以資料採礦技術進行太陽能全日照量預測

楊琇如[1]　　尹相志[2]
核能研究所[1]　　亞洲資採[2]

摘　要

為進行資料採礦對日照預測之可行性研究，本研究使用資料採礦技術對氣象資料庫進行資料挖掘，利用資料採礦的方法對資料庫變數進行整理與變量篩選，整個資料採礦的步驟包含變數檢定、預測模型建立及模型評估驗證。

本篇重點於介紹資料採礦之過程及發現，首先收集核能研究所自行架設於桃園龍潭區的氣象環境資料、行政院環保署之懸浮粒子資料、及中央氣象局日出日落之資料，經過變數檢定之過程後，篩選出對預測目標影響較顯著之變數進行日照預測模型建置。本研究分為兩階段建模，第一階段模型採用羅吉斯迴歸統計模型建模，且利用全日照達標率分佈、Lorenz's Curve、及K-S Test模型評估技術評估模型效度，再經過反覆調校模型參數，以產生最佳效度日照預測模型。最後將模型規則轉換為評分卡方式，方便使用者了解，且便於後續之模型自動化佈署，可實際應用於太陽日照達標機率之預測。

第二階段模型採用線性迴歸(Linear Regression)技術建置日照量數值之推估(Estimation)模型，並採用皮爾森相關係數(Pearson Correlation)指標做為評估效度之依據，用以檢定各變數之預測力，同時透過不同的轉換函數，以檢視不同轉換模式下的變數預測力強弱。為可以確保模型的穩定性不至於因為連續變數間高度重疊造成偏誤，在第二階段模型中，引入「因子分析」的概念來進行變數處理，以建置應用於太陽能全日照量之預測模型。

關鍵字：資料探勘、羅吉斯迴歸、全日照達標率分佈、羅倫茲曲線、克默果夫－史密諾夫檢定、線性迴歸、皮爾森相關係數、因子分析

一、前言

由於核能發電造成環境的影響及能源逐漸耗竭所造成能源價格居高不下等問題，讓各國開始積極的發展再生能源，而目前太陽能發電將會是未來發電的趨勢。

太陽能發電之所以會廣為各國所發展，原因在於陽光取之不盡用之不竭，另外，煤炭、石油等礦物燃料會產生有害氣體和廢渣，而使用太陽能不會消耗其他地球的資源也不會排放出任何對環境不良影響的物質而造成溫室效應與環境的污染，是一種清潔的能源。且核能發電會有核洩漏的危險，一旦核洩漏便會造成極大的生態危機，相較之下太陽能發電的使用較其他發電方式安全。

因此，為了能使智慧電網更有效率配置能源，本研究藉由日照預測模型利用歷史氣象相關資料的挖掘與分析，是能有效的預測全日照值是否達標準的重要工具，並且要進一步根據達標之案例，同樣運用氣候等資訊，以評估日照量的絕對數值。

日照預測模型是根據歷史氣象資料，運用統計模式建立預測模型，主要分為兩階段建模，第一階段是預測是否達標，第二階段則是根據達標案例推估日照量。其中運用現場各種傳感器(Sensor)每分鐘所收集之氣象數據，運用大數據統計模式建立分類及推估模型，以客觀的方式預測是否達發電標準並推估未來15分鐘之日照量。運用此法具有以下好處：減少人為評估的主觀偏差、預測基準一致化以及自動化流程、 加速智慧電網自動化調節評估流程。

二、研究範圍

本研究範圍摘要第一階段分類模型為：（一）本次日照預測(Sunshine forecast)採取觀測點綜合指標、日出日落時間、以及懸浮微粒資料為主要資料源，用以進行太陽能預測技術研發（二）全日照值是否達標定義為PSP(全日照值)>=0.3 kW/㎡（三）採用羅吉斯迴歸(Logistics Regression)技術建置日照預測模型（四）採用WOE (Weight of Evidence)以及IV (Information Value)兩項指標，用以檢定各變數之預測力，同時用以決定連續變數之切割點（五）協助計算出全日照達標機率值；第二階段推估模型為：（一）日照預測(Sunshine forecast)採取觀測點綜合指

