題目：台灣牧場乳量預測

組員：洪廷維、

1. Introduction

本專題的目的是實現牧場乳量預測，我們被給予四份關於乳牛的資料庫：birth.csv、breed.csv、report.csv、spec.csv，接著必須自己進行資料處理，篩選出有用的資料，並利用機器學習的方式預測台灣不同地區牧場生產的乳量，以掌握乳量生產的關鍵，使台灣酪農業受益。

1. Framework
   1. 概述

本專題使用Python及其相關套件（numpy, pandas, keras, xgboost, scikit-learn）實現乳量預測。軟體主架構如圖1所示，可將整體分成兩個部分：資料前處理、機器學習模型，首先透過資料前處理將四份資料庫轉為training data及test data，接著用training data訓練模型，最後把test data丟進訓練好的模型以產出預測結果。

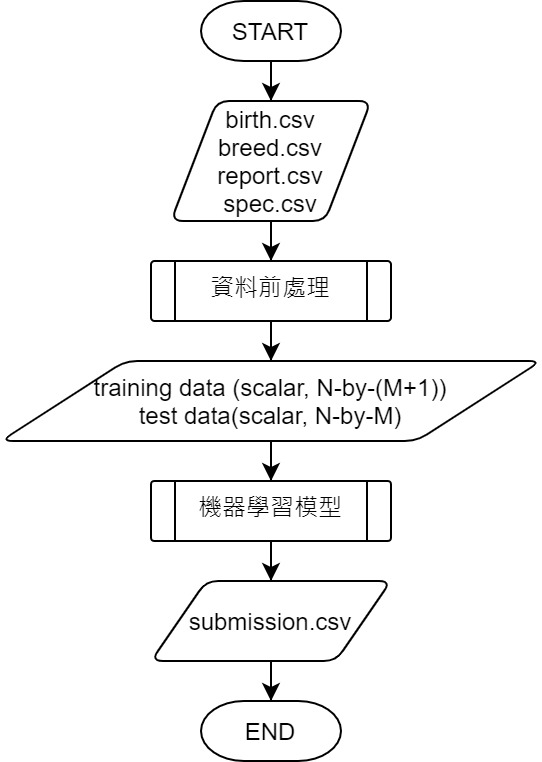


圖1

* 1. 資料前處理

資料前處理的目的是提取資料庫中有用的資料，使預測準確度提升，其基本概念可分為資料清理、資料整合、資料轉換，下文會分別介紹。資料前處理架構如圖2所示，一開始先讀檔，接著每次選一個特徵進行資料整合，直到沒有有用的特徵後進行資料轉換，最後再拆成training data、test data。

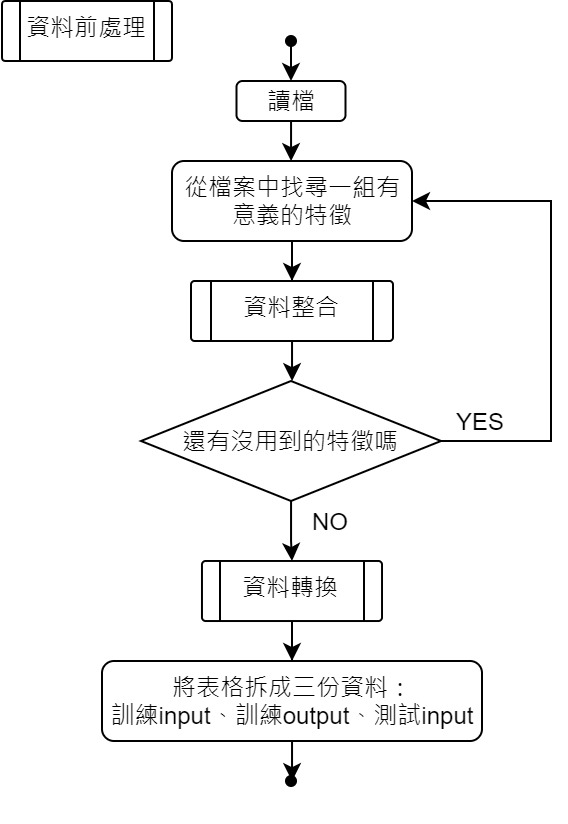


圖2

* + 1. 資料清理

資料清理是指清除與乳量無關聯的資料，或是清除、填補缺漏的資料，由於資料清理是很基本的動作，在每個步驟都可能執行，所以未在圖2中畫出。我們將資料清理分為橫向（一個特徵）與縱向（一筆資料）清理。

