嘉南藥理科技大學專題研究計畫成果報告

計畫編號：CNHA94-04
計畫名稱：資料採擷中的統計思維
(Statistical Thinking in Data Mining)

執行期間：94 年 1 月 1 日至 94 年 12 月 31 日

□整合型計畫
計畫總主持人：
子計畫主持人：

■個別型計畫
計畫主持人：陳俞成

中華民國 95 年 02 月 28 日

嘉南藥理科技大學專題研究計畫成果報告

資料採擷中的統計思維

Statistical Thinking in Data Mining

計畫編號：CNHA-94-04

執行期限：94 年 1 月 1 日至 94 年 12 月 31 日

主持人：陳俞成　嘉南藥理科技大學醫務管理系

Email:ycchen@mail.chna.edu.tw

一、中文摘要

統計界時常面臨科學及工業界提出問題之挑戰。早期問題大多源自農業與工業實驗，而且問題範圍相對較小；隨著計算機科學與資訊時代的進步，統計問題在大小與複雜度上均同時面臨高速的膨脹。統計學者的工作即在於由龐大的資料中抽取出重要的形態與趨勢，並瞭解資料解說出什麼知識；稱此過程為「自資料中學習」。大略可將學習問題分成監督式(supervised)或非監督式(unsupervised)兩類。在監督式學習的問題中，是想探討如何由某些獨立變數來預測依變數之結果；而非監督式學習的問題中，並沒有依變數，是想描述所蒐集之變數間的關聯與型態。工業革命時期，工廠以大量生產的特性對經濟造成重大影響；隨著資訊時代來臨，已由技術導向轉為市場或消費導向。近年來，客戶關係管理(CRM, Customer Relationship Management)及一對一行銷成了相當熱門的話題。完善的客戶關係管理模式與機制，有助於提高客戶忠誠度，可以降低管理行銷成本，讓客戶購買更多產品，以提高收益。欲完成上述目的，必須從瞭解客戶著手。要瞭解客戶得先從資料的收集與有效率的組織管理開始；資料來源很多，例如：量販大賣場帳單記錄，信用卡消費記錄，申請表，電話記錄等等。將收集的資料進行有效率組織管理的步驟，須靠資料倉儲(Data Warehouse)。資料倉儲是資料庫技術的一個新主題，由於資訊科技之進步，利用電腦幫助我們有效率的操作、計算和儲存大量資料。對於資料的分析、瞭解，進而轉換為有用知識或訊息，就得靠資料採擷(Data Mining)。資料採擷可以說是從巨大的資料倉儲中找出有用資訊之過程與技術，主要的技術工具是由機械學習、人工智慧與統計等學門發展而來。一般來說，資料採擷包含下列功能：分類(Classification)、推估(Estimation)、預測(Prediction)、關聯分組或購物籃分析(Affinity Grouping or Market Basket Analysis)、同值分組(Clustering)與描述(Description)。

關鍵詞：資料倉儲，資料採擷，統計

Abstract

Science and industry always bring problem to challenge the field of statistics. In the early days, these problems often came from agricultural and industrial experiments and were relatively small in scope of data. With the advent of computers and the information age, statistical problems have exploded both in size and complexity. The job of statistician is to extract important patterns and trends, and understand “what the data says.” We call this “learning from data.” The learning problems can be roughly categorized as either supervised or unsupervised. In supervised learning problems, the goal is to predict the value of a dependent variable based on a number of independent variables; in unsupervised learning problems, there is no dependent variable, and the goal is to describe the associations and patterns among a set of collected variables. In Industrial Revolution age, factories make mass production and put great effect to economics. As information age coming, technology directed has been turned about market or consumer directed. Recently, CRM(Customer Relationship Management) and one-to-one marketing are popular issues. A good model of CRM can help improve the royalty of customers, low down the marketing fee, make customers to buy more and get more profit. It needs to understand the customers to complete the goal of CRM. Collecting data and efficiently organizing data are the first step to understand the customers. There are many sources of data, for example: shopping center transaction records, consumer records of credit card, application forms, telephone records, etc. Data Warehouse is the process to efficiently organize the collecting data. Data Warehouse is a new issue of database technology. For the advancement of information technology, we can use computer efficiently to operate, compute and store mass data. From analyzing and understanding data to useful knowledge and information, it should be based on Data Mining. Data Mining is the process and technology to find useful information from mass Data Warehouse. The majority technology tools were formed from machine learning, artificial intelligence and statistics. Data mining includes the following functions: Classification, Estimation, Prediction, Affinity Grouping or Market Basket Analysis, Clustering and Description. We will study what the role statistics is in the Data Mining.

