Tiếp cận nội dung để phát hiện thư rác
Mô hình túi từ và giải thuật học Boosting-RODS

Đỗ Thanh Nghị

BM. Khoa học máy tính
Khoa Công nghệ thông tin
Số 1 Lý Tự Trọng, Ninh Kiều, Cần Thơ
dtnghi@cit.ctu.edu.vn

Ngày 28 tháng 9 năm 2011

Nội dung

Dịch vụ liên lạc phổ biến

Dịch vụ thư điện tử

- ưu điểm: đơn giản, nhanh chóng, chi phí thấp
- nhiều người sử dụng

Vấn nạn thư rác

- quảng cáo, khiêu dâm, phản động, thậm chí là những đoạn mã độc hại đính kèm
- gây lãng phí, phiền toái

Ví dụ minh họa

5-12 Free Bonus Viagra100mg pills with every order. All Major Credit Cards Accepted. Cialis20mg $2.97, Viagra100mg $1.87 37w Spam

Tawana Celestine <tawana_celestinebj@charleskendall.com> to show details Aug 17 Reply

Foreign Pharmacy | Discount & Cheap Drugstore

An Online Pharmacy provides Cheap drugs at your doorstep

CialisPlusViagra Powerpack special price
ViagraAs low as $1.85
ViagraPlus as low as $2.95
ViagraProfessional as low as $3.95
ViagraSuperForce as low as $4.99
CialisAs low as $2.40
CialisSuperActive+ as low as $3.33
LevitraAs low as $2.50
SomaAs low as $1.06

Order with us and save your medical bills up to 80-90%.

We have worldwide customers ...

Thiệt hại

Thống kê về thiệt hại do thư rác

- nghiên cứu cho thấy tổn thất năm 2002 ước tính khoảng 2 tỷ đô la (Sung-jin, 2003)
- một nghiên cứu khác về tổn thất tháng 10 năm 2003 ước tính khoảng 10,4 tỷ đô la (Mi2g, 2003)
- theo (Doug, 2003): nếu một người phát tán thư rác thu được lợi 10 ngàn đô la trong 1 tháng thì tổn thất do họ gây ra là 100 ngàn đô la.

