Вероятностный подход к поиску поведенческих паттернов*

Вишневский В. В., Ветров Д. П.
valera.vishnevskiy@yandex.ru, vetrovd@yandex.ru
Москва, Москвоский Государственный Университет им. М. В. Ломносова

В данной работе предложен новый метод для поиска скрытых закономерностей (паттернов) в последовательностях событий, основанный на вероятностном представлении P-Паттернов(probabilistic pattern) в дискретных последовательностях событий. Поиск производится снизу вверх: сначала находятся простые закономерности, потом, путем их соединения, образуются более сложные паттерны. Рассматривается применение данного алгоритма для анализа поведения мышей. Найденные паттерны используются для классификации животных. Проведено сравнение реализованного алгоритма с существующими аналогами, показавшее, что предложенный метод более устойчив к шуму в исходных данных.

Задача поиска закономерностей(стереотипов, паттернов, шаблонов — здесь синонимы) в поведении животных и людей крайне важна в современной нейробиологии и когнитивных науках. Выделив характерные паттерны, можно, например, делать выводы о сложности поведения различных особей, определять изменения в поведении наблюдаемых процессов, другими словами, решив задачу поиска паттернов, мы можем определенным образом *измерять* поведение особи, или группы особей.

В данной работе мы будем рассматривать временны́е паттерны в «структурном», или «эффектном» описании [3, с. 57]. Исходными данными будет размеченное поведение особи, то есть последовательность пар «момент времени», «поведенческий акт». Неформально можно сказать, что интересующие нас паттерны — это упорядоченная последовательность поведенческих актов, следующие один за другим через *относительно инвариантные* временны́е интервалы. Причем, этот паттерн должен повторяться в исходных данных *достаточно* часто.

Несмотря на то, что описанные выше паттерны широко распространены в описании поведения, стандартные статистические методы не подходят для их поиска: эти методы либо не учитывают всю сложность паттернов(например, периодические орбиты [4]), либо оперируют такими понятиями как циклы, волны, тренды, что невозможно напрямую использовать для поиска интересующих нас паттернов.

На сегодняшний день, для анализа таких поведенческих закономерностей наиболее широкое распространение получил метод поиска Т-Паттернов(temporal patterns), предложенный в 2000-ом году Магнусом Магнуссоном в [2].

Основные определения. Понятие Т-Паттерна

Пусть время наблюдения разбито на N_t интервалов. В каждый момент *периода наблюдения* $[1, N_t]$ может произойти некоторое событие $\mathbf{e}$(*действие, поведенческий акт, event*) [1] из множества допустимых событий $\mathcal{E}$. Соответственно, каждому типу события сопоставляется множество моментов времени $TS(\mathbf{e})$, когда это событие имело место:

$$TS(\mathbf{e}) = \{t_1^{\mathbf{e}}, \ldots, t_{N_\mathbf{e}}^{\mathbf{e}}\},$$
$$\mathbf{e} \in \mathcal{E}, \qquad 0 \leqslant t_i^{\mathbf{e}} \leqslant N_t, \quad (i = 1, \ldots, N_t),$$

здесь $N_\mathbf{e}$ — количество появлений события $\mathbf{e}$ в данных.

Понятие Т-Паттерна включает в себя определение модели связи последовательных событий и способа их появления в данных(т.е. определения, что данная структура не случайна, а является *закономерностью*). Каждое событие паттерна определяется фиксированным временны́м интервалом, в течение которого это событие должно присутствовать после предыдущего события. Другими словами, расстояния между событиями моделируются равномерным распределением.

Основной недостатк метода поиска Т-Паттернов заключается, во-первых, в том, что само определение Т-Паттерна не позволяет ему иметь пропуски событий. По этой причине метод становится крайне чувствителен к шуму в исходных данных, из-за чего можно пропустить информативные длинные и сложные паттерны. Во-вторых, полученные Т-Паттерны сильно специфичны особи, в поведении которой они были найдены.