標、日出日落時間以及氣候觀測數據為主要資料源，用以預估未來15分鐘之日照量，以便進行相關之電網調節作業(二)採用大數據分析之推估技術(包括線性迴歸、決策樹、類神經網路等)建置日照量推估模型。為了方便核能研究所進行自動化預測，因此模型結果產出以線性迴歸結果為主（三）採用皮爾森相關係數(Pearson Correlation)指標，用以檢定各變數之預測力，同時透過不同的轉換函數，以檢視不同轉換模式下的變數預測力強弱（四）透過因子分析來進行資料降維提升預測準確度（五）評估模型效益。本研究將根據核能研究所自行建置之氣象環境設備所收集之觀測點氣象資料、行政院環保署收集之懸浮粒子資料、及中央氣象局收集之日出日落資料，建立預測模型，研究範圍包括規劃模型建置取樣流程、選取樣本以供回溯預測時之全日照達標狀況及推估未來15分鐘全日照量、建置太陽能全日照量預測模型、評估全日照達標機率及推估全日照量之模型準確度[1,2]。

三、研究方法

本研究第一階段分類模型，根據氣象每半小時預測一次時間進行模型的切割，依據樣本時窗設定，開始建立回溯資料樣本。首先對樣品進行時窗切割，在時窗切割過程中，我們必須將時間點區隔為兩個部份[3,4,5]：

（一） 抽樣時窗(Sample Windows)：進行預測時，必須回溯多久以前的氣象歷史資訊。

（二） 觀察時窗(Performance Windows)：進行預測時，要預估未來多久氣象日照的結果。

分割後之樣本結構如表1。

接著採用WOE及IV方法進行變數檢定，WOE公式如下：

- WOE= ln(%Good / %Bad)

首先根據變數選項(僅能處理類別變數，因此連續變數必須事先切割)統計各級距之達發電效率與未達發電效率數量，同時計算該級距達發電效率佔總體達發電效率之比率及未達發電效率佔總體未達發電效率之比率。其中，WOE越大，代表該級距之資料品質越好，WOE越小(或是負數)則代表該級距的資料品質越差，如果WOE接近於零，表示接近平均水準，因此我們可以使用WOE作為判斷變數選項重要性之依據。

另外採用Information Value作為判斷變數整體重要性的指標，Information Value公式如下：

- Information Value= 100* Σ{(%Good-%Bad)*WOE}

Information Value是大於零的正數，當值越大時則代表越顯著，當值越接近零時，代表該變數預測力越弱，通常我們可以依照以下規則來判斷該變數之預測力強弱：

- <2： 無預測力
- 2~10: 弱
- 10~30： 中等
- 30+： 強

最後使用羅吉斯迴歸(Logistics Regression)技術建置模型，羅吉斯迴歸類似線性迴歸模式。迴歸分析是描述一個依變數與一個或多個預測變數之間的關係，然而一般迴歸分析時，依變數與自變數通常均為連續變數，但羅吉斯迴歸所探討結果的依變數是離散型，應用於二元分類模型。

利用羅吉斯迴歸的目的是在於建立一個最精簡和最能配適(fit)的分析結果，而且在實用上合理的模式，建立模式後可用來預測依變數與一組預測變數之間的關係，羅吉斯迴歸特性如下，線性方程式如圖1：

（一） Ln(P/1-P)=線性方程式

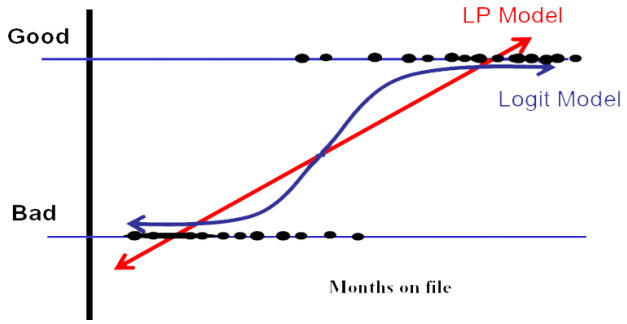
（二） P/(1-P)=odds rate

（三） 二元分類

（四） 輸入變數必須符合迴歸方程式，呈現單調遞增或是單調遞減趨勢

表1 分類模型樣本結構表

	切分時點	Total
訓練組	2011/01/01 07:00:00 ~ 2013/05/31 18:30:00	21,359
測試組	2013/06/01 05:30:00 ~ 2014/06/30 18:30:00	9,672



$$\ln\left[p/(1-p)\right] = a_0 + a_1x_1 + \cdots + a_kx_k$$

圖1 羅吉斯分析概念

透過上述之變數檢定過程產出之分類模型建模變數如表2。

表2 分類模型建模變數

使用變數	IV值
AirTempC	66.81
AvgUVB24hour	26.38
DiffAirTempC24hour	27.07
DiffRH30min	47.4
DiffUVB60min	93.86
MaxRH30min	77.64
AvgRH24hour	24.34
DiffAirTempC60min	157.43
MinPSP30min	332.7
StdevNIP24hour	88.46
StdevNIP30min	262.61
MinuteToSunrise	105.95

