使用繼承式基因演算法預測空氣品質

江承憲
國立虎尾科技大學自動化工程研究所 研究生
ten7728@hotmail.com

吳家豪
國立虎尾科技大學自動化工程研究所 研究生
ab112004ab@yahoo.com.tw

張原誠
國立虎尾科技大學自動化工程研究所 研究生
seaderly@gmail.com

何信璋
國立虎尾科技大學自動化工程研究所 教授
sjho@nfu.edu.tw

摘要

就歷年監測資料，台灣空氣污染指標以懸浮微粒為主。懸浮微粒是直徑小於10微米的可吸入顆粒物，對身體健康產生威脅。本研究蒐集影響空氣品質最有相關性的四種因素(氣溫、風速、風向、相對溼度)。把資料庫中有異常值去掉並擷取有價值的因素資料後，所用資料及合計有109訓練樣本及78測試樣本。再用繼承式基因演算法(IBCGA)及支持向量機(SVM)來預測空氣品質，從中選出影響懸浮微粒最重要的特徵，以建立懸浮微粒的模型，然後可預測空氣品質。經實驗結果分析，溫度和相對溼度為影響空氣品質最重要的兩個因素，而預測準確率可達74.5%。由本文提出的方法可降低空氣品質監測站所耗費的時間和人力等資源，目的使空品站的效率提高，形成一個完整的防護網路。

關鍵字：繼承式基因演算法、空氣品質監測站、支持向量機、預測、數學建模

Abstract

According to the statistics from sensing data for numerous years, the concentration of aerosols is the major index of Taiwan's air pollution. The aerosols refer to suspended particles smaller than 10 microns in diameter which are respirable particulate matter, a threat to health. In this study, we collected four most relevant factors (temperature, wind speed, wind direction, and relative humidity) affecting the air quality. The used dataset consists of 109 and 78 samples for training and test respectively, obtained by removing outliers and extracting informative factors in database. The study applied an inheritable genetic algorithm (IBCGA) and support vector machine (SVM) to establish a mathematic model of suspended particles for predicting air quality and analyzing the most important factors. The prediction accuracy of independent test was 74.5%. Using factor analysis, the most important two factors affecting air quality were temperature and relative humidity. The proposed method can help the air quality monitoring stations reduce time cost, manpower, and other resources to advance efficiency of the air quality monitoring stations in Taiwan, and establish a complete air protection network.

Keywords: Genetic algorithm, air quality monitoring station, support vector machine, prediction, mathematic modeling.

1. 前言

懸浮微粒是指懸浮在空氣中的固體顆粒，是空氣汙染的主要來源。直徑小於10微米的可吸入顆粒物，即被認定有害於人體。顆粒物能夠在大氣中停留很長時間，並可隨時呼吸進入體內，積聚在氣管或肺中，影響身體健康。由於顆粒物對健康影響巨大，各國政府均設置標準，歐盟標準為日平均值50 $\mu g/m^3$，台灣標準為日平均值125 $\mu g/m^3$。

尤其在近年來受到大陸沙塵暴的影響，造成台灣地區在冬末春初空氣品質的急劇惡化，已為行政院環境保護署所重視。。

2. 基因演算法

IBCGA 為一種尋找最佳化參數的工具，可以在大量的參數中挑選較少的參數並得到最大的適應性。IBCGA 主要由 IGA 與繼承機制構成。此智慧型基因演算法使用分割解決問題的策略及直交表，針對基因的互換步驟達到最佳化的問題[4]。

2.1 SVM 分類器

SVM (Support vector machine)分類器為一

種處理分類問題的學習模型（如圖一所示）。SVM在處理分類問題時，會尋找一個超平面(hyperplane)來將不同類的資料作區分，並且此超平面和資料群會有最大的距離空間。而為了使不同類型的資料能夠更有效的被分開，SVM使用了幾種不同的核心方程式(kernel function)，來讓資料解搜尋空間，從低維度提升到高維度，讓尋找出來的超平面可以更有效的分隔不同類型的資料[3]。

2.2 智慧型基因演算法(IGA)

IGA是在交配的地方加上直交表的應用，類似於使用多點交配，並且再選取出較好的染色體，來繼續跟下一代交配，來達到有較優良基因的子代。而IGA與GA的差別在於跑同樣的代數(Generation)中IGA得到的fitness函數，能夠優於GA，而對於有大量參數計算時，不管SGA跑了幾代，更可以明顯的看出結果的差異[5]。

2.2.1 基因演算法

基因演算法(Genetic Algorithm, GA)，是1975年John Holland提出以基因遺傳演算法來解決數學最佳化的問題。基因演算法主要啟發自達爾文的進化論：物競天擇、適者生存的定律；模擬生物基因(gene)有擇優(Selection)、交配(Crossover)及突變(Mutation)的能力，產生更優秀的新生代的程序；將要解的最佳化問題轉化為染色體及基因的模型，再藉著基因擇優、交配及突變這三個最重要的過程去尋求最佳解[6]。

