

# Giải thuật ArcX4-iODT cho phân lớp dữ liệu gen

TS. Đỗ Thanh Nghị

Khoa CNTT-TT, ĐH. Cần Thơ

Số 1 Lý Tự Trọng, Ninh Kiều, Cần Thơ

Email: [dtngghi@cit.ctu.edu.vn](mailto:dtngghi@cit.ctu.edu.vn)

Site: <http://www.cit.ctu.edu.vn/~dtngghi>

Ngày 12 tháng 8 năm 2011



# Nội dung

Giới thiệu

Giải thuật ArcX4-iODT

Kết quả thực nghiệm

Kết luận, hướng phát triển



# Top 10 vấn đề khó của khai mỏ dữ liệu (Yang & Wu, 06)

## Phân lớp dữ liệu số chiều lớn

- ▶ dữ liệu hoàn toàn dễ tách biệt
- ▶ có nhiều lựa chọn mô hình
- ▶ mô hình tốt: dự báo tốt trong tương lai (tập kiểm tra)

## Phân lớp dữ liệu gien

- ▶ số phần tử rất ít: vài chục hoặc vài trăm
- ▶ số thuộc tính (chiều) rất lớn: hàng ngàn đến hàng chục ngàn
- ▶ mô hình học thường cho độ chính xác 100% trong tập học
- ▶ nhưng dự báo tập kiểm tra không tốt



# ArcX4-iODT cho phân lớp dữ liệu gen

## Giải thuật ArcX4-iODT

- ▶ Rừng cây ngẫu nhiên xiên phân (Do et al., 09)
- ▶ phân hoạch đơn thuộc tính  $\Rightarrow$  phân hoạch đa thuộc tính
- ▶ xử lý tốt dữ liệu có số chiều lớn
- ▶ điều khiển độ phức tạp của mô hình ArcX4-iODT
- ▶ mô hình mạnh với nhiễu
- ▶ kết quả thực nghiệm trên dữ liệu gen (Jinyan & Huiqing, 02)
- ▶ ArcX4-iODT tốt hơn rừng ngẫu nhiên RF-C4.5 (Breiman, 01), (Quinlan, 93) và máy học véc-tơ hỗ trợ SVM (Vapnik, 95)

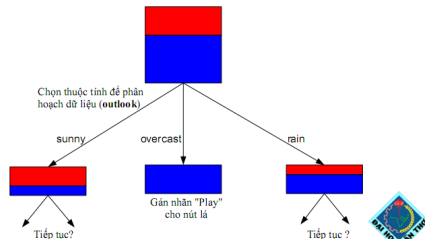


# Cây quyết định C4.5 (Quinlan, 93), CART (Breiman, 84)

## Giải thuật học

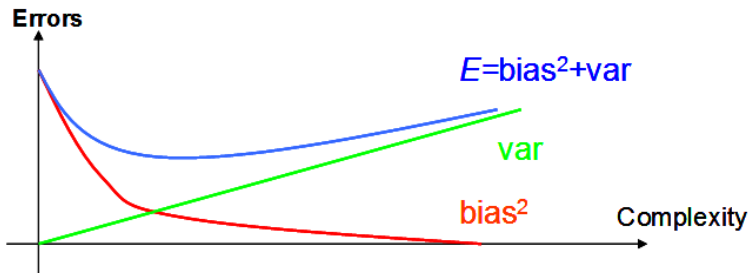
- ▶ xây dựng cây: **hàm phân hoạch**
- ▶ cắt nhánh: tránh học vệt

| outlook  | temp. | hum. | windy | Play, Don't Play |
|----------|-------|------|-------|------------------|
| sunny    | 85    | 85   | false | Don't Play       |
| sunny    | 80    | 90   | true  | Don't Play       |
| overcast | 83    | 78   | false | Play             |
| rain     | 70    | 96   | false | Play             |
| rain     | 68    | 80   | false | Play             |
| rain     | 65    | 70   | true  | Don't Play       |
| ...      | ...   | ...  | ...   | ...              |
| ...      | ...   | ...  | ...   | ...              |

