**2.1 Разработка методов выделения строгой стереотипии в поведении**

**2.1.1 Обзор реализованного метода поиска Т-Паттернов**

Был реализован и доработан алгоритм поиска Т-Паттернов, предложенный М.С.Магнуссоном.

Изначальным для нашего метода, является понятие поведенческого акта(события). Исследователь определяется множество поведенческих актов, которые его интересуют и, впоследствии, будут фиксироваться в экспериментах.

Исходные данные представляют собой последовательность пар: время и тип поведенческого акта, который имел начало в этот момент времени. Таким образом, каждому событию сопоставляется определенное множество моментов времени. Далее, для удобства, поведенческие акты будут обозначаться следующим образом A, B, C, D… Общую продолжительность наблюдений будем обозначать как Nt.

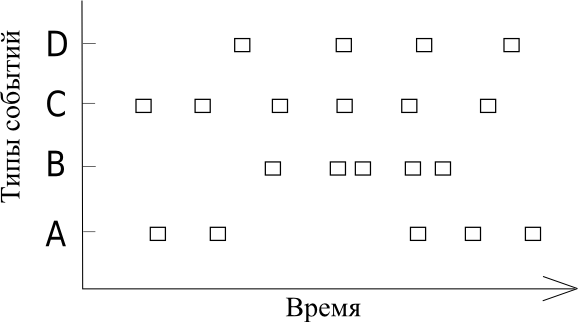


Рис 2.1.1 Пример графического изображения исходных данных.

Стоит отметить, что мы имели дело с исходными данными, в которых каждое следующее событие обозначало конец предыдущего поведенческого акта. Существует и иной способ кодирования поведенческих данных с сохранением информации о продолжительности отдельных актов: для выделенного множества интересующих поведенческих актов A, B, C… создать по 2 события, определяющих начало и конец поведенческого акта, A\_starts, A\_ends, B\_starts, B\_ends, и т.д. Описанный далее алгоритм способен одинаково обрабатывать данные, закодированные обоими способами.

Говоря о поведенческих закономерностях, следует определить более формально, в математических терминах, структуру, которую мы пытаемся найти. Однако сам способ задания исходных данных концептуально определяет понятие паттерна: паттерн – это упорядоченная последовательность событий, где каждые два последовательные события соединены относительно инвариантным временным промежутком. Более формально, понятие Т-Паттерна будет раскрыто далее в этой работе. Важно, что сама потребность более формального задания паттернов обусловлена необходимостью подсчета математических оценок для *процесса поиска* паттернов в исходных данных.

Магнуссон в своей работе [1], определил следующий тип паттернов, поиск которых осуществляет алгоритм. Т-Паттерн – это упорядоченная последовательность поведенческих актов, соединенных фиксированными, инвариантными во времени, критическими интервалами. Далее, два события, или псевдопаттерна, связаны отношением критического интервала(например, A[dL, dR]B ), если, после появления события A в момент времени t, интервал [t+dL,t+dR], (0≤dL ≤dR), содержит событие B, чаще, чем это ожидается из предположения о независимости событий(см. Рис. 2.1.2). Таким образом определенные структуры искомых паттернов адекватно описывают интересующие нас биологические закономерности(более подробно см. [1]).

Сама процедура поиска паттернов заключается в итеративном повторении двух шагов:

*Шаг 1. Конструирование новых паттернов*

Для каждых двух элементов из текущего множества паттернов, определить связывающий их критический интервал. Если такой существует, то добавить новый паттерн в текущее множество паттернов.

*Шаг 2. Удаление неполных копий и дубликатов*

Для каждых двух элементов из текущего множества паттернов проверить, не является ли один из элементов неполной копией или дубликатом другого. Если это так, то такой паттерн должен быть отброшен.

Данный процесс повторяется, пока множество текущих паттернов не перестанет изменяться. Результатом работы алгоритма и будет это множество. Более подробное описание данной итеративной процедуры может быть найдено в [1] или [2].

