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 HIỆNXÂM 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

TT	Viết tắt	Nội dung
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

DANH MỤC HÌNH VẼ

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 cónhiềuthuận 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 đangđặt 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 yếutố 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 mềmđể nhằm đáp ứng việc đảm bảo an toàn an ninh thông tin. Tuy nhiên đối với những giảipháp đó các tổ chức, đơn vị đều phải thực hiện cân đối về chính sách tài chínhhằng năm với mục đích làm sao cho giải pháp an toàn thông tin là tối ưu và cóđược chi phí rẻ nhất và đảm bảo thông tin trao đổi được an toàn, bảo vệ thôngtin của đơn vị mình trước những tấn công của tội phạm công nghệ từ bên ngoàido vậy mà đề tài Kỹ thuật phân cụm dữ liệu trong phát hiện xâm nhập trái phép dựa trên mã nguồn mở đượcphát triển giúp được phần nào yêu cầu của các tổ chức, đơn vị về an toàn thông tinvà 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.

CHƯƠNG I: TỔNG QUAN VỀ TẤN CÔNG MẠNG MÁY TÍNH VÀ CÁC PHƯƠNG PHÁP PHÁT HIỆN

1.1.Các kỹ thuật tấn công mạng máy tính

Hiện nay vẫn chưa có định nghĩa chính xác về thuật ngữ "tấn công" (xâm nhập, công kích). Mỗi chuyên gia trong lĩnh vực ATTT luận giải thuật ngữ này theo ý hiểu của mình. Ví dụ, "xâm nhập - là tác động bất kỳ đưa hệ thống từ trạng thái an toàn vào tình trạng nguy hiểm". Thuật ngữ này có thể giải thích như sau: "xâm nhập - đó là sự phá huỷ chính sách ATTT" hoặc "là tác động bất kỳ dẫn đến việc phá huỷ tính toàn vẹn, tính bí mật, tính sẵn sàng của hệ thống và thông tin xử lý trong hệ thống".

Tấn công (attack) là hoạt động có chủ ý của kẻ phạm tội lợi dụng các thương tổn của hệ thống thông tin và tiến hành phá vỡ tính sẵn sàng, tính toàn vẹn và tính bí mật của hệ thống thông tin .

Tấncông(*attack,intrustion*)mạnglàcáctácđộnghoặclàtrìnhtựliênkếtgiữacáctácđộngvớinhauđểpháhuỷ,dẫnđếnviệchiệnthựchoácácnguycơbằngcáchlợi dụngđặctínhdễbịtổnthươngcủacáchệthốngthôngtinnày.Cónghĩalà,nếucóthểbàitrừnguycơthươngtổncủacáchệthôngtinchínhlàtrừbỏkhảnăngcóthểthựchiệntấncông.

Để thực hiện được tấn công mạng, thì người thực hiện tấn công phải có sự hiểu biết về giao thức TCP/IP, có hiểu biêt về hệ điều hành và sử dụng thành thạo một số ngôn ngữ lập trình. Khi đó kẻ tấn công sẽ xác định phương hướng tấn công vào hệ thống.

1.1.1. Một số kiểu tấn công mạng

Có rất nhiều dạng tấn công mạng đang được biết đến hiện nay, dựa vào hành động tấn công được phân thành các loại là tấn công thăm dò, tấn công sử dụng mã độc, tấn công xâm nhập mạng và tấn công từ chối dịch vụ.

Hoặc chúng ta có thể chia thành 2 loại tấn công chung nhất là tấn công chủ động và tấn công thụ động.

- Tấn công chủ động (active attack): Kẻ tấn công thay đổi hoạt động của hệ thống và hoạt động của mạng khi tấn công và làm ảnh hưởng đến tính toàn vẹn, sẵn sàng và xác thực của dữ liệu.

- Tấn công bị động (passive attack): Kẻ tấn công cố gắng thu thập thông tin từ hoạt động của hệ thống và hoạt động của mạng làm phá vỡ tính bí mật của dữ liệu.

Dựa vào nguồn gốc của cuộc tấn công thì có thể phân loại tấn công thành 2 loại hình tấn công bao gồm: tấn công từ bên trong và tấn công từ bên ngoài, tấn công trực tiếp.

- Tấn công bên trong bao gồm những hành vi mang tính chất xâm nhập hệ thống nhằm mục đích phá hoại. Kẻ tấn công bên trong thường là những người nằm trong một hệ thống mạng nội bộ, lấy thông tin nhiều hơn quyền cho phép.

Tấncôngkhôngchủý:Nhiềuhưhạicủamạngdongười dùngtrongmạngvôýgâynên.Nhữngngườinàycóthểvôýđểhackerbênngoàihệthống lấyđượcpasswordhoặclàmhỏngcáctàinguyêncủamạngdothiếuhiểubiết.

Tấncôngcóchủý:Kẻtấncóchủýchốnglạicácquitắc,cácquiđịnhdocácchínhsá channinhmạngđưara.

- Tấn công bên ngoài là những tấn công xuất phát từ bên ngoài hệ thống như Internet hay các kết nối truy cập từ xa; gồm có:

+ Kẻ tấn công nghiệp dư (“script-kiddy”): Dùng các script đã tạo sẵn và có thể tạo nên các các thiệt hại đối với mạng.

+ Kẻ tấn công đích thực (“true- hacker”): Mục đích chính của nhóm người này khi thực hiện các tấn công mạng là để mọi người thừa nhận khả năng của họ và để được nổi tiếng.

+ Kẻ tấn công chuyên nghiệp (“the elite”): Thực hiện các tấn công mạng là để thu lợi bất chính.

Tấn công bên ngoài có thể là những dạng tất công trực tiếp, các dạng tấn công này thông thường là sử dụng trong giai đoạn đầu để chiếm quyền truy cập. Phổ biến nhất vẫn là cách dò tìm tên người sử dụng và mật khẩu. Tội phạm mạng có thể sử dụng những thông tin liên quan đến chủ tài khoản như ngày tháng năm sinh, tên vợ (chồng) hoặc con cái hoặc số điện thoại để dò tìm thông tin tài khoản và mật khẩu với mục đích chiếm quyền điều khiển của một tài khoản, thông thường đối với những tài khoản có mật khẩu đơn giản thì tội phạm mạng chỉ dò tìm mật khẩu qua thông tin chủ tài khoản, một cách tiếp cận việc chiếm quyền truy nhập bằng cách tìm tài khoản và mật khẩu tài khoảng khác là dùng chương trình để dò tìm mật khẩu. Phương pháp này trong một số khả năng hữu dụng thì có thể thành công đến 30%. Một kiểu tấn công bên ngoài khác được đề cập đến nữa chính là hình thức nghe trộm, việc nghe trộm thông tin trên mạng có thể đưa lại những thông tin có ích như tên, mật khẩu của người sử dụng, các thông tin mật chuyển qua mạng. Việc nghe trộm thường được tiến hành ngay sau khi kẻ tấn công đã chiếm được quyền truy nhập hệ thống, thông qua các chương trình cho phép đưa card giao tiếp mạng (Network Interface Card-NIC) vào chế độ nhận toàn bộ các thông tin lưu truyền trên mạng. Những thông tin này cũng có thể dễ dàng lấy được trên Internet.

- Một số các lỗi khác liên quan đến con người, hệ thống cũng là những kiểu tấn công trực tiếp từ bên ngoài nhưng có mức độ phức tạp và khó khăn

hơn, nguy hiểm nhất là yếu tố con người bởi nó là một trong nhiều điểm yếu nhất trong bất kỳ hệ thống bảo mật nào.

- Khi một mạng máy tính bị tấn công, nó sẽ bị chiếm một lượng lớn tài nguyên trên máy chủ, mức độ chiếm lượng tài nguyên này tùy thuộc vào khả năng huy động tấn công của tội phạm mạng, đến một giới hạn nhất định khả năng cung cấp tài nguyên của máy chủ sẽ hết và như vậy việc từ chối các yêu cầu sử dụng dịch vụ của người dùng hợp pháp bị từ chối. Việc phát động tấn công của tội phạm mạng còn tùy thuộc vào số lượng các máy tính ma mà tội phạm mạng đó đang kiểm soát, nếu khả năng kiểm soát lớn thì thời gian để tấn công và làm sập hoàn toàn một hệ thống mạng sẽ nhanh và cấp độ tấn công sẽ tăng nhanh hơn, tội phạm mạng có thể một lúc tấn công nhiều hệ thống mạng khác nhau tùy vào mức độ kiểm soát chi phối các máy tính ma như thế nào.

1.1.2. Phân loại các mối đe dọa trong bảo mật hệ thống

a) Mối đe dọa bên trong

Thuật ngữ mối đe dọa bên trong được sử dụng để mô ta một kiểu tấn công được thực hiện từ một người hoặc một tổ chức có quyền truy cập vào hệ thống mạng. Các cách tấn công từ bên trong được thực hiện từ một khu vực được coi là vùng tin cậy trong hệ thống mạng. Mối đe dọa này có thể khó phòng chống hơn vì các nhân viên hoặc những tổ chức có quyền hạn trong hệ thống mạng sẽ truy cập vào mạng và dữ liệu bí mật của doanh nghiệp. Phần lớn các doanh nghiệp hiện nay đều có tường lửa ở các đường biên mạng và họ tin tưởng hoàn toàn vào các ACL (Access Control List) và quyền truy cập vào server để qui định cho sự bảo mật bên trong. Quyền truy cập server thường bảo vệ tài nguyên trên server nhưng không cung cấp bất kì sự bảo vệ nào cho

mạng. Mối đe dọa ở bên trong thường được thực hiện bởi các nhân viên, tổ chức bất bình, muốn "quay mặt" lại với doanh nghiệp. Nhiều phương pháp bảo mật liên quan đến vành đai của hệ thống mạng, bảo vệ mạng bên trong khỏi các kết nối bên ngoài, như là truy cập Internet. Khi vành đai của hệ thống mạng được bảo mật, các phần tin cậy bên trong có khuynh hướng bị bớt nghiêm ngặt hơn. Khi một kẻ xâm nhập vượt qua vỏ bọc bảo mật cứng cáp đó của hệ thống mạng, mọi chuyện còn lại thường là rất đơn giản. Các mạng không dây giới thiệu một lĩnh vực mới về quản trị bảo mật. Không giống như mạng có dây, các mạng không dây tạo ra một khu vực bao phủ có thể bị can thiệp và sử dụng bởi bất kì ai có phần mềm đúng và một adapter của mạng không dây. Không chỉ tất cả các dữ liệu mạng có thể bị xem và ghi lại mà các sự tấn công vào mạng có thể được thực hiện từ bên trong, nơi mà cơ sở hạ tầng dễ bị nguy hiểm hơn nhiều. Vì vậy, các phương pháp mã hóa mạnh luôn được sử dụng trong mạng không dây.

b) Mối đe dọa từ bên ngoài

Mối đe dọa ở bên ngoài là từ các tổ chức, chính phủ, hoặc cá nhân cố gắng truy cập từ bên ngoài mạng của doanh nghiệp và bao gồm tất cả những người không có quyền truy cập vào mạng bên trong. Thông thường, các kẻ tấn công từ bên ngoài cố gắng từ các server quay số hoặc các kết nối Internet. Mối đe dọa ở bên ngoài là những gì mà các doanh nghiệp thường phải bỏ nhiều hầu hết thời gian và tiền bạc để ngăn ngừa.

c) Mối đe dọa không có cấu trúc

Mối đe dọa không có cấu trúc là mối đe dọa phổ biến nhất đối với hệ thống của một doanh nghiệp. Các hacker mới vào nghề, thường được gọi là script kiddies, sử dụng các phần mềm để thu thập thông tin, truy cập hoặc thực hiện một kiểu tấn công DoS vào một hệ thống của một doanh nghiệp.

Script kiddies tin tưởng vào các phần mềm và kinh nghiệm của các hacker đi trước.

Khi script kiddies không có nhiều kiến thức và kinh nghiệm, họ có thể tiến hành phá hoại lên các doanh nghiệp không được chuẩn bị. Trong khi đây chỉ là trò chơi đối với các kiddie, các doanh nghiệp thường mất hàng triệu đô la cũng như là sự tin tưởng của cộng đồng. Nếu một web server của một doanh nghiệp bị tấn công, cộng đồng cho rằng hacker đã phá vỡ được sự bảo mật của doanh nghiệp đó, trong khi thật ra các hacker chỉ tấn công được một chỗ yếu của server. Các server Web, FTP, SMTP và một vài server khác chứa các dịch vụ có rất nhiều lổ hổng để có thể bị tấn công, trong khi các server quan trọng được đặt sau rất nhiều lớp bảo mật. Cộng đồng thường không hiểu rằng phá vỡ một trang web của một doanh nghiệp thì dễ hơn rất nhiều so với việc phá vỡ cơ sở dữ liệu thẻ tín dụng của doanh nghiệp đó. Cộng đồng phải tin tưởng rằng một doanh nghiệp rất giỏi trong việc bảo mật các thông tin riêng tư của nó.

d) Mối đe dọa có cấu trúc

Mối đe dọa có cấu trúc là khó ngăn ngừa và phòng chống nhất vì nó xuất phát từ các tổ chức hoặc cá nhân sử dụng một vài loại phương pháp luận thực hiện tấn công. Các hacker với kiến thức, kinh nghiệm cao và thiết bị sẽ tạo ra mối đe dọa này. Các hacker này biết các gói tin được tạo thành như thế nào và có thể phát triển mã để khai thác các lỗ hổng trong cấu trúc của giao thức. Họ cũng biết được các biện pháp được sử dụng để ngăn ngừa truy cập trái phép, cũng như các hệ thống IDS và cách chúng phát hiện ra các hành vi xâm nhập. Họ biết các phương pháp để tránh những cách bảo vệ này. Trong một vài trường hợp, một cách tấn công có cấu trúc được thực hiện với sự trợ giúp từ một vài người ở bên trong. Đây gọi là mối đe dọa có cấu trúc ở bên

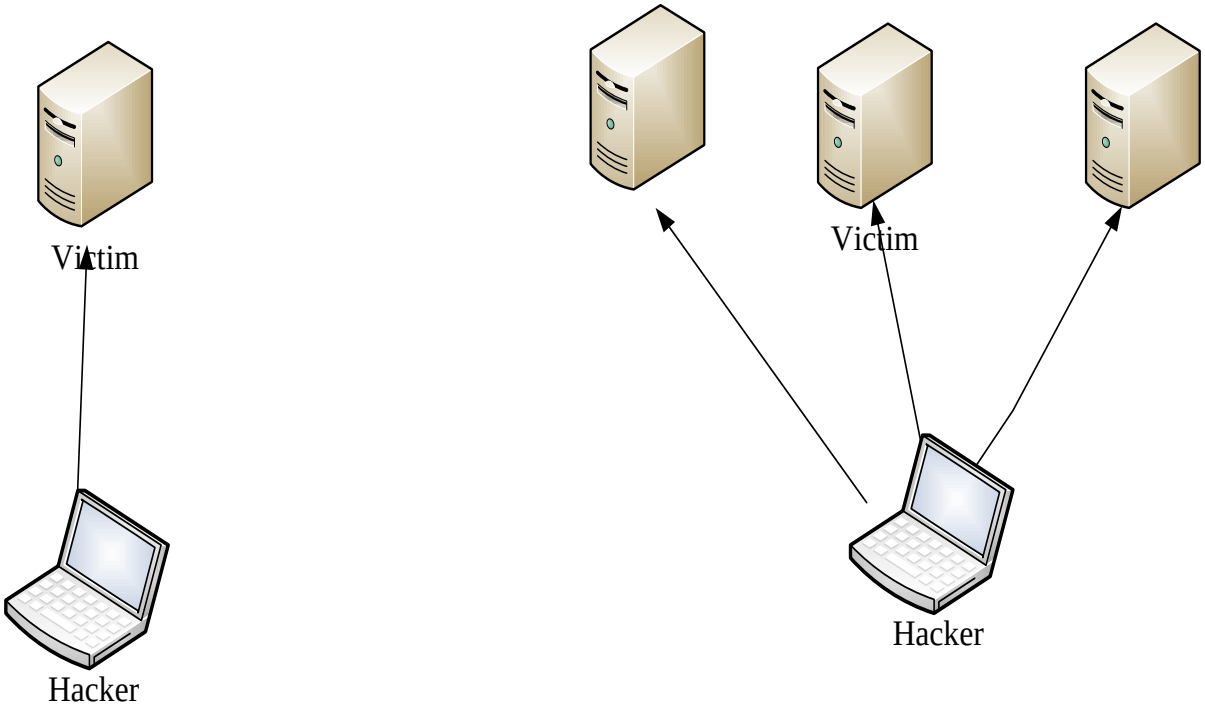
trong. Cấu trúc hoặc không cấu trúc có thể là mối đe dọa bên ngoài cũng như bên trong.

1.1.3.Các mô hình tấn công mạng

a)Mô hình tấn công truyền thống

Mô hình tấn công truyền thống được tạo dựng theo nguyên tắc "một đến một" hoặc "một đến nhiều", có nghĩa là cuộc tấn công xảy ra từ một nguồn gốc. Mô tả: Tấn công "một đến một"

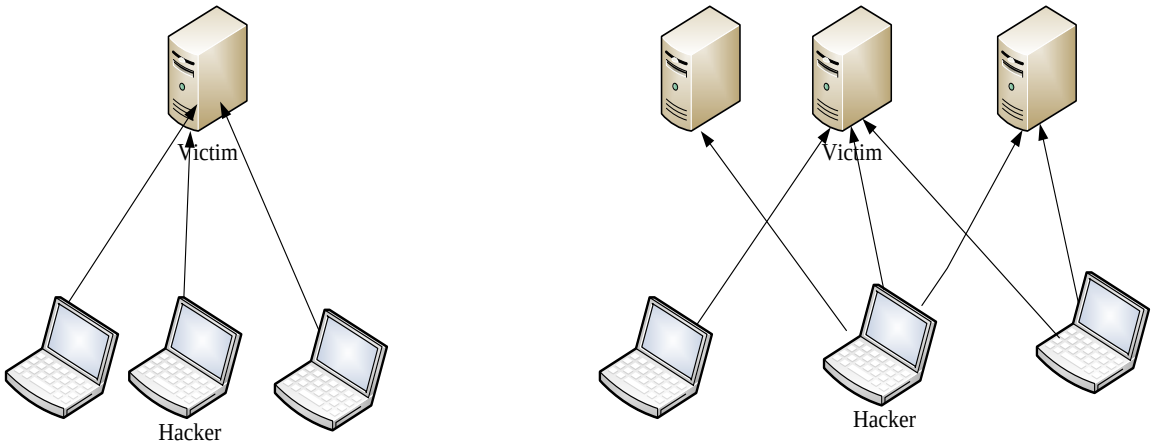
Hình 1.1: Mô hình tấn công truyền thống



b) Mô hình tấn công phân tán

Khác với mô hình truyền thống trong mô hình tấn công phân tán sử dụng quan hệ "nhiều đến một" và "nhiều đến nhiều". Tấn công phân tán dựa trên các cuộc tấn công "cổ điển" thuộc nhóm "từ chối dịch vụ", chính xác hơn là dựa trên các cuộc tấn công như Flood hay Storm (những thuật ngữ trên có thể hiểu tương đương như "bão", "lũ lụt" hay "thác tràn").

Hình 1.2:Mô hình tấn công phân tán



c)Các bước tấn công mạng

Hình 1.3: Các bước tấn công mạng

Các kiểu tấn công có nhiều hình thức khác nhau, nhưng thông thường đều thực hiện qua các bước theo hướng mô tả sau:

+ Xác định mục tiêu tấn công: Xác định rõ mục tiêu cần tấn công, nơi chuẩn bị tấn công.

+ Thu thập thông tin, tìm lỗ hổng: Khảo sát thu thập thông tin về nơi chuẩn bị tấn công bằng các công cụ để tìm hiểu đầy đủ về hệ thống cần tấn công. Sau khi đã thu thập đủ thông tin, kẻ tấn công sẽ dò tìm những thông tin về lỗ hổng của bảo mật hệ thống dựa trên những thông tin đã tìm được, phân tích điểm yếu của hệ thống mạng, sử dụng các bộ công cụ để dò quét tìm lỗi trên hệ thống mạng đó.

+ Lựa chọn mô hình tấn công: Khi đã có trong tay những điểm yếu của hệ thống mạng, kẻ tấn công sẽ lựa chọn công cụ để tấn công vào hệ thống như làm tràn bộ đệm hoặc tấn công từ chối dịch vụ…

+ Thực hiện tấn công: Sử dụng các công cụ để tấn công vào hệ thống. Sau khi đã tấn công thành công và khai thác được hệ thống rồi sẽ thực hiện việc duy trì với mục đích khai thác và tấn công trong tương lai gần. Chúng có thể sử dụng những thủ thuật như mở cửa sau (backdoor) hoặc cài đặt một trojan để nhằm mục đích duy trì sự xâm nhập của mình. Việc duy trì và làm chủ một hệ thống tạo cho kẻ tấn công có đủ những điều kiện để khai thác, phục vụ những nhu cầu về thông tin. Ngoài ra, hệ thống mạng này khi bị chiếm quyền điều khiển cũng sẽ trở thành nạn nhân của một hệ thống botnet được sử dụng trong các cuộc tấn công khác mà cụ thể là tấn công từ chối dịch vụ đến một hệ thống mạng khác.

+ Xóa dấu vết: Khi một kẻ tấn công đã tấn công thành công sẽ cố gắng duy trì sự xâm nhập này. Bước tiếp theo là chúng phải làm sao xóa hết dấu vết để không còn chứng cứ pháp lýtấn công. Kẻ tấn công phải xóa các tập tin log, xóa các cảnh báo từ hệ thống phát hiện xâm nhập. Ở các giai đoạn thu thập thông tin và dò tìm lỗ hổng trong bảo mật, kẻ tấn công thường làm lưu lượng kết nối mạng thay đổi khác với lúc mạng bình thường rất nhiều, đồng thời tài nguyên của hệ thống sẽ bị ảnh hưởng đáng kể.

