ПОЛИТЕХ
Санкт-Петербургский политехнический университет Петра Великого

Поиск ассоциативных правил

Тушканова Ольга Николаевна

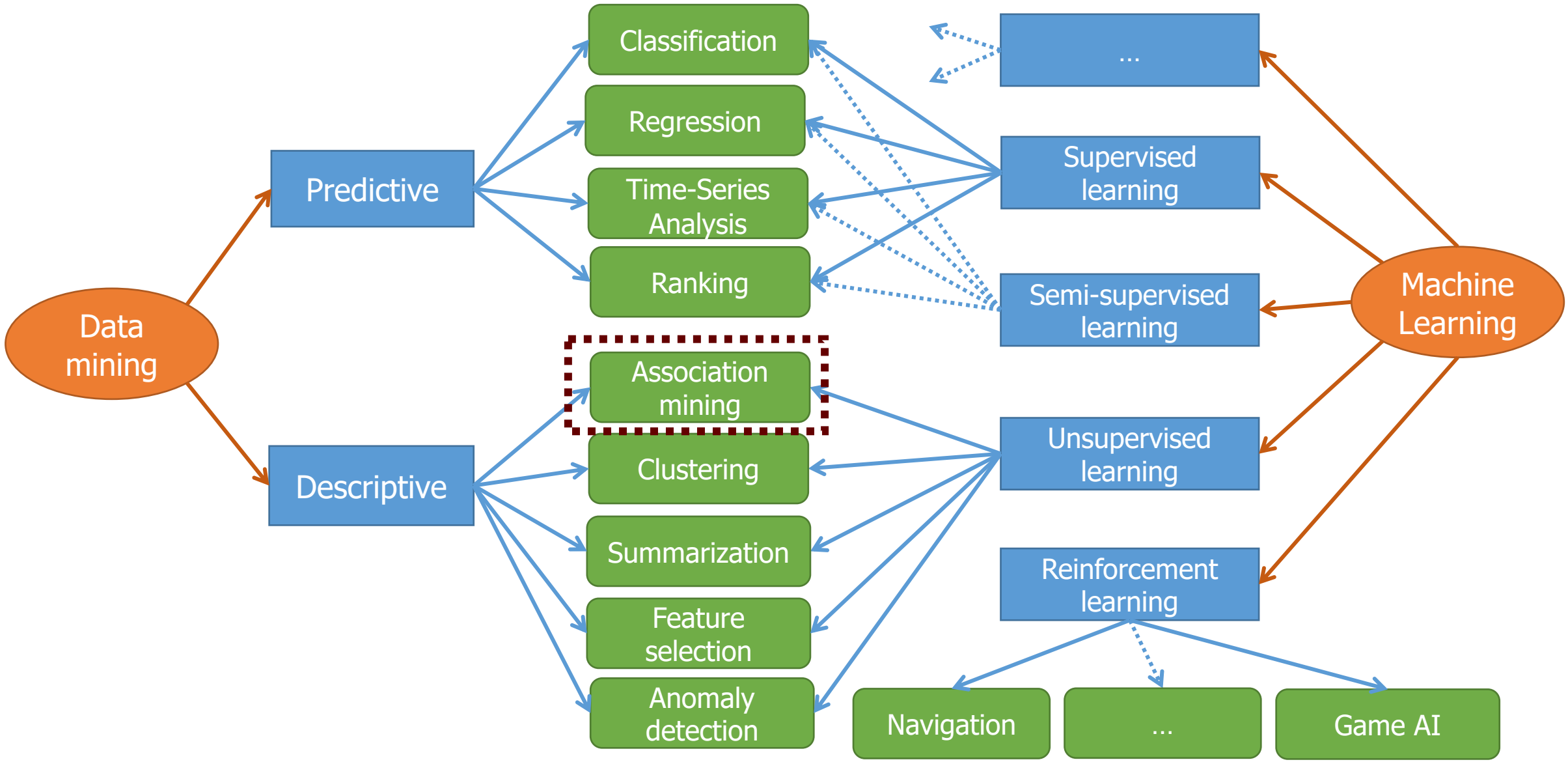
Санкт-Петербург

2021

Российская Академия Наук

Содержание

Основные задачи DM и ML



Market basket analysis – первое приложение Data Mining

IF		THEN	Confidence
People who lives at Distrito Federal	→	Buy preparation courses for national exams	83,8%
People who are Graduated	→	Buy preparation courses for national exams	72,04%
People who are male, lives at Minas Gerais and buys preparation courses for national exams.	→	Are Graduated.	73,6%

