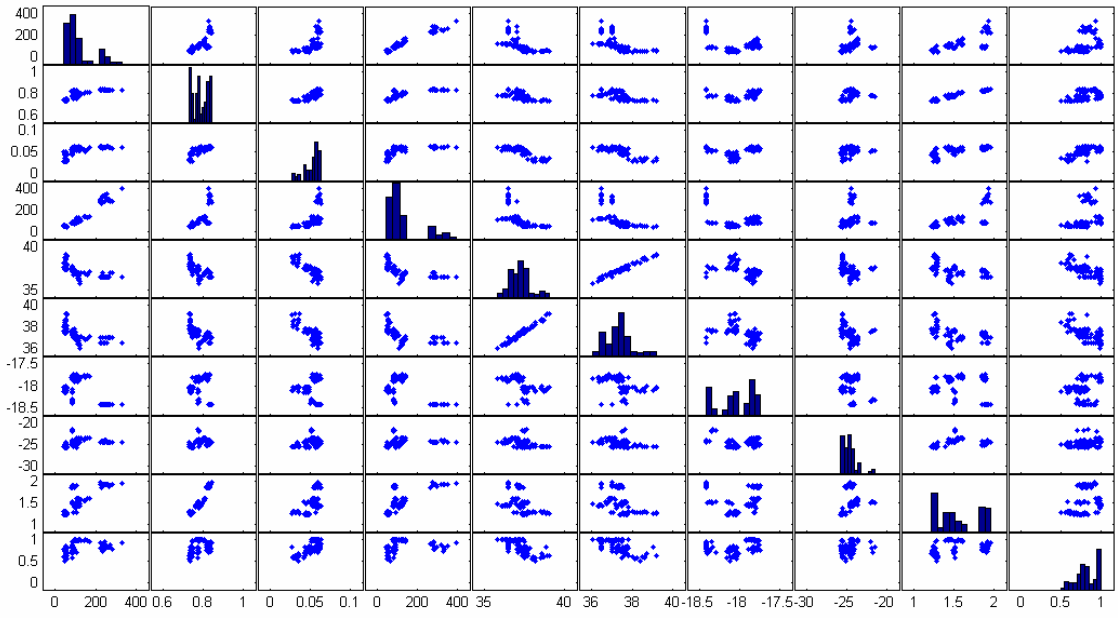


圖 5.1　晶圓資料散佈圖

5.2 變數分群結果

變數分群目的在於探討一個系統中變數間的關聯性以及變數群聚的情形，要評價此問題結果的優劣，最好的辦法就是跟專家知識做比較。但難處就在於我們找不到這樣的專家。因此我們以因素負荷當作對照參考，試著去解釋之間的異同。表 5.1 是經過最大變異旋轉後的因素負荷，標示出絕對值大於 0.7 的值 (粗體)。與同一因素高相關的變數被視爲同一群，第九個因素之後的因素負荷皆小於 0.7。即使列出了八個因素仍有變數與這八個因素的相關係數均小於 0.7，如 var12、var13，故 var12、var13 不屬於任何一群。

var	因素1	因素2	因素3	因素4	因素5	因素6	因素6	因素7
1	**-0.751**	0.294	-0.029	-0.025	0.089	-0.225	-0.185	-0.450
2	**-0.781**	0.270	-0.060	0.009	0.023	-0.241	-0.167	-0.421
3	**-0.795**	0.248	-0.033	0.040	0.027	-0.249	-0.118	-0.443
4	**-0.811**	0.292	0.026	0.058	0.078	-0.234	-0.033	-0.360
5	**-0.831**	0.275	0.013	-0.005	0.077	-0.190	-0.061	-0.325
6	**-0.839**	0.222	0.065	0.076	0.029	-0.239	-0.052	-0.348
7	**-0.837**	0.384	-0.041	0.020	0.068	-0.239	0.009	-0.236
8	**-0.844**	0.315	-0.012	0.084	0.039	-0.248	-0.083	-0.285
9	**-0.850**	0.251	-0.015	0.154	0.013	-0.256	-0.037	-0.274
10	-0.537	0.005	-0.113	0.115	0.066	**-0.748**	-0.144	-0.304
11	-0.493	0.064	-0.169	0.109	0.053	**-0.770**	-0.124	-0.312
12	-0.381	-0.063	0.226	0.154	0.057	-0.523	-0.104	-0.611
13	-0.386	-0.045	0.142	0.198	0.094	-0.489	-0.174	-0.602
14	**-0.838**	0.363	-0.008	0.061	0.050	-0.262	-0.058	-0.268
15	**-0.858**	0.299	-0.008	0.130	0.025	-0.250	-0.035	-0.253
16	**-0.820**	0.405	-0.019	0.013	0.046	-0.227	-0.058	-0.276
17	**-0.824**	0.391	-0.022	0.009	0.059	-0.228	-0.064	-0.286
18	-0.043	0.091	**-0.932**	0.110	0.041	-0.180	-0.038	0.271
19	0.424	0.041	-0.085	-0.114	-0.033	0.195	0.020	**0.864**
20	0.402	0.036	-0.093	-0.128	-0.056	0.175	0.013	**0.870**
21	0.501	0.010	-0.102	-0.124	-0.014	0.214	0.067	**0.795**
22	0.442	0.104	-0.120	-0.013	-0.040	0.100	0.218	**0.826**
23	0.096	-0.071	0.035	-0.084	**-0.982**	0.083	0.041	0.072
24	0.384	**-0.899**	0.012	-0.113	-0.033	0.134	0.040	0.050
25	0.316	**-0.930**	0.074	-0.020	-0.034	-0.022	-0.074	-0.106
26	0.329	**-0.937**	0.043	-0.051	-0.020	-0.007	-0.024	-0.052
27	0.332	**-0.937**	0.008	-0.019	-0.032	-0.016	-0.037	-0.054
28	-0.077	0.155	-0.113	**0.921**	0.100	-0.130	-0.182	-0.221
29	0.206	0.142	0.059	-0.308	-0.071	0.373	**0.797**	0.243
30	-0.498	0.083	-0.170	0.026	0.071	**-0.803**	-0.162	-0.178

表 5.1　因素分析的結果

由於 k-中心的輸入是未經任何處理的兩兩變數間的互信息，隨機產生的群中心初始值會影響分群結果，因此我們使用蒙特卡羅方法 (Monte Carlo method) 執行 100 次 k-中心分群，藉由影值來選擇結果最好的那次。令 $s^m(i)$ 表示變數 i 在第 m 次分群結果的影值，兩種選擇分群結果的準則如下：

$$\max_m \sum_{s^m(i)<0} s^m(i) \tag{4.13}$$

$$\max_m \sum_i s^m(i) \tag{4.14}$$

(4.13)式選擇負的影值和最大的那次分群結果，(4.14)式則是選擇全部的影值和最大的那次。ARACNE 和 MRNET 也如法炮製地重複執行譜分群多次*，因爲譜分群已經轉換到另一個較利於分群的特徵向量空間，所以我們只跑 10 次。選擇分群結果的準則一樣如(4.13)式和(4.14)式，最後結果如圖 5.2 和圖 5.3。

