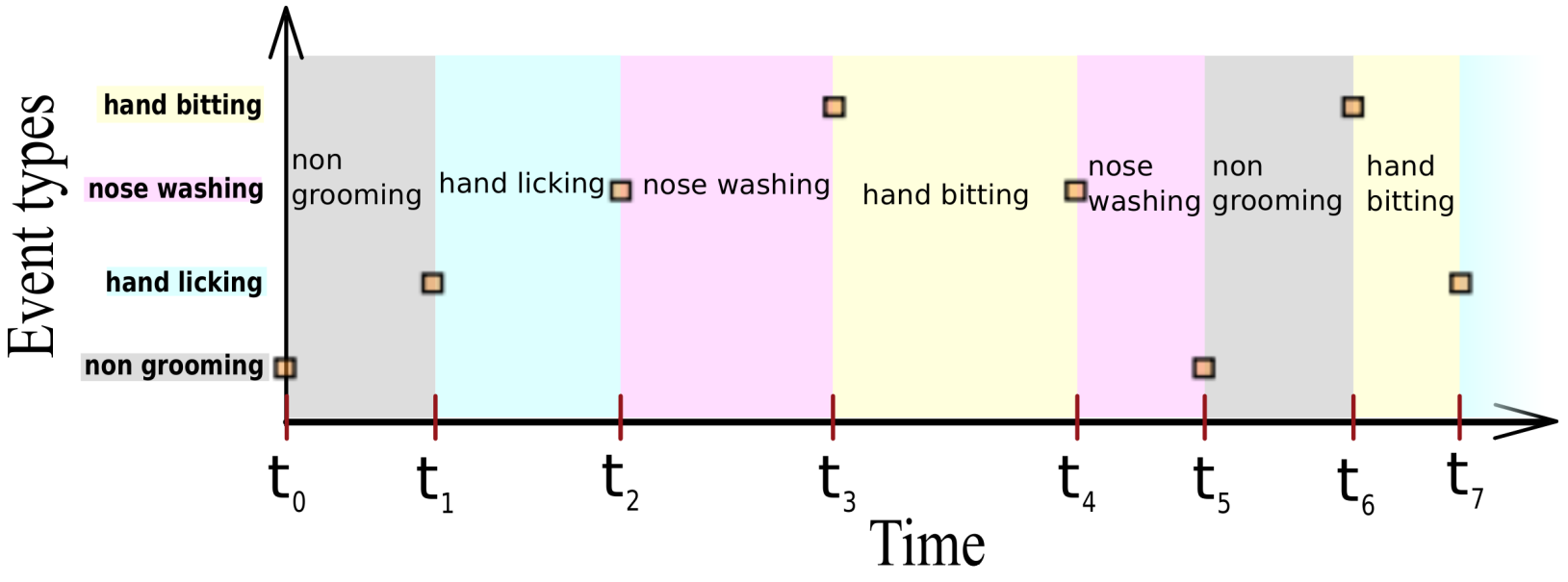
Вероятностный подход к поиску поведенческих паттернов

Вишневский В.В.

25 мая 2011 г.

Входные данные



Подход к поиску паттернов

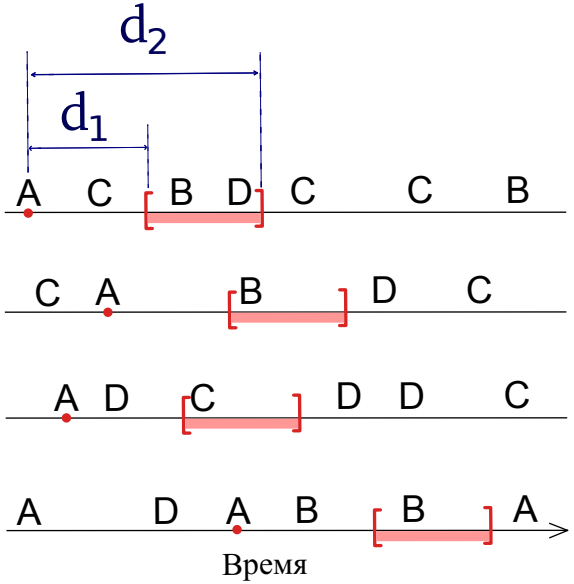
Паттерн — это часто встречающаяся последовательность событий(поведенческих актов), возникающих один за другим через определенные промежутки времени.

Инициализируем множество паттернов поведенческими актами. Потом итеративно повторяем:

- **Конструирование:**Для всех пар паттернов проверить, повторяется ли один за другим достаточно часто. Если да, то получаем новый паттерн.
- **Редукция:** Удалить одинаковые паттерны, которые были сконструированы по-разному.

Понятие Т-Паттерна(M.S. Magnusson)

- События соединеятся критическими интервалами. $A[dA_l, dA_r]B[dB_l, dB_r]C \ldots F$.
- Критический интервал $(A[d_1, d_2]B)$ – это связь между двумя паттернами, означающая, что паттерн B появляется в промежутке $[d_1, d_2]$ после паттерна A *чаще, чем ожидается*.

