Решетки замкнутых множеств для наук о данных

Сергей Олегович Кузнецов

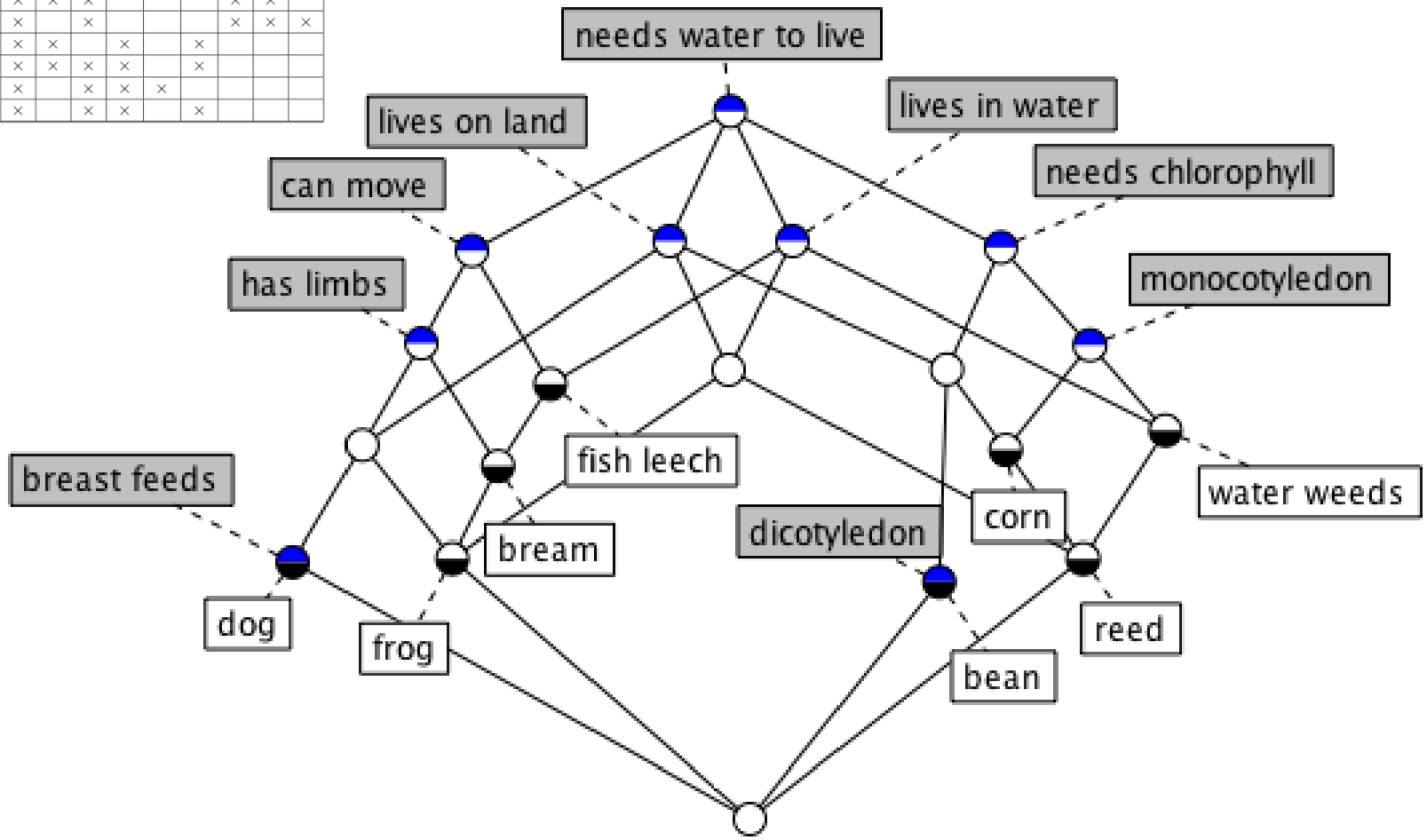
Департамент анализа данных и искусственного интеллекта,
Факультет Компьютерных Наук,
Национальный исследовательский университет Высшая школа экономики,
Москва, 2014

Пример. "Жизнь в воде"

	needs water to live	lives in water	lives on land	needs chlorophyll	dicotyledon	monocotyledon	can move	has limbs	breast feeds
fish leech	×	×					×		
bream	×	×					×	×	
frog	×	×	×				×	×	
dog	×		×				×	×	×
water weeds	×	×		×		×			
reed	×	×	×	×		×			
bean	×		×	×	×				
corn	×		×	×		×			

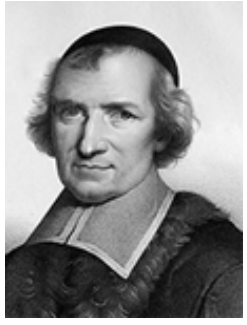
Пример. "Жизнь в воде". Диаграмма

	needs water to live	lives in water	lives on land	needs chlorophyll	dicotyledon	monocotyledon	can move	has limbs	breast feeds
fish leech	×	×					×		
bream	×	×					×	×	
frog	×	×	×				×	×	
dog	×		×				×	×	×
water weeds	×	×		×		×			
reed	×	×	×	×		×			
bean	×		×	×	×				
corn	×		×	×		×			



Решетки замкнутых множеств. Предыстория.

- A.Arnauld, P.Nicole, Logique de Port-Royal (1662)



Antoine Arnauld

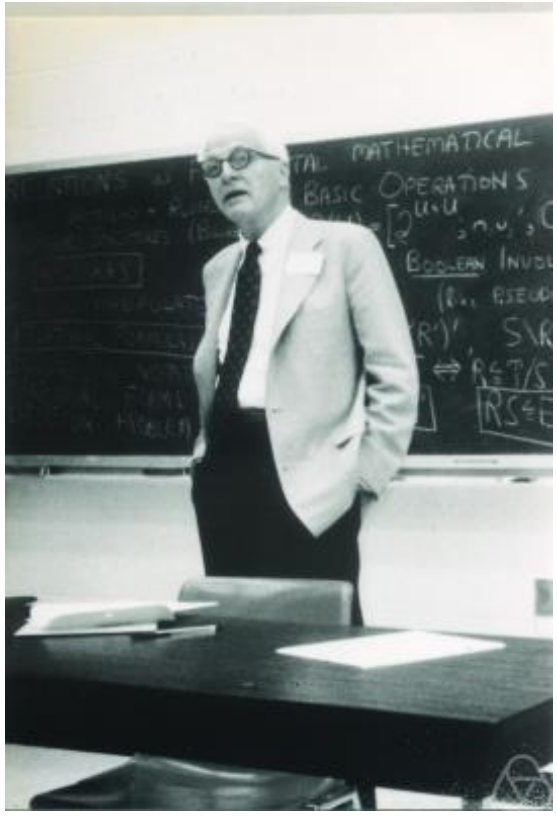
Pierre Nicole

Решетки замкнутых множеств. Предыстория.

- A.Arnauld, P.Nicole, Logique de Port-Royal (1662)
- E. Galois (1811-1832)

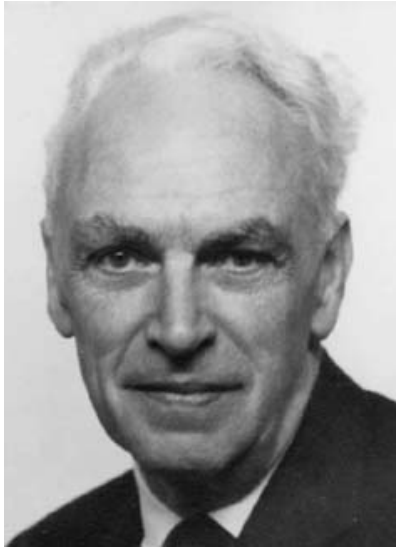
Решетки замкнутых множеств. Предыстория.

- A.Arnauld, P.Nicole, Logique de Port-Royal (1662)
- E. Galois (1811-1832)
- G. Birkhoff, начиная с 1930х



Решетки замкнутых множеств. Предыстория.

- A.Arnauld, P.Nicole, Logique de Port-Royal (1662)
- E. Galois (1811-1832)
- G. Birkhoff, начиная с 1930х
- O. Øre, начиная с 1930х



Решетки замкнутых множеств. Предыстория.

- A.Arnauld, P.Nicole, Logique de Port-Royal (1662)
- E. Galois (1811-1832)
- G. Birkhoff, начиная с 1930х
- O. Øre, начиная с 1930х
- M. Barbut, B. Monjardet, Ordre et classification, Hachette, Paris, 1970

Bernard Monjardet

Анализ формальных понятий (Formal Concept Analysis).

- R. Wille, Restructuring lattice theory: An approach based on hierarchies of concepts, 1982

Анализ формальных понятий (Formal Concept Analysis).

- R. Wille, Restructuring lattice theory: An approach based on hierarchies of concepts, 1982
- B. Ganter, R. Wille, Formale Begriffsanalyse, Springer, 1996
- B. Ganter, R. Wille, Formal Concept Analysis, Springer, 1999
- Глава в книге B. Davey, H. Priestly, Introduction to Order and Lattices, 1990.
- Глава в книге G. Grätzer (Ed.), General Lattice Theory.
- Concept Data Analysis, C.Carpineto, G. Romano, 2004.
- Galois Connections and Applications, K. Denecke, M. Erné, S. L. Wismath (Eds.), Springer Science & Business Media, 2004

FCA. Основные конференции

- International Conference on Conceptual Structures (ICCS), FCA participation starting from 1996 (Proceedings in LNAI, Springer)

- International Conference on Formal Concept Analysis (ICFCA), from 2003 года (Proceedings in LNAI, Springer)

- International Conference on Concept Lattices and Their Applications (CLA), from 2006, special issues

Решетки

Частично-упорядоченное множество $(L, \leq)$ называется **решеткой**, если для любой пары элементов $x, y \in L$ существуют супремум $\sup\{x, y\}$ и инфимум $\inf\{x, y\}$.

Решетки

Частично-упорядоченное множество $(L, \leq)$ называется **решеткой**, если для любой пары элементов $x, y \in L$ существуют супремум $\sup\{x, y\}$ и инфимум $\inf\{x, y\}$.

диаграмма частичного порядка **не** являющегося решеткой

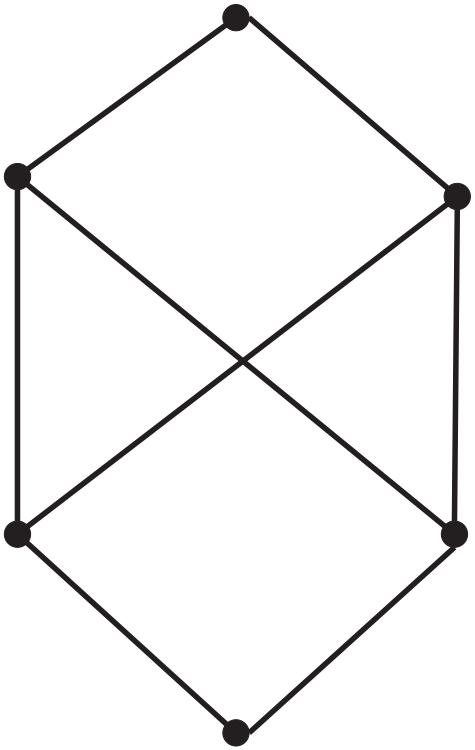
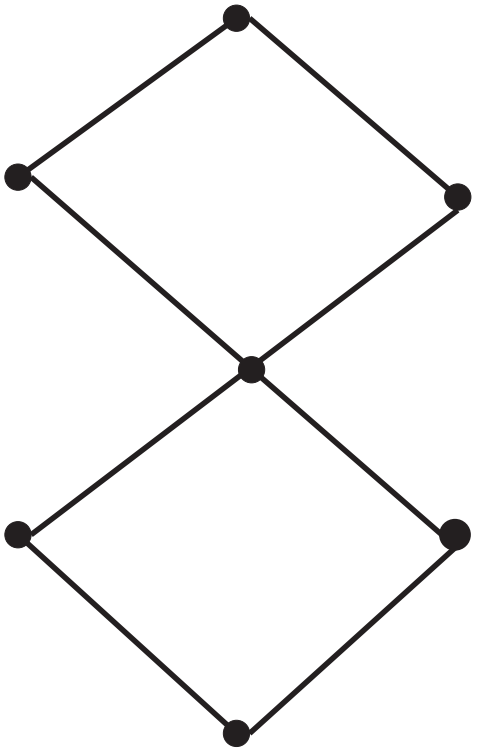
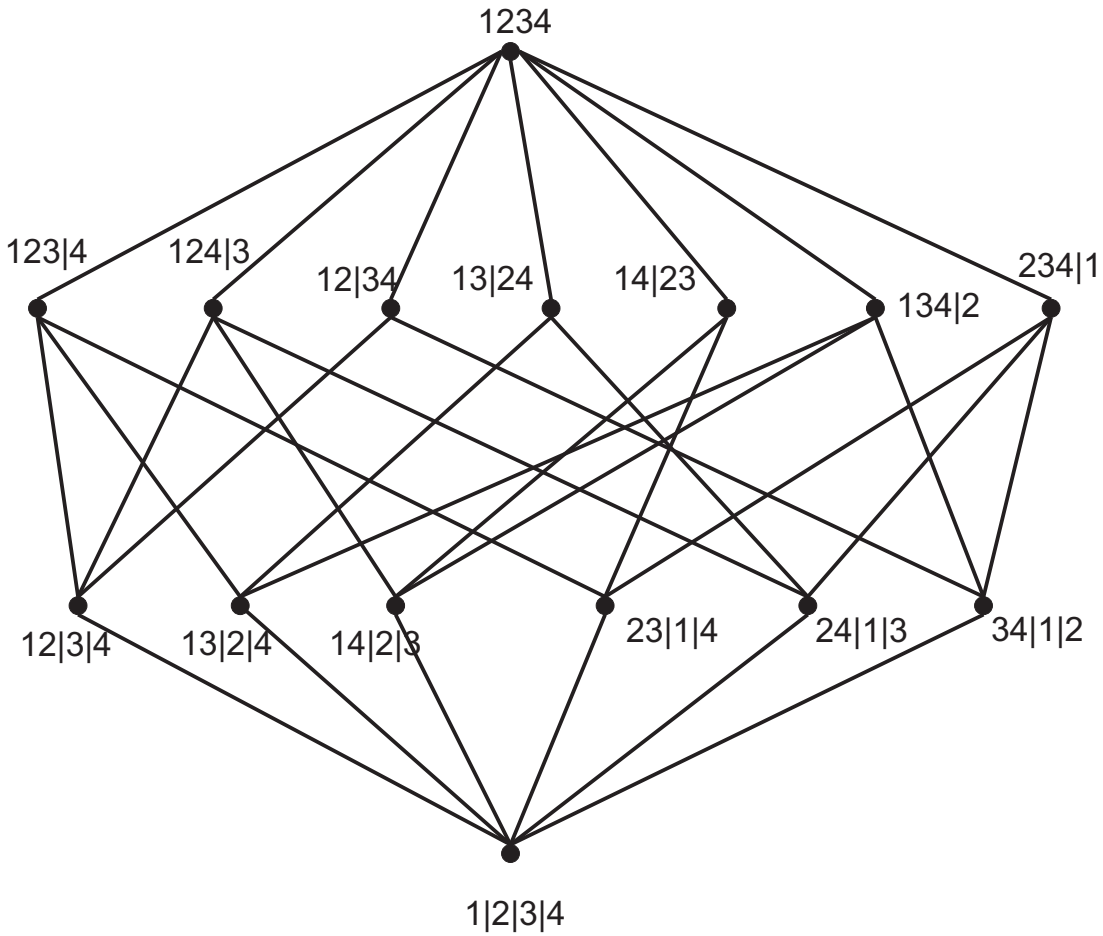


диаграмма частичного порядка являющегося решеткой



Решетка разбиений 4-х элементного множества

$A = \{1, 2, 3, 4\}$



Решетка. Другое определение

Теорема. Произвольное множество L является решеткой тогда и только тогда когда на нем определимы две операции $\vee$ и $\wedge$, удовлетворяющие следующим свойствам для любых $x, y, z \in L$:

L1	$x \vee x = x, \quad x \wedge x = x$	(идемпотентность)
L2	$x \vee y = y \vee x, \quad x \wedge y = y \wedge x$	(коммутативность)
L3	$x \vee (y \vee z) = (x \vee y) \vee z,$ $x \wedge (y \wedge z) = (x \wedge y) \wedge z$	(ассоциативность)
L4	$x = x \wedge (x \vee y) = x \vee (x \wedge y)$	(поглощение)

Решетка. Другое определение

Теорема. Произвольное множество L является решеткой тогда и только тогда когда на нем заданы две операции $\vee$ и $\wedge$, удовлетворяющие следующим свойствам для любых $x, y, z \in L$:

L1	$x \vee x = x, \quad x \wedge x = x$	(идемпотентность)
L2	$x \vee y = y \vee x, \quad x \wedge y = y \wedge x$	(коммутативность)
L3	$x \vee (y \vee z) = (x \vee y) \vee z,$ $x \wedge (y \wedge z) = (x \wedge y) \wedge z$	(ассоциативность)
L4	$x = x \wedge (x \vee y) = x \vee (x \wedge y)$	(поглощение)

Теорема позволяет рассматривать решетку как алгебру $(L, \vee, \wedge)$ со свойствами L1-L4. **Естественным порядком** решетки, задаваемой таким образом, называется отношение "$\leq$" $\subseteq L \times L$, определяемое как $x \leq y \stackrel{\text{def}}{=} x \wedge y = x$ (или, эквивалентно, $x \vee y = y$).

Полные решетки

Решетка называется **полной** если у любого подмножества ее элементов (в том числе пустого) есть супремум и инфимум.

$$\bigvee \emptyset = \mathbf{0} \quad \bigwedge \emptyset = \mathbf{1}$$

Все конечные решетки полны.

Для произвольного подмножества элементов полной решетки можно писать

$$\bigvee X, \quad \bigwedge X$$

в силу ассоциативности и коммутативности операций $\vee$ и $\wedge$.

Дистрибутивность

Решетка, в которой выполняются условия

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$$

$$x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$$

называется **дистрибутивной**.

Дистрибутивность

Решетка, в которой выполняются условия

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$$

$$x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$$

называется **дистрибутивной**.

Пример. Кольцо множеств - семейство F подмножеств множества I, содержащее вместе с любыми двумя множествами S и T их теоретико-множественные пересечение $S \cap T$ и объединение $S \cup T$.

Модулярность

Решетка, в которой выполняется условие

$$\text{если } x \leq z, \text{ то } x \vee (y \wedge z) = (x \vee y) \wedge z$$

называется **модулярной**.

Анализ Формальных Понятий(АФП). 1

[R.Wille, 1982], [B.Ganter, R.Wille, 1996, 1999]

Пусть даны множества G, M.
Пусть $I \subseteq G \times M$ - бинарное отношение между множествами G и M.
Тройка $\mathbb{K} := (G, M, I)$ называется **(формальным) контекстом**.

Интерпретация:

M – множество **признаков**, G – множество **объектов**

gIm или $(g, m) \in I \iff$ объект g обладает признаком m.

