

ĐỀ TÀI KC.01.15/06-10

Nhận dạng tấn công mạng với mô hình trực quan cây quyết định

Đỗ Thanh Nghị, Lê Quyết Thắng

Khoa Công nghệ thông tin, ĐHCT
Số 1 Lý Tự Trọng, Ninh Kiều, Cần Thơ
dtngghi@cit.ctu.edu.vn

Ngày 13 tháng 3 năm 2011



Giới thiệu

Cây quyết định cho phát hiện tấn công mạng

Kết quả thực nghiệm

Kết luận và hướng phát triển



Nội dung công việc thực hiện

Thành phần an toàn hệ thống

- ▶ nghiên cứu phát hiện tấn công mạng
- ▶ xây dựng dịch vụ web hỗ trợ phát hiện tấn công
- ▶ phát hiện tấn công mạng với mô hình cây quyết định
- ▶ đơn giản, hiệu quả (chính xác cao)
- ▶ trích luật dự báo => cấu hình hệ thống SNORT hay OSSEC



Phát hiện tấn công mạng

Nghiên cứu liên quan

- ▶ kỳ thi KDDCup 99
- ▶ tài trợ của DARPA, AFRL và MIT Lincoln Lab.
- ▶ tập dữ liệu khoảng 5 triệu nối kết, 41 thuộc tính, học để phát hiện tấn công mạng (từ chối dịch vụ, etc.)
- ▶ tiếp cận máy học: cây quyết định, máy học vectơ hỗ trợ, boosting, rừng ngẫu nhiên, mạng nơron, etc.
- ▶ thắng cuộc: Bagged boosting (rừng cây quyết định)



Phát hiện tấn công mạng

Tại sao KDDCup 99 khó?

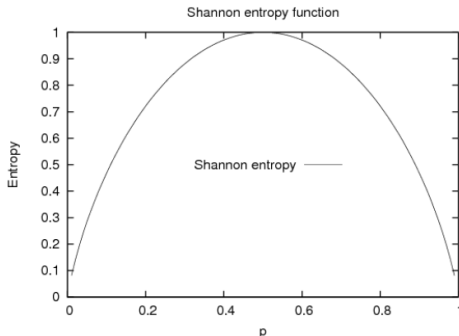
- ▶ dữ liệu không cân bằng: top 10 vấn đề khó
- ▶ các giải thuật được thiết kế: độ chính xác toàn cục
- ▶ dự báo không tốt cho lớp nhỏ
- ▶ lớp quan tâm: lớp nhỏ (tấn công)
- ▶ giá phải trả cho dự báo sai lớp nhỏ là cao



Cây quyết định cho phát hiện tấn công mạng

Giải thuật cây quyết định C4.5 (Quinlan, 1993)

- xây dựng cây: hàm phân hoạch sử dụng Shannon entropy ¹



¹không tính đến sự mất cân bằng về lớp (luôn cực đại tại $p=0.5$)

Cây quyết định cho phát hiện tấn công mạng

Cải tiến giải thuật cây quyết định C4.5

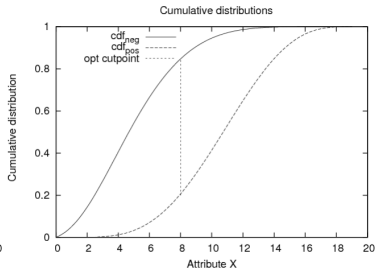
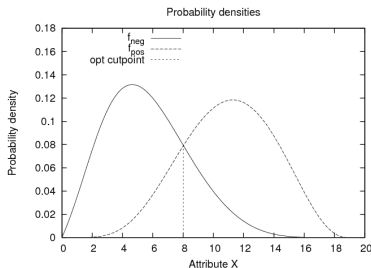
- ▶ thay đổi hàm phân hoạch dữ liệu ²
- ▶ với **Kolmogorov-Smirnov**
- ▶ cải thiện dự báo lớp nhỏ
- ▶ không làm mất đáng kể dự báo lớp lớn
- ▶ **tốt hơn cả người thắng cuộc KDDCup 99**
- ▶ trích luật dự báo => cấu hình hệ thống SNORT hay OSSEC

²C4.5 sử dụng Shannon entropy, không tính đến sự mất cân bằng về lớp



Hàm phân hoạch cho dữ liệu không cân bằng

Kolmogorov-Smirnov



Ví dụ với hàm phân hoạch dựa trên Kolmogorov-Smirnov

Bảng: Tập dữ liệu không cân bằng

Attribute X	1	2	3	4	5	6	7	8	9	10
#ind. of positive (minority)	2	2	2	2	1	1	0	0	0	0
#ind. of negative (majority)	0	0	0	10	0	20	40	10	10	10

