BÀI TẬP THỰC HÀNH 2
PHÂN LỚP

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 bao gồm "contact-lenses", "iris", và "soybean", ba tập dữ liệu này được cung cấp sẵn trong gói phần mềm Weka và có định dạng .arff.

- Contact-lenses: 24 mẫu, 5 thuộc tính đều rời rạc, thuộc tính cuối cùng là thuộc tính phân lớp.
- Iris: 150 mẫu, 4 thuộc tính liên tục và 1 thuộc tính phân lớp dạng rời rạc
- Soybean: 683 mẫu, 36 thuộc tính đều rời rạc, thuộc tính cuối cùng là thuộc tính phân lớp. Khác với 2 tập dữ liệu đầu, "soybean"

có chứa giá trị thiếu.

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;**. Với mỗi phương pháp áp dụng trên từng 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; và e) tỉ lệ mẫu được phân lớp đúng.

A. Phân lớp dữ liệu trên tập "contact-lenses.arff" 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 "iris.arff" thành 10 giỏ có độ rộng bằng nhau: sử dụng chức năng "Filter" trong cửa sổ "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 "iris.arff" 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 các thuật toán phân lớp và từng chiến lược đánh giá đã nêu bên trên.

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 "iris"? 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 Word "Observations.doc".

Qui định nộp bài

- Hình thức: thư mục bài làm có tên là **<MSSV>**, bao gồm:
  - Results.xls: chứa kết quả tóm tắt các lượt chạy trong thực nghiệm A-C.
  - Observations.doc: trả lời các câu hỏi và quan sát của sinh viên

Tài liệu tham khảo

[1] Slide lý thuyết bài 4

[2] Textbook: J. Han and M. Kamber: Data Mining, Concepts and Techniques, Second Edition - Chapter 6: Classification and Prediction (6.3.1 – 6.3.3, 6.4.1 – 6.4.2, 6.5.1 – 6.5.3, 6.12.1).