Lọc thư rác

- giảm bớt tổn thất

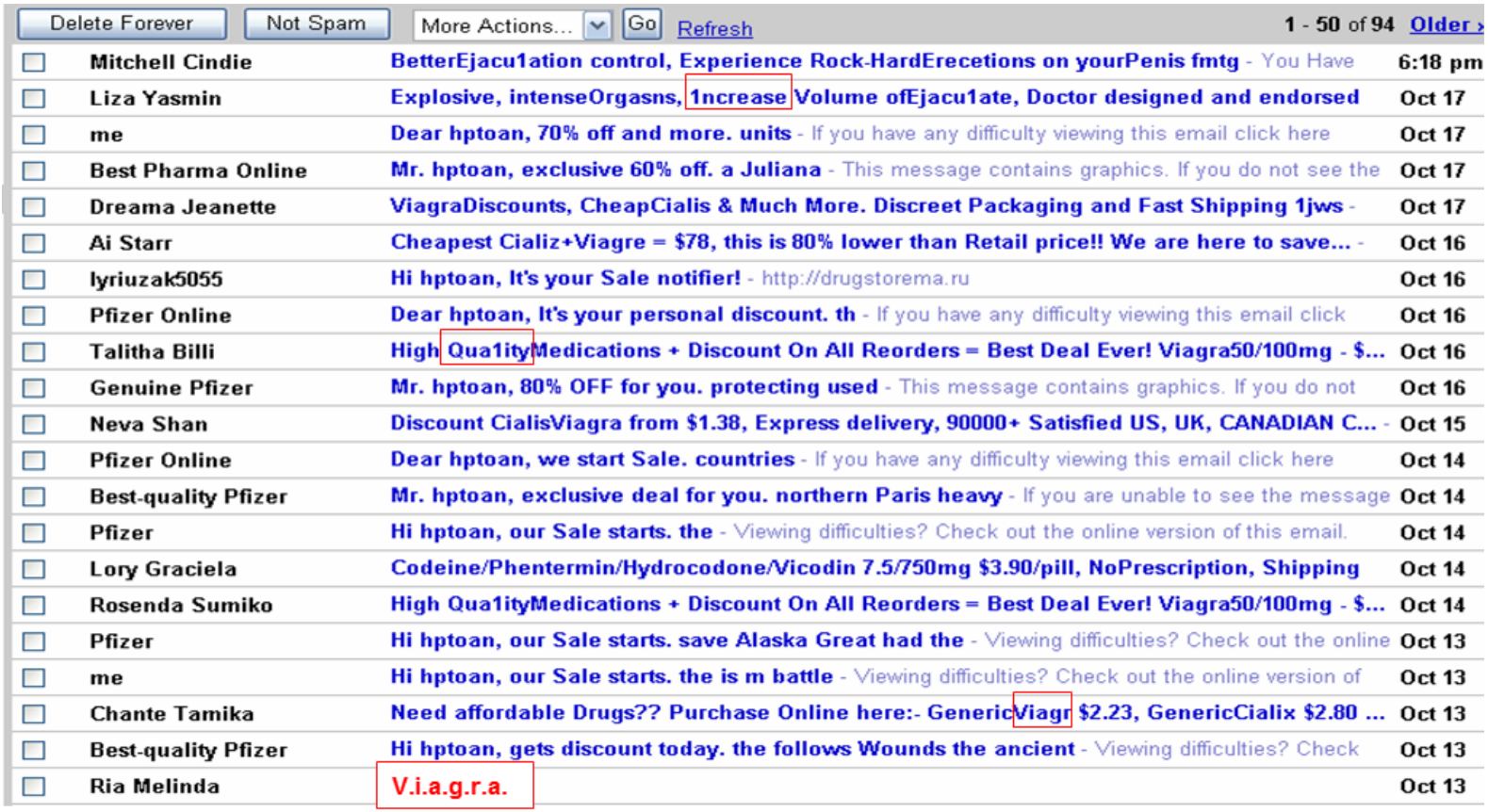
Lọc thư rác dựa trên địa chỉ người gửi

Dễ bị qua mặt: sinh ngẫu nhiên địa chỉ người gửi

S: 220 crepes.fr
C: **HELO** cit.ctu.edu.vn
S: 250 Hello cit.ctu.edu.vn, pleased to met you
C: **MAIL FROM**: <alice@cit.ctu.edu.vn>
S: 250 alice@cit.ctu.edu.vn … Sender ok
C: **RCPT TO**: <peter@crepes.fr>
S: 250 peter@crepes.fr … Recipient ok
C: **DATA**
S: 354 Enter mail, end with "." On a line by itself
C: How are you?
C: Do you like internet?
C: .
S: 250 Message accepted for delivery
C: **QUIT**

Lọc thư rác dựa trên từ khóa

Dễ bị qua mặt: biến thể



Delete Forever | Not Spam | More Actions... | Go | Refresh | 1 - 50 of 94 | Older »

Mitchell Cindie	BetterEjacu1ation control, Experience Rock-HardErecetions on yourPenis fmtg - You Have	6:18 pm
Liza Yasmin	Explosive, intenseOrgasns, 1ncrease Volume ofEjacu1ate, Doctor designed and endorsed	Oct 17
me	Dear hptoan, 70% off and more. units - If you have any difficulty viewing this email click here	Oct 17
Best Pharma Online	Mr. hptoan, exclusive 60% off. a Juliana - This message contains graphics. If you do not see the	Oct 17
Dreama Jeanette	ViagraDiscounts, CheapCialis & Much More. Discreet Packaging and Fast Shipping 1jws -	Oct 17
Ai Starr	Cheapest Cializ+Viagre = $78, this is 80% lower than Retail price!! We are here to save... -	Oct 16
lyriuzak5055	Hi hptoan, It's your Sale notifier! - http://drugstorema.ru	Oct 16
Pfizer Online	Dear hptoan, It's your personal discount. th - If you have any difficulty viewing this email click	Oct 16
Talitha Billi	High Qua1ityMedications + Discount On All Reorders = Best Deal Ever! Viagra50/100mg - $...	Oct 16
Genuine Pfizer	Mr. hptoan, 80% OFF for you. protecting used - This message contains graphics. If you do not	Oct 16
Neva Shan	Discount CialisViagra from $1.38, Express delivery, 90000+ Satisfied US, UK, CANADIAN C... -	Oct 15
Pfizer Online	Dear hptoan, we start Sale. countries - If you have any difficulty viewing this email click here	Oct 14
Best-quality Pfizer	Mr. hptoan, exclusive deal for you. northern Paris heavy - If you are unable to see the message	Oct 14
Pfizer	Hi hptoan, our Sale starts. the - Viewing difficulties? Check out the online version of this email.	Oct 14
Lory Graciela	Codeine/Phentermin/Hydrocodone/Vicodin 7.5/750mg $3.90/pill, NoPrescription, Shipping	Oct 14
Rosenda Sumiko	High Qua1ityMedications + Discount On All Reorders = Best Deal Ever! Viagra50/100mg - $...	Oct 14
Pfizer	Hi hptoan, our Sale starts. save Alaska Great had the - Viewing difficulties? Check out the online	Oct 13
me	Hi hptoan, our Sale starts. the is m battle - Viewing difficulties? Check out the online version of	Oct 13
Chante Tamika	Need affordable Drugs?? Purchase Online here:- GenericViagr $2.23, GenericCialix $2.80 ...	Oct 13
Best-quality Pfizer	Hi hptoan, gets discount today. the follows Wounds the ancient - Viewing difficulties? Check	Oct 13
Ria Melinda	V.i.a.g.r.a.	Oct 13

Lọc thư rác dựa trên nội dung

Mô hình túi từ và học máy để phát hiện thư rác

- nội dung thư (không có cấu trúc): biểu diễn về cấu trúc bảng
- mô hình túi từ: thư điện tử biểu diễn dạng véctơ có giá trị thành phần thứ **i** là tần số xuất hiện từ thứ **i** trong thư
- bỏ qua các từ không chứa nhiều thông tin để nhận dạng thư rác, quy về từ gốc
- tập thư điện tử: bảng (ma trận), mỗi dòng là một thư, mỗi cột tương ứng với một từ trong tự điển => số cột (chiều) rất lớn đến vài chục ngàn
- học máy: xây dựng mô hình dự báo thư rác