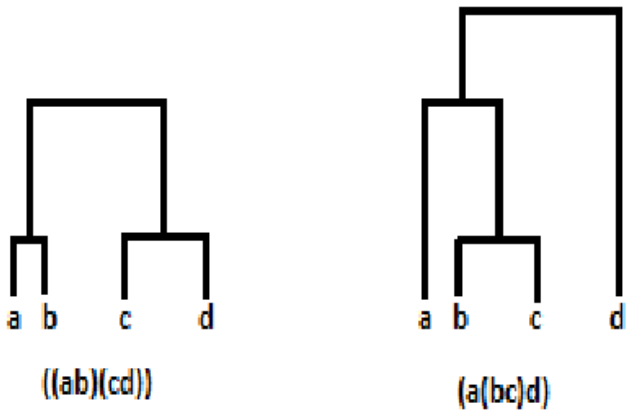


чаще, чем ожидается

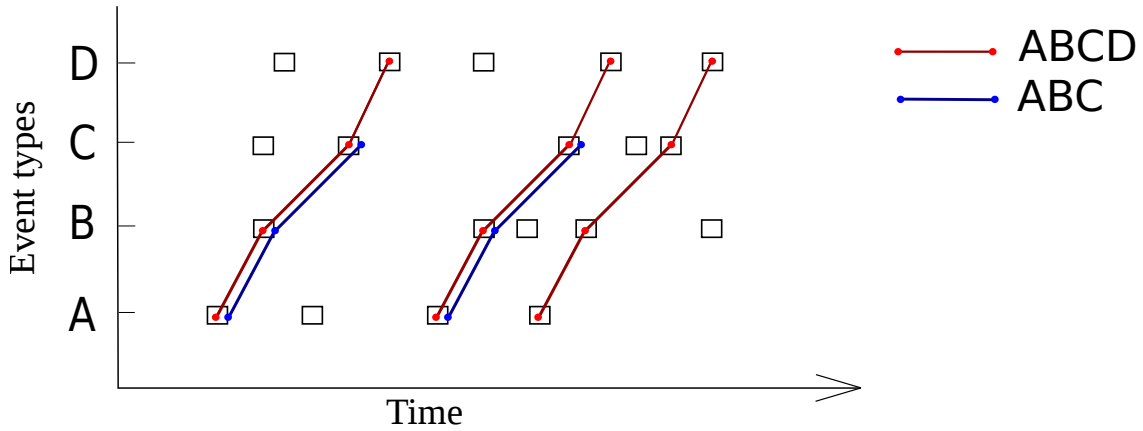
- Гипотеза H_0: события распределены равномерно и независимо. То есть закономерностей нету.
- Считаем вероятность входных данных при условии H_0.
- Если эта вероятность мала(меньше α) $\Longrightarrow$ мы отвергаем H_0 и считаем, что существует закономерность.

Типы «лишних» паттернов

- **Дубликаты:** $(AB)(CD)$ и $(A(BC))D$



- **Неполные копии:** (BCD) не встречается вне $(ABCD)$

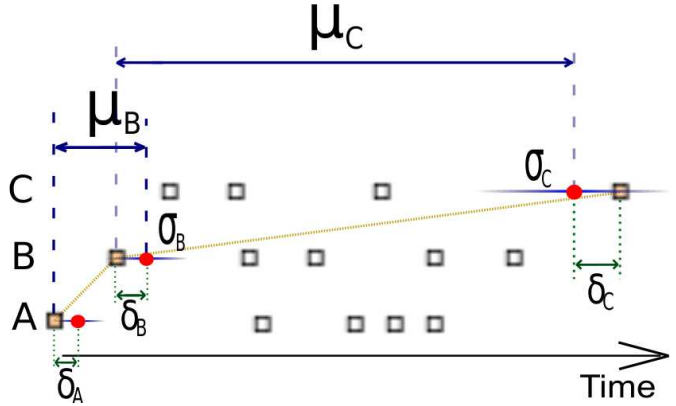


Недостатки Т-Паттернов

- Чувствительность к шуму и пропускам в исходных данных.
- Часто выделяется слишком много похожих паттернов.
- Закрытые исходные коды.

Вероятностная модель P-Паттерна

- $P = A[\mu_A, \sigma_A]B[\mu_B, \sigma_B]C[\mu_C, \sigma_C]$



- Функция потерь:

$$f_{LOSS}(x, N) = \begin{cases} \exp\left(-\frac{\lambda x}{N}\right), & x < N, \\ 0, & x = N. \end{cases}$$

- Правдоподобие паттерна:

$$L_P(\varepsilon) = f_{LOSS}(N_-, N_P) \prod_{i=1}^{N_P} \left(\frac{1}{\sqrt{2\pi}\,\sigma_i}\right) \prod_{i \in \mathcal{N}_+} \exp\left(-\frac{\delta_i^2}{2\sigma_i^2}\right)$$

