ПОИСК СЛОЖНЫХ НЕПЕРИОДИЧЕСКИХ ШАБЛОНОВ В ПОСЛЕДОВАТЕЛЬНОСТЯХ ЧИСЕЛ И СИМВОЛОВ МЕТОДАМИ ЛОКАЛЬНОЙ ГЕОМЕТРИИ

В. А. Дюк

Санкт-Петербургский институт информатики и автоматизации РАН
199178, Санкт-Петербург, 14-я линия В.О., д.39
v_duke@spiiras.nw.ru

УДК 681.3.01
В. А. Дюк. **Поиск сложных непериодических шаблонов в последовательности чисел и символов методами локальной геометрии** // Труды СПИИРАН. Вып. 1, т. 2 — СПб: СПИИРАН, 2002.
Аннотация: *Приводится краткое описание новой технологии обнаружение логических закономерностей на базе представлений локальной геометрии. Описываются особенности ее применения для поиска сложных непериодических шаблонов в последовательностях чисел и символов. Приводятся примеры.* — Библ. 7 назв.

UDC 681.3.01
V.A. Duke. **Search of complex acyclic patterns in a sequence of numbers and symbols by methods of local geometry** // SPIIRAS Proceedings. Issue 1, v. 2. — SPb: SPIIRAS, 2002.
Abstract: *The brief description of new technology detection of logic laws is resulted on the basis of representations of local geometry. The features of its application for search of complex acyclic patterns in sequences of numbers and symbols are described. The examples are resulted.* — Bibl.7 items.

Методы анализа последовательностей — временных или иных рядов чисел и символов — в настоящее время испытывают определенные затруднения. Специалисты отмечают, что несколько основных моделей, используемых при таком анализе, оказались плохо совместимыми друг с другом по базовым посылкам. Например, для числовых рядов Фурье-анализ требует отсутствия непериодических составляющих, методы Бокса чувствительны к виду одномерных распределений и т.д. Алгоритмы поиска закономерностей в последовательностях символов основываются на переборах, которые можно реализовать только в очень ограниченных вариантах, либо опираются на сильные эвристические допущения.

Продуктивным направлением анализа временных рядов сегодня является подход, связанный с преобразованием временного ряда в матрицу с помощью однопараметрической сдвиговой процедуры «Гусеница» [1]. Этот подход независимо разрабатывался в России (Санкт-Петербург, Москва) и США (там его аналог получил название SSA — Singular Spectrum Analysis) и показал себя мощным средством исследования временных рядов (в основном в метеорологии, гидрологии, климатологии). Алгоритм преобразования временного ряда в матрицу данных состоит в следующем.

Анализу подвергается временной ряд $\{x_i\}_{i=1}^{N}$, образованный последовательностью N равноотстоящих значений некоторой (возможно, случайной) функции $f(t)$:

$$x_i = f((i-1)\Delta t), \text{ где } i = 1, 2, \ldots, N.$$

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Выбирают некоторое число $M<N$, называемое *длиной гусеницы*, и первые M значений последовательности f представляют в качестве первой строки матрицы $\boldsymbol{X}$. В качестве второй строки матрицы берут значения последовательности с x_2 по x_{M+1}. Последнюю строку с номером $k = N - M + 1$ составляют последние M элементов последовательности.

Построенную матрицу, элементы которой равны $x_{ij} = x_{i+j-1}$, можно рассматривать как M-мерную выборку объема k или M-мерный временной ряд, которому соответствует M-мерная траектория (ломаная в M-мерном пространстве из k-1 звена. Матрица $\boldsymbol{X}$ (ее называют матрицей ряда) представлена в традиционном для прикладной статистики виде «строка — объект, столбец — признак». Для ее дальнейшей обработки теперь можно применять различные методы из богатого арсенала математического аппарата многомерного анализа.

Хорошо разработанным является исследование матрицы с помощью анализа главных компонент. Результатом такого исследования служит разложение временного ряда на простые компоненты: медленные тренды, сезонные и другие периодические или колебательные составляющие, а также шумовые компоненты.

Сегодня для анализа закономерностей временного ряда все чаще стали применяться методы Data Mining, предназначенные для обнаружения различных шаблонов (паттернов) во временном ряде [2, 3]. При этом особую ценность в обнаружении таких шаблонов имеют логические методы. Эти методы позволяют находить логические if-then правила1, характерные для строк M-мерной временной матрицы и не характерные для случайно сгенерированной M-мерной матрицы данных. Они пригодны для анализа как числовых, так и символьных последовательностей, и их результаты имеют прозрачную интерпретацию.

