LAB03 – Classification

Mục tiêu của bài tập

- Trải nghiệm tác vụ phân lớp dữ liệu bằng cách áp dụng các giải thuật phân lớp được hỗ trợ bởi công cụ WEKA và tự cài đặt giải thuật ID3.
- Rèn luyện kỹ năng phân tích dữ liệu thông qua việc tiến hành thực nghiệm và nhận xét trên kết quả thu được.

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 **⟨ID nhóm⟩** , 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. Ngôn ngữ: **Python**. Các ngôn ngữ khác tối đa được 80% điểm.
- 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 điểm 20 rồi quy đổi về tỉ lệ 30% điểm thực hành.

A – Nội dung thực hiện báo cáo viết (10 điểm)

Dữ liệu thực nghiệm

Sinh viên làm việc trên tập dữ liệu gồm các bài viết đăng trên Twitter (còn gọi là tweet). Tập dữ liệu này được trích từ nguồn "ngữ liệu Edinburgh Twitter" trong công trình khoa học của Petrovic và cộng sự (2010). Tweets được sử dụng rộng rãi trong bài toán phân tích ý kiến (sentiment analysis), và nhà kiến tạo ngữ liệu đã đưa ra lý do như sau: "Dịch vụ blog mini Twitter đã và đang trở thành công cụ phổ biến để thể hiện ý kiến, loan truyền tin tức hay đơn giản là liên lạc với bạn bè. Người ta thường bình luận về những sự kiện trong thời gian thực, với hàng trăm bài viết nhỏ (tweets) được đăng mỗi giây cho những sự kiện quan trọng." (Petrovic và cộng sự, 2010). Bài thực hành này chỉ thao tác trên một tập dữ liệu con nhỏ hơn của toàn bộ ngữ liệu, vốn được tạo ra nhằm mục đích học tập. Tập dữ liệu nhỏ ở định dạng ARFF, bao gồm 100 tweets dương (positive) và 100 tweet âm (negative).

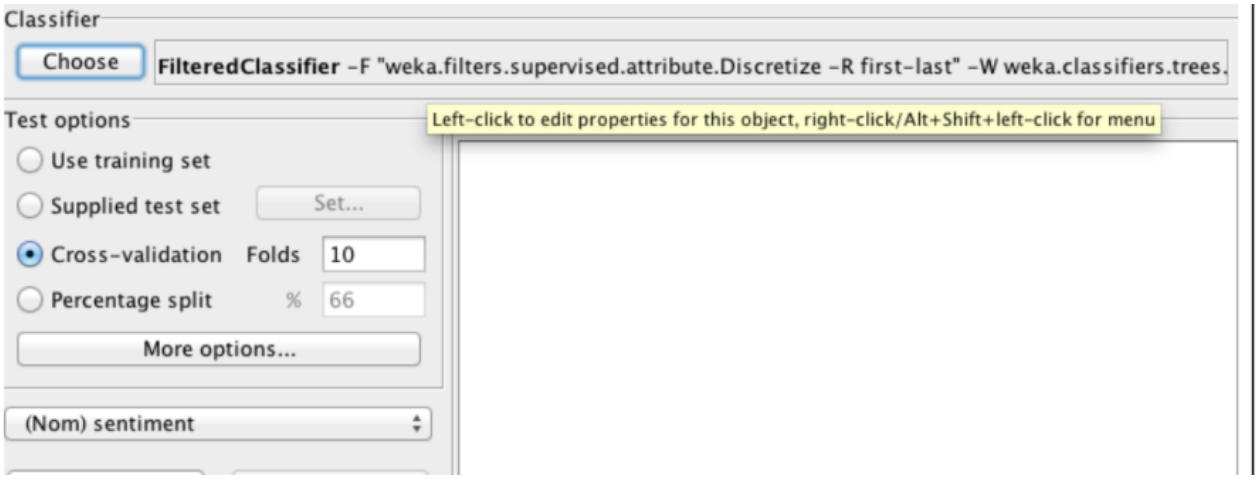
Tập dữ liệu gồm hai thuộc tính, tweet_body (kiểu dữ liệu string, chứa nội dung văn bản của mỗi tweet) và class (kiểu dữ liệu nominal, mang một trong hai giá trị, dương (postive, pos) hoặc âm (negative, neg)).

```
@relation tweets

@attribute tweet_body string
@attribute sentiment {pos,neg}

@data
'anyone feel motivated the fri afternoon prior to a holiday? wanted to get lots do
&lt;3 her ',pos
'seriously, do you have to rub it in maggie!!!!  ',pos
'if i\'m not wrong.. Alt is when image can\'t be displayed.. Tooltip is the \'titl
'I don\'t like social karma  much. Would rather skip it, but can\'t afford to piss
'I\'d be happy to review the Iomega if EMC send me one!!! ',pos
'Something  I have wanted to make for a while now... finally done   URL',pos
'you\'re so sweet!  proving me right again...  the Dutch are the Best!   ',pos
'Look for us on the back of your Pepsi can! Our Pepsi can offer hit\nstores today!
'well rob i have to admit that you have to admit that you feel cool for being on t
'big skype call at 1010! msg me if you want in ',pos
```

