frlily / August 24, 2015 09:33PM

如何做好資料挖掘模型的9條經驗總結

資料挖掘是利用業務知識從資料中發現和解釋知識（或稱為模式）的過程，這種知識是以自然或者人工形式創造的新知識。

當前的資料挖掘形式，是在20世紀90年代實踐領域誕生的，是在集成資料挖掘演算法平台發展的支撐下適合商業分析的一種形式。也許是因為資料挖掘源於實踐而非 理論，在其過程的理解上不太引人注意。20世紀90年代晚期發展的CRISP-DM，逐漸成為資料挖掘過程的一種標準化過程，被越來越多的資料挖掘實踐者 成功運用和遵循。

雖然CRISP-DM能夠指導如何實施資料挖掘，但是它不能解釋資料挖掘是什麼或者為什麼適合這樣做。在本文中我將闡述我提出資料挖掘的九種準則或「定律」（其中大多數為實踐者所熟知）以及另外其它一些熟知的解釋。開始從理論上（不僅僅是描述上）來解釋資料挖掘過程。

我的目的不是評論CRISP-DM，但CRISP-DM的許多概念對於理解資料挖掘是至關重要的，本文也將依賴於CRISP-DM的常見術語。CRISP-DM僅僅是論述這個過程的開始。

第一，目標律：業務目標是所有資料解決方案的源頭。
它定義了資料挖掘的主題：資料挖掘關註解決業務業問題和實現業務目標。資料挖掘主要不是一種技術，而是一個過程，業務目標是它的的核心。
沒有業務目標，沒有資料挖掘（不管這種表述是否清楚）。因此這個準則也可以說成：資料挖掘是業務過程。

第二，知識律：業務知識是資料挖掘過程每一步的核心。
這裡定義了資料挖掘過程的一個關鍵特徵。CRISP-DM的一種樸素的解讀是業務知識僅僅作用於資料挖掘過程開始的目標的定義與最後的結果的實施，這將錯過資料挖掘過程的一個關鍵屬性，即業務知識是每一步的核心。

為了方便理解，我使用CRISP-DM階段來說明：

商業理解必須基於業務知識，所以資料挖掘目標必須是業務目標的映射（這種映射也基於資料知識和資料挖掘知識）；

資料理解使用業務知識理解與業務問題相關的資料，以及它們是如何相關的；

資料預處理就是利用業務知識來塑造資料，使得業務問題可以被提出和解答（更詳盡的第三條—準備律）；

建模是使用資料挖掘演算法創建預測模型，同時解釋模型和業務目標的特點，也就是說理解它們之間的業務相關性；

評估是模型對理解業務的影響；

實施是將資料挖掘結果作用於業務過程；

總之，沒有業務知識，資料挖掘過程的每一步都是無效的，也沒有「純粹的技術」步驟。 業務知識指導過程產生有益的結果，並使得那些有益的結果得到認可。資料挖掘是一個反覆的過程，業務知識是它的核心，驅動著結果的持續改善。

這背後的原因可以用「鴻溝的表現」（chasm of representation）來解釋（Alan Montgomery在20世紀90年代對資料挖掘提出的一個觀點）。Montgomery指出資料挖掘目標涉及到現實的業務，然而資料僅能表示現實的一 部分；資料和現實世界是有差距（或「鴻溝」）的。在資料挖掘過程中，業務知識來彌補這一差距，在資料中無論發現什麼，只有使用業務知識解釋才能顯示其重要 性，資料中的任何遺漏必須通過業務知識彌補。只有業務知識才能彌補這種缺失，這是業務知識為什麼是資料挖掘過程每一步驟的核心的原因。

第三，準備律：資料預處理比資料挖掘其他任何一個過程都重要。
這是資料挖掘著名的格言，資料挖掘項目中最費力的事是資料獲取和預處理。非正式估計，其佔用項目的時間為50%-80%。最簡單的解釋可以概括為「資料是困 難的」，經常採用自動化減輕這個「問題」的資料獲取、資料清理、資料轉換等資料預處理各部分的工作量。雖然自動化技術是有益的，支持者相信這項技術可以減
少資料預處理過程中的大量的工作量，但這也是誤解資料預處理在資料挖掘過程中是必須的原因。

