Соловьев С.Ю.

Постановки задач современной информатики

Задачи анализа данных

2015 – 2021

Напоминание: **З а д а ч а**

Дано	Известно
Исходные данные	***Свойства исх. данных***

Алгоритм / Метод / Способ / Схема

Требуется	такое, что
Результирующие данные	***Свойства рез. данных***

Анализ данных – раздел информатики, ориентированный на выявление и описание связей признаков, заданных в количественных и качественных шкалах.

data mining;
интеллектуальный анализ данных

Основные разделы анализа данных

Корреляционный анализ ⇨ Статистический анализ

Регрессионный анализ ⇨ Статистический анализ

Факторный анализ ⇨

Многомерное шкалирование ⇨ ⇨

Дисперсионный анализ

Дискриминантный анализ

Планирование эксперимента

Анализ временных рядов

Поиск ассоциативных правил ⇨ ⇨ ⇨

Кластерный анализ ⇨ Распознавание образов

Предистория анализа данных

Школы: Ю.И.Журавлев,
Н.Г.Загоруйко,
А.Д.Закревский,
В.К.Финн

Конференции

Машинные методы обнаружения закономерностей

Новосибирск 1976, Рига 1979, Новосибирск 1980

Термин Data Mining 1989 г.р. (Г. Пятецкий-Шапиро)

Факторный анализ

Факторный анализ – раздел анализа данных, в котором разрабатываются методы выявления скрытых факторов, отвечающих за наличие корреляций между наблюдаемыми признаками.

многомерного статистического анализа

Факторный анализ. *Пролог*

X:

	Двоеборие	Рывок	Толчок	Возраст
1.	400	180	220	24
2.	420	195	225	27
3.	440	200	240	21
4.	435	195	240	26
5.	465	205	260	22
выб.ср.=	432	195	237	24

← *Переменные:* $V_1 \dots V_m$

Ковариационная матрица **Cov(X)=**

466	170	296	-24
170	70	100	-7
296	100	196	-17
-24	-7	-17	5.2

Корреляционная матрица **Cor(X)=**

1	0.94	0.98	-0.48
	1	0.85	-0.36
		1	-0.52
			1

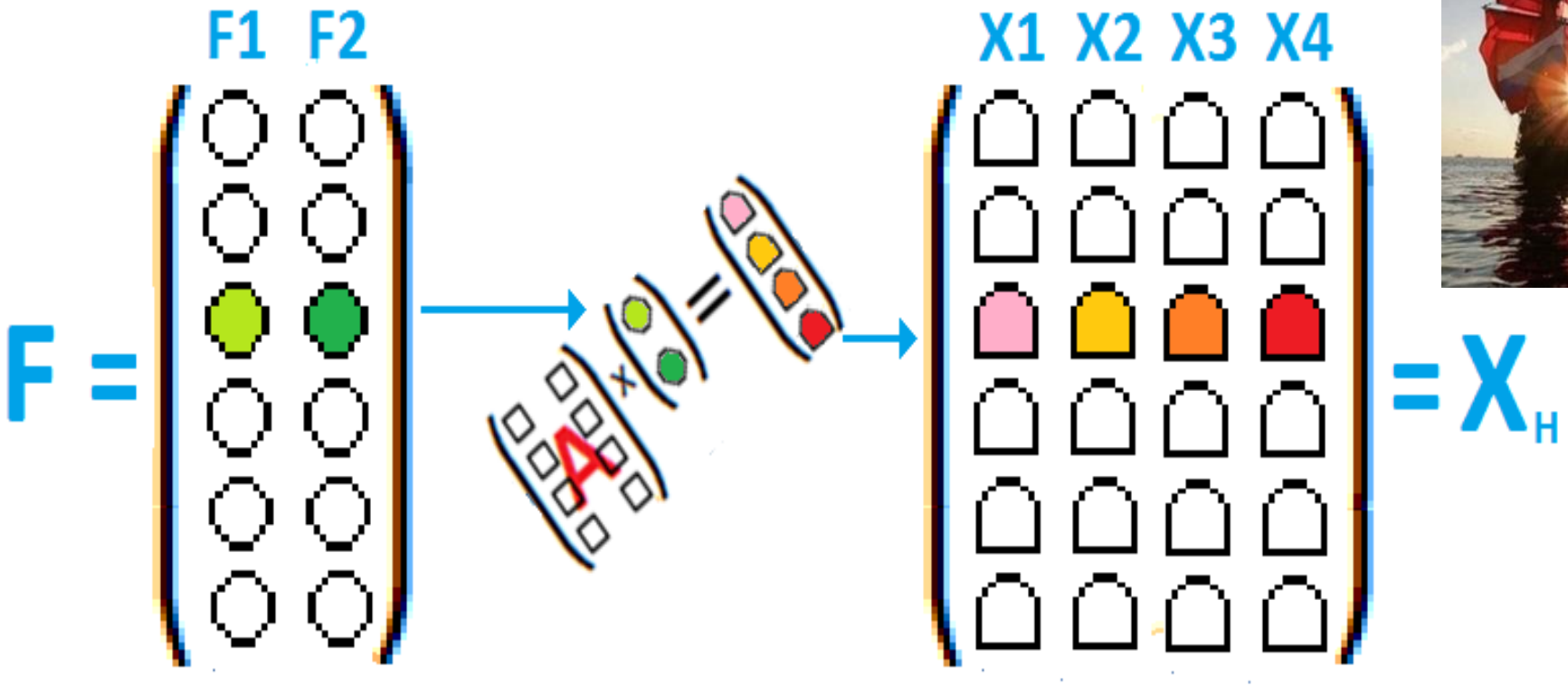
Факторный анализ. $X \to X_ц, X_н$

X:	400	180	220	24	Cov(X)	Cor(X)
	420	195	225	27		
	440	200	240	21		
	435	195	240	26		
	465	205	260	22		
выб.ср.=	432	195	237	24		

$X_ц$:	-32	-15	-17	0	$Cov(X_ц)$=	Cov(X)
	-12	0	-12	3		$Cor(X_ц)$ = Cor(X)
	8	5	3	-3		
	3	0	3	2		
	33	10	23	-2		
сквотк.=	21.58	8.37	14	2.28		

$X_н$:	-1.48	-1.79	-1.21	0	$Cov(X_н)$=	$Cor(X_н)$=	Cor(X)
	-0.56	0	-0.86	1.32			
	0.37	0.60	0.21	-1.32			
	0.14	0	0.21	0.88			
	1.53	1.19	1.65	-0.88			