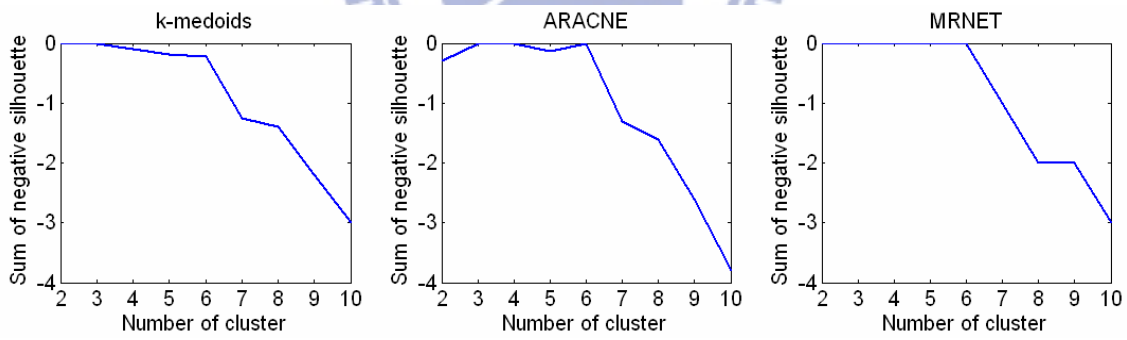


圖 5.2　k-中心、ARACNE 和 MRNET 在不同分群數下的負影值和

* 實際上是執行譜分群的最後一個步驟 k-平均。

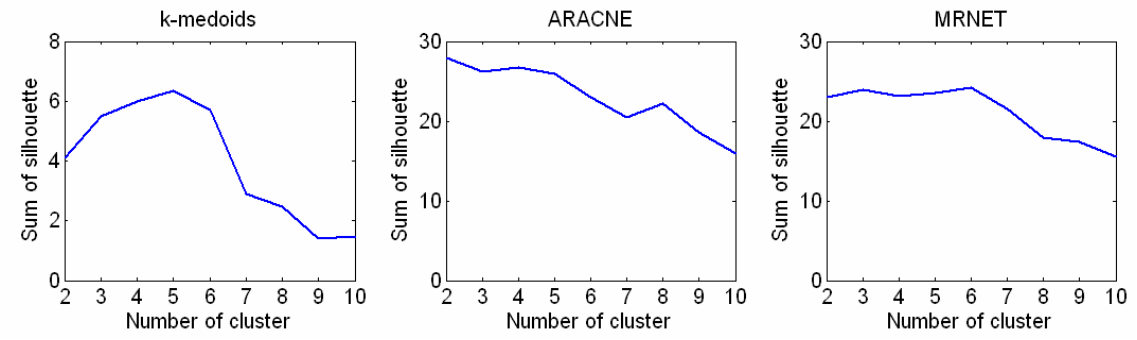


圖 5.3　k-中心、ARACNE 和 MRNET 在不同分群數下的影值和

由圖 5.2 我們可看出三者一致地當分群數大於六之後，負的影值便快速增加，由此判斷分六群會是一個不錯的選擇。圖 5.3 爲全部的影值和，注意 k-中心縱軸的刻度與其他兩張圖不同，原因是未經處理的互信息中存在許多虛假相關，這會使得影值計算中的 $b(i)$ 項 ((4.1)式) 變大，結果就是整體的影值變小。ARACNE 和 MRNET 爲消除虛假相關後的互信息網路，影值和比 k-中心大了不少。k-中心和 MRNET 在圖 5.2、圖 5.3 中的表現是一致的，最佳分群數在五、六左右。ARACNE 則沒有這個現象，在圖 5.2 中的最佳分群數是六，但在圖 5.3 中分兩群就達到峰值了。MRNET 在圖 5.3 中的曲線也不像 k-中心那麼明顯，分群數二到六之間影值和的差異不大。或許這就是互信息網路的特性，刪除掉一些虛假相關的邊使得利用影值來找尋最佳分群數目時沒有明顯的結果。但在另一方面，簡潔的互信息網路可用於變數關係的視覺化，使我們清楚知道哪些變數有真正的直接關係，如圖 5.4。圖 5.4 爲 ARACNE 互信息網路分五群的結果，同顏色代表同一群，邊上的權重爲互信息。

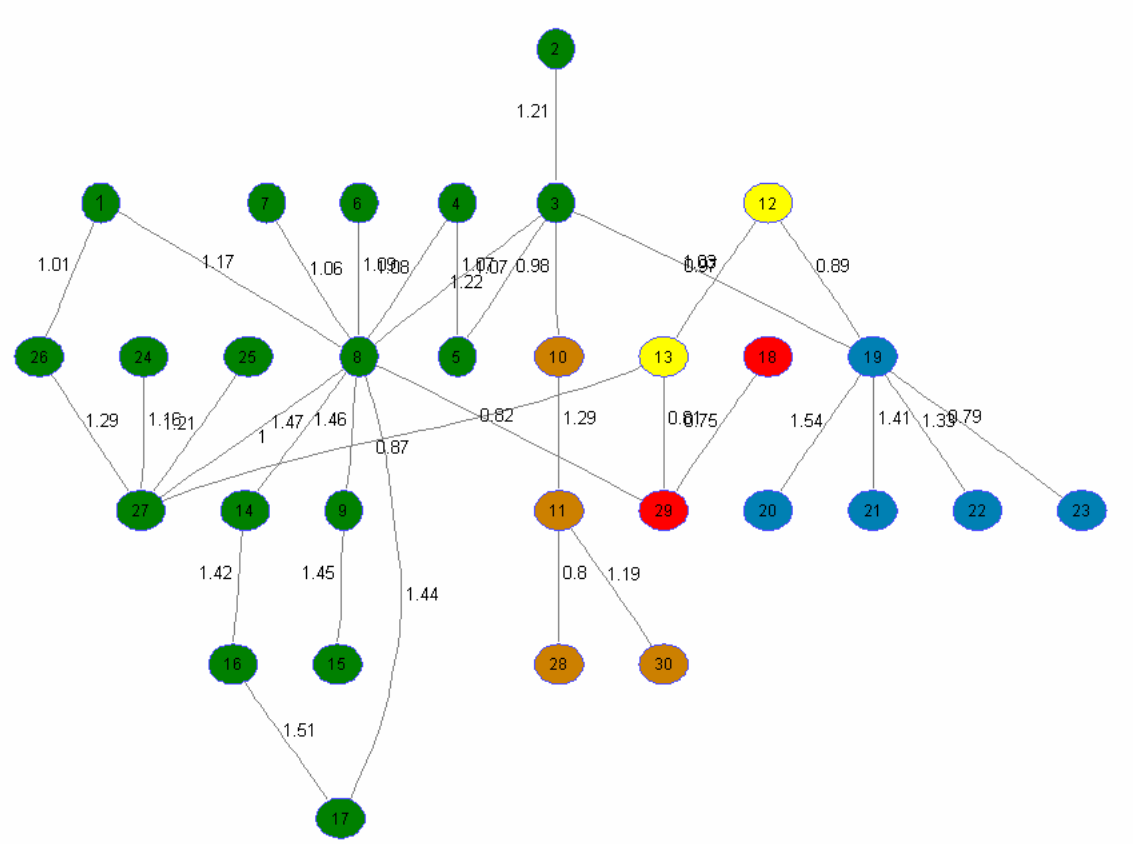


圖 5.4　ARACNE 互信息網路分五群

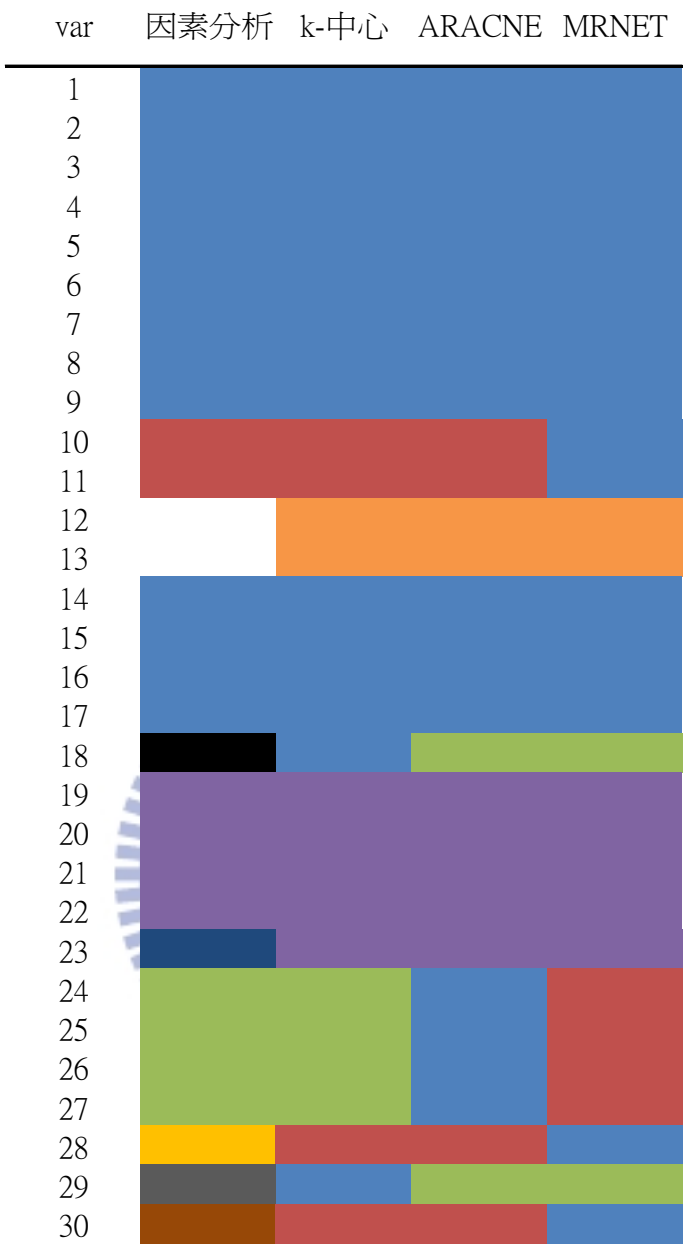


圖 5.5 分群結果*

圖 5.5 爲 k-中心、ARACNE 和 MRNET 分五群和用因素分析分八群的結果。k-中心、ARACNE 和 MRNET 這三者的分群非常類似但與因素負荷的結果略有不

