國立交通大學

工業工程與管理學系

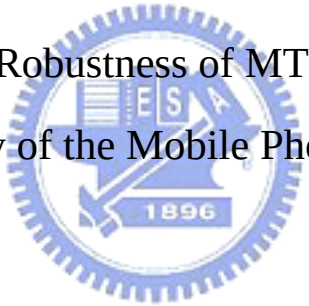
碩士論文

應用 MTS 於非平衡資料分析之穩健性研究

—以行動電話檢測流程為例

An Evaluation of the Robustness of MTS for Imbalanced Data

—A Case Study of the Mobile Phone Test Process



研 究 生：蕭宇翔

指導教授：蘇朝墩 教授

沙永傑 教授

中華民國九十四年五月

國立交通大學

工業工程與管理學系

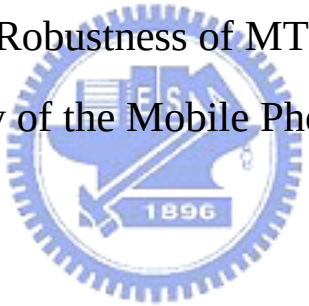
碩士論文

應用 MTS 於非平衡資料分析之穩健性研究

—以行動電話檢測流程為例

An Evaluation of the Robustness of MTS for Imbalanced Data

—A Case Study of the Mobile Phone Test Process



研 究 生：蕭宇翔

指導教授：蘇朝墩 教授

沙永傑 教授

中華民國九十四年五月

應用 MTS 於非平衡資料分析之穩健性研究

—以行動電話檢測流程為例

An Evaluation of the Robustness of MTS for Imbalanced Data

—A Case Study of the Mobile Phone Test Process

研 究 生：蕭宇翔　　Student: Yu-Hsiang Hsiao

指導教授：蘇朝墩 教授　　Advisor: Prof. Chao-Ton Su

沙永傑 教授　　Prof. David Yung-Jye Sha

國立交通大學

工業工程與管理學系

碩士論文

A Thesis

Submitted to Department of Industrial Engineering and Management
College of Management
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of Master Science
in
Industrial Engineering
May 2005
Hsinchu, Taiwana

中華民國九十四年五月

應用 MTS 於非平衡資料分析之穩健性研究
—以行動電話檢測流程為例

研 究 生：蕭宇翔　　指導教授：蘇朝墩 教授

沙永傑 教授

國立交通大學工業工程與管理學系碩士班

摘要

分類為資料探勘的主要任務之一，在分類模型建構過程中亦時常融合特徵選取，藉以提高分類效率。就二元分類問題而言，分析資料的類別數量比例通常是影響分類法能否正確學習分類模型的因素之一。我們稱一組在類別數量上呈現差距的資料為非平衡資料，此差距將可能導致分類模型學習過程發生偏差，並降低未來在少量類別上的判別敏感度，而這樣的情形並不容許於現實的應用環境中。MTS 為田口玄一博士針對多變量資料所提出的診斷與預測新技術，相異於其它分類法，MTS 在分類模型的建構過程是透過量測尺度的建立，而非對分析資料的學習，因此較不受資料分佈型態的影響。本研究以 MTS 及若干分類法對非平衡資料進行分類縮減模型的建構與類別預測，結果發現，MTS 在處裡非平衡資料的分類問題上確實有較穩健、出色的結果。此外，本研究亦根據柴比雪夫定理提出機率閾值來作為 MTS 的分類依據，並且有不錯的表現。最後，以台灣某高科技公司的行動電話 RF 檢測流程為研究對象，該流程所呈現之資料即為非平衡型態，透過 MTS 分析，所得結果顯著地減少原有的測試屬性，並仍然保有高檢測正確性。

關鍵字：資料探勘、分類、特徵選取、非平衡資料、馬氏-田口系統、MTS、機率閾值、行動電話

An Evaluation of the Robustness of MTS for Imbalanced Data — A Case Study of Mobile Phone Test Process

Student: Yu-Hsiang Hsiao Advisor: Prof. Chao-Ton Su

Prof. David Yung-Jye Sha

Department of Industrial Engineering and Management

National Chiao Tung University

Abstract

Classification is one of the main tasks of data mining. To execute classification efficiently, feature selection is usually merged into establishing a classification model. In binary classification problems, the ratio of the number of examples belonging to two classes in training data set is an important factor that impacts the effective learning of the classification model. If a data set contains several examples from one class and few examples from the other, we call it imbalanced data. There will be bias in the classification model that is learned from imbalanced training data set and this will result in lower sensitivity of detecting the class which has few examples in training data set. MTS is a new diagnosis and forecasting technique for multivariate data. MTS establishes a classification model by constructing a continuous measurement scale rather than learning from training data set. Therefore, MTS is not influenced by data distribution. This study compared MTS with other classification techniques and found that MTS is an outperforming and robust technique for imbalanced data. In addition, this study proposed a probabilistic threshold according to Chebyshev's theorem for MTS and probabilistic threshold derives good classification performance. Finally, MTS was employed to analyze the RF test process in mobile phone manufacture. The data coming from RF test process is typically imbalanced type. Implementation results showed that the test attributes have been significantly reduced and RF test process could also maintain high inspection accuracy.

Keyword: data mining, classification, feature selection, imbalanced data, MTS, Mahalanobis-Taguchi System, probabilistic threshold, mobile phone

致 謝

轉眼間，兩年的碩士班生涯就要結束了，一份喜悅與沈重摻雜的情緒不禁湧上心頭。喜悅的是，隨著碩士論文的提交，代表了這兩年多來的努力將要劃上一道完整且充實的休止符；沈重的是，結束了碩士的修業課程，代表著人生新一階段的開始，我們必須更加懷著戰戰兢兢、如履薄冰的心情迎接未來的挑戰。

在此，感謝父母親一路辛苦栽培，並默默地為我豎立起最堅實的一道後盾，讓我得以在父母親溫暖、舒適的臂膀中，無慮地全心投入學業。如今順利至碩士班畢業，我想，這不但是對自己負責，而更是對父母親最有效的回報與安慰了！

此外，最感謝的人莫過於恩師：蘇朝墩教授，學生於這兩年中，無論在學業或生活方面皆受蘇教授指導甚多，不僅從教授身上學得從事研究工作所應有的專業知識與態度，並更深刻感受到其嚴謹與負責的處事作風，使學生在學業或人生的道路上皆有所啟發並獲益良多。感激之情，溢於言表！

當然，忘不了實驗室裡的成員，許志華學長、許俊欽學長、王慧君學姊、陳隆昇學長、楊健炘學長、林敬森學長、周家任學長及彭加景同學，感謝你們平時在生活及課業上的照顧與提攜，甚至在畢業後仍然給予我相當大的協助，使我在這兩年內順利修必課程並完成碩士論文，雖然平時與你們打打鬧鬧，但在此我要由衷的對你們說聲謝謝，我會永記這份情誼！

最後，謝謝姿蓉不斷給予我精神上的鼓勵，以及弟弟宇惟、妹妹淳勻的貼心關懷！我要將這份畢業的喜悅化為滿心的感謝，獻給周遭所有的人，感謝你們為我所做的付出！

宇翔 謹誌

民國九十四年五月 於 交通大學

目　錄

中文摘要 iii

英文摘要 iv

誌謝 v

目錄 vi

表目錄 ix

圖目錄 xi

第一章 緒論 1

1.1 研究背景及動機 1

1.2 研究目的 3

1.3 論文架構 4

第二章 相關研究 5

2.1 資料庫中的知識發現 5

2.2 資料探勘 6

2.2.1 探勘任務 7

2.2.2 探勘技術 9

2.3 資料探勘中的分類任務 12

2.3.1 分類過程 12

2.3.2 分類技術 13

2.3.3 分類法評估 18

第三章 馬氏-田口系統 19

3.1 MTS 之構成要素 19

3.1.1 多變量診斷系統 19

3.1.2 馬氏距離..... 20
3.1.3 田口之穩健工程..... 21
3.2 MTS 方法論..... 22
3.2.1 馬氏距離分類法..... 22
3.2.2 MTS 閾值..... 24
3.2.3 特徵變數之篩選..... 26
3.3 MTS 之執行步驟..... 30
3.4 MTS 之特點..... 33

第四章 馬氏-田口系統之穩健性評估..... 35
4.1 威斯康辛乳癌..... 36
4.1.1 MTS 分析結果..... 36
4.1.2 SDA 分析結果..... 38
4.1.3 DT 分析結果..... 38
4.1.4 BPN 分析結果..... 39
4.1.5 結論..... 40
4.2 英文字母辨識..... 41
4.2.1 MTS 分析結果..... 42
4.2.2 SDA 分析結果..... 43
4.2.3 DT 分析結果..... 44
4.2.4 BPN 分析結果..... 45
4.2.5 結論..... 46
4.3 心臟病..... 47
4.3.1 MTS 分析結果..... 48
4.3.2 SDA 分析結果..... 49
4.3.3 DT 分析結果..... 50
4.3.4 BPN 分析結果..... 51
4.3.5 結論..... 52
4.4 評估結論..... 53

第五章 實例研究..... 55
5.1 案例描述..... 55
5.2 MTS 之執行..... 58
5.3 改善效益..... 65
5.4 分類法比較..... 65
5.4.1 MTS 分析結果..... 66
5.4.2 SDA 分析結果..... 67
5.4.3 DT 分析結果..... 68
5.4.4 BPN 分析結果..... 69
5.4.5 分類法比較結果..... 70

第六章 結論..... 72
6.1 研究結論與貢獻..... 72
6.2 未來研究建議..... 73

參考文獻 74

表　目　錄

表 3.1　L_{12}直交表配置 27

表 4.1　訓練及測試集之數量分佈（威斯康辛乳癌） 36

表 4.2　MTS 特徵選取結果（威斯康辛乳癌） 37

表 4.3　不同閾值下之 MTS 縮減模型分類結果（威斯康辛乳癌） 37

表 4.4　MTS 縮減模型分類結果之比較（威斯康辛乳癌） 37

表 4.5　SDA 特徵選取結果（威斯康辛乳癌） 38

表 4.6　SDA 縮減模型分類結果之比較（威斯康辛乳癌） 38

表 4.7　DT 特徵選取結果（威斯康辛乳癌） 39

表 4.8　DT 縮減模型分類結果之比較（威斯康辛乳癌） 39

表 4.9　BPN 之最佳網路架構（威斯康辛乳癌） 39

表 4.10　BPN 特徵選取結果（威斯康辛乳癌） 40

表 4.11　BPN 縮減模型分類結果之比較（威斯康辛乳癌） 40

表 4.12　訓練及測試集之數量分佈（英文字母辨識） 42

表 4.13　MTS 特徵選取結果（英文字母辨識） 42

表 4.14　不同閾值下之 MTS 縮減模型分類結果（英文字母辨識） 43

表 4.15　MTS 縮減模型分類結果之比較（英文字母辨識） 43

表 4.16　SDA 特徵選取結果（英文字母辨識） 44

表 4.17　SDA 縮減模型分類結果之比較（英文字母辨識） 44

表 4.18　DT 特徵選取結果（英文字母辨識） 45

表 4.19　DT 縮減模型分類結果之比較（英文字母辨識） 45

表 4.20　BPN 之最佳網路架構（英文字母辨識） 45

表 4.21　BPN 特徵選取結果（英文字母辨識） 46

表 4.22　BPN 縮減模型分類結果之比較（英文字母辨識） 46

表 4.23　訓練及測試集之數量分佈（心臟病） 48

表 4.24　MTS 特徵選取結果（心臟病） 49

表 4.25　不同閾值下之 MTS 縮減模型分類結果（心臟病） 49

表 4.26　MTS 縮減模型分類結果之比較（心臟病） 49

表 4.27 SDA 特徵選取結果（心臟病）..... 50
表 4.28 SDA 縮減模型分類結果之比較（心臟病）..... 50
表 4.29 DT 特徵選取結果（心臟病）..... 50
表 4.30 DT 縮減模型分類結果之比較（心臟病）..... 51
表 4.31 BPN 之最佳網路架構（心臟病）..... 51
表 4.32 BPN 特徵選取結果（心臟病）..... 52
表 4.33 BPN 縮減模型分類結果之比較（心臟病）..... 52
表 5.1 RF 功能檢測項目..... 56
表 5.2 RF 功能檢測屬性..... 57
表 5.3 樣本資料（RF 功能檢測）..... 58
表 5.4 訓練集之正常樣本原始數據（RF 功能檢測）..... 58
表 5.5 訓練集之正常樣本標準化數據及馬氏距離（RF 功能檢測）... 59
表 5.6 訓練集之正常樣本相關反矩陣（RF 功能檢測）..... 59
表 5.7 直交表配置與 SN 比（RF 功能檢測）..... 61
表 5.8 特徵選取結果（RF 功能檢測）..... 63
表 5.9 系統改善結果（RF 功能檢測）..... 63
表 5.10 效果增量之結果比較（RF 功能檢測）..... 64
表 5.11 訓練及測試集之數量分佈（RF 功能檢測）..... 66
表 5.12 MTS 特徵選取結果（RF 功能檢測）..... 66
表 5.13 不同閾值下之 MTS 縮減模型分類結果（RF 功能檢測）..... 67
表 5.14 MTS 縮減模型分類結果之比較（RF 功能檢測）..... 67
表 5.15 SDA 特徵選取結果（RF 功能檢測）..... 68
表 5.16 SDA 縮減模型分類結果之比較（RF 功能檢測）..... 68
表 5.17 SDA 特徵選取結果（RF 功能檢測）..... 68
表 5.18 SDA 縮減模型分類結果之比較（RF 功能檢測）..... 69
表 5.19 BPN 之最佳網路架構（RF 功能檢測）..... 69
表 5.20 BPN 特徵選取結果（RF 功能檢測）..... 69
表 5.21 BPN 縮減模型分類結果之比較（RF 功能檢測）..... 70

圖　目　錄

圖 2.1　KDD 流程…… 6
圖 2.2　貝式網路 [7]…... 9
圖 2.3　分類任務流程圖…... 13
圖 2.4　決策樹 [12]…... 15
圖 2.5　多層前饋類神經網路…... 16
圖 3.1　多變量診斷系統 [2]…... 19
圖 3.2　馬氏距離與歐氏距離 [2]…... 20
圖 3.3　修改後的多變量診斷系統…... 21
圖 3.4　正常樣本與異常樣本之馬氏距離分配…... 24
圖 3.5　機率閾值訂定步驟一…... 25
圖 3.6　機率閾值訂定步驟二…... 25
圖 3.7　MTS 流程圖…... 32
圖 4.1　各分類法測試總準確率比較（威斯康辛乳癌）…... 41
圖 4.2　各分類法測試相對敏感度比較（威斯康辛乳癌）…... 41
圖 4.3　各分類法測試總準確率比較（英文字母辨識）…... 47
圖 4.4　各分類法測試相對敏感度比較（英文字母辨識）…... 47
圖 4.5　各分類法測試總準確率比較（心臟病）…... 53
圖 4.6　各分類法測試相對敏感度比較（心臟病）…... 53
圖 5.1　行動電話製造流程…... 56
圖 5.2　訓練樣本之馬氏距離分布圖（RF 功能檢測之完整模型）…... 60
圖 5.3　測試樣本之馬氏距離分配圖（RF 功能檢測之完整模型）…... 60
圖 5.4　屬性效果增量圖（RF 功能檢測）…... 62
圖 5.5　訓練樣本之馬氏距離分布圖（RF 功能檢測之縮減模型）…... 63
圖 5.6　測試樣本之馬氏距離分布圖（RF 功能檢測之縮減模型）…... 64
圖 5.7　各分類法測試總準確率比較（RF 功能檢測）…... 71
圖 5.8　各分類法測試相對敏感度比較（RF 功能檢測）…... 71

第一章　緒論

1.1　研究背景及動機

拜科技進步之賜，一個現代化組織相較於過去大大的增加了對資料取得、蒐集與整理的能力，然而，如何有效的從資料堆中獲取有用的知識，擺脫「空有大量資料，卻資訊匱乏」的窘境，才是所當關心的課題。為此，多數的組織藉由資料探勘（data mining）技術，來萃取出潛藏於雜亂資料中的各種有益資訊，並作為決策上的支援。

分類（classification）或類別預測為資料探勘領域的主要任務之一，用來萃取描述多變量資料（multivariate data）的類別模型，以便能夠預測類別標記未知的物件類別，有關這方面的實務相關應用相當多，例如：產品測試、聲音辨識、疾病診斷、字體辨識、信用評比等。二元分類問題（binary classification problems）是屬於分類問題中的一部分，其關注於二類別資料的分析。此外，在分類模型萃取過程裡，通常融合了維度縮減（dimension reduction）技術，以相對具有關鍵性影響及較原始資料少的特徵變數來建立分類模型，如此可減少資料收集成本、提升分類的效率，這樣的過程在機器學習（machine learning）中稱之為特徵選取（feature selection）。相對於原始的完整模型（full model），經由特徵選取後所建構者，稱為縮減模型（reduced model）。目前，若干的統計、數學、機器學習（machine learning）及人工智慧（artificial intelligence）法皆可用於解決二元分類問題，並達成特徵選取，常用的如：決策樹（decision tree, DT）、類神經網路（artificial neural networks, ANN）、逐步判別分析（stepwise discriminant analysis, SDA）等。然而，決策樹演算法雖然在相對小的資料集分析上相當有效，但應用於現實世界中，往往因為資料量或特徵變數的增大而增加了分析時的繁瑣；對類神經網路而言，由於整個網路猶如「黑箱」作業，因此常因其解釋性差而受到批評；應用統計方法

於現實資料時，往往因為資料型態不完全符合統計分析上所需的一般假設或要求，因而必須對原始資料進行處理或轉換，也因此徒增資料分析上的複雜度。

另外，分類方法，如：決策樹、類神經網路、逐步判別分析等，是在訓練樣本（training samples）上「學習」出可區分類別的模型。然而，在二元分類裡的訓練資料集中，屬於類別 1 和類別 2 的資料數量比例通常是影響這些分類方法是否能有效學習的因素之一，如果訓練資料裡，屬於類別 1 的數量相當多，而類別 2 的數量相當少，則此種稱之為「非平衡（imbalance）」的資料型態將使學習過程產生偏差，而減低未來在判斷類別 2 上的敏感度（sensitivity），因此導致分類方法的可靠度降低 [8, 9]。基於這樣的理由，如果所要探勘的資料集在兩類別數量間呈現懸殊的差異時，使用決策樹、類神經網路和逐步判別分析是不適當的。但，值得注意的，在相當多的領域裡，由於其特性使然，使得所收集到的資料正好呈現以上情形，例如：在醫學疾病診斷上，患有某病症的病患通常僅為健檢者中的少數，其餘則皆為健康者；在量產產品的品質測試上，不良品也只佔所有成品的極少部分，尤其在現今講求六標準差（Six Sigma）品質的高科技產業中更顯如此。然而，正確診斷出這些少數的疾病患者、不良品，對一個類別預測系統而言卻是較為緊要的，唯有透過正確的診斷，才能對疾病患者採取適當的醫療行動；才能避免不良商品流入市場。因此，一個類別預測模型若因為不平衡的訓練資料而降低對這些「關鍵少數」的預測敏感度，將會造成無法預期的損失。於是，在每個方法皆宣稱自己的分類能力的同時，一個真正好的、能為使用者所信任的方法，必然不因類別分布型態而左右其類別預測能力，因此，我們需要的是一套穩健的方法，即在面對訓練資料的類別分布不平衡時，甚至是數量差距懸殊時，仍然能建構出高總準確率及高敏感度的類別預測模型，而這樣的方法也更能符合實務應用上對簡單、快速、不受限的期待，並獲得青睞。

馬氏-田口系統（Mahalanobis-Taguchi System, MTS）是近年來由田口玄一（Genichi Taguchi）博士針對多變量資料所提出的診斷（diagnosis）與預測（forecasting)新技術,此法包含數學、統計之概念及穩健工程(robust engineering）的原理，結合了馬氏距離（Mhalanobis distance, MD)、直交表（orthogonal arrays, OAs)、SN 比（signal-to-noise ratio）[1, 2]。馬氏距離被用來建構多變量系統的測量尺度(measurement scale)，而直交表及 SN 比則被用於系統最佳化。雖然 MTS 同樣具有分類及特徵選取的能力,但其可更進一步的利用馬氏距離來衡量觀察樣本相對於參照群體(reference group)的異常程度，而不僅限於一般的分類功能。田口博士認為，MTS 並非基於機率上的推論方法，而是一種資料解析，因此沒有任何的假設限制或分配理論 [1]。此外，在建模過程中，MTS 並非透過對整體訓練資料的「學習」來萃取分類模型或規則,而是僅利用事先定義的正常(normal)類別作為參照群體，並建立量測尺度的基準，新樣本則以距離基準點的遠近來判斷其類別。因此，訓練資料的類別分布所呈現的型態，對於 MTS 建立分類模型過程的影響是較小的,這將大幅的提升 MTS 在實務上的適用性。不論如何,MTS 在田口博士的鼓吹下，日漸受到重視，然而，其方法論與相關之應用仍有待進一步的探討。

1.2 研究目的

本研究目的包含兩個層面。在方法論層面上，主要針對多變量資料的二元分類問題，探討非平衡的訓練資料對 MTS 在類別預測上的影響，並和決策樹、類神經網路、逐步判別分析比較之。藉由評估各方法所建立的縮減模型（reduced model）在分類預測上的準確率（accuracy）、敏感度（sensitivity）及特效性（specificity）指標，來了解這些分類技術對於處理非平衡資料的能力，並期望能突顯 MTS 的穩健性及在現今高科技環境中的適用性。此外，本研究提出以柴比雪夫定理（Chebyshev's theorem）為基礎的機率閾值（probabilistic threshold)，

作為 MTS 在類別判斷時的依據。

在 MTS 應用層面上，以台灣某高科技公司的行動電話檢測製程為研究案例，希望藉由 MTS 及工程上的背景知識，來改善行動電話檢測製程中，最為耗時、繁雜的無線頻率功能檢測流程（radio frequency functional test process）。本個案研究期望能夠在保有良品與不良品的判別準確率下，縮減測試流程中冗餘的測試項目，以提供一個更符合經濟效益的測試流程，藉此減少生產工時，提升生產能量，也因此將更能適應手機市場上快速求新與求變的需求環境，並提升公司在產業中的競爭力。

1.3 論文架構

本論文包含六章。第一章緒論，介紹本研究之背景、動機及目的。第二章對資料探勘中的任務及技術等相關領域進行概述，並簡單介紹本研究所用到的分類法。第三章介紹 MTS 的概念及方法，其中包含本研究所提出的機率閾值訂定步驟。第四章以 UCI 資料庫（data bank）上的資料為例，探討訓練資料的不平衡對 MTS 及其他方法在建構類別預測模型上的影響，並比較之；同時亦驗證本研究所提出的機率閾值是否有較好的判別能力。第五章收集台灣某高科技公司的行動電話檢測數據，依 MTS 的架構進行分析，改善測試流程。第六章說明本研究的結論與未來研究建議。

