

Hàm entropy tổng quát cho giải thuật cây quyết định

Đỗ Thanh Nghị¹, Phạm Nguyên Khang²

Khoa Công Nghệ Thông Tin, Trường ĐH Cần Thơ
Số 1 – đường Lý Tự Trọng, Ninh Kiều, TP Cần Thơ

¹dtnghe@cit.ctu.edu.vn

²pnkhang@cit.ctu.edu.vn

Tóm tắt

Cây quyết định là một trong những giải thuật phân lớp được sử dụng nhiều và thành công nhất trong cộng đồng khai mô dữ liệu. Việc phân nhánh của cây quyết định thường dựa trên hàm entropy không tham số, chẳng hạn như Shannon entropy, không có tính tổng quát. Trong bài viết, chúng tôi trình bày những kết quả nghiên cứu từ việc thay đổi entropy không tham số bởi những hàm entropy tổng quát có tham số kiểu α , β . Sự thay thế hàm entropy cho phép cải thiện được mô hình học của cây quyết định. Kết quả thực nghiệm trên 10 tập dữ liệu từ kho dữ liệu UCI chỉ ra rằng giải thuật cây quyết định C4.5 sử dụng hàm entropy tổng quát cho kết quả tốt hơn so với việc sử dụng các hàm entropy không tham số, bao gồm Shannon entropy và chỉ số Gini.

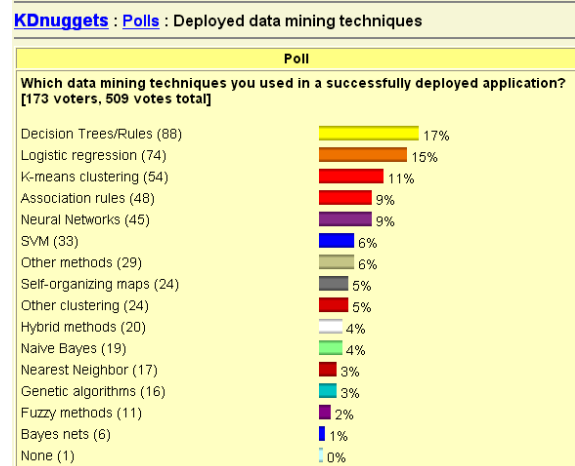
Từ khóa : Cây quyết định, Shannon entropy, chỉ số Gini, Entropy kiểu α , Entropy kiểu β .

1. Giới thiệu

Trong suốt nhiều năm qua, cộng đồng khai mô dữ liệu đã có những thống kê về các phương pháp sử dụng nhiều nhất và thành công trong lãnh vực khám phá tri thức và khai mô dữ liệu [4]. Kết quả trình bày trong hình 1 cho thấy rằng giải thuật cây quyết định vẫn là giải thuật quan trọng, hiệu quả và được sử dụng nhiều nhất. Khác với các mô hình học khác như mạng nơ ron hay máy học véc tơ hỗ trợ [13], mô hình học của cây quyết định đơn giản, nhanh, cũng cho kết quả tốt, điều đặc biệt quan trọng là kết quả sinh ra của cây quyết định là tập hợp các luật đơn giản dễ diễn dịch. Mô hình cây quyết định có cấu trúc dạng cây mà ở đó, nút lá được gán nhãn tương ứng với lớp của dữ liệu, nút trong được tích hợp với điều kiện kiểm tra để rẽ nhánh. Một luật quyết định có dạng IF-THEN được tạo ra từ việc thực hiện AND trên các điều kiện theo đường dẫn từ nút gốc đến nút lá.

Việc phân nhánh của giải thuật cây quyết định dựa trên hàm đo độ hỗn loạn hay độ lợi thông tin. Nếu xét qua hai giải thuật tiêu biểu cây quyết định, C4.5 [7] và CART [2] đều sử dụng những hàm entropy không tham số như Shannon entropy [9] hay chỉ số Gini không có tính tổng quát.