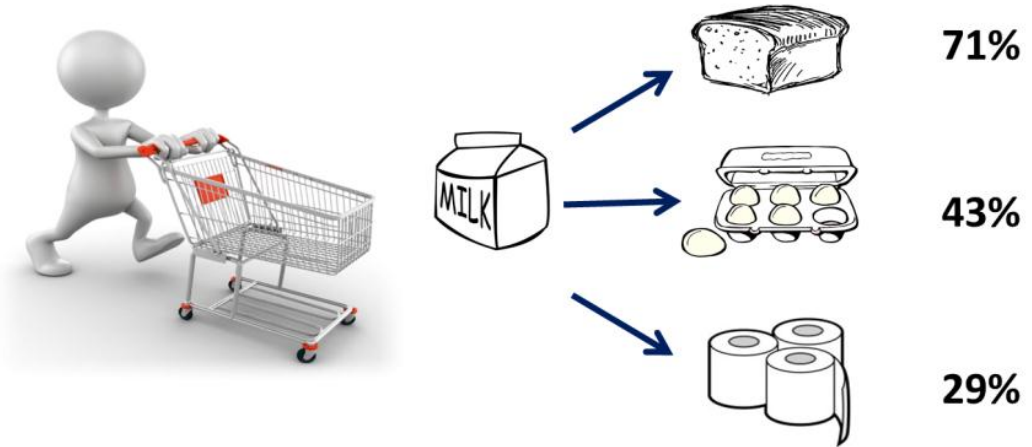
Ассоциативное правило

Цель поиска ассоциативных правил – нахождение закономерностей между связанными событиями.

Ассоциативное правило отражает связь между значениями атрибутов в данных.

посылка $\longrightarrow X \rightarrow Y \longleftarrow$ следствие

Если произошло событие ***X***, то с **некоторой вероятностью** произойдет и событие ***Y.***



Поддержка правила $\mathbf{X} \to \mathbf{Y}$

TID	Товары	Символы
1	Хлеб, молоко, печенье	a, b, c
2	Молоко, сметана	b, d
3	Хлеб, молоко, печенье, сметана	a, b, c, d
4	Сметана, колбаса	d, e
5	Хлеб, молоко, печенье сметана	a, b, c, d
6	Конфеты	f

{Хлеб, молоко} → {Печенье}

{a, b} → {c}

n = 6 n_{XY} = 3

supp (X →Y) = 3 / 6 = 0.5

Пусть $\mathbf{E}$ – транзакционное множество данных,

$\mathbf{D}$ – множество символов, которые используются для обозначения объектов ({a, b, c, d, e, f}),

$\mathbf{X}, \mathbf{Y}$ – подмножества символов из множества $\mathbf{D}$, будем называть такие подмножества **паттернами**

$\Lambda(\mathbf{X} + \mathbf{Y})$ – подмножество множества транзакций $\mathbf{E}$, которые содержат паттерны $\mathbf{X}$ и $\mathbf{Y}$.

Тогда отношение

$$supp(X \to Y) = \frac{|\Lambda(X + Y)|}{|E|} = \frac{n_{XY}}{n}$$

называют поддержкой (support) правила $\mathbf{X} \to \mathbf{Y}$ на множестве $\mathbf{E}$,
| | - мощность некоторого множества (кол-во элементов в нем).

Величина поддержки соответствует оценке вероятности совместного появления паттернов $\mathbf{X}$ и $\mathbf{Y}$ в множестве $\mathbf{E}$:

$$p_{XY} = supp(X \to Y) = \frac{n_{XY}}{n}$$

Поддержка (support) – $supp(X \to Y)$

Уверенность (confidence) – $conf(X \to Y)$

Уверенность правила $\mathbf{X} \to \mathbf{Y}$

TID	Товары	Символы
1	Хлеб, молоко, печенье	a, b, c
2	Молоко, сметана	b, d
3	Хлеб, молоко, печенье, сметана	a, b, c, d
4	Сметана, колбаса	d, e
5	Хлеб, молоко, печенье сметана	a, b, c, d
6	Конфеты	f

{Хлеб, молоко} → {Печенье}

{a, b} → {c}

$n_{XY} = 3$ $n_X = 3$

conf (X →Y) = 3 / 3 = 1

Пусть $\mathbf{E}$ – транзакционное множество данных,

$\mathbf{D}$ – множество символов, которые используются для обозначения объектов ({a, b, c, d, e, f}),

$\mathbf{X}, \mathbf{Y}$ – подмножества символов из множества $\mathbf{D}$, будем называть такие подмножества **паттернами**

$\Lambda(\mathbf{X} + \mathbf{Y})$ – подмножество множества транзакций $\mathbf{E}$, которые содержат паттерны $\mathbf{X}$ и $\mathbf{Y}$.

