



**Khoa Công Nghệ Thông Tin
Trường Đại Học Cần Thơ**



Đánh giá hiệu quả của giải thuật học

Đỗ Thanh Nghi
dtnghi@cit.ctu.edu.vn

Cần Thơ
12-02-2019

Nội dung

- Phương pháp cổ điển
- ROC

Nghi thức kiểm tra

- nếu dữ liệu có 1 tập học và 1 tập kiểm tra sẵn dùng
 - dùng dữ liệu học để xây dựng mô hình, dùng tập kiểm tra để đánh giá hiệu quả của giải thuật
- nếu dữ liệu không có 1 tập kiểm tra sẵn
 - sử dụng nghi thức k-fold : chia tập dữ liệu thành k phần (fold) bằng nhau, lặp lại k lần, mỗi lần sử dụng k-1 folds để học và 1 fold để kiểm tra, sau đó tính trung bình của k lần kiểm tra
 - nghi thức hold-out : lấy ngẫu nhiên 2/3 tập dữ liệu để học và 1/3 tập dữ liệu còn lại dùng cho kiểm tra, có thể lặp lại quá bước này k lần rồi tính giá trị trung bình

10-fold



■ it 1 :

test	train	train	train	train	train	train	train	train	train
------	-------	-------	-------	-------	-------	-------	-------	-------	-------

■ it 2 :

train	test	train	train	train	train	train	train	train	train
-------	------	-------	-------	-------	-------	-------	-------	-------	-------

■ it 10 :

train	train	train	train	train	train	train	train	train	test
-------	-------	-------	-------	-------	-------	-------	-------	-------	------

Nghi thức kiểm tra

- nếu dữ liệu có số phần tử lớn hơn 300
 - sử dụng nghi thức k-fold với $k = 10$
- nếu dữ liệu có số phần tử nhỏ hơn 300
 - sử dụng nghi thức leave-1-out (k-fold với $k = \text{số phần tử}$)

Nội dung

- **Phương pháp cổ điển**
- ROC

Confusion matrix (C) cho k lớp

dự đoán =>	1	...	k
1			
...			
k			

- ❑ $C[i, j]$: số phần tử lớp **i** được giải thuật dự đoán là lớp **j**
- ❑ $C[i, i]$: số phần tử phân lớp đúng
- ❑ Độ chính xác lớp **i**: $C[i, i] / C[i,]$
- ❑ Độ chính xác tổng thể: $\sum C[i, i] / C$

Confusion matrix (C) cho 2 lớp (+/-)

dự đoán =>	dương	âm
dương	TP	FN
âm	FP	TN

$$prec = \frac{tp}{tp + fp}$$

$$rec = \frac{tp}{tp + fn} = \frac{tp}{pos}$$

$$acc = \frac{tp + tn}{tp + fn + tn + fp} = \frac{tp + tn}{pos + neg}$$

$$F1 = \frac{2 \times prec \times rec}{prec + rec}$$

$$bep = \frac{prec + rec}{2}$$

Dữ liệu không cân bằng

- nếu dữ liệu không cân bằng
 - lớp positive có số lượng rất nhỏ so với lớp negative
 - ví dụ : positive = 5%, negative = 95%
 - một giải thuật học có thể cho kết quả 95% độ chính xác khi phân loại, nhưng chúng ta có thể mất hoàn toàn lớp positive
- khả năng tách lớp positive từ lớp negative

Nội dung

- Phương pháp cổ điển
- **ROC**

ROC (Provost & Fawcett, 1997)

■ Receiver Operating Characteristic

- khả năng tách positive từ negative
- đồ thị 2 chiều : tpr, fpr
- $O(0, 1)$: lý tưởng
- đường chéo ($tpr = fpr$) : ngẫu nhiên
- sử dụng diện tích dưới đường ROC để đánh giá hiệu năng

ROC (Provost & Fawcett, 1997)

