



Khoa Công Nghệ Thông Tin
Trường Đại Học Cần Thơ



Phương pháp tập hợp Ensemble-based methods

Đỗ Thanh Nghi
dtngghi@cit.ctu.edu.vn

Cần Thơ
12-02-2019

Nội dung

- Giới thiệu về Ensemble-based
- Bagging, Random forests, Boosting
- Kết luận và hướng phát triển

Nội dung

- **Giới thiệu về Ensemble-based**
- Bagging, Random forests, Boosting
- Kết luận và hướng phát triển

Ensemble-based

■ phương pháp ensemble-based

- xây dựng tập hợp các mô hình cơ sở dựa trên tập học
- kết hợp các mô hình khi phân loại cho độ chính xác cao
- dựa trên cơ sở bias/variance
- bagging, random forests, boosting
- áp dụng cho nhiều giải thuật cơ sở khác nhau như cây quyết định, SVM, naive Bayes, etc.
- giải quyết các vấn đề về phân loại, hồi quy, gom nhóm, etc.
- cho kết quả tốt, tuy nhiên không thể dịch được kết quả sinh ra
- được ứng dụng thành công trong hầu hết các lĩnh vực tìm kiếm thông tin, nhận dạng, phân tích dữ liệu, etc.

- **Giới thiệu về Ensemble-based**
- Bagging, Random forests, Boosting
- Kết luận và hướng phát triển

Top 10 DM algorithms (2015)



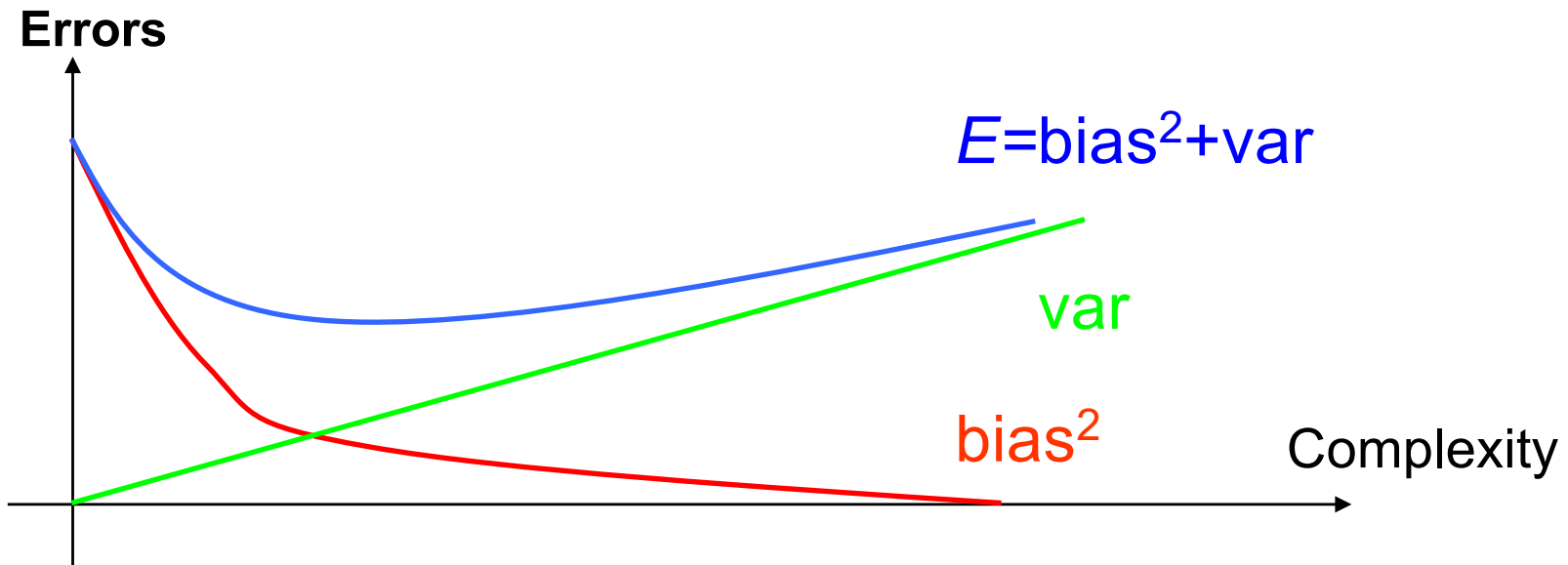
Here are the algorithms:

- 1. C4.5
- 2. k-means
- 3. Support vector machines
- 4. Apriori
- 5. EM
- 6. PageRank
- 7. AdaBoost
- 8. kNN
- 9. Naive Bayes
- 10. CART

Ensemble-based

■ hiệu quả giải thuật học

- bias : thành phần lỗi độc lập với mẫu dữ liệu học
- variance : thành phần lỗi do biến động liên quan đến sự ngẫu nhiên của tập học



Nội dung

- Giới thiệu về Ensemble-based
- **Bagging, Random forests, Boosting**
- Kết luận và hướng phát triển

Ensemble-based

■ Nguyên lý

- giảm thành phần lỗi bias/variance
- kết hợp nhiều bộ phân lớp yếu (nhưng không quá tệ) khác nhau và đa dạng
- bộ phân lớp yếu: cây quyết định, Bayes thơ ngây, etc.
- Bagging (Breiman, 1996)
- Boosting (Freund & Schapire, 1995), (Breiman, 1997)
- Forêts aléatoires (Breiman, 2001)

Ensemble-based

■ averaging technique

- averaging technique
- xây dựng tập hợp các mô hình cơ sở độc lập nhau
- kết hợp sự phân loại của các mô hình
- bagging và random forests
- giảm variance

■ boosting technique

- xây dựng tập hợp các mô hình cơ sở tuần tự (tập trung lên các lỗi sinh ra từ các mô hình trước)
- AdaBoost và arcing
- giảm bias

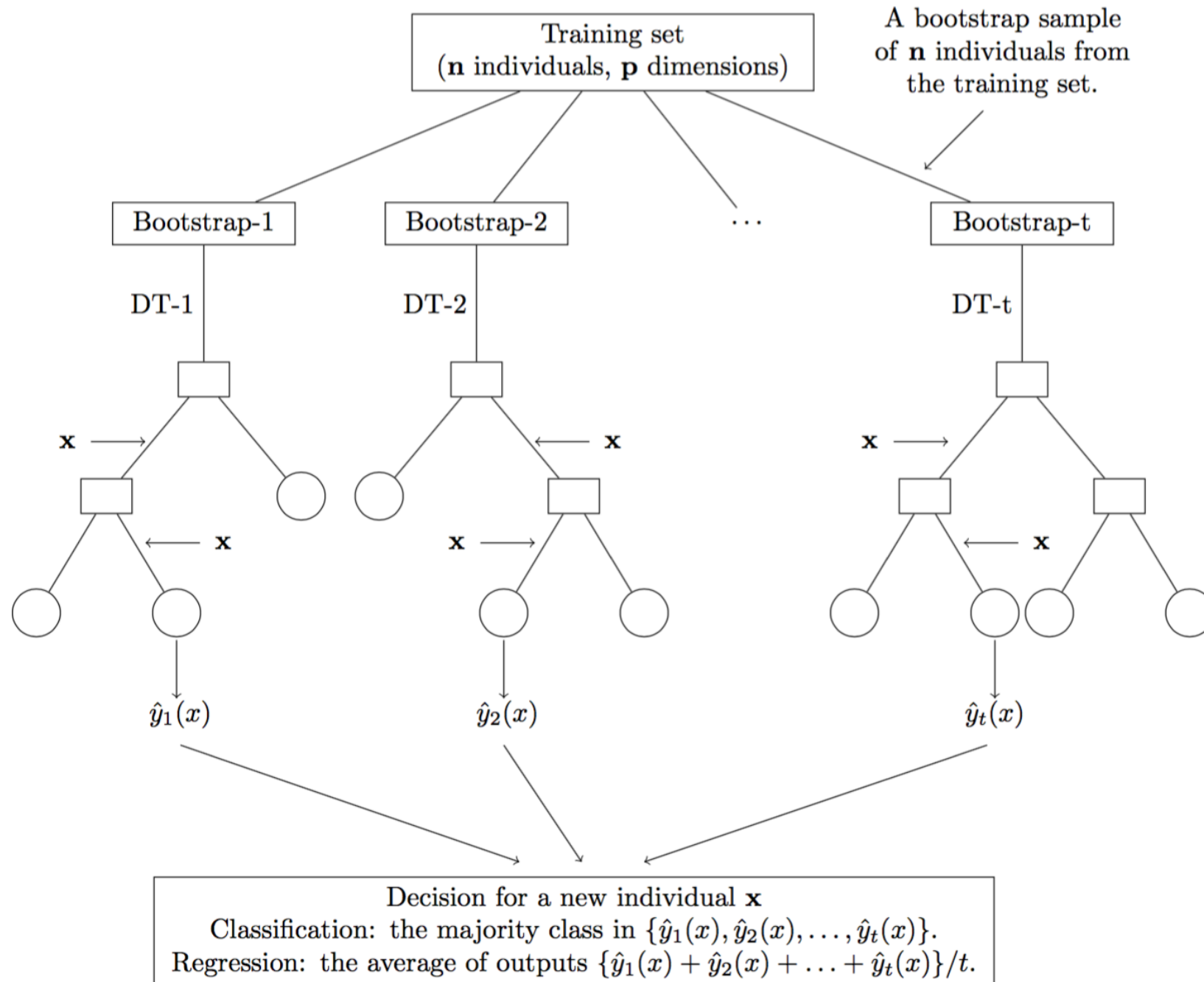
Bagging (Breiman, 1996)