Тогда отношение

$$conf\ (X \to Y) = \frac{|\Lambda(X + Y)|}{|\Lambda(X)|} = \frac{n_{XY}}{n_X}$$

называют уверенностью (confidence) правила $\mathbf{X} \to \mathbf{Y}$ на множестве $\mathbf{E}$.

Величина уверенности соответствует оценке условной вероятности появления паттерна $\mathbf{Y}$ в некотором другом паттерне $\mathbf{Z}$ в множестве $\mathbf{E}$ при условии, что в паттерне $\mathbf{Z}$ также присутствует паттерн $\mathbf{X}$:

$$p_{Y|X} = conf(X \to Y) = \frac{n_{XY}}{n_X}$$

Модель поддержка-уверенность

Говорят, что выражение вида $\mathbf{X} \to \mathbf{Y}$ – это ассоциативное правило с порогом поддержки σ и порогом уверенности γ (σ, γ – ассоциативное правило), если

$$supp(X \to Y) \geq \sigma \text{ и } conf(X \to Y) \geq \gamma$$

Задача поиска ассоциативных правил:

Для заданных значений порогов поддержки σ и уверенности γ и для заданной базы обучающих данных $\mathbf{E}$ найти все пары паттернов $\mathbf{X}$ и $\mathbf{Y}$, для которых отношение $\mathbf{X} \to \mathbf{Y}$ является σ, γ - ассоциативным правилом.

Другие меры оценки

$PS(X \to Y) = p(Y|\mathrm{X}) - \mathrm{P}(\mathrm{Y})$ – интерес или мера Г. Пятецкого-Шапиро

$V(X \to Y) = \frac{p(\mathrm{X}) \cdot \mathrm{p}(\bar{Y})}{p(\mathrm{X}\bar{Y})}$ – убежденность (англ. conviction)

$lift(X \to Y) = \frac{p(Y|\mathrm{X})}{P(XY)}$ – лифт (англ. lift)

$F(X \to Y) = \frac{p(Y|\mathrm{X}) - p(Y)}{1 - p(Y)}$ – фактор определенности (англ. certainty factor)

N	Measure	Formula	Range
1	ϕ-coefficient	$\frac{p_{AB} - p_A p_B}{\sqrt{p_A p_B (1-p_A)(1-p_B)}}$	[-1; 1]
2	Odds Ratio (α)	$\left(p_{AB} p_{\overline{A}\overline{B}}\right) / \left(p_{A\overline{B}} p_{\overline{A}B}\right)$	[0; ∞]
3	Yule's Q	$\frac{p_{AB} p_{\overline{A}\overline{B}} - p_{A\overline{B}} p_{\overline{A}B}}{p_{AB} p_{\overline{A}\overline{B}} + p_{A\overline{B}} p_{\overline{A}B}}$	[-1; 1]
4	Yule's Y	$\frac{\sqrt{p_{AB} p_{\overline{A}\overline{B}}} - \sqrt{p_{A\overline{B}} p_{\overline{A}B}}}{\sqrt{p_{AB} p_{\overline{A}\overline{B}}} + \sqrt{p_{A\overline{B}} p_{\overline{A}B}}}$	[-1; 1]
5	Kappa (κ)	$\frac{p_{AB} + p_{\overline{A}\overline{B}} - p_A p_B - p_{\overline{A}} p_{\overline{B}}}{1 - p_A p_B - p_{\overline{A}} p_{\overline{B}}}$	[-1; 1]
6	J-measure (J)	$p_{AB} \log(p_{B\|A} / p_B) + p_{A\overline{B}} \log(p_{\overline{B}\|A} / p_{\overline{B}})$	[0; 1]
7	Gini Index (G)	$p_A \left(p^2_{B\|A} + p^2_{\overline{B}\|A}\right) + p_{\overline{A}} \left(p^2_{B\|\overline{A}} + p^2_{\overline{B}\|\overline{A}}\right) - p_B^2 - p_{\overline{B}}^2$	[0; 1]
8	Support (sup)	p_{AB}	[0; 1]
9	Confidence (conf)	$p_{B\|A}$	[0; 1]
10	Laplace (L)	$(np_{AB} + 1) / (np_B + 2)$	[0; 1]
11	Conviction (V)	$p_A p_{\overline{B}} / p_{A\overline{B}}$	[0.5; ∞]
12	Interest (I)	$p_{AB} / p_A p_B$	[0; ∞]