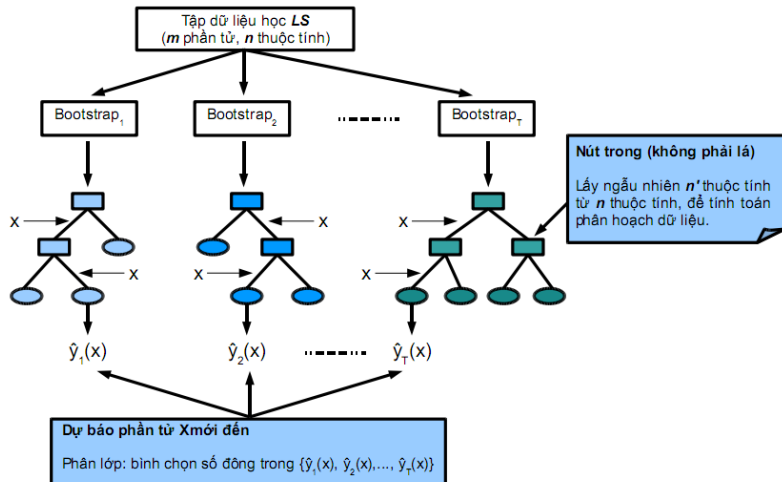


# Phân tích hiệu quả của giải thuật học (Breiman, 1996)

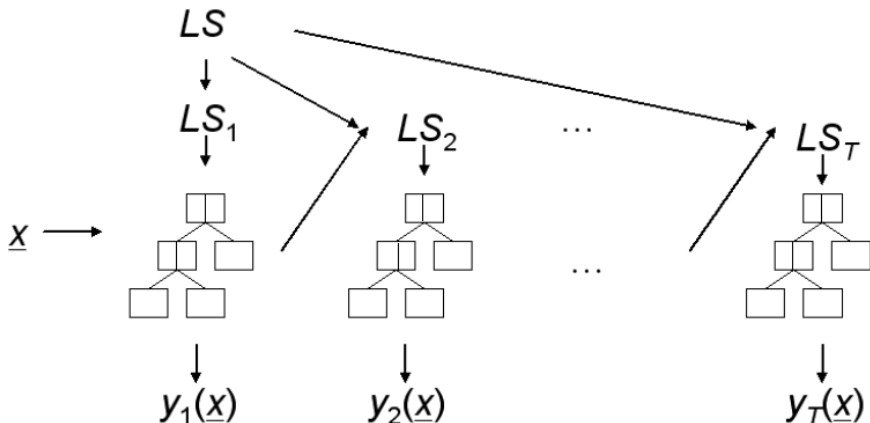
Lỗi của giải thuật học:  $error = bias^2 + variance$



# Bagging (Breiman, 96), Random forests (Breiman, 01) : giảm variance

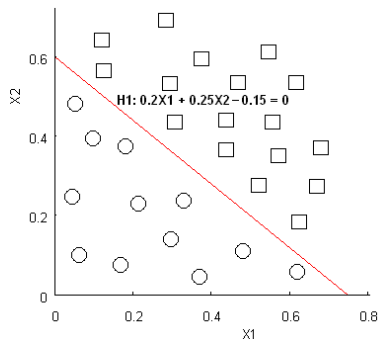
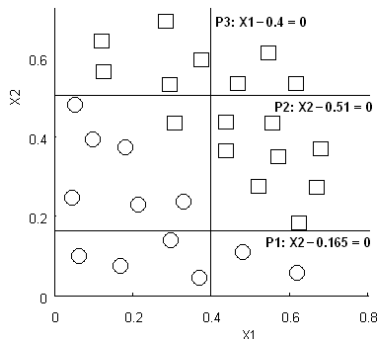


# ArcX4 (Breiman, 97), AdaBoost (Freund & Schapire, 95) : giảm bias



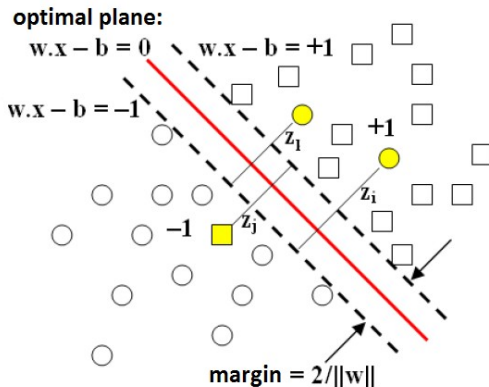
# Cây ngẫu nhiên xiên phân (ODT)

