6

Lab

MÁY HỌC TRONG IDS

Thực hành

Hệ thống tìm kiếm, phát hiện và ngăn ngừa xâm nhập

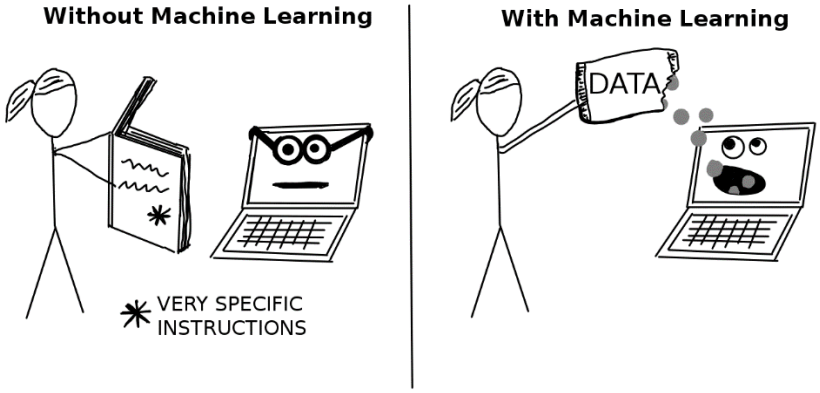
Lưu hành nội bộ

A. TỔNG QUAN

A.1 Học máy – Machine learning là gì?

A.1.1 Giới thiệu về học máy

Học máy (Machine learning) là một lĩnh vực nghiên cứu cung cấp cho các máy tính khả năng học và cải tiến từ kinh nghiệm mà không cần được lập trình sẵn, nó tập trung vào khả năng phát triển của các chương trình có thể sử dụng các dữ liệu để tự khám phá. Học máy là một sự thay đổi mô hình từ ***lập trình bình thường***, trong đó tất cả các dòng lệnh phải được cung cấp rõ ràng cho máy tính sang ***lập trình gián tiếp*** diễn ra thông qua việc cung cấp dữ liệu.



Quá trình học được bắt đầu với việc quan sát hoặc được cấp dữ liệu để tìm ra các mẫu trong dữ liệu và đưa ra dự đoán dựa trên các dữ liệu được cung cấp. Mục đích chính là cho phép các máy tính học mà không cần sự hỗ trợ của con người.

Có nhiều loại học máy:

- ***Các giải thuật học máy có giám sát***: các thông tin đã có trước sẽ được học để dự đoán các thông tin trong tương lai với các dữ liệu đã được dán nhãn.
- ***Các giải thuật học máy không có giám sát***: các thông tin đã có trước chưa được dán nhãn hay phân loại, các giải thuật phải tự phân tích và gom nhóm các thông tin dựa trên các mẫu, độ tương đồng và khác biệt, …
- ***Các giải thuật học máy bán giám sát***: một phần thông tin có trước đã được gán nhãn, một phần thì không.

