Ho Chi Minh University of Science
Information Technology (HQ Program)

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

PROJECT 2 – Classification and Clustering

Lớp: 19KHMT – **Nhóm:** 17

December 26, 2021

PROGRESS: 100%

GIẢNG VIÊN

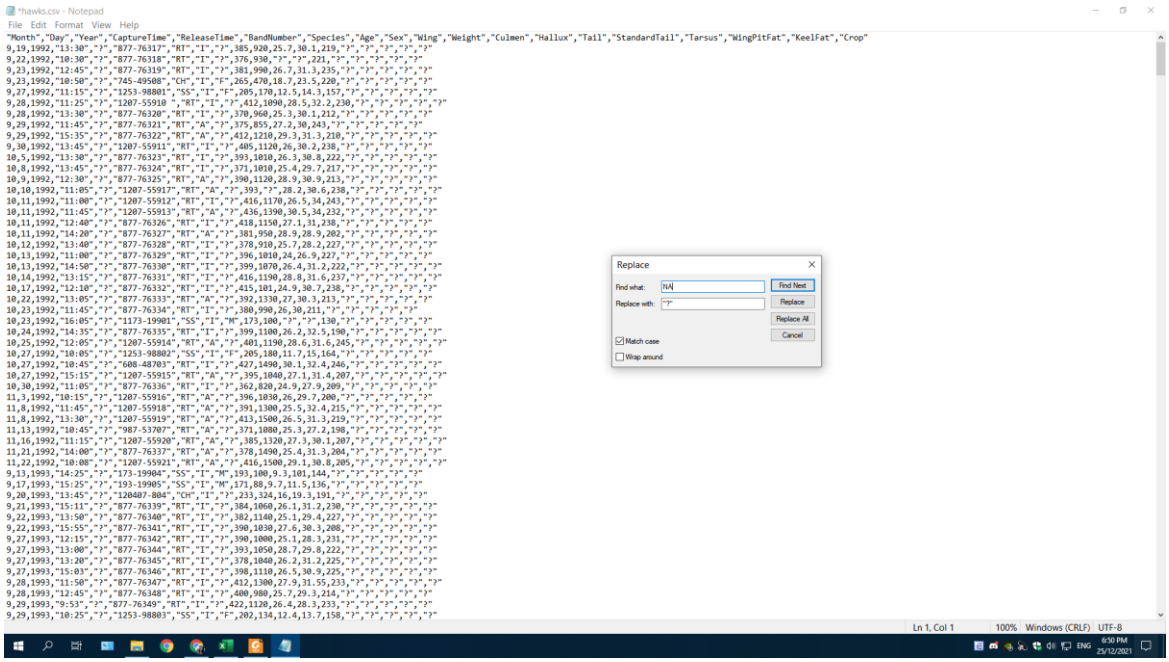
Lê Hoài Bắc

Nguyễn Khánh Toàn

Project 2	Classification			Clustering	Report	Progress
	Explorer	Experimenter	Observation			
Ngô Văn Anh Kiệt 19127191	x	x	x		x	100%
Triệu Nguyên Phát 19127505				x	x	100%