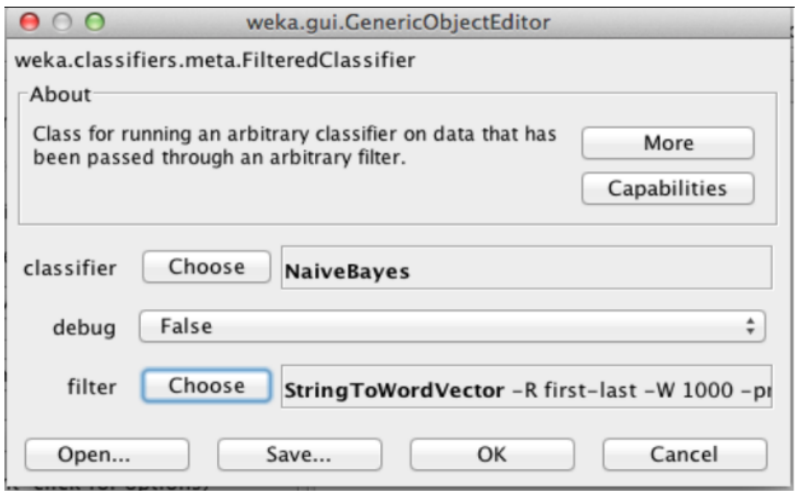
Tải dữ liệu tại link bên dưới hoặc sử dụng tập tin cung cấp sẵn trên hệ thống Moodle.