Работа выполнена при финансовой поддержке РФФИ(проект №08-01-00405), гранта Президента РФ(МК3827.2010.9), и федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009–2013 годы(контракт №П1265).
Авторы статьи выражают благодарность членам «Лаборатории Нейробиологии Памяти» при Институте Нормальной Физиологии П.К. Анохина, возглавляемой членом-корреспондентом РАН К.В. Анохиным. Отдельно благодарим Ирину Зарайскую, предоставившую нам экспериментальные данные по поведению.

[1] Чаще всего понимается, что в этот момент времени имеет место *начало* действия

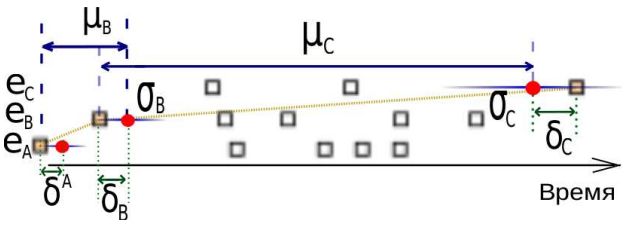

Рис. 1. Вхождение нечеткого паттерна $\mathbf{P} = \mathbf{e}_A[\mu_A, \sigma_A]\mathbf{e}_B[\mu_B, \sigma_B]\mathbf{e}_C[\mu_C, \sigma_C]$. в данные. Маркеры-кружки соответствуют ожидаемым позициям событиям, закрашенные маркеры-квадраты соответствуют позиции реальной позиции соответствующего события в данных.

Вероятностная модель паттерна. P-Паттерны

Определение 1. *Нечетким паттерном, или P-Паттерном* $\mathbf{P}$ *длины* $N_\mathbf{P}$ *назовем упорядоченную последовательность событий* $\mathbf{e}_i, (i = 1, \ldots, N_\mathbf{P})$, *где каждое событие паттерна характеризуется смещением и разбросом от предыдущего события. Будем записывать паттерн* $\mathbf{P}$ *в следующем виде:*

$$\mathbf{P} = [\mu_1, \sigma_1]\mathbf{e}_1[\mu_2, \sigma_2]\mathbf{e}_2 \ldots [\mu_{N_\mathbf{P}}, \sigma_{N_\mathbf{P}}]\mathbf{e}_{N_\mathbf{P}}, \quad \mu_1 = 0.$$

Здесь μ_i *и* σ_i *— метематическое ожидание и корень из дисперсии нормального распределения, моделирующего величину времени, прошедшего между событиями.*

Представление P-Паттерна иллюстрировано на рис. 1.

Далее, чтобы иметь возможность обрабатывать пропуски в P-Паттернах, введем понятие *функции потерь*, которая определяет «штраф» за пропуск m событий в паттерне длины $N_\mathbf{P}$ следующим образом:

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

Здесь λ является структурным параметром, определяющим уровень «нечеткости» паттернов. Если этот параметр велик, то мы, по сути, запрещаем реализациям паттерна иметь пропуски. Если выставить этот параметр слишком малым, то будут обнаруживаться паттерны, не разу полностью не встречающиеся в данных, то есть закономерности могут быть найдены даже в случайных данных.

Определение 2. *Правдоподобие паттерна* $\mathbf{P}$ *— это функция, определенная в каждый момент времени наблюдения* ε $(\varepsilon = 1, \ldots, N_t)$ *следующим образом:*

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

где δ_i *— расстояния от ожидаемой позиции события в P-Паттерне до ближайшего события в данных(более наглядно см. рис. 1). Т.е.:*

$$\delta_i = \min_{x \in TS(\mathbf{e}_i)} \left| \underbrace{\varepsilon + \sum_{j=1}^{i-1}(\mu_j + \delta_j) + \mu_i}_{\text{ожидаемая позиция события}} - x \right|,$$

здесь, если событие было пропущено, то соответствующее $\delta_i = 0$. *Далее,* N_- *— количество пропущенных событий в паттерне, а* $\mathcal{N}_+$ *— множество индексов присутствующих в паттерне событий. Событие считается пропущенным, если* $\exp\left(-\frac{\delta_i^2}{2\sigma_i^2}\right) < \exp\left(-\frac{\lambda}{N_\mathbf{P}}\right)$.

