研究成果

機器學習與分類

張復（資訊科學研究所副研究員）

分類是很有趣的題目，早就獲得哲學家的青睞。分類之所以有趣，可能是因為我們每個人都知道如何分類，卻說不出所以然來。柏拉圖認為，分類的基礎在於某種理想型態。比如，圓形的理想型態是與某個點保持固定距離的平面點集合。在現實世界所看到的圓並不完全遵守這個規則，只是這個理想型態的某種變形。這想法最大的難題在於並不是所有的類別都可以用這麼乾淨的數學公式來表示。比如椅子，它的數學公式是什麼？即使有，要找出它確實的形式也是很麻煩的事。

亞理斯多德認為，類別只是所有相同型態個別物件的集合。這樣的說法好像只是循環定義，因為在定義項中出現了「同型態」的字詞。然而，這個想法很有用。它告訴我們，即使我們無法定義類別，但可以學習類別。也就是說，你只要把同一類別的個例收集起來，就可以讓別人甚至機器（電腦）來學習這個類別。難以置信的是，在 70 年代，人們真的使用這個方法來訓練電腦去辨識文字。辨識的方法是，將每個待辨識的文字影像與事先收集好的文字影像（稱為訓練樣本）做逐一的比對。這工作雖然費時（即使是電腦），但非常有效。後來，有人進一步證明，這種方法所可能犯的錯誤，隨著訓練樣本的增加，會逐漸收斂到最低的比率。至於，比對時應該採取什麼方法，這就牽扯到下一位哲學家的看法。

維根斯坦（Ludwig Wittgenstein）也許是第一個願意面對下面事實的哲學家：你可能永遠無法使用單一的屬性將不同類別的物件區分。我們會將一群物件當作同一類別對待，也許是因為它們之間有某種如家族般的相似性。我們今天知道，同一個家族的成員並不擁有完全相同的基因，而只是一套重疊性很高的基因。等一下我還會回到這題目上，因為它與生醫所潘文涵及我所做的族群分類研究息息相關。這就是維根斯坦的想法對應於分類領域裡的特徵向量（feature vector）的想法。亦即我們把每一個物件（訓練樣本或測試樣本）都表示為一套特徵的組合。這樣的組合在數學上被稱做向量。

你可以把向量想像成 N-維空間中的一個點，N 就是你表示這物件所使用的特徵的數量。如此想法有什麼好處呢？假定你只使用兩個或三個特徵來表示一個物件。這樣，你就可以在二維或三維空間中看到它們。你可以看到，那些被我們歸為同類的物件會經常群聚在一起，請參考圖 1。在這圖裡面，我們把所有的個例都表示為二維空間中的一個點。 1(a) 或 1(b) 裡面一共只有兩個類別，分別以 A 與 B 來命名。你看到的這兩個例子是人造的，但是它們對應於實際應用裡會常看到的狀況。

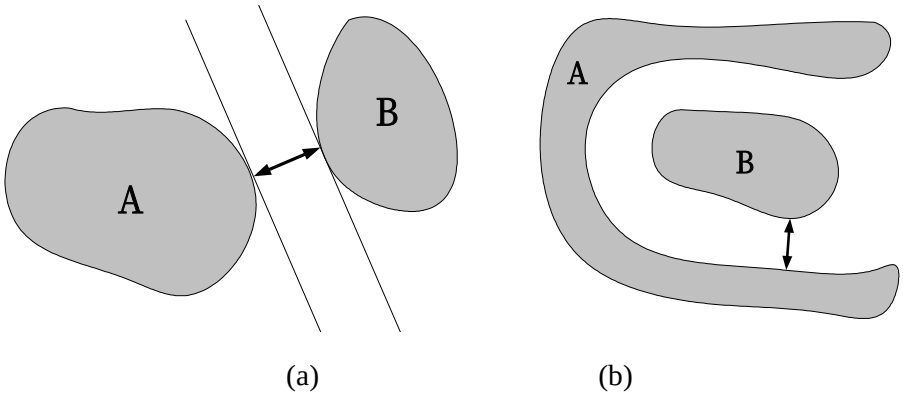

圖 1、 (a) 兩個群組之間有一條鴻溝。 (b) 兩個群組之間有一定的距離。

看到圖 1，你可能會恍然大悟。你會說，雖然同一類別的個例並沒有凝聚在同一點上，它們卻群聚在一起，而且足以與其他類別的群聚分離開來。沒有錯，剩下的問題只是怎樣把這個分離的狀態表達出來。其中一種方法，是把 A 類與 B 類之間最大的鴻溝找出來。要找出最大的鴻溝，你可以旋轉一對平行線，讓它們各自貼著 A 類或 B 類最突出的點。你可以不斷旋轉它們，並調整之間的距離，直到最大的鴻溝出現。數學上有一套方法可以讓你這麼做，稱為支持向量機器（support vector machines, SVM）。這個方法表面看起來不夠好，因為它只能分離 1(a)裡的兩個類別，而無法分離 1(b)裡的類別。但是，如果我們將原空間裡的點透過某種轉換函數將它們映射到高維空間去，甚至是無窮維空間去，不同類別的點就變得容易分離了。

