

# Giải thuật rừng ngẫu nhiên xiên phân (RF-ODT) cho phân lớp dữ liệu có số chiều lớn

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

ĐHBK Tp.HCM, 27/03/2014



# Nội dung

Giới thiệu

Rừng ngẫu nhiên xiên phân (RF-ODT)

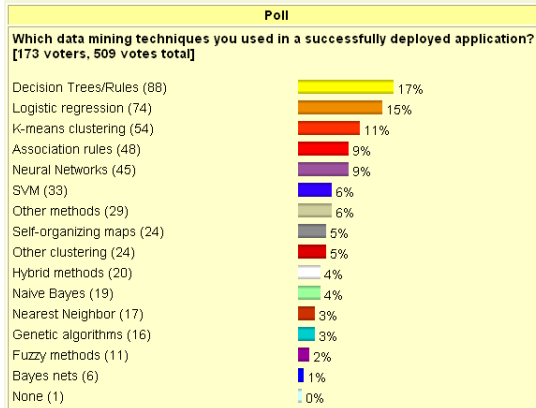
Kết quả thực nghiệm

Hướng phát triển



# Cây quyết định: top 10 giải thuật khai mở dữ liệu hiệu quả (Wu et al., 08)

## KDnuggets : Polls : Deployed data mining techniques



# Cây quyết định: top 10 giải thuật khai mỏ dữ liệu hiệu quả (Wu et al., 08)

## Ưu điểm của cây quyết định

- ▶ thời gian huấn luyện nhanh
- ▶ xử lý được dữ liệu liên tục, rời rạc
- ▶ mô hình dễ diễn dịch (luật **if ... then ...**)
- ▶ kết quả tốt cho phân lớp, hồi quy

## Khuyết điểm của cây quyết định

- ▶ hàm phân hoạch: đơn biến
- ▶ không hiệu quả cho vấn đề phức tạp: số chiều rất lớn (nhiều), mất cân bằng, phi tuyến



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

## Phân lớp dữ liệu có số chiều lớn

- ▶  $\# \text{individus} \ll \# \text{dimensions}$
- ▶ vài trăm phần tử với hàng nghìn chiều
- ▶ dữ liệu hoàn toàn dễ tách biệt
- ▶ có nhiều lựa chọn mô hình
- ▶ mô hình học thường cho độ chính xác 100% trong tập học
- ▶ nhưng dự báo tập kiểm tra không tốt
- ▶ mô hình tốt: dự báo tốt trong tương lai



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

## Phân lớp dữ liệu không cân bằng

- ▶ lớp quan tâm chiếm tỷ lệ rất thấp (lớp dương, thiểu số)
- ▶ lớp còn lại chiếm tỷ lệ rất cao (lớp âm, đa số)
- ▶ giải thuật phân lớp: độ chính xác toàn cục
- ▶ không dự đoán được lớp thiểu số (dương)
- ▶ ví dụ: tỷ lệ lớp dương là 5% và lớp âm là 95%
- ▶ mô hình học **M** không dự đoán được lớp dương, chỉ luôn dự đoán là lớp âm, thì độ chính xác toàn cục vẫn là 95%



# Rừng ngẫu nhiên xiên phân (RF-ODT)

## Phân lớp dữ liệu có số chiều lớn

- ▶ xem xét sự phụ thuộc của các biến
- ▶ hàm phân hoạch: đa biến (xiên)
- ▶ phân lớp rất hiệu quả dữ liệu có số chiều lớn

## Mở rộng cho các vấn đề

- ▶ phân lớp dữ liệu không cân bằng
- ▶ phân lớp đa lớp
- ▶ tăng khả năng chịu đựng nhiễu

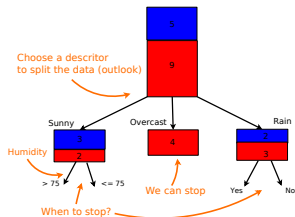


# Cây quyết định (Breiman et al., 84), (Quinlan, 93)

## Học phân lớp dữ liệu

- ▶ xây dựng cây: phân hoạch dữ liệu (gini, entropy)
- ▶ gán nhãn ở nút lá: bình chọn số đông
- ▶ cắt nhánh: tránh học vẹt

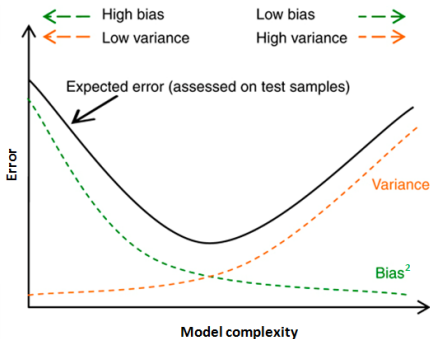
| outlook  | temp. | hum. | windy | Play, Don't Play |
|----------|-------|------|-------|------------------|
| sunny    | 85    | 85   | false | Don't Play       |
| sunny    | 80    | 90   | true  | Don't Play       |
| overcast | 83    | 78   | false | Play             |
| rain     | 70    | 96  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