Những dấu hiệu này rất có ích cho người quản trị mạng có thể phân tích và đánh giá tình hình hoạt động của hệ thống mạng. Hầu hết các cuộc tấn công đều tiến hành tuần tự như các bước đã nêu trên. Làm sao để nhận biết hệ thống mạng đang bị tấn công, xâm nhập ngay từ hai bước đầu tiên là hết sức quan trọng. Ở giai đoạn xâm nhập, bước này không dễ dàng đối với kẻ tấn công. Do vậy, khi không thể xâm nhập được vào hệ thống, để phá hoại có nhiều khả năng kẻ tấn công sẽ sử dụng tấn công từ chối dịch vụ để ngăn cản không cho người dùng hợp lệ truy xuất tài nguyên hệ thống.

1.2.Một số kỹ thuật tấn công mạng

1.2.1. Tấn công thăm dò

Thăm dò là việc thu thập thông tin trái phép về tài nguyên, các lỗ hổng hoặc dịch vụ của hệ thống.

Tấn công thăm dò thường bao gồm các hình thức:

- Sniffing
- Ping Sweep
- Ports Scanning

1.2.2. Tấn công xâm nhập

Tấn công xâm nhập là một thuật ngữ rộng miêu tả bất kỳ kiểu tấn công nào đòi hỏi người xâm nhập lấy được quyền truy cập trái phép của một hệ thống bảo mật với mục đích thao túng dữ liêu, nâng cao đặc quyền.

Tấn công truy nhập hệ thống: Là hành động nhằm đạt được quyền truy cập bất hợp pháp đến một hệ thống mà ở đó hacker không có tài khoản sử dụng.

Tấn công truy nhập thao túng dữ liệu: Kẻ xâm nhập đọc, viết, xóa, sao chép hay thay đổi dữ liệu.

1.2.3. Tấn công từ chối dịch vụ

Về cơ bản, tấn công từ chối dịch vụ là tên gọi chung của cách tấn công làm cho một hệ thống nào đó bị quá tải không thể cung cấp dịch vụ, làm gián đoạn hoạt động của hệ thống hoặc hệ thống phải ngưng hoạt động. Tùy theo phương thức thực hiện mà nó được biết dưới nhiều tên gọi khác nhau. Mục đích là lợi dụng sự yếu kém của giao thức TCP (Transmision Control Protocol) để thực hiện tấn công từ chối dịch vụ DoS (Denial of Service), mới hơn là tấn công từ chối dịch vụ phân tán DdoS (Distributed DoS), mới nhất là tấn công từ chối dịch vụ theo phương pháp phản xạ DRDoS (Distributed ReflectionDoS).

1.2.4. Tấn công từ chối dịch vụ cổ điển

a) DoS (Denial of Service): gồm

- Bom thư
- Đăng nhập liên tiếp
- Làm ngập SYN (Flooding SYN)
- Tấn công Smurf
- Tấn công gây lụt UDP
- Tấn công ping of death
- Tấn công tear drop

b) SYN Attack

Được xem là một trong những kiểu tấn công DoS kinhđiển nhất. Lợi dụng sơ hở của thủ tục TCP khi "bắt tay ba bước", mỗi khi client muốn thực hiện kết nối với server thì nó thực hiện việc bắt tay ba bước thông qua các gói tin (packet).

Bước 1: client sẽ gửi gói tin (packet chứa SYN=1) đến máy chủ để yêu cầu kết nối.

Bước 2: khi nhận được gói tin này, server gửi lại gói tin SYN/ACK để thông báo cho client biết là nó đã nhận được yêu cầu kết nối và chuẩn bị tài nguyên cho việc yêu cầu này. Server sẽ dành một phần tài nguyên để nhận và truyền dữ liệu. Ngoài ra, các thông tin khác của client như địa chỉ IP và cổng (port) cũng được ghi nhận.

Bước 3: cuối cùng client hoàn tất việc bắt tay ba bước bằng cách hồi âm lại gói tin chứa ACK cho server và tiến hành kết nối.

Do TCP là thủ tục tin cậy trong việc giao nhận nên trong lần bắt tay thứ hai, server gửi gói tin SYN/ACK trả lời lại client mà không nhận lại được hồi âm của client để thực hiện kết nối thì nó vẫn bảo lưu nguồn tài nguyên chuẩn bị kết nối đó và lặp lại việc gửi gói tin SYN/ACK cho client đến khi nhận được hồi đáp của client. Điểm mấu chốt ở đây là làm cho client không hồi đáp cho Server, và có càng nhiều, càng nhiều client như thế trong khi server vẫn lặp lại việc gửi packet đó và giành tài nguyên để chờ trong lúc tài nguyên của hệ thống là có giới hạn. Các hacker tấn công sẽ tìm cách để đạt đến giới hạn đó.

1.2.5. Tấn công dịch vụ phân tán DdoS

Xuất hiện vào năm 1999, so với tấn công DoS cổ điển, sức mạnh của DDoS cao hơn gấp nhiều lần. Hầu hết các cuộc tấn công DDoS nhằm vào

việc chiếm dụng băng thông (bandwidth) gây nghẽn mạch hệ thống dẫn đến hệ thống ngưng hoạt động. Để thực hiện thì kẻ tấn công tìm cách chiếm dụng và điều khiển nhiều máy tính/mạng máy tính trung gian (đóng vai trò zombie) từ nhiều nơi để đồng loạt gửi ào ạt các gói tin với số lượng rất lớn nhằm chiếm dụng tài nguyên và làm tràn ngập đường truyền của một mục tiêu xác định nào đó.

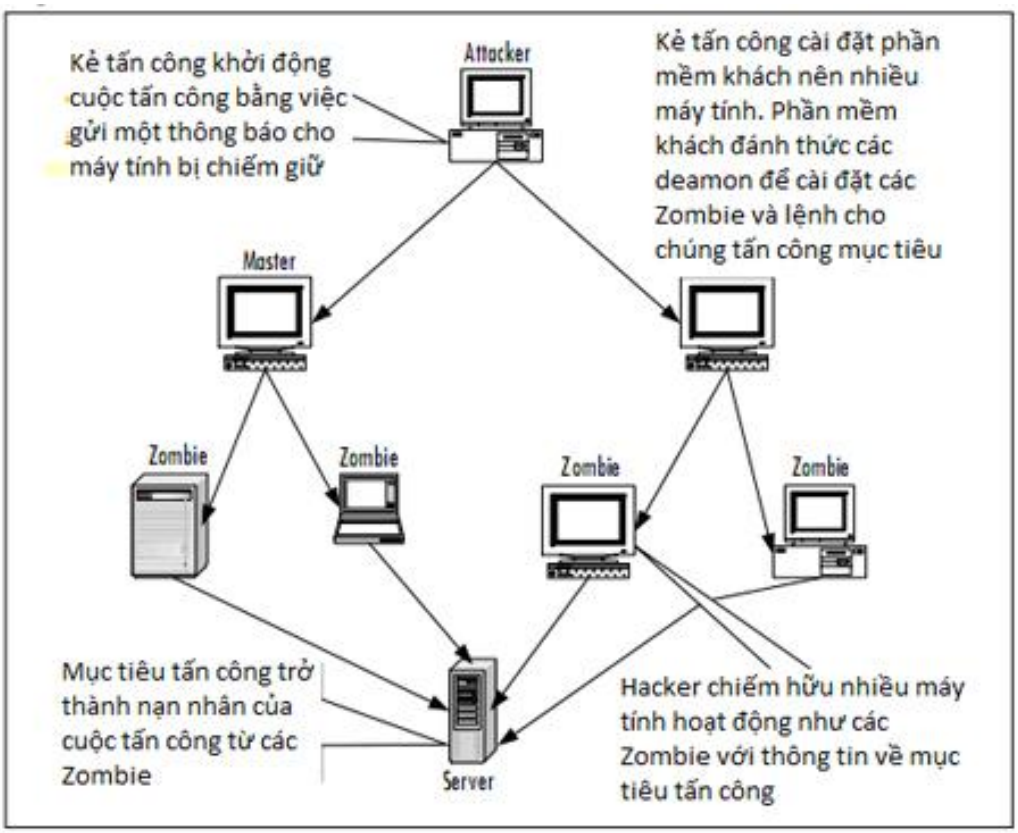
Các giai đoạn của một cuộc tấn công kiểu DDoS:

- Giai đoạn chuẩn bị.
- Giai đoạn xác định mục tiêu và thời điểm.
- Phát động tấn công.
- Xoá dấu vết.

Mặc dù một số dạng tấn công DoS có thể được khuếch đại bởi nhiều trung gian, nhưng xuất phát của DoS vẫn là bắt nguồn từ một máy tính đơn lẻ. Tuy nhiên DDoS đã được phát triển xa hơn ngoài cuộc tấn công một tầng (lũ SYN) và hai tầng (Smurf). Tấn công DDoS ra đời là bước tiếp theo của DoS, khắc phục được nhiều thiếu xót mà DoS chưa đáp ứng được. Đây là phương pháp tấn công hiện đại có sự kết hợp của nhiều tầng tính toán phân tán. Khác biệt đáng chú ý trong phương pháp tấn công này là nó bao gồm hai giai đoạn khác nhau. Giai đoạn đầu tiên, thủ phạm bố trí các máy tính phân tán trên Internet và cài đặt các phần mềm chuyên dụng trên các máy chủ để hỗ trợ tấn công. Giai đoạn thứ hai, máy tính bị xâm nhập (được gọi là Zombie) sẽ cung cấp thông tin qua kẻ trung gian (được gọi là Master) để bắt đầu cuộc tấn công. Hàng trăm, có thể hàng ngàn, các zombie có thể được chọn đồng thời tham gia vào các cuộc tấn công của Hacker. Với việc sử dụng phần mềm điều khiển, các Zombie này sẽ thực thi cuộc tấn công DDoS hướng vào mục tiêu. Hiệu quả cộng dồn của tấn công Zombie làm hủy hoại nạn nhân với sự ồ ạt

của một lượng lớn tin truyền tải làm tắc nghẽn thông tin hoặc làm cạn kiệt nguồn tài nguyên. Ngoài ra, với kiểu tấn công như vậy sẽ giấu đi thông tin của kẻ tấn công thực sự: chính là kẻ đưa ra các lệnh điều khiển cho các Zombie.Mô hình đa cấp của các cuộc tấn công DDoS cộng với khả năng giả mạo của các gói tin và mã hóa thông tin đã gây nhiều khó khăn cho quá trình tìm kiếm, phát hiện kẻ tấn công thực sự. Các cấu trúc lệnh hỗ trợ một cuộc tấn công DDoS khá phức tạp (xem hình 1.4) và khó có thể đưa ra một thuật ngữ để mô tả chính xác. Dưới đây là những cái tên quy ước cho dễ hiểu hơn của cấu trúc và các thành phần tham gia trong cuộc tấn công DDoS:

Hình 1.4: Tổng quan về một sơ đồ hình cây của tấn công DDoS.



- **Client**: Phần mềm khách được Hacker sử dụng để bắt đầu cuộc tấn công. Phần mềm khách sẽ gửi các chuỗi lệnh đến các máy chủ dưới quyền.

- **Daemon:** Các chương trình đang chạy trên một Zombie sẽ nhận chuỗi lệnh đến từ Client và thực thi các lệnh đó.Daemon sẽ chịu trách nhiệm thực thi chi tiết cuộc tấn công từ các dòng lệnh. Các máy chủ tham gia vào cuộc tấn công DDoS bao gồm:
  - **Master:** Một máy tính chạy các phần mềm khách.
  - **Zombie:** Một máy tính cấp dưới chạy quá trình Daemon.
  - **Target:** Mục tiêu của cuộc tấn công.

Trên sơ đồ hình 1.4 nhận thấy rằng, để bắt đầu cho cuộc tấn công, tin tặc sẽ tìm kiếm mục tiêu trên Internet là những máy tính lỏng lẻo trong bảo mật. Tin tặc sử dụng cả hai kỹ thuật kiểm tra tự động và bằng tay để lần ra lỗ hổng của hệ thống mạng và máy chủ .Tin tặc sử dụng các tập lệnh để rà quét tự động các máy không an toàn, từ đó có thể được phát hiện được một cách chính xác cơ sở hạ tầng bảo mật của mạng máy tính nội bộ. Tùy thuộc vào trình độ của Hacker mà có thể gây ra những khó khăn cho việc tìm kiếm và xác định danh tính của kẻ tấn công bởi kể tấn công sẽ tím cách thích ứng với cách tiếp cận của mình, việc chiếm quyền điều khiển một máy tính cúng tốn khá nhiều thời gian. Sau khi các máy tính bảo mật kém đã được xác định, kẻ tấn công tìm cách xâm nhập hệ thống. Hacker có thể truy cập được vào máy chủ bằng nhiều cách (thường thông qua các tài khoản máy chủ hoặc tài khoản quản trị), hầu hết các phương pháp xâm nhập này đều có thể phòng ngừa được.Nhiệm vụ đầu tiên là Hacker sẽ phải đảm bảo việc xóa sạch bằng chứng cho thấy hệ thống đã bị xâm nhập và cũng đảm bảo rằng các máy chủ bị xâm nhập sẽ thông qua được các công cụ kiểm tra. Công cụ sử dụng để đảm bảo các nhiệm vụ này sẽ thành công được gọi chung là các rootkits. Những máy chủ đã bị chiếm trở thành Master, còn các máy khác sẽ đóng vai trò Zombie. Master được cài đặt với một bản sao của phần mềm khách và được sử dụng

làm trung gian giữa những kẻ tấn công và các Zombie. Các Master nhận thông tin rồi chuyển qua các Zombie mà chúng phụ trách. Băng thông mạng cho các máy master không phải là một tham số ưu tiên hàng đầu, bởi các máy master chỉ chịu trách nhiệm gửi và nhận các đoạn tin điều khiển ngắn nên có thể tiến hành trong cả mạng có băng thông thấp . Trên máy tính không được chỉ định làm Master, Hacker cài đặt các phần mềm Daemon để gửi ra các luồng tấn công, và các máy tính này được gọi là Zombie. Chương trình daemon chạy trên nền của zombie, đợi một thông báo để kích hoạt và phát động cuộc tấn công nhắm vào nạn nhân đã được chỉ định. Một chương trình daemon có thể khởi động nhiều loại tấn công, chẳng hạn như UDP hoặc lũ SYN. Kết hợp với khả năng sử dụng sự giả mạo, các daemon có thể chứng minh là một công cụ tấn công rất linh hoạt và mạnh mẽ. Sau khi kẻ tấn công đã chuẩn bị được đầy đủ những gì cần thiết số Zombie, cũng như đã xác định được nạn nhân của mình, kẻ tấn công có thể liên hệ với các master (hoặc thông qua các phương pháp riêng hoặc với một chương trình đặc biệt giành riêng cho DDoS) và chỉ thị chúng để khởi động cuộc tấn công. Các Zombie sẽ bắt đầu tấn công sau khi nhận lệnh từ master. Chỉ mất vài giây để khởi động và phân tán rộng cuộc tấn công, với tốc độ như vậy, thì hacker có thể ngưng cuộc tấn công. Việc sử dụng và phát triển các phương pháp ttân công DoS và DDoS đã tạo được sự quan tâm của chính phủ, các doanh nghiệp, và các chuyên gia bảo mật, do nó đưa ra một phương thức tấn công mới cực kỳ hiệu quả , trong khi rất khó trong việc tìm kiếm thông tin kẻ tấn công thực sự.

1.3.Hệ thống phát hiện xâm nhập trái phép

1.3.1. Khái niệm về hệ thống phát hiện xâm nhập trái phép

Phát hiện xâm nhập là tiến trình theo dõi các sự kiện xảy ra trên một hệ thống máy tính hay hệ thống mạng, phân tích chúng để tìm ra các dấu hiệu

“xâm nhập bất hợp pháp”. Xâm nhập bất hợp pháp được định nghĩa là sự cố gắng tìm mọi cách để xâm hại đến tính toàn vẹn, tính sẵn sàng, tính có thể tin cậy hay là sự cố gắng vượt qua các cơ chế bảo mật của hệ thống máy tính hay mạng đó. Việc xâm nhập có thể là xuất phát từ một kẻ tấn công nào đó trên mạng Internet nhằm giành quyền truy cập hệ thống, hay cũng có thể là một người dùng được phép trong hệ thống đó muốn chiếm đoạt các quyền khác mà họ chưa được cấp phát [2]. Như đã đề cập ở trên, hệ thống phát hiện xâm nhập là hệ thống phần mềm hoặc phần cứng có khả năng tự động theo dõi và phân tích để phát hiện ra các dấu hiệu xâm nhập.

- ✓ Network IDS hoặc NIDS

Là các hệ thống phát hiện tấn công, nó có thể bắt giữ các gói tin được truyền trên các thiết bị mạng (cả hữu tuyến và vô tuyến) và so sánh chúng với cơ sở dữ liệu các tín hiệu.

- ✓ Host IDS hoặc HIDS

Được cài đặt như là một tác nhân trên máy chủ. Những hệ thống phát hiện xâm nhập này có thể xem những tệp tin log của các trình ứng dụng hoặc của hệ thống để phát hiện những hành động xâm nhập.

- ✓ Signature

Là những phần mà ta có thể thấy được trong một gói dữ liệu. Nó được sử dụng để phát hiện ra một hoặc nhiều kiểu tấn công. Signature có thể có mặt trong các phần khác nhau của một gói dữ liệu. Ví dụ ta có thể tìm thấy các tín hiệu trong header IP, header của tầng giao vận (TCP, UDP header) hoặc header tầng ứng dụng. Thông thường, IDS ra quyết định dựa trên những tín hiệu tìm thấy ở hành động xâm nhập. Các nhà cung cấp IDS cũng thường xuyên cập nhật những tín hiệu tấn công mới khi chúng bị phát hiện ra.

- ✓ Alert

Là những lời thông báo ngắn về những hành động xâm nhập bất hợp pháp. Khi IDS phát hiện ra kẻ xâm nhập, nó sẽ thông báo cho người quản trị bảo mật bằng alert. Alert có thể hiện ngay trên màn hình, khi đăng nhập hoặc bằng mail và bằng nhiều cách khác. Alert cũng có thể được lưu vào file hoặc vào cơ sở dữ liệu để các chuyên gia bảo mật có thể xem lại.

- ✓ Log
    - Thông thường, những thông tin mà IDS thu được sẽ lưu lại trong file. Chúng có thể được lưu lại dưới dạng text hoặc dạng nhị phân. Tốc độ lưu lại thông tin ở dạng nhị phân sẽ nhanh hơn ở dạng text.
- ✓ False Alarm

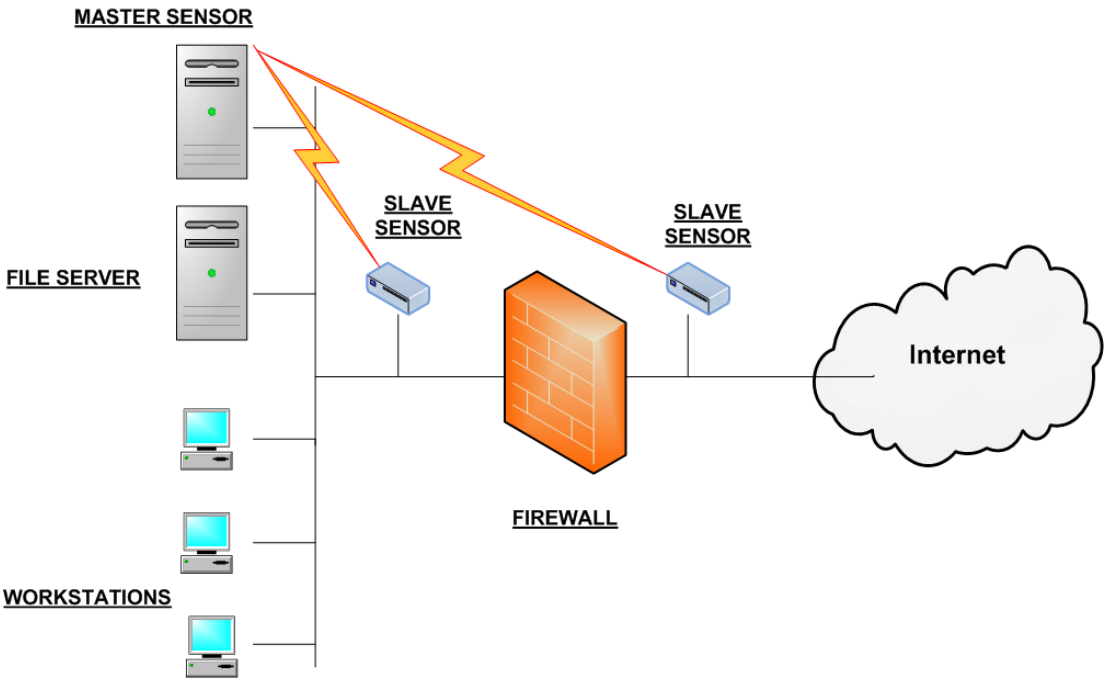
Là những thông báo đúng về một dấu giống dấu hiệu xâm nhập nhưng hành động.

- ✓ Sensor

Là những thiết bị mà hệ thống phát hiện xâm nhập chạy trên nó bởi vì nó được sử dụng như các giác quan trên mạng. Cũng tương tự như các sensor trong các tài liệu kỹ thuật khác, sensor dùng để bắt tín hiệu âm thanh, màu sắc, áp xuất... thì sensor ở đây sẽ bắt các tín hiệu có dấu hiệu của xâm nhập bất hợp pháp.

