

Design of New Medicine Development Based on Conjoint Analysis and Rough Sets

杉原 一臣 河崎 誠 石井 博昭
Kazutomi Sugihara Makoto Kawasaki Hiroaki Ishii

大阪大学 大学院工学研究科 応用物理学専攻
Dept. of Applied Physics, Graduate School of Engineering, Osaka University

1 はじめに

近年、製薬会社では新薬開発の効率化が求められている。例えば、最も効果の高い成分の組み合わせを推定するために、少ないサンプルから効果の高い成分を特定することなどが必要となっている。そこで、開発の効率化を図るため多変量解析の活用がなされている。コンジョイント解析 (Conjoint Analysis) は、予め用意された諸要因の組合せに対する全体評価から、各要因の部分効用値を求める多変量解析モデルである。すなわち、このモデルでは最も効果高い要因の組合せを推定することができる。部分効用値を求める手法については、部分効用値から全体評価を構成するための結合法則に基づき、様々な手法が提案されているが、その最も代表的な手法に MONANOVA (Monotone Analysis of Variance) 法がある。河崎 [1] は MONANOVA 法を用いて、新薬開発へのコンジョイント解析の応用手法を提案している。しかし、MONANOVA 法では部分効用値の和を全体評価とする線形関数を仮定している。そのため、成分の組合せによる相互作用を反映しないという問題がある。

本論文では、ラフ集合 (Rough Sets) [2] を用いたコンジョイント解析を試みる。ラフ集合はデータ解析のための近似概念である。ラフ集合では、質的データを含むデータベース (情報システム) から IF-Then 形式の相関ルールが抽出される。要因の組合せというのは一種の質的データとして解釈できるので、コンジョイント解析にラフ集合を取り入れることは有効である。またデータの特徴を IF-Then 形式で抽出できるので、より柔軟な特徴表現が可能となる。また本提案手法では、Greco らが提案した Dominance relation [3] によるラフ集合を採用している。これは「新薬開発のための有望な成分の組合せを探す」という目的から、従来の同値関係によるラフ集合よりも Dominance relation によるラフ集合の方が適用しやすいからである。

2 情報システムの作成

Dominance relation によるデータ解析を行うには、属性値が順序関係を持っていなければならない。そこで、与えられた質的データから順序データに変換した情報システムの作成方法を説明する。区間回帰分析 [4] を用いて、各属性のカテゴリーについて順位づけを行う。そのため、以下のような区間モデルを考える。

$$Z = \sum_{i=1}^M (a_i^t x_i + c_i^t x_i) \quad (1)$$

ここで、 (a_i, c_i) ($i = 1, \dots, M$) は属性 i の区間効用ベクトルであり、それぞれ中心 $a_i = (a_{i1}, \dots, a_{in_i})^t$ と幅 $c_i = (c_{i1}, \dots, c_{in_i})^t$ を表している。 x_i はそのサンプルに属性 i で該当しているカテゴリー $k_i \in (1, \dots, n_i)$ の成分のみ 1 をとり、それ以外の値は 0 をとる 0-1 ベクトルである。また n_i は属性 i のカテゴリー数を表わす。

J 個の入力データ x_{ij} と出力データ (評価値) Z_j が与えられたとき、次の線形計画問題により属性 i でカテゴリー k_i の区間効用値 $(a_{ik_i}, c_{ik_i}) = [a_{ik_i} - c_{ik_i}, a_{ik_i} + c_{ik_i}]$ を求める。

$$\begin{aligned} \min_{a_i, c_i} \quad & \sum_{j=1}^J \sum_{i=1}^M c_i^t x_{ij} \\ \text{s.t.} \quad & \sum_{i=1}^M a_i^t x_{ij} - c_i^t x_{ij} \leq Z_j \\ & Z_j \leq \sum_{i=1}^M a_i^t x_{ij} + c_i^t x_{ij} \\ & a_i, c_i \geq 0 \quad (i = 1, \dots, M) \end{aligned} \quad (2)$$

