HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

ĐỖ THỊ LƯƠNG

NGHIÊN CỨU MỘT SỐ THUẬT TOÁN HỌC MÁY ĐỂ PHÂN LỚP DỮ LIỆU VÀ THỬ NGHIỆM

Chuyên ngành: Hệ Thống thông tin
Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - NĂM 2019

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **Tiến sỹ Vũ Văn Thỏa**

Phản biện 1: ……..

Phản biện 2: ……..

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng ..... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Trong thời gian gần đây, sự phát triển mạnh mẽ của công nghệ thông tin và các dịch vụ liên quan đã làm số lượng thông tin được trao đổi trên mạng Internet tăng một cách đáng kể. Số lượng thông tin được lưu trữ trong các kho dữ liệu cũng tăng với một tốc độ chóng mặt. Đồng thời, tốc độ thay đổi thông tin là cực kỳ nhanh chóng. Theo thống kê của Broder et al (2003), cứ sau 9 tháng hoặc 12 tháng lượng thông tin được lưu trữ, tìm kiếm và quản lý lại tăng gấp đôi. Hiện nay, loài người đang bước vào kỷ nguyên IoT (Internet of Things – Internet kết nối vạn vật). Thông qua internet, người dùng có nhiều cơ hội để tiếp xúc với nguồn thông tin vô cùng lớn. Tuy nhiên, cùng với nguồn thông tin vô tận đó, người dùng cũng đang phải đối mặt với sự quá tải thông tin. Đôi khi, để tìm được các thông tin cần thiết, người dùng phải chi phí một lượng thời gian khá lớn.

Với số lượng thông tin đồ sộ như vậy, một yêu cầu cấp thiết đặt ra là làm sao tổ chức, tìm kiếm và khai thác thông tin (dữ liệu) một cách hiệu quả nhất. Một trong các giải pháp được nghiên cứu để giải quyết vấn đề trên là xây dựng các mô hình tính toán dựa trên các phương pháp học máy nhằm phân loại, khai thác thông tin một cách tự động và trích xuất các tri thức hữu ích. Trong đó, bài toán phân lớp (Classification) dữ liệu có ý nghĩa hết sức quan trọng. Phân lớp dữ liệu là việc xếp các dữ liệu vào những lớp đã biết trước. Ví dụ: Phân lớp sinh viên theo kết quả học tập, phân lớp các loài thực vật, …

Bài toán phân lớp dữ liệu thường được giải quyết bằng cách sử dụng một số kỹ thuật học máy như: Thuật

toán Bayes (Naive Bayes), Cây quyết định (Decision Tree), Máy vector hỗ trợ (Support Vector Machine), Mạng Nơ-ron nhân tạo (Artificial Neural Network), …

Xuất phát từ những lý do trên, học viên chọn thực hiện đề tài luận văn tốt nghiệp chương trình đào tạo thạc sĩ có tên "**Nghiên cứu một số thuật toán học máy để phân lớp dữ liệu và thử nghiệm**".

Mục tiêu của luận văn là nghiên cứu các kỹ thuật học máy để giải quyết bài toán phân lớp dữ liệu nói chung và thử nghiệm đánh giá hiệu năng của chúng trên bộ dữ liệu KDD cup 99.

Nội dung của luận văn được trình bày trong ba chương nội dung chính như sau:

Chương 1: Tổng quan về phân lớp dữ liệu và học máy.

Nội dung chính của chương 1 là khảo sát tổng quan về bài toán phân lớp dữ liệu, học máy và các vấn đề liên quan.

Chương 2: Nghiên cứu một số thuật toán học máy

Nội dung chính của chương 2 là nghiên cứu chi tiết một số kỹ thuật học máy để giải quyết bài toán phân lớp dữ liệu và một số vấn đề liên quan.

Chương 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

Nội dung chính của chương 3 là thực hiện thử nghiệm và đánh giá các mô hình phân lớp dữ liệu dựa trên các phương pháp học máy đã nghiên cứu trong chương 2 cho bộ dữ liệu KDD cup 99.

CHƯƠNG 1. TỔNG QUAN VỀ PHÂN LỚP DỮ LIỆU VÀ HỌC MÁY

1.1. Giới thiệu bài toán phân lớp dữ liệu và các vấn đề liên quan

1.1.1. Khái niệm về phân lớp dữ liệu và bài toán phân lớp dữ liệu

Phân lớp (classification) dữ liệu là một tiến trình xử lý nhằm xếp các mẫu dữ liệu hay các đối tượng vào một trong các lớp đã được định nghĩa trước. Các mẫu dữ liệu hay các đối tượng được xếp vào các lớp dựa trên giá trị của các thuộc tính (attributes) của mẫu dữ liệu hay đối tượng. Quá trình phân lớp dữ liệu kết thúc khi tất cả các dữ liệu đã được xếp vào các lớp tương ứng. Khi đó, mỗi lớp dữ liệu được đặc trưng bởi tập các thuộc tính của các đối tượng chứa trong lớp đó.

Bài toán phân lớp dữ liệu có thể được mô tả như hình 1.1 dưới đây.