Правдоподобие

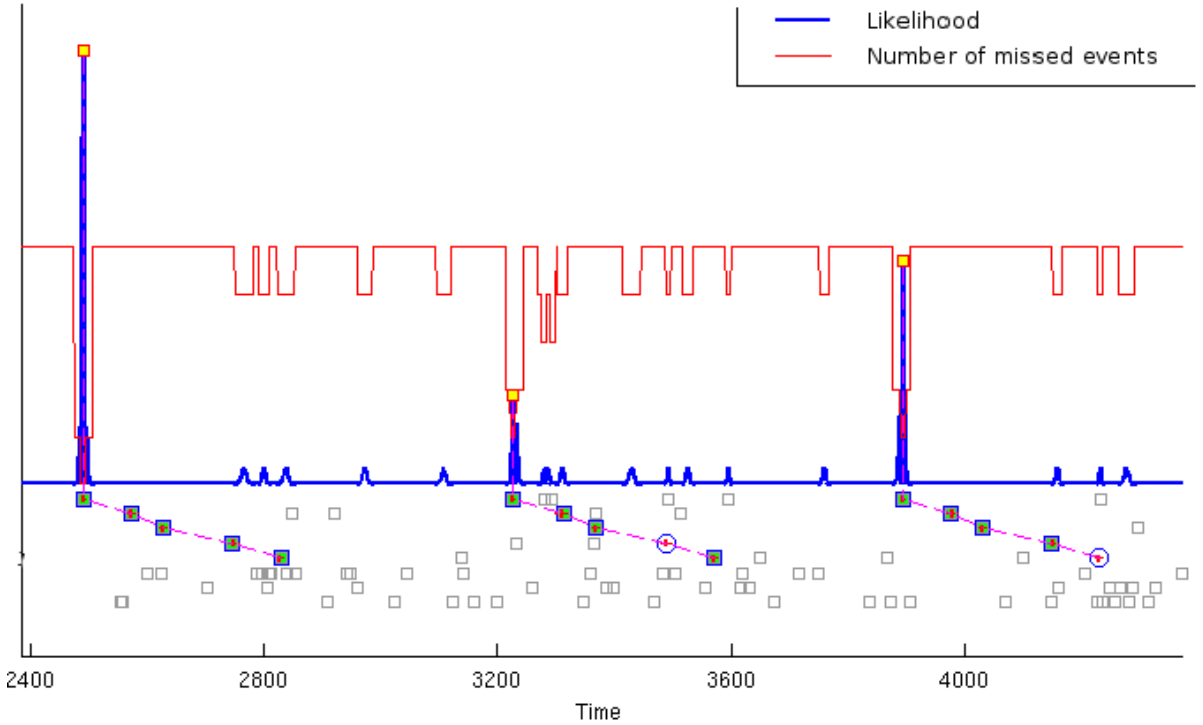


Рис.: Пример функции правдоподобия паттерна. Желтыми маркерами с красной границей изображены максимумы функции правдоподобия: моменты времени, когда мы считаем, что паттерн имеет место. В нижней части рисунка закрашенными квадратами показаны присутствующие события, полыми кружками — пропущенные события в паттерне. Полые серые квадраты соответствуют наблюдаемым поведенческим актам.

Конструирование новых паттернов

- Вводится модель связи событий:

$$g_{\mu,\sigma}(x_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right).$$

- Тестируется против гипотезы о равномерном, случайном, независимом распределении событий.
- Долгий перебор по μ и σ.

Удаление паттернов

Корреляция векторов значений функции правдоподобия:

$$cor\left(\overrightarrow{L_1}, \overrightarrow{L_2}\right) = \frac{\overrightarrow{L_1}^{\top}\overrightarrow{L_2}}{\sqrt{\overrightarrow{L_1}^{\top}\overrightarrow{L_1}}\sqrt{\overrightarrow{L_2}^{\top}\overrightarrow{L_2}}} \in [0, 1]$$

— коэффициент корреляции между двумя P-Паттернами. Чем он ближе к 1, тем два паттерна более близки друг к другу.

Параметры предложенного метода

Пар-р	Возможные значения	defaults	На что влияет
α	$[0, 1]$	0.001	Уровень значимости паттерна
N_{min}	$[0, +\infty]$	3	Минимальное количество появлений паттерна в данных
λ	$[0, +\infty]$	8	Допустимая степень нечеткости паттерна
ν	$[0, 1]$	0.6	Минимальная степень похожести паттернов для удаления
γ	$[0, 1]$	0.4	Чувствительность к отклонению от ожидаемого правдоподобия

T-Паттерны и P-Паттерны

- T-Паттерны распараллеливаем на SMP с помощью OpenMP. Тестирование на 4-х ядерном CPU.
- P-Паттерны распараллеливаем на GPU с помощью CUDA. Тестирование на GF 8800GTX, 128 потоковых процессора.

Ускорение OpenMP

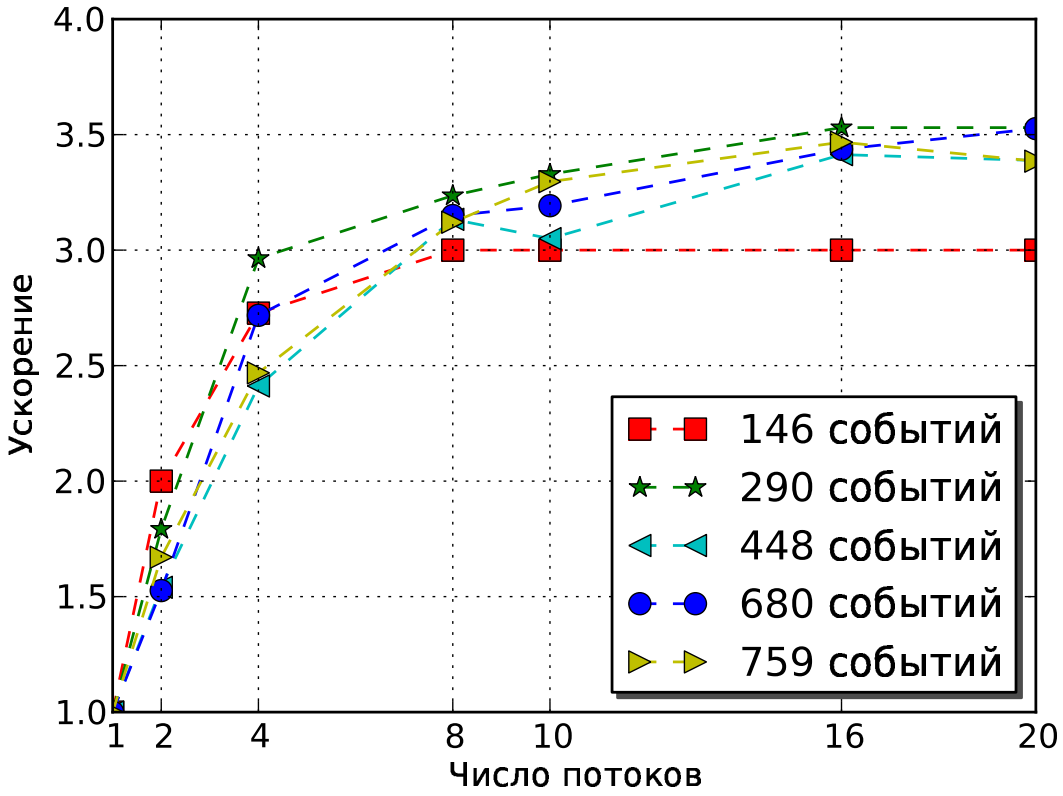


Рис.: Ускорение алгоритма поиска T-Паттернов на 4-х ядерном процессоре.