橫向清理是對特徵作清理，首先我們利用產業知識判斷該特徵對乳量多寡的重要性，這部分參考了農委會[1]~[2]、維基百科、各式部落格網站等等，記載了大量關於影響乳牛產乳量的知識，接著也參考相關乳量預測論文[3]~[6]中的特徵挑選（及他們的Model Inputs是什麼）。選定數個特徵後開始對每個特徵清理，我們的清理方式分為以下三種：

|  |  |  |
| --- | --- | --- |
| 種類 | 該特徵中的資料 | 清理方式 |
| 第一種 | 缺一點 | 刪除缺少該特徵的那幾筆資料 |
| 第二種 | 缺一些 | 補平均值 |
| 第三種 | 缺很多 | 刪除該特徵 |

如下圖，假設特徵2之中有缺項(NAN)，若為第一、三種情況，其對應刪除方向如圖所示，若為第二種，把NAN換成特徵2的平均值，分成三種的目的是為了在資料筆數、特徵原始程度、特徵數量之間權衡，我們希望保留一定量的資料筆數，越原始的特徵（而不是幾乎都是填補的平均值）以及盡量多的特徵數量。另外，判別種類的方式沒有進一步分析和測試，若要，可以透過輸出乳量的RMSE為指標校正。

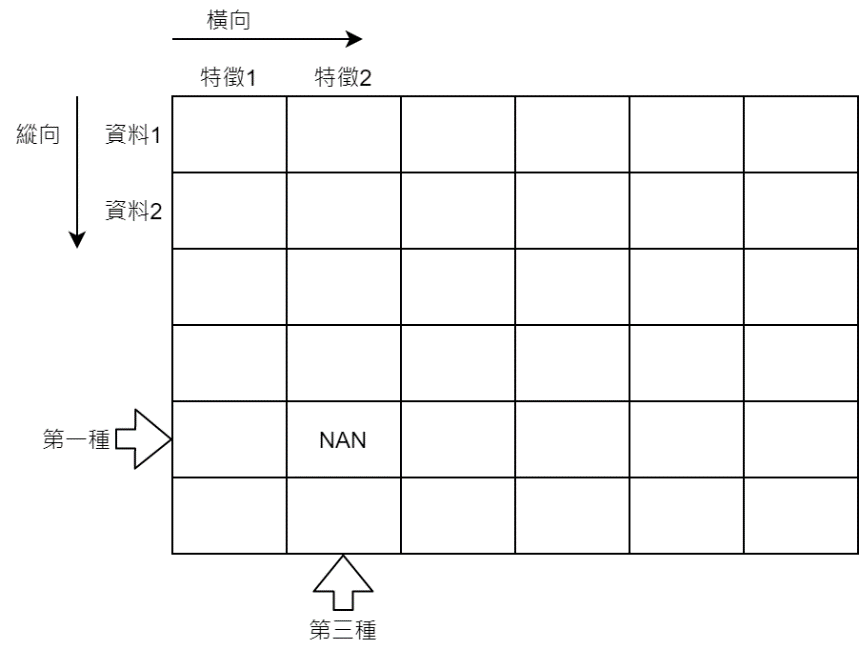


圖3

縱向清理是對資料作清理，我們透過特徵的平均值與標準差為基準判斷資料中的離群值，但我們是針對人為勘誤作刪除，所以並不是將資料視為N維向量，以範數為度量計算其偏離平均值的程度並加以刪除，因為這樣可能不小心刪除到有用的資料，而是，一次一個特徵，判斷該特徵中是否有不合理的值（應該是正數但卻是負數、超出好幾個標準差之外），並刪除此筆資料。

