

Tìm kiếm chuyên gia với phản hồi từ người dùng và kNN-C4.4, kNN-RF-C4.4

Đỗ Thanh Nghị, Văn Thị Xuân Hồng

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

Ngày 13 tháng 4 năm 2012



Nội dung

Giới thiệu

Giải thuật kNN-C4.4, kNN-RF-C4.4

Kết quả thực nghiệm

Kết luận, hướng phát triển



Tìm kiếm thông tin

Tìm kiếm chuyên gia

- ▶ xét duyệt dự án
- ▶ xét duyệt bài báo khoa học
- ▶ hợp tác, cộng tác

Tiếp cận học để xếp hạng

- ▶ dữ liệu chuyên gia dạng văn bản: mô hình túi từ
- ▶ số thuộc tính (chiều) rất lớn: hàng ngàn đến hàng chục ngàn
- ▶ giải thuật máy học trên dữ liệu văn bản: xếp hạng



Tìm kiếm chuyên gia: xét duyệt bài báo hội thảo EGC'11

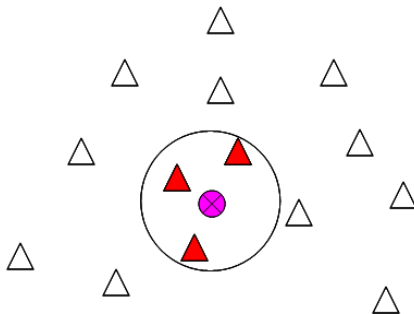
Máy học: kNN-C4.4, kNN-RF-C4.4

1. tìm k chuyên gia gần với tóm tắt bài báo
2. trưởng ban chương trình: chọn vài chuyên gia gần được xem là lớp +1, số chuyên gia còn lại là lớp -1
3. học cây quyết định cho xếp hạng: C4.4 hay RF-C4.4
4. lặp lại bước 2 cho đến khi thỏa mãn yêu cầu (~ 3 bước lặp)



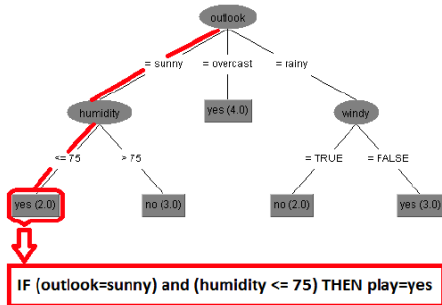
Giải thuật k láng giềng (Fix & Hodges, 1952)

Trả về k chuyên gia gần với tóm tắt bài báo cần xét



Cây quyết định C4.5 (Quinlan, 93)

Học để phân lớp (chơi golf hay không)



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
overcast	64	65	true	Play
sunny	72	95	false	Don't Play
sunny	69	70	false	Play
rain	75	80	false	Play
sunny	75	70	true	Play
overcast	72	90	true	Play
overcast	81	75	false	Play
rain	71	80	true	Don't Play



Cây quyết định C4.4 (Provost & Domingos, 03)

Học để ước lượng xác suất (xếp hạng)

- ▶ học như giải thuật C4.5: phân hoạch dữ liệu dựa vào entropy
- ▶ không cắt nhánh: cây lớn tối đa
- ▶ ước lượng xác suất tại nút lá
- ▶ nút lá thường có rất ít phần tử
- ▶ sử dụng ước lượng Laplace: tăng độ tin cậy



Cây quyết định C4.4 (Provost & Domingos, 03)

Ước lượng xác suất tại nút lá

- ▶ nút lá L có N phần tử trong đó có k phần tử thuộc lớp $+1$:
$$P_L(+1) = \frac{k}{N}$$

Xét ví dụ với dữ liệu chỉ có 2 lớp ± 1

- ▶ nút lá L_1 chỉ chứa 2 phần tử thuộc lớp $+1$:
$$P_{L_1}(+1) = \frac{2}{2} = 1.0$$
- ▶ nút lá L_2 chứa 18 trên tổng số 20 phần tử thuộc lớp $+1$:
$$P_{L_2}(+1) = \frac{18}{20} = 0.9$$
- ▶ ước lượng xác suất tại nút lá L_1 lớn hơn L_2 : ít tin cậy



Cây quyết định C4.4 (Provost & Domingos, 03)