Khi xử lý dữ liệu gien có số chiều lớn, [6] đã đề nghị thay thế Shannon entropy của giải thuật cây quyết



Hình 1. Giải thuật khai mô dữ liệu phổ dụng

định bởi hàm Tsallis entropy [10] và hàm entropy kiểu α [8], kết quả cho thấy việc thay thế này giúp cho xử lý tốt hơn so với dùng hàm Shannon entropy. Wehenkel [12] cung cấp các kiểm chứng lý thuyết về việc sử dụng hàm entropy tổng quát kiểu β [3] và cũng chỉ ra được sự tương đương khi dùng hàm Shannon entropy và chỉ số Gini. Lenca [5] cũng đề nghị lớp hàm off-centered entropy của các hàm Shannon entropy, entropy tổng quát, dùng cho phân lớp dữ liệu không cân bằng (lớp rất bé so với lớp khác).

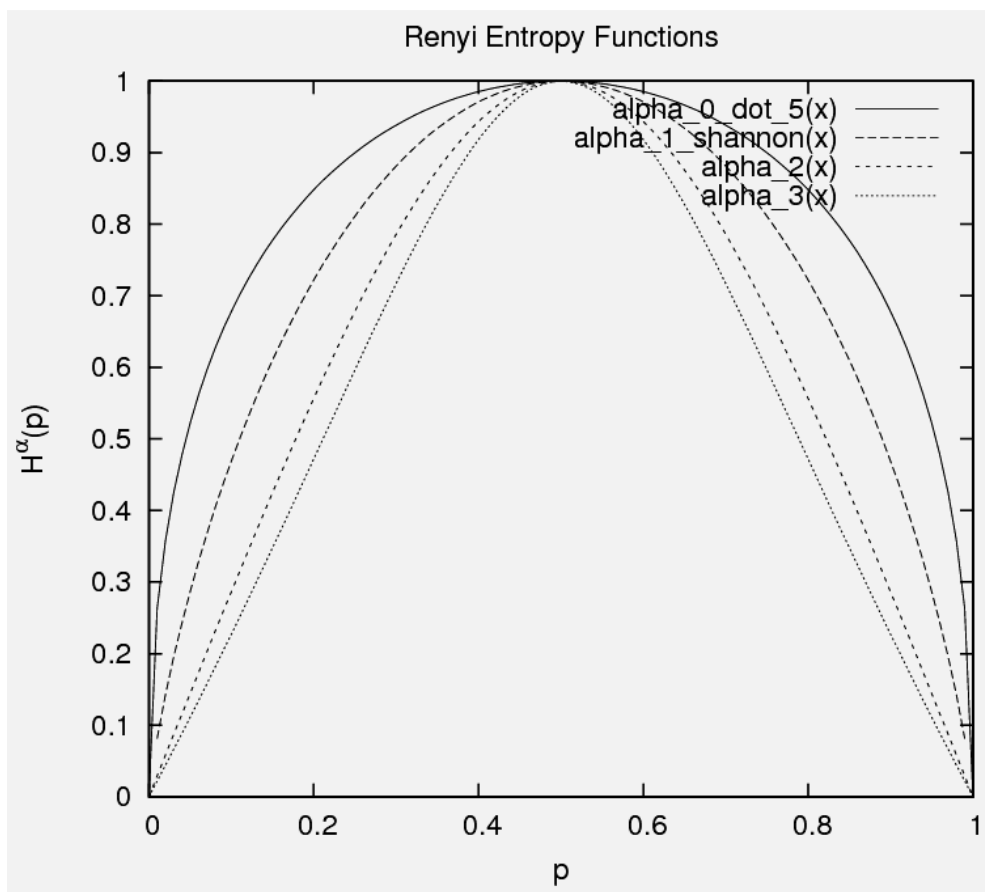
Trong bài viết, chúng tôi trình bày những kết quả nghiên cứu từ việc thay đổi entropy không tham số bởi những hàm entropy tổng quát có tham số kiểu α [8],