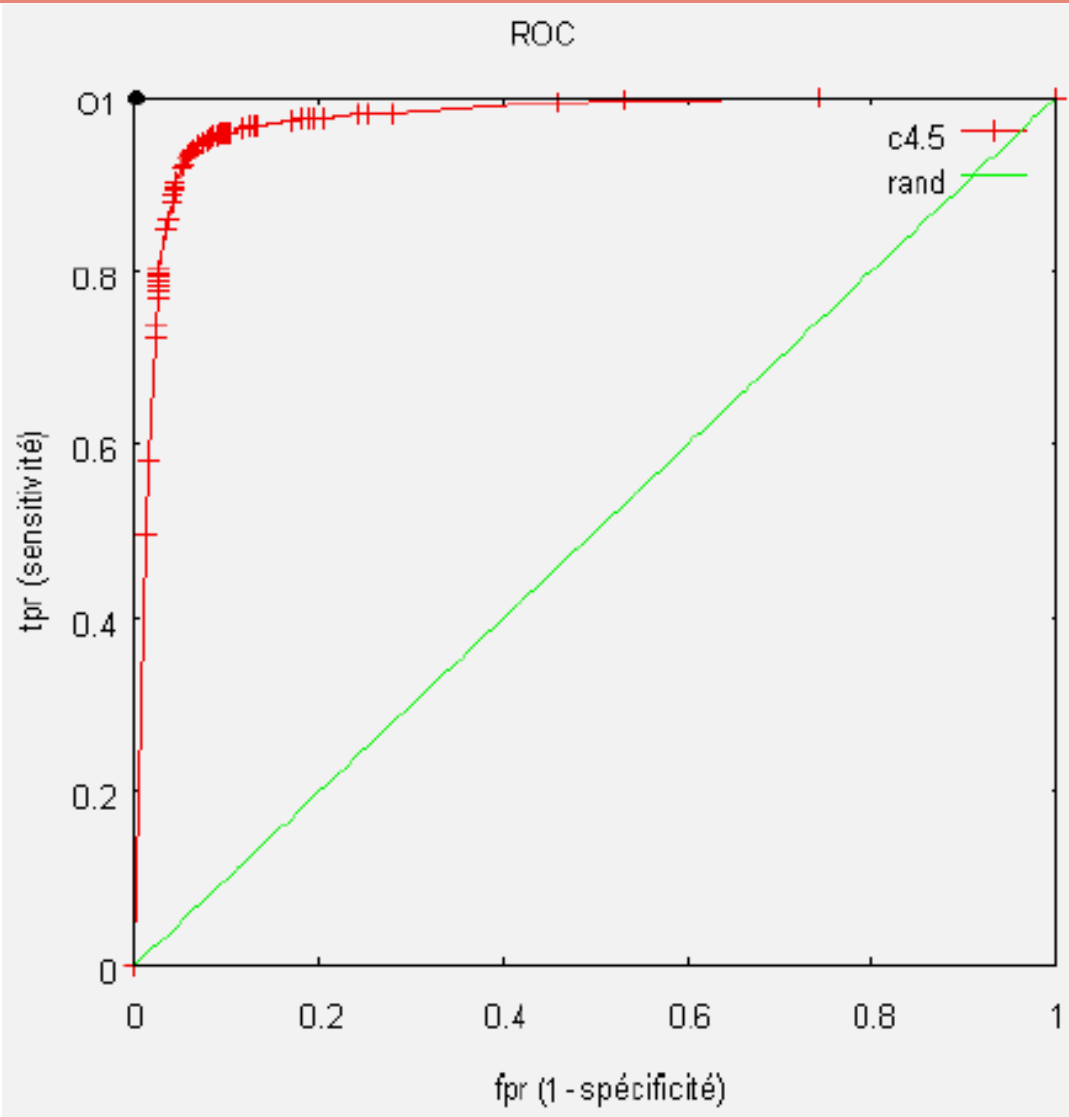
■ Receiver Operating Characteristic

- khả năng tách positive từ negative
- đồ thị 2 chiều : tpr, fpr
- O(0, 1) : lý tưởng
- đường chéo (tpr = fpr) : ngẫu nhiên
- sử dụng diện tích dưới đường ROC để đánh giá hiệu năng

