

УДК 004.42:004.67

ФОРМИРОВАНИЕ И КОМПАКТИФИКАЦИЯ ХРАНЕНИЯ ГЕОМЕТРИЧЕСКИХ ПАТТЕРНОВ ПСЕВДОПЕРИОДИЧЕСКИХ ДАННЫХ

И.Ю. Колдин, Л.И. Сучкова

Алтайский государственный технический университет им. И.И.Ползунова
г. Барнаул

Статья посвящена алгоритмам обработки данных мониторинга на базе геометрических паттернов. Рассматривается выделение периодических частей, формирование геометрических паттернов, хранение базы паттернов в структуре PGG.

Ключевые слова: обработка данных мониторинга, геометрический паттерн, структура PGG.

В настоящее время актуально развитие методов обработки и анализа измерительной информации с целью идентификации аварийных ситуаций и их предвестников [1-5].

Одним из подходов к формализации данных измерений является использование геометрических паттернов. Геометрическому паттерну визуально соответствует кривая зависимости измеряемых данных от времени, описываемая функцией, характеризующейся типом и набором коэффициентов, соответствующих типу. Пример геометрического паттерна приведен на рисунке 1.

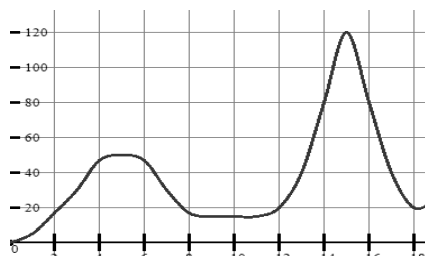


Рисунок 1 – Сформированный геометрический паттерн

Представляют интерес геометрические паттерны, описывающие периодические данные, повторяющиеся на промежутках времени – периодах, причем допускается как вариабельность данных внутри периода, так и вариабельность границ периода для различных временных интервалов [6-10].

Основными этапами при анализе данных измерений с применением геометрических

паттернов являются:

1. Сбор данных за некоторый промежуток времени.
2. Выявление в них закономерностей и построение геометрического паттерна.
3. Анализ паттерна на уникальность и обеспечение его компактного хранения.

Основным методом при выделении периодов из потока данных является разделение данных измерения, опираясь на впадины – точки минимума потока данных. Фиксирование таких точек будет определяться по установленному пользователем значению. В итоге границы периодов устанавливаются в тех точках минимума, которые меньше выставленного значения. Также предлагается автоматически обновлять это значение во время работы системы по формуле:

$$U = \frac{\alpha(\sum_{i=0}^N V_i)}{N}, \quad (1)$$

где U – значение границы разделения, N – число последних зафиксированных точек минимума, V – значение точки минимума, α – поправочный коэффициент, позволяет улучшить гибкость алгоритма. Значение α должно быть подобрано эмпирически, если оно слишком мало, то может генерироваться слишком много ложноположительных результатов, если оно слишком большое – ложноотрицательных. Пример работы данного алгоритма представлен на рисунке 2.

