

Giải thuật rừng ngẫu nhiên với luật gán nhãn cục bộ cho phân lớp

Đỗ Thanh Nghị, Phạm Nguyên Khang
Nguyễn Hữu Hòa, Nguyễn Minh Trung

Khoa CNTT-TT, ĐH. Cần Thơ
dtnghi@cit.ctu.edu.vn



Nội dung

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển

Nội dung

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển

Giới thiệu

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển

- Phân lớp dữ liệu có số chiều lớn
 - Top 10 vấn đề khó của máy học (Yang & Wu, 06)
 - Dữ liệu hoàn toàn dễ tách biệt
 - Có nhiều lựa chọn mô hình
 - Mô hình máy học thường phân lớp chính xác 100% trên tập học, nhưng dự đoán sai trên tập kiểm tra
 - Mô hình tốt: phân lớp tốt trên tập kiểm tra (mới)
- Mô hình phân lớp dữ liệu có số chiều lớn
 - Máy học véc-tơ hỗ trợ SVM (Vapnik, 95)
 - Rừng ngẫu nhiên (Breiman, 01)

Giới thiệu

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển

■ Rừng ngẫu nhiên (Breiman, 01)

- Phân lớp hiệu quả dữ liệu có số chiều lớn
- Thời gian huấn luyện mô hình nhanh
- Khả năng chịu đựng nhiễu
- Không bị học vẹt

■ Luật gán nhãn cục bộ

- Luật bình chọn số đông tại các nút lá của rừng ngẫu nhiên làm giảm độ chính xác khi phân lớp
- Đề xuất: luật gán nhãn cục bộ
- Cải tiến kết quả phân lớp của rừng ngẫu nhiên trên 10 tập dữ liệu y-sinh (Jinyan & Huiqing, 02)

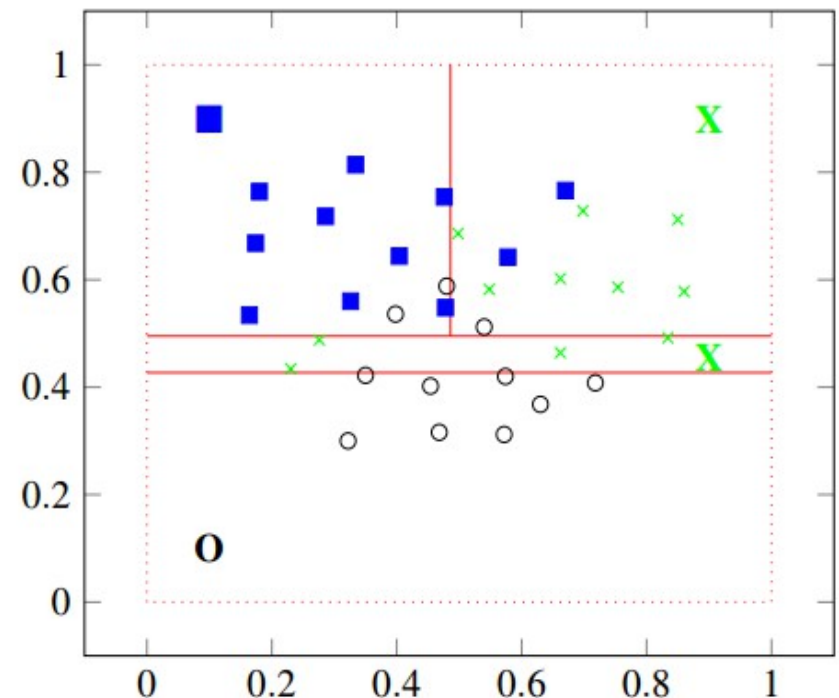
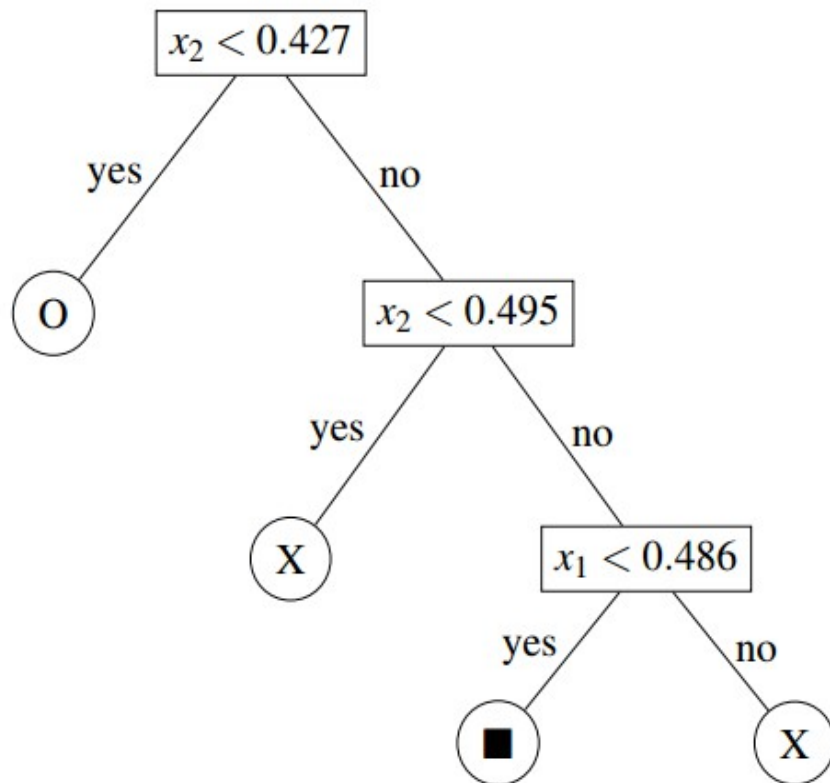
Nội dung