Hình 1.1. Bài toán phân lớp dữ liệu

Quy trình giải quyết bài toán phân lớp dữ liệu

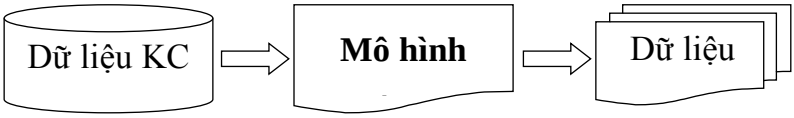
(1) Giai đoạn huấn luyện

Quá trình thực hiện giai đoạn học được mô tả trong hình 1.2.

Hình 1.2. Giai đoạn xây dựng mô hình phân lớp dữ liệu

(2) Giai đoạn kiểm chứng

Quá trình thực hiện giai đoạn phân lớp thử nghiệm được mô tả trong hình 1.3.



Hình 1.3. Quá trình kiểm tra đánh giá mô hình phân lớp dữ liệu

1.1.2. Các độ đo đánh giá mô hình phân lớp dữ liệu

(1) Độ đo Precision (Mức chính xác)

- Định nghĩa: Precision = TP / (TP + FP).

- Ý nghĩa: Giá trị Precision càng cao thể hiện khả năng càng cao để một kết quả phân lớp dữ liệu được đưa ra bởi bộ phân lớp là chính xác.

(2) Độ đo Recall (Độ bao phủ, độ nhạy hoặc độ triệu hồi)

- Định nghĩa: Recall = TP / (TP + FN).

- Ý nghĩa: Giá trị Recall càng cao thể hiện khả năng kết quả đúng trong số các kết quả đưa ra của bộ phân lớp càng cao.

(3) Độ đo Accuracy (Độ chính xác)

- Định nghĩa: Accuracy = (TP + TN) / (TP + TN + FP + FN) * 100%.

- Ý nghĩa: Accuracy phản ánh độ chính xác chung của bộ phân lớp dữ liệu..

(4) Độ đo F-Measure

- Định nghĩa: F-Measure = 2.(Precision.Recall) / (Precision + Recall).

- Ý nghĩa: F-Measure là độ đo nhằm đánh giá độ chính xác thông qua quá trình kiểm chứng dựa trên sự xem xét đến hai độ đo là Precision và Recall. Giá trị F-Measure càng cao phản ánh độ chính xác càng cao của bộ

phân lớp dữ liệu. Có thể coi độ đo F-Measure là trung bình điều hoà của hai độ đo Precision và Recall.

(5) Độ đo Specitivity (Độ đặc hiệu)

- Định nghĩa: Specitivity = TN/(TN+FP).

- Ý nghĩa: Độ đo Specitivity đánh giá khả năng một dữ liệu là phần tử âm được bộ phân lớp cho ra kết quả chính xác.

1.1.3. Các phương pháp đánh giá mô hình phân lớp dữ liệu

Phương pháp Hold-out

Phương pháp k-fold cross validation

1.1.4. Các ứng dụng của bài toán phân lớp dữ liệu

1.2. Tổng quan về học máy

1.2.1. Khái niệm về học máy và phân loại các kỹ thuật học máy

a. Khái niệm về học máy

Học máy là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể .

b. Phân loại các kỹ thuật học máy

Học có giám sát

Học không giám sát

Học bán giám sát

1.2.2. Ứng dụng học máy xây dựng mô hình phân lớp dữ liệu

1.3. Giới thiệu chung về học sâu

1.3.1. Khái niệm về học sâu

Học sâu là một chi của ngành học máy dựa trên một tập hợp các thuật toán để cố gắng mô hình dữ liệu trừu tượng hoá ở mức cao bằng cách sử dụng nhiều lớp xử

lý với cấu trúc phức tạp, hoặc bằng cách khác bao gồm nhiều biến đổi phi tuyến.

Các quá trình học sâu có thể mô tả như trong hình 1.4

Hình 1.4. Các quá trình học sâu

1.3.2. Hướng tiếp cận học sâu

Hướng tiếp cận học sâu đầu tiên thường được kể đến là các mạng nơ-ron sâu. Dưới đây, luận văn liệt kê một số dạng mạng nơ-ron sâu tham khảo trên mạng Internet.

Mạng nơ-ron tích chập

Mạng nơ-ron lặp

Mạng nơ-ron chuyển đổi

Học tăng cường

1.4. Kết luận chương 1

Trong chương 1 của luận văn đã giới thiệu bài toán phân lớp dữ liệu và khảo sát quy trình phân lớp dữ liệu cũng như các độ đo đánh giá các mô hình phân lớp dữ liệu và các ứng dụng khác nhau của phân lớp dữ liệu.

Trong chương này luận văn cũng trình bày tổng quan về các học máy và giới thiệu về học sâu.

Trong chương tiếp theo luận văn sẽ nghiên cứu ba thuật toán học máy để xây dựng mô hình phân lớp là cây quyết định, Bayes và máy vectơ hỗ trợ.