Факторный анализ



$$cor(F) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\begin{aligned} X_1 &= a_{11}F_1 + a_{12}F_2 \\ X_2 &= a_{21}F_1 + a_{22}F_2 \\ X_3 &= a_{31}F_1 + a_{32}F_2 \\ X_4 &= a_{41}F_1 + a_{42}F_2 \end{aligned}$$

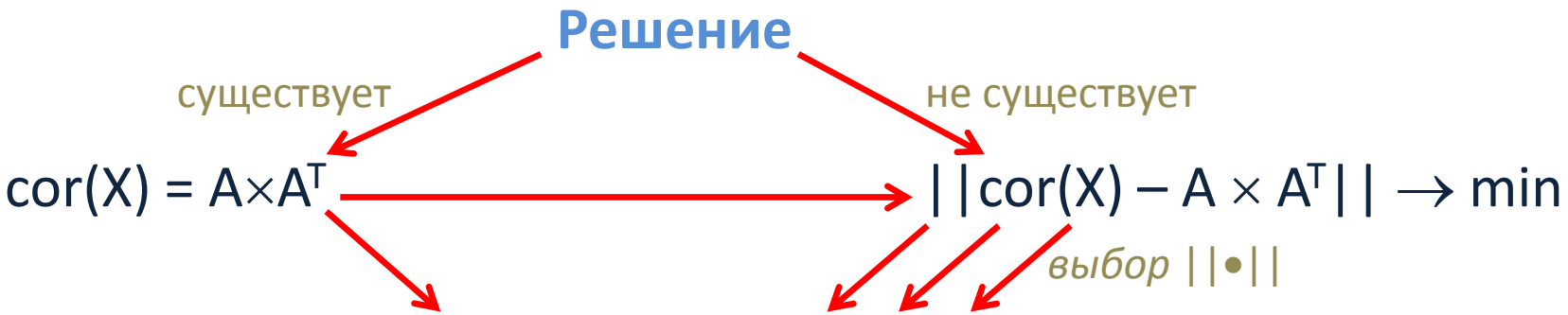
$$cor(X) = \begin{pmatrix} 1 & 0.94 & 0.98 & -0.48 \\ & 1 & 0.85 & -0.36 \\ & & 1 & -0.52 \\ & & & 1 \end{pmatrix}$$

Исследовательская задача факторного анализа

Дано	Известно
матрица **X ; ; ;**	$\mathbf{X_{m\times n}}$ – нормированная **; ; ;**
число **k**	**k < n**

Исследование решения: *существование, единственность*

Требуется	такие, что
матрицы $\mathbf{A_{n\times k}}$, $\mathbf{F_{n\times k}}$	$\mathbf{A\times F^T = X^T}$ **&** **cor(F) = E**



если A – решение, то AV – также решение

Задача факторного анализа

Дано	Известно
матрица **X ; ; ;** число **k ; ; ; ; ;** **А**лгоритм выбора канонического решения	$\mathbf{X}_{m\times n}$ – нормированная **; ; ;** **k < n ; ; ; ; ; ; ; ; ; ; ; ; ; ;** **||•||**

Метод(**A**, •)

Требуется	такие, что
матрица $\mathbf{A}_{n\times k}$ **; ; ;** **ее Оценка**	**Оценка =** $\|\|\mathbf{cor(X) - A \times A^T}\|\| \rightarrow \mathbf{min}$

Методы(A,●) факторного анализа

Метод главных факторов
Метод наименьших квадратов
Метод максимального правдоподобия
Альфа-факторный анализ
Факторизация образов

	Метод глав. факторов	Метод макс. правдоподоб	Альфа факторный	Анализ образов
$A_{6\times2}$	0.73 — 0.32	0.75 — 0.30	0.70 0.44	0.58 0.13
	0.64 — 0.28	0.70 — 0.27	0.59 0.38	0.54 0.14
	0.55 — 0.24	0.60 — 0.18	0.50 0.33	0.48 0.13
	0.51 0.47	0.43 0.36	0.59 — 0.38	0.37 — 0.27
	0.44 0.41	0.51 0.61	0.50 — 0.33	0.39 — 0.26
	0.37 0.34	0.53 0.25	0.42 — 0.27	0.29 — 0.24

≈Задача определения минимального числа общих факторов

Дано: упорядоченные по важности факторы **F1, …, Fm**

Пример

Фактор	Важность
1	51.4
2	17.2
3	10.3
4	7.7
5	3.9
6	3.3
7	2.8
8	2.1
9	0.8
10	0.5

Методы*)

число **k**	**k** наиболее важных

*)Если Важность=С.зн, то **k** = К-воС.зн. > 1.

Если Важность=Дисп, то **k** : сумма Дисп ≈ 90%.

Статистический критерий Бартлетта-Уилкса

≈Задача определения минимального числа общих факторов. *Принцип x/100-x*

Фактор	Важность
1	51.4
2	17.2
3	10.3
4	7.7
5	3.9
6	3.3
7	2.8
8	2.1
9	0.8
10	0.5

Вращение факторов

“В хорошем факторном решении каждая переменная характеризуется преобладающим влиянием некоторого одного фактора.”

Корреляционная матрица						A	
1	0,50	0,32	0,20	0,22	0,12	0,6	0,6
0,50	1	0,21	0,15	0,16	0,07	0,4	0,4
032	0,21	1	0,44	0,67	0,41	0,7	0,2
0,20	0,15	0,44	1	0,56	0	0,6	-0,2
0,22	0,16	0,67	0,56	1	0,51	0,8	-0,4
0,12	0,07	0,41	0,33	0,51	1	0,5	-0,3

$$\begin{bmatrix} 0,6 & 0,6 \\ 0,4 & 0,4 \\ 0,7 & -0,2 \\ 0,6 & -0,2 \\ 0,8 & -0,4 \\ 0,5 & -0,3 \end{bmatrix} \cdot \begin{bmatrix} 0,87 & 0,50 \\ -0,50 & 0,87 \end{bmatrix} = \begin{bmatrix} 0,22 & 0,82 \\ 0,15 & 0,55 \\ 0,71 & 0,18 \\ 0,64 & 0,12 \\ 0,90 & 0,05 \\ 0,58 & -0,01 \end{bmatrix}$$