次に、得られた区間効用値より各カテゴリーごとに項目の順位づけを行う。区間の順位づけには以下の区間選好関係の定義 [5] を用いる。すなわち、区間値 $A = [a, \bar{a}]$, $B = [b, \bar{b}]$ が与えられたとき、

$$A \succeq B \leftrightarrow a \geq b, \bar{a} \geq \bar{b}.$$

なお、半順序関係の場合は「2-3 位」というように複数の順位で表現する。各属性のカテゴリーに順位がつけられた後、その順位に対応して属性値を与える。すなわち、属性のカテゴリー数 N 個中「 $n - n'$ 位」であるカテゴリーの属性値を $[N - n' + 1, N - n + 1]$ という区間または実数値で与えるものとする。例えば、 $A : 1$ 位, $B : 2-3$ 位, $C : 2-3$ 位, $D : 4$ 位という順位であったとすると、この場合は $A : 4$, $B : [2, 3]$, $C : [2, 3]$, $D : 1$ という属性値が割り当てられる。

次に意思決定者が出力データを幾つかのクラスに分類する。 R 個のクラスに分類したときの r 番目のクラスを Cl_r ($r = 1, 2, \dots, R$) と表わし、以下の選好関係を満たしているとする。

$$Cl_R \succ \dots \succ Cl_r \succ \dots \succ Cl_1 \quad (3)$$

以上の操作により、情報システム $S = \langle U, C \cup \{d\}, V_C \cup V_d, f \rangle$ が形成される。ただし、 U はサンプルの集合、 C は属性の集合、 $\{d\}$ は決定を表わし、ここでは評価値を分類して割り当てたクラスを意味する。また、属性 $q (q \in C)$ に関する属性値集合を V_q 、決定 (クラス) の集合を V_d を表す。 f は $f : U \times (C \cup \{d\}) \rightarrow (V_C \cup V_d)$ となる関数である。

3 Dominance Relation によるラフ集合を用いたデータ解析

Dominance relation によるラフ集合を用いて、作成された情報システムのデータ解析を行う。Dominance relation は以下のように定義されている。

定義 1. (Dominance relation[3])

部分集合 $P \subseteq C$ について、"y が x を支配している" という関係を $yD_P x$ と表し、以下のように定義する。

$$yD_P x \leftrightarrow \forall q \in P, f(y, q) \geq f(x, q) \quad (4)$$

ただし、 $f(x, q)$ はサンプル x が持つ属性 q の属性値を表している。ここでは $f(x, q)$ が区間値をとることもあるので、(4) 式を次のように書き換える。

$$yD'_P x \leftrightarrow \forall q \in P, \min_{u \in f(y, q)} u \geq \max_{v \in f(x, q)} v \quad (5)$$

D'_P より、 x に対する支配集合 $D_P^+(x)$ と被支配集合 $D_P^-(x)$ が次のように定義できる。

$$D_P^+(x) = \{y \in U | yD'_P x\} \quad (6)$$

$$D_P^-(x) = \{y \in U | xD'_P y\} \quad (7)$$

次にクラスの累積集合を定義する。

定義 2. 累積集合 (Cumulative set[3])

$$Cl_t^{\geq} = \bigcup_{s \geq t} Cl_s \quad (8)$$

$$Cl_t^{\leq} = \bigcup_{s \leq t} Cl_s \quad (9)$$

$Cl_t^{\geq}$ と $Cl_t^{\leq}$ をそれぞれ上方累積集合 (Upward cumulative set), 加法累積集合 (Downward cumulative set) と呼ぶ。累積集合はそれぞれ以下のように解釈されている。

$$x \in Cl_t^{\geq} \leftrightarrow "x \text{ は少なくともクラス } t \text{ に属している}"$$

$$x \in Cl_t^{\leq} \leftrightarrow "x \text{ はたかだかクラス } t \text{ に属している}"$$

$D_P^+(x)$ と $D_P^-(x)$ により累積集合 $Cl_t^{\geq}, Cl_t^{\leq}$ を近似する。 $D_P^+(x)$ による $Cl_t^{\geq}$ の上界・下界近似集合は以下の通りである。

$$P(Cl_t^{\geq}) = \{x \in U | (D_P^+(x) \cup \{x\}) \subseteq Cl_t^{\geq}\} \quad (10)$$

$$\bar{P}(Cl_t^{\geq}) = \{x \in U | (D_P^-(x) \cup \{x\}) \cap Cl_t^{\geq} \neq \emptyset\} \quad (11)$$