■ Bootstrap AGGREGatING

- từ tập học LS có N phần tử
- xây dựng tập hợp T mô hình cơ sở độc lập nhau
- mô hình thứ i được xây dựng trên tập mẫu bootstrap
- 1 bootstrap : lấy mẫu N phần tử có hoàn lại từ tập LS
- khi phân loại : sử dụng majority vote
- hồi quy : tính giá trị trung bình của dự đoán của các mô hình

Bagging (Breiman, 1996)

- Giới thiệu về Ensemble-based
- **Bagging, Random forest, Boosting**
- Kết luận và hướng phát triển



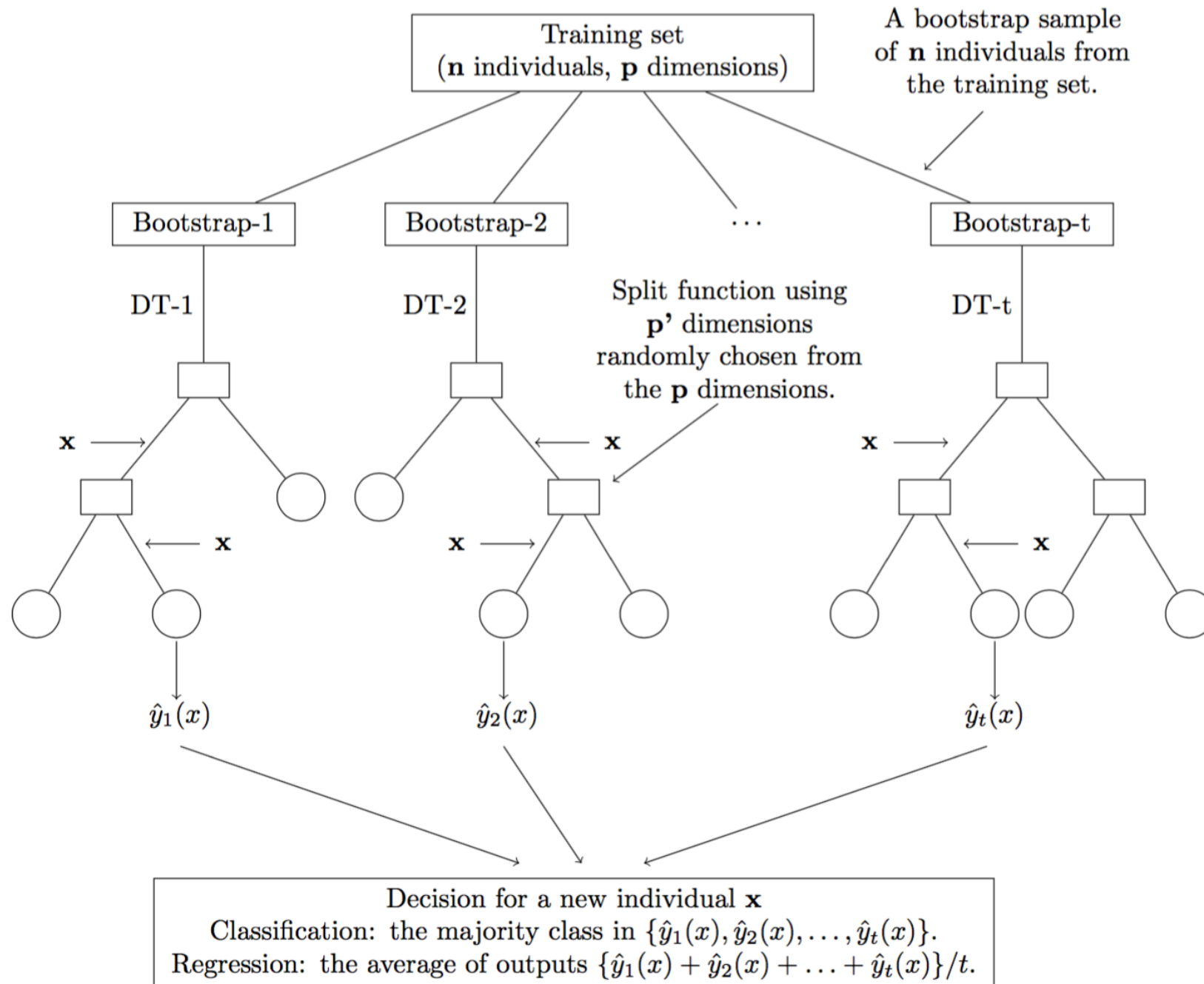
Random forests (Breiman, 2001)

