# MỘT HƯỚNG TIẾP CẬN ONE-VERSUS-ALL CHO PHÂN LỚP ĐA LỚP

**ThS. Huỳnh Lê Uyên Minh**
Khoa Sư phạm Toán-Tin, Trường Đại học Đồng Tháp
Email: uyenminhdhdt@gmail.com

**Tóm tắt**. *Trong bài báo này, chúng tôi đi xây dựng giải thuật phân lớp đa lớp bằng việc mở rộng giải thuật RF-ODT theo hướng tiếp cận phương pháp One-Versus-All để đưa vấn đề đa lớp về thành các vấn đề nhị phân, sau đó dùng phương pháp phân tích biệt lập tuyến tính tìm siêu phẳng tối ưu tách dữ liệu, gọi là giải thuật RF-ODT-OVA. Kết quả thực nghiệm cho thấy RF-ODT-OVA cho độ chính xác cao khi phân lớp đa lớp (96,85%), cao hơn cả rừng ngẫu nhiên và thấp hơn một ít so với máy học véctơ hỗ trợ. Đây cũng là một kết quả có ý nghĩa quan trọng cho việc học tập và nghiên cứu trong lĩnh vực khai khoáng dữ liệu.*

## 1. Mở đầu

Những năm gần đây, vấn đề khai khoáng dữ liệu đã trở thành một trong những hướng nghiên cứu chính trong công nghệ tri thức. Nó được xem như là một môn học được đưa vào giảng dạy ở cấp bậc đại học và sau đại học. Khai khoáng dữ liệu tập trung giải quyết các vấn đề cơ bản như: phân lớp, hồi quy, gom nhóm, luật kết hợp. Đặc biệt với phân lớp dữ liệu có số chiều lớn, vấn đề khó khăn thường gặp là dữ liệu thường tách rời nhau trong không gian có số chiều lớn, nên người ta mong muốn tìm ra được giải thuật có thể phân lớp tốt các dữ liệu như vậy.

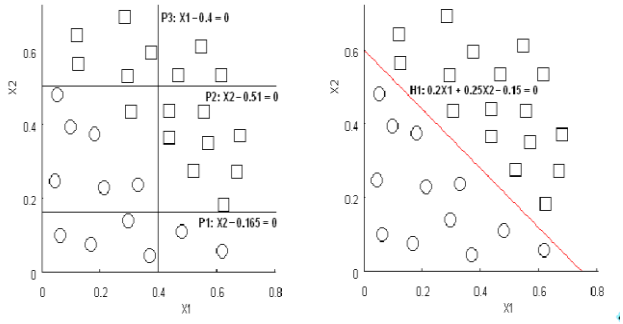
Hiện nay thì giải thuật rừng ngẫu nhiên và máy học SVM được xem là lựa chọn hợp lý, trong đó giải thuật rừng ngẫu nhiên (Breiman, 2001) là một trong những phương pháp tập hợp mô hình thành công nhất. Tuy nhiên, giải thuật rừng ngẫu nhiên chỉ xây dựng các cây quyết định thông thường và chỉ chọn một thuộc tính dùng để phân hoạch tại mỗi nút, vì thế cá nhân mỗi cây kém hiệu quả khi làm việc với dữ liệu có sự phụ thuộc nhau giữa các thuộc tính, thường gặp ở những dữ liệu có số chiều rất lớn. Do đó, hiện đã có giải thuật rừng ngẫu nhiên xiên phân RF-ODT được đánh giá là cho kết quả phân lớp hiệu quả hơn rừng ngẫu nhiên. Điểm hạn chế của RF-ODT là chỉ phân lớp hai lớp nên chúng tôi xây dựng giải thuật rừng ngẫu nhiên xiên phân RF-ODT-OVA cho phân lớp đa lớp theo hướng mở rộng giải thuật RF-ODT, đồng thời sẽ đánh giá hiệu quả của RF-ODT-OVA để làm cơ sở cho việc nghiên cứu, học tập trong lĩnh vực khai khoáng dữ liệu.