* 因素分析中 var12、var13 空白表示沒被分群。

同，我們簡單討論不同處如下：

var12、var13 未被分群：

從散佈圖 (圖 5.6) 上來看，var12 與 var13 大致上呈線性關係，不論互信息或相關系數， var12 和 var13 都表現地非常一致，且和對方的互信息與相關系數都相對高(圖 5.7)，因此姑且不管其他變數， var12、var13 應該要被分在一群。

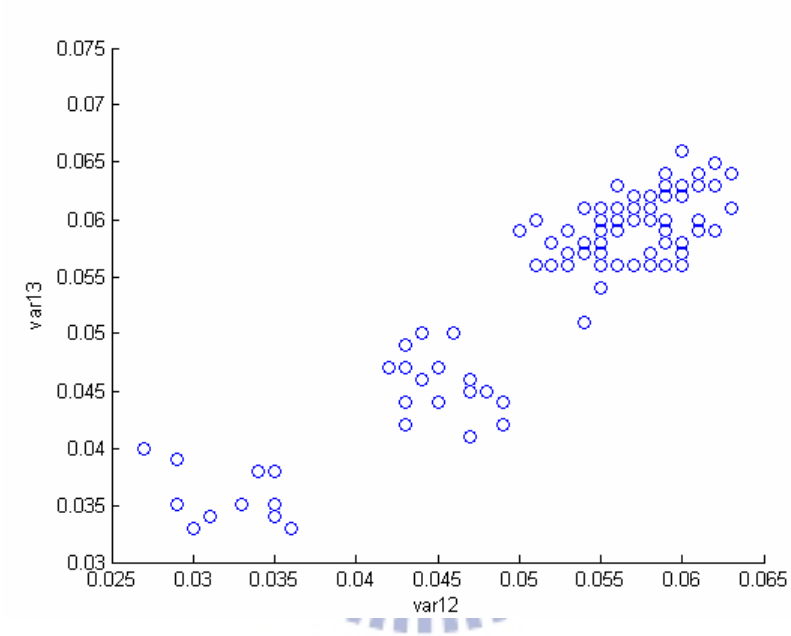


圖 5.6 var12 和 var13 的散佈圖

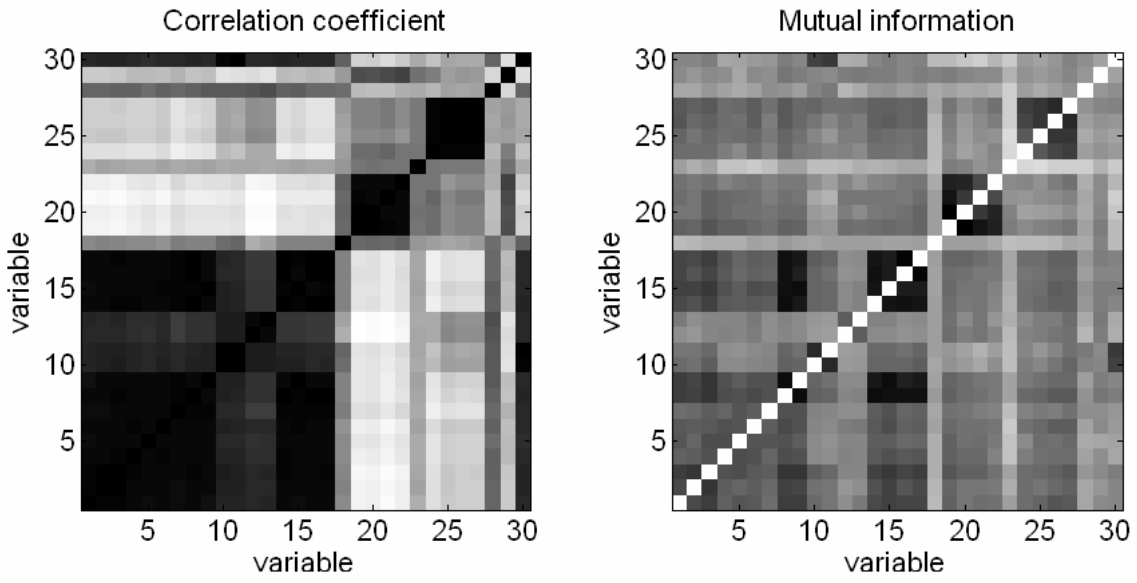


圖 5.7　變數間的相關系數 (絕對值) (左) 和互信息 (右)*

var28 與 var10、var11、var30：

var10、var11 是兩個與其他變數都不太相關的變數，除了和彼此的互信息相對大外，和 var30 的互信息也相對大，圖 5.8 中 var10、var11 和 var30 的關係大致爲線性，故它們三者應爲同一群。var28 和其他變數的相關係數都很低，幾乎都在 0.2 到 0.4 之間 (絕對值)，互信息最高的三個就是和 var10、var11、var30，分別爲 0.85、0.88、0.81，其餘幾乎都在 0.7 以下。var28 和 var10、var11、var30 也的確不是線性關係，但若將 var28 和 var10、var11、var30 的散佈圖分別用二次曲線去近似 (圖 5.9)，可發現他們都呈一種類似的二次關係，故將 var10、var11、var28、var30 分在同一群。