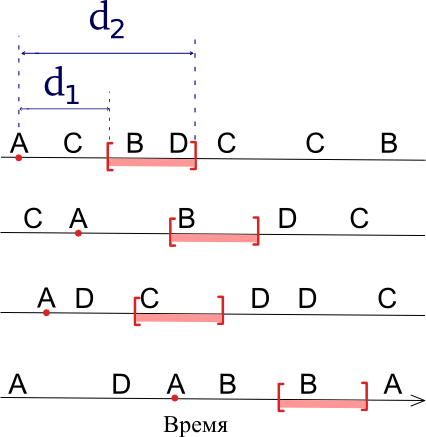


Рис.2.1.2. Пример критической связи A[d1, d2]B

**2.1.2 Параметры метода**

Реализованный алгоритм имеет большее число настраиваемых параметров, чем исходный метод Магнуссона. Сразу отметим, что устанавливая всем параметрам их значения по умолчанию, мы получим алгоритм, аналогичный Магнуссоновскому, то есть новые настраиваемые параметры не усложняют задачу экспериментатору. Выбор значений параметров должен основываться на априорных данных о специфике наблюдаемых процессов и ожидаемых результатов.

**Уровни значимости**

Данные параметры являются основными параметрами метода и являются так же настраиваемыми в методе Магнуссона. Первый из них – минимальны уровень значимости α, задает минимальный уровень неслучайности и выраженности паттерна(более формально см. выше). Чем меньше значение этого параметра, тем более ярко выраженные паттерны будут найдены. Вторым параметром(Nmin) является минимальное количество появлений паттерна в исходных данных. Эти параметра в паре могут задавать уровни значимости для паттернов разной длины, когда для выделения паттернов разной длины, используются различный набор этих параметров. Например, зачастую, разумно более строго выделять короткие паттерны, а длинным паттернам позволять быть немного размытым. То есть использовать примерно следующий набор параметров: α=0.0001 Nmin=5 для паттернов длины от 2 до 4, α=0.05 Nmin=3 для паттернов длины 5 и более.

**Запрет на повторяющиеся события**

Если при поиске поведенческих закономерностей известно, что в один и тот же паттерн не могут входить два одинаковых события, то существует возможность сообщить это алгоритму. Это условие позволяет отбросить множество лишних паттернов, и найти более длинные и ценные закономерности.

**Стратегия выбора критического интервала**

Во время поиска связи критического интервала A[dL, dR]B между двумя паттернами, вообще говоря, проверяются все возможные интервалы [dL, dR]. На практике, не является редкостью случай, когда для двух паттернов существует несколько пар dL и dR удовлетворяющих отношению критического интервала. Для выбора конкретных значений dL и dR, предлагается использовать одну из нижеописанных стратегий. Для каждой стратегии представлен результат работы алгоритма на тестовых данных. Данные содержат паттерн длины 7, встречающийся 8 раз.

*Выбор кратчайшего критического интервала:*

При использовании данной стратегии, на каждом шаге выбирается критический интервал, имеющий наименьшую длину d=dR-dL+1. Такой подход позволяет уменьшить длину критических связей, тем самым выявляя более выраженные и “стройные” паттерны. Одним из недостатков данного метода является эффект “расщепления”. Поясним данный эффект на примере: пусть в исходных данных существует критическая связь A[4, 20]B, наблюдаемая 15 раз. Данная связь настолько ярко выражена, что алгоритму не требуется подбирать границы критического интервала, чтобы B появлялось после A все 15 раз. Алгоритм, скорее, выделит два(или даже больше) критических интервала: A[4, 12]B, наблюдаемый 6 раз, и A[13, 18]B, наблюдаемый 7 раз. Таким образом, один ярко выраженный паттерн будет распадаться на несколько более редких, что в свою очередь, может помешать дальнейшему выявлению закономерностей. Ниже описанный метод позволяет избавиться от такой проблемы.