β [3]. Sự thay đổi này cho phép chuyên gia phân tích dữ liệu có cơ hội điều chỉnh các tham số của các hàm entropy tổng quát để thu được mô hình cây quyết định tốt hơn so với việc sử dụng các hàm entropy không tham số. Chúng tôi cũng cung cấp kết quả thực nghiệm của giải thuật cây quyết định sử dụng các hàm entropy khác nhau trên 10 tập dữ liệu từ kho dữ liệu UCI [1]. Trong tất cả các trường hợp, sự thay thế hàm entropy tổng quát cho kết quả tốt hơn so với việc sử dụng các hàm entropy không tham số, bao gồm cả Shannon entropy và chỉ số Gini.

Phần tiếp theo của bài viết được tổ chức như sau. Trong mục 2, chúng tôi trình bày hai hàm entropy tổng

2. Hàm entropy tổng quát cho giải thuật cây quyết định

Giải thuật học cây quyết định, CART [2], C4.5 [7] được xây dựng đệ quy, chia để trị từ trên-xuống (top-down recursive divide-and-conquer). Bắt đầu, tất cả dữ liệu học đều ở nút gốc. Dữ liệu được phân hoạch một cách đệ quy, bằng cách chọn một thuộc tính của dữ liệu và giá trị cắt tốt nhất để phân hoạch dữ liệu. Sự lựa chọn được thực hiện dựa trên hàm đo độ hỗn loạn hay độ lợi thông tin. Giải thuật cây quyết định C4.5 của cộng đồng máy học sử dụng hàm Shannon entropy [9]



Hình 2. Hàm entropy kiểu α

quát có tham số kiểu α , β . Mục 3 sẽ trình bày kết quả thực nghiệm và đánh giá kết quả thu được từ việc sử dụng các hàm entropy. Cuối cùng là kết luận và hướng phát triển được trình bày ở mục 4 trước khi kết thúc bởi các tài liệu tham khảo.

để phân hoạch dữ liệu, trong khi chỉ số Gini lại được sử dụng bởi giải thuật CART.

Chúng tôi đề nghị thay thế entropy không tham số bởi những hàm entropy tổng quát có tham số kiểu α , β cho phép cải tiến hiệu quả phân hoạch của mô hình cây quyết định. Sự thay đổi này xuất phát từ việc các giải thuật cây quyết định chỉ sử dụng những hàm entropy không tham số như là một heuristics để đo độ hỗn loạn thông tin. Đối với một phân phối dữ liệu, hàm entropy

không tham số có tốc độ tăng/giảm không đổi sử dụng (hình 2), nhưng chúng ta lại không biết được tốc độ tăng/giảm nhanh hay chậm của hàm entropy khi nào là tối ưu cho một tập dữ liệu. Chính vì lý do này, chúng tôi đề nghị sử dụng những hàm entropy có tham số kiểu α, β cho phép người sử dụng có cơ hội điều chỉnh các tham số của các hàm entropy (điều khiển tốc độ tăng/giảm nhanh hay chậm) để thu được mô hình cây quyết định tốt hơn so với việc sử dụng các hàm entropy không tham số.

2.1. Hàm entropy kiểu α

Hàm entropy tổng quát kiểu α được đề nghị bởi Rényi [8]. Đây là sự mở rộng của hàm Shannon entropy bằng việc thêm vào một tham số α . Với phân phối xác suất rời rạc $(p_1, \dots, p_m)$ của các sự kiện, hàm entropy kiểu α (α dương, khác 1) được định nghĩa như sau :

$$H^\alpha = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^m p_i^\alpha \right) \quad (1)$$

Hàm entropy kiểu α , $H^\alpha(p_1, \dots, p_m)$ có các tính chất cơ bản của một hàm entropy theo Shannon [9] đề nghị như :

