實作及評估以AUC當作分支準則的分類樹

實作方法來源-Ferri, C., Flach, P., & Hernandez-Orallo, J. (2002). Learning decision trees using the area under the roc curve. Proc. ICML

陳育婷Yu Ting Chen

2018/10/30 3rd meeting report

目錄

1.Review論文提出的分支準則

1.Review論文提出的分支準則

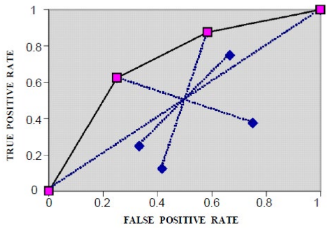


Figure 1. ROC points and convex hull of set Λ.

$$r_i = \frac{E_i^+}{E_i^+ + E_i^-}$$

Definition 2 (Optimal labellings). Given a decision tree for a problem with 2 classes formed by n leaves $\{l_1, l_2, \ldots, l_n\}$ *ordered* by local positive accuracy, i.e, $r_1 \geq r_2, \ldots, r_{n-1} \geq r_n$, we define the set of optimal labellings $\Gamma = \{S_0, S_1, \ldots, S_n\}$ where each labelling S_i ($0 \leq i \leq n$) is defined as: $S_i = \{A^1{}_i, A^2{}_i, \ldots, A^n{}_i\}$ where $A^j{}_i = (j,+)$ if $j \leq i$ and $A^j{}_i = (j,-)$ if $j > i$.

Theorem 6. Given a decision tree for a problem of 2 classes with n leaves, the convex hull of the 2^n possible labellings is formed by exactly those ROC points corresponding to the set of optimal labellings Γ, removing repeated leaves with the same local positive accuracy.

⟶ 計算 一顆決策樹 2^n 種可能的標籤方式(labellings) 之convex hull 相當於將葉節點根據local positive accuracy 排序

	+	–	S_0	S_1	S_2	S_3
LEAF 1	5	1	-	+	+	+
LEAF 2	4	2	-	-	+	+
LEAF 3	3	5	-	-	-	+

Definition 7 (AUC). Let Γ be the set of optimal labellings of a decision tree with n leaves, then the AUC metric is defined as

$$\mathrm{AUC}(\Gamma) = \sum_{i=1..n} A(P_{i-1}, P_i) = \sum_{i=1..n} \frac{E_i^-}{x_t} \cdot \frac{2y_{i-1} + E_i^+}{2y_t} =$$

$$\frac{1}{2x_t y_t} \sum_{i=1..n} E_i^- \left[\left(\sum_{j=1..i-1} 2E_j^+ \right) + E_i^+ \right]$$

Definition 8 (AUCsplit). Given several splits s_j, each one formed by n_j leaves $\{l^j{}_1, l^j{}_2, \ldots, l^j{}_{nj}\}$, then the best split is the one that maximises:

$$AUCsplit(s_j) = \sum_{i=1..n_j} A(P_{i-1}^j, P_i^j)$$

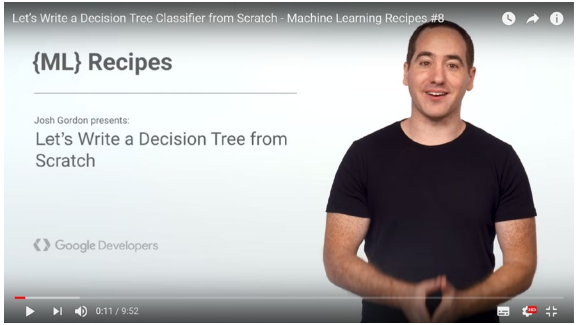
“

2.介紹用於實作之分類樹程式

2.介紹用於實作之分類樹程式

- **程式模板來源**

✓ 程式語言：Python
✓ 提供者：Google Developers

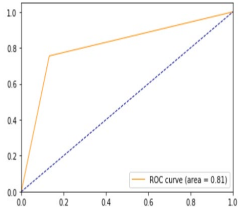


https://github.com/random-forests/tutorials/blob/master/decision_tree.ipynb

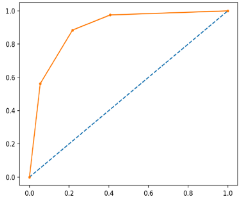
● 程式修改添加	原始	修改添加
分支數目	二元分支	屬質型變數：分支數目為水準數目 屬量型變數：維持二元分支
停止分支條件	1.節點樣本變數值一樣	2.節點樣本目標類別屬於同一類 3.節點樣本個數 < 3
分支及切割準則	1.Gini係數	2.論文方法-該分支所有節點之AUC計算 3.Gain Ratio
結果評估	無	1.4折交叉驗證 2.Index : accuracy, true positive (tp),rate false positive(fp)rate , precision, F1-measure
分類域值標準	節點中比例最高的類別	為了目標類別分佈不平衡(class imblance problem)而修正

2.介紹用於實作之分類樹程式

- 分支準則:二元→多元



→



(0,0)-(fp rate, tp rate)-(1,1)

AUC=(tp rate-fp rate+1)/2

=average of positive and negative accuracy

- 分支及切割準則:Gain Ratio,Gini,AUC

1. $$Ent(D) = -\sum_{k=1}^{|y|} p_k \, log_2 p_k \qquad Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

$$Gain\_ratio = \frac{Gain(D,a)}{IV(a)}, \quad where\ IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{D} log_2 \frac{|D^v|}{D}$$