Анализ Формальных Понятий(АФП). 2

[R.Wille, 1982], [B.Ganter, R.Wille, 1996, 1999]

Дан контекст $\mathbb{K} := (G, M, I)$, рассмотрим отображения $\varphi\colon 2^G \to 2^M$ и $\psi\colon 2^M \to 2^G$:

$$\varphi(A) \stackrel{\text{def}}{=} \{m \in M \mid gIm \text{ для всех } g \in A\},$$

$$\psi(B) \stackrel{\text{def}}{=} \{g \in G \mid gIm \text{ для всех } m \in B\}.$$

Для любых $A_1, A_2 \subseteq G$, $B_1, B_2 \subseteq M$

1. $A_1 \subseteq A_2 \Rightarrow \varphi(A_2) \subseteq \varphi(A_1)$
2. $B_1 \subseteq B_2 \Rightarrow \psi(B_2) \subseteq \psi(B_1)$
3. $A_1 \subseteq \psi\varphi(A_1)$ и $B_1 \subseteq \varphi\psi(B_1)$

Отображения φ и ψ задают **соответствие Галуа** между $(2^G, \subseteq)$ и $(2^M, \subseteq)$.

(Абстрактное) соответствие Галуа

Отображения $\varphi : P \to Q$ и $\psi : Q \to P$ задают **соответствие Галуа** между частично упорядоченными множествами $(P, \leq_p)$ и $(Q, \leq_q)$ если

$$x \leq_p \psi(y) \iff \varphi(x) \geq_q y.$$

Полярности

[Г.Биркгоф, Теория решеток, 1984]

Пусть ρ - некоторое бинарное отношение между элементами двух классов I и J. Для любых подмножеств $X \subset I$ и $Y \subset J$ определим $X^* \subset J$ ("поляра" для X) как множество всех $y \in J$ таких, что $x \rho y$ для всех $x \in X$, и определим $Y^+ \subset I$ ("поляра" для Y) как множество всех $x \in I$ таких, что $x \rho y$ для всех $y \in Y$.
Тогда

- если $X \subset X_1$, то $X^* \supset X_1^*$;
- если $Y \subset Y_1$, то $Y^* \supset Y_1^+$;
- $X \subset (X^*)^+$ и $Y \subset (Y^+)^*$

Отношения $\swarrow$, $\nearrow$, $\swarrow\!\!\!\nearrow$

$$g \swarrow m :\Longleftrightarrow \begin{cases} g \not I m \text{ и} \\ \text{если } g' \subseteq h' \text{ и } g' \neq h', \text{ тогда } h I m, \end{cases}$$

$$g \nearrow m :\Longleftrightarrow \begin{cases} g \not I m \text{ и} \\ \text{если } m' \subseteq n' \text{ и } m' \neq n', \text{ тогда } g I n, \end{cases}$$

$$g \swarrow\!\!\!\nearrow m :\Longleftrightarrow \quad g \swarrow m \text{ и } g \nearrow m$$

Отношения $\swarrow$, $\nearrow$, $\swarrow\!\!\!\nearrow$

$$g \swarrow m :\Longleftrightarrow \begin{cases} g \not| m \text{ и} \\ \text{если } g' \subseteq h' \text{ и } g' \neq h', \text{ тогда } h | m, \end{cases}$$

$$g \nearrow m :\Longleftrightarrow \begin{cases} g \not| m \text{ и} \\ \text{если } m' \subseteq n' \text{ и } m' \neq n', \text{ тогда } g | n, \end{cases}$$

$$g \swarrow\!\!\!\nearrow m :\Longleftrightarrow \quad g \swarrow m \text{ и } g \nearrow m$$

Теорема

[R.Wille, B.Ganter, 1996]

Решетка понятий дистрибутивна тогда и только тогда в каждой строке и каждом столбце контекста есть ровно одна стрелка $\swarrow\!\!\!\nearrow$.

Анализ Формальных Понятий(АФП). 3

Пусть дан контекст $\mathbb{K} := (G, M, I)$. По традиции АФП вместо φ и ψ используется единое обозначение $(\cdot)'$, так что для произвольных $A \subseteq G$, $B \subseteq M$

$$A' \stackrel{\text{def}}{=} \{m \in M \mid gIm \text{ для всех } g \in A\},$$

$$B' \stackrel{\text{def}}{=} \{g \in G \mid gIm \text{ для всех } m \in B\}.$$

(Формальное) понятие есть пара (A, B):

$$A \subseteq G,\ B \subseteq M,\ A' = B,\ \text{ и }\ B' = A.$$

A называется **объемом** (extent), а B называется **содержанием** (intent) понятия (A, B).

Понятия частично-упорядочены отношением $(A_1, B_1) \geq (A_2, B_2) \iff A_1 \supseteq A_2 \quad (B_2 \supseteq B_1)$.

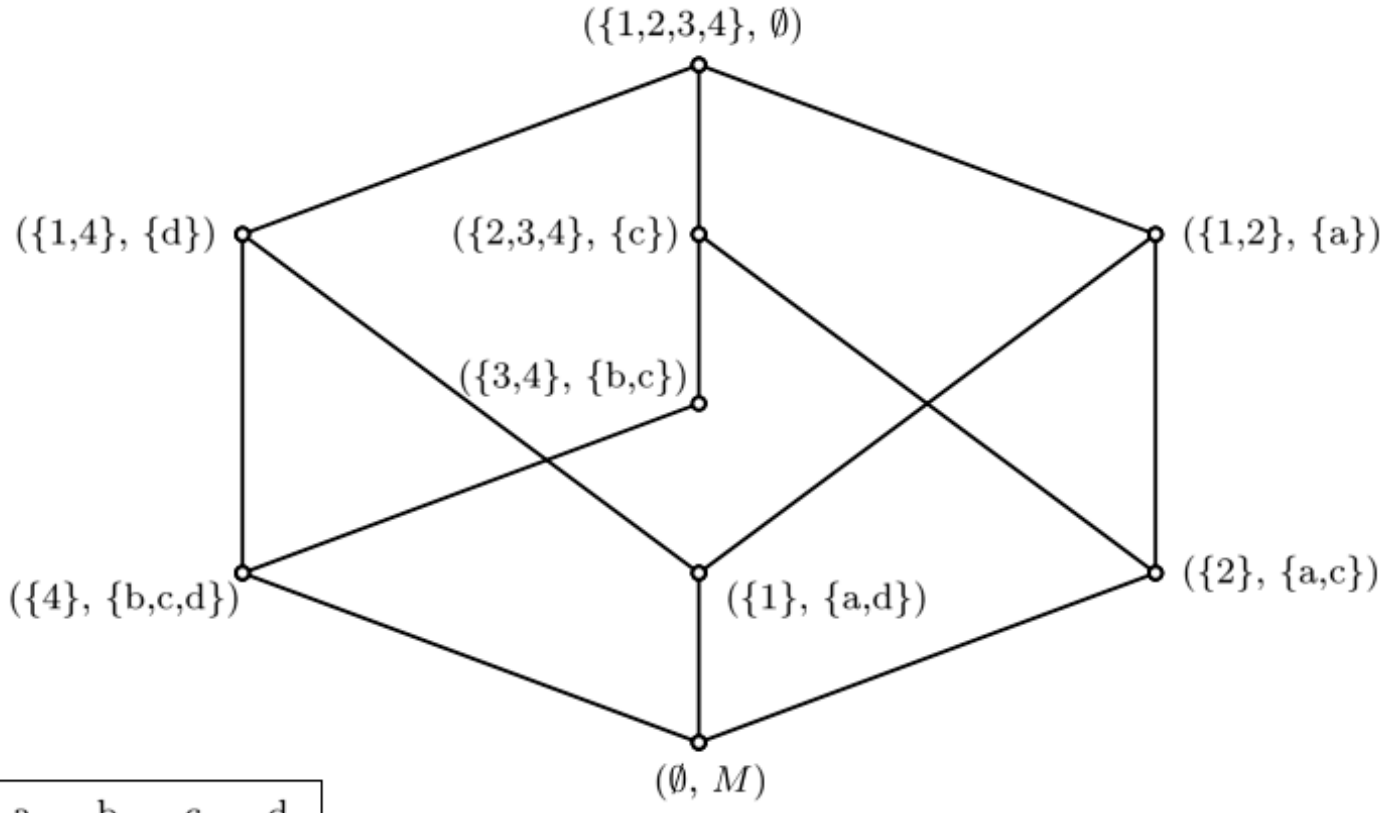
Свойства операций $(\cdot)'$

Пусть (G, M, I) — формальный контекст, $A, A_1, A_2 \subseteq G$ — множества объектов, $B \subseteq M$ — множество признаков, тогда

1. Если $A_1 \subseteq A_2$, то $A_2' \subseteq A_1'$;
2. Если $A_1 \subseteq A_2$, то $A_1'' \subseteq A_2''$;
3. $A \subseteq A''$;
4. $A''' = A'$ и, следовательно, $A'''' = A''$;
5. $(A_1 \cup A_2)' = A_1' \cap A_2'$;
6. $A \subseteq B' \Leftrightarrow B \subseteq A' \Leftrightarrow A \times B \subseteq I$.

Аналогично для подмножеств признаков.

Пример. Формальный контекст и решетка понятий.



	G \ M	a	b	c	d
1	△	×			×
2	◺	×		×	
3	▭		×	×	
4	□		×	×	×

a – ровно 3 вершины,
b – ровно 4 вершины,
c – имеет прямой угол,
d – все стороны равны

Оператор замыкания на множестве

Оператором замыкания на множестве G называется отображение $\varphi\colon \mathcal{P}(G) \to \mathcal{P}(G)$, которое каждому подмножеству $X \subseteq G$ сопоставляет его **замыкание** $\varphi X \subseteq G$, обладающее следующими свойствами:

1. $\varphi\varphi X = \varphi X$ (**идемпотентность**)
2. $X \subseteq \varphi X$ (**экстенсивность**)
3. $X \subseteq Y \Rightarrow \varphi X \subseteq \varphi Y$ (**монотонность**)

Элемент $X \subseteq G$ называется **замкнутым** если $\varphi X = X$.

Пример. Пусть дан контекст (G, M, I), тогда операторы $(\cdot)'' \colon 2^G \to 2^G$, $(\cdot)'' \colon 2^M \to 2^M$ являются операторами замыкания.

Супремум- и инфинум-плотные подмножества

Подмножество $X \subseteq L$ элементов решетки $(L, \leq)$ называется **супремум-плотным**, если любой элемент решетки $v \in L$ представим как

$$v = \bigvee \{x \in X \mid x < v\}.$$

Двойственно для **инфинум-плотных** подмножеств.

Основная теорема Анализа Формальных Понятий

[R.Wille 1982]

Решетка понятий $\underline{\mathfrak{B}}(G, M, I)$ есть полная решетка. Для произвольного множества формальных понятий

$$\{(A_j, B_j) \mid j \in J\} \subseteq \underline{\mathfrak{B}}(G, M, I)$$

точные нижняя и верхняя грани задаются как

$$\bigwedge_{j \in J}(A_j, B_j) = (\bigcap_{j \in J} A_j, (\bigcup_{j \in J} B_j)'')$$

$$\bigvee_{j \in J}(A_j, B_j) = ((\bigcup_{j \in J} A_j)'', \bigcap_{j \in J} B_j).$$

Полная решетка V изоморфна решетке $\underline{\mathfrak{B}}(G, M, I)$ тогда и только тогда когда существуют отображения $\gamma\colon G \to V$ и $\mu\colon M \to V$ такие, что $\gamma(G)$ супремум-плотно в V, $\mu(M)$ инфимум-плотно в V, а $gIm \Leftrightarrow \gamma g \leq \mu m$ для всех $g \in G$ и всех $m \in M$.

В частности, $V \cong \underline{\mathfrak{B}}(V, V, \leq)$.

Неразложимые элементы решетки

Пусть $(L, \leq)$ - полная решетка. Элемент $x \in L$ называется **супремум-неразложимым**, если

$$x \neq \bigvee \{y \mid y < x\}.$$

Двойственно, элемент $x \in L$ называется **инфинум-неразложимым**, если

$$x \neq \bigwedge \{y \mid y > x\}.$$

Множества супремум- и инфинум-неразложимых элементов решетки $(L, \leq)$ обозначаются, соответственно, через $J(L)$ и $M(L)$. Множество $J(L)$, а также любое его надмножество - супремум-плотные. Множество $M(L)$, а также любое его надмножество - инфинум-плотные.

Из основной теоремы АФП следует, что

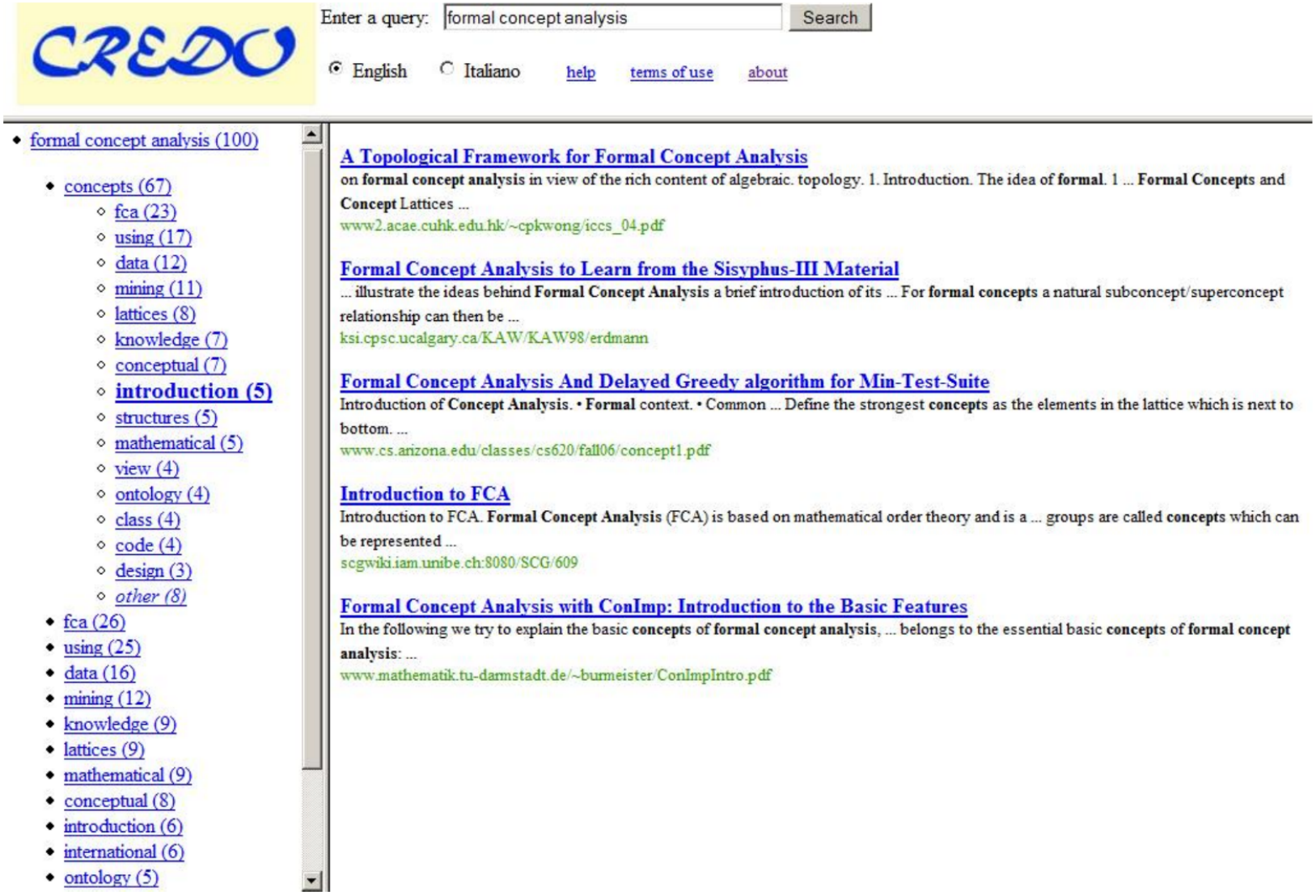
$$L \cong \underline{\mathfrak{B}}(J(L), M(L), \leq).$$

CREDO

Программная система "Conceptual REorganization of DOcuments", разработанная Carpineto и Romano, Fondazione Ugo Bordoni, Italy

- отображает верхнюю часть (два уровня от верхнего элемента) айсберга решетки формальных понятий (термы добавляются вниз по дереву) в виде дерева
- поддерживает онлайн навигацию по формальным понятиям, позволяя сужать пространство поиска
- Carpineto C., Romano G.: Exploiting the Potential of Concept Lattices for Information Retrieval with CREDO. J. Universal Computer Science 10(8)(2004), 985Ц1013

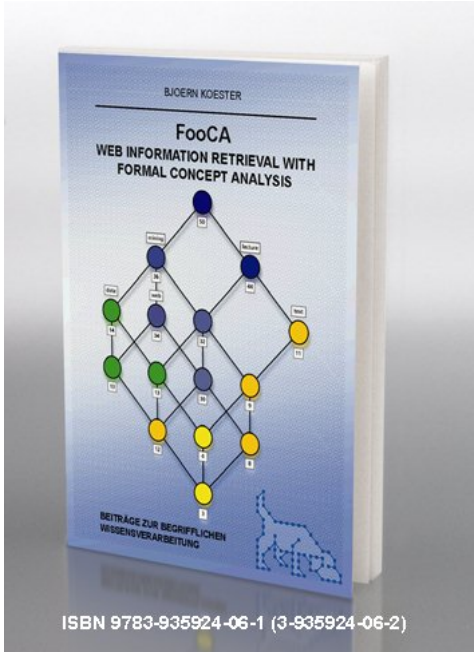
CREDO


CREDO
Enter a query: formal concept analysis
Search
English
Italiano
help
terms of use
about
formal concept analysis (100)
concepts (67)
fca (23)
using (17)
data (12)
mining (11)
lattices (8)
knowledge (7)
conceptual (7)
introduction (5)
structures (5)
mathematical (5)
view (4)
ontology (4)
class (4)
code (4)
design (3)
other (8)
fca (26)
using (25)
data (16)
mining (12)
knowledge (9)
lattices (9)
mathematical (9)
conceptual (8)
introduction (6)
international (6)
ontology (5)
A Topological Framework for Formal Concept Analysis
on formal concept analysis in view of the rich content of algebraic. topology. 1. Introduction. The idea of formal. 1 ... Formal Concepts and Concept Lattices ...
www2.acae.cuhk.edu.hk/~cpkwong/iccs_04.pdf
Formal Concept Analysis to Learn from the Sisyphus-III Material
... illustrate the ideas behind Formal Concept Analysis a brief introduction of its ... For formal concepts a natural subconcept/superconcept relationship can then be ...
ksi.cpsc.ucalgary.ca/KAW/KAW98/erdmann
Formal Concept Analysis And Delayed Greedy algorithm for Min-Test-Suite
Introduction of Concept Analysis. • Formal context. • Common ... Define the strongest concepts as the elements in the lattice which is next to bottom. ...
www.cs.arizona.edu/classes/cs620/fall06/concept1.pdf
Introduction to FCA
Introduction to FCA. Formal Concept Analysis (FCA) is based on mathematical order theory and is a ... groups are called concepts which can be represented ...
scgwiki.iam.unibe.ch:8080/SCG/609
Formal Concept Analysis with ConImp: Introduction to the Basic Features
In the following we try to explain the basic concepts of formal concept analysis, ... belongs to the essential basic concepts of formal concept analysis: ...
www.mathematik.tu-darmstadt.de/~burmeister/ConImpIntro.pdf

FooCA

- FCA + Google, программная система, разработанная Bjoern Koester, Webstrategy GmbH, Darmstadt и TU Dresden, Germany
- результаты поиска представлены напрямую в виде **формального контекста** (документы $\times$термы), а также **диаграммой решетки понятий** (переход к адресованным документам осуществляется по клику на таблице или вершине диаграммы)
- онлайн навигация по формальным понятиям - добавление или удаление признаков запускает новый поиск и перестраивает иерархию формальных понятий
- B. Koester: FooCA - Web Information Retrieval withFormal Concept Analysis. Verlag Allgemeine Wissenschaft, Mhltal, 2006. ISBN 9783-935924-06-1.