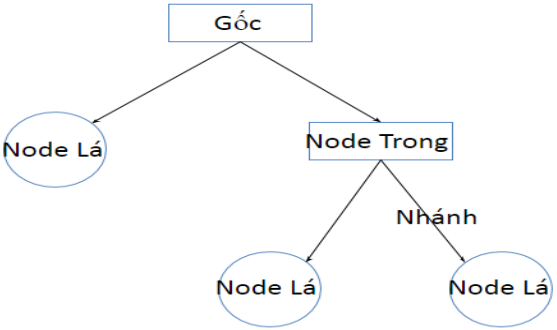
CHƯƠNG 2. NGHIÊN CỨU MỘT SỐ THUẬT TOÁN HỌC MÁY

2.1. Khảo sát thuật toán cây quyết định và các vấn đề liên quan

2.1.1. Giới thiệu phương pháp

Cây quyết định là một cấu trúc ra quyết định có dạng cây. Cây quyết định nhận đầu vào là một bộ giá trị các thuộc tính mô tả một đối tượng hay một tình huống và trả về một giá trị rời rạc. Mỗi bộ thuộc tính đầu vào được gọi là một mẫu hay một ví dụ, đầu ra gọi là lớp hay nhãn phân lớp. Khi đó, với tập thuộc tính đầu vào được cho dưới dạng véc tơ x, nhãn phân lớp đầu ra được ký hiệu là y thì cây quyết định có thể xem như một hàm f(x) = y.

Cây quyết định được biểu diễn dưới dạng một cấu trúc cây như trong Hình 2.1 dưới đây.



Hình 2.1. Mô hình cây quyết định

2.1.2. Xây dựng cây quyết định dựa trên Entropy

2.1.3. Đánh giá phương pháp

Mô hình phân lớp dữ liệu sử dụng cây quyết định có các ưu điểm sau đây.

- Cây quyết định tự giải thích và khi được gắn kết lại, chúng có thể dễ dàng tự sinh ra.

- Cây quyết định có thể xử lý được nhiều kiểu các thuộc tính đầu vào. Cây quyết định được xem như là một phương pháp phi tham số.

Bên cạnh đó, cây quyết định cũng có những nhược điểm

2.2. Khảo sát thuật toán Bayes và các vấn đề liên quan

2.2.1. Giới thiệu phương pháp

Ý tưởng cơ bản của cách tiếp cận phân lớp dữ liệu Bayes là sử dụng công thức Bayes về xác suất có điều kiện để lựa chọn kết quả phân lớp là sự kiện có xác suất lớn nhất.

Công thức Bayes:

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)} \qquad (2.2)$$

Trong đó:

- H (Hypothesis) là giả thuyết và E (Evidence) là chứng cứ hỗ trợ cho giả thuyết H.
- P(E|H): xác suất E xảy ra khi H xảy ra (xác suất có điều kiện, khả năng của E khi H đúng) thường gọi là xác suất tiên nghiệm.
- P(H|E): xác suất hậu nghiệm của H nếu biết E.

2.2.2. Thuật toán Naïve Bayes

Thuật toán phân lớp Naive Bayes (Naive Bayes Classification - NBC) thường được gọi ngắn gọn là thuật toán là Naive Bayes [19]. Thuật toán Naive Bayes dựa trên định lý Bayes (2.2) để đưa ra các phán đoán cũng như phân loại dữ liệu dựa trên các dữ liệu được quan sát và thống kê.

2.2.3. Mạng Bayes

2.2.4. Đánh giá phương pháp

So với các phương pháp khác, phương pháp phân lớp dữ liệu Bayes lập luận theo kinh nghiệm được tích lũy và áp dụng vào mô hình phân lớp đối tượng khá linh hoạt và phù hợp với đặc trưng của bài toán cụ thể. Các cơ chế ước lượng trong phương pháp này cũng gần gũi với cách suy luận thông thường. Phương pháp phân lớp dữ liệu Bayes được ứng dụng rất rộng rãi bởi tính dễ hiểu và dễ triển khai.

Tuy nhiên, phương pháp phân lớp dữ liệu Bayes cho hiệu quả không cao trong trường hợp tập dữ liệu mẫu có độ phức tạp lớn và các thuộc tính của dữ liệu mẫu có quan hệ phụ thuộc hoặc không đầy đủ. Trong những trường hợp này, có thể sử dụng mạng Bayes.

2.3. Khảo sát thuật toán máy vectơ hỗ trợ và các vấn đề liên quan

2.3.1. Giới thiệu phương pháp

Máy vector hỗ trợ (Support Vector Machines - SVM) được Cortes và Vapnik giới thiệu vào năm 1995 trên cơ sở mở rộng từ chuyên đề lý thuyết học thống kê (Vapnik 1982), dựa trên nguyên tắc tối thiểu rủi ro cấu trúc (structural risk minimization). Ý tưởng chính của SVM để giải quyết bài toán phân lớp (1.1)-(1.2) là ánh xạ tập dữ liệu mẫu thành các vector điểm trong không gian vector R^d và tìm các siêu phẳng có hướng để chia tách chúng thành các lớp khác nhau.

2.3.2. Thuật** toán **SVM tuyến tính với tập dữ liệu phân tách được

2.3.3. Thuật toán SVM tuyến tính với tập dữ liệu không phân tách được

2.3.4. Thuật toán SVM phi tuyến phân lớp nhị phân

2.3.5. Thuật toán tối thiểu tuần tự SMO

2.3.6. Thuật toán SVM phân lớp đa lớp

2.3.7. Đánh giá phương pháp

Ưu điểm nổi bật của phương pháp SVM là thực hiện tối ưu toàn cục cho mô hình phân lớp. Do đó, mô hình SVM có chất lượng cao, chịu đựng được nhiễu. Mặt khác, SVM là một phương pháp tốt (phù hợp) đối với những bài toán phân lớp có không gian biểu diễn thuộc tính lớn. Các đối tượng cần phân lớp được biểu diễn bởi một tập rất lớn các thuộc tính.