# Phân tích hiệu quả của giải thuật học (Breiman, 96)

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



# Tập hợp mô hình

## Nguyên lý

- ▶ giảm thành phần lỗi bias và/hoặc variance
- ▶ chú ý: bias và variance tỷ lệ nghịch
- ▶ kết hợp các mô hình học yếu (không quá tệ) và đa dạng
- ▶ Bagging (Breiman, 96)
- ▶ Boosting (Freund & Schapire, 95), (Breiman, 97)
- ▶ Rừng ngẫu nhiên (Breiman, 01)



# Rừng ngẫu nhiên (Breiman, 01)

## Nguyên lý

- ▶ xây dựng tập cây quyết định đa dạng, không cắt nhánh
- ▶ không cắt nhánh: giữ bias thấp
- ▶ tính đa dạng (ngẫu nhiên): giảm variance

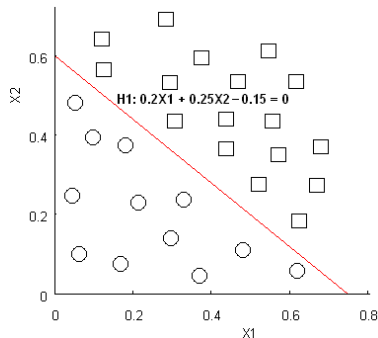
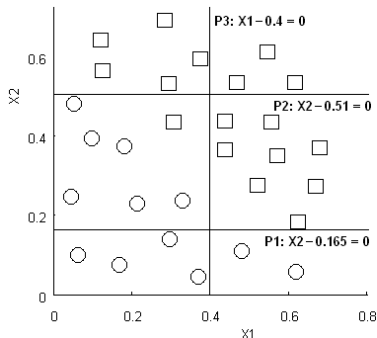
## Hiệu năng

- ▶ huấn luyện: nhanh, đơn giản
- ▶ kết quả tốt khi so với Boosting, SVM
- ▶ khả năng chịu đựng nhiễu cao



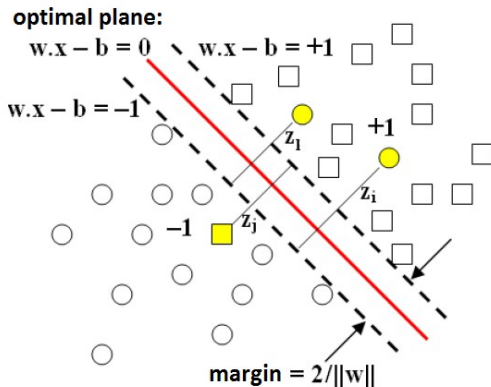
## Cây xiên phân (ODT)

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

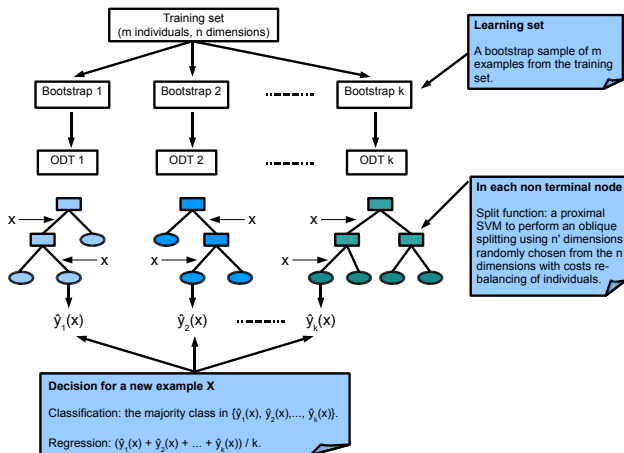


# ODT: siêu phẳng phân hoạch tối ưu

Máy học véc-tơ hỗ trợ - SVM (Vapnik, 95)



# Rừng ngẫu nhiên xiên phân RF-ODT (Do et al., 09)



# Đánh giá hiệu quả của giải thuật RF-ODT

## So sánh RF-ODT với RF-C4.5 và Lib-SVM

- ▶ cài đặt giải thuật RF-ODT: C/C++, ATLAS, BLAS/LAPACK
- ▶ cài đặt giải thuật RF-C4.5: C
- ▶ sử dụng dụng cụ LibSVM (Chang & Lin, 01)