Keywords: Data Warehouse，Data Mining，Statistics

二、緣由與目的

統計界時常面臨科學及工業界提出問題之挑戰。早期問題大多源自農業與工業實驗，而且問題範圍相對較小；隨著計算機科學與資訊時代的進步，統計問題在大小與複雜度上均同時面臨高速的膨脹。統計學者的工作即在於由龐大的資料中抽取出重要的形態與趨勢，並瞭解資料解說出什麼知識；稱此過程為「自資料中學習」。大略可將學習問題分成監督式(supervised)或非監督式(unsupervised)兩類。在監督式學習的問題中，是想探討如何由某些獨立變數來預測依變數之結果；而非監督式學習的問題中，並沒有依變數，是想描述所蒐集之變數間的關聯與型態。

工業革命時期，工廠以大量生產的特性對經濟造成重大影響，促進世界繁榮進步；隨著資訊時代來臨，已由技術導向轉為市場或消費導向。近年來，客戶關係管理(CRM, Customer Relationship Management)及一對一行銷成了相當熱門的話題。完善的客戶關係管理模式與機制，有助於提高客戶忠誠度，可以降低管理行銷成本，讓客戶購買更多產品，以提高收益。欲完成上述目的，必須從瞭解客戶著手。要瞭解客戶得先從資料的收集與有效率的組織管理開始；資料來源很多，例如：量販大賣場帳單記錄，信用卡消費記錄，申請表，電話記錄等等。將收集的資料進行有效率組織管理的步驟，須靠資料倉儲(Data Warehouse)。資料倉儲是資料庫技術的一個新主題，由於資訊科技之進步，利用電腦幫助我們有效率的操作、計算和儲存大量資料。對於資料的分析、瞭解，進而轉換為有用知識或訊息，就得靠資料採擷(Data Mining)。

在 1960 年代電腦發明階段，電腦主要功能是作為資料儲存的工具，儲存硬體以磁帶為主，其資料存取的速度無法應付即時性的需求，因而主要提供歷史性的靜態資料。在 1970 年代，儲存硬體由循序存取的磁帶進步到隨機存取的磁碟，以及關聯式資料庫系統(RDBMS)與結構性查詢語言(SQL)的發展，即時性的單層次動態資料傳遞遂成可行。在 1990 年代，即時線上分析處理(OLAP)的資料倉儲系統可傳遞歷史性多層次動態資料。到了現在，透過統計、人工智慧與機器學習等不同學門之專家學者的投入，所關住的是經過整理、運算大量資訊所得到的知識。對於資料倉儲（Data Warehouse）和資料挖掘（Data Mining）有許多人時常混淆，不知如何分辨。其實，資料倉儲是資料庫技術的一個新主題，由於資訊科技之進步，利用電腦系統幫助我們操作、計算和儲存資料，讓作業方式改變，決策方式也跟著改變。此外，決策支援系統和主管資訊系統也日漸普遍，它們操作資料的方式不盡相同，因而有必要把作業性資料庫和資料倉儲分隔開來，利用不同資料庫系統與技術操作，才能達系統最佳化。由於關聯式資料庫、平行處理及分散式資料庫技術的進步，不論是主從式架構或主機型架構的資料庫系統，資料倉儲技術皆可以利用原有作業中或已有的系統，進而提供一個穩固的基礎以支持決策支援系統（DSS）。資料倉儲本身是一個非常大的資料庫，它儲存著由組織作業資料庫中整合而來的資料，特別是指從線上處理系統（OLTP）所得來的資料。將這些整合過的資料置放於資料倉儲中，由決策者則利用這些資料作決策；這個轉換及整合資料的過程是建立資料倉儲最大的挑戰。因為將作業中的資料轉換成有用的的策略性資訊是整個資料倉儲的重點。也就是，資料倉儲應該具有這樣的資料：整合性資料（integrated data）、詳細和彙總性的資料(detailed and summarized data)、歷史資料、中繼資料（Metadata）。如果資料倉儲集合具有完整地探測世界的資料，則採擷出有用的決策資料與知識，是建立資料倉儲與使用資料採擷的最大目的。而從資料倉儲採擷有用的資料，則是資料採擷的研究重點，兩者的本質與過程是兩碼事。換句話說，資料倉儲應先行建立完成，資料採擷才能有效率的進行，因為資料倉儲本身所含資料是「乾淨」(不會有錯誤的資料參雜其中)、完整的，而且是整合在一起的。