  B. Koester: Conceptual Knowledge Retrieval with FooCA: Improving Web Search Engine Results with Contexts and Concept Hierarchies. Proc. ICDM 2006, Springer-Verlag, Berlin, 2006.



FooCA

Search Formal Concept Analysis Yahoo English FooCA

Retrieval results 10 Min. objects per attribute 2 Min. attribute length 3

☐ Stemming ☑ Stopwords ☑ Clarify context ☑ Context refinement
☑ Attribute ranking ☐ Show original results ☐ Show extracted attributes

Your FooCA search for Formal Concept Analysis brought these results:

G/M	✓ X (10) analysis + - concept + - formal + -	✓ X (6) concepts + -	✓ X (3) method + -	✓ X (3) data + -	✓ X (2) conference + - held + - international + -	✓ X (2) lattices + -
1	x	x	x	x		
2	x	x	x	x		
3	x					
4	x	x	x			
5	x	x				
6	x	x				x
7	x			x	x	
8	x				x	
...						
10	x	x				x

http://www.kvocentral.org/resources/fca.html

6 out of 88 attributes selected. Export the Formal Context (CXT) FlashLattice. « [1..10] »

About FooCA and Terms of Use. FooCA is powered by Yahoo! Search

SearchSleuth

- разработан Peter Eklund и Jon Ducrou, University of Wollongong, Australia
- отображает **соседей и сиблингов для формального понятия поискового запроса** (прямое обобщение запроса, уточнение и категоризация) в виде текстовых ярлыков(ссылок) на термы/признаки, задающие формальные понятия
- онлайн навигация, расширение контекста с помощью последовательного запуска поиска для кажого из соседей и сиблингов поискового запроса
- J. Ducrou, P. Eklund: SearchSleuth: The Conceptual Neighbourhood of an Web Query. Proc. CLA 2007, LIRMM & University of Montpellier II, 2007.

SearchSleuth

-analysis

formal concept analysis ~[formal concept fca] ~[formal concept context]

+fca +data +lattice +context +based +mathematics +mining +theory +method +conceptual

1. Formal Concept Analysis Homepage
Formal Concept Analysis is a method of conceptual knowledge representation and data analysis. ... Christian Lindig's Concepts, (in C, older version: TkConcept? ...
www.upriss.org.uk/fca/fca.html

2. Formal concept analysis - Wikipedia, the free encyclopedia
... example concepts satisfy the formal definitions; the ... describing formal concept analysis for computer scientists. A Formal Concept Analysis Homepage ...
en.wikipedia.org/wiki/Formal_concept_analysis

3. Formal Concept Analysis
Formal Concept Analysis is a branch of applied mathematics. ... Several books on Formal Concept Analysis have appeared, among them the first ...
www.math.tu-dresden.de/~ganter/fba.html

4. Formal Concept Analysis
Formal Concept Analysis (FCA) is a method mainly used for the analysis ... into units which are formal abstractions of concepts of human thought, allowing ...
www.cs.cmu.edu/afs/cs.cmu.edu/project/jair/pub/volume24/cimiano05a-html/node3...

5. Linguistic Applications of Formal Concept Analysis
scribes the role that formal concept analysis can play in the automated or ... Associative and Formal Concepts. In: Priss; Corbett; Angelova (eds.), Con ...
www.upriss.org.uk/papers/fcaic03.pdf

Число понятий (размер решетки)

- Число понятий может быть экспоненциально от размера контекста: $(A, A, \neq)$.
- Задача вычисления числа всех понятий $\#P$-полна.

Вспомним

$\#P$ есть класс перечислительных задач, связанных с задачами распознавания из класса NP: Задача лежит в классе #P если есть недетерминированная полиномиальная машина Тьюринга, которая для каждого случая *I* данной задачи имеет число принимающих вычислений, равное числу различных решений этого частного случая *I* [Valiant 1979].

Задача #P-**полна**, если она лежит в классе #P и является #P-**трудной**, то есть любая задача из #P может быть сведена к ней по Тьюрингу за полиномиальное время.

Если #P = P, то NP = P.

Примеры #P-полных задач

Вычисление

- перманента матрицы,
- числа совершенных паросочетаний в двудольном графе,
- числа выполняющих наборов КНФ,
- числа вершинных покрытий в графе,
- ...

Сложность алгоритмов порождения понятий

Все понятия контекста могут быть порождены алгоритмом с временной сложностью

$$O(min\{|G|, |M|\} \times |G| \times |M| \times |C|)$$

где |G| - число объектов, |M| - число признаков, |C| - число понятий.

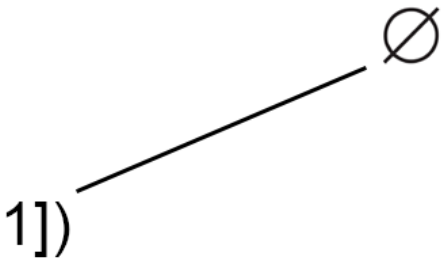
Алгоритм 3О: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×

$\varnothing$

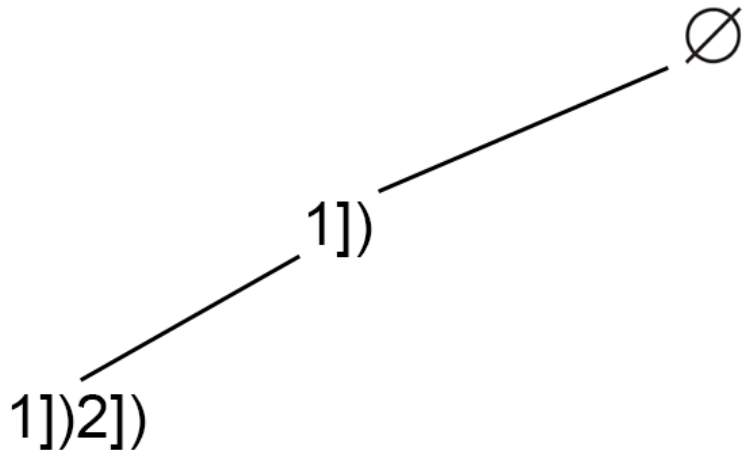
Алгоритм 3О: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



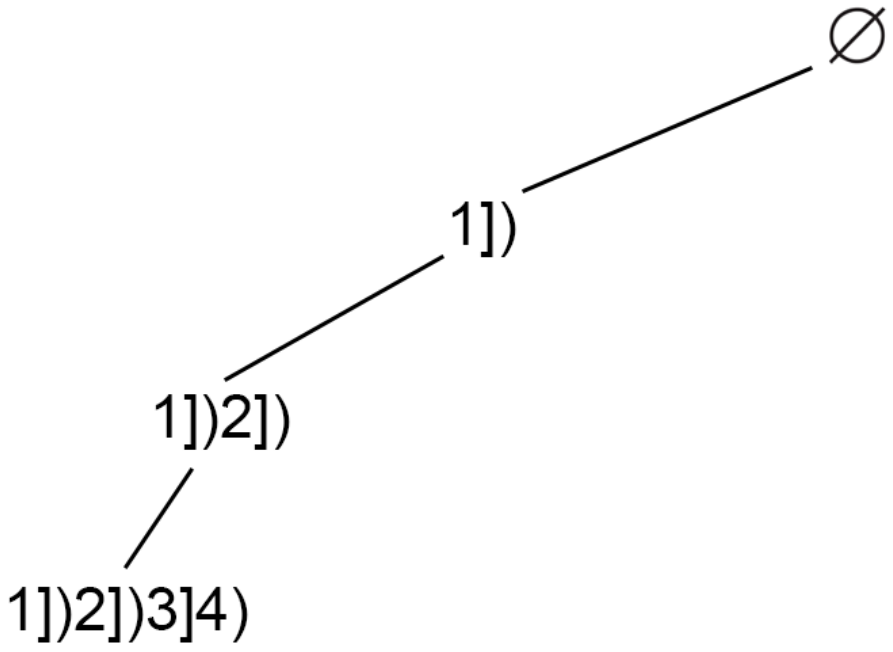
Алгоритм 3O: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



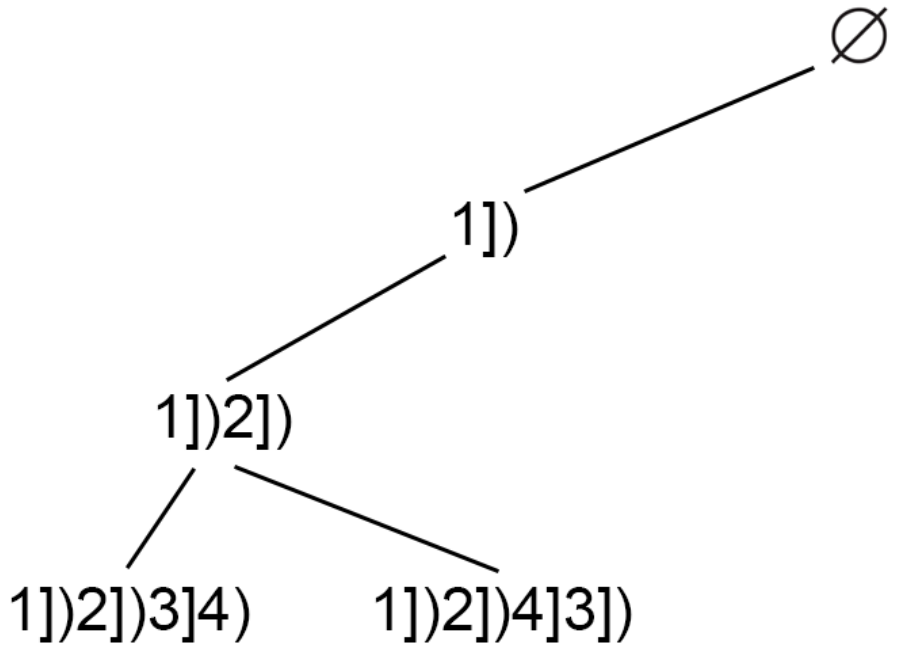
Алгоритм 3O: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



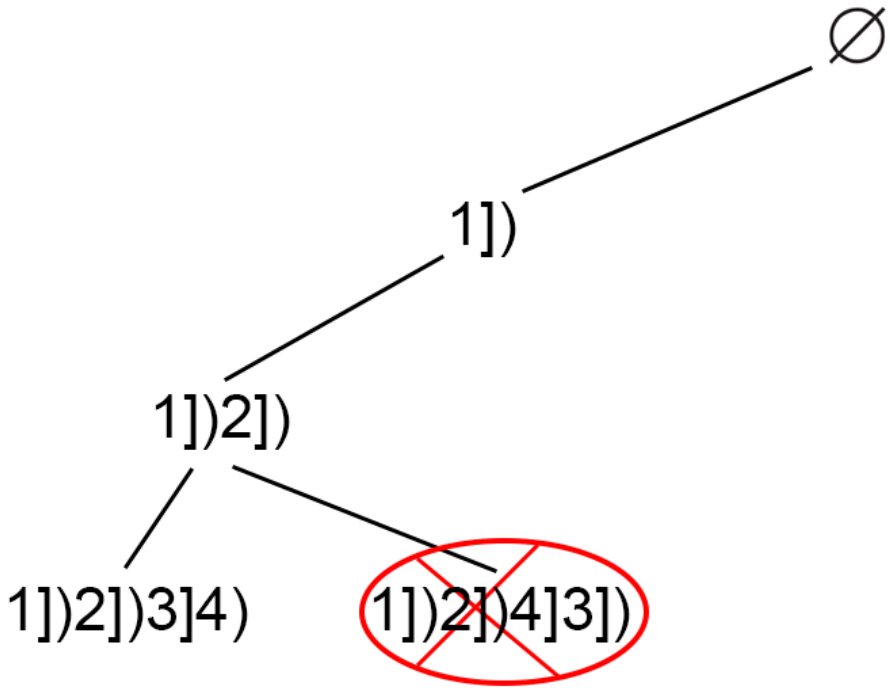
Алгоритм 3O: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



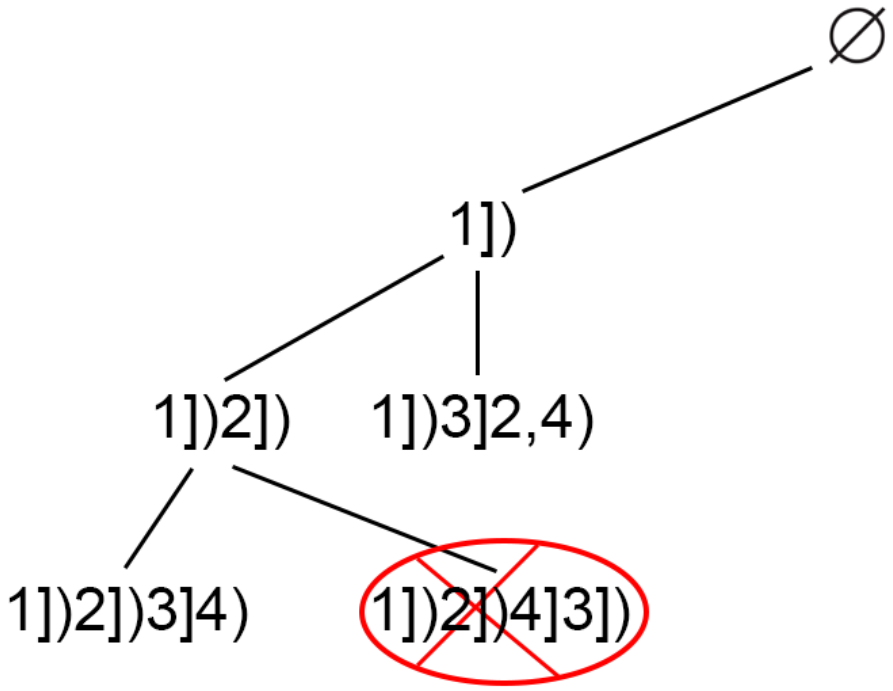
Алгоритм 3O: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



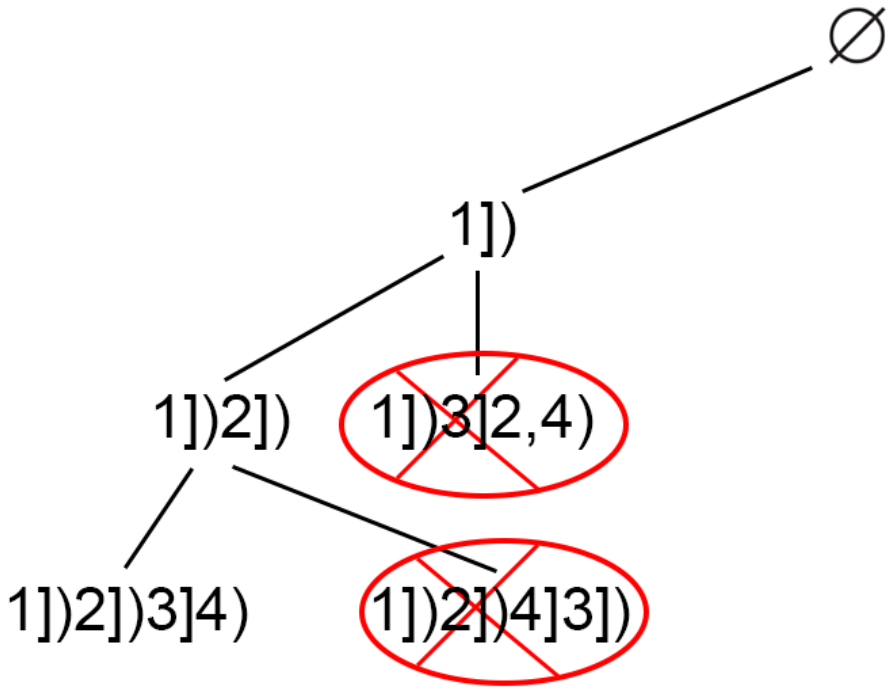
Алгоритм 3O: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



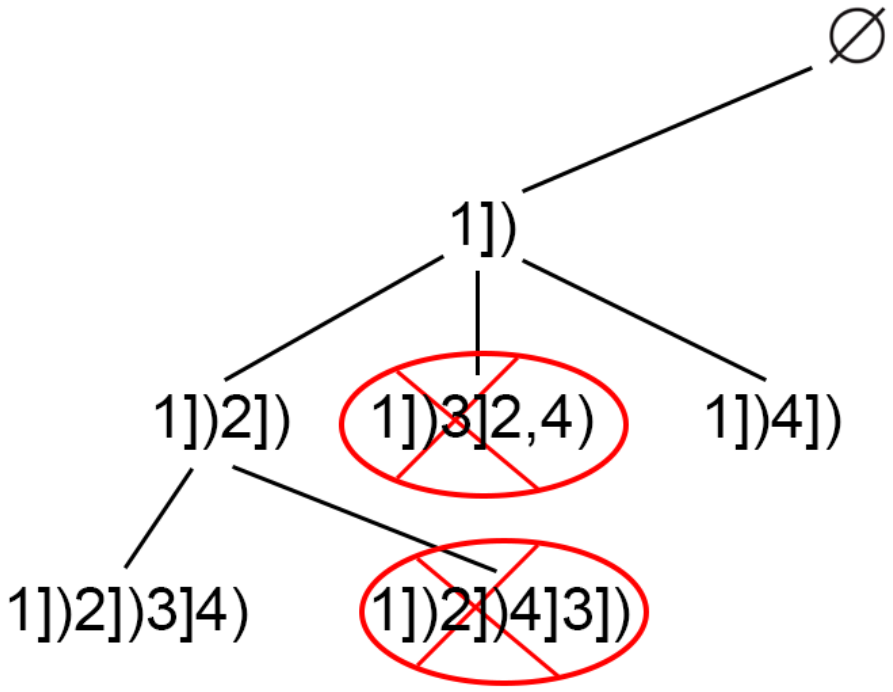
Алгоритм 3O: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