第二章　相關研究

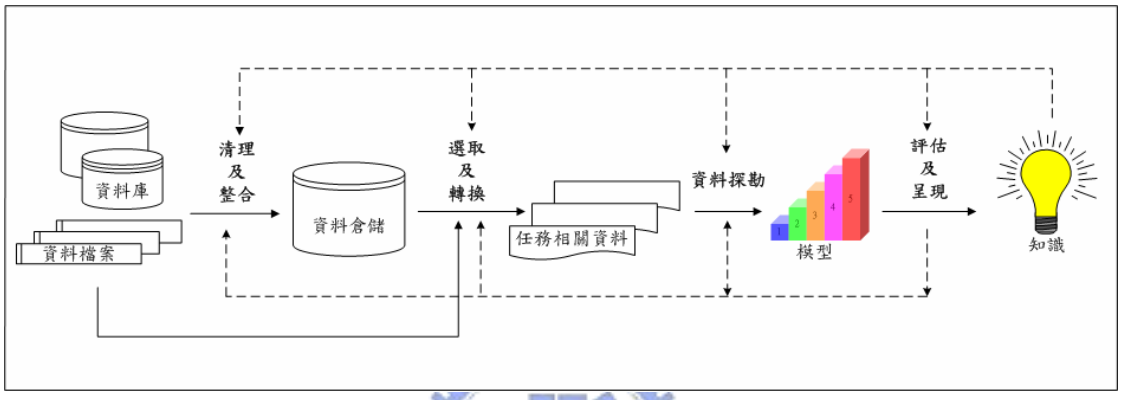
2.1 資料庫中的知識發現

由於資料儲存及管理技術的發展，大量資料的取得與儲存對現在的科學、商業、工業或其他領域而言並非難事，然而，人們對於資料的處理及理解能力卻遠落後於資料的收集速度。因此，豐富的資料將勢必引發對強而有力之資料分析的需求。

資料庫中的知識發現（knowledge discovery in database, KDD）是一個在資料中探尋有效的、新穎的、潛藏有用的，並且最終可以被理解的模式之有價值的過程 [10]。它是一門交叉性學科，涉及了資料庫（databases）、機器學習（machine learning）、模式識別（pattern recognition）、統計學（statistics）、人工智慧（artificial intelligence）、不確定性推論（reasoning with uncertainty）、專家系統（expert system）、資料形象化（data visualization）、機器發現（machine discovery）、科學發現（scientific discovery）、資訊檢索（information retrieval）及高效能運算（high-performance computing）等各領域的技術 [11]。KDD 過程如圖 2.1 所示，並簡述如下 [7]：

1. 首先在事前必須針對此次任務的應用範疇、研究對象、相關背景知識，及最後欲達成的目標進行通盤而細密的了解。
2. 資料清理：對各來源資料庫、資料文件填補遺失值、平滑雜亂資料、辨識孤立點，並解決資料不一致的情形。
3. 資料整合：假設欲分析的資料出自多個來源及形式，則便涉及到整合多個資料庫或資料文件於統一的資料儲存中。
4. 資料選取：從整合後的資料庫或資料倉儲中取得與任務相關 (task-relevant data) 的資料。

5. 資料轉換：將選取資料轉換或統一成適合探勘應用的形式，如：資料廣義化（generalization）、資料標準化（normalization）、資料壓縮（compression）。
6. 資料探勘：包含欲探勘的知識類型、探勘方法的選擇及執行。
7. 模型評估：衡量資料探勘後所發現的模型或知識是否能夠真正帶來貢獻。
8. 知識呈現：使用視覺化和知識表達技術，向使用者或決策者呈現探勘結果。



2.1 KDD 流程

KDD 自從 1989 年 8 月在美國底特律召開的第 11 屆人工智能聯合會議的議題討論會裡首次被提出以來，由於它在資料處理、分析並挖掘知識上的顯著能力，及在資訊管理、決策支援、錯誤偵測、健康照護、市場策略、財務預測、流程控制等方面的廣泛應用，使得這些年來無論學術界或業界紛紛投以對 KDD 流程的研究與新技術的開發，並推動此領域不斷發展。

2.2 資料探勘

資料探勘是 KDD 過程的核心，牽涉到從已觀察的資料來推斷隱藏在大量資料背後的知識。從資料分析的角度而言，資料探勘可分為描述式（descriptive）和預測式（predictive）。描述式資料探勘以簡潔概要的方式描述資料，並提供資料的一般性質。預測式資料探勘則藉由分析資料，並建立一個或一組模型，來試圖預測新資料集的行為。因此，基於我們所欲探尋的模型及應用上的需求，資料

探勘任務主要可區分為:概念描述(concept description)、分類與預測(classification and prediction)、叢集（clustering）、關聯（association）和演化分析（evolution analysis）；而執行這些任務的探勘技術領域則包含：統計方法（statistical approaches）、機器學習法（machine learning approaches）、資料庫導向法（database-oriented approaches）及其它若干方法 [7, 10, 11, 12, 13, 16]。

2.2.1 探勘任務

1. 概念描述（concept description）

概念描述包含資料特徵化（characterization）與比較（comparison）。資料庫通常存放著大量的細節資料，然而，使用者通常希望以簡潔的描述形式來觀察匯總的資料集，這種資料描述可以提供目標類別的大致面貌，例如：某年利用信用卡消費 100000 元以上之客戶的特徵匯總描述，其特徵為年齡介於 30 至 40 歲、年收入 100 萬以上、高學歷、職銜為主管等。或者能將它與對比類別相區別開來，例如：比較某年以信用卡消費 100000 元以上及 10000 元以下之客戶的一般特徵。此外，使用者亦可方便、靈活地以不同細微度(granularity)和不同的角度來描述資料集，例如：在高學歷屬性上，觀察具有碩士學位與博士學位的客戶。

2. 分類與預測（classification and prediction）

分類是針對類別標籤作預測，為監督式學習（supervised learning）問題，即仰賴已知類別的訓練樣本來探索輸入屬性及輸出類別間的關係，以建構出類別預測模型，並用來預測新未來樣本的類別，屬於示例式學習（learning by examples）。預測則是建立連續值函數模型，可被視為建構和使用模型來評估給定樣本可能具有的屬性值或值域，例如，依公司、職位、服務年資等，來預測某人的薪資收入。

3. 叢集（clustering）

叢集是在類別未知的情況下，將實體或抽象物件的集合分組成由類似物件為一群的過程。不同於分類，類別標記並不會出現在訓練樣本中，因此，叢集為非監督式學習（unsupervised learning）的一個範例，即不依賴欲先定義的類別和已知類別的訓練樣本來建構預測模型，是屬於觀察式學習（learning by observations）。由叢集所生成的群體是一組資料物件的集合，這些物件與同一群體中的物件彼此有高度的相似，而與其他群體中的物件相異。在許多應用中，可將一個叢集群體中的資料物件視為一體。例如：以叢集來劃分市場上的顧客群，叢集裡為需求特徵相似的顧客，可作為提供何種商品資訊的參考。

4. 關聯（association）

關聯是針對物件間的關係進行分析，這種物件間的關係稱為關聯規則。一個關聯規則透露物件間的存在關係，即在一資料庫中，一個物件集合的出現，可能強烈關係到另一個物件集合的出現與否。例如：在 3C 產品銷售中，透過關聯分析發現以下規則，20 %的消費者年齡在 20 至 29 歲、並且月收入介於 20000 至 30000，而在這樣的年齡及月收入組合下的消費者會購買 mp3 隨身聽的可能性有 60 %。

5. 演化分析（evolution analysis）

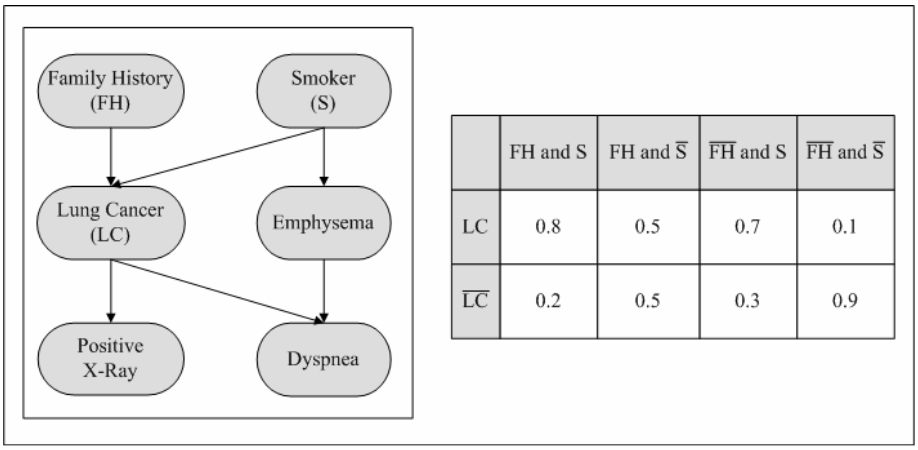
演化分析描述行為隨時間變化的物件之規律或趨勢，並對其塑模。這類型的分析包括有趨勢分析（trend analysis）、時間序列分析（time-series analysis）和週期分析（periodicity analysis）。例如：根據股票市場過去幾年的主要股票資料，透過演化分析可以辨識整個股票市場和特定公司的股票演化規律，而這種規律可以幫助預測股票價格的未來走向，並作為投資決策支援。

2.2.2 探勘技術

1. 統計方法（statistical approaches）

許多的統計方法被應用於資料探勘，包含貝氏網路（Bayesian network）、迴歸分析（regression analysis）、相關分析（correlation analysis）、叢集分析（cluster analysis）等。通常，統計方法利用訓練資料集來建構統計模型，並在假設空間裡搜尋統計衡量指標上的最佳者。

貝氏網路可用來預測某條件下事件發生的機率，其由兩部分定義，第一部分是有向圖（directed graph），其中的每個節點代表一個變數或狀態，每條弧代表一個機率性的依賴，第二部分是每個屬性事件發生的條件機率或聯合條件機率表，圖 2.2 顯示一個肺癌的貝式網路，及導致肺癌的條件機率表，網路中的節點表示變數或狀態，有向弧則代表因果關係。迴歸分析利用過去所觀察的資料來導出一個方程式，而此方程式可將物件的屬性集適配到一個輸出變數，例如：線性迴歸（linear regression）、判別分析（discriminant analysis）。相關分析是用來研究屬性或變數間的相互關係，例如：χ^2相關性檢定。叢集分析則是基於物件間的距離量測，來發現物件集裡的群體關係，例如：k-means、k-medoids。



	FH and S	FH and $\overline{S}$	$\overline{FH}$ and S	$\overline{FH}$ and $\overline{S}$
LC	0.8	0.5	0.7	0.1
$\overline{LC}$	0.2	0.5	0.3	0.9

圖 2.2 貝式網路 [7]

2. 機器學習法（machine learning approaches）

在資料探勘中，最常見的機器學習法包括決策樹歸納法（decision tree induction）、類神經網路（artificial neural networks）、歸納式學習（inductive concept learning）及概念式叢集（conceptual clustering）。

決策樹是一個分類樹，決策樹歸納法建構一個類似於流程圖的結構，其中每個內部節點表達一個屬性上的測試，演算法會在每個節點選擇最佳分類的屬性，並依其將資料劃分類別，每個分枝對應於測試的一個輸出值，每個外部節點（樹葉）則表達一個類別，而整棵樹即表示分類上的規則。類神經網路是仿生物神經網路的資訊處理系統，它由大量簡單的神經元（neuron），及介於神經元間的訊號傳遞連結（connection）所構成，人工神經元是生物神經元的簡單模擬，它從外界環境或其它人工神經元取得資訊，透過非常簡單的運算，再輸出其結果到外界環境或者其它人工神經元，藉由類神經網路可建構輸入屬性與輸出屬性間的關係。

歸納式學習是從一些資料進行分析，並歸納出具有一般性的概念，此法無須先前的知識，端看所給的訓練資料是否足夠用來學習和歸納分析；簡言之，是一種由學習主體對學習客體做歸納的學習方式，對某一個概念，由施教者提供適當份量的訓練資料，並且告知受教者該資料的正確值，使受教者得到一系列的「輸入－輸出」序對，此時受教者便得以建立或逐步修正自己對此觀念的認知模型，藉由反覆進行這些步驟，使受教者的認知模型逐步趨近於正確。概念叢集與傳統的叢集不同，它是一個兩步驟的過程，首先確定相似物件的分群，接著為每群對象發現了特徵描述，也就是每群物件代表了一個概念或類別。

3. 資料庫導向法（database-oriented approaches）

不同於上述的兩項領域，資料庫導向法並不搜尋最佳模型，而是利用資料塑模或特定的啟發式方法（heuristics）來發現手中資料的特性，如：屬性導向歸納法（attribute-oriented induction）、重複掃描資料庫中頻繁項目集合（iterative

database scanning for frequent item sets）等。

屬性導向歸納法是資料庫查詢導向的（database query oriented）、基於廣義化的（generalization-based）線上資料分析處理技術，首先使用資料庫查詢收集任務相關資料，然後通過考察資料中每個屬性之不同值的個數來進行廣義化，使屬性值個數落在一定的範圍內，其中廣義化可透過屬性刪除或屬性廣義化來達成。重複掃描資料庫法被用來搜尋交易資料庫中的頻繁項目集合，而項目間的關聯規則就從這些頻繁項目集合裡推導出，如：常用於搜尋頻繁項目集合的 Apriori 演算法。

4. 其他技術

其它技術如：基因演算法（genetic algorithms）、粗略集合（rough sets）、模糊集合（fuzzy sets）、形象化（visualization）等，亦被應用於資料探勘。基因演算法試圖結合自然演算的想法，是利用電腦模擬染色體的基因結合、突變及自然選擇過程的最佳化技術。粗略集合理論可以用來近似地定義那些根據屬性無法區分的類別。模糊集合利用隸屬函數（membership function）的表達來取代以往對互斥事件的處理觀點。而形象化技術將資料轉化為點、線、面的形象物件，例如資料散佈圖（scatter plot）、3D圖等，使得資料分析者可以快速的發現資料的顯著特性。

目前，這些資料探勘技術時常被整合、結合來解決複雜的問題，或提供另一套解答方法，舉簡單例子：我們時常以圖、表等形象化技術來表現其他探勘技術的匯總；而屬性導向歸納法則時常作為關聯規則探勘前的資料屬性處理。這樣的做法不僅幫助探勘能力的提升，並更有助於未來的開發研究，因此愈來愈多的資料探勘系統試著融合多樣化的探勘技術，來處理不同的資料、不同的探勘任務，及不同的應用領域。

2.3 資料探勘中的分類任務

分類是針對類別標籤作預測。仰賴已知類別的訓練樣本來探索輸入屬性及輸出類別間的關係，以建構出類別預測模型，並用來預測未來樣本的類別。目前已有數學、統計學、機器學習、專家系統和神經生物學方面的研究學者提出許多分類方法，並在信用核證、醫療診斷、效能預測和選擇性行銷等方面有廣泛的應用。

2.3.1 分類過程

分類是一個兩步驟的過程，如圖 2.4。第一步驟，透過分析由屬性所描述的樣本或物件建立一個模型，用來描述預定的資料類別或概念集。為了建立模型而被分析的資料稱為訓練樣本（training samples），它是隨機地由任務相關資料中選出。由於提供了每個訓練樣本類別標籤，因此此步驟也稱為監督式學習（supervised learning），通常，經由訓練樣本學習出來的模型，是以分類規則、決策樹或數學公式的形式表達。例如，給定一個顧客信用資訊的資料庫，分類規則可被學習，並用來確認顧客的信譽是優良或普通，而這些規則可被用來為以後的資料作分類，同時也能增進對資料庫內容更好的理解。

第二步驟，我們使用步驟一所建構的模型來進行分類。由於學習模型往往會對資料有過度適配（overfitting）的現象，因此若使用訓練樣本來導出分類法，並評估正確性，可能會錯誤地導致太過樂觀的估計。此步驟中，測試樣本（test samples）將會被使用，這些樣本是隨機選取，並且與訓練樣本無關的。首先，將評估模型的預測準確率（accuracy），即對於每個測試樣本，我們將已知的類別標籤與該樣本透過學習模型所預測的類別作比較，而模型準確率是定義為被模型正確分類的測試樣本百分比。如果準確率是可以被接受的，則此模型就可以被用來對類別標籤未知的資料或物件進行分類。例如，透過分析現有顧客資料所學習

到的分類規則，在測試準確率可接受下，可以被利用來預測新的或未來顧客的信譽。

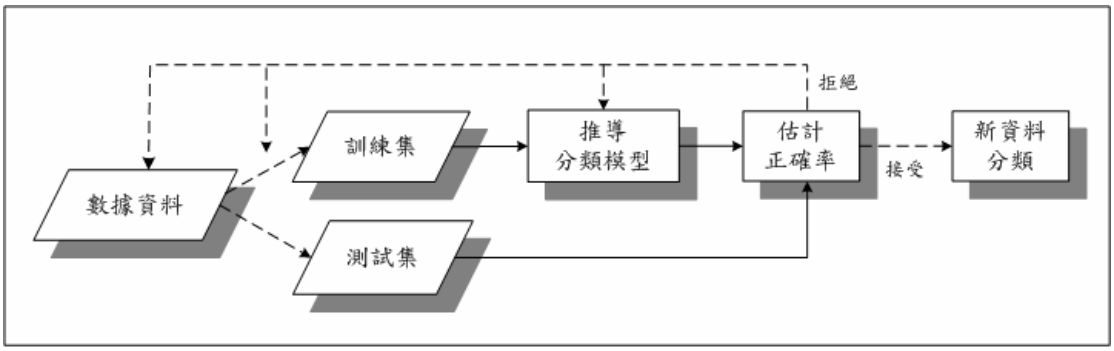


圖 2.3 分類任務流程圖

2.3.2 分類技術

至今，許多研究利用不同的演算方法及技術來處理多變量資料的分類問題，而基於先前探勘技術的討論，在此介紹幾個常見的分類技術：統計方法中的逐步判別分析、機器學習法中的決策樹歸納法，及類神經網路。

1. 判別分析（discriminant analysis）

判別分析為一個簡單的參數型統計方法，用在對多變量資料的分類。此法必須滿足三個統計上的假設：（1）群體必須滿足多元常態分配（multivariate normal distribution）；（2）群體的共變數矩陣（covarianc matrices）相等；（3）群體的平均值向量（mean vectors）、共變數矩陣及先驗機率（prior probabilities）已知。判別分析包含三程序 [14, 15]：

（1） 確認最佳變數集：找出可以用來清楚分辨訓練樣本類別的最佳變數，而這些變數稱為判定變數（discriminator variables）。

（2） 確認新座標軸：使用（1）中的判定變數，將訓練樣本的多個判定變數值投影到一新的座標軸上，使得到一個新變數，再根據這些訓練樣本的新變數，找出最能區分類別的座標軸。

（3） 分類：使用（2）所找出的座標軸發展分類規則，對新的樣本作分類。

由以上三個程序可以了解，判別分析經由所有判定變數的線性組合來發展一個新變數，而此線性組合必須能最大化新變數在組間與組內平方和上的比值（between-group to within-group sum of squares ratio, SSb/SSw），這將使在不同類別間具有最佳的判別。此線性組合稱為線性判別方程式（linear discriminant function），而經過判別方程式所計算的新變數值稱為判別得分（discriminant score），並構成判別空間（discriminant space）。最後，決定判別空間中最能區分類別的閾值，建立分類規則。

在此之前，我們假設判定變數的最佳集合是已知的，並且利用已知的判定變數來建立判別方程式，然而，實際情形並非如此。逐步判別分析（stepwise discriminant analysis, SDA）將所有變數依據指定的篩選指標，利用過濾的手法，個別逐步的從目前的判定變數集外導入較重要者，或從中剔除不必要者於判別方程式中，直到沒有變數滿足指標閾值為止，藉此尋求分類能力最佳的變數集合，作為最終的判定變數。

2. 決策樹（decision tree）

決策樹是一個類似於流程圖的樹狀結構，其中每個內部節點表示在某個屬性上的測試，最頂層的節點稱為根節點（root node），每個分枝則代表一個測試結果，而每個最底層的節點稱為樹葉（leaf），代表類別或類別分布（class distributions）。一個由根到葉的路徑存放著對樣本的分類規則，圖 2.4 即為一典型的決策樹，他表示了一部汽車的最高時速高低（high, medium, low）的分類規則，內部節點以橢圓形來表示，而矩形則代表類別樹葉。決策樹中每個節點上的屬性是使用基於熵理論（entropy theory）的信息增益（information gain）度量來決定，這種度量稱作屬性選擇度量（attribute selection measure）或分裂優良性度量（measure of the goodness of split）[7, 16, 19]。

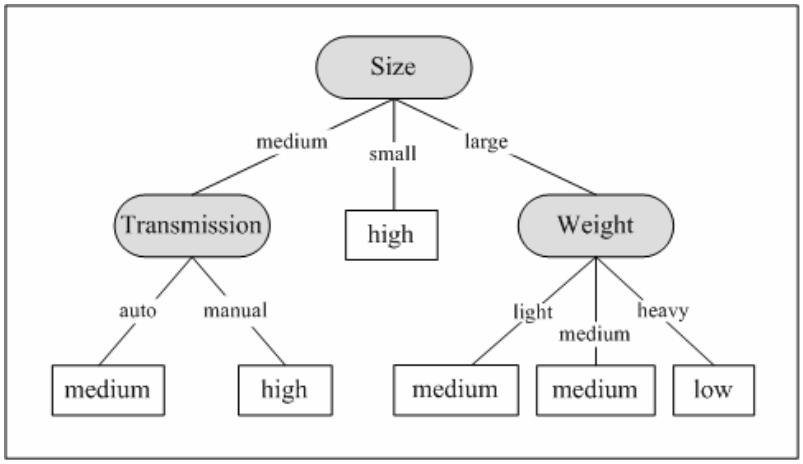


圖 2.4 決策樹 [12]

決策樹歸納法的基本演算步驟如下：

（1） 決策樹從代表所有訓練樣本的單個節點開始。

（2） 如果節點中的樣本皆屬於同一類別，則該節點成為決策樹葉，並標記類別。

（3） 如果節點中的樣本不屬於同一類別，演算法使用信息增益的度量作為資訊，並選擇使信息增益最大的屬性，做為該節點的測試或決策屬性。

（4） 對測試屬性的每個已知值各建立一個分枝，並藉此來劃分樣本，使進入下一層節點。

（5） 演算法使用同樣的過程，在每個分枝的樣本上，以剩餘的屬性，重複步驟（1）到（4），遞迴地形成決策樹的伸展。

（6） 重複遞迴的步驟在以下條件下即停止，並形成樹葉：

（a） 該節點的樣本皆屬於同一類別。

（b） 沒有剩餘的屬性可被用來進一步劃分樣本。通常，以該節點中的多數類別為決策樹葉的標記，或者以該節點中的類別分布表示。

當決策樹建立時，由於資料中的雜亂，使得許多分枝是反應訓練樣本中的異常資料。透過修剪（pruning），利用統計度量（statistical measures）減去不可靠的分枝，可以處理這種對訓練樣本過度適配的現象。執行決策樹修剪，將可獲得較快的分類，並提高決策樹的正確分類能力 [7]。另外，不出現在決策

