應用機器學習於颱風風浪分類之研究

許振嘉
國立臺灣海洋大學大專生
e-mail：gary920310@gmail.com

魏志強
國立臺灣海洋大學副教授
e-mail：ccwei77@gmail.com

摘要

臺灣為海島型國家，北迴歸線經過且是熱帶與副熱帶氣候的交界處，在此緯度亦為颱風之路徑，在臺灣較常受到颱風侵襲之季節為夏、秋兩季。颱風帶來豐沛的雨量，為臺灣帶來不可或缺的水資源，但暴風雨所帶來的災害，則嚴重影響人們生命財產的安全。又因每次颱風路徑不逕相同，且從不同方向登陸，受地形阻擋的影響，使得風向、風速或海面風浪也有不同的變化，在颱風侵襲期間，若能準確地預測浪高，則可以降低沿岸災害及保障人身安全，達到未雨綢繆的效果。本研究利用邏輯斯回歸分類模型及機器學習之隨機森林法建立颱風警報時期的波浪預測，本研究以臺灣東北部宜蘭、彭佳嶼測站和龜山島浮標為測試例。本研究收集中央氣象局所屬的氣象站及浮標歷史觀測資料與颱風警報資料，資料年限為西元 2005 至 2012 年，選定海象、氣象與颱風相關的參數結合三種演算法建立整體學習。

關鍵詞：颱風、波浪、隨機森林、邏輯斯回歸、預測

1. 前言

颱風從太平洋海面上逐漸靠近臺灣時，強烈的暴風吹襲海面，因為風速大加上風吹時間很長，會產生週期較大且波高較高的浪，加上海底地形影響，靠岸時的浪可能更高，對於沿岸附近生態活動的危害，低窪地區可能遭受海水倒灌，附近的漁船也可能被能量極強的浪破壞毀損，損失慘重。

東北海岸曾因颱風波浪影響的災害有碼頭損壞、海堤倒塌、風景區景觀生態破壞、海水倒灌等。例如 2015 年 9 月底的杜鵑颱風，28 日 17 點 40 分從宜蘭南澳登陸，但在 15 點時蘇澳港地區因海水倒灌，道路淹水可達 70 公分，甚至 1 公尺高，蘇澳區漁會理事長表示，此次災害是 50 年來最嚴重的海水倒灌。

隨機森林演算法在過去十多年來在機器學習運用上非常受到青睞，分類效能高、擴展性高和容易使用為其特色，另外它還擔任整體學習中重要的方法，結合數棵決策樹演算法以達到較個別分類器更好的預測效能。

2. 資料來源

圖 1 氣象站(紅點)與浮標位置示意圖

本研究蒐集 2005 年至 2010 年中央氣

象局發布之海上及陸上颱風警報之颱風警報單以及龜山島浮標氣象觀測資料和波高資料實測值及中央氣象局所屬地面氣象站觀測資料(龜山島、彭佳嶼兩站)。上述所蒐集之資料皆以 1 小時為觀測資料區間。本研究共蒐集 34 場颱風事件，龜山島測站 1713 筆小時資料。

3. 演算法及研究工具

3.1 隨機森林(Random Forest)

隨機森林被歸屬於機器學習的一種演算法，是一種基於決策樹改良的演算法，於 2001 年由 Breiman 和 Cutler 學者借鑒貝爾實驗室的 Ho 所提出的隨機決策森(Ho,1995,1998)的方法，在決策樹基礎上開發成隨機森林演算法，隨機森林是一種整體學習方式，透過集成學習的思想將多棵決策樹集成的一種算法，它的基本單元是決策樹，而它的本質屬於機器學習的一大分支—集成學(Ensemble Learning)方法，從直觀角度來解釋，每顆決策樹都是一個分類器，對於一個輸入樣本，N 棵樹會有 N 個分類結果，而隨機森林集成了所有分類投票結果，將投票次數最多的類別指定為最終的輸出，這就是一種最簡單的 Bagging 思想。而隨機森林與其他演算法相較之下具有較好的準確率，能夠有效的運行在大數據集上，且能夠處理高維特徵的輸入樣本而不需要降維[1,2]。

