知識天地

是「舊瓶裝新酒」？還是「新瓶裝舊酒」？

張源俊（統計科學研究所副研究員）

2002 年麻省理工學院有一篇關於機器學習博士的論文，標題爲"Everything old is new again"（Rifkin, 2002） 引起我的高度興趣。因爲它一點都不像論文。「資料探勘」（Data Mining） 這個首先由資訊管理的專家們於九零年代提出的名詞，在當時僅僅是希望從既有的資料庫中，挖掘出有用的資訊，而目前它幾乎已經成了資料分析的代名詞。在資訊管理的商業包裝下，資料探勘成爲話題，統計也因此更受重視。然而這是一個新方法嗎？由於統計科學多樣化的特質，統計學者們或直接或間接地必須與計算及資料分析有密切的關係。因此，爲表現統計的「專業」，對此一以「資料」爲中心的活動，更有一分不能「置身事外」的「使命感」。

從統計學者的觀點，資料探勘以及後來所謂的「資訊及知識萃取」（Information and Knowledge Extraction） 並非指一個單一工具或方法；比較貼切的說法是「資料探勘」及「資訊及知識萃取」是在資工或資管包裝及串聯下的統計資料分析過程。這樣的說法並無貶低資工或刻意凸顯統計的意味。相反的這正強調了跨領域合作的必要性。

由於資料量龐大，資料在能進行分析之前，必得先做一些前置處理。而這些繁複的工作必得藉助統計方法及資訊工程的技術幫助。這些工作雖未必涉及高深的資訊工程技術或統計方法，但有經驗的工程師及統計專家往往能夠提出有利於事後分析的建議。而分析過程中所包含的分析方法更是五花八門，包括敘述性統計、統計模型建構、資訊的呈現（如資料視覺化等）、資料庫、網路、資訊安全、圖像辨識等等的技術。從傳統的分類、分群、回歸分析、資訊視覺化、多變量分析等的各種方法或工具，甚至目前相當熱門的統計（機器）學習的方法不一而足。大概沒有任何一個人可以熟悉這麼多工具或具備如此廣泛的知識。過程中需要統計學家與資訊工程師們充分合作，更得融入該科學主題的專業知識，但往往彼此都認爲對方的「專業」只是「技術細節」，合作成功的例子因而並不多見，跨領域的合作仍待加強，尤其是想把此一概念運用於「科學資料」時，團隊合作的需求將更高。

美國統計學會會長（1997）Jon Kettening 曾說：「我喜歡把統計想成一個是從資料中學習的科學，…」（I like to think of statistics as the science of learning from data …）。而很巧的是「從資料中學習」（learning from data）這一段文字幾乎出現在所有資料探勘及機器學習的教科書上，統計與資訊工程的重疊由此可見。我們常常必須在被服務者與服務者（client and servant）的角色之間轉換，故如何與其他領域的學者合作是當前統計學者們的必修課。著名的機率學者鍾開萊教授（Kai Lai Chung）即曾經在他書中的前言寫到：「某些人的技術細節是其他人的專業領域」（"One man's technicality is another's professionalism."-- from "Lectures from Markov Processes to Brownian Motion", 1980）。

我們以目前最受矚目的支持向量機（support vector machine）在資料探勘的應用爲例。在資料探勘的統計分類工具中，因爲核化（kernelization）的算則（algorithm）有處理大量資料的演算能力，故常被用於資料探勘中的分類問題。從技術方面而言，核化的另一好處是，對許多以「內積（inner product）」演算爲主的傳統統計方法（如 PCA, FDA, CCA），核化提供了一個相對容易的擴充平台（framework），在幾乎不更改原有程式的情況下，核化算則可以提供非線性的性質 （Chang, Lee, Pao, Lee and Huang, 2006）。就統計概念而言，似乎了無新意。但它卻開了一扇通往非線性的門。較以往透過函數估計的方法而言，維度的詛咒（curse of dimensionality）對此類方法似乎弱了許多。

這些研究大多是受支持向量機興起的影響。然而支持向量機卻非新議題。從發展的歷史來看，支持向量機起源於 Rosenblatt（1958）的 Perceptron；而 Perceptron 的發展更可回溯到更早 McCulloch and Water Pitts 於 1943 年的論文 "A Logical Calculus of Ideas Immanent in Nervous Activity"。而此一發展的誘因，甚至可追溯到西班牙及英國的神經學學者 Dr. Cajal 和 Dr. Sherrington。其中，Dr. Cajal 更是 1906 年的諾貝爾獎得主。

學門之間的互相激盪往往有出乎意料的發展，尊重專業是學門之間合作的唯一途徑。是「舊瓶裝新酒」？還是「新瓶裝舊酒」？這似乎不重要了。最後僅以愛因斯坦對加州理工學院（California Institute of Technology）學生的演講與大家互相勉勵。

> *It is not enough that you should understand about applied science in order that your work may increase man's blessings. Concern for man himself and his fate must always form the chief interest of all technical endeavors, concern for the great unsolved problems of the organization of labor and the distribution of goods in order that the creations of our mind shall be a blessing and not a curse to mankind. Never forget this in the midst of your diagrams and equations.*
>
> Albert Einstein,
>
> Address to the student body of California Institute of Technology

參考文獻

1. Yuan-chin Ivan Chang, Yuh-Jye Lee, Hsing-Kuo Pao, Mei-Hsien Lee, and Su Yun Huang. Data Visualization via Kernel Machines, Technical report C-2006-04, Institute of Statistical Science, Academia Sinica, to appear in Handbook of Computational Statistics (Volumn III)- Data Visualization Ed. by Chun-houh Chen, Wolfgang Hardle and Antony Unwin, 2006.
2. McCulloch W. S., Pitts W. " *A Logical Calculus of the Ideas Immanent in Nervous Activity*" Bulletin of Mathematical Biophysics, vol. 5, pp. 115-133, 1943.
3. R. Rifkin. Everything Old Is New Again: A Fresh Look at Historical Approaches in Machine Learning. PhD thesis, MIT, Cambridge, MA, 2002.
4. F. Rosenblatt, The perceptron: **A** probabilistic model for information storage and organization in the brain. *Psychological Review***,** Vol. 65, pp. 386-408, 1958.
5. B. Scholkopf, A. Smola, and K.-R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.
6. Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.