

**ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

ĐINH MẠNH CƯỜNG

**PHÁT HIỆN XÂM NHẬP DỰA TRÊN THUẬT TOÁN
K-MEANS**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên 2015

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

ĐINH MẠNH CƯỜNG

PHÁT HIỆN XÂM NHẬP DỰA TRÊN THUẬT TOÁN
K-MEANS

Chuyên ngành: Khoa học máy tính

Mã số: 60.48.01.01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC

PGS.TS. NGUYỄN VĂN TAM

Thái Nguyên 2015

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn “**Phát hiện xâm nhập dựa trên thuật toán K-Means**” là công trình nghiên cứu do tôi thực hiện dưới sự hướng dẫn của PGS.TS. Nguyễn Văn Tam. Các nội dung được trình bày trong luận văn là những kết quả đạt được trong thời tôi gian thực đề tài dưới sự hướng của tập thể giáo viên hướng dẫn, tôi không sao chép nguyên bản lại kết quả của các nghiên cứu đã từng được công bố và đây cũng là kết quả của quá trình nghiên cứu, học tập và làm việc nghiêm túc của tôi trong quá trình học cao học. Bên cạnh đó, trong một số nội dung luận văn là kết quả phân tích, nghiên cứu, tổng hợp từ nhiều nguồn tài liệu khác. Các thông tin tổng hợp hay các kết quả lấy từ nhiều nguồn tài liệu khác đã được tôi trích dẫn một cách đầy đủ và hợp lý. Nguồn tài liệu tham khảo có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Các số liệu và thông tin sử dụng trong luận văn này là trung thực.

Thái Nguyên, ngày 20 tháng 07 năm 2015

Người cam đoan

Đinh Mạnh Cường

LỜI CẢM ƠN

Tôi xin chân thành cảm ơn các thầy, cô trong Viện Công nghệ thông tin, Trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên đã tham gia giảng dạy, giúp đỡ tôi trong suốt quá trình học tập nâng cao trình độ kiến thức để phục vụ cho công tác giảng dạy của tôi hiện tại và sau này.

Tôi xin bày tỏ lòng biết ơn chân thành tới PGS.TS. Nguyễn Văn Tam, các Thầy đã tận tình hướng dẫn hướng dẫn tôi trong suốt thời gian thực hiện luận văn.

Vì điều kiện thời gian và trình độ có hạn nên luận văn cũng không thể tránh khỏi những thiếu sót. Tôi xin kính mong các Thầy, Cô giáo, các bạn đồng nghiệp đóng góp ý kiến để đề tài được hoàn thiện hơn.

Tôi xin chân thành cảm ơn!

MỤC LỤC

MỞ ĐẦU	4
Chương 1: KHÁI QUÁT BÀI TOÁN PHÁT HIỆN XÂM NHẬP	4
1.1. Định nghĩa về phát hiện xâm nhập.....	4
1.1.1. Định nghĩa.	4
1.1.2. Sự khác nhau giữa IDS/IPS.	4
1.2. Các thành phần và chức năng của hệ thống phát hiện thâm nhập.	5
1.2.1. Thành phần thu thập gói tin.....	6
1.2.2. Thành phần phát hiện gói tin.....	6
1.2.3. Thành phần phản hồi.....	6
1.3. Phân loại phát hiện xâm nhập.....	7
1.3.1. Network based IDS – NIDS7	
1.3.2. Host based IDS – HIDS9	
1.4. Các phương pháp phát hiện xâm nhập.....	11
1.4.1. Mô hình phát hiện sự lạm dụng11	
1.4.2. Mô hình phát hiện sự bất thường.....12	
1.4.3. So sánh giữa hai mô hình15	
Chương 2: PHÁT HIỆN XÂM NHẬP DỰA TRÊN THUẬT TOÁN K-MEANS .17	
2.1 Thuật toán K-means17	
2.1.1 Các khái niệm17	
2.1.2 Thuật toán.....20	
2.1.3 Nhược điểm của K-Means và cách khắc phục.....35	
2.2. Thuật toán K-means với phát hiện xâm nhập.....35	
2.2.1 Phân tích tập dữ liệu kiểm thử.....35	
2.2.2 Mô hình phát hiện bất thường dựa trên thuật toán K-means.....39	
Chương 3: XÂY DỰNG CHƯƠNG TRÌNH PHÁT HIỆN XÂM NHẬP DỰA TRÊN THUẬT TOÁN K-MEANS47	
3.1 Mô tả bài toán47	
3.2 Mô tả dữ liệu đầu vào.....47	
3.2.1 Mô tả các thuộc tính trong file dữ liệu đầu vào48	
3.2.2 Giảm số lượng bản ghi trong dữ liệu đầu vào:50	
3.3 Cài đặt thuật toán K-Means và thử nghiệm trong phân cụm phân tử dị biệt .53	
3.3.1 Giới thiệu về môi trường cài đặt53	
3.3.2 Các chức năng của chương trình53	
3.4. Nhận xét, đánh giá chương trình thử nghiệm.....59	
KẾT LUẬN VÀ HƯỚNG NGHIÊN CỨU60	
TÀI LIỆU THAM KHẢO61	
PHẦN PHỤ LỤC62	