- Giới thiệu
- **Giải thuật rừng ngẫu nhiên**
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển

Cây quyết định

■ CART (Breiman et al., 84), C4.5 (Quinlan, 93)

- Xây dựng cây: phân hoạch dữ liệu
- Cắt nhánh: tránh học vẹt



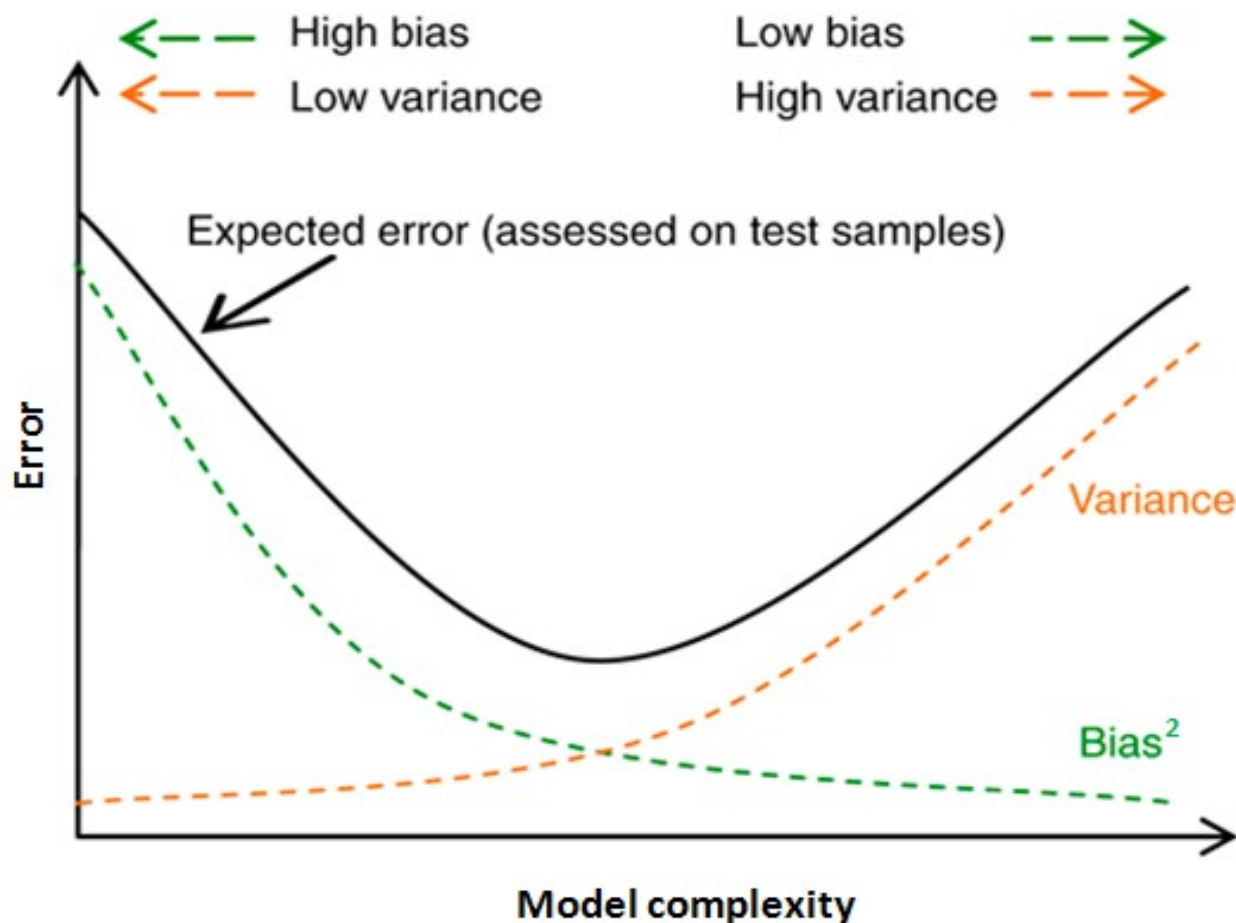
- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển

Hiệu quả giải thuật học

■ Phân tích của (Breiman, 96)