## Tập dữ liệu

- ▶ số chiều lớn: 20 tập dữ liệu y sinh (Jinyan & Huiqing, 02)
- ▶ đa chiều: 10 tập dữ liệu từ kho UCI (Asuncion & Newman, 07)



# Tập dữ liệu y sinh có số chiều lớn

**Bảng:** Mô tả 20 tập dữ liệu y sinh

| 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  | *MLL-Leukemia                                   | 72    | 12582  | MLL, rest                 | trn-tst    |
| 4  | Breast Cancer                                   | 97    | 24481  | relapse, non-relapse      | trn-tst    |
| 5  | Duke Breast Cancer                              | 42    | 7129   | cancer, normal            | loo        |
| 6  | Prostate Cancer                                 | 136   | 12600  | cancer, normal            | trn-tst    |
| 7  | Lung Cancer BrighamWomen Hospital               | 181   | 12533  | cancer, normal            | trn-tst    |
| 8  | Central Nervous System                          | 60    | 7129   | positive, negative        | loo        |
| 9  | Translation Initiation Site                     | 13375 | 927    | positive, negative        | 10-fold    |
| 10 | Ovarian Cancer NCI-PBSII                        | 253   | 15154  | cancer, normal            | loo        |
| 11 | Diffuse Large B-Cell Lymphoma                   | 47    | 4026   | germinal, activated       | loo        |
| 12 | *Subtypes of Acute Lymphoblastic (Hyperdip)     | 327   | 12558  | Hyperdip, rest            | trn-tst    |
| 13 | *Subtypes of Acute Lymphoblastic (TEL-AML1)     | 327   | 12558  | TEL-AML1, rest            | trn-tst    |
| 14 | *Subtypes of Acute Lymphoblastic (T-ALL)        | 327   | 12558  | TEL-ALL, rest             | trn-tst    |
| 15 | *Subtypes of Acute Lymphoblastic (Others)       | 327   | 12558  | Others, diagnostic groups | trn-tst    |
| 16 | Ovarian Cancer NCI-QStar                        | 216   | 373410 | cancer, normal            | loo        |
| 17 | Lung Cancer DanaFarberCancer Institute (ADEN)   | 203   | 12600  | ADEN, rest                | loo        |
| 18 | Lung Cancer DanaFarberCancer Institute (SQUA)   | 203   | 12600  | SQUA, rest                | loo        |
| 19 | Lung Cancer DanaFarberCancer Institute (COLD)   | 203   | 12600  | COLD, rest                | loo        |
| 20 | Lung Cancer DanaFarberCancer Institute (normal) | 203   | 12600  | normal, rest              | loo        |





# Tiêu chí đánh giá

Bảng: Confusion matrix

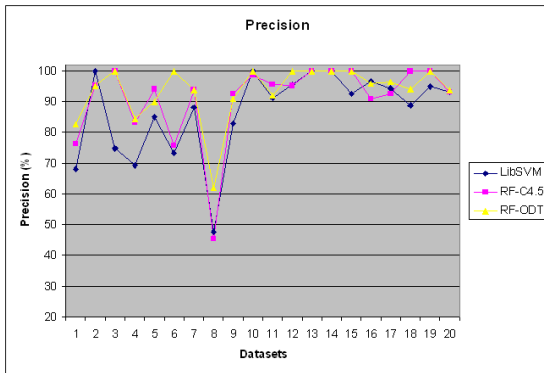
|            |          | Prediction |          |
|------------|----------|------------|----------|
|            |          | positive   | negative |
| True class | positive | tp         | fn       |
|            | negative | fp         | tn       |

## Recall, Precision, F1, Accuracy

- ▶  $Recall = \frac{tp}{tp+fn} = \frac{tp}{pos}$
- ▶  $Precision = \frac{tp}{tp+fp}$
- ▶  $F1 = \frac{2 \times prec \times rec}{prec+rec}$
- ▶  $Accuracy = \frac{tp+tn}{tp+fn+tn+fp} = \frac{tp+tn}{pos+neg}$
- ▶ thời gian huấn luyện: PC Pentium 2,4 GHz, 1 GB RAM, Linux



# Kết quả phân lớp (Precision)



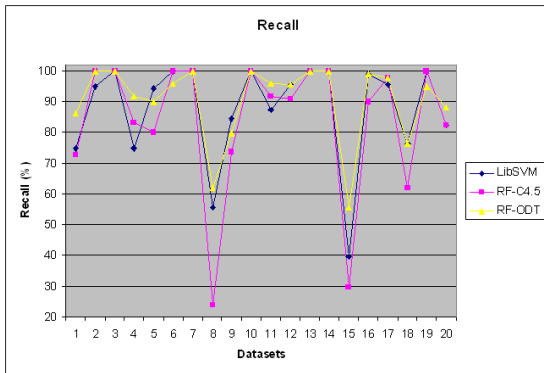
## Kết quả phân lớp (Precision)

