# Bộ dữ liệu breast-cancer

## Mô tả dữ liệu

* Link data: https://archive.ics.uci.edu/ml/datasets/breast+cancer
* Nguồn: Matjaz Zwitter & Milan Soklic (physicians) Institute of Oncology University Medical Center Ljubljana, Yugoslavia.
* Kiểu dữ liệu: nominal
* Số thuộc tính: 9
* Số lượng mẫu: 286
* Missing Values: yes
* Số lượt truy cập: 561671
* Lĩnh vực: Y tế
* Đây là một trong ba tên miền do Viện Ung thư cung cấp, đã nhiều lần xuất hiện trong tài liệu học máy. (Xem thêm nội dung hạch bạch huyết và khối u nguyên phát.) Tập dữ liệu này bao gồm 201 cá thể của một lớp và 85 thể hiện của một lớp khác. Các trường hợp được mô tả bằng 9 thuộc tính, một số thuộc tính tuyến tính và một số thuộc tính danh nghĩa
* Thông tin thuộc tính

1. Class: no-recurrence-events, recurrence-events

2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.

3. menopause: lt40, ge40, premeno.

4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.

5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.

6. node-caps: yes, no.

7. deg-malig: 1, 2, 3.

8. breast: left, right.

9. breast-quad: left-up, left-low, right-up, right-low, central.

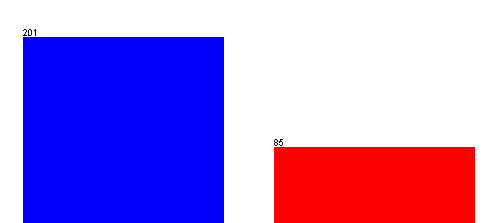
10. irradiat: yes, no.

## Tiền xử lý dữ liệu

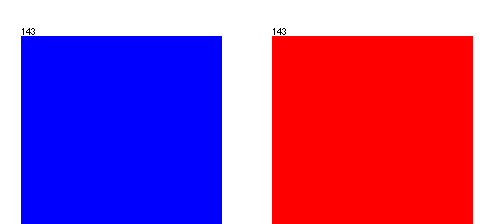
Dữ liệu không cần bằng giữa hai lớp phát bệnh và không phát bệnh

Vì vậy em sẽ tăng dữ liệu ở lớp không phát bệnh lên bằng kỹ thuật ClassBalancer đơn giản là tái lại trọng số sao cho tổng trọng số của tất cả các mẫu của lớp trong dữ liệu là như nhau. Không có trường hợp nào bị xóa hoặc thêm, vì vậy tổng số lượng mẫu của hai lớp vẫn không thay đổi.

Trước khi chuẩn hoá



Sau khi chuẩn hoá



## So sánh kết quả của các thuật toán

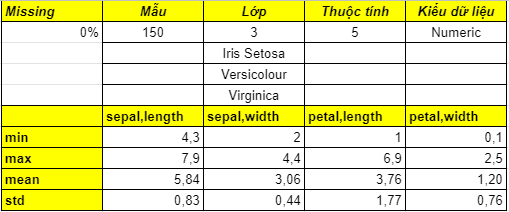
|  | Test mode: 10-fold cross-validation | | |
| --- | --- | --- | --- |
| Tham số | J48 | ANN | KNN |
| Đặc trưng | cây đã được tỉa | 1 lớp ẩn có 26 nơ ron | k = 11 (tốt nhất) |
| Số lá: 27 | tốc độ học: 0,3 |  |
| Kích thước cây: 31 | quán tính: 0,2 |  |
|  | batchsize: 100 |  |
|  | epoch: 500 |  |
| Số mẫu phân lớp đúng | 170 | 176 | 190 |
| Số mẫu phân lớp sai | 116 | 110 | 96 |
| Thống kê kappa | 0,187 | 0,2336 | 0,3272 |
| Sai số trung bình tuyệt đối | 0,4728 | 0,3825 | 0,4317 |
| Lỗi bình phương trung bình gốc | 0,5192 | 0,5838 | 0,4735 |
| Sai số tuyệt đối tương đối | 94,54% | 76,4855% | 86,331 % |
| Ma trận nhầm lẫn | a b <-- classified as  102.45 40.55 | a = no-recurrence-events  75.71 67.29 | b = recurrence-events | a b <-- classified as  107.43 35.57 | a = no-recurrence-events  74.02 68.98 | b = recurrence-events | a b <-- classified as  112.41 30.59 | a = no-recurrence-events  65.61 77.39 | b = recurrence-events |
|
|