N	Measure	Formula	Range
13	Cosine (IS)	$p_{AB}/\sqrt{p_A p_B}$	[0; 1]
14	Piatetsky-Shapiro's (PS)	$p_{AB} - p_A p_B$	[-0.25; 0.25]
15	Certainty Factor (F)	$(p_{B\|A} - p_B)/(1 - p_B)$	[-1; 1]
16	Added Value (AV)	$p_{B\|A} - p_B$	[-0.5; 1]
17	Collective Strength (S)	$\frac{p_{AB} + p_{\overline{AB}}}{p_A p_B + p_{\overline{A}} \cdot p_{\overline{B}}} \times \frac{1 - p_A p_B - p_{\overline{A}} p_{\overline{B}}}{1 - p_{AB} - p_{\overline{AB}}}$	[0; ∞]
18	Jaccard (ζ)	$p_{AB}/(p_A + p_B - p_A p_B)$	[0; 1]
19	Klosgen (K)	$\sqrt{p_{AB}} \cdot p_{B\|A} - p_B$	$[\sqrt{(2/\sqrt{3}-1)} \cdot 2 - \sqrt{3} - 1/\sqrt{3}, 2/3\sqrt{3}]$
20	Information Gain (IG)	$\log(p_{AB}/(p_A \cdot p_B))$	$[-\infty; \log(1/p_A)]$
21	Sebag and Schoenauer's (SEB)	$p_{AB}/p_{A\overline{B}}$	[0; ∞]
22	Regression coefficient	$(p_{AB} - p_A p_B)/(p_A \cdot (1 - p_A))$	[-1, 1]

TID	Товары	Символы
1	Хлеб, молоко, печенье	a, b, c
2	Молоко, сметана	b, d
3	Хлеб, молоко, печенье, сметана	a, b, c, d
4	Сметана, колбаса	d, e
5	Хлеб, молоко, печенье сметана	a, b, c, d
6	Конфеты	f

Алгоритмы поиска ассоциативных правил

- **Алгоритм AIS.** Первый алгоритм поиска ассоциативных правил, был разработан сотрудниками исследовательского центра IBM Almaden (Agrawal, Imielinski and Swami) в 1993 году.
- **Алгоритм Apriori.** Предложен Agrawal и Srikant в 1994 году. Основан на принципе поиска в ширину и антимонотонности значения поддержки.
- **Алгоритм FP-growth** (FP - frequent pattern). Базируется на построении дерева паттернов возрастающей длины (FP-tree) и на последующем извлечении из него паттернов, для которых значение функции поддержки не меньше заданного порога.
- **Алгоритм Eclat** (Equivalence Class Transformation). Предложен Zaki, Parthasarathy, Li и Ogihara в 1997 году. Основан на принципе поиска в глубину.

Схема поиска правил

E – транзакционное множество данных

D – множество всех возможных символов из множества **E**,

$\mathbf{X_i}$, $\mathbf{Y_j}$ – подмножества символов (паттерны) из множества **D**