Kết hợp nhiều chiều xử lý dữ liệu có số chiều lớn, phụ thuộc

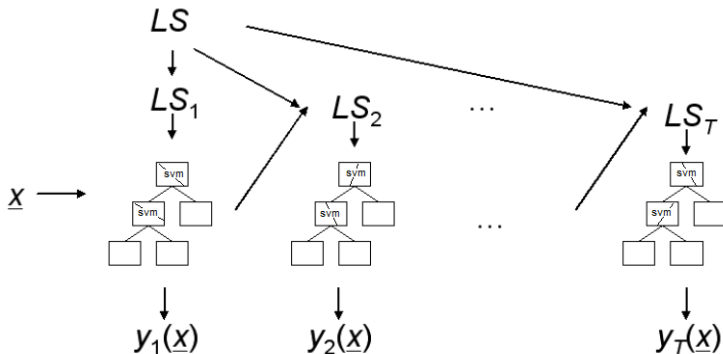


# ODT: siêu phẳng phân hoạch tối ưu

## Máy học véc-tơ hỗ trợ - SVM



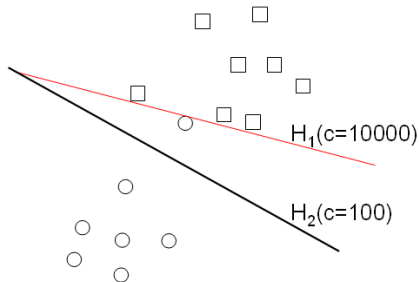
# Giải thuật ArcX4-iODT



# Giải thuật ArcX4-iODT

## Điều khiển độ phức tạp

- ▶ giới hạn độ sâu cây, sử dụng tham số điều chỉnh lề của SVM
- ▶ hiệu quả: xử lý dữ liệu số chiều lớn, khả năng chịu đựng nhiễu



## Thực nghiệm trên 10 tập dữ liệu gen

Lấy từ Bio-medical Dataset Repository (Jinyan & Huiqing, 02)

- ▶ số phần tử ít, số chiều lớn

| ID | Ensemble                                    | #ind | #dim  | Classes                   | Validation |
|----|---------------------------------------------|------|-------|---------------------------|------------|
| 1  | Colon Tumor                                 | 62   | 2000  | tumor, normal             | loo        |
| 2  | ALL-AML-Leukemia                            | 72   | 7129  | ALL, AML                  | trn-tst    |
| 3  | Breast Cancer                               | 97   | 24481 | relapse, non-relapse      | trn-tst    |
| 4  | Prostate Cancer                             | 136  | 12600 | cancer, normal            | trn-tst    |
| 5  | Lung Cancer BrighamWomen Hospital           | 181  | 12533 | cancer, normal            | trn-tst    |
| 6  | Central Nervous System                      | 60   | 7129  | positive, negative        | loo        |
| 7  | Diffuse Large B-Cell Lymphoma               | 47   | 4026  | germinal, activated       | loo        |
| 8  | *Subtypes of Acute Lymphoblastic (Hyperdip) | 327  | 12558 | Hyperdip, rest            | trn-tst    |
| 9  | *Subtypes of Acute Lymphoblastic (TEL-AML1) | 327  | 12558 | TEL-AML1, rest            | trn-tst    |
| 10 | *Subtypes of Acute Lymphoblastic (Others)   | 327  | 12558 | Others, diagnostic groups | trn-tst    |



# Chuẩn bị chương trình

## Chương trình dùng cho thực nghiệm

- ▶ giải thuật ArcX4-iODT: C/C++, ATLAS/LAPACK
- ▶ sử dụng chương trình RF-C4.5 từ (Do et al., 08)
- ▶ máy học véc-tơ hỗ trợ SVM: LibSVM (Chang & Lin, 01)