| Thuật toán | Tên phương pháp | TP Rate | FP Rate | Precision | Recall | F - Measure | ROC Area | Class |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DT | J48 | 0,594 | 0,406 | 0,6 | 0,594 | 0,587 | 0,565 | Weighted Avg |
| ANN | MultilayerPerceptron | 0,751 | 0,518 | 0,592 | 0,751 | 0,662 | 0,646 | no-recurrence-events |
|  | 0,482 | 0,249 | 0,66 | 0,482 | 0,557 | 0,646 | recurrence-events |
|  | 0,617 | 0,383 | 0,626 | 0,617 | 0,61 | 0,646 | Weighted Avg |
| KNN | IBk | 0,664 | 0,336 | 0,674 | 0,664 | 0,659 | 0,692 | Weighted Avg |

KL: Với bộ dữ liệu này khi xử lý cân bằng dữ liệu bằng class balance và thực hiện thử nghiệm trên một vài thuật toán thì ta thu được kết quả trên. Thuật toán tốt nhất trong trường hợp này là KNN với k = 11

# Bộ dữ liệu iris

## Mô tả dữ liệu



## Tiền xử lý dữ liệu

1. Làm sạch dữ liệu (data cleaning/cleansing): loại bỏ nhiễu (remove noise), hiệu chỉnh những phần dữ liệu không nhất quán (correct data inconsistencies)
2. Tích hợp dữ liệu (data integration): trộn dữ liệu (merge data) từ nhiều nguồn khác nhau vào một kho dữ liệu
3. Biến đổi dữ liệu (data transformation): chuẩn hoá dữ liệu (data normalization)
4. Thu giảm dữ liệu (data reduction): thu giảm kích thước dữ liệu (nghĩa là giảm số phần tử) bằng kết hợp dữ liệu (data aggregation), loại bỏ các đặc điểm dư thừa (redundant features) (nghĩa là giảm số chiều/thuộc tính dữ liệu), gom cụm dữ liệu

## So sánh kết quả của các thuật toán

|  |  | Thuật toán | | | |
| --- | --- | --- | --- | --- | --- |
| Độ đo | Tham số | MLP | SMO | IBK | J48 |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | Số mẫu phân lớp đúng | 50 | 49 | 49 | 49 |
| Số mẫu phân lớp sai | 1 | 2 | 2 | 2 |
| % Phân lớp đúng | 98.0392 % | 96.0784 % | 96.0784 % | 96.0784 % |
| % Phân lớp sai | 1.9608 % | 3.9216 % | 3.9216 % | 3.9216 % |
| Số lớp ẩn | 1 lớp ẩn có 3 nơ ron | Không có | | |
| Tốc độ học | 0,3 |
| Động lượng | 0,2 |
| batchsize | 100 |
| epoch | 500 |
| Ma trận nhầm lẫn | a b c <-- classified as  15 0 0 | a = Iris-setosa  0 19 0 | b = Iris-versicolor  0 1 16 | c = Iris-virginica | a b c <-- classified as  15 0 0 | a = Iris-setosa  0 19 0 | b = Iris-versicolor  0 2 15 | c = Iris-virginica | | |
|
|
|

KL: Thuật toán MLP cho kết quả phân lớp tốt nhất và chính xác nhất trên bộ dữ liệu iris. Ta thấy kết quả phân lớp qua ma trận nhầm lẫn có 15 mẫu phân lớp đúng vào lớp a, 19 mẫu phân lớp đúng vào lớp b và 16 mẫu phân lớp đúng vào lớp c nhưng chỉ có 1 mẫu của lớp c bị phân nhầm vào lớp b