同様に、 $D_P^-(x)$ による $Cl_t^{\leq}$ の上界・下界近似集合は以下の通りである。

$$P(Cl_t^{\leq}) = \{x \in U | (D_P^-(x) \cup \{x\}) \subseteq Cl_t^{\leq}\} \quad (12)$$

$$\bar{P}(Cl_t^{\leq}) = \{x \in U | (D_P^+(x) \cup \{x\}) \cap Cl_t^{\leq} \neq \emptyset\} \quad (13)$$

これらの下界近似集合から以下のような3つのタイプの IF-Then ルールが得られる。

- If $\min_{u_1 \in f(x, q_1)} u_1 \geq \max_{v_1 \in f(x^*, q_1)} v_1$ and $\dots$ and $\min_{u_M \in f(x, q_M)} u_M \geq \max_{v_M \in f(x^*, q_M)} v_M$, then $x \in Cl_t^{\geq}$.

- If $\max_{u_1 \in f(x, q_1)} u_1 \leq \min_{v_1 \in f(x^*, q_1)} v_1$ and $\dots$ and $\max_{u_M \in f(x, q_M)} u_M \leq \min_{v_M \in f(x^*, q_M)} v_M$, then $x \in Cl_t^{\leq}$.

- If $\min_{u_1 \in f(x, q_1)} u_1 \geq \max_{v_1 \in f(x^*, q_1)} v_1$ and $\dots$ and $\min_{u_{p'} \in f(x, q_{p'})} u_{p'} \geq \max_{v_{p'} \in f(x^*, q_{p'})} v_{p'}$ and $\max_{u_{p'+1} \in f(x, q_{p'+1})} u_{p'+1} \leq \min_{v_{p'+1} \in f(x^*, q_{p'+1})} v_{p'+1}$ and $\dots$ and $\max_{u_M \in f(x, q_M)} u_M \leq \min_{v_M \in f(x^*, q_M)} v_M$, then $x \in Cl_s \cup Cl_{s+1} \cup \dots \cup Cl_t$.

ただし、 $O = \{q_1, \dots, q_{p'}\}$, $O' = \{q_{p'+1}, \dots, q_M\}$ とすると、 $O \cup O' = C$ かつ $O \cap O' \neq \emptyset$ である。また、 x^* は得られたルールを支持するサンプルを表す。

4 数値例

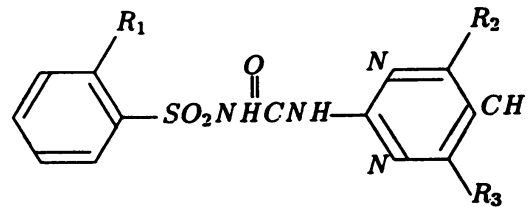


図 1. 母体となる化合物

図 1 のような化合物を母体とした新薬開発問題を考える。この化合物の特定の位置 R_1, R_2, R_3 に様々な置換機を結合させる。22 のサンプルについて、表 1 のような実験結果が得られた。この結果から、どのような組合せが有望であるかを推定する。

区間回帰分析を用いて得られた区間効用値と、その順位に基づいて与えられた属性値を表 2-4 に示す。また、意思決定者が活性値について表 5 のような 5 つのクラスに分類したとする。得られた属性値とクラスにより、表 6 のような情報システムが作成された。

$D_P^+(x), D_P^-(x)$ による累積集合 $Cl_t^{\geq}, Cl_t^{\leq}$ の上下界近似集合は以下の通りである。

$$P(Cl_2^{\geq}) = U - \{17, 18, 19, 20\}$$

$$P(Cl_3^{\geq}) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 21\}$$

$$P(Cl_4^{\geq}) = \{1, 2, 3, 4, 5, 6, 7\}$$

$$P(Cl_5^{\geq}) = \{1, 3\}$$

$$\bar{P}(Cl_2^{\geq}) = U - \{17, 18, 19, 20\}$$

$$\bar{P}(Cl_3^{\geq}) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 21\}$$

$$\bar{P}(Cl_4^{\geq}) = \{1, 2, 3, 4, 5, 6, 7, 8, 10, 21\}$$

$$\bar{P}(Cl_5^{\geq}) = \{1, 2, 3, 4\}$$

$$P(Cl_1^{\leq}) = \{17, 18, 19, 20\}$$

$$P(Cl_2^{\leq}) = \{12, 13, 14, 15, 16, 17, 18, 19, 20, 22\}$$

$$P(Cl_3^{\leq}) = \{9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22\}$$