* x 軸與 y 軸同樣爲 30 個變數，顏色從白到黑代表值從小到大。

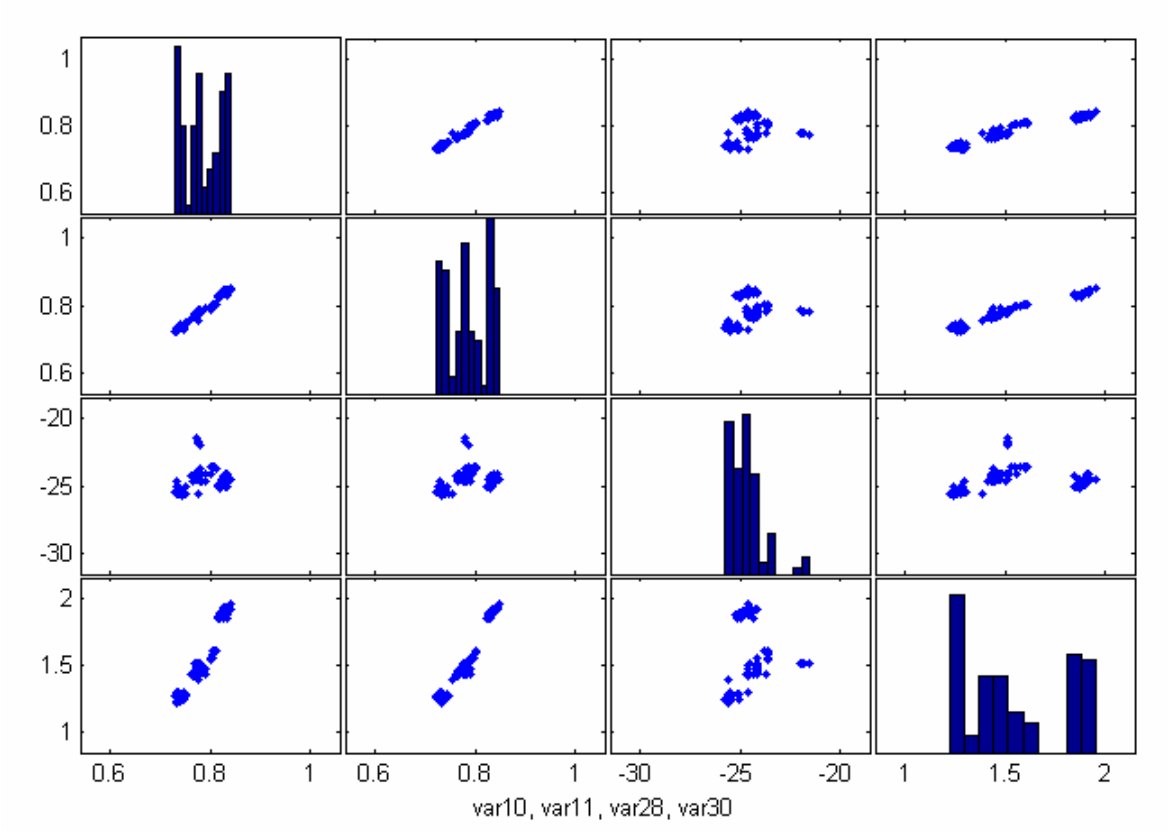


圖 5.8　var10、var11、var28 與 var30 的散佈圖

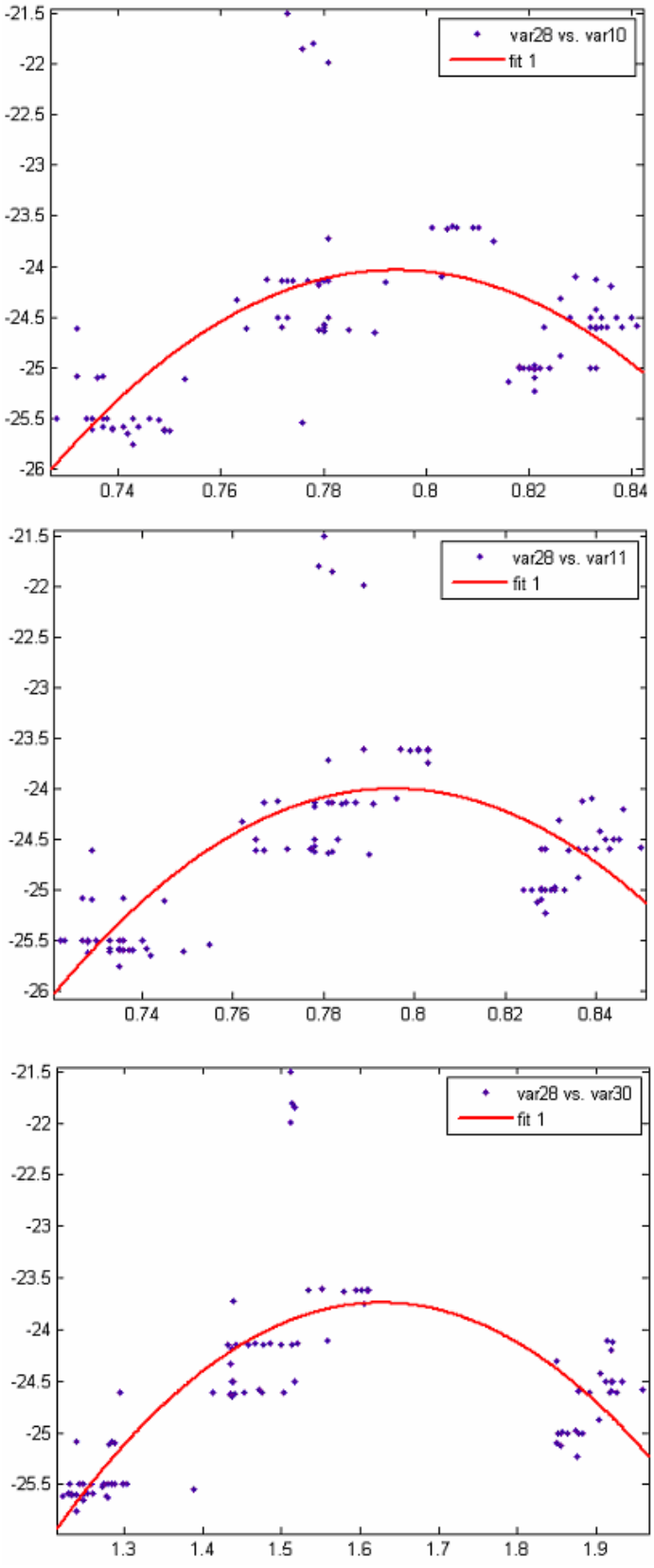


圖 5.9　var28 與 var10、var11、var30 的二次回歸

5.3 變數選取結果

針對簡化系統的變數選取，我們以 k-中心的群中心或是互信息網路中度中心度最高的變數作爲群的代表變數，每群選出一個代表變數，故分幾群就有幾個代表變數。全部 30 個變數可以簡化成以這些變數來代表。實驗資料共有 151 筆取樣，取其中的 101 筆爲訓練資料 (training data)，另外 50 筆爲測試資料 (test data)，以訓練資料選取變數，測試資料做驗證。我們分別選出三個、五個和七個代表變數，結果如表 5.2。

因素分析	15	18	26				
k-中心	8	19	29				
ARACNE	8	11	19				
MRNET	8	19	27				
因素分析	15	26	18	28	23		
k-中心	8	10	13	19	27		
ARACNE	8	11	13	19	29		
MRNET	8	13	19	27	29		
因素分析	15	28	23	26	28	29	30
k-中心	3	8	10	19	25	27	29
ARACNE	3	8	9	11	19	23	29
MRNET	8	11	19	23	27	28	29

表 5.2　簡化系統的變數選取結果

我們仿照 MRMR 設計出一個指標來做這部分的驗證

$$score = -\sum_{v_j \in \bar{S}} \min_{v_i \in S} \frac{H(v_j \mid v_i)}{H(v_j)} + \frac{1}{|S|-1} \sum_{v_i \in S, v_k \in S, i \neq k} \frac{H(v_k \mid v_i)}{H(v_k)} \qquad (4.15)$$

S 爲代表變數的集合，$\overline{S}$ 是 S 以外的其他變數集合，$|\cdot|$ 表示集合的元素個數，$H(\cdot)$ 和 $H(\cdot|\cdot)$ 分別爲熵和條件熵。表 5.2 的任何一組代表變數都可以代進(4.15)式計算出一個分數，此分數望大，表示其他變數可被此組代表變數解釋且代表變數之間的相關性不大。(4.15)式的第一項的意義爲群內相關 (intra-dependence)，是一個熵的比例，分母爲 $\overline{S}$ 內某一個變數 v_j 的熵，分子爲已知 v_i 後 v_j 的熵， v_i 爲某一個代表變數。此比例越小表示在 v_i 爲已知的情況下，v_j 的混亂度下降很多，意即 v_i 可以解釋 v_j 。取最小值的意義在於只要 S 中的某一個 v_i 可以解釋 v_j 即可，實際實驗時我們發現使 v_j 的混亂度下降最多的 v_i 通常就是 v_j 所屬那群的代表，故稱“群內”相關。第二項亦是一個熵的比例，表示群間相關 (inter-dependence)，不同的是 v_i 和 v_k 均屬於 S ，此比例越大表示 v_i 和 v_k 越無法互相解釋， v_i 和 v_k 越獨立。我們希望其他變數都能夠被代表變數解釋且代表變數之間盡量獨立，因此第一項望小，第二項望大，第一項加上負號故相加後的分數是望大。

	三個代表變數	五個代表變數	七個代表變數
因素分析	-2.04	-1.30	1.21
k-中心	-2.68	1.83	3.12
ARACNE	-1.76	0.71	2.00
MRNET	-2.80	1.26	2.15

表 5.3　簡化系統變數選取的驗證結果

表 5.3 是利用(4.15)式所得到的分數，選取三個代表變數時 ARACNE 的得分最高，五個、七個代表變數時則是 k-中心的得分最高，顯示我們選出的代表變數

較因素分析選出的具系統代表性。

針對解釋目標變數的變數選取，我們以良率作爲目標變數，根據選出的解釋變數對良率進行預測 (分類和迴歸)，並與 MRMR、逐步選取法所選出的變數作比較。分類和迴歸都分別各以五種方法測試，分類用線性鑑別分析 (linear discriminant analysis, LDA)、k-最近鄰居分類 (k-nearest neighbor classification)、學習向量量化 (learning vector quantization, LVQ)、類神經網路 (neural network, NN)及分類迴歸樹 (classification and regression tree, CART)；迴歸用線性迴歸 (linear regression)、k-最近鄰居迴歸 (k-nearest neighbor regression)、輻射基底函數 (radial basis function, RBF)、類神經網路和分類迴歸樹。以上方法的簡介請見附錄。