2. $$Gini(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|y|} p_k^2 \quad Gini\_index(D,a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v)$$

3. $$\mathrm{AUC}(\Gamma) = \frac{1}{2x_t y_t} \sum_{i=1..n} E_i^- \left[\left(\sum_{j=1..i-1} 2E_j^+ \right) + E_i^+ \right]$$

- 結果評估：accuracy, tp rate, fp rate,precision,F1-measure

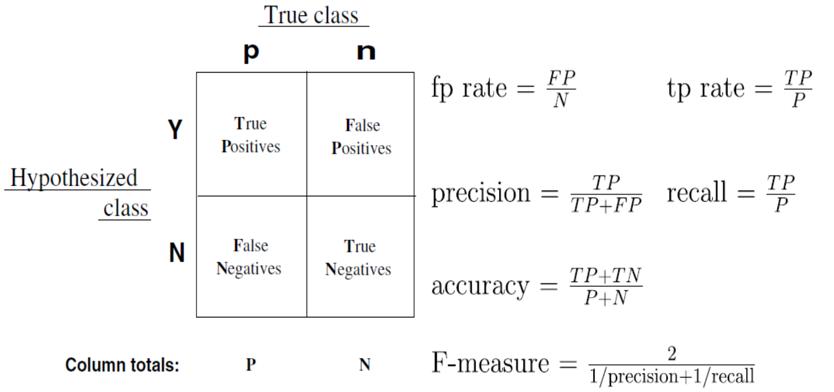


Fig. 1. Confusion matrix and common performance metrics calculated from it.

- 分類域值選擇標準 : class imbalance problem-調整分類方法規則性

Number of Negative Instances:Positive Instances = a : b

→ if $P(+|x) > \frac{b}{a+b}$, $then\ assign +, else\ assign -$

3.介紹用於實作之資料集

3.介紹用於實作之資料集
-Post Oprerative Patient Data

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Browse Through: **8** Data Sets

Table View List View

Default Task - Undo
- Classification (8)
- Regression (2)
- Clustering (0)
- Other (5)

Attribute Type - Undo
- Categorical (14)
- Numerical (41)
- Mixed (8)

Data Type
- Multivariate (8)
- Univariate (0)
- Sequential (0)
- Time-Series (0)
- Text (0)
- Domain-Theory (0)
- Other (0)

Area
- Life Sciences (4)
- Physical Sciences (0)
- CS / Engineering (2)
- Social Sciences (0)
- Business (0)
- Game (1)
- Other (1)

 # Attributes - Undo
- Less than 10 (8)
- 10 to 100 (25)
- Greater than 100 (2)

 # Instances
- Less than 100 (1)
- 100 to 1000 (3)
- Greater than 1000 (4)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Acute Inflammations	Multivariate	Classification	Categorical, Integer	120	6	2009
Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
Chess (King-Rook vs. King)	Multivariate	Classification	Categorical, Integer	28056	6	1994
Contraceptive Method Choice	Multivariate	Classification	Categorical, Integer	1473	9	1997
Mechanical Analysis	Multivariate	Classification	Categorical, Integer, Real	209	8	1990
Post-Operative Patient	Multivariate	Classification	Categorical, Integer	90	8	1993
Teaching Assistant Evaluation	Multivariate	Classification	Categorical, Integer	151	5	1997

3.介紹用於實作之資料集
-Post Oprerative Patient Data

Abstract: Dataset of patient features

Data Set Characteristics:	Multivariate	**Number of Instances:**	90	**Area:**	Life
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	8	**Date Donated**	1993-06-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	81672

Source:

Creators:

Sharon Summers, School of Nursing, University of Kansas
Medical Center, Kansas City, KS 66160
Linda Woolery, School of Nursing, University of Missouri,
Columbia, MO 65211

Donor:
Jerzy W. Grzymala-Busse (jerzy '@' cs.ukans.edu) (913)864-4488

Data Set Information:

The classification task of this database is to determine where patients in a postoperative recovery area should be sent to next. Because hypothermia is a significant concern after surgery (Woolery, L. et. al. 1991), the attributes correspond roughly to body temperature measurements.

Results:
-- LERS (LEM2): 48% accuracy

3.介紹用於實作之資料集
-Post Oprerative Patient Data

Attribute Information:8 attributes,90 instances

1.L-CORE (patient's internal temperature in C):
high (> 37), mid (>= 36 and <= 37), low (< 36)

2.L-SURF (patient's surface temperature in C):
high (> 36.5), mid (>= 36.5 and <= 35), low (< 35)

3.L-O2 (oxygen saturation in %):
excellent (>= 98), good (>= 90 and < 98),
fair (>= 80 and < 90), poor (< 80)

4.L-BP (last measurement of blood pressure):
high (> 130/90), mid (<= 130/90 and >= 90/70), low (< 90/70)

5.SURF-STBL (stability of patient's surface temperature):
stable, mod-stable, unstable

6.CORE-STBL (stability of patient's core temperature)
stable, mod-stable, unstable

7.BP-STBL (stability of patient's blood pressure)
stable, mod-stable, unstable

8.COMFORT (patient's perceived comfort at discharge : measured as integer ∈ [0, 20]

Class:Distribution: I (2) S (24) A (64)

decision ADM-DECS (discharge decision) :

I (patient sent to Intensive Care Unit)

S (patient prepared to go home)

A (patient sent to general hospital floor)

```
['mid', 'low', 'good', 'high', 'unstable', 'stable', 'stable', 10, 'A']
```