Lọc thư rác dựa trên nội dung

Mô hình túi từ và học máy để phát hiện thư rác

- đề xuất giải thuật học Boosting of Random Oblique Stump (Do et al., 2009)
- đơn giản, nhanh, cho độ chính xác cao
- kết quả thử nghiệm với 1921 thư điện tử: giải thuật của chúng tôi đạt được độ chính xác đến 97.45%
- tốt hơn giải thuật Naive Bayes (Good, 1965), C4.5, AdaBoost-C4.5 (Quinlan, 1993, 1999), SVM (Vapnik, 1995)

Các giải thuật học hiện có

Phân lớp với các giải thuật

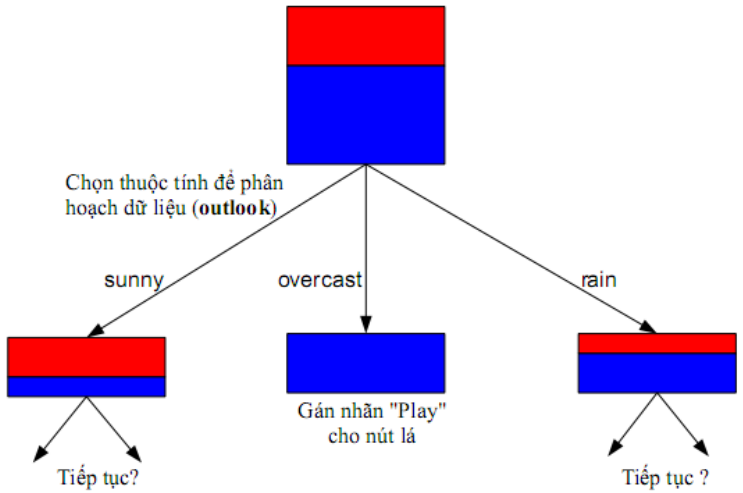
- cây quyết định C4.5 (Quinlan, 1993)
- AdaBoost-C4.5 (Quinlan, 1999)
- Naive Bayes (Good, 1965)
- máy học véctơ hỗ trợ - SVM (Vapnik, 1995)