訓練資料用來選取解釋變數以及建立迴歸、分類的模型。測試資料用來測試選取的變數對良率(目標變數)的解釋能力。對於分類，良率以 0.9 作爲分界點，高於 0.9 的爲一類 (高良率)，低於 0.9 的爲一類 (低良率)。分類結果以正確率 (accuracy)表示。對於迴歸，我們比較 *SSE* (sum of squared error)

$$SSE = \sum_i (y_i - \hat{y}_i)^2 \tag{4.16}$$

y_i 和 $\hat{y}_i$ 分別爲實際值與估計值，*SSE* 越低代表迴歸結果越準確。

我們分別選擇三個、五個和七個解釋變數來對良率進行分類和迴歸，除 MRMR 選出的變數有順序外，其餘方法所選的變數均無關次序。 k-中心、ARACNE 和 MRNET 也分別分三群、五群和七群，從每群選出和良率最相關 (互信息最大) 的變數作爲解釋變數，結果如表 5.4。

MRMR	6	18	30				
逐步選取法	20	25	28				
k-中心	6	11	20				
ARACNE	6	20	20				
MRNET	6	20	20				

MRMR	6	18	30	20	4		
逐步選取法	6	13	25	28	30		
k-中心	6	11	12	20	27		
ARACNE	6	11	12	18	20		
MRNET	6	12	18	20	27		

MRMR	6	18	30	20	4	12	5
逐步選取法	6	13	20	21	25	28	30
k-中心	5	6	11	12	18	20	27
ARACNE	4	6	11	15	18	20	23
MRNET	6	11	18	20	23	27	28

表 5.4　解釋良率的變數選取結果

三個解釋變數	LDA	kNN	LVQ	NN	CART	平均
MRMR	0.76	0.92	0.88	0.82	0.94	0.864
逐步選取法	0.8	0.96	0.88	0.86	0.94	0.888
k-中心	0.8	0.96	0.86	0.82	0.92	0.872
ARACNE	0.8	0.96	0.88	0.86	0.98	0.896
MRNET	0.74	0.88	0.86	0.86	0.88	0.844

五個解釋變數	LDA	kNN	LVQ	NN	CART	平均
MRMR	0.78	0.96	0.92	0.92	0.96	0.908
逐步選取法	0.9	0.96	0.88	0.88	0.92	0.908
k-中心	0.92	0.96	0.86	0.86	0.98	0.916
ARACNE	0.82	0.96	0.88	0.84	0.98	0.896
MRNET	0.8	0.96	0.86	0.86	0.98	0.892

七個解釋變數	LDA	kNN	LVQ	NN	CART	平均
MRMR	0.82	0.96	0.9	0.9	0.96	0.908
逐步選取法	0.96	0.96	0.84	0.84	0.96	0.912
k-中心	0.92	0.96	0.92	0.88	0.94	0.924
ARACNE	0.78	0.96	0.8	0.8	0.94	0.856
MRNET	0.96	0.94	0.82	0.82	0.98	0.904

表 5.5　分類結果 (正確率)

表 5.5 是分類結果，我們可看出選擇三個變數時，ARACNE 有最高的平均正確率；五個和七個變數時，k-中心選出的變數有最高的平均正確率。其中 kNN 和 CART 這兩種分類方法的分類效果較好。

三個解釋變數	linear Reg	kNN	RBF	NN	CART	平均
MRMR	0.752	0.312	1.012	0.772	0.374	0.644
逐步選取法	0.483	0.305	0.396	0.474	0.287	0.389
k-中心	0.567	0.229	0.470	0.579	0.254	0.420
ARACNE	0.560	0.247	0.683	0.554	0.209	0.451
MRNET	0.598	0.434	0.504	0.645	0.433	0.523

五個解釋變數	linear Reg	kNN	RBF	NN	CART	平均
MRMR	0.542	0.233	0.541	0.535	0.275	0.425
逐步選取法	0.391	0.258	0.553	0.387	0.475	0.413
k-中心	0.328	0.268	0.503	0.327	0.281	0.341
ARACNE	0.497	0.242	0.747	0.484	0.229	0.440
MRNET	0.607	0.264	0.752	0.629	0.415	0.534

七個解釋變數	linear Reg	kNN	RBF	NN	CART	平均
MRMR	0.603	0.229	0.697	0.599	0.442	0.514
逐步選取法	0.315	0.274	0.577	0.394	0.272	0.366
k-中心	0.373	0.222	0.542	0.466	0.355	0.392
ARACNE	0.570	0.248	0.672	0.767	0.369	0.525
MRNET	0.336	0.278	0.697	0.337	0.213	0.372

表 5.6　迴歸結果(SSE)*

迴歸結果如表 5.6，選擇一個、七個解釋變數時，逐步選取法選出的變數有最低的 SSE；五個變數時，k-中心勝過逐步選取法且其 *SSE* 是所有情況中最低的。根據圖 5.3 可知 k-中心最佳的分群數爲五，由於我們是採分幾群就選幾個變數的模式，圖 5.3 可以當作選取變數數目的一個參考。另一方面，逐步選取法是內嵌於線性迴歸內的變數選取方法，選取變數的機制就是專爲迴歸而設計的，或許因爲這樣使得逐步選取法在迴歸上有不錯的表現，而我們的方法並沒有特別針對分類

* linear Reg = linear regression。

或迴歸。

綜合表 5.5 分類結果和表 5.6 迴歸結果，k-中心選出的變數對良率的解釋能力大致上比其他方法好。互信息網路搭配譜分群 (ARACNE、MRNET) 所選出的變數雖然結果不如 k-中心和逐步選取法，但依然有它存在的價值。互信息網路旨在闡明變數間真正的關係，但變數間的關係網路並不是只有互信息網路，也可能來自專家知識或貝氏網路的結構學習等。假使我們已經由某種方法得到變數的關係網路，那麼就可以省略前面計算互信息和互信息網路的步驟，同樣可以做變數分群和變數選取。

5.4 離散資料的變數分群與選取結果

如第一章中所述，使用互信息的優點就是可以處理離散資料，在此情況下我們依然可以完成互信息的計算、變數分群和變數選取。爲簡單起見，我們使用同一包資料，先對每個變數作離散化處理：將每個變數的數值用 k-平均分成高、中、低三類，之後變數在取樣中的值就只以高、中、低這三個類別表示。當然，經過此一轉換後互信息、變數分群和變數選取的結果必定會和原本的不同。

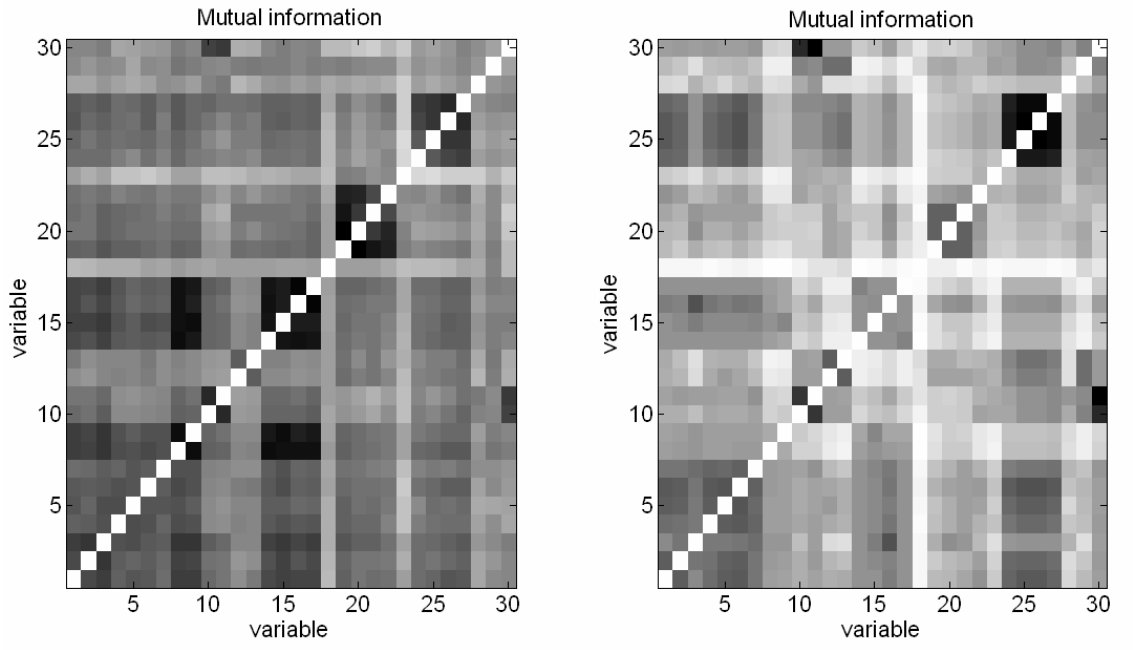


圖 5.10 連續資料(a)和離散資料(b)的*互*信息

圖 5.10 是*互*信息的比較，(a)是用原本連續資料計算出的*互*信息，(b)是用經過上述轉換後的離散資料計算出的*互*信息。可發現(b)中的*互*信息整體下降，更重要的是有些變數間的關係結構也被破壞掉了，圖上來看最明顯的就是 var14、var15、var16、var17 這四個變數，原本四個變數之間非常相關 (*互*信息很大)，離散化後他們之間的關係不復存在。var14、var15、var16、var17 與 var8、var9 的情形也是如此。以 var17 和 var8 爲例，圖 5.11 是 var17 與 var8 的散佈圖，可看出這兩個變數呈線性正相關，相關系數高達 0.979，*互*信息 1.437 也算大。圖上的分隔線是用 k-平均離散化後類別之間的分界線，共分成九個小區域，圖 5.12 爲對應每個小區域內的取樣個數統計，可以將圖 5.12 當成是資料離散化後 var17 與 var8 的散布圖，已經看不出他們之間的線性關係了，大約五分之四的取樣都擠在左下的區域。離散化後的*互*信息降爲 0.401。