*Выбор самого длинного критического интервала:*

Как следует из названия, при использовании этой стратегии, среди всех значимых dL и dR, выбирается интервал, соответствующий наибольшему значению длины d. В результате, некоторые паттерны “загрубляются”, и алгоритм стремится найти максимальное количество вхождений каждого критического интервала.

*Выбор самого значимого критического интервала:*

При этом подходе выбирается критический интервал, имеющий максимальный уровень значимости, тем самым выделяя самые значимые связки.

В общем случае, нет каких-либо рекомендаций по выбору стратегии поиска критического интервала. По умолчанию, как и в методе Магнуссона, используется стратегия выбора самого длинного интервала. Можно отметить, что если данные не зашумлены и паттерны достаточно ярко выражены, то лучше использовать стратегию выбора самого значимого критического интервала.

|  |  |  |
| --- | --- | --- |
| **Стратегия выбора интервала** | **Найдено паттернов** | **Найдено вхождений целевого паттерна** |
| Кратчайший | 40 |  |
| Самый длинный | 5 | 8 |
| Самый значимый | 7 | 7 |

Табл. 2.1.1. Сравнение стратегий выбора интервала на экспериментальных данных

**2.1.3 Эксперименты и сравнение реализации с программой THEME**

Было произведено сравнение результатов работы реализованного нами алгоритма с оригинальной реализацией метода поиска Т-Паттернов в программе THEME. Исходными данными для экспериментов были записи груминга хомячков.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Найдено паттернов, содержащих более 3-х событий** | | | |
| **Эксп.№1** | **Эксп.№2** | **Эксп.№3** | **Эксп.№4** |
| **THEME** | 8 | 16 | 14 | 21 |
| **Наш алгоритм** | 12(7) | 14(14) | 17(13) | 18(17) |

Табл. 2.1.2. Сравнение результатов работы алгоритмов на реальных данных. В скобках указано количество паттернов, соответствующих паттернам, которые выделила программа THEME.

Как видно из Таб. 2.1.2, реализованный нами метод, справляется с поставленной задачей. У рассмотренных методов, множества найденных паттернов, содержащих более трех поведенческих актов, имеют значительное пересечение. Более того, для самых длинных паттернов(7-8 событий, такие паттерны представляют наибольший интерес для исследователей), множества найденных закономерностей почти полностью совпадали и иногда наш метод выделял более сложные закономерности, расширяя результаты работы THEME. Благодаря возможности более тонкой настройки алгоритма, реализованный нами метод, превосходит метод поиска паттернов программы THEME в производительности.

**2.1.4 Детали реализации метода**

Алгоритм был реализован на языке Си без использования сторонних библиотек. Для более удобного интерфейса работы с алгоритмом был создан интерфейс для среды MATLAB.

Matlab-реализация алгоритма состоит из следующих файлов:

**mexPattern.mex:** откомпилированный mex-файл, реализующий алгоритм поиска паттернов.

**mexPattern.m:** объявление mex-функции.

**T\_DRAW\_PATTERNS.m:** графический вывод найденных паттернов.

**T\_GENERATE\_PATTERNS.m:** создание искусственных паттернов во временных рядах.

**T\_LOAD\_FILE.m:** загрузка временного ряда из файла для дальнейшей работы с ним.

**T\_STAT\_VALIDATE.m:** процедура статистической валидации.

**test.m:** пример использования модуля.

**2.1.5 Список используемой литературы**

[1] Magnus Magnusson (2000) «Discovering hidden time patterns in behavior: Tpatterns and their detection» University of Iceland.

[2] Anolli, L (2005) «The Detection of the Hidden Design of Meaning. In The Hidden Structure of Interaction: From Neurons to Culture Patterns»

[3] Dirk-Jan Povel. «Internal Representation of Simple Temporal Patterns».

.