Ước lượng xác suất tại nút lá: ước lượng Laplace

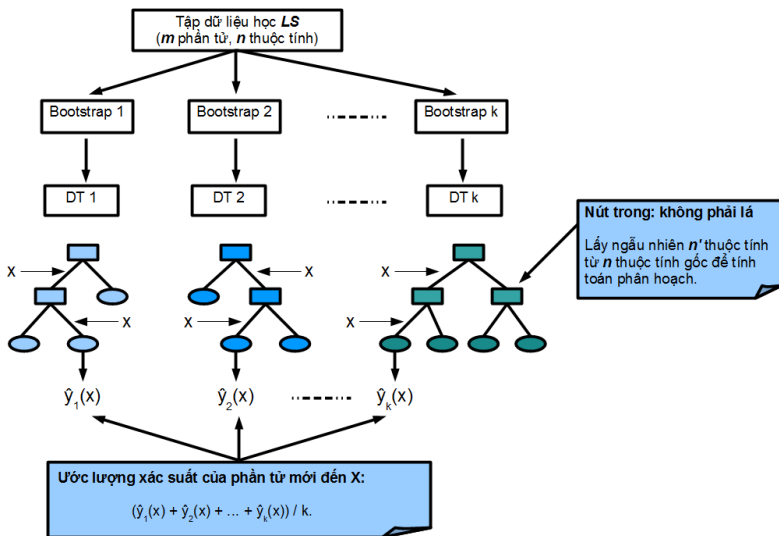
- ▶ nút lá L có N phần tử thuộc c lớp, trong đó có k phần tử thuộc lớp $+1$: $P_L(+1) = \frac{k+1}{N+c}$

Xét ví dụ với dữ liệu chỉ có 2 lớp ± 1

- ▶ nút lá L_1 chỉ chứa 2 phần tử thuộc lớp $+1$:
 $P_{L_1}(+1) = \frac{2+1}{2+2} = 0.75$
- ▶ nút lá L_2 chứa 18 trên tổng số 20 phần tử thuộc lớp $+1$:
 $P_{L_2}(+1) = \frac{18+1}{20+2} = 0.86$
- ▶ ước lượng xác suất tại nút lá L_1 nhỏ hơn L_2 : hợp lý



Rừng ngẫu nhiên cây quyết định C4.4 (RF-C4.4)



Hội thảo EGC'2011: French speaking - KDD

Tạo dữ liệu chuyên gia

- ▶ ban chương trình: 87 thành viên
- ▶ thu thập các tóm tắt công trình của 87 thành viên
- ▶ chỉ sử dụng công trình từ năm 2008-2010: DBLP
- ▶ tiền xử lý, mô hình túi từ: BoW (McCallum, 98)
- ▶ bảng dữ liệu: mỗi chuyên gia tương ứng với 1 dòng, mỗi từ vựng được xem là thuộc tính, có tổng cộng 9441 từ vựng



Chuẩn bị chương trình

Chương trình dùng cho thực nghiệm

- ▶ cài đặt giải thuật C4.4, RF-C4.4 bằng ngôn ngữ C
- ▶ xử lý văn bản: BoW (McCallum, 98)
- ▶ giao diện chương trình: TCL/TK



Tiêu chí đánh giá

Bảng: Confusion matrix

		Prediction	
		positive	negative
True class	positive	tp	fn
	negative	fp	tn

Recall, Precision, F1, Precision@k

- ▶ $Recall = \frac{tp}{tp+fn} = \frac{tp}{pos}$
- ▶ $Precision = \frac{tp}{tp+fp}$
- ▶ $F1 = \frac{2 \times prec \times rec}{prec + rec}$
- ▶ $Precision@k = \frac{tp}{k}$