Ускорение алгоритма поиска P-Паттернов. CUDA

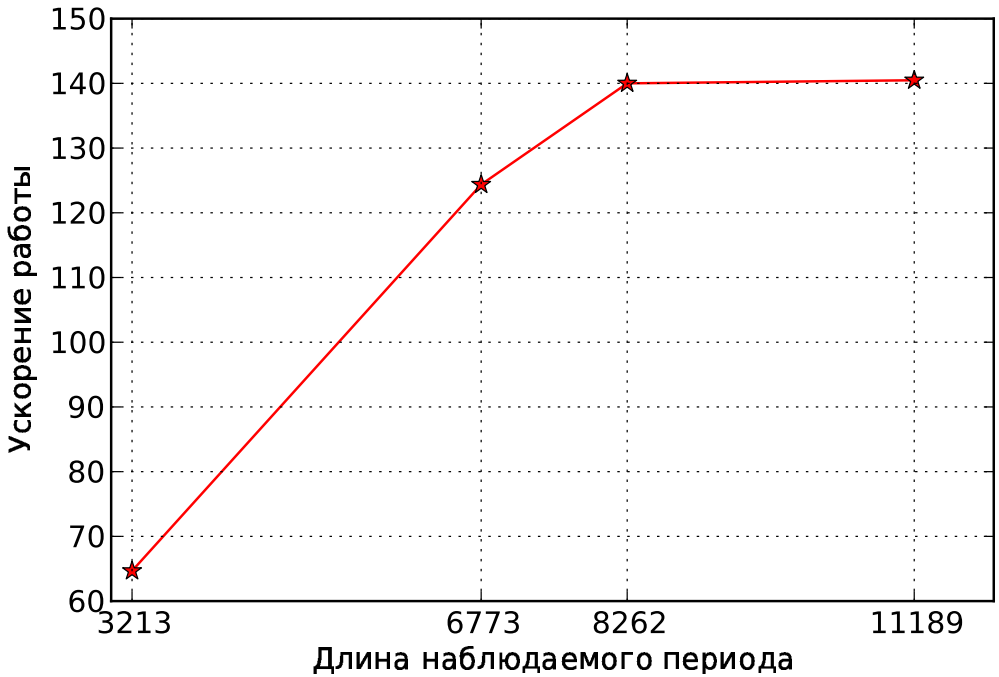


Рис.: Ускорение стадии подсчета правдоподобия паттернов в зависимости от размера входных данных.

Ускорение алгоритма поиска P-Паттернов. CUDA

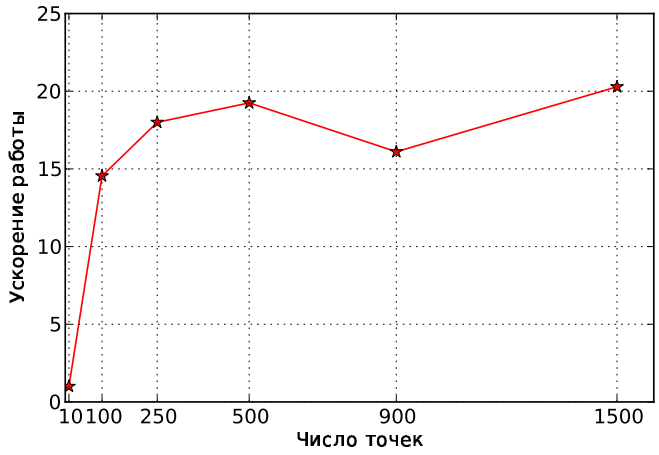
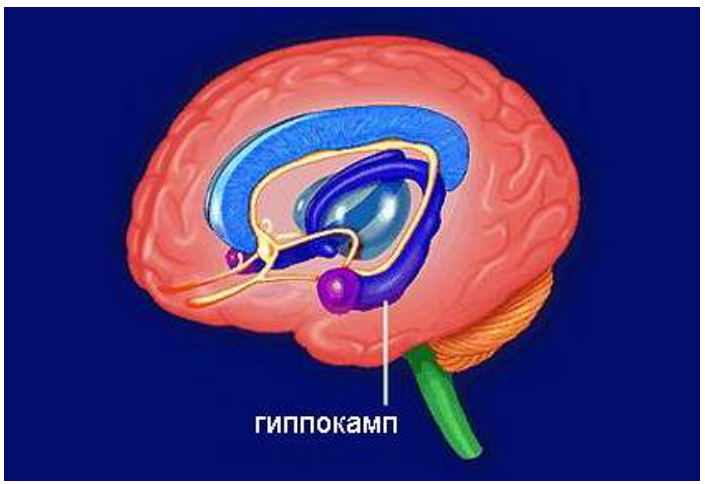


Рис.: Ускорение стадии конструирования паттернов в зависимости от размера входных данных.

Ускорение метода в целом $\sim$ 40 раз.
Типичные экспериментальные данные: 12 секунд на GPU, 470 секунд на CPU(1 поток).
Утилизация GPU $\sim$ 230 GFLOPS (Заявленная производительность 518 GFLOPS)

Эксперимент с грызунами без гиппокампа

- Гиппокамп — отдел головного мозга. Его функции связывают с механизмами работы памяти, обучением, пространственной навигацией.
- Две группы: контрольная(12 особей) и грызуны без гиппокампа(12 особей).
- Определить по поведению к какой группе относится особь.



Результаты экспериментов

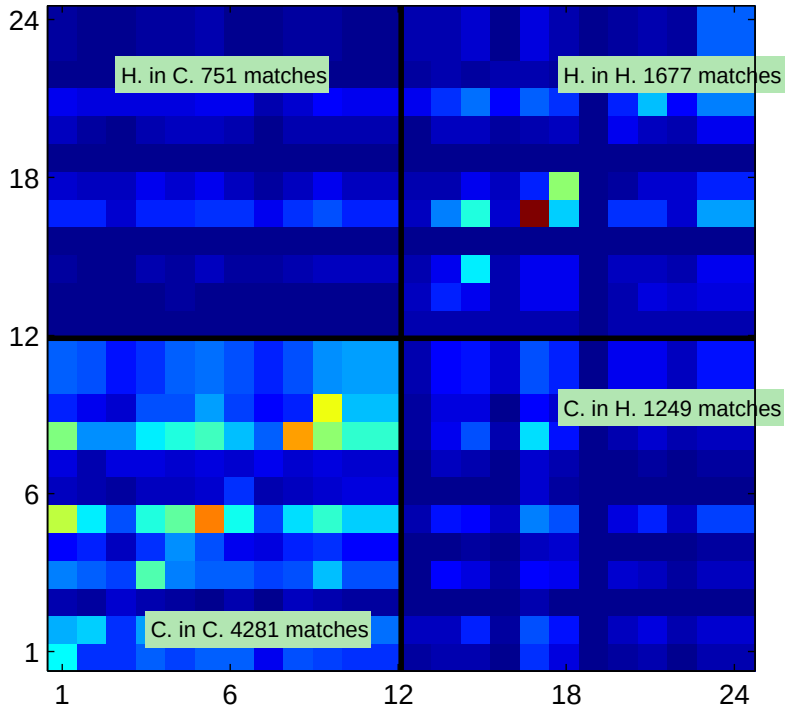


Рис.: Таблица соответствий паттернов. Неформально: по вертикали *откуда* берутся паттерны, по горизонтали — *где* ищутся вхождения этих паттернов; например, в ячейке $(3, 10)$ записано число соответствий паттернов третей особи в поведении десятой.

