資料探勘

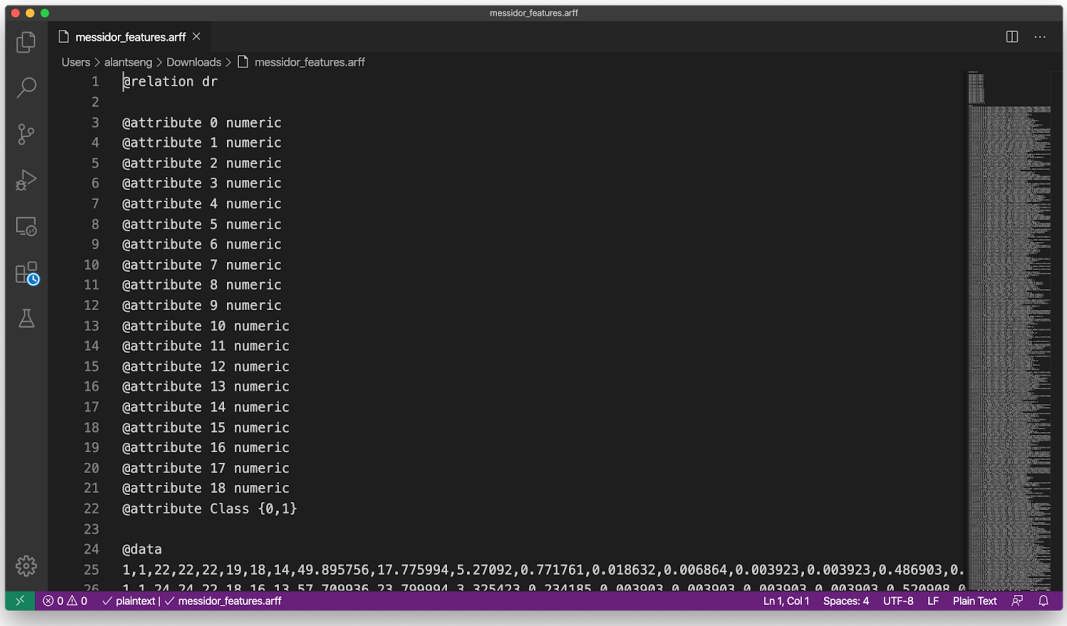
06160485

曾宏鈞

1. 資料

1-1 來源

[UCI 糖尿病性的視網膜病變資料集]
格式：.arff
打開可發現
@relation：代表有一個資料表叫做dr。
@attribute：共有20個屬性，有19個特徵及1個答案，型別為數字與{0,1}的兩個類別。
@data：資料的本身，很像CSV使用逗點隔開。



1-2 說明

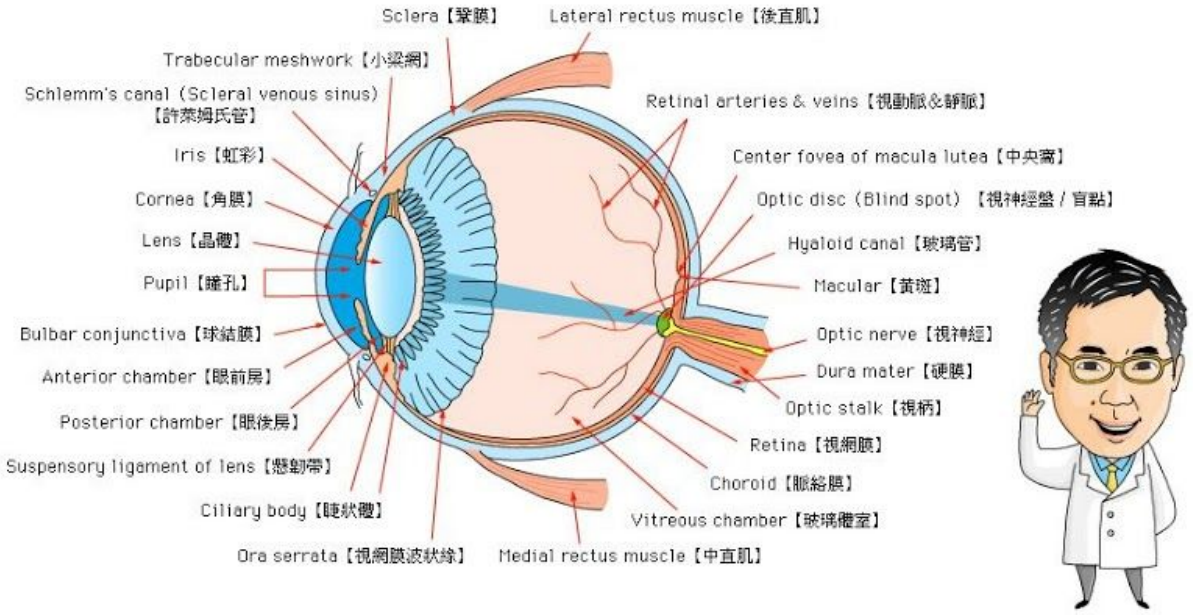
取自Messidor資料集，預測圖像是否包含糖尿病性視網膜病變的徵兆。使用在Balint Antal, Andras Hajdu：基於集合體的糖尿病性視網膜病變自動篩查系統，所有特徵都代表檢測到的病變資訊，且使用圖像分析和特徵提取以及分類。