http://stp.lingfil.uu.se/~santinim/sais/2016/100pos_100neg_tweets.arff

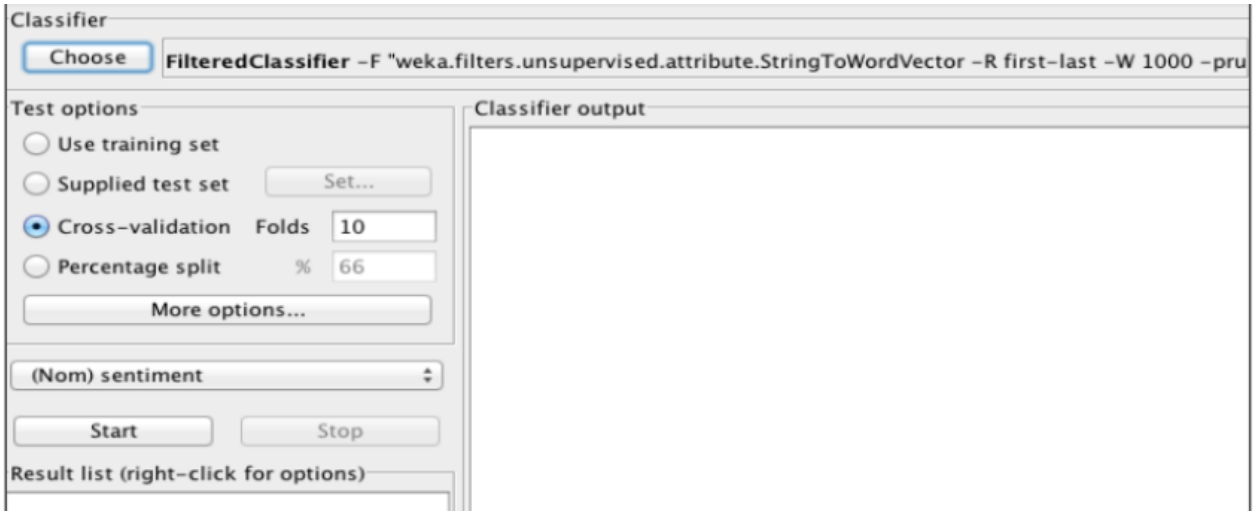
Khởi động Weka và chọn giao diện Explorer. Mở tập dữ liệu thực nghiệm trong Weka (Preprocess → Open File). Di chuyển đến tab Classify. Nhấn nút Choose và tiếp đó chọn Meta → FilteredClassifier. Nhấn lên tên của FilteredClassifier để hiển thị cửa sổ tham số.



Trong cửa sổ này, bạn có thể chọn bộ phân lớp và bộ lọc rút trích đặc trưng tương ứng. Chọn bộ phân lớp naïve Bayes (Classifier → bayes → NaiveBayes) và bộ lọc StringToWordVector (filter → unsupervised → StringToWordVector). StringToWordVector có chức năng chuyển đổi chuỗi ký tự (tức là nội dung của tweet) thành vector từ khóa.

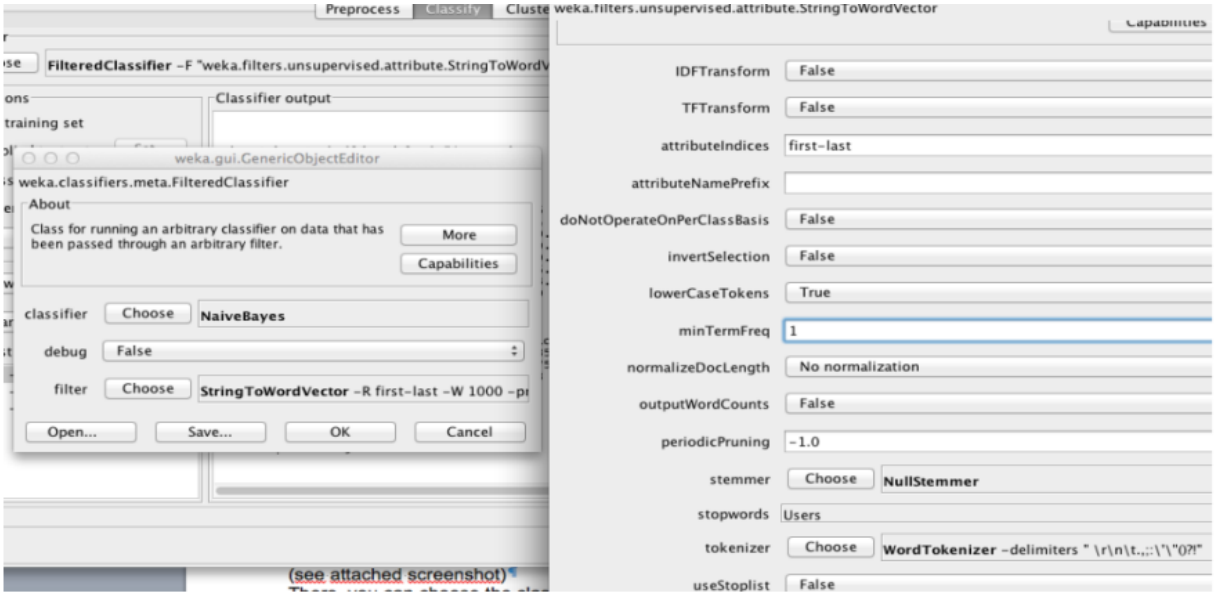


Nhấn OK và tiếp đó nhấn Start.



