간결화 알고리즘과 규칙 관계 분석

Concise Algorithm and Analysis of Rule Relations

홍정식 교수[1], 김재호 석사과정[2], 이정언 석사과정[2], 황인서 학사과정[2]

[1]서울과학기술대학교 산업공학과,

[2]서울과학기술대학교 일반대학원 데이터사이언스학과

[1]hong@seoultech.ac.kr,
[2]{doe2x2, jeongeon.lee}@seoultech.ac.kr, nayaseo98@gmail.com

본 연구는 2022년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행되었음 (P0017123, 2022년 산업혁신인재성장지원사업)

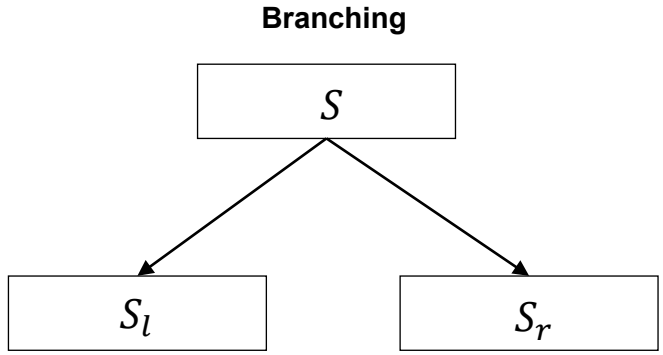
2022년 춘계공동학술대회

Table of Contents

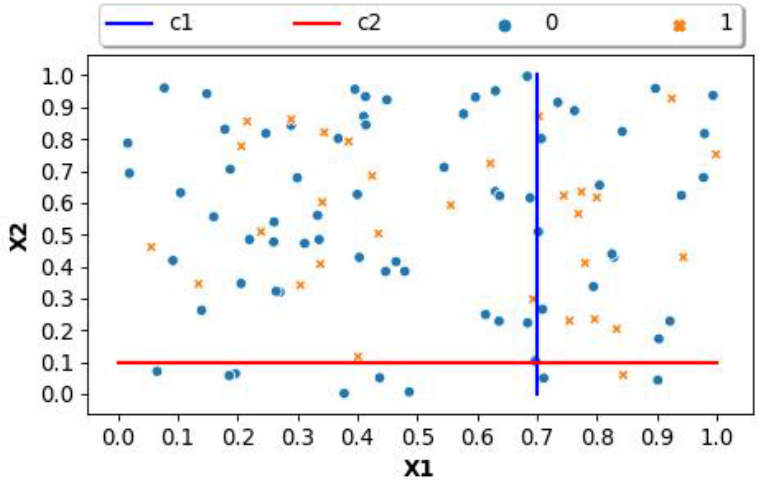
1. Motivation

❖ DT 알고리즘의 문제점 ① – Greedy splitting

- Impurity of the next level nodes in CART
  - Gini index : $g(S) = 1 - \sum {p_k}^2$
  - Gini gain : $g(S_l, S_r) = \frac{|S_l|}{|S|} g(S_l) + \frac{|S_r|}{|S|} g(S_r)$



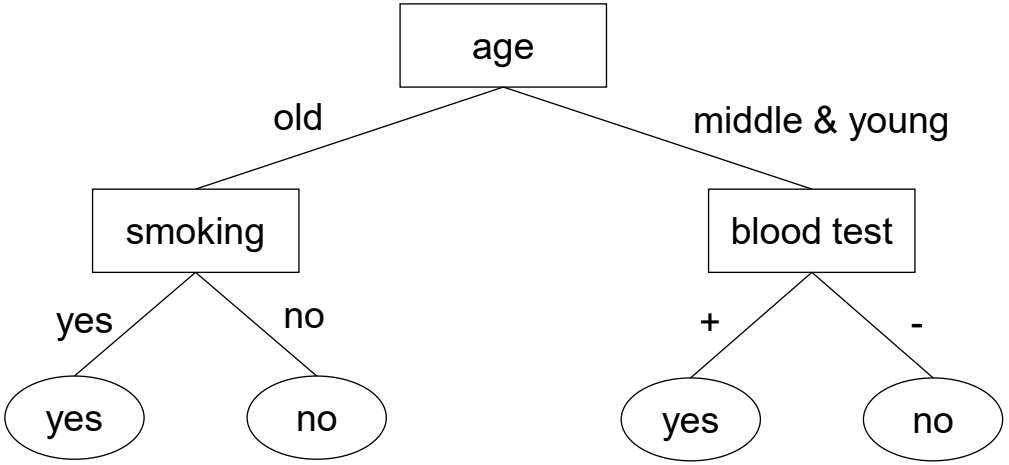
- 문제점
  - 양쪽 노드의 불순도 가중 합이 최소화되는 지점으로 분기함 → 두 개의 노드를 동시에 고려함으로 문제가 생김
  - CART는 양쪽 노드의 동질성을 고려해 c1으로 분기하나,
    한쪽 노드만 고려해보면 c2가 더 높은 동질성



