ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG

NGUYỄN VIỆT THÀNH - 15520814
LÊ HOÀNG TUẤN - 15520967

ĐỒ ÁN CHUYÊN NGÀNH
PHÁT HIỆN TRAFFIC BẤT THƯỜNG BẰNG MÁY HỌC

GIẢNG VIÊN HƯỚNG DẪN
ThS. UNG VĂN GIÀU

TP. HỒ CHÍ MINH, 2018

Mục lục

Danh sách hình vẽ

Danh sách bảng

TÓM TẮT ĐỒ ÁN

Trong bài báo cáo này chúng tôi trình bày về vấn đề ứng dụng công nghệ máy học (hay còn gọi là học máy - machine learning) vào việc phân tích và phát hiện các luồng traffic nào là bình thường và bất thường. Trong phạm vi của đồ án, chúng tôi tập trung vào phân tích các HTTP request từ tập dữ liệu CSIC 2010. Kết quả đạt được của chúng tôi là phân tích các traffic nào là bất thường và bình thường, thuật toán mà chúng tôi chọn là Decision Tree.

Nội dung của bài báo cáo này gồm 2 phần chính là:

- **Cơ sở lý thuyết** - phần này chúng tôi sẽ giới thiệu một số lý thuyết toán học liên quan. Sau đó chúng tôi giới thiệu sơ lược về máy học, các khái niệm, và phân loại. Tiếp theo chúng tôi cũng phân tích thuật toán mà chúng tôi chọn sử dụng - Decision Tree. Bên cạnh đó, chúng tôi cũng tiếp cận một số lỗ hổng bảo mật web phổ biến.
- **Áp dụng thuật toán vào phân tích tập dữ liệu** - phần này là kết quả của nhóm chúng tôi đạt được, phần nãy sẽ tập trung vào tập dữ liệu và thuật toán mà chúng tôi chọn sử dụng.

MỞ ĐẦU

Lý do chọn đề tài

Nhiều ứng dụng web ngày nay gặp vấn đề bảo mật, nguyên nhân nó từ các nhà phát triển ứng dụng web, muốn tạo ra sản phẩm nhanh, không quan tâm cũng như kiến thức liên quan đến bảo mật. Để khắc phục vấn đề bảo mật. Nhà phát triển web cần tìm ra một công cụ để giảm thiểu rủi ro bảo mật. Phát hiện xâm nhập là một công cụ mạnh mẽ để nhận diện và ngăn chặn tấn công tới hệ thống. Hầu hết những công nghệ phát hiện xâm nhập hệ thống web hiện nay không có khả năng giải quyết các tấn công web phức tạp, những kiểu tấn công mới chưa từng biết trước đó.

Tuy nhiên, với việc áp dụng máy học (tiếng anh: **machine learning**), ta có thể xây dựng những mô hình giúp phát hiện những kiểu tấn công đã biết hoặc chưa biết. Như chúng ta đã biết, machine learning gây nên cơn sốt công nghệ trên toàn thế giới trong vài năm nay. Trong giới học thuật, mỗi năm có hàng ngàn bài báo khoa học về đề tài này. Trong giới công nghiệp, từ các công ty lớn như Google, Facebook, Microsoft đến các công ty khởi nghiệp đều đầu tư vào machine learning. Hàng loạt các ứng dụng sử dụng machine learning ra đời trên mọi lĩnh vực của cuộc sống, từ khoa học máy tính đến những ngành ít liên quan hơn như vật lý, hóa học, y học, chính trị.

Chính vì những điều trên đã thôi thúc chúng tôi tiến hành tiếp cận máy học trong lĩnh vực phát hiện tấn công web.

Mục đích thực hiện đề tài

Khi thực hiện đề tài, nhóm chúng tôi mong muốn được tiếp cận nghiên cứu và tìm hiểu về lĩnh vực máy học. Và từ đó vận dụng vào ngành mà chúng tôi đang học - An toàn thông tin.

Hai mục tiêu mà chúng tôi hướng đến để đạt được trong đề tài này là:

- Thứ nhất, sẽ có kiến thức cơ bản về máy học và lý thuyết liên quan.
- Thứ hai, tìm hiểu và chọn được một thuật toán để vận dụng phân tích một tập dữ liệu cho trước để phát hiện luồng traffic nào là bình thường và bất thường.