樹中的所有屬性皆假設是不相關的，因此決策樹歸納法亦可用在特徵選取。

3. 類神經網路（artificial neural network）

類神經網路是一組連接的輸入/輸出單元，其中每個連接都與一個權重相關聯，在學習訓練階段，透過調整類神經網路的權重，使得能夠正確預測輸入樣本的類別來學習。多層前饋（multilayer feed-forward）類神經網路如圖 2.5 所示，由輸入層（input layer）輸入訓練樣本的每個屬性值，這些輸入單元的加權值依次同時地提供給第一隱藏層（hidden layer）作為隱藏單元的輸入，再透過一激發函數（activation function）的轉換後，產生隱藏單元的輸出值，該隱藏層的加權輸出，做為下一隱藏層的輸入，由此下去，最後一個隱藏層的輸出構成輸出層（output layer）的單元輸入，而輸出層的輸入值經過激發函數轉換後，發布給定樣本的網路預測。隱藏層和輸出層的單元稱為神經元（neuron），一個輸出單元可以用來表現兩個類別，如果多於兩個類別，則每個類別分別使用一個輸出單元。如果權重都不送回前一輸出單元，則此類神經網路是前饋的。網路的設計是一個試誤（try and error）的過程，對於最佳的隱藏層單元數決定，並沒有明確的規則可供使用，而權重的初始值也可能會影響分類結果的準確性，因此一但網路經過訓練，並且其分類準確率無法被接受時，通常會選用不同的隱藏層單元數或不同的初始權重，來重複此訓練過程。

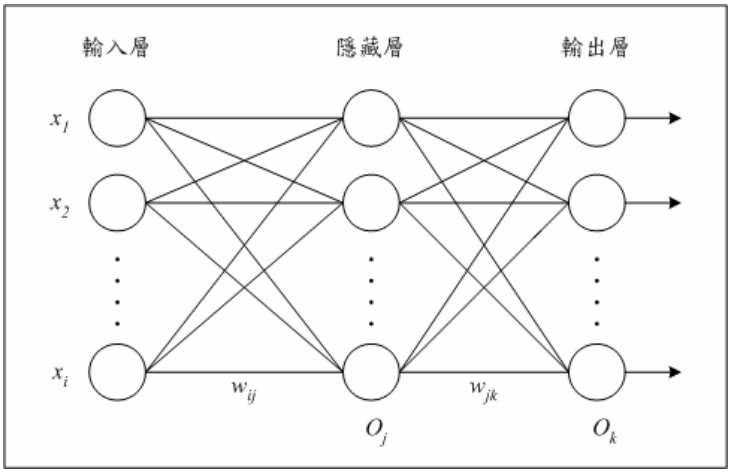


圖 2.5 多層前饋類神經網路

倒傳遞神經網路（back-propagation neural network, BPN），在多層前饋神經網路上學習，為目前最常使用於分類的類神經網路 [17, 18]。倒傳遞神經網路透過疊代的處理一組訓練樣本，將每個樣本的預測與實際知道的類別標籤做比較來進行學習。在訓練過程中，以 sigmoid 函數作為激發函數，並會透過不斷修正其權重，使網路預測和實際類別間的誤差最小，而這種權重修改是「倒向」進行的，也就是由輸出層到第一個隱藏層，因此稱為「倒傳遞」。一般而言，權重最終將會收斂，而學習過程即停止。

倒傳遞神經網路之基本演算步驟如下：

（1） 網路的權重首先會被初始化為很小的隨機變數。

（2） 由訓練樣本提供給網路的輸入層，向前以加權組合的 sigmoid 函數值逐層傳播輸入，並在輸出層輸出預測結果。

（3） 向後傳播預測結果與實際類別的誤差，並由輸出層逐層向後計算、更新權重。

（4） 重複（2）、（3）直到滿足停止條件。

（5） 停止條件如下：

（a） 權重收斂，新舊權重差距小於某個指定的閾值。

（b） 類別預測的正確率達到某個指定的閾值。

（c） 超過預先定的疊代次數。

此外，倒傳遞神經網路的特徵選取可由以下步驟執行：

（1） 在每個輸入單元，計算其所連接的每對「輸入-隱藏」權重和「隱藏-輸出」權重之絕對值乘積，並相加之。

（2） 將每個輸入單元在（1）所求得的值以遞減方式排序。排序愈後面，表示該輸入單元愈不重要。

（3） 由（2）可篩選出較重要的輸入單元。

（4） 以篩選的特徵變數重新訓練網路，並比較結果。

2.3.3 分類法評估

一般來說，分類方法可以根據下列標準進行比較和評估 [7]：

1. 分類準確率（accuracy）：為最常用的評估指標，這涉及到模型是否能正確預測新的或先前未見資料的類別標籤之能力。
2. 速度（speed）：這涉及到產生和使用模型的計算效率。
3. 穩健性（robustness）：涉及到給定雜亂（noise）、不平衡（imbalance）或具有遺失值資料時，模型是否能正確預測的能力
4. 可擴延性（scalability）：對於給定大量資料，分類技術能否有效建構模型的能力。
5. 可解釋性（interpretability）：關於所建構的模型是否可以提供使用者理解和洞察的資訊。

至今已有許多關於不同分類法的比較研究 [16, 18, 19]，而且到目前為止仍然是一個值得研究的課題。研究中尚未發現對於上述五項評估指標皆表現優異的分類法，也因此在比較分類法時，必須視分類問題的使用目的及環境，來對分類法在評估指標的表現能力上進行取捨。

第三章　馬氏-田口系統

3.1 MTS 之構成要素

馬氏-田口系統（MTS）為針對多變量資料所發展的診斷及預測技術，它以考量變數間相關性的馬氏距離作為多變量系統的量測尺度，並以穩健工程之原理，執行系統最佳化的過程。

3.1.1 多變量診斷系統

一個典型的多變量診斷系統如圖 3.1 所示，圖中 X_1、$X_2 \cdots X_k$ 代表 k 個特徵變數，可提供資訊給決策者以制定決策；輸入信號（M）為系統狀態的真值，通常信號因子與系統輸出間具有互動關係，例如：汽車方向盤的轉向角度為一信號因子，它可以改變汽車的回轉半徑；雜音因子隨使用環境而異，是無法控制的參數，並會影響系統而造成偏差。在多變量診斷系統中，決策者並無法獨立地觀察每個特徵變數來制定正確的決策，因為變數間總潛在著未知的相關性，因此，在建構系統時，決策者必須將特徵變數間所存在的關係結構納入考量。

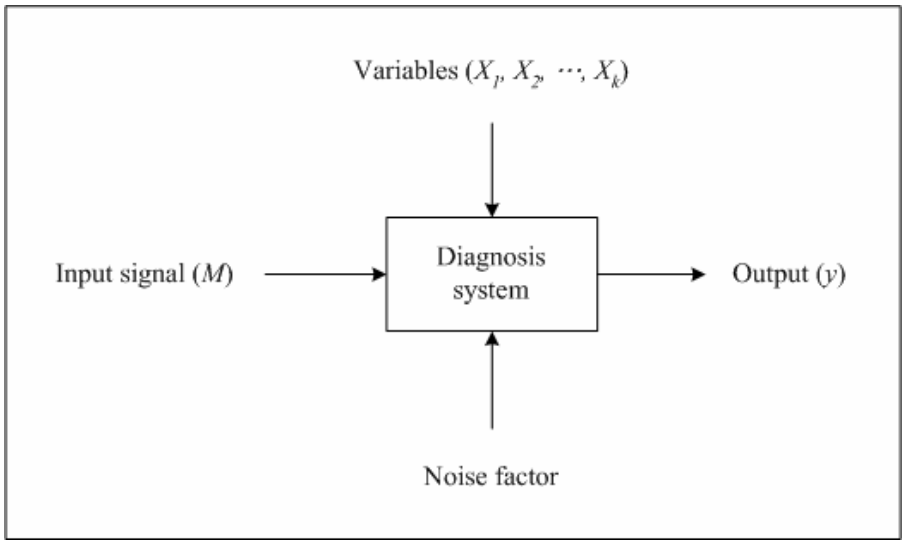


圖 3.1 多變量診斷系統 [2]

3.1.2 馬氏距離

馬氏距離是在 1936 年由印度統計學家 P. C. Mahalanobis 所提出，為考慮特徵變數關係結構的一種統計距離，它整合多變量系統中之不同變數資訊，使成為一體的、綜合性的評價指標。馬氏距離對於參照群體（reference group）的特性變數關係結構非常敏感，在典型的方法中，馬氏距離被用來量測一未知樣本點相較於每個群體中心點的遠近程度，並將未知樣本點歸類為距離較近的群體。相較於歐氏距離（Euclidean distance）而言，雖然皆是衡量未知樣本點與群體的距離，但馬氏距離更將變數間的關係性納入考量。圖 3.2 比較馬氏距離和歐氏距離的差別，假設參照群體的樣本點包含 X_1 和 X_2 兩個變數，橢圓虛線表示馬氏邊界，而圓形虛線代表歐氏邊界，X 為參照群體的中心點，稱為基準點或參照點，A、B 為兩個樣本點，並且 A 點的表現較接近於參照群體的分布趨向。當以歐氏距離來衡量 A、B 兩點的情況時，由圖中可明顯發現 B 點距離歐氏邊界較近，因此我們會判定 B 較 A 相似於參照群體；但若改以馬氏距離來觀察，則 A 點比 B 點更靠近馬氏邊界，即 A 點較相似於參照群體。由此可知，在多變量分析上，變數間的相關性是不容忽視的，因此利用考量關係結構的馬氏距離作為衡量指標，將會獲得較正確的結果。

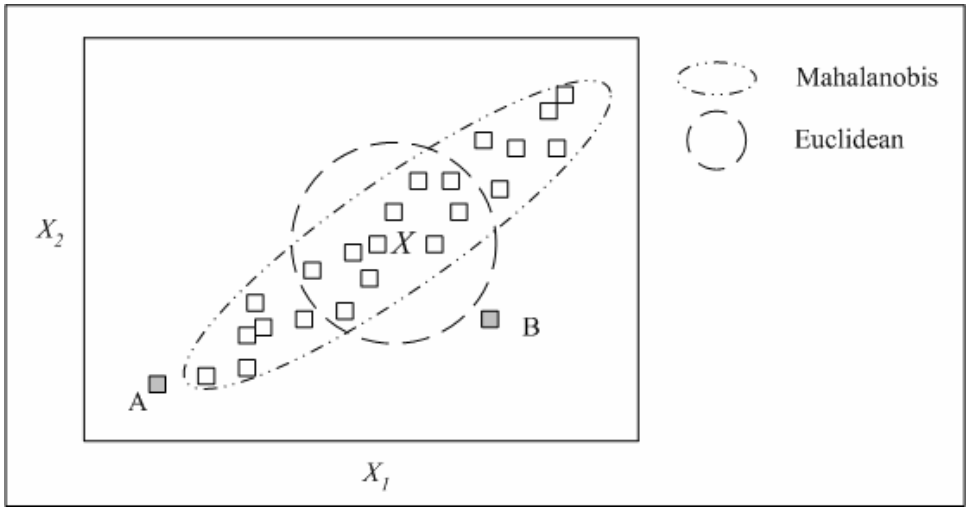


圖 3.2 馬氏距離與歐氏距離 [2]

在 MTS 中，馬氏距離被以適當的尺度作修改，並作為未知樣本點是否相似於參照群體的量測尺度。參照群體裡所有樣本點的馬氏距離構成馬氏空間（Mahalanobis space, MS），馬氏空間可說是包含參照群體裡所有變數的平均值、變異數及關係結構的一個資料庫。圖 3.3 顯示一個修改後的多變量診斷系統。

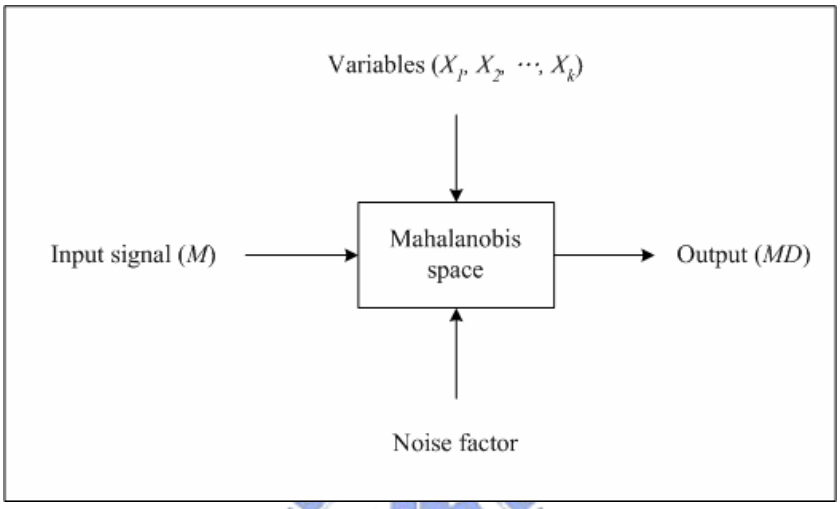


圖 3.3 修改後的多變量診斷系統

3.1.3 田口之穩健工程

田口玄一博士在 1950 年代提出的穩健工程概念，其目的在於提供具成本效率的改善方法，以提升產業在全球化市場中的競爭力。田口所提的穩健工程是以工程的角度事先了解品質問題，並利用社會損失成本作為衡量產品品質得依據。田口專注於工程品質的改善，包含：缺陷（defects）、故障（failures）、噪音（noise）、震動（vibrations）、污染（pollution）等，可用距離理想狀態的偏差來衡量之。改善工程品質的主要工具為直交表和 SN 比，其所強調的重點是在產品或製程設計時就考慮品質問題，亦即如何降低品質績效的變異。直交表用在實驗設計的配置，可以最小化所需的實驗次數，並且減少雜音因子所帶來的影響。SN 比則是用來衡量系統的功能性，SN 比愈大者愈佳。田口的穩健工程依成本效益的概念，應用直交表與 SN 比找出最佳的參數水準組合，這觀念和傳統的實驗設計完全依循統計原理，強調模式的確立，有很大的不同。

在 MTS 中，基於田口的穩健工程之觀點，馬氏距離可被視為一工程品質，因為它可以用來衡量一個未知樣本距離參照群體（馬氏空間）的異常（abnormality）程度。

3.2 MTS 方法論

在很多情況下，一系統的產品或品質是由許多不同的特徵（characteristics）所表現出來，這些不同特徵間可能是彼此獨立或互有影響的。但，這些關係往往不易釐清的，因此，整合系統中之不同特性資訊，使成為一綜合性的整體評價指標，對系統分析而言是比較有幫助的。田口博士採用馬氏距離為多元特性資料的評價指標，利用馬氏距離來判斷多元資料間是否為同質（homogeneous）或異質（heterogeneous），並利用 MTS 進行特徵變數的篩選。

3.2.1 馬氏距離分類法

MTS 的主要目的之一是導入一個以考量特徵變數之相互關係為基礎的尺度（scale）來量測樣本的異常（abnormality）程度，並作為分類診斷上的依據。以醫學診斷為例，其目標是基於這樣的尺度來量測出某人體是否患有疾病，及患病的嚴重度。馬氏距離為考慮多元變數之共變數矩陣（covariance matrix）的一種統計距離，被利用來建構 MTS 的量測尺度。在 MTS 中所用的馬氏距離是由馬氏原先所定義的馬氏距離除以特徵變數的數量。假設一多變量樣本集有 k 個特徵變數，包含 n 個樣本，其馬氏距離計算式如下：

$$MD_j = D_j^{\,2} = (\frac{1}{k}) \cdot Z_{ij}^{\,T} \cdot C^{-1} \cdot Z_{ij} \quad i = 1...k,\ j = 1...n \qquad (1)$$

其中，$Z_{ij} = (z_{1j}, z_{2j}, ..., z_{kj})$，表示 $x_{ij}(i = 1...k)$ 的標準化值之標準向量；

$z_{ij} = (x_{ij} - \overline{x_i}) / s_i$；

x_{ij} 為第 j 個樣本的第 i 個特徵變數值；

$\overline{x_i}$ 為第 i 個特徵變數的平均值；

s_i 為第 i 個特徵變數之標準差；

T 為轉置向量；

C^{-1} 為相關反矩陣；

k 為特徵變數的各數。

執行 MTS 的第一步，我們必須先定義出正常狀態（normal condition），並從中選取作為參照的正常（normal）群體，來建構馬氏空間（Mahalanobis space, MS）。在醫學診斷上，正常狀態即為健康者；在製造檢驗系統裡，正常狀態則為高品質的產品。在正常群體裡，所有樣本的馬氏距離構成一馬氏空間，馬氏空間可被視為一包含正常群體之所有特徵變數平均值、標準差及關係反矩陣的資料庫。當特徵變數的數目很大時，馬氏空間中的馬氏距離之分配會近似於 F 分配（分子自由度 k、分母自由度 ∞），或是卡方分配（自由度為 k）。我們可以很容易證明，馬氏空間的變數向量（經標準化後）之平均值為 0，而馬氏距離的平均趨近於 1，因此，馬氏空間又稱為基準空間（base space）。

當要判別一個未知樣本是否屬於正常狀態時，我們只需要利用馬氏空間中的平均值、標準差對其特徵變數值進行標準化，並以關係反矩陣來計算馬氏距離，即可獲得解答。我們稱一個非來自正常狀態的樣本為異常（abnormal）樣本，通常其馬氏距離會變得頗大，而馬氏距離愈大，則代表該樣本與正常狀態間有愈顯著的不同。圖 3.4 顯示正常群體及異常樣本的馬氏距離分配示意圖，圖中藉由分類閾值（threshold）的決定，來作為分類、診斷和預測上的依據。當有一樣本針對其 k 項特徵進行檢驗後，利用馬氏空間所計算的馬氏距離若大於閾值則判其為異常，反之則為正常。

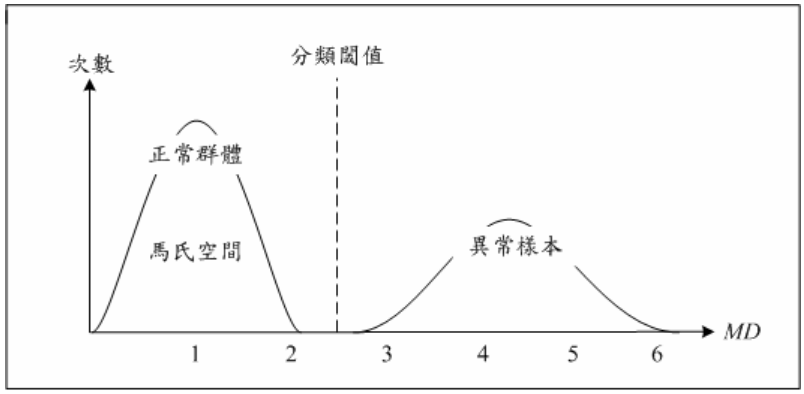


圖 3.4 正常樣本與異常樣本之馬氏距離分配

3.2.2 MTS 閾值

在 MTS 中，由於正常樣本與異常樣本之馬氏距離分布時常會發生重疊的現象，因此，如何有效決定其分隔上的閾值，是一直以來仍未獲得圓滿解答的問題。過去在 MTS 的使用上，通常以試誤（try and error）的方式在訓練樣本集中找到可以最大化總準確率的馬氏距離做為分類閾值，然而這樣的方式不僅毫無效率，並且容易產生對訓練資料的過度適配，因而分類能力無法在測試上重現，尤其在重疊現象現顯著的發生時，其能力就更值得商榷。此外，對於分類診斷問題而言，正確診斷出會影響績效的異常樣本並進行修正是較為重要的，因此在閾值的設定上應當以儘可能保護異常樣本的正確辨識為考量。本研究為此提出機率閾值（probabilistic threshold），即以柴比雪夫定理（Chebyshev's theorem）為基礎，在有效辨識異常樣本下，預估正常樣本的馬氏距離分佈行為，並作為訂定 MTS 閾值的依據。

1. 柴比雪夫定理

設 X 為一隨機變數，其平均數為 μ_X ，變異數為 $\sigma_X{}^2$ ，對任意正數 θ 而言，存在下列機率關係：

$$P(|X-\mu_X| \le \theta \cdot \sigma_X) \ge 1-\frac{1}{\theta^2} \qquad (2)$$

式（2）稱為柴比雪夫不等式，經過移項、變形後，下式仍然成立：

$$P(X \le \mu_X + \theta \cdot \sigma_X) \ge 1 - \frac{1}{\theta^2} \qquad (3)$$

2. 機率閾值訂定步驟

利用「訓練集」來訂定馬氏距離分類上的閾值，其步驟說明如下：

- 步驟一：將正常樣本之馬氏距離視為一隨機變數 X ，並以其平均值 $\overline{md}$ 估計 μ_X，標準差 s_{md} 估計 σ_X。為了避免離群值的影響，平均值 $\overline{md}$ 及標準差 s_{md} 以雙邊截尾後的樣本計算之，由於馬氏空間的馬氏距離分佈較為集中，因此建議雙尾各截 3％ 至 5％即可。
- 步驟二：對異常樣本的馬氏距離，除去左尾離群值，如圖 3.5。

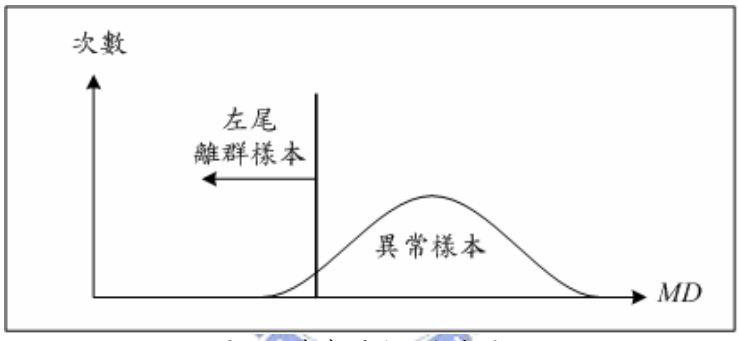


圖 3.5 機率閾值訂定步驟一

- 步驟三：比較異常樣本截尾後之馬式距離（步驟二）與所有正常樣本之馬氏距離，並計算正常樣本在非重疊部分所佔的樣本數量比率 h％。例如一訓練集中共有 15 個正常樣本，其馬氏距離分佈區間為 0.45 至 2.1，而異常樣本在截尾後之馬氏距離最小值為 1.7，如圖 3.6，正常樣本在非重疊部分所佔的比例為 $\frac{13}{15}$，即 86.67%（$h = 86.67$）。

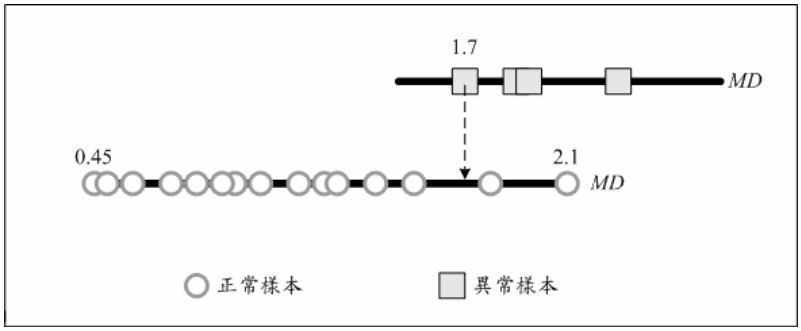


圖 3.6 機率閾值訂定步驟二

- 步驟四：根據式（3）之不等式，並令下式成立：

$$P(X \le \mu_X + \theta \cdot \sigma_X) = P(X \le \overline{md} + \theta \cdot s_{md}) \ge 1 - \frac{1}{\theta^2} = h\% - \lambda\% \qquad (4)$$