■ Random forests

- từ tập học LS có N phần tử
- xây dựng tập hợp T mô hình cơ sở độc lập nhau
- mô hình thứ i được xây dựng trên tập mẫu bootstrap, chú ý
 - tại nút trong, chọn ngẫu nhiên n' thuộc tính ($n' \ll n$) và tính toán phân hoạch tốt nhất dựa trên n' thuộc tính này
 - cây được xây dựng đến độ sâu tối đa không cắt nhánh
- 1 bootstrap : lấy mẫu N phần tử có hoàn lại từ tập LS
- khi phân loại : sử dụng majority vote
- hồi quy : tính giá trị trung bình của dự đoán của các mô hình

Random forests (Breiman, 2001)

- Giới thiệu về Ensemble-based
- Bagging, Random forest, Boosting
- Kết luận và hướng phát triển



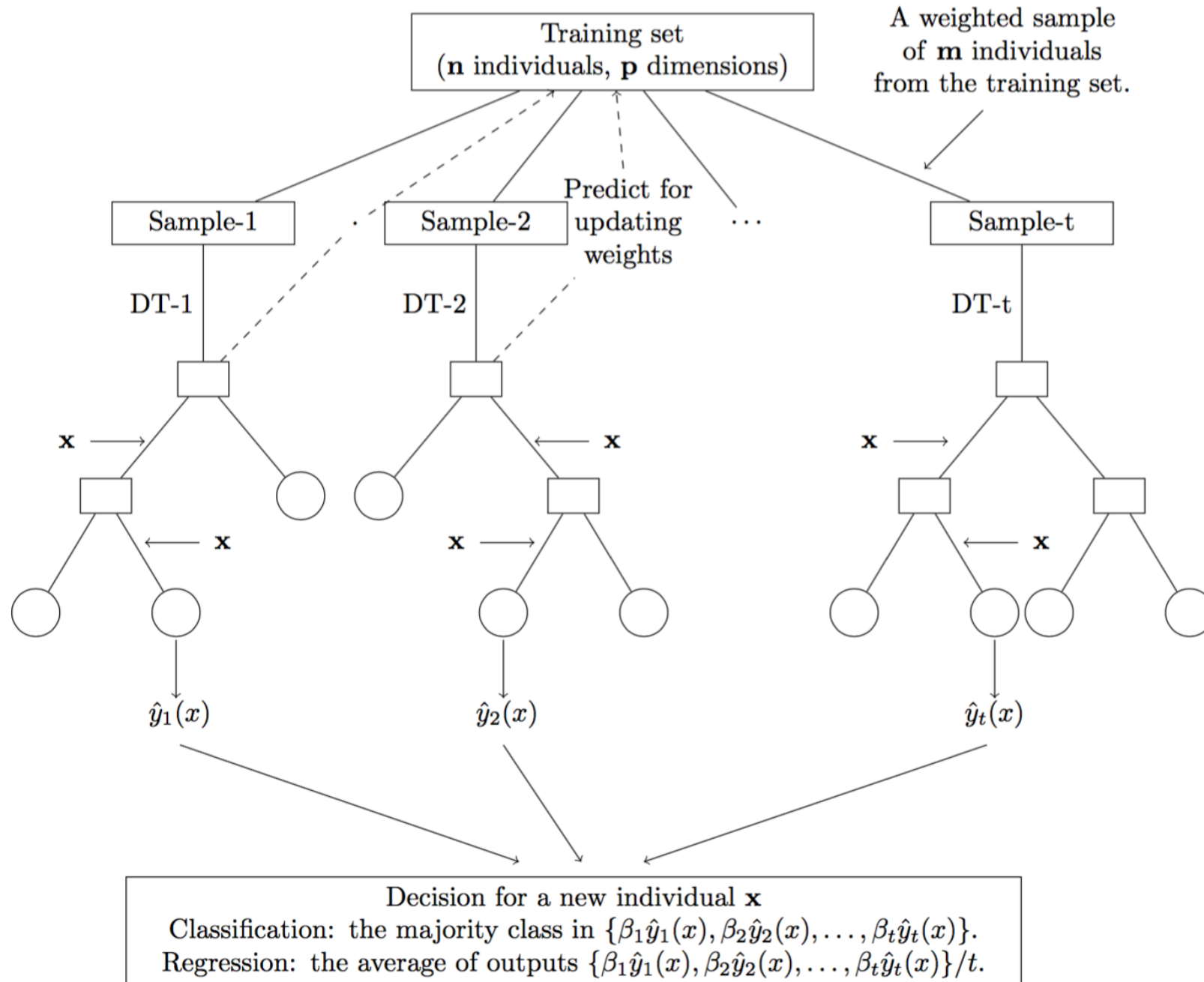
Boosting (Freund & Schapire, 1995)