Cây quyết định C4.5

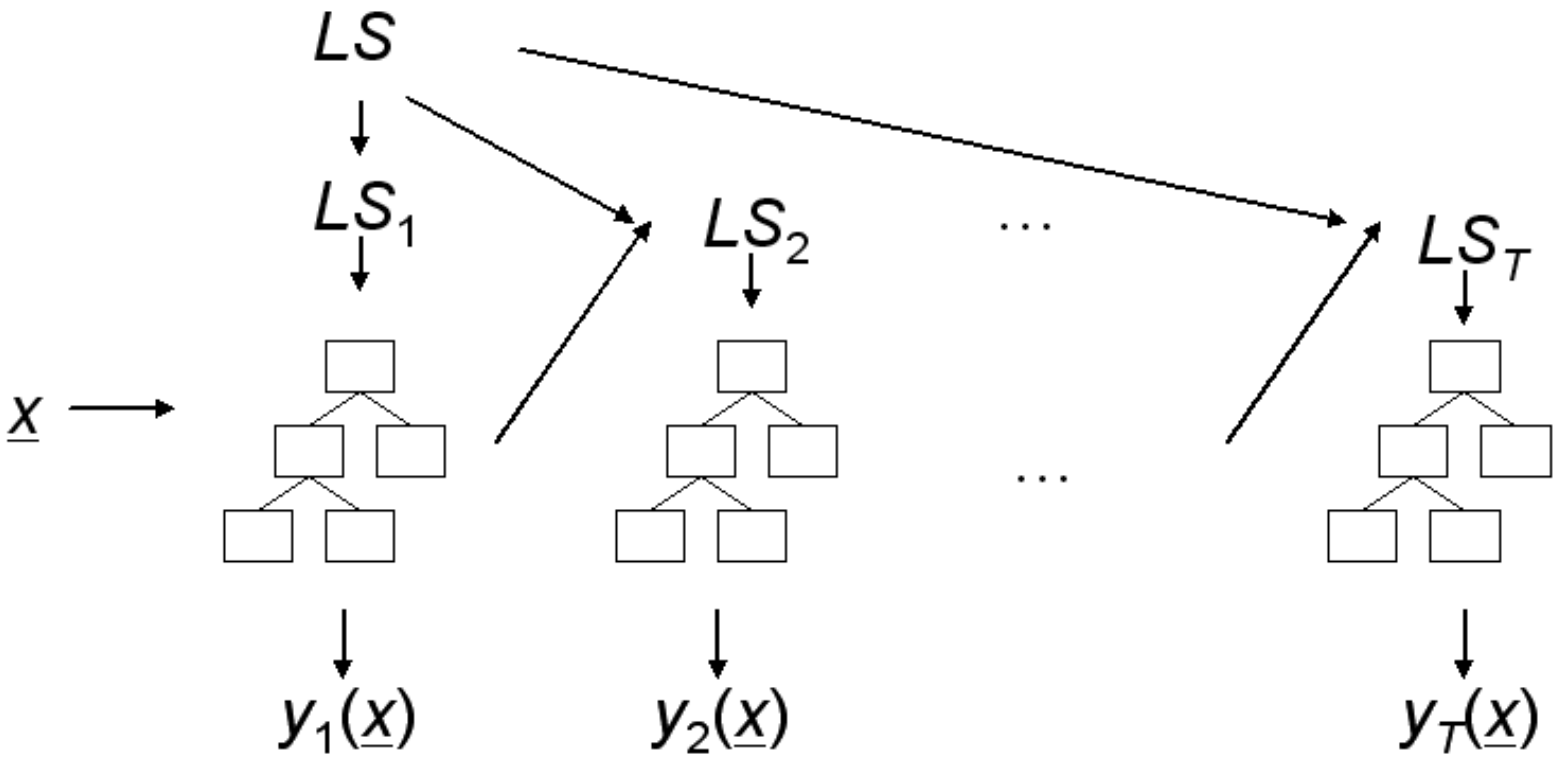
Giải thuật học

- xây dựng cây: hàm phân hoạch
- cắt nhánh: tránh học vẹt

outlook	temp.	hum.	windy	Play, Don't Play
sunny	85	85	false	Don't Play
sunny	80	90	true	Don't Play
overcast	83	78	false	Play
rain	70	96	false	Play
rain	68	80	false	Play
rain	65	70	true	Don't Play
...	...	...	...	...
...	...	...	...	...



AdaBoost-C4.5



Giải thuật Naive Bayes

Mô hình dự báo Bayes dựa trên định lý xác suất Bayes như sau:

$$\Pr[H \mid E] = \frac{\Pr[E \mid H]\Pr[H]}{\Pr[E]} \quad (3.1)$$

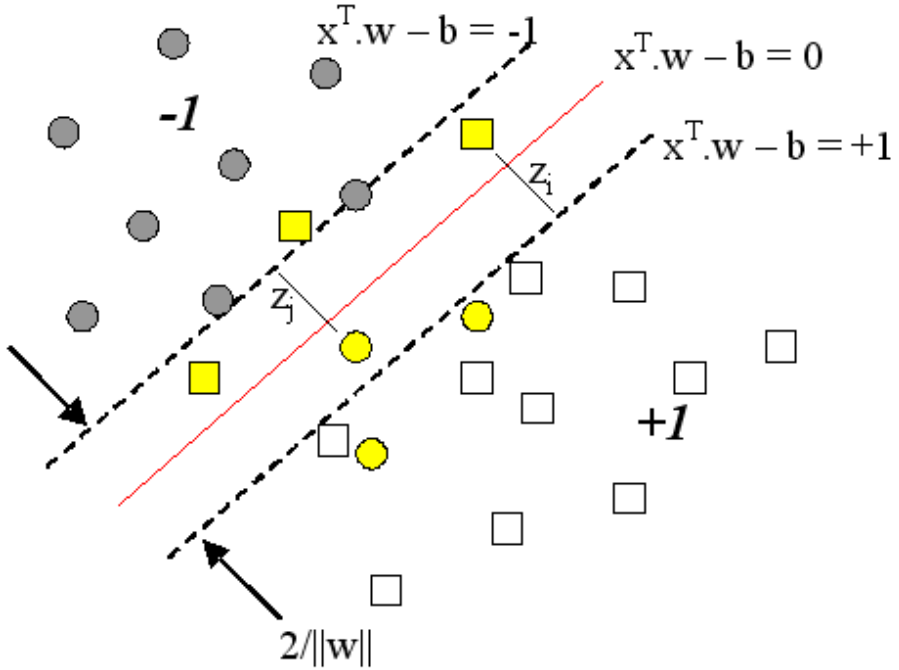
trong đó $\boldsymbol{E}$ được biết như là giá trị dữ liệu cần dự báo và $\boldsymbol{H}$ chính là lớp (nhãn) dự báo.

Với giả thiết là các thuộc tính độc lập nhau, nên cách dự báo trong mô hình Bayes thơ ngây được tính như 3.2.

$$\Pr[H \mid E] = \frac{\Pr[E_1 \mid H]\Pr[E_2 \mid H]\ldots\Pr[E_n \mid H]\Pr[H]}{\Pr[E]} \quad (3.2)$$