2.2.2 直交表

如表一所示，使用直交式實驗中之直交表中的因素分析，可以有效的分析數個因素。直交式實驗是屬於部分因子實驗，有效的在因素中找出近似最配方的水準。在表二為L_4直交表，列為實驗次數，行可以放兩因素，如果要做完全因子實驗須做八次實驗，而用直交表只需四次就可以推理出近似最佳解的因素配方[2]。

3. IBCGA 流程

IBCGA流程圖如圖二所示。

3.1 流程步驟

第一步：初始化(Initiation)，經由亂數產生初始族群。

第二步：評估(Evaluation)，利用f(X)來評估每條染色體的值。

第三步：選擇 (Selection)，隨機選取兩條較優良的染色體放入交配池。

第四步：交配(Crossover)，將染色體進行交配，並加入直交表，選出較優良的基因，加入子代中。

第五步：突變(Mutation)，隨機選擇p_m×N個子代染色體來進行交換式突變(swap mutation)以產生新的子代染色體，為pm突變發生率。在突變的過程中，為了避免最好的結果消失掉，因此繼承式雙目標基因演算法會先把最好的子代染色體挑出來，使其不參與突變步驟。

第六步：結束測試(Termination test)，若停止條件獲得滿足，則輸出最佳解，否則回到第二步。

第七步：繼承(Inheritance)，當r <r_{end}時，在每條染色體中隨機選擇一個基因，使基因的值從0突變成1，以增加r的數量成r+1並回到第二步，否則停止演算法[3,4, 7]。

3.2 IBCGA 實驗方法

從文獻參考取出原始資料，並根據原始方式做出第一組資料，原始資料之訓練109筆，測試78筆，訓練資料比例 0類:1類為53:53約為1:1，測試data 0類:1類為56:22約為2.55:1；第二組修正原始訓練與測試資料集之0類與1類比例，使得訓練與測試資料集之比例一致，訓練129筆，測試55筆，訓練data 0類:1類為76:53約為1.43:1，測試data 0類:1類為33:22約為1.5:1。

此兩種實驗方法之目的在於找出一個最佳化之建模方式，以期能找出一個預測最高之系統。執行方式以兩組資料為基準，先將訓練資料正規化，其後將測試資料以訓練資料為基準進行正規化。完成後，每組資料各執行30個IBCGA的最佳化特徵選取(feature selection)，每個最佳化流程設定為每組最佳化跑60代(generation)，每組裡面分別以10-fold cross-validation方式進行機器訓練，特徵選取的目標從9個特徵自動尋找至1個特徵，每個

最佳化的特徵集合目標設定為尋找最高之預測準確度。在 30 個最佳化流程結束後，從其中挑選出一個出現頻率最高之特徵集合。最後再使用最終之特徵集合進行 SVM 預測分類之動作。

4. 結果與討論

結果顯示，兩種實驗之結果均選取 TEMP(氣溫)與 RH(相對溼度)是最佳的兩項參數，也是決定空氣品質最重要之參數。

利用 SPSS 統計軟體跑多變量 Logistic 迴歸，探討四個變量的顯著程度，所選取分析的資料集為訓練資料集，以方便與 IBCGA 訓練結果比較。

結果顯示，最顯著與最重要的變數可以從表格中看出為 RH(相對溼度) 。

以預測準確度角度來看，task 1 的測驗結果之準確率(ACC, accuracy)是 66.67%，task 2 的測驗結果之準確率(ACC, accuracy)是 74.55%。且 task 1 訓練與測驗的差距遠大於 task 2 訓練與測驗的差距，此結果代表若訓練與測驗的比例一致，則 task 2 表現較佳。

從 SEN、SPE 與 PRE 三方面來看，由於 task 1 的測驗結果之 Sensitivity (SEN)為 68.18%、Specificity (SPE)為 66.07%與 Precision (PRE)為 44.12%；task 2 的測驗結果之 Sensitivity (SEN)為 59.09%、Specificity (SPE)為 84.85%與 Precision (PRE)為 72.22。總體說來 task 2 的結果較好。

雖然 task 1 訓練資料集為 balanced 架構(意即 0 類:1 類=1:1)，但是由於 task 2 訓練資料集與測驗資料集的比例一致，因此 task 2 的結果較好。不過兩個 task 中，IBCGA 最佳組合均是 TEMP(氣溫)與 RH(相對溼度)，也代表此兩個參數之表現是很穩定的。

5. 結論

本研究使用繼承式基因演算法結合 SVM 來預測空氣品質，在根據氣溫、風速、風向、相對溼度的影響因素下來預測空氣品質及分析上述四項影響因子哪一個所佔的影響效果比較大。IBCGA 是以 IGA 加入繼承式機制，使得輸出更快得到收斂的結果，並結合適應性指標使每個染色體經過 SVM 以達到最佳化的結果。經本文的研究溫度和相對溼度為影響空氣品質最重要的因素，IBCGA 方法的預測可以使空氣品質預測之準確率達 74.5%。由本文提出的方法可使空氣品質監測站的效率提高，形成一個完整的防護網路。

