frlily / October 21, 2015 05:45PM

統如何分辨出垃圾郵件? 資料挖掘演算法與現實生活中的應用案例

本文，主要想簡單介紹下資料挖掘中的演算法，以及它包含的類型。然後，通過現實中觸手可及的、活生生的案例，去詮釋它的真實存在。

一、資料挖掘的演算法類型

資料挖掘

[img]http://www.finereport.com/tw/wp-content/themes/BusinessNews/images/15102101.jpg[/img]

一般來說，資料挖掘的演算法包含四種類型，即分類、預測、聚類、關聯。前兩種屬於有監督學習，後兩種屬於無監督學習，屬於描述性的模式識別和發現。

（一）有監督學習

有監督的學習，即存在目標變數，需要探索特徵變數和目標變數之間的關係，在目標變數的監督下學習和優化演算法。例如，信用評分模型就是典型的有監督學習，目標變數為「是否違約」。演算法的目的在於研究特徵變數（人口統計、資產屬性等）和目標變數之間的關係。

（1）分類演算法

分類演算法和預測演算法的最大區別在於，前者的目標變數是分類離散型（例如，是否逾期、是否腫瘤細胞、是否垃圾郵件等），後者的目標變數是連續型。一般而言，具體的分類演算法包括，邏輯回歸、決策樹、KNN、貝葉斯判別、SVM、隨機森林、神經網路等。

（2）預測演算法

預測類演算法，其目標變數一般是連續型變數。常見的演算法，包括線性回歸、回歸樹、神經網路、SVM等。

（二）無監督學習

無監督學習，即不存在目標變數，基於資料本身，去識別變數之間內在的模式和特徵。例如關聯分析，通過資料發現項目A和項目B之間的關聯性。例如聚類分析，通過距離，將所有樣本劃分為幾個穩定可區分的群體。這些都是在沒有目標變數監督下的模式識別和分析。

（1）聚類分析

聚類的目的就是實現對樣本的細分，使得同組內的樣本特徵較為相似，不同組的樣本特徵差異較大。常見的聚類演算法包括kmeans、系譜聚類、密度聚類等。

（2）關聯分析

關聯分析的目的在於，找出項目（item）之間內在的聯繫。常常是指購物籃分析，即消費者常常會同時購買哪些產品（例如游泳褲、防晒霜），從而有助於商家的捆綁銷售。

二、基於資料挖掘的案例和應用

上文所提到的四種演算法類型（分類、預測、聚類、關聯），是比較傳統和常見的。還有其他一些比較有趣的演算法分類和應用場景，例如協同過濾、異常值分析、社會網路、文本分析等。下面，想針對不同的演算法類型，具體的介紹下資料挖掘在日常生活中真實的存在。下面是能想到的、幾個比較有趣的、和生活緊密關聯的例子。

資料挖掘

[img]http://www.finereport.com/tw/wp-content/themes/BusinessNews/images/15102102.jpg[/img]

（一）基於分類模型的案例

這裡面主要想介紹兩個案例，一個是垃圾郵件的分類和判斷，另外一個是在生物醫藥領域的應用，即腫瘤細胞的判斷和分辨。

資料挖掘

（1）垃圾郵件的判別

郵箱系統如何分辨一封Email是否屬於垃圾郵件？這應該屬於文本挖掘的範疇，通常會採用樸素貝葉斯的方法進行判別。它的主要原理是，根據郵件正文中的單詞，是否經常出現在垃圾郵件中，進行判斷。例如，如果一份郵件的正文中包含「報銷」、「發票」、「促銷」等辭彙時，該郵件被判定為垃圾郵件的概率將會比較大。

一般來說，判斷郵件是否屬於垃圾郵件，應該包含以下幾個步驟。

第一，把郵件正文拆解成單片語合，假設某篇郵件包含100個單詞。

第二，根據貝葉斯條件概率，計算一封已經出現了這100個單詞的郵件，屬於垃圾郵件的概率和正常郵件的概率。如果結果表明，屬於垃圾郵件的概率大於正常郵件的概率。那麼該郵件就會被劃為垃圾郵件。

（2）醫學上的腫瘤判斷
如何判斷細胞是否屬於腫瘤細胞呢？腫瘤細胞和普通細胞，有差別。但是，需要非常有經驗的醫生，通過病理切片才能判斷。如果通過機器學習的方式，使得系統自動識別出腫瘤細胞。此時的效率，將會得到飛速的提升。並且，通過主觀（醫生）+客觀（模型）的方式識別腫瘤細胞，結果交叉驗證，結論可能更加靠譜。