Đọc kỹ nội dung trong bảng Classifier output và trả lời những câu hỏi dưới đây.

1. Bộ lọc StringToWordVector chuyển chuỗi ký tự thành nhiều thuộc tính số (@attribute). Bạn đếm được bao nhiêu thuộc tính số trong bảng classifier output?
2. Thuộc tính class (tức là "ý kiến" của mỗi tweet) có bị ảnh hưởng bởi bộ lọc không?
3. Ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Tóm lại, bạn nhận thấy bộ phân lớp đã thực thi như thế nào? Bạn có hài lòng với kết quả phân lớp này không? Tại sao?
4. Nhấn StringToWordVector để hiển thị cửa số chứa nhiều tùy chọn.



Các tùy chọn này là tham số ảnh hưởng đến hành vi của bộ lọc và do đó cũng ảnh hưởng đến bộ phân lớp về mặt tổng thể. Nhấn More và đọc mô tả của các tham số. Sau khi đã đọc hiểu mọi tham số, bạn hãy tập trung vào tham số minTermFreq. Hiệu chỉnh giá trị của tham số này. Đầu tiên đặt giá trị bằng 5. Chạy lại bộ phân lớp, phân tích kết quả đầu ra, và ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Bạn nhận thấy bộ phân lớp đã thực thi như thế nào?

5. Tiếp đó đặt giá trị của tham số minTermFreq bằng 10. Chạy lại bộ phân lớp, phân tích kết quả đầu ra, và ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Bạn nhận thấy bộ phân lớp đã thực thi như thế nào?
6. Bạn có thể giải thích chức năng của tham số minTermFreq thông qua cách thức mà tham số này tác động đến hiệu quả phân lớp?
7. Phục hồi giá trị của tham số minTermFreq về 1. Tải tập tin hư từ (stopword) về máy tính từ địa chỉ sau, http://stp.lingfil.uu.se/~santinim/sais/2016/stopwords_eng.txt.

Thiết lập tham số useStoplist thành True và chỉ định các tập tin stopwords _eng.txt vào trường stopwords. Đọc kỹ nội dung của bảng classifier output. Bạn đếm được bao nhiêu thuộc tính trong bảng classifier output?

8. Ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Bộ phân lớp hoạt động như thế nào so với kết quả thực thi trong những câu hỏi trước?
9. Bạn sẽ làm thế nào để tăng sức ảnh hưởng của danh sách hư từ lên việc phân lớp? Hãy đưa ra một vài kiến nghị (ví dụ thêm nhiều từ trong tweet vào tập tin danh sách hư từ, hoặc giảm số từ trong tập tin, loại bỏ/thêm vào/xử lý phủ định, v.v.)
10. Bạn được tùy chọn một tham số từ danh sách tham số của bộ lọc, ngoài những tham số bạn đã trải nghiệm trong các câu hỏi bên trên. Mô tả tham số và giải thích lý do bạn chọn tham số này. Ghi nhận vào báo cáo các giá trị accuracy, TP-Rate, FP-Rate, Precision, Recall, F-Measure và kết quả khảo sát confusion matrix. Bộ phân lớp hoạt động như thế nào với cấu hình tham số mà bạn đã chọn? So sánh với các lượt chạy trước đó.

Trình bày nội dung trả lời vào báo cáo viết và chụp lại màn hình (kèm những đánh dấu cần thiết) trong quá trình thực hiện bài làm.

Lưu ý , mọi lựa chọn hay kết luận đều phải có lý giải đi kèm (và có thể cả hình ảnh minh họa), cũng như phải cung cấp nguồn tài liệu nếu trích dẫn thông tin từ tài liệu khác.

