RapidMiner 資料探勘於澎湖海域水溫趨勢之初步應用

呂逸林、林志遠、林金榮
水產試驗所澎湖海洋生物研究中心

前言

由於資訊、通訊與感測器技術的進步和小型化，無線感測網路 (wireless sensor network, WSN) 已被廣泛的應用在各個領域，協助進行環境資訊的蒐集，作為各類管理決策的依據。在海洋環境的應用方面，WSN 技術因具有低成本、低功耗特性，國外曾用來研究水溫變化對於藻類生長程度之評估 (Bondarenko, 2007; Sieber, 2008)，亦有科學家利用其遠距控制的功能，應用於美國蒙特利灣 (Monterey Bay)，進行珊瑚礁 (Matt, 2007) 或海域環境的長期監測 (Lu et al., 2012)。

另一方面，資料探勘也已被廣泛的應用在各類需要處理大量資料的領域，包括商業、網路、衛星影像分析、分子技術、環境管理等。例如：Patricia 等 (2011) 以 GIS 和資料探勘技術進行環境監測與風險評估。Can and Saban (2009) 以 WSN 技術結合 GIS 分析，來瞭解土壤中水氣的分布情形。海洋方面亦有愈來愈多的學者將資料探勘應用在相關的研究，例如 Gilberto 等 (2008) 對巴西沿岸湧昇流生態系統 (upwelling ecosystem) 進行研究，整合了時間與空間的環境與生態因子，以資料探勘技術探討生物多樣性 (biodiversity) 和環境因子的關係，作為海域環境診斷 (diagnostic) 與管理的工具。Su 等 (2004) 以資料探勘結合 GIS 探討黃海海域魚類族群的組成與變動和環境因子之間的關係。Chauand Mutti (2007) 利用資料探勘和多變量統計分析香港吐露港 (Toloharbour) 水文與生態資料，探討港區各測站營養鹽變動的機制。許 (2008) 運用影像分析與資料探勘技術，探討橈足類與纖毛蟲之間的攝食行為。李和郭 (2013) 利用倒傳遞類神經網路 (back propagation network, BPN) 的資料探勘技術，分析熱帶海域水表面溫度，有效的改善了海面溫度估計的準確度。資料探勘技術在海洋領域的應用，顯得多元而廣泛。

近十年來，澎湖海域受到極端氣候的影響，分別在 2008 與 2011 年出現寒害，對澎湖海域生物棲地造成極大的威脅（李等，2009；Hsieh, 2008），同時導致箱網養殖業者嚴重的損失。為了掌握異常水溫的發生，並預測水溫可能的走向，讓民間業者與政府有較充裕的時間，採取適當的因應措施減少損失，本所 2010－2012 年已分年於澎湖海域建置了三處底碇式及二組浮標式的 WSN 水質水溫觀測系統，並依系統觀測結果，發布低溫特報或電子警示訊息 (林等，2012；Lu et. al., 2012；呂等，2014)。由於測站運作會

產生大量的資訊，除了對蒐集資訊品質的控制外，如何善用資訊，並即時加以分析應用與呈現，成為大數據時代的挑戰與機會。

本文嘗試利用部分 WSN 測站所蒐集的水文資訊，結合中央氣象局澎湖氣象站的氣象資料，以資料探勘技術探討澎湖海域溫度的走向，並作為低溫預警發布的參考工具。

方法與資料

過去要處理大量不同來源的資訊，並進行統計分析與結果呈現，並不容易做到，但隨著資訊技術的進步，我們開始有能力結合統計方法、資料庫、機器學習 (machine learning)、專家系統 (expert system)、資料視覺化技術 (data visualization techniques)，以及高效的電腦處理技術等，來處理大量的資料 (Fayyad et al., 1996)。整合不同來源資料庫的資料，從中擷取或探勘出所需知識的過程稱之為資料探勘 (data mining) (Han and Kamber, 2008)。在程序上，會重複地對資料進行清除 (cleaning)、整合 (integration)、選擇 (selection)、轉換 (transformation) 等動作，再利用分類 (classification)、推估 (estimation)、預測 (prediction)、關聯分組 (affinity grouping)、同質分組 (clustering) 等資料探勘技術，得到不同的有趣樣式 (interesting patterns)，並進行評估 (pattern evaluation) 找到最適結果，最後將結果表示 (knowledge presention) 出來，所進行的流程如圖 1 所示。

RapidMiner 原名 Yale，是用於資料探勘、機器學習、商業預測分析的以 Java 語言設計與 XML 文件描述的一種開源軟體，其程序包括：數據加載和轉換、數據預處理和可視化、建模、評估和部署，可與 R 語言進行協同工作。RapidMiner 中的功能均是通過連接各類運算元 (operataor) 來形成流程 (process)。不同運算元有不同的特性，共包含五大類：流程控制類 (循環和條件功能)、資料輸出入類 (資料交換功能)、資料轉換類 (資料抽取、清洗整理功能)、建模類 (分類回歸建模、關聯分析、聚類分析、集成學習等功能)、評估類 (多重交叉檢驗、自助法檢驗等功能) (Xccds, 2011)。