"

4.實作結果呈現

4.實作結果呈現-分類樹demo :最後1折(第4折)

–# of training data : 66, # of testing data:22

```
# Demo:Find the best question to ask first for ourdataset.
num,branch,col,cutpoint,best_gain, best_question = find_best_split(training_data,data)
print(num,branch,col,cutpoint,best_gain,best_question)

num,branch,col,cutpoint,best_gain, best_question = find_best_splitratio(training_data,data)
print(num,branch,col,cutpoint,best_gain,best_question)

num,branch,col,cutpoint,best_gain, best_question = find_best_splitauc(training_data,data)
print(num,branch,col,cutpoint,best_gain,best_question)
```

```
False 3 0 ['high', 'low', 'mid'] 0.022721464806949088 What's the L-CORE of instances?
True 2 7 10 0.08894064707447578 Is COMFORT >= 10?
False 3 1 ['low', 'high', 'mid'] 0.6054421768707483 What's the L-SURF of instances?
```

```
def build_treeauc(rows,data):
    num,branch,column,cutpoint,gain,question = find_best_splitauc(rows,data)
    # Base case: no further gain(3 stopping criterions met,or gain is truly 0)
    if gain == 0:
        return Leaf(rows)
```

```
my_treeauc = build_treeauc(training_data,data)
print_tree(my_treeauc)
```

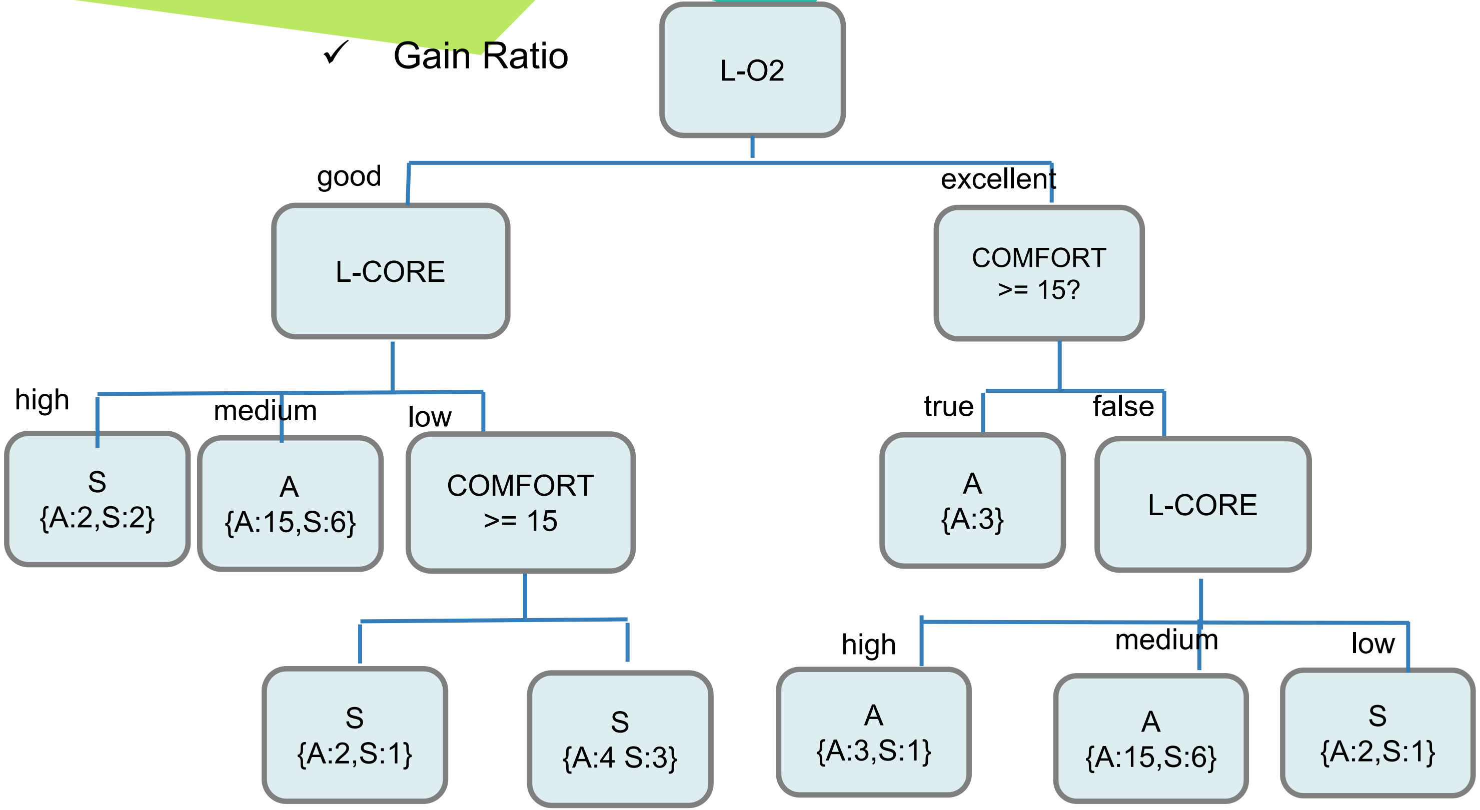
```
What's the L-SURF of instances?
-->low
  Predict {'A': 12, 'S': 6}
-->high
  Predict {'S': 1, 'A': 11}
-->mid
  Is COMFORT >= 15?
  --> True:
    Predict {'A': 6, 'S': 2}
  --> False:
    What's the L-CORE of instances?
    -->high
      Predict {'A': 3}
    -->low
      Predict {'S': 3, 'A': 2}
    -->mid
      Predict {'A': 15, 'S': 6}
```

```
Actual: A. Predicted: A   Actual: S. Predicted: A
Actual: S. Predicted: A   Actual: S. Predicted: A
Actual: A. Predicted: A   Actual: A. Predicted: A
Actual: A. Predicted: A   Actual: S. Predicted: A
Actual: A. Predicted: A   Actual: A. Predicted: A
Actual: A. Predicted: A   Actual: A. Predicted: A
Actual: A. Predicted: S   Actual: S. Predicted: A
Actual: A. Predicted: A   Actual: A. Predicted: S
Actual: A. Predicted: A   Actual: S. Predicted: A
Actual: A. Predicted: A   Actual: A. Predicted: A
Actual: A. Predicted: A
```

```
Accuracy: 0.45454545454545453. TP rate: 0.8571428571428571. FP rate: 0.7333333333333333. Precision: 0.35294117647058826. F1:
```

