TeraSoft 鈦思科技

利用 MATLAB 資料解析功能進行以數據主導的洞察力分析: 能源負載預測為例

By Seth DeLand and Adam Filion, MathWorks

不管是能源製造商、電網營運商或貿易商，做決策時都必須考慮電網的未來預估負載量，因此，精準的能源負載預測變成一種必須性且具有商業價值。

現今，對於龐大資料數據的易於取得性，使得我們可以建立高準確度的預測模型。然而在發展資料解析(data analytics)流程的過程中，困難之處在於如何把原始資料轉變成可執行的洞悉觀點(insight)。典型的資料解析工作流程包含四個步驟，每步驟都有其挑戰：

1. 匯入不同來源的資料，例如網路檔案、資料庫、電子表格。
2. 剔除異常數值、雜訊，及彙整這些資料集。
3. 使用機器學習技術，根據彙總的資料開發準確的預測模型。
4. 把模型套用到實際生產環境。

在這篇文章中，我們將利用 MATLAB® 來完成一個電力負載預測應用的完整資料解析工作流程。在本應用中，電力事業分析師可以選擇紐約州的任何一個地區，查看該區過去的能源負載情況並預測未來的負載(圖 1)。根據這個結果，他們可以了解天氣對於能源負載的影響進而決定需要產生或購買多少的電力。光是紐約州一州每年就花費數十億美元在電力上，足足可見這個研究結果對於電力公司所可能帶來的效益。

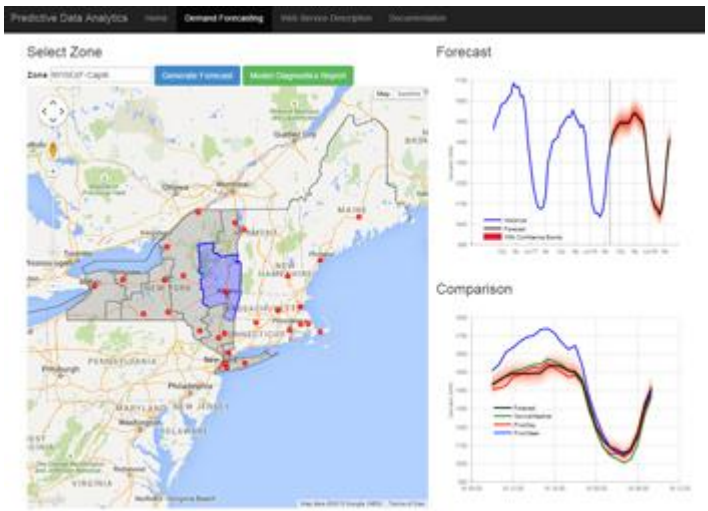


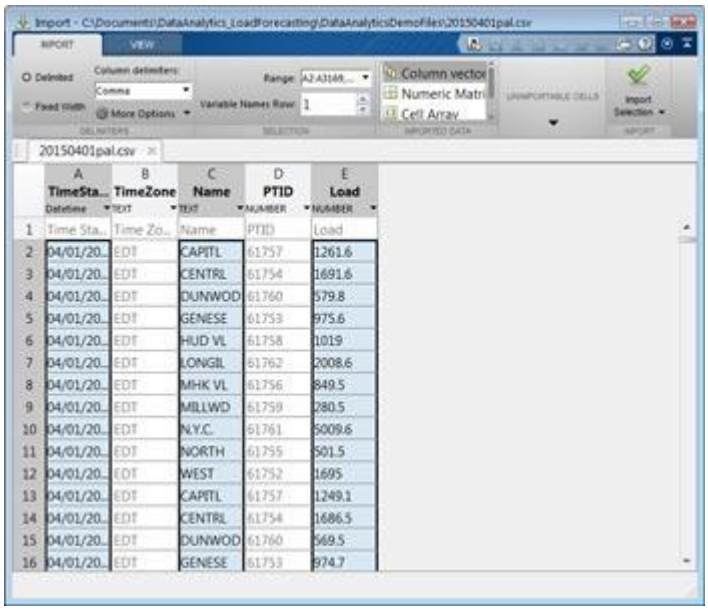
圖 1. MATLAB 在紐約州能源需求預測上的應用範例

資料的匯入與匯出

這個案例使用了兩個資料集：紐約獨立系統管理網站(NYISO, New York Independent System Operator)的電力負載資料以及美國國家氣象資料中心(National Climatic Data Center)的天氣資料(本案例使用溫度及降露點資料)。

在 NYISO 每個月出版的電力資料壓縮檔中，每一天都有一個單獨的 CSV 檔。一般來說，處理這類分散在多個檔案資料的作法為下載一個範例檔，從範例檔中找出需要分析的數據值，再將這些數據值匯入到完整的資料集。