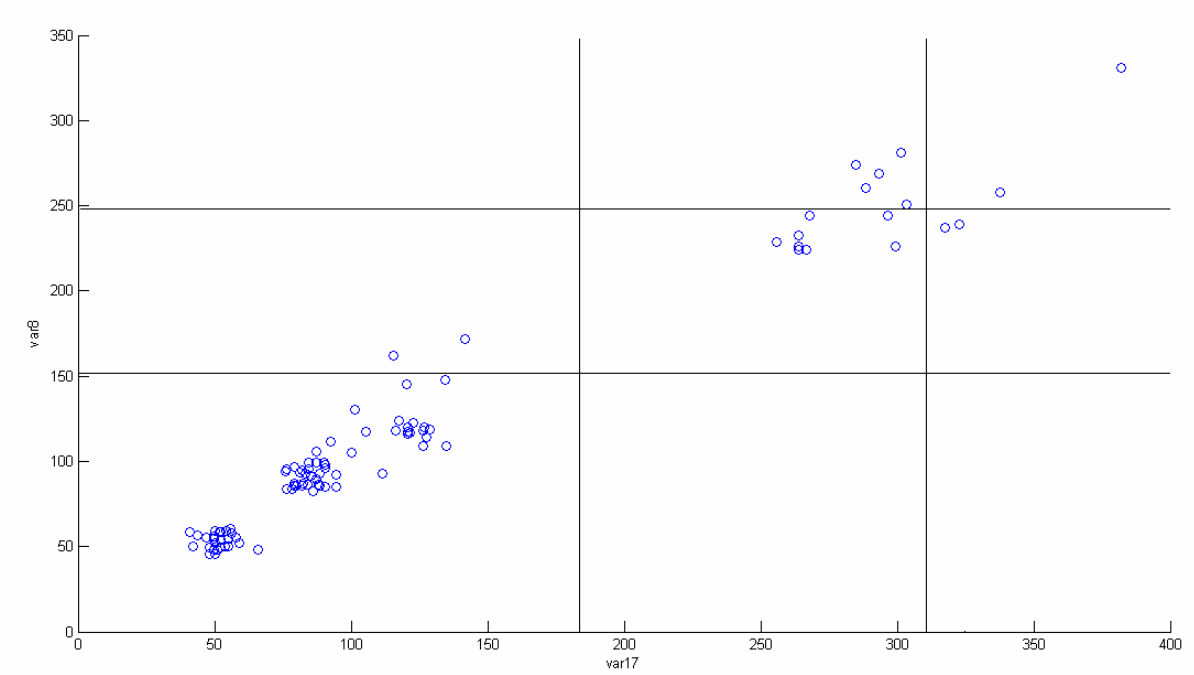


圖 5.11 var17 與 var8 的散佈圖

0	5	2
2	8	2
82	0	0

圖 5.12 對應圖 5.11 的取樣個數統計圖

上面的例子在探討資料離散化後變數之間關係的失真情形，有可能我們一開始得到的資料就是離散的，也就沒有所謂失不失真的問題。我們所要強調的是，即便資料是離散的，我們的方法依然可以進行，變數選取結果如表 5.7 和表 5.8。

k-中心	3	21	30				
ARACNE	3	21	30				
MRNET	3	18	21				
k-中心	3	8	15	21	30		
ARACNE	3	18	21	25	30		
MRNET	3	18	21	23	30		
k-中心	3	8	15	17	21	25	30
ARACNE	3	8	18	23	28	29	30
MRNET	3	15	18	21	23	29	30

表 5.7　簡化系統的變數選取結果 (離散資料)

MRMR	30	22	18				
k-中心	21	25	30				
ARACNE	12	21	30				
MRNET	18	21	30				
MRMR	30	22	18	25	19		
k-中心	8	15	21	25	30		
ARACNE	8	12	18	21	30		
MRNET	12	18	21	23	30		
MRMR	30	22	18	25	19	28	12
k-中心	3	8	12	15	16	21	30
ARACNE	12	18	21	23	28	29	30
MRNET	12	18	15	21	23	29	30

表 5.8　解釋良率的變數選取結果 (離散資料)

第六章　　結論

本論文針對傳統變數分群與變數選取方法的限制 (變數間呈線性關係、連續資料、資料呈常態分布) 與缺失提出一套基於互信息理論的新方法。互信息的優點在於它可量化變數間的關聯程度，包括線性與非線性；離散資料和混合型態的資料 (部分連續，部分離散) 下也可計算，同時對於資料的分布也沒有設限。除使用互信息來克服以上所述限制外，我們提出的方法還有一個優點有別於傳統方法，那就是：傳統方法的目標大都是單一的，例如階層式分群就只能做變數分群，逐步選取法就針對目標變數挑選解釋變數，而我們的方法是全面的，可以同時達到變數分群、簡化系統的變數選取和解釋目標變數的變數選取這三種目標。

根據實驗結果總結出以下幾點：

1. 變數分群沒有一個絕對標準的答案，我們將分群結果和因素分析的結果對照比較，發現因素分析傾向於將變數分較多群，縱使已經做了因素旋轉。我們進一步由散布圖、二次迴歸等不同的層面來解釋分群結果的差異。
2. 針對簡化系統的變數選取，我們選擇具有高中心度的變數，因爲這些變數與系統中其他變數的關係密切，以網路的觀點來看，這些變數爲訊息流通的樞紐。我們仿照 MRMR 設計出一個指標來驗證所選的代表變數，結果顯示我們所選的變數在這個指標下優於因素分析。
3. 針對解釋目標變數的變數選取，我們用不同的方法對良率進行分類和迴歸，以驗證解釋變數對良率的解釋能力，同時與逐步選取法和 MRMR 所選出的變數作比較。結果顯示 k-中心所選出的變數在分類上勝過其他方法，迴歸則是逐步

選取法較具優勢。互信息網路搭配譜分群 (ARACNE、MRNET) 所選出的變數雖然比不上 k-中心和逐步選取法，但可應用於已知的變數關係網路。

4. 離散資料我們的方法依然可以運行，將同一包資料離散化後來實驗，因離散化後計算出的互信息會有失真，所以變數分群和選取的結果也和資料連續時不同。

本論文的改進能從以下幾點著手：

1. 實驗中我們是採變數分幾群就選幾個變數的模式 (每群只選出一個代表變數或解釋變數)，事實上不一定要如此，可能考量成本後希望選取較多的變數以掌握較完整的資訊。但變數並不一定適合分那麼多群，此時便可用前述的方法找出最佳的分群數，再從每一群中選出一個以上的變數以滿足需求。
2. 互信息網路搭配譜分群的最佳分群數指標不像 k-中心那麼清楚，且選出來的變數無論是簡化系統或解釋良率都比不上 k-中心。選取變數的方式都一樣，所以問題可能出在變數分群，或許可以嘗試其他的圖分割方法，如凱尼漢-林演算法 (Kernighan-Lin algorithm)。
3. 本論文提出的方法建立在互信息之上，資料取樣的多寡、不同的估計方法、參數所計算出的互信息會有差異，互信息的變異對變數分群、變數選取的影響以及影響的程度也是一個值得探討的問題。

附錄

線性鑑別分析 (linear discriminant analysis, LDA)