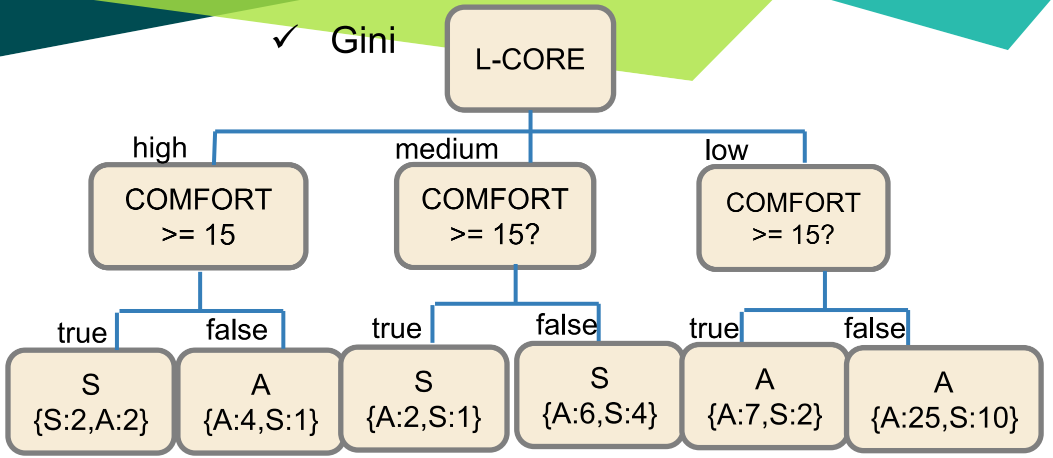
4.實作結果呈現-分類樹demo fold1

✓ Gain Ratio

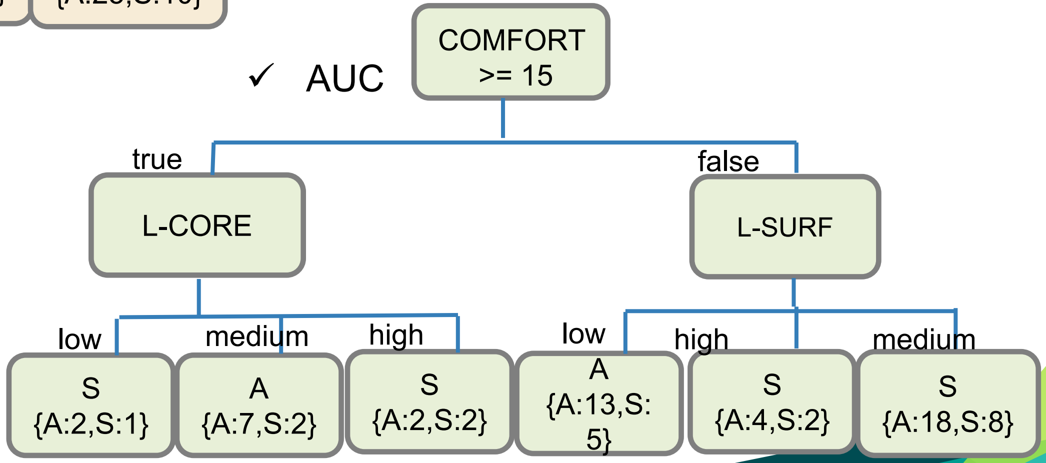


4.實作結果呈現-分類樹demo fold1

✓ Gini

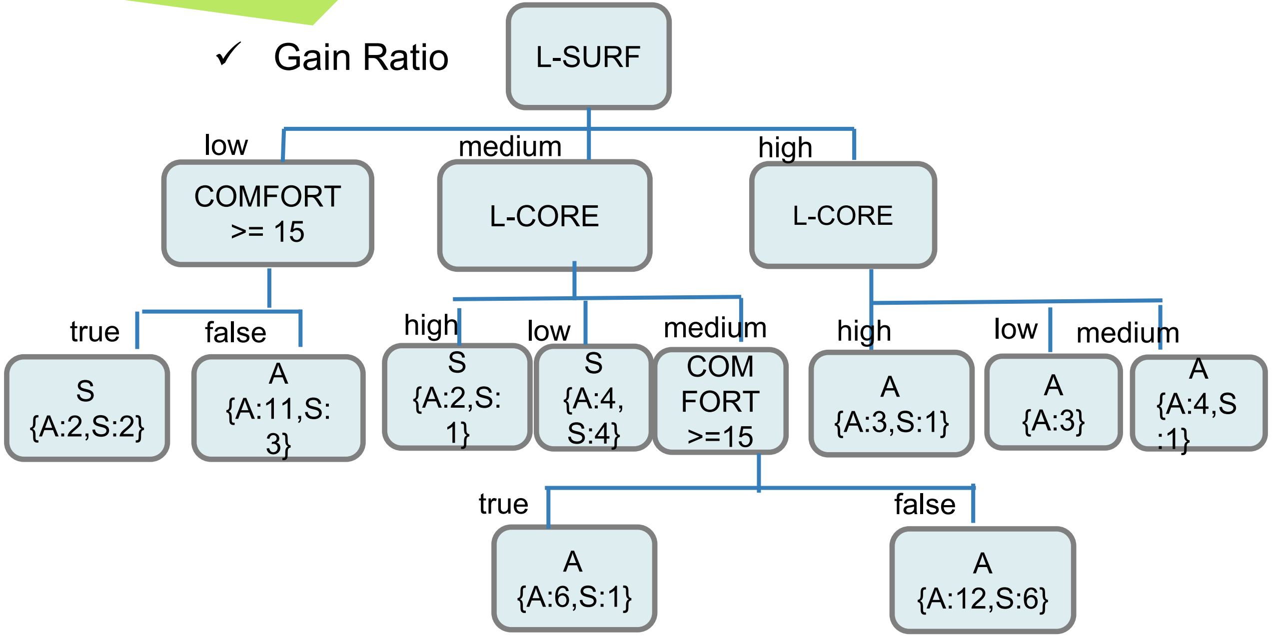


✓ AUC



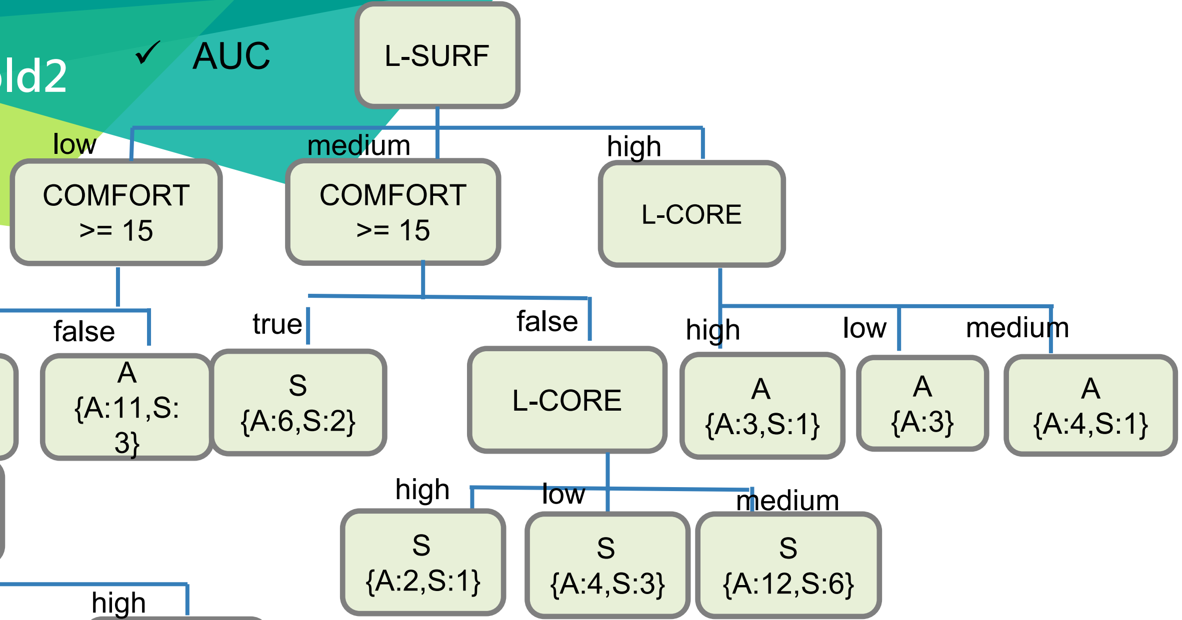
4.實作結果呈現-分類樹demo fold2

✓ Gain Ratio

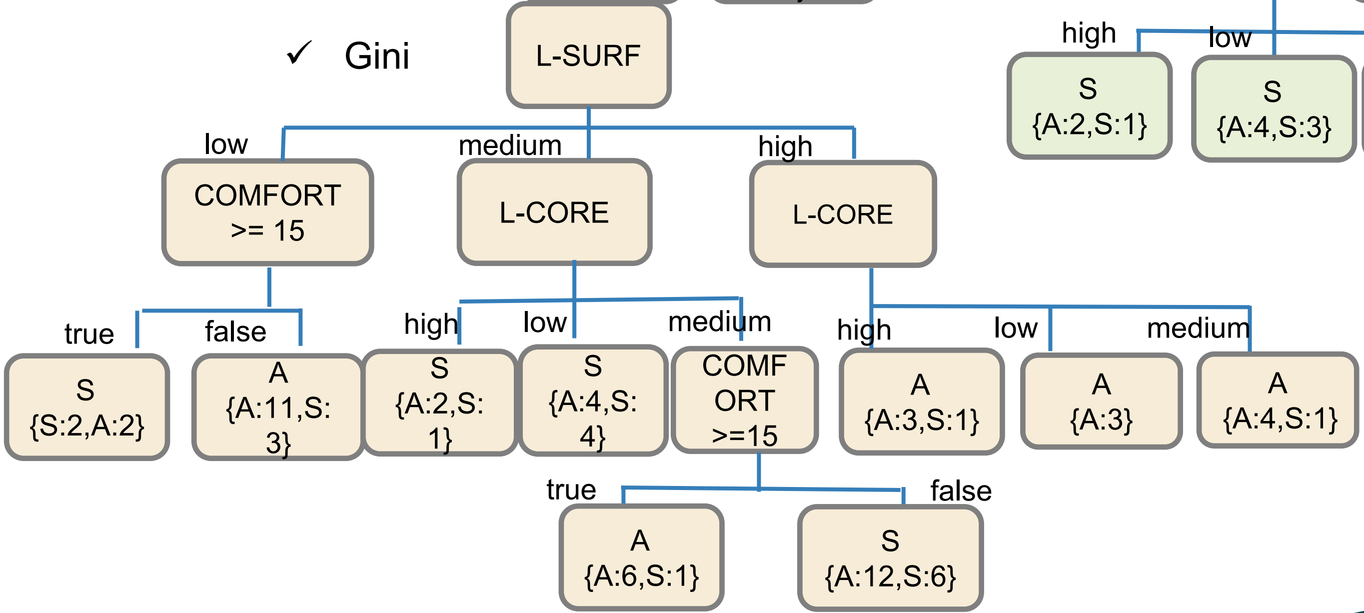


4.實作結果呈現-分類樹demo fold2

✓ AUC

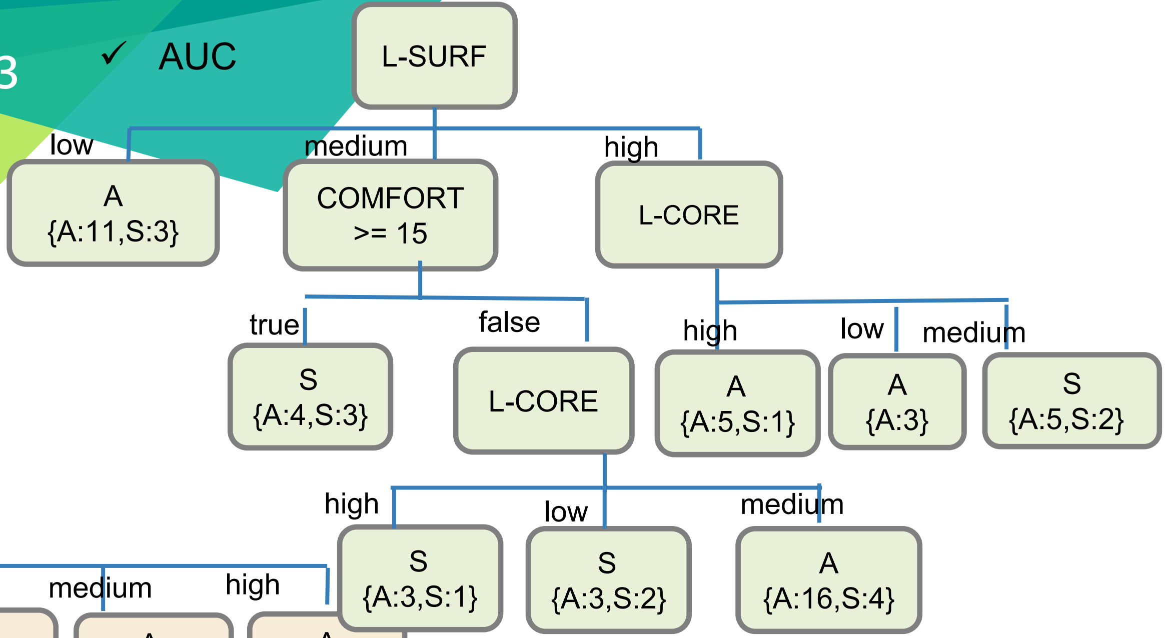


✓ Gini

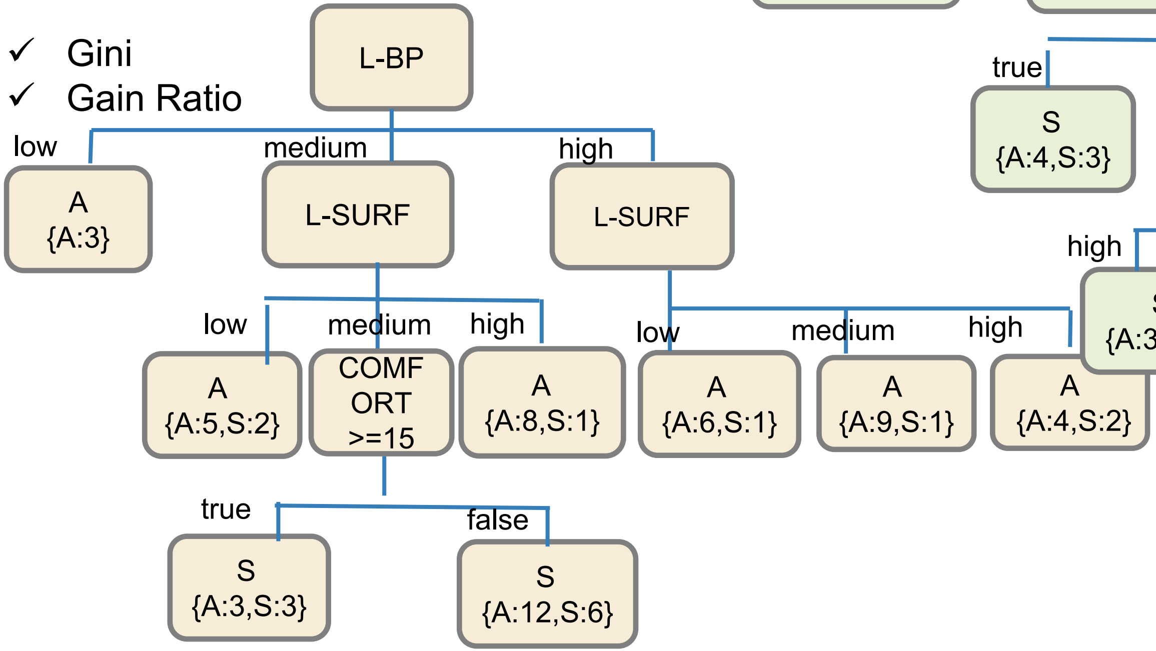


4.實作結果呈現-分類樹demo fold3

✓ AUC

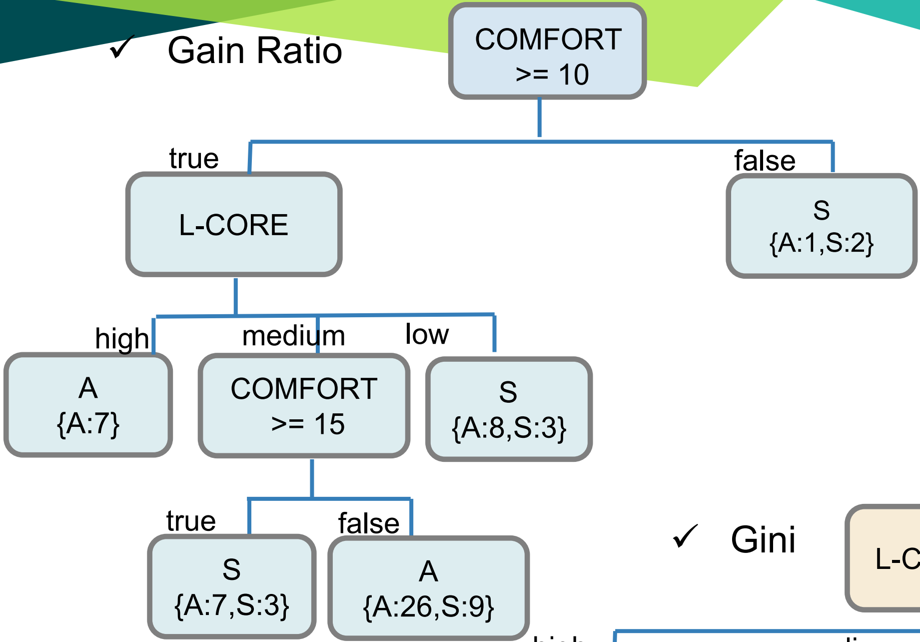


✓ Gini
✓ Gain Ratio

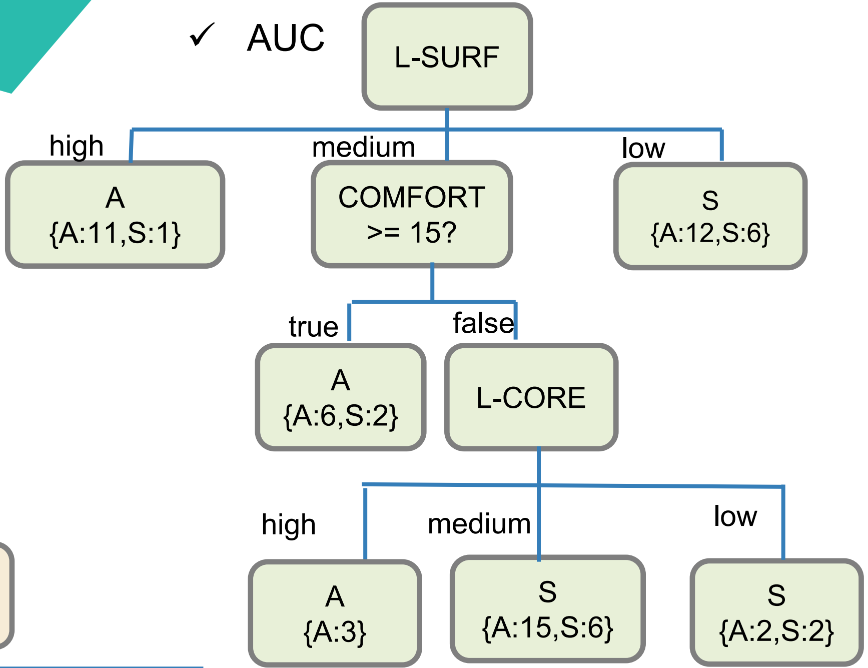


4.實作結果呈現-分類樹demo fold4