σ, γ – параметры алгоритма

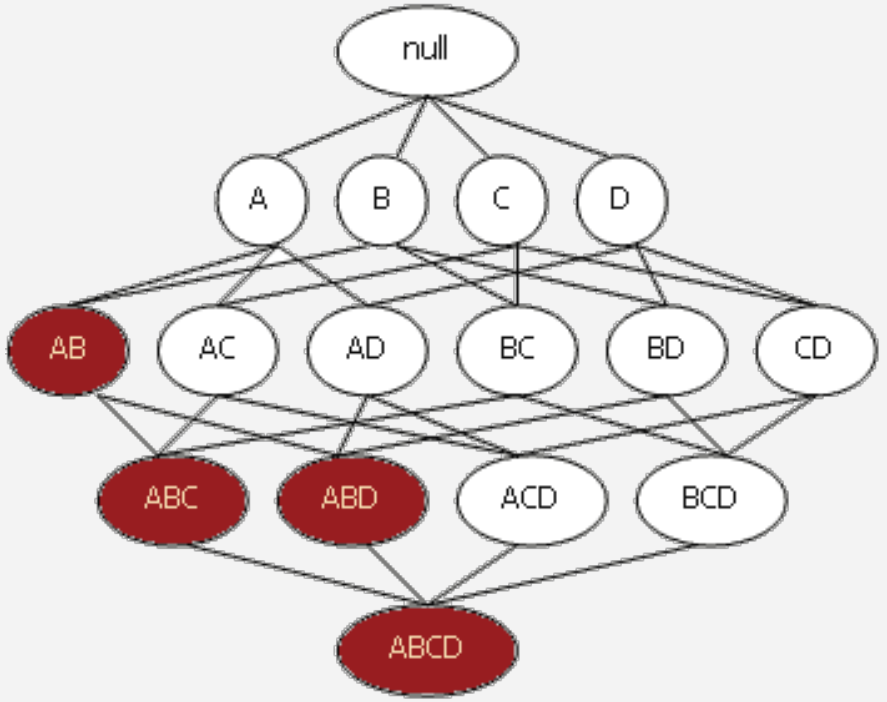
1. Найти все паттерны $\mathbf{D_k} \in \mathbf{D}$ с поддержкой не менее чем σ, т.е. те, для которых $supp(D_k) \geq \sigma$. Такие паттерны будем называть **часто встречающимися паттернами**.
2. Среди всех часто встречающихся паттернов $\mathbf{D_k} \in \mathbf{D}$, найденных на шаге 1, сгенерировать ассоциативные правила $\mathbf{X_i} \rightarrow \mathbf{Y_j}$, такие, что:
   1) $\mathbf{X_i + Y_j = D_k}$, т. е. часто встречающийся паттерн $\mathbf{D_k}$ состоит из символов, входящих в множества $\mathbf{X_i}$ и $\mathbf{Y_j}$
   2) $conf(\mathbf{X_i} \rightarrow \mathbf{Y_j}) \geq \gamma$

Основная идея алгоритма Apriori

Антимонотонность функции sup:

Для двух паттернов $D_k \in \mathbf{D}$ и $D_r \in \mathbf{D}$, таких, что $D_k \subset D_r$, справедливо

$$supp(D_k) \geq supp(D_r)$$



Паттерн D_k обязательно входит во все те транзакции, в которые входит паттерн D_r

Паттерн D_k может входить также и в другие транзакции

Если $supp(D_k) < \sigma$, то $supp(D_k + ?) < \sigma$

Паттерны $D_k + ?$ можно исключить из генерации ассоциативных правил

Пример работы

TID	Товары	Символы
1	Хлеб, молоко, печенье	a, b, c
2	Молоко, сметана	b, d
3	Хлеб, молоко, печенье, сметана	a, b, c, d
4	Сметана, колбаса	d, e
5	Хлеб, молоко, печенье, сметана	a, b, c, d
6	Конфеты	f

$\sigma = 0.5 \quad n = 6$

1-элементные паттерны

Паттерн	n_i
a	3
b	4
c	3
d	4
e	1
f	1

$\sigma_i \geq 0.5$

1-элементные ЧВП

Паттерн	n_i
a	3
b	4
c	3
d	4

2-элементные паттерны

Паттерн	n_i
ab	3
ac	3
ad	2
bc	3
bd	3
cd	2

$\sigma_i \geq 0.5$

2-элементные ЧВП

Паттерн	n_i
ab	3
ac	3
bc	3
bd	3

3-элементные паттерны

Паттерн	n_i
abc	3

$\sigma_i \geq 0.5$

3-элементные ЧВП

Паттерн	n_i
abc	3

Паттерн	n_i	conf
a → b	3	1
b → a	3	0.75
a → c	3	1
c → a	3	1
b → c	3	0.75
c → b	3	1
b → d	3	0.75
d → b	3	0.75
a, b → c	3	1
a, c → b	3	1
b, c → a	3	1
c → a, b	3	1
b → a, c	3	0.75
a → b, c	3	1

Алгоритм Apriori: недостатки

- Процесс генерации кандидатов в часто встречающиеся паттерны - **узкое место** в алгоритме Apriori
- Если база транзакций содержит 100 предметов, то потребуется сгенерировать $\mathbf{2^{100} \sim 10^{30}}$ кандидатов в ЧВП
- Требует **многократного** сканирования базы транзакций (столько раз, сколько элементов содержит самый длинный паттерн)

Алгоритм FP-growth

Алгоритм Frequent Pattern-Growth – “выращивание” часто встречающихся паттернов.

Позволяет избежать затратной процедуры генерации кандидатов и уменьшить необходимое число проходов по данным **до 2**.

В основе метода лежит процедура преобразования БД в компактную древовидную структуру Frequent-Pattern Tree – дерево часто встречающихся паттернов.