表 1: 置換基の組合せと実験結果

サンプル	R_1	R_2	R_3	活性値
1	COOEt	Cl	OCH ₃	7.5751
2	CH ₃	OCH ₃	OCH ₃	7.0700
3	COOCH ₃	OCH ₃	H	6.5638
4	COOCH ₃	CH ₃	OCH ₃	6.4353
5	COOEt	OCH ₃	H	6.2700
6	Cl	CH ₃	OCH ₃	6.1040
7	COOEt	OCH ₃	H	6.0991
8	NO ₂	OCH ₃	H	5.8961
9	CH ₃	CH ₃	H	5.1475
10	COOEt	CH ₃	H	5.1290
11	Cl	CH ₃	H	4.9066
12	CH ₃	Cl	H	4.2874
13	CH ₃	H	H	4.0726
14	Cl	H	H	3.9666
15	CH ₃	H	H	3.9547
16	CH ₃	H	CH ₃	3.7696
17	COOEt	H	CH ₃	3.4685
18	NO ₂	OCH ₃	H	3.4293
19	CH ₃	OCH ₃	H	3.3925
20	Cl	OCH ₃	H	3.1278
21	NO ₂	CH ₃	H	6.3840
22	COOEt	H	H	4.2757

$$P(Cl_4^{\leq}) = U - \{1, 2, 3, 4\}$$

$$\bar{P}(Cl_1^{\leq}) = \{17, 18, 19, 20\}$$

$$\bar{P}(Cl_2^{\leq}) = \{12, 13, 14, 15, 16, 17, 18, 19, 20, 22\}$$

$$\bar{P}(Cl_3^{\leq}) = U - \{1, 2, 3, 4, 5, 6, 7\}$$

$$\bar{P}(Cl_4^{\leq}) = U - \{1, 3\}$$

ただし, $P(Cl_1^{\leq}), \bar{P}(Cl_1^{\leq}), P(Cl_2^{\leq}), \bar{P}(Cl_2^{\leq})$ は常に

$$P(Cl_1^{\leq}) = \bar{P}(Cl_1^{\leq}) = U,$$

$$P(Cl_2^{\leq}) = \bar{P}(Cl_2^{\leq}) = U$$

である。

本論文では, 下界近似集合 $P(Cl_1^{\leq}), P(Cl_2^{\leq})$ より得られるルールのみを用いる。下界近似集合から得られた全てのルールを表 7 と表 8 に示す。例えば, 表 7 の " $R_1 : \geq 5, R_2 : \geq 3, R_3 : \geq 3, d : Cl_5^{\leq}$ " は " R_1 の属性値が 5 以上, R_2 の属性値が 3 以上, R_3 の属性値が 3 以上のカテゴリーの成分であれば, その活性値は少なくともクラス 5 である" というルールを表している。また, 表 8 の " $R_1 : \leq 4, R_2 : \leq 2, R_3 : \leq 3, d : Cl_5^{\leq}$ " は " R_1 の属性値が 5 以下, R_2 の属性値が 3 以下, R_3 の属性値が 3 以下のカテゴリーの成分であれば, その活性値はたかだかクラス 5 である" というルールを表している。このように, 各々のサンプルから IF-THEN ルールが得られる。

本論文の目的は効果の高い成分の組み合わせを探すことであるので, $Cl_4^{\leq}$ や $Cl_5^{\leq}$ といった高いクラスに属するという結論を導くルールを調べる。すなわち, 表 7 から次のようなルールが挙げられる。

• If $f(R_1, x) \geq 5$ and $f(R_2, x) \geq 3$ and $f(R_3, x) \geq 3$, then $x \in Cl_5^{\leq}$.

• If $f(R_1, x) \geq 2$ and $f(R_2, x) \geq 3$ and $f(R_3, x) \geq 2$, then $x \in Cl_4^{\leq}$.

• If $f(R_1, x) \geq 5$ and $f(R_2, x) \geq 3$ and $f(R_3, x) \geq 2$, then $x \in Cl_5^{\leq}$.