Đối tượng và phạm vi nghiên cứu của đề tài

Đối tượng và phạm vi nghiên cứu của chúng tôi tập trung vào hai điểm chính:

- **Tập dữ liệu được sử dụng để training** - ở đây chúng tôi chọn tập dữ liệu **HTTP DATASET CSIC 2010**. Lý do vì sao chúng tôi chọn tập dữ liệu này sẽ được trình bày chi tiết trong phần sau của báo cáo.
- **Thuật toán được sử dụng** - thuật toán mà nhóm chúng tôi chọn là Decision Tree. Lý do nhóm chọn cũng sẽ được giới thiệu chi tiết trong phần nội dung của bài báo cáo

Trong phạm vi của đồ án, nhóm chúng tôi chỉ tập trung vào việc phân tích GET và POST trong thành phần HTTP Header của các gói tin.

Chương 1

CƠ SỞ LÝ THUYẾT

1.1 Một số lý thuyết về xác suất

Có thể nói một điều rằng lý thuyết xác suất là một trong những lý thuyết quan trọng nhất của khoa học hiện đại và đặc biệt là **Machine Learning** bởi vì đa phần các thuật toán của Machine Learning đều có cơ sở dựa trên xác suất.

1.1.1 Không gian xác suất

Khi nói đến xác suất là người ta nói đến các lý thuyết toán học về sự *bất định - uncertainly* hay nói một cách khác, xác suất biểu thị khả năng xảy ra của các *sự kiện - event* trong một môi trường bất định nào đó. Ví dụ chúng ta xét về xác suất có mưa hay không có mưa vào thứ hai tuần tới, xác suất tỏ tình thành công hay thất bại của cậu bạn thân,... Tóm lại cứ nói đến xác suất là đề cập đến sự không chắc chắn hay bất định đó.

Về mặt toán học, người ta kí hiệu một **không gian xác suất - probability space** bao gồm 3 thành phần (Ω, F, P) như sau:

- Ω (có thể đọc là "Ô-me-ga") chính là tập các giá trị **có thể xảy ra - possible outcome** với sự kiện trong không gian xác suất. Người ta còn gọi nó là **không gian mẫu**.
- $F \subseteq 2^{\Omega}$ là tập hợp các sự kiện có thể xảy ra trong không gian xác suất.

- P là xác suất (hoặc phân phối xác suất) của sự kiện. P ánh xạ một sự kiện $E \in F$ vào trong một giá trị thực $p \in [0; 1]$. Ở đây chúng ta gọi $p = P(E)$ là xác suất của sự kiện E.

Ví dụ minh họa

Chúng ta cùng nhau xem xét một ví dụ khá kinh điển trong lý thuyết xác suất đó chính là ví dụ **tung xúc sắc**.

Giả sử rằng chúng ta tung một con xúc sắc 6 mặt. Không gian các **outcomes** có thể xảy ra trong trường hợp này là $\Omega = \{1, 2, 3, 4, 5, 6\}$ - chúng ta không tính đến các trường hợp xúc sắc rơi lơ lửng tức là không thuộc mặt nào. Không gian các sự kiện F sẽ tùy thuộc vào sự định nghĩa của chúng ta. Ví dụ chúng ta định nghĩa sự kiện xúc sắc là mặt chẵn hoặc mặt lẻ thì không gian sự kiện $F = \{\varnothing, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}$ trong đó $\varnothing$ là sự kiện có xác suất 0 - hay còn gọi là biến cố *không thể có*. Ω là sự kiện có xác suất 1 - hay còn gọi là *biến cố chắc chắn*.

1.1.2 Các tính chất xác suất

Giống như ví dụ ở phía trên, khi *không gian mẫu - outcomes space* là hữu hạn thì chúng ta thường lựa chọn không gian sự kiện $F = 2^{\Omega} = \{\varnothing, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}$. Cách tiếp cận này chưa hẳn đã tổng quát hóa cho mọi trường hợp tuy nhiên nó đủ dùng trong các bài toán thực tế, tất nhiên là với giả thiết không gian mẫu của chúng ta là **hữu hạn**. Khi không gian mẫu là **vô hạn - infinite** chúng ta phải hết sức cẩn thận trong việc lựa chọn không gian sự kiện F. Khi đã định nghĩa được không gian sự kiện F thì hàm xác suất của chúng ta bắt buộc phải thỏa mãn các tính chất sau đây:

- **Không âm - non-negativity** - xác suất của mọi sự kiện là không âm, tức là với mọi $x \in F, \;\; P(x) \geq 0$
- **Xác suất toàn cục - trivial event** $P(\Omega) = 1$

- **Tính cộng - additivity** tức là với mọi $x, y \in F$ nếu như $x \cap y = \varnothing$ thì ta có $P(x \cup y) = P(x) + P(y)$

1.1.3 Biến ngẫu nghiên

Biến ngẫu nghiên (Random Variables) là một thành phần quan trọng trong lý thuyết xác suất. Nó biểu diễn giá trị của các đại lượng không xác định, thông thường nó được coi như một ánh xạ từ tập các **outcomes** trong không gian mẫu thành các giá trị thực.