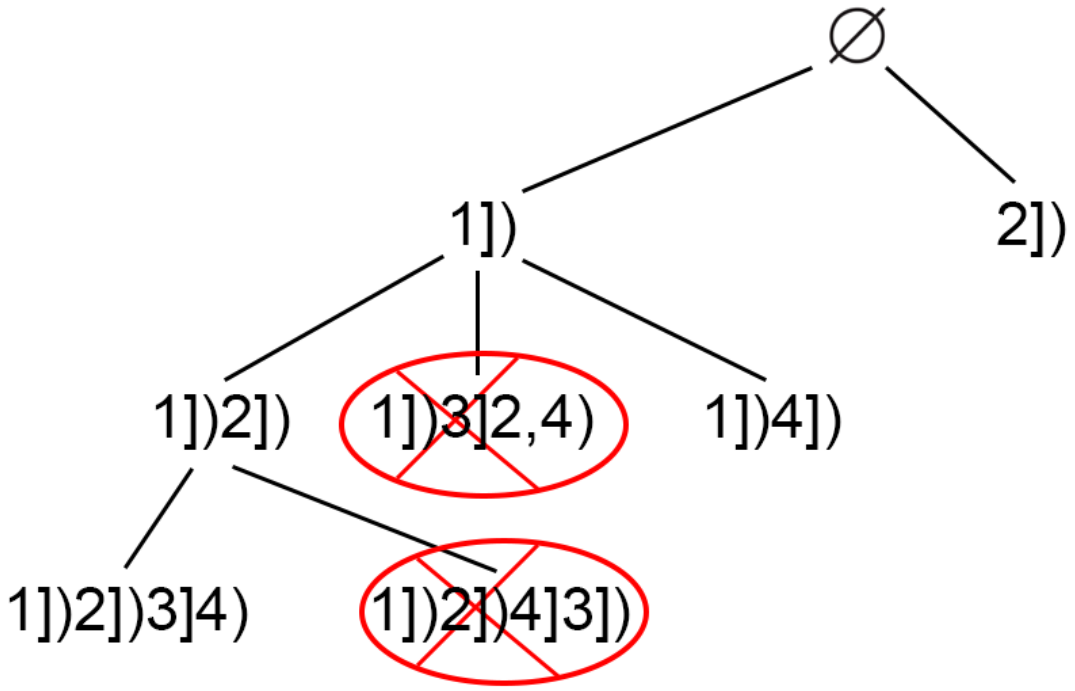
Алгоритм 3O: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



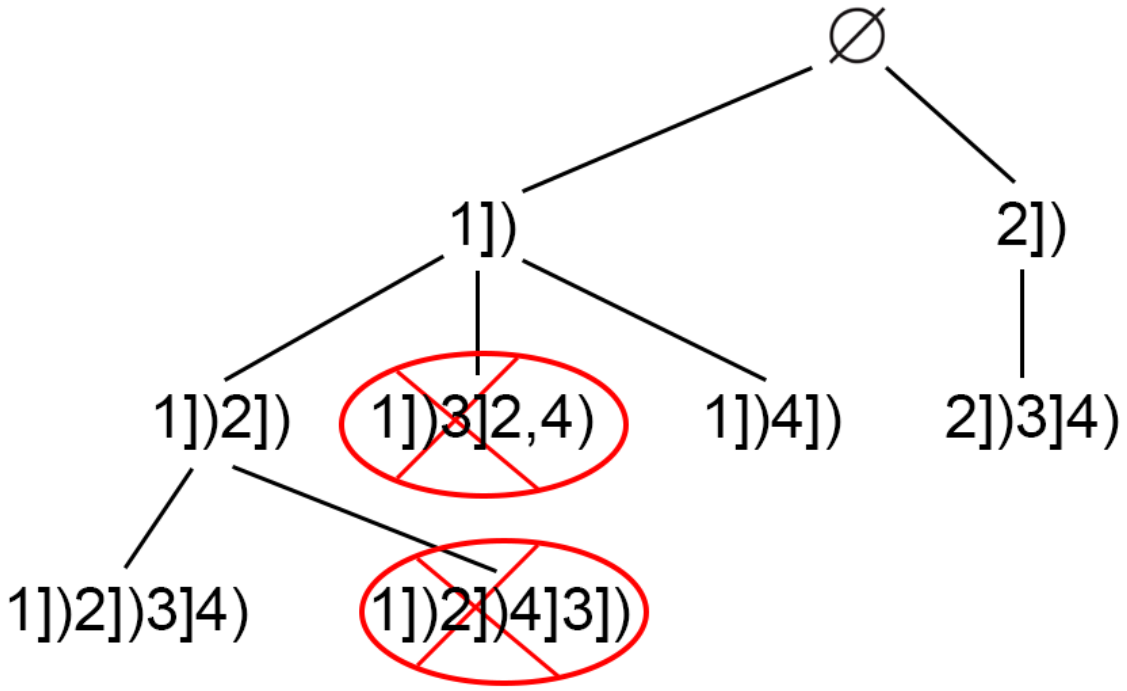
Алгоритм 3O: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



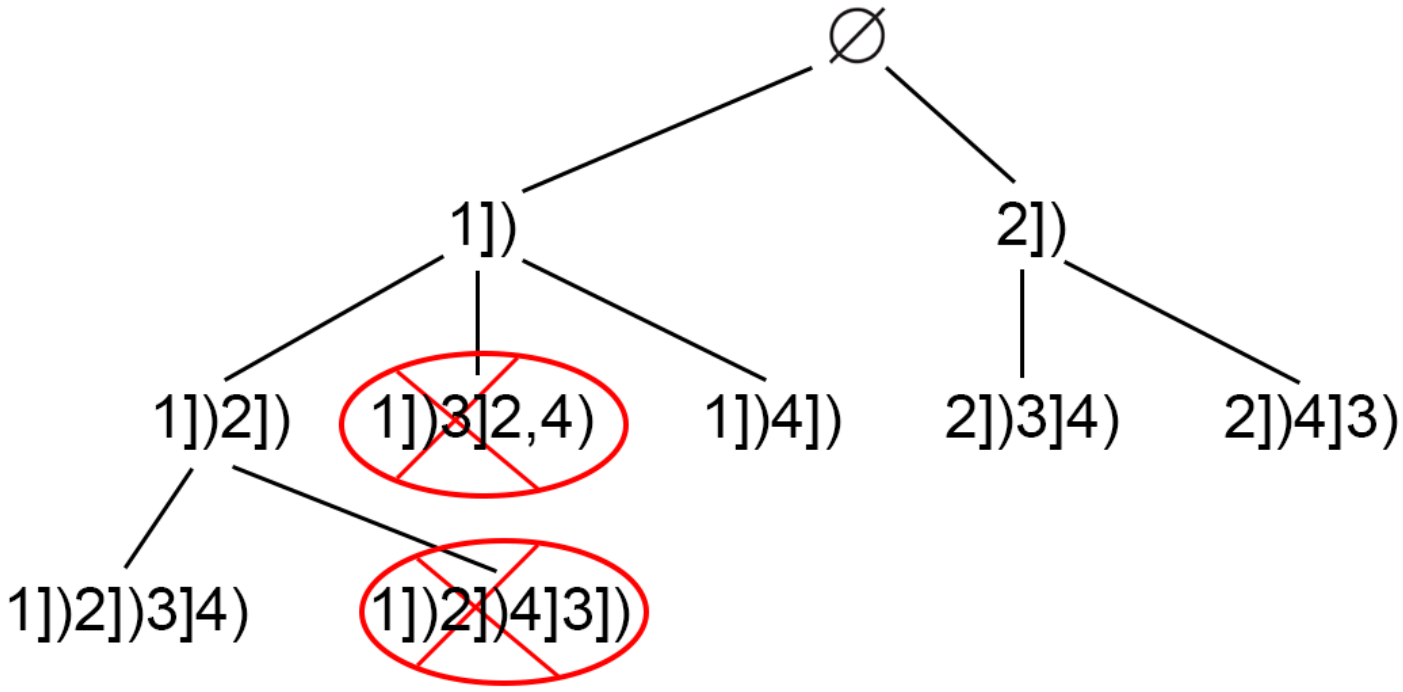
Алгоритм 3O: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



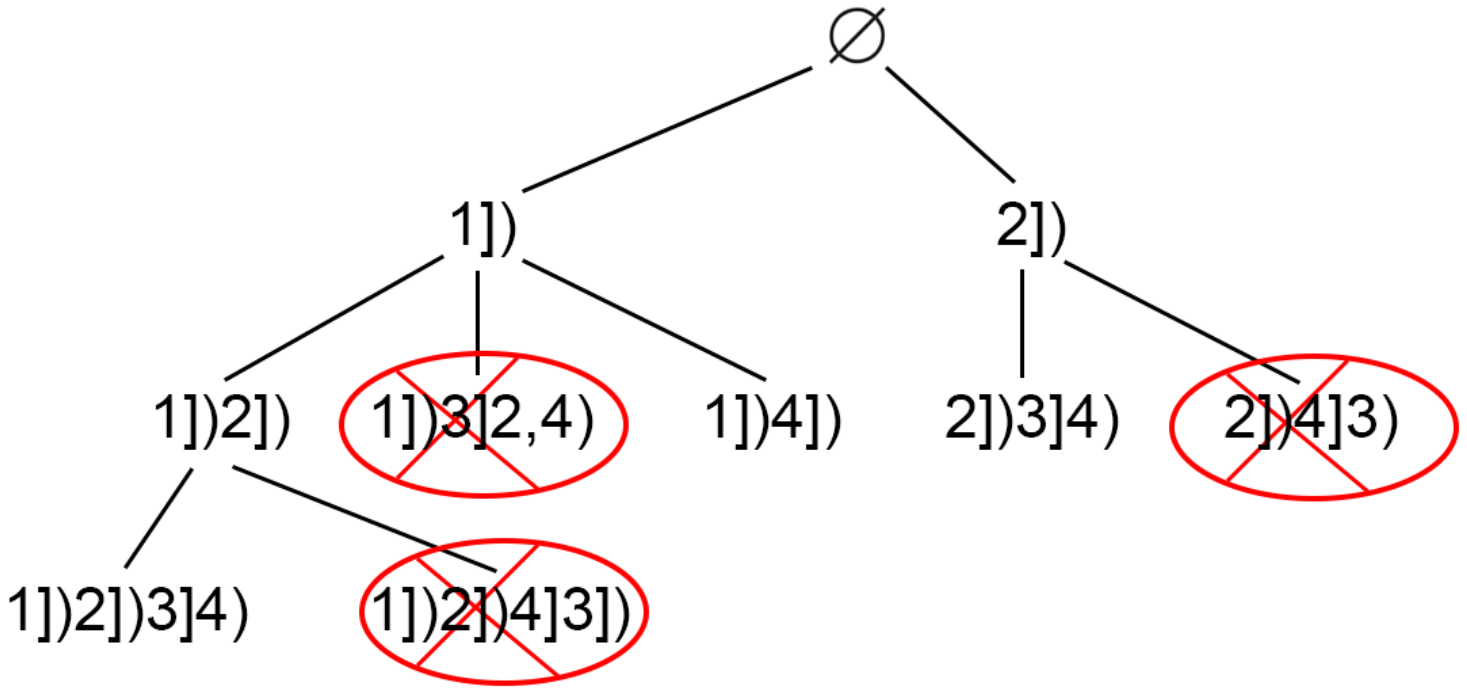
Алгоритм 3O: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



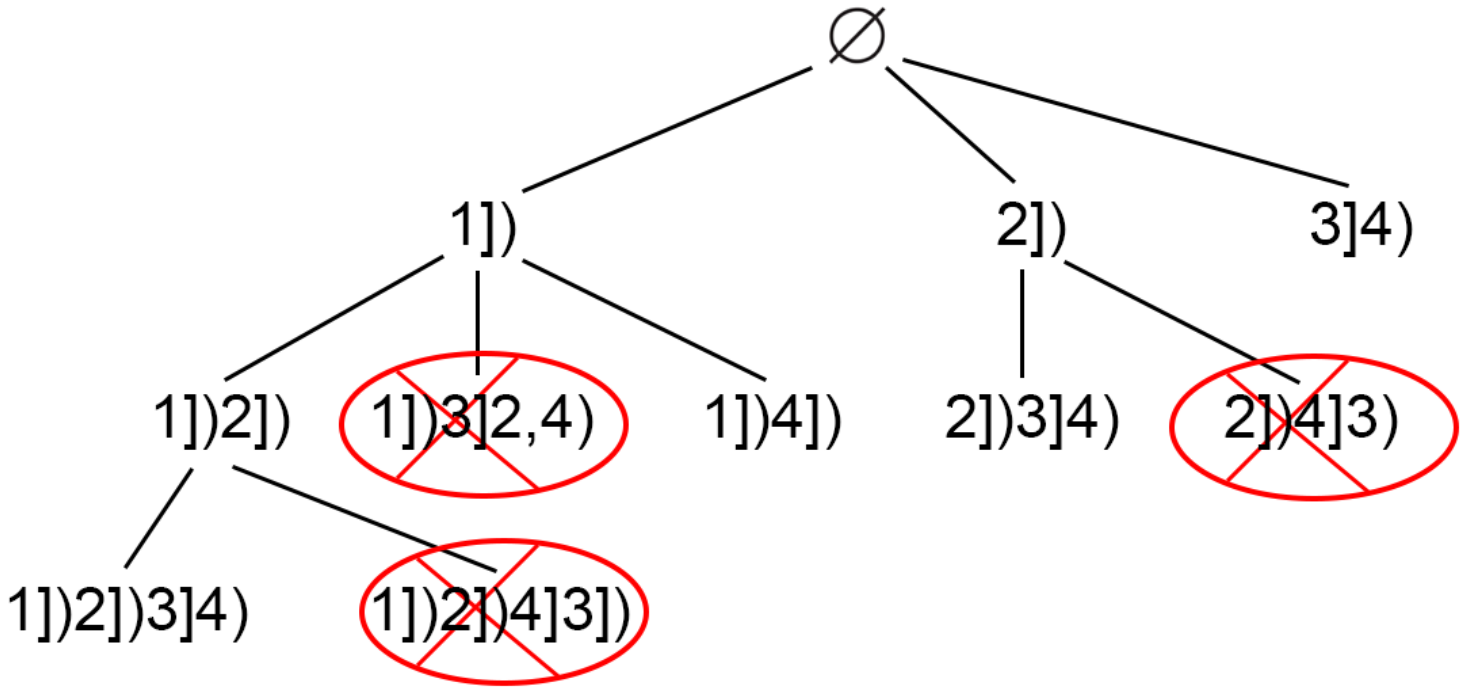
Алгоритм 3O: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



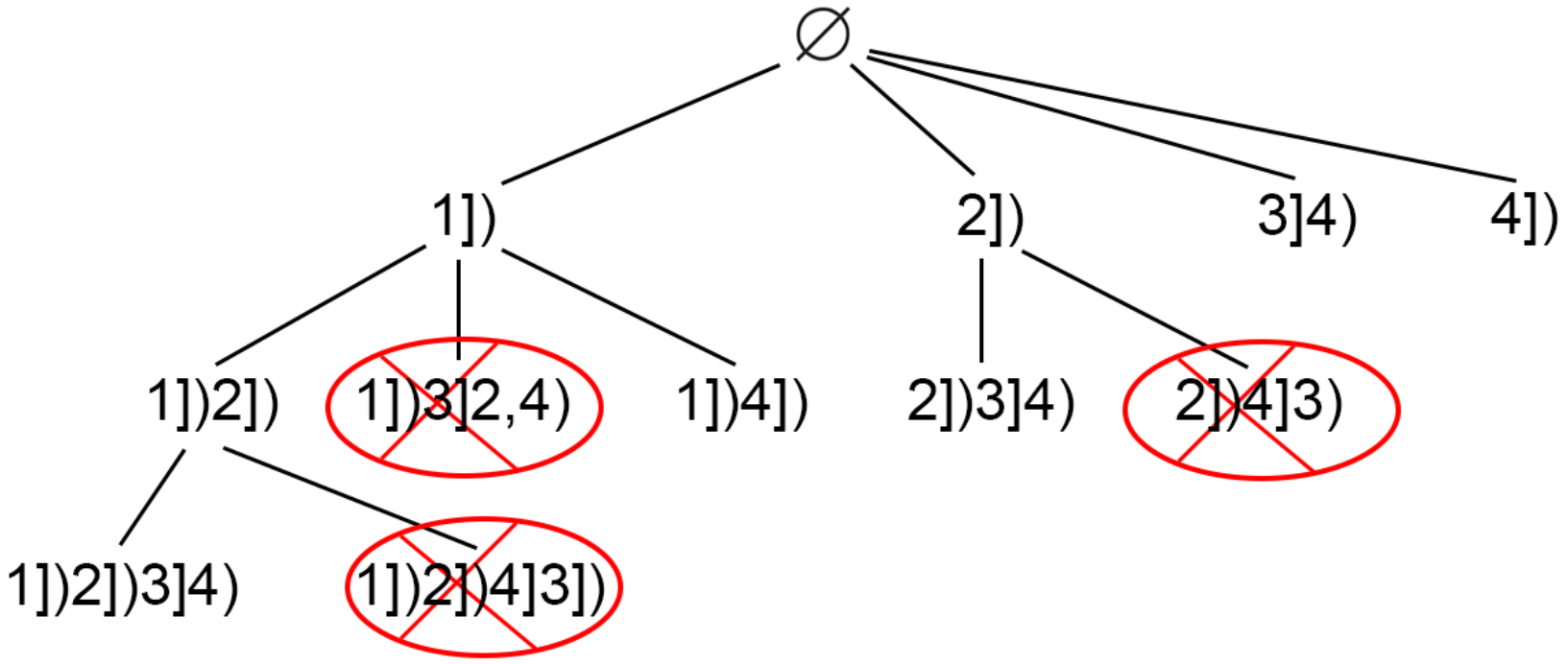
Алгоритм 3O: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



Алгоритм 3O: использование лексикографического порядка

G \ M	a	b	c	d
1	×			×
2	×		×	
3		×	×	
4		×	×	×



Импликации и исследование признаков

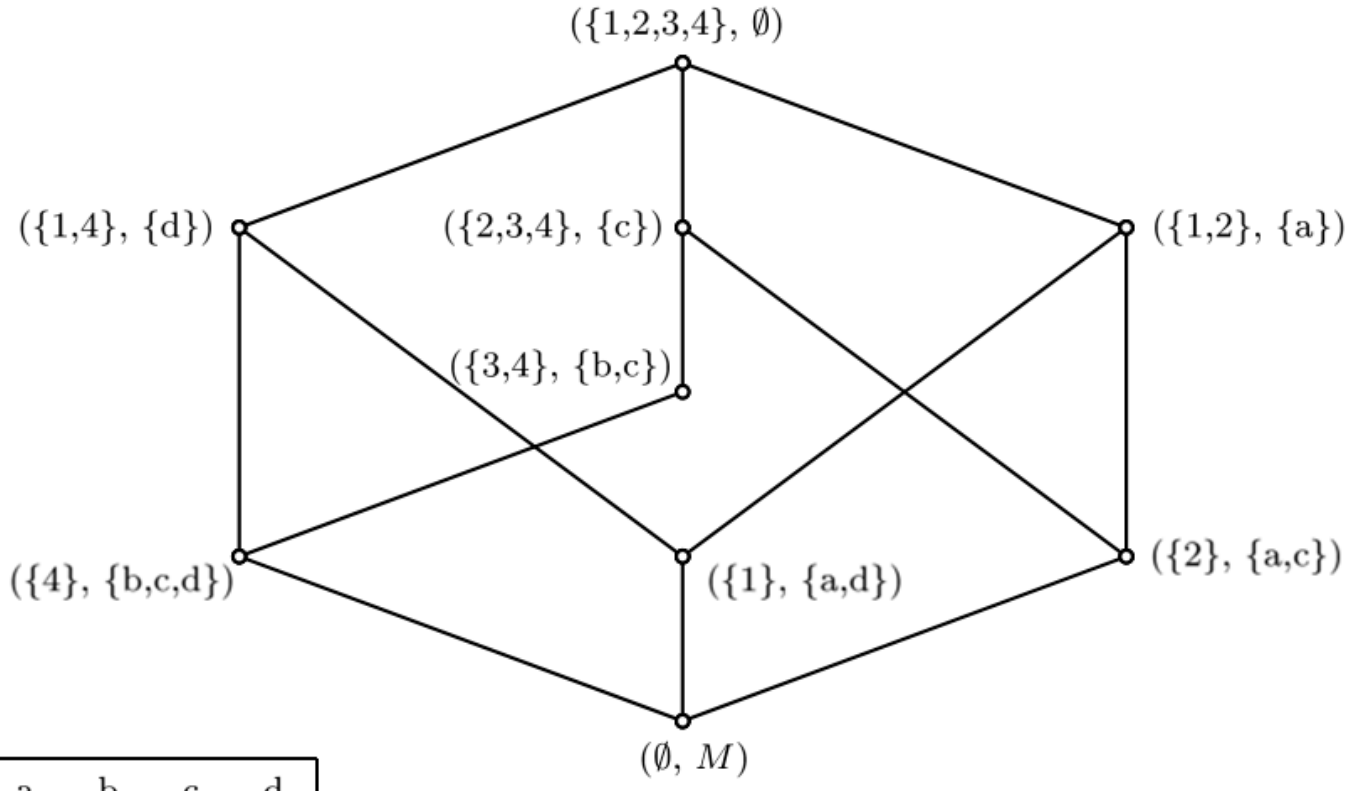
Импликация $A \to B$, где $A, B \subseteq M$, имеет место если $A' \subseteq B'$, т.е. каждый объект, обладающий всеми признаками из A, также обладают всеми признаками из B.