DANH MỤC HÌNH ẢNH

Hình 1.1: Các vị trí đặt IDS trong mạng.....	4
Hình 1.2: Mô hình kiến trúc hệ thống phát hiện xâm nhập (IDS).....	5
Hình 1.3: Mô hình NIDS	7
Hình 2.1 Ví dụ về phân nhóm đối tượng.....	17
Hình 2.2: Các thiết lập để xác định ranh giới các cụm ban đầu	18
Hình 2.3: Mô tả độ đo khoảng cách giữa các đối tượng.	19
Hình 2.4: Sơ đồ thuật toán phân nhóm K-Means.....	21
Hình 2.5: Mô tả trực quan quá trình phân cụm dữ liệu.	22
Hình 2.6: Biểu diễn các đối tượng trên mặt phẳng tọa độ x, y	25
Hình 2.7: Biểu diễn các đối tượng, phân tử trung tâm trên mặt phẳng tọa độ x, y	26
Hình 2.8: Biểu diễn các đối tượng, phân tử trung tâm trên mặt phẳng tọa độ x, y (Vòng lặp 1).....	29
Hình 2.9: Biểu diễn các đối tượng, phân tử trung tâm trên mặt phẳng tọa độ x, y (Vòng lặp 2)	31
Hình 2.10: Biểu diễn các đối tượng, phân tử trung tâm trên mặt phẳng tọa độ x, y (Vòng lặp 3)	33
Hình 2.11: Mô hình hệ thống phát hiện bất thường sử dụng thuật toán K-means ..	40
Hình 2.12: Bốn quan hệ của một cuộc tấn công	42
Hình 2.13: Mô tả hoạt động của môđun tổng hợp	44
Hình 3.1: Giảm số bản ghi cho file đầu vào của chương trình.....	51
Hình 3.2: Xem và chỉnh sửa cho file đầu vào của chương trình nếu cần.	52
Hình 3.3: Dữ liệu của chương trình mở bằng Notepad.	52
Hình 3.5: Giao diện chọn bộ dữ liệu.....	54
Hình 3.6: Hiển thị chi tiết dữ liệu đầu vào.....	55
Hình 3.7: Form thực hiện thuật toán K-Means.....	56
Hình 3.8: Kết quả thực hiện thuật toán K-Means.	57
Hình 3.9: Số bản ghi kết nối thuộc về mỗi cụm.....	58
Hình 3.10: Kết quả thực hiện thuật toán K-Means với bộ dữ liệu có 494020 bản ghi kết nối.	58