■ Boosting

- từ tập học LS có N phần tử
- xây dựng tập hợp T mô hình cơ sở tuần tự
- mô hình thứ i được xây dựng trên tập mẫu lấy từ LS, tập trung vào các phần tử bị phân loại sai bởi mô hình thứ i-1 trước đó
- khi phân loại : sử dụng majority vote có trọng số
- hồi quy : tính giá trị trung bình của dự đoán của các mô hình có sử dụng trọng số

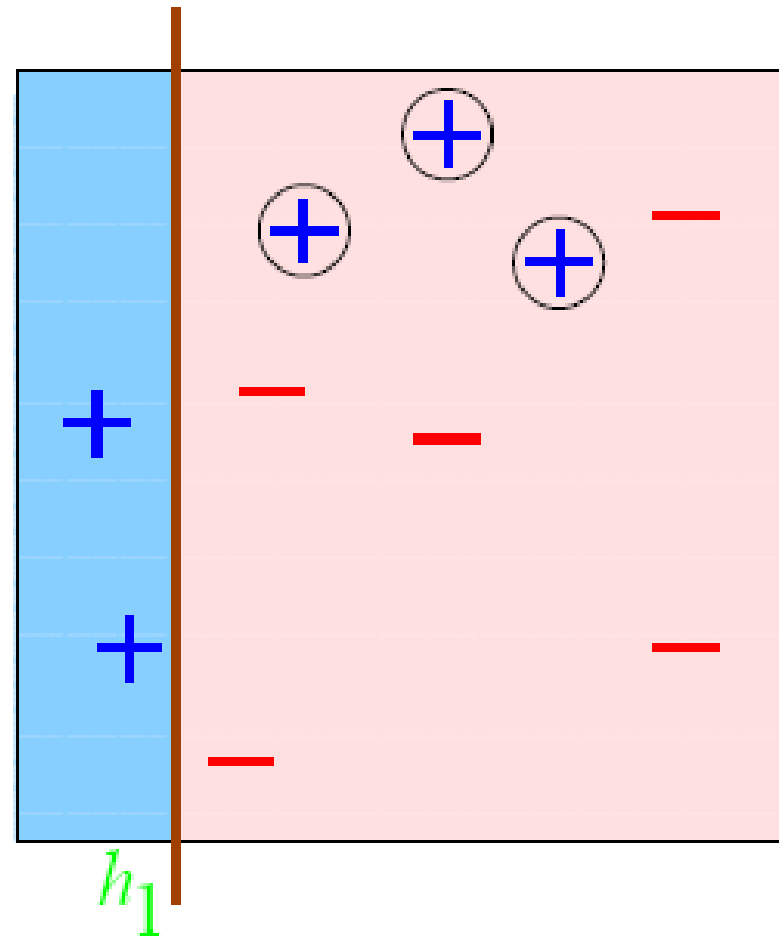
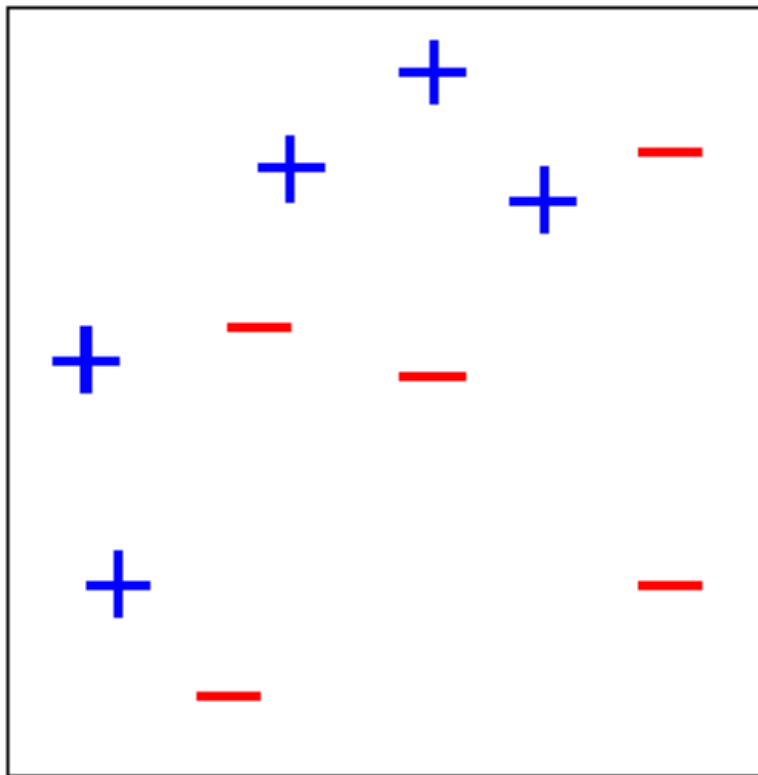
Boosting (Freund & Schapire, 1995)

