Deep Learning Final Project -臺灣海洋廢棄物預測

109061517 邱俊嘉, 109061520 陳俊宇, 109061807 吳亞澤 Group 1

Introduction

Background

海洋廢棄物污染場域廣大，相較於一般空氣、水體或廢棄物污染，更不容易找出污染源與污染區域之間的直接關聯；人造固體廢棄物因外觀、尺寸、重量、材質之間具有高度差異，相對於重金屬或農藥等化學污染，目前無法用單一檢測儀器或程序測量。人造廢棄物的源頭減量才是最根本的治理辦法，但是改變源頭不易，且需較長的時間，因此做為末端補救的淨灘相當重要。

海岸廢棄物快篩調查可在短時間內做大範圍的抽樣調查，並量化廢棄物，可作為測量的方法之一，供淨灘選址參考。快篩的抽樣方式為於海岸線每隔 10 公里取一測站，以臺灣本島 1,210 公里海岸線為母體，即有 121 個測站。希望藉由測站資訊預測相近測站的資訊，以達到減少測站和人力。

Dataset

海洋廢棄物快篩數據訓練集

- 34維的特徵。
- 訓練集：319 筆資料.
- 測試集：163 筆資料
- 10 種海廢等級

level	number
1	11
2	46
3	29
4	28
5	43
6	48
7	51
8	34
9	20
10	9
total	319

Data preprocess

1. 缺失值處理

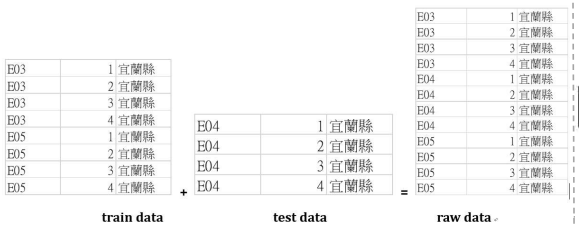
以同一個地點的其他3個season的眾數來填補缺失值

1暴露岩岸	2暴露人造	3暴露岩盤
0	0	1
0	0	1
0	0	1

2. 特徵值若為中文或英文就以正整數代號取代

3. 將是否為同一個系列測站的資訊納入考慮

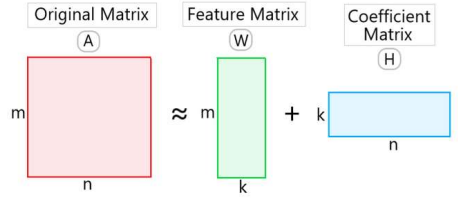
4. 將train data和test data還原成raw data



5. 處理同一個地點，但是County或Location不一樣的情況

6. 將所有feature做個別的Min-Max Normalization

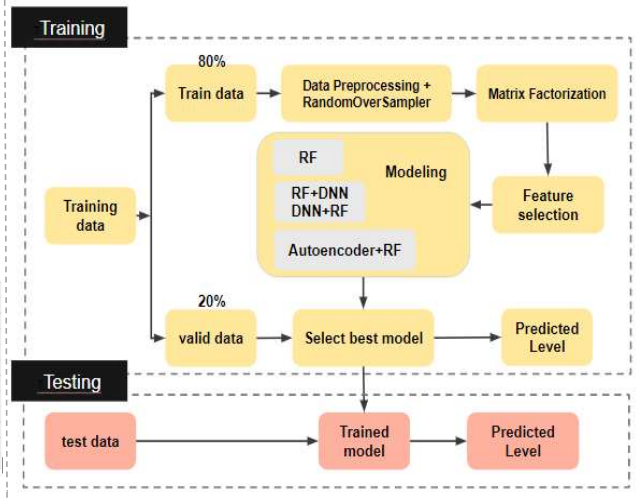
7. Non-Negative Matrix Factorization



8. 做feature selection找出關鍵feature

Index	feature	importance	normalized_importance	cumulative_importance
0	Lat	586	0.167118	0.167118
1	Lon	560.6	0.159875	0.326993
2	County	324.5	0.0925424	0.419535
3	Seat	289.2	0.0824754	0.502011
4	Season	246	0.0701554	0.572166
5	Foam material	215.6	0.0614858	0.633652
6	Fishing nets and ropes	189.8	0.054128	0.68778
7	縣市(county_2)	156.5	0.0446314	0.732411
8	Location	127.8	0.0364466	0.768858
9	Substrate type	118.5	0.0337944	0.802652
10	Plastic bottle container	109	0.0310851	0.833737
11	4沙灘(coastal landform 4)	94	0.0268074	0.860545
12	Shore shape	90.5	0.0258092	0.886354
13	5砂礫混合灘(coastal landform 5)	84.8	0.0241837	0.910538
14	2暴露人造結構物(coastal landform 2)	71.6	0.0204192	0.930957
15	Float	66	0.0188222	0.949779