$H^\alpha(p_1, \dots, p_m)$ bất biến khi bị hoán vị các đối số $p_1, \dots, p_m$ của nó;

$0 \leq H^\alpha(p_1, \dots, p_m) \leq H^\alpha(1/m, \dots, 1/m)$, hàm H^α phải có giá trị lớn nhất là $\log(m)$ tương ứng với phân phối ngẫu nhiên đều của các sự kiện, hàm H^α có giá trị 0 khi phân phối của sự kiện là chắc chắn;

$H^\alpha(p_1, \dots, p_m)$ là hàm lõm nghiêm ngặt (strictly concave).

Để biết thêm kiểm chứng của các tính chất trên, độc giả có thể tham khảo thêm trong tài liệu [8].

Tại giá trị $\alpha = 1$, hàm H^α không xác định, tuy nhiên nó tồn tại giới hạn của hàm như sau :

$$\lim_{\alpha \rightarrow 1} H^\alpha = - \sum_{i=1}^m p_i \log p_i \quad (2)$$

Từ công thức (2) cho thấy rằng, khi giá trị tham số α tiến về 1 thì hàm entropy kiểu α trở thành hàm Shannon entropy.

Giả sử chúng ta có hai sự kiện (2 lớp dữ liệu với phân phối xác suất $p_1 = p$ và $p_2 = 1-p$), hình 2 là đồ thị thu được của hàm entropy kiểu α với các giá trị lần lượt là 0.5, 1 (Shannon entropy), 2, 3.

Khi thay đổi các giá trị của tham số α , chúng ta có thể điều khiển được tốc độ tăng/giảm nhanh hay chậm của hàm entropy dùng để đo độ hỗn loạn của thông tin tại mỗi nút phân hoạch của cây quyết định. Điều này

cho phép thu được mô hình học của cây quyết định phù hợp cho từng tập dữ liệu hơn là chỉ sử dụng hàm entropy không tham số.

2.2. Hàm entropy kiểu β

Khái niệm hàm entropy tổng quát kiểu β được Daróczy giới thiệu lần đầu tiên trong [3]. Entropy kiểu β (β dương và khác 1) của một phân phối xác suất rời rạc $(p_1, \dots, p_m)$ được định nghĩa như sau :

$$H^\beta(p_1, \dots, p_m) = \frac{2^{\beta-1}}{2^{\beta-1} - 1} \left(1 - \sum_{i=1}^m p_i^\beta \right) \quad (3)$$

Hàm $H^\beta(p_1, \dots, p_m)$ cũng có các tính chất cơ bản của một hàm entropy như :

$H^\beta(p_1, \dots, p_m)$ bất biến với việc hoán vị các đối số của nó;

$0 \leq H^\beta(p_1, \dots, p_m) \leq H^\beta(1/m, \dots, 1/m)$, điều này có nghĩa là hàm H^β đạt cực đại bằng 1 ứng với trường hợp phân phối đều và hàm H^β đạt cực tiểu bằng 0 khi phân phối chắc chắn;

$H^\beta(p_1, \dots, p_m)$ là hàm lõm nghiêm ngặt.

Chứng minh của các tính chất này được trình bày trong [3, 12].

Hình 3 mô tả đồ thị của hàm H^β trong trường hợp 2 lớp (trong đó $p_1 = p$ và $p_2 = 1-p$), với một vài giá trị β khác nhau. Đặc biệt, nếu đặt $\beta = 2$ trong công thức (3) ta sẽ thu được (4), đây là chỉ số Gini được sử dụng trong giải thuật CART :

$$H^\beta(p_1, \dots, p_m) = 2 \left(1 - \sum_{i=1}^m p_i^2 \right) \quad (4)$$

Hàm H^β không xác định tại giá trị $\beta = 1$. Tuy nhiên, hàm H^β vẫn tồn tại giới hạn khi $\beta \rightarrow 1$:

$$\lim_{\beta \rightarrow 1} H^\beta = - \sum_{i=1}^m p_i \log p_i \quad (5)$$

Công thức (5) chính là entropy Shannon [9] được dùng trong giải thuật C4.5.