□ $\text{Error} = \text{Bias}^2 + \text{Variance}$

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển



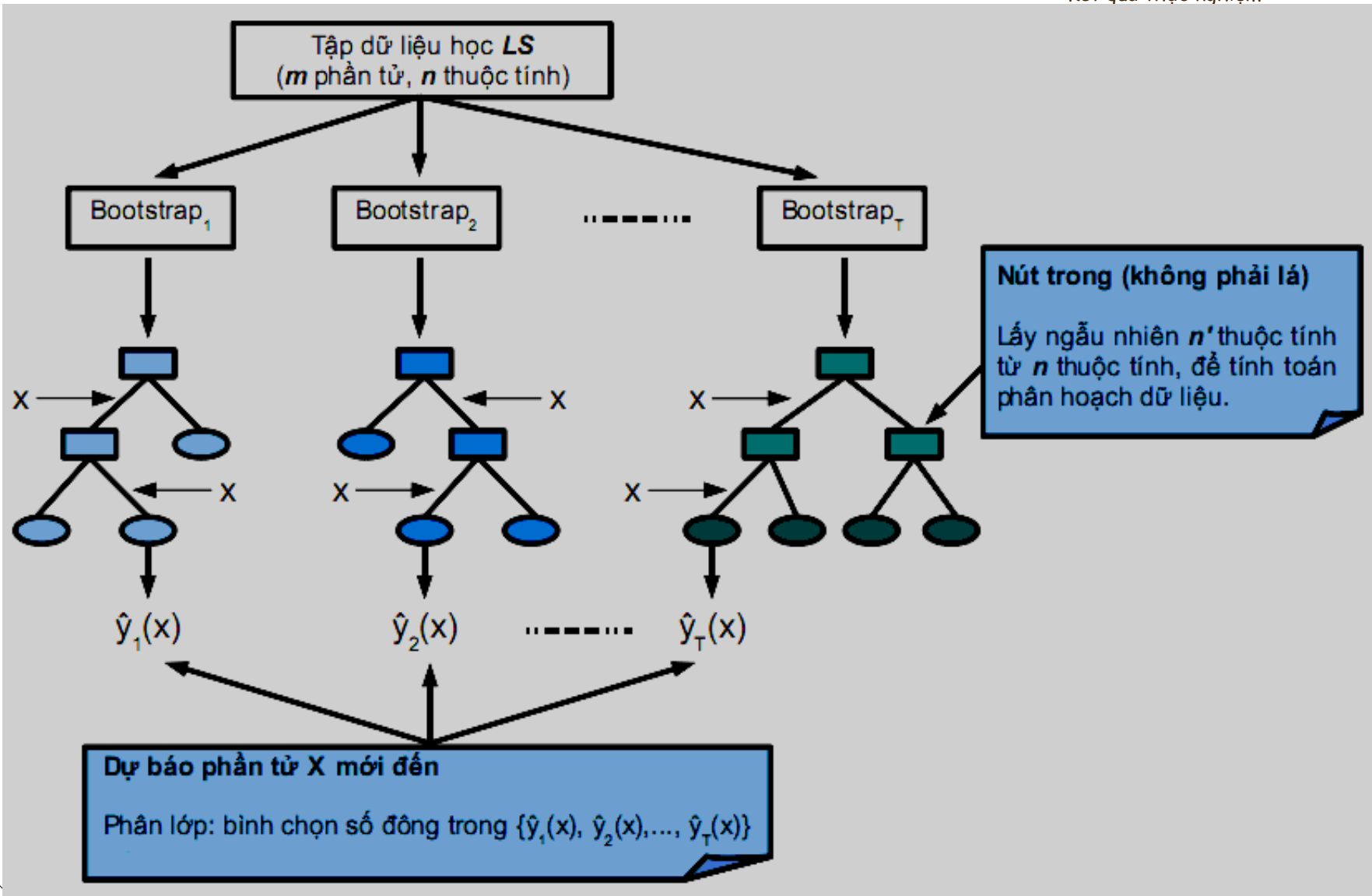
- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển

Rừng ngẫu nhiên (Breiman, 01)

- Xây dựng tập hợp các cây quyết định
 - Giữ thành phần lỗi bias thấp: cây không cắt nhánh
 - Giảm thành phần lỗi variance: tương quan giữa các cây giảm thấp nhờ vào cách xây dựng ngẫu nhiên
- Xây dựng cây
 - Bootstrap: ngẫu nhiên có hoàn lại từ tập huấn luyện
 - Nút trong: lấy ngẫu nhiên tập con thuộc tính để tính toán phân hoạch
 - Cây xây dựng không cần cắt nhánh

Rừng ngẫu nhiên (Breiman, 01)

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm

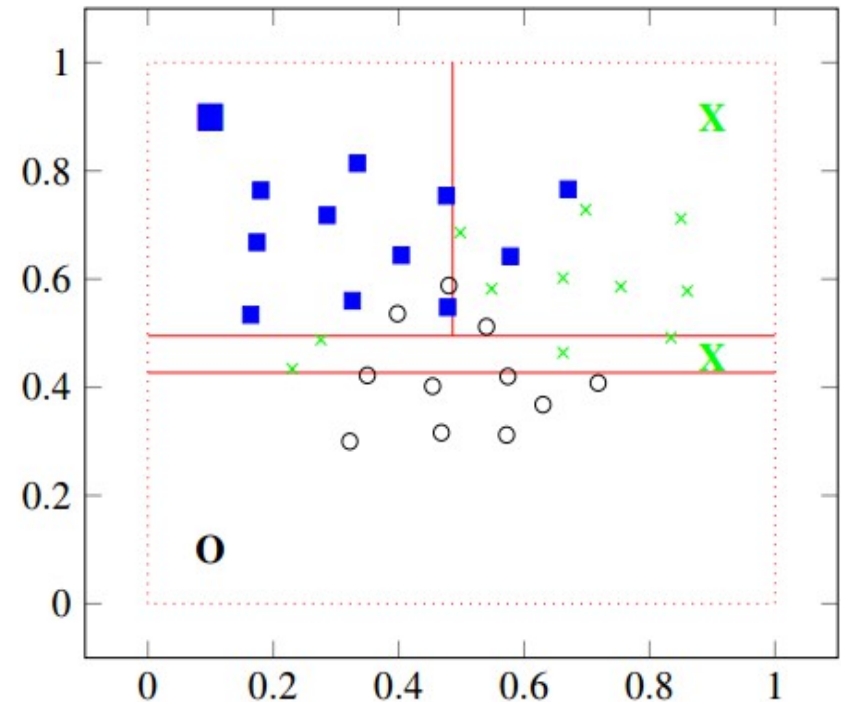
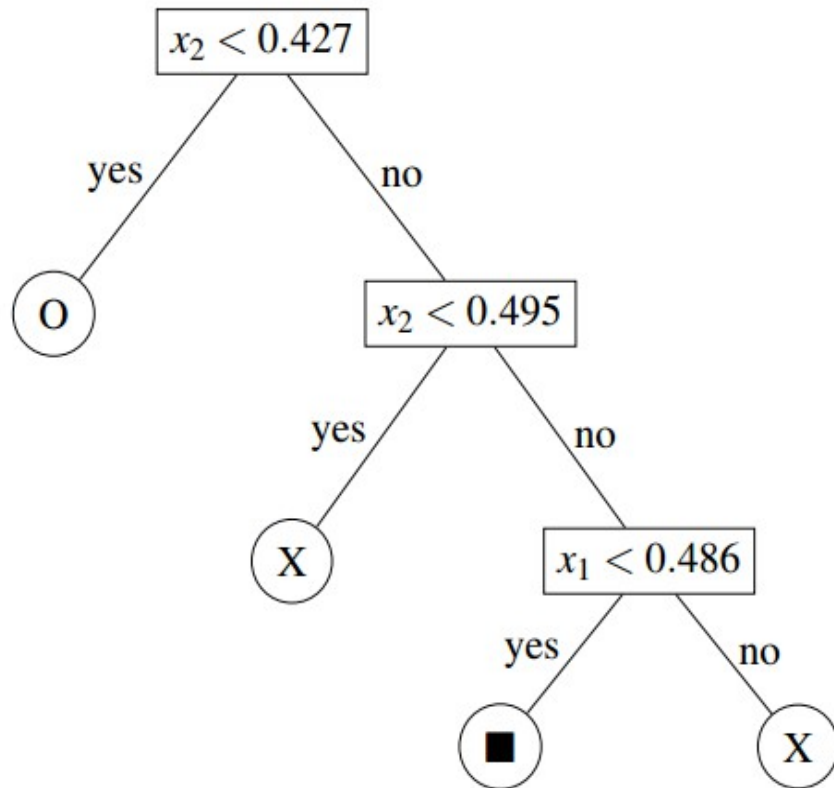