資料預處理的目的是把資料挖掘問題轉化為格式化的資料，使得[url=http://www.finereport.com/tw/]資料分析[/url]技術（如資料挖掘演算法）更容易利用它。資料任何形式的變化（包括清理、最大最小值轉換、增長等）意味著問題空間的變化，因此這種分析必須是探索性的。
這是資料預處理重要的原因，並且在資料挖掘過程中佔有如此大的工作量，這樣資料挖掘者可以從容地操縱問題空間，使得容易找到適合分析他們的方法。

有兩種方法「塑造」這個問題 空間。第一種方法是將資料轉化為可以分析的完全格式化的資料，比如，大多數資料挖掘演算法需要單一表格形式的資料，一個記錄就是一個樣例。資料挖掘者都知道 什麼樣的演算法需要什麼樣的資料形式，因此可以將資料轉化為一個合適的格式。第二種方法是使得資料能夠含有業務問題的更多的信息，例如，某些領域的一些資料 挖掘問題，資料挖掘者可以通過業務知識和資料知識知道這些。
通過這些領域的知識，資料挖掘者通過操縱問題空間可能更容易找到一個合適的技術解決方案。

因此，通過業務知識、資料知識、資料挖掘知識從根本上使得資料預處理更加得心應手。
資料預處理的這些方面並不能通過簡單的自動化實現。

這個定律也解釋了一個有疑義的現象，也就是雖然經過資料獲取、清理、融合等方式創建一個資料倉庫，但是資料預處理仍然是必不可少的，仍然佔有資料挖掘過程一 半以上的工作量。此外，就像CRISP-DM展示的那樣，即使經過了主要的資料預處理階段，在創建一個有用的模型的反覆過程中，進一步的資料預處理的必要的。

第四，試驗律（NFL律：No Free Lunch）：對於資料挖掘者來說，天下沒有免費的午餐，一個正確的模型只有通過試驗（experiment）才能被發現。機器學習有一個原則：如果我們充分了解一個問題空間（problem space），我們可以選擇或設計一個找到最優方案的最有效的演算法。一個卓越演算法的參數依賴於資料挖掘問題空間一組特定的屬性集，這些屬性可以通過分析發現或者演算法創建。但是，這種觀點來自於一個錯誤的思想，在資料挖掘過程中資料挖掘者將問題公式化，然後利用演算法找到解決方法。事實上，資料挖掘者將問題公式化和尋找解決方法是同時進行的——演算法僅僅是幫助資料挖掘者的一個工具。

有五種因素說明試驗對於尋找資料挖掘解決方案是必要的：

資料挖掘項目的業務目標定義了興趣範圍（定義域），資料挖掘目標反映了這一點；

與業務目標相關的資料及其相應的資料挖掘目標是在這個定義域上的資料挖掘過程產生的；

這些過程受規則限制，而這些過程產生的資料反映了這些規則；

在這些過程中，資料挖掘的目的是通過模式發現技術（資料挖掘演算法）和可以解釋這個演算法結果的業務知識相結合的方法來揭示這個定義域上的規則；

資料挖掘需要在這個域上生成相關資料，這些資料含有的模式不可避免地受到這些規則的限制。

在這裡強調一下最後一點，在資料挖掘中改變業務目標，CRISP-DM有所暗示，但經常不易被覺察到。廣為所知的CRISP-DM過程不是下一個步驟僅接著上一個步驟的「瀑布」式的過程。事實上，在項目中的任何地方都可以進行任何CRISP-DM步驟，同樣商業理解也可以存在於任何一個步驟。業務目標不是簡 單地在開始就給定，它貫穿於整個過程。這也許可以解釋一些資料挖掘者在沒有清晰的業務目標的情況下開始項目，他們知道業務目標也是資料挖掘的一個結果，不是靜態地給定。