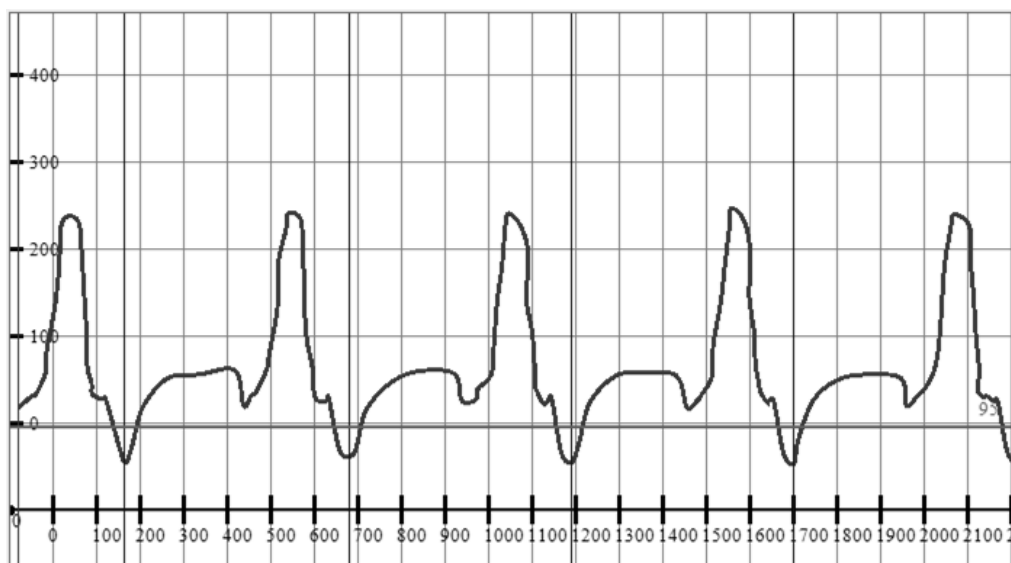


Рисунок 2 – Работа алгоритма разделения данных на периоды

После обработки данных мониторинга формируется набор выделенных периодов, для каждого из которых строится геометрический паттерн, представляющий собой усредненный шаблон. Для формирования паттерна используется кусочно-линейная аппроксимация, то есть сначала применяется алгоритм для оптимального разделения периода на сегменты, а после аппроксимируется каждая его часть по формуле прямой:

$$y = kx + b, \quad (2)$$

где k и b – ее коэффициенты. Для разделения периода на сегменты используется алго-

ритм раздвижного окна. Алгоритм последовательно соединяет точки в один сегмент и описывает его с помощью кусочно-линейной аппроксимации. При каждом добавлении точки в сегмент рассчитывается остаточная ошибка, с помощью которой определяется, когда нужно закончить один сегмент и начать новый. Результаты работы данного метода при разных значениях допустимой погрешности представлены на рисунке 3, где Data – выделенный период из потока данных, PLR – сформированный паттерн с помощью кусочно-линейной аппроксимации.

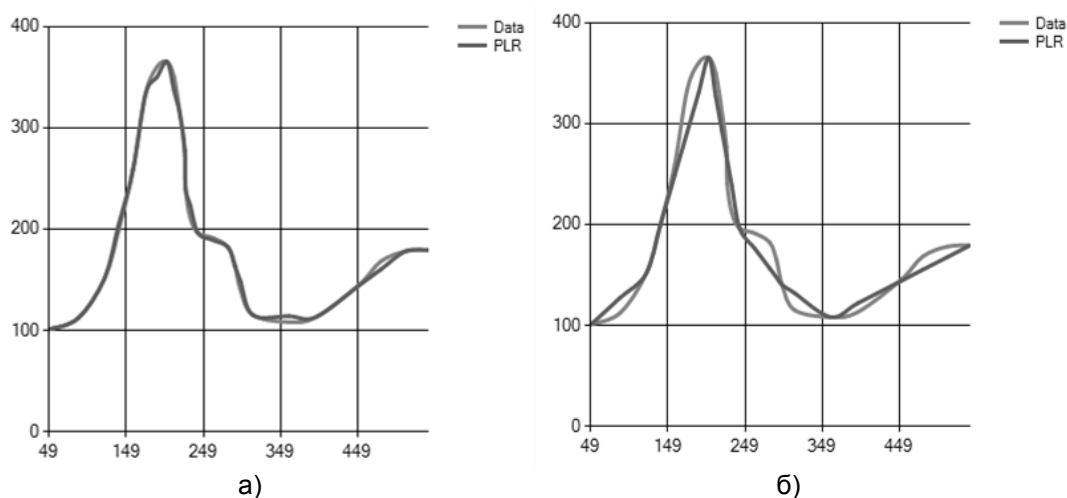


Рисунок 3 – Формирование геометрического паттерна с помощью кусочно-линейной аппроксимации: а) допустимая погрешность 5%, б) допустимая погрешность 14%

В итоге паттерн представляется в виде

последовательности сегментов, для каждого

ФОРМИРОВАНИЕ И КОМПАКТИФИКАЦИЯ ХРАНЕНИЯ ГЕОМЕТРИЧЕСКИХ ПАТТЕРНОВ ПСЕВДОПЕРИОДИЧЕСКИХ ДАННЫХ

из которых хранятся аппроксимирующие коэффициенты k и b , границы сегмента и допустимая погрешность. Проведенными экспериментальными исследованиями установлено, что данный способ обеспечивает сжатие объема анализируемых данных и существенный выигрыш в скорости обработки, что очень важно при работе системы в реальном времени.