LDA 爲一種監督式的分類方法，對於一筆待分類的取樣 $\mathbf{x} \in R^p$，我們將它分到 c 類根據

$$\arg\max_c \; p(\mathbf{x}\,|\,c)\,p(c) \tag{7.1}$$

$p(\mathbf{x}\,|\,c)$ 和 $p(c)$ 分別爲 $\mathbf{x}$ 是 c 類的可能性 (likelihood) 和 c 類的事前機率 (prior probability)。我們假設每類中的資料均呈多維高斯分佈 (multivariate normal distribution) 且每類的共變異矩陣 (covariance) $\mathbf{\Sigma}$ 均相同

$$p(\mathbf{x}\,|\,c) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_c)\right) \tag{7.2}$$

$\boldsymbol{\mu}_c$ 爲 c 類平均。根據(7.1)式，我們目的是要找出使得 $p(\mathbf{x}\,|\,c)\,p(c)$ 最大的類別 c，等同於要讓 $\ln(p(\mathbf{x}\,|\,c)\,p(c))$ 最大

$$\ln(p(\mathbf{x}\,|\,c)\,p(c)) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_c) - \frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{p}{2}\ln(2\pi) + \ln(p(c)) \tag{7.3}$$

扣除常數項後(7.1)式可重新寫成

$$\arg\max_c \left((\mathbf{x}^T \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\boldsymbol{\mu}_c{}^T \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_c + \ln(p(c))\right) \tag{7.4}$$

線性迴歸 (linear regression)

統計上，迴歸分析是利用一組解釋變數（$\mathbf{x}_i$, i=1,…,p）來預測目標變數（$\mathbf{y}$）的方法。線性迴歸令 $\mathbf{y}$ 的估計值 $\hat{\mathbf{y}}$ 爲解釋變數的線性組合

$$\hat{\mathbf{y}} = \beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_p \mathbf{x}_p \tag{7.5}$$

$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \cdots & \beta_p \end{bmatrix}^T$ 爲線性迴歸模型中的參數，我們的目標就是利用訓練資料來估計 $\boldsymbol{\beta}$ ， $E(\boldsymbol{\beta})$ 爲 RSS (residual sum of squares)

$$\begin{aligned} &\arg\min_{\boldsymbol{\beta}} E(\boldsymbol{\beta}) \\ &E(\boldsymbol{\beta}) = \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 \end{aligned} \tag{7.6}$$

$\|\cdot\|$ 爲向量的 2-模 (2-norm)， $\mathbf{X}$ 和 $\mathbf{y}$ 爲訓練資料， $\mathbf{X}$ 的列代表取樣，行代表不同的解釋變數，爲了湊合常數 β_0 ，必須在原有資料前加上一行 1，始成爲 $\mathbf{X}$ 。爲了找到使 $E(\boldsymbol{\beta})$ 最小的 $\boldsymbol{\beta}$

$$\frac{\partial E}{\partial \boldsymbol{\beta}} = 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^T\mathbf{y} = 0 \tag{7.7}$$

得到正規方程式 (normal equation)

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \tag{7.8}$$

於是

$$\boldsymbol{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{7.9}$$

將 $\boldsymbol{\beta}$ 帶入(7.5)即得到線性迴歸模型。

k-最近鄰居 (k-nearest neighbor, kNN)

kNN 是一個簡單卻有效的演算法，可用於分類和迴歸，概念如下：對於一筆待分類的取樣 $\mathbf{x}$ ，我們先計算 $\mathbf{x}$ 與訓練資料中每一筆取樣的距離 (實驗使用歐幾里得距離) 並將距離排序，取出距離最近的前 k 筆取樣，由這 k 個最近鄰居所屬的類別來投票決定 $\mathbf{x}$ 該被分到哪一類，爲避免有相同票數的情形發生，k 通常定爲奇數。迴

歸時，y 的估計值爲 k 個最近鄰居的平均。

學習向量量化 (learning vector quantization, LVQ)

LVQ 是一種基於原型的 (prototype-based)監督式分類演算法，步驟如下

1. 訓練資料中每個類別都任意取出數個取樣當作該類別的原型 (prototype)。
2. 從剩下的訓練資料中任意取出一筆取樣 $\mathbf{x}$，令 $\mathbf{m}$ 爲最靠近 $\mathbf{x}$ 的原型，移動 $\mathbf{m}$ 根據：若 $\mathbf{x}$ 和 $\mathbf{m}$ 屬於同一類別，則 $\mathbf{m}_{new} = \mathbf{m}_{old} + \eta(\mathbf{x} - \mathbf{m}_{old})$；若 $\mathbf{x}$ 和 $\mathbf{m}$ 屬於不同類別，則 $\mathbf{m}_{new} = \mathbf{m}_{old} - \eta(\mathbf{x} - \mathbf{m}_{old})$。$\eta$ 爲學習速率 (learning rate)，通常 $0.01 < \eta < 0.1$。
3. 重複步驟 2.直到所有訓練資料中的取樣都被使用過，在每一次的遞迴過程中，調整 η 使其越來越小。
4. 將某一筆測試資料分到和它最近的原型的那類。

LVQ 依據最終原型的位置來得到分類邊界，利用訓練資料的類別決定移動的方向，學習速率決定移動量的大小。步驟 2.中，若原型和 $\mathbf{x}$ 同類別，則將該原型往 $\mathbf{x}$ 的方向拉近，反之則往 $\mathbf{x}$ 的相反方向推遠。LVQ 的結果通常比 k-平均好。

輻射基底函數 (radial basis function, RBF)[11]

RBF 可視爲一種函數近似 (function approximation) 的方法，寫成以下形式

$$\hat{f}(\mathbf{x}) = \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}) \tag{7.10}$$

$\hat{f}(\mathbf{x})$ 是函數近似的結果。ϕ_j 爲基底函數 (basis function)，共有 M 個。w_j 是 ϕ_j 的權

重。常用的基底函數如高斯

$$\phi_j(\mathbf{x}) = \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \Sigma_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right\} \tag{7.11}$$

$\boldsymbol{\mu}_j$ 和 Σ_j 分別爲某一類別的平均和共變矩陣；對訓練資料執行非監督式分類 (如 k-平均，類別數目已知)，分類後即可求得第 j 類的 $\boldsymbol{\mu}_j$ 和 Σ_j。最小化 SSE (sum of squared error) 可求得權重 w_j (類似線性迴歸中求 $\boldsymbol{\beta}$ 的方法)

$$\begin{aligned} &\min\ SSE \\ &SSE = \sum_{i=1}^{N}(y_i - \hat{f}(\mathbf{x}_i))^2 = \sum_{i=1}^{N}(y_i - \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}_i))^2 \end{aligned} \tag{7.12}$$

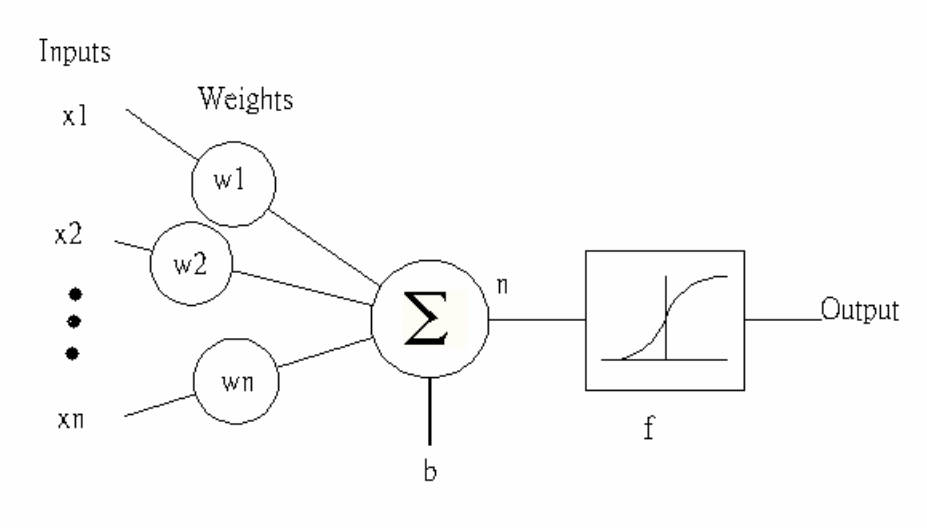
故

$$\mathbf{w} = \left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{y} \tag{7.13}$$

$\mathbf{x}_i$、y_i 爲訓練資料中的取樣，y_i 爲目標，共有 N 筆取樣。$\mathbf{w} = \begin{bmatrix} w_1 & \cdots & w_M \end{bmatrix}^T$，$\mathbf{y} = \begin{bmatrix} y_1 & \cdots & y_N \end{bmatrix}^T$，$\left(\Phi_{N\times M}\right)_{ij} = \phi_j(\mathbf{x}_i)$。

