О поиске ассоциативных правил в небинарных данных

Генрихов Игорь Евгеньевич
Дюкова Елена Всеволодовна

ООО Мобайл парк ИТ
ВЦ имени А. А. Дородницына ФИЦ ИУ РАН

19-я Всероссийская конференция с международным участием
«Математические методы распознавания образов»

Москва, 2019

Задача поиска ассоциативных правил

P – множество, элементы которого называются атрибутами.

D – база данных, содержащая некоторые наборы атрибутов (транзакции).

Ассоциативное правило – пара непересекающихся наборов атрибутов X и Y, которые одновременно содержатся минимум в одной транзакции.

Ассоциативное правило, порождаемое X и Y, обозначается через $X \Rightarrow Y$.

***Поддержкой* (*support*)** правила $X \Rightarrow Y$ называется отношение числа транзакций, содержащих $X \cup Y$, к числу всех транзакций.

***Достоверностью* (*confidence*)** правила $X \Rightarrow Y$ называется отношение числа транзакций, содержащих $X \cup Y$, к числу транзакций, содержащих X.

Требуется найти ассоциативные правила с поддержкой не менее s, $s \in [0,1]$, и с достоверностью не менее c, $c \in [0,1]$.

Транзакции с бинарными данными

Поиск ассоциативных правил обычно осуществляется в два этапа.

I. Находятся все s-частые наборы атрибутов. Набор атрибутов Z называется s-частым, если отношение числа транзакций, содержащих Z, к числу всех транзакций не менее s (в противном случае Z называется s-нечастым).

II. Для каждого найденного s-частого набора Z путем разбиения Z на два непересекающихся подмножества X и Y строятся ассоциативные правила вида $X \Rightarrow Y$ с достоверностью не менее c .

Общий случай: транзакции с небинарными данными

Каждый атрибут имеет некоторое множество числовых значений и вместо наборов атрибутов рассматриваются наборы их значений.

Поиск ассоциативных правил обычно осуществляется в три этапа.

I. Для каждого небинарного атрибута задается некоторый числовой порог, позволяющий перекодировать исходные небинарные данные в бинарные.

II. Находятся все s-частые наборы бинарных атрибутов.

III. Для каждого найденного s-частого набора строятся ассоциативные правила с достоверностью не менее c .

Результат существенно зависит от выбора варианта бинаризации.

Перебор по всем вариантам бинаризации требует больших временных затрат.

Постановка задачи поиска ассоциативных правил в случае частично упорядоченных данных

$P = P_1 \times \cdots \times P_n$, где $P_1, \ldots, P_n$ – конечные частично упорядоченные множества

Элемент $y = (y_1, \ldots, y_n)$ из P ***следует*** за элементом $x = (x_1, \ldots, x_n)$ $(x \preccurlyeq y)$, если y_i следует за x_i при $i = 1, \ldots, n$.

Элементы $x, y \in P$ ***сравнимы***, если либо $x \preccurlyeq y$ (y следует за x), либо $y \preccurlyeq x$ (x следует за y), иначе ***несравнимы***.

$l_i -$ ***наименьший*** элемент в P_i, если $\forall x_i \in P_i$: $l_i \preccurlyeq x_i$ (если наименьший элемент в P_i отсутствует, то P_i дополняется таким элементом)

$x_i -$ ***существенное*** значение в $x = (x_1, \ldots, x_n)$, если $x_i \neq l_i$

Постановка задачи поиска ассоциативных правил в случае частично упорядоченных данных

$S_D(x)$, $x \in P$, $-$ число транзакций z в D таких, что $x \preccurlyeq z$

Элемент $x \in P$ является ***s -частым***, если $S_D(x)/|D| \geq s$.

Пара несравнимых элементов $x = (x_1, \dots, x_n)$ и $y = (y_1, \dots, y_n)$ из P $-$ ***непересекающаяся***, если $\forall i \in \{1, \dots, n\}$ хотя бы один из элементов x_i и y_i равен l_i.

Пусть x, y – непересекающиеся. Через $x \odot y = (u_1, \dots, u_n)$ обозначим элемент из P, в котором $u_i = l_i$, если $x_i = y_i = l_i$, иначе либо $u_i = x_i$, если x_i – существенное значение для x, либо $u_i = y_i$, если y_i – существенное значение для y.