- Giới thiệu về Ensemble-based
- Bagging, Random forest, Boosting
- Kết luận và hướng phát triển



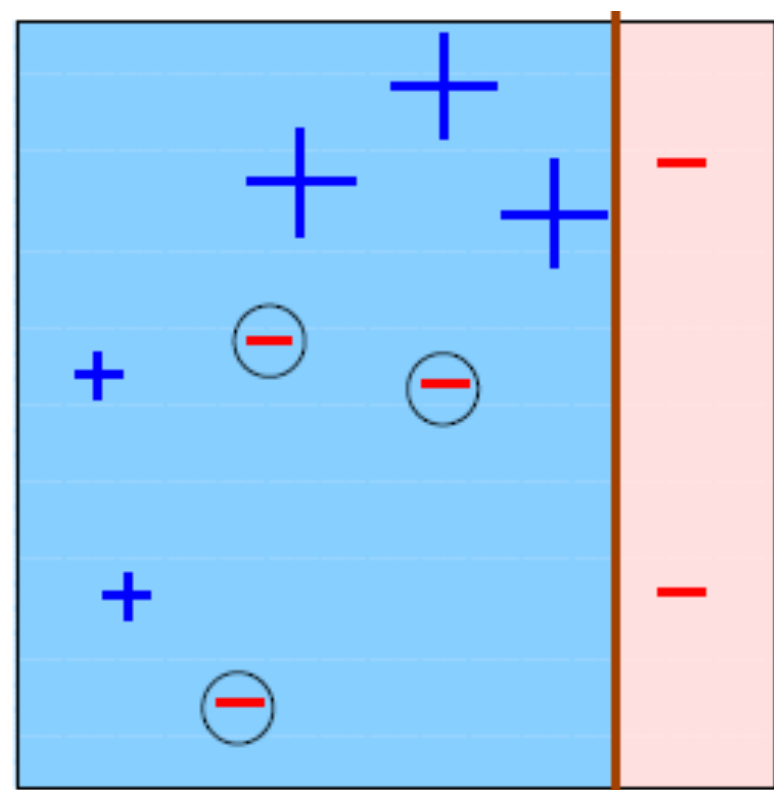
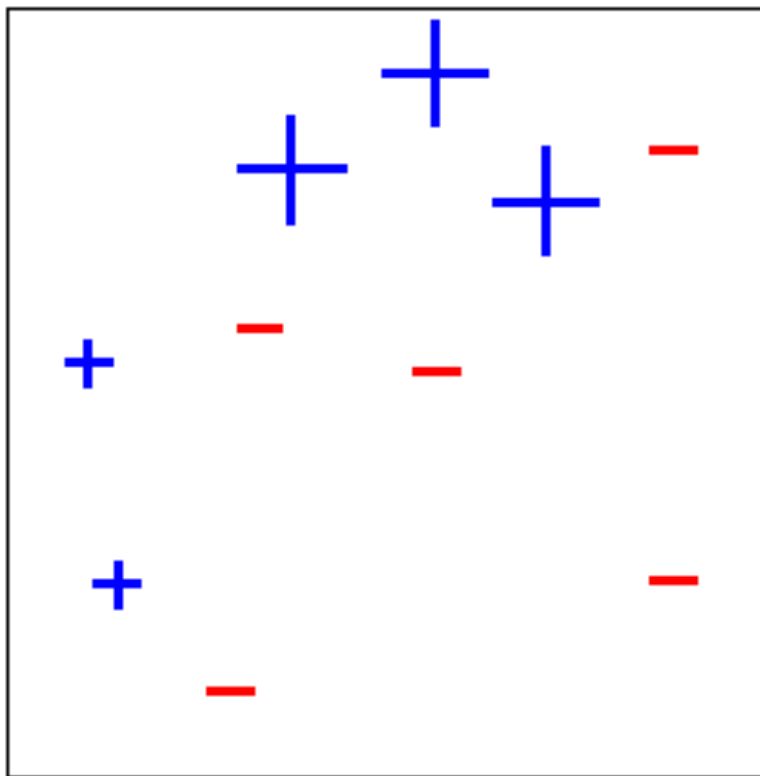
- Giới thiệu về Ensemble-based
- **Bagging, Random forest, Boosting**
- Kết luận và hướng phát triển