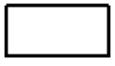
Импликации удовлетворяют **правилам Армстронга**:

$$\frac{}{X \to X}, \quad \frac{X \to Y}{X \cup Z \to Y}, \quad \frac{X \to Y, Y \cup Z \to W}{X \cup Z \to W}$$

относительно которых существуют **покрытия** и **базисы**.

Пример: решетка понятий и импликации



	G \ M	a	b	c	d
1		×			×
2		×		×	
3			×	×	
4			×	×	×

a – ровно 3 вершины,
b – ровно 4 вершины,
c – имеет прямой угол,
d – все стороны равны

Импликации

$abc \to d$
$b \to c$
$cd \to b$

Покрытия и базисы

Покрытием множества импликаций называется такое его подмножество, из которого все импликации выводимы с помощью правил Армстронга.
Базисом называется неизбыточное покрытие, то есть покрытие, любое подмножество которого не является покрытием.

Дополнительные правила

Из правил Армстронга выводятся следующие полезные соотношения

$$\frac{X \subseteq Y}{Y \to X}$$

$$\frac{X_1 \to Y_1, \quad X_2 \to Y_2}{X_1 \cup X_2 \to Y_1 \cup Y_2}$$

$$\frac{X \to Y}{X \to Y \setminus X}, \qquad \frac{X \to Y \setminus X}{X \to Y}$$

Многозначный контекст (реляционная таблица)

Многозначный контекст это четверка (G, M, W, I), где W есть множество значений признаков, $I \subseteq G \times M \times W$, такое что $(g, m, w) \in I$ и из $(g, m, v) \in I$ следует $w = v$.

Признак m **полный** если для всех $g \in G$ существует $w \in W$ такое что $(g, m, w) \in I$. Многозначный контекст полон если все его признаки полны.

Для полных многозначных контекстов значение признака m на объекте g обозначается через $m(g)$; таким образом $(g, m, m(g)) \in I$.

Функциональные зависимости

Функциональная зависимость $X \Rightarrow Y$ имеет место в полном многозначном контексте (G, M, W, I) если для каждой пары объектов $g, h \in G$ имеет место

$$(\forall m \in X \quad m(g) = m(h)) \rightarrow (\forall n \in Y \quad n(g) = n(h)).$$

Функциональные зависимости и декомпозируемость

Пример. Многозначный контекст (реляционная таблица) $K = (G, M, I)$, в котором **есть** функциональная зависимость $b \to c$:

G \ M	a	b	c
1	a_1	b_1	c_1
2	a_2	b_1	c_1
3	a_1	b_2	c_2
4	a_2	b_3	c_2

$=$

G \ M	a	b
1	a_1	b_1
2	a_2	b_1
3	a_1	b_2
4	a_2	b_3

$\bowtie$

G \ M	b	c
1	b_1	c_1
2	b_2	c_2
3	b_3	c_2

Функциональные зависимости и декомпозируемость

Пример. Многозначный контекст (реляционная таблица)
$K = (G, M, I)$, в котором **нет** функциональной зависимости $b \to c$:

G \ M	a	b	c
1	a_1	b_1	c_1
2	a_2	b_1	c_2
3	a_1	b_2	c_2
4	a_2	b_3	c_2

$\neq$

G \ M	a	b
1	a_1	b_1
2	a_2	b_1
3	a_1	b_2
4	a_2	b_3

$\bowtie$

G \ M	b	c
1	b_1	c_1
2	b_2	c_2
3	b_3	c_2

Импликации и функциональные зависимости. 2

По полному многозначному контексту $K = (G, M, W, I)$ можно определить контекст $K_N := (\mathcal{P}_2(G), M, I_N)$, где $\mathcal{P}_2(G)$ есть множество пар различных объектов из G, а I_N определяется как

$$\{g, h\} I_N m :\Leftrightarrow m(g) = m(h).$$

В многозначном контексте K имеет место функциональная зависимость $X \Leftrightarrow Y$ тогда и только тогда когда в контексте K_N имеет место импликация $X \to Y$. Имеет место и обратная сводимость.

Теорема Для контекста $K = (G, M, I)$ можно построить многозначный контекст K_W такой что импликация $X \to Y$ имеет место в K тогда и только тогда когда в K_W имеет место функциональная зависимость $X \Leftrightarrow Y$ in K_W.

Импликации и функциональные зависимости. 3

Пример. Рассмотрим следующий формальный контекст $K = (G, M, I)$:

G \ M	a	b	c	d
1	×	×		
2	×		×	×
3		×		×
4	×		×	

Многозначный контекст где функциональные зависимости “совпадают" с импликациями контекста K:

G \ M	a	b	c	d
1	×	×	1	1
2	×	2	×	×
3	3	×	3	×
4	×	4	×	4
5	×	×	×	×

Генераторный базис импликаций

Подмножество признаков $D \subseteq M$ есть **генератор** замкнутого подмножества признаков $B \subseteq M$, $B'' = B$ если $D \subseteq B$, $D'' = B = B''$.

Подмножество $D \subseteq M$ есть **минимальный генератор**, если для любого $E \subset D$ имеет место $E'' \neq D'' = B''$.

Генератор $D \subseteq M$ называется **нетривиальным** если $D \neq D'' = B''$.

Множество всех нетривиальных минимальных генераторов B обозначим $\mathrm{nmingen}(B)$.

Генераторный базис импликаций выглядит следующим образом: $\{F \to (F'' \setminus F) \mid F \subseteq M, F \in \mathrm{nmingen}\,(F'')\}$.

Базис собственных посылок

Для множества признаков $B \subseteq M$ определим B^* как множество признаков из $M \setminus B$ таких, что они следуют из B, но не из его собственных подмножеств:

$$B^* = B'' \setminus (B \cup \bigcup_{S \subseteq B} S'')$$

Если $B^* \neq \emptyset$, то B называется собственной посылкой (proper premise).

Множество импликаций $\{B \to B'' \mid B$ - собственная посылка $\}$ является базисом, который называется базисом собственных посылок.

Канонический базис: базис Дюкенна-Гига

Подмножество признаков $P \subseteq M$ называется **псевдосодержанием**, если $P \neq P''$ и для любого псевдосодержания Q, такого что $Q \subset P$, имеет место $Q'' \subseteq P$.

Множество $\{P \to (P'' \setminus P) \mid P$ — псевдосодержание$\}$ образует базис импликаций, имеющий минимальный размер (по числу импликаций). Этот базис называется **базисом Дюкенна-Гига** (Duquenne-Guigues basis, stembase, canonical base).

Долгое время считалось, что размер базиса не может быть "большим"...

Канонический базис: базис Дюкенна-Гига

Подмножество признаков $P \subseteq M$ называется **псевдосодержанием**, если $P \neq P''$ и для любого псевдосодержания Q, такого что $Q \subset P$, имеет место $Q'' \subseteq P$.

Множество $\{P \to (P'' \setminus P) \mid P$ — псевдосодержание$\}$ образует базис импликаций, имеющий минимальный размер (по числу импликаций). Этот базис называется **базисом Дюкенна-Гига** (Duquenne-Guigues basis, stembase, canonical base).

Долгое время считалось, что размер базиса не может быть "большим"...

Но к сожалению это не так.
Задача подсчета псевдосодержаний #P-трудна

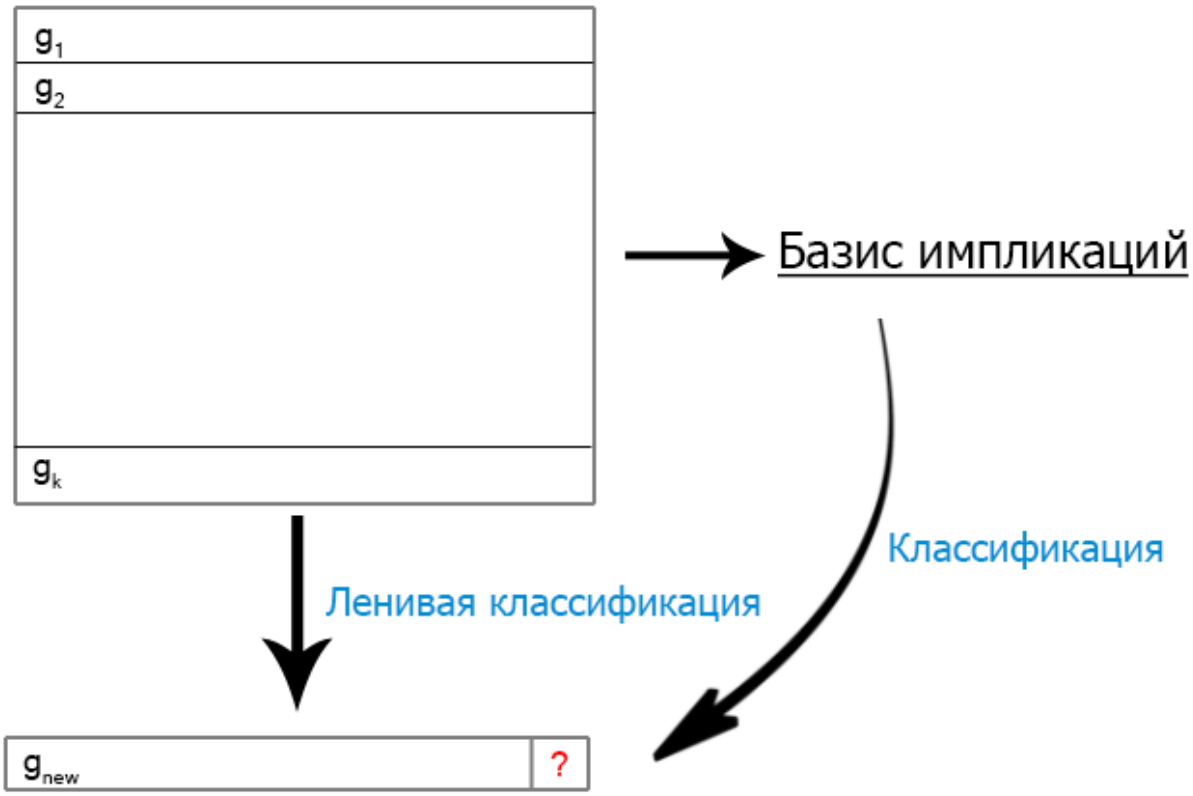
Сложность порождения базиса

- Число псевдосодержаний может быть экспоненциальным от размера контекста [СК 2004]
- Задача подсчета псевдосодержаний #P-трудна [СК 2004]
- Задача определения того, является ли подмножество признаков псевдосодержанием, coNP-полна [МБ,СК,СО 2006-2010]
- Задача порождения псевдосодержаний в произвольном порядке TRANSHYP-трудна [B.Sertkaya 2009]
- Порождение псевдосодержаний в обратном лексикографическом порядке с полиномиальной задержкой возможна только если P=NP [F.Distel, 2010]
- Порождение псевдосодержаний в лексикографическом порядке с полиномиальной задержкой возможна только если P=NP [МБ, СК 2010]
- Является ли задача о возможности перечисления псевдо-содержаний в произвольном порядке с полиномиальной задержкой NP-трудной?

Альтернатива базисам: ленивая классификация

Базис импликаций с короткими посылками

$$
\begin{aligned}
a &\to \dots \\
ab &\to \dots \\
abc &\to \dots \\
abcd &\to \dots
\end{aligned}
$$



Минимальные генераторы

Подмножество признаков $D \subseteq M$ есть **генератор** замкнутого подмножества признаков $B \subseteq M$, $B'' = B$, если $D \subseteq B$, $D'' = B = B''$.

При этом D есть **минимальный генератор**, если для любого $E \subset D$ имеет место $E'' \neq D'' = B''$.

Множество всех минимальных генераторов B обозначим $\mathrm{mingen}(B)$.

Решетки в Data Mining. Ассоциативные правила

В середине 1990х гг. в работах R. Agrawal и др. по ассоциативным правилам были переоткрыты частичные импликации из Анализа Формальных Понятий.

$A \to_{c,s} B$ - частичная импликация (ассоциативное правило) контекста (G, M, I)

- $c, s \in [0, 1]$;
- $c = \frac{|(A \cup B)'|}{|A'|}$ - **достоверность** (confidence, conf);
- $s = \frac{|(A \cup B)'|}{|G|}$ - **поддержка** (support, supp).

Базис ассоциативных правил

Как минимизировать множество ассоциативных правил, из которого (с помощью допустимых преобразований) можно получить все остальные правила?

Базис ассоциативных правил

Как минимизировать множество ассоциативных правил, из которого (с помощью допустимых преобразований) можно получить все остальные правила?

Рассмотрим ассоциативное правило $A \to_{c,s} B$ и при фиксированных достоверности $c = \frac{|(A \cup B)'|}{|A'|}$ и поддержке $s = \frac{|(A \cup B)'|}{|G|}$ попробуем уменьшить посылку и увеличить заключение.

1. **Уменьшение посылки.** При фиксированных c и s уменьшать посылку можно до некоторого $D \subseteq A$ такого, что $(D \cup B)' = (A \cup B)' = A' \cap B' = D' \cap B'$, то есть $D' = A' = A'''$. Таким образом, минимальное такое D есть минимальный генератор A'', т.е. $D \in \text{mingen}(A'')$.

Базис ассоциативных правил

Как минимизировать множество ассоциативных правил, из которого (с помощью допустимых преобразований) можно получить все остальные правила?

Рассмотрим ассоциативное правило $A \to_{c,s} B$ и при фиксированных достоверности $c = \frac{|(A \cup B)'|}{|A'|}$ и поддержке $s = \frac{|(A \cup B)'|}{|G|}$ попробуем уменьшить посылку и увеличить заключение.

1. **Уменьшение посылки.** При фиксированных c и s уменьшать посылку можно до некоторого $D \subseteq A$ такого, что $(D \cup B)' = (A \cup B)' = A' \cap B' = D' \cap B'$, то есть $D' = A' = A'''$. Таким образом, минимальное такое D есть минимальный генератор A'', т.е. $D \in \text{mingen}(A'')$.

2. **Увеличение заключения.** Заключение можно увеличить на множество Δ, такое что $(A \cup B)' = (A \cup B \cup \Delta)' = (A \cup B)' \cap \Delta'$, что возможно только когда $(A \cup B)' \subseteq \Delta'$, что эквивалентно $A \cup B \to \Delta$, а также $\Delta \subseteq (A \cup B)''$. Таким образом, заключение ассоциативного правила можно увеличить до $(A \cup B)''$.

Таким образом, можно хранить и порождать лишь правила вида $D \to (A \cup B)''$, где $D \in \text{mingen}(A'')$.

Базис ассоциативных правил

Рассмотрим ассоциативное правило $D \to (A \cup B)''$, где $D \in \text{mingen}(A'')$. Это правило в диаграмме соответствует пути вниз из понятия (A', A'') в понятие $((A \cup B)', (A \cup B)'')$.

Если $(A', A'') \not\succ ((A \cup B)', (A \cup B)'')$, т.е. соответствующие вершины не является соседями в диаграмме, то найдется понятие (E', E''), такое что $(A', A'') \succ (E', E'') > ((A \cup B)', (A \cup B)'')$.

Базис ассоциативных правил

Рассмотрим правила $D \to E''$, где $D \in \text{mingen}(A'')$ и $F \to (A \cup B)''$, где $F \in \text{mingen}(E'')$.

Достоверность первого правила - $c_1 = \frac{|E'|}{|A'|}$, а второго - $c_2 = \frac{|(A \cup B)'|}{|E'|}$.

Достоверность же исходного правила есть

$$c = \frac{|(A \cup B)'|}{|A'|} = \frac{|E'|}{|A'|} \cdot \frac{|(A \cup B)'|}{|E'|} = c_1 \cdot c_2.$$

Это показывает, что достаточно хранить правила вида

$$\{F \to ('' \setminus F'') \mid F \subseteq M, F \in \text{ mingen } (F''), (F', F'') \succ (E', E'')\},$$

соответствующие ребрам диаграммы - поддержки остальных можно получить перемножением поддержек по соответствующим путям в диаграмме.