Вместе с тем, при выборе того или иного метода поиска следует опираться на критерий, отражающий его способность выявлять *наиболее полные и точные if-then правила для каждой строчки* временной матрицы за приемлемое время. К сожалению, известные методы в слабой степени способны удовлетворять этому критерию.

Так, деревья решений принципиально не способны находить «лучшие» комбинации в данных. Они реализуют наивный принцип последовательного просмотра признаков и выявляют фактически «осколки» настоящих закономерностей, создавая лишь иллюзию логического вывода.

Критерий отбора «хромосом» в генетических алгоритмах и используемые процедуры являются эвристическими и далеко не гарантируют нахождения «лучшего» решения. Как и в реальной жизни, эволюцию может «заклинить» на какой-либо непродуктивной ветви. И, наоборот, можно привести примеры, как два неперспективных родителя, которые будут исключены из эволюции генетическим алгоритмом, оказываются способными произвести высокоэффективного

---

[1] Каждую строчку матрицы данных можно «покрыть» множеством различных правил вида:

$$\textbf{IF}\ \underbrace{\text{(условие 1) и (условие 2) и ... (условие N)}}_{\textbf{A}}\ \textbf{THEN}\ \underbrace{\text{(условие M)}}_{\textbf{B}}$$

Примеры условий: $X = C_1$; $X < C_2$; $X > C_3$; $C_4 < X < C5$ и др., где X обозначает какой-либо столбец матрицы, C_I — константы. Точность правила — это доля случаев B среди случаев A. Полнота правила — это доля случаев A среди случаев B.

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

потомка. Это особенно становится заметно при решении высокоразмерных задач со сложными внутренними связями.

Трудоемкость переборных алгоритмов не нуждается в комментариях. Известные коммерческие системы (например, WizWhy) ограничиваются анализом комбинаций до 6-10 элементов.

Указанные недостатки усугубляются тем, что все рассмотренные алгоритмы совершают серьезную ошибку уже на первом шаге, когда происходит определение исходных элементарных событий на основании анализа отдельно взятых признаков.

Вместе с тем, эффективным для анализа закономерностей временного ряда оказалось применение технологии обнаружения логических закономерностей на основе представлений локальной геометрии [4-7]. За счет свойств локальных пространств процедура поиска логических закономерностей в данных имеет геометрическое истолкование. Перебор вариантов при поиске «лучших» if-then правил методами локальной геометрии практически отсутствует. Поиск осуществляется с помощью модифицированного аппарата линейной алгебры с применением средств интерактивной графики. Также важным моментов является возможность распараллеливания многих операций, лежащих в основе применяемых алгоритмов.

На рис. 1 приведена иллюстрация результатов поиска шаблона методами локальной геометрии в небольшом фрагменте ДНК e-coli (кишечной палочки). Этому фрагменту соответствует первая строка таблицы. Правая часть таблицы (выделенный прямоугольник), обозначенный как «область поиска», — *M*-мерная матрица ряда (*M* = 23). К этой части была также присоединена не показанная на рисунке случайно сгенерированная таблица такого же размера (в ней вероятности появления каждого из четырех символов A, C, T, G одинаковы и равны 0,25).

В рассмотренном примере во фрагменте ДНК удалось найти шаблон **A**X**AA**X**A**, где в позиции X может стоять любой из четырех символов A, C, T, G. Обращает на себя внимание то, что найденный шаблон появляется в последовательности ДНК через различные по длине интервалы (8, 8, 7, 8). Имеется ряд других более сложных примеров поиска непериодических шаблонов в ДНК и других последовательностях символов, которые не приводятся здесь из-за ограниченного объема статьи.

Следующий пример относятся к области анализа электроэнцефалограмм (ЭЭГ). В этом примере ставилась задача найти отличия в ЭЭГ биологических объектов, подвергавшихся определенным видам информационного воздействия, от ЭЭГ этих же биологических объектов в обычной информационной обстановке (фоновых ЭЭГ).