Vì vậy, một giải thuật cây quyết định sử dụng entropy tổng quát kiểu β cho điều kiện phân nhánh sẽ đảm bảo tính tổng quát hóa tốt nhất. Các chuyên gia phân tích dữ liệu có thể nâng cao kết quả bằng cách điều chỉnh tham số β ví dụ như đặt $\beta = 2$ (chỉ số Gini) hoặc cho $\beta \rightarrow 1$ kết quả thu được sẽ tiến tới kết quả của entropy Shannon), v.v.

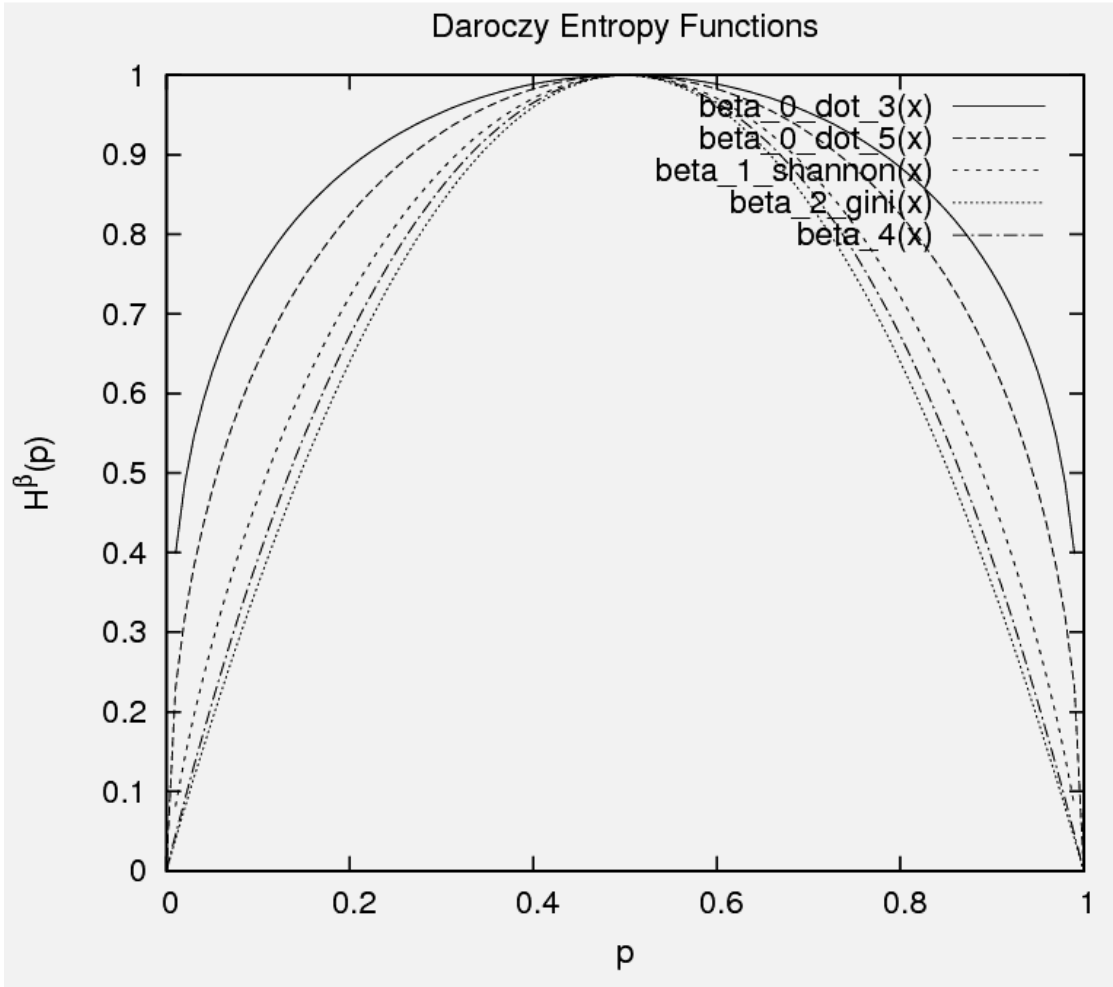
Tương tự như trên, bằng việc thay đổi giá trị của tham số β chúng ta có thể điều khiển được tốc độ tăng/giảm nhanh hay chậm của hàm entropy dùng để đo độ hỗn loạn của thông tin tại mỗi nút phân hoạch của

