

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

---



**Đỗ Xuân Cường**

**KỸ THUẬT PHÂN CỤM DỮ LIỆU TRONG  
PHÁT HIỆN XÂM NHẬP TRÁI PHÉP**

Chuyên ngành: Khoa học máy tính  
Mã số: 60 48 0101

**LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH**

**NGƯỜI HƯỚNG DẪN KHOA HỌC: TS LƯƠNG THẾ DŨNG**

## LỜI CẢM ƠN

Đầu tiên em xin gửi lời cảm ơn sâu sắc nhất tới TS Lương Thế Dũng, người hướng dẫn khoa học, đã tận tình chỉ bảo, giúp đỡ em thực hiện luận văn.

Em xin cảm ơn các thầy cô trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên đã giảng dạy và truyền đạt kiến thức cho em.

Em xin trân thành cảm ơn các đồng chí Lãnh đạo Sở Thông tin và Truyền thông và các đồng nghiệp đã tạo mọi điều kiện giúp đỡ em hoàn thành nhiệm vụ học tập.

Em cũng xin bày tỏ lòng biết ơn đối với gia đình, bạn bè và người thân đã động viên khuyến khích và giúp đỡ trong suốt quá trình hoàn thành luận văn này.

Mặc dù đã hết sức cố gắng hoàn thành luận văn với tất cả sự nỗ lực của bản thân, nhưng luận văn vẫn còn những thiếu sót. Kính mong nhận được những ý kiến đóng góp của quý Thầy, Cô và bạn bè đồng nghiệp.

***Em xin trân thành cảm ơn!***

**LỜI CAM ĐOAN**

Luận văn là kết quả nghiên cứu và tổng hợp các kiến thức mà bản thân đã thu thập được trong quá trình học tập tại trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên, dưới sự hướng dẫn, giúp đỡ của các thầy cô và bạn bè đồng nghiệp, đặc biệt là sự hướng dẫn của TS Lương Thế Dũng – Trưởng khoa An toàn thông tin, Học viện Kỹ thuật Mật mã.

Em xin cam đoan luận văn không phải là sản phẩm sao chép của bất kỳ công trình khoa học nào.

*Thái Nguyên, ngày tháng năm 2015*

**HỌC VIÊN**

**Đỗ Xuân Cường**

## MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT	v
DANH MỤC CÁC BẢNG	vi
DANH MỤC HÌNH VẼ	vii
LỜI NÓI ĐẦU	1
CHƯƠNG I: TỔNG QUAN VỀ TẤN CÔNG MẠNG MÁY TÍNH VÀ CÁC PHƯƠNG PHÁP PHÁT HIỆN	3
1.1. Các kỹ thuật tấn công mạng máy tính	3
1.1.1. Một số kiểu tấn công mạng.....	3
1.1.2. Phân loại các mối đe dọa trong bảo mật hệ thống .....	6
1.1.3. Các mô hình tấn công mạng	9
1.2. Một số kỹ thuật tấn công mạng	12
1.2.1. Tấn công thăm dò.....	12
1.2.2. Tấn công xâm nhập.....	12
1.2.3. Tấn công từ chối dịch vụ.....	13
1.2.4. Tấn công từ chối dịch vụ cổ điển.....	13
1.2.5. Tấn công dịch vụ phân tán DDoS.....	14
1.3. Hệ thống phát hiện xâm nhập trái phép	18
1.3.1. Khái niệm về hệ thống phát hiện xâm nhập trái phép .....	18
1.3.2. Các kỹ thuật phát hiện xâm nhập trái phép.....	21
1.3.3. Ứng dụng kỹ thuật khai phá dữ liệu cho việc phát hiện xâm nhập trái phép.....	24
CHƯƠNG II: MỘT SỐ KỸ THUẬT PHÂN CỤM DỮ LIỆU	26
2.1. Phân cụm phân hoạch	26
2.1.1. Thuật toán K-means .....	27
2.1.2. Thuật toán CLARA.....	30
2.1.3. Thuật toán CLARANS.....	31
2.2. Phân cụm phân cấp	33
2.2.1. Thuật toán CURE.....	34