適用問題	分類
資料筆數	1151
屬性數量	20
缺值	無

1-3 屬性介紹

特徵

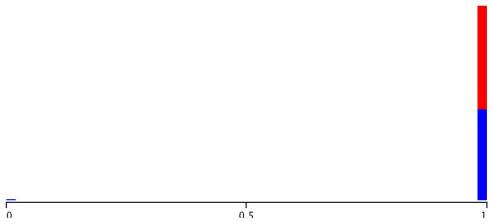
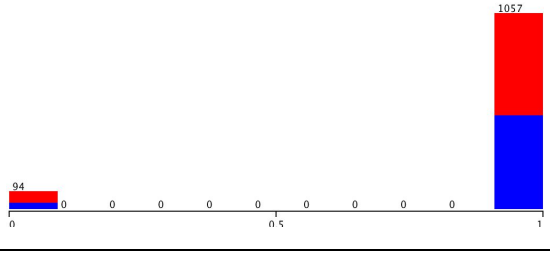
0. 檢測質量	0:差, 1:好
1. 預篩選的結果(嚴重視網膜異常)	1:有, 0:沒有
2-7. MA檢測結果，每一項特徵代表在信心水準(alpha)下0.5… 1.0	數字
8-15. 包含2-7的滲出液訊息，由病變數/ROI直徑歸一化特徵，以縮減不同大小的差距	數字
16. 黃斑部到盲點的歐式距離，也使用ROI直徑標準化	數字
17. 盲點直徑	數字
18. 基於AM/FM的分類結果	0:沒有糖尿病視網膜病變(DR), 1,2,3:代表不同的DR標籤



1-4 屬性資訊

我們將原始資料讀入weka並視覺化結果

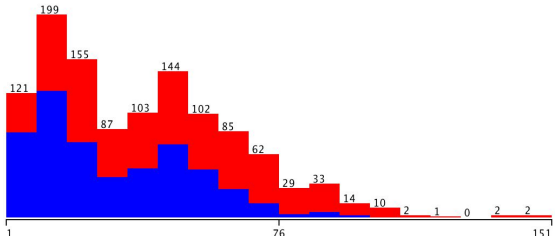
特徵

0	Name: 0 Type: Numeric Missing: 0 (0%) Distinct: 2 Unique: 0 (0%) Statistic | Value Minimum | 0 Maximum | 1 Mean | 0.997 StdDev | 0.059	
1	Name: 1 Type: Numeric Missing: 0 (0%) Distinct: 2 Unique: 0 (0%) Statistic | Value Minimum | 0 Maximum | 1 Mean | 0.918 StdDev | 0.274	

2

Name: 2 Type: Numeric
Missing: 0 (0%) Distinct: 110 Unique: 11 (1%)

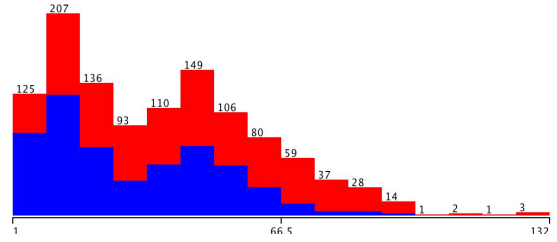
Statistic	Value
Minimum	1
Maximum	151
Mean	38.428
StdDev	25.621



3

Name: 3 Type: Numeric
Missing: 0 (0%) Distinct: 104 Unique: 8 (1%)

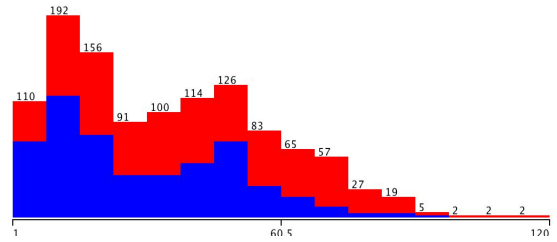
Statistic	Value
Minimum	1
Maximum	132
Mean	36.91
StdDev	24.106



4

Name: 4 Type: Numeric
Missing: 0 (0%) Distinct: 99 Unique: 13 (1%)

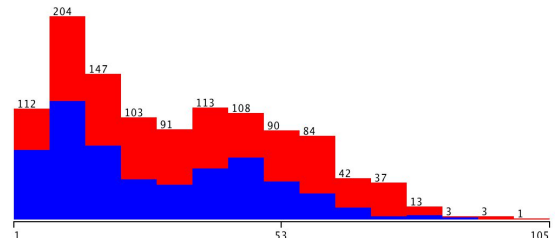
Statistic	Value
Minimum	1
Maximum	120
Mean	35.141
StdDev	22.805



5

Name: 5 Type: Numeric
Missing: 0 (0%) Distinct: 91 Unique: 9 (1%)

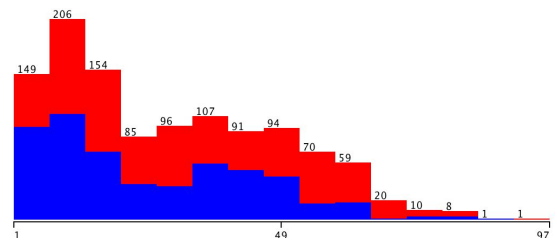
Statistic	Value
Minimum	1
Maximum	105
Mean	32.297
StdDev	21.115