Vị trí của sensor phụ thuộc vào mô hình của hệ thống mạng. Ta có thể đặt ở một hoặc nhiều nơi, nó phụ thuộc vào loại hoạt động mà ta muốn giám sát (internal, external hoặc cả 2). Ví dụ, nếu ta muốn giám sát hành động xâm nhập từ bên ngoài và ta chỉ có một router kết nối với internet thì nơi thích hợp nhất là đặt phía sau thiết bị router (hay firewall). Nếu ta có nhiều đường kết nối với Interrnet thì ta có thể đặt sensor tại mỗi điểm kết nối với Internet. Ta có thể hình dung qua hình vẽ sau:

Hình 1.5: Đặt một sensor phía sau hệ thống Firewall



1.3.2. Các kỹ thuật phát hiện xâm nhập trái phép

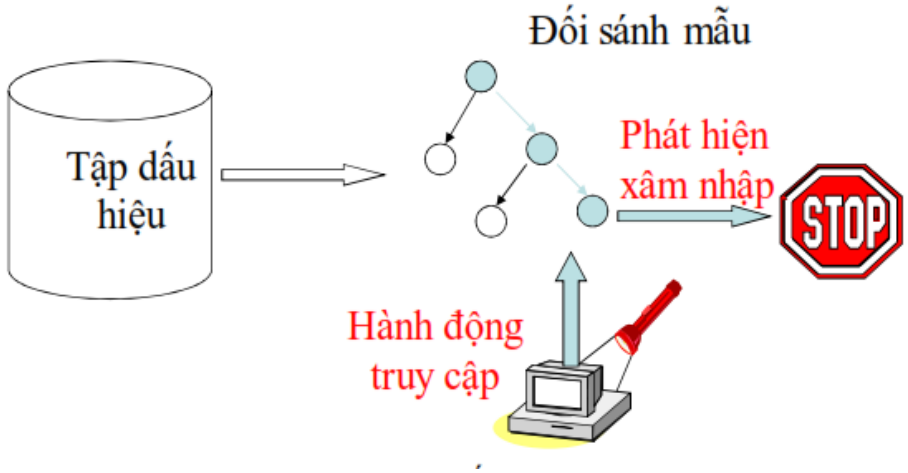
IDS sử dụng nhiều kỹ thuật khác nhau để phát hiện các hành động xâm nhập hệ thống trái phép. Những kỹ thuật cơ bản như: Dựa trên dấu hiệu, sự kiện bất thường và dựa trên mô hình. Thông thường IDS sử dụng nhiều phương pháp phát hiện xâm nhập và đôi khi cũng sử dụng phương pháp riêng lẻ hay kết hợp nhằm phát hiện chính xác các hành động xâm nhập.

a) Phát hiện dựa vào dấu hiệu

Dấu hiệu là một mẫu tương ứng với các đe dọa đã biết được thống kê các đặc trưng và lưu lại trên hệ thống. Hệ thống sẽ thu thập các thông tin liên quan và so sánh với các dấu hiệu tấn công được lưu trữ trong cơ sở dữ liệu để xác định xem hành động đó có nguy hiểm hay không. Ví dụ sau đây mô tả cách IDS phát hiện xâm nhập dựa vào dấu hiệu:

Thư điện tử có tiêu đề “Free pictures!” đính kèm file “freepics.exe” và file này có đặc điểm của mã độc hại đã biết.

Hình 1.6: Môtảdấuhiệuxâmnhập



Kỹ thuật này rất hiệu quả trong việc phát hiện các đe dọa đã biết nhưng lại không hiệu quả trong việc phát hiện những nguy cơ chưa được biết. Ví dụ kẻ tấn công sửa tên file thành “freepic2.exe”, thì việc tìm kiếm dấu hiệu trên với mã độc hại này sẽ không hiệu quả.

Phát hiện dựa vào dấu hiệu là một kỹ thuật đơn giản vì nó chỉ so sánh hành động hiện tại với danh sách dấu hiệu đã biết bằng cách so sánh các toán tử. Kỹ thuật này ít được dùng trong mô hình mạng lớn hay các giao thức ứng dụng bởi vì nó không thể theo dõi và hiểu được trạng thái của tất cả các thành phần phức tạp trong hệ thống. Bên cạnh đó kỹ thuật này không có khả năng ghi nhớ những yêu cầu trước đó khi có một yêu cầu hiện tại. Do đó việc phát hiện tấn công dựa trên phương pháp này có độ tin cậy không cao.

b) Phát hiện dựa trên sự bất thường

Phát hiện dựa vào sự bất thường là quá trình so sánh hành động được coi là bình thường với các sự kiện đang diễn ra nhằm phát hiện ra sự bất thường. Với kỹ thuật này IDS dựa vào profile miêu tả hành động bình thường của nhiều đối tượng như người dùng, máy chủ, các kết nối mạng, hay ứng dụng. Profile này được tạo ra bằng cách giám sát các hành động thông thường trong một khoảng thời gian để đưa ra đặc điểm nổi bật của hành động

đó.

Kỹ thuật này chỉ có độ chính xác cao khi IDS được gắn vào một hệ thống mạng cụ thể và có thời gian đủ lâu để học tất cả các hành động bình thường của hệ thống.

c) Kỹ thuật phát hiện dựa vào phân tích trạng thái giao thức

Phân tích trạng thái giao thức là quá trình phân tích hành vi của giao thức được sử dụng trên cơ sở đã biết các định nghĩa về hoạt động hợp lệ của giao thức để nhận ra hành vi tấn công. Kỹ thuật này dựa vào profile liên quan đến giao thức mà IDS hỗ trợ. "Trạng thái" trong phân tích trạng thái giao thức nghĩa là IDS có khả năng hiểu và theo dõi trạng thái của mạng, truyền tải và các giao thức ứng dụng.

Điều ngăn cản chính của phương pháp này chính là việc tập trung tài nguyên, bởi vì sự phức tạp trong quá trình phân tích và thực hiện giám sát trạng thái cho nhiều phiên làm việc đồng thời. Một vấn đề khác là phương pháp này không thể phát hiện được các tấn công có đặc trưng mà hành vi thông thường của giao thức được thừa nhận, như việc thực hiện nhiều hành động trong một khoảng thời gian ngắn như tấn công từ chối dịch vụ. Hơn nữa, chuẩn giao thức được sử dụng trong IDS có thể xung đột với cách thực hiện của giao thức hiện có trong mạng.

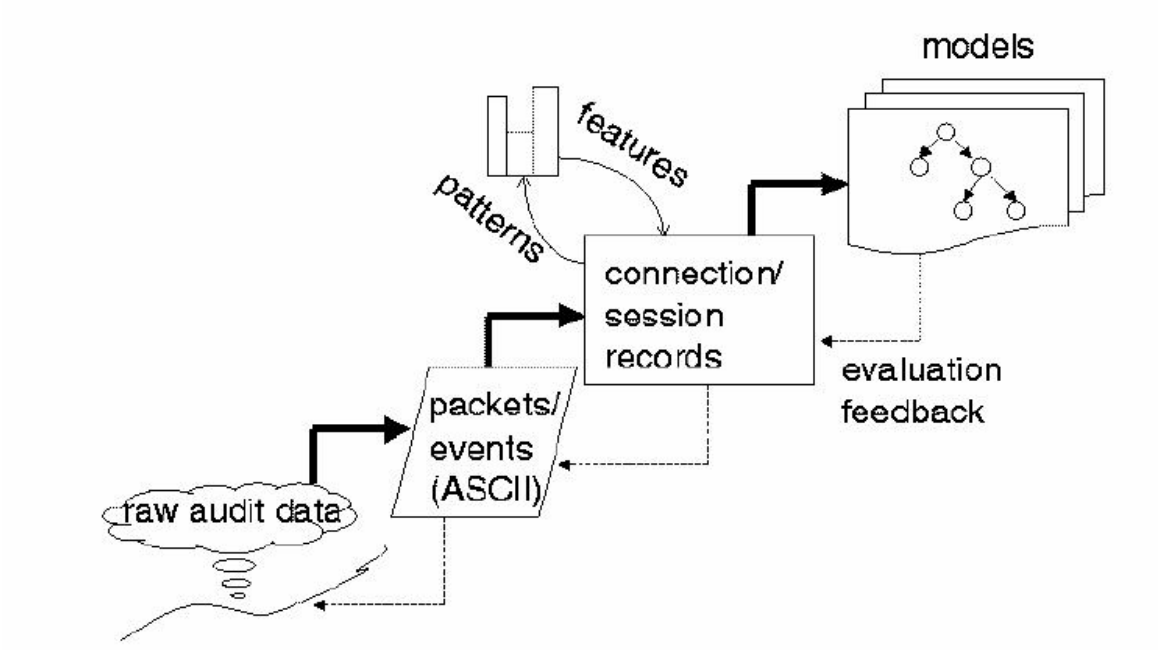
d) Phát hiện dựa trên mô hình

Phương pháp phát hiện dựa trên mô hình sử dụng các kỹ thuật học máy, khai phá dữ liệu, trí tuệ nhân tạo để xây dựng các mô hình, các luật phát hiện tấn công một cách tự động từ các tập dữ liệu mô phỏng tấn công. Sau đó các mô hình được sử dụng trong các hệ thống IDS để dự đoán các tấn công mới. Phương pháp này có ưu điểm là cho phép phát hiện được các tấn công mới, tuy nhiên hạn chế của nó là đưa ra nhiều cảnh báo nhầm hơn các phương pháp trên.

1.3.3. Ứng dụng kỹ thuật khai phá dữ liệu cho việc phát hiện xâm nhập trái phép

Khai phá dữ liệu là một phướng pháp tiếp cận tương đối mới trong việc phát hiện xâm nhập. Khai phá dữ liệu được định nghĩa cụ thể theo "Sự khám phá ra các mẫu, các mối quan hệ, các biến đổi, những sự bất thường, những qui luật, những cấu trúc và sự kiện quan trọng mang tính chất thống kê trong dữ liệu". Trong đó tồn tại nhiều kiểu thuật toán khai phá dữ liệu khác nhau như: phân lớp, phân tích hồi quy, phân cụm, khai phá luật kết hợp [3]….. Công việc khai phá dữ liệu trong phát hiện xâm nhập trái phép là để trích lọc tri thức từ một tập dữ liệu lớn của các thông tin truy cập trên mạng, để phân tích biểu diễn nó thành mô hình phát hiện xâm nhập trái phép. Phương pháp tiếp cận này xét về việc phát hiện xâm nhập như là tiến trình phân tích dữ liệu, trong khi đó các phương pháp tiếp cận trước là những quá trình kỹ nghệ tri thức.

Hình 1.7: Quá trình khai phá dữ liệu của việc xây dựng mô hình PHXN



Phương pháp khai phá dữ liệu để phát hiện xâm nhập lần đầu tiên được

thực hiện bởi MADAMID (Mining Audit Data for Automated Models for Instruction Detection: Khai phá dữ liệu được sử dụng trong mô hình tự động để phát hiện xâm nhập) [4]. Quá trình khai phá dữ liệu trong việc xây dựng những mô hình phát hiện xâm nhập: Dữ liệu thô đầu tiên được chuyển đổi thành thông tin gói dữ liệu mạng với mã ASCII mà lần lượt nó được chuyển đổi thành thông tin ở mức truy cập; Những bản ghi ở mức truy cập này chứa trong đó những thuộc tính kết nối như là dịch vụ, thời gian kết nối… Thuật toán khai phá dữ liệu được áp dụng cho những dữ liệu này để tạo ra các mô hình phát hiện xâm nhập.

CHƯƠNG II: MỘT SỐ KỸ THUẬT PHÂN CỤM DỮ LIỆU

Kỹ thuật phân cụm dữ liệu có rất nhiều cách tiếp cận và các ứng dụng trong thực tế, nó đều hướng tới hai mục tiêu chung đó là chất lượng của các cụm khám phá được và tốc độ thực hiện của thuật toán. Hiện nay, các kỹ thuậtphân cụm có thể phân loại theo các phương pháp tiếp cận chính như sau: phân cụm phân hoạch (Partitioning Methods); phân cụm phân cấp (Hierarchical Methods); phân cụm dựa trên mật độ (Density-Based Methods); phân cụm dựa trên lưới (Grid-Based Methods); phân cụm dựa trên mô hình(Model-Based Clustering Methods) và phân cụm có dữ liệu ràng buộc (Binding data Clustering Methods) [5][6].

2.1.Phân cụm phân hoạch

Phân cụm phân hoạch (Partitioning Methods): Kỹ thuật này phân hoạch một tập hợp dữ liệu có n phần tử cho trước thành k nhóm dữ liệu cho đến khi xác định số các cụm được thiết lập. Số các cụm được thiết lập là các đặc trưng được lựa chọn trước. Phương pháp này là tốt cho việc tìm các cụm hình cầu trong không gian Euclidean. Ngoài ra, phương pháp này cũng phụ thuộc vào khoảng cách cơ bản giữa các điểm để lựa chọn các điểm dữ liệu nào có quan hệ là gần nhau với mỗi điểm khác và các điểm dữ liệu nào không có quan hệ hoặc có quan hệ là xa nhau so với mỗi điểm khác. Tuy nhiên, phương pháp này không thể xử lí các cụm có hình dạng kỳ quặc hoặc các cụm có mật độ các điểm dầy đặc. Các thuật toán phân hoạch dữ liệu có độ phức tạp rất lớn khi xác định nghiệm tối ưu toàn cục cho vấn đề phân cụm dữ liệu, do nó phải tìm kiếm tất cả các cách phân hoạch có thể được. Chính vì vậy, trên thực tế thường đi tìm giải pháp tối ưu cục bộ cho vấn đề này bằng cách sử dụng một hàm tiêu chuẩn để đánh giá chất lượng của cụm cũng như để hướng dẫn cho

quá trình tìm kiếm phân hoạch dữ liệu. Như vậy, ý tưởng chính của thuật toán phân cụm phân hoạch tối ưu cục bộ là sử dụng chiến lược ăn tham (Greedy) để tìm kiếm nghiệm.

Điển hình trong phương pháp tiếp cận theo phân cụm phân hoạch là các thuật toán như: K-means, K-medoids, CLARA (Clustering Large Applications), CLARANS (Clustering Large Applications based on RAndomized Search) . . .

2.1.1. Thuật toán K-means

ThuậttoánphânhoạchK-means [7]doMacQueenđềxuấttronglĩnhvựcthốngkênăm1967.Thuậttoándựatrênđộđo khoảngcáchcủacácđốitượngdữliệutrongcụm.Trongthựctế,nóđokhoảngcáchtới giátrịtrungbìnhcủacácdữliệutrongcụm.Nóđượcxemnhưlàtrungtâmcủacụm.Nh ưvậy,nócầnkhởi tạomộttậptrungtâmcáctrungtâm cụmbanđầuvàthôngquađónólặplạicácbướcgồmgánmỗiđốitượngtớicụmmàtrun gtâmgầnvàtínhtoánlạitrungtâmcủamỗicụmtrêncơsởgánmớichocácđốitượng.Q uátrìnhlặpnàydừngkhicác trungtâmcụmhộitụ.

Mục đíchcủathuậttoánK-meanslà sinh*k*cụmdữliệu{C_1,C_2,…, C_k}từmộttậpdữliệuchứa*n*đốitượngtrongkhônggian*d*chiềuX_i={x_{i1},x_{i2},…, x_{id}}, i=1÷n, saochohàmtiêuchuẩn:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} D^2 \left(x - m_i \right)$$

đạtgiátrịtốithiểu,trongđó$m_i$làtrọngtâmcủacụmC_i,Dlàkhoảngcáchgiữa haiđốitượng.

Thuật toán K-means bao gồm các bước sau:

Input: Số cụm k và các trọng tâm cụm $\{m_j\}_{j=1}^{k}$

Output: các cụm C[i]($1 \le i \le k$) và hàm tiêu chuẩn E đạt giá trị tối thiểu.

Begin

Bước 1 : Khởi tạo

Chọn k trọng tâm $\{m_j\}_{j=1}^{k}$ban đầu trong không gian R^d (d là số chiều của dữ liệu). Việc lựa chọn này có thể là ngẫu nhiên hoặc theo kinh nghiệm.

Bước 2 : Tính toán khoảng cách $D_{j=1}^{k} = \sqrt{\sum_{i=0}^{n} (x_i - m_j)^2}$

Đối với mỗi điểm x_i ($1 \le i \le k$), tính toán khoảng cách của nó tới mỗi trọng tâm m_j ($1 \le j \le k$). Sau đó tìm trọng tâm gần nhất đối với điểm.

Bước 3: Cập nhật lại trọng tâm

Đối với mỗi ($1 \le j \le k$), cập nhật trọng tâm cụm m_j bằng cách xác định trung bình cộng các véctơ đối tượng dữ liệu.

Bước 4: Gán lại các điểm gần trung tâm với nhóm mới

Nhóm các đối tượng vào nhóm gần nhất dựa trên trọng tâm của nhóm.

Điều kiện dừng:

Lặp lại các bước 2 và 3 cho đến khi các trọng tâm của cụm không thay đổi.

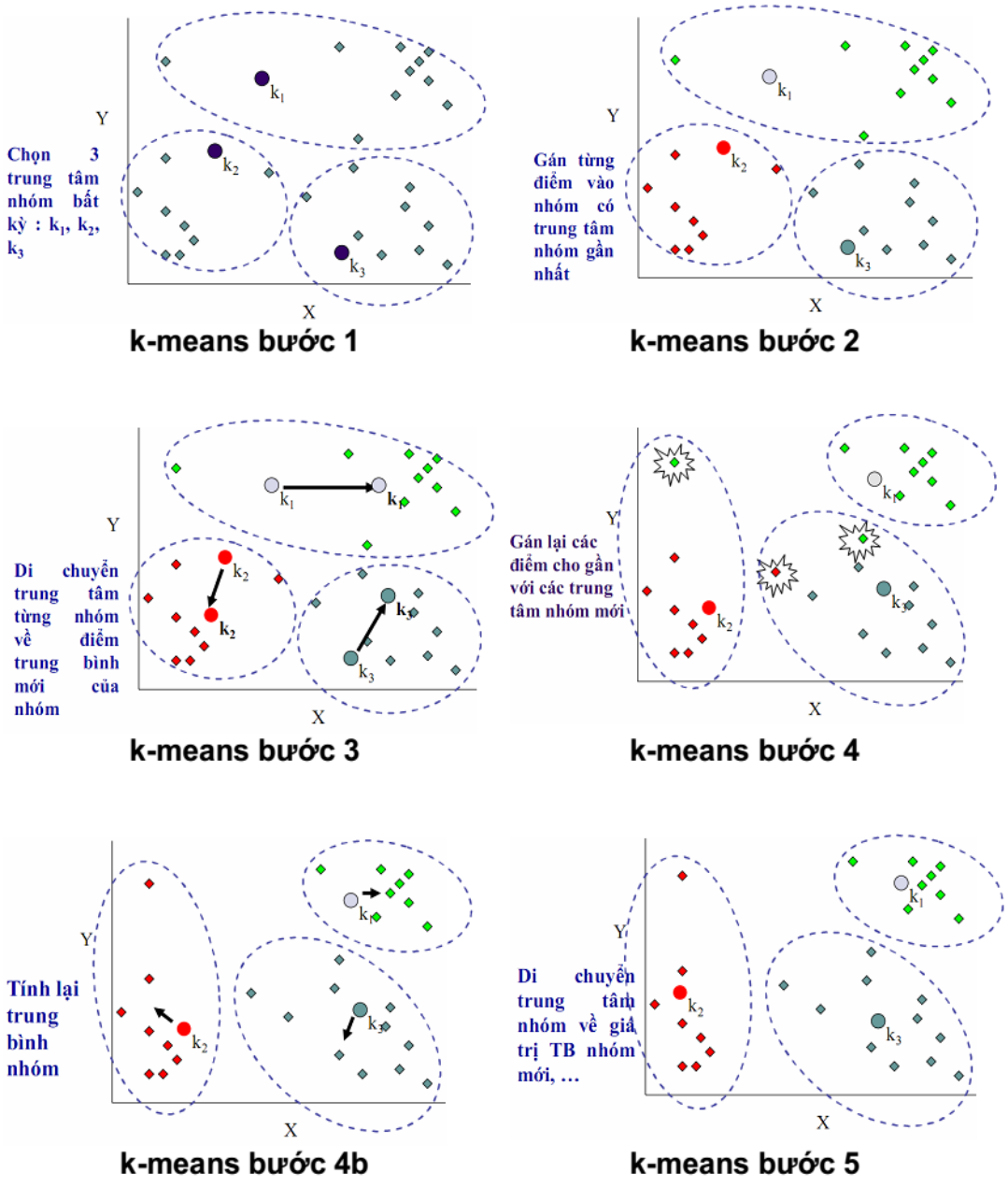
End.

Nhậnxét:

Độphứctạpcủathuật toán theo thời gian là *O(nkl)*với n là số đốitượng dữ liệuđưa vào, k là số cụm dữ liệu, l là sốvòng lặp …

DoK-meansphân tích cụm đơn giảnnên có thể áp dụng đối vớitập dữ liệu lớn. Tuynhiên,nhượcđiểm của K-meanslàchỉ áp dụngvới dữ liệu có thuộctínhsố và khám phá ra các cụm có dạnghình cầu, K-meanscòn rất nhạy cảm với nhiễu và các phần tử ngoạilai trong dữliệu.

Hình 2.1Ví dụ các bước của thuật toán k-means



Ví dụ: Giả sử có một tập đối tượng được định vị trong hệ trục toạ độX, Y. Cho k =3 tức người dùng cần phân các đối tượng vào trong 3 cụm. Theo giải thuật, ta chọn ngẫu nhiên 3 trung tâm cụm ban đầu (Hình kmeans bước 1). Sau đó, mỗi đối tượng được phân vào trong các cụm đã chọn dựa trên tâm cụm gần nhất (Hình k-means bước 2).

Cập nhật lại các tâm (Hình k-means bước 3). Đó là giá trị trung bình của mỗi cụm được tính toán lại dựa trên các đối tượng trong cụm. Tuỳ theo các tâm mới này, các đối tượng được phân bố lại vào trong các cụm dựa trên tâm cụm gần nhất (Hình k-means bước 4).