其中，λ%（$h > \lambda > 0$）為一校正值，考慮使用者對該模型是否能正確辨識異常樣本之信心水準，通常建議設為 5%。

- 步驟五：根據式（4）求得θ，並計算隨機變數 X 之上界 $(\overline{md} + \theta \cdot s_{md})$，作為 MTS 中的分類閾值（$T$）。

3. 機率閾值計算式

MTS 中之分類閾值（T）計算公式如下：

$$T = \overline{md} + \sqrt{\frac{100}{100 + \lambda - h}} \cdot s_{md} \qquad (5)$$

其中，$\overline{md}$ 為馬氏空間中正常樣本之馬氏距離截尾平均；

s_{md} 為馬氏空間中正常樣本之馬氏距離截尾標準差；

λ 為校正值，建議設為 5；

h 為在非重疊部分所佔的樣本數量百分比數，如圖 3.6 所示。

3.2.3 特徵變數之篩選

MTS 的第二個目的為替多變量系統刪除不重要的特徵變數，以降低系統成本或加速資料處理，而直交表及 SN 比在確認重要變數上是很有效的。

1. 直交表之配置

在直交表裡，每項特徵變數或因子會被個別配置到不同的行，而每一列則為變數或因子的不同水準之組合，代表一種實驗組合。利用直交表，我們可以研究每個特徵變數對系統輸出的影響。今假設一多變量系統有 k 項特徵變數，並設定每項特徵變數為 2 水準：

水準 1=使用此項特徵變數；

水準 2=不使用此項特徵變數；

接著選用適合k項因子配置的直交表進行實驗。

在此假設系統中有$k=10$項特徵變數（$C_1, C_2, \ldots, C_{10}$）待分析，因而我們選用L_{12}直交表，其實驗配置如表 3.1。其中，Run1 的各特性變數水準皆為 1（皆使用），此時需利用所有的k項特徵變數建立馬氏空間；而Run2 實驗則僅使用水準為 1 的特徵變數（C_1, C_2, C_3, C_4, C_5）來製作馬氏空間，依此類推。

表 3.1 L_{12}直交表配置

Run	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}		$M_1, M_2, \cdots, M_d$	SN 比
	1	2	3	4	5	6	7	8	9	10	11		
1	1	1	1	1	1	1	1	1	1	1	1	$MD_1, MD_2, \cdots, MD_d$	η_1
2	1	1	1	1	1	2	2	2	2	2	2		η_2
3	1	1	2	2	2	1	1	1	2	2	2		η_3
4	1	2	1	2	2	1	2	2	1	1	2		η_4
5	1	2	2	1	2	2	1	2	1	2	1		η_5
6	1	2	2	2	1	2	2	1	2	1	1		η_6
7	2	1	2	2	1	1	2	2	1	2	1		η_7
8	2	1	2	1	2	2	2	1	1	1	2		η_8
9	2	1	1	2	2	2	1	2	2	1	1		η_9
10	2	2	2	1	1	1	1	2	2	1	2		η_{10}
11	2	2	1	2	1	2	1	1	1	2	2		η_{11}
12	2	2	1	1	2	1	2	1	2	2	1		η_{12}

2. SN 比之使用

在根據直交表的特性變數組合建構馬氏空間後，接下來必須使用 SN 比來選取重要變數。在 MTS 中，多變量系統的輸入信號（M）即為異常樣本的嚴重水準（severity level），例如：不良產品所帶來的損失金額。田口建議可使用兩種不同的 SN 比：（1）望大特性 SN 比；（2）動態特性 SN 比。其使用方式及時機分述如下：

（1） 望大特性 SN 比：

當信號真值未知，即異常樣本的嚴重水準未知時，我們使用望大特性 SN

比來決定重要的特性變數。首先，收集不屬於正常狀態的d個樣本（d個異常樣本），因為異常樣本之馬氏距離必須與正常樣本之馬氏距離有很清楚的劃分，所以d個異常樣本以馬氏空間為參照的馬氏距離為越大越好，屬於望大特性，其 SN 比計算氏如下：

$$\eta = -10 \cdot \log_{10}\left[\frac{1}{d}\cdot\left(\frac{1}{MD_1}+\frac{1}{MD_2}+\cdots+\frac{1}{MD_d}\right)\right] \qquad (6)$$

（2） 動態特性 SN 比：

動態特性系統的輸出和輸入信號有著互動關係。在 MTS 中，動態特性 SN 比的計算有以下兩種情況：

- 情況 1：信號真值已知，即所有異常樣本的嚴重水準已知。

首先，收集不屬於正常狀態的d個異常樣本，並以 $y_i\ (i=1\ldots d)$ 表示其馬氏距離的平方根，作為系統的輸出，而 M_1、$M_2\cdots M_d$ 表示d個異常樣本的信號真值。輸入信號和系統輸出關係如下：

$$y_i = \beta M_i \qquad i=1\ldots d \qquad (7)$$

其中，$y_i = \sqrt{MD_i}$ ；

M_i 為第i個異常樣本的信號真值；

β 為斜率值（理想狀況下$\beta=1$）。

情況 1 之動態特性 SN 比計算式如下：

$$\eta = 10 \cdot \log_{10}\frac{\frac{1}{r}\cdot(S_\beta - V_N)}{V_N} \qquad (8)$$

其中，$S_T = y_1^{\,2} + y_2^{\,2} + \cdots + y_d^{\,2}$ ；

$S_\beta = \frac{1}{r}\cdot(M_1y_1 + M_2y_2 + \cdots + M_dy_d)^2$ ；

$r = M_1^{\,2} + M_2^{\,2} + \cdots + M_d^{\,2}$ ；

$$V_N = \frac{S_T - S_\beta}{d-1} \text{ 。}$$

- 情況 2：輸入信號真值未知。

在某些信號真值未知的案例中，使用動態特性 SN 比亦會有不錯的表現。首先，收集不屬於正常狀態的 d 組群體（d 組異常群體），組內皆包含 m 個異常樣本，且每組異常程度不同。以各異常樣本的馬氏距離平方根 $y_{ij}\ (i=1...d,\ j=1...m)$ 作為系統的輸出，而 $M_i\ (i=1...d)$ 則視為 d 組異常樣本的輸入信號真值，其計算式如下：

$$M_i = \frac{1}{m} \cdot \sum_{j=1}^{m} y_{ij} \qquad i=1...d \tag{9}$$

其中，$y_{ij} = \sqrt{MD_{ij}} \qquad i=1...d,\ j=1...m$ 。

情況 2 之動態特性 SN 比計算式如下：

$$\eta = 10 \cdot \log_{10} \frac{\frac{1}{r} \cdot (S_\beta - V_e)}{V_e} \tag{10}$$

其中，$Y_i = \sum_{j=1}^{m} y_{ij} \qquad i=1...d$ ；

$S_e = S_T - S_\beta$ ；

$V_e = \frac{S_e}{dm-1}$ ；

$S_T = \sum_{i=1}^{d} \sum_{j=1}^{m} y_{ij}^{\ 2}$ ；

$r = m \cdot \sum_{i=1}^{d} M_i^{\ 2}$ ；

$S_\beta = \frac{1}{r} \cdot (\sum_{i=1}^{d} M_i Y_i)^2$ 。

當直交表中每個實驗的 SN 比（例如表 3.1 中的 $\eta_1 \sim \eta_{12}$）計算完畢，對於一特徵變數 X_i，我們使用 $\overline{SN_i^+}$ 來表示使用（水準 1）此變數 X_i 的所有實驗之平

均 SN 比；相反的，以 $\overline{SN_i^-}$ 表示不使用（水準 2）此變數的所有實驗之平均 SN 比。以表 3.1 為例，特徵變數 C_1 之 $\overline{SN_1^+}$ 為 $\eta_1 \sim \eta_6$ 之平均，而 $\overline{SN_1^-}$ 為 $\eta_7 \sim \eta_{12}$ 之平均，依此類推。最後以效果增量（effect gain）表示兩者之間的差距，計算式如下：

$$Gain_i = \overline{SN_i^+} - \overline{SN_i^-} \qquad (11)$$

依增量越大越好的原則，決定特徵變數對於系統分類診斷的重要性，來選擇最佳條件。另外，可以挑選出的最佳變數建構縮減模型（reduced model），並利用其計算異常樣本的馬氏距離，求得一個 SN 比。倘若系統在經過 MTS 分析後獲得了改善，那麼縮減模型的 SN 比將會較完整模型（full model）所得到的 SN 比大，因此，我們可以 SN 比在分析前後的增量來評估系統在功能上的改善。最後，必須作測試，以確認縮減模型是否有足夠的分類診斷能力。

3.3 MTS 之執行步驟

MTS 的執行過程可分為四階段，分述如下 [2, 3, 20]：

1. 構建完整模型（full model）之量測尺度。

- 定義用來判定正常樣本的 k 項特徵變數。在健康檢查的例子中，醫生必須檢視所有會造成疾病的 k 個因子來判斷健檢者是否健康。
- 確立正常狀態（例如健康者），決定正常群體，並收集群內所有 n 個正常樣本的 k 項資料。
- 計算正常群體各特徵變數的平均值（$\overline{x_i}$）、標準差（s_i），並標準化之。利用標準化值計算相關反矩陣（C^{-1}）。
- 計算正常群體中各樣本的馬氏距離，構成一馬氏空間（基準空間）。利用這些馬氏距離，可以定義出原點（標準化變數向量平均為 0）及單位距離（馬

氏距離平均為 1)。

- 使用馬氏空間中的原點及單位距離作為量測尺度的參照基準。

2. 確認完整模型之量測尺度。

- 收集不屬於正常狀態的 d 個或 d 組異常樣本，並包含各樣本的 k 項特徵變數。在健康檢查的例子中，異常樣本是指患有任何病症的健檢者。
- 計算這些異常樣本的馬氏距離來確認尺度。異常樣本中的特徵變數值以階段 1 中正常群體的平均值（$\overline{x_i}$）、標準差（s_i）標準化之，並且利用正常群體的關係反矩陣（C^{-1}）計算異常樣本的馬氏距離。
- 如果階段 1 所構建的尺度是好的，那麼異常樣本的馬氏距離將會較大。透過這個準則來確認量測尺度。

3. 確認重要特徵變數

- 將 k 項特徵變數視為控制因子，每因子設 2 水準：水準 1 為使用該項，水準 2 為不使用該項；並將控制因子配置於適當的直交表。
- 使用直交表及 SN 比找出重要特徵變數。利用 d 個或 d 組異常樣本的馬氏距離，選擇適當方法計算 SN 比，作為直交表中每個實驗的回應值。重要特徵變數則以效果增量越大越好的原則來評估。

4. 利用重要的特徵變數作分類診斷

- 確認縮減模型（reduced model）的量測尺度。正常群體以上步驟所決定的重要特徵變數建立馬氏空間，並用異常樣本的馬氏距離作尺度確認。
- 評估系統改善。利用異常樣本在縮減模型及完整模型上所得的 SN 比來評估系統在功能上的改善。
- 訂定 MTS 分類閾值。本研究使用所提出的機率閾值。
- 進行測試，以驗證所建立的縮減模型是否有足夠的判別能力。

圖 3.5 顯示利用 MTS 確認重要變數並作分類診斷預測的流程圖。

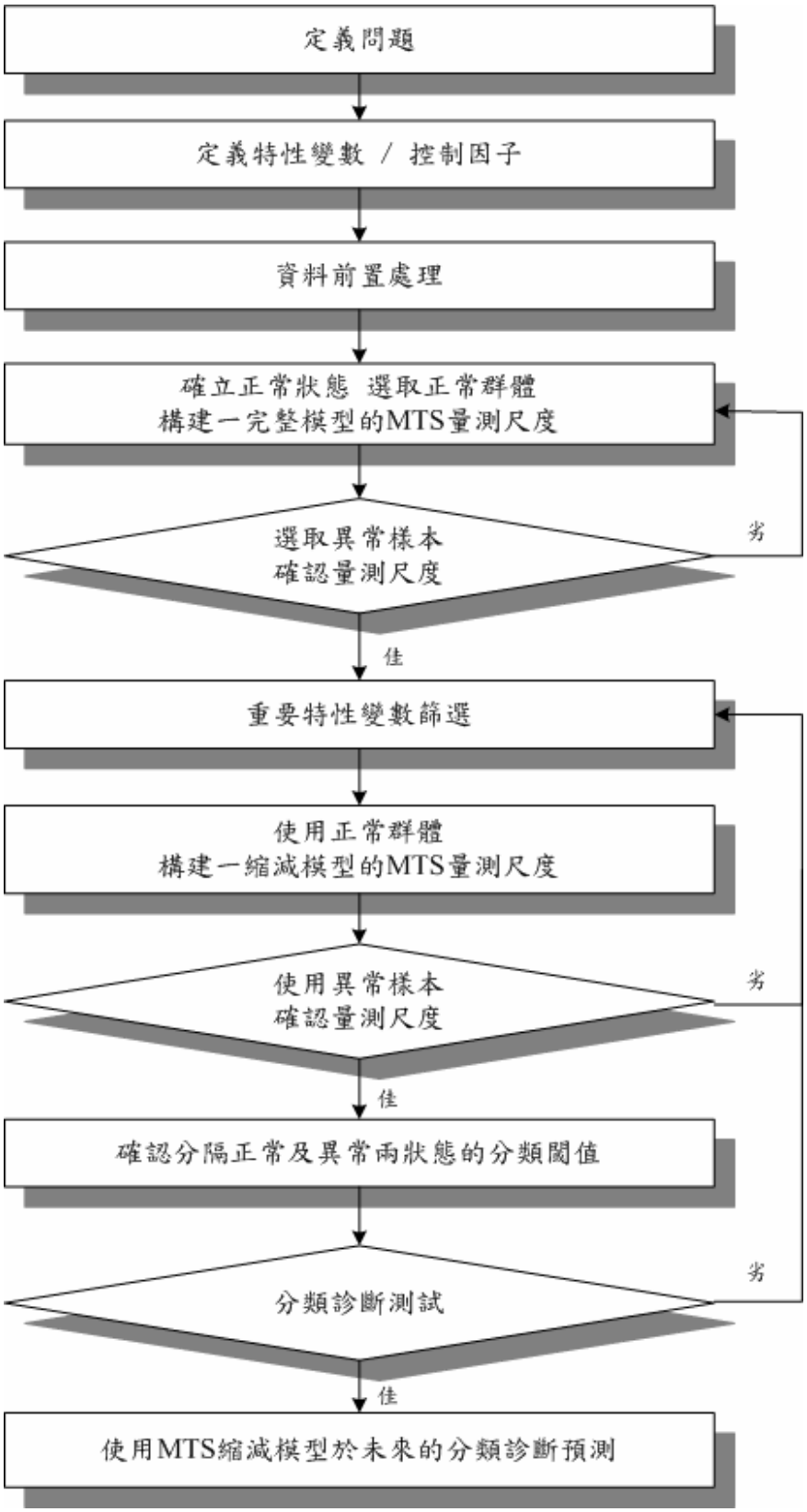


圖 3.7　MTS 流程圖

3.4 MTS 之特點

綜合先前的介紹與討論，以下列出 MTS 與統計或其他常用分類技術有所相異的四大特點：

1. 量測尺度

對一般可用於分類的技術而言，如判別分析、複迴歸分析、決策樹、類神經網路等，其目標不外乎是準確的將未知樣本分類至所屬的不同群體或母體（populations）。不同地，MTS 的主要目的是提供一個量測尺度，可以使我們在連續的尺度上量測一樣本的異常程度。這樣的尺度不僅可以用來達成分類的工作，更能夠基於樣本的異常度，幫助決策者針對該樣本制定適當的對應行動。

2. 樣本群體

在一般的分類技術中，通常視正常樣本及異常樣本來自兩個分隔的母體。然而，在 MTS 中，是沒有所謂的母體的。我們需要的只是一個稱作「正常」的群體，用來建立關係結構，和定義量測尺度的參照點。至於異常樣本（或正常狀態外的樣本），雖然我們在分析時將其同歸入異常群體中，但事實上在 MTS 的觀點裡，因其所呈現的細部狀態都不同，所以每個異常樣本皆是獨一無二的。就好比健康的人，這個群體在健檢項目的表現上都是差不多；但對於不健康的人而言，不合格的健檢項目可能因人而異，因此，每個不健康的人皆可視為不同的個體。

3. 資料分析

在典型的多變量方法裡，以機率或統計為基礎的推論經常被使用在多變量系統的分析上。分析的過程中經常需要事先知道資料的機率分配或對統計假設的要求，例如，最為常見的常態分配假設、變異數相等假設等。然而，現實的資料型態卻往往無法合乎這些要求。在 MTS 裡，我們將所使用的量測及程序視為一資料分析法，而非一般的機率推論，因此在 MTS 的使用上，並無任何統計假設限

制及前提。

4. 維度縮減

維度縮減對於一個多變量系統而言仍是一個挑戰。主成分分析（principle component analysis）雖然以若干個主成分的形式來包含並取代系統的變數，來達到維度縮減的目的，但並沒有因此而減少分析上所需的變數數量。逐步迴歸（stepwise regression）對於特徵變數的縮減上是相當有用的，但卻會隨變數數量的增大而增加其疊代運算次數，而顯得非常複雜。對此，MTS 使用 SN 比來確認重要變數，並減少系統的維度，此法在觀念及計算上不但簡單，且不需要任何關於特徵變數的先驗知識。

第四章　馬氏-田口系統之穩健性評估

本章以加州大學UCI資料庫（http://www.ics.uci.edu/~mlearn/MLSummary）的數據為例，利用馬氏-田口系統（MTS）、逐步判別分析（SDA）、決策樹分析（DT）和倒傳遞神經網路（BPN），在非平衡（imbalance）的訓練資料進行特徵選取及分類模型或規則的萃取，並探討其是否會對未來的預測產生偏差，最後比較之。在MTS的執行過程中，皆使用本研究所提出的機率閾值（probabilistic threshold），並將之與傳統試誤法（try and error）所訂定的閾值做一判斷效果的比較。在分析環境上，使用SPSS統計套裝軟體執行逐步判別分析，決策樹分析使用See5 套裝軟體，倒傳遞式神經網路則利用Professional II類神經網路套裝軟體進行分析。分析資料集分別為：（1）威斯康辛乳癌（Wisconsin breast cancer）；（2）字母辨識（letter recognition）；（3）心臟病（heart disease）。

為了表現各種分類法在面對不平衡資料時的穩健能力，本研究在每組資料集上，利用兩類別（正常及異常樣本）在訓練資料中所佔比例的不同，表現三種訓練資料的不平衡程度，並做為分析時的三個實驗，而分佈比例相差愈大則表示該訓練資料愈不平衡。首先，使用各分類法在不同分佈比例的訓練資料上進行特徵選取，並且分別建立類別預測的縮減模型，接著使用相同的測試資料做測試，並評估其個別表現。績效評估指標說明如下 [7]：

（1）特效性（specificity）：在多量類別上的預測準確率，用來評估分類法對屬於多量類別之樣本的辨識能力。

（2）敏感度（sensitivity）：在少量類別上的預測準確率，用來評估分類法對屬於少量類別之樣本的辨識能力。

（3）總準確率（total accuracy rate）：分類法在整體資料（包含二類別）上的類別預測準確率。

（4）相對敏感度（relative sensitivity）：為敏感度與特效性的比值，比值愈接近

1，表示此分類法在兩類別的預測能力上是愈相當的。若小於 1，表示該類別預測較無能力辨識少量類別。

4.1 威斯康辛乳癌

此資料集有 9 項特徵變數($V_1, V_2, \ldots, V_9$)，分為良性腫瘤及惡性腫瘤兩類別，良性腫瘤為正常狀態，惡性腫瘤則為異常。經過前處理後，任務相關資料包含正常樣本 444 筆，異常樣本 239 筆。本例所規劃的三個實驗乃由任務相關資料中隨機抽取數據，其訓練及測試集數量如表 4.1 所示。在本分析中，正常樣本屬於多量類別，並以特效性衡量其類別預測準確率，而異常樣本屬於少量類別，其類別預測準確率則由敏感度表示之。異常樣本在訓練集中所佔比例由實驗 1 至實驗 3 逐漸縮小。

表 4.1 訓練及測試集之數量分佈（威斯康辛乳癌）

訓練集分佈比例 正常：異常		訓練集			測試集		
		正常	異常	總和	正常	異常	總和
實驗 1	2 : 1	66	33	99	22	11	33
實驗 2	5 : 1	165	33	198	22	11	33
實驗 3	11 : 1	363	33	396	22	11	33

4.1.1 MTS 分析結果

透過訓練集的分析，每個實驗由 MTS 所篩選的特徵變數如表 4.2，我們將利用這些變數建立類別預測縮減模型。表 4.3 顯示在各實驗中，利用機率閾值及試誤法來決定縮減模型分類閾值的類別預測比較結果。試誤法在訓練過程中搜尋可最大化總準確率的閾值，而機率閾值則利用柴比雪夫定理來計算，最後經過測試集的測試，顯示兩者的判斷效果是相當的。表 4.4 為 MTS 所建構的類別預測縮減模型（利用機率閾值）之訓練及測試結果。可以發現，MTS 在三個實驗裡皆縮減了一半的特徵變數，而所建構的縮減模型用在測試集的類別預測時，無論在

特效性、敏感度或總準確率上皆有不錯的表現。另外，三個實驗的測試結果在相對敏感度上都接近 1，此訊息代表不平衡的訓練資料對 MTS 建構分類預測模型的過程並無顯著的負面影響，即不因訓練資料在兩類別上的數量差距而降低對少量類別的診斷能力。

表 4.2　MTS 特徵選取結果（威斯康辛乳癌）

MTS 特徵選取		
實驗	特徵數	特徵變數
實驗 1	5	V_1、V_3、V_5、V_6、V_8
實驗 2	4	V_3、V_5、V_6、V_8
實驗 3	5	V_2、V_3、V_4、V_6、V_7

表 4.3 不同閾值下之 MTS 縮減模型分類結果（威斯康辛乳癌）

實驗 1	閾值	訓練集（%）			測試集（%）		
		特效性	敏感度	總準確率	特效性	敏感度	總準確率
機率閾值	5.515	96.97	100.00	97.98	95.45	100.00	96.97
試誤法	7.15	100.00	100.00	100.00	100.00	100.00	100.00
實驗 2	閾值	訓練集（%）			測試集（%）		
		特效性	敏感度	總準確率	特效性	敏感度	總準確率
機率閾值	5.516	95.76	100.00	96.46	95.45	90.91	93.94
試誤法	6.80	99.39	100.00	99.49	95.45	90.91	93.94
實驗 3	閾值	訓練集（%）			測試集（%）		
		特效性	敏感度	總準確率	特效性	敏感度	總準確率
機率閾值	5.978	96.42	100.00	96.72	100.00	100.00	100.00
試誤法	9.75	100.00	100.00	100.00	100.00	90.91	96.97