1) 吉尼指數(Gini Coefficient)

在了解隨機森林前，要先知道隨機森林的基礎，分類回歸樹是決策樹的一種，以樹狀的結構，從上而下由根節點(SI)，許多分岔的路徑(a1,a2…a10)與眾多的子節點(A1,A2….A4)組成，並在最後產生葉節點(B1,B2…B6)，將要分類的樣本由根節點輸入，而後在每個節點選用一個特徵變量對節點內的樣本進行分割，經由分叉路徑到達下一個子節點，每個分岔路徑對應一個特徵變量，而最後到達葉節點代表代表著每個分類，而分割時所選擇的特徵變量是採用吉尼指數(Gini Index)的計算。

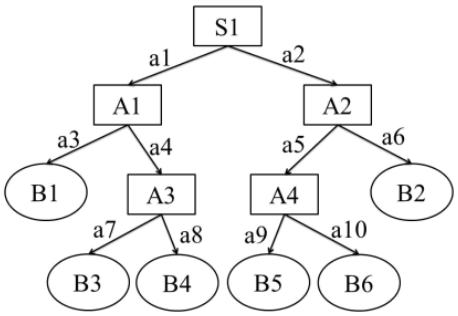

圖 2 決策樹原理

吉尼指數是從吉尼係數(Gini Coefficient)演算而來，假設集合 S 包含了 N 個類別的樣本，則吉尼指數為

$$\text{gini(S)} = \sum_{J=1}^{N} P_j\left(1-P_j\right) = \sum_{J=1}^{N}(P_j - P_j^{\,2}) = \sum_{J=1}^{N} P_j - \sum_{J=1}^{N} P_j^{\,2} = 1 - \sum_{J=1}^{N} P_j^{\,2} \quad (1)$$

吉尼指數所算得即是一個集合中的資料純度，因此在輸入樣本資料時，須給定以知之類別項目，即輸入訓練樣本之資料，依此樣本所提供之類別項目，計算分割之吉尼指數，找尋最佳之特徵變量直，吉尼指數的值越大，該集合純度越不高，反之指數越小純度越高。

2) 純度

集合 S 中，所有資料都是同一個類別，帶入吉尼指數公式，即此類別 j 在集合 S 中出現的機率P_j會等於 100%，而該集合 S 只有一個類別，因此算式為

$$\text{gini(S)} = 1\text{-}1^2=0 \quad (2)$$

若是集合 S 中有兩個類別，且所占比

例分別為 80%與 20%，將此帶入公式後，則為

$$gini(S) = 1-(0.8^2+0.2^2)=0.32 \quad (3)$$

另外，S 中同樣有 2 個類別，但所占比例為 50%，也就是個一半，則

$$gini(S) = 1-(0.5^2+0.5^2)=0.5 \quad (4)$$

由此可知，當集合 S 中都為同一個類別，即純度最大時，吉尼指數為 0，而在多個類別，當一個類別所占比例越高，吉尼指數會越小，因此分類回歸樹枝分割方式，是以吉尼指數最小值來求最佳分割解。上式街只是計算一個集合，而分類回歸樹每次皆將一個集合分成 2 個集合(N1,N2)，則吉尼指數為

$$gini_{split}(S) = \frac{N1}{N2} gini(S1) + \frac{N1}{N2} gini(S2) \quad (5)$$

3) 隨機森林

隨機森林即是以上述理論為基礎，將一顆顆的決策樹，組合成決策樹森林，隨機森林可解釋若干特徵變量(X1,X2…..XK)對因變量 Y 的作用。如果因變量 Y 有 n 個樣本資料，有 K 個特徵變量與之相關，在建構決策樹的時候，隨機森林採用重複抽樣的方法，會隨機的在原訓練樣本資料中選擇 n 個樣本，其中有的樣本會被抽樣多次，簡單說，隨機森林依據下列步驟：

