Thiết kế ngữ nghĩa hình thang cho hệ phân lớp dựa trên luật mờ

Phạm Đình Phong
NCS K19, Trường Đại học Công nghệ, ĐHQGHN

Tóm tắt

Hệ phân lớp dựa trên luật mờ được nghiên cứu rộng rãi do người dùng cuối có thể sử dụng những tri thức dạng luật được trích rút từ dữ liệu có tính dễ hiểu, dễ sử dụng đối với con người như là những tri thức của họ. Đại số gia tử (ĐSGT) là một cách tiếp cận mới cho việc xử lý miền giá trị của biến ngôn ngữ, khai thác tính chất sánh được của các giá trị ngôn ngữ, là cơ sở xác định ngữ nghĩa định lượng từ ngữ nghĩa định tính, đã cho phép tạo ra các ràng buộc về ngữ nghĩa và đã được ứng dụng hiệu quả trong quá trình tìm kiếm, thiết kế tập giá trị ngôn ngữ cùng với ngữ nghĩa dựa trên tập mờ của chúng cho bài toán xây dựng tự động cơ sở luật cho hệ phân lớp dựa trên luật mờ. Tuy nhiên, với tiếp cận ĐSGT truyền thống $\mathcal{AX}$, ngữ nghĩa lõi và ngữ nghĩa dựa trên tập mờ hình thang của các từ ngôn ngữ chưa được mô hình hóa. Nghiên cứu này đề xuất mở rộng lý thuyết ĐSGT biểu diễn được ngữ nghĩa của các từ ngôn ngữ phụ thuộc ngữ cảnh và đề xuất thay đổi phương pháp lượng hóa ĐSGT để mô hình hóa ngữ nghĩa định tính của các từ ngôn ngữ phù hợp với ngữ cảnh mới nhằm cung cấp một cơ chế hình thức cho việc sinh lõi ngữ nghĩa và ngữ nghĩa tính toán dựa trên tập mờ hình thang của khung nhận thức ngôn ngữ. Nghiên cứu này cũng chứng tỏ khả năng ứng dụng hiệu quả của ĐSGT mở rộng $\mathcal{AX}^{mr}$ trong thiết kế tự động hệ phân lớp dựa trên luật mờ.

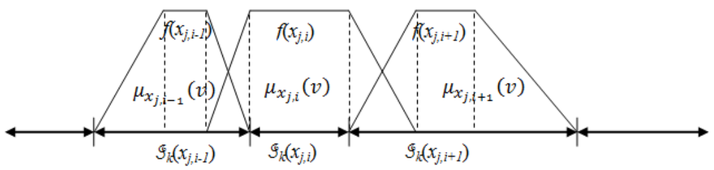
Mục tiêu nghiên cứu

Xây dựng một cơ chế hình thức toán học cho việc sinh tự động ngữ nghĩa tính toán dựa trên tập mờ hình thang từ ngữ nghĩa định tính của các từ ngôn ngữ cho bài toán thiết kế tự động cơ sở luật cho hệ phân lớp dựa trên luật ngôn ngữ mờ.

Phương pháp

Thiết kế ngữ nghĩa dựa trên tập mờ của các từ ngôn ngữ

- Nhờ bổ sung một gia từ đặc biệt h_0, lõi ngữ nghĩa của các từ ngôn ngữ được mô hình hóa và là cơ sở hình thức cho việc sinh ngữ nghĩa dựa trên tập mờ hình thang của các từ ngôn ngữ.
- Mỗi thuộc tính j của tập dữ liệu được liên kết với một ĐSGT $\mathcal{AX}^{mr}_j$. Khi cho các giá trị cụ thể của các tham số ngữ nghĩa, các từ ngôn ngữ, các khoảng tính mờ và lõi ngữ nghĩa của chúng được xây dựng với thứ tự tương đồng với thứ tự ngữ nghĩa vốn có của các từ ngôn ngữ. Các khoảng tính mờ $\Im\downarrow k\downarrow j\ (x\downarrow j,i\ )$ tạo thành một phân hoạch nhị phân trên khoảng [0,1]. Các tập mờ hình thang được xây dựng dựa trên các lõi ngữ nghĩa của các từ ngôn ngữ.



Ngữ nghĩa dựa trên tập mờ hình thang của các từ ngôn ngữ

- Các tập mờ hình thang có thể được cấu trúc hóa dưới dạng *đơn thể hạt* hoặc *đa thể hạt*. Trong các thực nghiệm cho thấy thiết kế *đa thể hạt* thường cho hiệu quả phân lớp tốt hơn.