第二階段推估模型，根據氣象每15分鐘預測一次(預報未來15分鐘後GHI)時間進行模型的切分，因此根據時窗設定，開始建立回溯資料樣本 ，樣本結構如表2：

表2 推估模型樣本結構表

	切分時點	Total	平均 PSP
訓練組	2011/01/01 07:00:00 ~2013/05/31 23:59:50	14139	0.563
測試組	2013/06/01 00:00:00 ~2014/06/30 18:30:00	7169	0.560

第二階段透過大數據分析技術建置全天空輻射量推估模型，由於預測標的是連續變數，將使用推估(Estimation)類的演算法來執行分析預測。為了確保模型的穩定性與預測力，因此將先透過變數檢定技術找出關鍵變數。

在統計學中，可以運用皮爾森相關係數的概念來檢定兩個變數之間的關連性，簡單來說，相關係數就等於兩個變數的共變數除以兩者的標準差，公式如下。

- XY的相關係數
  R= COVXY/(STDEVX *STDEVy)
- XY的共變異數
  $COV_{XY} = \frac{\sum XY}{n} - \overline{X}\overline{Y}$
- X的標準差
  $STDEV_X$=平方根($(\Sigma X^2/n)-x^2$)

相關係數結果通常介於正負 1 之間，數值越接近 1，則代表兩者越呈現正相關；反之則是負相關，若數值接近零，則代表兩者的關連性較為薄弱。雖然兩者之間相關性高，但究竟實際值與預測值之間有多接近？此時可以利用 R2 來作為客觀的評估指標，其公式如下：

$$R^2=\frac{\text{利用迴歸線解釋的變異 數}}{\text{總體的變異數}}=\frac{SSR}{SST}$$

SSR (Regression sum of square)= $\sum(\hat{Y}-\bar{Y})^2$ 預測值與平均值的差平方

SST (Total sum of squares) = $\sum(Y-\bar{Y})^2$ 實際值與平均值的差平方

變數檢定部份，在篩選變數處理過程中，由於並非所有的輸入變數都與全天空輻射量呈現線性關係，為了確保最佳的預測能力，利用透過不同的線性轉換模式，將輸入變數進行轉換，轉換模式如下：

（一） 原始數據
（二） 開根號轉換
（三） 開對數轉換(僅針對值域較高，且無負值與零的變數轉換)
（四） 倒數轉換(僅針對相關係數為負值者進行轉換)
（五） Logit轉換(僅針對分布高度左偏的案例進行轉換)

根據以上資料預處理流程以及變數篩選標準，可以產出初步之候選變數，以供作為預測模型之使用。

透過上述推估模型之變數檢定結果，可產生第二階段推估模型之建模變數如表3。

表3 推估模型建模變數

衍生變數	相關係數
AvgPSP5min	0.926
AvgPSP5min_LN	0.926
AvgPSP5min_LOGIT	0.926
AvgPSP5min_EXP	0.912
AvgPSP5min_SQ	0.907
Factor1	0.892
TempPSP5min	0.889
AvgUV5min	0.887
AvgUV5min_SQ	0.872
RHPSP5min	0.855
PSPUV5min	0.836
AvgUV5min_LN	0.81
RHUV5min	0.808
AvgNIP5min_LOGIT	0.727
AvgNIP5min_LN	0.726
TempUV5min	0.726
AvgNIP5min_SQ	0.725

透過表 3 之建模變數表，建置推估未來 15 分鐘之全日照量模型。

四、研究結果

第一階段分類模型，本研究採用採用以下技術進行模型評估：

（一） 全日照達標率分佈

全日照達標與全日照未達標在經過建模之後，應該呈現兩個分離的常態分配，全日照達標的平均分數會高於全日照未達標，如圖2所示，當這兩個常態分配分離的越開，就代表模型鑑別全日照達標案件的效果越顯著。

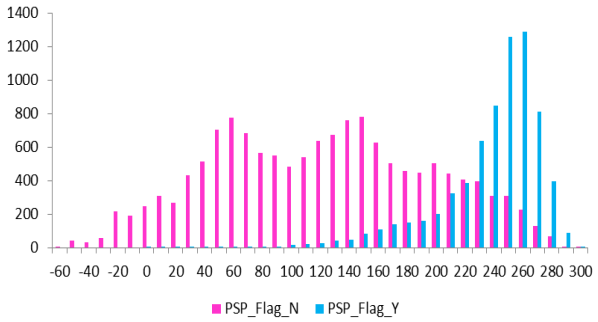


圖2 全日照達標案件分佈(橫軸:分數;縱軸:案件數)

