葡萄酒品質分類

報告者:錢冠邑

Outline

- 資料簡介

- 透過不同分類方法將資料分類
  1. 所有資料視為Training Data
  2. 分為Training Data與Testing Data

- 總結

資料簡介

- 資料來源: UCI Machine Learning Repository*
- 資料量:
  - 白葡萄酒4898筆

葡萄牙 “Vinho Verde” 地區的葡萄酒資料

- 資料變數皆使用儀器測量的資料

由葡萄酒的化學性質來分辨葡萄酒的好壞

- 不包含品牌、價格、葡萄類型
- 好壞以酒的品質分數評估(0~10分)

- 品質分數分類：
  - 分兩類: High(≥6)、Low(<6)
  - 分七類

*P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. (2009). Modeling wine preferences by data mining from physicochemical properties. *In Decision Support Systems, Elsevier* 47, 547-553.

Unsupervised 分類結果

Kmeans

分兩類

	Low	High
G1	747	2051
G2	893	1207

華德法

	Low	High
G1	1129	1744
G2	511	1514

分七類

	3	4	5	6	7	8	9
G1	6	8	161	151	17	6	0
G2	3	37	174	452	218	43	1
G3	0	7	118	396	222	49	3
G4	3	34	360	328	58	14	0
G5	2	27	236	283	117	14	1
G6	2	11	305	439	182	36	0
G7	4	39	103	149	66	13	0

	3	4	5	6	7	8	9
G1	4	7	133	129	13	6	0
G2	2	20	177	438	214	46	2
G3	2	7	230	336	139	33	1
G4	4	43	102	132	52	9	0
G5	2	29	132	383	198	39	1
G6	1	19	280	387	171	24	1
G7	5	38	403	393	93	18	0

透過不同分類方法將資料分類

- 將資料取後不放回的分法，分為:
  1. Training Data: 70%
  2. Testing Data: 30%
- 分類方法: CART、C5.0、Random Forest、SVM、KSVM
- 重複次數: 100次
- PCA為選取6個主成份
- 評估指標
  1. 前五個重要變數
  2. Accuracy
  3. f1=Precision×Recall
  4. dev.norm=$-2 \times \log(L)\ /n$
  - accuracy and f1 ⇒ higher; deviance ⇒ lower

實際類別 \ 預測類別	Class 1 (T)	Class 2 (F)
Class 1 (T)	TP (True Positive)	FN (False Negative)
Class 2 (F)	FP (False Positive)	TN (True Negative)

準確率 (Precision)	預測類別中，有多少比率的資料剛好屬於該類別	$p = \frac{TP}{TP + FP}$
回想率 (Recall)	實際為某類別，且被判為該類別的比率	$r = \frac{TP}{TP + FN}$

CART

- 比較在有修剪的模型下，有無PCA的表現

Without PCA				
Data	Accuracy	f1	dev.norm	Node
Training	0.8243	0.7611	0.8773	45.17
Testing	0.7622	0.6842	1.1941	

with PCA				
Data	Accuracy	f1	dev.norm	Node
Training	0.8069	0.7453	0.9361	61.11
Testing	0.7289	0.6520	1.3470	

- 評估指標中，不做PCA效果較佳，而分支數也比較多，但數值結果差異並不太大

SVM & KSVM with PCA

- 兩個方法在Testing Data上的表現差不多
- 而比較不做PCA的SVM & KSVM評估指標，
  不做PCA的SVM & KSVM表現較佳

SVM			
Data	Accuracy	f1	dev.norm
Training	0.7317	0.6855	14.83
Testing	0.7224	0.6751	15.34

KSVM_au			
Data	Accuracy	f1	dev.norm
Training	0.7523	0.6966	13.69
Testing	0.7327	0.6754	14.77

註: paper上accuracy: 0.5030~0.8680

paper為一選模方法，該方法不同參數有不同accuracy

Summary

- 從Kmeans與華德法可以發現從化學性質做分類，與品酒師的品質分數不盡相同
- 這筆資料在PCA的表現不太好，使得透過演算法分類後的結果較不佳
- 在重要變數的選擇上，酒精濃度皆為最重要，而揮發酸度與游離二氧化硫也是重要的變數
- 三項評估指標最佳皆為Random Forest
- 最少分支數則為修剪後的CART

	CART_cut	C5.0	R.F.	SVM	KSVM_au	KSVM_gv
Accuracy	0.7622	0.7615	0.8166	0.7817	0.7794	0.7689
f1	0.6842	0.6802	0.7502	0.7117	0.7091	0.7376
dev.norm	1.1941	1.091	0.8777	12.01	12.19	12.77
Node	45.17	88.42	535.6			

	CART	C5.0	Random Forest
1st	alcohol	alcohol	alcohol
2nd	density	free.sulfur.dioxide	volatile.acidity
3rd	volatile.acidity	volatile.acidity	density
4th	free.sulfur.dioxide total.sulfur.dioxide	Fixed.acidity	free.sulfur.dioxide
5th	free.sulfur.dioxide total.sulfur.dioxide	residual.sugar	total.sulfur.dioxide