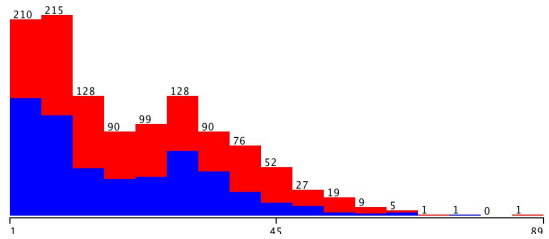
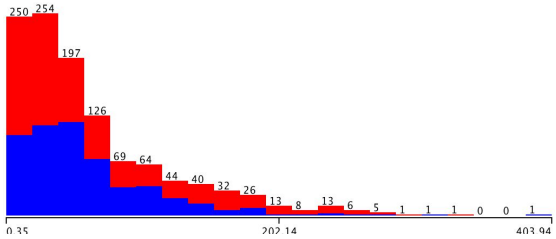
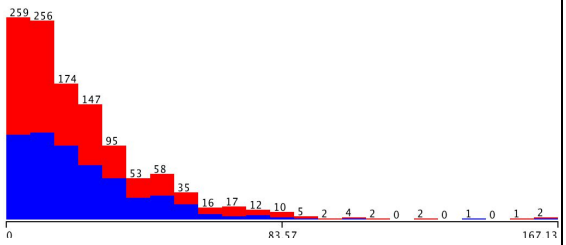
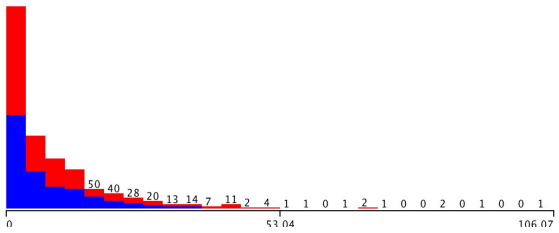
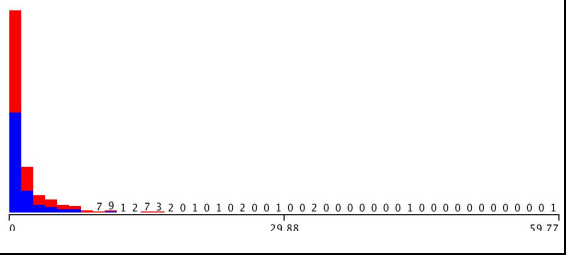


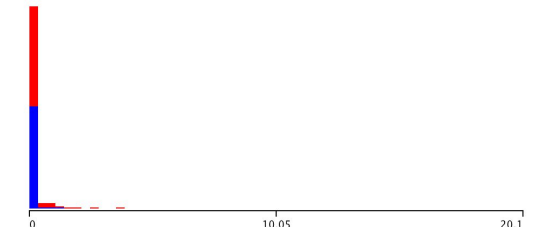
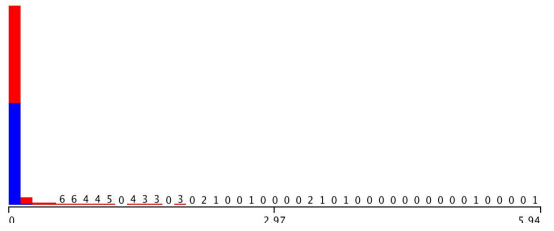
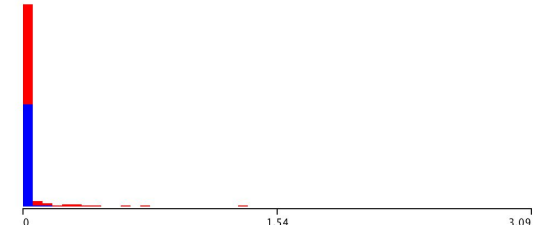
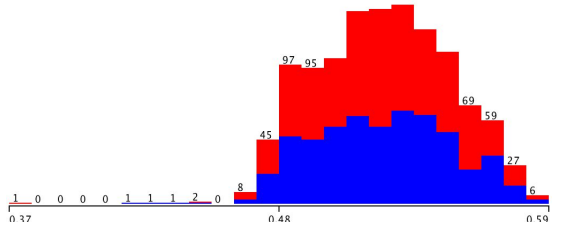
6

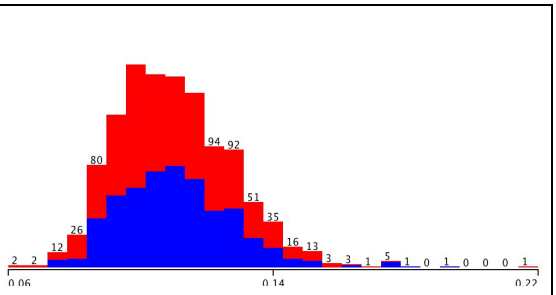
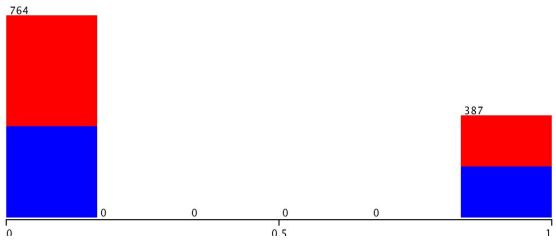
Name: 6 Type: Numeric
Missing: 0 (0%) Distinct: 84 Unique: 9 (1%)