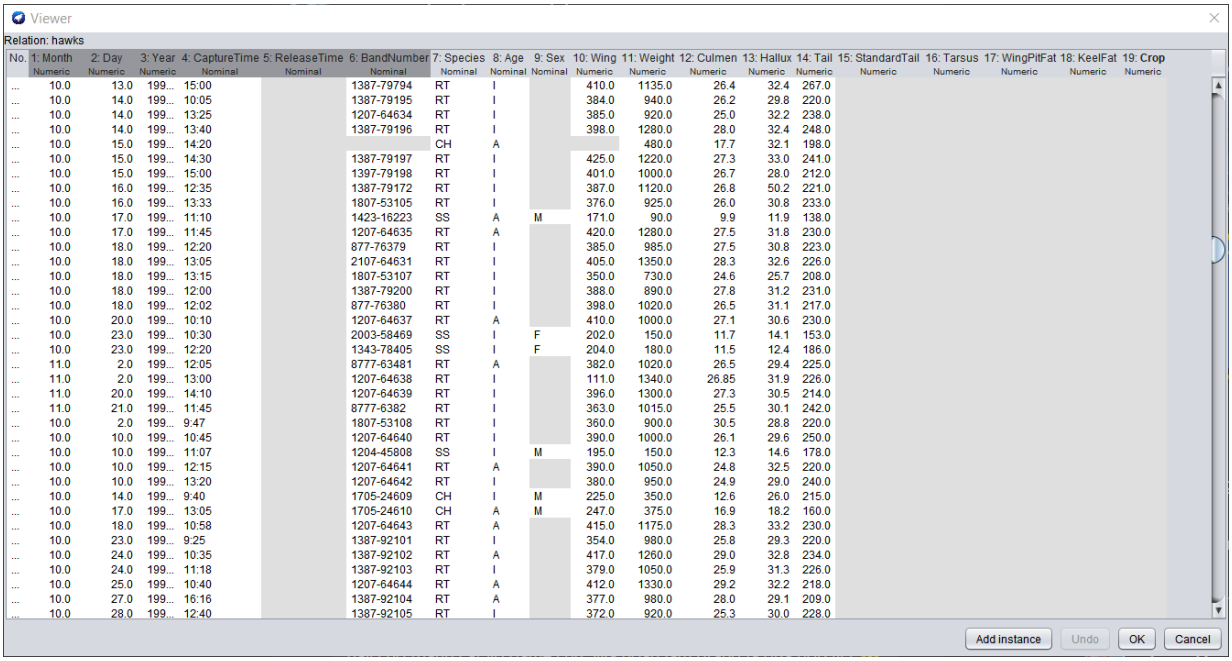
Preprocessing hawks dataset with Weka

- **Bước 1**: Bởi vì các dữ liệu bị thiếu trong file “hawks.csv” được ghi bằng chuỗi “NA”, “” hoặc “ “ nên ta cần phải chuyển về cùng một định dạng đánh dấu dữ liệu bị thiếu mà Weka có thể đọc được là “?”.
  Do file data này tương đối nhẹ, ta chỉ cần bật file csv bằng Notepad (Windows 10) hoặc các trình soạn thảo văn bản tương tự để dùng chức năng Replace All thay đổi tất cả chuỗi không hợp định dạng thành “?”.



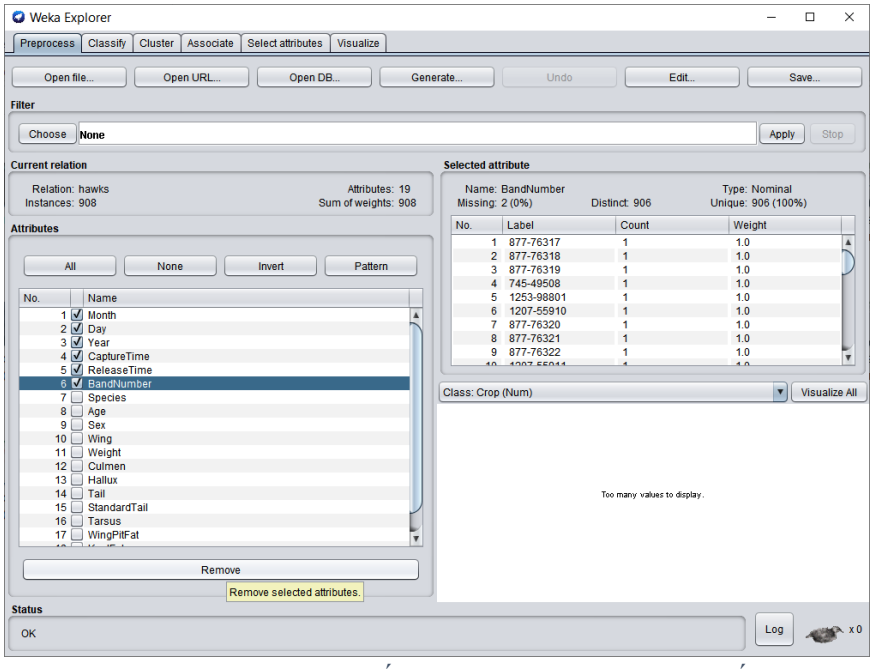
Hình 1. Thay đổi các giá trị bị thiếu từ NA thành "?"