Nội dung

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển

Luật gán nhãn (số đông)

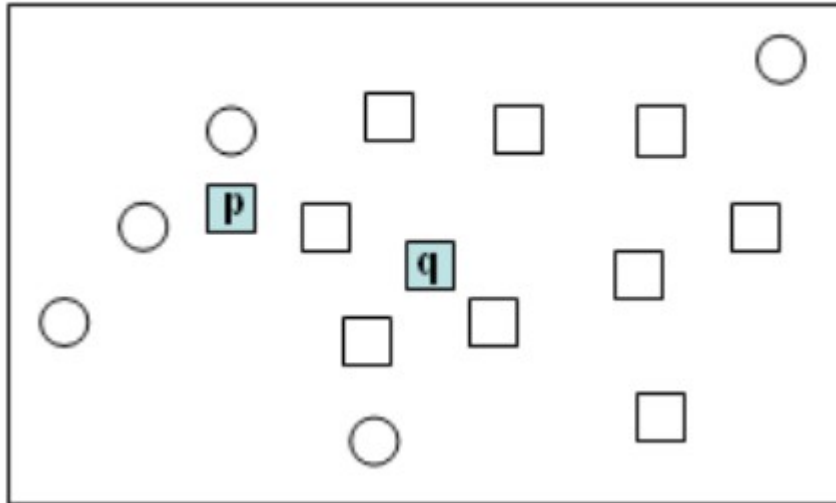
- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển



Luật gán nhãn

số đông

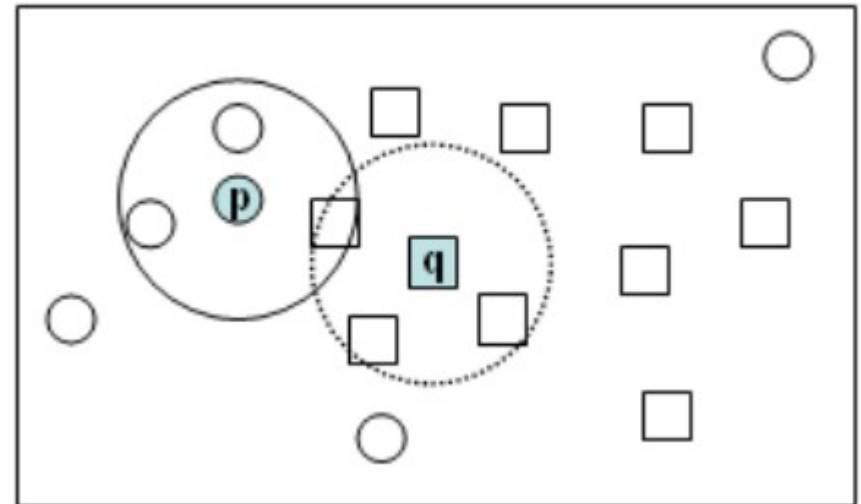
Bình chọn số đông: nhấn nút lá là ☐



- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển

cục bộ

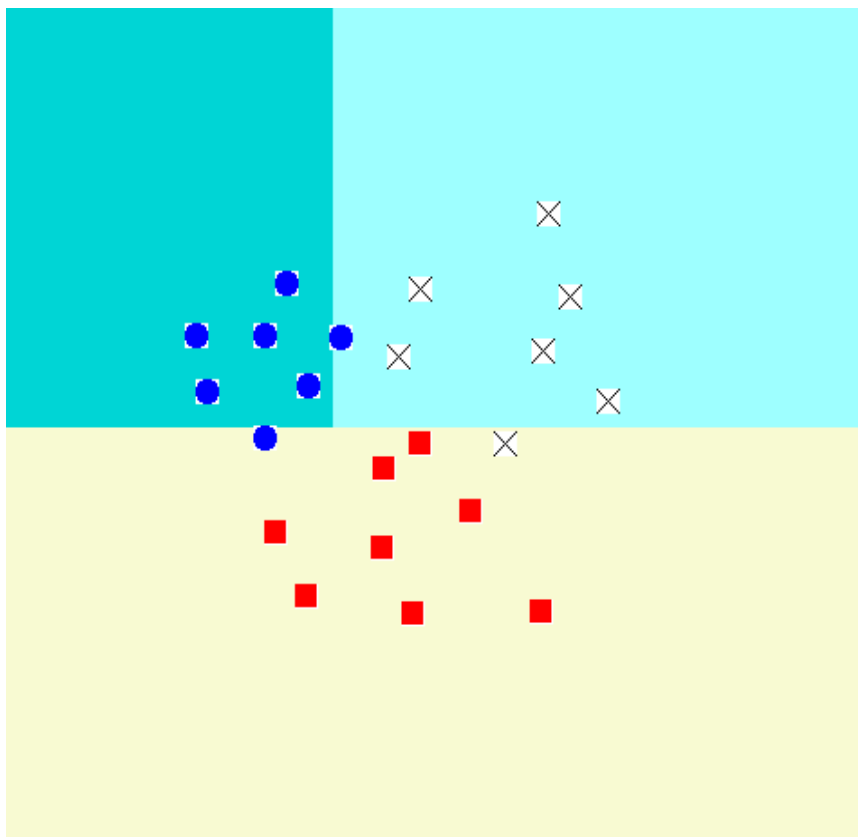
Luật cục bộ: không gán nhãn nút lá



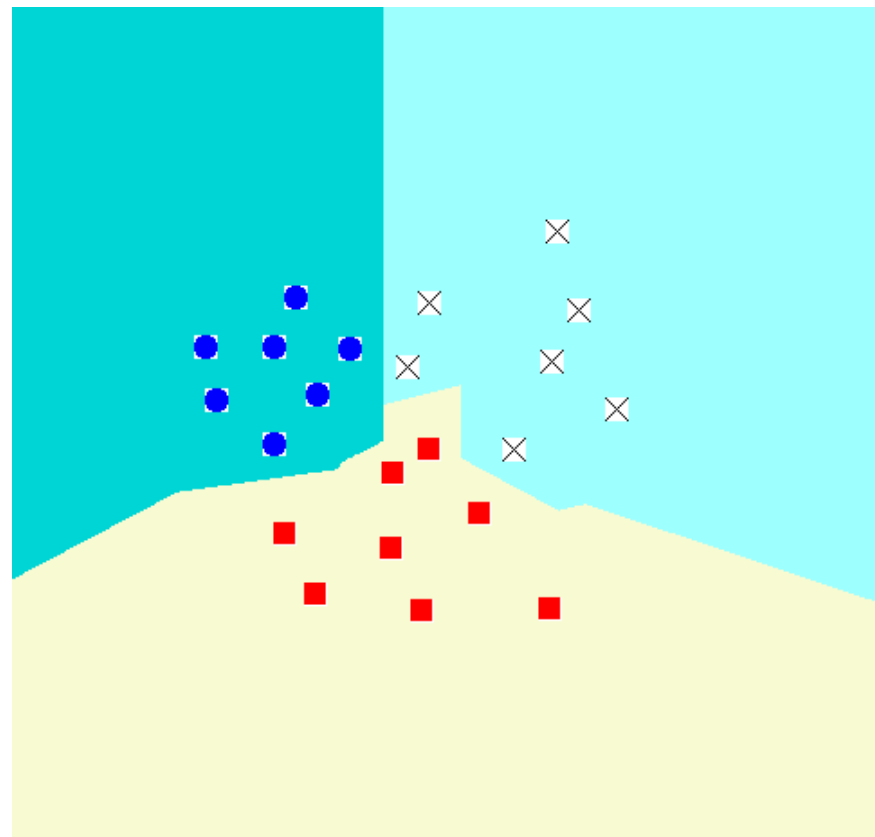
Luật gán nhãn

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển

số đông



cục bộ



Nội dung

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- **Kết quả thực nghiệm**
- Kết luận và hướng phát triển

Thực nghiệm

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển

■ Chương trình

- Cài đặt rừng ngẫu nhiên RF-C4.5 (Maj), dựa trên mã chương trình C4.5 (Quinlan, 93)
- Cài đặt rừng ngẫu nhiên với luật cục bộ RF-C4.5 (KNN)
- Máy học véc-tơ hỗ trợ LibSVM (Chang & Lin, 11)

■ Máy tính PC Linux Mandriva 2010.2

- Intel Core2 Duo 2.4 GHz
- 4GB RAM

Thực nghiệm

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- **Kết quả thực nghiệm**
- Kết luận và hướng phát triển

■ 10 tập dữ liệu γ -sinh (Jinyan & Huiqing, 02)

ID	Tập dữ liệu	Số phần tử	Số chiều	Lớp	Nghi thức
1	ALL-AML-Leukemia	72	7129	ALL, AML	trn-tst
2	MLL-Leukemia	72	12582	MLL, rest	trn-tst
3	Breast Cancer	97	24481	relapse, non-relapse	trn-tst
4	Prostate Cancer	136	12600	cancer, normal	trn-tst
5	Lung Cancer	181	12533	cancer, normal	trn-tst
6	Diffuse Large B-Cell Lymphoma	47	4026	germinal, activated	loo
7	Subtypes of Acute Lymphoblastic (Hyperdip)	327	12558	Hyperdip, rest	trn-tst
8	Subtypes of Acute Lymphoblastic (TEL-AML1)	327	12558	TEL-AML1, rest	trn-tst
9	Subtypes of Acute Lymphoblastic (T-ALL)	327	12558	TEL-ALL, rest	trn-tst
10	Subtypes of Acute Lymphoblastic (Others)	327	12558	Others, diagnostic groups	trn-tst

Thực nghiệm

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- **Kết quả thực nghiệm**
- Kết luận và hướng phát triển