MATLAB 的匯入工具可以讓我們選取 CSV 檔裡的欄位，再將這些選取的資料匯入成各式各樣的 MATLAB 資料結構，像是向量、矩陣、資料格(cell)陣列及表格。我們下載的這個能源負載 CSV 檔包括了時間標記、地區名稱、地區的負載等資料；透過 MATLAB 的匯入工具，我們選取 CSV 檔的欄位以及想要的格式，即可以從範例檔直接匯入資料，或產生一個 MATLAB 函式來匯入所有符合範例檔格式的檔案(圖 2)。接下來，我們可以編寫一個指令來調用這個 MATLAB 函式，透過程式語言從資料來源匯入所有的數據。



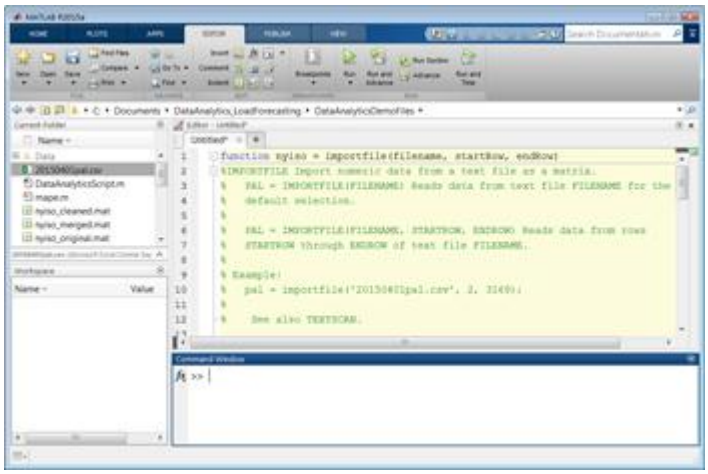


圖 2. 上: 選取欲匯入的 CVS 數據

下: 自動產生 MATLAB 資料匯入函式

完成資料匯入後，我們可以製作初步的曲線圖來了解趨勢，重新編排時間與日期標示，並執行轉換，例如資料表中欄與列的交換。

資料清理及彙整

現實世界中大多數的資料都參雜一些遺失或錯誤的數值，在進一步探討數據之前，這些數值必須要先被辨識及處理。在將 NYISO 數據重新格式化並繪製成圖形之後，我們可以看到電力負載的峰值落在需求正常週期浮動的外側(圖 3)。我們必須判斷這些峰值是否為資料模型中可以忽略的異常數值，還是它們反映出某些應該要列入模型中考量的現象。現在，我們暫先選擇正常週期內的行為來檢視；峰值則待之後我們發現這些現象有必要被列入考量的時候再來處理。

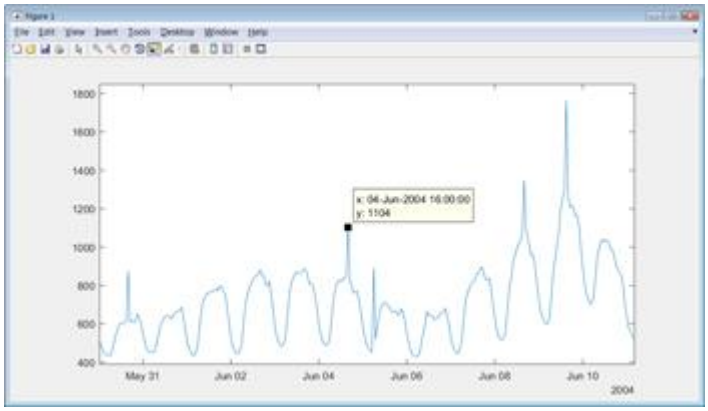


圖 3. 能源負載需求曲線上的異常峰值

有幾種方法可以把峰值的辨識自動化。比方說，我們可以使用樣條平滑(smoothing spline)並透過計算原始曲線與經過平滑處理的曲線之間的差異來定義出峰值精確位置(圖 4)。