Bảng: So sánh kết quả phân lớp (Precision)

| Precision          | LibSVM | RF-C4.5 | RF-ODT | RF-ODT vs LibSVM | RF-ODT vs RF-C4.5 |
|--------------------|--------|---------|--------|------------------|-------------------|
| mean               | 86.88  | 91.14   | 93.56  | <b>6.68</b>      | 2.42              |
| standard deviation | 13.80  | 13.01   | 9.10   | 8.44             | 6.95              |
| student ratio      |        |         |        | 3.54             | 1.56              |
| p-value            |        |         |        | <b>0.0022</b>    | 0.1357            |
| result of RF-ODT   |        |         |        | gain**           |                   |
| RF-ODT win         |        |         |        | 15               | 9                 |
| RF-ODT tie         |        |         |        | 3                | 7                 |
| RF-ODT defeat      |        |         |        | 2                | 4                 |
| p-value            |        |         |        | <b>0.0012</b>    | 0.1334            |
| result of RF-ODT   |        |         |        | gain**           |                   |



# Kết quả phân lớp (Recall)



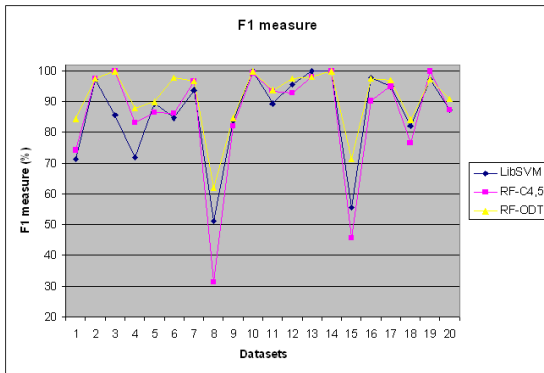
## Kết quả phân lớp (Recall)

Bảng: So sánh kết quả phân lớp (Recall)

| Recall             | LibSVM | RF-C4.5 | RF-ODT | RF-ODT vs LibSVM | RF-ODT vs RF-C4.5 |
|--------------------|--------|---------|--------|------------------|-------------------|
| mean               | 87.77  | 83.90   | 90.45  | 2.68             | <b>6.55</b>       |
| standard deviation | 16.62  | 22.59   | 12.89  | 6.38             | 10.41             |
| student ratio      |        |         |        | 1.88             | 2.81              |
| p-value            |        |         |        | 0.0759           | <b>0.0111</b>     |
| result of RF-ODT   |        |         |        |                  | gain*             |
| RF-ODT win         |        |         |        | 8                | 11                |
| RF-ODT tie         |        |         |        | 8                | 7                 |
| RF-ODT defeat      |        |         |        | 4                | 2                 |
| p-value            |        |         |        | 0.1938           | 0.0112            |
| result of RF-ODT   |        |         |        |                  | gain*             |



# Kết quả phân lớp (F1)



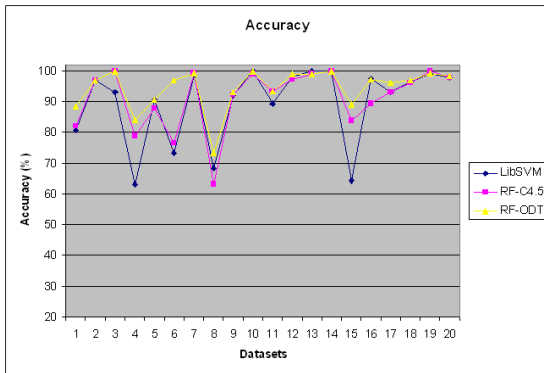
## Kết quả phân lớp (F1)

Bảng: So sánh kết quả phân lớp (F1)

| F1-measure         | LibSVM | RF-C4.5 | RF-ODT | RF-ODT vs LibSVM | RF-ODT vs RF-C4.5 |
|--------------------|--------|---------|--------|------------------|-------------------|
| mean               | 86.50  | 85.88   | 91.49  | <b>5.00</b>      | <b>5.62</b>       |
| standard deviation | 14.15  | 18.13   | 10.14  | 6.20             | 8.60              |
| student ratio      |        |         |        | 3.61             | 2.92              |
| p-value            |        |         |        | <b>0.0019</b>    | <b>0.0088</b>     |
| result of RF-ODT   |        |         |        | gain**           | gain*             |
| RF-ODT win         |        |         |        | 15               | 14                |
| RF-ODT tie         |        |         |        | 3                | 5                 |
| RF-ODT defeat      |        |         |        | 2                | 1                 |
| p-value            |        |         |        | 0.0012           | 0.0005            |
| result of RF-ODT   |        |         |        | gain**           | gain***           |