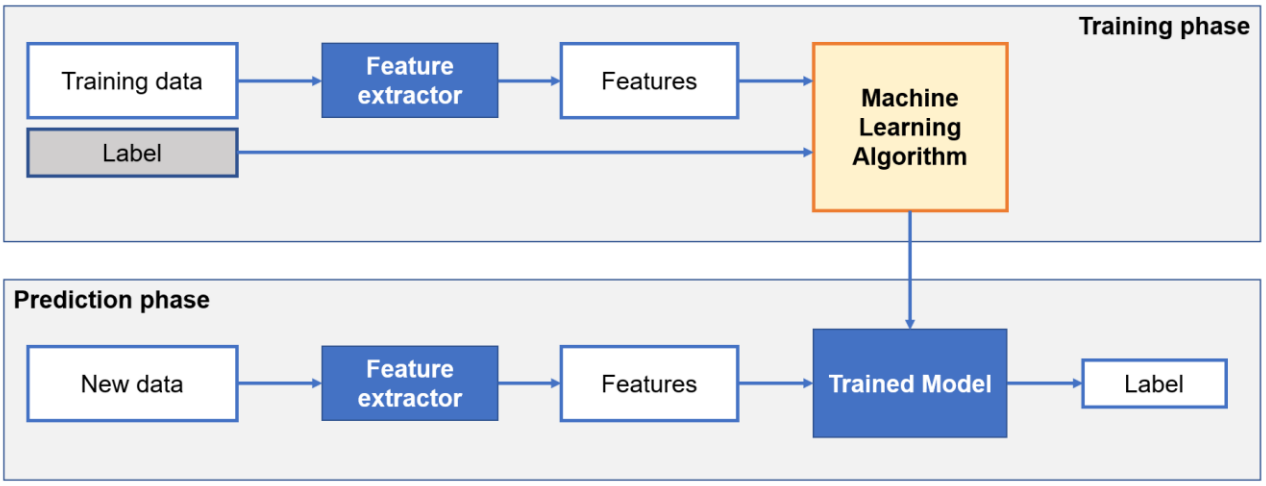
A.1.2 Các khái niệm trong học máy

- Giải thuật (algorithm) là một tập các rule mà máy tính sẽ làm theo để đạt được một mục đích nào đó, trong đó định nghĩa đầu vào, đầu ra và các bước cần thiết để từ đầu vào thu được đầu ra.
- Học máy (machine learning) là một tập các phương thức cho phép máy tính học từ dữ liệu để đưa ra và cải thiện các dự đoán.

- **Giải thuật học máy** (machine learning algorithm) là một chương trình dùng để học được một mô hình học máy từ dữ liệu.
- **Mô hình học máy** (machine learning model) là chương trình đã được học, trong đó có ánh xạ đầu vào với dự đoán tương ứng.
- **Tập dữ liệu** (dataset) là một bảng gồm dữ liệu mà máy tính sẽ học, có chứa các đặc tính và mục tiêu cần dự đoán.
    - Đặc tính (feature) là danh sách các đầu vào để dự đoán hoặc phân loại, tương ứng với các cột trong bảng dataset.
    - Thực thể (instance) là một dòng của bảng dataset, gồm các giá trị cụ thể cho từng đặc tính.
    - Mục tiêu (target) là thông tin mà máy tính cần học để dự đoán được.

Dataset được dùng để tạo ra mô hình học máy được gọi là tập huấn luyện (training data).

Trong phạm vi bài thực hành sẽ tìm hiểu, cài đặt và thử nghiệm nhóm các giải thuật học máy có giám sát, với các tập dữ liệu train đã được phân tích và gán nhãn, với quy trình minh họa như bên dưới.



A.2 Học máy trong IDS

Đặt trong ngữ cảnh hỗ trợ các IDS phát hiện và ngăn chặn tấn công, tập dữ liệu dùng để train các mô hình học máy sẽ là các lưu lượng hoặc thông tin phân tích được từ lưu lượng mạng để từ đó đưa ra được dự đoán cuối cùng là tấn công gì. Khi đó, các thành phần trong mô hình học máy như sau:

- Dataset: lưu lượng mạng đã được phân tích thành các đặc tính và gán nhãn. Trong bài thực hành này sẽ sử dụng tập dữ liệu **KDD Cup 1999** bao gồm các lưu lượng mạng đã được thu thập, phân tích và dán nhãn tấn công hay bình thường tương ứng, được dùng để train cho các IDS.
- Giải thuật học máy: các giải thuật học máy có giám sát.

- Mục tiêu dự đoán: dán nhãn tấn công hoặc bình thường, nếu có tấn công thì đưa ra dự đoán kiểu tấn công.

A.3 Cài đặt môi trường

- Môi trường thực hiện: Windows hoặc Linux (Nên sử dụng Ubuntu).

B. THỰC HÀNH

B.1 Tìm hiểu về tập dữ liệu KDD Cup 1999[1]

> **Yêu cầu 1.1** Sinh viên tìm hiểu về tập dữ liệu KDD Cup 1999 và điền các kết quả tìm hiểu được vào form bên dưới.

Bộ dataset KDD Cup 1999: https://kdd.org/kdd-cup/view/kdd-cup-1999/Data

Lưu ý: nên sử dụng file kddcup.data_10_percent.zip để không gây quá tải cho máy.

TÌM HIỂU VỀ TẬP DỮ LIỆU KDD CUP 1999

Dữ liệu trong bộ dữ liệu KDD Cup 1999 là lưu lượng mạng đã được thu thập, phân tích, xử lý để lấy các thuộc tính và từ đó gán nhãn tương ứng với loại tấn công hoặc dữ liệu bình thường. Sinh viên tìm hiểu các phần sau:

1. Số nhóm tấn công: .....
   Kể tên các nhóm tấn công: .....
2. Số kiểu tấn công: .....
   Kể tên các kiểu tấn công được gán nhãn: .....
3. Mỗi instance trong tập dữ liệu KDD Cup 1999 bao gồm ..... thuộc tính, cụ thể gồm các thuộc tính: .....