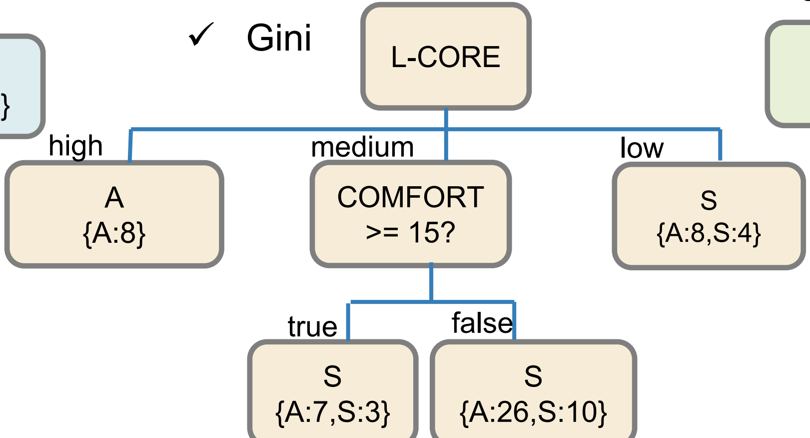
✓ Gain Ratio



✓ AUC



✓ Gini



4.實作結果呈現-結果評估

1	Gini	Gain Ratio	AUC	2	Gini	Gain Ratio	AUC	3	Gini	Gain Ratio	AUC	4	Gini	Gain Ratio	AUC
Accuracy	**0.682**	**0.636**	0.318		0.409	0.409	**0.455**		**0.5**	**0.5**	0.455		0.272	**0.5**	**0.318**
tp rate	0.25	0.25	**0.75**		**0.2**	**0.2**	**0.2**		**0.75**	**0.75**	0.125		**0.571**	0.286	**0.429**
fp rate	**0.222**	**0.278**	0.778		0.529	0.529	**0.471**		0.643	0.643	**0.357**		0.867	**0.4**	**0.733**
Precision	**0.20**	0.167	**0.176**		0.1	0.1	**0.111**		**0.4**	**0.4**	0.167		**0.235**	**0.25**	0.214