1.1.4 Xác suất có điều kiện

1.1.5 Quy tắc Bayes

1.1.6 Kỳ vọng

1.1.7 Một vài phân phối xác suất thường gặp

1.2 Giới thiệu máy học

1.2.1 Khái niệm

1.2.2 Phân loại

1.3 Thuật toán Decision Tree

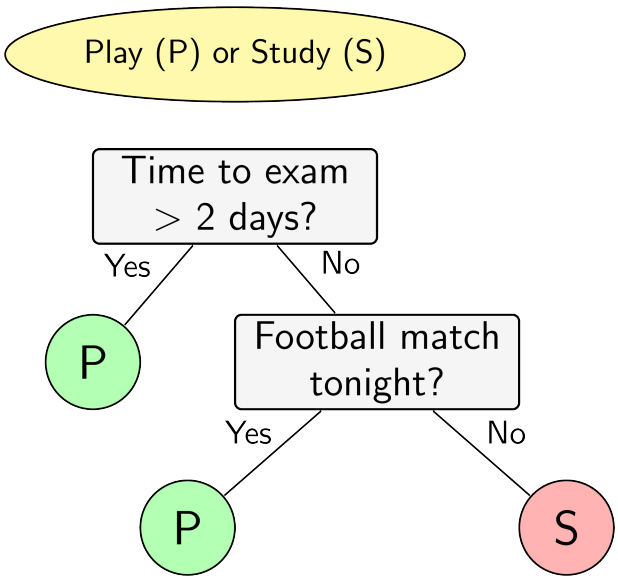
1.3.1 Giới thiệu

Sắp đến kỳ thi, một cậu sinh viên tự đặt ra quy tắc *học hay chơi* của mình như sau:

- Nếu còn nhiều hơn hai ngày tới ngày thi, cậu sẽ đi chơi.

- Nếu còn không quá hai ngày và đêm hôm đó có một trận bóng đá, cậu sẽ sang nhà bạn chơi và cùng xem bóng đêm đó.
- Cậu sẽ chỉ học trong các trường hợp còn lại.

Việc ra quyết định của cậu sinh viên này có thể được mô tả trên sơ đồ trong hình 1.1.



Hình 1.1: Một ví dụ về việc đưa ra các quyết định dựa trên câu hỏi

Hình ellipse nền vàng thể hiện quyết định cần được đưa ra. Quyết định này phụ thuộc vào các câu trả lời của các câu hỏi trong các ô hình chữ nhật màu xám. Dựa trên các câu trả lời, quyết định cuối cùng được cho trong các hình tròn màu lục (chơi) và đỏ (học).

Việc quan sát, suy nghĩ và ra các quyết định của con người thường được bắt đầu từ các câu hỏi. Machine learning cũng có một mô hình ra quyết định dựa trên các câu hỏi. Mô hình này có tên là *cây quyết định (decision tree)*.

Trong **decision tree**, các ô màu xám, lục, đỏ trên hình 1.1 được gọi là các node. Các node thể hiện đầu ra (màu lục và đỏ) được gọi là *node lá* (*leaf node* hoặc *terminal node*). Các node thể hiện câu hỏi là các *non-leaf node*. Non-leaf node trên cùng (câu hỏi đầu tiên) được gọi là node gốc (*root node*). Các non-leaf node thường có hai hoặc nhiều node con (*child node*). Các child node này có thể là một leaf node hoặc một non-leaf node khác. Các child node có

cùng bố mẹ được gọi là *sibling node*. Nếu tất cả các non-leaf node chỉ có hai child node, ta nói rằng đó là một *binary decision tree* (cây quyết định nhị phân). Các câu hỏi trong binary decision tree đều có thể đưa được về dạng câu hỏi đúng hay sai. Các decision tree mà một leaf node có nhiều child node cũng có thể được đưa về dạng một binary decision tree. Điều này có thể đạt được vì hầu hết các câu hỏi đều có thể được đưa về dạng câu hỏi đúng sai.

Ví dụ, ta có thể xác định được tuổi của một người dựa trên nhiều câu hỏi đúng sai dạng: tuổi của bạn lớn hơn x đúng không? (Đây chính là thuật toán *tìm kiếm nhị phân – binary search*.)