I. 用 N 來代表訓練樣本數目，M 代表變量的數量。
II. 必須設定一個數字 m，要用來決定在一個分裂節點上進行分裂時，需使用到多少個變量，而且 m 應當小於 M。
III. 在 N 個訓練樣本資料裡利用重複取樣方式，取樣 N 次後，形成一組訓練樣本，而且會使用各棵樹未被取樣之樣本資料作為測試樣本，將測試樣本放回決策樹進行分類，來評估誤差。
IV. 在每一個分類節點上，隨機選擇 m 個基於此節點上之變量，而後依據這 m 個變量，透過吉尼指數運算，找尋最佳分裂方式。
V. 每顆決策樹都任其完整生長而不進行剪枝。

隨機森林理論依據是大數定律，森林由 k 棵決策樹$\{h(X,\theta_K); k = 1,2 \ldots k\}$之基本分類器，利用集成學習後所成為的組合分類器，由於每棵決策樹的建構過程中，隨機的用重複抽樣來抽取訓練樣本集合和隨機選擇特徵變量的過程都是獨立運行的，因此θ_K是獨立同分布的隨機變量。之所以引入隨機變量θ_K是為了控制每棵樹的生長，通常對第 k 棵決策樹引進隨機變量θ_K，利用訓練樣本集合和θ_K來生成第 k 棵決策樹，也就等於生成一個分類器$h(X, \theta_K)$，其中 X 是一個輸入的樣本向量。隨機森林的第 K 棵決策樹建構的過程如圖 8 所示。

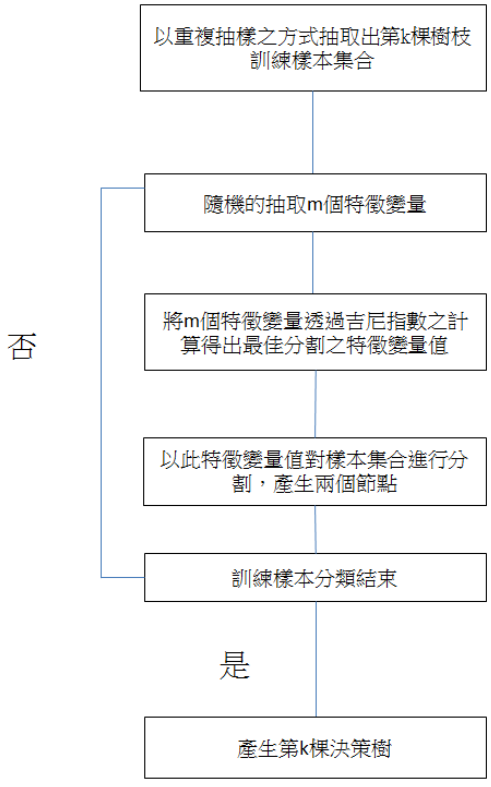

圖 3 隨機森林第 K 棵樹生成過程

以上圖方式重複產生 K 棵決策樹，最後將這 K 棵樹組合起來，便可以產生一個隨機森林，而後以此訓練完成之隨機森林，將未分類之樣本資料輸入時，由這 K 棵樹對此未分類樣本進行分類，而分類之結果即是將這 k 棵樹對此未分類樣本之分類結果，每棵樹指頭一票給它認為最適合分類的類別，最後進行簡單的多數決方式，選擇投票最多的分類類別作為該位分類樣本之分類結果。

3.2 邏輯斯回歸分析 (Logistic Regression)

在統計學上，回歸分析是一種很常見的分析方法，主要是用來了解兩個或多個變數間的相關程度，並建立模型來預測未知的樣品。當應變數(Dependent Variable)為連續型變數時，通常會使用線性迴歸(Linear Regression)來進行分析；若應變數為類別變數時(特別是兩分類的變數)，則會使用邏輯斯迴歸來做分析[3]。

1) 假設函數(Hypothesis)

首先在我們處理回歸問題無論是線性或邏 輯回歸，我們要先定義假設函數：

$$\text{Hypothesis}: h_\theta(x) = \theta^T X \quad (6)$$

上述是線性回歸形式的 Hypothesis，但是使用線性方程式處理分類問題會因為連續型變數數據擬合時會因為數據的分布而產生數據擬合上的誤差，所以線性方程式用來處理分類問題是有缺缺陷的，故我們需定義出邏輯回歸的假設函數。