Máy học véctơ hỗ trợ - SVM

SVM tuyến tính

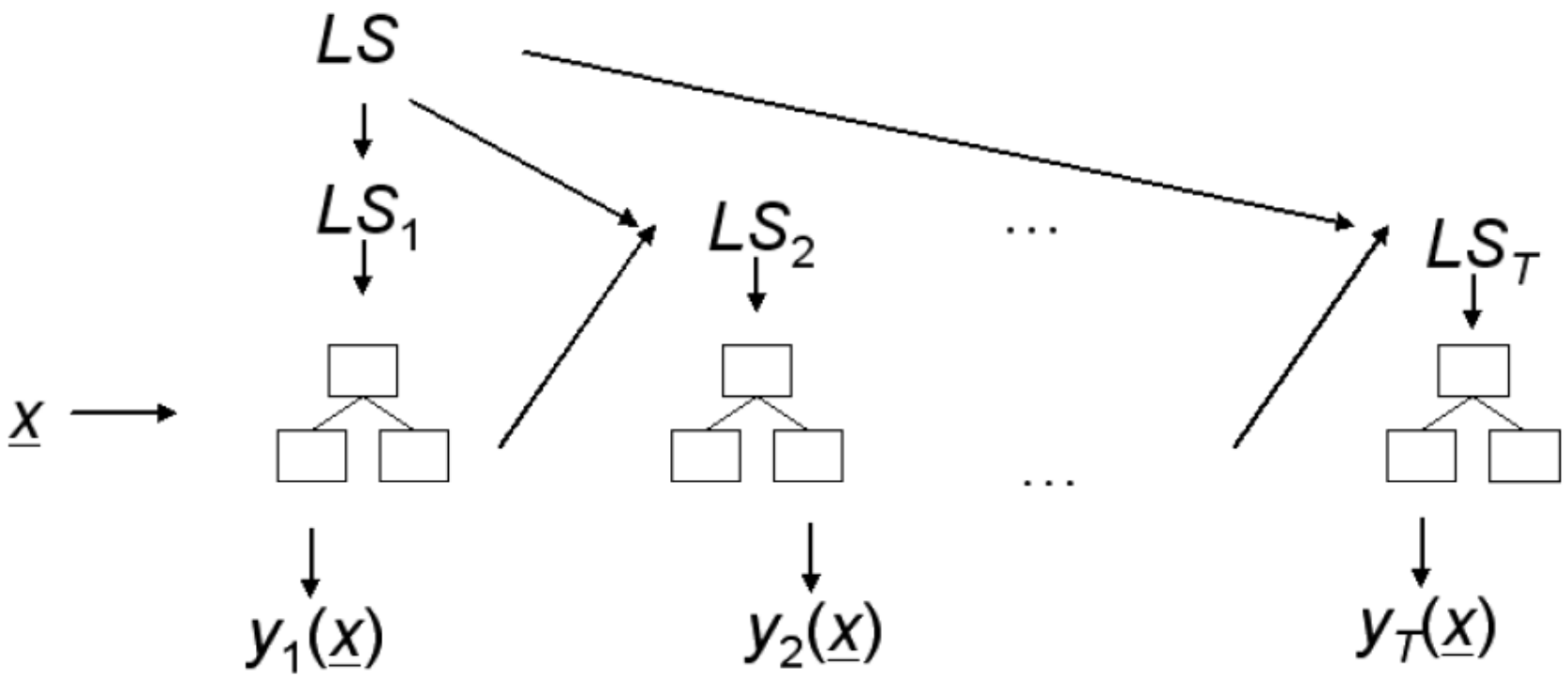


Boosting of Random Oblique Stump

Giải thuật

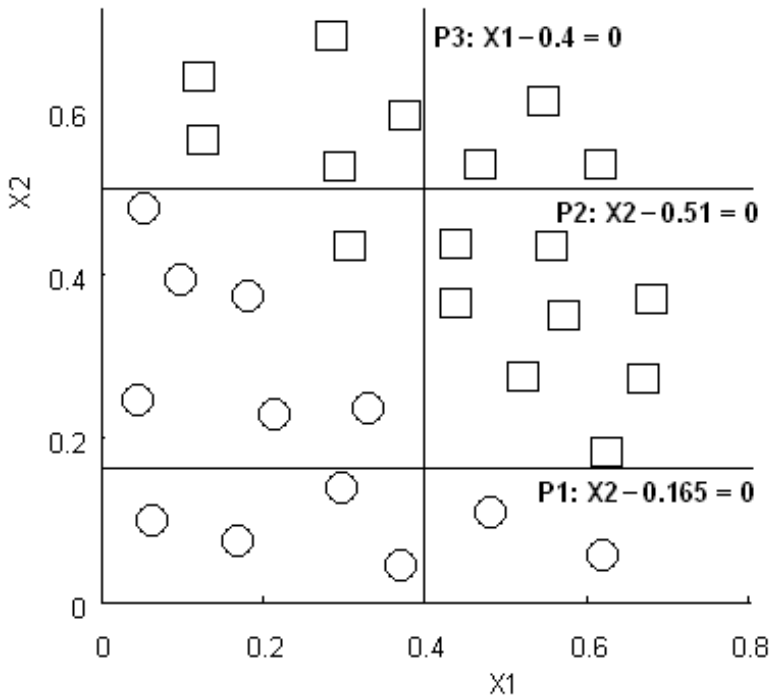
- xây dựng tuần tự tập cây xiên phân ngẫu nhiên đơn giản
- cây chỉ có 3 nút (1 nút gốc và 2 nút lá)
- nút gốc: siêu phẳng phân hoạch tối ưu (SVM)
- xây dựng cây tập trung vào khắc phục lỗi từ các mô hình xây dựng trước đó
- nhanh: cây đơn giản có 3 nút
- hiệu quả: phân hoạch đa thuộc tính + boosting