DANH MỤC BẢNG

Bảng 2.1: Danh mục các đối tượng	24
Bảng 2.2: Bảng biểu diễn các thuộc tính trên mặt phẳng x,y	24
Bảng 2.3: Khởi tạo các phân tử trọng tâm	25
Bảng 2.4: Bảng khoảng cách Euclidean (vòng lặp 1)	28
Bảng 2.5: Tìm khoảng cách min giữa các khoảng cách (Vòng lặp 1)	28
Bảng 2.6: Kết quả phân nhóm các đối tượng (vòng lặp 1)	28
Bảng 2.7: Phân tử trọng tâm (vòng lặp 1)	29
Bảng 2.8: Bảng khoảng cách Euclidean (Vòng lặp 2)	30
Bảng 2.9: Tìm khoảng cách min giữa các khoảng cách (Vòng lặp 2)	30
Bảng 2.10: Kết quả phân nhóm các đối tượng (vòng lặp 2)	31
Bảng 2.11: Phân tử trọng tâm (vòng lặp 2)	31
Bảng 2.12: Bảng khoảng cách Euclidean (vòng lặp 3)	32
Bảng 2.13: Tìm khoảng cách min giữa các khoảng cách (vòng lặp 3)	32
Bảng 2.14: Kết quả phân nhóm các đối tượng (vòng lặp 3)	33
Bảng 2.15: Phân tử trọng tâm (vòng lặp 3)	33
Bảng 2.16: Kết quả phân nhóm các đối tượng (vòng lặp 4)	34
Bảng 2.17: Bảng kết quả phân nhóm thuốc	34
Bảng 2.18: danh sách các cảnh báo chưa rút gọn	45
Bảng 2.19: Danh sách các cảnh báo sau khi đã rút gọn	46
Bảng 3.1: Các thuộc tính cơ bản (nhóm này chứa tất cả các thuộc tính có được từ một kết nối TCP / IP)	48
Bảng 3.2: Các thuộc tính lưu thông (nhóm này bao gồm các thuộc tính mà nó được tính toán với khoảng thời gian một cửa sổ)	49
Bảng 3.2: Các thuộc tính nội dung	49

BẢNG TỪ VIẾT TẮT

IDS	Intrusion Detection System
IPS	Intrusion Prevention Systems
HIDS	Host-based IDS
NIDS	Network-based IDS

MỞ ĐẦU

Ngày nay, hệ thống mạng máy tính đã trở nên rất phổ biến và được ứng dụng trong hầu hết các hoạt động kinh tế-xã hội của nước ta. Tuy nhiên, mạng máy tính cũng phải đương đầu với nhiều thách thức, đặc biệt là vấn đề an toàn và bảo mật dữ liệu trên mạng. Trong các mối đe dọa đối với an ninh mạng thì việc xâm nhập mạng để thay đổi thông tin, lấy cắp dữ liệu và phá hoại hạ tầng mạng là nghiêm trọng nhất. Chính vì vậy, việc phát hiện và ngăn chặn xâm nhập mạng máy tính là chủ đề đang được quan tâm nghiên cứu và phát triển ứng dụng mạnh mẽ hiện nay. Phát hiện và ngăn chặn được hiểu là xác định xâm nhập và ngăn chặn một cách nhanh nhất khi nó xảy ra. Hiện nay không có phương pháp phát hiện truy nhập trái phép nào là hoàn hảo bởi các kỹ thuật xâm nhập ngày càng tinh vi và luôn luôn được đổi mới. Khi phương pháp phát hiện xâm nhập được biết đến thì những kẻ xâm nhập sẽ sửa những chiến lược và thử một kiểu xâm nhập mới.

Trong quá trình học tập và nghiên cứu tôi đã tìm hiểu được một số kiến thức về phát hiện xâm nhập và tìm hiểu các thuật toán để phát hiện xâm nhập đó nhằm đảm bảo thông tin trao đổi được đảm bảo an toàn. Chính vì thế tôi đã lựa chọn chủ đề “**Phát hiện xâm nhập dựa trên thuật toán K-means.**” là đề tài nghiên cứu cho luận văn của mình.