F1	**0.056**	0.05	**0.071**		0.033	0.033	**0.036**		**0.13**	**0.13**	0.036		**0.083**	0.067	**0.071**
Testing data:S:4,A:18				Testing data:S:4,A:18				Testing data:S:8,A:14				Testing data:S:8,A:14			

LERS (LEM2) : 48%
R-rpart package : 54% (4 fold)

Index	Accuracy	tp rate	fp rate	precision	F1
Gini	**0.466**(0.172)	**0.443(0.263)**	**0.449(0.181)**	**0.234**(0.125)	**0.076**(0.042)
Gain Ratio	**0.511(0.094)**	**0.372(0.255)**	**0.463(0.158)**	**0.229**(0.129)	**0.07**(0.042)
AUC	0.387**(0.079)**	0.276(0.281)	0.585(0.203)	0.131**(0.049)**	0.053**(0.021)**

5.結論

5.結論

1. 論文提出的AUC分支準則實際上可以實行
2. 不管是屬質、屬量型變數均可以處理
3. 此次實作是根據UCI上的1筆資料集

Table 3. Accuracy and AUC for Gain Ratio and AUCsplit.

	Gain Ratio		AUCsplit		Better?	
Set	Acc.	AUC	Acc.	AUC	Acc.	AUC
1	90.7±6.6	83.6±11.8	96.5±3.9	94.3±6.7	✓	✓
2	57.7±6.5	61.1±7.9	56.0±6.2	56.7±8.0	x	x
3	97.6±7.8	97.4±8.5	99.1±1.1	99.1±1.4	✓	✓
4	78.9±4.6	79.8±7.2	77.6±4.7	76.9±6.5	x	x
5	95.8±2.6	95.2±3.1	95.8±2.6	95.2±3.1		
6	1±0	1±0	1±0	1±0		
7	92.5±4.1	91.5±6.1	92.9±3.7	94.7±4.6		✓
8	72.1±10.2	61.3±16.9	69.5±10.6	59.3±16.2	x	
9	92.0±4.7	90.4±7.0	89.6±5.0	89.7±6.7	x	
10	62.6±8.8	64.2±10.6	64.0±9.0	65.8±10.1		
11	73.3±5.7	76.6±6.9	72.5±5.1	76.7±6.0		
12	99.1±2.3	99.5±1.6	99.2±0.6	99.5±0.6		
13	68.2±10.2	67.4±11.9	71.0±10.4	73.6±11.0	✓	✓