cây quyết định để thu được mô hình học của cây quyết định phù hợp cho từng tập dữ liệu.

3. Kết quả thực nghiệm

Để có thể đánh giá tính hiệu quả của các loại entropy, chúng tôi đã tích hợp entropy tổng quát kiểu α, β vào giải thuật cây quyết định C4.5 [7].

Trong các thí nghiệm, chúng tôi nghiên cứu hiệu quả của giải thuật cây quyết định khi được tích hợp với các kiểu entropy khác nhau. Chúng tôi quan tâm tới việc so sánh hiệu quả của entropy tổng quát kiểu α, β với các entropy khác được sử dụng trong lĩnh vực phân loại dữ liệu như Shannon entropy và chỉ số Gini.



Hình 3. Hàm entropy kiểu β

Bảng 1. Mô tả các tập dữ liệu

ID	Tập dữ liệu	Số chiều	Số phần tử	Số lớp	Nghi thức kiểm tra
1	Pima	8	768	2	10-fold
2	Bupa	6	345	2	10-fold
3	Adult	110	48842	2	32561 Trn – 16281 Tst
4	Breast Cancer	30	569	2	10-fold
5	German	20	1000	2	10-fold
6	Abalone	8	4177	2	10-fold
7	Spambase	57	4601	2	10-fold
8	Tic-tac-toe	9	958	2	10-fold
9	Satimage	36	6435	7	4435 Trn – 2000 Tst
10	Segment	19	2310	7	10-fold

Bảng 2. Kết quả phân lớp của cây quyết định sử dụng những hàm entropy khác nhau

ID	Tập dữ liệu	Độ chính xác (%)			
		Shannon entropy	Gini index	α entropy	β entropy
1	Pima	<u>74.55</u>	72.73	74.42 ($\alpha = 1.2$)	74.94 ($\beta = 1.1$)
2	Bupa	69.11	<u>69.41</u>	68.24 ($\alpha = 0.7$)	71.47 ($\beta = 1.4$)
3	Adult	85.50	85.73	<u>85.79</u> ($\alpha = 1.5$)	85.84 ($\beta = 2.5$)
4	Breast Cancer	93.33	92.63	<u>95.44</u> ($\alpha = 0.5$)	96.32 ($\beta = 0.5$)
5	German	71.15	73.02	75.00 ($\alpha = 1.5$)	<u>73.54</u> ($\beta = 1.5$)
6	Abalone	90.91	91.05	91.77 ($\alpha = 1.5$)	<u>91.53</u> ($\beta = 1.5$)
7	Spambase	92.96	92.35	93.55 ($\alpha = 0.5$)	93.59 ($\beta = 0.5$)
8	Tic-tac-toe	93.23	93.44	93.65 ($\alpha = 1.5$)	93.65 ($\beta = 1.5$)
9	Satimage	85.65	<u>86.20</u>	85.65 ($\alpha = 1.8$)	86.90 ($\beta = 1.9$)
10	Segment	<u>96.62</u>	95.58	96.49 ($\alpha = 0.5$)	96.84 ($\beta = 0.5$)