Wolpert的「沒有免費的午餐」理論已經應用於機器學習領域，無偏的狀態好於（如一個具體的演算法）任何其他可能的問題（資料集）出現的平均狀態。這是因為，如果我們考慮所有可能的問題，他們的解決方法是均勻分布的，以至於一個演算法（或偏倚）對一個子集是有利的，而對另一個子集是不利的。這與資料挖掘者所知的具有驚人的相似性，沒有一個演算法適合每一個問題。但是經 過資料挖掘處理的問題或資料集絕不是隨機的，也不是所有可能問題的均勻分布，他們代表的是一個有偏差的樣本，那麼為什麼要應用NFL的結論？答案涉及到上 面提到的因素：問題空間初始是未知的，多重問題空間可能和每一個資料挖掘目標相關，問題空間可能被資料預處理所操縱，模型不能通過技術手段評估，業務問題本身可能會變化。由於這些原因，資料挖掘問題空間在資料挖掘過程中展開，並且在這個過程中是不斷變化的，以至於在有條件的約束下，用演算法模擬一個隨機選擇的資料集是有效的。對於資料挖掘者來

說：沒有免費的午餐。

這大體上描述了資料 挖掘過程。但是，在有條件限制某些情況下，比如業務目標是穩定的，資料和其預處理是穩定的，一個可接受的演算法或演算法組合可以解決這個問題。在這些情況下，
一般的資料挖掘過程中的步驟將會減少。 但是，如果這種情況穩定是持續的，資料挖掘者的午餐是免費的，或者至少相對便宜的。像這樣的穩定性是臨時的，因為
對資料的業務理解（第二律）和對問題的理解（第九律）都會變化的。

第五，模式律（大衛律）：資料中總含有模式。
這條規律最早由David Watkins提出。 我們可能預料到一些資料挖掘項目會失敗，因為解決業務問題的模式並不存在於資料中，但是這與資料挖掘者的實踐經驗並不相關。

前文的闡述已經提到，這是因為：在一個與業務相關的資料集中總會發現一些有趣的東西，以至於即使一些期望的模式不能被發現，但其他的一些有用的東西可能會被 發現（這與資料挖掘者的實踐經驗是相關的）；除非業務專家期望的模式存在，否則資料挖掘項目不會進行，這不應感到奇怪，因為業務專家通常是對的。

然而，Watkins提出一個更簡單更直接的觀點：「資料中總含有模式。」這與資料挖掘者的經驗比前面的闡述更一致。這個觀點後來經過Watkins修正，基於客戶關係的資料挖掘項目，總是存在著這樣的模式即客戶未來的行為總是和先前的行為相關，顯然這些模式是有利可圖的（Watkins的客戶關係管理定律）。但是，資料挖掘者的經驗不僅僅局限於客戶關係管理問題，任何資料挖掘問題都會存在模式（Watkins的通用律）。

Watkins的通用律解釋如下：

資料挖掘項目的業務目標定義了興趣範圍（定義域），資料挖掘目標反映了這一點；

與業務目標相關的資料及其相應的資料挖掘目標是在這個定義域上的資料挖掘過程產生的；

這些過程受規則限制，而這些過程產生的資料反映了這些規則；

在這些過程中，資料挖掘的目的是通過模式發現技術（資料挖掘演算法）和可以解釋這個演算法結果的業務知識相結合的方法來揭示這個定義域上的規則；

資料挖掘需要在這個域上生成相關資料，這些資料含有的模式不可避免地受到這些規則的限制。

總結這一觀點：資料中總存在模式，因為在這過程中不可避免產生資料這樣的副產品。為了發掘模式，過程從（你已經知道它）——業務知識開始。

利用業務知識發現模式也是一個反覆的過程；這些模式也對業務知識有貢獻，同時業務知識是解釋模式的主要因素。在這種反覆的過程中，資料挖掘演算法簡單地連接了業務知識和隱藏的模式。