Ассоциативным $(\boldsymbol{s}, \boldsymbol{c})$ ***-правилом***, $s \in [0,1]$, $c \in [0,1]$, называется пара непересекающихся элементов x и y множества P, таких что элемент $x \odot y$ – s *-частый* и $S_D(x \odot y)/S_D(x) \geq c$.

Поиск пороговых ассоциативных правил

Сведём задачу поиска ассоциативных правил в частично упорядоченных данных к бинарному случаю задав для каждого P_i порог перекодировки p_i, $p_i \in P_i$.

В $x = (x_1, \dots, x_n) \in D$ число x_i заменим на 1, если $p_i \preccurlyeq x_i$, иначе заменим на 0.

Для набора порогов $H = (p_1, \dots, p_n)$ вместо P получаем множество $P_H = \{0,1\}^n$, $0 \preccurlyeq 1$, $0 \neq 1$, с базой D_H, состоящей из элементов множества P_H.

Ассоциативные правила, найденные по базе D_H, называются ***пороговыми***.

Порогу p_i, $p_i \in P_i$, поставим в соответствие элемент $x = (x_1, \dots, x_n) \in P$, в котором $x_i = p_i$ и $x_j = l_j$ при $j \neq i$. Тогда **порог** p_i называется ***значимым***, если x – s -*частый*.

Набор порогов $H = (p_1, \dots, p_n)$ называется ***значимым***, если $\forall i \in \{1, 2, \dots, n\}$: p_i – значимый.

Поиск пороговых ассоциативных правил

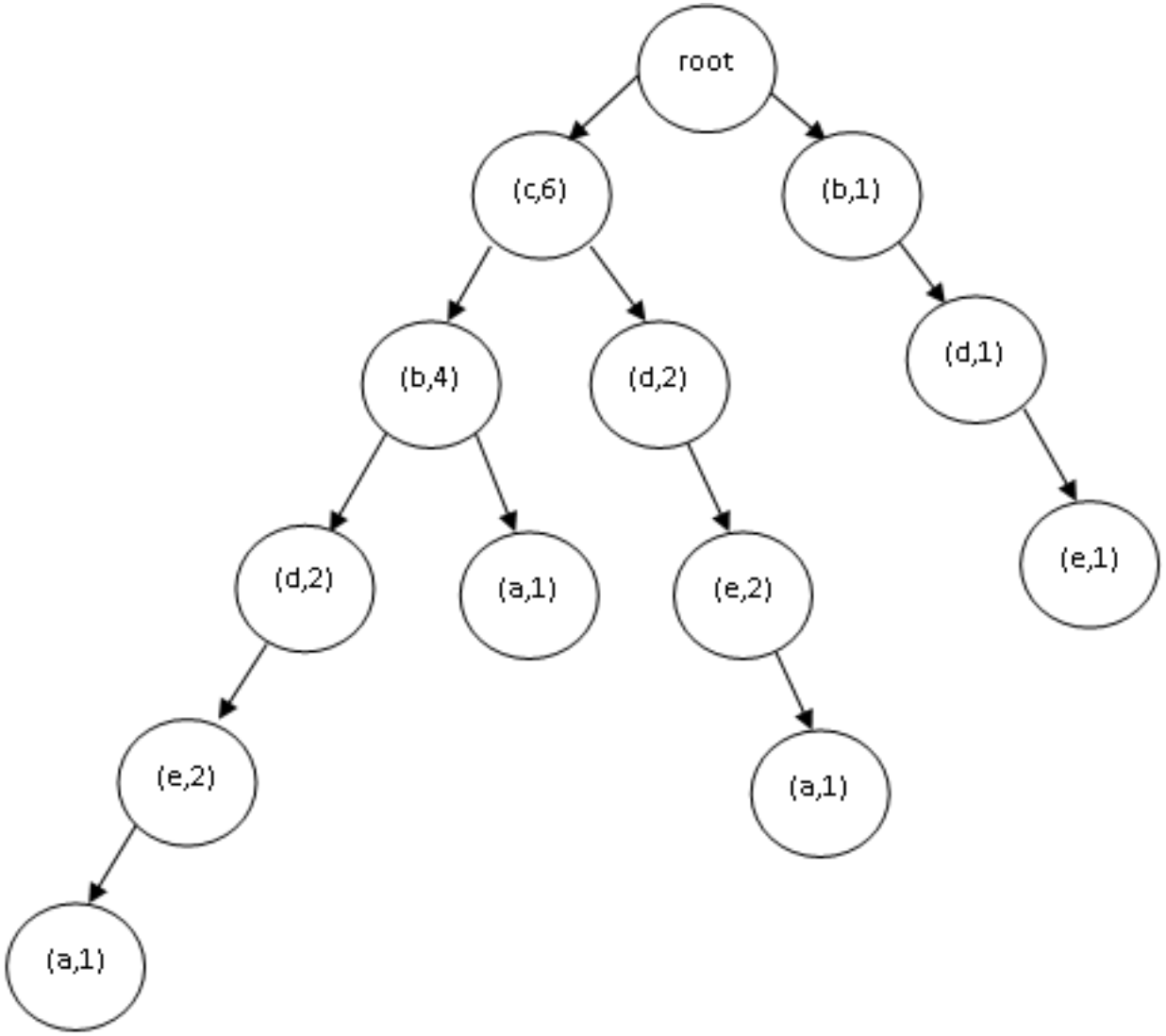
Задача: найти все пороговые ассоциативные правила, которые порождаются значимыми наборами порогов.

