Lab03- Classification & Clustering

Data Mining -Term I /2020-2021

GVHD : TA Dương Nguyễn Thái Bảo

Thông tin thành viên

Họ Tên	MSSV
Trần Ngọc Tịnh	18120597
Nguyễn Ngọc Năng Toàn	18120600

1 Báo cáo về phần tiền xử lí

Sử dụng các kiến thức ở Lab1 để tiền xử lí dữ liệu cho phù hợp & có thể thích hợp để chạy trên Weka

Ở file data.csv chúng ta chỉ xử lí làm sạch dữ liệu

1. *Xóa các cột bị thiếu giá trị thuộc tính ở hơn 50% số mẫu .*
2. *Điền các giá trị thiếu .*
3. *Xóa các mẫu bị trùng lặp .*
4. *Điền các giá trị thiếu bằng mean (cho thuộc tính numeric) và mode cho thuộc tính categorical .*
5. *Thiết lập cột Species làm cột cuối cùng để Weka tự hiểu là thuộc tính lớp .*

Ở file data1.csv chúng ta sau khi làm sạch dữ liệu ta sẽ phân lớp để có thể chạy được các thuật toán yêu cầu không có thuộc tính có giá trị là numeric

1. *Xóa các cột bị thiếu giá trị thuộc tính ở hơn 50% số mẫu .*
2. *Điền các giá trị thiếu .*
3. *Xóa các mẫu bị trùng lặp .*
4. *Điền các giá trị thiếu bằng mean (cho thuộc tính numeric) và mode cho thuộc tính categorical .*

5. *Thiết lập cột Species làm cột cuối cùng để Weka tự hiểu là thuộc tính lớp .*
6. *Rời rạc hóa các giá trị numeric cụ thể như sau :*
   + Chuyển thuộc tính month từ numeric thành object . vd : 1 -> Jan , 2-> Feb …
   +Chuyển thuộc tính day từ numeric thành object bằng cách dùng Tuần . vd 1 -> Week1 , 27 -> Week 4
   +Chuyển Year từ số thành chuỗi bằng cách thêm chữ Year phía trước . VD : 1992 -> Year-1992…
   +Các thuộc tính còn lại Wing , Weight , Culmen , Hallux , Tail , StandardTail , KeelFat , Crop ta sẽ rời rạc từng thuộc tính thành các phần bằng nhau .

2 Báo cáo về phần đánh giá

Cảm nhận , quan sát của bản thân :

Phân lớp là hình thức phân tích dữ liệu để rút ra các mô tả các lớp dữ liệu quan trọng .Không có thuật toán nào vượt trội nhất cho mọi tập dữ liệu. Mỗi thuật toán có mỗi phạm vi ứng dụng vào các cơ sở dữ liệu có kích thước và loại dữ liệu khác nhau . Trong thực tế ta có thể phải cần áp dụng nhiều thuật toán để chọn ra mô hình phù hợp nhất cho bài toán của mình .

Dựa vào kết quả ở bảng Result.xls , trả lời các câu hỏi yêu cầu :

- ***Phương pháp phân lớp nào thường cho kết quả tốt nhất ?***

  Trả lời : *Phương pháp J48 thường cho kết quả tốt và ổn định ở 4 loại thực nghiệm A , B , C , D và 2 tập dữ liệu data.csv,data1.csv và các dữ liệu thu được sau khi rời rạc hóa data.csv nhờ vào filters của Weka .*

- *Phương pháp nào không thực hiện tốt và tại sao ?*

  Trả lời : *2 Phương pháp NaiveBayesSimple và ID3*

  + Phương pháp NaiveBayesSimple thực hiện tốt trên tập dữ liệu data.csv (là tập dữ liệu

chỉ mới giải quyết missing + xóa vài cột => có rất nhiều thuộc tính numeric) nhưng lại cho kết quả khá thấp rõ rệt khi sử dụng tập dữ liệu data1 (Rời rạc hóa bằng python) và dữ liệu data với các thực nghiệm B , C . Kết luận : Phương pháp NaiveBayesSimple thực hiện tốt trên các thuộc tính liên tục và độ chính xác càng giảm khi khái quát hóa dữ liệu lên mức cao hơn .

+ Phương pháp ID3 không thực hiện được trên dữ liệu khi có thuộc tính là numeric . Thoạt nhìn thì thấy phương pháp này khá hiệu quả khi ở phương pháp đánh giá UsingTrainingTest có độ chính xác 100% , nhưng 2 phương pháp đánh giá còn lại thì độ chính xác lại thấp 1 cách rõ rệt . Và phương pháp đánh giá UsingTrainingTest là một phương pháp đánh giá Overestimate mà mình sẽ trình bày ở phía câu hỏi thứ 5 .

- *Tại sao ta sử dụng phiên bản đã rời rạc hóa của tập dữ liệu nếu tập dữ liệu đã được rời rạc hóa ?*

  Trả lời : *Việc khái quát hóa dữ liệu lên mức khái niệm cao hơn đôi khi là cần thiết trong quá trình tiền xử lý . Việc này đặc biệt hữu ích với những thuộc tính liên tục . Việc khái quát hóa làm cô đọng giá trị nguyên thủy , vì vậy các thao tác vào ra liên quan đến quá trình học sẽ giảm .*

- *Việc rời rạc và cách rời rạc có ảnh hưởng đến cách quả phân lớp hay không , nếu có thì ảnh hưởng như thế nào ?*

  Trả lời : *Có . Ở file data1.csv là tập dữ liệu do em rời rạc bằng tay , em đã rời rạc thành các giỏ (trung bình chỉ có 3-5 giỏ) quá ít do đó kết quả sau khi chạy có tỉ lệ mẫu phân lớp đúng bị giảm.*

  Việc rời rạc hóa tập dữ liệu một cách hợp lí sẽ cho kết quả tốt về mặt thời gian lẫn độ chính xác. Việc rời rạc tập giá trị giúp việc thao tác trở nên dễ dàng hơn .

- *Chiến lược nào trong ba chiến lược đánh giá đã quá cao (Overestimate) độ chính xác và tại sao ?*

  Trả lời : *Chiến lược Using Training Test đã đánh giá quá cao độ chính xác . Do đã sử dụng tập training và tập test chung , đây là một cách đánh giá khá tệ . Ví dụ : Một người đi thi đã gặp các câu hỏi giống 100% khi ở nhà làm bài , đưa một bài toán tương tự thì người đó sẽ có thể không làm được .*

- *Chiến lược nào đánh giá thấp (underestimate) độ chính xác và tại sao ?*

  Trả lời : *Chiến lược đánh giá Percentage Split với 66% đã đánh giá thấp độ chính xác vì thuật toán này sẽ chia tập dữ liệu thành 2 phần , tập huấn luyện và tập kiểm thử theo tỉ lệ % . Mà ở đây phần tập huấn luyện chỉ chiếm 66% (quá thấp) và không cần thiết phải cho phần trăm tập kiểm thử tới 34 % => Thuật toán không được tối ưu độ chính xác vì thiếu dữ liệu .*

Tài liệu tham khảo :

http://uet.vnu.edu.vn/~thuyhq/Student_Thesis/K46_Nguyen_Thi_Thuy_Linh_Thesis.pdf

https://viblo.asia/p/thuat-toan-phan-lop-naive-bayes-924lJWPm5PM

https://ongxuanhong.wordpress.com/2015/08/25/ap-dung-cac-phuong-phap-phan-lop-classification-tren-tap-du-lieu-mushroom/

https://ongxuanhong.wordpress.com/2015/08/25/danh-gia-mo-hinh-model-evaluation/