1. Motivation

❖ DT 알고리즘의 문제점 ② – Unnecessary Condition

- Rule interdependence of DT
  - 위계적인 구조로 인해 하위 노드에서 불필요한 규칙을 무조건 포함하게 될 수 있음

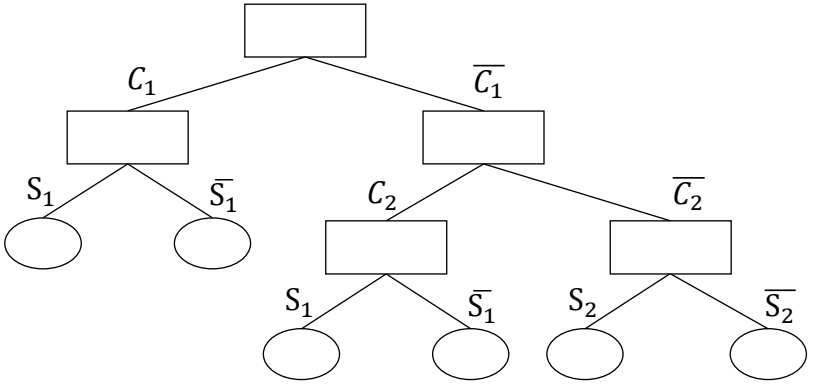


- 폐암 진단 예시
  - 뿌리 노드의 기준인 나이와 상관 없이, 폐암은 혈액 검사만으로도 진단이 될 수 있는데 위계적인 구조로 인해 나이가 항상 규칙에 포함됨.
  - 불필요한 규칙이 늘어나 사용자가 해석을 오도할 가능성이 있고 해석 상의 어려움이 발생

1. Motivation

❖ **DT 알고리즘의 문제점 ③ – All leaf nodes : mutually exclusive**

- 하나의 object를 규정하는 특징이 여러 개 일 수 있음 (Cause1 and Symptom1, Cause2 and Symptom2)
- DT의 모든 규칙은 negative association 문제
- 독립적인 규칙, positive association 관계의 규칙이 배제됨



❖ **DT 알고리즘의 문제점의 해결**

① Greedy splitting ⇒ OSM Tree

② Unnecessary Condition ⇒ Concise Rule Induction

③ All leaf nodes : mutually exclusive ⇒ Rule 규칙관계 분석

2. Backgrounds

❖ **Two types of Machine learning models (ML models)**

(1) White-box model (Interpretable model, transparent model)

- Decision Tree (DT)

- Rule Induction (RI)

- Linear regression (LR)

(2) Black-box model (Opaque model)

- Neural Network (NN)

- Ensemble Model (EM)

2. Backgrounds

SEOULTECH

❖ Concise Rule 관련연구

- DT & RI은 if-then rules로 표현됨
- DT Rule은 불필요한 조건들을 포함
- RI Rule separate & conquer
- Rule pruning은 predictive accuracy 관점에서 수행됨

❖ 규칙관계분석 관련연구

- Rule overlapping is avoided

(H.Lakkaraju – Interpretable Decision Sets: A Joint Framework for Description and Prediction 2016)

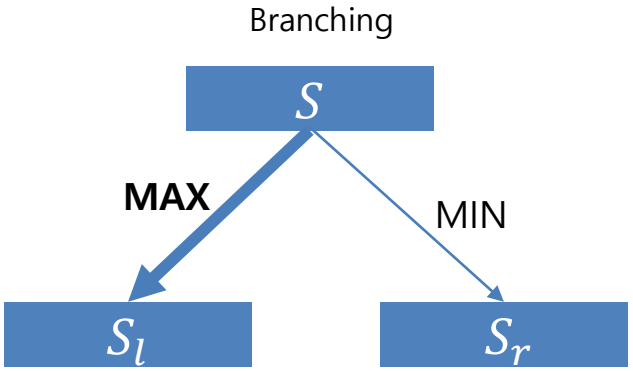
- Non – overlapping rules : independent rule (?)

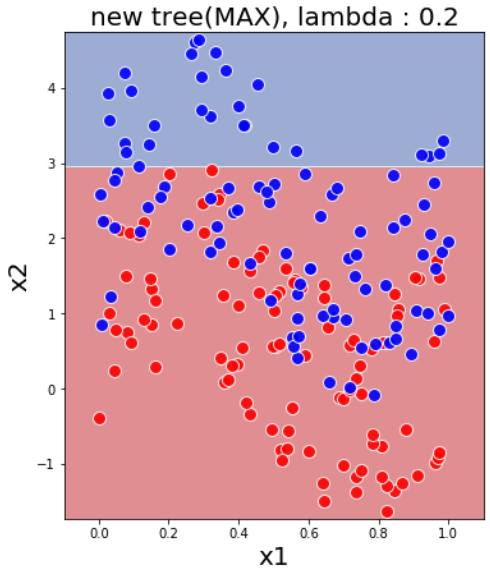
3. Concise Algorithm

❖ **새로운 분기 기준 - OSM(One-Sided-Maximum)**

New Splitting Criterion, S Hwang et. al, 2020

- 분기 시 한쪽 노드의 동질성만을 고려
- The study proposes a splitting criterion including the number of samples in each child node as following:
  - Proposed : $\max\left(\left(\frac{|S_l|}{|S|}\right)^{\lambda} g(S_l), \left(\frac{|S_r|}{|S|}\right)^{\lambda} g(S_r)\right)$, where $\lambda \in [0,1]$
- The hyperparameter λ controls the relative importance of the homogeneity and the sample coverage.