При построении FP-дерева используется подход “разделяй и властвуй” (англ. divide and conquer) - декомпозиция одной сложной задачи на множество более простых.

Построение FP-дерева

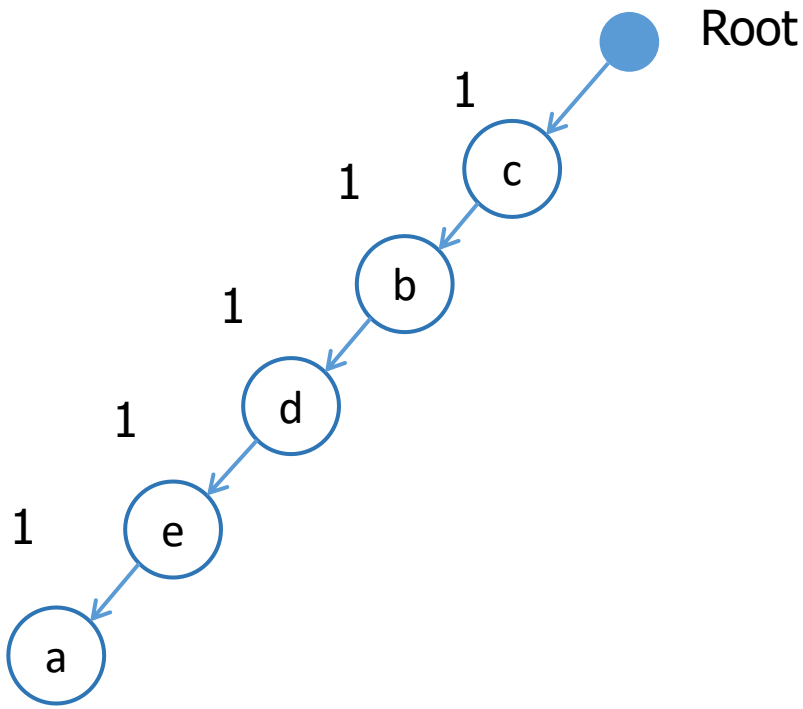
TID	Транзакции
1	a, b, c, d, e
2	a, b, c
3	a, c, d, e
4	b, c, d, e
5	b, c
6	b, d, e
7	c, d, e

σ = 3 / 7
n = 7

Символ	n_i
c	6
b	5
d	5
e	5
a	3

Упорядоченные транзакции

TID	Наборы
1	c, b, d, e, a
2	c, b, a
3	c, d, e, a
4	c, b, d, e
5	c, b
6	b, d, e
7	c, d, e



Построение FP-дерева

TID	Транзакции
1	a, b, c, d, e
2	a, b, c
3	a, c, d, e
4	b, c, d, e
5	b, c
6	b, d, e
7	c, d, e