Общая задача поиска ассоциативных правил

Найти все "частые" (с поддержкой не ниже порога) ассоциативные правила со степенью достоверности не ниже порога

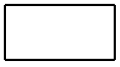
Путь решения

- Достаточно найти все частые замкнутые множества признаков (то есть все частые содержания контекста)

Базис ассоциативных правил (базис Люксенбургера)

- Остовное дерево диаграммы решетки понятий
- Базис импликаций Дюкенна-Гига

Ассоциативные правила

	G \ M	a	b	c	d
1		×			×
2		×		×	
3			×	×	
4			×	×	×

Объекты:

1 – равносторонний треугольник

2 – прямоугольный треугольник,

3 – прямоугольник,

4 – квадрат,

Признаки:

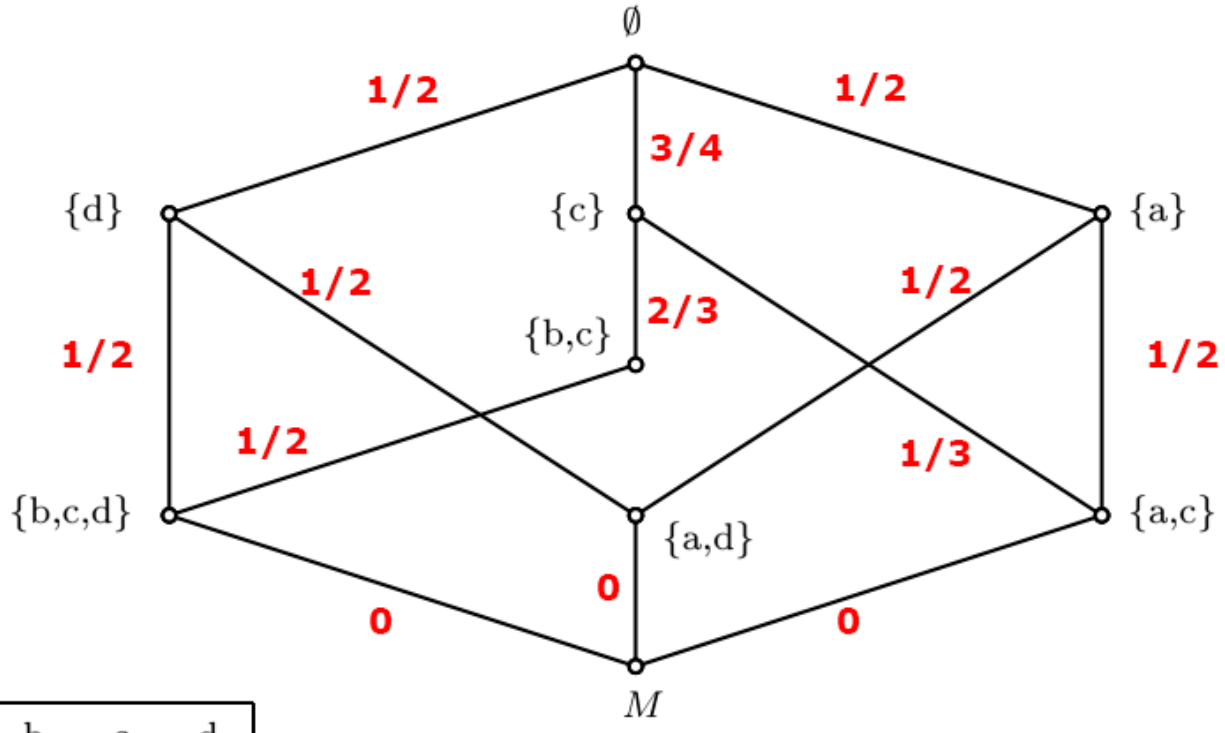
a – ровно 3 вершины,

b – ровно 4 вершины,

c – имеет прямой угол,

d – все стороны равны

Пример: достоверность

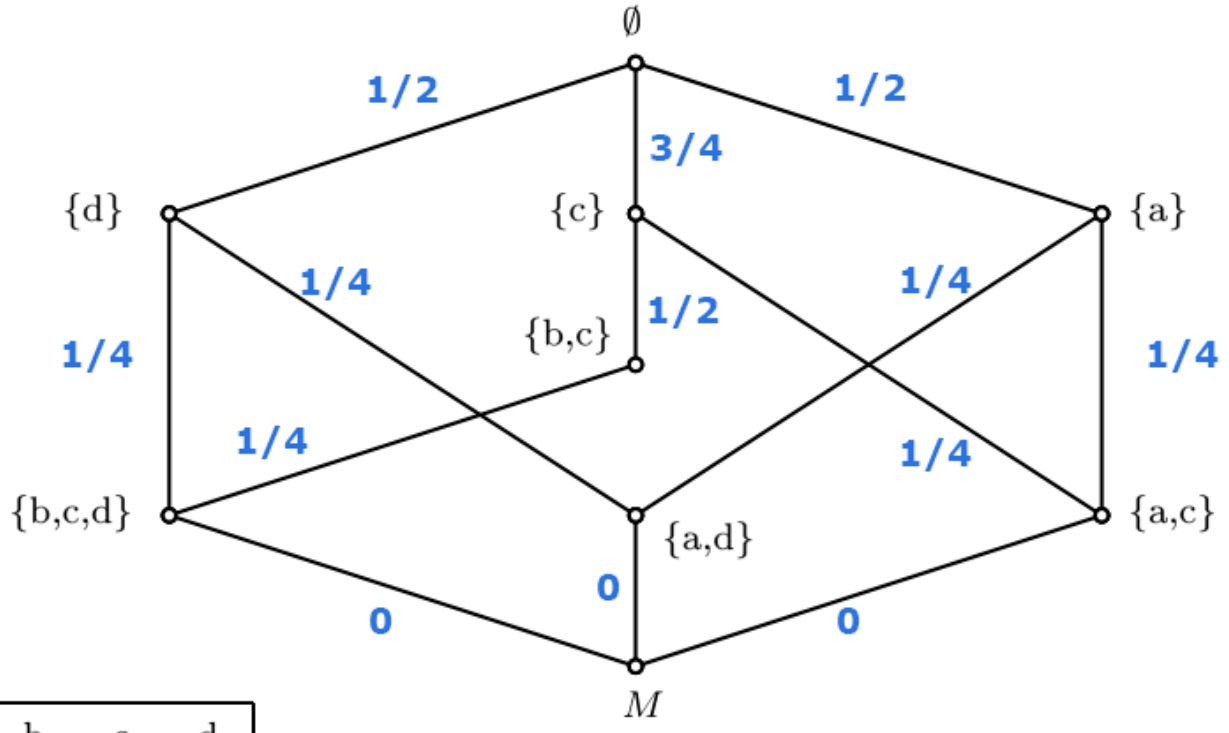


	G \ M	a	b	c	d
1	△	×			×
2	◺	×		×	
3	▭		×	×	
4	□		×	×	×

Пусть **хорошие правила** задаются условиями $\sup \geq 1/2$ и $\mathrm{conf} \geq 2/3$. Тогда их всего два:

1. $\emptyset \to c$, $\quad \sup(\emptyset \to c) = \mathrm{conf}(\emptyset \to c) = 3/4$;
2. $c \to b$, $\quad \sup(c \to b) = 1/2$, $\mathrm{conf}(c \to b) = 2/3$.

Пример: поддержка



	G \ M	a	b	c	d
1	△	×			×
2	◺	×		×	
3	▭		×	×	
4	□		×	×	×

Пусть **хорошие правила** задаются условиями $\sup \geq 1/2$ и $\mathrm{conf} \geq 2/3$. Тогда их всего два:

1. $\emptyset \to c$, $\quad \sup(\emptyset \to c) = \mathrm{conf}(\emptyset \to c) = 3/4$;
2. $c \to b$, $\quad \sup(c \to b) = 1/2$, $\mathrm{conf}(c \to b) = 2/3$.

Полурешетка антиунификаций в машинном обучении

Антиунификация для конечных термов была предложена в работах G. Plotkin и J. C. Reynolds. Так в работе

> J. C. Reynolds, Transformational systems and the algebraic structure of atomic formulas, *Machine Intelligence*, vol. 5, pp. 135-151, Edinburgh University Press, 1970.

антиунификация исследовалась как операция супремума в решетке термов.

Пример:

$$\text{Для } A = P(a, x, f(x)) \text{ и } B = P(y, f(b), f(f(b))) \text{ имеем}$$
$$\wedge(A, B) = P(z_1, z_2, f(z_2)).$$

В работе

> G.D. Plotkin, A Note on inductive generalization, *Machine Intelligence*, vol. 5, pp. 153-163, Edinburgh University Press, 1970.

антиунификация применялась как основа индуктивного обобщения. Позже эта идея использовалась в методах Индуктивного Логического Программирования (ILP).

ДСМ-метод

Одной из первых моделей машинного обучения, неявно использовавших решетки (системы замыканий или семейства Мура) был ДСМ-метод, предложенный впервые в

> В. К. Финн, О машинно-ориентированной формализации правдоподобных рассуждений в стиле Ф. Бэкона – Д.С. Милля \\ Семиотика и Информатика, **20** (1983), 35-101

Метод сходства (Первое правило индуктивной логики):

> "Если два или большее число примеров исследуемого явления обладают только одним общим признаком, то ... [этот признак] есть причина (или следствие) данного явления."
>
> John Stuart Mill, *A System of Logic, Ratiocinative and Inductive*, London, 1843

В ДСМ-методе гипотезы относительно причины явления ищутся среди пересечений описаний положительных примеров явления. На пересечения могут быть наложены различные дополнительные условия.

Логические средства ДСМ-метода:
Многозначное многосортное расширение логики предикатов первого порядка с помощью кванторов по кортежам переменной длины (слабая логика предикатов второго порядка).

Пример: Формализация миллевского **метода сходства**:

$$\mathcal{M}^{+}_{a,n}(V,W) := \exists k\, \widetilde{\mathcal{M}}^{+}_{a,n}(V,W,k),$$

$$\widetilde{\mathcal{M}}^{+}_{a,n}(V,W,k) := \exists Z_1 \ldots \exists Z_k\, \exists U_1 \ldots \exists U_k\, (\mathop{\&}_{i=1}^{k} J_{\langle 1,n\rangle}(Z_i \Rightarrow_1 U_i)\,\&$$

$$\&\,\forall U(J_{\langle 1,n\rangle}(Z_i \Rightarrow_1 U) \to \mathcal{U} \subset U_i))\,\&\,\&(Z_1 \cap \ldots \cap Z_k) =$$

$$V\,\&\,V \neq \emptyset\,\&\,W \neq \emptyset\,\&$$

$$\&\,\forall i\,\forall j((i \neq j)\,\&\,1 \leq i,j \leq k) \to Z_i \neq Z_j)\,\&\,\&\,\forall X\,\forall Y((J_{\langle 1,n\rangle}(X \Rightarrow_1 Y)\,\&$$

$$\&\,\forall \mathcal{U}(J_{\langle 1,n\rangle}(X \Rightarrow_1 U) \to \mathcal{U} \subseteq Y)\,\&\,\&\,V \subset X) \to$$

$$(W \subseteq Y\,\&(\mathop{\vee}_{i=1}^{k}(X = Z_i))))\,\&\,k \geq 2)$$

Этим задается система замыканий относительно операции $\bigcap$ “сходства” объектов.

ДСМ-метод в терминах АФП

Помимо признаков из множества M имеется **целевой признак** $w \notin M$, относительно которого все объекты разделяются следующим образом:

- **положительные примеры**: Множество $G_+ \subseteq G$ объектов, про которые известно, что они обладают целевым признаком w,
- **отрицательные примеры**: Множество $G_- \subseteq G$ объектов, про которые известно, что они не обладают целевым признаком w,
- **недоопределенные примеры**: Множество $G_\tau \subseteq G$ объектов, про которые неизвестно, обладают ли они целевым признаком или нет.

Возникают три подконтекста: $\mathbb{K}_\varepsilon := (G_\varepsilon, M, I_\varepsilon)$, $\varepsilon \in \{-, +, \tau\}$.

ДСМ-метод в терминах АФП

В подконтекстах $\mathbb{K}_\varepsilon := (G_\varepsilon, M, I_\varepsilon)$, $\varepsilon \in \{-, +, \tau\}$
операторы Галуа и соответствующие операторы замыкания обозначаются через $(\cdot)^\varepsilon$, $(\cdot)^{\varepsilon\varepsilon}$, например, X^+, X^{++} и т.д.

Формальное содержание $H \subseteq M$ контекста $\mathbb{K}_+$ есть **положительная гипотеза** если H не является подмножеством содержания ни одного отрицательного примера $g \in G_-$:

$$H^{++} = H, \quad \forall g \in G_- \quad H \not\subseteq g^-.$$

Отрицательные гипотезы определяются симметрично (с заменой + на -)

Формальное содержание $H \subseteq M$ контекста $\mathbb{K}_-$ есть **отрицательная гипотеза** если H не является подмножеством содержания ни одного положительного примера $g \in G_+$:

$$H^{--} = H, \quad \forall g \in G_+ \quad H \not\subseteq g^+.$$

Пример обучающей выборки

G \ M	цвет	жесткий	гладкий	форма	**фрукт**
яблоко	желтое	нет	да	круглое	+
грейпфрут	желтый	нет	нет	круглый	+
киви	зеленое	нет	нет	овальное	+
слива	синяя	нет	да	овальная	+
кубик	зеленый	да	да	кубический	−
яйцо	белое	да	да	овальное	−
теннисный мяч	белый	нет	нет	круглый	−

Естественное шкалирование выборки

G \ M	w	y	g	b	f	$\bar{\text{f}}$	s	$\bar{\text{s}}$	r	$\bar{\text{r}}$	**фрукт**
яблоко		×				×	×		×		+
грейпфрут		×				×		×	×		+
киви			×			×		×		×	+
слива				×		×	×			×	+
кубик			×		×		×			×	−
яйцо	×				×		×			×	−
теннисный мяч	×					×		×	×		−

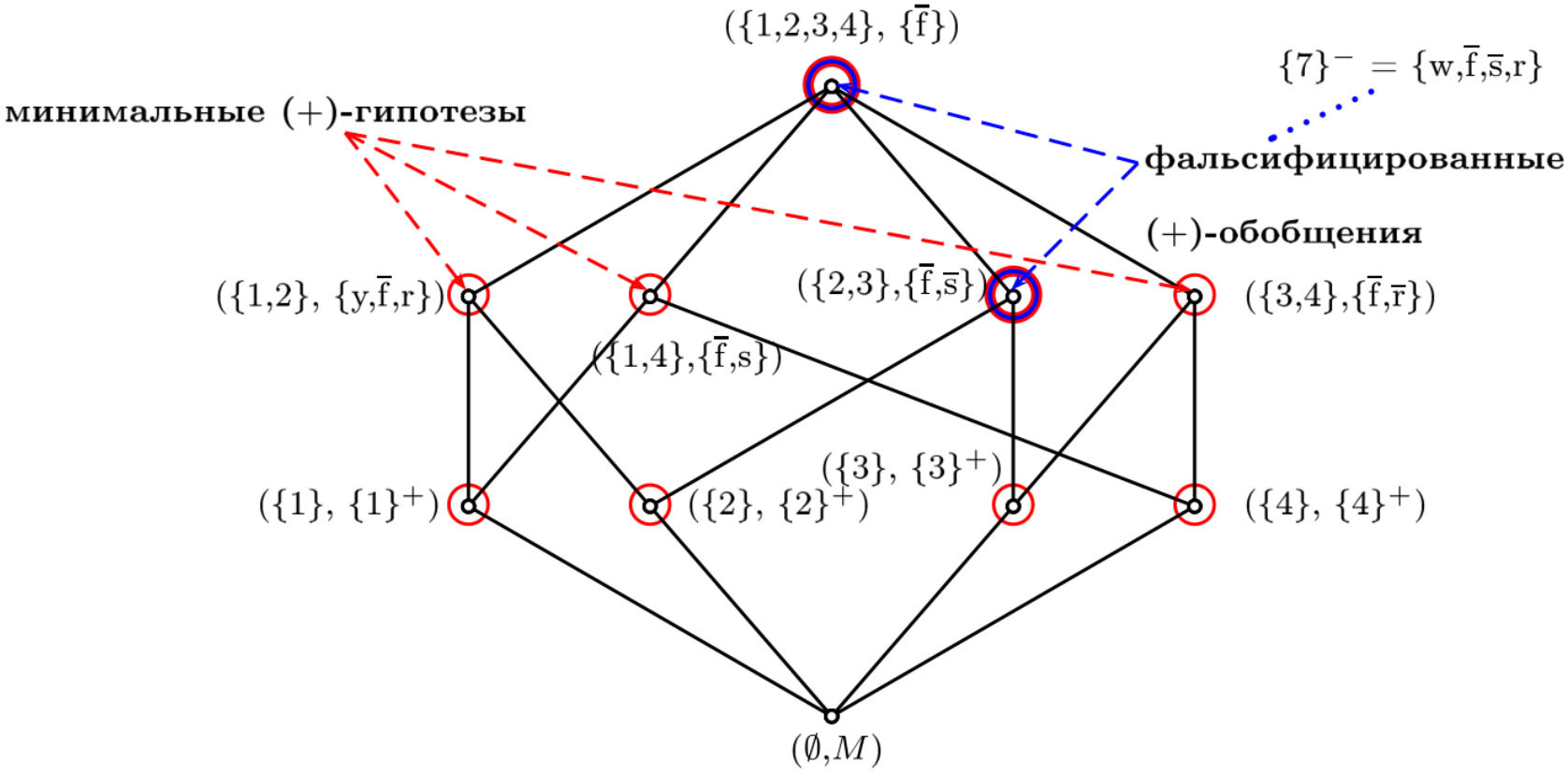
Сокращения:

"g" - зеленый, "y" - желтый, "w" - белый, "f" - твердый, "$\bar{\text{f}}$" - нетвердый,

"s" - гладкий, "$\bar{\text{s}}$" - негладкий, "r" - круглый,

"$\bar{\text{r}}$" - некруглый.

Решетка понятий положительного контекста



G \ M	w	y	g	b	f	f̄	s	s̄	r	r̄	фрукт
1		×				×	×		×		+
2		×				×		×	×		+
3			×			×		×		×	+
4				×		×	×			×	+
5			×		×		×			×	−
6	×				×		×			×	−
7	×					×		×	×		−

Классификация недоопределенного примера g^τ

- Если g_τ^τ содержит в качестве подмножества положительную гипотезу и не содержит ни одной отрицательной гипотезы, то g_τ **классифицируется положительно** (предсказывается наличие целевого признака w).
- Если g_τ^τ содержит в качестве подмножества отрицательную гипотезу и не содержит ни одной положительной гипотезы, то g_τ **классифицируется отрицательно** (предсказывается отсутствие целевого признака w).
- Если g_τ^τ содержит в качестве подмножеств гипотезы обоих знаков или если g_τ^τ вообще не содержит в качестве подмножеств ни положительных ни отрицательных гипотез, то классификация объекта, соответственно, **противоречива** или **недоопределенна**.

Как следует из определения, для классификации достаточно иметь множество всех **минимальных** (относительно $\subseteq$) гипотез.

Классификация недоопределенного примера манго

	G\M	w	y	g	b	f	$\bar{\text{f}}$	s	$\bar{\text{s}}$	r	$\bar{\text{r}}$	**фрукт**
1	яблоко		X				X	X		X		+
2	грейпфрут		X				X		X	X		+
3	киви			X			X		X		X	+
4	слива				X		X	X			X	+
5	кубик			X		X		X			X	−
6	яйцо	X				X		X			X	−
7	теннисный мяч	X					X		X	X		−
8	манго		X				X	X			X	τ

Объект **манго** классифицируется положительно, поскольку:

- для $(+)$-гипотезы $\{\bar{\text{r}}, \bar{\text{f}}\}$:
  $\{\bar{\text{r}}, \bar{\text{f}}\} \subseteq$ манго$^\tau = \{\text{y}, \bar{\text{f}}, s, \bar{\text{r}}\}$;
- для $(-)$-гипотез $\{\text{w}\}$ и $\{\text{f, s, } \bar{\text{r}}\}$:
  $\{\text{w}\} \not\subseteq$ манго$^\tau$, $\{\text{f, s, } \bar{\text{r}}\} \not\subseteq$ манго$^\tau$.

Классификация недоопределенного примера мыло

	G\M	w	y	g	b	f	$\bar{f}$	s	$\bar{s}$	r	$\bar{r}$	**фрукт**
1	яблоко		X				X	X		X		+
2	грейпфрут		X				X		X	X		+
3	киви			X			X		X		X	+
4	слива				X		X	X			X	+
5	кубик			X		X		X			X	−
6	яйцо	X				X		X			X	−
7	теннисный мяч	X					X		X	X		−
8	мыло	X				X		X		X		τ

Объект **мыло** классифицируется отрицательно, поскольку:

- для $(-)$-гипотезы $\{\mathrm{w}\}$:
  $\{\mathrm{w}\} \subseteq$ **мыло**$^{\tau} = \{\mathrm{w}, \mathrm{f}, \mathrm{s}, \mathrm{r}\}$,
- ни одна $(+)$-гипотеза не является подмножеством множества **мыло**$^{\tau} = \{\mathrm{w}, \mathrm{f}, \mathrm{s}, \mathrm{r}\}$.