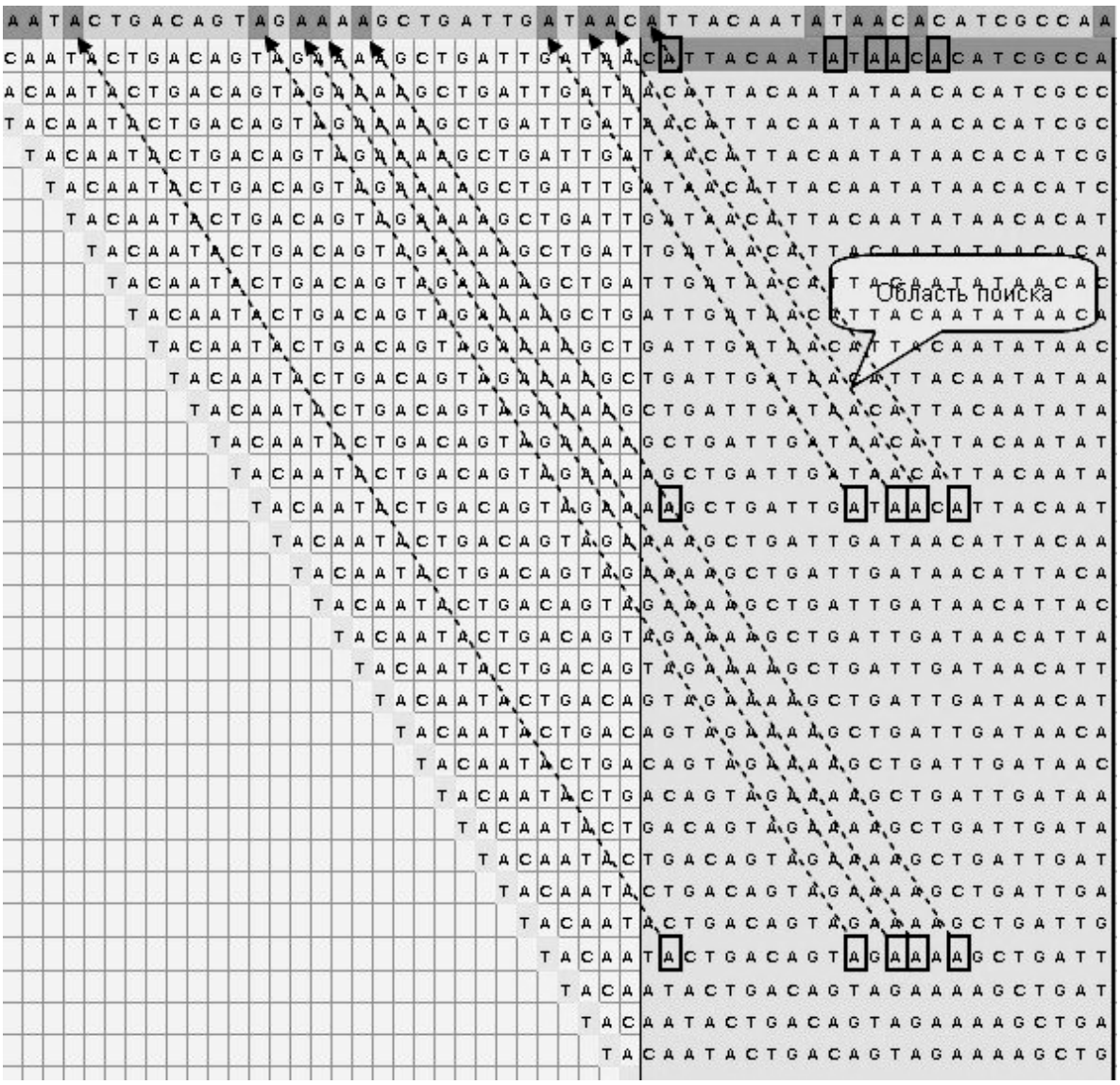
На первом этапе анализа ЭЭГ подвергались предобработке. Сначала производилось сглаживание ЭЭГ по методу скользящего среднего. Затем участки сигнала с положительной первой производной заменялись на 1, остальные на 0. Таким образом, вместо исходного сигнала обработке подвергалась последовательность, состоящая из 0 и 1.

На втором этапе производился анализ полученных последовательностей методами локальной геометрии. В результате применения описанной выше процедуры «Гусеница» и поиска логических закономерностей методами локальной геометрии были выявлены паттерны, характерные для ЭЭГ при информационном воздействии (такие паттерны не встречаются в фоновой ЭЭГ). Два из них представлены на рис. 2-3.