# Tiêu chí đánh giá

Bảng: Confusion matrix

|            |          | Prediction |          |
|------------|----------|------------|----------|
|            |          | positive   | negative |
| True class | positive | tp         | fn       |
|            | negative | fp         | tn       |

## Recall, Precision, F1, Accuracy

- ▶  $Recall = \frac{tp}{tp+fn} = \frac{tp}{pos}$
- ▶  $Precision = \frac{tp}{tp+fp}$
- ▶  $F1 = \frac{2 \times prec \times rec}{prec+rec}$
- ▶  $Accuracy = \frac{tp+tn}{tp+fn+tn+fp} = \frac{tp+tn}{pos+neg}$

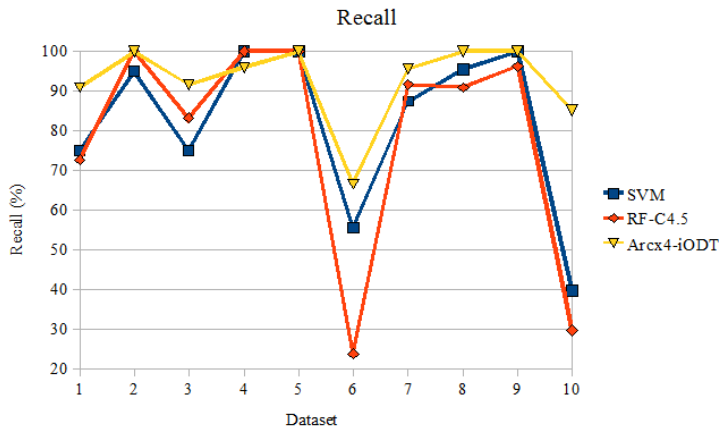


# Kết quả phân lớp của LibSVM, RF-C4.5 và ArcX4-iODT

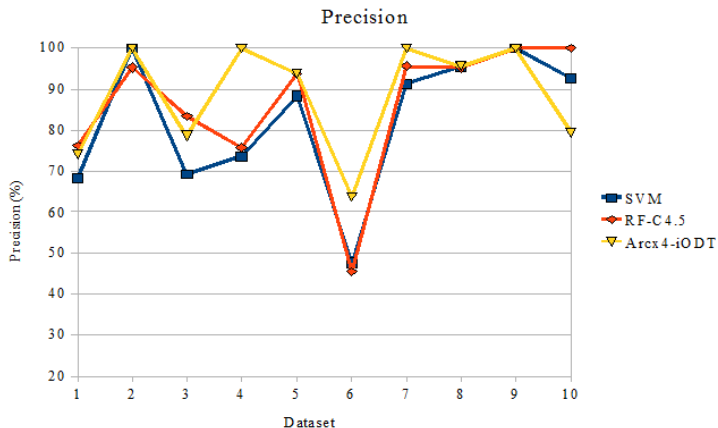
| ID | Precision  |              |              | Recall     |            |              | F1-measure |         |              |
|----|------------|--------------|--------------|------------|------------|--------------|------------|---------|--------------|
|    | Lib-SVM    | RF-C4.5      | Arcx4-iODT   | Lib-SVM    | RF-C4.5    | Arcx4-iODT   | Lib-SVM    | RF-C4.5 | Arcx4-iODT   |
| 1  | 68.18      | <b>76.19</b> | 74,07        | 75.00      | 72.73      | <b>90,91</b> | 71.43      | 74.42   | <b>81,63</b> |
| 2  | <b>100</b> | 95.24        | <b>100</b>   | 95.00      | <b>100</b> | <b>100</b>   | 97.44      | 97.56   | <b>100</b>   |
| 3  | 69.23      | <b>83.33</b> | 78,57        | 75.00      | 83.33      | <b>91,67</b> | 72.00      | 83.33   | <b>84,62</b> |
| 4  | 73.53      | 75.76        | <b>100</b>   | <b>100</b> | <b>100</b> | 96           | 84.75      | 86.21   | <b>97,98</b> |
| 5  | 88.26      | <b>93.75</b> | <b>93,75</b> | 100        | 100        | <b>100</b>   | 93.75      | 96.77   | <b>96,77</b> |
| 6  | 47.62      | 45.46        | <b>63,64</b> | 55.56      | 23.81      | <b>66,67</b> | 51.28      | 31.25   | <b>65,11</b> |
| 7  | 91.30      | 95.65        | <b>100</b>   | 87.50      | 91.67      | <b>95,65</b> | 89.36      | 93.62   | <b>97,78</b> |
| 8  | 95.46      | 95.24        | <b>95,65</b> | 95.46      | 90.91      | <b>100</b>   | 95.46      | 93.02   | <b>97,78</b> |
| 9  | <b>100</b> | <b>100</b>   | <b>100</b>   | <b>100</b> | 96.30      | <b>100</b>   | <b>100</b> | 98.11   | <b>100</b>   |
| 10 | 92.59      | <b>100</b>   | 79,31        | 39.68      | 29.63      | <b>85,19</b> | 55.56      | 45.71   | <b>82,14</b> |