* + 1. 資料整合

圖2中的資料整合用於整合資料庫中的特徵及資料，包含生成新的特徵（依月份生成對應該地區的平均氣溫）、特徵間的運算（日期相減生成間隔）以及不同資料庫間的合併（不同資料庫會有不同的紀錄方式，所以需透過相關的索引為指標合併成相同格式），詳細架構如圖4。

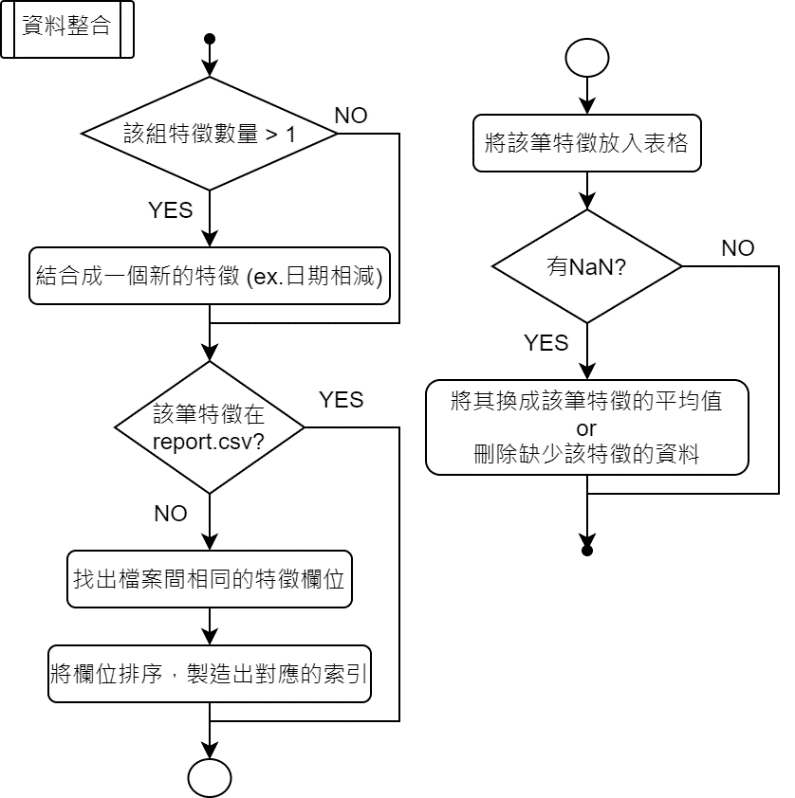


圖4

* + 1. 資料轉換

圖2中的資料轉換架構如下，進行資料轉換時，代表已經由前面的步驟產生出一個含有我們要的所有特徵的表格，此部分會對類別資料作one-hot coding轉換，非類別資料作Z-score正規化

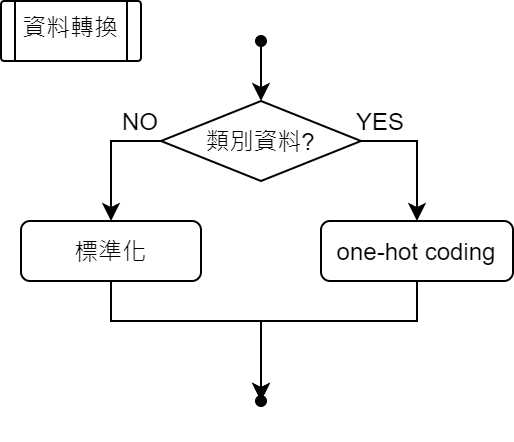


圖5

* 1. 機器學習模型

1. Results
2. Summary
3. Reference
4. <https://www.coa.gov.tw/ws.php?id=2501744>
5. <https://medium.com/jameslearningnote/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E7%AC%AC2-4%E8%AC%9B-%E8%B3%87%E6%96%99%E5%89%8D%E8%99%95%E7%90%86-missing-data-one-hot-encoding-feature-scaling-3b70a7839b4a>
6. <https://www.sciencedirect.com/science/article/pii/S0022030214002690>
7. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.626.3829&rep=rep1&type=pdf>
8. <http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S0375-15892012000300010>
9. <https://www.sciencedirect.com/science/article/pii/S0168169906000998>