# Kết quả phân lớp (Accuracy)





# Kết quả phân lớp (Accuracy)

**Bảng:** So sánh kết quả phân lớp(Accuracy)

| Accuracy           | LibSVM | RF-C4.5 | RF-ODT | RF-ODT vs LibSVM | RF-ODT vs RF-C4.5 |
|--------------------|--------|---------|--------|------------------|-------------------|
| mean               | 89.70  | 91.38   | 94.64  | 4.94             | 3.27              |
| standard deviation | 12.53  | 9.91    | 6.73   | 8.27             | 5.08              |
| student ratio      |        |         |        | 2.67             | 2.87              |
| p-value            |        |         |        | <b>0.0151</b>    | <b>0.0097</b>     |
| result of RF-ODT   |        |         |        | gain*            | gain*             |
| RF-ODT win         |        |         |        | 13               | 13                |
| RF-ODT tie         |        |         |        | 5                | 6                 |
| RF-ODT defeat      |        |         |        | 2                | 1                 |
| p-value            |        |         |        | 0.0037           | 0.0009            |
| result of RF-ODT   |        |         |        | gain**           | gain**            |



# Tập dữ liệu từ kho UCI

**Bảng:** Mô tả 10 tập dữ liệu đa chiều từ kho UCI

| ID | Ensemble      | #Ind  | #Dim | Classes   | Validation |
|----|---------------|-------|------|-----------|------------|
| 21 | Bupa          | 345   | 6    | 1, 2      | 10-fold    |
| 22 | Breast cancer | 569   | 30   | M, B      | 10-fold    |
| 23 | Pima          | 768   | 8    | 1, 2      | 10-fold    |
| 24 | Segment       | 2310  | 19   | 1, rest   | 10-fold    |
| 25 | Spambase      | 4601  | 57   | spam, non | 10-fold    |
| 26 | Optdigits     | 5620  | 64   | 9, rest   | trn-tst    |
| 27 | Satimage      | 6435  | 36   | 4, rest   | trn-tst    |
| 28 | Pendigits     | 10992 | 16   | 0, rest   | trn-tst    |
| 29 | Letters       | 20000 | 16   | A, rest   | 3-fold     |
| 30 | Shuttle       | 58000 | 9    | 1, rest   | trn-tst    |



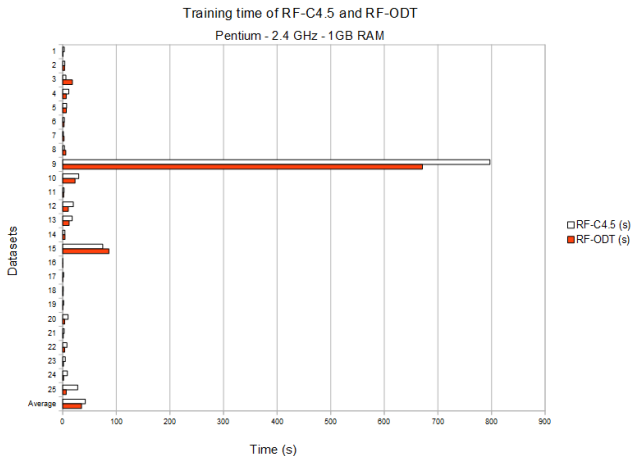
# Kết quả phân lớp

**Bảng:** So sánh kết quả phân lớp (Accuracy)

|                    |                    |
|--------------------|--------------------|
| Accuracy           | RF-ODT vs RF-C4.5  |
| mean               | <b>0.69</b>        |
| standard deviation | 3.10               |
| student ratio      | 0.70               |
| p-value            | <b>0.5001</b>      |
| result of RF-ODT   | non significant    |
| RF-ODT win         | 8                  |
| RF-ODT tie         | 0                  |
| RF-ODT defeat      | 2                  |
| p-value            | 0.1094             |
| result of RF-ODT   | almost significant |



# Thời gian huấn luyện (secs)



# Thời gian huấn luyện (secs)

## Phân lớp tập dữ liệu có nhiều phần tử

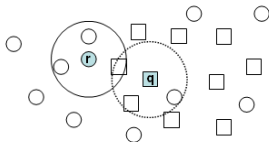
- ▶ Forest Cover Types từ kho UCI
- ▶ 500k ind., 54 dim.
- ▶ LibSVM: huấn luyện hơn 1 tháng
- ▶ RF-ODT: 801,61 giây
- ▶ RF-ODT nhanh hơn RF-C4.5 khoảng 22 lần
- ▶ RF-ODT chính xác hơn RF-C4.5: 0.41%