B.2 Công cụ WEKA[2]

WEKA là một công cụ mã nguồn mở có hỗ trợ nhiều giải thuật học máy để khai thác dữ liệu, bao gồm các giải thuật classification, regression, clustering, ... Công cụ này có 2 cách sử dụng: thông qua command line và giao diện người dùng.

B.2.1 Tìm hiểu chung về WEKA

> **Yêu cầu 2.1** Sinh viên cài đặt WEKA, tìm hiểu và load một tập dữ liệu có định dạng .arff đơn giản có sẵn của WEKA.

Yêu cầu:

- Cài đặt được WEKA.
- Hiểu được cấu trúc của file **.arff** định nghĩa các thông tin của tập dữ liệu gồm thuộc tính, instance, mục tiêu cần dự đoán như thế nào?
- Giải thích được các thông tin hiển thị khi load tập dữ liệu vào WEKA.

Hướng dẫn load tập dữ liệu có sẵn lên WEKA:

[1] https://www.kdd.org/kdd-cup/view/kdd-cup-1999
[2] https://waikato.github.io/weka-wiki/

- **Bước 1: Cài đặt WEKA**

Sinh viên tham khảo tải công cụ WEKA tại đường dẫn:

https://www.cs.waikato.ac.nz/ml/weka/downloading.html và tham khảo các hướng dẫn cài đặt trên các hệ điều hành cụ thể.

- **Bước 2: Chạy WEKA**

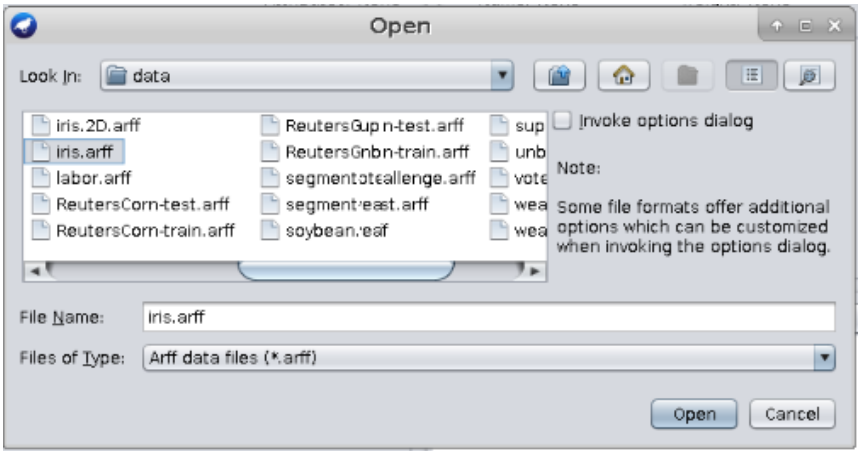
Sinh viên có thể lựa chọn 2 cách chạy bằng command line hoặc sử dụng giao diện. Với lựa chọn command line, chạy lệnh sau:

```
java -jar <path to weka>/weka.jar
```