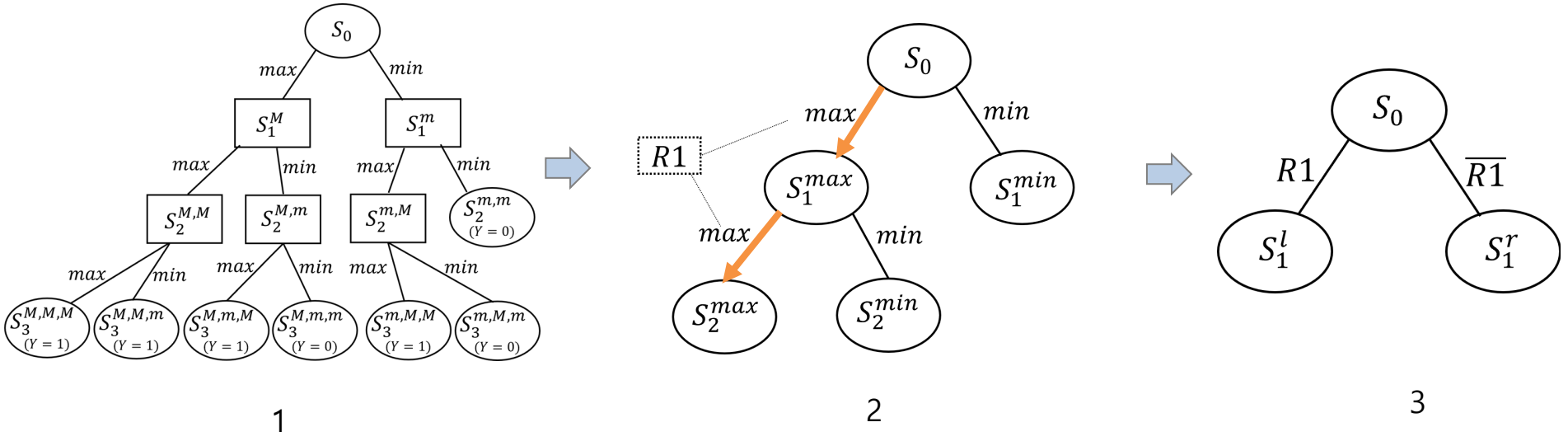
3. Concise Algorithm

❖ Rule Generation

- Concise Rule Induction consists of the set of serial algorithms (rule generation, concise, and then ensemble)
- **Rule Generation** (Separate and Conquer)

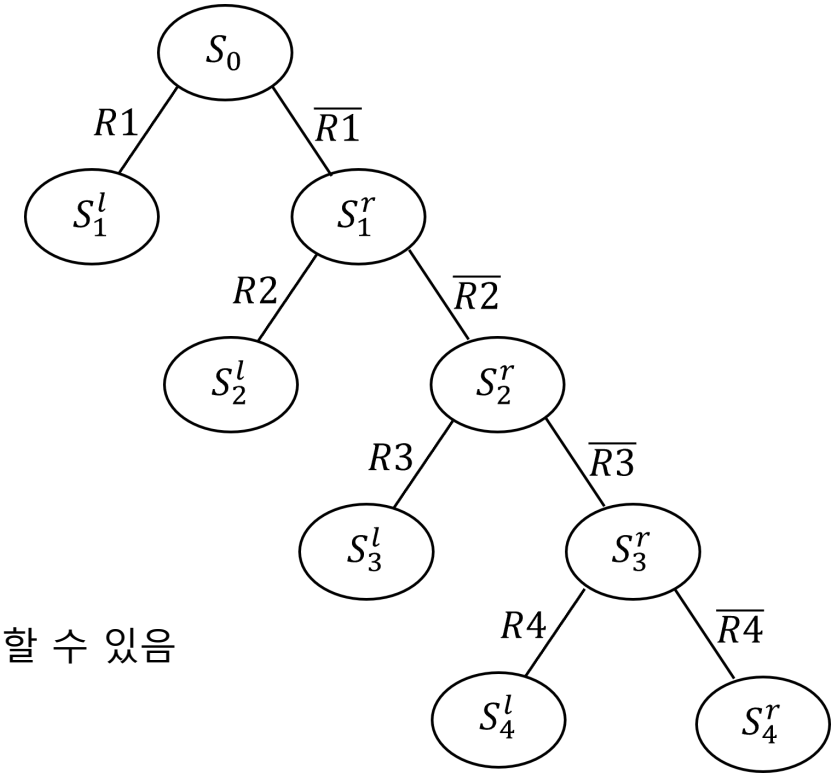
  - Max로 분기되지 않은 min은 해석력이 떨어지므로 루트노드로부터 max로만 분기된 규칙을 유도

    1. Generates a decision tree using OSM splitting criteria.
    2. We pick only one leaf node that is generated by following all maximal sides (R1).
    3. Repeat the above process for subset (S_1^r) that are not included in R1.