表 4.4　MTS 縮減模型分類結果之比較（威斯康辛乳癌）

MTS 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	96.97	100.00	97.98	1.03	95.45	100.00	96.97	1.05
實驗 2	95.76	100.00	96.46	1.04	95.45	90.91	93.94	0.95
實驗 3	96.42	100.00	96.72	1.04	100.00	100.00	100.00	1.00

4.1.2 SDA 分析結果

各實驗所篩選的特徵變數及縮減模型之訓練及測試結果分別如表 4.5、4.6。逐步判別分析所篩選的特徵變數較 MTS 稍多，用於測試集的類別預測時，雖然特效性皆達到百分之百，並且總準確率亦有九成以上，但敏感度的表現卻在實驗 2 及實驗 3 稍低。這表示在訓練學習的過程中，由於使用的異常樣本數量明顯少於正常樣本，因此所訓練的預測模型對異常樣本的診斷能力較差。

表 4.5　SDA 特徵選取結果（威斯康辛乳癌）

SDA 特徵選取		
實驗	特徵數	特徵變數
實驗 1	6	V_1、V_3、V_5、V_6、V_8、V_9
實驗 2	5	V_1、V_3、V_4、V_6、V_8
實驗 3	6	V_2、V_5、V_6、V_7、V_8、V_9

表 4.6　SDA 縮減模型分類結果之比較（威斯康辛乳癌）

SDA 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	100.00	90.91	96.97	0.91	100.00	100.00	100.00	1.00
實驗 2	100.00	96.96	99.49	0.97	100.00	81.82	93.94	0.82
實驗 3	100.00	100.00	100.00	1.00	100.00	81.82	93.94	0.82

4.1.3 DT 分析結果

表 4.7 為特徵變數篩選結果，由表 4.8 可以發現，在分類縮減模型的測試上，隨著異常樣本在訓練資料中所佔的比例逐漸縮小，萃取出的分類模型對異常樣本的判別敏感度亦逐漸降低。因此縱使決策樹分析在特徵選取的能力上表現不錯，但其對處理非平衡資料的穩健性是不足的。

表 4.7　DT 特徵選取結果（威斯康辛乳癌）

DT 特徵選取		
實驗	特徵數	特徵變數
實驗 1	2	V_2、V_8
實驗 2	3	V_2、V_3、V_6
實驗 3	2	V_2、V_7

表 4.8　DT 縮減模型分類結果之比較（威斯康辛乳癌）

DT 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	98.48	96.97	97.98	0.98	100.00	90.91	96.97	0.91
實驗 2	100.00	100.00	100.00	1.00	95.49	63.64	84.85	0.67
實驗 3	99.72	100.00	99.75	1.00	100.00	45.45	81.82	0.45

4.1.4　BPN 分析結果

倒傳遞神經網路之網路架構經過試誤，最佳完整及縮減模型架構（RMSE 最小）如表 4.9。利用最佳完整模型之網路架構進行特徵選取，其選取效果不如 MTS 或決策樹顯著，結果如表 4.10。觀察表 4.11，在測試結果上，相對敏感度指標有隨訓練資料之不平衡現象的顯著而逐漸降低的趨勢。此外，雖然實驗 3 測試結果的總準確率高達九成以上，但卻犧牲了異常樣本的判斷敏感度，而這樣的結果並不是我們所樂見的。

表 4.9　BPN 之最佳網路架構（威斯康辛乳癌）

BPN 網路架構						
實驗	完整模型			縮減模型		
	學習率	momentum	網路架構	學習率	momentum	網路架構
實驗 1	0.1	0.95	9-5-1	0.2	0.95	6-4-1
實驗 2	0.1	0.8	9-9-1	0.2	0.85	5-4-1
實驗 3	0.15	0.9	9-5-1	0.2	0.9	5-7-1

表 4.10　BPN 特徵選取結果（威斯康辛乳癌）

BPN 特徵選取		
實驗	特徵數	特徵變數
實驗 1	6	V_2、V_3、V_4、V_5、V_6、V_8
實驗 2	5	V_3、V_4、V_5、V_6、V_8
實驗 3	5	V_2、V_5、V_6、V_7、V_8

表 4.11　BPN 縮減模型分類結果之比較（威斯康辛乳癌）

BPN 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
實驗 2	95.45	95.45	95.45	1.00	91.67	77.78	84.85	0.85
實驗 3	100.00	100.00	100.00	1.00	100.00	81.82	93.94	0.82

4.1.5 結論

在威斯康辛乳癌的應用裡，MTS 有效的將 9 個特徵變數刪減為 4 至 5 個，而其在類別預測總準確率的表現上與各方法比較起來皆具相當的水準，此外，在 MTS 閾值的訂定上，採用機率閾值亦得到不錯的結果，如圖 4.1。圖 4.2 顯示各分類法在三個實驗中相對敏感度指標的表現，可以很清楚看到，在實驗 1 的測試結果中，四種分類法的相對敏感度皆接近 1，顯示被訓練的分類模型在判斷正常或異常樣本的能力上是相當接近的。然而，除了 MTS 外，逐步判別分析、決策樹及倒傳遞神經網路在實驗 2 及實驗 3 的相對敏感度皆逐漸下降，其中決策樹分析尤其明顯，此現象表示這些方法在類別預測模型建構的過程中，容易受到訓練資料分佈型態不平衡所影響，因為訓練資料中的異常樣本數量遠少於正常樣本，因此使訓練過程發生偏差，而減少對異常樣本的判斷敏感度。由本例分析結果可見，MTS 在處理不平衡資料的分類問題上是比較穩健的。

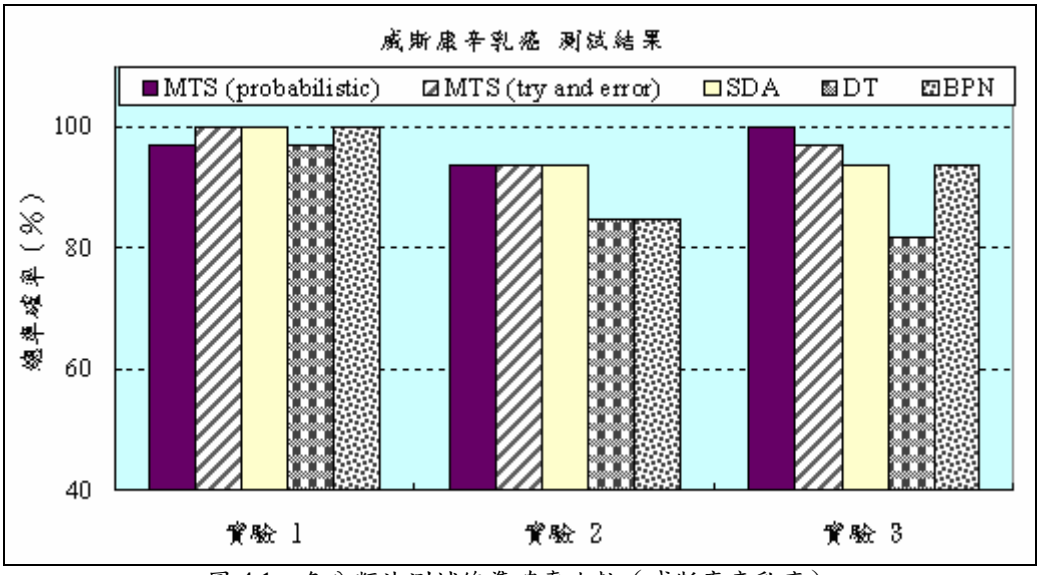


圖 4.1　各分類法測試總準確率比較（威斯康辛乳癌）

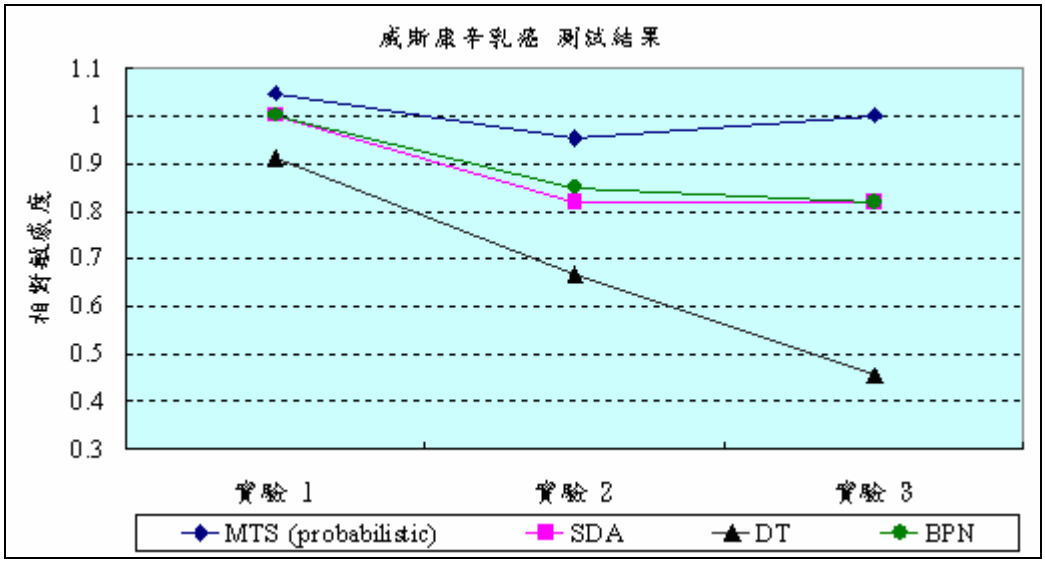


圖 4.2 各分類法測試相對敏感度比較（威斯康辛乳癌）

4.2 英文字母辨識

此資料集有 16 項特徵變數（$V_1, V_2, \ldots, V_{16}$），類別為 A 至 Z 的 26 個英文字母，本研究將字母 A 視為正常狀態，字母 A 以外的其他 25 個字母視為異常狀態。因此，本例之目的在有效辨識字母 A。經過前處理，任務相關資料包含正常樣本（字母 A）789 筆及異常樣本若干筆。本例所規劃的三個實驗乃由任務相關資料中隨機抽取數據，其訓練及測試集數量如表 4.12 所示。在本分析中，正常樣本

屬於多量類別，並以特效性衡量其類別預測準確率，而異常樣本屬於少量類別，其類別預測準確率則由敏感度表示之。異常樣本在訓練集中所佔比例由實驗 1 至實驗 3 逐漸縮小。

表 4.12 訓練及測試集之數量分佈（英文字母辨識）

訓練集分佈比例 正常：異常		訓練集			測試集		
		正常	異常	總和	正常	異常	總和
實驗 1	2 : 1	90	45	135	30	15	45
實驗 2	9 : 1	405	45	450	30	15	45
實驗 3	15 : 1	675	45	720	30	15	45

4.2.1 MTS 分析結果

三個實驗由 MTS 所篩選的特徵變數如表 4.13，並利用這些變數建立類別預測縮減模型。表 4.14 為使用兩種閾值的類別預測結果，測試結果顯示兩者的判斷結果是相當的。表 4.15 為 MTS 所建構的類別預測縮減模型（利用機率閾值）之訓練及測試結果。MTS 在三個實驗裡縮減了半數的特徵變數，而縮減模型的測試結果仍然在各評估指標保有高正確性，並且在相對敏感度指標上都接近於 1，可見 MTS 在建構預測模型時，抗拒不平衡訓練資料的穩定性。

表 4.13 MTS 特徵選取結果（英文字母辨識）

MTS 特徵選取		
實驗	特徵數	特徵變數
實驗 1	7	V_5、V_6、V_8、V_9、V_{10}、V_{15}、V_{16}
實驗 2	8	V_8、V_9、V_{10}、V_{11}、V_{12}、V_{13}、V_{15}、V_{16}
實驗 3	7	V_9、V_{10}、V_{11}、V_{12}、V_{13}、V_{15}、V_{16}

表 4.14　不同閾值下之 MTS 縮減模型分類結果（英文字母辨識）

實驗 1	閾值	訓練集（%）			測試集（%）		
		特效性	敏感度	總準確率	特效性	敏感度	總準確率
機率閾值	3.700	98.89	100.00	99.26	100.00	100.00	100.00
試誤法	3.15	98.89	100.00	99.26	100.00	100.00	100.00
實驗 2	閾值	訓練集（%）			測試集（%）		
		特效性	敏感度	總準確率	特效性	敏感度	總準確率
機率閾值	4.113	99.26	100.00	99.33	96.67	93.33	95.56
試誤法	5.15	99.75	97.78	99.56	100.00	80.00	93.33
實驗 3	閾值	訓練集（%）			測試集（%）		
		特效性	敏感度	總準確率	特效性	敏感度	總準確率
機率閾值	4.291	98.67	97.78	98.61	93.33	93.33	93.33
試誤法	4.45	99.11	97.78	99.03	96.67	93.33	95.56

表 4.15　MTS 縮減模型分類結果之比較（英文字母辨識）

MTS 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	98.89	100.00	99.26	1.01	100.00	100.00	100.00	1.00
實驗 2	99.26	100.00	99.33	1.01	96.67	93.33	95.56	0.97
實驗 3	98.67	97.78	98.61	0.99	93.33	93.33	93.33	1.00

4.2.2　SDA 分析結果

各實驗利用逐步判別分析所篩選的特徵變數及類別預測之訓練及測試結果分別如表 4.16、4.17。由表 4.16 可以觀察到，選取的特徵數量並不穩定，並且在實驗 3 中多達 12 項。在縮減模型的測試方面，三個實驗的總準確率皆低於 MTS 的測試結果。另外，由相對敏感度指標可以明顯觀察到大幅遞減的現象，這表示類別預測模型對異常樣本的辨別能力已隨訓練過程中異常樣本所佔比例的下降而逐漸低落。

表 4.16　SDA 特徵選取結果（英文字母辨識）

SDA 特徵選取		
實驗	特徵數	特徵變數
實驗 1	8	V_6、V_8、V_9、V_{10}、V_{11}、V_{12}、V_{15}、V_{16}
實驗 2	9	V_6、V_7、V_8、V_9、V_{10}、V_{11}、V_{12}、V_{13}、V_{16}
實驗 3	12	V_1、V_2、V_3、V_6、V_8、V_9、V_{10}、V_{11}、V_{12}、V_{13}、V_{14}、V_{16}

表 4.17　SDA 縮減模型分類結果之比較（英文字母辨識）

SDA 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	100.00	93.33	97.78	0.93	93.33	100.00	95.56	1.07
實驗 2	99.26	93.33	98.67	0.94	100.00	86.67	95.56	0.87
實驗 3	98.22	88.89	97.76	0.91	96.67	53.33	82.22	0.55

4.2.3　DT 分析結果

表 4.18 為決策樹分析的特徵選取結果，表中顯示特徵變數有漸增的情形，與 MTS 的篩選結果比較起來，是相對較不穩定的。表 4.19 可以發現，在分類縮減模型的測試上，總準確率的表現不如前兩種方法佳。另外，實驗 1 雖然在特效性上表現不若實驗 2 及實驗 3 好，但在與敏感度比較後，可以明顯感受到實驗 1 對於兩類樣本的正確判斷能力有較相近的績效表現，而實驗 2 及實驗 3 就因為訓練資料中的異常樣本比例更加縮小，使得萃取的分類縮減模型僅能正確辨識正常樣本，對異常樣本的辨識敏感度則呈現低落的情形。同樣的，這樣的資訊仍可從相對敏感度指標的呈現上獲得，實驗 1 的相對敏感度接近 1，表示在兩類別上的辨識能力相當，而實驗 2 及實驗 3 則遠小於 1，表示對異常樣本的辨識能力較差。

表 4.18　DT 特徵選取結果（英文字母辨識）

DT 特徵選取		
實驗	特徵數	特徵變數
實驗 1	5	V_9、V_{11}、V_{14}、V_{15}、V_{16}
實驗 2	7	V_7、V_9、V_{11}、V_{12}、V_{13}、V_{15}、V_{16}
實驗 3	9	V_2、V_4、V_5、V_7、V_9、V_{10}、V_{13}、V_{14}、V_{16}

表 4.19　DT 縮減模型分類結果之比較（英文字母辨識）

DT 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	100.00	97.78	99.26	0.98	80.00	86.67	82.22	1.08
實驗 2	99.75	97.78	99.51	0.98	100.00	80.00	93.33	0.80
實驗 3	100.00	88.89	99.31	0.89	100.00	60.00	86.67	0.60

4.2.4　BPN 分析結果

倒傳遞神經網路之網路架構經過試誤，最佳完整及縮減模型架構（RMSE 最小）如表 4.20。利用最佳完整模型之網路架構進行特徵選取，結果如表 4.21，其所篩選的數量較 MTS 稍多。表 4.22 為縮減模型的訓練及測試結果，觀察測試結果，在總準確率上，倒傳遞神經網路仍然具有相當的水準，但在相對敏感度指標確有逐漸小幅下滑的趨向。

表 4.20　BPN 之最佳網路架構（英文字母辨識）

BPN 網路架構						
實驗	完整模型			縮減模型		
	學習率	momentum	網路架構	學習率	momentum	網路架構
實驗 1	0.1	0.85	16-11-1	0.1	0.8	8-4-1
實驗 2	0.1	0.85	16-11-1	0.25	0.95	9-5-1
實驗 3	0.1	0.95	16-14-1	0.2	0.95	7-6-1

表 4.21　BPN 特徵選取結果（英文字母辨識）

BPN 特徵選取		
實驗	特徵數	特徵變數
實驗 1	8	V_1、V_6、V_7、V_8、V_9、V_{10}、V_{11}、V_{16}
實驗 2	9	V_2、V_6、V_7、V_8、V_9、V_{10}、V_{11}、V_{12}、V_{16}
實驗 3	7	V_5、V_6、V_8、V_9、V_{12}、V_{13}、V_{16}

表 4.22　BPN 縮減模型分類結果之比較（英文字母辨識）

BPN 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	100.00	95.56	98.52	0.96	90.00	93.33	91.11	1.04
實驗 2	100.00	95.56	99.56	0.96	100.00	93.33	97.78	0.93
實驗 3	100.00	84.44	99.03	0.84	100.00	86.67	95.56	0.87

4.2.5 結論

在英文字母的辨識應用上，MTS 在三個實驗中皆將變數刪減至半數以下，並且與其他分類法的特徵選取比較起來，有較穩定的結果。此外，在 MTS 閾值的訂定上，機率閾值與傳統試誤法比較後，依然在測試上的各項指標有相當好的表現。圖 4.3 為各方法在 3 個實驗中的測試總準確率之比較。圖 4.4 綜合各分類法在三個實驗的測試中，相對敏感度指標的趨勢表現，對實驗 1 的測試結果而言，四種分類法的相對敏感度皆接近 1，顯示當訓練資料中的正常與異常樣本比例呈現 2 比 1 時，對類別預測縮減模型的建構過程仍無造成偏差，因而在判斷正常或異常的能力上皆是相當不錯的，但隨實驗 2 和實驗 3 在訓練類別的數量差距上逐漸拉大，MTS 依然在相對敏感度指標上保持相當穩健，不致有太大的變動，相反地，倒傳遞神經網路則呈現逐漸的下滑，而逐步判別分析及決策樹更是在此指標上明顯遞減。此現象味著，除了 MTS 外，其他分類法所建構的類別預測縮減模型對異常樣本的辨識能力顯著劣於對正常樣本的判斷。因此，即使倒傳遞神經網路在實驗 2 及實驗 3 的測試總準確率比 MTS 略佳，但卻犧牲對異常樣本的

辨識敏感度，而這樣偏差的預測結果並非所期望的。

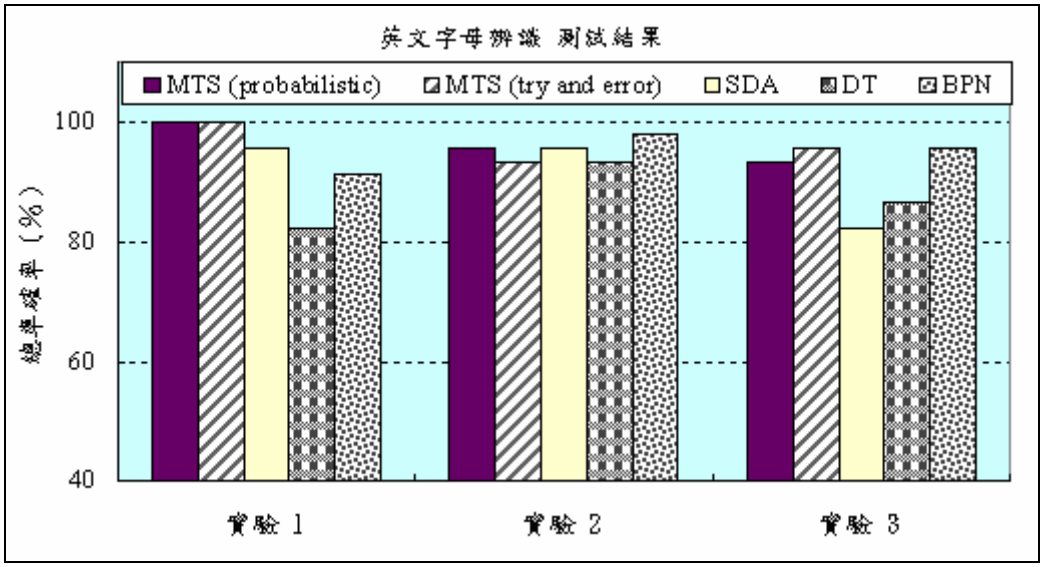


圖 4.3 各分類法測試總準確率比較（英文字母辨識）

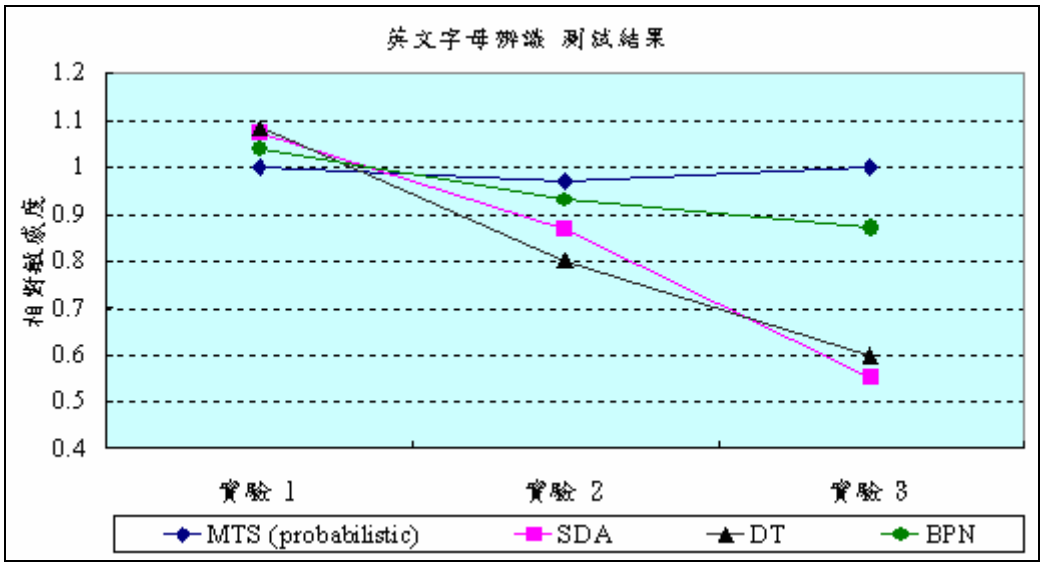


圖 4.4 各分類法測試相對敏感度比較（英文字母辨識）

4.3 心臟病

此資料集原有 13 項特徵變數，分為健康及患有心臟病兩類別，由於在所有樣本的 13 項特徵變數中，有 3 項幾近完全遺漏，因此本例將該 3 項刪除。經過整理，本例之任務相關資料包含 10 項特徵變數（$V_1, V_2, \ldots, V_{10}$）及兩類別，其中，健康的正常樣本為 364 筆，患有心臟病的異常樣本 141 筆。本例所規劃的三個實