■ Nghi thức kiểm tra

- Training set: xây dựng mô hình
- Testing set: đánh giá mô hình
- Leave-1-out

■ Tiêu chí đánh giá

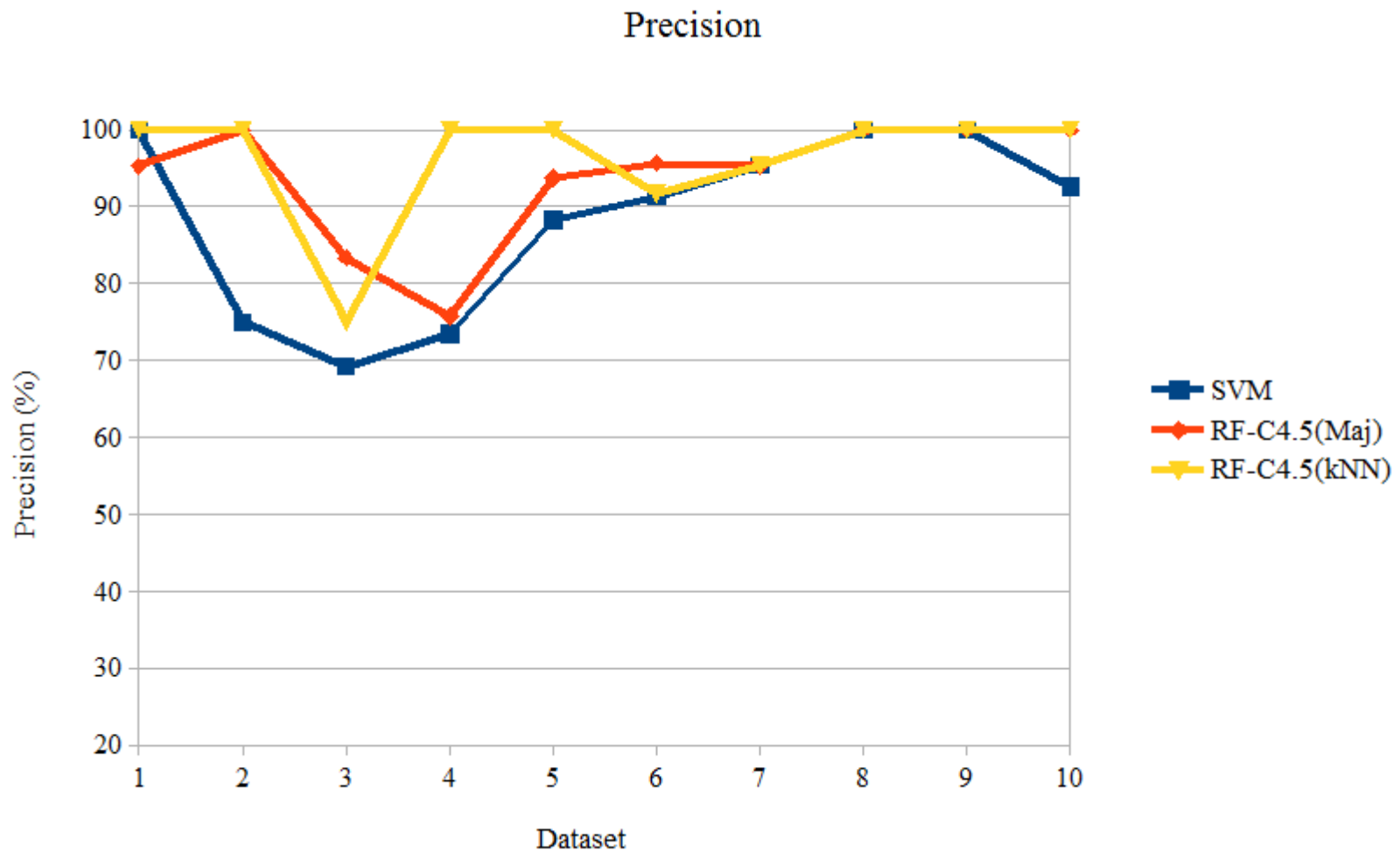
- Precision, Recall, F1, Accuracy

■ Các tham số

- Xây dựng 200 cây, sử dụng ngẫu nhiên 1/3 dim.
- Luật cục bộ: 1 láng giếng
- SVM: hàm nhân tuyến tính với $cost = 100000$