本文使用 2012 年底至 2013 年底兩組

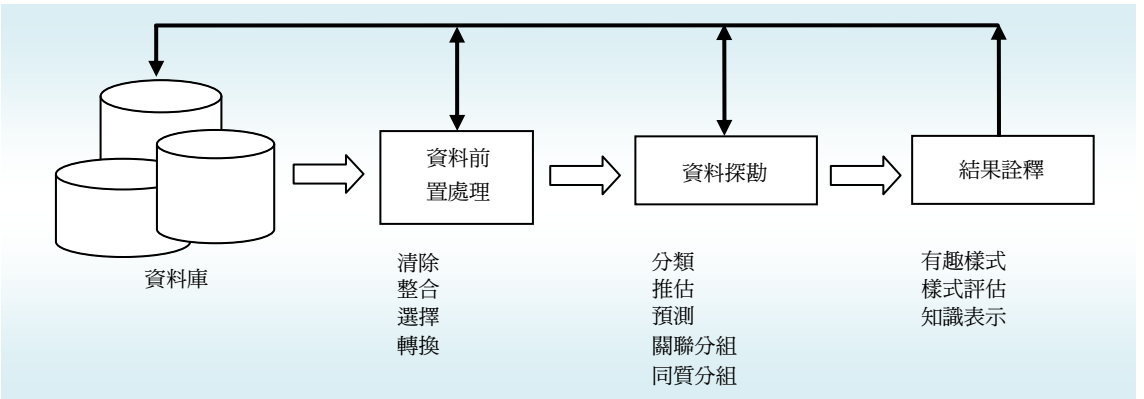

圖 1　資料探勘程序與方法

WSN 測站 30 分鐘間隔之觀測水溫，與馬公氣象站東吉測站同時段之氣溫、風速資料，利用上述 RapidMiner 軟體，設定各類運算元及其相關參數來進行澎湖地區海、氣象模式的初步探勘運算。

結果與討論

本文希望藉由資料探勘的技術針對水溫的走向進行預估，根據之前本研究利用主成分分析 (principal component analysis, PCA) 結果，以及陳永明等 (2008) 在 2008 年針對澎湖寒害的分析討論，採用了平均風速與氣溫兩個氣候參數。此外，以數值預測常用的線性迴歸模型 (linear regression) 作為預測模型，利用統計方式決定本文 2 個變因 (平均風速與氣溫) 對獨立變數 (水溫) 的影響力，以進行預測模式之推估，而不同的訓練資料集對於迴歸所得到的模型可能會造成不同的結果。經 RapidMiner 軟體先進行運算元、參數及流程等設定後，以可視化方式來表現資訊處理的過程。

首先利用擷取 (retrieve) 運算元由 excel 檔中，取得東吉嶼 WSN 測站的水溫資料，再針對水溫資料中因儀器通訊問題而遺失資料，以取代 (replace) 運算元，利用前後的水溫平均值取代之，再以整合 (join) 運算元將東吉氣象站的逐時氣溫與風速和水溫資料整併後，得到東吉嶼的海氣象資料集，再利用這個資料集進行線性迴歸模型推導，得到預測模型後，利用運用模型 (apply model) 運算元，以另一組實測的海氣象資料集，進行模型預測結果測試，最後以表現 (performance) 運算元對預測模式進行評估。

過程中，採用兩組訓練資料，分別是 2012 年 12 月 25 日至 2013 年 2 月 10 日，這段時間是澎湖冬季的海氣象資料集，以及 2013 年 1 月 1 日至 2013 年 12 月 31 日為全年海氣象資料集，進行迴歸分析，利用兩組資料是希望能比較全年的訓練資料與冬季訓練資料何者得到的模型較佳。結果得到水溫預測的模型，分別為：

模型一：

prediction W_temp =0.396 × A_temp + 0.071 × Ave. WS + 13.942

模型二：

prediction W_temp =0.563 × A_temp + 0.041 × Ave. WS + 11.339

prediction W_temp：預測水溫；A_temp：大氣溫度 (℃)；Ave. WS：平均風速 (m/s)。

最後本研究再以 2014 年 1 月 1 日至 3 月 31 日期間，東吉測站每小時平均氣溫和風速的資料集，分別投入模型一和模型二進行水溫預測，並將預測水溫與東吉 WSN 測站得到的實際水溫 (W_temp) 比較，進行表現評估，評估結果如表所示，我們發現不論由均方根差、絕對誤差或相對誤差，均可看出利用冬季資料集 (模型一) 所得到的預測結果，比整年的資料集 (模型二) 得到的預測結果佳。