Рассматриваются два способа решения задачи.

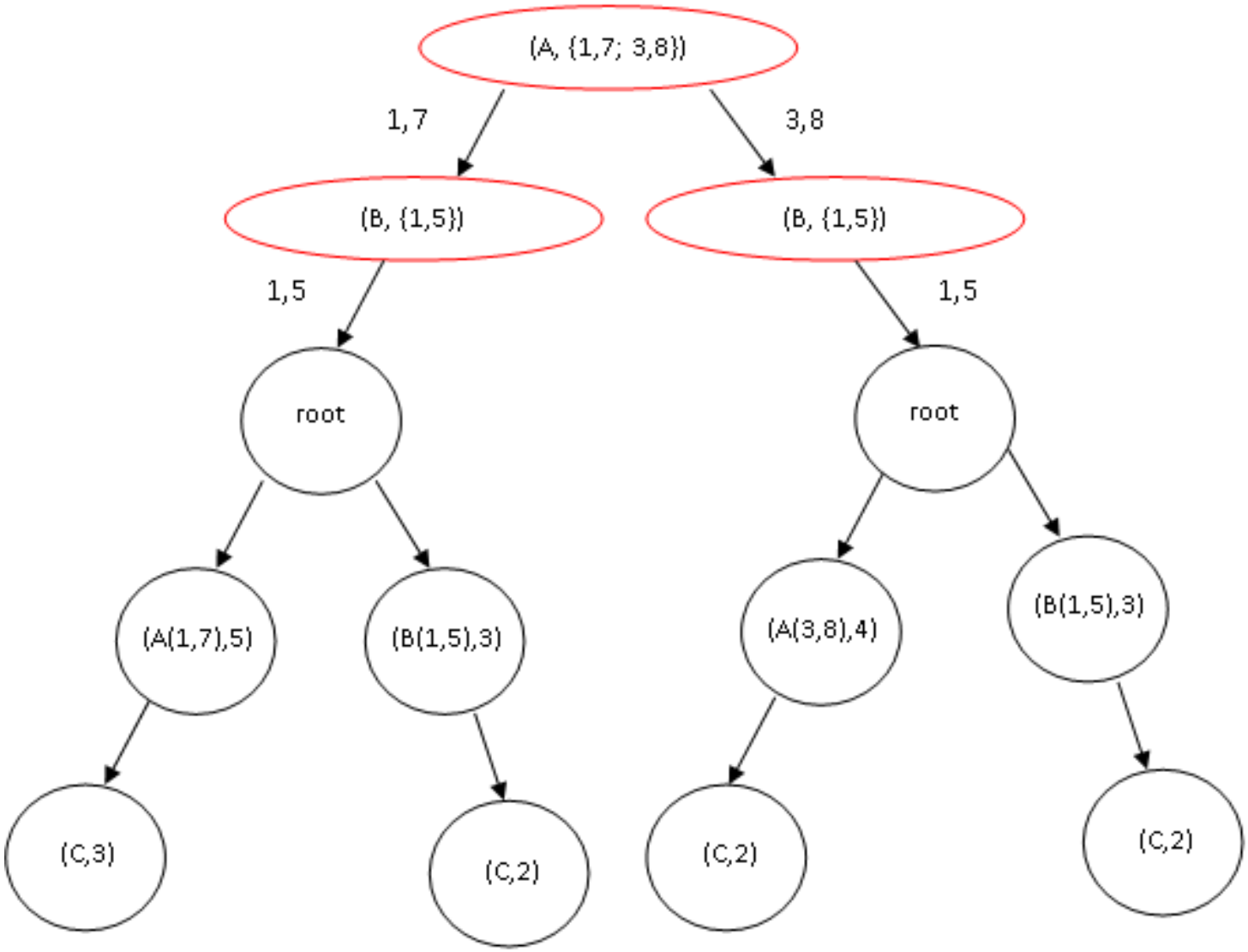
Первый способ основан на последовательном нахождении значимых наборов порогов и построении для каждого такого набора классического бинарного FP-дерева.