Decision tree là một mô hình *supervised learning*, có thể được áp dụng vào cả hai bài toán *classification và regression*. Việc xây dựng một decision tree trên dữ liệu huấn luyện cho trước là việc đi xác định các câu hỏi và thứ tự của chúng. Một điểm đáng lưu ý của decision tree là nó có thể làm việc với các đặc trưng (trong các tài liệu về decision tree, các đặc trưng thường được gọi là thuộc tính – *attribute*) dạng *categorical*, thường là rời rạc và không có thứ tự. Ví dụ, mưa, nắng hay xanh, đỏ, v.v. Decision tree cũng làm việc với dữ liệu có vector đặc trưng bao gồm cả thuộc tính dạng categorical và liên tục (numeric). Một điểm đáng lưu ý nữa là decision tree ít yêu cầu việc chuẩn hoá dữ liệu.

1.3.2 Phân loại

Có 3 loại decision trees phổ biến sau:

- **ID3 (Iterative Dichotomiser 3)** - Tạo cây nhiều chiều, tìm cho mỗi node một đặt tính phân loại sao cho đặt tính này có giá trị "information gain" lớn nhất. Cây được phát triển tới mức tối đa kích thước. Sau đó áp dụng phương thức cắt tỉa cành để xử lý những dữ liệu chưa nhìn thấy.
- **C4.5** - Kế thừa từ ID3 nhưng loại bỏ hạn chế về việc chỉ sử dụng đặc tính có giá trị phân loại bằng cách tự động định nghĩa một thuộc tính rời rạc. Dùng để phân chia những thuộc tính liên tục thành những tập rời rạc.
- **CART (Classification and Regression Trees)** - Tương tự như C4.5, nhưng nó hỗ trợ

thêm đối tượng dự đoán là giá trị số (*Regression*). Cấu trúc CART dạng cây nhị phân, mỗi node sử dụng một ngưỡng để đạt được "information gian" lớn nhất.

Bảng 1.1 so sánh giữa các loại thuật toán decision tree.

1.3.3 Ưu và nhược điểm của thuật toán

Tùy vào loại Decision tree sử dụng mà ta có ưu nhược điểm riêng. Nhưng nhìn chung thuật toán có những ưu nhược điểm chung như sau:

Về ưu điểm

- Decision tree thường mô phỏng cách suy nghĩ con người. Vì vậy nó đơn giản để hiểu và diễn giải dữ liệu.
- Giúp ta nhìn thấy được sự logic của dữ liệu (không như thuật toán phần laoij SVM, KNN . . .)
- Có khả năng chọn được những features tốt nhất.
- Phân loại dữ liệu không cần tính toán phức tạp.
- Giải quyết vấn đề nhiễu và thiếu dữ liệu.
- Có khả năng xử lý dữ liệu có biến liên tục và rời rạc.

Về nhược điểm

- Tỉ lệ tính toán tăng theo hàm số mũ còn vấn đề ngày càng lớn hơn.
- Dễ bị vấn đề overfitting và high bias khi tập dữ liệu huấn luyện nhỏ.

Trong bài báo cáo này chúng tôi sử dụng loại **CART**. Do tính đơn giản, dể tiếp cận của nó, cũng như những giá trị feature mà ta sử dụng là kiểu dữ liệu biến liên tục không phải phân loại nên không dùng **ID3** được. Và đây là loại decision tree được thư viện *scikit-learn* chọn sử dụng.

1.3.4 Làm sạch dữ liệu

1.3.5 Quá trình xây dựng cây

1.4 Một số lỗ hổng tấn công web phổ biến

1.4.1 SQL Injection

1.4.2 Cross-Site Scripting (XSS)

Name	Splitting criteria	Attribute type	Missing values	Pruning Strategy	Outlier Detection
ID3	Information Gain	Handles only Categorical value	Do not handles missing values	No pruning is done	Susceptible to outliers
C4.5	Towing Criteria	Handles both Categorical & Numeric value	Handles missing values	Cross-Complexity is used	Can handle Outliers
CART	Gain Ratio	Handles both Categorical & Numeric value	handles missing values	Error Based pruning is used	Susceptible to outliers

Bảng 1.1: Bảng so sánh một số loại thuật toán Decision Trees

Chương 2

VẬN DỤNG THUẬT TOÁN VÀO PHÂN TÍCH TẬP DỮ LIỆU

2.1 Tập dữ liệu được sử dụng để training

2.2 Thực hiện áp dụng thuật toán vào tập dữ liệu

TÀI LIỆU THAM KHẢO