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Резюме

1. Поиск сложных непериодических шаблонов в последовательностях чисел и символов методами локальной геометрии представляет интерес для целого ряда областей, связанных с анализом временных и иных рядов в биологии, медицине, технике и экономике.
2. Особую ценность данные методы, по-видимому, имеют в современных молекулярно-генетических исследованиях, в которых наступил этап выяснения функционального смысла различных участков секвенированной ДНК.
3. Методы локальной геометрии продемонстрировали принципиальную возможность получения новых результатов при анализе электрофизиологических измерений.



Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Рис. 1. Методами локальной геометрии во фрагменте ДНК найден шаблон с изменяющимся периодом

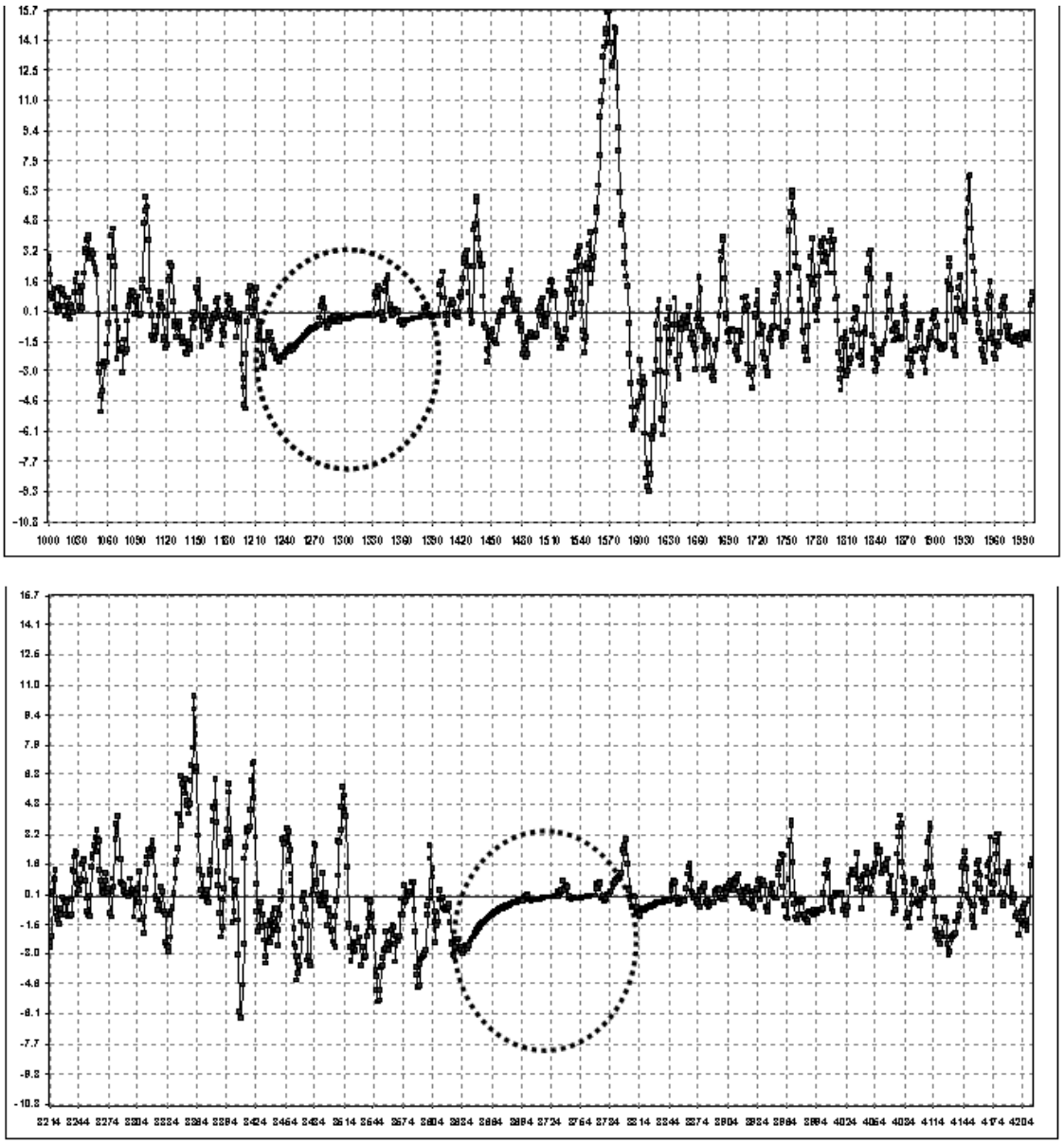


Рис. 2. Первый наиболее «сильный» паттерн в ЭЭГ

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

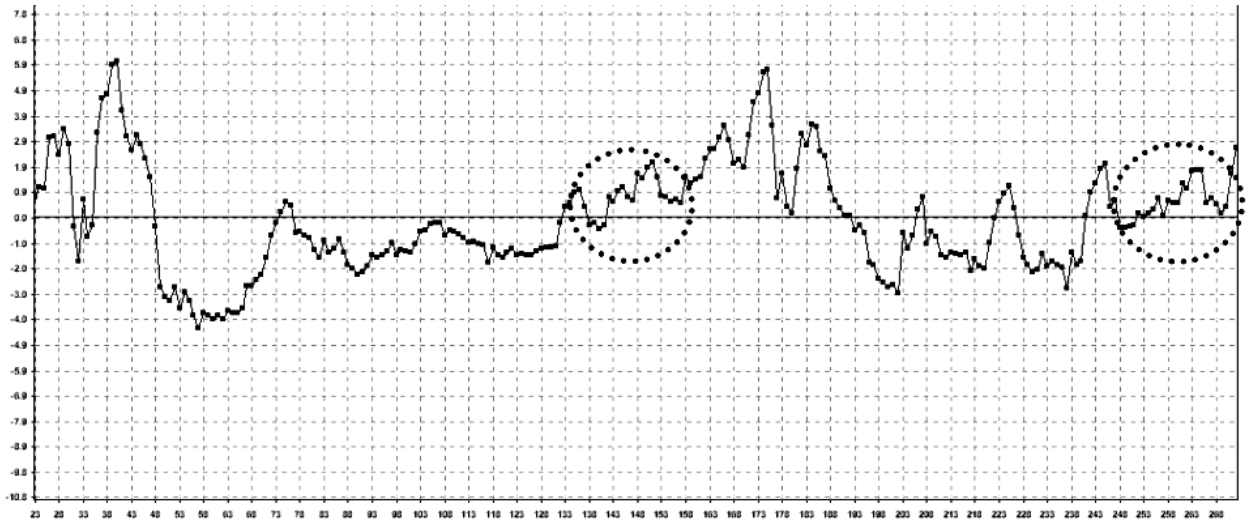


Рис. 3. Второй наиболее «сильный» паттерн в ЭЭГ

Литература

[1] Главные компоненты временных рядов: метод «Гусеница» (Под ред. Д. Л. Данилова и А. А. Жиглявского). — Санкт-Петербург Государственный университет, 1997.