Statistic	Value
Minimum	1
Maximum	97
Mean	28.747
StdDev	19.509

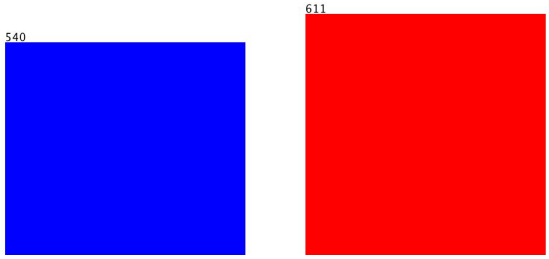


7	Name: 7 Type: Numeric Missing: 0 (0%) Distinct: 69 Unique: 8 (1%) Statistic | Value Minimum | 1 Maximum | 89 Mean | 21.151 StdDev | 15.102	
8	Name: 8 Type: Numeric Missing: 0 (0%) Distinct: 1141 Unique: 1131 (98%) Statistic | Value Minimum | 0.349 Maximum | 403.939 Mean | 64.097 StdDev | 58.485	
9	Name: 9 Type: Numeric Missing: 0 (0%) Distinct: 1141 Unique: 1131 (98%) Statistic | Value Minimum | 0 Maximum | 167.131 Mean | 23.088 StdDev | 21.603	
10	Name: 10 Type: Numeric Missing: 0 (0%) Distinct: 1130 Unique: 1119 (97%) Statistic | Value Minimum | 0 Maximum | 106.07 Mean | 8.705 StdDev | 11.568	
11	Name: 11 Type: Numeric Missing: 0 (0%) Distinct: 1032 Unique: 1022 (89%) Statistic | Value Minimum | 0 Maximum | 59.766 Mean | 1.836 StdDev | 3.923	

12	Name: 12 Missing: 0 (0%) Distinct: 795 Type: Numeric Unique: 771 (67%) Statistic | Value Minimum | 0 Maximum | 51.423 Mean | 0.561 StdDev | 2.484	
13	Name: 13 Missing: 0 (0%) Distinct: 579 Type: Numeric Unique: 547 (48%) Statistic | Value Minimum | 0 Maximum | 20.099 Mean | 0.212 StdDev | 1.057	
14	Name: 14 Missing: 0 (0%) Distinct: 415 Type: Numeric Unique: 385 (33%) Statistic | Value Minimum | 0 Maximum | 5.938 Mean | 0.086 StdDev | 0.399	
15	Name: 15 Missing: 0 (0%) Distinct: 351 Type: Numeric Unique: 325 (28%) Statistic | Value Minimum | 0 Maximum | 3.087 Mean | 0.037 StdDev | 0.179	
16	Name: 16 Missing: 0 (0%) Distinct: 1132 Type: Numeric Unique: 1113 (97%) Statistic | Value Minimum | 0.368 Maximum | 0.592 Mean | 0.523 StdDev | 0.028	

17	Name: 17 Missing: 0 (0%) Distinct: 1110 Type: Numeric Unique: 1070 (93%) Statistic | Value Minimum | 0.058 Maximum | 0.219 Mean | 0.108 StdDev | 0.018	
18	Name: 18 Missing: 0 (0%) Distinct: 2 Type: Numeric Unique: 0 (0%) Statistic | Value Minimum | 0 Maximum | 1 Mean | 0.336 StdDev | 0.473	

答案

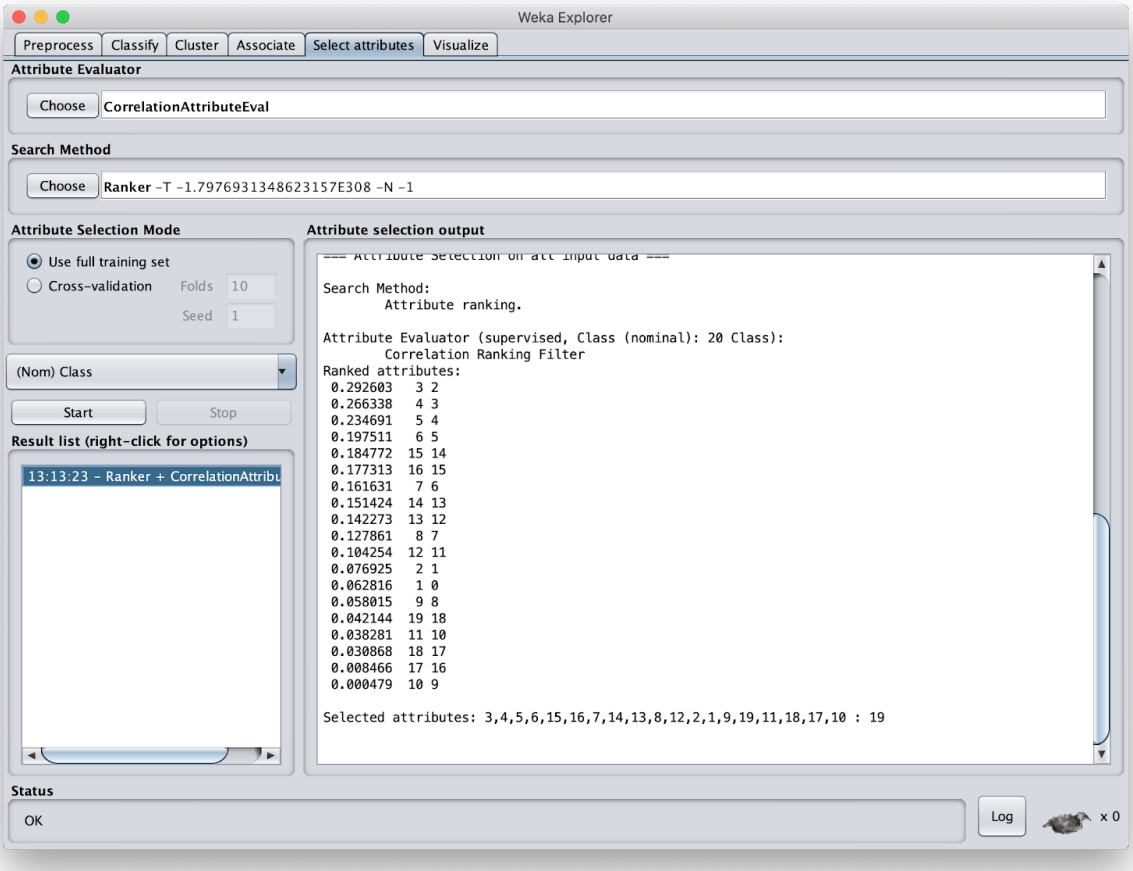
Name: Class Missing: 0 (0%) Distinct: 2 Type: Nominal Unique: 0 (0%) No. | Label | Count | Weight 1 | 0 | 540 | 540.0 2 | 1 | 611 | 611.0	