2.1.2. Thuật toán CLARA

CLARA (Clustering LARge Application) được Kaufman và Rousseeuw đề xuất năm 1990, thuật toán này nhằm khắc phục nhược điểm của thuật toán PAM trong trường hợp giá trị của k và n lớn. CLARA tiến hành trích mẫu cho tập dữ liệu có n phần tử và áp dụng thuật toán PAM cho mẫu này và tìm ra các các đối tượng medoid của mẫu này. Người ta thấy rằng, nếu mẫu dữ liệu được trích một cách ngẫu nhiên, thì các medoid của nó xấp xỉ với các medoid của toàn bộ tập dữ liệu ban đầu. Để tiến tới một xấp xỉ tốt hơn, CLARA đưa ra nhiều cách lấy mẫu rồi thực hiện phân cụm cho mỗi trường hợp này và tiến hành chọn kết quả phân cụm tốt nhất khi thực hiện phân cụm trên các mẫu này. Để cho chính xác, chất lượng của các cụm được đánh giá thông độ phi tương tự trung bình của toàn bộ các đối tượng dữ liệu trong tập đối tượng ban đầu. Kết quả thực nghiệm chỉ ra rằng, 5 mẫu dữ liệu có kích thước 40+2k cho các kết quả tốt. Các bước thực hiện của thuật toán CLARA như sau:

Thuật toán CLARA:

Input: CSDL gồm n đối tượng, số cụm k.

Output: k cụm dữ liệu

1. For i = 1 to 5 do

Begin

2. Lấy một mẫu có 40 + 2k đối tượng dữ liệu ngẫu nhiên từ tập dữ liệu và áp dụng thuật toán PAM cho mẫu dữ liệu này nhằm để tìm các đối tượng medoid đại diện cho các cụm.

3. Đối với mỗi đối tượng Oj trong tập dữ liệu ban đầu, xác định đối tượng medoid tương tự nhất trong số k đối tượng medoid.

4. Tính độ phi tương tự trung bình cho phân hoạch các đối tượng dành ở bước trước, nếu giá trị này bé hơn giá trị tối thiểu hiện thời thì sử dụng giá trị này thay cho giá trị tối thiểu ở trạng thái trước, như vậy tập k đối tượng medoid xác định ở bước này là tốt nhất cho đến thời điểm hiện tại.

End;

Độ phức tạp tính toán của thuật toán là $O(k(40+k)^2 + k(n-k))$, và CLARA có thể thực hiện đối với tập dữ liệu lớn. Chú ý đối với kỹ thuật tạo mẫu trong PCDL: kết quả phân cụm có thể không phụ thuộc vào tập dữ liệu khởi tạo nhưng nó chỉ đạt tối ưu cục bộ.

2.1.3. Thuật toán CLARANS

Thuật toán CLARANS (A Clustering Algorithm based on RANdomized Search) được Ng & Han đề xuất năm 1994, nhằm để cải tiến chất lượng cũng như mở rộng áp dụng cho tập dữ liệu lớn. CLARANS là thuật toán PCDL kết hợp thuật toán PAM với chiến lược tìm kiếm kinh

nghiệm mới. Ý tưởng cơ bản của CLARANS là không xem xét tất cả các khả năng có thể thay thế các đối tượng tâm medoids bởi một đối tượng khác, nó ngay lập tức thay thế các đối tượng medoid này nếu việc thay thế có tác động tốt đến chất lượng phân cụmchứ không cần xác định cách thay thể tối ưu nhất. Một phân hoạch cụm pháthiện được sau khi thay thế đối tượng trung tâm được gọi là một láng giềng của phân hoạch cụm trước đó. Số các láng giềng được hạn chế bởi tham số do người dùng đưa vào là Maxneighbor, quá trình lựa chọn các láng giềng này là hoàntoàn ngẫu nhiên. Tham số Numlocal cho phép người dùng xác định số vòng lặp tối ưu cục bộ được tìm kiếm. Không phải tất các láng giềng được duyệt mà chỉ có Maxneighbor số láng giềng được duyệt.

Thuật toán CLARANS có thể được diễn tả như sau:

Input: Tập dữ liệu gồm n đối tượng, số cụm k, O, dist, numlocal, maxneighbor;

Output: k cụm dữ liệu;

For i=1 to numlocal do

Begin

Khởi tạo ngẫu nhiên k medois

j = 1;

while j < maxneighbor do

Begin

Chọn ngẫu nhiên một láng giềng R của S.

Tính toán độ phi tương tự về khoảng cách giữa 2 láng giềng S và R. Nếu R có chi phí thấp hơn thì hoán đổi R cho S và j=1

ngược lại j++;

End;

Kiểm tra khoảng cách của phân hoạch S có nhỏ hơn khoảng cách nhỏ nhất không, nếu nhỏ hơn thì lấy giá trị này để cập nhật lại khoảng cách nhỏ nhất và phân hoạch S là phân hoạch tốt nhất tại thời điểm hiện tại.

End.

Như vậy, quá trình hoạt động của CLARANS tương tự với quá trình hoạt động của thuật toán CLARA. Tuy nhiên, ở giai đoạn lựa chọn các trung tâm medoid cụm dữ liệu, CLARANS lựa chọn một giải pháp tốt hơn bằng cách lấyngẫu nhiên một đối tượng của k đối tượng trung tâm medoid của cụm và cố gắng thay thế nó với một đối tượng được chọn ngẫu nhiên trong (n-k) đối tượng còn lại, nếu không có giải pháp nào tốt hơn sau một số cố gắng lựa chọn ngẫu nhiên xác định, thuật toán dừng và cho kết quả phân cụm tối ưu cục bộ.

Trong trường hợp xấu nhất, CLARANS so sánh một đối tượng với tất các đối tượng Medoid. Vì vậy, độ phức tạp tính toán của CLARANS là O(kn2), dovậy CLARANS không thích hợp với tập dữ liệu lớn. CLARANS có ưu điểm là không gian tìm kiếm không bị giới hạn như đối với CLARA và trong cùng một lượng thời gian thì chất lượng của các cụm phân được là lớn hơn so với CLARA.

2.2. Phân cụm phân cấp

Phân cụm phân cấp (Hierarchical Methods) xây dựng một phân cấp trên cơ sở các đối tượng dữ liệu đang xem xét. Nghĩa là sắp xếp một tập dữ liệu đã

cho thành một cấu trúc có dạng hình cây, cây phân cấp này được xây dựng theo kỹ thuật đệ quy. Có hai cách tiếp cận phổ biến của kỹ thuật này đó là: hòa nhập nhóm, thường được gọi là tiếp cận (Bottom-Up); phân chia nhóm, thường được gọi là tiếp cận (Top-Down)

Điển hình trong phương pháp tiếp cận theo phân cụm phân cấp là các thuật toán như: AGNES (Agglomerative Nesting), DIANA (Divisive Analysis), BIRCH (1996), CURE (1998), CHAMELEON (1999) . . . Thực tế áp dụng, có nhiều trường hợp kết hợp cả hai phương pháp phân cụm phân hoạch và phân cụm phân cấp, nghĩa là kết quả thu được của phương pháp phân cấp có thể cải tiến thông qua bước phân cụm phân hoạch. Phân cụm phân hoạch và phân cụm phân cấp là hai phương pháp phân cụm dữ liệu cổ điển, hiện đã có rất nhiều thuật toán cải tiến dựa trên hai phương pháp này đã được áp dụng phổ biến trong khai phá dữ liệu.

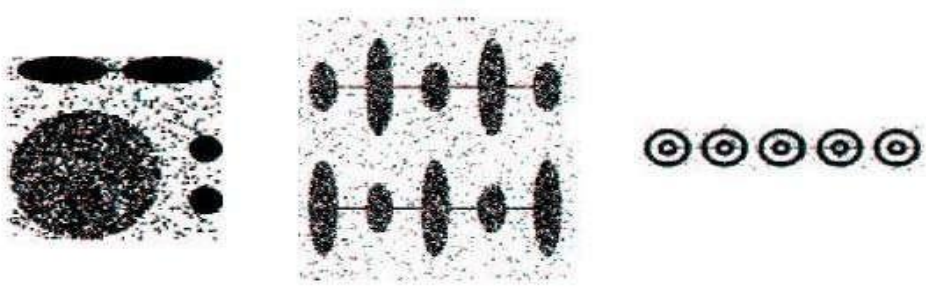
2.2.1. Thuật toán CURE

Trong khi hầu hết các thuật toán thực hiện phân cụm với các cụm hình cầu và kích thước tương tự, như vậy là không hiệu quả khi xuất hiện các phần tử ngoại lai. Thuật toán CURE khắc phục được vấn đề này và tốt hơn với các phần tử ngoại lai. Thuật toán này định nghĩa một số cố định các điểm đại diện nằm rải rác trong toàn bộ không gian dữ liệu và được chọn để mô tả các cụm được hình thành. Các điểm này được tạo ra nhờ lựa chọn các đối tượng nằm rải rác cho cụm và sau đó "co lại" hoặc di chuyển chúng về trung tâm cụm bằng nhân tố co cụm. Quá trình này được lặp lại và như vậy trong quá trình này, có thể đo tỉ lệ gia tăng của cụm. Tại mỗi bước của thuật toán, hai cụm có cặp các điểm đại diện gần nhau (mỗi điểm trong cặp thuộc về mỗi cụm khác nhau) được hòa nhập.

Nhưvậy,cónhiềuhơnmộtđiểmđạidiệnmỗicụmchophépCUREkhámphá

đượccác cụmcóhình dạng khôngphảilà hình cầu. Việcco lạicáccụmcótác dụng làmgiảmtácđộng củacácphần tửngoạilai. Nhưvậy,thuậttoánnàycókhảnăngxửlýtốttrongtrườnghợpcócácphầntửngoạilaiv àlàmchohiệuquảvớinhữnghìnhdạngkhôngphảilàhìnhcầuvàkíchthướcđộrộngbi ếnđổi.Hơnnữa,nótỉlệtốtvớicơsởdữliệulớnmàkhônglàmgiảmchấtlượngphâncụ m.

Hình 2.2: Các cụm dữ liệu được khám phá bởi CURE



Đểxửlýđượccáccơsởdữliệulớn,CUREsửdụngmẫungẫunhiênvàphânhoạc h,mộtmẫulàđượcxácđịnhngẫunhiêntrướckhiđượcphânhoạchvàsauđótiếnhànhp hâncụmtrênmỗiphânhoạch,nhưvậymỗiphânhoạch làtừngphầnđãđượcphâncụm,cáccụmthuđượclạiđượcphâncụmlầnthứhaiđểthuđ ượccáccụmconmongmuốn,nhưngmẫungẫunhiênkhôngnhấtthiếtđuramộtmôtả chotoànbộtậpdữliệu.

ThuậttoánCUREđượcthựchiệnquacácbướccơbảnsau:

Chọn một mẫu ngẫu nhiên S từ tập dữ liệu ban đầu.

Phân hoạch mẫu S thành các nhóm dữ liệu có kích thước bằng nhau: Ý tưởng chính ở đây là phân hoạch mẫu thành p nhóm dữ liệu bằng nhau, kích thước của mỗi phân hoạch là n'/p (n' là kích thước của mẫu).

Phân cụm các điểm của mỗi nhóm: thực hiện PCDL cho các nhóm cho đến khi mỗi nhóm được phân thành n'/pq cụm (với q>1).

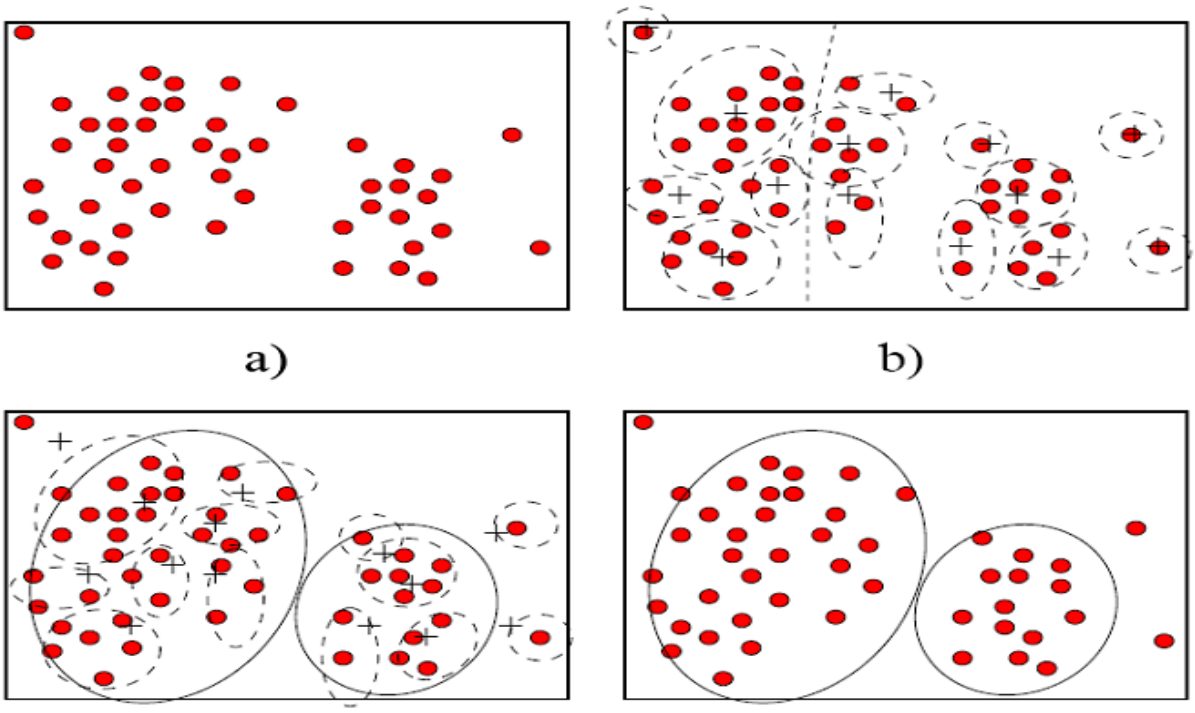
Loại bỏ các phần tử ngoại lai: trước hết, khi các cụm được hình thành cho đến khi số các cụm giảm xuống một phần so với số các cụm ban đầu. Sau đó, trong trường hợp các phần tử ngoại lai được lấy mẫu cùng với quá trình pha khởi tạo mẫu dữ liệu, thuật toán sẽ tự động loại bỏ các nhóm nhỏ.

Phân cụm các không gian: các đối tượng đại diện cho các cụm di chuyển về hướng trung tâm cụm, nghĩa là chúng được thay thế bởi các đối tượng gần trung tâm hơn.

Đánh dấu dữ liệu với các nhãn tương ứng.

Độphứctạptínhtoáncủathuậttoán CURE là O(n^2log(n)). CURE làthuậttoánt incậytrongviệckhámphára cáccụmcóhìnhthùbấtkỳvàcóthểápdụngtốtđốivớidữliệucóphầntửngoạilaivàtrênc áctậpdữliệuhaichiều. Tuynhiên, nólạirấtnhạycảmvớithamsốnhưsốcácđốitượngđ ạidiện, tỉlệ co củacácphầntử đạidiện.

Hình 2.3: Ví dụ thực hiện phân cụm bằng thuật toán CURE



2.2.2. Thuật toán CHAMELEON

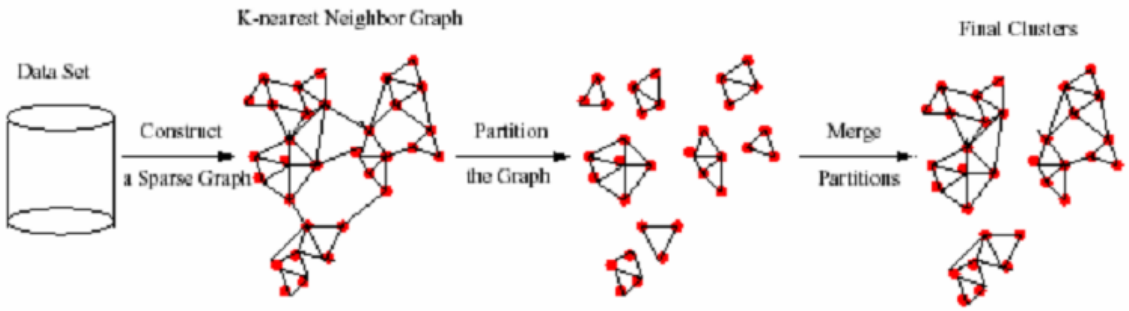
Một giải thuật phân cụm phân cấp sử dụng mô hình động gọi là CHAMELEON, nó khảo sát môhình hoá động trong phân cụm phân cấp, được phát triển bởi Karypis, Han vàKumar (1999). Khi xử lý phân cụm, 2 cụm được hoà nhập nếu liên kết nối và độchặt (độ gần) giữa hai cụm được liên kết cao với liên kết nối và độ chặt nội tạicủa các đối tượng nằm trong phạm vi các cụm. Xử lý hoà nhập dựa trên mô hìnhđộng tạo điều kiện thuận lợi cho sự khám phá ra các cụm tự nhiên và đồng nhất,nó áp dụng cho tất cả các kiểu dữ liệu miễn là hàm tương đồng được chỉ định.

CHAMELEON có được dựa trên quan sát các yếu điểm của hai giải thuật phân cụm phân cấp: CURE và ROCK. CURE và các lược đồ quan hệ bỏ qua thông tin về liên kết nối tổng thể của các đối tượng trong 2 cụm; ngược lại,

ởROCK, các lược đồ quan hệ lờ đi thông tin về độ chặt của 2 cụm trong khi nhấn mạnh liên kết nối của chúng.

CHAMELEON trước tiên sử dụng một giải thuật phân chia đồ thị để phân cụm các mục dữ liệu vào trong một số lượng lớn các cụm con tương đối nhỏ. Sau đó dùng giải thuật phân cụm phân cấp tập hợp để tìm ra các cụm xác thực bằng cách lặp lại việc kết hợp các cụm này với nhau. Để xác định các cặp cụm con giống nhau nhất, cần đánh giá cả liên kết nối cũng như độ chặt của các cụm, đặc biệt là các đặc tính nội tại của bản thân các cụm. Do vậy nó không tuỳ thuộc vào một mô hình tĩnh được cung cấp bởi người dùng và có thể tự động thích ứng với các đặc tính nội tại của các cụm đang được hoà nhập.

Hình 2.4: Mô hình CHAMELEON, Phân cụm phân cấp dựa trên k-láng giềng gần và mô hình hóa động



Như trong hình, CHAMELEON miêu tả các đối tượng dựa trên tiếp cận đồ thị được dùng phổ biến: k-láng giềng gần nhất. Mỗi đỉnh của đồ thị k-láng giềng gần nhất đại diện cho một đối tượng dữ liệu, tại đó tồn tại một cạnh giữa hai đỉnh (đối tượng), nếu một đối tượng là giữa k đối tượng giống nhau so với các đối tượng khác. Đồ thị k-láng giềng gần nhất Gk có được khái niệm láng giềngđộng: Bán kính láng giềng của một điểm dữ liệu được xác định bởi mật độ củamiền mà trong đó các đối tượng cư trú. Trong một miền

dày đặc, láng giềng được định nghĩa hẹp, và trong một miền thưa thớt, láng giềng được định rộng hơn. So sánh với mô hình định nghĩa bởi phương pháp dựa trên mật độ như DBSCAN, DBSCAN dùng mật độ láng giềng toàn cục, Gk có được láng giềng tự nhiên hơn. Hơn nữa, mật độ miền được ghi như trọng số của các cạnh. Cạnh của một miền dày đặc theo trọng số lớn hơn so với của một miền thưa thớt.

CHAMELEON chỉ rõ sự tương đồng giữa mỗi cặp các cụm Ci và Cj theo liên kết nối tương đối $RI(C_i,C_j)$ và độ chặt tương đối $RC(C_i,C_j)$ của chúng.

Liên kết nối tương đối $RI(C_i,C_j)$ giữa hai cụm C_i và C_j được định nghĩa như liên kết nối tuyệt đối giữa C_i và C_j đã tiêu chuẩn hoá đối với liên kết nối nội tạicủa hai cụm C_i và C_j. Đó là:

$$RI(C_i,C_j)=\frac{\left|EC_{\{C_i,C_j\}}\right|}{\frac{1}{2}\left(\left|EC_{C_i}\right|+\left|EC_{C_j}\right|\right)}$$

với $EC_{\{Ci,\ Cj\}}$, là cạnh cắt (edge-cut) của cụm chứa cả C_i và C_j để cụm này được rơi vào trong C_i và C_j, và tương tự như vậy, ECC_i (hay ECC_j) là kích thước của min-cut bisector (tức là tổng trọng số của các cạnh mà chia đồ thị thành hai phần thô bằng nhau).

2.3.Phân cụm dựa trên mật độ

Phân cụm dựa trên mật độ (Density-Based Methods) nhóm các đối tượng dữ liệu dựa trên hàm mật độ xác định, mật độ là số các đối tượng lân cận của một đối tượng dữ liệu theo một nghĩa nào đó. Trong cách tiếp cận này, khi một dữ liệu đã xác định thì nó tiếp tục được phát triển thêm các đối tượng dữ liệu mới miễn là số các đối tượng lân cận này phải lớn hơn một ngưỡng đã được xác định trước. Phương pháp phân cụm dựa trên mật độ của các đối tượng để xác định các cụm dữ liệu có thể phát hiện ra các cụm dữ liệu

với hình thù bất kỳ. Kỹ thuật này có thể khắc phục được các phần tử ngoại lai hoặc giá trị nhiễu rất tốt, tuy nhiên việc xác định các tham số mật độ của thuật toán là rất khó khăn, trong khi các tham số này lại có tác động rất lớn đến kết quả phân cụm.