Классификация примера шампиньон

	G\M	w	y	g	b	f	$\bar{\text{f}}$	s	$\bar{\text{s}}$	r	$\bar{\text{r}}$	**фрукт**
1	яблоко		X				X	X		X		+
2	грейпфрут		X				X		X	X		+
3	киви			X			X		X		X	+
4	слива				X		X	X			X	+
5	кубик			X		X		X			X	−
6	яйцо	X				X		X			X	−
7	теннисный мяч	X					X		X	X		−
8	шампиньон	X					X	X			X	τ

Объект **шампиньон** классифицируется противоречиво, поскольку:

- для $(+)$-гипотезы $\{\bar{\text{f}}, s\}$:
  $\{\bar{\text{f}}, s\} \subseteq$ **шампиньон**$^{\tau} = \{\text{w}, \bar{\text{f}}, s, \bar{\text{r}}\}$;
- для $(-)$-гипотезы $\{\text{w}\}$:
  $\{\text{w}\} \subseteq$ **шампиньон**$^{\tau} = \{\text{w}, \bar{\text{f}}, s, \bar{\text{r}}\}$.

Классификация примера арбуз

	G\M	w	y	g	b	f	$\bar{f}$	s	$\bar{s}$	r	$\bar{r}$	**фрукт**
1	яблоко		X				X	X		X		+
2	грейпфрут		X				X		X	X		+
3	киви			X			X		X		X	+
4	слива				X		X	X			X	+
5	кубик			X		X		X			X	−
6	яйцо	X				X		X			X	−
7	теннисный мяч	X					X		X	X		−
8	арбуз			X		X		X		X		τ

Объект **арбуз** классифицируется неопределенно, поскольку

- для $(+)$-гипотез $\{y, \bar{f}, r\}$, $\{\bar{f}, s\}$ и $\{\bar{f}, \bar{r}\}$:
  $\{y, \bar{f}, r\} \not\subseteq$ арбуз$^\tau = \{g, f, s, r\}$,
  $\{\bar{f}, s\} \not\subseteq$ арбуз$^\tau = \{g, f, s, r\}$, $\{\bar{f}, \bar{r}\} \not\subseteq$ арбуз$^\tau = \{g, f, s, r\}$.
- для $(-)$-гипотез $\{w\}$ и $\{f, s, \bar{r}\}$:
  $\{w\} \not\subseteq$ арбуз$^\tau = \{g, f, s, r\}$,
  $\{f, s, \bar{r}\} \not\subseteq$ арбуз$^\tau = \{g, f, s, r\}$.

Обучение в решетках на множествах графов

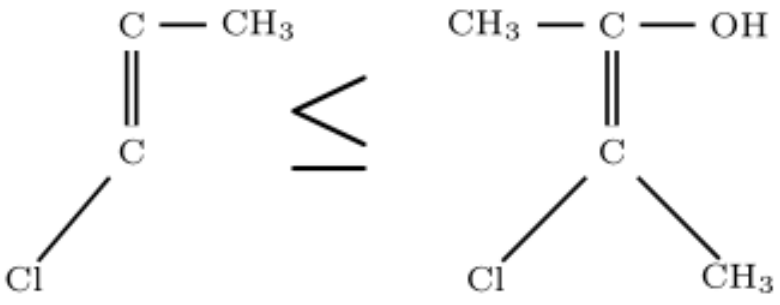
1. Частичный порядок на помеченных графах
2. (Полу)решетка на множествах графов
3. Машинное обучение по примерам, заданных графами

Порядок на помеченных графах

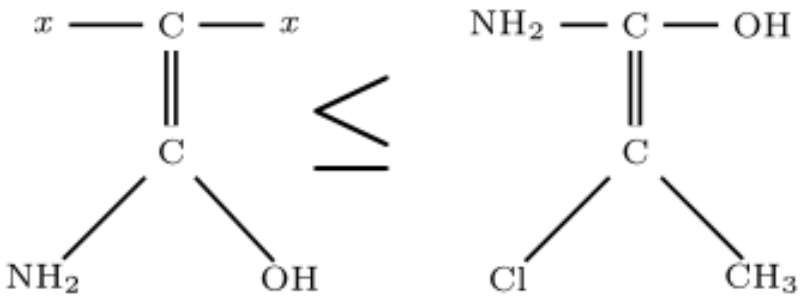
Граф $\Gamma_1 := ((V_1, l_1), E_1)$ **доминирует** над графом $\Gamma_2 := ((V_2, l_2), E_2)$ или $\Gamma_2 \leq \Gamma_1$ если существует взаимнооднозначное отображение $\varphi\colon V_2 \to V_1$, которое

- учитывает ребра: $(v, w) \in E_2 \Rightarrow (\varphi(v), \varphi(w)) \in E_1$,
- учитывает порядок на метках: $l_2(v) \leq l_1(\varphi(v))$.

Пример:



метки вершин неупорядоченны

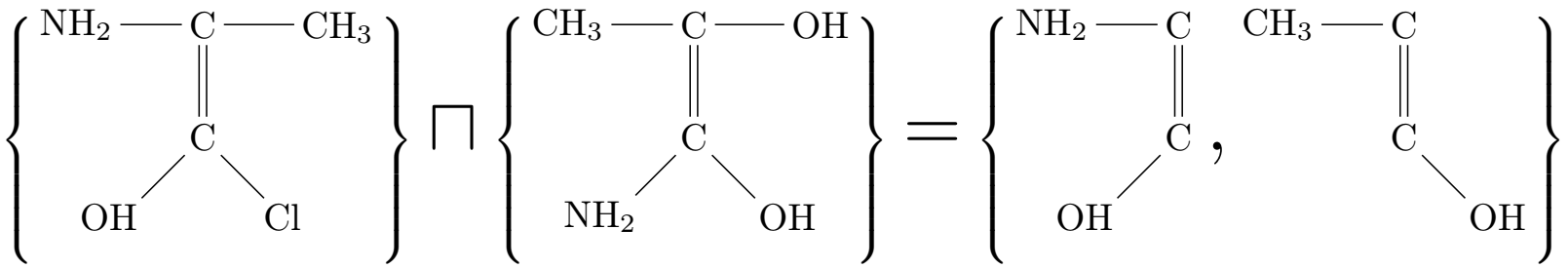


$x \preceq A$ для произвольной вершинной метки $A \in \mathcal{L}$

Полурешетка на множествах графов

$\{X\} \sqcap \{Y\} := \{Z \mid Z \leq X, Y, \quad \forall Z_* \leq X, Y \quad Z_* \not\geq Z\}$
= Множество всех максимальных общих подграфов графов Γ_1 и Γ_2.

Пример:



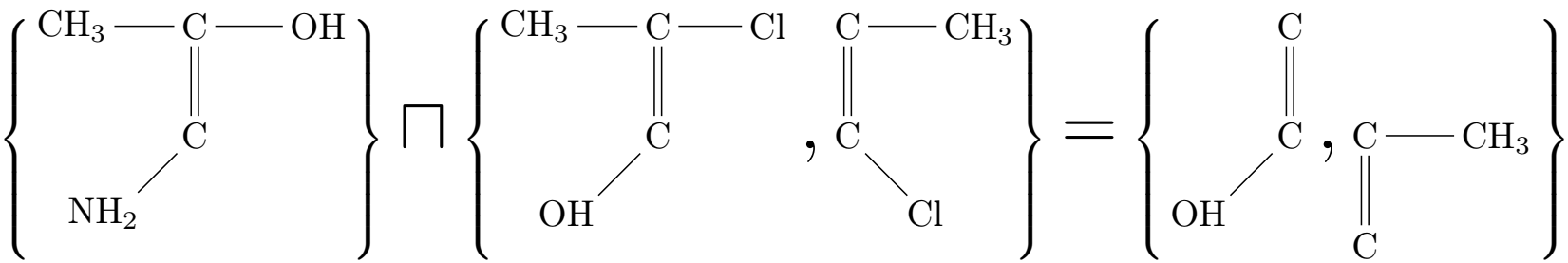
Пересечение множеств графов

Для множеств графов $\mathcal{X} = \{X_1, \ldots, X_k\}$ и $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ операция задается следующим образом

$$\mathcal{X} \sqcap \mathcal{Y} := \text{ MAX } (\bigcup_{i,j}(\{X_i\} \sqcap \{Y_j\}))$$

Операция $\sqcap$ идемпотентна, коммутативна и ассоциативна.

Пример:



Отношение вложения на множествах графов

Поскольку $\sqcap$ задает полурешетку на множестве помеченных графов, через нее как можно определить отношение естественного порядка следующим образом.

$$\mathcal{G} \sqsubseteq \mathcal{H} \iff \mathcal{G} \sqcap \mathcal{H} = \mathcal{G}$$

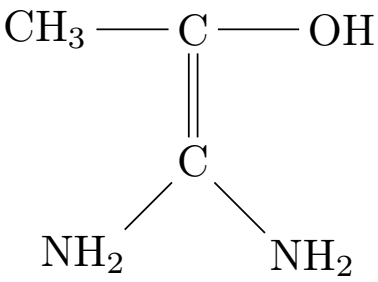
Заметим, что такое определение отношения $\sqsubseteq$ эквивалентно следующему:

$$\mathcal{G} \sqsubseteq \mathcal{H} \iff \forall g \in \mathcal{G}\ \exists h \in \mathcal{H} \text{ такой что } g \leq h$$

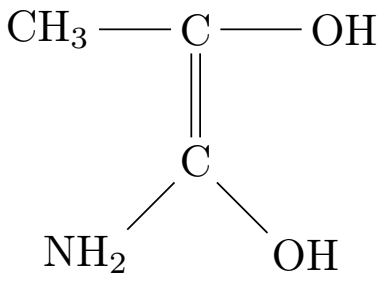
Обучающая выборка

Положительные примеры:

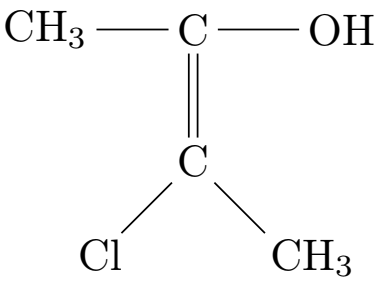
G_1 :



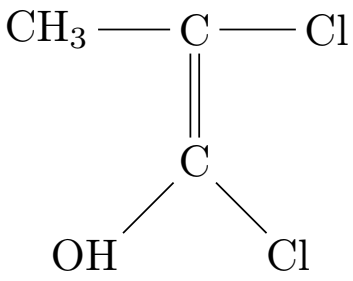
G_2 :



G_3 :



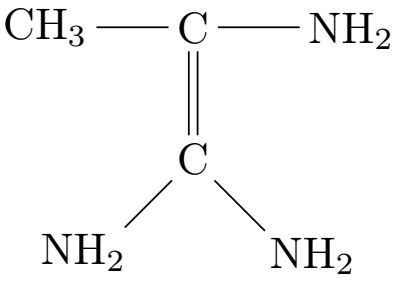
G_4 :



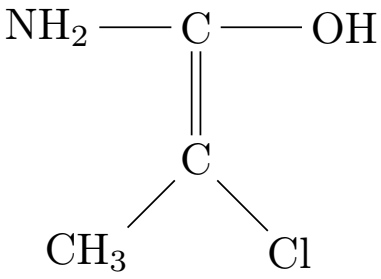
Обучающая выборка

Отрицательные примеры:

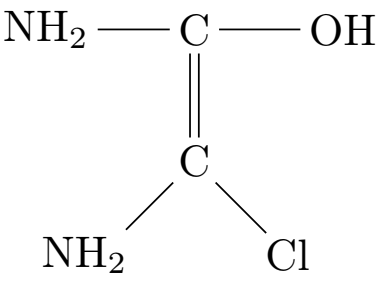
G_5 :



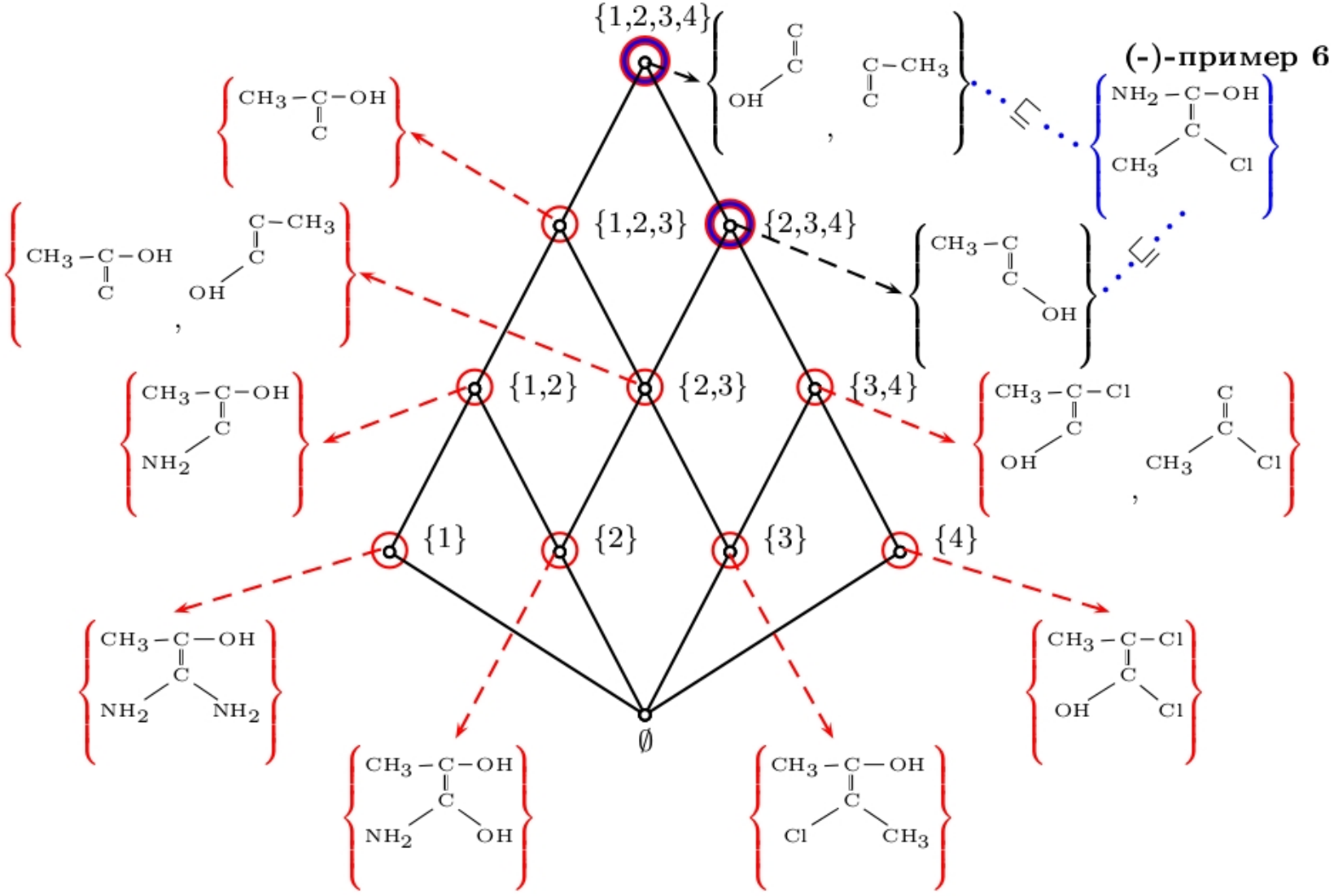
G_6 :



G_7 :



Решетка положительных сходств



положительные примеры 1, 2, 3, 4

Классификация неопределенных примеров

Классификация проводится точно так же как и в случае объектно-признакового представления (представления с помощью контекстов), с заменой теоретико-множественного отношения $\subseteq$ на $\sqsubseteq$:

Недоопределенный пример g_τ, представленный множеством графов $\mathcal{G}$

- **классифицируется положительно** (предсказывается наличие целевого признака w) если существует положительная гипотеза $\mathcal{H}_+$ такая что $\mathcal{H}_+ \sqsubseteq \mathcal{G}$ и ни для одной отрицательной гипотезы $\mathcal{H}_-$ не имеет места $\mathcal{H}_- \sqsubseteq \mathcal{G}$.
- **классифицируется отрицательно** (предсказывается отсутствие целевого признака w) если существует отрицательная гипотеза $\mathcal{H}_-$ такая что $\mathcal{H}_- \sqsubseteq \mathcal{G}$ и ни для одной положительной гипотезы $\mathcal{H}_+$ не имеет места $\mathcal{H}_+ \sqsubseteq \mathcal{G}$.

Классификация неопределенных примеров

Классификация проводится точно так же как и в случае объектно-признакового представления (представления с помощью контекстов), с заменой теоретико-множественного отношения $\subseteq$ на $\sqsubseteq$:

Недоопределенный пример g_τ, представленный множеством графов $\mathcal{G}$

- Если $\mathcal{G}$ содержит (в смысле отношения $\sqsubseteq$) в качестве подмножеств гипотезы обоих знаков или если $\mathcal{G}$ вообще не содержит в качестве подмножеств ни положительных ни отрицательных гипотез, то классификация объекта примера g_τ соответственно, **противоречива** или **недоопределенна**.

Проекции как средство приближения.1

Мотивация: Даже задача проверки отношения $\leq$ для помеченных графов NP-полна (эквивалентна задаче ИЗОМОРФИЗМ ПОДГРАФУ).

Отображение ψ называется **проекцией** упорядоченного множества $(D, \sqsubseteq)$ если ψ обладает следующими свойствами:

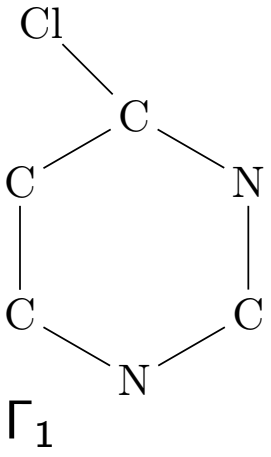
идемпотентностью: $\psi(\psi(x)) = \psi(x)$,

монотонностью: если $x \sqsubseteq y$, то $\psi(x) \sqsubseteq \psi(y)$,

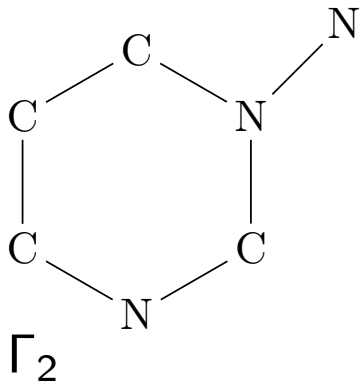
равномерное сжатие: $\psi(x) \sqsubseteq x$, $y \sqsubseteq \psi(x) \Rightarrow \exists z : y = \psi(z)$

Проекции как средство приближения.2

Пример: Проекции графов с пометками: $\psi_n(\Gamma)$ отображает граф Γ в множество его n-вершинных цепей, не доминируемых (в смысле определения отношения $\leq$) другими n-вершинными цепями. В данном примере $n = 3$.



Γ_1



Γ_2

$\psi_3(\Gamma_1)$:

Cl — C — C
Cl — C — N
Cl — C — N
C — C — C
C — C — N
C — N — C
N — C — N

$\psi_3(\Gamma_2)$:

Cl — N — C
C — C — C
C — C — N
C — N — C
N — C — N

Свойства проекций

Любая проекция полной полурешетки $(D, \sqcap)$ сохраняет операцию $\sqcap$, т.е. для любых $X, Y \in D$

$$\psi(X \sqcap Y) = \psi(X) \sqcap \psi(Y).$$

Пример. Проекция помеченных графов $\psi_n(\Gamma)$ отображает граф Γ в множество всех его n-вершинных цепей, не доминируемых другими n-вершинными цепями. В данном примере $n = 3$.