Второй способ, основанный на построении так называемого **полного FP-дерева**. Предложенная нами конструкция полного дерева является модификацией классического FP-дерева и позволяет просматривать все возможные варианты бинаризации исходных данных.

Пример FP-дерева



Пример полного FP-дерева



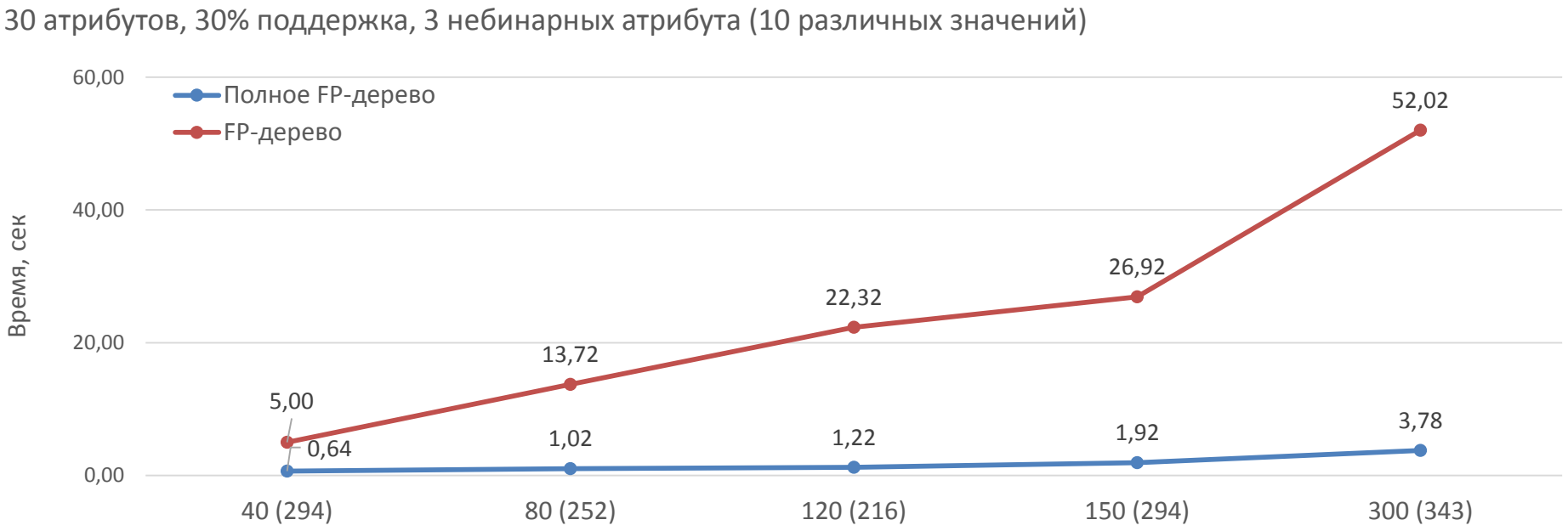
Свойства полного FP-дерева

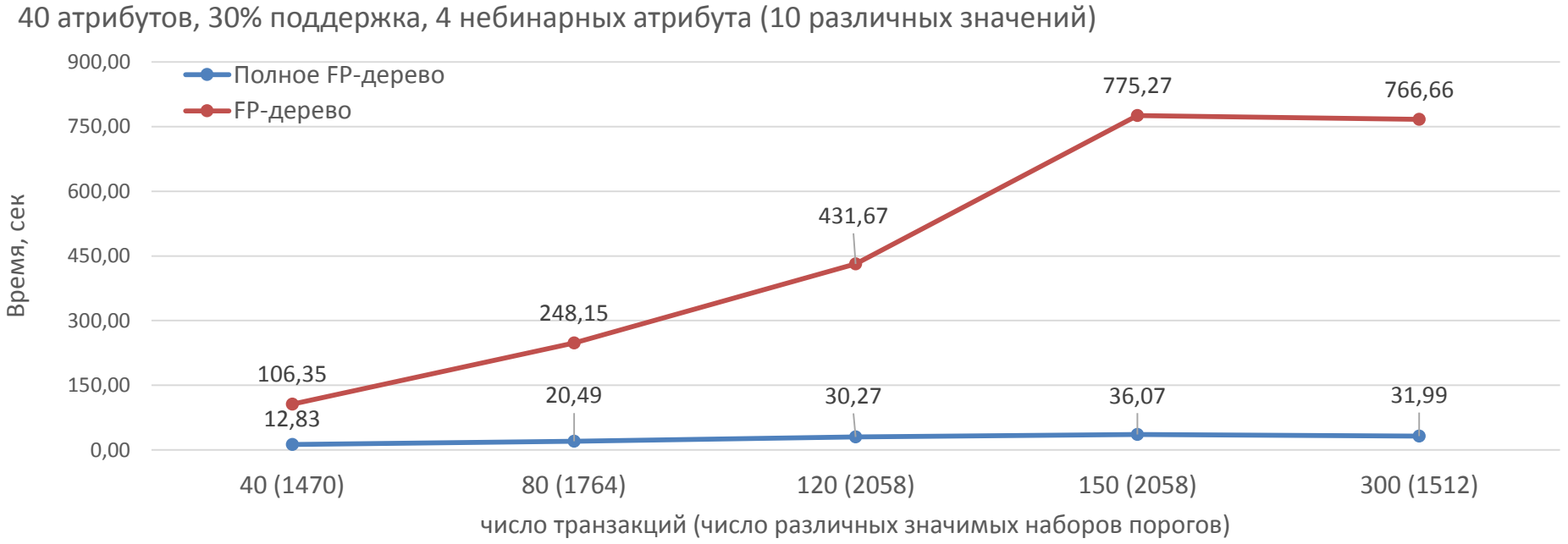
Пусть даны два значимых набора порогов: $H_1 = \{p_{1,\dots,}p_n\}$ и $H_2 = \{q_{1,\dots,}q_n\}$, отличные по одному порогу с индексом t. Пусть также порогу p_t соответствует элемент $x \in P$, а порогу q_t элемент $y \in P$.

Свойство 1. Пусть $R(H_1)$ – множество s-частых элементов из P_{H_1}, $R(H_2)$ – множество s-частых элементов из P_{H_2}, тогда $R(H_2) \subseteq R(H_1)$, если $S_D(x) \geq S_D(y)$.

Свойство 2. Пусть $R_t(H_1)$ – множество s-частых элементов из P_{H_1} в которых значение атрибута с индексом t равно нулю, $R_t(H_2)$ – множество s-частых элементов из P_{H_2} в которых значение атрибута с индексом t равно нулю, тогда $R_t(H_1) = R_t(H_2)$, если $S_D(x) \geq S_D(y)$.