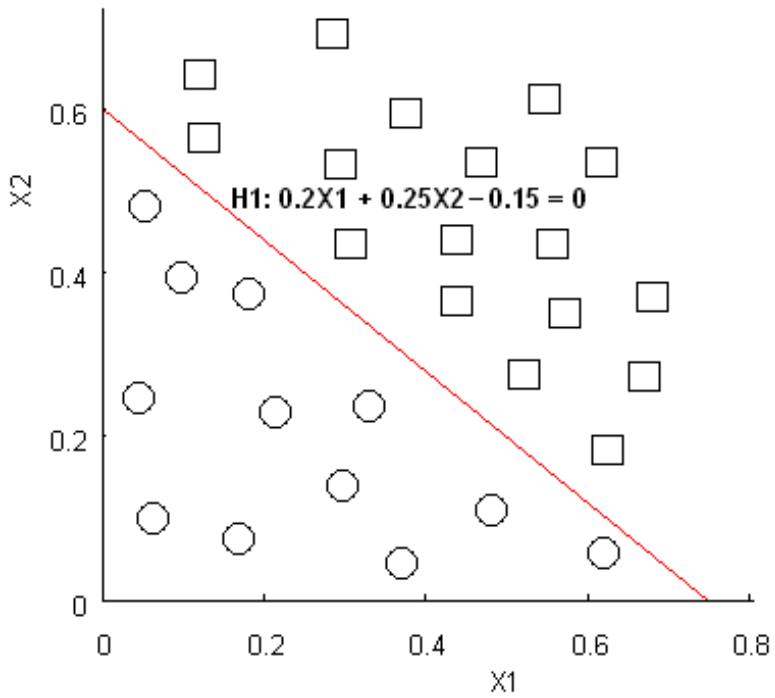
Boosting of Random Oblique Stump



Phân hoạch xiên của Random Oblique Stump

Kết hợp nhiều chiều xử lý dữ liệu có số chiều lớn, phụ thuộc





Chuẩn bị dữ liệu

Tạo dữ liệu

- thu thập 1921 thư (1143 thư rác và 778 không phải thư rác)
- tiền xử lý với BoW (McCallum, 1998): bỏ qua các từ không chứa nhiều thông tin để nhận dạng thư rác, quy về từ gốc
- mô hình túi từ: bảng dữ liệu, 1921 phần tử (thư), 28719 thuộc tính (từ) và 2 lớp (thư rác hay không phải thư rác)
- nghi thức kiểm tra chéo 10-fold

Chuẩn bị chương trình

Chương trình

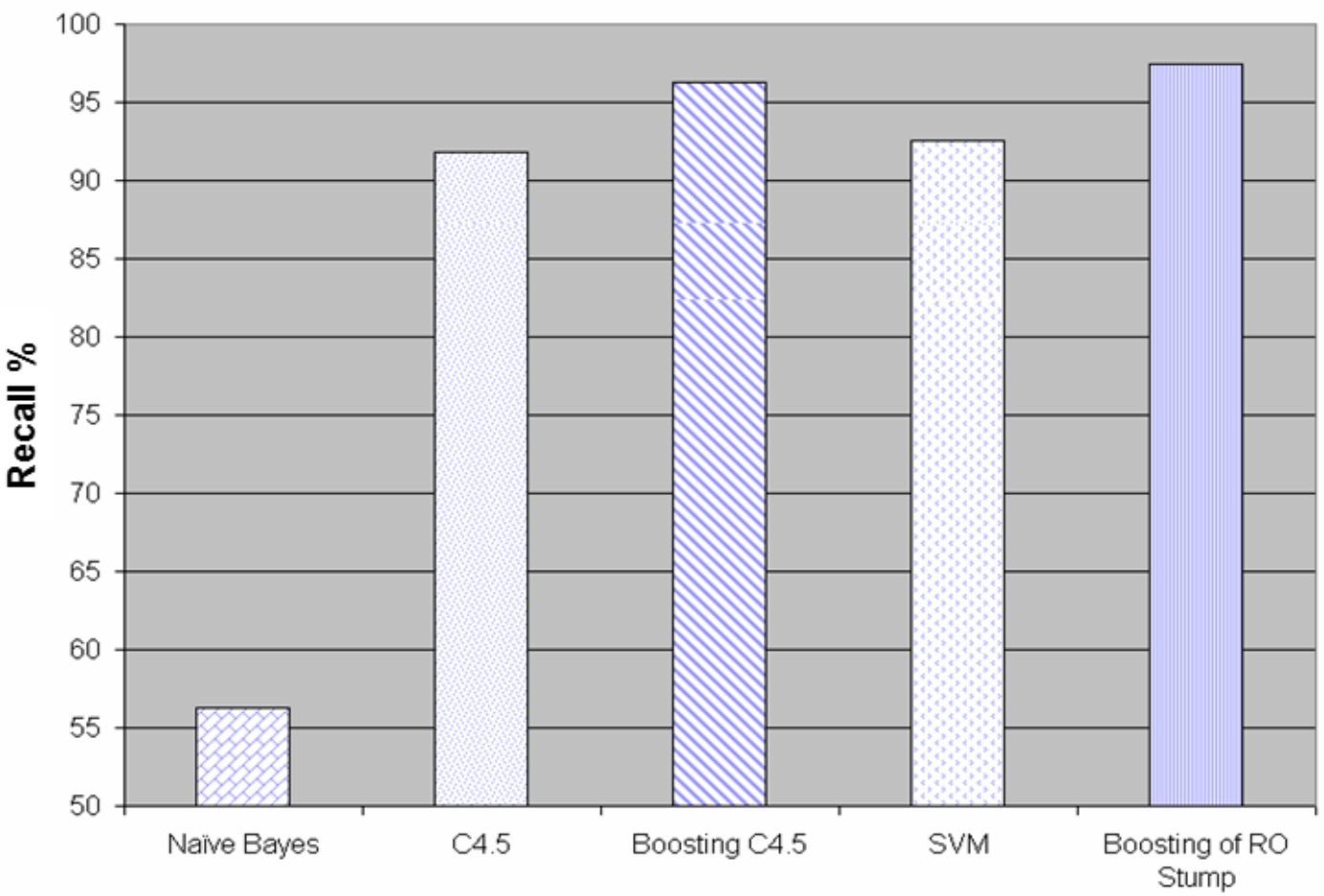
- giải thuật Boosting-RODS: C/C++, ATLAS/LAPACK
- giải thuật khác như C4.5, AdaBoost-C4.5, Naive Bayes: Weka (Witten & Frank, 2005)
- máy học SVM: LibSVM (Chang & Lin, 2001)