B – Nội dung thực hiện cài đặt (10 điểm)

Cài đặt chương trình đọc vào một tập dữ liệu bất kỳ có định dạng *.csv, xây dựng mô hình phân lớp bằng giải thuật ID3 và đánh giá giải thuật bằng phương pháp cross validation, rồi xuất ra tập tin kết quả.

(1.0đ) Chương trình nhận dữ liệu đầu vào là **tập tin *.csv** có cấu trúc như sau

- Giả sử tập dữ liệu có **N thuộc tính rời rạc** (thuộc tính phân lớp nằm cuối cùng) và M mẫu tương ứng với các thuộc tính này. Dữ liệu được tổ chức thành bảng có M+1 dòng và N cột.
- Dòng đầu tiên chứa tên của N thuộc tính, phân cách nhau bằng dấu phẩy (“,”). Tên thuộc tính không có khoảng trắng và ký tự đặc biệt.
- M dòng tiếp theo, mỗi dòng gồm N giá trị, phân cách nhau bằng dấu phẩy (“,”). Tên giá trị thuộc tính không có khoảng trắng và ký tự đặc biệt.

(3.0đ) Chương trình phát sinh dữ liệu đầu ra là tập tin **model.txt** chứa thông tin tương tự như trong phần văn bản của cửa số Classifier output (tab Classify – WEKA), bao gồm

- Mô hình cây quyết định ID3 rút ra từ toàn bộ tập dữ liệu (full training set). Lưu ý, cây này có thể khác với cây thu được trong mỗi lần chạy cross validation.
- Tiêu chí chọn thuộc tính tốt nhất (Entropy, Information Gain, Information Gain Ratio, và Gini Index).
- Số lượng mẫu phân lớp đúng/sai và tỉ lệ tương ứng
- Các giá trị TP Rate, FP Rate, Precision, Recall và F-Measure cho mỗi phân lớp.

Ví dụ: xét tập dữ liệu weather.nominal (có trong thư mục data của WEKA)

```
=== Classifier model (full training set) ===
outlook = sunny
|  humidity = high: no
|  humidity = normal: yes
outlook = overcast: yes
outlook = rainy
|  windy = TRUE: no
|  windy = FALSE: yes
Correctly Classified Instances          12               85.7143 %
Incorrectly Classified Instances         2               14.2857 %
===Best attribute criteria===
Entropy
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure
Class
yes              0.889     0.2         0.889     0.889     0.889
no               0.8       0.111       0.8       0.8       0.8
```

(1.0đ) Chương trình thực thi **giải thuật ID3** và đánh giá giải thuật bằng phương pháp **n-folds cross validation** với cú pháp tham số dòng lệnh như sau

<ID nhom> <input> <output> <folds> <best_att>

- <ID nhom>: tên của tập tin thực thi chương trình là ID của nhóm.
- <input>: tập tin dữ liệu đầu vào có định dạng *.csv
- <output FI>: tập tin đầu ra model.txt
- <folds>: số lượng fold chỉ định cho phương pháp cross validation.
- <best_att>: chiến lược chọn thuộc tính tốt nhất, 0: Entropy, 1: Information Gain, 2: Information Gain Ratio, và 3: Gini Index.

Chương trình **xử lý tuần tự các mẫu theo thứ tự từ trên xuống**. Cần thể hiện ra màn hình console cho người dùng biết chương trình đang xử lý đến giai đoạn nào. Ví dụ: đang xây dựng cây ID3, đang tính độ chính xác,…

Chương trình **xuất ra giá trị độ đo đánh giá thuộc tính theo chiến lược đã chọn** ra màn hình console trong quá trình tính toán.

(5.0đ) Tùy chọn 3 tập dữ liệu có quy mô nhỏ (~100 mẫu), trung bình (~500 mẫu), và lớn (~1000 mẫu). Chạy chương trình cài đặt với các tập dữ liệu đã chọn và đối chiếu kết quả phát sinh được với kết quả của WEKA ID3 trên cùng bộ tham số.