Для хранения геометрических паттернов предлагается использовать модификацию специальной структуры PGG (Pattern Growth Graph) [11]. PGG представляет собой особый контейнер, где каждый шаблон хранится в виде двунаправленного связанного списка с заголовочным узлом, в который записывается идентификатор шаблона, частота и время каждого его появления. Остальные узлы списка содержат информацию о его сегментах.

Опишем алгоритм проверки уникальности нового геометрического паттерна, сформированного путем обработки потока данных, с учетом хранения паттерна с использованием Pattern Growth Graph. Пусть на входе имеем полученный из потока данных период W и сформированный геометрический паттерн P' . Чтобы сравнить его с уже имеющимся в базе паттерном P , нужно вычислить погрешность $E1$ формирования паттерна по калибровочным данным и погрешность $E2$, равную среднему арифметическому значений погрешностей, полученных при сравнении сегментов паттернов P' и P . Если $(E1 - E2)/2$ меньше заранее установленного значения, то фиксируем совпадение паттернов. Иначе сравниваем данные периода W с паттерном P напрямую.

Для каждой точки периода W выполняем следующие действия. Пусть W_{xi} – время получения данных в i -той точке периода, W_{yi} – значение в i -той точке. Определяем сегмент S из паттерна P , в который попадает W_{xi} , затем в полученном сегменте находим значение S_y , совпадающее по времени с W_{xi} , и определяем погрешность E_i при сравнении W_{yi} и S_y по формуле:

$$E_i = \frac{W_{yi} - S_y}{W_{yi}}. \quad (3)$$

Общая погрешность вычисляется как среднее арифметическое из всех найденных погрешностей E_i . Блок-схема данного алгоритма представлена на рисунке 4.

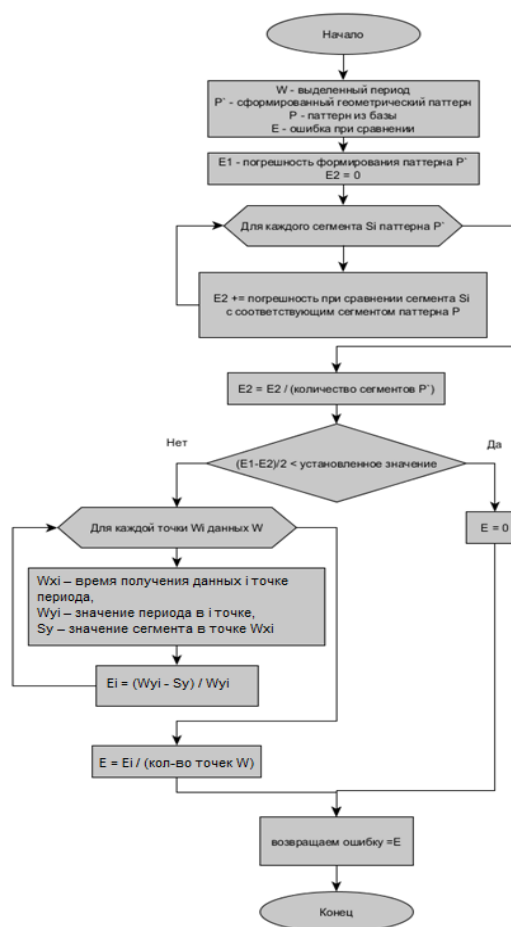


Рисунок 4 – Блок-схема алгоритма сравнения геометрических паттернов

При добавлении нового геометрического паттерна в PGG происходит его сопоставление с каждым паттерном в базе по вышеописанному алгоритму. После всех сравнений получаем один из трех возможных результатов и выполняем соответствующие ему действия:

1. Полное совпадение. В одном из сопоставлений алгоритм возвратил погрешность, равную нулю. В данном случае нет необходимости генерировать новый паттерн, поэтому только увеличиваем частоту и время появления соответствующего ему паттерна из PGG.