Boosting (Freund & Schapire, 1995)



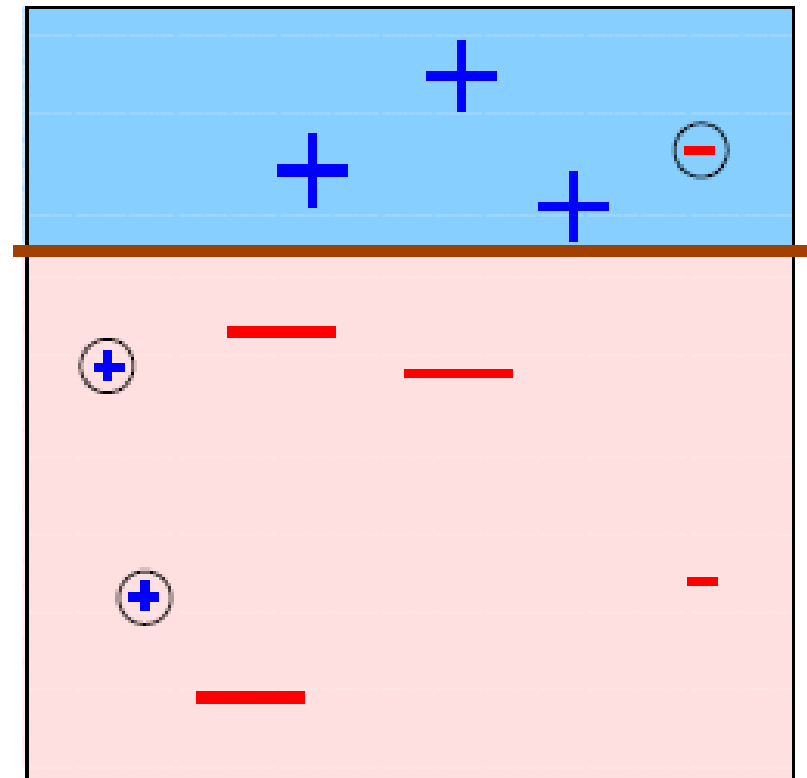
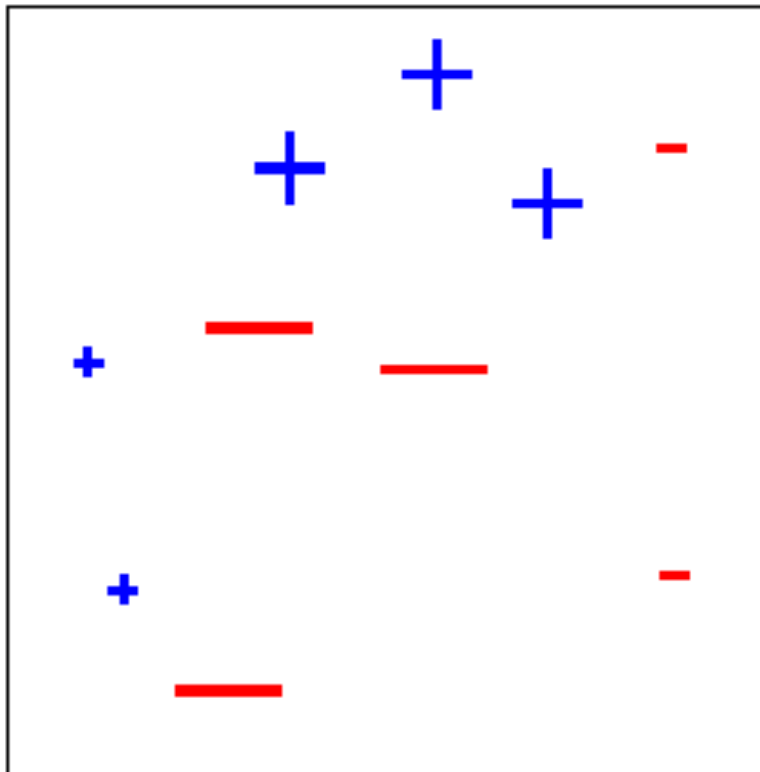
- Giới thiệu về Ensemble-based
- **Bagging, Random forest, Boosting**
- Kết luận và hướng phát triển