亦可將每個分數區間轉換為全日照達標率，由於分數是假設每隔20分Odds加倍，因此全日照達標率應該隨著分數減少而呈現指數的增加。從圖3評估結果，全日照達標率呈現正確的指數分佈。

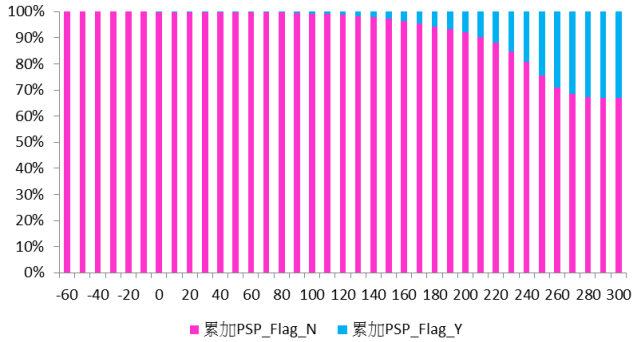


圖3 全日照達標百分比分佈(橫軸:分數;縱軸:案件累加百分比)

（二） Lorenz's Curve

Lorenz’ s Curve是用來評估模型分類效果的標準圖表，所謂的Lorenz’ s Curve的橫軸是根據分數由高到低，累計的全日照達標佔總全日照達標的比例，而縱軸則是分數由高到低，累計的全日照未達標佔總全日照未達標的比例。由於分數高者為達標率高之樣本，因此累積全日照未達標比例的成長速度會慢於累積全日照達標，因此Lorenz’ s Curve會呈現下凹的曲線。

圖4 Lorenz’ s Curve圖中的45度斜線代表隨機的模型，也代表沒有任何鑑別度。當模型越往右下方突出，則代表該預測模型成效越佳。

在Lorenz’ s Curve圖中，向右下突出的半月型面積除與45度線下方三角型面積的比值，我們稱之為Gini Coeficient，當Gini Coeeficient越接近1則代表越接近完美預測，當Gini係數接近零時則代表無預測力，實務上Gini 係數必須至少達到0.4 始具預測力。

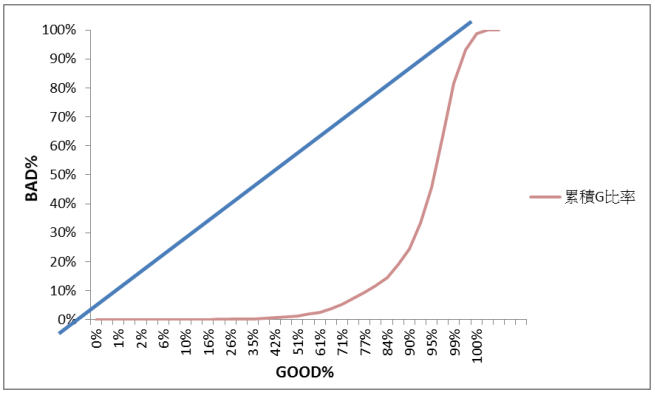


圖4 Lorenz’ s Curve

Gini係數是目前模型評估的標準指標，根據Gini係數能夠掌握模型之預測能力。在本研究中，經過模型多次調整與改版，其Gini係數提升至81.6%，如表4。

表4 模型Gini係數表

	Gini	K-S
模型效度	81.6%	67.4%

（三） K-S Test

K-S Test則是用來評估在那個區間能夠將全日照達標與全日照未達標分離的最開。

K-S Test的橫軸是模型分數，而圖5中三條線分別是分數由低到高累積全日照達標佔總體全日照達標的比率(圖5 紅色線)、分數由低到高累積全日照未達標佔總體全日照未達標的比率(圖5 藍色線)，以及全日照達標減去全日照未達標的比率差值(圖5綠色線)，其中機率差值這條線就是K-S值。