Классификация

- группа 1: контроль,
- группа 2: гиппокампальная,
- группа 3: шум с параметрами частоты и длины актов от группы 1,
- группа 4: шум с параметрами частоты и длины актов от группы 2,
- группа 5: данные содержащие 1 искусственный паттерн.

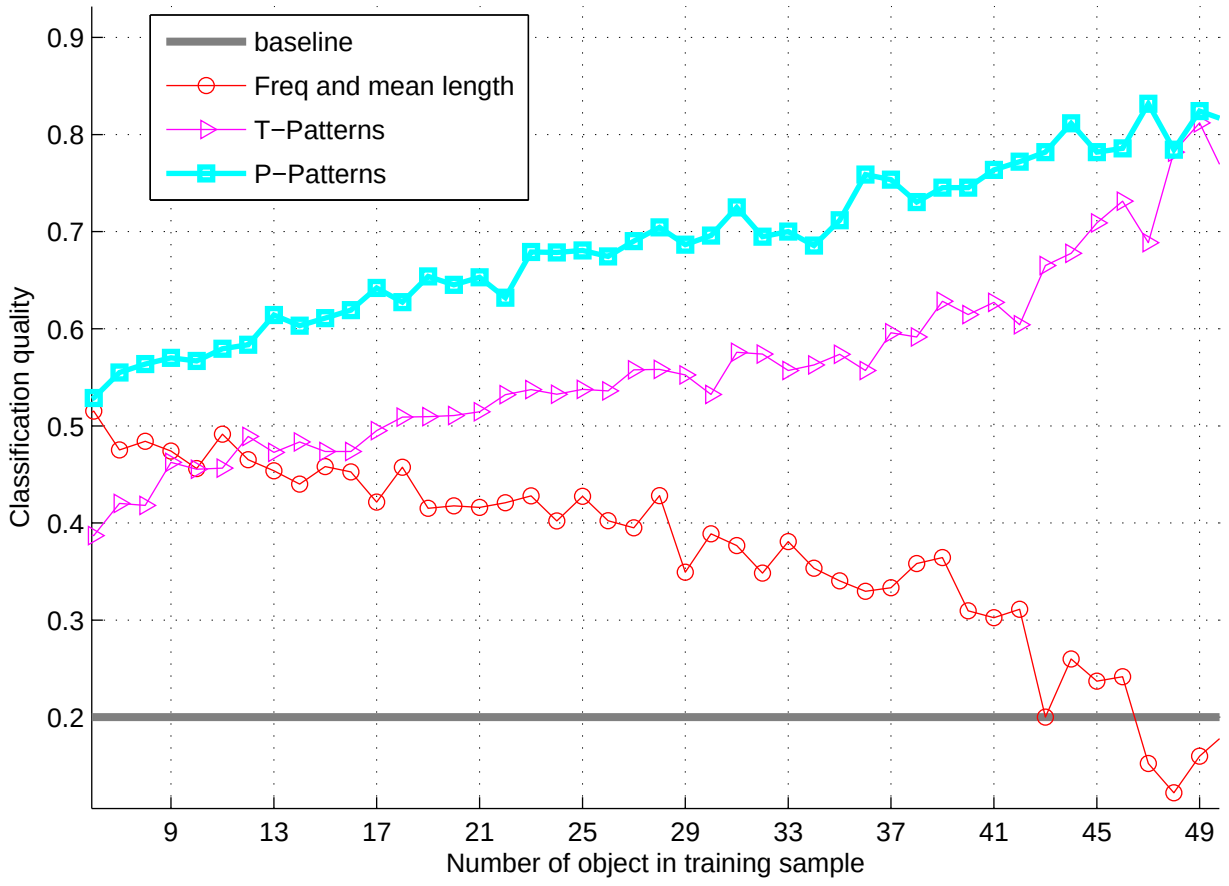
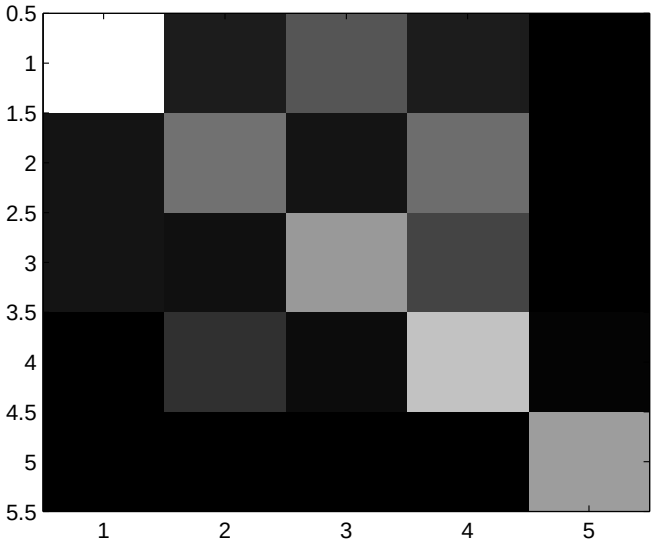


Рис.: Качество — средняя доля правильных классификаций. По горизонтали откладывалось количество объектов в обучении(для каждого значение качество усреднялось по ста повторениям с разными разбиениями для обучения).