3. Concise Algorithm

- Rule generation process is completed as follows:



- 생성된 규칙
  - 1st Rule : (R1)
  - 2nd Rule : ($\overline{R}$1, R2)
  - 3rd Rule : ($\overline{R}$1, $\overline{R}$2, R3)
  - 4th Rule : ($\overline{R}$1, $\overline{R}$2, $\overline{R}$3, R4)
  - 5th Rule : ($\overline{R}$1, $\overline{R}$2, $\overline{R}$3, $\overline{R}$4)

- 첫번째 규칙을 제외한 다른 규칙에서 분리조건들은 불필요할 수 있음
  - ➢ Concise algorithm

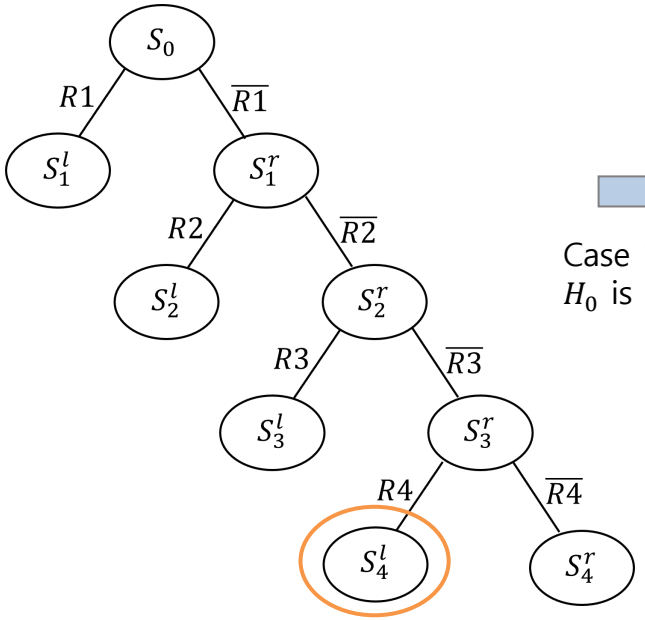
3. Concise Algorithm

- Concise algorithm
  - 불필요한 분리 조건들을 제거함으로써 규칙을 간결화
  - 분리 조건을 제거해본 후 통계적 검정을 통해 제거 여부를 판단

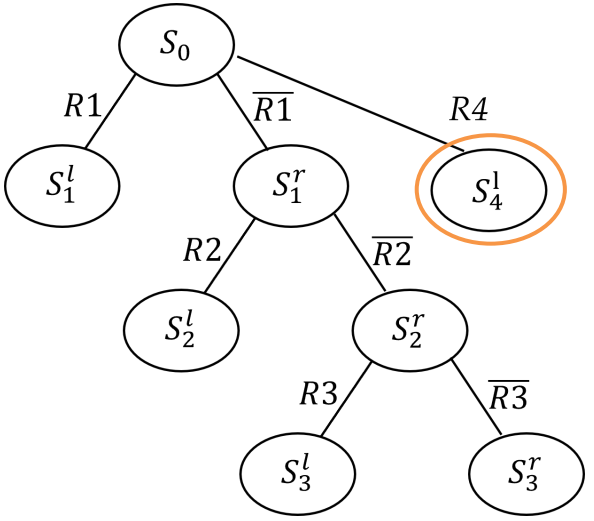
- Rule : $\bar{R}1$, $\bar{R}2$, $\bar{R}3$, … $\bar{R}(n-1)$, Rn
  - Step 1) 분리 조건 0개 test:
    - 통과 → 종료, Rule = Rn
    - 통과 실패 → step 2
  - Step 2) 분리 조건 1개 test:
    - 통과 조건 1개 → 종료, Rule = $\bar{R}i$ and Rn
    - 통과 조건 2개 → Coverage가 최대인 조건을 선택 -> 종료, Rule = $\bar{R}i^*$ and Rn
    - 통과 실패 → step 3
  - …
  - Step n-1) 분리 조건 n-2개 test : n-2개
    - 통과 조건 1개 → 종료, Rule = $\bar{R}i * \bar{R}j \ldots * Rn$
    - 통과 조건 2개 이상 → Coverage가 최대인 조건을 선택 → Rule = $\bar{R}i^* * \bar{R}j^* \ldots * Rn$
    - 통과 실패 → 종료, Rule = $\bar{R}1^* \bar{R}2 \ldots * Rn$

3. Concise Algorithm

- 예시 – 4th Rule의 분리 조건 제거 여부
  - 4th Rule : ($\overline{\text{R}}$1, $\overline{\text{R}}$2, $\overline{\text{R}}$3, R4), 분리 조건 : $\overline{\text{R}}$1, $\overline{\text{R}}$2, $\overline{\text{R}}$3
  - Let P[Y=y_0|$\overline{\text{R}}$1, $\overline{\text{R}}$2, $\overline{\text{R}}$3, R4] = p, y_0 is dominant response of the rule
  - Step 1) 0개 Test: Test on the necessity of ($\overline{\text{R}}$1, $\overline{\text{R}}$2, $\overline{\text{R}}$3)
    - H_0: P[Y = y_0|R4] ≥ p vs. H_1: not H_0
    - If H_0 is accepted(통과), then 4th Rule = R4
    - Else if H_0 is not accepted(통과 실패), then move step 2