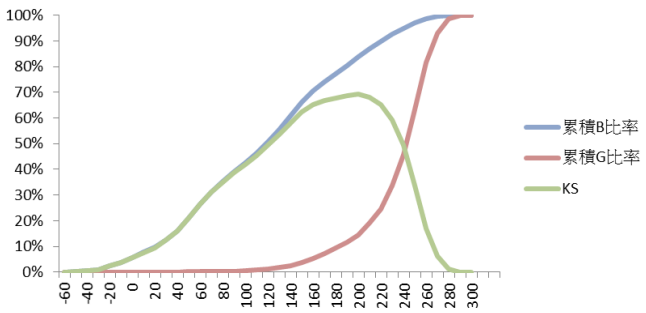


圖5 K-S Test

K-S值表示全日照達標案件與全日照未達標案件累積機率的差值，K-S越高，則代表兩者分離的越開，因此K-S曲線的最高點代表鑑別全日照達標案件與全日照未達標案件的最佳點。本研究中，K-S曲線最高點出現在67.4%。

第二階段之推估模型評估成果最簡單的模式就是將實計值與預測值對比進行繪圖，若是預測結果越接近，則圖形分佈會趨近於45度直線。如圖6即為未來15分鐘全日照量實際值與預測值之散佈圖。

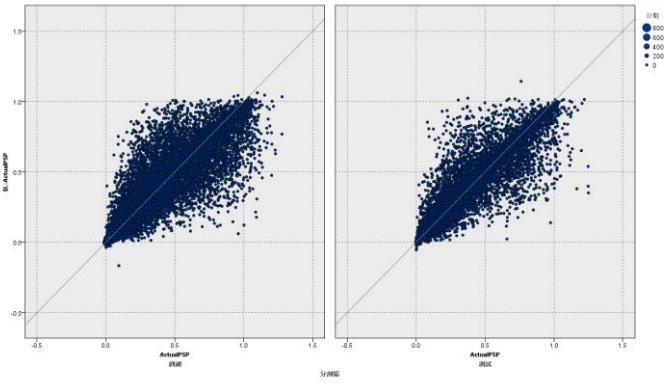


圖6 未來15分鐘預測全日照量效度評估圖(散布圖)

左圖為訓練組資料，右圖為測試組資料，分別相關係圖如下表5。

表5 推估模型之效度評估(樣本為全量樣本)

	訓練組	測試組
預測未來 5 分鐘	0.959	0.957
預測未來 15 分鐘	0.930	0.931

五、結論

在此次研究中採用二階段建模，第一階段建模為判斷未來全日照量是否達到PV太陽能模組穩定發電標準，第二階段建模為若未來日照量達發電標準即進行未來15分鐘全日照量之絕對值預測。

模型評估結果第一階段之分類模型，模型效度可達81.6%，第二階段推估模型，模型效度亦可達相關係數93%，預測效果可達強相關，因此利用資料採礦技術應用於全日照量預測以達預測太陽能模組發電量是未來可持續研究之方向。

另外發現「雲」亦是影響全日照量其中一項重要參數，未來可將天空雲量、雲密度、雲的移動方向等資訊納入模型之中，以提高模型預測之穩定度及準確度，亦可針對預判全日照量劇烈異動之個案加以研究，再根據此原因來找出是否可以進行變數擴充。

本研究案除了運用統計以及機器學習的技術之外，由於近日Google AlphaGo機器人與韓國棋王李世石對弈，首次在號稱是人類不可被攻破的堡壘「圍棋」上，人類首次吃了敗仗。也因此我們這邊也在思考是否有可能夠透過深度學習的技術來進行本研究案的預測。

深度學習之所以與傳統的機器學習不同，一方面它透過較為精巧的神經網路設計，實現了多層次的神經網絡，讓電腦可以有能力捕捉到較為複雜且抽象的規則。再者，相較於傳統機器學習技術需要透過人類去產生輸入變數、檢定有效變數，最後在進行建模的過程，深度學習可以僅憑藉原始數據，就能夠讓演算法自己去發掘有效特徵，並而找出有意義的模型。

經過深度學習初步雛型模型測試，透過神經網路方法建置之模型，以相同資料進行模型測試，模型效度亦可達96.6%。因此以深度學習方法進行全日照量預測，期望可提高模型預測之準確及穩定度未來將為本研究持續進行之方向。

參考文獻

[1] 古建華 顏秀珍、徐嘉連，2006，資料探勘在天氣預測之研究與應用。

[2] 資訊園，2012.10.03，資料採礦方法論之CRISP-DM。

[3] Allan Yiin asiaMiner，2014，核能研究所資料採礦課程講義。

[4] 楊琇如、裘尚立，2015，資料探勘對於日照預測之可行性分析變數檢定研究報告。

[5] 楊琇如、裘尚立、尹相志，2015，資料探勘對於日照預測之可行性分析研究報告。