$$tpr = \frac{tp}{tp + fn} = \frac{tp}{pos}$$

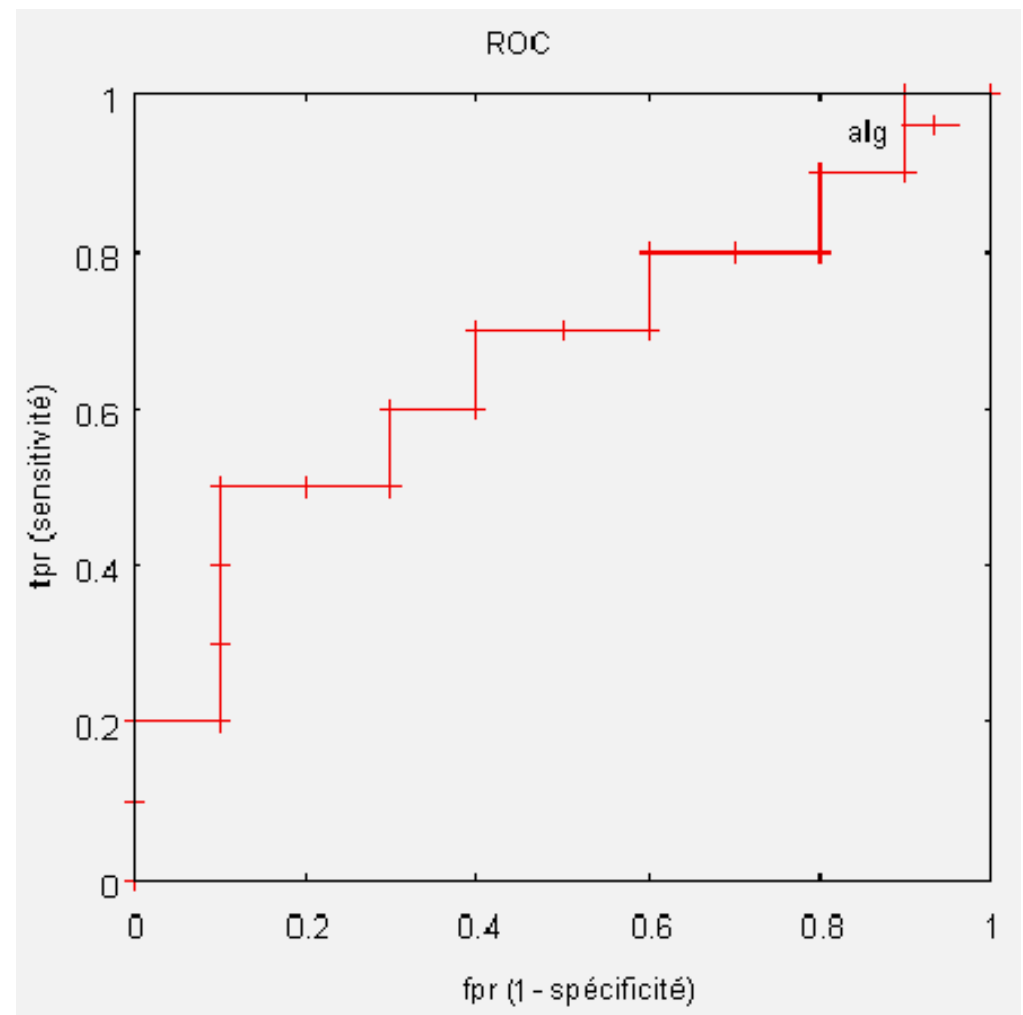
$$fpr = \frac{fp}{fp + tn} = \frac{fp}{neg}$$

ROC (Provost & Fawcett, 1997)



ROC (Provost & Fawcett, 1997)

Order	Classes	Discriminant threshold
1	+1	0.9
2	+1	0.8
3	-1	0.7
4	+1	0.6
5	+1	0.55
6	+1	0.54
7	-1	0.53
8	-1	0.55
9	+1	0.51
10	-1	0.505
11	+1	0.4
12	-1	0.39
13	+1	0.38
14	-1	0.37
15	-1	0.36
16	-1	0.35
17	+1	0.34
18	-1	0.33
19	+1	0.3
20	-1	0.1





Cám ơn !