Классификация



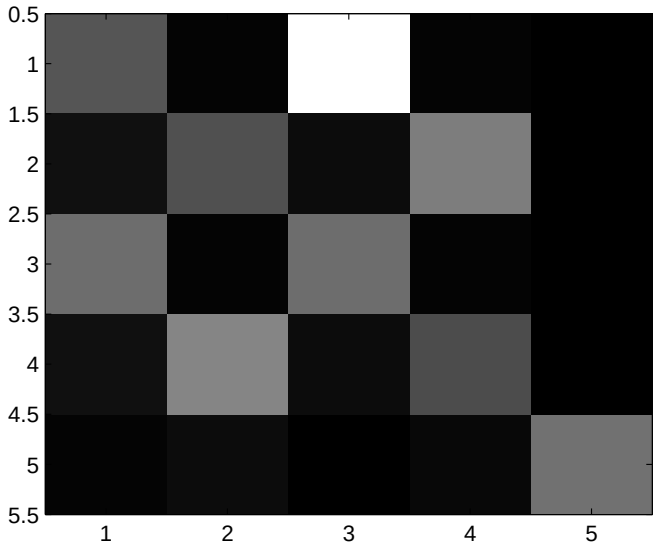


Рис.: Слева: confusion matrix для классификации по P-Паттернам, справа: по частотам и средним продолжительностям.

Общие паттерны

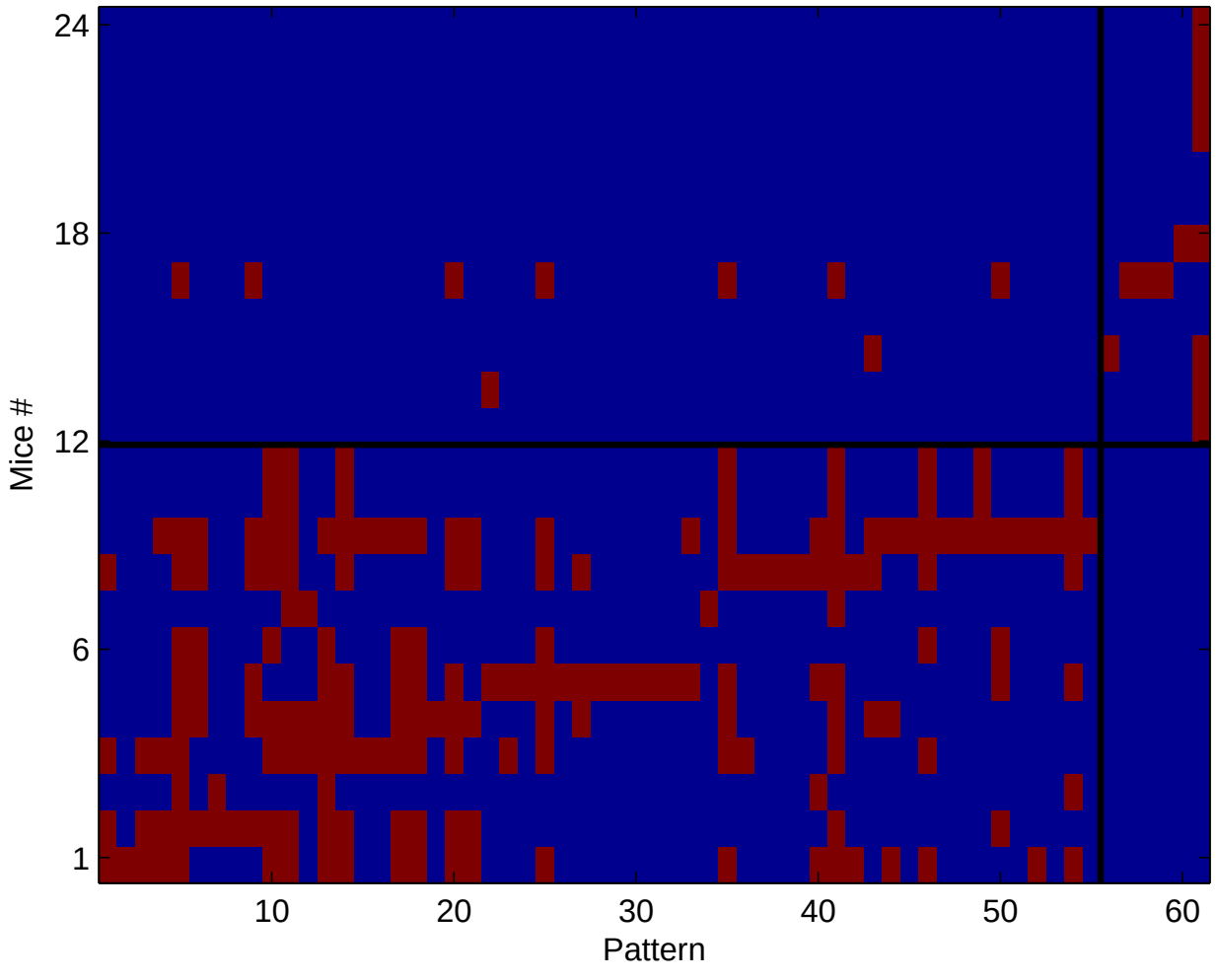
Параметры каждого паттерна подгоняются под конкретное наблюдаемое поведение.

- Вариант I. Ищем паттерны, найденные у одного животного в поведении других. Подсчет правдоподобия. Ослабить параметры.
- Вариант II. В процедуре конструирования паттренов рассматривать животных не по отдельности, а вместе.