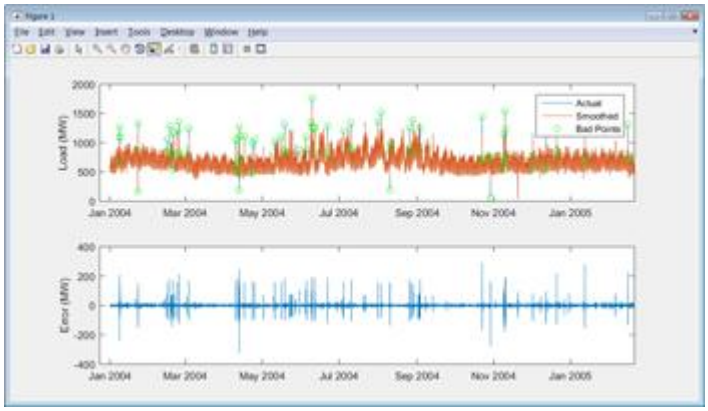


圖 4. 上: 實際負載、經過平滑處理的負載與異常值的標示圖

下: 實際數值與平滑處理數值的差異圖

把異常數值從資料中剔除後，接下來要決定的是如何處理因為刪除這些數值而遺失的數據點。第一種方法是直接忽略遺失的數值，這具有縮小數據集的優點。或者，我們也可以插入或從另一個可比較的樣本找出近似值來代替這些從 MATLAB 消失的數值，但千萬注意不要造成資料偏誤(bias)。基於估計負載量的目的，我們這次選擇忽略遺失的數值，但我們還是握有足夠的"好"數據來產生精確的模型。

接著用相同技巧篩選過氣溫和降露點數據之後，我們把這兩個資料集匯整，儲存在 MATLAB 的資料格式表中。藉由調用外接函式功能把各項選取的資料通通加入 MATLAB 表格，然後我們會看到各時間點的能源負載、氣溫、降露點等資訊都很輕易的被整合在同一個表格裡。

建立預測模型

MATLAB 提供許多建立模型資料的相關功能。如果我們已知不同參數如何影響電力負載，則可以用統計或曲線契合工具的線性或非線性迴歸分析把數據模型化。然而，如果存在多個變數、根本系統特別複雜，或支配的方程式未知，我們可以利用像是決策樹或類神經網路等機器學習功能來建立模型。

由於負載預測涉及到包含多個應考慮變數的複雜系統，我們選擇機器學習，更精確地來說--監督式學習(supervised learning)--來建立模型。監督式學習的模型是

依據歷史輸入數據(氣溫)及輸出數據(能源負載)來建立。模型經過訓練之後，便可以對未來的情況進行預測。在這個能源負載預測案例，我們可以使用類神經網路功能與類神經網路工具箱(Neural Network Toolbox™)來完成以下這些步驟：

I. 使用 MATLAB 中的類神經契合應用程式(Neural Fitting app)：

   a. 把我們認為與負載預測有關的變數具體化，包含每小時、每天的氣溫和降露點

   b. 選擇滯後指標，像是過去 24 小時內的負載

   c. 訂定目標，或想要預測的變數，在這個案例為電力負載

2. 選擇我們想要用來做模型訓練的數據集，還有想要保留來供測試用的數據集。

在這個案例，我們僅選擇單一個模型；然而在現實世界中大部分的應用，必須嘗試好幾個不同的機器學習模型來評估他們在訓練及測試數據時的表現。統計與機器學習工具箱(Statistics and Machine Learning Toolbox™)提供多種使用相似調用語法的機器學習方法，讓這些嘗試變得更加容易。這個工具箱同時包含了分類學習應用程式(Classification Learner app)，有助於監督式學習模型訓練的互動。

訓練完成之後，我們可以使用測試數據來檢測模型在執行新數據上的表現(圖5)。

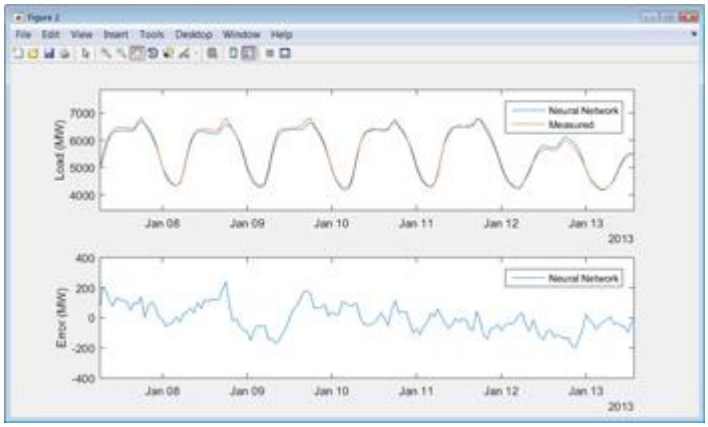


圖 5. 上：實測負載與類神經網路預測結果曲線比對圖

下：實測與預測值差異曲線圖

我們還可以使用類神經契合應用程式中，調用一個指令來產生MATLAB程式碼，讓設定、訓練、測試類神經網路的這些動作自動執行。

為了要測試這個訓練過的模型，我們執行它，並與之前所保留的、為了比對實際量測值與其預測結果的資料做比較，結果顯示這個類神經網路模型在測試數據的平均絕對值誤差率(mean absolute percent error, MAPE)低於 2%。