由於兩類別資料數目都差不多，因此不需針對答案去做抽樣

1-5 特徵工程

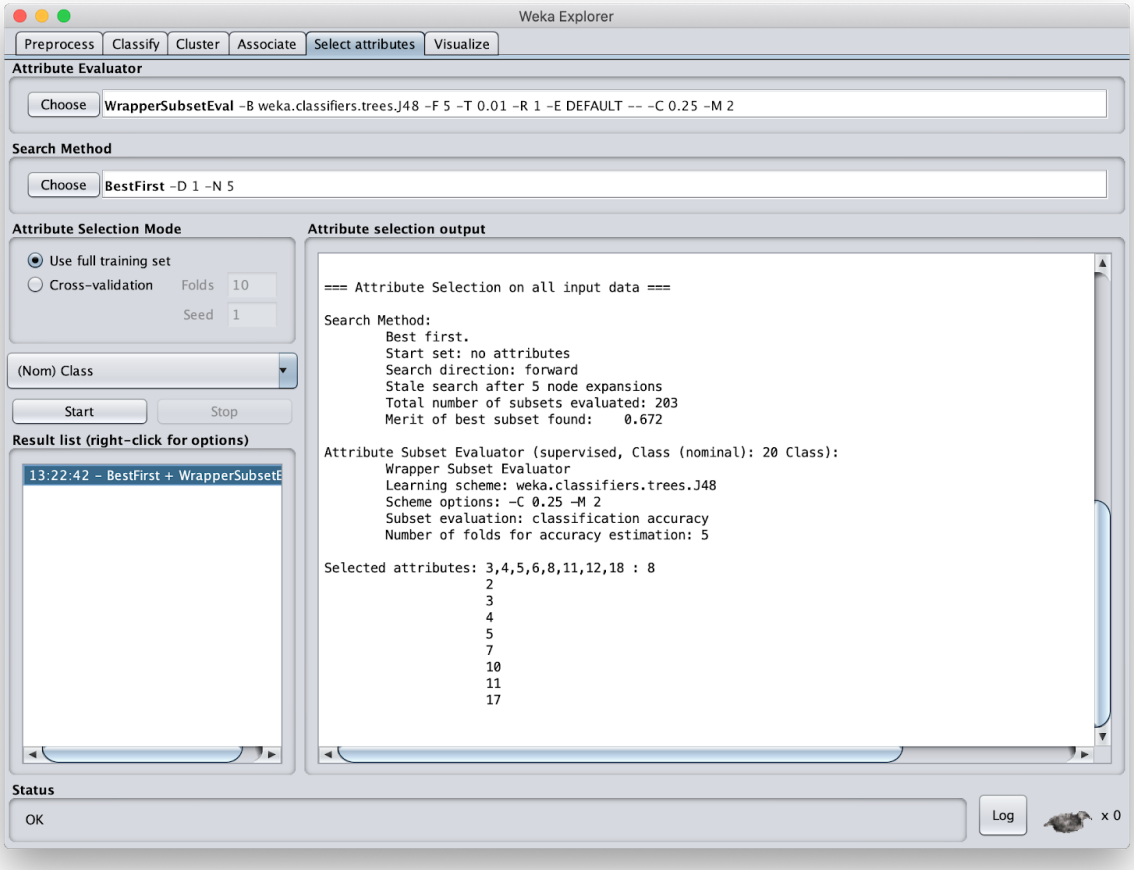
1-5-1 基於關聯性

- Weka操作步驟：
  Select attributes
  Attribute Evaluater > CorrelationAttributeEval
  Search Method > Ranker



1-5-2 基於學習

- Weka操作步驟：
  Select attributes
  Attribute Evaluater > WrapperSubsetEval > classifer > tree > J48
  Search Method > BestFirst



1-6 前處理原因

1. 觀察資料值域
   由1-4可知，特徵19個標籤最大值從0.219-403不等，因此資料前處理要做正規化的動作，避免訓練模型時，因特徵值域範圍不同造成某項特徵重要性權重的錯誤。
2. 抽樣
   我們將原始資料進行抽樣訓練，使得訓練時使用的時間及空間會較少。