• If $f(R_1, x) \geq 5$ and $f(R_2, x) \geq 4$ and $f(R_3, x) \geq 3$, then $x \in Cl_4^{\leq}$.

• If $f(R_1, x) \geq 5$ and $f(R_2, x) \geq 3$ and $f(R_3, x) \geq 2$, then $x \in Cl_4^{\leq}$.

• If $f(R_1, x) \geq 1$ and $f(R_2, x) \geq 4$ and $f(R_3, x) \geq 3$, then $x \in Cl_4^{\leq}$.

しかし R_1 の属性値が 5 以上であるため, $Cl_5^{\leq}$ に該当する組み合わせは存在しない。そこで $Cl_4^{\leq}$ という結論を導くルールに注目すると, R_1 の $\{Cl, COOEt, CH_3, COOCH_3, NO_2\}$ の内のいずれかと, R_2 の $\{CH_3\}$, R_3 の $\{OCH_3\}$ の組合せが有望であると推定される。すなわち, R_1 の属性値がどのような成分であっても, R_2, R_3 の属性値が高い成分との組み合わせであれば高い活性値が得られるということがわかる。

5 おわりに

本論文では, 新薬開発のためにラフ集合を用いたコンジョイント解析を行った。すなわち, 区間回帰分析により属性のカテゴリー間に順序選好関係を与えることで, 質的データを順序データに変換し, Dominance Relation によるラフ集合を用いて分析する手法を提案した。

参考文献

- [1] 河崎 誠 : 新薬開発のためのコンジョイント解析, 2000 年度大阪大学工学部応用物理学卒業論文
- [2] Z.Pawlak : Rough Classification, *International Journal of Man-Machine Studies*, 20, 469-483, 1984
- [3] S.Greco, B.Matarazzo and R.Slowinski : Rough sets theory for multicriteria decision analysis, *European Journal of Operational Research*, 129, 1-47, 2001
- [4] H.Tanaka and P.Guo : *Possibilistic Data Analysis for Operations Research*, Physica-Verlag, Heidelberg, 1999.
- [5] D.Dubois and H.Prade : *Systems of linear fuzzy constraints*, *International Journal of Fuzzy Sets and Systems*, 3, 37-48, 1980

表 2:得られた区間部分効用値と
与えられる属性値 (R_1)

カテゴリー	区間部分効用値	属性値
$COOEt$	[0.5857, 1.5086]	[4, 5]
CH_3	0.2647	2
$COOCH_3$	[0.3313, 1.8024]	[4, 5]
Cl	[0.0000, 0.0238]	1
NO_2	[0.3015, 1.5012]	3

表 3:得られた区間部分効用値と
与えられる属性値 (R_2)

カテゴリー	区間部分効用値	属性値
Cl	0.8949	[2, 3]
OCH_3	[0.0000, 1.6337]	[1, 3]
CH_3	[1.4155, 1.7550]	4
H	[0.5622, 0.8150]	[1, 2]

表 4:得られた区間部分効用値と
与えられる属性値 (R_3)

カテゴリー	区間部分効用値	属性値
OCH_3	[4.6885, 5.1717]	3
CH_3	[2.3206, 2.6899]	1
H	3.1278	2

表 5:活性値の分類クラス

クラス	範囲
Cl_5	[6.5-]
Cl_4	[5.5, 6.5]
Cl_3	[4.5, 5.5]
Cl_2	[3.5, 4.5]
Cl_1	[-3.5]

表 6:表 1 より作成された情報システム

サンプル	R1	R2	R3	クラス
1	[4, 5]	[2, 3]	3	Cl_5
2	2	[1, 3]	3	Cl_5
3	[4, 5]	[1, 3]	2	Cl_5
4	[4, 5]	4	3	Cl_4
5	[4, 5]	[1, 3]	2	Cl_4
6	1	4	3	Cl_4
7	[4, 5]	[1, 3]	2	Cl_4
8	3	[1, 3]	2	Cl_4
9	2	4	2	Cl_3
10	[4, 5]	4	2	Cl_3
11	1	4	2	Cl_3
12	2	[2, 3]	2	Cl_2
13	2	[1, 2]	2	Cl_2
14	1	[1, 2]	2	Cl_2
15	2	[1, 2]	2	Cl_2
16	2	[1, 2]	1	Cl_2
17	[4, 5]	[1, 2]	1	Cl_1
18	3	[1, 3]	2	Cl_1
19	2	[1, 3]	2	Cl_1
20	1	[1, 3]	2	Cl_1
21	3	4	2	Cl_4
22	[4, 5]	[1, 2]	2	Cl_2