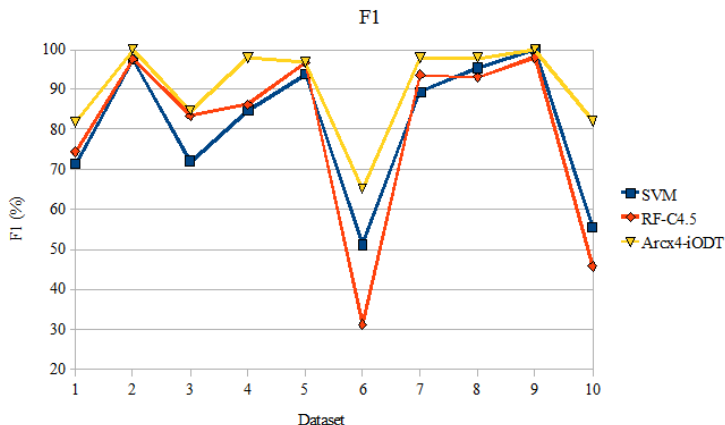
## So sánh kết quả theo tiêu chí Recall



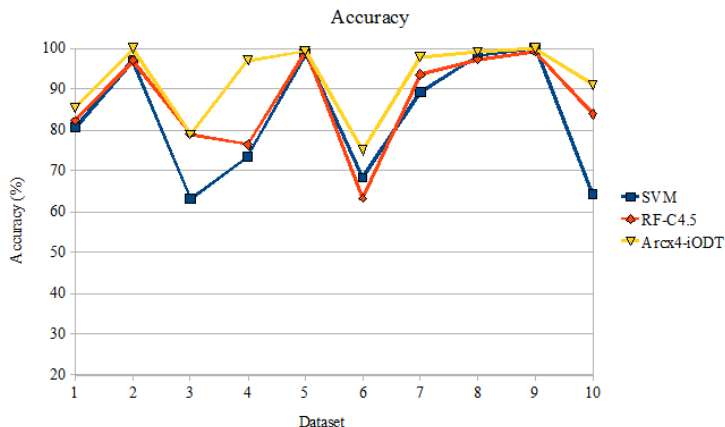
## So sánh kết quả theo tiêu chí Precision



## So sánh kết quả theo tiêu chí F1



## So sánh kết quả theo tiêu chí Accuracy



# Kết luận

## Giải thuật học mới ArcX4-iODT

- ▶ xây dựng tuần tự tập cây ngẫu nhiên xiên phân
- ▶ siêu phẳng phân hoạch tối ưu SVM
- ▶ xây dựng cây tập trung vào khắc phục lỗi từ các mô hình xây dựng trước đó
- ▶ điều khiển độ phức tạp : độ sâu cây, SVM
- ▶ xử lý dữ liệu số chiều lớn, khả năng chịu đựng nhiễu
- ▶ kết quả phân lớp 10 tập dữ liệu gen
- ▶ ArcX4-iODT chính xác hơn RF-C4.5 và SVM



# Hướng phát triển<sup>2</sup>

## Tiếp tục nghiên cứu

- ▶ phân lớp đa lớp, hồi quy
- ▶ cung cấp thêm kết quả thực nghiệm
- ▶ so sánh với các giải thuật phân lớp khác



## Cám ơn & câu hỏi thảo luận .....