Methodology



Experiment result

RandomForest				
Feature Name	Train Kappa	Train UAR	Train ACC	Test Kappa
no_remove	0.876	91.5 %	89.03 %	0.6523
remove_low_importance_and_high_correlation (RLIHC)	0.8618	90.4 %	87.77 %	0.6753
remove_low_importance	0.7235	78.94 %	75.55 %	0.7059
remove_zero_importance	0.6122	70.55 %	65.52 %	0.754
Non-Negative Matrix Factorization_RandomForest				
FA2, no_remove	0.4459	57.14 %	50.47 %	0.6612
FA2, RLIHC	0.5823	69.22 %	62.7 %	0.7789*
FA4, RLIHC	0.6626	74.96 %	69.91 %	0.7678
FA8, RLIHC	0.609	71.34 %	65.2 %	0.6505
RLIHC+upsampling				
RandomForest, n_iter=45	0.6707	75.55 %	70.85 %	0.6749
DecisionTreeClassifier, n_iter=40	0.9109	93.65 %	92.16 %	0.5305
Gradient Boosting Classifier, n_iter=100	0.603	70.38 %	64.89 %	0.5817
Non-Negative Matrix FA_RandomForest+upsample				
FA2, RLIHC	0.6673	75.59 %	70.53 %	0.6878
FA4, RLIHC	0.7089	77.62 %	74.29 %	0.6651
DNN(RELU) -> RandomForest				
FA2, no_remove	0.8727	91.95%	88.71%	0.6864
FA2, RLIHC	0.8063	87.82%	82.76%	0.7511
FA4, RLIHC	0.5923	69.79%	63.64%	0.6972
RandomForest -> DNN(GELU)				
FA2, RLIHC	0.5726	67.48 %	61.96 %	0.7643
FA2, RLIHC, weighted loss	0.5361	66.57 %	58.43 %	0.6896
Autoencoder (GELU) -> RandomForest				
FA2, RLIHC, Hidden dim=64	0.6246	73.46%	66.46%	0.5093
FA2, RLIHC, Hidden dim=128	0.4703	60.15%	52.66%	0.5527
FA2, RLIHC, Hidden dim=256	0.5746	69.05%	62.07%	0.5458
FA2, RLIHC, Hidden dim=512	0.5706	68.58%	61.76%	0.5861

Analysis

1. Feature selection (FS)效益討論：

由Table可得知，有採用feature selection的結果會比較好，其中又以remove_zero_importance的表現最為突出，相比完全沒用的**test kappa上升0.1017**

1. Matrix Factorization (FA)效益討論：

Factorization的結果則是remove low importance and high correlation (RLIHC)的平均表現較優秀，其中又以**降到兩維之後再取FS**的結果最為突出，比只做**FS再上升0.290**，**test kappa達0.7789**。這也是我們最好的結果。

1. 傳統機器學習模表現比較：

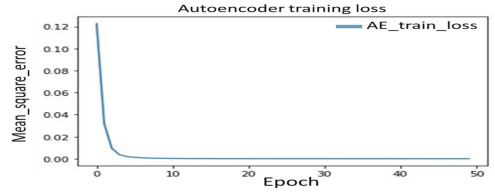
RandomForest (RF)、DecisionTreeClassifier、Gradient Boosting Classifier三者中，還是以**RandomForest的結果最好**。

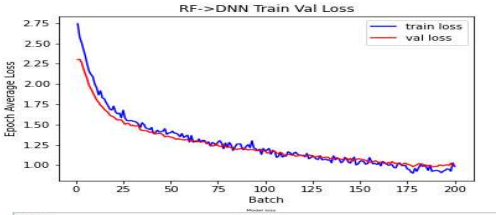
1. Upsampling 效益討論：

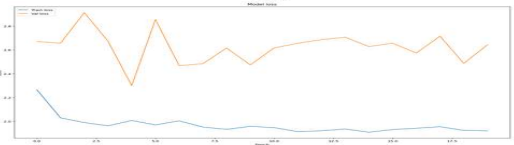
雖然資料類別數不平衡，但是採用upsampling或weighted cross entropy易造成**過擬和**，導致最終表現降低。

1. 類神經網路表現討論：

我們進行了三種NN架構對該任務進行不同方式的訓練，並把training loss於下方呈現。雖然參數皆有收斂，但表現都沒有優於原始的FA+FS+RF，由此可以推測此**資料集的數量可能過少**，且**類別過多**，若使用NN造成**模型過於複雜**，導致test kappa降低。







biic 人類行爲訊息暨互動計算研究室
Behavioral Informatics and Interaction Computation Lab

國立清華大學
NATIONAL TSING HUA UNIVERSITY