Case 1)
H_0 is accepted in step 1



3. Concise Algorithm

SEOULTECH

❖ **Concise Rule Algorithm**

Algorithm : Concise rule

Input: $R_1, \cdots, R_k$

$S(R_1), \cdots, S(R_k)$

$TS_1, \cdots, TS_k$

$P(R_1/TS_1), \cdots, P(R_k/TS_k) \quad (P(R_i/TS_i) = P(y = f_i/S(R_i/TS_i)))$

Output: $R_{1^*}, \cdots, R_{k^*}$

$S(R_1), \cdots, S(R_k)$

$TS_1, \cdots, TS_k$

$P(R_{1^*}/TS_{1^*}), \cdots, P(R_{k^*}/TS_{k^*})$

for $i = 1$ *to* k **do**

  for $j = i + 1$ *to* k **do**

    Compute $Z_{i,j} = \dfrac{P(R_j/TS_i) - P(R_j/TS_j)}{\sqrt{\frac{P(R_j/TS_j)(1-P(R_j/TS_j))}{|S(R_j/TS_i)|}}}$

  end

end

for $j = 1$ *to* k **do**

  $j^* = \min_{1 \le i \le j-1} I[Z_{ij}, z_\alpha]$

  $R_{j^*} = R_j \cdot \bar{R}_1 \cdot \bar{R}_2 \cdots \bar{R}_{j^*-1}$

  $P(R_{j^*}/TS_{j^*}) = P(y = f_j/S(R_j/TS_{j^*}))$

end

4. Analysis of Rule Relation

❖ **Definition of Rule**

(1) Function

Domain : Subset of data set

Co-Domain : Class value

ex) if 성별 = 남성, 수입 >= 2,400만원 , then 대출 = 승인

(2) Subset (Set)

Fix the value of class variable

“대출 = 승인” rules

R1 : { S(R1) / 성별 = 남성, 수입 >= 2,400만원 }

R2 : { S(R2) / 직업 = 근로소득자, 수입 >= 2,350만원 }

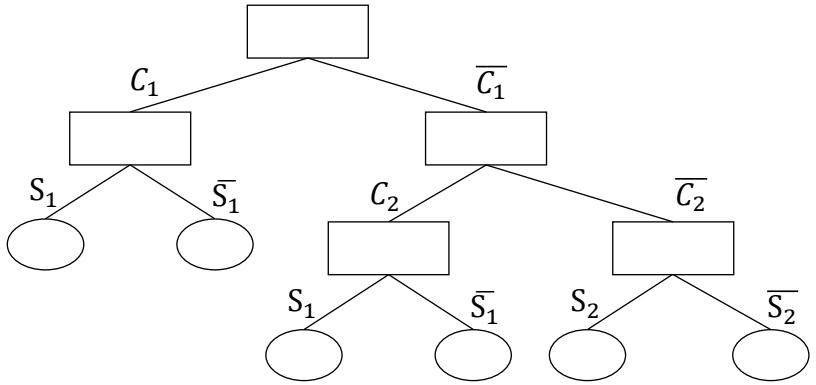
4. Analysis of Rule Relation

❖ **For a case of two causes and related symptoms**

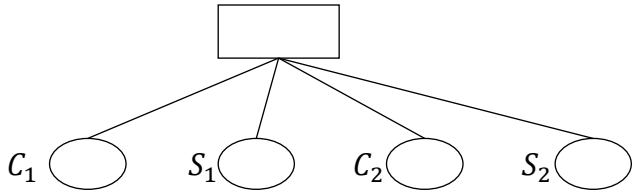
(Cause 1, Symptom 1)

(Cause 2, Symptom 2)

⇒ (C1 , S1) (C2 , S2) ⇒ Model



Possible DT presentation of 2 causes-symptoms



Possible RI presentation of 2 causes-symptoms

4. Analysis of Rule Relation

(2) Measure of Rule Association

(2.1) Overlapping Degree of Rules

$$L(R_i, R_j) = \frac{|S(R_i) \cap S(R_j)|}{|S(R_i) \cup S(R_j)|}$$

1

overlapping

non-overlapping ⇒ F-independent Rule

0

(2.2) Conditional Association Degree of Rules

$$A(R_i, R_j/y = y_k) = P(R_i \cap R_j/y = y_k) - P(R_i/y = y_K) \times P(R_j/y = y_k)$$

$A(R_i, R_j/y = y_k) \rightarrow$ - 0.25 R_i & R_j ⇒ Conditionally negative associative

$A(R_i, R_j/y = y_k) \rightarrow$ + 0.25 R_i & R_j ⇒ Conditionally positive associative

5. Experimental Result

❖ **Interpretability 비교 결과**

Dataset	Instances	Attributes		Classes
		Categorical	Numerical	
Adult	48,842	8	6	2
Australian	690		16	2
Avilar	20,867		10	12
BC-wisconsin	699	1	8	2
Car	1,728	6		4
Churn	5,000	4	15	2
Connect-4	67,557	42		3
EEG	14,980		14	2
Employee-attrition	1,470	26	8	2
GermanCredit	1,000	13	7	2
Image	2,310		19	7
Ionosphere	351		23	2
Letter-multi	20,000		16	26
Letter	20,000		16	2
Pima-indians-diabetes	768		8	2
Spambase	4,601		57	2
Splice-jxn	3,190	60		3
Waveform	5,000		21	3
Yeast	1,484	1	8	9