Bảng: Phân hoạch³ với Kolmogorov-Smirnov

Cutpoint ($\alpha = 5.5$) on attribute X	Left partition	Right partition
#ind. of positive (minority)	9	1
#ind. of negative (majority)	10	90

³Shannon entropy phân hoạch tại $\alpha = 3.5$



Thiết lập thực nghiệm

Chương trình, giải thuật

- ▶ cây quyết định C4.5 sử dụng Shannon entropy (C4.5-SE)
- ▶ C4.5 sử dụng Kolmogorov-Smirnov (C4.5-KS)
- ▶ các giải thuật khác từ thư viện Weka (Witten & Frank, 2000):
rừng ngẫu nhiên (Random forest) của (Breiman, 2001),
Bayes thơ ngây (Good, 1965)
- ▶ Bagged boosting của (Pfahringer, 1999): thắng cuộc
KDDCup 99



Thiết lập thực nghiệm

Dữ liệu

- ▶ tập dữ liệu, nghi thức: KDDCup 99
- ▶ tập học: 494 021 nối kết
- ▶ tập kiểm tra: 311 029 nối kết
- ▶ 41 thuộc tính: tcpdump
- ▶ 5 lớp: từ chối dịch vụ (DoS), dò tìm (Probe), etc.



Thiết lập thực nghiệm

Bảng: Mô tả tập dữ liệu KDDCup 99

STT	Lớp	Tập học	Tập kiểm tra
0	Normal	97278	60593
1	Probe	4107	4166
2	DoS	391458	229853
3	U2R	52	228
4	R2L	1126	16189



Thiết lập thực nghiệm

So sánh độ chính xác toàn cục, từng lớp và Cost

- M_{ij} : số nối kết lớp i bị dự báo là lớp j , có giá tương ứng C_{ij} và N : tổng số nối kết. $Cost = \frac{1}{N} \sum [M_{ij} C_{ij}]$

Bảng: Ma trận chi phí

Dự báo như $\Rightarrow$	0 (Normal)	1 (Probe)	2 (DoS)	3 (U2R)	4 (R2L)
0 (Normal)	0	1	2	2	2
1 (Probe)	1	0	2	2	2
2 (DoS)	2	1	0	2	2
3 (U2R)	3	2	2	0	2
4 (R2L)	4	2	2	2	0



So sánh kết quả

	C4.5-SE (%)	C4.5-KS (%)	KDDCup99 winning (%)	Random Forests (%)	Naive Bayes (%)
0:Normal	<u>99.20</u>	98.14	99.45	98.13	98.90
1:Probe	<u>94.46</u>	94.55	83.32	78.78	83.61
2:DoS	<u>97.22</u>	97.01	97.12	97.35	94.98
3:U2R	8.33	7.90	<u>13.16</u>	11.40	20.18
4:R2L	4.05	20.20	8.40	<u>11.73</u>	10.53
Overall	92.65	93.13	92.71	<u>92.73</u>	91.14
Cost	0.2414	0.2172	0.2331	<u>0.2180</u>	0.2546

C4.5 sử dụng Kolmogorov-Smirnov

- ▶ cải thiện dự báo lớp nhỏ, ảnh hưởng ít dự báo lớp lớn => **độ chính xác cao, cost thấp nhất**
- ▶ mô hình một cây quyết định đơn giản cho kết quả **tốt hơn cả người thắng cuộc KDDCup 99, rừng cây phức tạp**
- ▶ trích luật dự báo => cấu hình hệ thống SNORT hay OSSEC



Mô hình trực quan cây quyết định



[Mở cây]

[Đóng cây]

[Dự báo]

Sử dụng:

Để nút trích luật quyết định từ cây, chỉ cần nhấp chuột vào từng nút của cây. Nếu là nút lá thì ta có một luật quyết định là đường dẫn từ nút gốc (Root) đến nút lá. Nếu là nút trong ta có tập luật tương ứng với các nút lá của nút trong. Doubleclick menu Dự báo cho phép nhập dữ liệu cho dự báo.

1. NẾU

(count > 64) VÀ
(diff_srv_rate <= 0.23) VÀ
(dst_bytes <= 1) VÀ
(protocol_type = icmp)

THÌ

DoS (280711/1)

<= Tập luật



Kết luận

Nhận dạng tấn công mạng

- ▶ cải tiến giải thuật cây quyết định
- ▶ phân hoạch dữ liệu với Kolmogorov-Smirnov
- ▶ mô hình đơn giản nhưng cho kết quả **tốt hơn cả người thắng cuộc KDDCup 99**
- ▶ rút trích luật quyết định tương ứng dễ hiểu
- ▶ dự báo trực tuyến
- ▶ truy cập: **<http://www.phongchongdichhai.org.vn/sec>**



Hướng phát triển²

Nhận dạng tấn công mạng

- ▶ trích luật dự báo => cấu hình hệ thống SNORT hay OSSEC