表 7: D_p^+ による下界近似集合から
得られた IF-THEN ルール

ルール	R1	R2	R3	{d}
1	≥ 5	≥ 3	≥ 3	$Cl_5^{\geq}$
2	≥ 2	≥ 3	≥ 3	$Cl_4^{\geq}$
3	≥ 5	≥ 3	≥ 2	$Cl_5^{\geq}$
4	≥ 5	≥ 4	≥ 3	$Cl_4^{\geq}$
5	≥ 5	≥ 3	≥ 2	$Cl_5^{\geq}$
6	≥ 1	≥ 4	≥ 3	$Cl_4^{\geq}$
7	≥ 5	≥ 3	≥ 2	$Cl_4^{\geq}$
8	≥ 3	≥ 3	≥ 2	$Cl_5^{\geq}$
9	≥ 2	≥ 4	≥ 2	$Cl_5^{\geq}$
10	≥ 5	≥ 4	≥ 2	$Cl_5^{\geq}$
11	≥ 1	≥ 4	≥ 2	$Cl_5^{\geq}$
12	≥ 2	≥ 3	≥ 2	$Cl_5^{\geq}$
13	≥ 2	≥ 2	≥ 2	$Cl_5^{\geq}$
14	≥ 1	≥ 2	≥ 2	$Cl_5^{\geq}$
15	≥ 2	≥ 2	≥ 2	$Cl_5^{\geq}$
16	≥ 2	≥ 2	≥ 1	$Cl_5^{\geq}$
17	≥ 5	≥ 2	≥ 1	$Cl_5^{\geq}$
18	≥ 3	≥ 3	≥ 2	$Cl_5^{\geq}$
19	≥ 2	≥ 3	≥ 2	$Cl_5^{\geq}$
20	≥ 1	≥ 3	≥ 2	$Cl_5^{\geq}$
21	≥ 3	≥ 4	≥ 2	$Cl_5^{\geq}$
22	≥ 5	≥ 2	≥ 2	$Cl_5^{\geq}$

表 8: D_p^- による下界近似集合から
得られた IF-THEN ルール

ルール	R1	R2	R3	{d}
1	≤ 4	≤ 2	≤ 3	$Cl_5^{\leq}$
2	≤ 2	≤ 1	≤ 3	$Cl_5^{\leq}$
3	≤ 4	≤ 1	≤ 2	$Cl_5^{\leq}$
4	≤ 4	≤ 4	≤ 3	$Cl_5^{\leq}$
5	≤ 4	≤ 1	≤ 2	$Cl_4^{\leq}$
6	≤ 1	≤ 4	≤ 3	$Cl_4^{\leq}$
7	≤ 4	≤ 1	≤ 2	$Cl_4^{\leq}$
8	≤ 3	≤ 1	≤ 2	$Cl_4^{\leq}$
9	≤ 2	≤ 4	≤ 2	$Cl_5^{\leq}$
10	≤ 4	≤ 4	≤ 2	$Cl_4^{\leq}$
11	≤ 1	≤ 4	≤ 2	$Cl_5^{\leq}$
12	≤ 2	≤ 2	≤ 2	$Cl_5^{\leq}$
13	≤ 2	≤ 1	≤ 2	$Cl_5^{\leq}$
14	≤ 1	≤ 1	≤ 2	$Cl_5^{\leq}$
15	≤ 2	≤ 1	≤ 2	$Cl_5^{\leq}$
16	≤ 2	≤ 1	≤ 1	$Cl_5^{\leq}$
17	≤ 4	≤ 1	≤ 1	$Cl_5^{\leq}$
18	≤ 3	≤ 1	≤ 2	$Cl_5^{\leq}$
19	≤ 2	≤ 1	≤ 2	$Cl_5^{\leq}$
20	≤ 1	≤ 1	≤ 2	$Cl_5^{\leq}$
21	≤ 3	≤ 4	≤ 2	$Cl_5^{\leq}$
22	≤ 4	≤ 1	≤ 2	$Cl_5^{\leq}$