Boosting (Freund & Schapire, 1995)



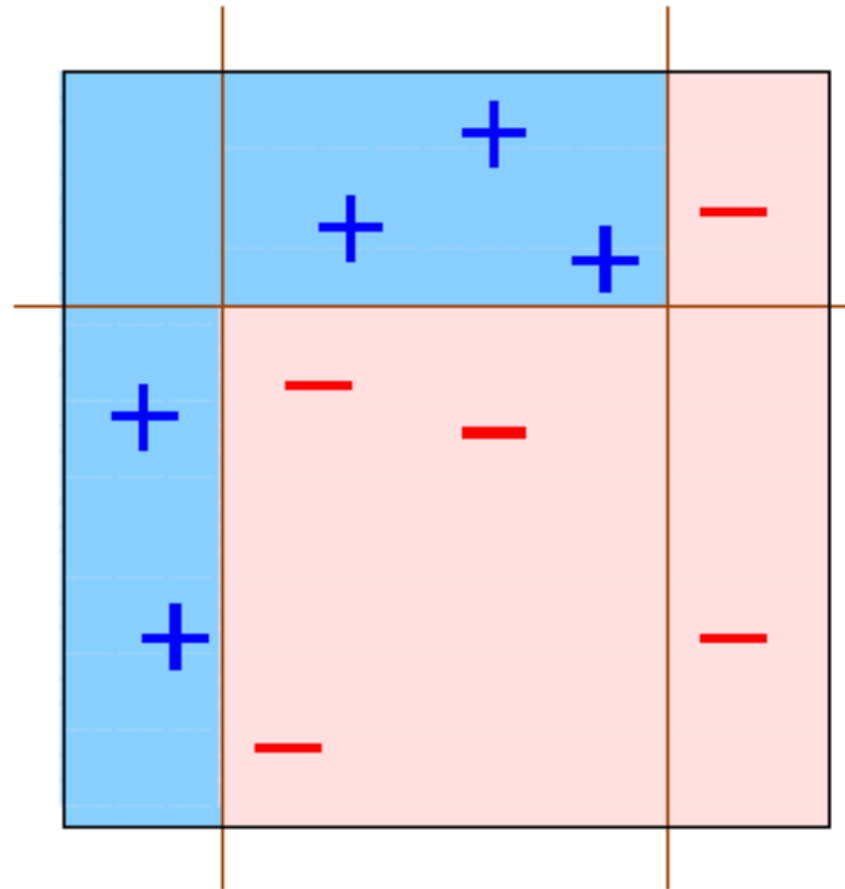
- Giới thiệu về Ensemble-based
- **Bagging, Random forest, Boosting**
- Kết luận và hướng phát triển

Boosting (Freund & Schapire, 1995)



- Giới thiệu về Ensemble-based
- **Bagging, Random forest, Boosting**
- Kết luận và hướng phát triển

Boosting (Freund & Schapire, 1995)



Nội dung

- Giới thiệu về Ensemble-based
- Bagging, Random forests, Boosting
- **Kết luận và hướng phát triển**

Phương pháp ensemble-based

- cải thiện rất tốt hiệu quả các phương pháp học thông thường như cây quyết định, naïve Bayes, SVM, etc.
 - dựa trên cơ sở bias/variance
 - xây dựng tập hợp các mô hình cơ sở dựa trên tập học
 - kết hợp các mô hình khi phân loại cho độ chính xác cao
 - kết quả rất khó diễn dịch, ví dụ như 1 rừng gồm hàng trăm cây quyết định

Hướng phát triển

- học trên dữ liệu không cân bằng
- diễn dịch kết quả sinh ra
- kiểm chứng sự hợp lệ của phương pháp



Cám ơn !