$\theta = 30°$

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

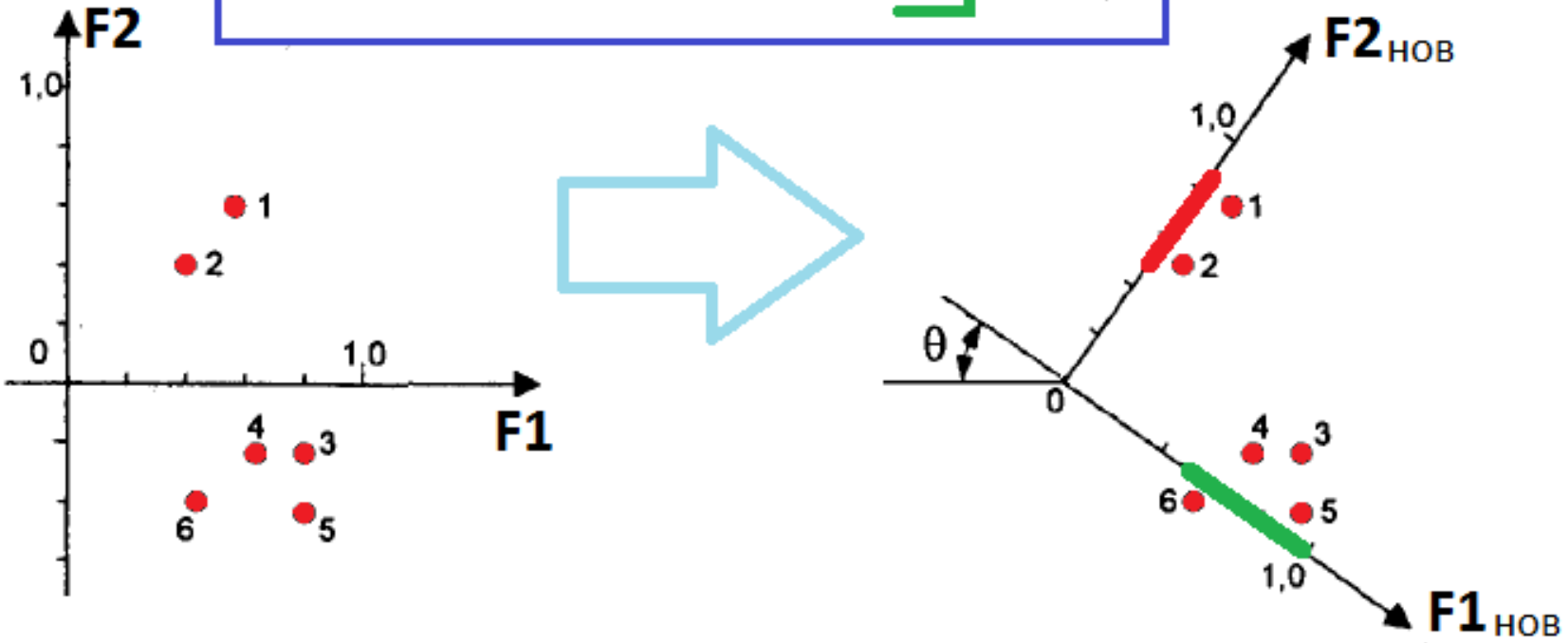
Методы ортогонального вращения:

- -- варимакс
- -- квартимакс
- -- эквимакс
- -- биквартимакс

Методы косоугольного вращения.

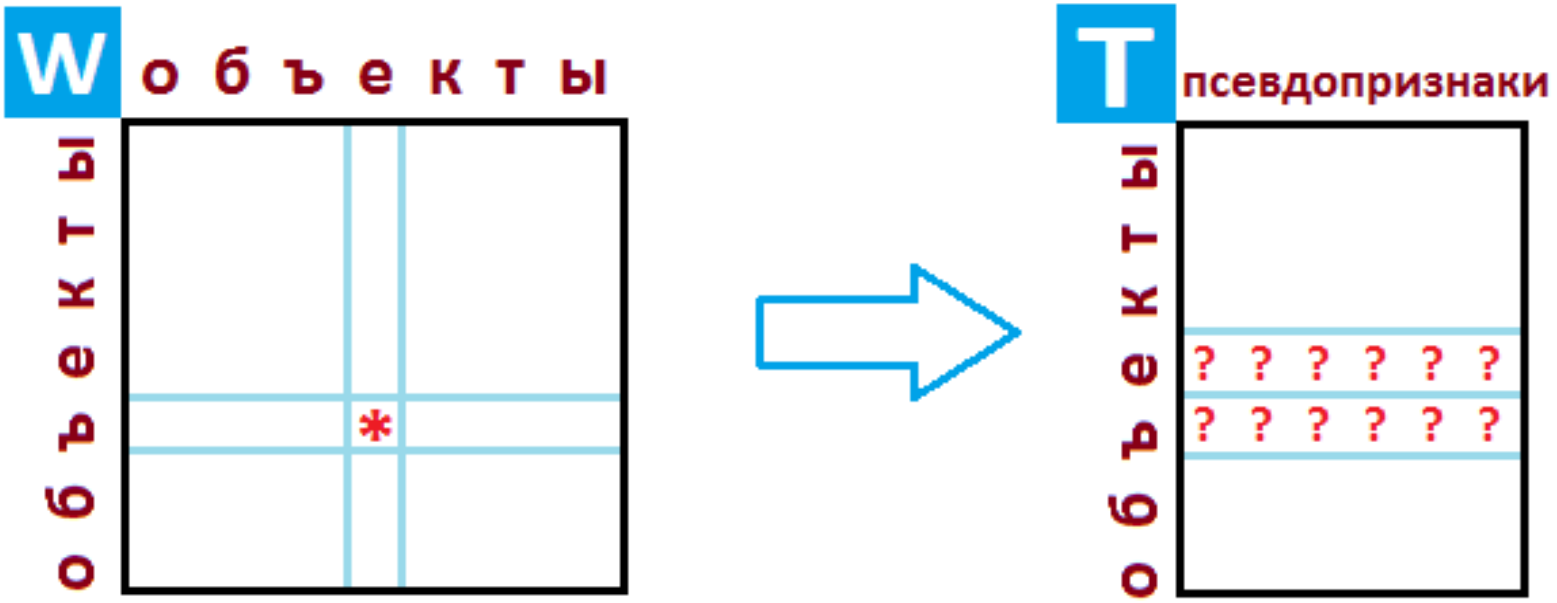
Вращение факторов. *Иллюстрация*

$$\begin{bmatrix} 0,6 & 0,6 \\ 0,4 & 0,4 \\ 0,7 & -0,2 \\ 0,6 & -0,2 \\ 0,8 & -0,4 \\ 0,5 & -0,3 \end{bmatrix} \cdot \begin{bmatrix} 0,87 & 0,50 \\ -0,50 & 0,87 \end{bmatrix} = \begin{bmatrix} 0,22 & 0,82 \\ 0,15 & 0,55 \\ 0,71 & 0,18 \\ 0,64 & 0,12 \\ 0,90 & 0,05 \\ 0,58 & -0,01 \end{bmatrix} \qquad \theta = 30°$$



Многомерное шкалирование

Многомерное шкалирование – раздел анализа данных, ориентированный на разработку методов сопоставления изучаемым объектам их описаний в многомерном псевдопризнаковом пространстве.