	n-th rule	CART	Hwang et al.(2020)	CN2	Proposed
# of Condition	1	1.49±0.55	**1.06±0.23**	2.19±1.20	1.69±1.23
	2	2.33±0.38	2.09±0.24	3.71±1.53	**1.73±0.88**
	3	3.05±0.33	3.14±0.35	4.04±1.70	**2.05±1.14**
	4	3.67±0.54	3.97±0.25	3.90±1.53	**2.54±1.70**
	5	4.15±0.74	4.85±0.47	4.02±1.67	**2.26±1.38**
	Average	2.94±0.95	3.02±1.34	3.57±0.70	**2.05±0.32**
Coverage	1	**0.28±0.19**	0.15±0.13	0.05±0.08	0.15±0.16
	2	**0.14±0.11**	0.12±0.10	0.05±0.01	0.14±0.10
	3	0.17±0.17	0.10±0.08	0.02±0.01	**0.17±0.15**
	4	0.08±0.04	0.07±0.04	0.01±0.01	**0.17±0.12**
	5	0.07±0.04	0.06±0.04	0.02±0.01	**0.20±0.15**
	Average	0.15±0.08	0.10±0.03	0.02±0.01	**0.17±0.02**
Homogeneity	1	0.90±0.09	0.92±0.12	**1.00±0.00**	0.93±0.11
	2	0.81±0.16	0.90±0.13	**1.00±0.00**	0.91±0.12
	3	0.81±0.11	0.88±0.12	**1.00±0.00**	0.88±0.12
	4	0.76±0.12	0.84±0.13	**1.00±0.00**	0.86±0.12
	5	0.70±0.13	0.82±0.13	**1.00±0.00**	0.83±0.13
	Average	0.79±0.06	0.87±0.04	**1.00±0.00**	0.88±0.04

6. Case Study (BC Data at UCI Repository)

❖ Wisconsin Breast Cancer Dataset Sample (Attribute 11개 x Record 683개)

Attribute	ID	Clump_ Thickness	Uniformity_of_ Cell_Size	Uniformity_of_ Cell_Shape	Marginal_ Adhesion	Single_Epithelial_ Cell_Size	Bare_ Nucleoli	Bland_ Chromatin	Normal_ Nucleoli	Mitoses	Class
Type	Integer (1-10)	Integer (1-10)	Integer (1-10)	Integer (1-10)	Integer (1-10)	Integer (1-10)	Integer (1-10)	Integer (1-10)	Integer (1-10)	Integer (1-10)	Integer (0:Benign, 1:Malignant)
1	1000025	5	1	1	1	2	1	3	1	1	0
2	1002945	5	4	4	5	7	10	3	2	1	0
⋮						⋮					
682	897471	4	8	6	4	3	4	10	6	1	1
683	897471	4	8	8	5	4	5	10	4	1	1

❖ Attribute Descriptions

No.	Attribute	Description
1	Clump_Thickness (덩어리 두께)	암세포의 그룹화를 나타내는 덩어리 두께
2	Uniformity_of_Cell_Size (세포 크기의 균일성)	세포 크기의 균일성 여부
3	Uniformity_of_Cell_Shape (세포 모양의 균일성)	세포 모양의 균일성 여부
4	Marginal_Adhesion (염색질의 유착성)	서로 분리된 세포가 부분적으로 유착되는 것으로, 유착의 기간이 오래되면 유방암 발병률이 높아짐
5	Single_Epithelial_Cell_Size (단일 상피세포 크기)	상피는 조직 표면에 있는 특수 세포의 얇은 층으로 단일 상피세포 크기가 커지면 악성 세포의 경우가 많음.
6	Bare_Nucleli (베어 핵)	세포질이 사실상 없는 세포학적 준비의 핵으로, 일반적으로 세포의 퇴화에서 볼 수 있음
7	Bland_Chromatin (블랜드 크로마틴)	보통 양성 세포에서 볼 수 있는 핵의 균일한(부드러운) "질감"을 이야기하며, 악성 세포(암)에서 염색질은 더 거칠어지는 경향이 있음
8	Normal_Nucleoli (정상핵)	정상 세포에서 핵은 보통 매우 작음
9	Mitoses (체세포분열)	세포가 분열하고 복제하는 과정으로 유사분열 수를 세어 암의 등급을 결정