Tuy nhiên, phương pháp SVM cũng có một số nhược điểm

2.4. Kết luận chương 2

Chương 2 đã khảo sát tương đối chi tiết các kỹ thuật học máy: phương pháp cây quyết định, phương pháp Bayes và phương pháp SVM. Đây là các kỹ thuật học máy thường được ứng dụng giải quyết bài toán phân lớp dữ liệu.

Trong chương 3 tiếp theo, luận văn sẽ áp dụng thử nghiệm các phương pháp trên cho bài toán phân loại tấn công mạng trên bộ dữ liệu KDD cup 99.

CHƯƠNG 3. THỬ NGHIỆM VÀ ĐÁNH GIÁ

3.1. Khảo sát và lựa chọn bộ dữ liệu để thử nghiệm

3.1.1. Giới thiệu chung

An ninh mạng là vấn đề an ninh phi truyền thống, còn khá mới mẻ nhưng ngày càng được thế giới và Việt Nam quan tâm cả cấp vĩ mô và vi mô.

Tại Việt Nam hiện có trên 55% dân số đang sử dụng điện thoại di động, trên 52% dân số sử dụng Internet [22]. Việt Nam đứng thứ 4 trên thế giới về thời gian sử dụng Internet và đứng thứ 22 trên thế giới tính theo dân số về số người sử dụng mạng xã hội. Hằng năm, Việt Nam phải chịu hàng ngàn cuộc tấn công mạng và Việt Nam đứng thứ 20 trên thế giới về xếp hạng các quốc gia bị tấn công mạng nhiều nhất, chịu thiệt hại lên tới 10.400 tỉ đồng riêng năm 2016 so với mức 8.700 tỉ đồng năm 2015 [17].

Trong năm 2017, Việt Nam đã hứng chịu rất nhiều các vụ tấn công mạng và để lại rất nhiều hậu quả nặng nề. Chỉ riêng quý 1 năm 2017, Việt Nam đã có gần 7700 sự cố tấn công mạng tại Việt Nam. Đến giữa tháng 9 số lượng các sự cố tấn công mạng đã lên đến gần 10000 [20] (số liệu của Trung tâm ứng cứu khẩn cấp máy tính Việt Nam – VNCERT). Trong đó có 1762 sự cố website lừa đảo, 4595 sự cố phát tán mã độc và 3607 sự cố tấn công thay đổi giao diện.

Theo báo cáo an ninh website của CyStack, chỉ trong quý 3 năm 2018 đã có 1.183 website của Việt Nam bị tin tặc tấn công và kiểm soát. Trong đó, các website giới thiệu sản phẩm và dịch vụ của doanh nghiệp là đối

tượng bị tin tặc tấn công nhiều nhất (chiếm 71,51%). Vị trí thứ hai là các website thương mại điện tử (chiếm 13,86%).

Tháng 11/2018, Diễn đàn RaidForums đã đăng tải thông tin được cho là dữ liệu của hơn 5 triệu khách hàng của chuỗi bán lẻ thiết bị Thế giới di động. Những thông tin bị rỏ rì bao gồm địa chỉ email, lịch sử giao dịch và thậm chí là cả số thẻ ngân hàng. Ngay sau đó, dữ liệu được cho là các hợp đồng trong chương trình F.Friends của FPT Shop cũng bị rò rỉ. Một số công ty Việt Nam như: Công ty cổ phần Con cưng, Ngân hàng hợp tác xã Việt Nam, ... cũng trở thành đích nhắm cho tin tặc.

Theo thống kê từ Trung tâm Giám sát an toàn không gian mạng quốc gia trực thuộc Cục An toàn thông tin (Bộ Thông tin và Truyền thông), có khoảng 4,7 triệu địa chỉ IP của Việt Nam thường xuyên nằm trong các mạng mã độc lớn (số liệu tháng 11/2018).

Trong quý I/2019, VNCERT ghi nhận có 4.770 sự cố tấn công mạng vào các trang web của Việt Nam. Cũng trong thời gian này hệ thống giám sát của VNCERT ghi nhận tổng cộng có hơn 78,3 triệu sự kiện mất an toàn thông tin tại Việt Nam.

Các thông tin và số liệu trên cho thấy một thực trạng đáng báo động về tấn công mạng tại Việt Nam hiện nay.

Như vậy, vấn đề phòng chống tấn công mạng đang là chủ đề nghiên cứu trở nên cấp thiết hơn trong bối cảnh bùng nổ cách mạng công nghệ truyền thông, Internet vạn vật và mạng xã hội gia tăng kết nối toàn cầu. Một trong những hướng nghiên cứu là xây dựng các hệ thống phòng chống tấn công mạng dựa trên các kỹ thuật học máy [16].

Từ những lý do trên, luận văn lựa chọn bộ dữ liệu về tấn công mạng KDD Cup 99 để thử nghiệm và đánh giá các mô hình phân lớp dữ liệu dựa trên các phương pháp học máy đã nghiên cứu trong chương 2.