3. 降低維度
   一般資料都有大量的特徵，我們使用主成分分析法(PCA)而不是單純的特徵選擇來選擇對模型比較有幫助的特徵，把資料集的複雜度變小有以下的優點：
   - 避免維度詛咒
   - 較好視覺化
   - 消除資料本身雜訊
   - 更容易解釋

2. 資料前處理

2-1 資料正規化

- Weka操作步驟：
  Preprocess>Filter
  filters>unsupervised>attribute>Normalize

<table>
<tr><th>before</th><th>after</th></tr>
<tr><td>
Name: 8 Type: Numeric
Missing: 0 (0%) Distinct: 1141 Unique: 1131 (98%)
<table>
<tr><th>Statistic</th><th>Value</th></tr>
<tr><td>Minimum</td><td>0.349</td></tr>
<tr><td>Maximum</td><td>403.939</td></tr>
<tr><td>Mean</td><td>64.097</td></tr>
<tr><td>StdDev</td><td>58.485</td></tr>
</table>
</td><td>
Name: 8 Type: Numeric
Missing: 0 (0%) Distinct: 1141 Unique: 1131 (98%)
<table>
<tr><th>Statistic</th><th>Value</th></tr>
<tr><td>Minimum</td><td>0</td></tr>
<tr><td>Maximum</td><td>1</td></tr>
<tr><td>Mean</td><td>0.158</td></tr>
<tr><td>StdDev</td><td>0.145</td></tr>
</table>
</td></tr>
<tr><td>
Name: 9 Type: Numeric
Missing: 0 (0%) Distinct: 1141 Unique: 1131 (98%)
<table>
<tr><th>Statistic</th><th>Value</th></tr>
<tr><td>Minimum</td><td>0</td></tr>
<tr><td>Maximum</td><td>167.131</td></tr>
<tr><td>Mean</td><td>23.088</td></tr>
<tr><td>StdDev</td><td>21.603</td></tr>
</table>
</td><td>
Name: 9 Type: Numeric
Missing: 0 (0%) Distinct: 1141 Unique: 1131 (98%)
<table>
<tr><th>Statistic</th><th>Value</th></tr>
<tr><td>Minimum</td><td>0</td></tr>
<tr><td>Maximum</td><td>1</td></tr>
<tr><td>Mean</td><td>0.138</td></tr>
<tr><td>StdDev</td><td>0.129</td></tr>
</table>
</td></tr>
<tr><td>
Name: 17 Type: Numeric
Missing: 0 (0%) Distinct: 1110 Unique: 1070 (93%)
<table>
<tr><th>Statistic</th><th>Value</th></tr>
<tr><td>Minimum</td><td>0.058</td></tr>
<tr><td>Maximum</td><td>0.219</td></tr>
<tr><td>Mean</td><td>0.108</td></tr>
<tr><td>StdDev</td><td>0.018</td></tr>
</table>
</td><td>
Name: 17 Type: Numeric
Missing: 0 (0%) Distinct: 1110 Unique: 1070 (93%)
<table>
<tr><th>Statistic</th><th>Value</th></tr>
<tr><td>Minimum</td><td>0</td></tr>
<tr><td>Maximum</td><td>1</td></tr>
<tr><td>Mean</td><td>0.313</td></tr>
<tr><td>StdDev</td><td>0.111</td></tr>
</table>
</td></tr>
</table>

- 說明：
  由上表可知，原始資料中特徵17的值域是 [0.058, 0.219]、特徵8的的值域是[0.349, 403.939]，標準差為0.018及58.485，平均為0.108、64.097，兩者相差640倍，有可能會因值域造成重要性的錯誤，因此我們將其正規化，使特徵的值域範圍在[0,1]之間，每一個特徵的重要性都相同。

2-2 資料抽樣

- Weka操作步驟：
  Preprocess>Filter
  filters>unsupervised>instance>resample
  sampleSizePresentage:50

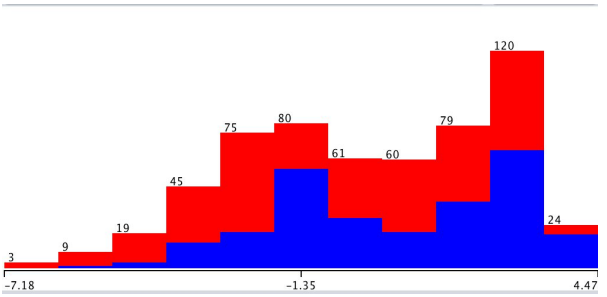
Current relation Relation: dr Instances: 1151 Attributes: 20 Sum of weights: 1151	**Current relation** Relation: dr-weka.filters.unsupervised.instance.Resa... Instances: 575 Attributes: 20 Sum of weights: 575

2-3 主成分分析法(PCA)

- Weka操作步驟：
  Preprocess>Filter
  filters>unsupervised>attribute>PrincipleComponent

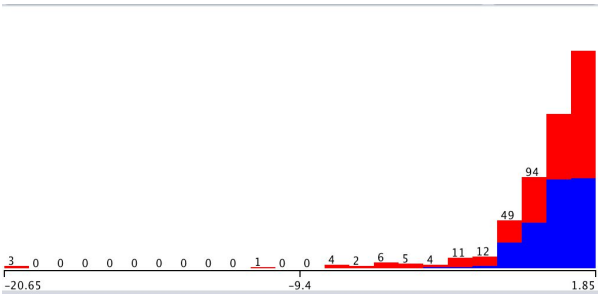
Name: -0.3944-0.3945-0.3913-0.3886-0.3882... Type: Numeric
Missing: 0 (0%) Distinct: 458 Unique: 356 (62%)