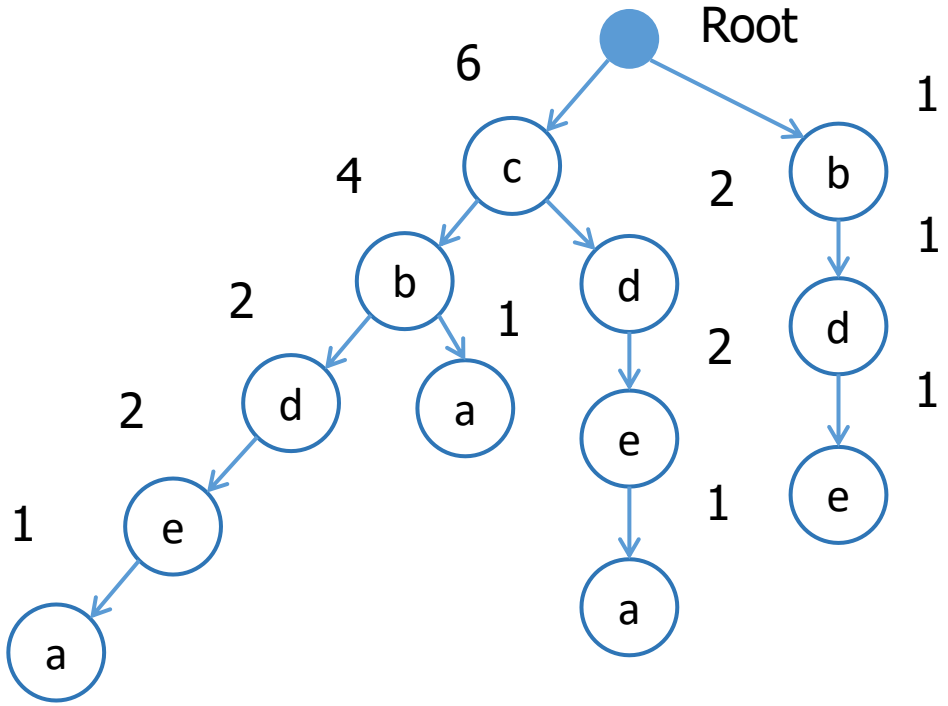
σ = 3 / 7
n = 7

Символ	n_i
c	6
b	5
d	5
e	5
a	3

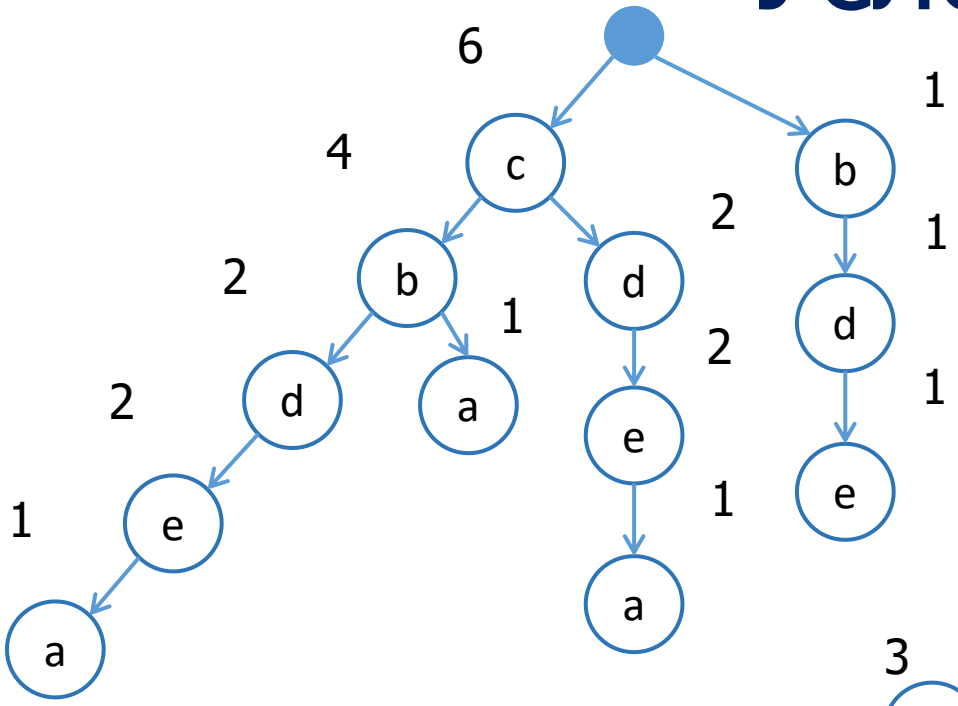
Упорядоченные транзакции

TID	Наборы
1	c, b, d, e, a
2	c, b, a
3	c, d, e, a
4	c, b, d, e
5	c, b
6	b, d, e
7	c, d, e