# Rừng ngẫu nhiên xiên phân (RF-ODT)

## Phân lớp dữ liệu không cân bằng

- ▶ thay đổi phân bố lực lượng mẫu: lấy mẫu giảm
- ▶ chi phí khác nhau: lỗi của lớp thiểu số cao hơn
- ▶ thay thế luật quyết định: luật cục bộ



# Phân lớp dữ liệu không cân bằng

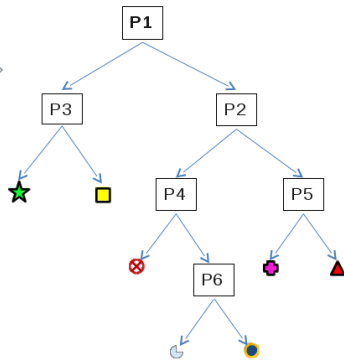
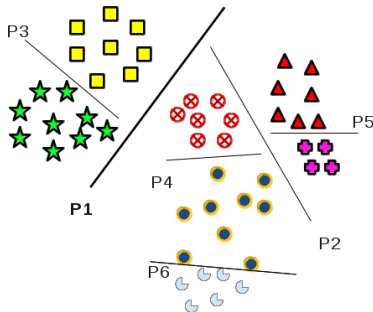
**Bảng:** So sánh kết quả phân lớp dữ liệu không cân bằng

| ID | Ensemble      | Methods | Acc. min class | Acc. maj class | Accuracy     | F1           | AUC          |
|----|---------------|---------|----------------|----------------|--------------|--------------|--------------|
| 1  | 20-newsgrp    | RF-C4.5 | 76,83          | <b>99,11</b>   | 97,97        | 79,51        | <b>99,65</b> |
|    |               | RF-ODT  | <b>99,12</b>   | 98,32          | <b>98,36</b> | <b>86,11</b> | 99,46        |
| 2  | Breast cancer | RF-C4.5 | 92,28          | 97,74          | 95,61        | 94,13        | 98,19        |
|    |               | RF-ODT  | <b>95,75</b>   | <b>98,56</b>   | <b>97,54</b> | <b>96,70</b> | <b>99,64</b> |
| 3  | Letters       | RF-C4.5 | 92,22          | <b>99,97</b>   | 99,65        | 95,59        | 99,97        |
|    |               | RF-ODT  | <b>96,30</b>   | <b>99,97</b>   | <b>99,82</b> | <b>97,74</b> | <b>100</b>   |
| 4  | Abalone       | RF-C4.5 | 23,93          | <b>98,48</b>   | <b>92,01</b> | 33,07        | 87,64        |
|    |               | RF-ODT  | <b>53,02</b>   | 93,75          | 90,22        | <b>48,41</b> | <b>89,26</b> |
| 5  | Pendigits     | RF-C4.5 | 93,75          | <b>99,81</b>   | 99,23        | 95,89        | 99,64        |
|    |               | RF-ODT  | <b>98,21</b>   | 99,56          | <b>99,43</b> | <b>97,06</b> | <b>99,82</b> |
| 6  | Page blocks   | RF-C4.5 | 86,19          | <b>99,00</b>   | <b>97,70</b> | <b>88,43</b> | <b>99,23</b> |
|    |               | RF-ODT  | <b>92,20</b>   | 97,90          | 97,31        | 87,54        | <b>99,23</b> |
| 7  | Pima          | RF-C4.5 | 59,94          | <b>84,63</b>   | 75,97        | 63,29        | <b>83,63</b> |
|    |               | RF-ODT  | <b>70,19</b>   | 82,91          | <b>78,57</b> | <b>69,42</b> | 82,49        |
| 8  | German        | RF-C4.5 | 46,56          | <b>89,26</b>   | 76,77        | 52,99        | 79,04        |
|    |               | RF-ODT  | <b>60,73</b>   | 84,36          | <b>77,50</b> | <b>60,32</b> | <b>79,52</b> |
| 9  | Segment       | RF-C4.5 | 97,74          | <b>99,95</b>   | 99,61        | 98,66        | 99,97        |
|    |               | RF-ODT  | <b>99,68</b>   | 99,90          | <b>99,87</b> | <b>99,57</b> | <b>100</b>   |
| 10 | Satimage      | RF-C4.5 | 59,72          | <b>98,49</b>   | <b>94,40</b> | 69,23        | 96,16        |
|    |               | RF-ODT  | <b>77,25</b>   | 95,03          | 93,13        | <b>70,41</b> | <b>96,53</b> |