參考文獻

[1] 林逸塵，類神經網路應用於空氣品質預測之研究，中山大學環境工程研究所碩士論文，2002。

[2] 梁宏維，智慧型基因演算法在粗糙度的建模與預測，虎尾科技大學機械與機電工程研究所碩士論文，2009。

[3] 許凱迪，利用V3環狀序列的物化特性預測HIV-1病毒類型，交通大學生物資訊及系統生物研究所碩士論文，2009。

[4] 許馨云，從X光影像預測肺結核纖維化等級，交通大學生物資訊及系統生物研究所碩士論文，2012。

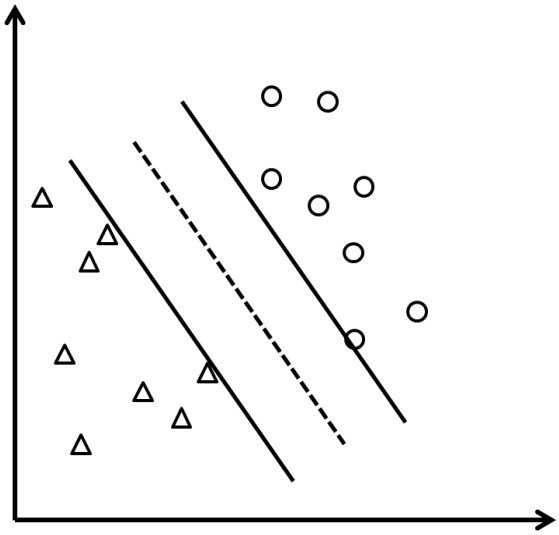
[5] 陳彥佑，應用智慧型基因演算法(IGA)於統測成績及大學生身高的建模與預測，虎尾科技大學自動化工程研究所碩士論文，2010。

[6] 黃忠雄，應用智慧型基因演算法於造血幹細胞與粗糙度的建模與預測，虎尾科技大學機械與機電工程研究所碩士論文，2009。

[7] 黃韻如，使用EMT相關基因預測肺癌轉移，交通大學生物資訊及系統生物研究所碩士論文，2012。

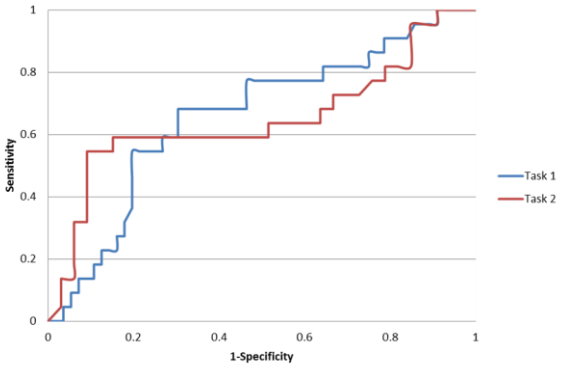
[8] 楊敦翔，以類神經網路與特徵選取技巧處理空氣能見度預測問題之研究，中山大學機械與機電工程學系碩士論文，2003。

[9] 簡家宏，應用基因類神經網路於空氣品質短期預測及監測資料異常值診斷之研究—以台中縣沙鹿測站為例，雲林科技大學環境與安全衛生工程研究所碩士論文，2004。

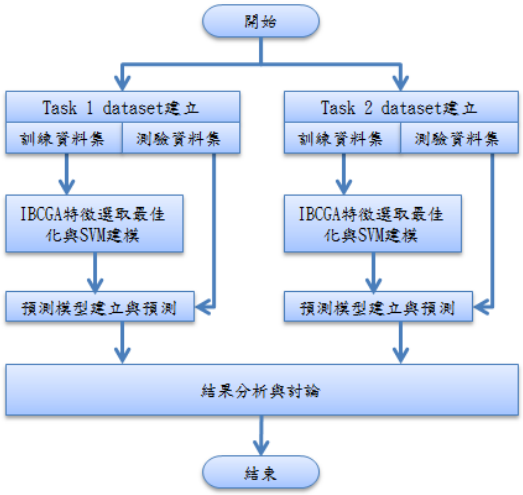


圖一：為二維SVM概念示意圖。

圖二：IBCGA 流程圖。



圖三：ROC 曲線比較。



圖四：實驗流程圖。

表一：L_4直交表。

實驗次數	X1	X2	X3
1	0	0	0
2	0	1	1
3	1	0	1
4	1	1	0

表二：第一組特徵重要性排名(knockout) 。

排名	欄位編號	欄位內容	Knockout 後之 ACC
1	1	TEMP(氣溫)	57.5472%
2	4	RH(相對溼度)	63.2075%

表三：第二組特徵重要性排名(knockout) 。

排名	欄位編號	欄位內容	Knockout 後之 ACC
1	1	TEMP(氣溫)	55.0388%
2	4	RH(相對溼度)	69.7674%

表四：數據結果。

	Task 1	Task 2
測試數據	106 (53:53)	129 (76:53)
驗證數據	78 (56:22)	55 (33:22)
靈敏度	0.6818	0.5909
特異性	0.6607	0.8485
準確率	0.6667	0.7455