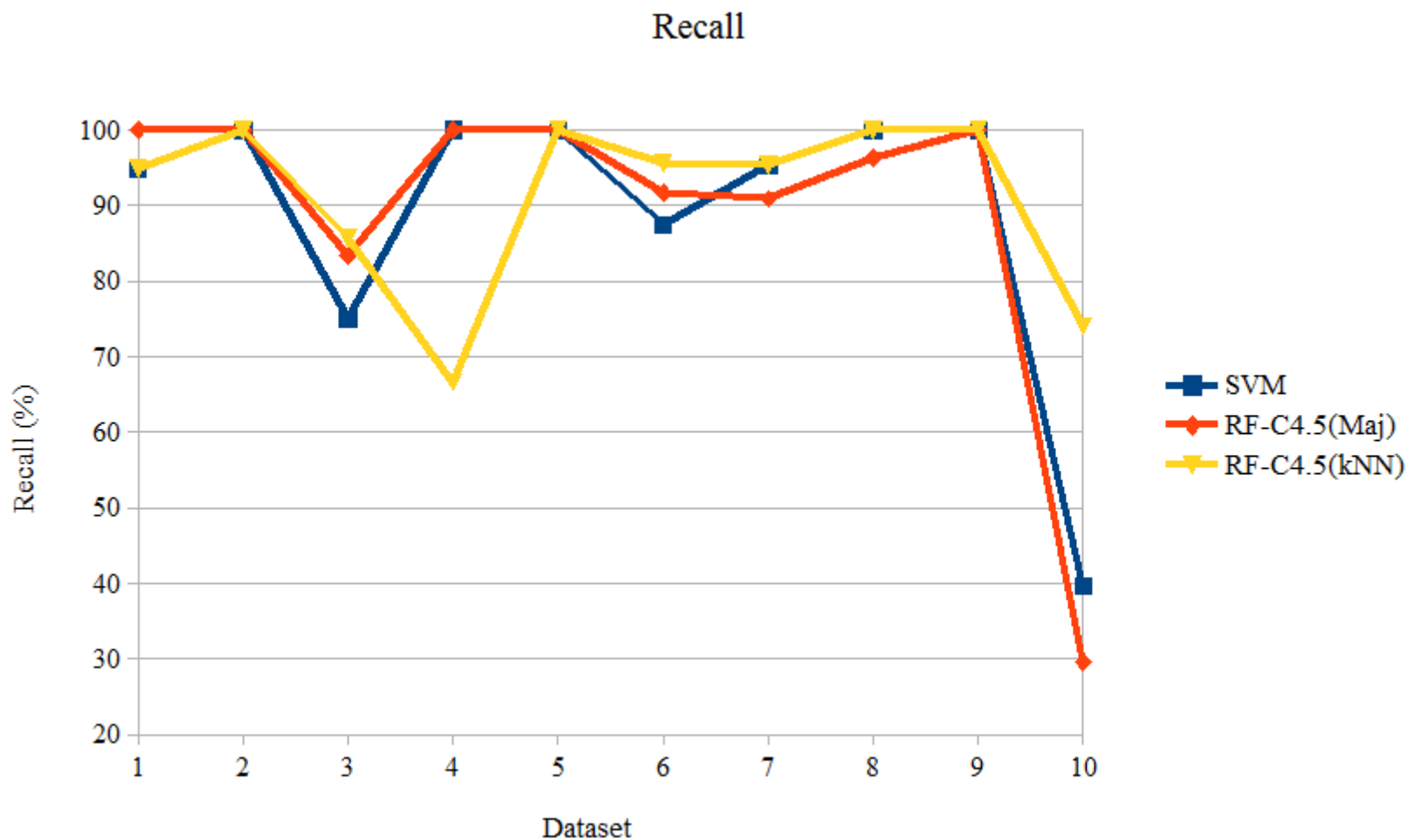
Kết quả thực nghiệm

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- **Kết quả thực nghiệm**
- Kết luận và hướng phát triển



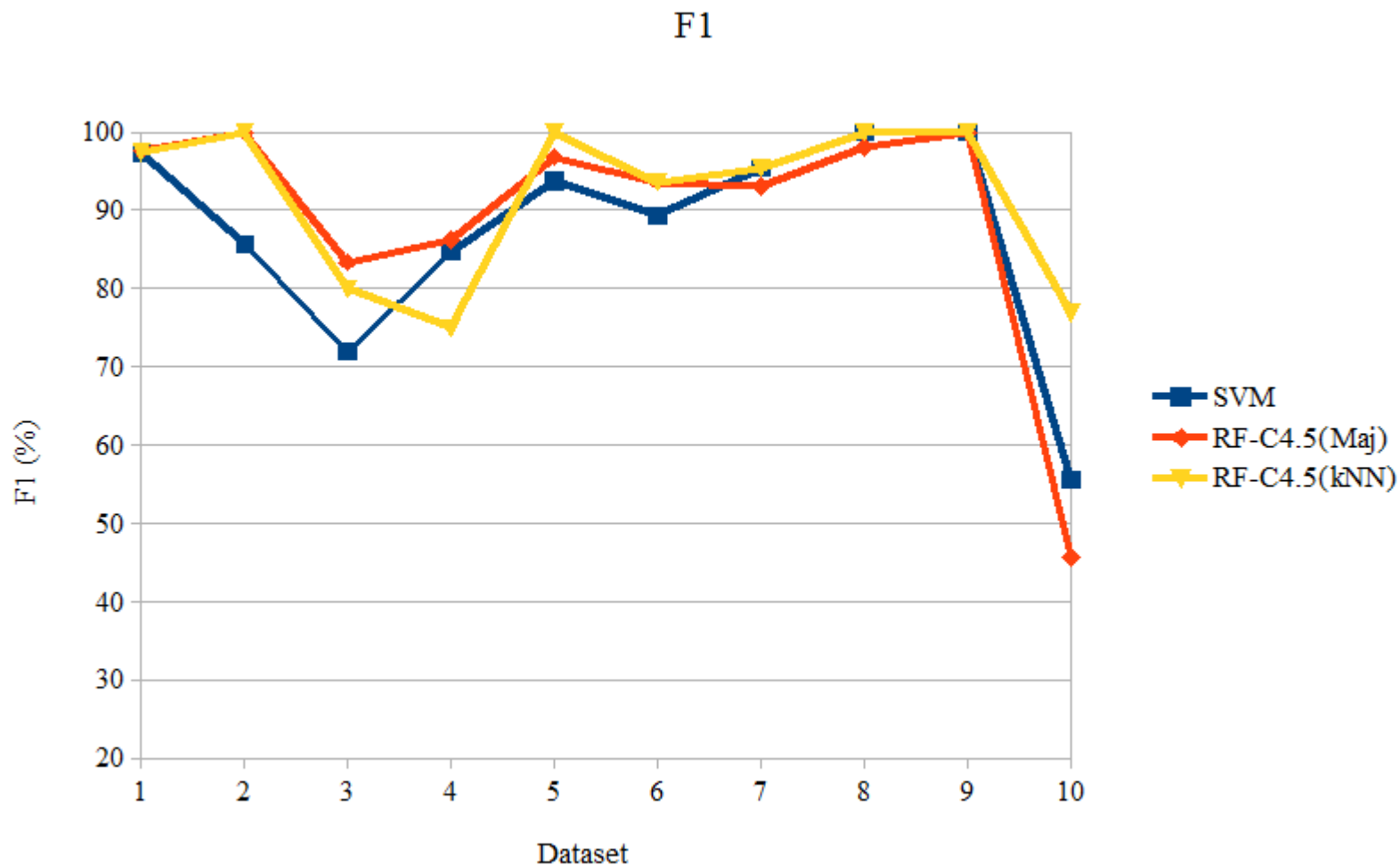
Kết quả thực nghiệm

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- **Kết quả thực nghiệm**
- Kết luận và hướng phát triển