Sinh tập luật khởi đầu từ dữ liệu

- Các khoảng tính mờ của tất cả các thuộc tính tạo thành các siêu hộp và mỗi siêu hộp chỉ chứa một mẫu dữ liệu. Chỉ sinh luật từ siêu hộp có chứa dữ liệu. Các luật có độ dài bằng số thuộc tính n.
- Sinh các luật ngắn hơn bằng cách bỏ bớt một số điều kiện luật.
- Áp dụng các tiêu chuẩn sàng để loạt bớt các luật ít quan trọng.

Tối ưu các tham số ngữ nghĩa và lựa chọn hệ luật tối ưu.

- Cho các tham số ngữ nghĩa tương tác với dữ liệu để tìm bộ tham số tối ưu và làm đầu vào cho tiến trình lựa chọn hệ luật tối ưu.
- Giải thuật tối ưu bầy đàn đa mục tiêu với hàm thích nghi chia sẻ được sử dụng trong nghiên cứu này.

Kết quả và đánh giá

Các kết quả thực nghiệm của hệ phân lớp dựa trên luật mờ với ngữ nghĩa dựa trên tập mờ hình thang của các từ ngôn ngữ (**FRBC_$\mathcal{AX}^{mr}$**) đối với 23 tập dữ liệu mẫu chuẩn UCI và so sánh với ngữ nghĩa dựa trên tập mờ tam giác được thiết kế dựa trên ĐSGT truyền thống $\mathcal{AX}$ (**FRBC_$\mathcal{AX}$**) được thể hiện trong bảng sau.

STT	Tập dữ liệu	FRBC_$\mathcal{AX}^{mr}$				FRBC_$\mathcal{AX}$				$\neq P_{te}$	$\neq R \times C$
		#R	#C	P_{tr}	P_{te}	#R	#C	P_{tr}	P_{te}		
1	App	3,70	**16,91**	91,30	**88,09**	4,00	21,32	92,28	87,55	0,54	-4,41
2	Aus	5,00	41,85	87,72	**86,86**	4,10	**36,20**	88,06	86,38	0,48	*5,65*
3	Ban	7,00	78,19	76,28	72,10	6,00	**52,20**	76,17	**72,80**	*-0,70*	*25,99*
4	Bup	8,97	**170,70**	77,54	**69,41**	8,83	187,20	78,13	68,09	1,32	-16,50
5	Cle	16,47	**640,19**	69,86	**63,40**	15,93	657,43	72,44	62,19	1,21	-17,24
6	Der	10,87	**189,46**	96,88	95,52	10,90	198,05	98,03	**96,07**	*-0,55*	-8,59
7	Gla	16,80	488,38	80,26	**72,78**	15,43	**343,60**	80,45	72,09	0,69	*144,78*
8	Hab	4,00	20,00	77,67	**77,43**	3,00	**10,20**	76,91	75,76	1,67	*9,80*
9	Hay	10,03	139,42	89,98	83,33	9,53	**122,27**	90,11	**84,17**	*-0,84*	*17,15*
10	Hea	8,03	**120,69**	88,07	**84,57**	7,67	122,72	89,63	84,44	0,13	-2,03
11	Hep	3,77	**25,75**	94,44	**89,17**	3,83	26,16	95,83	88,44	0,73	-0,41
12	Ion	8,63	**83,71**	94,67	**90,98**	8,97	90,33	95,35	90,22	0,76	-6,62
13	Iri	5,43	34,59	98,35	**96,67**	4,67	**26,29**	98,40	96,00	0,67	*8,30*
14	Mam	7,20	**82,08**	85,31	**84,46**	6,90	92,25	86,05	84,20	0,26	-10,17
15	New	5,07	**30,93**	96,30	**95,03**	6,00	45,18	97,02	94,42	0,61	-14,25
16	Pim	5,97	**50,33**	78,53	**76,66**	5,97	60,89	78,28	76,18	0,48	-10,57
17	Sah	6,26	**58,41**	74,55	**70,27**	6,30	86,75	76,35	69,33	0,94	-28,35
18	Son	5,97	**53,91**	86,84	**77,29**	6,80	79,76	88,39	76,80	0,49	-25,85
19	Tae	11,13	**163,61**	68,36	59,46	11,60	261,00	72,11	**59,47**	*-0,01*	-97,39
20	Veh	11,03	**216,19**	71,64	**68,12**	11,60	242,79	70,30	67,62	0,50	-26,60
21	Wdb	4,00	**23,08**	97,16	95,96	4,87	37,35	97,62	**96,96**	*-1,00*	-14,27
22	Win	5,87	42,09	100,00	**98,52**	5,57	35,82	99,88	98,30	0,22	*6,27*
23	Wis	6,93	**59,81**	97,20	96,51	6,93	74,36	97,81	**96,74**	*-0,23*	-14,55
Trung bình			**123.05**	**86,04**	**82,29**		**126,53**	**86,77**	**81,92**		