Statistic	Value
Minimum	-7.18
Maximum	4.474
Mean	-0
StdDev	2.493



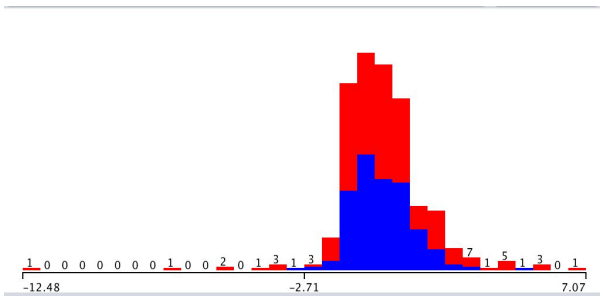
Name: -0.41111-0.3913-0.38312-0.36614-0.333... Type: Numeric
Missing: 0 (0%) Distinct: 458 Unique: 356 (6...

Statistic	Value
Minimum	-20.651
Maximum	1.846
Mean	-0
StdDev	2.212



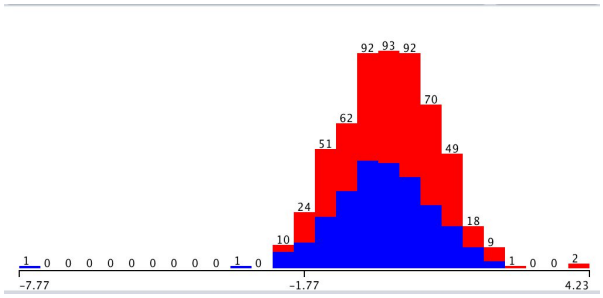
Name: 0.4559+0.42110+0.4098-0.34115-0.3241... Type: Numeric
Missing: 0 (0%) Distinct: 458 Unique: 356 (6...

Statistic	Value
Minimum	-12.477
Maximum	7.065
Mean	0
StdDev	1.463



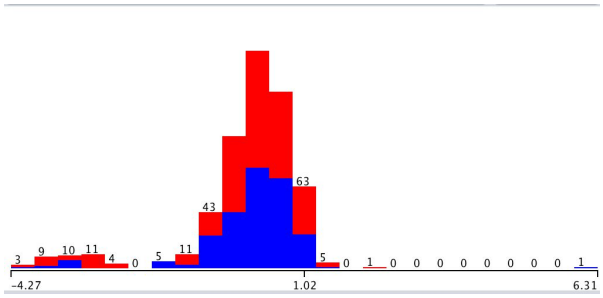
Name: 0.71816-0.54817+0.2830-0.2521+0.1161... Type: Numeric
Missing: 0 (0%) Distinct: 458 Unique: 356 (6...

Statistic	Value
Minimum	-7.765
Maximum	4.228
Mean	0
StdDev	1.07



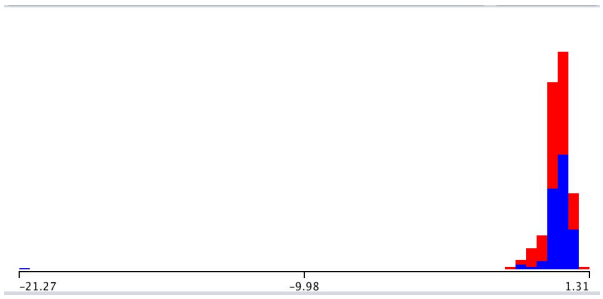
Name: 0.7931-0.4917-0.2530+0.12618-0.10211... Type: Numeric
Missing: 0 (0%) Distinct: 458 Unique: 356 (62%)

Statistic	Value
Minimum	-4.266
Maximum	6.312
Mean	0
StdDev	1.021



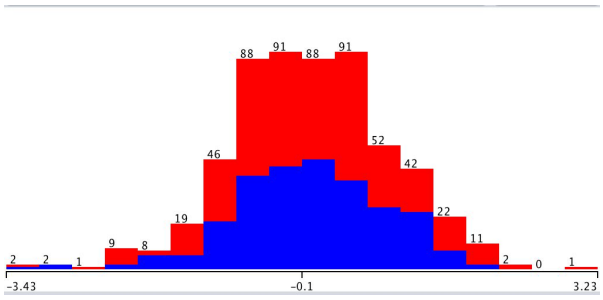
Name: 0.8930+0.3011-0.23918-0.18316+0.0671... Type: Numeric
Missing: 0 (0%) Distinct: 458 Unique: 356 (6...

Statistic	Value
Minimum	-21.268
Maximum	1.308
Mean	-0
StdDev	0.997



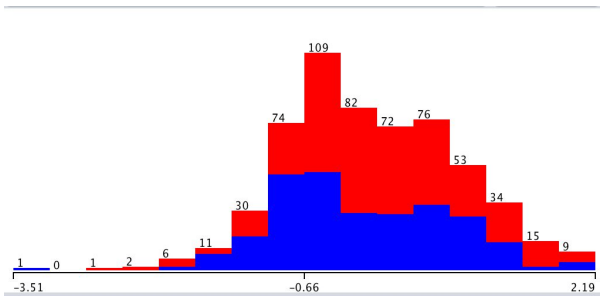
Name: -0.6417-0.54116-0.3818-0.3531-0.0790... Type: Numeric
Missing: 0 (0%) Distinct: 458 Unique: 356 (62%)