# Rừng ngẫu nhiên xiên phân (RF-ODT)

## Phân lớp đa lớp





# RF-ODT nhận dạng ký tự số viết tay MNIST

**Bảng:** So sánh kết quả nhận dạng ký tự số viết tay MNIST

| Methods                                                                      | Accuracy (%) |
|------------------------------------------------------------------------------|--------------|
| 1-layer Neural nets (LeCun et al., 1998)                                     | 88.00        |
| Nearest-neighbor (Euclidean L2) (LeCun et al., 1998)                         | 95.00        |
| Nearest-neighbor (Euclidean L2, deskewing) (LeCun et al., 1998)              | 97.60        |
| Convolution net LeNet-1 (Subsampling) (LeCun et al., 1998)                   | 98.30        |
| Convolution net LeNet-4 (LeCun et al., 1998)                                 | 98.90        |
| Convolution net LeNet-5 (Distortions) (LeCun et al., 1998)                   | 99.15        |
| Convolution net Boosted LeNet-4 (Distortions) (LeCun et al., 1998)           | 99.30        |
| Convolution net, cross-entropy (Affine distortions) (Simard et al., 2003)    | 99.40        |
| LibSVM (RBF, $\gamma = 0, 05$ , $c = 105$ )                                  | 98.37        |
| LibSVM (Poly, $\text{deg} = 5$ , $c = 105$ )                                 | 96.65        |
| SVM (Poly, $\text{deg} = 9$ , virtual SV, local) (Decoste & Scholkopf, 2002) | 99.44        |
| AdaBoost.M1 (100 trees with C4.5) (Kegl & Busa-Fekete, 2009)                 | 95.95        |
| Products of boosted stumps (haar) (Kegl & Busa-Fekete, 2009)                 | 99.12        |
| <b>RF-ODT (200 oblique decision trees, GIST)</b>                             | <b>99.20</b> |



# RF-ODT nhận dạng ký tự số viết tay USPS

**Bảng:** So sánh kết quả nhận dạng ký tự số viết tay USPS

| Methods                                                               | Accuracy (%) |
|-----------------------------------------------------------------------|--------------|
| Human Performance (Bromley & Sackinger, 1991)                         | 97.50        |
| Nearest-neighbor (Simard et al., 1993)                                | 94.10        |
| LeNet-1 (LeCun et al., 1989)                                          | 95.00        |
| Boosted neural nets (LeCun et al., 1989)                              | 97.40        |
| Tangent distance(Simard et al., 1993)                                 | 97.40        |
| AdaBoost.M1 (100 trees with C4.5) (Kegl & Busa-Fekete, 2009)          | 94.02        |
| Products of boosted stumps (Kegl & Busa-Fekete, 2009)                 | 95.76        |
| Products of boosted stumps (haar) (Kegl & Busa-Fekete, 2009)          | 96.26        |
| SVM (Poly, $deg = 9$ , virtual SV, local) (Decoste & Scholkopf, 2002) | 97.00        |
| SVM (RBF, $\gamma = 0.2$ , $c = 10^5$ )                               | 95.42        |
| SVM (Poly, $deg = 5$ , $c = 10^5$ )                                   | 95.72        |
| <b>RF-ODT (200 trees, GIST)</b>                                       | <b>96.71</b> |



# RF-ODT nhận dạng chữ viết tay New-bench-HCR

**Bảng:** So sánh kết quả nhận dạng chữ viết tay New-bench-HCR

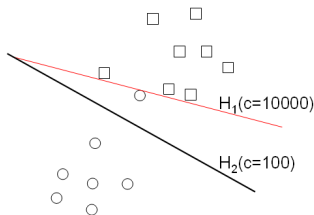
| Methods                                        | Accuracy (%) |
|------------------------------------------------|--------------|
| 1 Nearest-neighbor (van der Maaten, 09)        | 78,32        |
| 3 Nearest-neighbor (van der Maaten, 09)        | 79,21        |
| 5 Nearest-neighbor (van der Maaten, 09)        | 79,26        |
| Linear kernel classifiers (van der Maaten, 09) | 67,01        |
| <b>RF-ODT (100 trees, GIST)</b>                | <b>94,80</b> |



# Giải thuật ArcX4-iODT

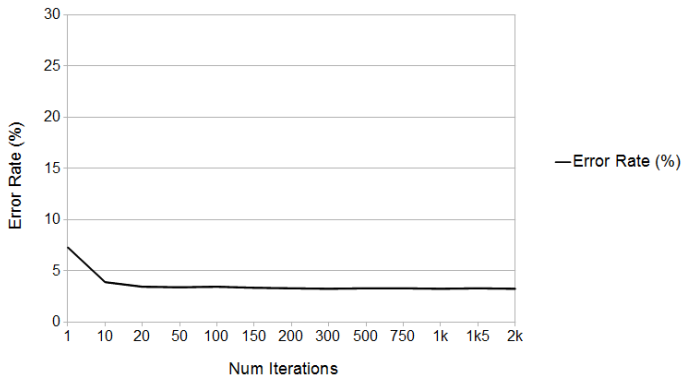
## Nguyên lý của ArcX4-iODT

- ▶ xây dựng tuần tự tập cây ngẫu nhiên xiên phân
- ▶ 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