## 2. Kết quả chính

### 2.1. Giới thiệu giải thuật rừng ngẫu nhiên xiên phân RF-ODT cho phân lớp hai lớp

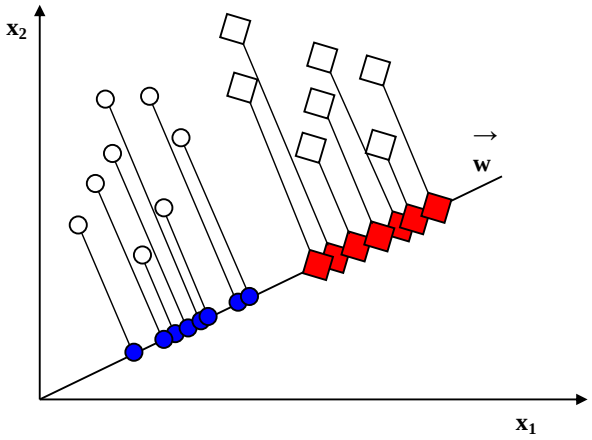
Giải thuật RF-ODT (Do et al., 2009) đi xây dựng một tập hợp các cây quyết định xiên tương tự như trong rừng ngẫu nhiên của Breiman (Breiman, 2001). Điểm

khác biệt trong quá trình xây dựng cây quyết định xiên ngẫu nhiên của RF-ODT là sử dụng phương pháp phân tích biệt lập tuyến tính của Fisher để phân hoạch đa thuộc tính tại các nút (Linear Discriminant Analysis - LDA).




**Hình 1: Phân hoạch đơn thuộc tính (trái) và đa thuộc tính (phải)**

Quá trình thực hiện của LDA dựa trên độ biệt lập tuyến tính của dữ liệu (Fisher, 1936). Ý tưởng chính của LDA là tìm véctơ (siêu phẳng) sao cho khi chiếu dữ liệu lên đó thì độ biệt lập giữa trung bình dữ liệu của 2 lớp là lớn nhất và độ chồng lấp giữa 2 lớp là nhỏ nhất.




**Hình 2: Minh hoạ véctơ (w) dùng để chiếu dữ liệu 2 thuộc tính (chiều)**

## 2.2. Xây dựng giải thuật RF-ODT-OVA cho phân lớp dữ liệu đa lớp

Trong giải thuật RF-ODT-OVA này, chúng tôi cũng sử dụng phương pháp phân tích biệt lập tuyến tính LDA để phân hoạch dữ liệu xiên phân tại mỗi nút của cây. Tuy nhiên, phương pháp LDA chỉ thực hiện cho vấn đề phân lớp nhị phân nên chúng tôi xây dựng giải thuật bằng cách mở rộng giải thuật RF-ODT theo tiếp cận OVA với tiêu chí cực đại hóa lề phân hoạch để phân lớp.

Ở đây chúng tôi dùng phương pháp OVA vì tính đơn giản của nó, nếu kết hợp với tính hiệu quả của giải thuật RF-ODT và việc sử dụng tuần tự LDA trong các mô hình đa lớp sẽ giúp khắc phục nhược điểm của OVA.

***Chi tiết của giải thuật RF-ODT-OVA được mô tả cụ thể như dưới dây:***

– Giải thuật sẽ xây dựng k mô hình phân lớp. (k: số lớp).

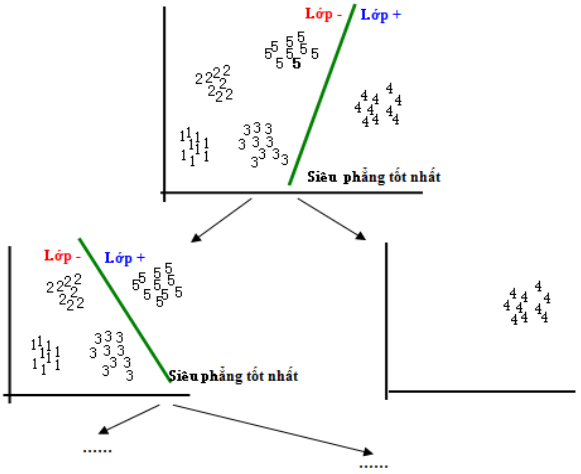
– Ta có k’ là số lớp của mô hình phân lớp thứ i: khởi tạo k’=k; sau mỗi mô hình phân lớp ta có số lớp k’=số lớp mô hình phân lớp trước -1