Виды задач многомерного шкалирования

Неметрическое шкалирование ➔

W – [неточная] матрица различий между объектами;

T – точки в метрическом пространстве.

Метрическое шкалирование ➔➔

W – [неточная] матрица расстояний в метрическом пространстве;

T – точки в метрическом пространстве.

---

*) Метрическое пространство = евклидово пространство

Задача неметрического шкалирования

преобразование неметрической информации в метрическую

объекты $B = \{ B_1 \dots B_n \}$; ; матрица различий W ; ; ; число k ; ; ; начальное приближение $T_{(0)}$	w_{ij} – показатель различия B_i от B_j ; ; ; $k < n-1$ *)

Методы неметрического шкалирования

Таблица T	$T = \{(b_{i1} \dots b_{ik}) \mid B_i \in B \} \in R^k$; ; ; $w_{ij} \le w_{pq} \Rightarrow d_{ij} \le d_{pq}$, где $d^2_{ij} = (b_{i1}-b_{j1})^2 + \dots + (b_{ik}-b_{jk})^2$

*) при k=n-1 решение существует (доказано)

при k<n-1 можно говорить о псевдорешении в смысле нек. функционала

Идея методов неметрического шкалирования

(А) **W** порождает эталонный порядок$_{(э)}$ на **B**

(Б) **T$_{(i)}$** порождает порядок$_{(i)}$ расстояний на **B** (i = 0, 1, …)

Процесс : **T$_{(0)}$ → T$_{(1)}$ … → T$_{(i)}$ → … → T$_{(стоп)}$,**

такой что порядок$_{(i)}$ → порядок$_{(э)}$

Соображение

Растянуть малые расстояния, соответствующие большим различиям.

Сжать большие расстояния, соответствующие малым различиям.

Методы неметрического шкалирования

1962 Анализ близостей Шепарда

1964 Монотонная регрессия Краскела

1968 Метод ранговых образов Гуттмана

1973 Попарное неметрическое шкалирование Джонсона

1974 Многомерное шкалирование Кумса и Холмана

1980 Метод последовательных интервалов Шрайбера

Терехина А.Ю. Анализ данных методами многомерного шкалирования. – М.: Наука, 1986.

Еще Метод неметрического развертывания Каменского

Метрическое шкалирование

метрическое пространство W

условие (1) $w_{ij} \geq 0, \quad w_{ii} = 0$

условие (2) $w_{ij} = w_{ji}$

условие (3) $w_{ik} + w_{kj} \geq w_{ij}$

def: $C = max\{w_{kj} - w_{ki} - w_{ij} \mid \forall\, i, j, k\}$

$W \rightarrow W'$: $w'_{ii} = w_{ii}$ & $w'_{ij} = w_{ij} + C \quad (i \neq j)$

Лемма. **Если W удовлетворяет (1) и (2),**

то W' удовлетворяет (1) – (3).

1/3 ≈Задача метрического шкалирования методом простой ординации Орлочи

объекты $B = \{ B_1 \dots B_n \}$;; матрица расстояний W	w_{ij} – расстояние от B_i до B_j

Метод простой ординации Орлочи*), 1966

Последовательность таблиц $T_1 \dots T_{Ko}$ и их оценок**).	$T_k = \{(b_{i1} \dots b_{ik}) \mid B_i \in B\}$;;; $T_k \in R^k \quad (k = 1 \dots Ko)$

---

*) первая ось – по max w_{ij}, вторая ось – из проекций и т.д.;
О применимости: для точной W, таблица T_{Ko} порождает W.

**) для последующего выбора наилучшего решения

Простая ординация Орлочи (W – точная матрица расстояний)

W

0	500	300	400
500	0	400	300
300	400	0	500
400	300	500	0

W_1

0	0	240	240
0	0	240	240
240	240	0	480
240	240	480	0

W_2

0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Первая ось: 1–2
Начало коо: 1

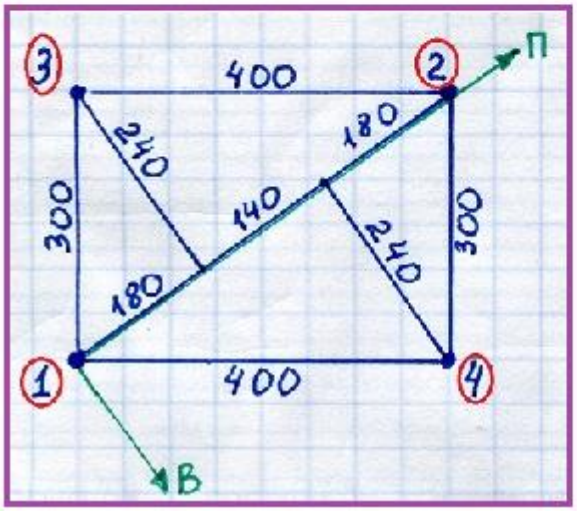
T	
1.	
2.	
3.	
4.	

Вторая ось: 1– 4

T_1	П
1.	0
2.	500
3.	180
4.	320

Стоп

T_2	П	В
1.	0	0
2.	500	0
3.	180	-240
4.	320	240



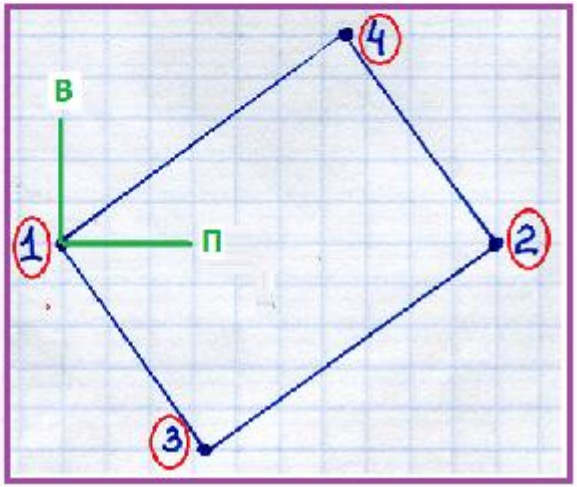
Оценка(T_1) =

$1 - (4*240^2 + 480^2) / (2*300^2 + 2*400^2 + 2*500^2)$

$= 0.54$ (54%)

Оценка(T_2) =

$= 1.00$ (100%)



Простая ординация Орлочи (W – проблемная матрица)

W

0	500	300	400
500	0	300	300
300	300	0	500
400	300	500	0

W_1

0	0	166	232
0	0	166	232
166	166	0	495
232	232	495	0

Объекты 2, 3 и 4 не образуют треугольник

Первая ось: 1–2
Начало коо: 1

Стоп

T
1.
2.
3.
4.