如果這個解釋是正確的，那麼大衛律是完全通用的。除非沒有相關的資料的保證，否則在每個定義域的每一個資料挖掘問題總是存在模式的。

第六，洞察律：資料挖掘增大對業務的認知。
資料挖掘是如何產生洞察力的？這個定律接近了資料挖掘的核心：為什麼資料挖掘必須是一個業務過程而不是一個技術過程。業務問題是由人而非演算法解決的。資料挖 掘者和業務專家從問題中找到解決方案，即從問題的定義域上達到業務目標需要的模式。資料挖掘完全或部分有助於這個認知過程。資料挖掘演算法揭示的模式通常不 是人類以正常的方式所能認識到的。綜合這些演算法和人類正常的感知的資料挖掘過程在本質上是敏捷的。在資料挖掘過程中，問題解決者解釋資料挖掘演算法產生的結 果，並統一到業務理解上，因此這是一個業務過程。

這類似於「智能放大器」的概念，在早期的人工智慧的領域，AI的第一個實際成果不是智能機器，而是被稱為「智能放大器」的工具，它能夠協助人類使用者提高獲取有效信息的能力。資料挖掘提供一個類似的「智能放大器」，幫助業務專家解決他們不能單獨完成的業務問題。

總之，資料挖掘演算法提供一種超越人類以正常方式探索模式的能力，資料挖掘過程允許資料挖掘者和業務專家將這

種能力融合在他們的各自的問題的中和業務過程中。

第七，預測律：預測提高了信息泛化能力。
「預測」已經成為資料挖掘模型可以做什麼的可接受的描述，即我們常說的「預測模型」和「預測分析」。這是因為許多流行的資料挖掘模型經常使用「預測最可能的結果」（或者解釋可能的結果如何有可能）。這種方法是分類和回歸模型的典型應用。

但是，其他類型的資料挖掘模型，比如聚類和關聯模型也有「預測」的特徵。這是一個含義比較模糊的術語。一個聚類模型被描述為「預測」一個個體屬於哪個群體，一個關聯模型可能被描述為基於已知基本屬性「預測」一個或更多屬性。

同樣我們也可以分析「預測」這個術語在不同的主題中的應用：一個分類模型可能被說成可以預測客戶行為—更加確切的說它可以預測以某種確定行為的目標客戶，即使不是所有的目標個體的行為都符合「預測」的結果。一個詐騙檢測模型可能被說成可以預測個別交易是否具有高風險性，即使不是所有的預測的交易都有欺詐行為。

「預測」這個術語廣泛的使用導致了所謂的「預測分析」被作為資料挖掘的總稱，並且在業務解決方案中得到了廣泛的應用。但是我們應該意識到這不是日常所說的「預測」，我們不能期望預測一個特殊個體的行為或者一個特別的欺詐調查結果。

那麼，在這個意義下的「預測」是什麼？分類、回歸、聚類和 關 聯演算法以及他們集成模型有什麼共性呢？答案在於「評分」，這是預測模型應用到一個新樣例的方式。模型產生一個預估值或評分，這是這個樣例的新信息的一部分；在概括和歸納的基礎上，這個樣例的可利用信息得到了提高，模式被演算法發現和模型具體化。值得注意的是這個新信息不是在「給定」意義上的「資料」，它僅 有統計學意義。

第八，價值律：資料挖掘的結果的價值不取決於模型的穩定性或預測的準確性。
準確性和穩定性是預測模型常用的兩個度量。準確性是指正確的預測結果所佔的比例；穩定性是指當創建模型的資料改變時，用於同一口徑的預測資料，其預測結果變 化有多大（或多小）。鑒於資料挖掘中預測概念的核心角色，一個預測模型的準確性和穩定性常被認為決定了其結果的價值的大小，實際上並非如此。

體現預測模型價值的有兩種方式：一種是用模型的預測結果來改善或影響行為，另一種是模型能夠傳遞導致改變策略的見解（或新知識）。

對於後者，傳遞出的任何新知識的價值和準確性的聯繫並不那麼緊密；一些模型的預測能力可能有必要使我們相信發現的模式是真實的。然而，一個難以理解的複雜的 或者完全不透明的模型的預測結果具有高準確性，但傳遞的知識也不是那麼有見地；然而，一個簡單的低準確度的模型可能傳遞出更有用的見解。

