TA: Dương Nguyễn Thái Bảo
duongbao9998@gmail.com

Lab 03 - Classification & Clustering

Data Mining - Term I/2020-2021

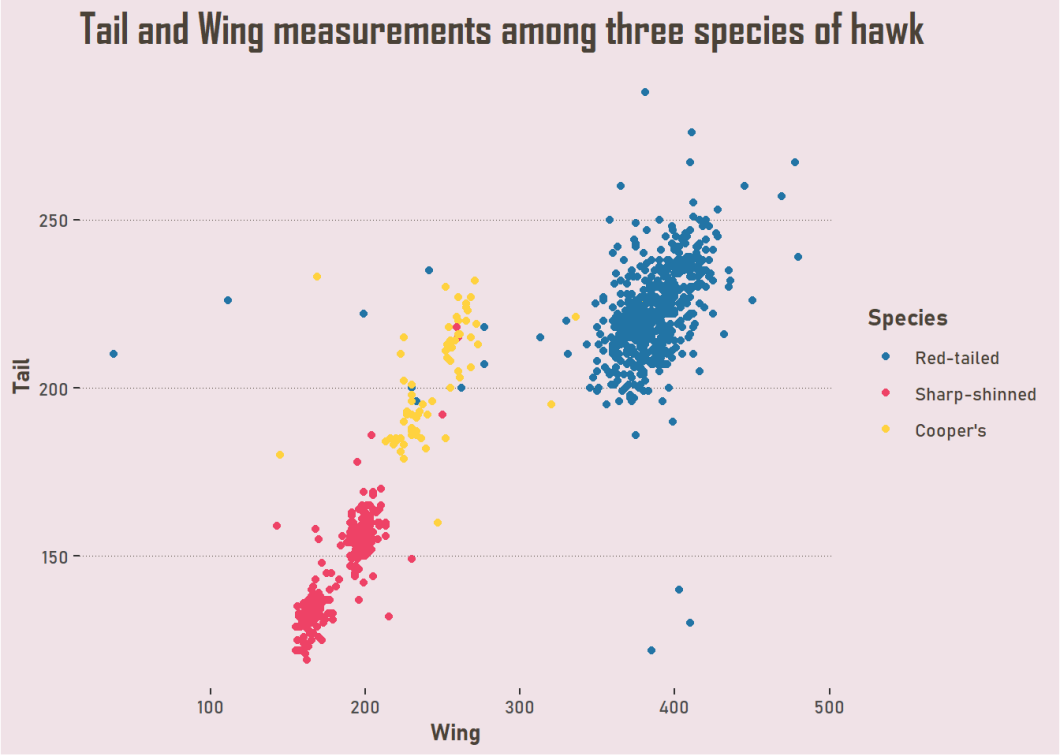
Mục tiêu của bài tập

- Trong bài tập này, sinh viên khảo sát các tiện ích phân lớp dữ liệu trong WEKA thông qua việc sử dụng hai chức năng Explorer và Experimenter. Dữ liệu sử dụng cho thực nghiệm trên tập dữ liệu "hawks".
- Bên cạnh đó, phát huy kỹ năng lập trình để tự cài đặt một số thuật toán gom cụm cơ bản.

Quy định

- Tối đa 2 thành viên/nhóm.
- Thời gian: 3 tuần (xem chi tiết trên Moodle).
- Thư mục bài làm: nếu nhóm có 1 sinh viên thì đặt tên là **<MSSV>**, nếu 2 sinh viên thì đặt tên là **<MSSV1>_<MSSV2>**, bao gồm các nội dung sau:
  - preprocess.py: file tiền xử lý dữ liệu (nếu cần).
  - Results.xls: chứa kết quả tóm tắt các lượt chạy trong thực nghiệm A-D.
  - RawResults.csv: chứa thông tin kết xuất khi chạy thực nghiệm D.
  - Observations.pdf: trả lời các câu hỏi và quan sát của nhóm.
  - Lab03-Clustering.ipynb: bài làm phần Clustering.
- Nén thư mục bài làm thành định dạng **zip** rồi nộp trên Moodle.
- **Bài làm giống nhau 0 điểm cả môn**.
- Ghi rõ nguồn tham khảo đầy đủ.

1 CLASSIFICATION



Thông tin chi tiết về tập dữ liệu hawks: Hawks: Measurements on Three Hawk Species in Stat2Data: Datasets for Stat2.