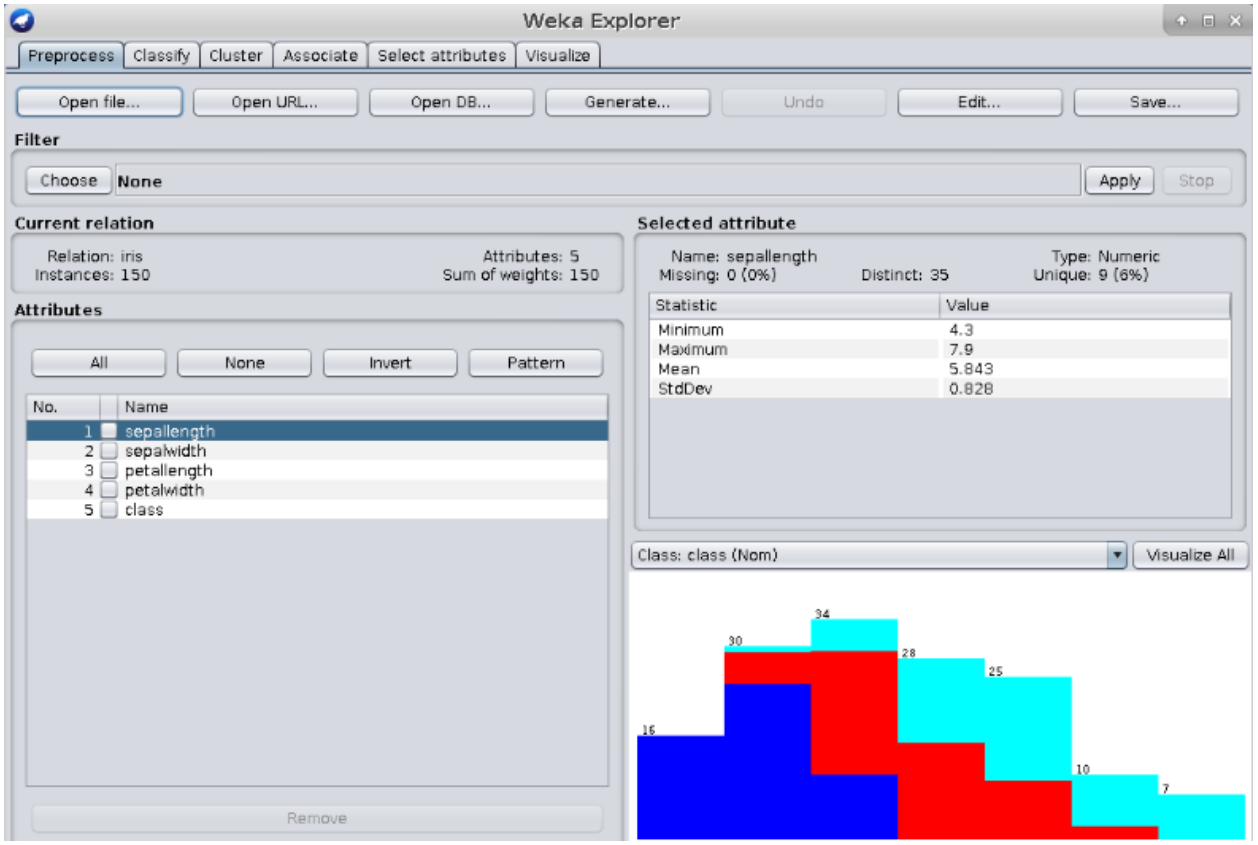
- **Bước 3:** Trong cửa sổ GUI Chooser, chọn **Explorer**

- **Bước 4: Load dữ liệu vào WEKA**

Ở tab **Preprocess** trong cửa sổ WEKA Explorer, click chọn **Open files...** để chọn file dữ liệu cần đưa vào. WEKA khi cài đặt có cung cấp sẵn một số tập dữ liệu đơn giản ở thư mục **data/** trong thư mục cài đặt của WEKA. Trong thư mục này bao gồm nhiều file có định dạng **.arff**, là các dataset có thể đưa vào sử dụng ngay trong WEKA. Lựa chọn một file bất kỳ, ở đây sử dụng file **iris.arff** là một dataset chứa dữ liệu thu thập được về đặc điểm của một số loại hoa diên vĩ (iris) khác nhau như độ dài và độ rộng của cánh hoa và thuộc tính phân lớp là loại hoa diên vĩ nào.



- **Bước 5:** Quan sát và giải thích các kết quả trong tab **Preprocess** của WEKA.



B.2.2 Sử dụng các bộ phân lớp trong WEKA

Quá trình xử lý của WEKA là việc thực thi các giải thuật machine learning trên bộ dữ liệu đầu vào để xây dựng được một mô hình (model) thực hiện chức năng phân lớp (classification) hoặc hồi quy (regression). Chức năng của mô hình là có thể đưa ra kết luận về thuộc tính phân lớp dựa trên các thuộc tính còn lại.

WEKA hỗ trợ nhiều giải thuật học máy khác nhau để áp dụng trên bộ dữ liệu đầu vào trong tab **Classify**. Bên cạnh đó, WEKA cũng cung cấp nhiều tùy chọn để định nghĩa tập huấn luyện để xây dựng mô hình học máy và tập kiểm tra từ tập dữ liệu đầu vào đã load lên WEKA (tìm hiểu về **Test options**).