6. Case Study (BC Data at UCI Repository)

SEOULTECH

❖ Dataset Summary by Class

Columns	Class	1	2	3	4	5	6	7	8	9	10	Total
Clump_Thickness	0 : Benign	136	46	92	67	83	15	1	4	0	0	444
	1: Maligant	3	4	12	12	45	18	22	40	14	69	239
	-	139	50	104	79	128	33	23	44	14	69	683
Uniformity_of_Cell_Size	0 : Benign	369	37	27	8	0	0	1	1	1	0	444
	1: Maligant	4	8	25	30	30	25	18	27	5	67	239
	-	373	45	52	38	30	25	19	28	6	67	683
Uniformity_of_Cell_Shape	0 : Benign	344	51	30	12	2	2	2	1	0	0	444
	1: Maligant	2	7	23	31	30	27	28	26	7	58	239
	-	346	58	53	43	32	29	30	27	7	58	683
Marginal_Adhesion	0 : Benign	363	37	31	5	4	3	0	0	0	1	444
	1: Maligant	30	21	27	28	19	18	13	25	4	54	239
	-	393	58	58	33	23	21	13	25	4	55	683
Single_Epithelial_Cell_Size	0 : Benign	43	355	28	7	5	1	2	2	0	1	444
	1: Maligant	1	21	43	41	34	39	9	19	2	30	239
	-	44	376	71	48	39	40	11	21	2	31	683
Bare_Nucleoli	0 : Benign	387	21	14	6	10	0	1	2	0	3	444
	1: Maligant	15	9	14	13	20	4	7	19	9	129	239
	-	402	30	28	19	30	4	8	21	9	132	683
Bland_Chromatin	0 : Benign	148	153	125	7	4	1	6	0	0	0	444
	1: Maligant	2	7	36	32	30	8	65	28	11	20	239
	-	150	160	161	39	34	9	71	28	11	20	683
Normal_Nucleoli	0 : Benign	391	30	11	1	2	4	2	3	0	0	444
	1: Maligant	41	6	31	17	17	18	14	20	15	60	239
	-	432	36	42	18	19	22	16	23	15	60	683
Mitoses	0 : Benign	431	8	2	0	1	0	1	1	0	0	444
	1: Maligant	132	27	31	12	5	3	8	7	0	14	239
	-	563	35	33	12	6	3	9	8	0	14	683

6. Case Study (BC Data at UCI Repository)

❖ **CART / CN2 / CRI _ Rule List (y = 1)**

R_i (y=1)	CRI (Concise Rule Induction)	CART	CN2
1	Uniformity_of_Cell_Size >= 4.5	Uniformity_of_Cell_Size > 2.5, Uniformity_of_Cell_Shape > 2.5, Marginal_Adhesion > 5.5	Uniformity_of_Cell_Size >= 5.0, Bland_Chromatin >= 5.0
2	Clump_Thickness >= 8.5	Uniformity_of_Cell_Size>2.5, Uniformity_of_Cell_Shape> 2.5, Marginal_Adhesion<= 5.5, Clump_Thickness> 6.5, Uniformity_of_Cell_Size> 4.5	Bare_Nucleoli >= 9.0, Uniformity_of_Cell_Shape >= 4.0
3	Uniformity_of_Cell_Size >= 1.5 Marginal_Adhension>= 6.5	Uniformity_of_Cell_Size>2.5, Uniformity_of_Cell_Shape> 2.5, Marginal_Adhesion<= 5.5, Clump_Thickness> 6.5, Uniformity_of_Cell_Size<= 4.5	Clump_Thickness >= 7.0, Clump_Thickness >= 9.0
4	Normal_Nucleoli >= 8.5	Uniformity_of_Cell_Size>2.5, Uniformity_of_Cell_Shape> 2.5, Marginal_Adhesion<= 5.5, Clump_Thickness<= 6.5, Bland_Chromatin> 4.5	Bare_Nucleoli >= 3.0, Clump_Thickness >= 5.0, Bare_Nucleoli <= 7.0

6. Case Study (BC Data at UCI Repository)

❖ **CART / CN2 / CRI _ Rule List (y = 0)**

R_i (y=0)	CRI (Concise Rule Induction)	CART	CN2
1	Uniformity_of_Cell_Shape < 1.5	Uniformity_of_Cell_Size <= 2.5, Bare_Nucleoli<= 2.5, Single_Epithelial_Cell_Size <= 2.5	Uniformity_of_Cell_Shape <= 3.0 Bare_Nucleoli <= 3.0, Bare_Nucleoli <= 2.0
2	Uniformity_of_Cell_Size < 1.5	Uniformity_of_Cell_Size <= 2.5, Bare_Nucleoli <= 2.5, Single_Epithelial_Cell_Size > 2.5	Normal_Nucleoli <= 3.0, Uniformity_of_Cell_Size <= 2.0, Clump_Thickness <= 4.0, Single_Epithelial_Cell_Size >= 2.0
3	Bland_Chromatin < 1.5	Uniformity_of_Cell_Size > 2.5 Uniformity_of_Cell_Shape <= 2.5	Marginal_Adhesion <= 7.0
4	Uniformity_of_Cell_Shape < 2.5	Uniformity_of_Cell_Size <= 2.5, Bare_Nucleoli > 2.5	-

6. Case Study (BC Data at UCI Repository)

❖ Rule Quality (y = 1)

- y=1 사전확률 : 239/683 = 0.409949

R_i (y=1)	Homogeneity			Coverage			Weighted relative Accuracy(WA)*			Justifiability		
	(1) CRI	(2)CART	(3)CN2	(1)CRI	(2)CART	(3)CN2	(1)CRI	(2)CART	(3) CN2	(1)CRI	(2)CART	(3)CN2
1	0.9828	1.0	1.0	0.2562	0.157	0.186	0.147	0.101	0.120	0	0	0
2	1.0	1.0	1.0	0.1215	0.084	0.17	0.072	0.054	0.110	0	1	0
3	1.0	0.90	1.0	0.1405	0.04	0.122	0.083	0.022	0.079	0	1	0
4	1.0	0.88	0.95	0.1098	0.032	0.088	0.065	0.017	0.053	0	1	1

❖ Rule Quality (y = 0)

