撰文／林守德

在數位大海中撈針

歧異點探測技術可以防止信用卡盜刷，還能協助尋找恐怖份子。

一個月前，英國《每日郵報》指出，1954年4月11日是個很特別的日子：它的特別之處，並非在於有什麼重要的事件發生，而是那天全球的新聞中，居然沒有任何值得報導的事件（《每日郵報》當天的頭版頭條，是一個小偷偷了一個杯子）——堪稱20世紀最平靜的一日。

這則新聞提出了一項弔詭的論述：通常「新聞」是指平常不會發生的事件，所以才值得報導。但是，因為每天都有「新聞」已經變成常態，反而讓平靜無波的日子顯得特別不尋常。

在網際網路風行的現代，取得資訊已非困難的事情；如何從大量資訊中找出有用的訊息，反而是一項挑戰。比如從人的行為資料中，找出可疑或危險人物；從密密麻麻的刷卡記錄中找出少數異常現象，防堵盜刷行為；或是從大量的檢驗資料中，自動找出某些傳染病的少數帶原者或傳播者。想達到這些目的，通常需要在大量資料中找出異於常態的部份。

這工作對人腦而言可是苦差事：首先必須仔細觀察資料，從資料中歸納出所謂的「正常」或是「頻繁」事件，然後再以這些事件為基礎，來篩選出「異常」事件。當資料量變多，分析的工作不僅耗時費神，正確率也會遞減。而電腦的快速計算與記憶的功能，正好很適合從事這類「大海撈針」的工作。在統計學以及計算機科學中，專門處理這工作的技術就叫「歧異點探測」。

歧異點探測的方法很多，但是基本上可以分為三大類：第一類是找發生機率很低的事件，第二類是用分類的方式，第三種是與同儕相比。

機率分佈的方法，主要是先將每筆資料出現的機率做統計，然後把統計上較不容易出現、卻真正發生過的資料抓出來。舉例而言，在男性身高的統計結果中，若平均為170公分，標準差是15，則我們可以得知身高超過215公分的男性機率非常低，這樣的個體就可以當成歧異點。

而分類法是利用分群演算法將資料分群，然後將無法歸屬於任何一群的點抓出來當成歧異點。比如說把重要新聞事件分為「政治」、「經濟」等不同類型，如此一來，1954年4月11日這類極少數無法被歸類的平靜日子，可能就會被認為是個歧異點。

與同儕相比的歧異點，基本概念很簡單：它希望能夠發現與同類的資料有所不同之處。比如說某個不擅於長打的第四棒、或是善於助攻的中鋒等奇特的事件。

如何將「歧異點探測」根據不同目的適切應用在生活上，則還有兩難之處。想一想：刷卡付帳時明明信用卡還有額度，卻無法刷卡，需要與銀行確認是本人消費才能刷卡，這多掃興？打開微網誌發現帳號或粉絲專頁被停權，又多惱人？這可能都是因為「歧異點探測」精確度太低造成的誤判。而銀行也可能因此損失信用卡客戶，臉書因此遭到使用者杯葛。

但是在國防、金融、詐騙等重要安全議題上，我們又希望「歧異點探測」可以做到滴水不漏，不管多小的問題點都能抓出來。畢竟誰都不希望存款消失、信用卡盜刷、車站爆炸這種事發生在自己身上。

另外有些特殊歧異點，即使利用電腦的幫助，還是不易找到。在911恐怖攻擊悲劇發生後，相關單位調查發現，這批恐怖份子不僅曾一起出現在某處，也有一些特殊的共同行為：例如，都曾在某幾所飛機駕訓學校上過課、總在最後時刻才買機票、而且都是用現金而非信用卡買機票。這些個別資訊乍看之下都不是很特別，因為總是有些人會在最後時間買機票；在國外上飛機駕訓課程者也大有人在。但是綜合起來，就會顯示出一些歧異點的蛛絲馬跡。可惜的是，這樣潛在的關係，往往要事後的分析才能夠發現不尋常的端倪，所以，如何在「事前」偵測出可疑的歧異事件，目前還是研究人員努力的方向。

林守德是台灣大學資訊工程系助理教授、中華民國人工智慧學會秘書長。

影像來源：林守德