當我們開始把測試數據集套入我們的模型，我們發現少數模型預測值與實際負載值明顯偏離：例如在假日附近會出現預測上的偏差。另外，我們也發現在 2012 年 10 月 29 日，紐約市的模型預測負載量與實際偵測值相差了數千兆瓦(圖 6)。經由網路搜尋，很快就可以知道那天正值颶風桑迪肆虐，整個地區的電網遭受嚴重損壞。把假日這種規律且可預測的變數加入模型中看來相當合理，但像桑迪颶風這種突然、偶發性的事件則難以列入評估。

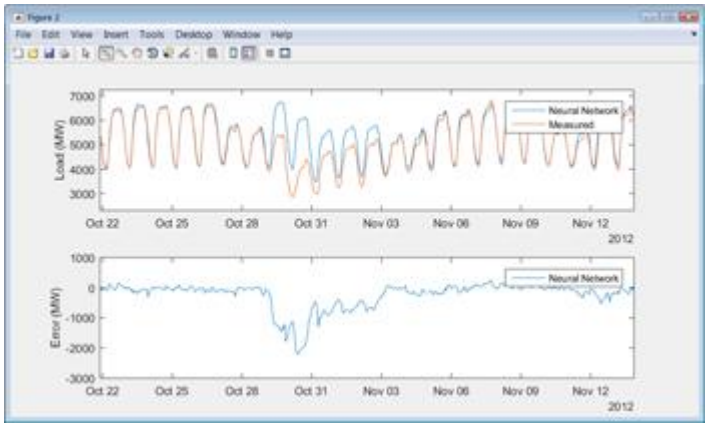


圖 6. 2012/10/29 紐約市的實際與預測電力負載值落差

開發、測試、及改善一個預測模型需要經過多次的重複作業。使用平行運算工具箱(Parallel Computing Toolbox™)在多核心處理器讓多個步驟同時運行可以縮短訓練及測試的時間。如果遇到非常大的數據集，您可以透過 MATLAB 分散式運算引擎(Distributed Computing Server™)在多台電腦上同時操作這些步驟。

將模型發展成實際之應用

當模型的準確度已經可以達到我們的要求，最後一個步驟便是把模型應用到實際的生產系統。我們有幾個做法可以選擇。首先可使用 MATLAB 編譯器(MATLAB Compiler™)產生一個可獨立執行的應用程式或內建的工作表格；其次，使用編譯器增益集 SDK (MATLAB Compiler SDK™)產生.NET 和 Java® 元件；或者是

使用 MATLAB 生產伺服器(MATLAB Production Server™)把應用程式套用到能同時服務大量用戶的生產環境。

為了完成我們的負載預測工具，我們在 MATLAB 透過 RESTful API 做了一個數據解析，可以回報預測的數字並繪製出圖表放在應用程式或報告中。有了生產伺服器編譯器應用程式，我們可以明確地定義出我們想要 MATLAB 函式。這個應用程式會自動執行相依分析並把必要的檔案打包到一個可套用元件。我們把這個元件當作處理引擎，套用到 MATLAB 生產伺服器，讓這個分析能夠適用於任何網內的軟體或裝置，包含網路應用程式，其他伺服器及行動裝置(圖 7)。

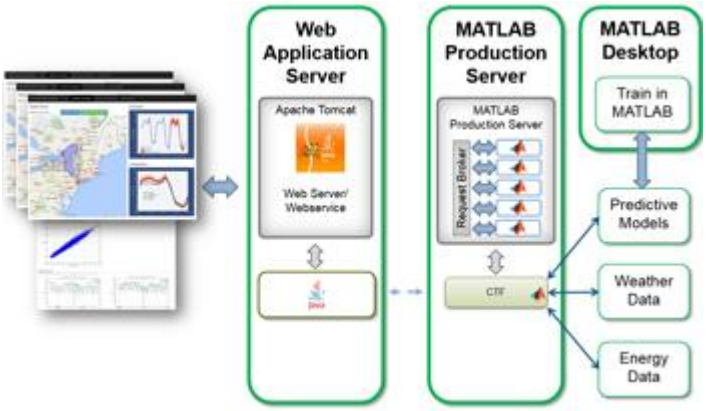


圖 7. 使用 Apache Tomcat 及 MATLAB 生產伺服器套用到生產環境的 MATLAB 數據解析

下一步

由於這個模型已經使用超過數個月的測試數據來做驗證，我們相信它可以提供誤差在 2%以內的 24 小時預測。決策者可以透過 Web 前端使用這個高度準確的能源負載預測模型。

這個模型還可以結合其他數據資源做延伸，像是加入節日日曆、劇烈天氣警報等。因為這整個數據分析工作流程都以 MATLAB 編碼記錄，來自外部資源的數據也可以很容易地跟現有數據合併、整合，進而重新訓練模型。當新模型被套用到 MATLAB 生產伺服器時，終端用戶甚至不需要重新整理網頁，嵌在負載預測應用程式裡的演算法就會自動地被更新。