Chúng tại có các thuật toán phân cụm dựa trên mật độ như: DBSCAN(KDD'96), DENCLUE (KDD'98), CLIQUE (SIGMOD'98)), OPTICS (SIGMOD'99)…

2.3.1. Thuật toán DBSCAN

Thuật toán DBSCAN (Density – Based Spatial Clustering of Applications with Noise) là một giải thuật phân cụm dựa trên mật độ, được phát triển bởi Ester, Kriegel, Sander và Xu năm 1996. Giải thuật này tăng trưởng các miền với mật độ cao vào trong các cụm và khám phá ra các cụm có hình dạng bất kỳ trong không gian cơ sở dữ liệu có nhiễu.

Ý tưởng cơ bản của phân cụm dựa trên mật độ: Đối với mỗi đối tượng của một cụm, láng giềng trong một bán kính cho trước (ε) (gọi là ε-láng giềng) phải chứa ít nhất một số lượng tối thiểu các đối tượng (MinPts).

Một đối tượng nằm trong một bán kính cho trước (ε) chứa không ít hơn một số lượng tối thiểu các đối tượng láng giềng (MinPts), được gọi là đối tượng nồng cốt (core object) đối với bán kính (ε) và số lượng tối thiểu các điểm (MinPts).

Một đối tượng p là mật độ trực tiếp tiến (directly density-reachable) từ đối tượng q với bán kính ε và số lượng tối thiểu các điểm MinPts trong một tập các đối tượng D nếu p trong phạm vi ε-láng giềng của q với q chứa ít nhất một số lượng tối thiểu điểm MinPts.

Một đối tượng p là mật độ tiến (density-reachable) từ đối tượng q với bán kính ε và MinPts trong một tập hợp các đối tượng D nếu như có một đối

tượng p_1, p_2, …,p_n, p_1=q và p_n=p với $1 \leq i \leq n$, $p_i \in D$ và p_{i+1} là mật độ trực tiếp tiến từ p_i đối với ε và MinPts

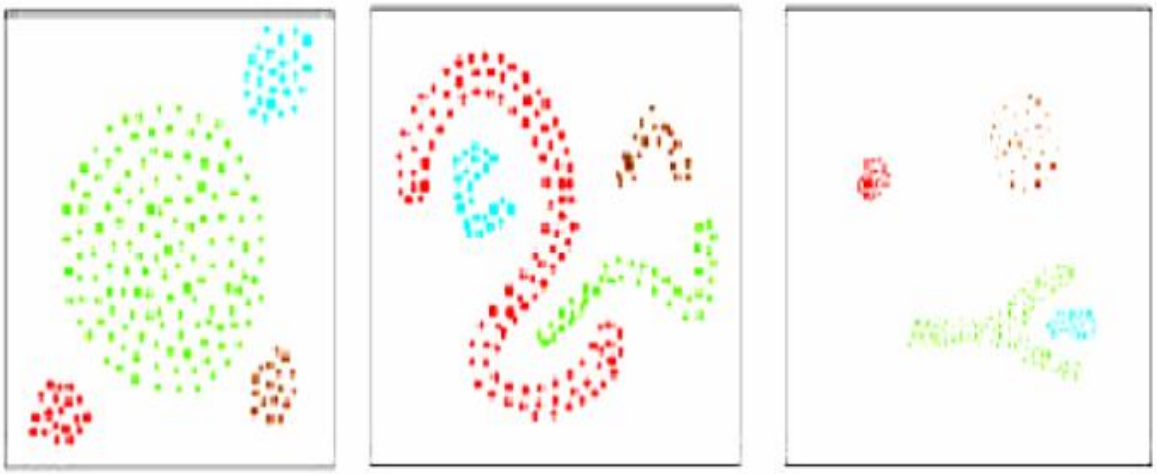
Một đối tượng p là mật độ liên kết với đối tượng q đối với ε và MinPts

trong một ập đối tượng D nếu như có một đối tượng o $\in$ D để cả p và q là mật độ tiến từ o đối với ε và MinPts.

DBSCAN có thể tìm ra các cụm với hình thù bất kỳ, trong khi đó tại cùng một thời điểm ít bị ảnh hưởng bởi thứ tự của các đối tượng dữ liệu nhập vào. Khi có một đối tượng được chèn vào chỉ tác động đến một láng giềng xác định. Mặt khác, DBSCAN sử dụng tham số ε và MinPts trong thuật toán để kiểm soát mật độ của các cụm. DBSCAN bắt đầu với một điểm tuỳ ý và xây dựng mật độ láng giềng có thể được đối với ε và MinPts. Vì vậy, DBSCAN yêu cầu người dùng xác định bán kính ε của các láng giềng và số các láng giềng tối thiểu MinPts, các tham số này khó mà xác định được tối ưu, thông thường nó được xác định bằng phép chọn ngẫu nhiên hoặc theo kinh nghiệm.

Độ phức tạp của DBSCAN là $O(n^2)$, nhưng nếu áp dụng chỉ số không gian để giúp xác định các láng giềng của một đối tượng dữ liệu thì độ phức của DBSCAN đã được cải tiến là O(nlogn). Thuật toán DBSCAN có thể áp dụng cho các tập dữ liệu không gian lớn đa chiều, khoảng cách Euclide được sử dụng để đo sự tương tự giữa các đối tượng nhưng không hiệu quả đối với dữ liệu đa chiều.

Hình 2.5:Hình dạng các cụm được khám phá bởi thuật toán DBSCAN



Thuật toán : DBSCAN khởi tạo điểm p tùy ý và lấy tất cả các điểm liên lạc mật độ từ p tới ε và MinPts. Nếu p là điểm nhân thì thủ tục trên tạo ra một cụm theo ε và MinPts, nếu p là một điểm biên, không có điểm nào liên lạc mật độ từ p và DBSCAN sẽ đi thăm điểm tiếp theo của tập dữ liệu.

Nếu sử dụng giá trị toàn cục ε và MinPts, DBSCAN có thể hoà nhập hai cụm thành một cụm nếu mật độ của hai cụm gần bằng nhau. Giả sử khoảng cách giữa hai tập dữ liệu S1 và S2 được định nghĩa là :

dist(S1, S2) = min{dist(p, q)} {p ∈ S1 và q ∈ S2}.

2.3.2. Thuật toán OPTICS

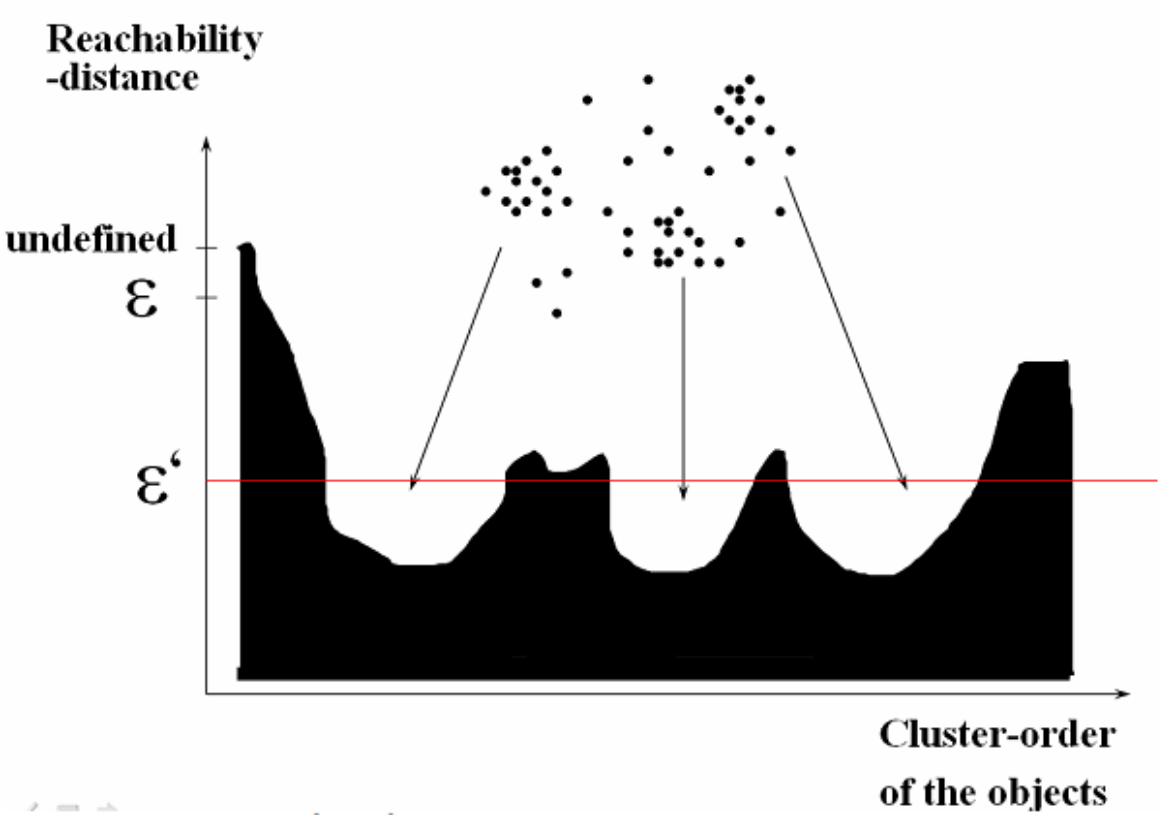
Mặc dù giải thuật phân cụm dựa trên mật độ DBSCAN có thể tìm ra cụm các đối tượng với việc lựa chọn các tham số đầu vào như ε và MinPts, người dùng vẫn chịu trách nhiệm lựa chọn các giá trị tham số tốt để tìm ra các cụm chính xác. Trên thực tế, đây là bài toán có sự kết hợp của nhiều giải thuật phân cụm khác. Các thiết lập tham số như vậy tương đối khó, đặc biệt trong thế giới thực, các tập dữ liệu có số chiều cao. Hầu hết các giải thuật rất nhạy với các tham số: các thiết lập có sự khác biệt nhỏ có thể dẫn tới các phân chia dữ liệu rất khác nhau. Hơn nữa, các tập dữ liệu thực số chiều cao thường

có phân bố rất lệch, thậm trí ở đó không tồn tại một thiết lập tham số toàn cục cho đầu vào.

Để khắc phục khó khăn này, một phương pháp sắp xếp cụm gọi là OPTICS (Ordering Point To Identify the Clustering Structuer) được phát triển bởi Ankerst, Breunig, Kriegel và Sander năm 1999, cải tiến bằng cách giảm bớt các tham số đầu vào. Thuật toán này không phân cụm các điểm dữ liệu mà thực hiện tính toán và sắp xếp trên các điểm dữ liệu theo thứ tự tăng dần nhằm tự động phân cụm dữ liệu và phân tích cụm tương tác hơn là đưa ra phân cụm một tập dữ liệu rõ ràng. Đây là thứ tự mô tả cấu trúc phân dữ liệu cụm dựa trên mật độ của dữ liệu, nó chứa thông tin tương ứng với phân cụm dựa trên mật độ từ một dãy các tham số được thiết lập và tạo thứ tự của các đối tượng trong cơ sở dữ liệu, đồng thời lưu trữ khoản cách lõi và khoảng cách liên lạc phù hợp của mỗi đối tượng. Hơn nữa, thuật toán được đề xuất rút ra các cụm dựa trên thứ tự thông tin. Như vậy thông tin đủ cho trích ra tất cả các cụm dựa trên mật độ khoảng cách bất kỳ ε’ mà nhỏ hơn khoảng cách ε được sử dụng trong sinh thứ tự.

Việc sắp xếp thứ tự được xác định bởi hai thuộc tính riêng của các điểm dữ liệu đó là khoảng cách nhân và khoảng cách liên lạc. Các phép đo này chính là kích thước mà có liên quan đến quá trình của thuật toán DBSCAN, tuy nhiên, chúng được sử dụng để xác định thứ tự của các điểm dữ liệu đã được xắp xếp. Thứ tự dựa trên cơ sở các điểm dữ liệu mà có khoảng cách nhân nhỏ nhất và tăng dần độ lớn. Điều duy nhất về phương pháp này là người sử dụng không phải xác định giá trị ε hoặc MinPts phù hợp.

Hình 2.6: Sắp xếp cụm trong OPTICS phụ thuộc vào ε [8]



Thuật toán này có thể phân cụm các đối tượng đã cho với các tham số đầu vào như ε và MinPts, nhưng nó vẫn cho phép người sử dụng tùy ý lựa chon các giá trị tham số mà sẽ dẫn đến khám phá các cụm chấp nhận được. Các thiết lập tham số thường dựa theo kinh nghiệm tập hợp và khó xác định, đặc biệt là với các tập dữ liệu đa chiều.

Tuy nhiên, nó cũng có độ phức tạp thời gian thực hiện như DBSCAN bởi vì có cấu trúc tương đương với DBSCAN: O(nlogn) với n là kích thước của tập dữ liệu. Thứ tự cụm của tập dữ liệu có thể được biểu diễn bằng đồ thị, và được minh họa hình sau, có thể thấy ba cụm, giá trị ε quyết định số cụm.

2.4. Phân cụm dựa trên lưới

Phân cụm dựa trên lưới (Grid-Based Methods) dựa trên lưới thích hợp với dữ liệu nhiều chiều, dựa trên cấu trúc dữ liệu lưới để phân cụm, phương

pháp này chủ yếu tập trung áp dụng cho lớp dữ liệu không gian. Mục tiêu của phương pháp này là lượng hóa dữ liệu thành các ô tạo thành cấu trúc dữ liệu lưới. Sau đó, các thao tác phân cụm chỉ cần làm việc với các đối tượng trong từng ô trên lưới chứ không phải các đối tượng dữ liệu. Cách tiếp cận dựa trên lưới này không di chuyển các đối tượng trong các ô mà xây dựng nhiều mức phân cấp của nhóm các đối tượng trong một ô. Phương pháp này gần giống với phương pháp phân cụm phân cấp nhưng chúng không trộn các ô, đồng thời giải quyết khắc phục yêu cầu đối với dữ liệu nhiều chiều mà phương pháp phân phân cụm dựa trên mật độ không giải quyết được. ưu điểm của phương pháp phân cụm dựa trên lưới là thời gian xử lí nhanh và độc lập với số đối tượng dữ liệu trong tập dữ liệu ban đầu, thay vào đó là chúng phụ thuộc vào số ô trong mỗi chiều của không gian lưới.

2.4.1. Thuật toán STING

STING (STatistical INformation Grid) (Wang, Yang và Munz 1997) là một tiếp cận đa phân giải dựa trên lưới. Trong đó vùng không gian dữ liệu được phân rã thành số hữu hạn các cells chữ nhật, điều này có ý nghĩa là các cells lưới được hình thành từ các cells lưới con để thực hiện phân cụm. Có nhiều mức của các cells chữ nhật tương ứng với các mức khác nhau của phân giải trong cấu trúc lưới, và các cells này hình thành cấu trúc phân cấp: mỗi cells ở mức cao được phân hoạch thành các số các cells nhỏ ở mức thấp hơn tiếp theo trong cấu trúc phân cấp. Các điểm dữ liệu được nạp từ CSDL, giá trị của các tham số thống kê cho các thuộc tính của đối tượng dữ liệu trong mỗi ô lưới được tính toán từ dữ liệu và lưu trữ thông qua các tham số thống kê ở các cell mức thấp hơn (điều này giống với cây CF). Các giá trị của các tham số thống kê gồm : số trung bình – mean, số tối đa – max, số tối thiểu – min, số đếm –count , độ lệch chuẩn –s,…

Các đối tượng dữ liệu lần lượt được chèn vào lưới và các tham số thống kê ở trên được tính trực tiếp thông qua các đối tượng dữ liệu này. Các truy vấn không gian được thực hiện bằng cách xét các cells thích hợp tại mỗi mức phân cấp. Một truy vấn không gian được xác định như là một thông tin khôi phục lại của dữ liệu không gian và các quan hệ của chúng. STING có khả năng mở rộng cao, nhưng do sử dụng phương pháp đa phân giải nên nó phụ thuộc chặt chẽ vào trọng tâm của mức thấp nhất. Đa phân giải là khả năng phân rã tập dữ liệu thành các mức chi tiết khác nhau. Khi hòa nhập các cells của cấu trúc lưới để hình thành các cụm, nó không xem xét quan hệ không gian giữa các nút của mức con không được hòa nhập phù hợp (do chúng chỉ tương ứng với các cha của nó) và hình dạng của các cụm dữ liệu khám phá là isothetic, tất cả ranh giới của các cụm có các biên ngang và dọc, theo biên của các cells và không có đường biên chéo được phát hiện ra.

Các lợi thế của cách tiếp cận này so với các phương pháp phân cụm dữ liệu khác:

- Tính toán dựa trên lưới là truy vấn độc lập vi thông tin thống kê được bảo quản trong mỗi cells đại diện nên chỉ cần thông tin tóm tắt của dữ liệu trong cells chứ không phải là dữ liệu thực tế và không phụ thuộc vào câu truy vấn.

- Cấu trúc dữ liệu lưới thuận tiện cho quá trình xử lý song song và cập nhật liên tục.

- Duyệt toàn bộ CSDL một lần để tính toán các đại lượng thống kê cho mỗi cells, nên nó hiệu quả và do đó độ phức tạp thời gian để tạo các cụm xấp xỉ O(n), trong đó n là tổng số các đối tượng. Sau khi xây dựng cấu trúc phân cấp, thời gian xử lý cho các truy vấn là O(g), trong đó g là tổng số cells lưới ở mức thấp (g<<n)

Các hạn chế của thuật toán này :

- Trong khi sử dụng cách tiếp cận đa phân giải để thực hiện phân tích cụm chất lượng của phân cụm STING hoàn toàn phụ thuộc vào tính chất hộp ở mức thấp nhất của cấu trúc lưới. Nếu tính chất hộp là mịn, dẫn đến chi phí thời gian xử lý tăng, tính toán trở nên phức tạp và nếu mức dưới cùng là quá thô thì nó có thể làm giảm bớt chất lượng và độ chính xác của phân tích cụm.

Thuật toán STING:

1. Xác định tầng để bắt đầu
2. Với mỗi cái của tầng này, tính toán khoảng tin cậy (hoặc ước lượng khoảng) của xác suất mà cells này liên quan tới truy vấn
3. Từ khoảng tin cậy của tính toán trên,gán nhãn cho là có liên quan hoặc không liên quan.
4. Nếu lớp này là lớp cuối cùng , chuyển sang Bước 6; nếu khác thì chuyển sang Bước 5
5. Duyệt xuống dưới của cấu trúc cây phân cấp một mức. Chuyển sang Bước 2 cho các cells mà hình thành các cells liên quan của lớp có mức cao hơn.
6. Nếu đặc tả được câu truy vấn, chuyển sang bước 8; nếu không thì chuyển sang bước 7.
7. Truy lục lại dữ liệu vào trong các cells liên quan và thực hiện xử lý. Trả lại kết quả phù hợp yêu cầu của truy vấn. Chuyển sang Bước 9.
8. Tìm thấy các miền có các cells liên quan. Trả lại miền mà phù hợp với yêu cầu của truy vấn. Chuyển sang bước 9.
9. Dừng.

2.4.2. Thuật toán CLIQUE

Trong không gian đa chiều, các cụm có thể tồn tại trong tập con của các chiều hay còn gọi là không gian con. Thuật toán CLIQUE là thuật toán

hữu ích cho PCDL không gian đa chiều trong các CSDL lớn thành các không gian con. Thuật toán này bao gồm các bước:

- Cho n là tập lớn của các điểm dữ liệu đa chiều; không gian dữ liệu thường là không giống nhau bởi các điểm dữ liệu. Phương pháp này xác định những vùng gần, thưa và "đặc" trong không gian dữ liệu nhất định, bằng cách đó phát hiện ra toàn thể phân bố mẫu của tập dữ liệu.

- Một đơn vị là dày đặc nếu phần nhỏ của tất cả các điểm dữ liệu chứa trong nó vượt quá tham số mẫu đưa vào. Trong thuật toán CLIQUE, cụm được định nghĩa là tập tối đa liên thông các đơn vị dày đặc.

Các đặc trưng của CLINQUE

- Tự động tìm kiếm không gian con của không gian đa chiều, sao cho mật độ đặc của các cụm tồn tại trong không gian con.

- Mẫn cảm với thứ tự của dữ liệu vào và không phù hợp với bất kỳ quy tắc phân bố dữ liệu nào.

- Phương pháp này tỷ lệ tuyến tính với kích thước vào và có tính biến đổi tốt khi số chiều của dữ liệu tăng.

Nó phân hoạch tập dữ liệu thành các hình hộp chữ nhật và tìm các hình hộp chữ nhật đặc, nghĩa là các hình hộp này chứa một số các đối tượng dữ liệu trong số các đối tượng láng giếng cho trước. Hợp các hình hộp này tạo thành các cụm dữ liệu. Tuy nhiên, CLINQUE được bắt đầu bằng cách tiếp cận đơn giản do đó chính xác của kết quả phân cụm có thể bị ảnh hưởng dẫn tới chất lượng của các phương pháp này có thể giảm.

Phương pháp bắt đầu nhận dạng các cells đặc đơn chiều trong không gian dữ liệu và tìm kiếm phân bố của dữ liệu, tiếp đến CLINQUE lần lượt tìm các hình chữ nhật 2 chiều, 3 chiều,… cho đến khi hình hộp chữ nhật đặc k chiều được tìm thấy, độ phức tạp tính toán của CLIQUE là O(n).

2.4.3. Thuật toán WaveCluster

Thuật toán WaveCluster (Sheikholeslami, Chatterjee và Zhang 1998) là một tiếp cận phân cụm đa phân giải. phương pháp này gần giống với STING, tuy nhiên thuật toán sử dụng phép biến đổi dạng sóng để tìm ô đặc trong không gian. Đầu tiên kỹ thuật này tóm tắt dữ liệu bằng việc tận dụng cấu trúc dạng lưới đa chiều lên trên không gian dữ liệu. Tiếp theo nó sử dụng phép biến đổi dạng sóng để biến đổi không gian có đặc trưng gốc, tìm kiếm ô trong không gian đã được biến đổi. Phương pháp này là phức tạp với các phương pháp khác chính là ở phép biến đổi.