3.1.2. Mô tả bộ dữ liệu KDD Cup 99

Dưới sự bảo trợ của Cơ quan Quản lý Nghiên cứu Dự Án Phòng Thủ Tiên tiến thuộc Bộ Quốc phòng Mỹ (DARPA) và phòng thí nghiệm nghiên cứu không quân (AFRL), năm 1998 phòng thí nghiệm MIT Lincoln đã thu thập và phân phối bộ dữ liệu được coi là bộ dữ liệu tiêu chuẩn cho việc đánh giá các nghiên cứu trong hệ thống phát hiện xâm nhập mạng máy tính. Dữ liệu được sử dụng trong cuộc thi KDD cup 99 là một phiên bản của bộ dữ liệu DARPA 98 [18].

Tập dữ liệu đầy đủ của bộ KDD cup 99 chứa 4.898.430 dòng dữ liệu, đây là một khối lượng dữ liệu lớn. Trong nghiên cứu và thử nghiệm, tập dữ liệu 10% của bộ KDD cup 99 thường được lựa chọn. Tập 10% của bộ KDD 99 tuy là tập con nhưng nó mang đầy đủ dữ liệu cho các loại hình tấn công khác nhau, đầy đủ thông tin quan trọng để thử nghiệm

Từ đó, các kiểu tấn công khác nhau trong bộ dữ liệu được nhóm thành 5 loại (gán nhãn lớp) của bộ dữ liệu KDD cup'99 bao gồm:

1. Normal: dữ liệu thể hiện loại kết nối TCP/IP bình thường;

2. DoS (Denial of Service): dữ liệu thể hiện loại tấn công từ chối dịch vụ;

3. Probe: dữ liệu thể hiện loại tấn công thăm dò;

4. R2L (Remote to Local): dữ liệu thể hiện loại tấn công từ xa khi hacker cố gắng xâm nhập vào mạng hoặc các máy tính trong mạng;

5. U2R **(**User to Root): dữ liệu thể hiện loại tấn công chiếm quyền Root (quyền cao nhất) bằng việc leo thang đặc quyền từ quyền người dùng bình thường lên quyền Root.

Trong bộ dữ liệu KDD cup 99, với mỗi kết nối TCP/IP có 41 thuộc tính số và phi số được trích xuất. Đồng thời, mỗi kết nối được gán nhãn (thuộc tính 42) giúp phân biệt kết nối bình thường (Normal) và các tấn công.

Xây dựng kịch bản và lựa chọn công cụ thử nghiệm

3.1.3. Xây dựng kịch bản thử nghiệm

Bài toán đặt ra là phân loại kiểu tấn công trong bộ dữ liệu KDD cup 99 nhằm hỗ trợ cho các hệ thống phát hiện xâm nhập mạng. Đây là bài toán được nhiều tác giả quan tâm nghiên cứu trong thời gian gần đây. Có thể tham khảo các kết quả nghiên cứu chi tiết trong các tài liệu [1], [2], [6], [8], [9], [11] và [16].

Trong mục này, luận văn sẽ thực hiện thử nghiệm với bài toán sau:

Đầu vào của bài toán:

(1) Bộ dữ liệu KDD cup 99;

(2) Các thuật toán thử nghiệm:

- Thuật toán Cây quyết định (Decision Tree);
- Thuật toán Bayes;
- Thuật toán máy vecto hỗ trợ (SMV).

Đầu ra của bài toán:

Các độ đo đánh giá hiệu năng các mô hình phân loại kiểu tấn công sử dụng các thuật toán thử nghiệm trên bộ dữ liệu KDD cup 99.

Luận văn sẽ tiến hành thử nghiệm theo hai kịch bản.

3.1.4. Lựa chọn công cụ thử nghiệm

Weka là một phần mềm miễn phí về học máy được viết bằng Java, phát triển bởi University of Wekato. Weka có thể coi như là bộ sưu tập các thuật toán về học máy dùng trong phân tích và khai phá dữ liệu. Các thuật toán đã được xây dựng sẵn và người dùng chỉ việc lựa chọn để sử dụng.

Các tính năng chính của Weka:

Các môi trường chính trong Weka:

3.2. Triển khai thử nghiệm và đánh giá kết quả

3.2.1. Mô tả thử nghiệm

3.2.2. Kết quả thử nghiệm

(1) Kết quả giai đoạn huấn luyện của các mô hình theo kịch bản 1

Bảng 3.1: Kết quả thử nghiệm 2 lớp của thuật toán j48

```
=== Detailed Accuracy By Class ===
TP Rate FP Rate  Precision  Recall   F-
Measure     Class
0.996   0.004    0.996      0.996     0.996
normal
0.996   0.004    0.995      0.996     0.995
anomaly
0.996   0.004    0.996      0.996     0.996
(Avg.)
=== Confusion Matrix ===
  a        b   <-- classified as
 13389    60 |     a = normal
    51 11692 |     b = anomaly
```