我們引進一個函數 g，令邏輯回歸的 Hypothesis 表示為

$$h_\theta(x) = g(\theta^T X) \quad (7)$$

$$Z = \theta^T X \quad (8)$$

$$g(z) = \frac{1}{1+e^{-z}}, \ Z \epsilon R \quad (9)$$

上式為邏輯回歸的假設函數，Z 值為實數，g(z) function 為 Sigmoid function 或 Logistic function。圖 4 為 Sigmoid function，此函數具有下列特性[4]：

I. 當 z 值趨近於無窮大，g(z)值漸近線漸近於 1；
II. 當 z 值趨近於負無窮大，g(z)值漸近線漸近於 0；
III. g(z)的值範圍皆在 0～1 之間，所以我們也得到 $0 \leq h_\theta \leq 1$。

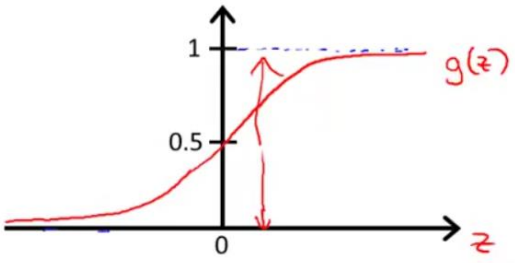

圖 4 Sigmoid function

邏輯回歸的假設函數定義為：

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T X}} \quad (10)$$

假設函數的輸出代表的是:$h_\theta(x)$ = 給定輸入 x，參數化的θ，y=1 or 0 的機率。數學式可表示如下

$$h_\theta(x) = P(y=1|x; \theta) \quad (11)$$

$$P(y=0|x;\theta) + P(y=1|x;\theta) = 1 \quad (12)$$

$$P(y=0|x;\theta) = 1 - P(y=1|x;\theta) \quad (13)$$

Sigmoid function 圖形中（即 g(z)），假設給定閾值（表界線範圍）是 0.5，當 $h_\theta(x) \geq 0.5$ 時, y = 1；$h_\theta(x) \leq 0.5$ 時, y = 0；$g(z) \geq 0.5$ 時，$z \geq 0$；對於 $h_\theta(x) = g(\theta^T X) \geq 0.5$ 則 $\theta^T X \geq 0$ 代表預測 y = 1；反之，$\theta^T X < 0$ 代表預測 y = 0。此時可以視 $\theta^T X = 0$ 為一決策邊界，當大於或小於 0 時，邏輯回歸分別預測不同分類結果。

2) 代價函數(Cost function)

根據線性回歸對於 Cost function 的定義為：

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2 \quad (14)$$

上式可將 $\frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2$ 簡寫為 Cost $(h_\theta(x^{(i)}, y))$，更進一步簡化為：Cost $(h_\theta(x)$, y) = $\frac{1}{2}(h_\theta(x) - y)^2$。

上述中，若方程式和線性回歸相似，取$h_\theta(x) = \frac{1}{1+e^{-\theta^T X}}$，會存在一個問題，也就是邏輯回歸的 Cost function 是非凸性的（圖 5）。

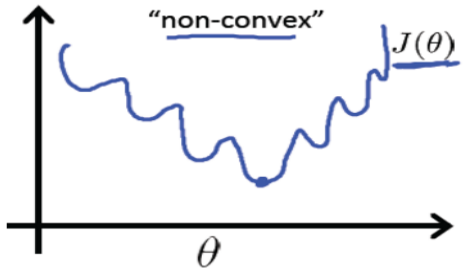

圖 5 非凸性 Cost function

對於凸函數來說局部最小值即為全局最小值的點（圖 6）。

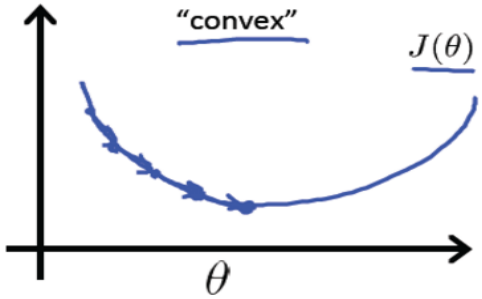

圖 6 凸性 Cost function

我們需要找出其他形式的 Cost function 來保證邏輯回歸的代價函數是凸函數。Cost function 定義方式是使用對數損失函數(logarithmic loss function)損失函數越小，模型就越好。所以使用對數損失函數的邏輯回歸 Cost Function 為