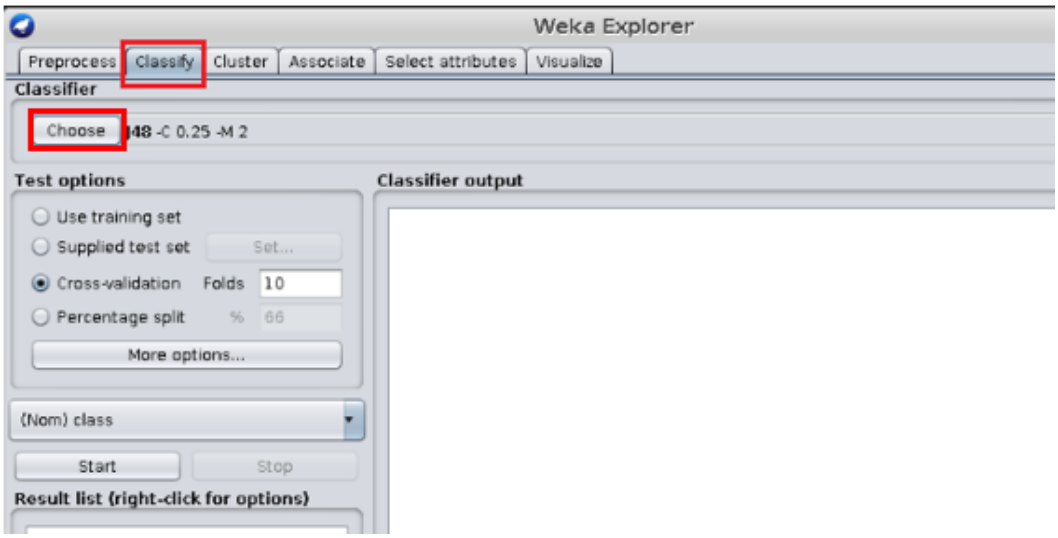
> **Yêu cầu 2.2** Sinh viên lựa chọn 01 bộ phân lớp bất kỳ và thực hiện khai thác trên tập dữ liệu đã chọn ở trên. Trình bày và giải thích kết quả.

Yêu cầu: cần giải thích test option lựa chọn và ý nghĩa của các thông số kết quả đầu ra sau khi chạy bộ phân lớp.

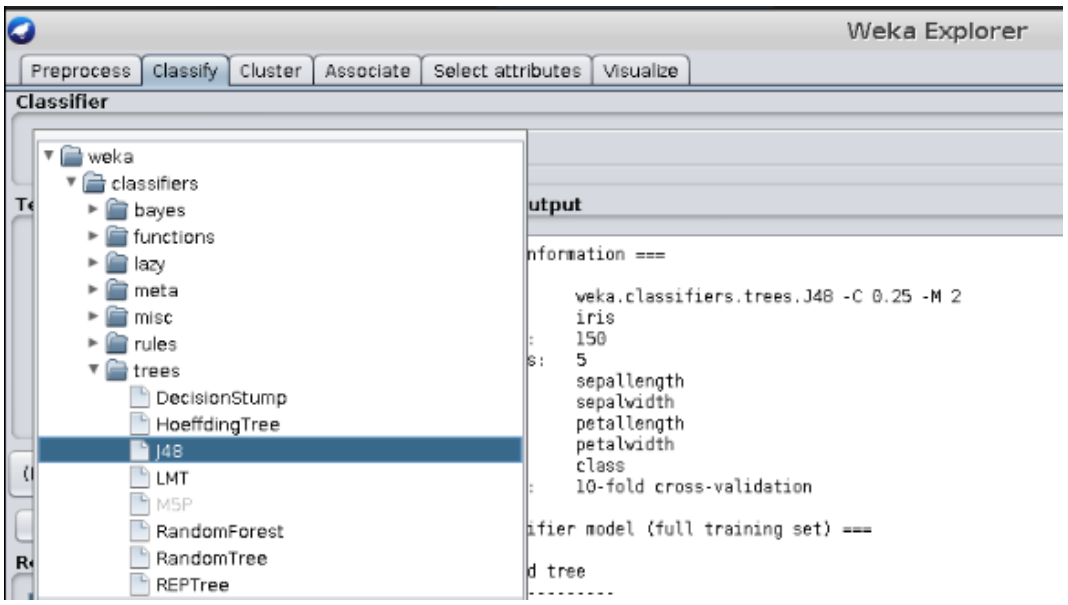
Hướng dẫn: Các bước sau thực hiện sau khi đã load tập dữ liệu lên WEKA.

- **Bước 1: Lựa chọn bộ phân lớp cho tập dữ liệu**

Trong cửa sổ WEKA Explorer, chọn tab **Classify**.



Trong khung lựa chọn **Classifier**, nhấp chọn **Choose** để hiển thị danh sách các bộ phân lớp được WEKA hỗ trợ. Tùy vào tập dữ liệu được load lên mà chỉ có những bộ phân lớp phù hợp có thể chạy trên đó mới khả dụng.