По сути, правдоподобие показывает насколько можно убыть уверенным, что данный P-Паттерн начинается в определенный момент времени ε.

Заметим, что правдоподобие P-Паттерна может быть отсчитано с конца, или с m-го события паттерна.

Утверждение 1. *Математическое ожидание функции правдоподобия* (1) *в момент времени* ε, *при условии, что в данный момент времени имеет место начало модельного P-Паттерна* $\mathbf{P}$ *вычисляется следующим образом:*

$$\mathsf{E}\left[L_\mathbf{P}(\varepsilon)\right] = \frac{1}{(2\sqrt{\pi})^{N_\mathbf{P}}\,\sigma_1 \ldots \sigma_{N_\mathbf{P}}}.$$

Доказательство приведено в [1].

Здесь используется тот факт, что межточечные расстояния между событиями в модельном P-Паттерне распределены по нормальному закону:

$$\delta_i \sim \mathcal{N}(0, \sigma_i), \quad (i = 1, \ldots, N_\mathbf{P}).$$

Теперь мы можем считать, что P-Паттерн $\mathbf{P}$ имеет место быть только в следующие моменты времени:

$$t\colon L_\mathbf{P}(t) \geqslant \gamma \mathsf{E}\left[L_\mathbf{P}(\varepsilon)\right], \quad (2)$$

где γ — заданная константа.

Конструирование P-Паттернов. Рассмотрим пару P-Паттернов $\mathbf{P}_L$(левый) и $\mathbf{P}_R$(правый).

Пусть $\{\alpha_i\}_{i=1,\dots,N_L}$ и $\{\beta_j\}_{j=1,\dots,N_R}$ — значения правдоподобия соответствующих Р-Паттернов в моменты времени, соответствующие паттерны имели вхождения (2). Важно, что правдоподобие левого Р-Паттерна отсчитывается с конца, так как мы ищем связь между концом левого паттерна и началом правого. Также пусть, $\{t_{L,i}\}_{i=1,\dots,N_L}$ и $\{t_{R,j}\}_{j=1,\dots,N_R}$ — моменты времени, когда эти Р-Паттерны имели место в смысле (2). N_L и N_R — количество вхождений паттернов $\mathbf{P}_L$ и $\mathbf{P}_R$, соответственно. Определим множество межточечных расстояний:

$$\rho = \{t_{R,j} - t_{L,i} \mid t_{R,j} \geqslant t_{L,i}\}.$$

Для каждого расстояния из этого множества введем соответствующий вес $w_l = \ln(1 + \alpha_i\beta_j), \quad l = 1, \dots, M$, где $M = |\rho|$.

Рассмотрим гипотезу H_0, что моменты времени и веса вхождения Р-Паттернов распределены независимо и равномерно на всем наблюдаемом промежутке. Тогда плотность распределение введенных выше межточечных расстояний $\{t\}$ имеет следующий вид [1]:

$$p_{LR}(t) = \begin{cases} (N_t - t)\,\frac{2}{N_t^2}, & t \in [0, N_t], \\ 0, & x \notin [0, N_t]. \end{cases} \quad (3)$$

Введем статистическую модель связи между паттернами(проверяемые параметры связи μ и σ фиксированы):

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

Рассмотрим следующую сумму:

$$k = \sum_{i=1}^{M} w_i g_{\mu,\sigma}(t_i), \quad (4)$$

где $t_i \sim p_{LR}$.