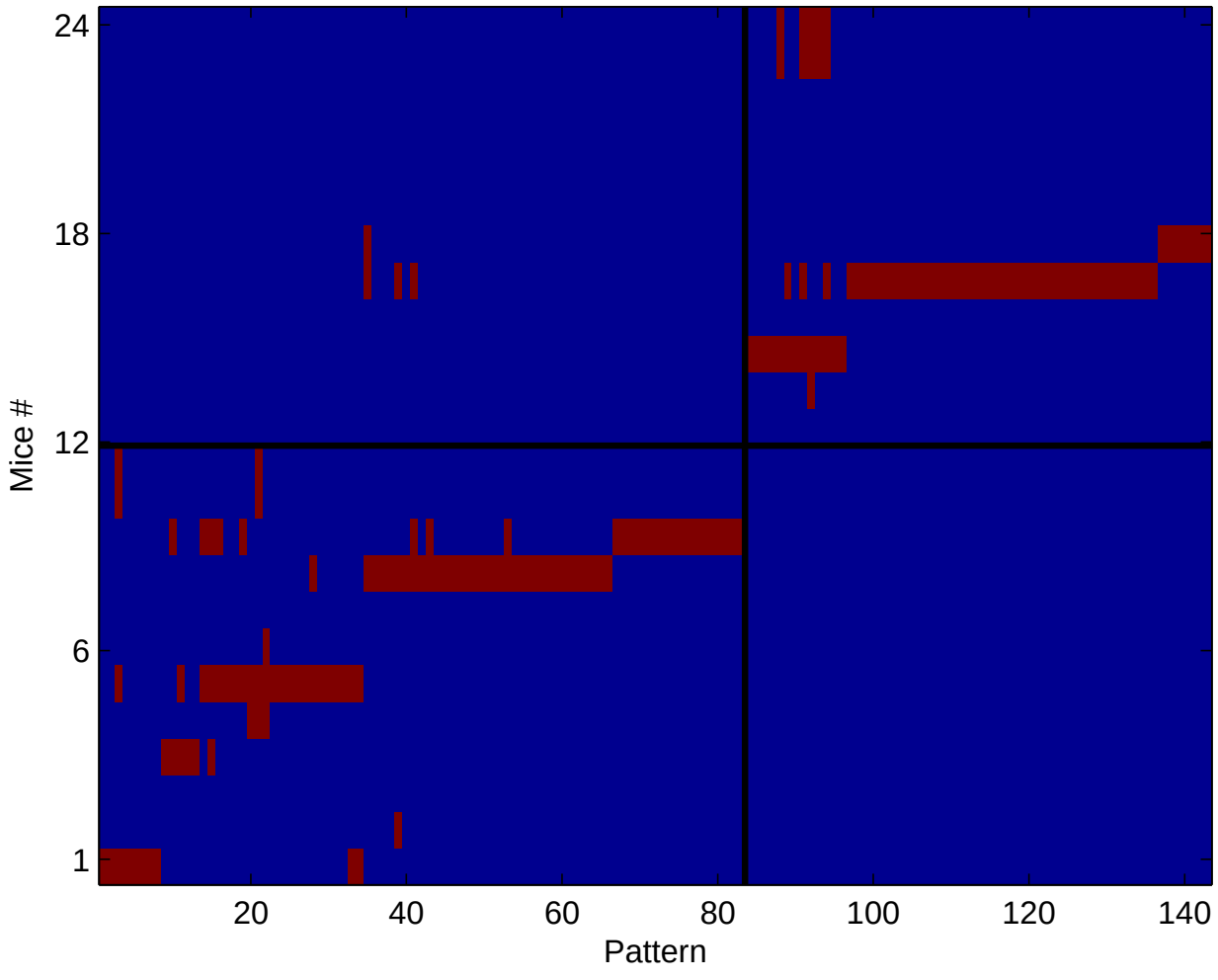
Характерные паттерны

- 2 группы: контроль(12 особей), гиппокампальная(12 особей).
- Всего в 24-ех файлах найдено 1150 P-Паттернов.
- Берем только P-Паттерны, содержащие 5 и больше событий.
- Для каждого такого P-Паттерна говорим, что он присутствует в поведении особи i, $(i = 1, \ldots, 24)$, если в этом поведении найдено больше, чем $N_{min} = 3$ экземпляра паттерна.
- P-Паттерн характерен для группы, если он встречается у многих особей из данной группы и редко встречается у особей из других групп.