Statistic	Value
Minimum	-3.434
Maximum	3.231
Mean	-0
StdDev	0.896



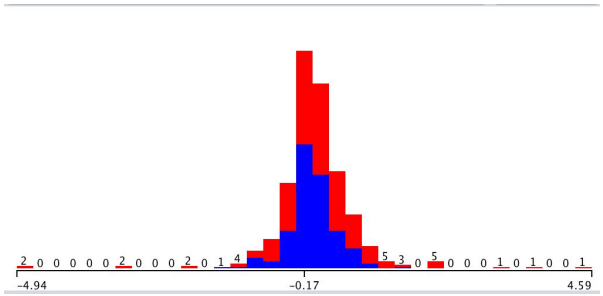
Name: 0.81918-0.35216-0.2151+0.2040-0.1661... Type: Numeric
Missing: 0 (0%) Distinct: 458 Unique: 356 (6...

Statistic	Value
Minimum	-3.509
Maximum	2.186
Mean	-0
StdDev	0.847



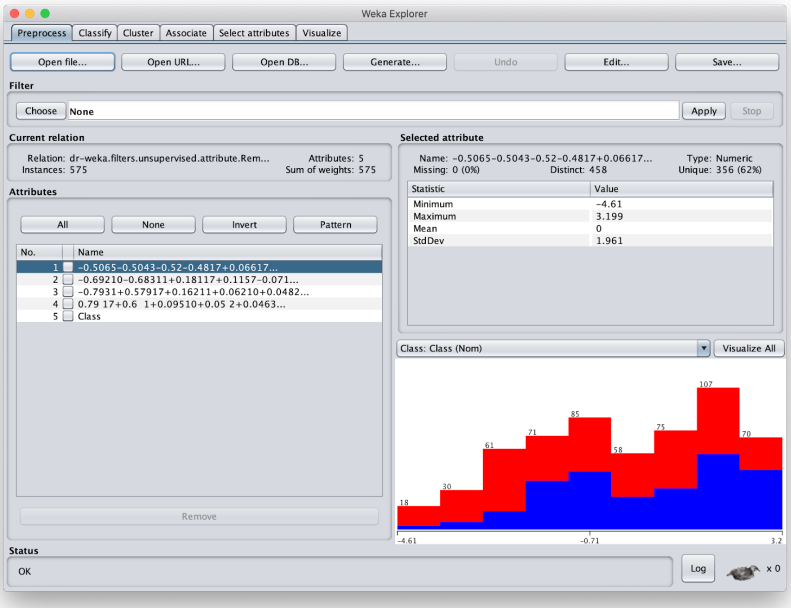
Name: -0.4912+0.47915+0.4198-0.40111+0.269... Type: Numeric
Missing: 0 (0%) Distinct: 458 Unique: 356 (6...

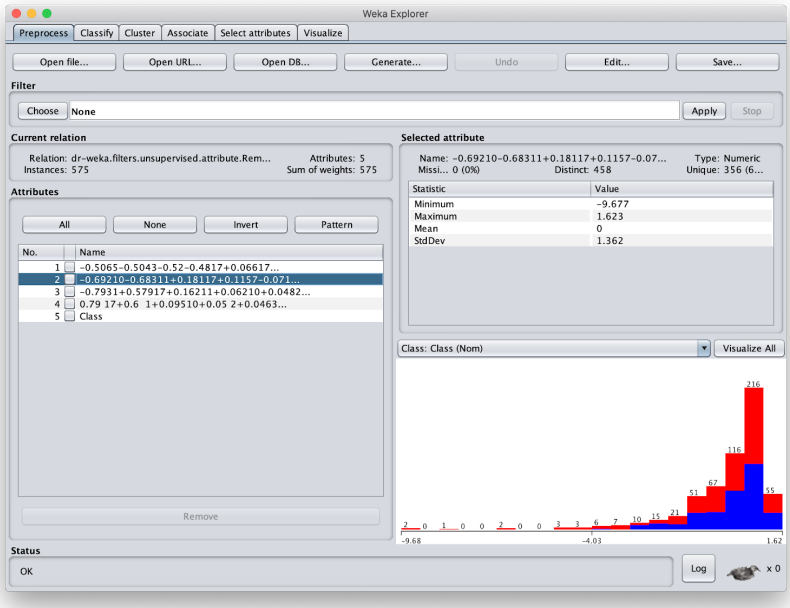
Statistic	Value
Minimum	-4.938
Maximum	4.589
Mean	0
StdDev	0.657



以上可知，我們使用PCA將19個特徵縮減到只有9個特徵(9個維度)

2-4 特徵選擇且資料處理過





3. 心得

在此次的期中作業中，我們使用了weka學習如何對資料做前處理。從此次練習中可以觀察到，weka是一個很簡單可以快速對資料做處理的資料探勘工具，對於現在正在學習資料探勘技術是非常有幫助的，除了上課學習的理論基礎，更有實作的配合。此外，以前在線性代數老師有提到的主成分分析法(PCA)對於當時的我覺得很抽象，但在做期中專案的時候就了解原來這個是多麼的重要，可以避免一些訓練上的問題，如維度詛咒等等。

綜合上述，我覺得此次期中作業練習除了了解整個做資料探勘的步驟，更對於課堂上的理論及更早線性代數的資料處理演算法有更一步的了解，透過實作能更了解每一個步驟的意義，因此此次作業對於資料探勘是非常重要且有意義的。

3-1 處理完的資料樣式