驗乃由任務相關資料中隨機抽取數據，其訓練及測試集數量如表 4.23 所示。在本分析中，正常樣本屬於多量類別，並以特效性衡量其類別預測準確率，而異常樣本屬於少量類別，其類別預測準確率則由敏感度表示之。異常樣本在訓練集中所佔比例由實驗 1 至實驗 3 逐漸縮小。

表 4.23 訓練及測試集之數量分佈（心臟病）

訓練集分佈比例 正常：異常		訓練集			測試集		
		正常	異常	總和	正常	異常	總和
實驗 1	2 : 1	60	30	90	30	15	45
實驗 2	6 : 1	180	30	210	30	15	45
實驗 3	10 : 1	300	30	330	30	15	45

4.3.1 MTS 分析結果

MTS 藉由分析每個實驗的訓練資料，篩選出的特徵變數如表 4.24，並利用這些變數建立類別預測縮減模型。在本例中，正常與異常樣本之馬氏距離分佈的重疊情形較先前兩個例子來的顯著，因此可預期 MTS 在總準確率的表現上將不如先前有百分之百正確的現象發生，也因此，如何在兩個重疊的分佈上決定分隔閾值是更為困難且重要的。表 4.25 為在 MTS 中使用兩種閾值訂定法的類別預測比較，在重疊現象較為顯著的本例中，利用所發展的機率閾值作為類別判斷依據，其測試結果明顯優於傳統試誤法，不僅在測試總準確率上表現較好，並且在測試特效性及敏感度上的能力亦是相當的。表 4.26 為 MTS 所建構的類別預測縮減模型（利用機率閾值）之訓練及測試結果，結果顯示實驗 1 的總準確率較低，但在相對敏感度上，三個實驗都相當接近 1，此表示 MTS 在面對不平衡的訓練資料時，並不影響分類模型建構的穩健性。

表 4.24　MTS 特徵選取結果（心臟病）

MTS 特徵選取		
實驗	特徵數	特徵變數
實驗 1	6	V_2、V_5、V_7、V_8、V_9、V_{10}
實驗 2	6	V_2、V_4、V_5、V_8、V_9、V_{10}
實驗 3	6	V_2、V_5、V_7、V_8、V_9、V_{10}

表 4.25 不同閾值下之 MTS 縮減模型分類結果（心臟病）

實驗 1	閾值	訓練集（%）			測試集（%）		
		特效性	敏感度	總準確率	特效性	敏感度	總準確率
機率閾值	1.610	90.00	80.00	86.87	83.33	86.67	84.44
試誤法	1.60	90.00	80.00	86.67	83.33	86.67	84.44
實驗 2	閾值	訓練集（%）			測試集（%）		
		特效性	敏感度	總準確率	特效性	敏感度	總準確率
機率閾值	1.887	91.10	86.67	90.48	90.00	93.33	91.11
試誤法	2.30	94.44	70.00	90.95	93.33	80.00	88.89
實驗 3	閾值	訓練集（%）			測試集（%）		
		特效性	敏感度	總準確率	特效性	敏感度	總準確率
機率閾值	1.964	90.33	83.33	89.70	93.33	93.33	93.33
試誤法	2.90	97.67	43.33	92.73	96.67	53.33	82.22

表 4.26　MTS 縮減模型分類結果之比較（心臟病）

MTS 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	90.00	80.00	86.87	0.89	83.33	86.67	84.44	1.04
實驗 2	91.10	86.67	90.48	0.95	90.00	93.33	91.11	1.04
實驗 3	90.33	83.33	92.73	0.92	93.33	93.33	93.33	1.00

4.3.2　SDA 分析結果

各實驗的特徵選取結果如表 4.27，類別預測縮減模型之訓練及測試結果如表 4.28。經由逐步判別分析所篩選的特徵變數較 MTS 少，但其所建構的縮減模型

在測試時，無論總準確率或相對敏感度指標皆不如 MTS 的表現。本例中，相對敏感度指標顯示，在愈不平衡的訓練資料中學習，將會產生對類別辨識愈大的偏差。

表 4.27　SDA 特徵選取結果（心臟病）

SDA 特徵選取		
實驗	特徵數	特徵變數
實驗 1	4	V_6、V_8、V_9、V_{10}
實驗 2	4	V_3、V_8、V_9、V_{10}
實驗 3	6	V_1、V_3、V_5、V_8、V_9、V_{10}

表 4.28　SDA 縮減模型分類結果之比較（心臟病）

SDA 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	85.00	83.33	84.44	0.98	86.67	73.33	82.22	0.85
實驗 2	90.00	83.33	89.05	0.93	86.67	66.67	80.00	0.77
實驗 3	90.67	80.00	89.70	0.88	86.67	53.33	75.56	0.62

4.3.3　DT 分析結果

表 4.29 為決策樹分析的特徵選取結果，在實驗 3 所篩選的變數明顯多於實驗 1 及實驗 2。表 4.30 中，在所萃取的縮減模型測試上，總準確率及相對敏感度的表現相當不好，此外，隨訓練資料的異常樣本所佔比例由實驗 1 至實驗 3 逐漸縮小，所萃取的分類縮減模型對異常樣本的判別敏感度亦逐漸降低，尤其實驗 3，幾乎已無法正確辨識異常樣本。

表 4.29　DT 特徵選取結果（心臟病）

DT 特徵選取		
實驗	特徵數	特徵變數
實驗 1	5	V_4、V_7、V_8、V_9、V_{10}
實驗 2	5	V_5、V_7、V_8、V_9、V_{10}
實驗 3	8	V_1、V_3、V_4、V_5、V_7、V_8、V_9、V_{10}

表 4.30　DT 縮減模型分類結果之比較（心臟病）

DT 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	96.67	83.33	92.22	0.86	96.67	46.67	80.00	0.48
實驗 2	98.33	83.33	96.19	0.85	96.67	20.00	71.11	0.21
實驗 3	90.00	86.67	97.88	0.96	96.67	13.33	68.89	0.14

4.3.4　BPN 分析結果

倒傳遞神經網路之網路架構經過試誤，最佳完整及縮減模型架構（RMSE 最小）如表 4.31。利用最佳完整模型之網路架構進行特徵選取，結果如表 4.32，所剩下的特徵變數較 MTS 結果略少，並相較於逐步判別分析及決策樹分析的結果而論，其在三個實驗中所篩選的變數並無過大的差異。然而，縱使在特徵選取上有不錯的結果，觀察表 4.33，類別預測縮減模型在測試時的相對敏感度指標卻隨訓練資料之不平衡現象的顯著而逐漸降低，並且在實驗 3 表現的相當差，這表示在倒傳遞神經網路的類別預測模型學習過程中，已受到訓練資料的不平衡分佈所影響。

表 4.31　BPN 之最佳網路架構（心臟病）

BPN 網路架構						
實驗	完整模型			縮減模型		
	學習率	momentum	網路架構	學習率	momentum	網路架構
實驗 1	0.1	0.8	10-9-1	0.1	0.85	5-4-1
實驗 2	0.25	0.9	10-10-1	0.1	0.85	4-7-1
實驗 3	0.15	0.95	10-4-1	0.3	0.9	5-4-1

表 4.32　BPN 特徵選取結果（心臟病）

BPN 特徵選取		
實驗	特徵數	特徵變數
實驗 1	5	V_5、V_6、V_8、V_9、V_{10}
實驗 2	4	V_3、V_8、V_9、V_{10}
實驗 3	5	V_3、V_5、V_8、V_9、V_{10}

表 4.33　BPN 縮減模型分類結果之比較（心臟病）

BPN 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	83.33	73.33	80.00	0.88	90.00	73.33	84.44	0.81
實驗 2	97.78	66.67	93.33	0.68	96.67	46.67	80.00	0.48
實驗 3	98.67	56.67	94.85	0.57	96.67	33.33	75.56	0.35

4.3.5 結論

在心臟病分析預測的應用裡，利用 MTS 所篩選的特徵變數於預測上，其總準確率的表現上十足優於其他三種方法。在 MTS 閾值訂定上，對於正常樣本與異常樣本之馬氏距離分佈有顯著的重疊現象時，採用本研究所提出的機率閾值，比起傳統的試誤法，將可避免對訓練資料的過渡適配，並得到相當好的結果。測試總準確率的匯總比較如圖 4.5。圖 4.6 顯示各分類法在三個實驗中相對敏感度指標的表現，可以很清楚看到，除了 MTS 無顯著的變化，並且皆相當接近 1 外，逐步判別分析、決策樹及倒傳遞神經網路的相對敏感度皆明顯的低於 1，並隨實驗逐漸下降。圖 4.6 中，決策樹分析和逐步判別分析在實驗 3 的表現，相當於只能辨識正常樣本，而對異常樣本的辨識能力極差。因此，除了 MTS，其他三種類別預測方法在模型建構的過程中，容易受到訓練資料分佈型態不平衡的影響，因此使訓練過程發生偏差。

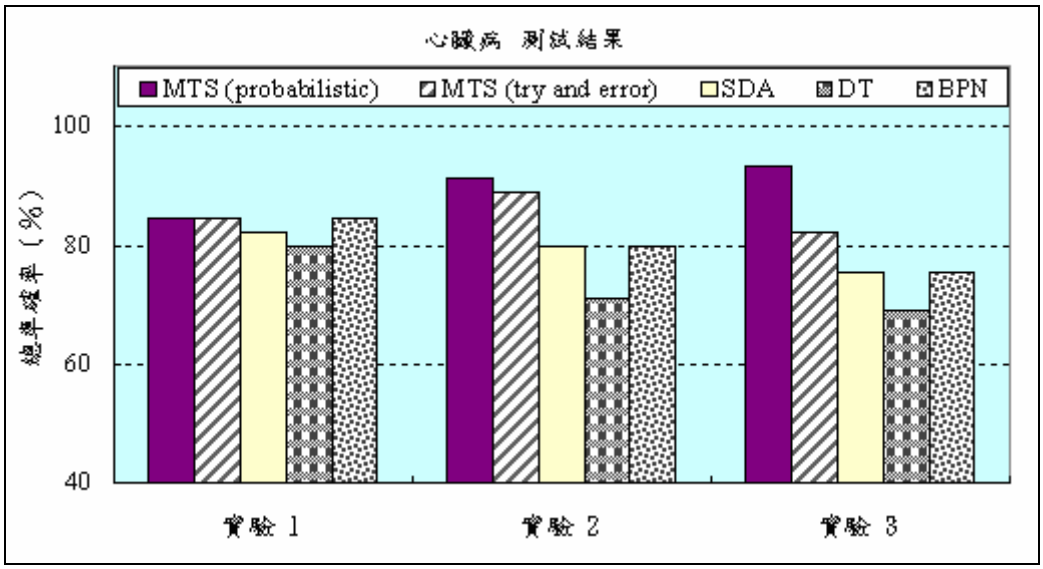


圖 4.5 各分類法測試總準確率比較（心臟病）

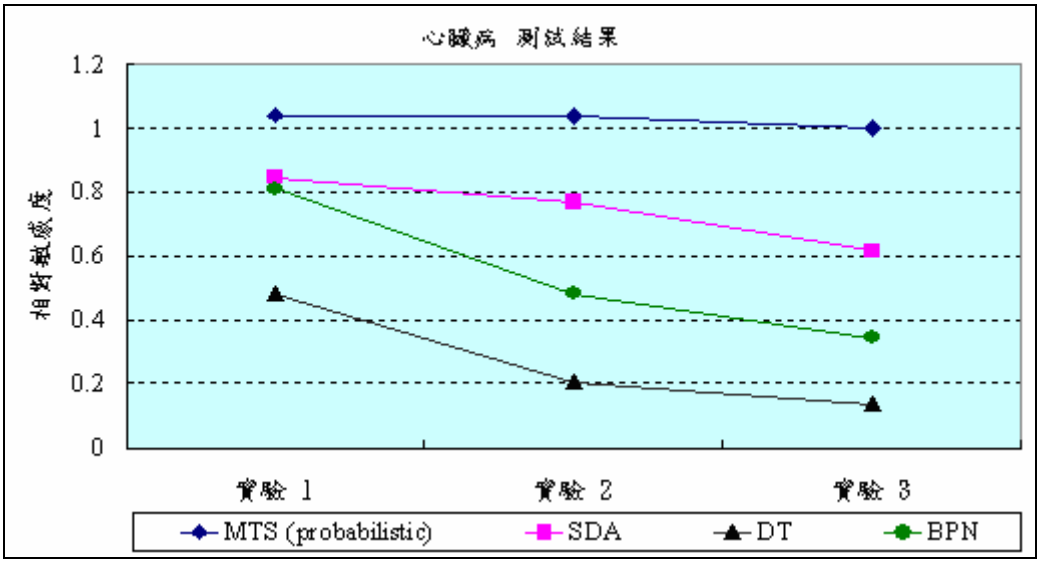


圖 4.6 各分類法測試相對敏感度比較（心臟病）

4.4 評估結論

本章旨在藉助 UCI 資料庫中的資料集來闡述訓練資料的不平衡對分類法在建構類別預測模型時所產生的偏差，以及對日後預測的影響，並同時驗證本研究所提出的 MTS 機率閾值是否可以有較佳的類別判斷效果。經過測試證實，機率閾值確實比一般使用的試誤法可得到較好的判別能力，並且發現，當三個具代表性的分類預測技術（逐步判別分析、決策樹分析及倒傳遞神經網路）在面對不平

衡訓練資料時，不論在特徵選取或類別預測模型的學習上皆較 MTS 容易受到影響，並因而降低其分類的穩健性。

透過觀察每項數據集的三個實驗測試結果，MTS 在相對敏感度指標上皆接近於 1，表示其分類模型在各類別的辨識能力上不因訓練資料分佈型態而干擾，並且在總準確率上，MTS 並不遜色於其他分類法，甚至有更好的表現。相對地，另外三種分類法在測試上的相對敏感度指標卻是隨訓練資料的不平衡程度擴大而下降，這意味著，其分類模型在學習的過程中已受到類別比例差距的影響，也就是在訓練資料中佔有比例愈小的類別，所學習的模型就在辨識該類別的能力上愈不佳。然而，在資料呈現不平衡狀態的領域中，佔有小比例的類別雖然對預測總準確率的評量而言無大幅影響，但卻可能是影響全系統績效的關鍵，例如生產檢測系統若無法正確辨識不良品，將使其流入市面，造成信譽上的損失。因此，一個類別預測技術必須要能夠克服資料的不平衡現象，不能夠因為分析資料的類別數量分佈狀態，而左右模型在建構、預測時的偏向，相對的，必須盡可能在各類別的辨識能力上求取有效且一致的穩健效果。

綜合以上結果，MTS 不因類別分布型態而左右其特徵選取及類別預測能力，即在面對訓練資料的類別分布不平衡時，甚至是數量差距懸殊時，仍能建構出高總準確率的類別預測縮減模型，並維持在各類別的辨識能力，也因此更能適用於各領域的類別預測。

第五章　實例研究

在通訊產業中，個人無線通訊市場正以驚人的速度不斷成長，而行動電信技術亦隨之快速增進。近幾年來，由於雙頻 (GSM/DCS) 行動電話使用者呈現穩定的增長趨勢，加上主要使用族群對其求新、求變的要求，因而促使了消費市場對行動電話的快速淘汰現象。因此，基於達到快速反應市場需求，並從中獲利的目標，行動電話製造商必須設法縮短生產週期時間，提高滿足需求的速度，並降低生產成本，提升市場競爭力。

5.1 案例描述

A 公司成立於民國七十三年，是台灣知名的電腦通訊製造商，該公司以專業的經營團隊、高品質產品的信譽和靈活反應市場的研發能力，使成為全球資訊市場的主要領導者。至民國九十二年止，該公司的營業額達到新台幣一千六百二十二億元，目前全球共有超過一萬名員工。其主要產品包括筆記型電腦、顯示器、行動通訊產品等，並從事代工生產 (ODM)，主要客戶遍及全球。

A 公司之雙頻行動電話生產流程如圖 5.1，其中顯示無線頻率 (radio frequency, RF) 功能檢測所需的作業時間為 190 秒鐘，遠高於其他製程。RF 功能檢測主要針對行動電話的接收/傳送信號進行檢驗，以確認其是否在不同波段 (channel) 及功率位準 (power level) 上滿足 ETI 協定。然而，為了確保行動電話的通訊品質，製造商通常會在製造流程中增加額外的檢測，例如在更多個不同頻率的波段和功率位準上作檢驗，因此導致測試時間擴大，並使該檢測試製程成為了整個生產流程中的瓶頸。

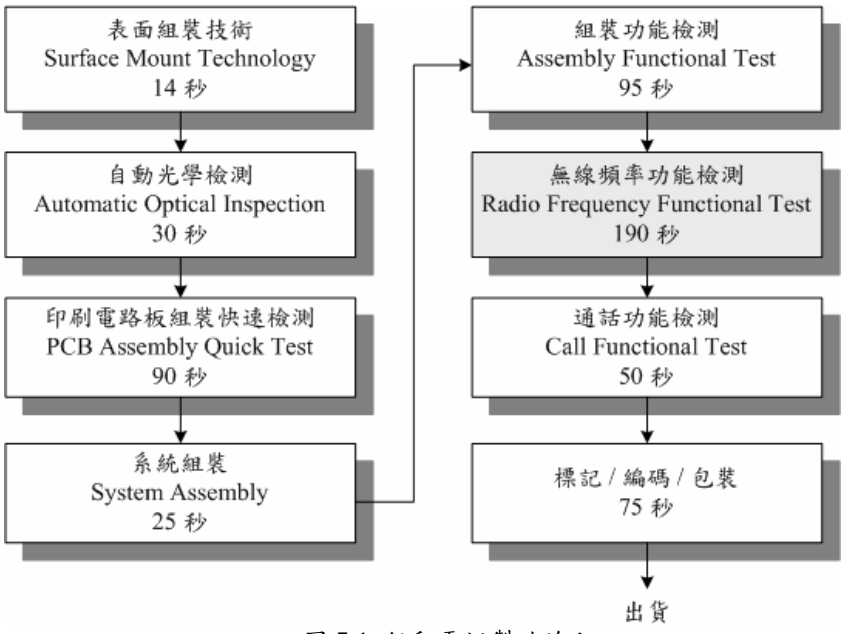


圖 5.1 行動電話製造流程

RF 功能檢測包含 8 個檢測項目，其名稱及代號如表 5.1。每個檢測項目根據不同的波段及功率位準擴充為數個檢測屬性，並以「檢測項目-波段-功率位準」的形式表達，最終檢測屬性為 62 項，其細目及代號如表 5.2，通過此 62 項檢測者即為良品。本案例之研究對象為 A 公司在台灣桃園廠的行動電話 RF 功能檢測製程，該製程的檢測結果在良品與不良品的平均比例上約為 25 比 1，此即為典型的非平衡型態資料。本案例的主要目的為利用馬氏-田口系統（MTS），針對該檢測製程進行研究，期望在保持原有檢測準確率下，縮減此製程原有的 62 項檢測屬性，以降低作業時間與成本。

表 5.1 RF 功能檢測項目

代號	檢測項目	代號	檢測項目
B	Power level (TXP)	F	ORFS-spectrum due to switching transient (ORFS_SW)
C	Phase error and frequency error (PEFR)	G	ORFS-spectrum due to modulation (ORFS_MO)
D	Bit error rate (-20) (BER-20)	H	Rx level report accuracy (RXP_Lev_Err)
E	Bit error rate (-102) (BER-102)	I	Rx level report quality (RXP_QUALITY)

表 5.2　RF 功能檢測屬性

代號	檢測項	檢測屬性	代號	檢測項	檢測屬性
C_1	TXP	B-10-5	C_{32}	ORFS_SW	F-965-5
C_2	PEFR	C-10-5	C_{33}	ORFS_MO	G-965-5
C_3	BER-20	D-10-5	C_{34}	RXP_Lev_Err	H-965-102
C_4	BER-102	E-10-5	C_{35}	RXP_QUALITY	I-965-102
C_5	ORFS_SW	F-10-5	C_{36}	TXP	B-522-0
C_6	ORFS_MO	G-10-5	C_{37}	PEFR	C-522-0
C_7	RXP_Lev_Err	H-10-102	C_{38}	BER-20	D-522-0
C_8	RXP_QUALITY	I-10-102	C_{39}	BER-102	E-522-0
C_9	TXP	B-72-5	C_{40}	ORFS_SW	F-522-0
C_{10}	PFER	C-72-5	C_{41}	ORFS_MO	G-522-0
C_{11}	BER-20	D-72-5	C_{42}	RXP_Lev_Err	H-522-102
C_{12}	BER-120	E-72-5	C_{43}	RXP_QUALITY	I-522-102
C_{13}	ORFS_SW	F-72-5	C_{44}	TXP	B-688-0
C_{14}	ORFS_MO	G-72-5	C_{45}	PFER	C-688-0
C_{15}	TXP	B-72-7	C_{46}	BER-20	D-688-0
C_{16}	TXP	B-72-11	C_{47}	BER-102	E-688-0
C_{17}	TXP	B-72-19	C_{48}	ORFS_SW	F-688-0
C_{18}	RXP_Lev_Err	H-72-102	C_{49}	ORFS_MO	G-688-0
C_{19}	RXP_QUALITY	I-72-102	C_{50}	TXP	B-688-3
C_{20}	TXP	B-114-5	C_{51}	TXP	B-688-7
C_{21}	PFER	C-114-5	C_{52}	TXP	B-688-15
C_{22}	BER-20	D-114-5	C_{53}	RXP_Lev_Err	H-688-102
C_{23}	BER-102	E-114-5	C_{54}	RXP_QUALITY	I-688-102
C_{24}	ORFS_SW	F-114-5	C_{55}	TXP	B-875-0
C_{25}	ORFS_MO	G-114-5	C_{56}	PEFR	C-875-0
C_{26}	RXP_Lev_Err	H-114-102	C_{57}	BER-20	D-875-0
C_{27}	RXP_QUALITY	I-114-102	C_{58}	BER-102	E-875-0
C_{28}	TXP	B-965-5	C_{59}	ORFS_SW	F-875-0
C_{29}	PFER	C-965-5	C_{60}	ORFS_MO	G-875-0
C_{30}	BER-20	D-965-5	C_{61}	RXP_Lev_Err	H-875-102
C_{31}	BER-102	E-965-5	C_{62}	RXP_QUALITY	I-875-102