# Giải thuật ArcX4-iODT: khi tăng số bước lặp

ArcX4-rODT for classifying USPS dataset



# ArcX4-iODT cho phân lớp dữ liệu y sinh

**Bảng:** So sánh kết quả phân lớp dữ liệu y sinh

| Ensemble<br>ID | Precision  |              |              | Recall       |            |              | F1         |              |              | Accuracy     |              |              |
|----------------|------------|--------------|--------------|--------------|------------|--------------|------------|--------------|--------------|--------------|--------------|--------------|
|                | LibSVM     | RF-C4,5      | ArcX4-rODT   | LibSVM       | RF-C4,5    | ArcX4-rODT   | LibSVM     | RF-C4,5      | ArcX4-rODT   | LibSVM       | RF-C4,5      | ArcX4-rODT   |
| 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> | 80,65        | 82,26        | <b>85,48</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>   | 97,06        | 97,06        | <b>100</b>   |
| 3              | 75,00      | 100          | <b>100</b>   | <b>100</b>   | <b>100</b> | <b>100</b>   | <b>100</b> | <b>100</b>   | <b>100</b>   | 93,33        | <b>100</b>   | <b>100</b>   |
| 4              | 69,23      | <b>83,33</b> | 78,57        | 75,00        | 83,33      | <b>91,67</b> | 72,00      | 83,33        | <b>84,62</b> | 63,16        | <b>78,94</b> | <b>78,94</b> |
| 5              | 85,00      | <b>94,12</b> | 90,00        | <b>94,44</b> | 80,00      | 90,00        | 89,47      | 86,49        | <b>90,00</b> | <b>90,48</b> | 88,10        | <b>90,48</b> |
| 6              | 73,53      | 75,76        | <b>100</b>   | <b>100</b>   | <b>100</b> | 96,00        | 84,75      | 86,21        | <b>97,96</b> | 73,53        | 76,47        | <b>97,06</b> |
| 7              | 88,26      | <b>93,75</b> | <b>93,75</b> | <b>100</b>   | <b>100</b> | <b>100</b>   | 93,75      | <b>96,77</b> | <b>96,77</b> | 98,66        | <b>99,33</b> | <b>99,33</b> |
| 8              | 47,62      | 45,46        | <b>63,64</b> | 55,56        | 23,81      | <b>66,67</b> | 51,28      | 31,25        | <b>65,11</b> | 68,33        | 63,33        | <b>75,00</b> |
| 9              | 83,13      | <b>92,58</b> | 90,24        | <b>84,42</b> | 73,83      | 80,69        | 83,77      | 82,15        | <b>85,20</b> | 92,15        | 92,30        | <b>93,27</b> |
| 10             | <b>100</b> | 98,78        | <b>100</b>   | <b>100</b>   | <b>100</b> | <b>100</b>   | <b>100</b> | 99,39        | <b>100</b>   | <b>100</b>   | 99,21        | <b>100</b>   |
| 11             | 91,30      | 95,65        | <b>100</b>   | 87,50        | 91,67      | <b>95,65</b> | 89,36      | 93,62        | <b>97,78</b> | 89,36        | <b>93,62</b> | <b>97,87</b> |
| 12             | 95,46      | 95,24        | <b>95,65</b> | <b>95,46</b> | 90,91      | <b>100</b>   | 95,46      | 93,02        | <b>97,78</b> | 98,21        | 97,32        | <b>99,11</b> |
| 13             | <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>   | <b>100</b>   | 99,11        | <b>100</b>   |
| 14             | <b>100</b> | <b>100</b>   | <b>100</b>   | <b>100</b>   | <b>100</b> | <b>100</b>   | <b>100</b> | <b>100</b>   | <b>100</b>   | <b>100</b>   | <b>100</b>   | <b>100</b>   |
| 15             | 92,59      | <b>100</b>   | 79,31        | 39,68        | 29,63      | <b>85,19</b> | 55,56      | 45,71        | <b>82,14</b> | 64,29        | 83,93        | <b>91,07</b> |



# Phát triển mở rộng của RF-ODT

## Nghiên cứu đang thực hiện

- ▶ các hàm phân hoạch đa biến khác
- ▶ các luật quyết định
- ▶ cài đặt song song
- ▶ hồi quy, gom nhóm
- ▶ phân loại ảnh, văn bản, etc.



## Cám ơn & câu hỏi thảo luận .....