Bảng 3.2: Kết quả thử nghiệm 2 lớp của thuật toán Naïve-Bayes

```
=== Detailed Accuracy By Class ===
TP Rate FP Rate  Precision  Recall   F-
Measure      Class
0,912   0,123    0,895      0,912     0,903
normal
0,877   0,088    0,897      0,877     0,887
anomaly
0,896   0,106    0,896      0,896     0,896
(Avg.)
=== Confusion Matrix ===
     a     b   <-- classified as
 12272  1177 |     a = normal
  1445 10298 |     b = anomaly
```

Bảng 3.3: Kết quả thử nghiệm 2 lớp của thuật toán Net-Bayes

```
=== Detailed Accuracy By Class ===
TP Rate FP Rate  Precision  Recall   F-
Measure     Class
0,991    0,064    0,947      0,991    0,969
normal
0,936    0,009    0,989      0,936    0,962
anomaly
0,966    0,038    0,967      0,966    0,966
(Avg.)
=== Confusion Matrix ===
  a       b   <-- classified as
 13330   119 |     a = normal
   747 10996 |     b = anomaly
```

Bảng 3.4: Kết quả thử nghiệm 2 lớp của thuật toán SMO

```
=== Detailed Accuracy By Class ===
TP Rate FP Rate  Precision  Recall   F-
Measure      Class
0.986   0.041    0.965      0.986
0.975          normal
0.959   0.014    0.984      0.959
0.971          anomaly
0.973   0.029    0.974      0.973
0.973          (Avg.)
=== Confusion Matrix ===
     a     b   <-- classified as
 13261   188 |     a = normal
   485 11258 |     b = anomaly
```

Bảng 3.5: Tổng hợp kết quả huấn luyện 2 lớp của các thuật toán thử nghiệm

Thuật toán	accuracy (%)	Normal			Anomaly		
		Pre	Rec	F1	Pre	Rec	F1
J48	99.55	99.6	99.6	99.6	99.5	99.6	99.5
NaiveBayes	89.59	89.5	91.2	90.3	89.7	87.7	88.7
BayesNet	96.56	94.7	99.1	96.9	98.9	93.6	96.2
SMO	97.32	96.5	98.6	97.5	98.4	95.9	97.1

(2) Kết quả giai đoạn kiểm chứng của các mô hình theo kịch bản 1

Kết quả kiểm chứng các mô hình được tổng hợp trong bảng 3.8.

Bảng 3.6: Tổng hợp kết quả kiểm chứng 2 lớp của các thuật toán thử nghiệm

Thuật toán	accuracy (%)	Normal			Anomaly		
		Pre	Rec	F1	Pre	Rec	F1
J48	63.97	32	87.3	46.8	95.4	58.8	72.8
NaiveBayes	55.77	24.3	67.8	35.8	88.2	53.1	66.3
BayesNet	51.68	25.7	87.8	39.8	94.2	43.7	59.7
SMO	52.7	22.7	66.9	33.9	87.1	49.5	63.2

(3) Kết quả giai đoạn huấn luyện thử nghiệm theo kịch bản 2

Bảng 3.7: Tổng hợp kết quả huấn luyện đa lớp của các thuật toán thử nghiệm

Các lớp	Các độ đo	Thuật toán			
		J48	NaiveBayes	BayesNet	SMO
Normal	Prec	99.40	95.80	97.80	97.60
	Rec	99.70	77.90	95.10	98.80
	F1	99.50	85.90	96.40	98.20
DoS	Prec	99.80	96.50	99.80	99.30
	Rec	99.90	95.00	93.60	98.00
	F1	99.80	95.80	96.60	98.70
U2R	Prec	42.90	0.60	3.40	66.70
	Rec	27.30	72.70	63.60	18.20
	F1	33.30	1.10	6.50	28.60
R2L	Prec	99.10	22.20	47.80	77.50
	Rec	82.80	52.20	94.30	64.10
	F1	86.70	31.10	63.40	70.20
Probe	Prec	99.10	61.40	79.20	96.80
	Rec	98.30	88.00	98.10	96.40

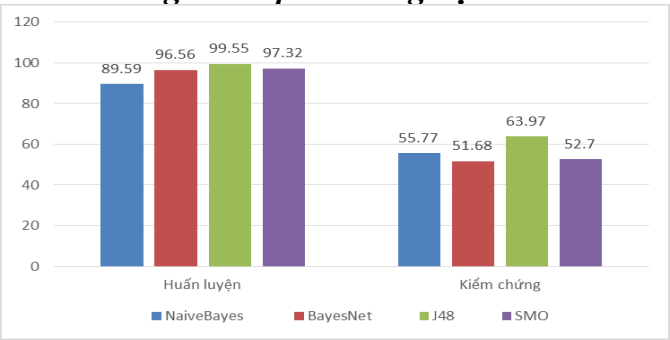
Các lớp	Các độ đo	Thuật toán			
		J48	NaiveBayes	BayesNet	SMO
	F1	98.70	72.40	87.70	96.60
accuracy (%)		99.44	84.86	94.79	97.98