T_1	П
1.	0
2.	500
3.	250
4.	320

Формальная оценка (T_1) = 0.56 (56%)

2/3 Задача метрического шкалирования методом главных компонент Тостерсена

объекты $B = \{ B_1 \dots B_n \}$;; матрица расстояний W	w_{ij} – расстояние от B_i до B_j

Метод главных компонент Тостерсена*), 1952

Последовательность таблиц $T_1 \dots T_{Ko}$ и их оценок**).	$T_k = \{(b_{i1} \dots b_{ik}) \mid B_i \in B\} \in R^k$ (k = 1 … Ko) ;;; Стресс-функционал $\sum_{i,j}(d_{ij} - w_{ij})^2 \to \min$, где $d^2_{ij} = (b_{i1}-b_{j1})^2 + \dots + (b_{ik}-b_{jk})^2$

---

*) ≈ метод главных компонент;

О применимости: для точной W, таблица T_{Ko} порождает W.

**) для последующего выбора наилучшего решения

3/3 Задача метрического шкалирования нелинейными методами

объекты $B = \{ B_1 \dots B_n \}$; ;	w_{ij} – расстояние от B_i до B_j ; ; ; ;
матрица расстояний W ; ; ; ;	$k < n$; ; ; ; ; ;
числа k, γ	*)

Среднеквадратический алгоритм

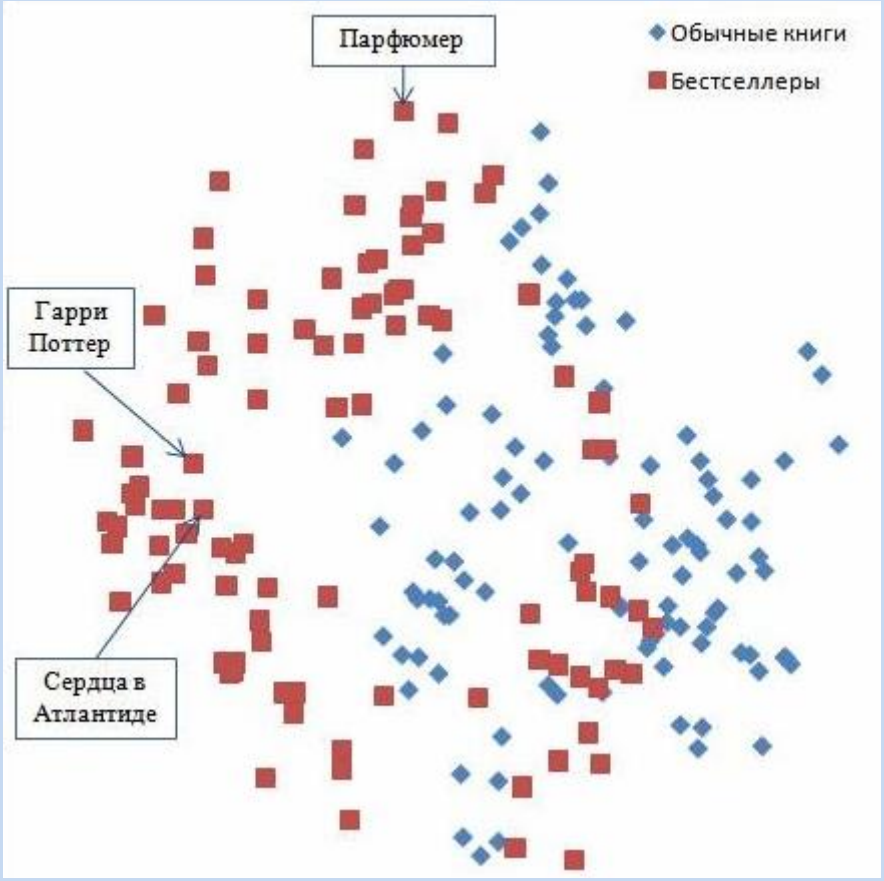
Векторы $(b_{i1} \dots b_{ik}) \in R^k$ для B_i	Стресс-функционал $\sum_{i,j} d^{\gamma}_{ij} (d_{ij} - w_{ij})^2 \to \min$, где $d^2_{ij} = (b_{i1}-b_{j1})^2 + \dots + (b_{ik}-b_{jk})^2$

*) Если **γ < 0**, то приоритет – более точному приближению малых расстояний. Если **γ > 0**, то приоритет – более точному приближению больших расстояний.

Задача метрического шкалирования

карта сходства – результат МШ при k = 2

Феномен Гарри Поттера в похожести на всех и наличии внутренней эволюции [Н.Бабич]



Результат МШ 200 книг; учитывались стиль, композиция, тема.

Поиск ассоциативных правил

аффинитивный анализ; affinity analysis

Поиск ассоциативных правил – раздел анализа данных, ориентированный на выявление закономерностей в заданной базе данных.

База данных = База транзакций

Универсальное множество объектов I

Транзакция (в ПАП) $T \subseteq I$

База транзакций $D = \{T_1, \ldots, T_m\}$

Виды баз:

№ транзакции	Наименование товара	• • •

1 ✓

Последовательность

$T_1 \leq \ldots \leq T_m$

№ транзакции	Наименование товара	Дата	• • •

2

№ транзакции	Наименование товара	Дата	ФИО	• • •

3

Параметр $\mathbf{Supp_{min}}$

Пусть $\mathbf{F \subseteq I}$.

Def: **Supp(F)** = $\left(\textit{к-во}\ T_i \in D : F \subseteq T_i\right)$ **/** m – поддержка **F**

Def: **F** – частый набор, если Supp(F) > $\mathbf{Supp_{min}}$.