Ở đây, mỗi cells lưới tóm tắt thông tin các điểm của một nhóm ánh xạ vào trong cells. Đây là thông tin tiêu biểu thích hợp đưa vào bộ nhớ chính để sử dụng phép biến đổi dạng sóng đa phân giải và tiếp theo là phân tích cụm.

Một phép biến đổi dạng sóng là kỹ thuật dựa trên cơ sở xử lý tín hiệu và xử lý ảnh bằng phân tích tín hiệu với tần số xuất hiện trong bộ nhớ chính. Bằng việc thực hiện một loạt các phép biến đổi ngược phức tạp cho nhóm này, nó cho phép các cụm trong dữ liệu trở thành rõ ràng hơn. Các cụm này có thể được xác định bằng tìm kiếm ô đặc trong vùng mới.

Phương pháp này phức tạp, nhưng lại có những lợi thế:

- Cung cấp cụm không giám sát, khử nhiễu các thông tin bên ngoài biên của cụm. Theo cách đó, vùng đặc trong không gian đặc trưng gốc hút các điểm ở gần và ngăn chặn các điểm ở xa. Vì vậy, các cụm tự động nổi bật và làm sạch khu vực xung quanh nó, do đó các kết quả tự động loại phần tử ngoại lai.

- Đa phân giải là thuộc tính hỗ trợ dò tìm các cụm có các mức biến đổi chính xác.

- Thực hiện nhanh với độ phức tạp của thuật toán là O(n), trong đó n là số đối tượng trong CSDL. Thuật toán có thể thích hợp với xử lý song song.

- Xử lý tập dữ liệu lớn có hiệu quả, khám phá các cụm có hình dạng bất kỳ, xử lý phần tử ngoại lai, mẫn cảm với thứ tự vào, và không phụ thuộc vào các tham số vào như số các cụm hoặc bán kính láng giềng.

Giải thuật phân cụm dựa trên wavelet phác thảo như sau: Giải thuật phân cụm dựa trên wavelet đối với phân cụm đa phân giải bằng phép biến đổi wavelet.

Đầu vào: Các vectơ đặc trưng của các đối tượng dữ liệu đa chiều

Đầu ra: Các đối tượng đã phân cụm

Giải thuật:

1) Lượng tử hoá không gian đặc trưng, sau đó phân các đối tượng vào các unit;

2) Áp dụng phép biến đổi wavelet trong không gian đặc trưng;

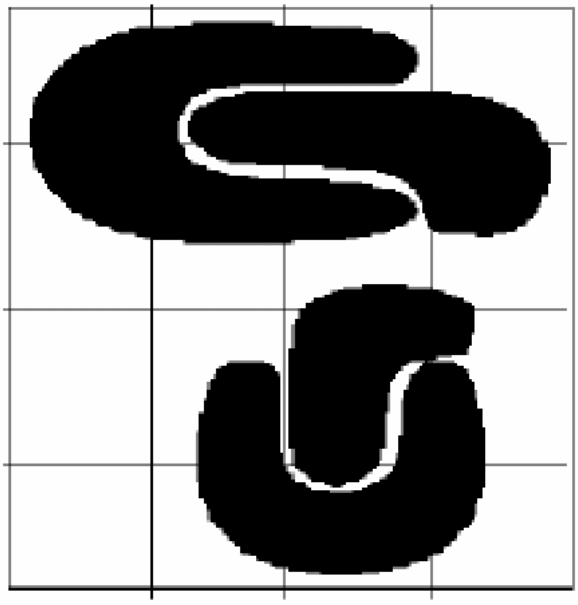
3) Tìm các phần hợp thành đã kết nối (các cụm) trong các dải con của không gian đặc trưng đã biến đổi tại các mức khác nhau;

4) Gắn các nhãn vào các unit;

5) Làm các bảng tra cứu và ánh xạ các đối tượng vào các cụm.

Độ phức tạp tính toán của giải thuật này là O(N) với N là số các đối tượng trong cơ sở dữ liệu.

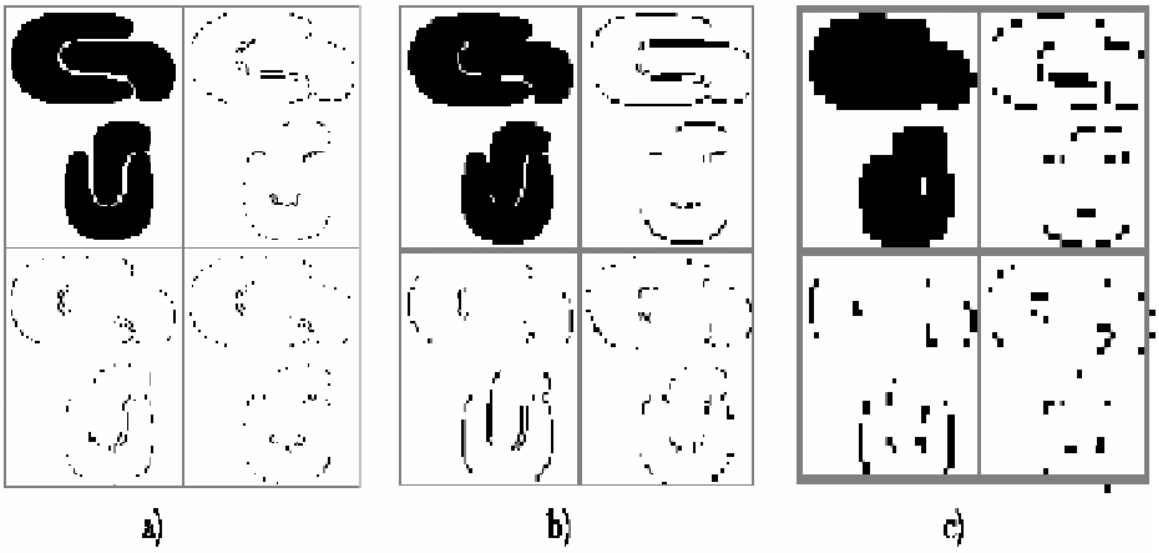
Hình 2.7: Một mẫu không gian đặc trưng 2 chiều



Ví dụ: Hình 2.7 cho thấy một mẫu không gian đặc trưng 2 chiều, tại đó, mỗi điểm trong ảnh đại diện cho các giá trị đặc trưng của một đối tượng trong các tập dữ liệu không gian.

Hình 2.8 cho thấy kết quả của các phép biến đổi wavelet tại các tỷ lệ khác nhau, từ mịn (tỷ lệ 1) cho tới thô (tỷ lệ 3). Tại mỗi mức, dải con LL (bình thường) chỉ ra tại cung phần tư phía trên bên trái, dải con LH (các cạnh nằm ngang) chỉ ra tại cung phần tư phía trên bên phải và dải con HL (các cạnh nằm dọc) chỉ ra tại cung phần tư phía dưới bên trái và dải con HH (các góc) chỉ ra tại cung phần tư phía dưới bên phải.

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.



2.5. Phân cụm dựa trên mô hình

Phân cụm dựa trên mô hình (Model-Based Clustering Methods) cố gắng khớp giữa các dữ liệu với mô hình toán học, nó dựa trên giả định rằng dữ liệu được tạo ra bằng hỗn hợp phân phối xác suất cơ bản. Các thuật toán phân cụm dựa trên mô hình có hai cách tiếp cận chính: mô hình thống kê và mạng nơron. Phương pháp này gần giống với phương pháp phân cụm dựa trên mật độ, vì chúng phát triển các cụm riêng biệt nhằm cải tiến các mô hình đã được xác định trước đó, nhưng đôi khi nó không bắt đầu với một số cụm cố định và không sử dụng cùng một khái niệm mật độ cho các cụm.

2.5.1. Thuật toán EM

Thuật toán EM (Expectation - Maximization) được nghiên cứu từ 1958 bởi Hartley và được nghiên cứu đầy đủ bởi Dempster, Laird và Rubin công bố năm 1977. Thuật toán này nhằm tìm ra sự ước lượng về khả năng lớn nhất của các tham số trong mô hình xác suất (các mô hình phụ thuộc vào các biến

tiềm ẩn chưa được quan sát), nó được xem như là thuật toán dựa trên mô hình hoặc là mở rộng của thuật toán k-means. EM gán các đối tượng cho các cụm đã cho theo xác suất phân phối thành phần của đối tượng đó. Phân phối xác suất thường được sử dụng là phân phối xác suất Gaussian với mục đích là khám phá lặp các giá trị tốt cho các tham số của nó bằng hàm tiêu chuẩn là hàm logarit khả năng của đối tượng dữ liệu, đây là hàm tốt để mô hình xác suất cho các đối tượng dữ liệu.

Thuật toán gồm 2 bước xử lý: Đánh giá dữ liệu chưa được gán nhãn (bước E) và đánh giá các tham số của mô hình, khả năng lớn nhất có thể xẩy ra (bước M).

Cụ thể thuật toán EM ở bước lặp thứ t thực hiện các công việc sau:

1) Bước E: Tính toán để xác định giá trị của các biến chỉ thị dựa trên mô hình hiện tại và dữ liệu:

$$z_{ij}^{(t)} = E_{\psi}(z_{ij} \mid x) = \Pr_{\psi}(z_{ij} = 1 \mid x) = \frac{f_j(x_i)\pi_j^{(t)}}{\sum_g^k 1 fg^{(x_i)\pi} g}$$

2) Bước M: Đánh giá xác suất π:

$$\pi_j^{(t+1)} = \sum_{i=1}^{n} z_{ij}^{(t)} / n$$

EM có thể khám phá ra nhiều hình dạng cụm khác nhau, tuy nhiên do thời gian lặp của thuật toán khá nhiều nhằm xác định các tham số tốt nên chí phí tính toán của thuật toán là khá cao. Đã có một số cải tiến được đề xuất cho EM dựa trên các tính chất của dữ liệu: có thể nén, có thể sao lưu trong bộ nhớ và có thể huỷ bỏ. Trong các cải tiến này, các đối tượng bị huỷ bỏ khi biết chắc chắn được nhãn phân cụm của nó, chúng được nén khi không bị loại bỏ và thuộc về một cụm quá lớn so với bộ nhớ và chúng sẽ được lưu lại trong các trường hợp còn lại.

2.5.2. Thuật toán COBWEB

COBWEB được đề xuất bởi Fisher năm 1987. Các đối tượng đầu vào của thuật toán được mô tả bởi cặp thuộc tính-giá trị, nó thực hiện phân cụm phân cấp bằng cách tạo cây phân lớp, các cấu trúc cây khác nhau.

Thuật toán này sử dụng công cụ đánh giá heuristic được gọi là công cụ phân loại CU (Category untility) để quản lý cấu trúc cây. Từ đó cấu trúc cây được hình thành dựa trên phép đo độ tương tự mà phân loại tương tự và phi tương tự, cả hai có thể mô tả phân chia giá trị thuộc tính giữa các nút trong lớp. Cấu trúc cây có thể hợp nhất hoặc phân tách khi chèn một nút mới vào cây.

Các bước chính của thuật toán:

1) Khởi tạo cây bắt đầu bằng một nút rỗng.

2) Sau khi thêm vào từng nút một và cập nhật lại cây cho phù hợp tại mỗi thời điểm.

3) Cập nhật cây bắt đầu từ lá bên phải trong mỗi trường hợp, sau đó cấu trúc lại cây.

4) Quyết định cập nhật dựa trên sự phân hoạch và các hàm tiêu chuẩn phân loại.

Tại mỗi nút, nó xem xét 4 khả năng xảy ra (Insert, Create, Merge, Split) và lựa chọn một khả năng có hàm giá trị CU đạt được tốt nhất của quá trình.

Một số hạn chế của COBWEB là nó thừa nhận phân bố xác suất trên các thuộc tính đơn lẻ là độc lập thống kê và chi phí tính toán phân bố xác suất của các cụm khi cập nhật và lưu trữ là khá cao.

Các phương pháp cải tiến của thuật toán COBWEB là CLASSIT, AutoClass.

2.6.Phân cụm dữ liệu mờ

Thông thường, mỗi phương pháp PCDL phân một tập dữ liệu ban đầu thành các cụm dữ liệu có tính tự nhiên và mỗi đối tượng dữ liệu chỉ thuộc về một cụm dữ liệu, phương pháp này chỉ phù hợp với việc khám phá ra các cụm có mật độ cao và rời nhau. Tuy nhiên, trong thực tế, các cụm dữ liệu lại có thể chồng lên nhau (một số các đối tượng dữ liệu thuộc về nhiều các cụm khác nhau), người ta đã áp dụng lý thuyết về tập mờ trong PCDL để giải quyết cho trường hợp này, cách thức kết hợp này được gọi là phân cụm mờ. Trong phương pháp phân cụm mờ, độ phụ thuộc của đối tượng dữ liệu x_k tới cụm thứ i (u_{ik}) có giá trị thuộc khoảng [0,1]. Ý tưởng trên đã được giới thiệu bởi Ruspini (1969) và được Dunn áp dụng năm 1973 nhằm xây dựng một phương pháp phân cụm mờ dựa trên tối thiểu hoá hàm tiêu chuẩn. Bezdek (1982) đã tổng quát hoá phương pháp này và xây dựng thành thuật toán phân cụm mờ c-means có sử dụng trọng số mũ.

c-means là thuật toán phân cụm mờ (của k-means). Thuật toán c-means mờ hay còn gọi tắt là thuật toán FCM (Fuzzy c- means) đã được áp dụng thành công trong giải quyết một số lớn các bài toán PCDL như trong nhận dạng mẫu, xử lý ảnh, y học, … Tuy nhiên, nhược điểm lớn nhất của thuật toán FCM là nhạy cảm với các nhiễu và phần tử ngoại lai, nghĩa là các trung tâm cụm có thể nằm xa so với trung tâm thực tế của cụm.

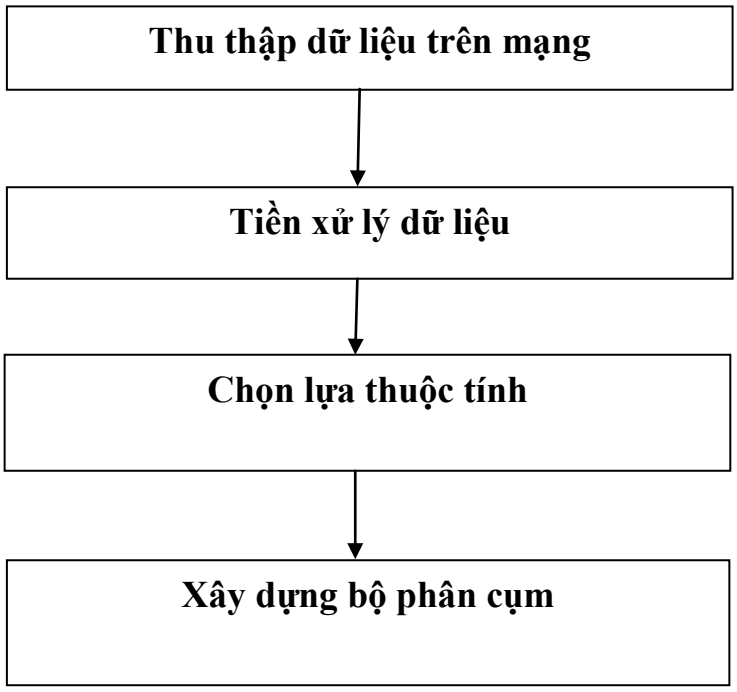
Đã có nhiều các phương pháp đề xuất để cải tiến cho nhược điểm trên của thuật toán FCM bao gồm: Phân cụm dựa trên xác suất (keller, 1993), phân cụm nhiễu mờ (Dave, 1991), Phân cụm dựa trên toán tử L_P Norm (Kersten, 1999). Thuật toán Insensitive Fuzzy c-means.

CHƯƠNG III: ỨNG DỤNG KỸ THUẬT PHÂN CỤM DỮ LIỆU TRONG PHÁT HIỆN XÂM NHẬP TRÁI PHÉP

3.1. Mô hình bài toán

Khai phá dữ liệu đang ngày càng được ứng dụng nhiều hơn trong việc phát hiện xâm nhập trái phép. Trong đó, các thuật toán phân cụm thường được sử dụng nhằm xây dựng các mô hình phát hiện xâm nhập trái phép, để dự đoán và phát hiện các tấn công mới. Mô hình bài toán phân cụm trong phát hiện xâm nhập trái phép dựa trên 5 bước cơ bản: bước đầu tiên là thu thập, thống kê tất cả các luồng dữ liệu, các gói tin TCP/IP trên mạng. Tiếp theo các gói tin được chuyển đổi thành các bản ghi dữ liệu có các giá trị thuộc tính như: dịch vụ, khoảng thời gian, số byte dữ liệu,…. Những thuộc tính quan trọng sau đó được lựa chọn và cuối cùng là sử dụng các thuật toán phân cụm để xây dựng ra các bộ phân cụm.

Hình 3.1: Các bước xây dựng mô hình phát hiện xâm nhập trái phép



3.1.1. Thu thập dữ liệu

Để thực hiện đánh giá các thuật toán phân cụm trong việc xây dựng các mô hình phát hiện xâm nhập trái phép, luận văn này sử dụng tập dữ liệu KDD Cup 1999 được xây dựng từ năm 1998 của tổ chức DARPA (cục quốc phòng Mỹ và quản lý bởi Trung tâm thí nghiệm MIT Lincoln) [9]. Đây là tập dữ liệu được trích rút từ gói tin có định dạng của giao thức TCP. Để thu thập được các dữ liệu tấn công, các cuộc tấn công đã được giả lập theo các hành động và mục tiêu cụ thể của kẻ tấn công.

Tập dữ liệu bao gồm một kiểu dữ liệu bình thường (normal) và 22 kiểu tấn công khác nhau được phân loại thành 4 lớp: từ chối dịch vụ (DoS), trinh sát hệ thống (Probe), chiếm quyền hệ thống (U2L) và khai thác điểm yếu (R2L).

- Lớp từ chối dịch vụ (DoS): Đây là loại tấn công có mục đích là làm giảm hoặc tê liệt khả năng cung cấp dịch vụ cho người sử dụng, của các ứng dụng hoặc các dịch vụ chạy trên mạng. Phương thức chung của các kiểu tấn công này là làm quá tải hệ thống và tràn ngập băng thông. Một số kiểu tấn công DOS phổ biến được liệt kê dưới bảng sau:

TT	Tên tấn công	Mô tả
1	Pod	Gửi các gói tin có kích thước lớn thông qua lệnh ping đến máy đích.
2	Smurf	Lợi dụng Router mạng để gửi Broadcast
3	Neptune	Đột nhập vào hệ thống
4	Teardrop	Gửi chồng chéo gói tin
5	Back	Tấn cống đến các đường định tuyến
6	Land	Làm chậm, hoặc bị treo hệ thống ứng dụng.

Bảng 3.1: Bảng mô tả lớp tấn công từ chối dịch vụ (DoS).

- Trinh sát hệ thống (Probe): Là loại tấn công có mục đích là thu thập các thông tin liên quan đến cấu hình của một hệ thống máy tính hoặc hệ thống mạng

nhằm mục đích phá hoại. Phương thức chung là sử dụng các công cụ dò quét, để tìm kiếm các cổng mở hay địa chỉ IP.

TT	Tên tấn công	Mô tả
1	Satan	Công cụ quét cổng và thăm dò
2	Portsweep	Sử dụng các gói kết nối để xác định cổng mở
3	Nmap	Công cụ quét cổng
4	Ipsweep	Sử dụng các gói ping để xác định ip

Bảng 3.2: Bảng mô tả lớp tấn công trinh sát(Probe)

- Lớp tấn công chiếm quyền hệ thống (U2R): Đây là loại tấn công nhằm chiếm đoạt quyền cao nhất hoặc chiếm dụng, kiểm soát một máy tính khi kẻ tấn công có một tài khoản đăng nhập bình thường (với quyền hạn chế). Phương thức của kiểu công này là truy cập vào hệ thống như một người dùng bình thường sau đó sử dụng các phương pháp leo thang đặc quyền để lấy quyền quản trị hệ thống, một số loại phổ biến được liệt kê trong bảng sau:

TT	Tên tấn công	Mô tả
1	buffer_overflow	Làm tràn bộ đệm.
2	Loadmodule	Lợi dụng các điểm yếu, để thực thi các module quản trị hệ thống .
3	Perl	Công cụ để tăng quyền user
4	Rootkit	Công cụ điều hành cao nhất trong hệ thống

Bảng 3.3: Bảng mô tả lớp tấn công leo thang đặc quyền (U2R).

- Lớp tấn công khai thác điểm yếu từ xa (R2L): Đây là kiểu tấn công khi kẻ tấn công gửi các gói tin hoặc đoạn mã đến một máy tính qua internet lợi dụng các điểm yếu của máy tính đó, từ đó khai thác các đặc quyền của các người dùng cục bộ.

TT	Tên tấn công	Mô tả
1	guess_passwd	Đoán password
2	Ftpwirte	Thay đổi quyền để truy cập vào FTP
3	Imap	Tấn công vào dịch vụ mail
4	Phf	Tấn công vào cơ sở dũ liệu web
5	Spy	Tấn công vào trình duyệt web bất kỳ.
6	Warezclient	Tấn công theo kiểu bom tấn
7	Warezmaster	Tấn công làm thay đổi ngay lập tức.
8	Multihop	Tấn công nhiều dạng cùng một lúc.

Bảng 3.4: Bảng mô tả lớp tấn công truy cập từ xa (R2L).

3.1.2. Trích rút và lựa chọn thuộc tính.

Dựa vào tập dữ liệu KDD 1999, luận văn đã lựa chon các thuộc tính cơ bản từ các gói tin kết nối đến của một giao thức TCP, chẳng hạn như khoảng thời gian kết nối, kiểu giao thức, số lượng byte dữ liệu, các cờ để chỉ ra tình trạng lỗi kết nối hoặc bình thường…. Các thuộc tính của một kết nối đơn được thông qua các lĩnh vực tri thức, và kể cả các các hoạt động tạo tập tin và một số hoạt động cố gắng truy cập vào hệ thống.