如何操作？通過分類模型識別。簡言之，包含兩個步驟。首先，通過一系列指標刻畫細胞特徵，例如細胞的半徑、質地、周長、面積、光滑度、對稱性、凹凸性等等，構成細胞特徵的資料。其次，在細胞特徵寬表的基礎上，通過搭建分類模型進行腫瘤細胞的判斷。

（二）基於預測模型的案例
這裡面主要想介紹兩個案例。即通過化學特性判斷和預測紅酒的品質。另外一個是，通過搜索引擎來預測和判斷股價的波動和趨勢。

（1）紅酒品質的判斷
如何評鑒紅酒？有經驗的人會說，紅酒最重要的是口感。而口感的好壞，受很多因素的影響，例如年份、產地、氣候、釀造的工藝等等。但是，統計學家並沒有時間去品嘗各種各樣的紅酒，他們覺得通過一些化學屬性特徵就能夠很好地判斷紅酒的品質了。並且，現在很多釀酒企業其實也都這麼幹了，通過監測紅酒中化學成分的含量，從而控制紅酒的品質和口感。

那麼，如何判斷鑒紅酒的品質呢？

第一步，收集很多紅酒樣本，整理檢測他們的化學特性，例如酸性、含糖量、氯化物含量、硫含量、酒精度、PH值、密度等等。
第二步，通過分類回歸樹模型進行預測和判斷紅酒的品質和等級。

（2）搜索引擎的搜索量和股價波動
一隻南美洲熱帶雨林中的蝴蝶，偶爾扇動了幾下翅膀，可以在兩周以後，引起美國德克薩斯州的一場龍捲風。你在互聯網上的搜索是否會影響公司股價的波動？

很早之前，就已經有文獻證明，互聯網關鍵詞的搜索量（例如流感）會比疾控中心提前1到2周預測出某地區流感的爆發。

同樣，現在也有些學者發現了這樣一種現象，即公司在互聯網中搜索量的變化，會顯著影響公司股價的波動和趨勢，即所謂的投資者注意力理論。該理論認為，公司在搜索引擎中的搜索量，代表了該股票被投資者關注的程度。因此，當一隻股票的搜索頻數增加時，說明投資者對該股票的關注度提升，從而使得該股票更容易被個人投資者購買，進一步地導致股票價格上升，帶來正向的股票收益。這是已經得到無數論文驗證了的。

（三）基於關聯[url=http://www.finereport.com/tw/]資料分析[/url]的案例：沃爾瑪的啤酒尿布
啤酒尿布是一個非常非常古老陳舊的故事。故事是這樣的，沃爾瑪發現一個非常有趣的現象，即把尿布與啤酒這兩種風馬牛不相及的商品擺在一起，能夠大幅增加兩者的銷量。原因在於，美國的婦女通常在家照顧孩子，所以，她們常常會囑咐丈夫在下班回家的路上為孩子買尿布，而丈夫在買尿布的同時又會順手購買自己愛喝的啤酒。沃爾瑪從資料中發現了這種關聯性，因此，將這兩種商品並置，從而大大提高了關聯銷售。

啤酒尿布主要講的是產品之間的關聯性，如果大量的資料表明，消費者購買A商品的同時，也會順帶著購買B產品。那麼A和B之間存在關聯性。在超市中，常常會看到兩個商品的捆綁銷售，很有可能就是關聯分析的結果。

（四）基於聚類分析的案例：零售客戶細分
對客戶的細分，還是比較常見的。細分的功能，在於能夠有效的劃分出客戶群體，使得群體內部成員具有相似性，但是群體之間存在差異性。其目的在於識別不同的客戶群體，然後針對不同的客戶群體，精準地進行產品設計和推送，從而節約營銷成本，提高營銷效率。

例如，針對商業銀行中的零售客戶進行細分，基於零售客戶的特徵變數（人口特徵、資產特徵、負債特徵、結算特徵），計算客戶之間的距離。然後，按照距離的遠近，把相似的客戶聚集為一類，從而有效的細分客戶。將全體客戶劃分為諸如，理財偏好者、基金偏好者、活期偏好者、國債偏好者、風險均衡者、渠道偏好者等。
資料挖掘
[img]http://www.finereport.com/tw/wp-content/themes/BusinessNews/images/15102103.jpg[/img]

（五）基於異常值分析的案例：支付中的交易欺詐偵測
採用支付寶支付時，或者刷信用卡支付時，系統會實時判斷這筆刷卡行為是否屬於盜刷。通過判斷刷卡的時間、地點、商戶名稱、金額、頻率等要素進行判斷。這裡面基本的原理就是尋找異常值。如果您的刷卡被判定為異常，這筆交易可能會被終止。