$$\text{Cost}\,(h_\theta(x),\text{y}) = \begin{cases} -\log(h_\theta(x)) \; if \; y = 1 \\ -\log(1 - h_\theta(x)) \; if \; y = 0 \end{cases} \quad (15)$$

假設函數輸出只介於 0~1 之間，則圖形可表示如圖 7 和圖 8。

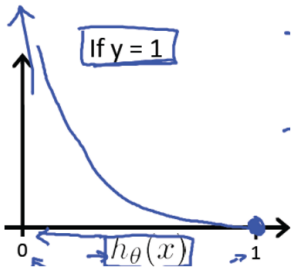

圖 7 漸減函數輸出介於 0~1 之間

直觀的來看，如果 y=1，hθ(x)＝ 1 ，則 Cost = 0，也就是當預測和實際的值完全相等時代價函數為零。但是若邏輯回歸預測出的 hθ(x) 越向 0 靠近，Cost→∞，代表需付出的代價函數越大。同理對於 y=0 的情況也適用。

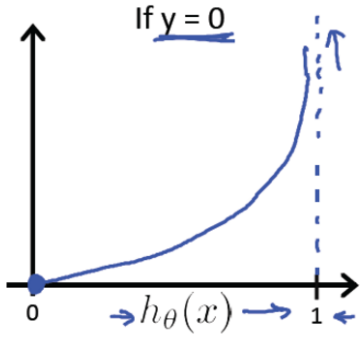

圖 8 漸增函數輸出介於 0~1 之間

上述邏輯回歸公式表示如下：

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x^{(i)}, y^{(i)}) \quad (16)$$

$$\text{Cost}(h_\theta(x),\text{y}) = \begin{cases} -\log(h_\theta(x)) \; if \; y = 1 \\ -\log(1 - h_\theta(x)) \; if \; y = 0 \end{cases} \quad (17)$$

以上公式可以整理為

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x^{(i)}, y^{(i)}) \quad (18)$$

$$J(\theta)=-\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)})\log(1 - h_\theta(x^{(i)}))] \quad (19)$$

3) 梯度下降法

為了求出最適合的參數θ來擬合數據，需要最小化 Cost Function 為: $\min_\theta J(\theta)$。一般可採用梯度下降算法來學習參數 θ，目標是最小化 J(θ)，數學式表示如下：

Want $\min_\theta J(\theta)$:

Repeat {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

(simultaneously update all θ_j)

} (20)

對 J(θ)求導後，算法如下：

Want $\min_\theta J(\theta)$:

Repeat {

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update all θ_j)

} (21)

4. 模式建立與結果

4.1 屬性資料選擇

本研究之屬性選擇使用中央氣象局發布之颱風警報單內颱風的中心氣壓(hPa)與中心最大風速(km/hr)，經度緯度，七級暴風半徑(km)，氣象資料使用中央氣象局所屬 1 個地面氣象站與龜山島浮標測站(彭佳嶼、龜山島)資料的最大平均風風速(m/s)、海面氣溫(℃)，海上氣壓(hpa)，共計 8 個屬性，波浪高度(m)實測值則為龜山島浮標資料。

表 1 2005 年至 2012 年發布警報之颱風

年份	名稱	年份	名稱
2005	Haitang Matsa	2008	Hagupit Jangmi
	Sanvu Talim Khanun Damrey Longwang	2009	Linfa Molave Morakot Parma
2006	Chanchu Ewiniar Billis Kaemi Bopha Saomai Shanshan	2010	Namtheun Lionrock Meranti Fanapi Megi
2007	Pabuk Wutip Sepat Wipha Krosa Mitag	2011	Aere Songda Meari Muifa Nanmadol
2008	Kalmaegi Fung-Wong Nuri Sinlaku	2012	Talim Doksuri Saola Haikui Kai-Tak Tembin Jelawat