14	95.4±2.5	96.3±2.5	96.2±2.5	97.6±2.1	✓	✓
15	86.4±14.2	85.1±17.9	83.4±14.0	63.5±22.3		x
16	98.0±10.9	84.6±13.1	98.6±0.8	94.8±5.6	✓	✓
17	95.2±1.4	92.6±3.5	96.7±1.2	95.1±3.1	✓	✓
18	71.4±12.4	61.5±20.8	68.9±11.6	59.8±21.3		
19	95.0±1.8	98.2±0.9	94.8±1.9	98.1±1.0		
20	1±0	1±0	1±0	1±0		
21	99.6±0.3	99.6±0.5	99.6±0.2	99.4±0.6		
22	96.8±0.9	93.3±4.7	96.8±0.2	95.1±6.9		✓
23	70.4±3.9	72.2±4.9	71.1±3.6	73.3±4.0		✓
24	99.5±0.2	98.9±1.4	99.5±0.1	99.3±0.7	✓	✓
25	98.9±1.8	94.2±19.4	99.5±0.3	98.5±1.8	✓	✓
M.	87.49	85.78	87.55	86.24		

- t test with level of confidence 0.9

✓ Accuracies : 8 wins/13 ties /4 loses

✓ AUC : 11 wins/11 ties/ 3 loses