Hệ phân lớp **FRBC_$\mathcal{AX}^{mr}$** có hiệu suất phân lớp trên tập kiểm tra cao hơn đối với 17 tập dữ liệu trong số 23 tập dữ liệu mẫu được thực nghiệm, đồng thời có độ phức tạp trung bình của hệ phân lớp thấp hơn so với phương pháp **FRBC_$\mathcal{AX}$**.

So sánh hiệu suất phân lớp sử dụng phương pháp kiểm định Wilcoxon Signed Rank với mức $\alpha = 0,05$ được thể hiện trong bảng dưới.

So sánh ($\alpha = 0,05$)	R^+	R^-	Exact P-value	Asymp. P-value	Hypothesis
FRBC_$\mathcal{AX}^{mr}$ vs **FRBC_$\mathcal{AX}$**	207,0	69,0	0,03544	0,034529	Rejected

So sánh độ phức tạp của phân lớp sử dụng kiểm định Wilcoxon Signed Rank với mức $\alpha = 0,05$ được thể hiện trong bảng dưới

So sánh ($\alpha = 0,05$)	R^+	R^-	Exact P-value	Asymp. P-value	Hypothesis
FRBC_$\mathcal{AX}^{mr}$ vs **FRBC_$\mathcal{AX}$**	193,0	83,0	0,09798	0,091405	Not Rejected

Giả thuyết sự tương đương về hiệu suất phân lớp của hai hệ phân lớp được so sánh bị bác bỏ và giả thuyết tương đương về độ phức tạp của hai hệ phân lớp không bị bác bỏ, do đó phương được đề xuất có hiệu suất phân lớp tốt hơn nhưng không làm tăng phức tạp của hệ phân lớp.

Kết luận

- Lý thuyết ĐSGT được mở rộng nhằm mô hình hóa lõi ngữ nghĩa và ngữ nghĩa tính toán dựa trên tập mờ hình thang của các từ ngôn ngữ của biến ngôn ngữ trên cơ sở bổ sung một gia tử đặc biệt h_0.
- Dựa trên lõi ngữ nghĩa của các từ ngôn ngữ, một cơ chế hình thức sinh ngữ nghĩa dựa trên tập mờ hình thang của các từ ngôn ngữ từ ngữ nghĩa định tính của chúng được đề xuất.
- Ngữ nghĩa dựa trên tập mờ hình thang được ứng dụng trong biểu diễn ngữ nghĩa của các từ ngôn ngữ trong cơ sở luật của hệ phân lớp dựa trên luật mờ cho hiệu suất phân lớp tốt hơn so với ngữ nghĩa dựa trên tập mờ tam giác và không làm tăng độ phức tạp của hệ phân lớp.
- Mở ra khả năng ứng dụng hiệu quả của ngữ nghĩa hình thang trong các khác nhau như bài toán điều khiển, thao tác cơ sở dữ liệu mờ và nhận dạng hệ mờ nhằm tăng tính hiệu quả và tính linh hoạt trong biểu diễn ngữ nghĩa.

Tài liệu tham khảo

- Ho N. C., Wechle W. (1990), "Hedge algebras: an algebraic approach to structures of sets of linguistic domains of linguistic truth values", *Fuzzy Sets and Systems* 35 (3), pp. 281–293.
- Ho N. C., Wechler W. (1992), "Extended algebra and their application to fuzzy logic", *Fuzzy Sets and Systems* 52, pp. 259–281.
- Ho N. C., Pedrycz W., Long D. T., Son T. T. (2013), "A genetic design of linguistic terms for fuzzy rule based classifiers", *International Journal of Approximate Reasoning* 54 (1), pp. 1– 21.
- Ishibuchi H., Yamamoto T. (2004), "Fuzzy Rule Selection by Multi-Objective Genetic Local Search Algorithms and Rule Evaluation Measures in Data Mining", *Fuzzy Sets and Systems* 141 (1), pp. 59–88.