準確性和價值之間的分離在改善行為的情況下並不明顯，然而一個突出問題是「預測模型是為了正確的事，還是為了正確的原因?」 換句話說，一個模型的價值和它的預測準確度一樣，都源自它的業務問題。例如，客戶流失模型可能需要高的預測準確度，否則對於業務上的指導不會那麼有效。相 反的是一個準確度高的客戶流失模型可能提供有效的指導，保留住老客戶，但也僅僅是最少利潤客戶群體的一部分。如果不適合業務問題，高準確度並不能提高模型的價值。

模型穩定性同樣如此，雖然穩定性是預測模型的有趣的度量，穩定性不能代替模型提供業務理解的能力或解決業務問題，其它技術手段也是如此。

總之，預測模型的價值不是由技術指標決定的。資料挖掘者應該在模型不損害業務理解和適應業務問題的情況下關注預測準確度、模型穩定性以及其它的技術度量。

第九，變化律：所有的模式因業務變化而變化。
資料挖掘發現的模式不是永遠不變的。資料挖掘的許多應用是眾所周知的，但是這個性質的普遍性沒有得到廣泛的重視。

資料挖掘在市場營銷和CRM方面的應用很容易理解，客戶行為模式隨著時間的變化而變化。行為的變化、市場的變化、競爭的變化以及整個經濟形勢的變化，預測模型會因這些變化而過時，當他們不能準確預測時，應當定期更新。

資料挖掘在欺詐模型和風險模型的應用中同樣如此，隨著環境的變化欺詐行為也在變化，因為罪犯要改變行為以保持領先於反欺詐。欺詐檢測的應用必須設計為就像處理舊的、熟悉的欺詐行為一樣能夠處理新的、未知類型的欺詐行為。

某些種類的資料挖掘可能被認為發現的模式不會隨時間而變化，比如資料挖掘在科學上的應用，我們有沒有發現不變的普遍的規律？也許令人驚奇的是，答案是即使是這些模式也期望得到改變。理由是這些模式並不是簡單的存在於這個世界上的規則，而是資料的反應—-這些規則可能在某些領域確實是靜態的。

然而，資料挖掘發現的模式是認知過程的一部分，是資料挖掘在資料描述的世界與觀測者或業務專家的認知之間建立的一個動態過程。因為我們的認知在持續發展和增 長，所以我們也期望模式也會變化。明天的資料表面上看起來相似，但是它可能已經集合了不同的模式、（可能巧妙地）不同的目的、不同的語義；分析過程因受業務知識驅動，所以會隨著業務知識的變化而變化。基於這些原因，模式會有所不同。

總之，所有的模式都會變化，因為他們不僅反映了一個變化的世界，也反映了我們變化的認知。

後記：

這九條定律是關於資料挖掘的簡單的真知。這九條定律的大部分已為資料挖掘者熟知，但仍有一些不熟悉（例如，第五、第六、第七）。大多數新觀點的解釋都和這九條定律有關，它試圖解釋眾所周知的資料挖掘過程中的背後的原因。

我們為什麼何必在意資料挖掘過程所採用的形式呢？除了知識和理解這些簡單的訴求，有實實在在的理由去探討這些問題。

資料挖掘過程以現在的形式存在是因為技術的發展—機器學習演算法的普及以及綜合其它技術集成這些演算法的平台的發展，使得商業用戶易於接受。我們是否應該期望因技術的改變而改變資料挖掘過程？最終它會改變，但是如果我們理解資料挖掘過程形成的原因，然後我們可以辨別技術可以改變的和不能改變的。

一些技術的發展在預測分析領域具有革命性的作用，例如資料預處理的自動化、模型的重建以及在部署的框架里通過預測模型集成業務規則。資料挖掘的九條定律及其 解釋說明：技術的發展不會改變資料挖掘過程的本質。這九條定律以及這些思想的進一步發展，除了有對資料挖掘者的教育價值之外，應該被用來判別未來任何資料挖掘過程革命性變化的訴求。