- **Bước 2**: Ta mở file csv bằng Weka Explorer, vào giao diện Edit của tab Preprocessing để kiểm tra lại các cột có đúng kiểu dữ liệu được mô tả trong trang web nguồn của tập dữ liệu hawks.

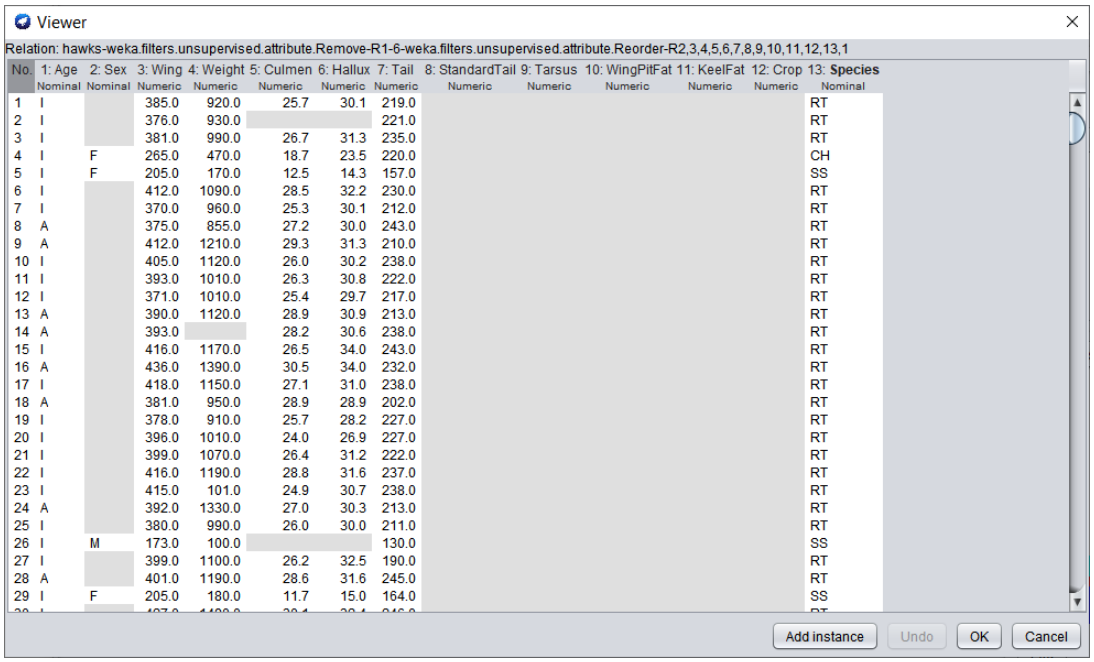


Hình 2. Các thuộc tính đúng kiểu dữ liệu được mô tả và thấy rõ các cột thiếu dữ liệu

- **Bước 3**: Nhận thấy trong các thuộc tính của tập dữ liệu hawks, có một vài thuộc tính không mang nhiều ý nghĩa cho việc phân loại chủng diều hâu, ví dụ như: ngày, tháng, năm ghi dữ liệu, thời điểm bắt, thả chim, mã số của chim. Ta sẽ loại bỏ các thuộc tính đó ra khỏi tập dữ liệu cần dùng, đồng thời đặt thuộc tính "Species" là thuộc tính lớp cần phân loại.