SVM 是很強健的分類方法。然而它有一項缺陷：你一次只能用 SVM 來處理兩個類別。如果要做很多類別的區分，像中文文字，其常用字便有五千種以上，使用 SVM 無論在訓練或實作上就變得非常耗時。我們最近發展出一種新的分類方法，可以同時針對所有的類別一起作處理。我們拿 1(b) 的例子來說明這個方法。你可以在 A 類中選個樣本，做為 A 類的原型，B 類中也選個樣本，做為 B 類的原型。然後你試著用這兩個原型去做分類。接近 A 原型的便歸類為 A 類，接近 B 原型的就歸類為 B 類。這樣做當然不保證所有的歸類都正確。從歸類錯誤的 A 類樣本裡，我們可以選裡面的一個做第二個原型。我們也可以在 B 類裡做同樣的事。有了原型以後，我們還可以使用群組計算法則（clustering algorithm）將原型的位置做適當的調整，讓它們變得更有代表性。然後，我們使用新得到的原型再去做歸類的嘗試。如此反覆不斷地去做，只要有分類錯誤，我們就增加新的原型，並且使用群組計算法則調整它們的位置。

上述方法，我們稱為調適性的原型學習（adaptive prototype learning, APL）方法。APL 有什麼好處呢？第一，它可以同時針對所有的類別做學習。學習之後，又可以針對所有的類別一起做區分。對於類別種類繁多的問題，APL 提供了極大的便利。第二，在理論上，APL 能夠適用於 1(b)的例子。所以，它不需要把樣本映射到高維空間就能得到很好的分類效果。第三、我們可以在理論上證明，APL 可以在比 SVM 要求較低的條件下達到漸進式的最低錯誤率。事實上，我們最近在一些標準的資料上測試，顯示出 APL 普遍高於 SVM 的分類效果。

不管使用什麼方法來分類，分類在許多研究領域裡都有極成功的應用。剛才我們提到的族群分類便是一個例子。在這個應用裡，我們把人類的 STRP markers 做向量化的處理，然後使用 APL 做分類。APL 的分類效果很好。而且，在區分族群時，使用 markers 數量的多寡（只要不低於某個閥值）不會影響到 APL 所製造的原型數量。原型數量的常態可以說是族群區別的一種 expression。這對於從事實驗的研究者是個福祉，因為他們只需要收集少數 markers 的數據。

在影像文字的辨識上，APL 也能夠提供很大的助益，我們在上面已經提到了。除此之外，在文件影像的分析與辨識上，我們陸續使用機器學習與分類的方法獲得不錯的成果。比如說，我們可以使用 APL 做語文分類，也就是說，我們不需要判別每個影像文字為哪個字，只需要判定它為哪種語言（比如：中文、英文、日文、…）的文字。APL 在這方面表現得非常好，它所建立的原型數量遠比為了要識別個別文字所需要建立的數量少。反之，如果使用 SVM 來從事這個工作，它所需要儲存的支持向量（support vectors）就遠比 APL 原型來的多。這意味著在實際做分類的工作時，SVM 必須花費更多的時間來計算，以達成判定結論。機器學習與分類還可以應用在灰階或彩色文件的二色化工作（圖 2），以及文件的排版分析（圖 3）。表面上，這兩個問題跟分類好像沒有什麼關係。但是，我們在這兩個問題上都需要做很多判斷。過去，研究者傾向於建立規則來從事判斷。這樣的方法陷

入了某些哲學家所犯的錯誤：他們會傾向於尋找一兩種特徵來完成目標。改用機器學習的方法，我們可以納入更多的特徵，並且使用自動學習的方式制定分類的機制，因此可以更簡易地得到高正確率的判斷決策。

藉由經驗使我們相信，機器學習與分類是值得繼續研究與應用的題目。我們下一個應用的目標是生物資訊（bioin- formatics）。同樣的經驗也顯示，不適當的特徵不可能製造機器學習與分類的奇蹟，反而要等待奇蹟出現。所以，在生物資訊方面，我們必須跟生物學者攜手合作，才會有突破的機會。

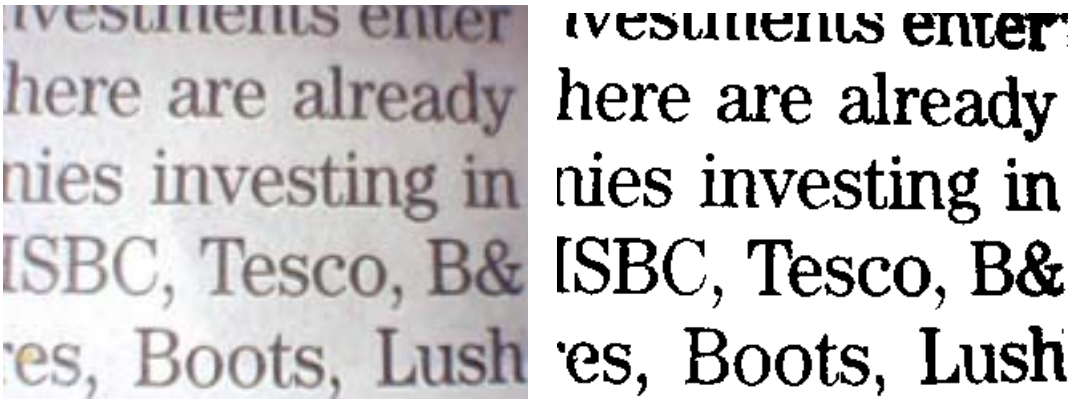

圖 2、 左：照相機所拍攝的文件影像。右：二色化的結果。

圖 3、文件的排版分析結果。