Def: **L = L(D)** – множество частых наборов

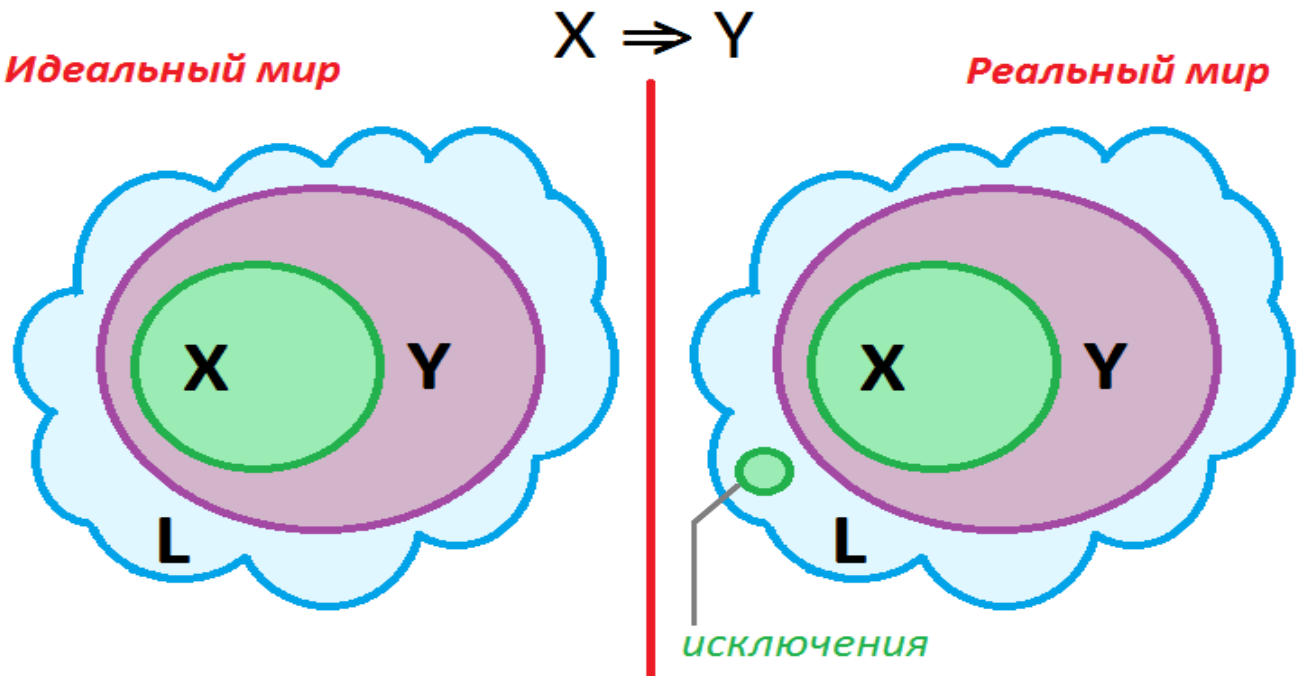
L – строительный материал
для построения ассоциативных правил.

Ассоциативные правила

Пусть $X,Y \in L(D)$, $X \cap Y = \varnothing$

Def: $X \Rightarrow Y$ – ассоциативное правило т. и т. т.

если из $X \subseteq T_i \in D$ следует $Y \subseteq T_i$.



Определение?

Оценки ассоциативных правил

Def: поддержка $\mathbf{Supp_{X \Rightarrow Y} = Supp(X \cup Y)}$

Def: достоверность $\mathbf{Conf_{X \Rightarrow Y} = Supp_{X \Rightarrow Y} / Supp(X)}$

Def: лифт $\mathbf{Lift_{X \Rightarrow Y} = Conf_{X \Rightarrow Y} / Supp(Y)}$

В идеальном случае $\mathbf{Conf_{X \Rightarrow Y}=1}$*, в реальности* $\mathbf{Conf_{X \Rightarrow Y} \leq 1}$.

Если $\mathbf{Lift_{X \Rightarrow Y} > 1}$*, то правило* $\mathbf{X \Rightarrow Y}$ *лучше случайного выбора.*

Альтернативное определение [≈Воронцов]

Для базы транзакций **D** и чисел $\mathbf{Supp_{min}}$ и $\mathbf{Conf_{min}}$ **N.B.**

ассоциативным правилом $\mathbf{X \Rightarrow Y}$ называется

пара $\mathbf{X,Y \in L(D)}$ такая, что

$$\mathbf{Supp_{X \Rightarrow Y} \geq Supp_{min}} \quad \& \quad \mathbf{Conf_{X \Rightarrow Y} \geq Conf_{min}}.$$

Задача поиска ассоциативных правил

База транзакций D ;;;;	D – неупорядоченное множество транзакций
Параметр $Supp_{min}$;;;;	
Параметр $Conf_{min}$;;;;	

Метод поиска ассоциативных правил*)

Множество правил вида $X \Rightarrow Y$	$Supp_{X \Rightarrow Y} \geq Supp_{min}$ & $Conf_{X \Rightarrow Y} \geq Conf_{min}$

*) **AIS • SETM • Apriori • DHP • PARTITION • DIC**

Apriori: (1) D, $Supp_{min}$ → L(D),
(2) L(D), $Conf_{min}$ → Правила

Алгоритм Apriori (пример)

№ тр	Тр
1	A,C,D
2	B,C,E
3	A,B,C,E
4	B,E

F1	Supp
{A}	0.5
{B}	0.75
{C}	0.75
{D}	0.25
{E}	0.75

L1	Supp
{A}	0.5
{B}	0.75
{C}	0.75
{E}	0.75

F2	Supp
{A,B}	0.25
{A,C}	0.5
{A,E}	0.25
{B,C}	0.5
{B,E}	0.75
{C,E}	0.5

L2	Supp
{A,C}	0.5
{B,C}	0.5
{B,E}	0.75
{C,E}	0.5

F3	Supp
{B,C,E}	0.5

L3	Supp
{B,C,E}	0.5

$Supp_{min}=0.5$

$Conf_{min}=0.8$

L = L2 ∪ L3	Supp
{A,C}	0.5
{B,C}	0.5
{B,E}	0.75
{C,E}	0.5
{B,C,E}	0.5

Правило	Conf
A⇒C	1.00
C⇒A	0.67
B⇒C	0.67
C⇒B	0.67
B⇒E	1.00
E⇒B	1.00
C⇒E	0.67
E⇒C	0.67

+

Правило	Conf
B,C⇒E	1.00
E⇒B,C	0.67
B,E⇒C	0.67
C⇒B,E	0.67
C,E⇒B	1.00
B⇒C,E	0.67

Правило	Conf
A⇒C	1.00
B⇒E	1.00
E⇒B	1.00
B,C⇒E	1.00
C,E⇒B	1.00

B и E эквивалентны

тривиальны

Самплинг

для построения множества частых наборов **L**

(1) Выделить часть $\mathbf{D' \subseteq D}$.

(2) Построить $\mathbf{L = L(D'; Supp_{min}-\delta)}$, для нек. $\mathbf{\delta > 0}$.