Trích chọn thuộc tính (Feature Selection, Feature Extraction) là nhiệm vụ rất quan trọng trong giai đoạn tiền xử lý dữ liệu khi triển khai các mô hình khai phá dữ liệu. Một vấn đề gặp phải là các tập dữ liệu dùng để xây dựng các mô hình khai phá dữ liệu thường chứa nhiều thông tin không cần thiết (thậm chí gây nhiễu) cho việc xây dựng mô hình. Chẳn hạn, một tập dữ liệu gồm hàng trăm thuộc tính dùng để mô tả về khách hàng của một doanh nghiệp được thu thập, tuy nhiên khi xây dựng một mô hình khai phá dữ liệu nào đó chỉ cần khoảng 50 thuộc tính từ hàng trăm thuộc tính đó. Nếu ta sử dụng tất cả các thuộc tính (hàng trăm, hàng ngàn) của khách hàng để xây dựng mô hình thì ta cần bộ xử lý CPU phải đủ lớn, nhiều bộ nhớ trong quá trình học mô hình, thậm chí các thuộc tính không cần thiết đó làm giảm độ chính xác của mô hình và gây khó khăn trong việc phát hiện tri thức.

Trong tập dữ liệu này có 41 thuộc tính được trích chọn. Bảng thuộc tính mô tả như sau:

TT	Tên thuộc tính	Mô tả
1	Duration	Khoảng thời gian (số giây) của kết nối.
2	protocol_type	Kiểu giao thức (TCP, UDP, ICMP).
3	Service	Các dịch vụ trên mạng.
4	Flag	Tình trạng bình thường hay lỗi kết nối.
5	src_bytes	Số lượng byte dữ liệu từ nguồn tới đích.
6	dst_bytes	số lượng byte dữ liệu từ đích đến nguồn.
7	Land	1 nếu kết nối đến máy chủ, 0 ngược lại.
8	wrong_fragment	Số sai trong phân mảnh.
9	Urgent	Số lượng gói tin khẩn cấp.
10	Hot	Số lượng "nóng" các chỉ số.
11	num_failed_logins	Số lần đăng nhập thất bại.
12	logged_in	1 nếu thành công, 0 nếu thất bại.
13	num_compromised	Số điều kiện thoả hiệp.
14	root_shell	1 nếu gốc đạt được, 0 ngược lại.
15	su_attempted	1 nếu là quyền root, 0 ngược lại.
16	num_root	Số root truy cập.
17	num_file_creations	Số lượng tạo tập tin.
18	num_shells	Số lượng cảnh báo.
19	num_access_files	Số hoạt động trên các tập tin kiểm soát truy cập.
20	num_outbound_cmd	Số các lệnh gửi đi trong một phiên ftp.
21	Is_host_login	1 nếu đăng nhập vào thuộc danh sách nóng, 0 ngược lại.
22	Is_guest_login	1 đăng nhập là một khách, 0 ngược lại.
23	Count	Số lượng kết nối cùng một máy chủ cùng 2

TT	Tên thuộc tính	Mô tả
		giây.
24	srv_count	Số lượng kết nối cùng một dịch vụ trong 2 giây.
25	serror_rate	% các kết nối "SYN" lỗi.
26	srv_serror_rate	% các kết nối "SYN" lỗi.
27	rerror_rate	% của các kết nối "REJ" lỗi.
28	srv_serror_rate	% của các kết nối "REJ" lỗi.
29	same_srv_rate	% kết nối các dịch vụ tương tự.
30	diff_srv_rate	% các kết nối đến các dịch vụ khác nhau.
31	srv_diff_host_rate	% Các kết nối đến các máy chủ khác nhau.
32	dst_host_count	Số lượng kết nối đến máy chủ nguồn.
33	dst_host_srv_count	Số lượng kết nối từ nguồn đến đích.
34	dst_host_same_srv_rate	% kết nối máy chủ đích đến nguồn các dịch vụ tương tự
35	dst_host_diff_srv_rate	% máy chủ kết nối từ đích đến nguồn qua các dịch vụ khác nhau.
36	dst_host_same_srv_port_rate	% kết nối máy chủ đích đến nguồn các dịch vụ tương tự qua cổng.
37	dst_host_srv_diff_host_rate	% máy chủ kết nối từ đích đến nguồn qua các dịch vụ khác nhau.
38	dst_host_serror_rate	% của các kết nối máy chủ đích "SYN" lỗi
39	dst_host_srv_serror_rate	% của các kết nối máy chủ đích đến nguồn "SYN" lỗi.
40	dst_host_rerror_rate	% của các kết nối máy chủ đích "REJ" lỗi
41	dst_host_srv_rerror_rate	% của các kết nối máy chủ đích đến nguồn "REJ" lỗi.

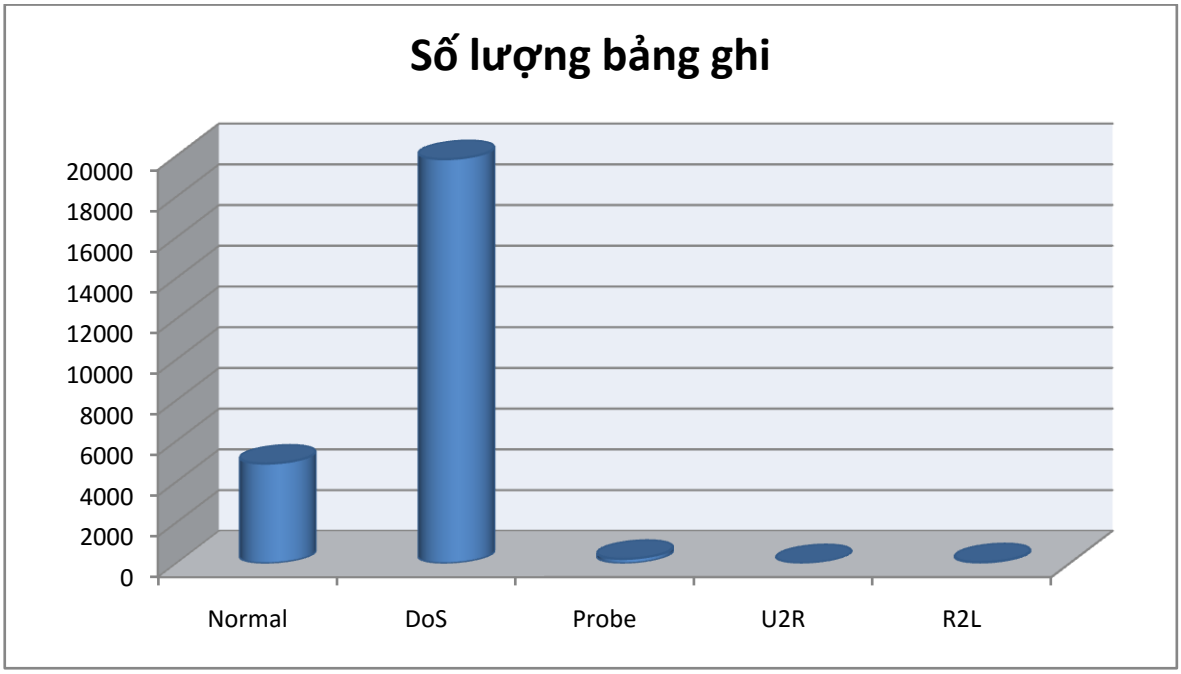
Bảng 3.5: Bảng mô tả 41 thuộc tính của tập dữ liệu KDD Cup 1999

- Trong tập dữ liệu KDD Cup 1999 ta trích chọn một phần dữ liệu để làm thực nghiệm. Bao gồm 25.000 bảng ghi và có 41 thuộc tính. Phân phối của các bản ghi như sau:

Lớp	Số lượng bảng ghi (dataset)	Tỉ lệ (%)
Normal	4893	19.572
DoS	19843	79.372
Probe	214	0.856
U2R	2	0.008
R2L	48	0.192
Tổng cộng	25000	100

Bảng 3.6: Bảng phân phối số lượng bản ghi.

Hình 3.2: Số lượng bản ghi có trong tập dữ liệu thực nghiệm



3.1.3. Xây dựng bộ phân cụm

Luận văn thực hiện các thực nghiệm để xây dựng các mô hình phát hiện xâm nhập trái phép dựa trên các thuật toán phân cụm. Tập dữ liệu thực nghiệm bao gồm

25.000bản ghi, 41 thuộc tính và 16 kiểu tấn công khác nhau được sử dụng. Trên cơ sở tập dữ liệu xây dựng để thực nghiệm, luận văn tập trung phân tích kỹ thuật phân cụm khác nhau giữa các cụmtrong tập dữ liệu, đưa ra phương án có độ chính xác cao nhất và thời gian thực hiện giữa các cụm.

Các bước xây dựng bộ phân cụm:

Bước 1. Loại bỏ các thuộc tính lớp của tập dữ liệu

Bước 2. Sử dụng tập dữ liệu để áp dụng các thuật toán phân cụm như K-means, EM,… để xây dựng ra các cụm dữ liệu.

Bước 3. Gắn lại các thuộc tính lớp vào các đối tượng đã được phân cụm.

Bước 4. Sử dụng tập dữ liệu đã phân cụm để đánh giá độ chính xác của quá trình tấn công sử dụng các cụm đã có.

3.2.Xây dựng các thực nghiệm phát hiện xâm nhập trái phép

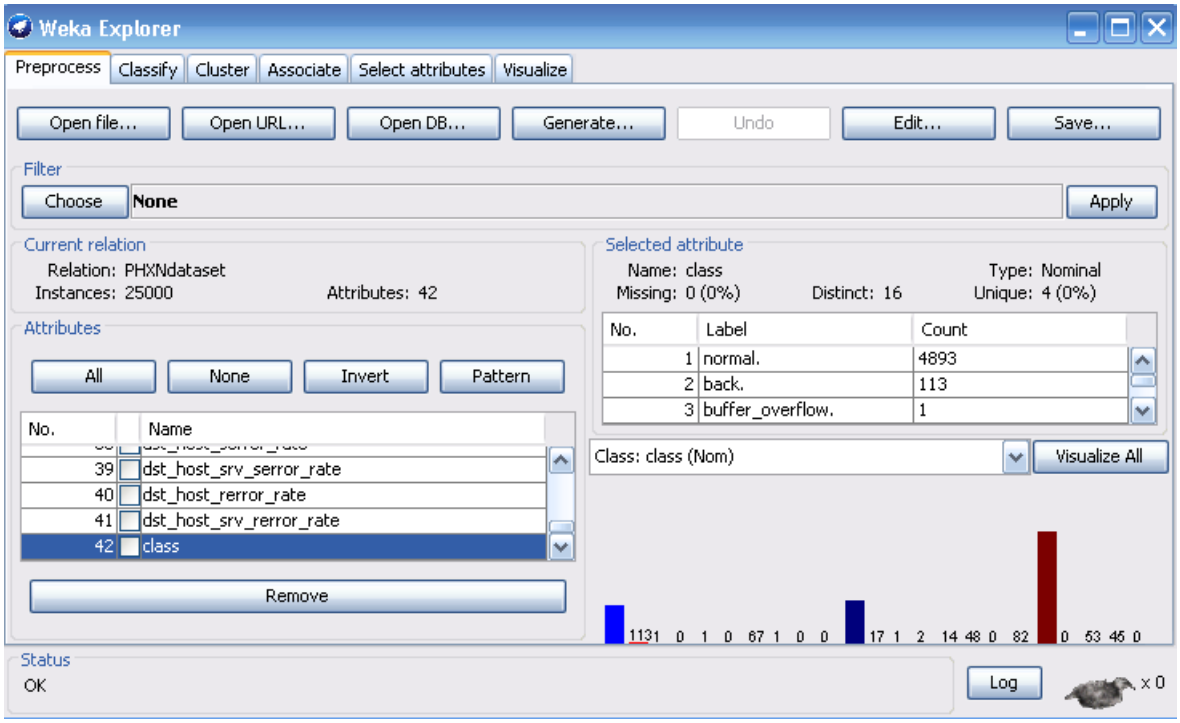
3.2.1. Môi trường và công cụ thực nghiệm

Luận văn sử dụng phần mềm mã nguồn mở WEKA (WaikatoEnviromentforKnowledge Analysis) được cài đặt trên máy tính với hệ điều hành window XP 32bits, bộ xử lý Core dual 1.8GHz, bộ nhớ Ram 1Gb.

Để cung cấp một môi trường tính toán và xây dựng đồ họa cho việc phân tích dữ liệu từ các tập dữ liệu thu thập được, luận văn đưa các tập dữ liệu và cài đặt các bước thuật toán trên công cụ Weka Explore [10][11] để thực hiện phân cụm và đánh giá độ chính xác, thời gian thực hiện.

Ngoài ra, luận văn sử dụng chương trình hiển thị kết quả Treeview với nguồn dữ liệu sau khi phân cụm là Cluster 3.0 để trực quan thấy được cụ thể kết quả phân cụm của các kiểu tấn công.

Hình 3.3: Tập dữ liệu đưa vào phân cụm qua Weka Explorer

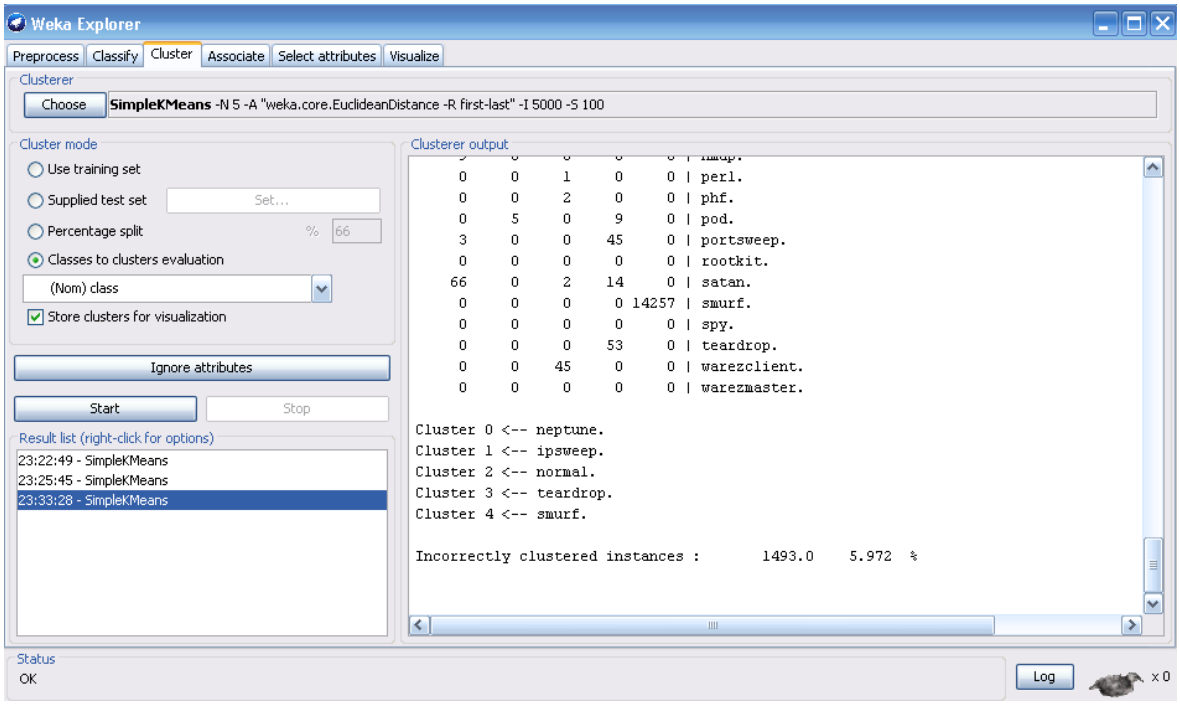


3.2.2. Tiến hành các thực nghiệm và kết quả đạt được

3.2.2.1.Phân cụm K-Means

Phân cụm K-means trong Weka có thuật toán Simple K-means [12][13], thuật toán hỗ trợ hai hàm để đo khoảng cách giữa các điểm là hàm Euclidean, Manhattan. Trong thực nghiệm này luận văn sử dụng hàm Euclidean. Tham số seed được sử dụng để sinh ra số ngẫu nhiên chọn các tâm cụm ban đầu để khởi tạo thuật toán. Trong thuật toán này luận văn sử dụng số seed cố định bằng 100 và thay đổi số cụm.

Hình 3.4: Tham số cài đặt phân cụm K-meansvới Weka Explorer



Kết quả phân cụm K-means với các cụm 3, 4, 5 như sau:

Phân cụm K-means	Độ chính xác (%)	Thời gian (Giây)
K=3	98.07%	9.19
K=4	93.88%	10.02
K=5	94.03%	23.61

Bảng 3.7: Kết quả phân cụm K-meansvới các cụm k khác nhau

Theo Bảng kết quảphân cụmK-means với các cụm k khác nhau thìkhi k=3 cho tỷ lệ độ chính xác là cao nhất và thời gian là ít nhất.

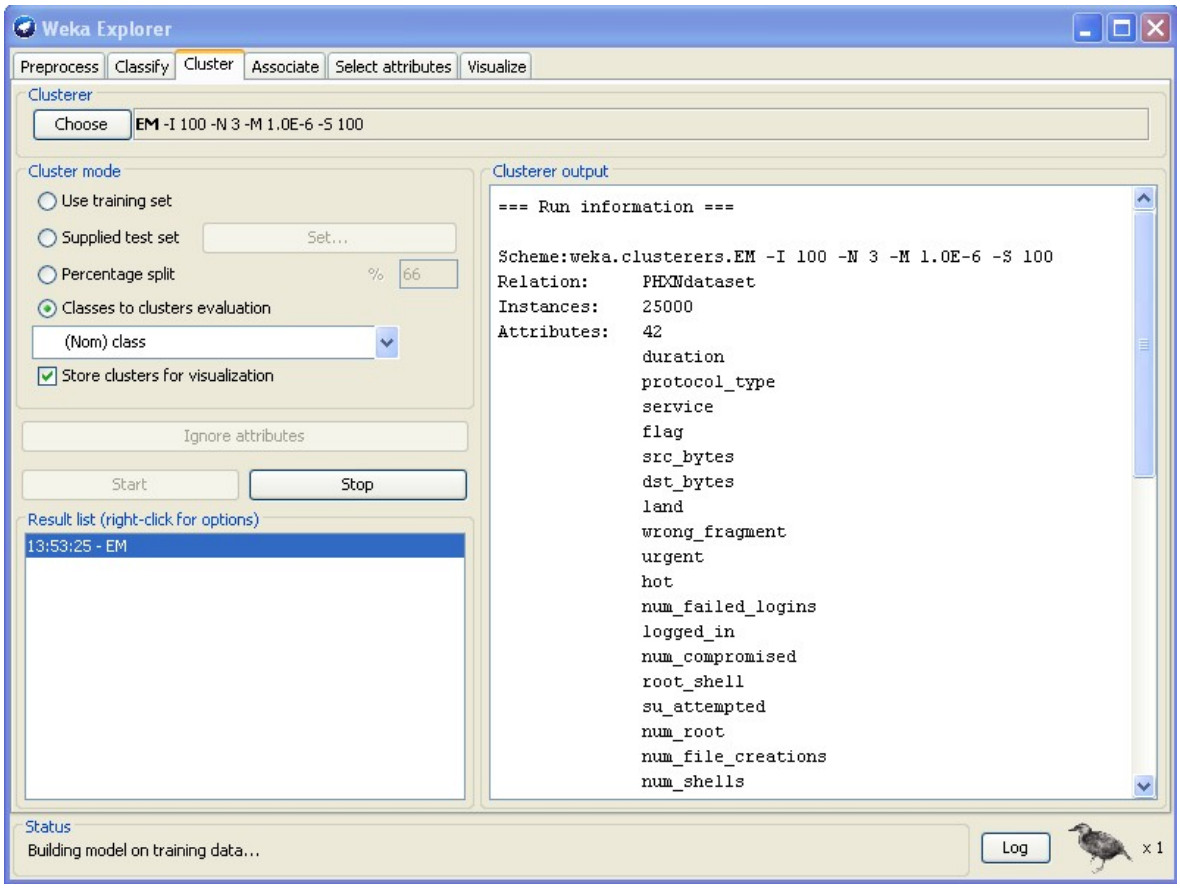
3.2.2.2. Phân cụm EM

- Thuật toán EM: EM cũng là một thuật toán quan trọng trong khai phá

dữ liệu. Chúng ta sử dụng thuật toán này khi chúng ta không thỏa mãn với kết quả của thuật toán K-Means. Bản chất của thuật toán EM là một thuật toán lặp nhằm tìm ra độ đo likelihood lớn nhất hoặc tối đa ước tính các thông số trong mô hình thống kê, nơi các mô hình phụ thuộc vào các biến tiềm ẩn không quan sát được.

