LAB03 – Classification

Mục tiêu của bài tập

- Sử dụng công cụ WEKA để khảo sát thực nghiệm về hiệu quả của các giải thuật phân lớp trên nhiều tập dữ liệu khác nhau
- Nâng cao năng lực lập trình thông qua việc tự cài đặt giải thuật phân lớp cơ bản

Quy định

- Thời gian thực hiện: **2 tuần** (xem ngày cụ thể trên Moodle)
- Tổ chức thư mục bài làm: thư mục có tên là **⟨tên nhóm⟩** (nếu tên nhóm có dấu và khoảng trắng thì bỏ dấu và viết dính liền), bao gồm các tài liệu sau
    - Báo cáo viết trả lời các câu hỏi tự luận và hướng dẫn sử dụng chương trình tự cài đặt, định dạng **pdf**. Trang đầu tiên ghi rõ thông tin nhóm, tỉ lệ thực hiện bài tập của mỗi thành viên (nếu không ghi, mặc định tỉ lệ tương đương), những phần chưa thực hiện được (để giáo viên tránh chấm sót).
    - Dữ liệu thu được trong quá trình thực nghiệm (nếu có)
    - Mã nguồn của chương trình tự cài đặt (nếu có). Ngôn ngữ: **C++/Python**, không chấp nhận các ngôn ngữ khác.
- Nén thư mục bài làm theo định dạng **zip** hoặc **rar** và nộp theo link Moodle
- **Bài làm cần tuân thủ nghiêm ngặt đặc tả của đề bài, mọi sự khác biệt sẽ không được xét tính điểm.**
- Bài làm được đánh giá trên thang 15 rồi quy đổi về tỉ lệ tương ứng điểm thực hành.

Dữ liệu thực nghiệm

Bài tập này làm việc trên ba tập dữ liệu, bao gồm

- mushroom.arff: 812 mẫu, 23 thuộc tính, mỗi mẫu gồm 22 đặc trưng hình dạng và đặc tính sinh học của một loài nấm để phân loại nấm độc hay nấm ăn được.
- zoo.arff: 812 mẫu, 17 thuộc tính, mỗi mẫu gồm 16 đặc trưng hình dạng và đặc tính sinh học của một động vật và phân loại vào một trong 7 lớp động vật.
- letter.arff: 16000 mẫu, 17 thuộc tính, mỗi mẫu gồm 16 thông số hình học của một chữ cái trong bộ chữ cái tiếng Anh 26 chữ.

Các tập dữ liệu được cung cấp trên Moodle kèm với đề bài. Ngoài ra, sinh viên có thể truy cập dữ liệu gốc tại trang web của UCI: https://archive.ics.uci.edu/ml/datasets.php

Nội dung thực hiện báo cáo với ứng dụng WEKA

Đọc tập dữ liệu **mushroom.arff** vào WEKA Explorer tại tab Preprocess. Nhiệm vụ khai thác dữ liệu là xác định một mẫu là nấm ăn được hay có độc từ các đặc trưng của mẫu.

Tại tab Classify, thử nghiệm các bộ phân lớp sau

- Logistic Regression (LR) (classifiers.functions.Logistic).
- J48 (classifiers.trees.J48): cây quyết định C4.5 có/không tỉa nhánh.
- IBk (classifiers.lazy.IBk): k-Nearest Neighbor. Thay đổi số lượng láng giềng lần lượt là **KNN = 1** và **KNN = 4**.

Sử dụng **10-fold cross-validation** để ước lượng hiệu quả của mỗi giải thuật.

1. (3.0đ) Thiết lập bảng thống kê độ chính xác phân lớp của mỗi giải thuật trên dữ liệu được cho. Các thông số này được lấy từ phần văn bản trong cửa số Classifier output.