4.2 目標資料區間化

本研究使用隨機森林處裡分類資料，而資料前收集整理之波浪目標資料皆為數值模式，故使用 EXCEL 內建函式，將龜山島浮標波浪高度(m)區間化轉為類別分類，浪高大於 250 (m)為 L 級大浪；浪高介於 150~250 (m)之間為 M 級中浪；浪高低於 150 (m)為 S 級小浪。

4.3 資料前處理

本研究使用 Microsoft Azure 雲端平台中的 Machine Learning 工具來完成建模

（如圖 10 至圖 12），選取 Clean Missing Data 模塊將所有有遺失的特徵資料進行整列刪除的清理動作，接著使用 Split Data 模塊將所有特徵資料和目標資料的 75%也就是 1284 筆資料取出來作為訓練資料，而剩下的 25%作為測試資料。

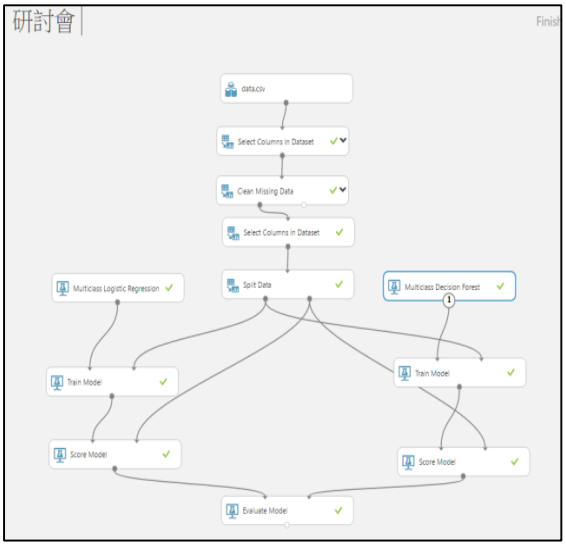

圖 10 Azure Machine Learning 圖示

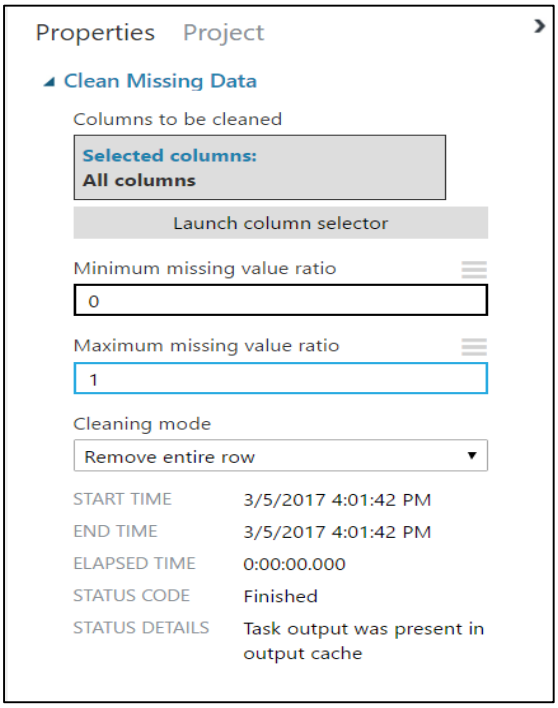

圖 11 Azure Machine Learning 執行畫面

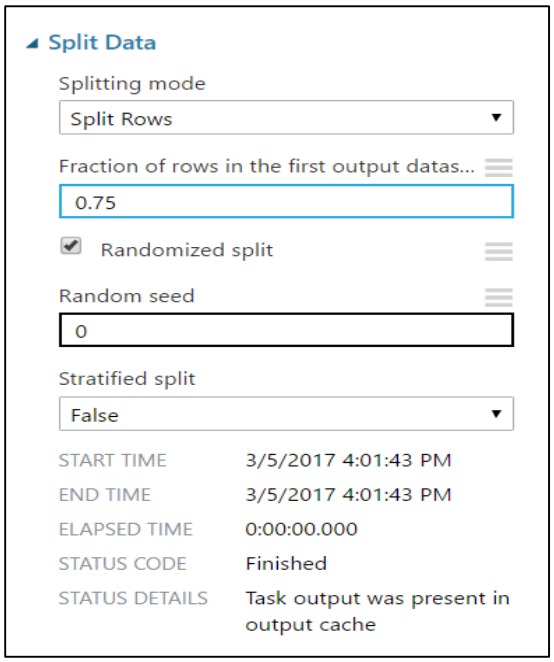

圖 12 Azure Machine Learning 參數設定

4.4 隨機森林(Random Forest)參數檢定

隨機森林是一種整體學習方式，透過集成學習的思想將多棵決策樹集成的一種法，它的基本單元是決策樹，而它的本質屬於機器學習的一大分支—集成學習(Ensemble Learning)方法，從直觀角度來解釋，每顆決策樹都是一個分類器，對於一個輸入樣本，N 棵樹會有 N 個分類結果，而隨機森林集成了所有分類投票結果，將投票次數最多的類別指定為最終的輸出，這就是一種最簡單的 Bagging 思想。而隨機森林與其他演算法相較之下具有較好的準確率，能夠有效的運行在大數據集上，且能夠處理高維特徵的輸入樣本而不需要降維。