異常值的判斷，應該是基於一個欺詐規則庫的。可能包含兩類規則，即事件類規則和模型類規則。第一，事件類規則，例如刷卡的時間是否異常（凌晨刷卡）、刷卡的地點是否異常（非經常所在地刷卡）、刷卡的商戶是否異常（被列入黑名單的套現商戶）、刷卡金額是否異常（是否偏離正常均值的三倍標準差）、刷卡頻次是否異常（高頻密集刷卡）。第二，模型類規則，則是通過演算法判定交易是否屬於欺詐。一般通過支付資料、賣家資料、結算資料，構建模型進行分類問題的判斷。

（六）基於協同過濾的案例：電商猜你喜歡和推薦引擎
電商中的猜你喜歡，應該是大家最為熟悉的。在京東商城或者亞馬遜購物，總會有「猜你喜歡」、「根據您的瀏覽歷史記錄精心為您推薦」、「購買此商品的顧客同時也購買了**商品」、「瀏覽了該商品的顧客最終購買了**商品」，這些都是推薦引擎運算的結果。

這裡面，確實很喜歡亞馬遜的推薦，通過「購買該商品的人同時購買了**商品」，常常會發現一些質量比較高、較為受認可的書。
一般來說，電商的「猜你喜歡」（即推薦引擎）都是在協同過濾演算法（Collaborative Filter）的基礎上，搭建一套符合自身特點的規則庫。即該演算法會同時考慮其他顧客的選擇和行為，在此基礎上搭建產品相似性矩陣和用戶相似性矩陣。基於此，找出最相似的顧客或最關聯的產品，從而完成產品的推薦。

（七）基於社會網路分析的案例：電信中的種子客戶
種子客戶和社會網路，最早出現在電信領域的研究。即，通過人們的通話記錄，就可以勾勒出人們的關係網路。電信領域的網路，一般會分析客戶的影響力和客戶流失、產品擴散的關係。

基於通話記錄，可以構建客戶影響力指標體系。採用的指標，大概包括如下，一度人脈、二度人脈、三度人脈、平均通話頻次、平均通話量等。基於社會影響力，分析的結果表明，高影響力客戶的流失會導致關聯客戶的流失。其次，在產品的擴散上，選擇高影響力客戶作為傳播的起點，很容易推動新套餐的擴散和滲透。

此外，社會網路在銀行（擔保網路）、保險（團伙欺詐）、互聯網（社交互動）中也都有很多的應用和案例。
資料挖掘
（八）基於文本分析的案例
這裡面主要想介紹兩個案例。一個是類似「掃描王」的APP，直接把紙質文檔掃描成電子文檔。相信很多人都用過，這裡準備簡單介紹下原理。另外一個是，江湖上總是傳言紅樓夢的前八十回和後四十回，好像並非都是出自曹雪芹之手，這裡面準備從統計的角度聊聊。

（1）字元識別：掃描王APP
手機拍照時會自動識別人臉，還有一些APP，例如掃描王，可以掃描書本，然後把掃描的內容自動轉化為word。這些屬於圖像識別和字元識別（Optical Character
Recognition）。圖像識別比較複雜，字元識別理解起來比較容易些。

查找了一些資料，字元識別的大概原理如下，以字元S為例。

第一，把字元圖像縮小到標準像素尺寸，例如12*16。注意，圖像是由像素構成，字元圖像主要包括黑、白兩種像素。

第二，提取字元的特徵向量。如何提取字元的特徵，採用二維直方圖投影。就是把字元（12*16的像素圖）往水平方

向和垂直方向上投影。水平方向有12個維度，垂直方向有16個維度。這樣分別計算水平方向上各個像素行中黑色像素的累計數量、垂直方向各個像素列上的黑色像素的累計數量。從而得到水平方向12個維度的特徵向量取值，垂直方向上16個維度的特徵向量取值。這樣就構成了包含28個維度的字元特徵向量。

第三，基於前面的字元特徵向量，通過神經網路學習，從而識別字元和有效分類。

（2）文學著作與統計：紅樓夢歸屬
這是非常著名的一個爭論，懸而未決。對於紅樓夢的作者，通常認為前80回合是曹雪芹所著，後四十回合為高鶚所寫。其實主要問題，就是想確定，前80回合和後40回合是否在遣詞造句方面存在顯著差異。

這事讓一群統計學家比較興奮了。有些學者通過統計名詞、動詞、形容詞、副詞、虛詞出現的頻次，以及不同詞性之間的相關係做判斷。有些學者通過虛詞（例如之、其、或、亦、了、的、不、把、別、好），判斷前後文風的差異。有些學者通過場景（花卉、樹木、飲食、醫藥與詩詞）頻次的差異，來做統計判斷。總而言之，主要通過一些指標量化，然後比較指標之間是否存在顯著差異，藉此進行寫作風格的判斷。