2.2.2. Thuật toán CHAMELEON .....	37
2.3. Phân cụm dựa trên mật độ	39
2.3.1. Thuật toán DBSCAN .....	40
2.3.2. Thuật toán OPTICS.....	42
2.4. Phân cụm dựa trên lưới	44
2.4.1. Thuật toán STING.....	45
2.4.2. Thuật toán CLIQUE .....	47
2.4.3. Thuật toán WaveCluster.....	49
2.5. Phân cụm dựa trên mô hình	52
2.5.1. Thuật toán EM.....	52
2.5.2. Thuật toán COBWEB .....	54
2.6. Phân cụm dữ liệu mờ	55
CHƯƠNG III: ỨNG DỤNG KỸ THUẬT PHÂN CỤM DỮ LIỆU TRONG PHÁT HIỆN XÂM NHẬP TRÁI PHÉP	56
3.1. Mô hình bài toán	56
3.1.1. Thu thập dữ liệu .....	56
3.1.2. Trích rút và lựa chọn thuộc tính.....	59
3.1.3. Xây dựng bộ phân cụm .....	62
3.2. Xây dựng các thực nghiệm phát hiện xâm nhập trái phép	63
3.2.1. Môi trường và công cụ thực nghiệm.....	63
3.2.2. Tiến hành các thực nghiệm và kết quả đạt được.....	64
KẾT LUẬN	71

**DANH MỤC CÁC TỪ VIẾT TẮT**

<b>TT</b>	<b>Viết tắt</b>	<b>Nội dung</b>
1.	CNTT	Công nghệ thông tin
2.	ATTT	An toàn thông tin
3.	CSDL	Cơ sở dữ liệu
4.	IDS	Hệ thống phát hiện xâm nhập
5.	PHXN	Phát hiện xâm nhập
6.	KDD	Khám phá tri thức trong cơ sở dữ liệu
7.	KPDL	Khai phá dữ liệu
8.	PCDL	Phân cụm dữ liệu
9.	PAM	Thuật toán phân cụm phân hoạch

**DANH MỤC CÁC BẢNG**

Bảng 3.1: Bảng mô tả lớp tấn công từ chối dịch vụ (DoS).....	57
Bảng 3.2: Bảng mô tả lớp tấn công trình sát(Probe).....	58
Bảng 3.3: Bảng mô tả lớp tấn công leo thang đặc quyền (U2R). ....	58
Bảng 3.4: Bảng mô tả lớp tấn công truy cập từ xa (R2L).....	59
Bảng 3.5: Bảng mô tả 41 thuộc tính của tập dữ liệu KDD Cup 1999 .....	61
Bảng 3.6: Bảng phân phối số lượng bản ghi. ....	62
Bảng 3.7: Kết quả phân cụm K-means với các cụm k khác nhau .....	65
Bảng 3.8: Kết quả phân cụm EM với các cụm k khác nhau .....	67
Bảng 3.9: Bảng so sánh kết quả phân cụm thuật toán K-means và EM .....	70

## DANH MỤC HÌNH VẼ

Hình 1.1: Mô hình tấn công truyền thống.....	9
Hình 1.2: Mô hình tấn công phân tán.....	10
Hình 1.3: Các bước tấn công mạng.....	10
Hình 1.4: Tổng quan về một sơ đồ hình cây của tấn công DDoS.....	16
Hình 1.5: Đặt một sensor phía sau hệ thống Firewall.....	21
Hình 1.6: Mô tả dấu hiệu xâm nhập.....	22
Hình 1.7: Quá trình khai phá dữ liệu của việc xây dựng mô hình PHXN.....	24
Hình 2.1 Ví dụ các bước của thuật toán k-means .....	29
Hình 2.2: Các cụm dữ liệu được khám phá bởi CURE .....	35
Hình 2.3: Ví dụ thực hiện phân cụm bằng thuật toán CURE .....	37
Hình 2.4: Mô hình CHAMELEON, Phân cụm phân cấp dựa trên k-láng giềng gần và mô hình hóa động .....	38
Hình 2.5: Hình dạng các cụm được khám phá bởi thuật toán DBSCAN .....	42
Hình 2.6: Sắp xếp cụm trong OPTICS phụ thuộc vào $\epsilon$ [8] .....	44
Hình 2.7: Một mẫu không gian đặc trưng 2 chiều .....	51
Hình 2.8: Đa phân giải của không gian đặc trưng trong hình 2.7. a) Tỷ lệ 1; b) Tỷ lệ 2; c) Tỷ lệ 3.....	52
Hình 3.1: Các bước xây dựng mô hình phát hiện xâm nhập trái phép .....	56
Hình 3.2: Số lượng bản ghi có trong tập dữ liệu thực nghiệm.....	62
Hình 3.3: Tập dữ liệu đưa vào phân cụm qua Weka Explorer .....	64
Hình 3.4: Tham số cài đặt phân cụm K-means với Weka Explorer .....	65
Hình 3.5: Tham số cài đặt phân cụm EM với Weka Explorer.....	66
Hình 3.6: Trực quan kết quả sau khi phân cụm (k=5) với Weka Explorer.....	67
Hình 3.7: Phân cụm k-means trong Cluster 3.0.....	68
Hình 3.8: Mô hình đồ họa trực quan kết quả sau các kiểu tấn công.....	69
Hình 3.9: Biểu đồ so sánh kết quả phân cụm thuật toán K-means và EM .....	70