(3) Вычислить реальные значения наборов из **L** на полной базе **D**.

О количестве ассоциативных правил

33 ≈транзакции

197 гипотез

(≈ассоциативных правил)

Гаек П., Гавранек Т.
Автоматическое образование гипотез:
Математические основы теории.
– М.: Наука,1984(1978)

➔ **Задача удобного представления результатов поиска АП**

Представление ассоциативных правил

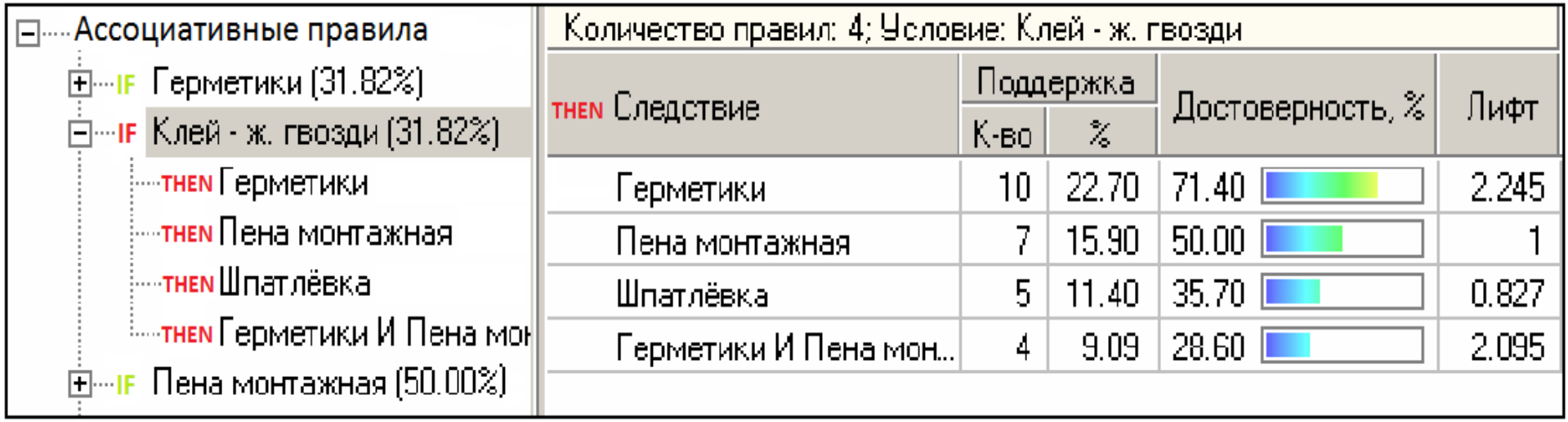
Таблицы правил

Правил: 63

Номер правила	Условие	Следствие	Поддержка %	Достоверность	Лифт
60	Клей - ж. гвозди	Герметики	4.55	40.00	2.933
	Шпатлёвка	Пена монтажная			
57	Герметики	Клей - ж. гвозди	4.55	33.33	2.933
	Пена монтажная	Шпатлёвка			
59	Герметики	Клей - ж. гвозди	4.55	40.00	2.514
	Шпатлёвка	Пена монтажная			

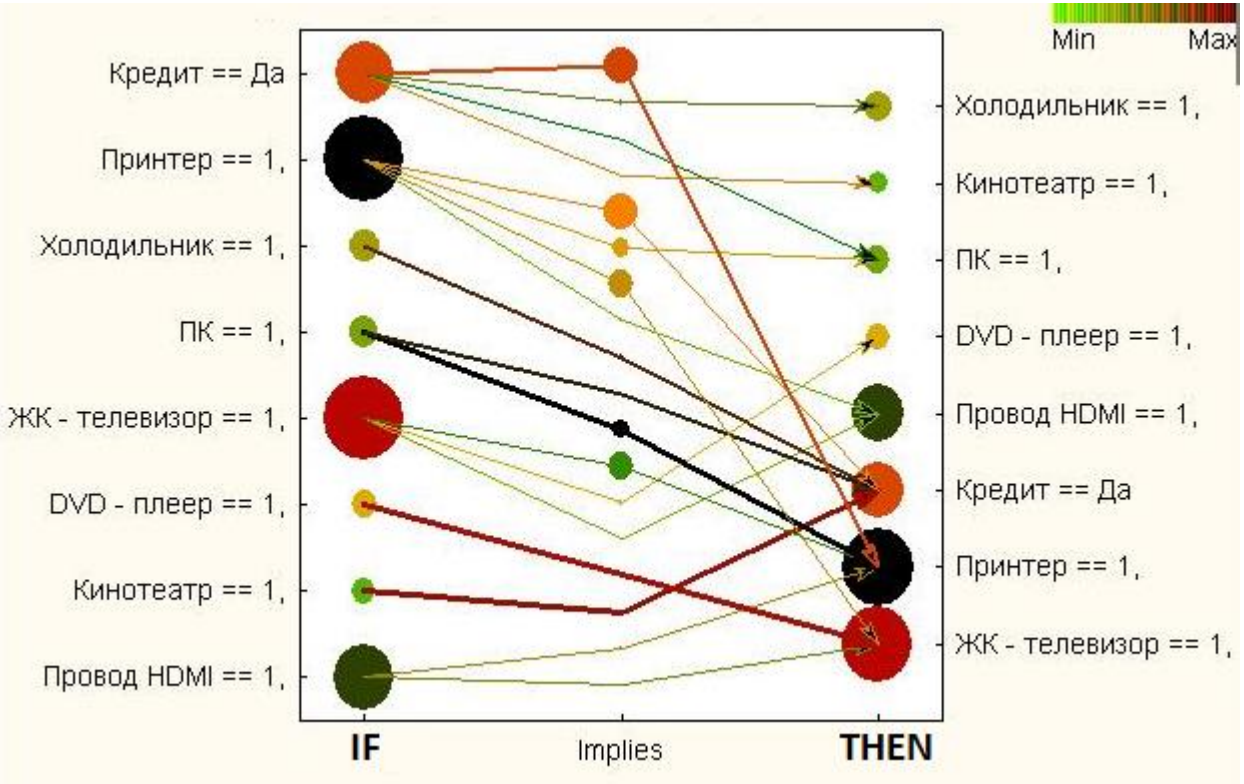
Представление ассоциативных правил

Дерево правил



Представление ассоциативных правил

Сеть правил



Линия, соединяющая круг из причины (IF) с кругом из следствия (THEN), означает одно ассоциативное правило. Чем толще линия и темнее соединения, тем выше достоверность правила. Чем больше и темнее размер круга, тем выше уровень поддержки. Размер круга, соответствующего причине (IF) или следствию (THEN), означает частоту встречаемости причины или следствия. Величина совместной поддержки отображается через размер и цвет круга посредине (Implies). [statsoft.ru]