資料採擷可以提昇企業競爭優勢，到底應該如何進行呢？根據 Glymour 等人的研究，提出一個參考的進行步驟如下：

1. 理解資料與進行的工作
2. 獲取相關知識與技術（Acquisition）
3. 融合與查核資料（Integration and checking）
4. 去除錯誤或不一致的資料（Data cleaning）
5. 發展模式與假設（Model and hypothesis development）
6. 實際資料挖掘工作

7. 測試與檢核所挖掘的資料（Testing and verfication）

8. 解釋與使用資料（Interpretation and use）

從八個步驟來看，資料採擷牽涉大量的規劃與準備，有專家聲稱高達 80%的過程花在準備資料階段，這包括表格的連結以及可能相當大量的資料轉換。從這個角度看，資料採擷只是知識發掘過程中的一個步驟而已，而達到這個步驟前還有許許多多的工作要完成。

資料採擷可以說是從巨大的資料倉儲中找出有用資訊之過程與技術，主要的技術工具是由機械學習、人工智慧與統計等學門發展而來。一般來說，資料採擷包含下列功能：分類(Classification)、推估(Estimation)、預測(Prediction)、關聯分組或購物欄分析(Affinity Grouping or Market Basket Analysis)、同值分組(Clustering)與描述(Description)。資料採擷有兩種基本模式，一種由上而下，去檢是我們的想法是否成立，是謂假設檢定；另一種由下而上，以原始資料分析找出原先不知道的事實，稱為知識發掘。知識發掘可以區分為監督式與非監督式；監督式知識發掘透過比較找出特定變數的數值意義，選定輸出變數，由電腦協助我們評估、分類或預測；而非監督式知識發掘，沒有輸出變數，利用電腦在原始資料中找出有意義之型態。

統計學最初是為了解釋現象，不論是自然科學或社會科學，透過資料的蒐集，描述性統計給現象賦予概括的意義，以中位數和平均數代表資料之中心，以變異數區分資料分散之程度，以隨機分佈刻劃現象變化之全貌。迴歸分析是希望透過模型來歸納資料，以利現象之解釋、預測或控制。

線性模式(Linear Models)是統計界最主要用於監督式學習的工具；而其用來估計的方法則可分成最小平方法(Least Squares Methods)與最近鄰域法(Nearest Neighbor Methods)。通常最小平方法需要比較強的假設，所得結果較穩定，但是偏誤可能會比較大；而最近鄰域法需要的假設較弱，所得結果較不穩定，但是偏誤可能會比較小。如何在偏誤與穩定性之間取得平衡，往往是一個藝術問題。此外，高維度資料也與低為度資料存在很大的特性差異。非監督式學習所面對的資料維度通常較高，我們感興趣的性質也比較複雜。在低維度問題(比如三維以下)，有一些無母數統計的方法可以直接給出有效的估計。一但遇到高維度資料，就需要是當的修正方法。關聯規則、集群分析、自我組織圖、主成分分析、多元尺度分析是常用的工具。

四、計畫成果自評

本研究內容與原計畫相符程度大約為百分之七十五左右，達成預期目標情況約為百分之七十左右，於應用上提出一個實務可行之解。

五、參考文獻

1. 彭文正(2001), 資料探勘-顧客關係管理暨電子行銷之應用, 數博網資訊股份有限公司。
2. 曾新穆、李建億(2003), 資料探勘, A Tutorial-Based Primer, 東華書局。
3. 曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯(2005), 資料探勘, 旗標出版股份有限公司。
4. Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). Classification and Regression Trees, Wadsworth.
5. Cherkassky, V. and Mulier, F. (1998). Learning from Data, Wiely, New York.
6. Cox, D. and Hinkley, D. (1974). Theoretical Statistics, Chapman and Hall, London.
7. Hastie, T., Tibshirani, R. and Friedman, J. (2001). The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Springer Verlag.