## LỜI NÓI ĐẦU

Công nghệ thông tin liên tục phát triển và thay đổi, nhiều phần mềm mới ra đời mang đến cho con người nhiều tiện ích hơn, lưu trữ được nhiều dữ liệu hơn, tính toán tốt hơn, sao chép và truyền dữ liệu giữa các máy tính nhanh chóng thuận tiện hơn,... Hệ thống mạng máy tính của các đơn vị được trang bị nhưng vẫn tồn tại nhiều lỗ hổng và các nguy cơ về mất an toàn thông tin. Các vụ xâm nhập mạng lấy cắp thông tin nhạy cảm cũng như phá hủy thông tin diễn ra ngày càng nhiều, thủ đoạn của kẻ phá hoại ngày càng tinh vi.

Công nghệ phát hiện xâm nhập trái phép hiện nay hầu hết dựa trên phương pháp đối sánh mẫu, phương pháp này cho kết quả phát hiện khá tốt, tuy nhiên nó đòi hỏi các hệ thống phát hiện xâm nhập trái phép phải xây dựng được một cơ sở dữ liệu mẫu khổng lồ và liên tục phải cập nhật. Vì vậy hiện nay lĩnh vực nghiên cứu để tìm ra các phương pháp phát hiện xâm nhập trái phép hiệu quả hơn đang được rất nhiều người quan tâm. Trong đó, một hướng quan trọng trong lĩnh vực này dựa trên các kỹ thuật khai phá dữ liệu [1].

Hiện nay hầu hết các cơ quan, tổ chức, doanh nghiệp đều có hệ thống mạng máy tính riêng kết nối với mạng Internet và ứng dụng nhiều các chương trình, phần mềm CNTT vào các hoạt động sản xuất kinh doanh. Việc làm này đã góp phần tích cực trong quản lý, điều hành, kết nối, quảng bá và là chìa khoá thành công cho sự phát triển chung của họ và cộng đồng. Trong các hệ thống mạng máy tính đó có chứa rất nhiều các dữ liệu, các thông tin quan trọng liên quan đến hoạt động của các cơ quan, tổ chức, doanh nghiệp.

Sự phát triển mạnh của hệ thống mạng máy tính cũng là một vùng đất có nhiều thuận lợi cho việc theo dõi và đánh cắp thông tin của các nhóm tội phạm tin học, việc xâm nhập bất hợp pháp và đánh cắp thông tin của các tổ

chức, đơn vị đang đặt ra cho thế giới vấn đề làm thế nào để có thể bảo mật được thông tin của tổ chức, đơn vị mình. Phát hiện xâm nhập bảo đảm an toàn an ninh mạng là những yếu tố được quan tâm hàng đầu trong các các tổ chức, đơn vị. Đã có những đơn vị thực hiện việc thuê một đối tác thứ 3 với việc chuyên đảm bảo cho hệ thống mạng và đảm bảo an toàn thông tin cho đơn vị mình, cũng có những đơn vị đưa ra các kế hoạch tính toán chi phí cho việc mua sản phẩm phần cứng, phần mềm để nhằm đáp ứng việc đảm bảo an toàn an ninh thông tin. Tuy nhiên đối với những giải pháp đó các tổ chức, đơn vị đều phải thực hiện cân đối về chính sách tài chính hàng năm với mục đích làm sao cho giải pháp an toàn thông tin là tối ưu và có được chi phí rẻ nhất và đảm bảo thông tin trao đổi được an toàn, bảo vệ thông tin của đơn vị mình trước những tấn công của tội phạm công nghệ từ bên ngoài do vậy mà đề tài Kỹ thuật phân cụm dữ liệu trong phát hiện xâm nhập trái phép dựa trên mã nguồn mở được phát triển giúp được phần nào yêu cầu của các tổ chức, đơn vị về an toàn thông tin và đảm bảo an toàn cho hệ thống mạng.

Đề tài “Kỹ thuật phân cụm dữ liệu trong phát hiện xâm nhập trái phép” học viên thực hiện với mong muốn xây dựng một cách hệ thống về các nguy cơ tiềm ẩn về xâm nhập trái phép vào mạng máy tính, các phương pháp phân cụm dữ liệu và cụ thể cách thức để ứng dụng kỹ thuật phân cụm dữ liệu trong phát hiện xâm nhập trái phép, đảm bảo an toàn an ninh thông tin cho tổ chức, đơn vị.