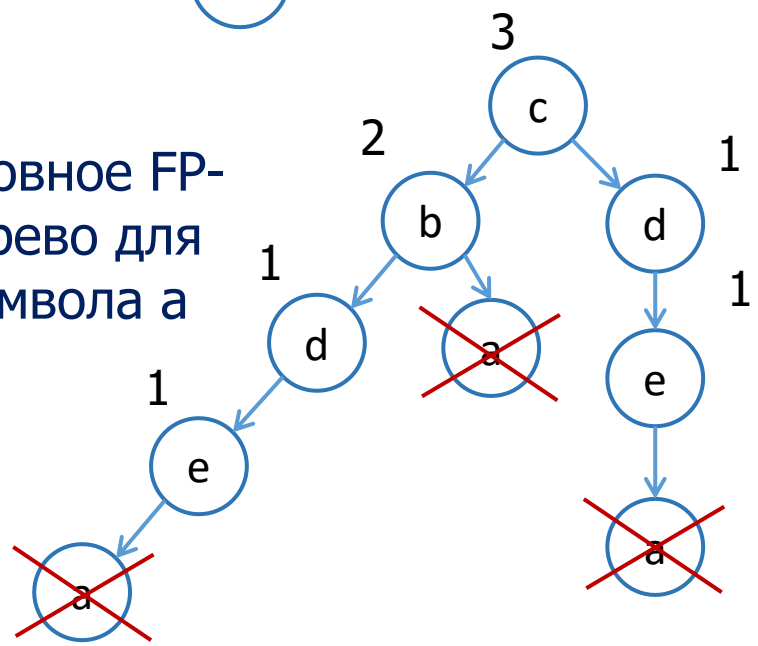
Итоговое FP-дерево



Условные FP-деревья



Условное FP-дерево для символа a



1. Выбираем символ (например, a) и все ветви, ведущие к нему:
   (c b d e a : 1), (c b a : 1), (c d e a : 1)
2. Удалить символ a:
   (c b d e : 1), (c b : 1), (c d e : 1)
3. Построить условное дерево.
4. Число вхождений дополняющих признаков:
   (c: 3), (b: 2), (d: 2), (e: 2)
5. Найти частые наборы:
   (c a : 3)

Выбор частых наборов

	Пути	Частоты	ЧВП
a	(c b d e a), (c b a), (c d e a)	(c : 3), (b : 2), (d : 2), (e : 2)	(c a : 3)
b	(c b)	(c : 4)	(c b : 4)
c	-	-	-
d	(c b d), (c d), (b d)	(c : 4), (b : 3)	(c d : 4), (b d: 3)
e	(c b d e), (c d e), (b d e)	(c, 4), (b, 3), (d, 5)	(c a : 3), (c b : 4), (c d : 4), (b d : 3), (d e : 5), (d c e : 4), (d b e : 3)

Apriori vs FP-growth

1. Проход по исходным данным.
2. Память.
3. Время.

Параллелизация
Apriori: возможно и нужно
FP-growth: только условные деревья

	Apriori	**FP-Growth**
1.	По числу элементов в самой большой транзакции	Два прохода
2.	Хранение транзакций в необработанном виде	Компактно, специальная структура
3.	Генерация кандидатов на частые наборы	Заполнение, проход по дереву (специальной структуре)

От транзакций к плоской таблице

TID	Товары	Символы
1	Хлеб, молоко, печенье	a, b, c
2	Молоко, сметана	b, d
3	Хлеб, молоко, печенье, сметана	a, b, c, d
4	Сметана, колбаса	d, e
5	Хлеб, молоко, печенье, сметана	a, b, c, d
6	Конфеты	f

TID	Хлеб	Молоко	Печенье	Сметана	Колбаса	Конфеты
1	1	1	1	0	0	0
2	0	1	0	1	0	0
3	1	1	1	1	0	0
4	0	0	0	1	1	0
5	1	1	1	1	0	0
6	0	0	0	0	0	1

Не только транзакции

No.	1: outlook	2: temperature	3: humidity	4: windy	5: **play**
	Nominal	Nominal	Nominal	Nominal	Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

			support	confidence
1. outlook = overcast	==>	play = yes	4	100%
2. temperature = cool	==>	humidity = normal	4	100%
3. humidity = normal & windy = false	==>	play = yes	4	100%
4. outlook = sunny & play = no	==>	humidity = high	3	100%
5. outlook = sunny & humidity = high	==>	play = no	3	100%
6. outlook = rainy & play = yes	==>	windy = false	3	100%
7. outlook = rainy & windy = false	==>	play = yes	3	100%
8. temperature = cool & play = yes	==>	humidity = normal	3	100%
9. outlook = sunny & temperature = hot	==>	humidity = high	2	100%
10. temperature = hot & play = no	==>	outlook = sunny	2	100%

Замечания

- Уменьшение минимальной поддержки - увеличение количества потенциально интересных правил, но требует существенных вычислительных ресурсов. Слишком маленькая поддержка правила делает его статистически необоснованным.
- Уменьшение порога уверенности - увеличение количества потенциально интересных правил. Но оправила со слишком низкой уверенностью нельзя учитывать.
- Правила с очень большой поддержкой более ценны, но, скорее всего, либо всем известны, либо элементы, присутствующие в них, – это лидеры продаж (низкая практическая ценность).
- Правила с очень большой уверенностью также практической ценности не имеют, т.к. элементы, входящие в следствие, средний покупатель скорее всего уже купил.

Применение

- Сегментация покупателей по поведению при покупках
- Анализ предпочтений клиентов (рекомендательные системы)
- Планирование расположения товаров в супермаркетах
- Кросс-продажи
- Адресная рассылка
- Медицина
- Анализ посещений веб-страниц
- Анализ данных по переписи населения
- Анализ и прогнозирование сбоев оборудования
- Биоинформатика

Спасибо за внимание!