類神經網路 (neural network)[12]

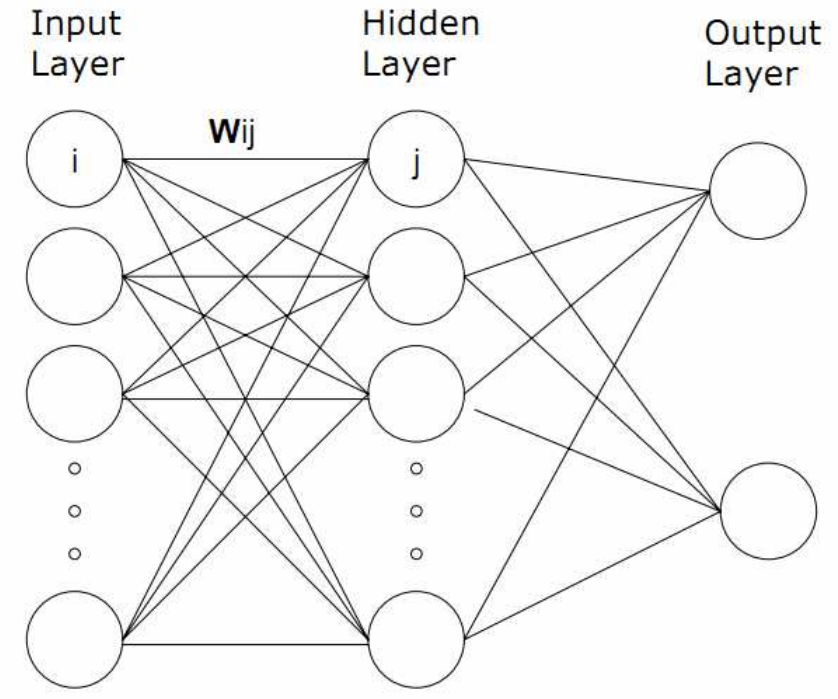
類神經網路爲模仿生物神經網路的資訊處理系統，有許多優點例如具有學習能力、儲存能力和容錯能力，因本質上即是屬於平行處理的架構，故運算速度非常快速。類神經網路可藉助學習和訓練的過程歸納出輸入資料中的隱含關係，學習完成後，儲存各神經元間連結的強弱程度，之後若遇到相似的輸入，就會依照學習的結果產生最近似的輸出，就算是資料不完整或是受到雜訊干擾，類神經網路也會作出最佳的預測。其基本的運作原理是以大量的、簡單的處理單元——神經元 (neuron) 互相連接，神經元的模型如下



$$n = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + b$$
$$Output = f(n) \tag{7.14}$$

$x_1, \ldots, x_n$ 爲輸入，來自外界環境或其他神經元。 $w_1, \ldots, w_n$ 爲權重，代表前一層或外界輸入對該神經元的相對強度，權重在類神經網路中相當重要，因爲網路的學習即是透過不斷地調整權重以減少誤差，使誤差收斂。 b 是某一設定的偏差 (bias)。 f 是轉移函數 (transfer function)，常用的如雙彎曲函數 (sigmoid function)，可產生 0 到 1 之間的輸出

$$f(n) = \frac{1}{1+e^{-n}} \tag{7.15}$$



上圖爲類神經網路的基本架構，其中隱藏層 (hidden layer) 的數目不一定只有一層，端看問題複雜度而定，類神經網路的訓練學習常使用誤差倒傳 (error back-propagation) 的方法來調整權重，訓練過程可分爲兩部分

1. 向前傳遞 (forward propagation)：計算每個神經元的輸出，即輸入訊號向前往各層傳遞，最後在輸出層得到輸出，對於每個輸出神經元，其實際輸出值與期望輸出值作比較，而產生一誤差函數。
2. 向後傳遞 (back propagation)：使用梯度下降法 (gradient descent) 來找尋誤差函數的局部最小值 (local minimum)。由後往前調整各神經元的權重，權重的變化量爲誤差函數對該神經元權重之偏微分乘上一學習速率，意即將誤差值往各層後傳，以縮小輸出值與實際輸出值之誤差，而使誤差函數到達局部最小值。

以上稱爲一個學習循環 (learning cycle)，一個網路架構可以將訓練資料反覆數個學習循環，直到網路收斂爲止，但不宜訓練太久，以免過度符合訓練資料的特性，導致測試資料輸入時無法正確推論。

分類迴歸樹 (classification and regression tree, CART)

分類迴歸樹是一種二元的資料分割方法，最後會歸納出一套“if-then”的規則，結果以二元樹 (binary tree) 呈現。遞迴的分割方式是從樹根 (root) 開始，在樹的每一個節點 (node) 都將資料分爲兩個無交集的子集合。任意一個節點 m 都可依不同的指標訂出不純度 (impurity) $Q(m)$ ，不同的指標如下

分類錯誤率 (misclassification error)：

$$1-\hat{p}_k \tag{7.16}$$

吉尼指標 (Gini index)：

$$\sum_k p_k(1-p_k) \tag{7.17}$$

熵：

$$\sum_k p_k \log p_k \tag{7.18}$$

p_k 爲資料屬於 k 類別的比例， $0 \le p_k \le 1$， $\hat{p}_k = \arg\max_k p_k$ 。

在某一個節點要對資料進行分割時會先檢查節點中的資料是否屬於同一個類別，若是，則此節點不需在分割，若此節點中有兩個以上的類別時，CART 會測試所有的變數，依照不同的變數值將資料分成兩個子集合，並計算分割後兩個子節點的不純度，最後根據以下準則來分割

$$\max\left(Q(m)-(Q(R1)+Q(R2))\right) \tag{7.19}$$

$Q(m)$ 是原有節點的不純度，$Q(R1)+Q(R2)$ 爲分割後兩個子節點的不純度之和，意即我們希望分割後的子節點越純越好，而每一個節點的的不純度減少同時也帶表整體分類迴歸樹的不純度的減少，其分類迴歸的能力會越好。找尋分割條件使樹的生長持續進行，直到節點無法再找到下一個分割條件使不純度降低，或節點內的所有資料均屬於同一類別爲止。但是樹太大會有過度配適 (overfitting) 的問題，所以建構完成後通常會對樹進行修剪，修剪準則 (pruning criteria) 如最小成本複雜度 (minimum cost complexity)

$$\min\left(\sum_{m=1}^{|T|}Q(m)+\alpha|T|\right) \tag{7.20}$$

$Q(m)$ 是第 m 個樹葉節點 (leaf node) 的不純度，$|T|$ 是樹葉節點的個數，α 是某個設定參數。

參考文獻

[1] 郭宇豪, "圖形化高斯模型應用於自動化生產資料之關聯性分析," 碩士論文, 電機與控制工程學系, 國立交通大學, 民國九十三年.

[2] H. Abdi, "Factor Rotations in Factor Analyses," *Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA*, pp. 792-795, 2003.

[3] P. E. Meyer, *et al.*, "minet: A r/bioconductor package for inferring large transcriptional networks using mutual information," *BMC bioinformatics*, vol. 9, p. 461, 2008.

[4] W. Zhao, *et al.*, "Inferring Connectivity of GeneticRegulatory Networks Using Information-Theoretic Criteria," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 262-274, 2007.

[5] C. Hsun-Hsien and R. Marco, "Transcriptional Network Cassifiers," *BMC bioinformatics*, vol. 10.

[6] A. Margolin, *et al.*, "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context," *BMC bioinformatics*, vol. 7, p. S7, 2006.

[7] P. E. Meyer, *et al.*, "Information-Theoretic Inference of Large Transcriptional Regulatory Networks," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, pp. 8-8, 2007.

[8] U. Von Luxburg, "A Tutorial on Spectral Clustering," *Statistics and Computing*, vol. 17, pp. 395-416, 2007.

[9] L. Hagen and A. B. Kahng, "New Spectral Methods for Ratio Cut Partitioning

and Clustering," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,* vol. 11, pp. 1074-1085, 1992.

[10] L. C. Freeman, "Centrality in Social Networks Conceptual Clarification," *Social networks,* vol. 1, pp. 215-239, 1979.

[11] 施昱安, "給定資料不同損失涵式的提升演算法," 碩士論文, 電機與控制工程學系, 國立交通大學, 民國九十二年.

[12] 羅華強, *類神經網路-MATLAB 的應用*: 清蔚科技, 2001.