Giải thuật	Accuracy	Detailed Accuracy By Class							
		Class edible				Class poisonous			
		TP Rate	FP Rate	Precision	Recall	TP Rate	FP Rate	Precision	Recall
LR									
J48									
IBk (KNN=1)									
IBk (KNN=4)									

2. (2.0đ) Giả sử WEKA cho báo cáo hiệu quả của hai giải thuật A và B như sau

Giải thuật A		
Accuracy: 99.8768%		
Class	TP Rate	FP Rate
edible	1.0	0.003
poisonous	0.997	0.0

Giải thuật B		
Accuracy: 99.8768%		
Class	TP Rate	FP Rate
edible	0.998	0
poisonous	1	0.002

Cả hai giải thuật có cùng độ chính xác nhưng TP Rate và FP Rate của chúng khác nhau. Đối với bài toán **nhận diện nấm độc** thì theo bạn giải thuật nào tốt hơn, giải thuật A, giải thuật B hay cả hai giải thuật đều như nhau? Giải thích sự lựa chọn của bạn.

Đọc tập dữ liệu **zoo.arff** vào WEKA Explorer tại tab Preprocess. Áp dụng phương pháp cây quyết định J48 trong trường hợp có tỉa nhánh và không tỉa nhánh. Thay đổi tham số **numFolds từ 2 đến 10.**

3. (2.0đ) Vẽ đồ thị thể hiện độ chính xác phân lớp (trục tung) theo sự biến thiên của numFolds (trục hoành), xét cả trường hợp có tỉa nhánh và không tỉa nhánh.
4. (1.0đ) Tham số numFolds có vai trò gì trong J48? Mô tả sự tác động của tham số này đến hiệu quả của cây quyết định xây dựng được khi thay đổi giá trị tham số.
5. (2.0đ) Dựa vào đồ thị hãy bình luận về hiệu quả tỉa nhánh giảm lỗi trên cây quyết định thu được. Tại sao tỉa nhánh giúp cây quyết định cải thiện độ chính xác (tức là, hiện tượng máy học nào đang xảy ra)?

Đọc tập dữ liệu **letter.arff** vào WEKA Explorer tại tab Preprocess. Bạn cần tìm ra mô hình phân lớp tốt nhất trên dữ liệu được cho. Bạn được phép chọn bất kỳ giải thuật phân lớp nào mà WEKA hỗ trợ (chứ không giới hạn trong những phương pháp đã học) và bất kỳ chiến lược đánh giá hiệu quả giải thuật nào (cross validation, percentage split, v.v.). Ngoài ra, với giải thuật đã chọn, bạn cũng cần thử các bộ tham số khác nhau vì một số bộ phân lớp hoạt động rất đa dạng tùy cấu hình tham số.

6. (1.0đ) Đánh giá độ chính xác của mô hình được chọn (tức là mô hình có hiệu quả thực nghiệm tốt nhất) bằng 10-fold cross validation và báo cáo các giá trị độ đo như trong bảng ở Câu 1.
7. (2.0đ) Mô tả ngắn gọn về mặt lý thuyết giải thuật được chọn để xây dựng mô hình. Vì sao bạn cho rằng đây là giải thuật tốt nhất so với những giải thuật khác?
8. (2.0đ) Báo cáo bộ tham số được dùng để thực nghiệm. Cấu hình tham số ảnh hưởng như thế nào đến hiệu quả của giải thuật?

Giáo viên sẽ sử dụng một tập dữ liệu độc lập để đánh giá mô hình được chọn dựa vào bộ tham số đã báo cáo. 3 nhóm có độ chính xác cao nhất sẽ được cộng 2.0đ.

Tài liệu tham khảo

[1] Slide bài giảng lý thuyết lý thuyết

[2] Trang chủ của WEKA: http://www.cs.waikato.ac.nz/ml/weka/

[3] Textbook: J. Han and M. Kamber: Data Mining, Concepts and Techniques, Second Edition - Chapter 6: Classification and Prediction.

[4] I. H. Witten and E. Frank: Data mining, Practical Machine Learning Tools and Techniques