不同模型預測結果比較

項　　目	模 型 一	模 型 二
均方根差	0.842 ± 0.000	1.310 ± 0.000
絕對誤差	0.640 ± 0.547℃	0.997 ± 0.850℃
相對誤差	2.99% ± 2.64%	4.62% ± 3.97%

再將實際水溫分別和模型一與模型二的預測水溫做成時間序列圖來看 (圖 2、3)。由圖中，我們可以發現模型一與模型二大致能反應水溫變動的趨勢。但利用冬季水溫資料集得到模型一 (圖 2) 的表現，明顯優於利用全年度訓練資料集的模型二 (圖 3)。兩個模型除了溫度變動的幅度不同外，兩個模型對於水溫預測的走勢相當一致；另外兩個模型在低水溫回溫時，均出現較明顯的偏差，而模型二又比模型一嚴重，這應該與參數大氣溫度在模型中的貢獻有關，因氣溫快速回升後，快速拉抬了預測水溫值所致。

由上述的討論，我們發現以迴歸分析所得到的預測模式，對海水溫度變動的趨勢提供基本預測能力，但準確度仍有改進的空間。而不同的訓練資料集，對模型的預測能力有明顯的影響，這或許與影響海氣象之間能量交換的主要因素可能會隨季節而改變，因此未來可以考量依季節不同進行模式的探討。另外，水溫的變動不若氣溫快速，尤其當氣溫由低溫回溫時，水溫的速度一般比較緩慢，因此在模型中是否要再增加參數，值得進一步的討論。

資料探勘是一個愈來愈受重視的工具，尤其在邁向生產 4.0 的路上，能夠掌握大量資訊中隱含的知識，對於決策與管理先機的取得至為重要。在全球環境變遷的影響下，異常的氣候所帶來災害的降低，有賴於對環境動態的掌握，本研究希望最終能得到較佳的預測效果，以降低澎湖養殖產業在冬天時可能面臨的環境風險，增加他們的經濟收益。

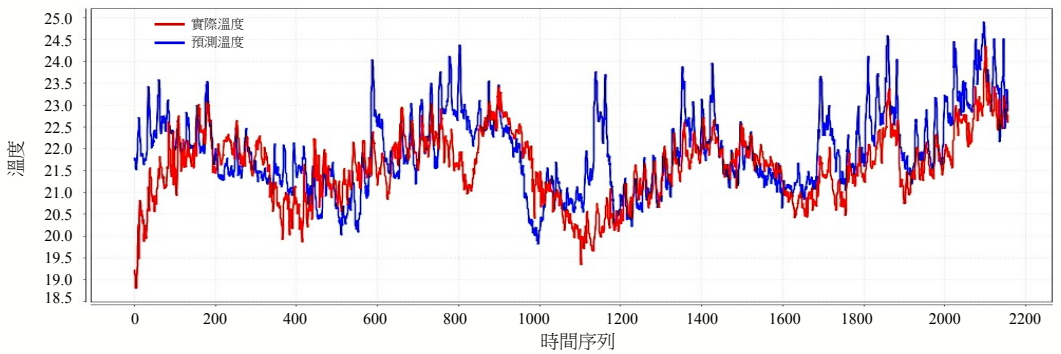

圖 2　澎湖海水溫度變動預測模型一測試結果

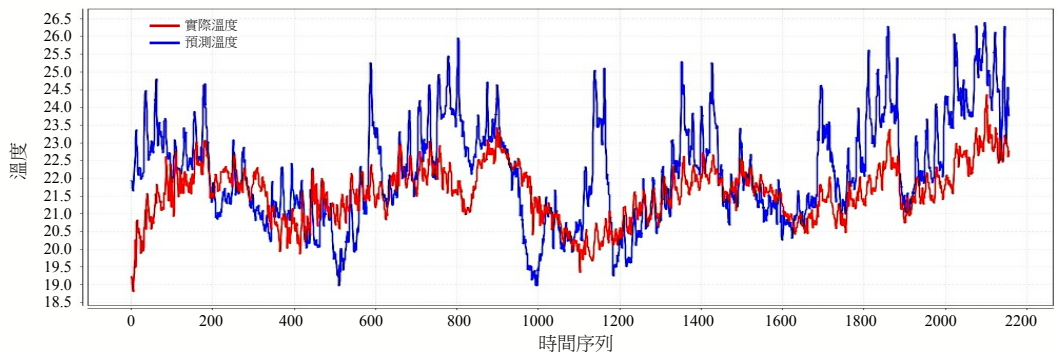

圖 3　澎湖海水溫度變動預測模型二測試結果