Теорема 2. *При выполнении H_0 статистика k распределена по нормальному закону:*

$$\begin{aligned} & k \sim \mathcal{N}\left(\mu_*, \sigma_*^2\right), \text{ где} \\ & \mu_* \approx M\,\mathsf{E}[w]\,\frac{2}{N_t}\left(1 - \frac{\mu}{N_t}\right) \\ & \sigma_*^2 \approx \frac{M}{N_t}\left(1 - \frac{\mu}{N_t}\right)\left[\left(\frac{1}{\sqrt{\pi}\,\sigma} - \frac{\mu}{N_t}\left(1 - \frac{\mu}{N_t}\right)\right) \times \right. \\ & \times \quad \left.\left((\mathsf{E}[w])^2 + \mathsf{D}[w]\right) + \frac{4\mathsf{D}[w]}{N_t}\left(1 - \frac{\mu}{N_t}\right)\right], \end{aligned} \quad (5)$$

здесь $\mathsf{E}[w]$ и $\mathsf{D}[w]$ — выборочное среднее и дисперсия весов, соответственно.

Доказательство приведено в [1].

Замечание 1. Приближенные формулы (5) дают удовлетворительный результат для задачи поиска паттернов, однако, точные формулы для значений μ_* и σ_*, можно найти в [1].

Теперь для различных пар μ и σ можно вычислить статистику (4), сравнить ее с α-квантилью распределения (5). Если гипотеза H_0 о «случайности» данных будет отвергнута односторонним критерием, то считается, что соответствующие Р-Паттерны образуют новый паттерн с параметрами μ и σ. Если существует несколько пар μ и σ, для которых отвергается гипотеза H_0, то для конструирования Р-Паттернов берутся непересекающиеся[2] параметры соответствующие максимальным значениям k.

Более подробно о процессе формирования нового Р-Паттерна можно найти в [1].

Редукция Р-Паттернов. Для удаления [1] паттернов-дубликатов и неполных копий анализируется коэффициент корреляции функций правдоподобия. Пусть $\overrightarrow{L_{\mathbf{P},i}}$ —вектор-столбец значений функции правдоподобия, отсчитанной от i-го события во всех моментах времени наблюдения.

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

— коэффициент корреляции между двумя Р-Паттернами.

Проверяются все пары паттернов, если все поведенческие акты, присутствующие в паттерне $\mathbf{P}_L$ также присутствуют в $\mathbf{P}_R$ с учетом порядка, и

$$\exists m\colon\ cor\left(\overrightarrow{L_{P_L,1}}, \overrightarrow{L_{P_R,m}}\right) > \nu,$$

тогда паттерн $\mathbf{P}_L$ удаляется из множества найденных паттернов.

Алгоритм поиска Р-Паттернов

1. Инициализировать текущее множество Р-Паттернов событиями(паттерны длины 1).
2. Для всевозможных пар Р-Паттернов из текущего множества провести процедуру *конструирования*.
3. Провести процедуру редукции паттернов.
4. Если текущее множество паттернов изменилось, перейти к п.2.

Параметры алгоритма и способы их настройки описаны в [1].

Сложность предложенного алгоритма — $O(n^3)$, где n — общее количество событий во временном ряде. В [1] представлена параллельная реализация данного метода на GPU, что позволило применять алгоритм поиска Р-Паттернов на реальных данных.

[2] $[\,\mu' - 3\sigma', \mu' + 3\sigma'] \cap [\,\mu'' - 3\sigma'', \mu'' + 3\sigma''] = \varnothing$

Эксперименты на реальных данных

Описанный ниже эксперимент демонстрирует способ применения предложенного метода на реальных поведенческих данных. Целью эксперимента является анализ того, как влияет отсутствие гиппокампа на поведение. Гиппокамп — один из древнейших отделов головного мозга млекопитающих, его функции связывают с механизмами работы памяти , обучением, пространственной навигацией.

Особи были разделены на 5 групп.