- y=0 사전확률 : 444/683 = 0.761578

Ri (y=0)	Homogeneity			Coverage			Weighted relative Accuracy(WA)			Justifiability		
	(1)CRI	(2)CART	(3)CN2	(1)CRI	(2)CART	(3)CN2	(1)CRI	(2)CART	(3) CN2	(1)CRI	(2)CART	(3)CN2
1	0.9942	1.0	0.99	0.5065	0.529	0.582	0.118	0.1870	0.204	0	0	0
2	0.9892	0.95	1.0	0.5461	0.036	0.41	0.124	0.0110	0.145	0	1	1
3	0.9866	0.84	0.73	0.2196	0.035	0.884	0.049	0.0070	0.074	0	1	0
4	0.9772	0.65	-	0.5915	0.048	-	0.128	0.0003	-	0	1	-

* Weighted relative Accuracy(WA) = Coverage x (Homogeneity – 사전확률)

6. Case Study (BC Data at UCI Repository)

❖ Overlapping Degree of Rules

R_i & R_j (y=1)	$R_i \cup R_j$	$R_i \cap R_j$	Overlapping Degree	R_i & R_j (y=0)	$R_i \cup R_j$	$R_i \cap R_j$	Overlapping Degree
R_1 & R_2	195	60	0.308	R_1 & R_2	391	322	0.824
R_1 & R_3	188	80	0.426	R_1 & R_3	371	121	0.326
R_1 & R_4	184	63	0.342	R_1 & R_4	395	344	0.871
R_2 & R_3	152	27	0.178	R_2 & R_3	390	127	0.326
R_2 & R_4	134	24	0.179	R_2 & R_4	418	346	0.828
R_3 & R_4	133	38	0.286	R_3 & R_4	408	135	0.331

* Overlapping Degree : $P(R_i \cap R_j) / P(R_i \cup R_j)$

❖ Conditional Association Degree of Rules

Y (n=683)	Rule	Conditions	Y_k	$P(R_i/y=Y_k)$
y=1 (n=239)	Rule1	Uniformity_of_Cell_Size >= 4.5	172	0.720
	Rule2	Clump_Thickness >= 8.5	83	0.347
	Rule3	Uniformity_of_Cell_Size >= 1.5 Marginal_Adhension>= 6.5	96	0.402
	Rule4	Normal_Nucleoli >= 8.5	75	0.314
y=0 (n=444)	Rule1	Uniformity_of_Cell_Shape < 1.5	344	0.775
	Rule2	Uniformity_of_Cell_Size < 1.5	369	0.831
	Rule3	Bland_Chromatin < 1.5	148	0.333
	Rule4	Uniformity_of_Cell_Shape < 2.5	395	0.890

R_i & R_j (y=1)	Rules Association	R_i & R_j (y=0)	Rules Association
R_1 & R_2	0.001	R1 & R2	0.081
R_1 & R_3	0.046	R1 & R3	0.014
R_1 & R_4	0.038	R1 & R4	0.086
R_2 & R_3	-0.027	R2 & R3	0.009
R_2 & R_4	-0.009	R2 & R4	0.040
R_3 & R_4	0.033	R3 & R4	0.008

* Rules Association : $A(R_i, R_j/y = y_k) = P(R_i \cap R_j/y = y_k) - (P(R_i/y = y_K) \times P(R_j/y = y_K))$

7. Conclusion

❖ Experimental Result

- CART, Hwang, CN2, CRI 알고리즘을 통해 19개 데이터셋을 통해 생성된 Rule의 Condition 개수와, Coverage, Homogeneity를 비교
- CRI 알고리즘의 Condition 수는 2.05개로 다른 알고리즘과 비교했을 때 가장 작았음을 확인
- CRI 알고리즘의 Coverage가 0.17로 다른 알고리즘과 비교했을 때 가장 높았음을 확인
- CRI 알고리즘의 Homogeneity는 0.88로 CN2 보다 낮았지만, CART, Hwang 보다는 높았음을 확인

	CART	Hwang et al	CN2	Proposed
#of Condition	2.94±0.95	3.02±1.34	3.57±0.70	2.05±0.32
Coverage	0.15±0.08	0.10±0.03	0.02±0.01	0.17±0.02
Homogeneity	0.79±0.06	0.87±0.04	1.00±0.00	0.88±0.04

❖ BC Case Study Result

y	Homogeneity			Coverage			WA			Justifiability		
	(1) CRI	(2)CART	(3)CN2	(1)CRI	(2)CART	(3)CN2	(1)CRI	(2)CART	(3) CN2	(1)CRI	(2)CART	(3) CN2
1	0.996	0.945	0.988	0.157	0.078	0.142	0.092	0.049	0.091	0	1	1
0	0.987	0.860	0.907	0.466	0.162	0.625	0.105	0.051	0.141	0	1	1

- Concise 알고리즘의 효과 : y=1일 때 Rule당 Condition수가 CART 3.5개, CN2 2.4개, CRI 1.125개로 CRI알고리즘의 Condition수가 가장 적었음을 확인
- Rule association Quality 비교 : y=1인 Rule에서는 CRI 알고리즘의 Coverage, Homogeneity, Weight Accuracy가 제일 높았으며, 다른 알고리즘들과 달리 Rule의 Justifiability위반도 없었음을 확인