– Xét mỗi mô hình phân lớp thứ i:

- Giải thuật sẽ xây dựng k’ siêu phẳng (ứng với k’ lớp), kết hợp tính chỉ số Gini (hoặc độ lợi thông tin) cho mỗi siêu phẳng, mục đích để tìm ra lớp học có thể được tách hiệu quả nhất từ k’ lớp học.
- Ứng với mỗi siêu phẳng, tạm gọi là siêu phẳng thứ j sẽ tách lớp thứ j (với j Є [1..k']): đưa lớp thứ j về làm lớp dương (+), đưa các lớp còn lại làm lớp âm (-), kế tiếp dùng phương pháp LDA thực hiện việc phân hoạch xiên phân để tìm ra siêu phẳng tối ưu tách lớp thứ j từ các lớp còn lại.
- Sau đó chọn siêu phẳng có chỉ số Gini (hoặc độ lợi thông tin $E_{new}$) nhỏ nhất để tách lớp thứ j (lớp dương) tại nút.

Phương pháp LDA mà giải thuật sử dụng ở đây đã được điều chỉnh cách tính độ lệch b theo tham số α ϵ (0,1) theo như đề xuất trong (Do et al., 2009).

Minh họa quá trình phân lớp của giải thuật RF-ODT-OVA:




**Hình 3: Minh họa quá trình phân lớp đa lớp của giải thuật RF-ODT-OVA**

## 2.3. Kết quả nghiên cứu

Giải thuật RF-ODT-OVA được cài đặt bằng ngôn ngữ lập trình C/C++ dựa trên mã nguồn của giải thuật RF-ODT. Tất cả các giải thuật được thực hiện trên máy tính cá nhân chạy hệ điều hành Linux. Các dữ liệu đa lớp sử dụng trong thực nghiệm thuộc lĩnh vực nhận dạng ký tự số, ký tự viết tay và dữ liệu vân tay, chi tiết của các tập dữ liệu được mô tả trong bảng 1:

**Bảng 1: Mô tả các tập dữ liệu đa lớp**

| Tập dữ liệu | Số mẫu huấn luyện | Số mẫu kiểm tra | Số chiều | Số lớp | Nhãn |
|---|---|---|---|---|---|
| Opt | 3823 | 1797 | 64 | 10 | 0→9 |
| Usps | 7291 | 2007 | 256 | 10 | 1→10 |
| Letter | 13334 | 6666 | 16 | 26 | 0→25 |
| Fp-57 | 700 | 352 | 200 | 57 | 1→56 |
| Fp-78 | 950 | 422 | 200 | 78 | 0→77 |

Các tham số sử dụng cho RF-ODT-OVA được ghi nhận trong bảng 2:

**Bảng 2: Các tham số sử dụng cho RF-ODT-OVA**

| STT | Tập dữ liệu | Số chiều ngẫu nhiên | Số chiều được chọn | Số cây được chọn |
|---|---|---|---|---|
| 1 | Opt | 64 | 32 | 100 |
| 2 | Letter | 16 | 12 | 100 |
| 3 | Usps | 256 | 50 | 100 |
| 4 | Fp-57 | 200 | 110 | 100 |
| 5 | Fp-78 | 200 | 100 | 100 |

Để đánh giá hiệu quả của giải thuật, chúng tôi cũng so sánh kết quả của RF-ODT-OVA với giải thuật RF-CART, Lib-SVM (hai giải thuật hiệu quả hiện nay) có trong thư viện R. Trong đó tham số cho RF-CART được điều chỉnh tương ứng, riêng tham số cho Lib-SVM được mô tả như bảng 3:

**Bảng 3: Các tham số sử dụng cho Lib-SVM**

| STT | Tập dữ liệu | Hàm nhân | Gamma |
|---|---|---|---|
| 1 | Opt | Radial | 0.00085 |
| 2 | Usps | Radial | 0.0035 |
| 3 | Letter | Radial | 0.018 |
| 4 | Fp-57 | Linear | - |

| | | | |
|---|---|---|---|
| 5 | Fp-78 | Radial | 0.0001 |

Kết quả thực nghiệm được đánh giá dựa trên độ chính xác phân lớp – được tính bằng số điểm dữ liệu được phân lớp đúng của tất cả các lớp chia cho tổng số điểm dữ liệu. Chúng tôi thu được kết quả như trình bày trong bảng 4.

**Bảng 4: Kết quả phân lớp của RF-ODT-OVA so với RF-CART, Lib-SVM**

| Tên tập dữ liệu | Độ chính xác (%) | | |
|---|---|---|---|
| | **RF-ODT-OVA** | **RF-CART** | **Lib-SVM** |
| Opt | 97.42 | 95.47 | 98.44 |
| Letter | 96.85 | 96.00 | 97.40 |
| Usps | 94.25 | 92.97 | 95.12 |
| Fp-57 | 97.16 | 93.94 | 96.59 |
| Fp-78 | 98.58 | 95.73 | 98.58 |
| **TRUNG BÌNH** | **96.85** | **94.82** | **97.23** |

## 3. Kết luận

Qua các ghi nhận trên cho thấy giải thuật RF-ODT-OVA cho kết quả hoàn toàn tốt hơn so với giải thuật RF-CART (cao hơn 2.03%), thấp hơn một ít so với giải thuật Lib-SVM (thấp hơn 0.38%), đồng thời RF-ODT-OVA cũng đạt được độ chính xác cao (96.85%) khi phân lớp dữ liệu đa lớp. Đây cũng là một kết quả nghiên cứu có ý nghĩa quan trọng cho việc học tập, nghiên cứu trong lĩnh vực khai khoáng dữ liệu.

## TÀI LIỆU THAM KHẢO

1. Đ. Q. Bảo, Đ. T. Nhung, Đ. T. Nghị, L. Philippe, L. Stéphane, "Phân loại dữ liệu gien với rừng ngẫu nhiên xiên phân". *Tuyển tập công trình nghiên cứu công nghệ Thông tin và Truyền thông năm 2009*, trang 1-8.
2. Alexander Statnikov, Constantin F. Aliferis, "Are Random Forests Better than Support Vector Machines for Microarray-Based Cancer Classification?*"* Vanderbilt University, Nashville, TN, USA, 2007, pp.686-687.
3. Breiman, L., "Random Forests", *Machine Learning*, 45(1), 2001, pp.5-32.

4. Chaoyang Zhang, Peng Li, Arun Rajendran, Youping Deng, "Parallel Multicategory Support Vector Machines (PMC-SVM) for Classifying microcarray Data*", In.:Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS'06). IEEE, 2006.*
5. Fisher, R. A., "The use of multiple measurements in taxonomic problems. Annals of Eugenics", *Annals of Eugenics*, Vol. 7, Pt. II, 1936, pp.179-188.
6. Freund , Y., Schapire, R., "A decision-theoretic generalization of on-line learning and an application to boosting", *Computational Learning Theory*, 1995, pp. 23–37.
7. Hsu, C. W., Lin, C. J., "A Comparison of Methods for Multi-class Support Vector Machines", *IEEE Transactions on Neural Networks*, 13, pp. 415-425, 2002, pp.1045–1052.
8. T-N. Do, S. Lallich, N-K. Pham, P. Lenca, "Classifying very-high-dimensional data with random forests of oblique decision trees", in *Advances in Knowledge Discovery and Management*, H. Briand, F. Guillet, G. Ritschard, D. Zighed Eds, Springer-Verlag, 2009, pp. 39-55.
9. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.