[2] *Richard J. Povinelli.* Time series data mining: Identifying temporal patterns for characterization and prediction of time series events. — A Dissertation submitted of the Graduated School, Marguette University, in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy. — Milwaukee, Wisconsin. —- December, 1999.

[3] *Fayyad U. M., Piatetsky-Shapiro G., Smyth P., Uthrsamy R.* Advances in knowledge discovery and data mining. — Menlo Park, California: AAAI Press, 1996.

[4] *Дюк В. А*. Обработка данных на ПК в примерах. — СПб: "Питер", 1997.

[5] *Duke V. A*. Latent knowledge extraction by methods of local geometry: development of expert system for keen appendicitis diagnostics // Proc. Int. Conf. On Informatics and Control (ICI&C 97), St. Petersburg, Russia, 1997, vol. 2, p.p. 663–668.

[6] *Дюк В. А*. Формирование знаний в системах искусственного интеллекта: геометрический подход (ч. 4, глава 2) // Телемедицина. Новые информационные технологии на пороге XXI века. — СПб: «Анатолия», 1998. С. 367–389.

[7] *Дюк В. А.* От данных к знаниям — новые возможности обработки баз данных // Тр. Межд. науч. конф. «Интеллектуальные системы и информационные технологии управления (Псков, 19–23 июня 2000 г.). — СПб.: Изд-во СПбГТУ, 2000. — с. 438-440.