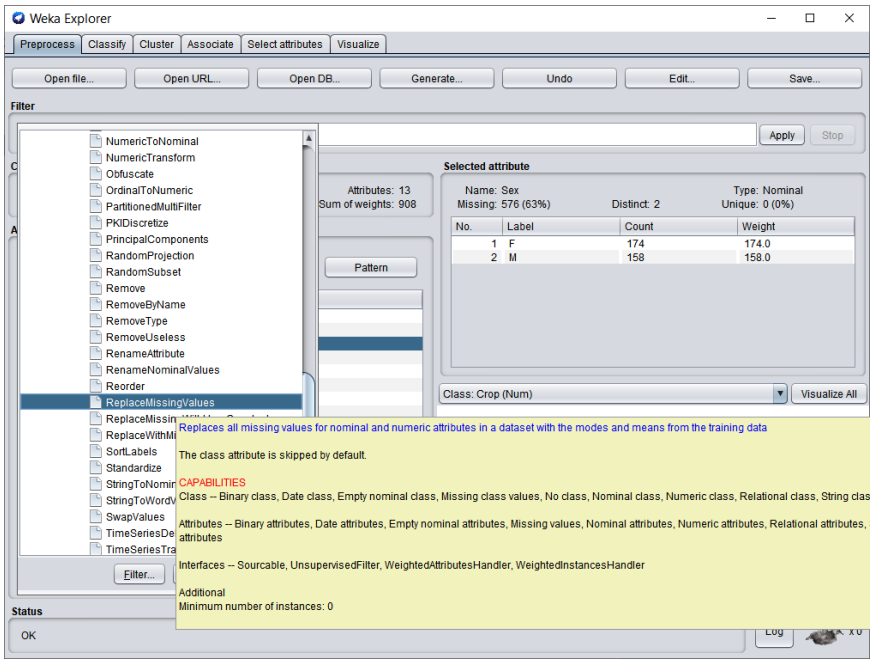


Hình 3. Loại bỏ một số thuộc tính khỏi tập dữ liệu gốc



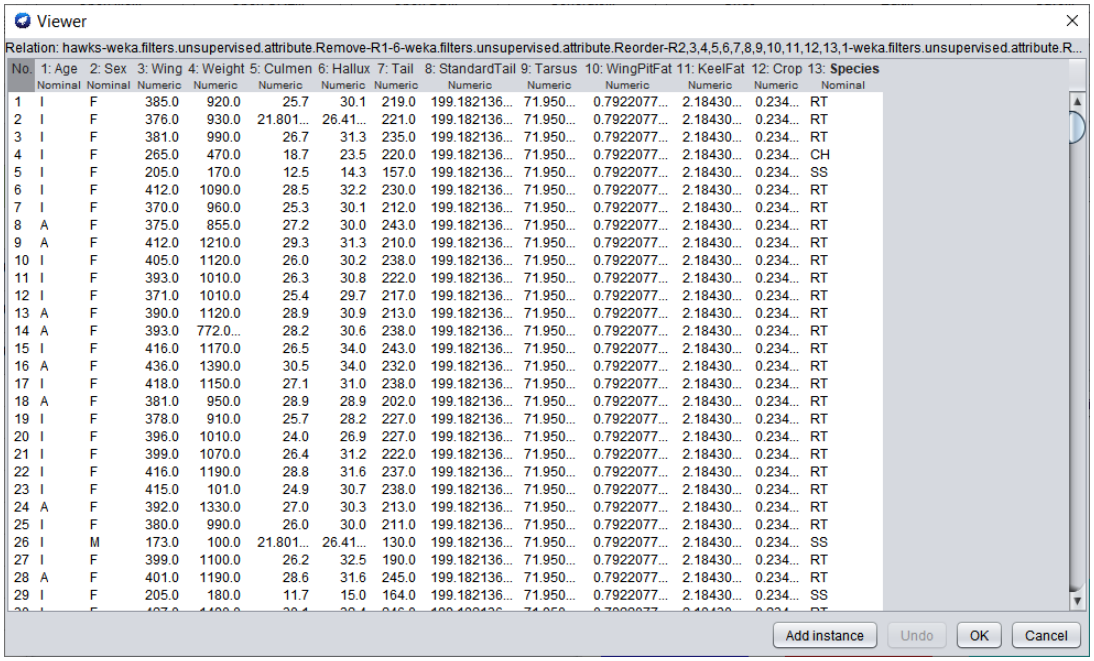
Hình 4. Đặt Species làm class

- **Bước 4**: Bây giờ, ta bắt đầu tiền xử lý các dữ liệu bị thiếu. Sử dụng bộ lọc weka.filters.unsupervised.attribute.ReplaceMissingValues để thay thế các dữ liệu bị thiếu bằng giá trị mean của thuộc tính (nếu là kiểu số) hoặc mode (nếu là kiểu định danh).



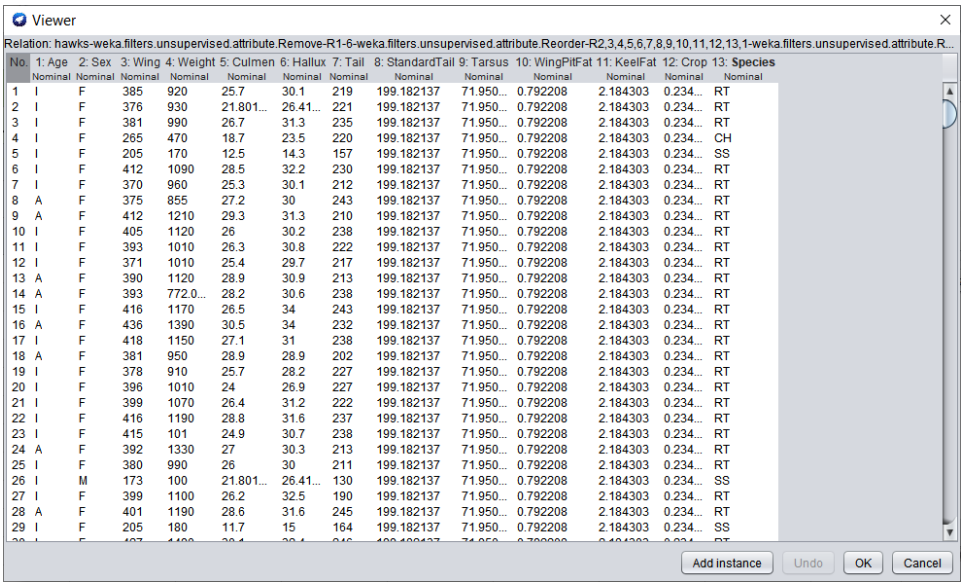
Hình 5. Sử dụng filter ReplaceMissingValues để điền dữ liệu thiếu

Như vậy, ta có bảng dữ liệu sau. Lưu vào file “preprocessed_hawks.arff”



Hình 6. Tập dữ liệu preprocessed_hawks.arff

- **Bước 5**: Do trong thí nghiệm A không sử dụng tập dữ liệu kiểu rời rạc, trong khi thuật toán ID3 thì lại không hoạt động được với dữ liệu liên tục, ta dùng bộ lọc weka.filters.unsupervised.attribute.NumericToNominal để tạm biến đổi các thuộc tính số thành thuộc tính định danh để có thể chạy ID3 cho phần này. Lưu tập dữ liệu này thành file "preprocessed_numeric2nominal_hawks.arff". Vì tập này chỉ được dùng tạm để chạy ID3 trong thí nghiệm A, kết quả của nó sẽ được ghi cùng dòng với tập "preprocessed_hawks.arff".



Hình 7. Thuộc tính numeric biến đổi thành nominal để chạy ID3 trong thí nghiệm A

Observation

Which classification method typically has the best result?

Phương pháp phân lớp thường cho kết quả tốt nhất là J48. Theo sau là Naïve Bayes Simple và cuối cùng là ID3.

Which method does not work well and why?