Задачи поиска ассоциативных правил

Задача поиска позитивных АП

Задача поиска негативных АП

Задача поиска обобщенных АП

Задача поиска численных АП

Задача поиска временнЫх АП

positive

negative

generalized

quantitative

temporal

Задача поиска негативных АП

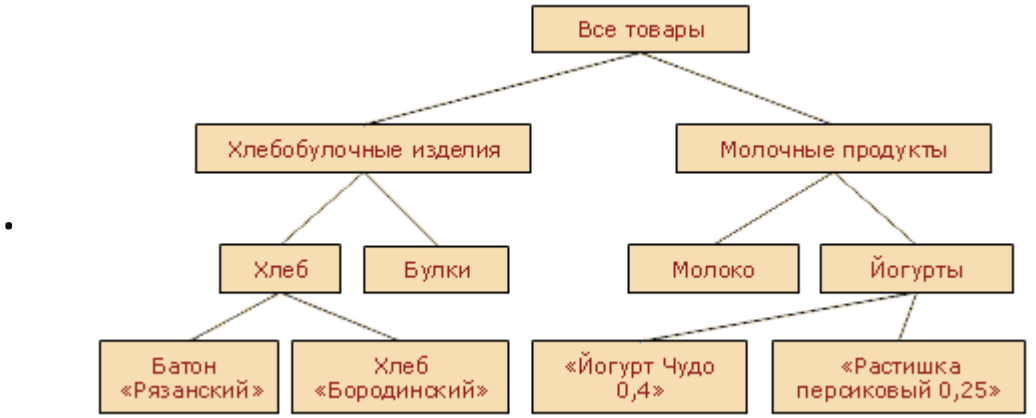
(+) Параметры выбора негативных правил	**(+)** В АП допускаются отрицания.

Метод

Множество АП	**(+)** критерий выбора негативных правил Пятецкого– Шапиро

Задача поиска обобщенных АП

(+) Иерархия объектов



Метод

Множество АП

(+) ограничение на обобщенные правила

Задача поиска численных АП

(+) числовые показатели в D	**(+)** псевдо-объекты вида $X \equiv (\Pi \in [\Pi_{min}, \Pi_{max}])$ ➔ Подзадача дискретизации диапазонов изменения переменных.

Метод

Множество АП	

Задача поиска временнЫх АП

Секвенциальный (последовательный) анализ

(+) временнЫе отметки транзакций	❷ **(+)** Два класса правил: 1. В диапазоне $\mathbf{Time_Z}$ имеет место АП **Z** 2. В диапазоне $\mathbf{Time_Z}$ с периодом $\mathbf{Period_Z}$ имеет место АП **Z**

Метод: AprioriAll и др.

Множество АП	

Data Mining --

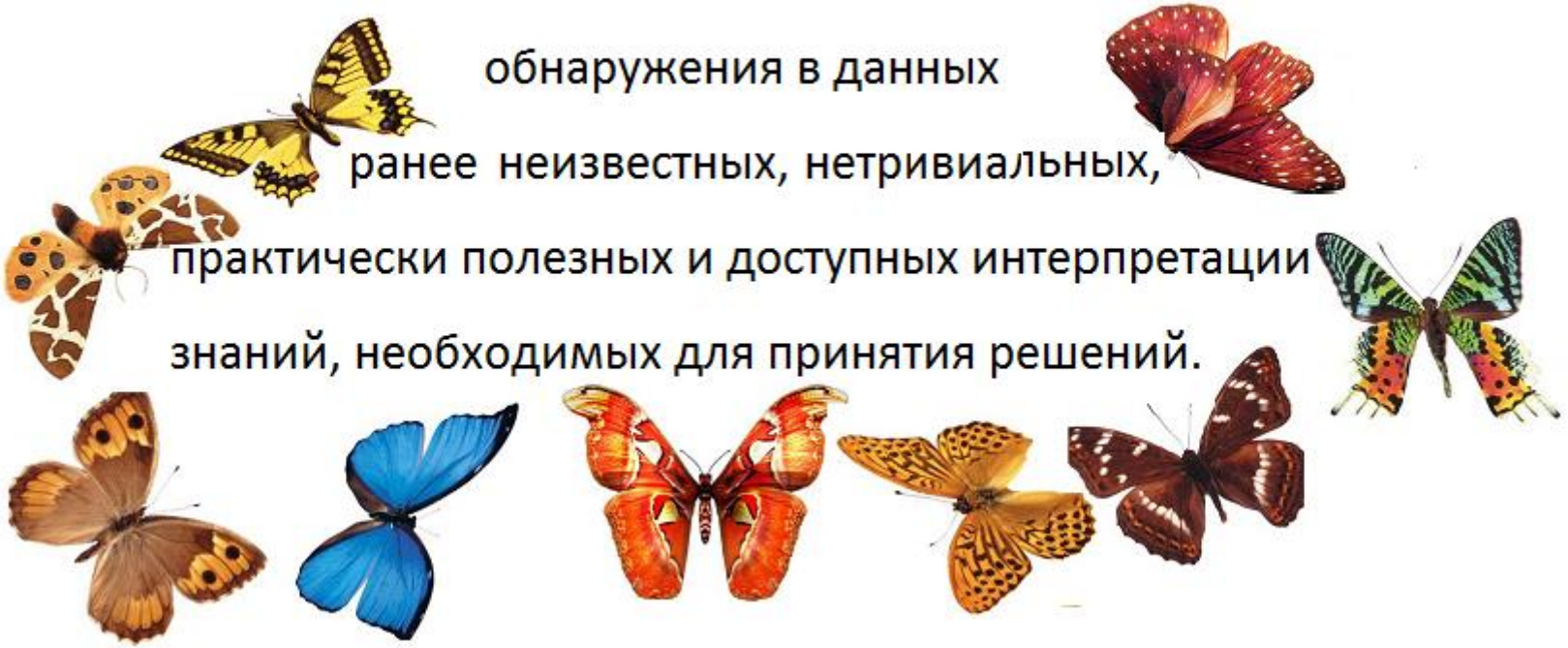
совокупность методов

обнаружения в данных

ранее неизвестных, нетривиальных,

практически полезных и доступных интерпретации

знаний, необходимых для принятия решений.



В о п р о с ы?

soloviev@glossary.ru