本研究使用 Azure Machine Learning 工具中的內建多類隨機森林套件(Multiclass Decision Forest)，來對於輸出目標做出分類決策，在參數的調整上，本研究調整了決策樹的使用棵樹，以 10 顆為單位取到 100 顆分類決策樹已比較出較準

確的決策樹數量，而決策樹的最大深度參數設定為 10，因為決策樹的深度過深會造成過度擬合的情況，故選擇 10 顆以避免過度擬合，表 2 為不同決策樹數量平均準確值數據表。圖 13 為不同決策樹數量準確度比較圖。

表 2 不同決策樹數量的平均準確值

決策樹(個)	10	20	30	40	50	60	70	80	90	100
Average accuracy	0.892523	0.900312	0.894081	0.9081	0.9081	0.911215	0.901869	0.901869	0.906542	0.909657

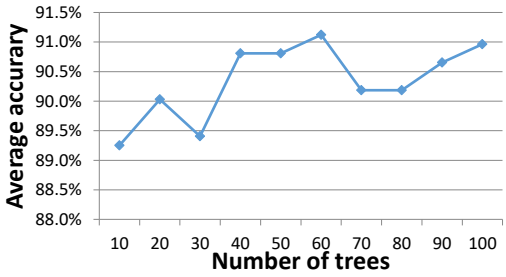

圖 13 不同決策樹數量準確度比較

由圖中觀察出決策樹數量取 60 顆時有較高的準確率，故使用 60 顆決策樹為參數下執行的隨機森林算法為結果，圖 14 則為三個浪級下之列聯表(Contingency table)。

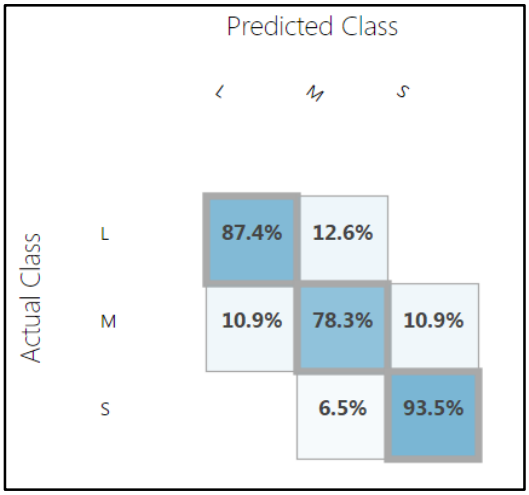

圖 14 隨機森林浪級準確度列聯表

4.5 邏輯斯回歸參數檢定和權重

本研究使用 Azure Machine Learning 工具中的內建邏輯斯回歸進行模式建立與分析)，來對於輸出目標做出分類工作。圖 15 為 Azure Machine Learning 參數設定，圖 16 為邏輯斯回歸各浪及分類之評估指數，圖 17 則為邏輯斯回歸浪級準確度列聯表。