Tiêu chí đánh giá

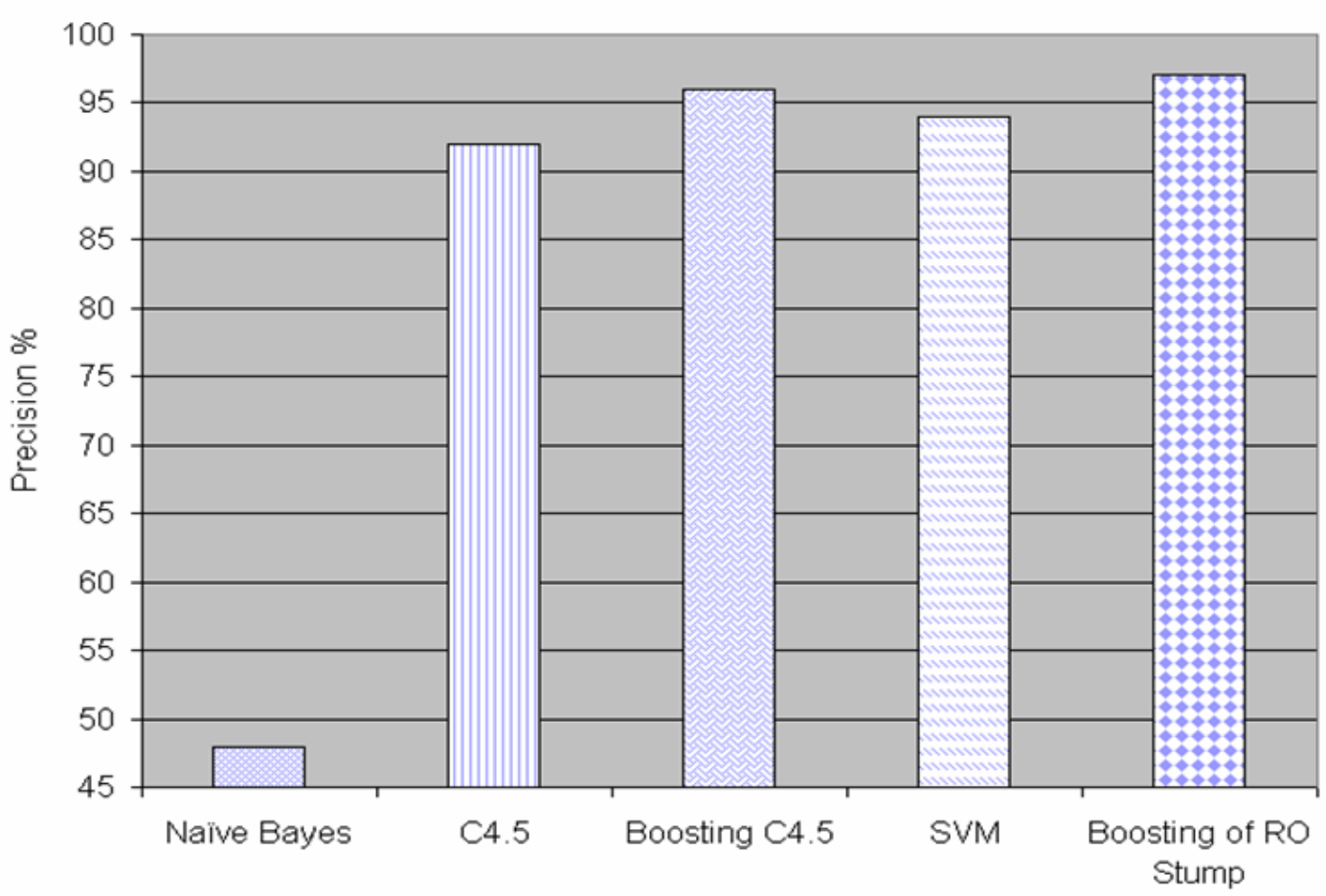
Recall, Precision, F1

- Recall: số thư rác được phát hiện đúng là thư rác chia cho tổng số thư rác
- Precision: số thư rác được phát hiện đúng là thư rác chia cho tổng số thư được dự báo là thư rác
- F1: trung bình điều hòa giữa Precision và Recall