2. Частично соответствует. При одном из сопоставлений алгоритм возвратил погрешность, которая меньше установленной. В таком случае совпавшие части будут нетронуты. Фиксируются лишь новые непохожие сегменты в виде разветвлений в узлах списка. Новый паттерн является наследником старого, и данная процедура соответствует росту паттерна.

3. Несоответствие. При всех сопоставлениях погрешность оказалась выше установленной. Не найдено ни одного даже частично похожего из базы, а значит, сформированный паттерн будет добавлен в PGG как новый.

Такой подход обработки потока данных позволяет компактно хранить и быстро использовать шаблоны для сравнения в режиме реального времени, а также хранить его историю изменений.

Применение новых алгоритмов обработки данных мониторинга, основанных на модификации PGG, позволяет охарактеризовать особенности данных измерений за определенный промежуток времени, помочь отличить значащие изменения данных от шума, а также обеспечить восстановление, обнаружение изменений потока данных, фиксирование аномальных ситуаций.

СПИСОК ЛИТЕРАТУРЫ

1. Клионский Д.М. Методы выявления аномальных событий в многокомпонентных измерительных сигналах на основе мультимасштабных и спектральных методов высокого разрешения. [Текст]: Автореф. дис. канд. техн. наук. — Санкт-Петербург: СПбГУ «ЛЭТИ» им. В.И. Ульянова (Ленина), 2012. — 18 с.
2. Марчук В.И., Токарева С.В. Способы обнаружения аномальных значений при анализе нестационарных случайных процессов// Монография. ГОУ ВПО «Южно-российский государственный университет экономики и сервиса», 2009. — 60 с.
3. Колдин И.Ю., Сучкова Л.И. Обработка периодических данных мониторинга с применением геометрического паттерна// И.Ю. Колдин, Л.И. Сучкова // Ползуновский альманах. — 2016. — №2. Режим доступа: http://elibrary.altstu.ru/elibrary/books/Files/pa2016_02/pdf/029_koldin.pdf.
4. Орешко Н.И., Гепленер В.В., Клионский Д.М. Применение гармонических вейвлетов в задачах

обработки осциллирующих сигналов // Цифровая Обработка Сигналов, № 2, 2012, С. 6-14.

5. Сучкова Л.И. Подход к прогнозированию нештатных ситуаций в системах мониторинга с использованием паттернов поведения группы временных рядов// Л.И. Сучкова // Ползуновский вестник. — 2013. — №2. Режим доступа: http://elibrary.altstu.ru/elibrary/books/Files/pv2013_02/pdf/088_suchkova.pdf.

Minos N. Garofalakis, Rajeev Rastogi, Kyuseok Shim. Data Mining and the Web: Past, Present and Future / VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases. — 1994. — С. 487-499.

6. Rakesh Agrawal, Ramakrishnan Srikant Fast Algorithms for Mining Association Rules / WIDM '99 Proceedings of the 2nd international workshop on Web information and data management. — 1999. — С. 43-47.

7. Sheng Ma and Joseph L. Hellerstein. «Mining Partially Periodic Event Patterns With Unknown Periods»// International Conference on Data Engineering. 2000.

8. Faraz Rasheed, Mohammed Alshalalfa, and Reda Alhajj, Associate Member, IEEE. «Efficient Periodicity Mining in Time Series Databases Using Suffix Trees»// IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 1, JANUARY 2011.

9. Johannes Assfalg, Thomas Bernecker, Hans-Peter Kriegel, Peer Kroger, Matthias Renz. «Periodic Pattern Analysis in Time Series Databases»// 14th International Conference, DASFAA'09, Brisbane, Australia, pp. 354-368, 2009.

10. L Tang, B Cui, H Li, G Miao, D Yang, X Zhou «Effective variation management for pseudo periodical streams» // Proceedings of the 2007 ACM SIGMOD international conference on Management of data, Brisbane, Australia, pp. 257-268, 2007.

Иван Юрьевич Колдин — магистрант, e-mail: vanya-koldin@yandex.ru; Лариса Иннокентьевна Сучкова — д.т.н., профессор, заместитель зав. кафедрой информатики, вычислительной техники и информационной безопасности, e-mail: lara8370@yandex.ru.