5.2 MTS 之執行

在本案例分析中，定義良品為正常狀態，不良品為異常狀態。由檢測製程中隨機抽取二組檢測結果數據，將第一組數據 300 筆作為訓練集，並藉以建立模型；第二組數據 100 筆作為測試樣本，以幫助瞭解該模型的類別預測效果。收集數據如表 5.3。

表 5.3 樣本資料（RF 功能檢測）

	正常	異常	總和
訓練集	270	30	300
測試集	90	10	100

1. 構建完整模型（full model）之量測尺度

將訓練集中的正常群體作為參照群體，並建構馬氏空間，馬氏空間可視為包含該正常群體之屬性平均值、標準差及相關反矩陣的資料庫。首先，計算訓練集中正常樣本之各特性變數的平均值、標準差，結果如表 5.4，隨後對每個屬性資料進行標準化，標準化後數據如表 5.5。

表 5.4 訓練集之正常樣本原始數據（RF 功能檢測）

樣本 \ 屬性	C_1	C_2	C_3	C_4	...	C_{61}	C_{62}
1	32.220	1.117	0	0	...	1	0
2	32.191	1.076	0	0.014	...	1	1
3	32.411	1.555	0	0.014	...	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
270	32.204	1.094	0	0.029	...	1	1
平均值 $\overline{x_i}$	32.22	1.467	0.003	0.021	...	0.652	0.511
標準差 s_i	0.082	0.24	0.019	0.017	...	0.493	0.523

表 5.5 訓練集之正常樣本標準化數據及馬氏距離（RF 功能檢測）

樣本 \ 屬性	C_1	C_2	C_3	C_4	...	C_{61}	C_{62}	MD
1	0	-1.458	-0.158	-1.235	...	0.706	-0.977	1.07470
2	-0.354	-1.629	-0.158	-0.412	...	0.706	0.935	0.91999
3	2.329	0.367	-0.158	-0.412	...	-1.323	0.935	1.09532
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
270	-0.195	-1.554	-0.158	0.471	...	0.706	0.935	0.99650

接著利用標準化值計算檢測屬性間的相關反矩陣，結果如表 5.6。最後以式（1）計算正常樣本之馬氏距離，例如正常樣本 1 之馬氏距離計算如下：

$$MD_1 = \frac{1}{62}\begin{bmatrix}0 & -1.458 & \cdots & -0.977\end{bmatrix}_{1\times 62}\begin{bmatrix}16.69 & -1.35 & \cdots & 0.008\\ -1.35 & 3.013 & \cdots & 0.214\\ \vdots & \vdots & \ddots & -0.069\\ 0.008 & 0.214 & \cdots & 1.457\end{bmatrix}_{62\times 62}\begin{bmatrix}0\\ -1.458\\ \vdots\\ -0.977\end{bmatrix}_{62\times 1}$$

$$=1.0747$$

表 5.6 訓練集之正常樣本相關反矩陣（RF 功能檢測）

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	...	C_{58}	C_{59}	C_{60}	C_{61}	C_{62}
C_1	16.69	-1.35	0.254	-0.033	0.182	0.008	-0.565	...	0.038	0.224	-0.311	-0.078	0.008
C_2	-1.35	3.013	-0.073	0.107	0.109	0.035	-0.203	...	0.032	-0.123	0.041	-0.021	0.214
C_3	0.254	-0.073	1.17	0.107	-0.096	0.07	0.006	...	5E-04	0.031	0.023	-0.057	-0.069
C_4	-0.033	0.107	0.107	1.301	-0.037	-0.158	-0.036	...	0.029	0.289	0.035	0.193	0.024
C_5	0.182	0.109	-0.096	-0.037	1.734	-0.206	-0.146	...	0.12	-0.099	-0.133	-0.054	-0.022
C_6	0.008	0.035	0.07	-0.158	-0.206	1.793	0.137	...	-0.054	-0.267	-0.191	-0.04	-0.160
C_7	-0.565	-0.203	0.006	-0.036	-0.146	0.137	1.575	...	0.169	0.048	-0.023	-0.235	0.075
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
C_{58}	0.038	0.032	5E-04	0.029	0.12	-0.054	0.169	...	1.565	0.238	-0.1	0.016	-0.474
C_{59}	0.224	-0.123	0.031	0.289	-0.099	-0.267	0.048	...	0.238	1.539	-0.1	0.25	-0.004
C_{60}	-0.311	0.041	0.023	0.035	-0.133	-0.191	-0.023	...	-0.1	-0.1	1.35	0.049	0.089
C_{61}	-0.078	-0.021	-0.057	0.193	-0.054	-0.04	-0.235	...	0.016	0.25	0.049	1.444	-0.018
C_{62}	0.008	0.214	-0.069	0.024	-0.022	-0.16	0.075	...	-0.474	-0.004	0.089	-0.018	1.457

2. 確認完整模型之量測尺度

將訓練集中的 30 筆異常樣本，利用馬氏空間進行標準化，並計算其馬氏距離。如果階段一所構建的尺度是好的，那麼異常樣本的馬氏距離將會較大。將訓練集的正常樣本 270 筆及異常樣本 30 筆之馬氏距離分別繪製於次數分配圖，如圖 5.2，由圖中的兩個分配可觀察到，異常樣本的馬氏距離確實遠大於正常樣本。同樣地，我們以相同方式，利用所建立的馬氏空間計算測試樣本的馬氏距離，並繪成次數分配圖，如圖 5.3，其異常樣本之馬氏距離亦遠大於正常樣本。透過此階段，證實階段一所建構的完整模型之量測尺度是有效的。

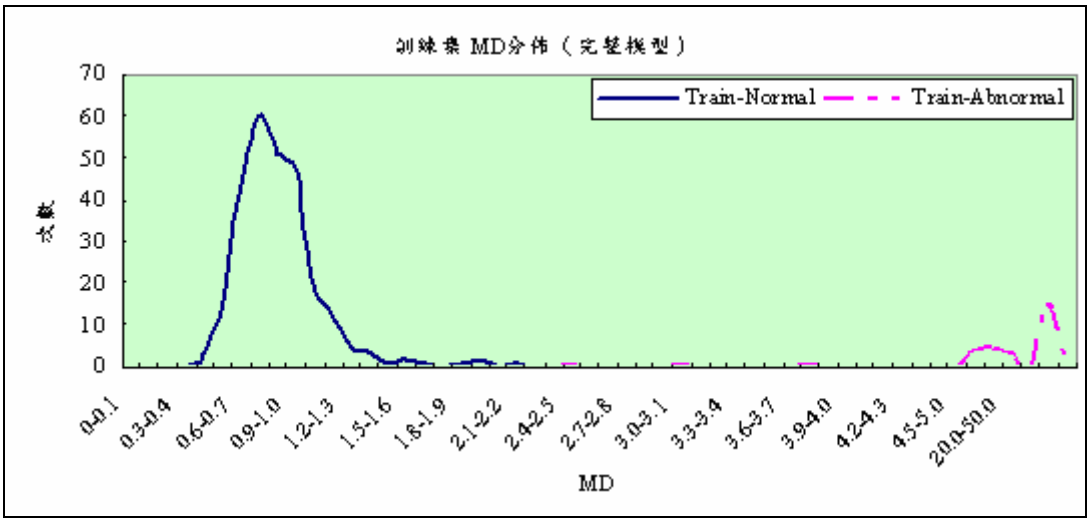


圖 5.2 訓練樣本之馬氏距離分布圖（RF 功能檢測之完整模型）

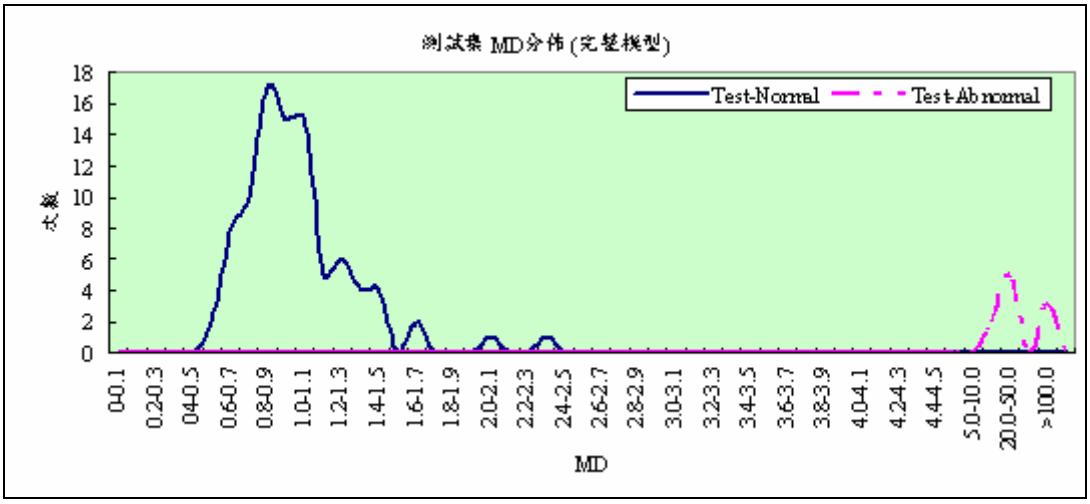


圖 5.3 測試樣本之馬氏距離分配圖（RF 功能檢測之完整模型）

3. 確認重要特性變數

將 62 項檢測屬性視為控制因子，每因子設 2 水準：水準 1 為使用該項屬性，水準 2 為不使用該項屬性；將之配置於 $L_{64}(2^{63})$ 直交表。接著，利用訓練集中的 30 筆異常樣本計算其對應的 30 個馬氏距離，並以式（6）計算望大特性SN比。因子配置及SN比如表 5.7，以Run1 為例，其 $MD_1 = 7.98674$ 乃是由訓練集中的第一筆異常樣本使用 62 項屬性所計算而得的馬氏距離，而SN比計算如下：

$$\eta_1 = -10 \cdot \log_{10}\left[\frac{1}{30}\cdot\left(\frac{1}{7.98674}+\cdots+\frac{1}{564802}\right)\right] = 18.02903$$

表 5.7 直交表配置與 SN 比（RF 功能檢測）

Run	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	...	C_{60}	C_{61}	C_{62}		MD_1	…	MD_{30}	SN ratio
	1	2	3	4	5	6	7	8	9	10	…	60	61	62	63				η（dB）
1	1	1	1	1	1	1	1	1	1	1	...	1	1	1	1	7.98674	…	564802	18.02903
2	1	1	1	1	1	1	1	1	1	1	...	2	2	2	2	12.9013	…	978550	7.608896
3	1	1	1	1	1	1	1	1	1	1	...	2	2	2	2	1.50223	…	339314	12.57003
4	1	1	1	1	1	1	1	1	1	1	...	1	1	1	1	1.60727	…	505956	14.05007
5	1	1	1	1	1	1	1	2	2	2	...	2	2	2	2	1.27893	…	752904	12.57109
6	1	1	1	1	1	1	1	2	2	2	...	1	1	1	1	1.60313	…	622553	11.40608
7	1	1	1	1	1	1	1	2	2	2	...	1	1	1	1	12.6725	…	39027.3	12.94864
8	1	1	1	1	1	1	1	2	2	2	...	2	2	2	2	11.873	…	188601	16.63614
9	1	1	1	2	2	2	2	1	1	1	...	2	2	2	2	1.42267	…	978514	12.81961
10	1	1	1	2	2	2	2	1	1	1	...	1	1	1	1	1.68229	…	117000	13.28187
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
60	2	2	1	2	1	1	2	1	2	2	...	2	1	1	2	1.01236	..	489627	10.97286
61	2	2	1	2	1	1	2	2	1	1	...	1	2	2	1	1.55941	…	778281	12.66432
62	2	2	1	2	1	1	2	2	1	1	...	2	1	1	2	1.34182	…	602506	13.4991
63	2	2	1	2	1	1	2	2	1	1	...	2	1	1	2	9.97475	…	38333	11.44389
64	2	2	1	2	1	1	2	2	1	1	...	1	2	2	1	9.29266	…	194704	13.35158

之後，利用式（11）計算各檢測屬性的效果增量，其結果如圖 5.4。以檢測屬性 C_1 的效果增量為例，計算如下：

$$\overline{SN_1^+} = \frac{1}{32}\cdot(18.02903+7.608896+\cdots+11.64836) = 12.91363$$

$$\overline{SN_1^-} = \frac{1}{32}\cdot(14.42624+12.68887+\cdots+13.35158) = 12.75443$$

$$Gain_1 = \overline{SN_1^+} - \overline{SN_1^-} = 12.91363-12.75443 = 0.1592$$

最後，依增量愈大愈好的原則來評估各檢測屬性的重要程度，並作為特徵選取的依據。

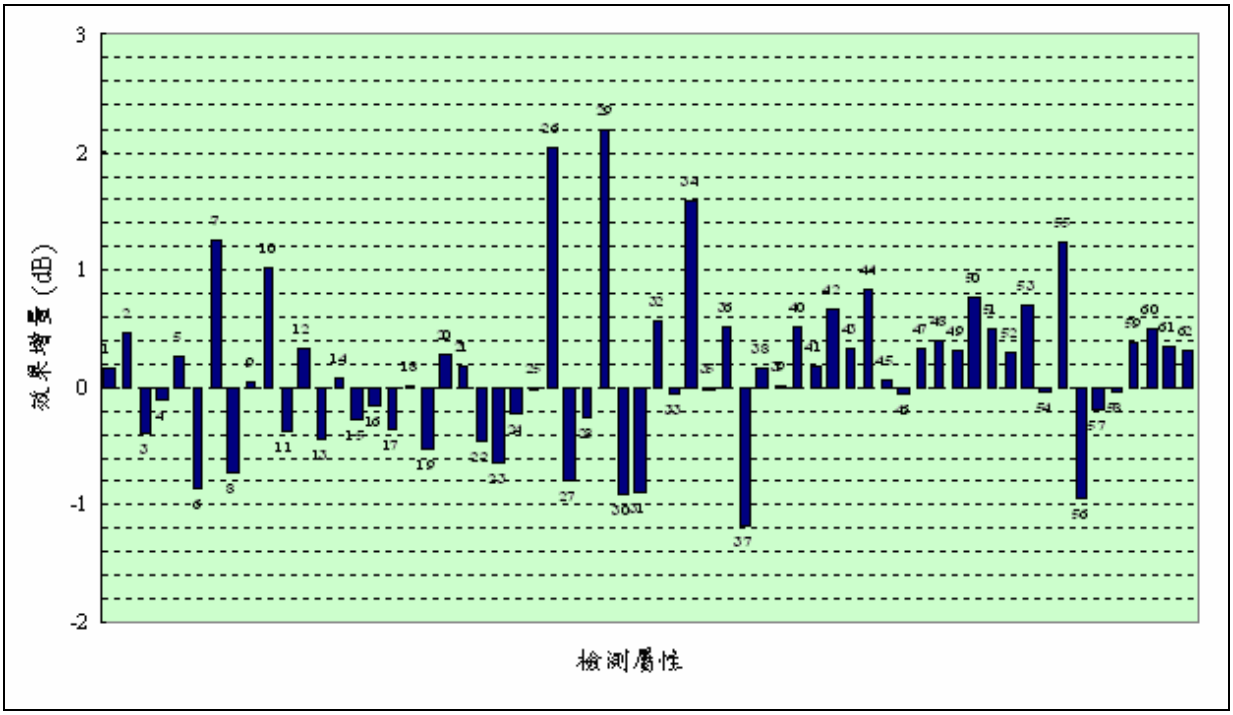


圖 5.4 屬性效果增量圖（RF 功能檢測）

4. 利用重要的特性變數作分類診斷

依據不同效果增量的篩選標準（>0、>0.1、>0.2、>0.3、>0.4、>0.5、>0.6、>0.7、>0.8、>0.9），原來的 62 項檢測屬性將依序減少為 36、31、27、24、17、14、10、8、7、6 項，再分別以其建立縮減模型，並選用類別預測總準確率最高者為最後特徵選取之結果。

以增量效果大於 0.5 為例，其檢測屬性縮減為 14 項，如表 5.8。在訓練集中，利用正常樣本的該 14 項屬性建立馬氏空間，並以異常樣本做縮減模型的尺度確認，圖 5.5 為訓練集利用該馬氏空間所得到的馬氏距離分佈，由兩群體的分離可以很容易的做尺度確認。此外，可以同樣 30 筆異常樣本計算在縮減模型的馬氏距離，並求得一個 SN 比，然後與原始完整模型比較之，以 SN 比在縮減模型的增量來評估系統在功能上的改善，表 5.9 顯示縮減模型所得到的 SN 比較完整模型高，即表示透過 MTS 的系統最佳化過程，使異常樣本與正常樣本的分離程度獲得了改善。在確認量測尺度為可接受後，計算機率閾值作為往後分類診斷的依據。根據式（5）計算得到機率閾值為 2.724，其計算過程如下：

$$T = \overline{md} + \sqrt{\frac{100}{100+\lambda-h}} \cdot s_{md} = 0.982 + \sqrt{\frac{100}{100+5-100}} \cdot 0.393 = 2.724$$

以此閾值對訓練集做類別判斷可得到 100%的正確性。最後，利用測試集進行測試，以驗證方才所建立的縮減模型是否有足夠的分類能力，圖 5.6 為測試樣本利用馬氏空間所得到的馬氏距離分佈，經過閾值的判斷，其分類結果之總準確率亦為 100%。

表 5.8 特徵選取結果（RF 功能檢測）

代號	檢測項	檢測屬性	代號	檢測項	檢測屬性
C_7	RXP_Lev_Err	H-10-102	C_{40}	ORFS_SW	F-522-0
C_{10}	PFER	C-72-5	C_{42}	RXP_Lev_Err	H-522-102
C_{26}	RXP_Lev_Err	H-114-102	C_{44}	TXP	B-688-0
C_{29}	PFER	C-965-5	C_{50}	TXP	B-688-3
C_{32}	ORFS_SW	F-965-5	C_{53}	RXP_Lev_Err	H-688-102
C_{34}	RXP_Lev_Err	H-965-102	C_{55}	TXP	B-875-0
C_{36}	TXP	B-522-0	C_{60}	ORFS_MO	G-875-0

表 5.9 系統改善結果（RF 功能檢測）

模型	屬性數	SN 比（dB）
縮減模型	14	19.31514
完整模型	62	18.02903
SN 比增量 ＝ +1.68211（dB）		

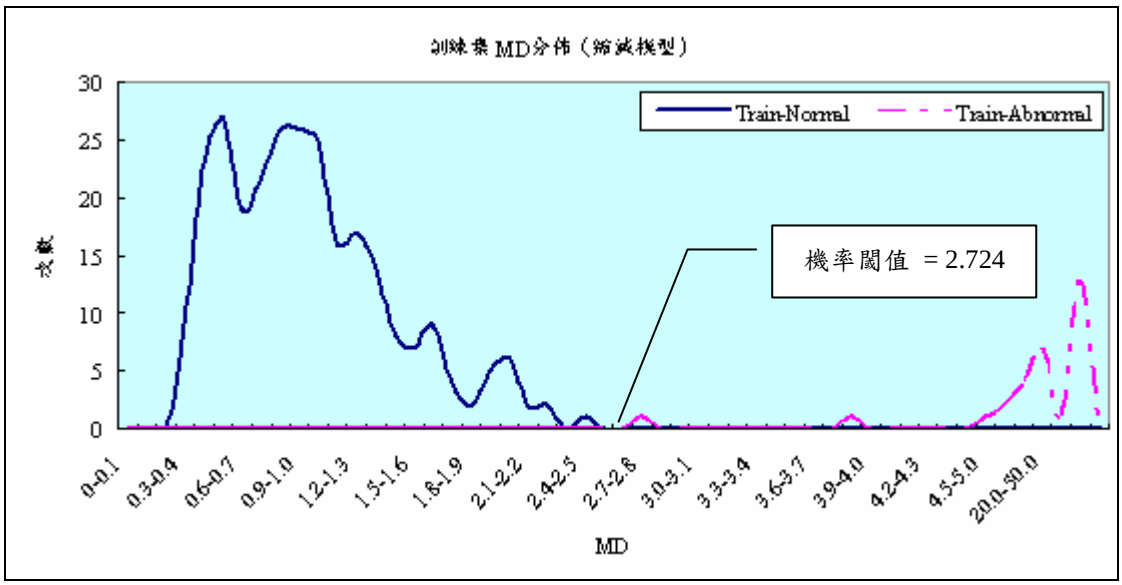


圖 5.5 訓練樣本之馬氏距離分布圖（RF 功能檢測之縮減模型）

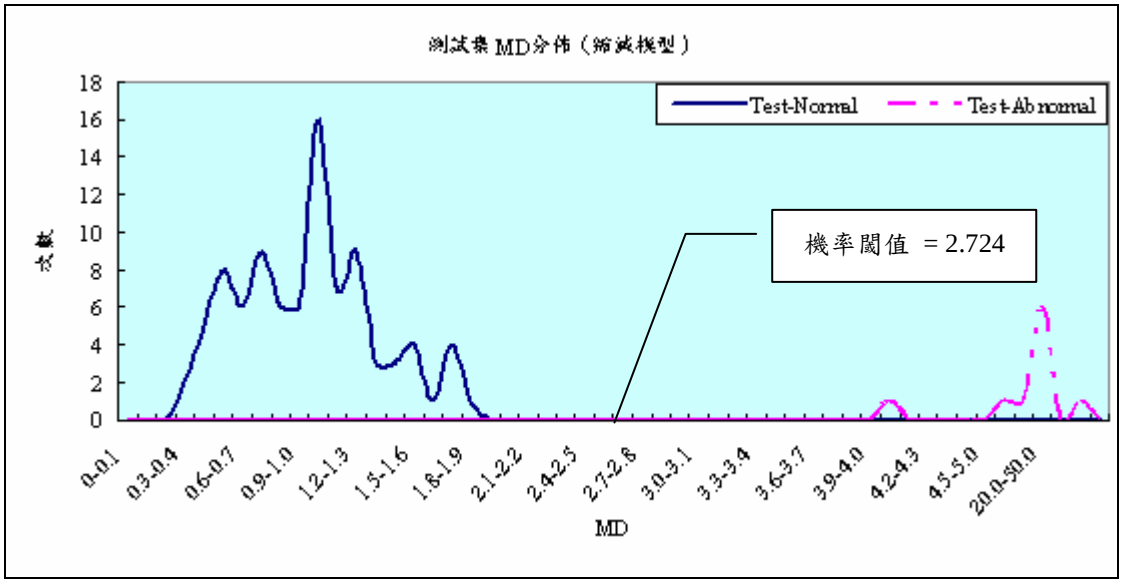


圖 5.6 測試樣本之馬氏距離分布圖（RF 功能檢測之縮減模型）

依不同效果增量篩選標準所得的完整結果與比較如表 5.10。顯而易見地，當本案例分析個別以效果增量大於 0.2、0.3、0.4 及 0.5 作為檢測屬性的篩選標準時，其訓練及測試總準確率皆高達百分之百。然而，本例應選用增量大於 0.5 為最後篩選標準，可使原始的 62 項檢測屬性減少為較少的 14 項。