(4) Kết quả giai đoạn kiểm chứng thử nghiệm theo kịch bản 2

Bảng 3.8: Tổng hợp kết quả kiểm chứng đa lớp của các thuật toán thử nghiệm

Các lớp	Các độ đo	Thuật toán			
		J48	NaiveBayes	BayesNet	SMO
Normal	Prec	58.00	40.30	53.20	41.50
	Rec	87.00	54.70	84.10	69.10
	F1	69.60	46.40	65.20	51.80
DoS	Prec	97.30	79.40	98.60	95.70
	Rec	96.80	69.10	61.30	86.50
	F1	97.00	73.90	75.60	90.90
U2R	Prec	76.50	2.40	9.70	83.30
	Rec	35.10	32.40	62.20	13.50
	F1	48.10	4.50	16.80	23.30
R2L	Prec	16.70	22.20	81.70	22.20
	Rec	0.10	1.00	19.90	0.20
	F1	0.20	1.80	32.00	0.30
Probe	Prec	83.80	71.50	68.70	51.60
	Rec	99.70	92.00	99.80	56.20
	F1	81.10	80.50	81.40	53.80
accuracy (%)		77.04	56.16	66.70	61.12

3.2.3. Đánh giá kết quả thử nghiệm

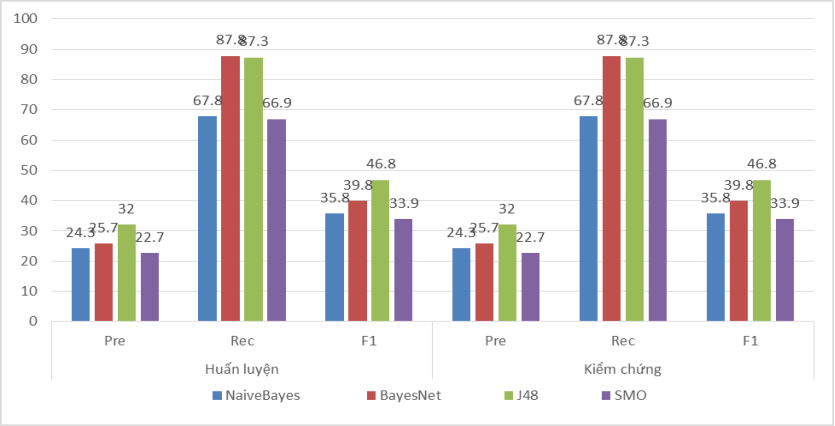


Hình 3.2 Biểu đồ so sánh độ chính xác của các thuật toán thử nghiệm 2 lớp

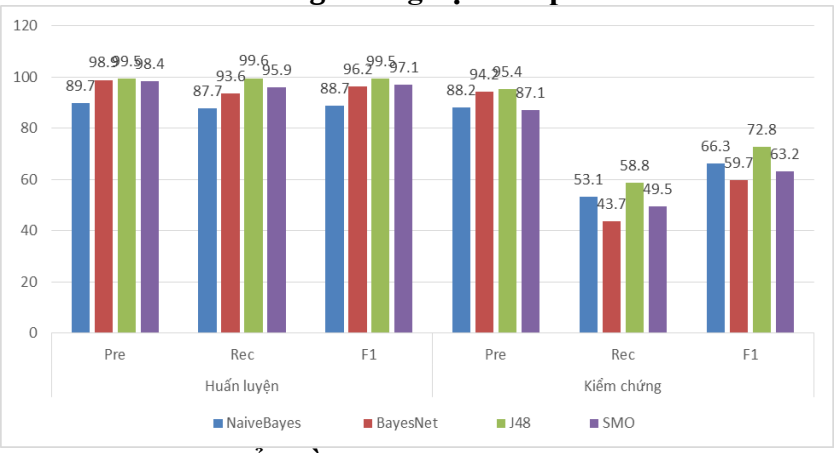
Quan sát biểu đồ trên hình 3.1 nhận thấy rằng, các thuật toán thử nghiệm đều cho kết quả có tỉ lệ phân loại chính xác cao trên tập huấn luyện (từ 90% trở lên).

Trong đó, mô hình cây quyết định (j48) có tỉ lệ phân loại chính xác cao nhất (99.55%) và mô hình Naïve Bayes tỉ lệ phân loại chính xác thấp nhất (89.59%).

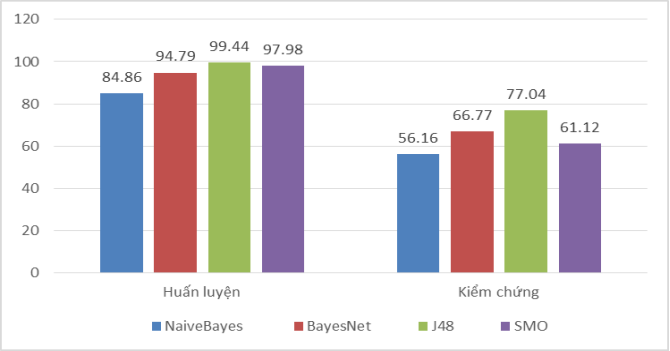
Tuy nhiên, khi thực hiện kiểm thử tỷ lệ phân loại chính xác bị sụt giảm rõ rệt chỉ còn trên 51%. Trong đó, mô hình cây quyết định (j48) có tỉ lệ phân loại chính xác cao nhất (63.97%) và mô hình Bayes Net tỉ lệ phân loại chính xác thấp nhất (51.68%).