Характерные P-Паттерны



Характерные T-Паттерны

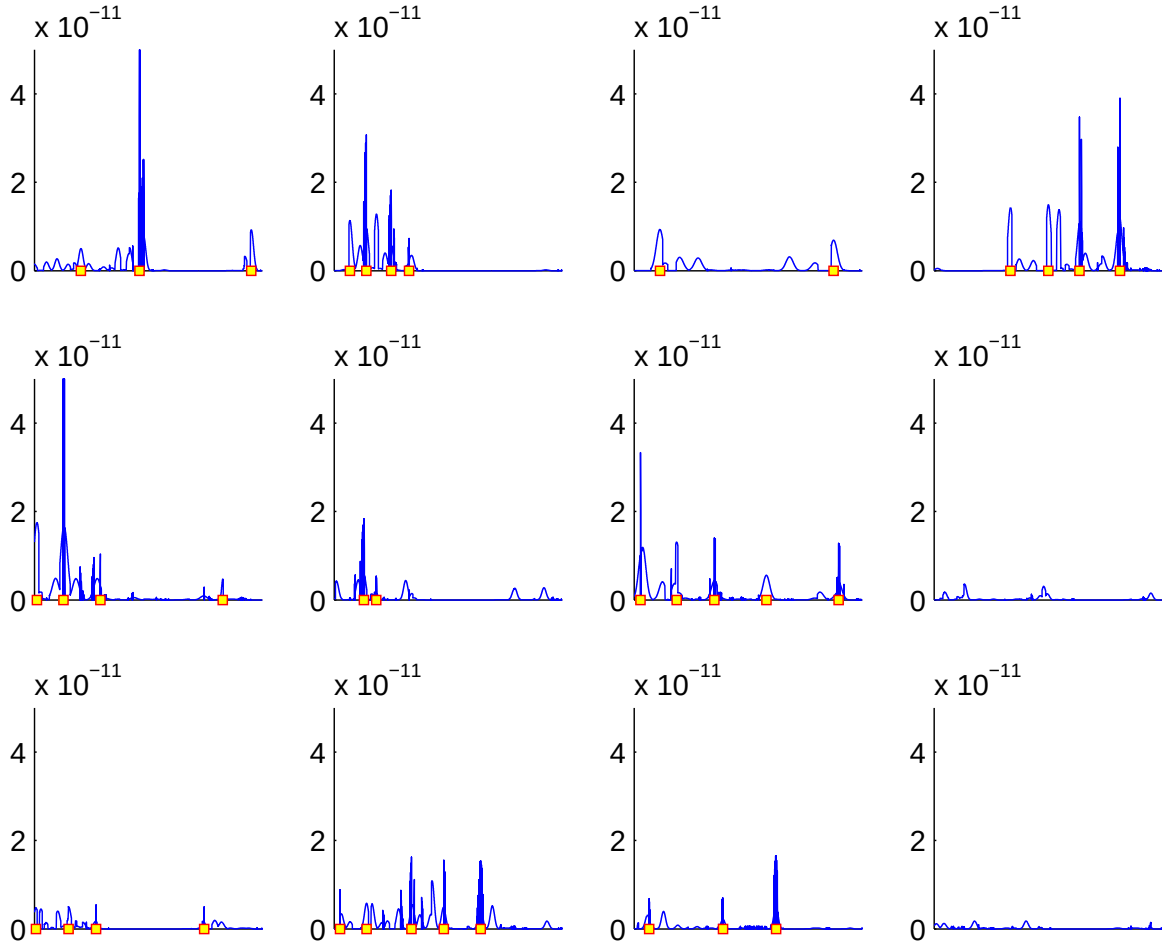


Один из характерных P-Паттернов контрольной группы

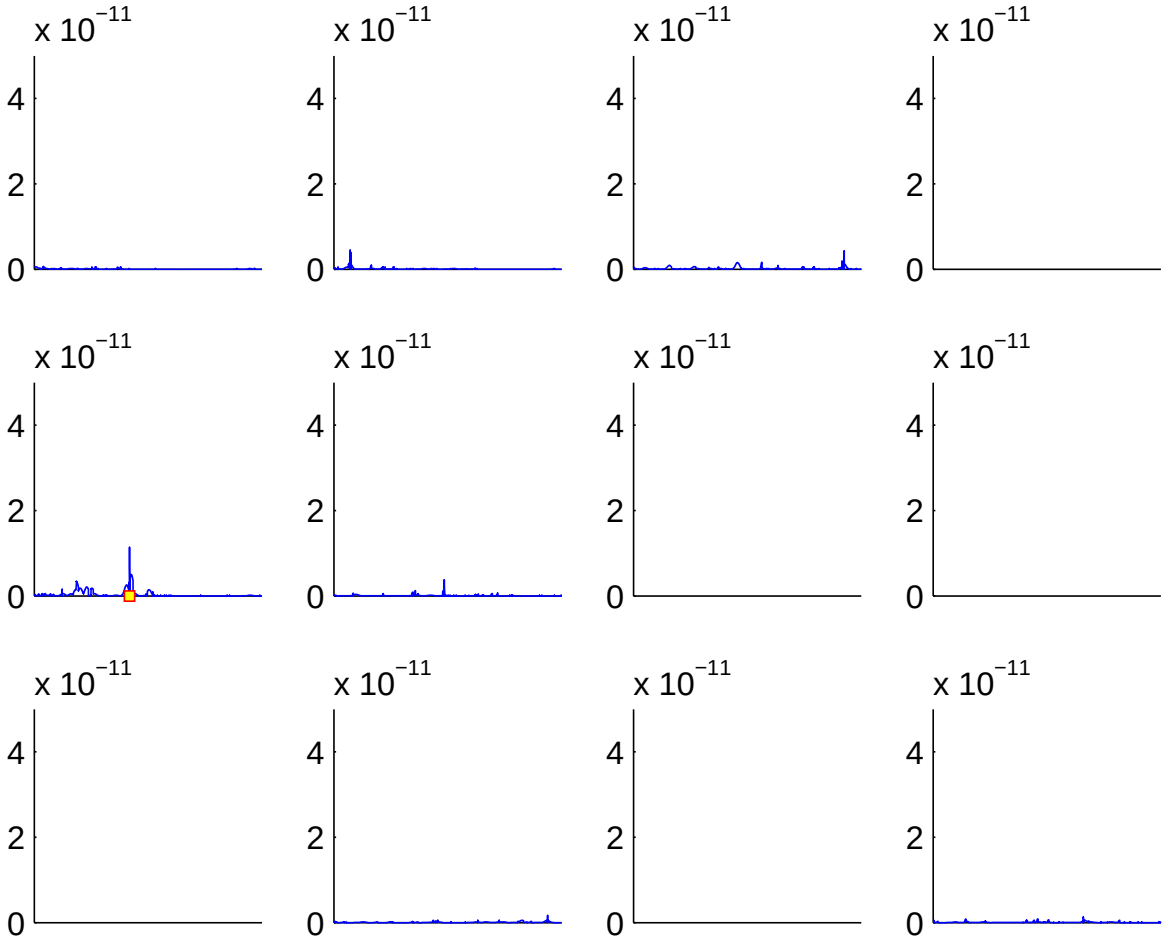
Вычесывание задними конечностями [22.9; 7.9] Вылизывание ладоней [1.1; 2.7] Быстрое умывание носа [0.4; 0.5] Умывание головы с ушами [3.2; 7.9] Умывание носа [17.0; 7.9] Вылизывание задних конечностей

Найден у 9 из 12 особей контрольной группы и ни разу не найден в гиппокампальной группе.

Отклик на P-Паттерн в контрольной группе



Отклик на P-Паттерн в гиппокампальной группе



Выводы

- Предложенный метод расширяет существующий подход к поиску паттернов.
- Устойчивость к шуму.
- Достигнуто ускорение параллельной версии на GPU в 40 раз.
- Качество классификации на экспериментальных данных $\sim 92\%$.
- Предложенный метод применим не только для анализа поведения животных(структура ДНК, спайковая активность нейронов, рынки, новостные тренды).

– Сложности на очень маленьких объемах данных.

– Долгое время работы на очень больших объемах данных.