表 5.10 效果增量之結果比較（RF 功能檢測）

增量	屬性數	機率閾值	總準確率（%）	
			訓練	測試
>0	36	1.994	99.67	99.00
>0.1	31	2.0345	99.00	99.00
>0.2	27	2.099	100.00	100.00
>0.3	24	2.1388	100.00	100.00
>0.4	17	2.5479	100.00	100.00
>0.5	14	2.724	100.00	100.00
>0.6	10	2.745	99.33	99.00
>0.7	8	3.4866	99.33	98.00
>0.8	7	1.791	87.00	90.00
>0.9	6	1.9867	87.67	89.00

5.3 改善效益

本案例藉助 MTS 在特徵選取與類別預測上的穩健能力，利用其分析流程來減少行動電話生產過程中冗餘的 RF 功能檢測屬性，以期降低生產時間，增加市場競爭力。經過 MTS 分析，RF 功能檢測屬性由原先的 62 項減少為 14 項，並且仍然在檢測上保有高準確率。透過屬性縮減，通過該製程所需時間亦由 190 秒降低為 95 秒，由於瓶頸產能的增加，生產線的產出由改善前的每小時 18 個提升為每小時 37 個，整體生產效率增為以往的 2 倍。此外，由於檢測屬性的減少，所需的檢測機台由原先的 8 台減少為 4 台，其除了節省機台設置成本約新台幣六百萬元外，更減少機台維護及相關人事費用。

透過對 RF 檢測製程的改善，對內不但降低生產作業費用、節省固定成本支出，並促使瓶頸產能增加、生產線流暢性提升，且促使減少在製品存貨成本；對外則由於生產週期時間縮短，將致使更能適應於行動電話的短生命週期特性及提高對市場需求的反應速度。

5.4 分類法比較結果

為了顯現 MTS 在實務應用上的穩健性，本節中同樣採用三組實驗來表現不同的資料不平衡程度，並將 MTS 的分析結果與逐步判別分析、決策樹分析和倒傳遞神經網路比較之，其中，實驗 1 的數據即為 5.2 節所使用的資料集。在本分析中，正常樣本屬於多量類別，並以特效性衡量其類別預測準確率，而異常樣本屬於少量類別，其類別預測準確率則由敏感度表示之。異常樣本在訓練集中所佔比例由實驗 1 至實驗 3 逐漸縮小。三實驗數據如表 5.11。

表 5.11 訓練及測試集之數量分佈（RF 功能檢測）

訓練資料分佈比例 正常：異常		訓練集			測試集		
		正常	異常	總和	正常	異常	總和
實驗 1	9 : 1	270	30	300	90	10	100
實驗 2	15 : 1	450	30	480	90	10	100
實驗 3	27 : 1	810	30	840	90	10	100

5.4.1 MTS 分析結果

表 5.12 為三實驗的屬性篩選結果，利用這些屬性來建立類別預測縮減模型，並建立機率閾值，而其表現在測試集的總準確率上較試誤法略佳，如表 5.13。觀察表 5.14 的測試結果，MTS 在三個實驗裡利用不平衡程度不等的訓練資料所建立的縮減模型，皆能有效的預測診斷出正常樣本與異常樣本，即使其間數量差距非常大，亦不對異常樣本的判斷造成影響。

表 5.12　MTS 特徵選取結果（RF 功能檢測）

MTS 特徵選取		
實驗	特徵數	特徵變數
實驗 1	14	C_7、C_{10}、C_{26}、C_{29}、C_{32}、C_{34}、C_{36}、C_{40}、C_{42}、C_{44}、C_{50}、C_{53}、C_{55}、C_{60}
實驗 2	15	C_7、C_{10}、C_{18}、C_{20}、C_{21}、C_{23}、C_{26}、C_{29}、C_{32}、C_{34}、C_{36}、C_{50}、C_{53}、C_{54}、C_{55}
實驗 3	15	C_2、C_7、C_{10}、C_{14}、C_{18}、C_{23}、C_{26}、C_{29}、C_{32}、C_{34}、C_{36}、C_{40}、C_{49}、C_{50}、C_{60}

表 5.13 不同閾值下之 MTS 縮減模型分類結果（RF 功能檢測）

實驗 1	閾值	訓練集（%）			測試集（%）		
		特效性	敏感度	總準確率	特效性	敏感度	總準確率
機率閾值	2.724	100.00	96.67	99.67	100.00	100.00	100.00
試誤法	2.50	100.00	100.00	100.00	100.00	100.00	100.00
實驗 2	閾值	訓練集（%）			測試集（%）		
		特效性	敏感度	總準確率	特效性	敏感度	總準確率
機率閾值	2.628	98.89	96.67	98.75	98.89	100.00	99.00
試誤法	3.50	99.63	93.33	99.00	100.00	90.00	99.00
實驗 3	閾值	訓練集（%）			測試集（%）		
		特效性	敏感度	總準確率	特效性	敏感度	總準確率
機率閾值	2.621	99.01	100.00	99.05	100.00	90.00	99.00
試誤法	3.00	99.63	100.00	99.67	100.00	80.00	98.00

表 5.14 MTS 縮減模型分類結果之比較（RF 功能檢測）

MTS 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
實驗 2	98.89	96.67	98.75	0.98	98.89	100.00	99.00	1.01
實驗 3	99.01	100.00	99.05	1.01	100.00	90.00	99.00	0.90

5.4.2 SDA 分析結果

表 5.15 為逐步判別分析在三組實驗數據中所篩選的重要檢測屬性，隨著資料不平衡程度的擴大，屬性數亦有明顯遞增的情形。利用篩選的屬性來建立類別預測縮減模型，其類別預測的測試結果如表 5.16，其中特效性指標顯示對正常樣本的判斷相當準確，而在敏感度指標上卻僅有四成的正確性，也就是該縮減模型對於異常樣本的診斷能力是不足的。

表 5.15　SDA 特徵選取結果（RF 功能檢測）

SDA 特徵選取		
實驗	特徵數	特徵變數
實驗 1	16	C_1、C_{14}、C_{19}、C_{24}、C_{26}、C_{29}、C_{32}、C_{34}、C_{41}、C_{42}、C_{44}、C_{45}、C_{46}、C_{50}、C_{56}、C_{61}
實驗 2	18	C_4、C_5、C_{10}、C_{22}、C_{23}、C_{26}、C_{27}、C_{32}、C_{34}、C_{36}、C_{37}、C_{42}、C_{47}、C_{48}、C_{50}、C_{54}、C_{55}、C_{61}
實驗 3	25	C_2、C_5、C_9、C_{14}、C_{15}、C_{17}、C_{19}、C_{23}、C_{26}、C_{29}、C_{30}、C_{34}、C_{35}、C_{36}、C_{41}、C_{45}、C_{46}、C_{49}、C_{50}、C_{51}、C_{54}、C_{55}、C_{56}、C_{57}、C_{60}

表 5.16　SDA 縮減模型分類結果之比較（RF 功能檢測）

SDA 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	100.00	80.00	98.00	0.80	100.00	40.00	94.00	0.40
實驗 2	100.00	80.00	98.00	0.80	96.67	40.00	91.00	0.41
實驗 3	100.00	86.67	99.52	0.87	100.00	40.00	94.00	0.40

5.4.3 DT 分析結果

決策樹分析在屬性篩選後所剩下的檢測屬性數量遠小於其它分類法，如表 5.17。觀察表 5.18，決策樹分析由三組不平衡程度不同的訓練資料來萃取分類縮減模型，在經過測試集的測試後發現，三個實驗對正常樣本的判斷正確性皆相當高，但對異常樣本的診斷能力僅在實驗 1 中達到六成的準確性，實驗 2 及實驗 3 卻只有三成。因此，縱使決策樹分析在特徵選取能力上表現不錯，但在面臨不平衡資料時，其萃取分類模型的穩健性仍是不足的。

表 5.17　SDA 特徵選取結果（RF 功能檢測）

SDA 特徵選取		
實驗	特徵數	特徵變數
實驗 1	6	C_1、C_7、C_{17}、C_{23}、C_{29}、C_{31}
實驗 2	8	C_{10}、C_6、C_{21}、C_{23}、C_{38}、C_{44}、C_{51}、C_{62}
實驗 3	8	C_1、C_{14}、C_{21}、C_{23}、C_{16}、C_{29}、C_{31}、C_{61}

表 5.18　SDA 縮減模型分類結果之比較（RF 功能檢測）

SDA 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	100.00	80.00	98.00	0.80	100.00	60.00	96.00	0.60
實驗 2	99.78	93.33	93.38	0.94	96.67	30.00	90.00	0.31
實驗 3	99.88	76.67	99.05	0.77	100.00	30.00	93.00	0.30

5.4.4 BPN 分析結果

倒傳遞神經網路之網路架構經過試誤，最佳完整及縮減模型架構（RMSE 最小）如表 5.19。利用最佳完整模型之網路架構進行屬性篩選後，其所剩於屬性量仍然多達 20 個，結果如表 5.20。表 5.21 為縮減模型的訓練及測試結果，觀察測試結果，在特效性指標上皆顯示對正常樣本的判斷準確率達到了百分之百，但在敏感度指標上卻相對偏低，顯示對異常樣本的判斷準確率最高僅有六成。

表 5.19　BPN 之最佳網路架構（RF 功能檢測）

BPN 網路架構						
實驗	完整模型			縮減模型		
	學習率	momentum	網路架構	學習率	momentum	網路架構
實驗 1	0.2	0.95	62-32-1	0.3	0.9	22-12-1
實驗 2	0.25	0.95	62-34-1	0.15	0.95	21-13-1
實驗 3	0.25	0.95	62-36-1	0.25	0.95	20-11-1

表 5.20　BPN 特徵選取結果（RF 功能檢測）

BPN 特徵選取		
實驗	特徵數	特徵變數
實驗 1	22	C_1、C_2、C_6、C_{10}、C_{14}、C_{17}、C_{21}、C_{24}、C_{26}、C_{29}、C_{34}、C_{41}、C_{42}、C_{43}、C_{44}、C_{45}、C_{46}、C_{53}、C_{55}、C_{56}、C_{60}、C_{61}
實驗 2	21	C_2、C_5、C_{10}、C_{13}、C_{15}、C_{17}、C_{23}、C_{26}、C_{28}、C_{29}、C_{32}、C_{33}、C_{34}、C_{38}、C_{39}、C_{45}、C_{52}、C_{53}、C_{54}、C_{55}、C_{61}
實驗 3	20	C_1、C_2、C_9、C_{10}、C_{16}、C_{17}、C_{21}、C_{23}、C_{24}、C_{26}、C_{28}、C_{29}、C_{32}、C_{34}、C_{42}、C_{45}、C_{46}、C_{53}、C_{55}、C_{60}

表 5.21　BPN 縮減模型分類結果之比較（RF 功能檢測）

BPN 縮減模型分類結果								
實驗	訓練結果				測試結果			
	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度	特效性（%）	敏感度（%）	總準確率（%）	相對敏感度
實驗 1	100.00	100.00	100.00	1.00	100.00	60.00	96.00	0.60
實驗 2	100.00	96.67	99.79	0.97	100.00	60.00	96.00	0.60
實驗 3	100.00	96.67	99.88	0.97	100.00	50.00	95.00	0.50

5.4.5 分類法比較結果

圖 5.7 為各方法在 3 個實驗中的測試總準確率比較，很明顯的，利用 MTS 可擁有較佳的類別預測結果。圖 5.8 表現各分類法在三個實驗的相對敏感度指標，由於此分析在實驗 1 中所取的訓練資料已相當不平衡（正常與異常比為 9 比 1），因此，除了 MTS 外，其它分類法的相對敏感度指標在實驗 1 上已表現得相當差，顯示所建構的分類縮減模型已受訓練資料的影響而偏重於對正常樣本的判斷。在實驗 3 中，訓練資料的類別分佈差距已擴大為 27 比 1，然而，根據圖 5.7 及 5.8，只有 MTS 在保持高總準確率下，亦能使相對敏感度指標維持於 1 附近，即兼具對正常樣本及異常樣本的辨識能力，其餘如決策樹分析已幾乎無法辨識不良品。此外，在 MTS 閾值的訂定上，採用本研究所提出機率閾值與傳統試誤法比較起來，在各項指標皆有較佳的表現。因此，藉由此實案分析，更加證實 MTS 對於不平衡資料的分析能力及分析上的穩健性，也因而適用於高良率的高科技產業。

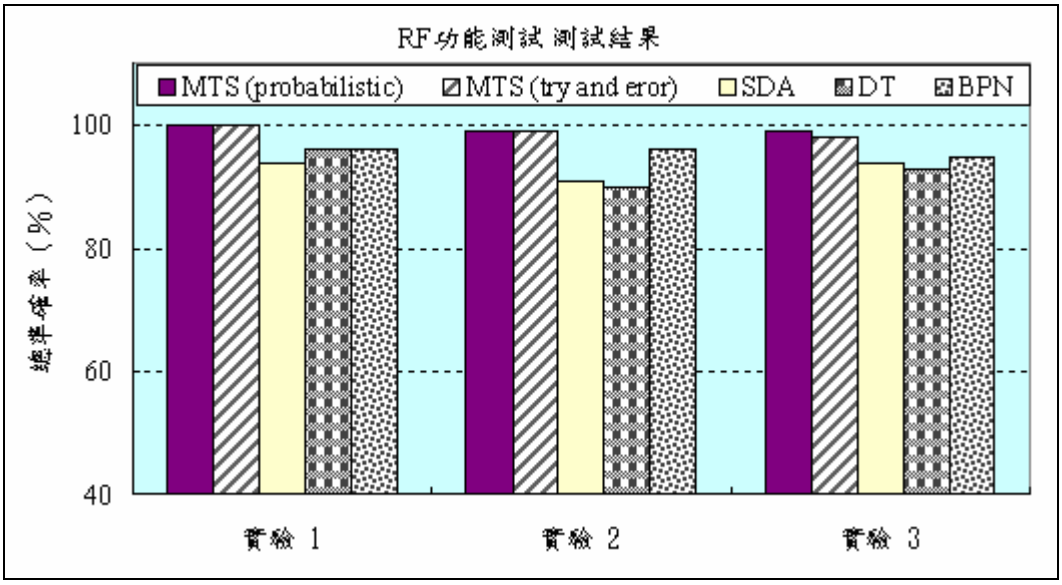


圖 5.7 各分類法測試總準確率比較（RF 功能檢測）

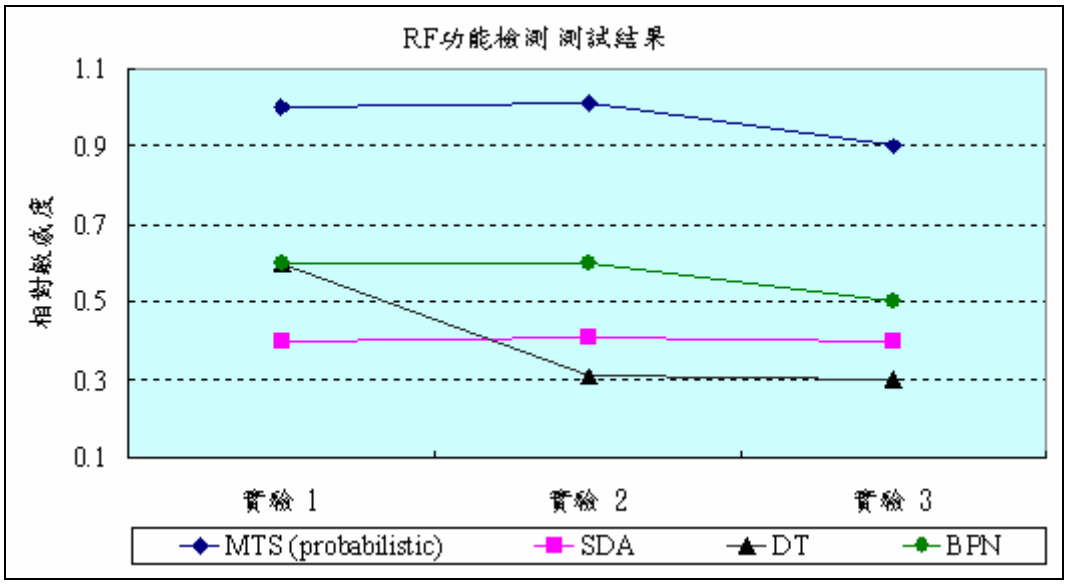


圖 5.8 各分類法測試相對敏感度比較（RF 功能檢測）

第六章　結論

6.1 研究結論與貢獻

針對多變量資料的二元分類問題而言，在欲被用以建構分類模型的訓練資料中，兩類別資料的數量比例通常是影響分類方法是否能有效學習的因素之一，本研究的主要目的之一在於探討兩類別的不平衡分佈對馬氏-田口系統（MTS）及其他分類技術（逐步別分析、決策樹分析、倒傳遞神經網路）所造成的影響。由分析結果發現，訓練資料的不平衡確實對其它分類技術的學習過程造成相當程度的偏差，致使所學習的類別預測縮減模型僅能維持對多量類別的判斷並降低對少量類別的辨識敏感度，而隨著不平衡程度的擴大，此現象更加趨於顯著。然而，在同樣的數據條件下，以 MTS 作為分析工具所得到的特徵選取及類別預測結果卻是相當穩健，無論訓練資料在兩類別比例呈現如何的不平衡，皆使所建構的分類縮減模型在測試上依然維持高特效性及高敏感度的類別預測能力。此外，MTS 執行過程中分類閾值的劃分方式是一直以來仍未獲得圓滿解答的問題，本研究藉由柴比雪夫定理的應用提出機率閾值，該閾值的訂定不但較一般常用的試誤法快速、明確，在經過測試比較後亦證實其不若試誤法容易造成對訓練資料過度適配的現象，並且可得到較好的判別能力。在最後的實案中，本研究以 MTS 的架構及手法對某高科技公司的行動電話 RF 檢測製程進行分析，試圖在維持良品與不良品的準確判斷下刪除冗餘的檢測屬性，以提升作業能量、增進市場競爭力。經由檢測製程的最佳化，不僅為該公司減少設備及相關費用支出，更由於生產效能的提升而加快對需求的反應速度，確保市場競爭中的生存條件。

本研究主要貢獻如下：

1. 對 MTS 的概念、原理及分析過程作一詳盡介紹，促進外界對此新技術的瞭解。

2. 經由系統化的分析，證實資料的不平衡確實會對特徵選取及分類模型的建構過程造成負面偏差，並發現 MTS 具有抗拒該偏差的穩健能力。因此，對於往後欲在資料型態呈現不平衡分佈的領域中，如醫學診斷、高科技產業等，進行特徵選取及類別預測分析者，此研究實可做為參考依據。
3. 在 MTS 執行過程中，利用本研究所提出的機率閾值，可以補強過去在 MTS 閾值劃分方法的不足與不確定。
4. 將 MTS 成功導入高科技生產線中品質檢測過程的分析改善，顯示 MTS 在實務上，尤其是現今高科技環境的適用能力。

6.2 未來研究建議

有關 MTS 的研究，未來仍有一些努力的空間。第一，本研究僅針對二元分類問題進行探討，但實際分類問題並不僅侷限於兩類別，因此未來可著重於多類別分類問題的研究，並將 MTS 的概念及手法擴展至其中。第二，利用正常樣本建構馬氏空間的過程中，必須在個別特徵變數上利用該變數的平均值及標準差進行標準化，然而，當正常樣本在某一變數上為同值時，該變數的標準化過程必然受阻礙，而造成必須以 0 來取代其標準化值，這勢必引起對原有變數資訊的損失，並且破壞馬氏空間中馬氏距離平均值趨近於 1 的假設，因此，這樣的問題仍有待未來研究的克服。

參考文獻

[1] Taguchi, G., Chowdhury, S. and Wu, Y. (2001). *The Mahalanobis-Taguchi System*, McGraw-Hill.

[2] Taguchi, G., Jugulum, R. (2002). *The Mahalanobis-Taguchi Strategy*, John Wiley & Sons, New York.

[3] Woodall, W. H., Koudelik, R., Tsui, K. L., Kim, S. B., Stoumbos, Z. G. and Carvounis, C. P. (2003). "A Review and Analysis of the Mahalanobis-Taguchi System," *Technometrics*, Vol. 45, pp. 1-15.

[4] Rajesh, J., Taguchi, G. and Taguchi, S. (2003). "Discussion - A Review and Analysis of the Mahalanobis-Taguchi System," *Technometrics*, Vol. 45, No. 1, pp. 16-21.

[5] Bovas, A. and Asokan Mulayath, V. (2003). "Discussion - A Review and Analysis of the Mahalanobis-Taguchi System," *Technometrics*, Vol. 45, No.1, pp. 22-25.

[6] Hawkins, D. M. (2003). "Discussion - A Review and Analysis of the Mahalanobis-Taguchi System," *Technometrics*, Vol. 45, pp. 25-29.

[7] Han, J. and Kamber, M. (2003). *Data Mining - Concepts and Techniques*, Academic Press.

[8] Pendharkar, P. C., Rodger, J. A., Yaverbaum, G. J., Herman, N. and Benner, M. (1993). "Association, statistical, mathematical and neural approaches for mining breast cancer patterns," *Expert System with Applications*, Vol. 17, pp. 223-232.

[9] An, A. and Wang, Y. (2001). "Comparisons of classification methods for screening potential compounds," *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, 29 Nov. - 2 Dec. 2001, pp.11-18.

[10] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). "From data mining to knowledge discovery: An overview." *Advances in Knowledge Discovery and Data Mining*, pp. 1-35. AAAI/MIT Press.

[11] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Communications of the ACM*, Vol. 39, pp. 27-34.

[12] Fu, Y. (1997). "Data mining," *IEEE Potentials*, Vol.16, Issue: 4, pp. 18-20.

[13] Olaru, C. and Wehenkel, L. (1999). "Data mining," *IEEE Computer Applications in Power* , Vol. 12, Issue: 3, pp.19-25.

[14] Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical*

Analysis, Prentice-Hall.

[15] Sharma, S. (1996). *Applied multivariate techniques*, New York: J. Wiley.

[16] Fdor, I. K. and Kamath, C. (2002). "Dimension reduction techniques and the classification of bent double galaxies," *Computational Statistics and Data Analysis*, 41, pp. 91-122.

[17] Hush, D. R. and Home, B. G. (1993). "Progress in Supervised Neural Network," *IEEE Signal Processing Magzine*, 10, pp. 8-39.

[18] Brown, D. E., Corruble, C. and Pittard, C. L. (1993). "A comparison of decision tree classifiers with backpropagation neural networks for multimodal classification problems," *Pattern Recognition*, 26, pp. 953-961.

[19] Curram, S. P. and Mingers, J. (1994). "Neural networks, decision tree induction and discriminant analysis: An empirical comparison," *Journal of the Operational Research Society*, 45, pp. 440-450.

[20] 蘇朝墩，2002，*品質工程*，中華民國品質學會。