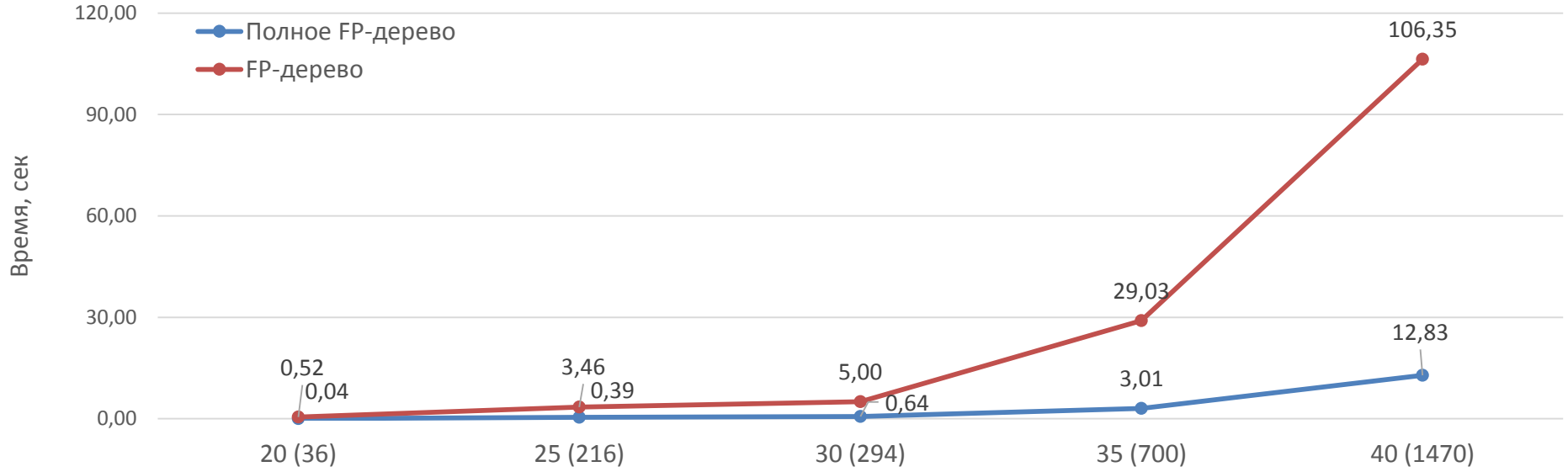
Численные эксперименты



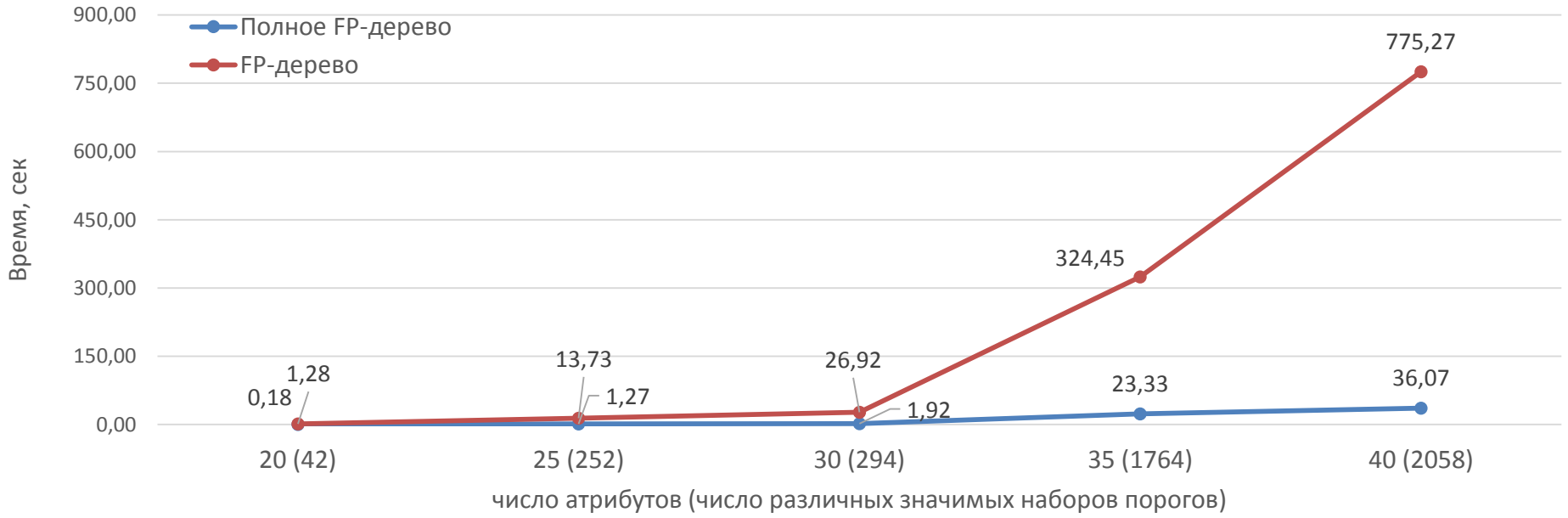


Численные эксперименты

40 транзакций, 30% поддержка, 90% бинарных атрибутов, 10 различных значений в небинарных атрибутах



150 транзакций, 30% поддержка, 90% бинарных атрибутов, 10 различных значений в небинарных атрибутах



Основные результаты

- Рассмотрены вопросы поиска ассоциативных правил в небинарных данных, представленных в виде произведения частичных порядков. Введено понятие порогового ассоциативного правила.

- Разработан эффективный подход к поиску всех возможных пороговых ассоциативных правил частично упорядоченных данных, основанный на построении полного FP-дерева. Обоснование подхода проведено на модельных данных.

Дальнейшие исследования

- Численные эксперименты на реальных данных.
- Реализация алгоритма синтеза полного FP-дерева с применением параллельных вычислений на основе технологии CUDA.
- Поиск пороговых ассоциативных правил специального вида, называемых неприводимыми ассоциативными правилами.