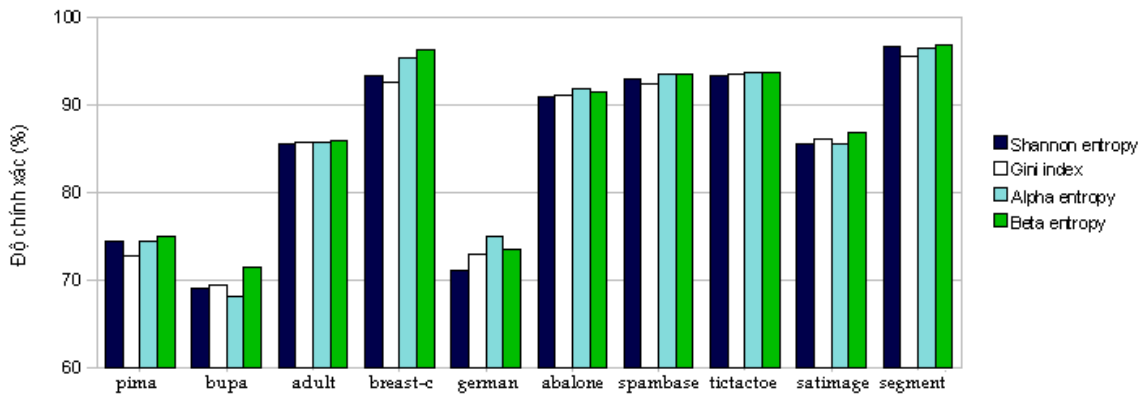
Chúng tôi sử dụng 10 tập dữ liệu của kho UCI [1] để thực hiện các thí nghiệm này. Bảng 1 mô tả chi tiết 10 tập dữ liệu. Trong bảng này, cột thứ 2 là tên của tập dữ liệu, cột 3 và cột 4 mô tả số chiều (thuộc tính) và số phần tử của dữ liệu. Số lớp được chỉ định trong cột 5.

Chúng tôi sử dụng tính chính xác (accuracy) làm tiêu chí để so sánh hiệu quả của việc sử dụng các hàm entropy khác nhau. Nghi thức kiểm tra đánh giá được trình bày trong cột cuối cùng của bảng 1. Trong các tập dữ liệu, có một số tập đã được chia sẵn thành tập huấn luyện (trn) và tập kiểm tra (tst). Đối với các tập dữ liệu này, chúng tôi sử dụng tập huấn luyện để xây dựng cây

Đối với các tập dữ liệu không có sẵn tập huấn luyện và tập kiểm tra, chúng tôi sử dụng nghi thức kiểm định chéo 10-fold (thường được dùng trong lĩnh vực thống kê và máy học). Dữ liệu được chia làm 10 tập con. Lần lượt ra lấy ra 1 tập con dùng làm tập kiểm tra, 9 tập con còn lại dùng làm tập huấn luyện. Lặp lại bước này đúng 10 lần và tính trung bình các kết quả thu được.

Bảng 2 trình bày kết quả phân lớp thu được với các entropy khác nhau. Kết quả tốt nhất được **in đậm**. Kết quả tốt thứ hai được gạch dưới. Chúng tôi cung cấp thêm đồ thị trong hình 4 để thuận lợi cho việc so sánh. Kết quả thực nghiệm cho thấy rằng entropy tổng quát

So sánh kết quả phân lớp
của các hàm entropy



Hình 4. So sánh kết quả phân lớp của các hàm entropy

quyết định. Chúng tôi cũng đã điều chỉnh các tham số đầu vào của giải thuật C4.5 nhằm đạt được kết quả cao nhất có thể. Chú ý rằng việc điều chỉnh tham số được thực hiện với tất cả các entropy. Sau khi xây dựng cây quyết định xong, chúng tôi dùng nó để phân loại tập kiểm tra.