$\psi_3(\Gamma_1)$:

Cl — C — C
Cl — C — N
Cl — C — N
C — C — C
C — C — N
C — N — C
N — C — N

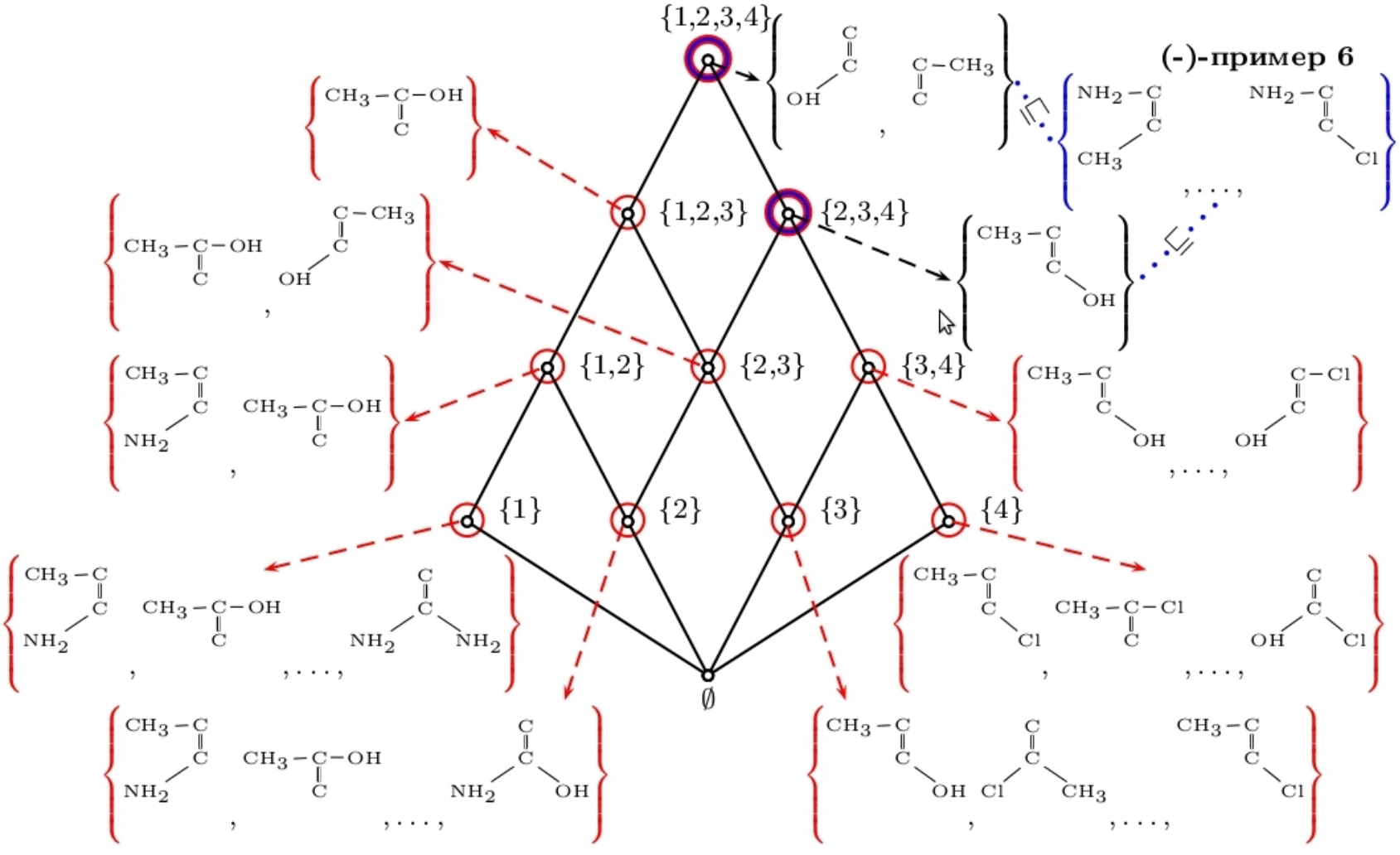
$\psi_3(\Gamma_2)$:

Cl — N — C
C — C — C
C — C — N
C — N — C
N — C — N

$\psi_3(\Gamma_1) \sqcap \psi_3(\Gamma_2)$:

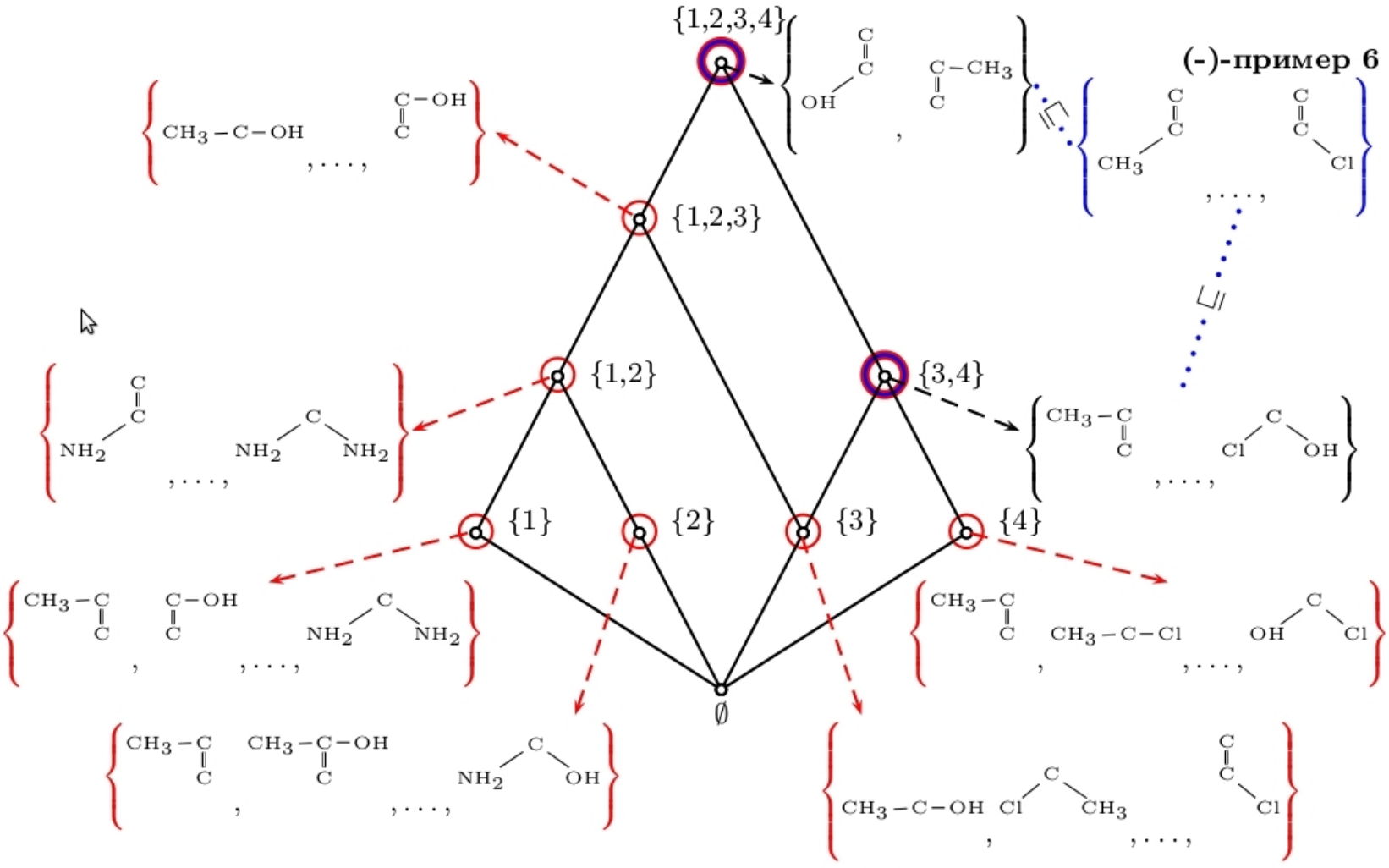
Cl — x — C
C — C — C
C — C — N
C — N — C
N — C — N

4-проекция



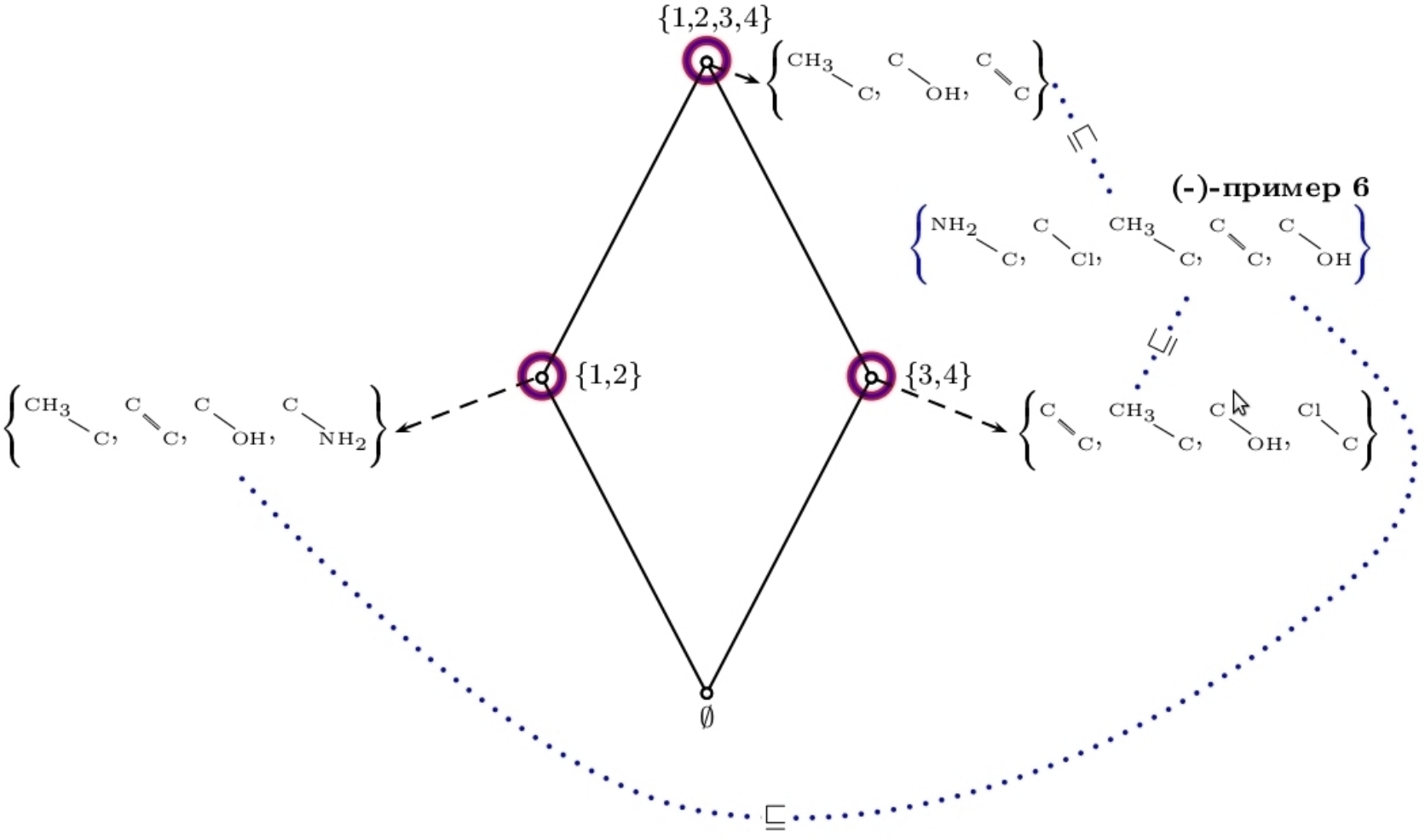
положительные примеры 1, 2, 3, 4

3-проекция



положительные примеры 1, 2, 3, 4

2-проекция



положительные примеры 1, 2, 3, 4

Анализ данных с помощью графовых моделей

- Изучение соотношений структура - токсичность спиртов
- Прогноз стратегий и путей биотрансорфмации веществ в организме человека
- Прогноз канцерогенности и хронической токсичности галогензамещенных алифатических углеводородов
- Прогноз канцерогенности полициклических ароматических углеводородов

Анализ токсичности с помощью ДСМ-метода

Bioinformatics, 19(2003)

V. G. Blinova, D. A. Dobrynin, V. K. Finn, S. O. Kuznetsov and E. S. Pankratova

Соревнование по предсказательной токсикологии: (PTC) Workshop at the joint 5th European Conference on Knowledge Discovery in Databases (KDD'2001) and the 12th European Conference on Machine Learning (ECML'2001), Freiburg.

Организаторы: Группы машинного обучения Университетов Фрайбурга, Оксфорда и Уэльса.

Экперты по токсикологии: US Environmental Protection Agency, US National Institute of Environmental, Health Standards.

Анализ токсичности с помощью ДСМ-метода. 2

Bioinformatics, 19(2003)

Обучающая выборка: Данные National Toxicology Program (NTP), включающие от 120 до 150 положительных примеров и от 190 до 230 отрицательных примеров токсичности: молекулярные графы с указанием того, является ли вещество токсичным для четырех поло-видовых групп:
$\{$самец, самка$\} \times \{$мыши, крысы$\}$.

Проверочная выборка: Данные Food and Drug Administration (FDA): около 200 химических веществ с известной молекулярной структурой, чья (не)токсичность была известна лишь для организаторов и должна была предсказываться участниками соревнования.

Участники: 12 групп исследователей (со всего мира), каждая из которых могла предоставить до 4 предсказательных моделей для каждой поло-видовой группы.

Анализ токсичности с помощью ДСМ-метода. 3

Средства сравнения результатов: ROC-диаграммы

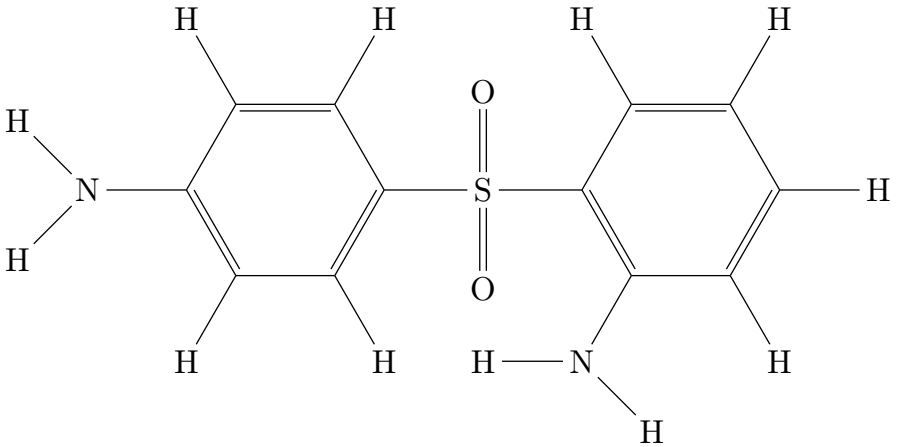
Этапы соревнования:

1. Кодировка химических структур в терминах признаков,
2. Порождение правил классификаций,
3. Проведение классификаций с помощью порожденных правил.

Организаторы соревнования публиковали результаты каждого этапа. В частности, на сайте соревнования размещались кодировки химических структур, порожденные правила классификаций и сами классификации.

Пример кодировки

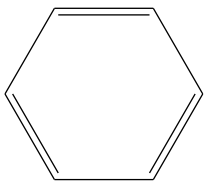
Химическая структура



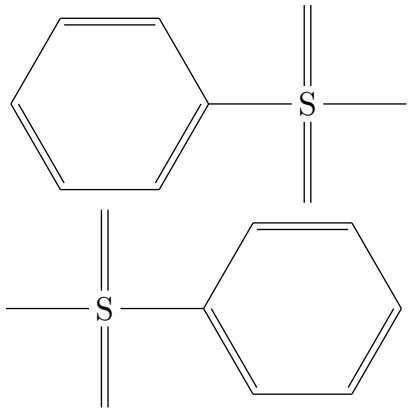
Полный список дескрипторов

6,06 x2
0200331 x2
1300241 x2
2400331 x2
0264241
0262241

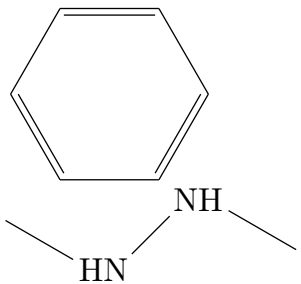
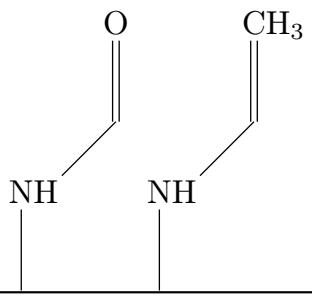
6,06 (циклические дескрипторы)



0200331 (линейные дескрипторы)



Некоторые положительные гипотезы

Молекулярный граф	Дескрипторы ФКСП (кодировка)	Число предсказаний в группах
	6,06 0200021	2FR
	0201131 0202410	1FR 1MM

Узорные структуры (Pattern Structures)

[Ganter, Kuznetsov 2001]

Тройка $(G, \underline{D}, \delta)$ есть **узорная структура** если

- G есть множество ("множество объектов");
- $\underline{D} = (D, \sqcap)$ есть полурешетка по операции пересечения;
- $\delta : G \to D$ есть отображение;
- множество $\delta(G) := \{\delta(g) \mid g \in G\}$ порождает полную подполурешетку $(D_\delta, \sqcap)$ полурешетки $(D, \sqcap)$.

Возможные источники происхождения операции $\sqcap$:

- Множество объектов G, каждый из которых имеет описание из упорядоченного множества P;
- Частично упорядоченное множество $(P, \leq)$ "описаний" ($\leq$ – отношение типа "быть более общим описанием" (отношение общности));
- (Дистрибутивная) решетка порядковых идеалов упорядоченного множества $(P, \leq)$.

Другой пример узорных структур: бикластеры сходных значений

Параметр близости $\varepsilon = 1$

	p_1	p_2	p_3	p_4
o_1	1	2	2	1
o_2	2	1	1	0
o_3	2	1	2	3
o_4	0	0	3	2

Бикластеры:

$$\langle \{o_1, o_2, o_3\}; \{p_1, p_2, p_3\} \rangle$$

$$\langle \{o_3, o_4\}; \{p_3, p_4\} \rangle$$

Заключение

Решетки замкнутых описаний - удобное средство кластеризации, построения таксономий предметных областей, точных и приближенных зависимостей в данных

Единые математические средства для разных данных: частичный порядок на описаниях, соответствия Галуа, замкнутые описания, пересечения описаний, понятия, импликации, ассоциативные правила.

Единые алгоритмические средства, средства приближений и ленивых классификаций.

Ближайшие задачи применительно к наукам о данных:

- быстрые параллельные алгоритмы ленивой классификации на данных сложной структуры,
- однопроходная n-кластеризация,
- порядковые и решеточные аналоги SVM на данных сложной структуры.

Спасибо за внимание!

Вопросы?