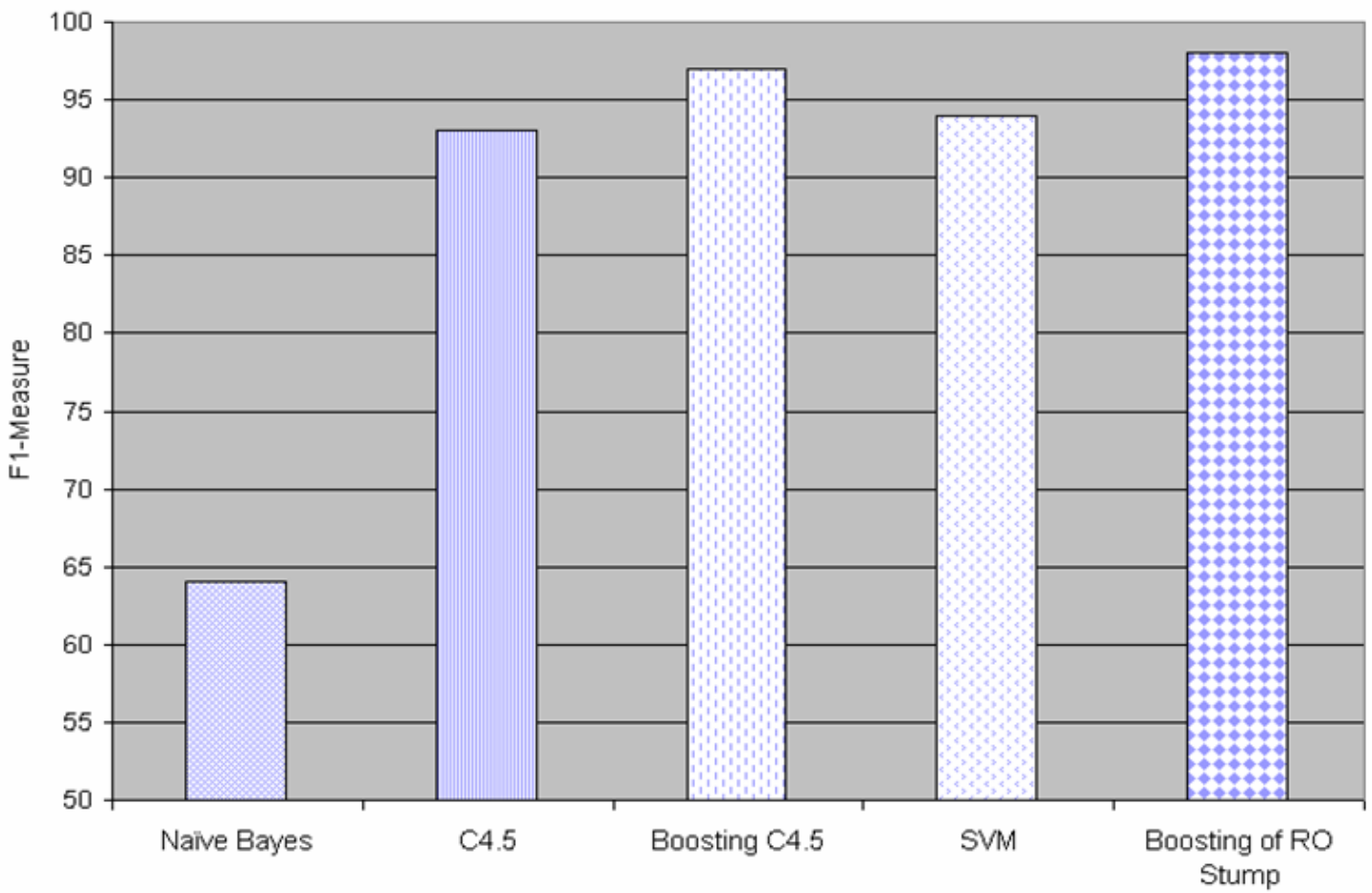
So sánh kết quả theo tiêu chí Recall



So sánh kết quả theo tiêu chí Precision



So sánh kết quả theo tiêu chí F1



Kết luận

Phát hiện thư rác

- tiếp cận: học từ nội dung
- mô hình túi từ: số chiều dữ liệu lớn
- phân lớp hiệu quả: Boosting of Random Oblique Stump
- phân hoạch đơn thuộc tính => phân hoạch xiên
- cây xiên phân đơn giản 3 nút: nhanh
- chính xác hơn C4.5, AdaBoost-C4.5, SVM
- chính xác hơn rất nhiều so với Naive Bayes thường được sử dụng để lọc thư rác

Phát triển

Tiếp tục nghiên cứu

- sưu tập thêm dữ liệu
- tích hợp vào hệ thống thư điện tử
- cải tiến tốc độ xử lý

Cám ơn & câu hỏi thảo luận .....