* Cấu trúc của luận văn bao gồm 3 chương như sau:

Chương 1: Chương này trình bày những kiến thức cơ bản về phát hiện xâm nhập như: định nghĩa, các thành phần và chức năng của hệ thống, phân loại, và các phương pháp phát hiện xâm nhập.

Chương 2: Chương này trình bày về việc phát hiện xâm nhập dựa trên thuật toán K-means. Nội dung của thuật toán, ví dụ minh họa thuật toán, tập dữ liệu kiểm thử và mô hình phát hiện xâm nhập dựa trên thuật toán K-means.

Chương 3: Chương này là kết quả cài đặt bài toán phát hiện xâm nhập dựa trên thuật toán k-means.

Chương 1

KHÁI QUÁT BÀI TOÁN PHÁT HIỆN XÂM NHẬP

1.1. Định nghĩa về phát hiện xâm nhập

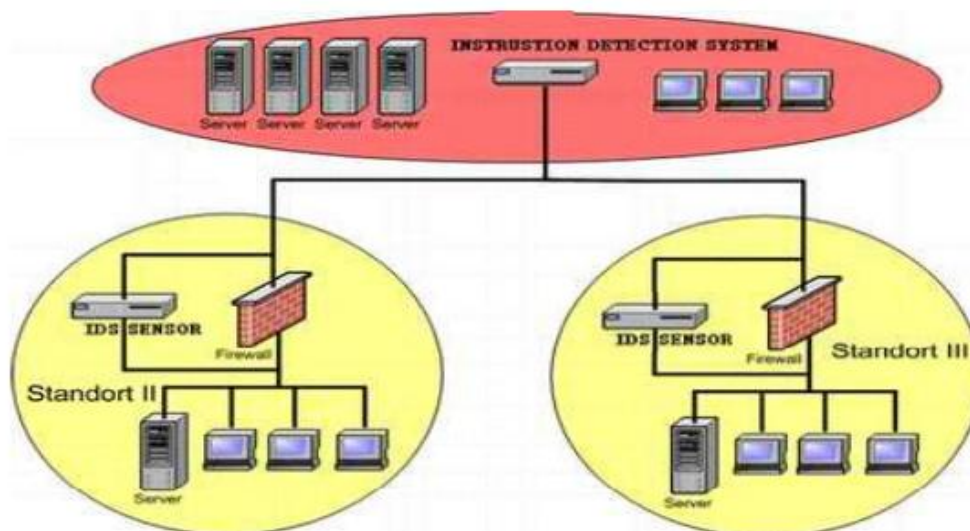
1.1.1. Định nghĩa

Hệ thống phát hiện xâm nhập (IDS) là hệ thống có nhiệm vụ theo dõi, phát hiện và (có thể) ngăn cản sự xâm nhập, cũng như các hành vi khai thác trái phép tài nguyên của hệ thống được bảo vệ mà có thể dẫn đến việc làm tổn hại đến tính bảo mật, tính toàn vẹn và tính sẵn sàng của hệ thống.[6]

Hệ thống IDS sẽ thu thập thông tin từ rất nhiều nguồn trong hệ thống được bảo vệ sau đó tiến hành phân tích những thông tin đó theo các cách khác nhau để phát hiện những xâm nhập trái phép.

Khi một hệ thống IDS có khả năng ngăn chặn các nguy cơ xâm nhập mà nó phát hiện được thì nó được gọi là một hệ thống phòng chống xâm nhập hay IPS.

Hình sau minh họa các vị trí thường cài đặt IDS trong mạng:



Hình 1.1: Các vị trí đặt IDS trong mạng.

1.1.2. Sự khác nhau giữa IDS/IPS

Có thể nhận thấy sự khác biệt giữa hai khái niệm ngay ở tên gọi: “phát hiện” và “ngăn chặn”. Các hệ thống IDS được thiết kế với mục đích chủ yếu là phát hiện và cảnh báo các nguy cơ xâm nhập đối với mạng máy tính nó đang bảo vệ trong khi đó, một hệ thống IPS ngoài khả năng phát hiện còn có thể tự hành động chống lại