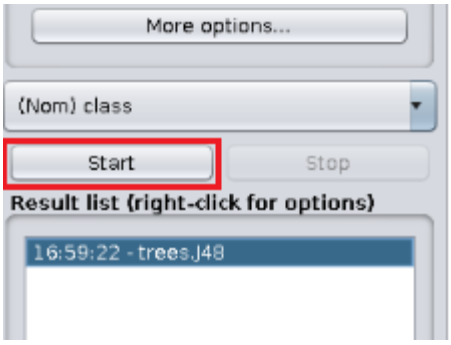


- **Bước 2: Lựa chọn Test options**

Đây là các tùy chọn được WEKA hỗ trợ để định nghĩa 2 tập dữ liệu huấn luyện và kiểm tra. Sinh viên tự tìm hiểu về phần này và lựa chọn tùy chọn phù hợp.

- **Bước 3: Chạy bộ phân lớp và quan sát kết quả**

Nhấn chọn **Start** để bắt đầu chạy bộ phân lớp.



- **Bước 4: Quan sát và giải thích kết quả**

Quan sát kết quả ở khung **Classifier output**, những kết quả này có ý nghĩa gì?

B.3 Sử dụng WEKA với tập dữ liệu KDD Cup 1999

Do tập dữ liệu KDD Cup có kích thước lớn, nên trong phạm vi bài thực hành chỉ sử dụng dataset KDD với phiên bản 10% (khoảng 75MB). Sinh viên tải dataset tại đường dẫn: https://www.kdd.org/kdd-cup/view/kdd-cup-1999/Data

> **Yêu cầu 3.1** Sinh viên lựa chọn 01 bộ phân lớp bất kỳ và thực hiện khai thác trên tập dữ liệu KDD Cup 1999. Trình bày và giải thích kết quả.

Yêu cầu: Giải thích test options, bộ phân lớp và các kết quả thu được.

Hướng dẫn sử dụng WEKA với KDD Cup 1999:

- **Bước 1:** Truy cập vào đường dẫn ở trên và tải các file cần thiết, các file tối thiểu cần tải là **kddcup.data_10_percent.zip** và **kddcup.names**.
- **Bước 2:** Giải nén file dataset **.zip**, thu được một file dạng text chứa các instance trên từng dòng, các giá trị thuộc tính phân cách với nhau bằng dấu phẩy (,).
- **Bước 3:** Mở file **kddcup.names** ta thu được danh sách nhiều kiểu tấn công ở đầu file và danh sách các thuộc tính kèm theo kiểu dữ liệu (phần khoanh đỏ).

Lọc lấy tên các thuộc tính và chuyển chúng thành một dòng lần lượt phân cách bằng dấu phẩy (,).

```
back,buffer_overflow,ftp_write,guess_passwd,imap,ipsweep,land,loadmodule,multihop,neptune,nmap,normal,perl,phf,pod,portsweep,rootki
t,satan,smurf,spy,teardrop,warezclient,warezmaster.
duration: continuous.
protocol_type: symbolic.
service: symbolic.
flag: symbolic.
src_bytes: continuous.
dst_bytes: continuous.
land: symbolic.
wrong_fragment: continuous.
urgent: continuous.
hot: continuous.
num_failed_logins: continuous.
logged_in: symbolic.
num_compromised: continuous.
root_shell: continuous.
su_attempted: continuous.
num_root: continuous.
num_file_creations: continuous.
num_shells: continuous.
num_access_files: continuous.
num_outbound_cmds: continuous.
is_host_login: symbolic.
is_guest_login: symbolic.
count: continuous.
srv_count: continuous.
serror_rate: continuous.
srv_serror_rate: continuous.
rerror_rate: continuous.
srv_rerror_rate: continuous.
same_srv_rate: continuous.
diff_srv_rate: continuous.
srv_diff_host_rate: continuous.
dst_host_count: continuous.
dst_host_srv_count: continuous.
dst_host_same_srv_rate: continuous.
dst_host_diff_srv_rate: continuous.
dst_host_same_src_port_rate: continuous.
dst_host_srv_diff_host_rate: continuous.
dst_host_serror_rate: continuous.
dst_host_srv_serror_rate: continuous.
dst_host_rerror_rate: continuous.
dst_host_srv_rerror_rate: continuous.
```

Chuyển thành dạng này:

```
duration,protocol_type,service,flag,src_bytes,dst_bytes,land,wrong_fragment,urgent,hot,num_failed_logins,logged_in,
```

- **Bước 4:** Chèn dòng tên các thuộc tính trên vào đầu file **kddcup.data_10_percent**