Hình 3.3 Biểu đồ so sánh độ chính xác của lớp Normal trong thử nghiệm 2 lớp



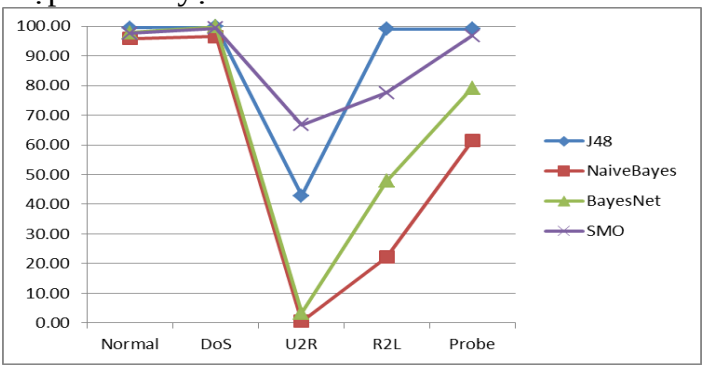
Hình 3.4 Biểu đồ so sánh độ chính xác của lớp Anomaltrong thử nghiệm 2 lớp



Hình 3.5 Biểu đồ so sánh độ chính xác của mô hình trong thử nghiệm đa lớp

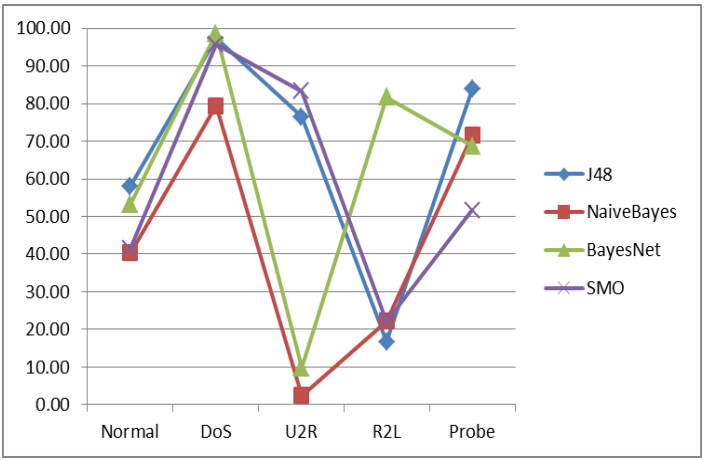
Quan sát trên hình 3.2, và 3.5 nhận thấy kết quả các mô hình khi thực hiện phân lớp đa lớp khi kiểm chứng cho kết quả độ chính xác cao hơn khi chỉ thực hiện phân lớp 2 lớp. Điều này có thể được lý giải là các mô hình khi thực hiện phân lớp đa lớp sẽ phù hợp hơn.

Hình 3.5 trình bày biểu đồ thống kê mức chính xác (Precision) theo từng lớp của các mô hình thử nghiệm đa lớp trên tập huấn luyện.



Hình 3.6 Mức chính xác theo lớp trong thử nghiệm đa lớp trên tập huấn luyện

Hình 3.7 trình bày biểu đồ thống kê mức chính xác (Precision) theo từng lớp của các mô hình thử nghiệm phân lớp đa lớp trên tập kiểm chứng.



Hình 3.7 Mức chính xác theo lớp trong thử nghiệm đa lớp trên tập kiểm chứng

Tóm lại, trong cả hai kịch bản thử nghiệm, mô hình cây quyết định và mô hình SVM có độ chính khá tốt. Điều này cũng phù hợp với thực tế là hai mô hình này thường được sử dụng để xây dựng các bộ phân lớp.

3.3. Kết luận chương 3

Trong chương 3 luận văn đã tiến hành thử nghiệm các thuật toán học máy nghiên cứu trong chương 2 cho bài toán phân loại tấn công mạng với bộ dữ liệu KDD cup 99.

Kết quả thử nghiệm bước đầu cho thấy các thuật toán học máy có thể triển khai trong thực tế và phù hợp với các yêu cầu đề ra cho bài toán phân lớp dữ liệu.

KẾT LUẬN

Các kết quả đạt được của luận văn:

Với mục tiêu nghiên cứu các thuật toán học máy cho bài toán phân lớp dữ liệu và thử nghiệm, luận văn đã đạt được một số kết quả sau đây:

- Nghiên cứu tổng quan về bài toán phân lớp dữ liệu và các vấn đề liên quan.
- Khảo sát tổng quan về học máy nhằm bài toán phân lớp dữ liệu.
- Giới thiệu chung về học sâu.
- Khảo sát chi tiết các phương pháp học máy: Cây quyết định, Bayes và SVM.
- Khảo sát bộ dữ liệu tấn công mạng KDD cup 99.
- Thực hiện thử nghiệm các thuật toán học máy j48, Naïve Bayes, Bayes Net và SMO để phân loại các kiểu tấn công mạng đối với bộ dữ liệu NSL-KDD.

Tuy nhiên, do hạn chế về mặt thời gian, luận văn chưa tiến hành thử nghiệm với các bộ dữ liệu lớn, Do đó, hiệu quả thử nghiệm chưa cao.

Hướng phát triển tiếp theo:

- Thực hiện xây dựng và triển khai hệ thống phân lớp dữ liệu sử dụng thuật toán học máy cho các bài toán thực tế.
- Nghiên cứu các kỹ thuật học sâu cho bài toán phân lớp dữ liệu.