Phương pháp phân lớp **ID3** cho kết quả không tốt, bởi vì tập dữ liệu ban đầu có rất nhiều thuộc tính kiểu số liên tục, trong khi ID3 chỉ hoạt động với kiểu rời rạc. Nếu ta dùng bộ lọc của Weka để chuyển kiểu số thành kiểu định danh (ở thí nghiệm A) hoặc làm rời rạc hóa dữ liệu với độ rộng của bin không đủ lớn (thí nghiệm B), ID3 sẽ bị overfitting rất nặng (kết quả luôn là 100% khi phương pháp test là dùng tập huấn luyện) hoặc fit không tốt như các thuật toán khác. Thuật toán ID3 phụ thuộc rất nhiều vào việc dữ liệu có rời rạc hay không, hoặc phân phối của dữ liệu khi chia bin bằng cách rời rạc hóa. Ngoài ra, do tập dữ liệu ban đầu bị thiếu rất nhiều dữ liệu nên khi rời rạc hóa, các điểm dữ liệu đó sẽ bị trùng nhau, dẫn đến việc chia rổ theo độ rộng trong rời rạc hóa sẽ khiến cho các điểm đó nằm cùng 1 rổ, gây ảnh hưởng rất lớn đến hiệu năng của thuật toán.

Why should we use the discretized version of the data set instead of the original one?

Việc rời rạc hóa các thuộc tính số sẽ giúp các thuộc tính được mô tả đúng với ý nghĩa của nó hơn. Ngoài ra, hiệu quả của các thuật toán như ID3 hay Naïve Bayes Simple phụ thuộc vào sự rời rạc của dữ liệu, nên ta cần phải thực hiện rời rạc hóa dữ liệu.

Do the discretization process and method affect the classification results? If yes then how?

Qua thí nghiệm B và C, ta nhận thấy việc rời rạc và cách rời rạc **có ảnh hưởng** đến kết quả phân lớp. Tuy nhiên, mức độ ảnh hưởng tùy theo từng phương pháp phân lớp.

Ở thí nghiệm A, do không rời rạc hóa dữ liệu nên ID3 không chạy tốt như các phương pháp khác. Ở thí nghiệm B và C thì ID3 đã làm tốt hơn, tuy nhiên thí nghiệm B vẫn bị overfitting khi thử nghiệm bằng tập huấn luyện, do tập dữ liệu gốc thiếu quá nhiều dữ

liệu nên các dữ liệu thiếu được điền bằng các giá trị giống nhau, dẫn đến việc rời rạc hóa theo độ rộng bin mất cân đối.

Thuật toán Naïve Bayes Simple và J48 cũng bị ảnh hưởng bởi việc rời rạc hóa, nhưng không nặng như ID3. Bởi vì Naïve Bayes Simple chỉ quan tâm đến xác suất của các lớp dữ liệu, còn J48 của Weka có thể tự rời rạc hóa dữ liệu trong quá trình chạy (nhưng không đảm bảo tốt hơn việc rời rạc hóa dữ liệu trước khi chạy).

Which evaluation strategy tends to overestimate the accuracy and why?

Phương pháp đánh giá thường hay đưa ra độ chính xác cao hơn là “Use training set”. Bởi vì chúng ta đã dùng tập dữ liệu huấn luyện để cho bộ phân loại học, nên hiển nhiên là bộ phân loại sẽ biết được nhãn của các điểm dữ liệu trong tập huấn luyện đó. Chính vì thế nên khi đánh giá mô hình, ta cần phải dùng một tập dữ liệu khác hoàn toàn với tập dữ liệu dùng để huấn luyện.

Which evaluation strategy tends to underestimate the accuracy and why?

Phương pháp đánh giá thường hay cho ra độ chính xác thấp nhất là “Percentage split 66%”. Điều này là do tập dữ liệu gốc bị thiếu quá nhiều dữ liệu, nên sau khi điền các giá trị bị thiếu, sẽ có rất nhiều điểm dữ liệu bị trùng. Sự thiếu đa dạng dữ liệu khiến cho việc tách bộ dữ liệu gốc gây ảnh hưởng nặng đến quá trình huấn luyện và thử nghiệm.