(dù là kiểu α hay β) luôn cho kết quả tốt hơn so với các entropy không tham số khác, bao gồm cả Shannon entropy và chỉ số Gini. Entropy kiểu β luôn cho kết quả tốt nhất. Độ chính xác do entropy kiểu α mang lại có kém hơn một chút so với entropy kiểu β nhưng vẫn còn tốt hơn nhiều so với entropy Shannon và chỉ số Gini.

Hơn nữa, với entropy tổng quát các chuyên gia phân tích dữ liệu còn có thể có cơ hội đạt được kết quả tốt hơn nữa bằng cách tinh chỉnh các tham số α , β thay vì phải chấp nhận một kết quả cố định trong trường hợp sử dụng entropy không tham số như là entropy Shannon hay chỉ số Gini.

4. Kết luận và hướng phát triển

Trong bài viết này, chúng tôi đã trình bày một phương pháp cải tiến việc phân nhánh của giải thuật cây quyết định cổ điển nhờ vào việc sử dụng entropy tổng quát. Ý tưởng chính là dùng entropy tổng quát kiểu α và β thay vì sử dụng entropy không tham số như entropy Shannon hay chỉ số Gini. Phương pháp đề nghị tạo điều kiện cho các chuyên gia phân tích dữ liệu có cơ hội điều chỉnh các tham số α và β cho phép điều khiển được tốc độ tăng/giảm nhanh hay chậm của hàm entropy dùng để đo độ hỗn loạn của thông tin tại mỗi nút phân hoạch của cây quyết định để thu được mô hình học của cây quyết định phù hợp cho từng tập dữ liệu. Kết quả thực nghiệm trên 10 tập dữ liệu của UCI cho thấy rằng việc tích hợp entropy tổng quát vào giải thuật cây quyết định C4.5 mang lại kết quả tốt hơn rất nhiều so với việc sử dụng entropy không tham số như entropy Shannon hay chỉ số Gini.

Hướng phát triển sắp tới của chúng tôi là sử dụng entropy kiểu α , β để học các luật quyết định. Chúng tôi cũng dự định thực hiện thêm thí nghiệm trên các tập dữ liệu khác và so sánh với các entropy khác nữa [10, 11] nhằm đánh giá sự ổn định của phương pháp được đề nghị.

5. Tài liệu tham khảo

[1] Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mllearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Breiman, L., J. H. Friedman, R. A. Olshen, et C. Stone (1984). *Classification and Regression Trees*. Wadsworth International.

[3] Daroczy, A. (1970). Generalized information functions. *Information and Control* (16), 36–51.

[4] Kdnuggets (2003-2007). Data mining methods. In KDNuggets Polls 2003-2007.

[5] Lenca, P. and Lallich, S., T. N. Do, et N. K. Pham (2008). A comparison of different off-centered entropies to deal with class imbalance for decision trees. In The Pacific-Asia Conference on Knowledge Discovery and DataMining, LNAI 5012, pp. 634–643. Springer-Verlag.

[6] Maszczyk, T., Duch, W. (2006). Comparison of Shannon, Renyi and Tsallis Entropy Used in Decision Trees. In The 9th international conference on Artificial Intelligence and Soft Computing, LNAI 5097, pp. 643–651. Springer-Verlag.

[7] Quinlan, J. R. (1993). *C4.5 : Programs for Machine Learning*. San Mateo, CA : Morgan Kaufmann.

[8] Renyi, A. (1961). On measures of entropy and information. In Proc. Fourth Berkeley Symp. Math. Stat. Prob., volume 1, page 547, Berkeley, *University of California Press*.

[9] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technological Journal* (27), 379–423, 623–656.

[10] Tsallis, C., Mendes, R., Plastino, A. (1998). The role of constraints within generalized nonextensive statistics. *Physica* 261A, 534–554

[11] Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology* (76), 103–154.

[12] Wehenkel, L. (1996). On uncertainty measures used for decision tree induction. In International Congress on Information Processing and Management of Uncertainty in Knowledge based Systems, pp. 413–418.

[13] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.