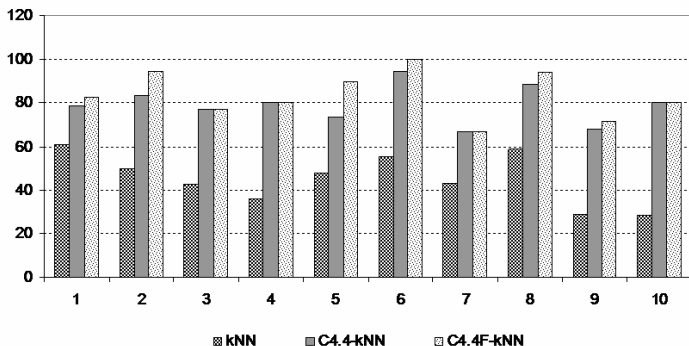


10 tóm tắt cần xét duyệt

ID	Tiêu đề câu truy vấn
1	Ordenez C. (2006), "Association rule discovery with the train and test approach for heart disease prediction", <i>Information Technology in Biomedicine</i> , vol. 10(2), pp. 334-343.
2	Chaiyaratana N., Zalzal A.M.S. (1997), "Recent developments in evolutionary and genetic algorithms: theory and applications", <i>Genetic Algorithms in Engineering Systems: Innovations and Applications 1997, GALESIA 97</i> , pp: 270-277.
3	Gao B. J. , Ester M. (2006), "Cluster Description Formats, Problems and Algorithms", <i>SDM 2006</i> .
4	Silva A., Lechevallier Y., Carvalho F. (2007), "Analyzing Distance Measures for Symbolic Data Based on Fuzzy Clustering", <i>Intelligent Systems Design and Applications 2007, ISDA 2007</i> , pp: 109-114.
5	Abdelhalim A., Traore I. (2009), "A New Method for Learning Decision Trees from Rules", <i>Machine Learning and Applications 2009, ICMLA '09</i> , pp: 693-698.
6	Chen B., Hoberock L.L. (1996), "A fuzzy neural network architecture for fuzzy control and classification", <i>Neural Networks 1996.</i> , pp: 1168-1173 vol.2.
7	Abidin S.Z.Z., Idris N.M., Husain A.H. (2010), "Extraction and classification of unstructured data in WebPages for structured multimedia database via XML", <i>Information Retrieval & Knowledge Management, (ICAMP), 2010</i> , pp: 44-49.
8	Ponmary Pushpa Latha D., Raj D.J.P., Sharmila D.J.S. (2007), "Generation of unified data structure and data warehouse for protein data banks", <i>Conference on Computational Intelligence and Multimedia Applications, 2007</i> , vol. 2, pp: 3-7.
9	Lamson B.G., Dimsdale B. (1996), "A natural language information Retrieval system", <i>Proceedings of the IEEE</i> , 54(12), pp: 1636-1640.
10	Kovalerchuk B. (2001), "Visualization and Decision-Making using Structural Information". <i>Conference on Imaging Science, Systems, and Technology (CISST'2001, June 25-28, 2001)</i> , Las Vegas, pp. 478-484.

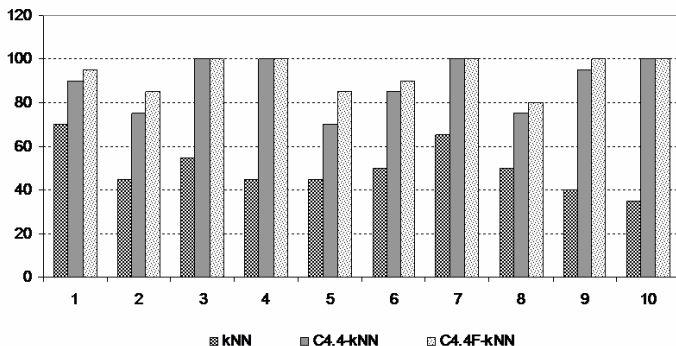


Kết quả ¹ từ top 20 chuyên gia trả về: Recall

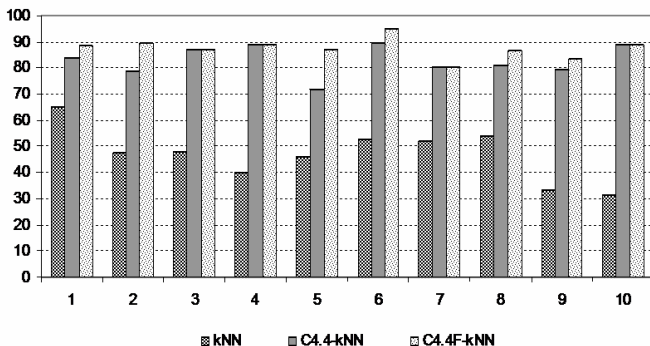


¹RF-C4.4: 50 cây C4.4 sử dụng 1000 thuộc tính ngẫu nhiên

Kết quả từ top 20 chuyên gia trả về: Precision



Kết quả từ top 20 chuyên gia trả về: F1



Kết quả theo tiêu chí Precision@5, 10, 15

ID	Precision@5 (%)	Precision@10 (%)	Precision@15 (%)	Số lần lặp
1	100.00	100.00	100.00	3
2	100.00	100.00	100.00	2
3	100.00	100.00	100.00	2
4	100.00	100.00	100.00	1
5	100.00	100.00	93.33	4
6	100.00	100.00	100.00	3
7	100.00	100.00	100.00	2
8	100.00	100.00	93.33	3
9	100.00	90.00	93.33	3
10	100.00	100.00	100.00	2
Avg	100.00	99.00	98.00	2.5



Kết luận

Tìm kiếm chuyên gia

- ▶ sử dụng phản hồi từ người dùng
- ▶ tiếp cận học xếp hạng kNN-C4.4, kNN-RF-C4.4
- ▶ thử nghiệm từ 87 chuyên gia của ban chương trình hội thảo khai mở dữ liệu EGC'2011 khối pháp ngữ
- ▶ kNN-C4.4 tìm được các chuyên gia phù hợp hơn kNN
- ▶ kNN-RF-C4.4 cho kết quả tốt hơn kNN-C4.4



Hướng phát triển

Tiếp tục nghiên cứu

- ▶ cung cấp thêm kết quả thực nghiệm
- ▶ so sánh với các giải thuật khác
- ▶ mở rộng ứng dụng



Cám ơn & câu hỏi thảo luận