Multiclass Logistic Regression Classifier

Settings

Setting	Value
Optimization Tolerance	1E-07
L1 Weight	1
L2 Weight	1
Memory Size	20
Quiet	True
Use Threads	True
Allow Unknown Levels	True
Random Number Seed	

圖 15 Azure Machine Learning 參數設定

Feature Weights

Feature	L	M	S
BMWS	5.53868	0	-4.73613
Bias	3.01739	0.939815	-3.95727
PRE	-3.14551	0	3.43831
East longitude	-1.52065	-0.329157	2.84046
TEMP	-1.31397	0	1.2839
L7 Radius	1.08587	-0.000162146	-0.343869
MIDPRE	-0.570668	0	0.960629
PredSpeed	-0.941265	0	0.721709
north latitude	0	-0.0786834	0.162821
BS	0	0.150877	0

圖 16 邏輯斯回歸分類結果

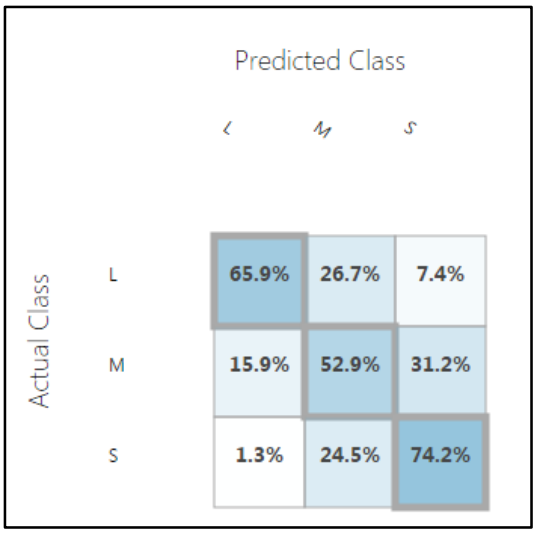

圖 17 邏輯斯回歸浪級準確度列聯表

4.6 模式分析結果

本研究由上述分析結果顯示：(1) 在浪高大於 250 公尺時（即 L 級大浪），隨機森林法準確率為 87.4%，邏輯斯回歸為 65.9%；(2) 在浪高介於 150 至 250 公尺時（即 M 級中浪），隨機森林法準確率為 78.3%，邏輯斯回歸為 52.99%；以及(3) 在浪高低於 150 公尺時（即 S 級小浪），隨機森林法準確率為 93.5%，邏輯斯回歸為 74.2%。綜合上述分析結果可知，隨機森林相較於邏輯回歸在分類問題上有著較高的準確率。

5. 結論與建議

本文利用機器學習法的隨機森林和邏輯斯回歸初步地分析颱風風浪分類之問題。本研究未來期望能將測站規模擴增到整個北海岸(基隆，彭佳嶼，宜蘭等)，藉由判斷侵台颱風之屬性來分系決策台灣北海岸可能發生之波浪災害，以期望能達到對於北海岸波浪災害有預防的效果。

參考文獻

[1] Ghosh, A., and Joshi, P. K., “A Comparison of Selected Classification Algorithms for Mapping Bamboo Patches in Lower Gangetic Plains Using Very High Resolution WorldView 2 Imagery,” ***International Journal of Applied Earth Observation and Geoinformation***, 26, 298 -311, 2014.

[2] Anita, P., and Poel, D. V. D., “Random Forests for Multiclass Classification: Random MultiNomial Logit,” ***Journal of Expert System with Applications***, 34, 1721 -1732, 2008.

[3] Arminger, G., Enache, D., and Bonne, T., “Analyzing Credit Risk Data: A Comparison of Logistic Discrimination Classification Tree Analysis and Feedforward Networks,” ***Computational Statistics***, 12, 293-310, 1997.

[4] Ashburn, P., ***The Vegetative Index Number and Crop Identification***, The LACIE Symposium, 1979.