Đối với thuật toán EM, luận văn sử dụng số seed bằng 100, số cụm thay đổi, tham số minStdDev = 1.0E-6, maxIterations = 100. Tiến hành thực nghiệm thuật toán EM trên Weka với tham số như hình dưới, ta thu được bảng dữ liệu sau:

Hình 3.5: Tham số cài đặt phân cụm EM với Weka Explorer



Kết quả phân cụm EM với các cụm 3, 4, 5 như sau

Phân cụm EM	Likelihood	Độ chính xác (%)	Thời gian (Giây)
K=3	41.435	98.13%	88.99
K=4	47.36	93.24%	94.55
K=5	42.83	88.49%	136.5

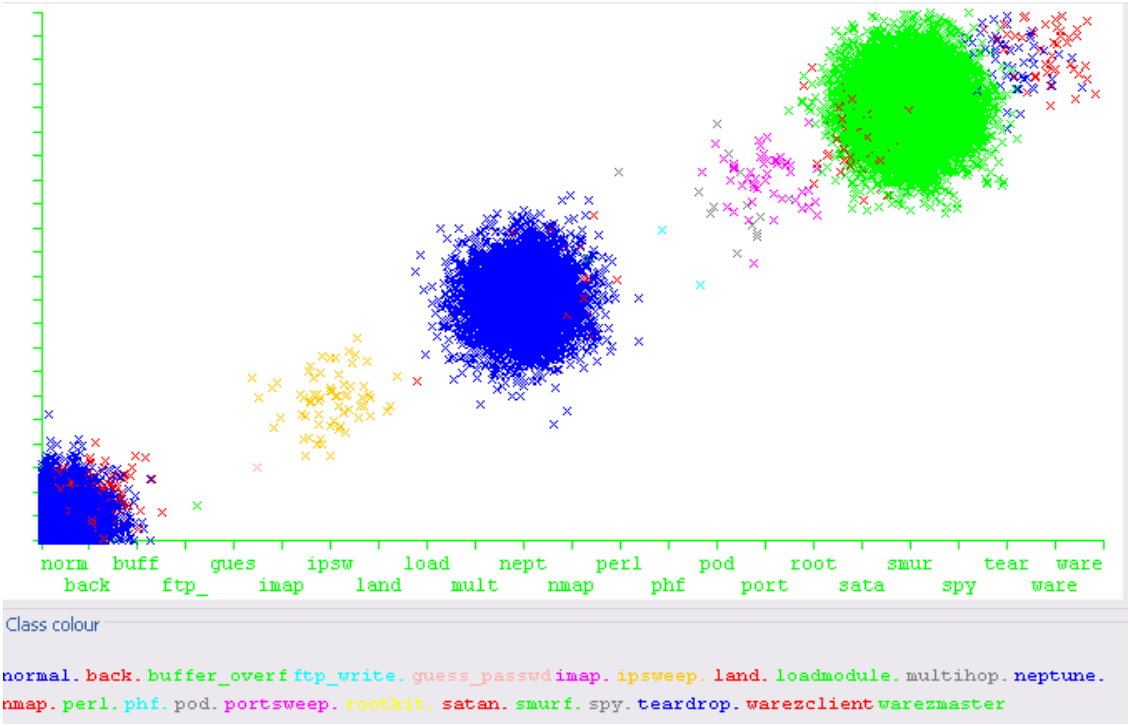
Bảng 3.8: Kết quả phân cụm EM với các cụm k khác nhau

So sánh với độ chính xác khi phân lớp thì số cụm cho giá trị likelihood tốt nhất chưa chắc đã cho giá trị độ chính xác tốt nhất. Độ chính xác tốt nhất của bộ dữ liệu của đề tài thực hiện tốt nhất đối với 3 cụm (k=3) là 98.13% và thời gian thực hiện nhanh nhất là 88.99 giây.

3.2.2.3. Đồ họa trực quan kết quả phân cụm

* Biểu diễn kết quả phân cụm theo Weka Explorer:

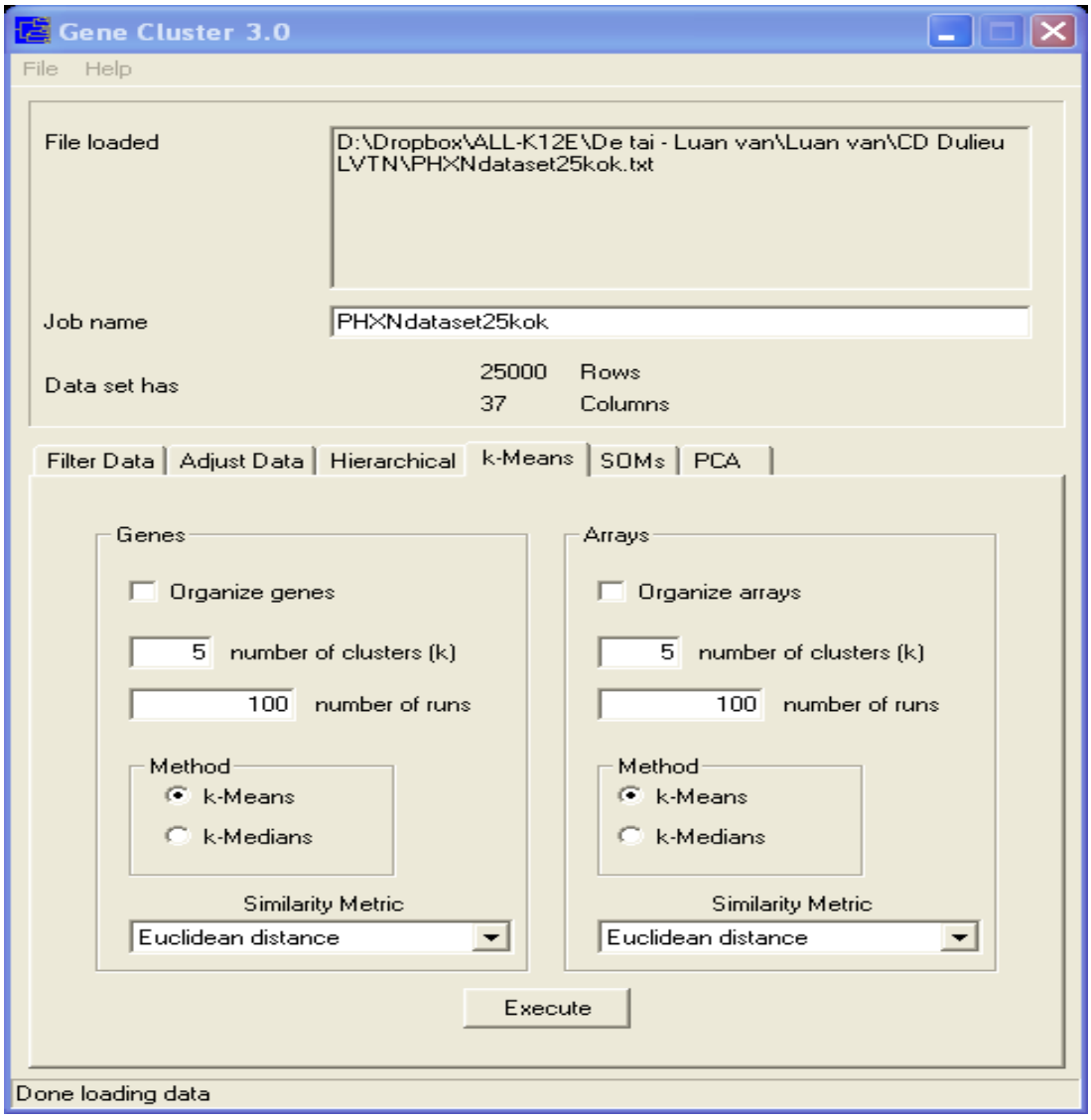
Hình 3.6: Trực quan kết quả sau khi phân cụm (k=5) với Weka Explorer



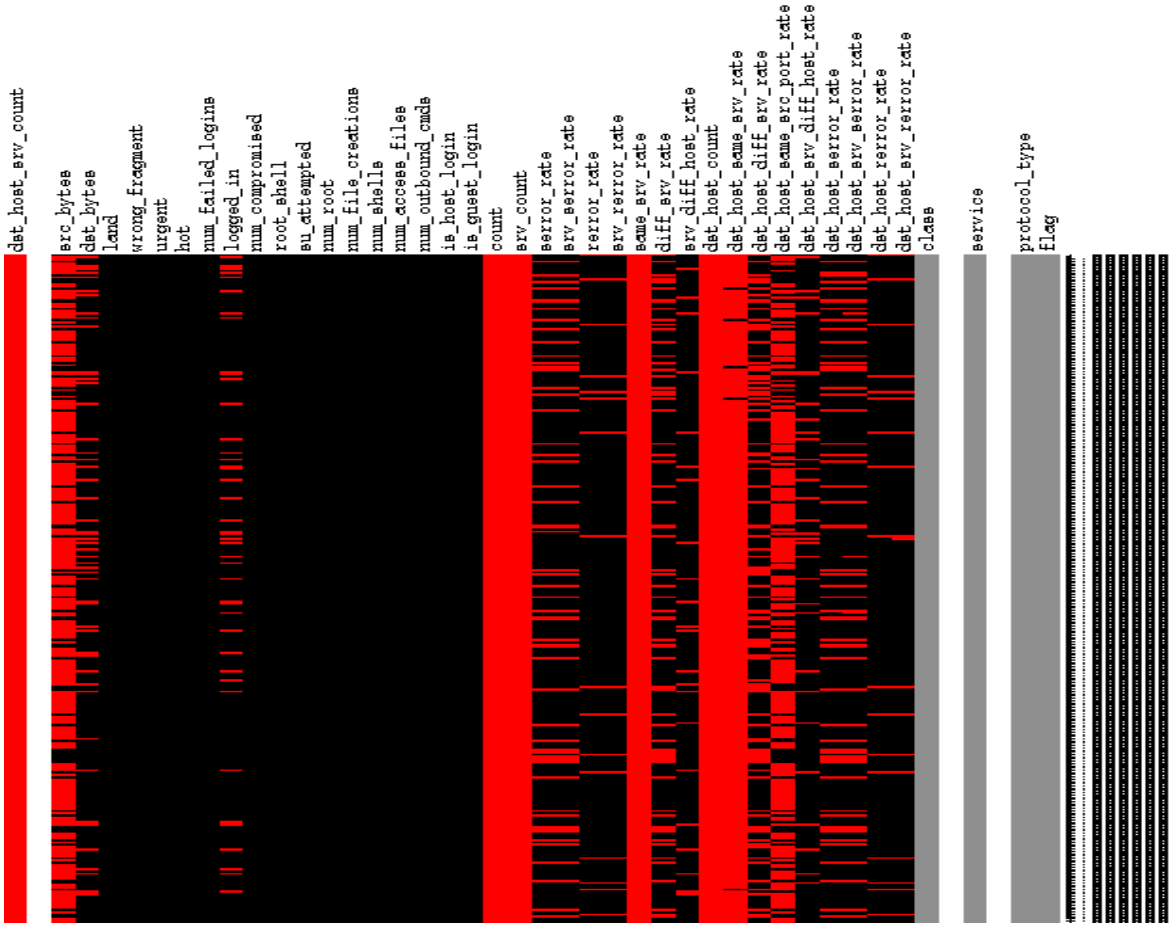
* Biểu diễn kết quả phân cụm theo Treeview: TreeView là phần mềm đọc các file có định dạng "* .CDT" và "* .GTR" được xuất ra bởi công cụ mã nguồn mở Cluster 3.0 [14]. Trước khi biểu diễn kết quả phân cụm theo TreeView, luận văn thực nghiệm tập dữ liệu ít hơn để xem xét tấn công qua Cluster 3.0 (hình 3.7).

Sau khi áp dụng phân cụm dữ liệuk=5 bằng cách sử dụng công cụ Cluster 3.0, kết quảsẽ được nhập vào chương trình TreeView [15][16] để hiển thị dữ liệu sau khi phân cụm (hình 3.8)

Hình 3.7: Phân cụm k-means trong Cluster 3.0



Hình 3.8: Mô hình đồ họa trực quan kết quả sau các kiểu tấn công



3.2.3.Phântíchvàđánh giá kết quả

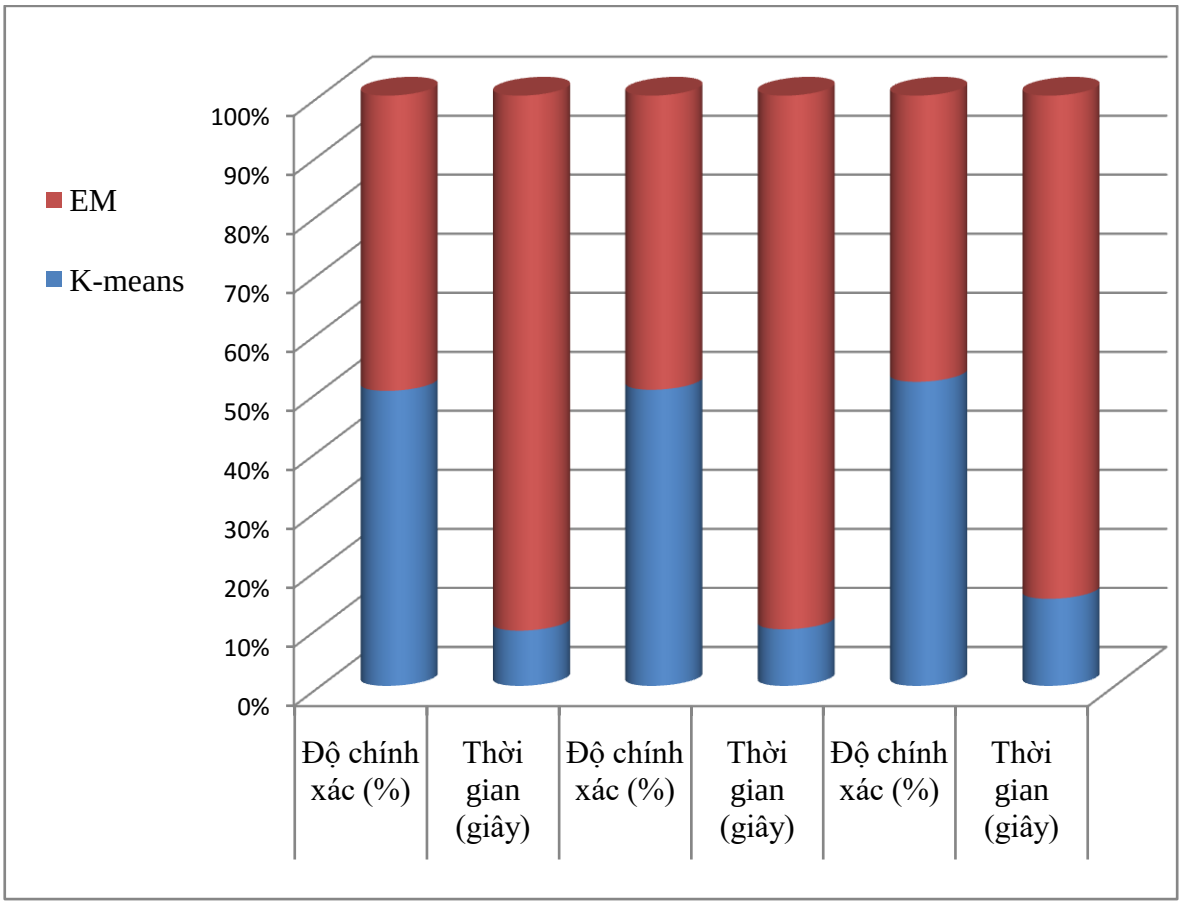
Sosánhmứcđộchínhxáccủacácbộphâncụmk-means, EM,cóthểnhậnthấy cácbộphâncụm EMchokếtquảtốtnhất về độ chính xác và thời gian huấn luyện lâu hơn so với phân cụm k-mean.Kếtquảsosánhđộchínhxác và thời gianđượcthểhiện thôngquabảng3.9 và hình 3.9.

Thuật toán	**Số cụm**					
	K =3		K =4		K=5	
	Độ chính xác (%)	Thời gian (giây)	Độ chính xác (%)	Thời gian (giây)	Độ chính xác (%)	Thời gian (giây)
K-means	98.07%	9.19	93.88%	10.02	94.03%	23.61

EM	98.13%	88.99	93.24%	94.55	88.49%	136.5

Bảng 3.9: Bảng so sánh kết quả phân cụm thuật toán K-means và EM

Hình 3.9: Biểu đồ so sánh kết quả phân cụm thuật toán K-means và EM



Với số cụm khác nhau thì các thuật toán cho kết quả vớiđộ chính xác và thời gian thực hiện khác nhau. Tuỳ từng hệ thống phát hiện xâm nhập trái phép mà ta sử dụng kết quả phân cụm cho mỗi hệ thống.

Thuật toán K-means cho thời gian chạy nhanh nhất tuy nhiên thuật toán EM lại cho độ chính xác tốt nhất đối với 03 cụm (k=03).

KẾT LUẬN

Qua quá trình nghiên cứu, thực hiện luận văn đã đạt được một số kết quả sau đây

- Luận văn đã trình bày tổng quan về tấn công mạng máy tính và các phương pháp phát hiện; trong đó nêu được các kỹ thuật tấn công mạng máy tính, các mô hình tấn công mạng, các kỹ thuật tấn công mạng, hệ thống phát hiện xâm nhập trái phép, các kỹ thuật phát hiện xâm nhập trái phép và ứng dụng kỹ thuật khai phá dữ liệu cho việc phát hiện xâm nhập trái phép… từ đó xác định và đưa ra phương án lựa chọn kỹ thuật phân cụm trong phát hiện xâm nhập trái phép.

- Trình bày chi tiết một số kỹ thuật phân cụm dữ liệu hiện nay như phân cụm phân hoạch (Partitioning Methods), phân cụm phân cấp (Hierarchical Methods), phân cụm dựa trên mật độ (Density-Based Methods), phân cụm dựa trên lưới (Grid-Based Methods), phân cụm dựa trên mô hình (Model-Based Clustering Methods), phân cụm dữ liệu mờ và đưa ra thuật toán cơ bản trong phân cụm dữ liệu.

- Luận văn thực hiện các thực nghiệm,ứng dụng thuật toán trong phân cụm dữ liệu đểxây dựng mô hình phát hiện xâm nhập trái phép với mức độ chính xác và thời gian thực hiện tối ưu nhất. Khai thác, ứng dụng thuật toán của phân cụm dữ liệu trong phần mềm Weka để tính toán, đưa ra được độ chính xác, thời gian thực hiện các loại tấn công.Ngoài ra, luận văn ứng dụng hiển thị kết quả qua chương trình Treeview với nguồn dữ liệu sau khi phân cụm là Cluster 3.0 để thấy được cụ thể kết quả phân cụm của các kiểu tấn công.

- Qua phân tích các kết quả thực nghiệm, luận văn đã lựa chọn được được kỹ thuật phân cụm EM đạt được độ chính xác tốt hơn so với thuật toán K-means.

Hướng phát triển:

Luận văn sẽ tiếp tục nghiên cứu một số ứng dụng của các thuật toán phân cụm và phát triển luận văn theo các hướng sau:

- Nghiên cứu thử nghiệm các thuật toán khai phá dữ liệu mới với tập dữ liệu lớn hơn, để đánh giá tìm ra các thuật toán tốt hơn.

- Xây dựng các hệ thống mạng mô phỏng để thử nghiệm các tấn công mới, nhằm thu thập các dấu hiệu tấn công phục vụ các nghiên cứu mới trong lĩnh này.

- Tích hợp các mô hình phân cụm, để xây dưng các hệ thống phát hiện xâm nhập trái phép, triển khai ứng dụng để đảm bảo an toàn cho các hệ thống mạng thực tế tại Việt Nam.

Trong quá trình hoàn thành đề tài này,mặc dù đã cố gắng, nỗ lực hết mình song do thời gian nghiên cứu, trình độ của bản thân có hạn và điều kiện nghiên cứu còn nhiều khó khăn nên không thểtránh khỏi những khuyết thiếu và hạn chế, bản thân rất mong nhận được những góp ý, nhận xét quý báu của quý thầy cô và bạn bè để kết quả của đề tài hoànthiện hơn.

TÀI LIỆU THAM KHẢO

Tài liệu tiếng Việt

[1] Nguyễn Hà Nam, Nguyễn Trí Thành, Hà Quang Thụy,*Khai phá dữ liệu*,NXB Đại học Quốc gia Hà Nội, 2013.

Tài liệu tiếng Anh

[2] George Danezis, *Designing and attacking anonymous communication systems*, July 2014, Cambridge.

[3] R.J Anderson, Security Enginnering – *A Guide to Building Dependable Distributed Systems*, Wiley 2001.

[4] Jiawei Han and Micheline Kamber, *Data Mining Concepts and Techniques,* Chapter 1 & Chapter 8 (Intelligent Database Systems Research Lab School of Computing Science Simon Fraser University, Canada), 2007.

[5] BS Everitt, *Cluster Analysis*, Edward Amold coblished by Haisted Press and imprint of john Wiley & Sons Inc, 3rd edition, 1993.

[6] Michael RAnderber, *Cluster analysis of application*, Academic Press, Inc, New York, 1973.

[7] S.Jain , M. Aalam , M.Doja , " K-means clustering using weka interface", Proceedings of the 4th National Conference; INDIACom, Computing For Nation Development, 2010.

[8] Daniel Barbara, Julia Couto, Sushil Jajodia, and Ningning Wu,*Adam: a testbed for exploring the use of data mining in intrusion detection*, ACM SIGMOD Record, volume 30, December 2001.

[9] Irvine, *KDD Cup Data*, October 29, 1999, http://kdd.ics.uci.edu/ databases/kddcup99/kddcup99.html

[10] Remco R.Bouckaert, *DocumentationWeka*, The University of Waikato, July 14, 2008.

[11] Mrs. Ghatge Dipali D, *Network Traffic Intrusion Detection System using Decision Tree & K-Means Clustering Algorithm*, International Journal of Emerging Trends & Technology in Computer Science, Volume 2, Issue 5, September – October 2013.

[12] Richa, Saurabh Mittal, *Data Mining Approach IDS K-Mean using Weka Environment*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, issure 8, August 2014.

[13]P. Divya, R. Priya, *Clustering Based Feature Selection and Outlier Analysis*, International Journal of Computer Science & Communication Networks, Vol.2 (6), p647-652.

[14] Michel de Hoon, *Clustering 3.0for Windows, Mac OS X, Linux, Unix*, Human Goneme Center, University of Tokyo, November 5, 2002.

[15] AJ Saldanha, *Java TreeView User's Manual*, National Center for Biotechnology Information, The United States National Library of Medicine, 2004.

[16] A.M.Riad, Ibrahim Elhenawy, Ahmed Hassan and Nancy Awadallah: *Visualize network anomaly Detection by using k-means clustering algorithm*, international Journal of Computer Network & Communications, Vol.5, No.5, September 2013.