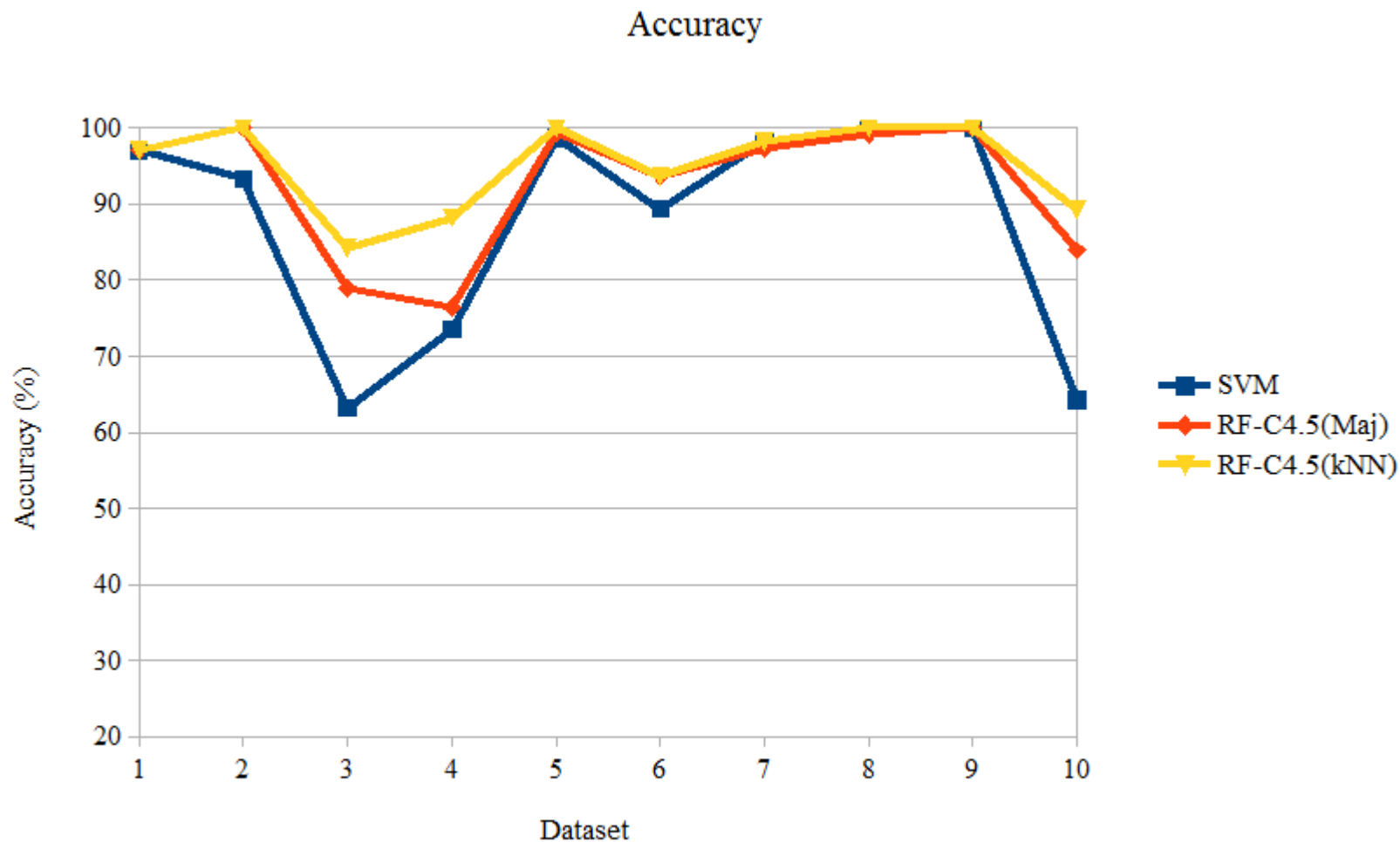
Kết quả thực nghiệm

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- **Kết quả thực nghiệm**
- Kết luận và hướng phát triển



Kết quả thực nghiệm

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- **Kết quả thực nghiệm**
- Kết luận và hướng phát triển



Nội dung

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển

Giới thiệu

- Giới thiệu
- Giải thuật rừng ngẫu nhiên
- Luật gán nhãn cục bộ
- Kết quả thực nghiệm
- Kết luận và hướng phát triển

- Rừng ngẫu nhiên với luật gán nhãn cục bộ
 - Thay thế luật gán nhãn bình chọn số đông
 - Giảm thành phần lỗi bias của cây
 - Cải tiến hiệu quả phân lớp dữ liệu có số chiều lớn
 - Đánh giá trên 10 tập dữ liệu γ -sinh
- Hướng phát triển
 - Luật gán nhãn cục bộ khác
 - Mô hình cục bộ
 - Xử lý song song

Cám ơn Quý Đại Biểu đã lắng nghe!!!