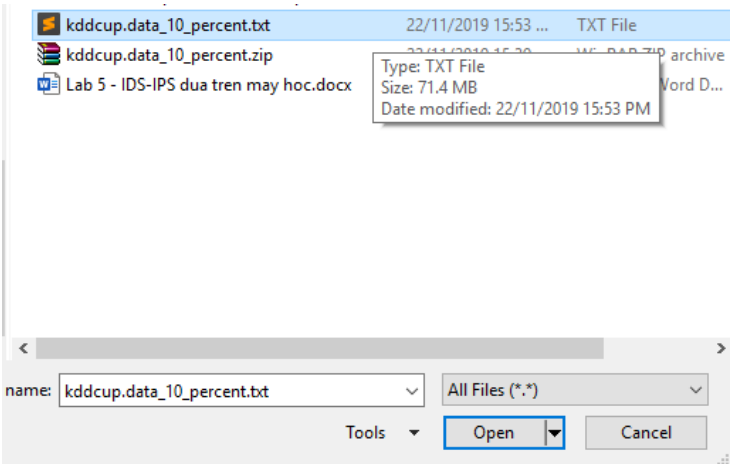
Lưu ý thêm một thuộc tính vào cuối danh sách là **result** tương ứng với cột phân loại tấn công.ẽ

kddcup.data_10_percent.txt

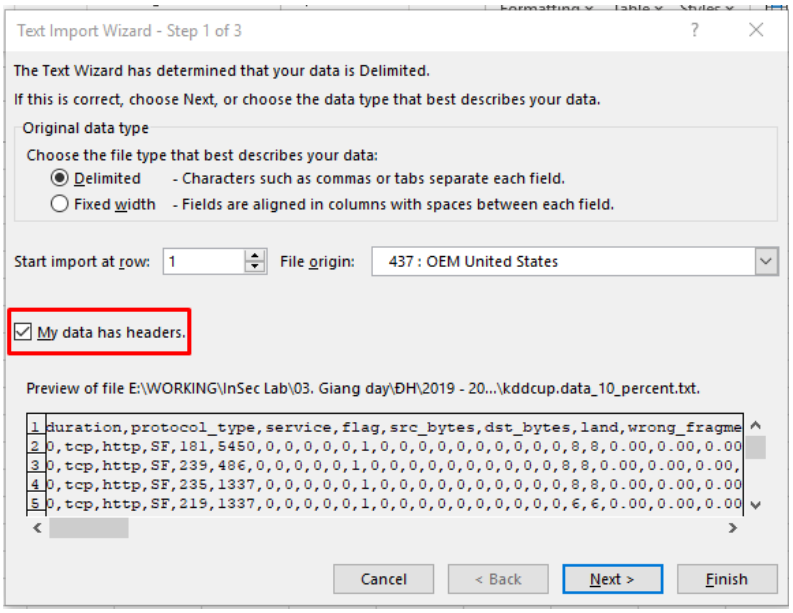
```
1 duration,protocol_type,service,flag,src_bytes,dst_bytes,land,wrong_fragment,urgent,hot,num_failed_logins,
  logged_in,num_compromised,root_shell,su_attempted,num_root,num_file_creations,num_shells,num_access_files
  ,num_outbound_cmds,is_host_login,is_guest_login,count,srv_count,serror_rate,srv_serror_rate,rerror_rate,s
  rv_rerror_rate,same_srv_rate,diff_srv_rate,srv_diff_host_rate,dst_host_count,dst_host_srv_count,dst_host_
  same_srv_rate,dst_host_diff_srv_rate,dst_host_same_src_port_rate,dst_host_srv_diff_host_ra
  te,dst_host_serror_rate,dst_host_srv_serror_rate,dst_host_rerror_rate,dst_host_srv_rerror_rate,result
2 0,tcp,http,SF,181,5450,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,9,9,1.00,0.
  00,0.11,0.00,0.00,0.00,0.00,0.00,normal.
3 0,tcp,http,SF,239,486,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,19,19,1.00,0
  .00,0.05,0.00,0.00,0.00,0.00,0.00,normal.
4 0,tcp,http,SF,235,1337,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,29,29,1.00,
  0.00,0.03,0.00,0.00,0.00,0.00,0.00,normal.
```