1.1 Phân lớp dữ liệu bằng Weka Explorer

Với mỗi thực nghiệm A-C bên dưới, sử dụng WEKA Explorer để tiến hành phân lớp dữ liệu bằng cách sử dụng các phương pháp phân lớp sau với tham số mặc định: 1) NaiveBayesSimple; 2) Id3; và 3) J48. Với mỗi phương pháp áp dụng trên tập dữ liệu, hãy sử dụng các phương pháp đánh giá sau (xem "Test options" trong cửa sổ "Classify" của Weka Explorer): a) "Use training set"; b) "Cross-validation" với 10 fold; và c) "Percentage split" với tỉ lệ 66%. Ghi nhận lại các kết quả của từng lượt chạy vào tập tin Excel "Result.xls", thông tin ghi nhận bao gồm:

(a) Loại thực nghiệm (A-C);

(b) Tên của tập tin dữ liệu đầu vào;

(c) Phương pháp phân lớp;

(d) Chiến lược đánh giá;

(e) Tỉ lệ mẫu được phân lớp đúng.

Các yêu cầu:

(A) Phân lớp dữ liệu trên tập sử dụng bốn phương pháp phân lớp và từng chiến lược đánh giá đã nêu bên trên.

(B) Rời rạc hóa mọi thuộc tính không phải là lớp trong tập dữ liệu thành 10 giỏ có độ rộng bằng nhau: sử dụng chức năng "Filter" trong cửa số "Preprocess" của Explorer, chọn 'filters' → 'unsupervised' → 'attribute' → 'Discretize'. Sử dụng tham số mặc định cho bộ lọc 'Discretize'. Sau khi đã bảo đảm được mọi thuộc tính không phải lớp đều là rời rạc, thực hiện phân lớp trên tập dữ liệu mới với 3 thuật toán phân lớp và từng chiến lược đánh giá đã nêu bên trên.

(C) Tiến hành rời rạc hóa mọi thuộc tính không phải là lớp trong tập dữ liệu thành 5 giỏ có độ sâu bằng nhau bằng cách chọn bộ lọc 'Discretize' và định tham số thích hợp. Sau khi đã bảo đảm được mọi thuộc tính không phải lớp đều là rời rạc, thực hiện phân lớp trên tập dữ liệu mới với 3 thuật toán phân lớp và từng chiến lược đánh giá đã nêu bên trên.

1.2 Phân lớp dữ liệu bằng Weka Experimenter

(D) Với thực nghiệm này, sử dụng WEKA Experimenter. Thực hiện phân lớp dữ liệu sử dụng các phương pháp phân lớp NaiveBayesSimple và J48 với tham số mặc định. Với từng phương pháp trên tập dữ liệu, chạy 10 lần phương pháp đánh giá cross validation với 10 fold. Ghi nhận lại kết quả vào tập tin "RawResult.csv". Từ những kết quả này, tính độ chính xác trung bình của mỗi phương pháp cho tập dữ liệu và thêm vào tập tin "Result.xls" (là tập tin trong thực nghiệm A-C) các thông tin sau: a) Thực nghiệm D; b) tên của tập dữ liệu, c) phương pháp phân lớp; d) chiến lược đánh giá; và e) tỉ lệ trung bình của các mẫu được phân lớp đúng sau 10×10 lượt chạy.

Để thực hiện các yêu cầu trên, bạn cần xử lý file .csv để Weka có thể đọc được tập dữ liệu. Bên cạnh đó tập dữ liệu cũng chứa missing values, do đó bạn cần tiền xử lý dữ liệu để có thể phân lớp (cách giải quyết tùy chọn, phần này không đặt nặng). Nếu các bạn tiền xử lý bằng python thì bạn nộp thêm file **preprocess.py** chứa cài đặt của 2 yêu cầu ở trên. Hoặc các bạn có thể tiền xử lý bằng Weka rồi báo cáo thêm các bước tiền xử lý của mình.

1.3 Đánh giá

Sau khi đã tiến hành thực nghiệm, ta cần bỏ một ít thời gian để đánh giá kết quả thu được. Một cách cụ thể, ít ra ta cũng phải trả lời được những câu hỏi sau:

- Phương pháp phân lớp nào thường cho kết quả cao nhất?
- Phương pháp nào không thực hiện tốt và tại sao?
- Tại sao ta sử dụng phiên bản đã rời rạc hóa của tập dữ liệu nếu tập dữ liệu đã được rời rạc hóa?
- Việc rời rạc hóa và cách rời rạc hóa có ảnh hưởng đến kết quả phân lớp hay không, nếu có thì ảnh hưởng thể nào?
- Chiến lược nào trong ba chiến lược đánh giá đã đánh giá quá cao (overestimate) độ chính xác và tại sao?
- Chiến lược nào đánh giá thấp (underestimate) độ chính xác và tại sao?

Trình bày những điều này và các quan sát khác vào tập tin "Observations.doc".

2 CLUSTERING

Trong phần này các bạn hãy làm quen với công cụ **Jupyter notebook**, sau đó mở file **Lab04-Clustering.ipynb** lên và làm theo hướng dẫn trong file.