1) Контрольная группа, содержащая разметку поведения здоровых мышей. 12 особей.
2) Гиппокампальная группа. Гиппокамп этих животных разрушали путем введения в эту структуру лидокаина, растворенного в искусственной спинномозговой жидкости (2 мкл. 4% раствора). 12 особей.
3) Шумовая группа, с параметрами частоты и продолжительности актов первой (контрольной) группы. 12 «особей».
4) Шумовая группа, с параметрами частоты и продолжительности актов второй (гиппокампальной) группы. 12 «особей».
5) Искусственные данные, содержащие один модельный P-Паттерн. 7 «особей».

Поведение каждой особи было представлено временным рядом длины $\sim$ 12 минут, всего было 24 детектируемых поведенческих актов, более подробно условия эксперимента описаны в [1]. Требовалось решить задачу классификации особи по поведению.

Далее, данные разбивались на обучение(N_l объектов) и контроль(N_c объектов). $N_l + N_c = 55$. Каждая особь описывалась вектором длины N_l, i-ое значение которого равняется количеству паттернов данной особи, также найденное в поведении i-ой особи из обучающей выборки.

На рис. 2 видно, что предложенный метод поиска P-Паттернов дает заметно лучшее качество классификации, чем метод поиска T-Паттернов. Наивная классификация на основе описания частот и средней продолжительности актов неприменима в данном эксперименте из-за двух шумовых классов.

Характерные паттерны присущие определенному классу также могут быть выделены в экспериментальных данных. Неформально, паттерн является характерным для заданного класса, если он присутствует в поведении многих особей этого класса и редко выявляется в поведении животных из других классов. Пример характерного для контрольной группы P-паттерна(в квадратных скобках указаны смещения и дисперсии в секундах между событиями):
«Вылизывание гениталий» [2,5; 6,2] «Вылизывание ладоней» [1,2; 7,9] «Умывание головы с ушами» [0,4; 5,7] «Вылизывание задних конечностей» [2,6; 7,9] «Умывание носа».

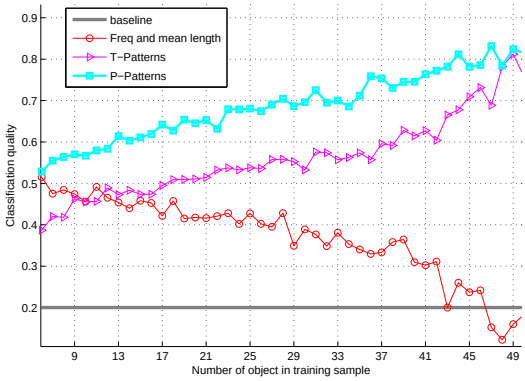

Рис. 2. Качество — средняя доля правильных классификаций. По горизонтали откладывалось количество объектов в обучении(качество усреднено по ста случайным разбиениям). Классификации по P- и T- паттернам производилась с помощью SVM. Классификации на основе частот и длин актов производилась с помощью решающих лесов.

Выводы

Представленный метод решает поставленные перед ним задачи и производит качественный поиск закономерностей как в синтетических временны́х рядах, так и в реальных поведенческих данных. В открытом доступе свободная, документированная, параллельная реализация представленного метода.

Главным преимуществом метода поиска P-Паттернов является их вариабельность: если на то есть предпосылки, то P-Паттерны найденные в поведении одной особи будут также найдены в поведении другой особи. Данный факт позволяет описывать поведение на основе найденных паттернов, и использовать стандартные алгоритмы машинного обучения для решения, например, задач классификации, кластеризации, или восстановления регрессии.

Литература

[1] В.В. Вишневский. «Параллельная реализация метода поиска закономерностей в последовательностях событий. — Дипломная работа. ВМиК МГУ, 2011.

[2] M.S. Magnusson. Discovering hidden time patterns in behavior: T-patterns and their detection. — Behavior Research Methods, Instruments, Computers 2000.

[3] P. Martin, P. Bateson. Measuring Behaviour: An Introductory Guide. — Cambridge University Press, second edition, 1993.

[4] R. Stoop, B. Arthur. Periodic orbit analysis demonstrates genetic constraints, variability, and switching in Drosophila courtship behavior. — Chaos – 2008 –vol.18/2.