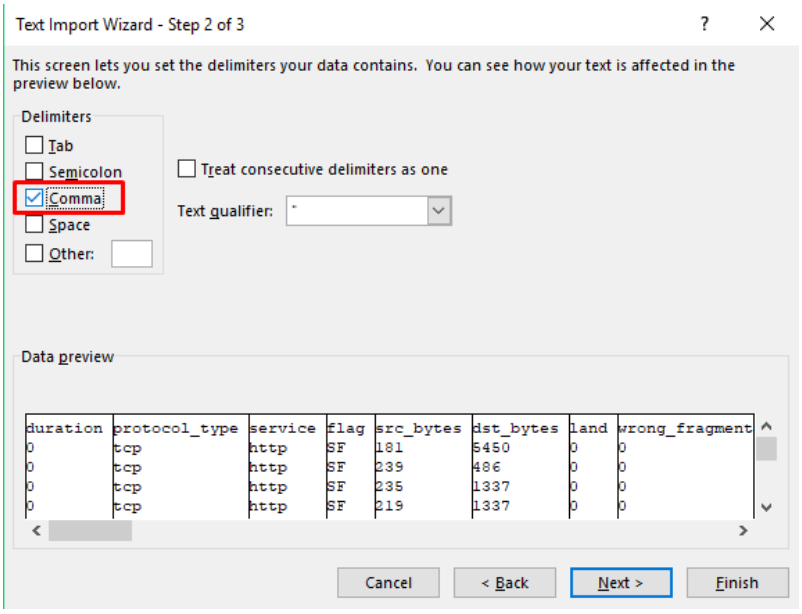
Cập nhật: sau khi thêm tên các thuộc tính, chú ý dòng dữ liệu thứ **485799** ***bị lỗi*** (có nhiều giá trị hơn số thuộc tính), sinh viên cần xóa dòng này khỏi file dataset.

- **Bước 5:** Mở Excel và mở tập dữ liệu đã thêm tên thuộc tính.

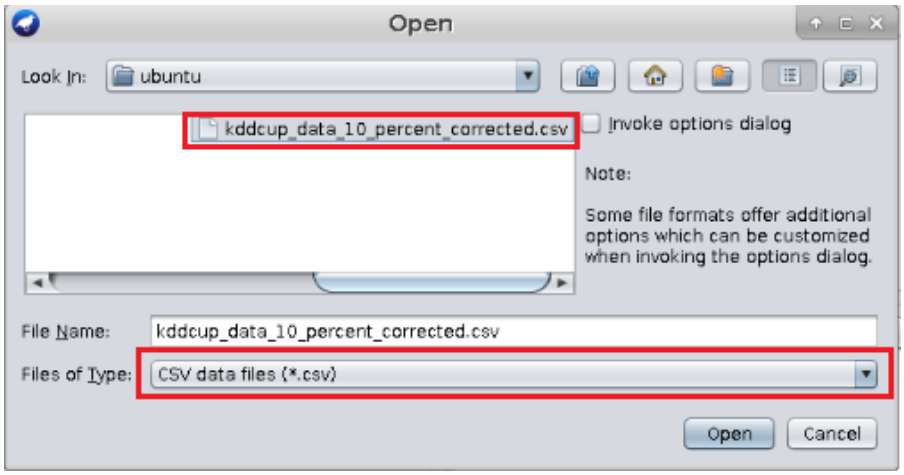


Lựa chọn các option **My data has headers** để xác định dòng đầu tiên trong file là tên các cột và chọn **Comma** để định nghĩa dấu phẩy (,) là ký tự phân tách các giá trị.





- **Bước 6:** Sau khi mở file thành công, lưu file dưới định dạng **.csv** bằng **Save as…**
- **Bước 7:** File .csv sau đó có thể được load vào WEKA tương tự như **Yêu cầu 2.1** với tùy chỉnh ở kiểu file như hình bên dưới.



Lưu ý: Khi chọn 01 thuật toán máy học để chạy bộ dữ liệu KDD Cup 1999, sinh viên cần giới thiệu về thuật toán đã chọn và giải thích tại sao sử dụng thuật toán này. Không chọn lại thuật toán J48 như trong phần trước.

C. YÊU CẦU

- Sinh viên tìm hiểu và thực hiện **theo nhóm đã đăng ký**.
- Sinh viên có thể chọn 1 trong 2 hình thức để báo cáo:
  - Hình thức 1: Báo cáo trên lớp trong buổi thực hành, GVTH sẽ chấm điểm trực tiếp. Dựa trên kết quả sinh viên thực hiện, có thể không cần nộp báo cáo.
  - Hình thức 2: Nộp báo cáo file **.docx** trên courses, gồm kết quả và chi tiết những việc (**Report**) mà nhóm đã tìm hiểu, thực hiện và kèm ảnh chụp màn hình kết quả (nếu có); giải thích cho quan sát (nếu có).
- Khuyến khích sinh viên báo cáo theo hình thức 1.

~HẾT~