Московский физико-технический институт
(государственный университет)
Кафедра интеллектуального анализа данных

Работа допущена к защите
зав. кафедрой

_____ Рудаков К.В.

«____» _____ 2014 г.

Выпускная квалификационная работа на степень бакалавра

Тема: **Проблема понижения размерности в задаче поиска аномалий в многомерных временных рядах**

Направление: 010900 – Прикладные математика и физика

Выполнил студент гр. 074 _____ Яшков Д.Д.

Научный руководитель,
д. ф.-м. н. _____ Воронцов К.В.

Москва – 2014

Оглавление

1. Введение 3
2. Обзор литературы 4
3. Постановка задачи 5
4. Понижение размерности по времени 6
   4.1. Дискретизация временных рядов 6
   4.2. Сегментация 7
   4.3. Кластеризация сегментов 9
5. Понижение размерности по компонентам многомерного ряда 13
6. Вычислительный эксперимент 15
7. Заключение 18

Список литературы 20

1. Введение

В данной работе рассматривается задача поиска аномалий в объектах, представленных многомерными временными рядами. Главной особенностью данной задачи является большой объем входных данных, а также отсутствие формального определения аномалии.

Предлагается способ уменьшения размерности входных данных, путем сведения её к задаче поиска аномалий в одномерных дискретных последовательностях. Это позволит применить большое число алгоритмов, работающих с дискретными последовательностями [1]. Мы рассматриваем общий случай, считая что исходные временные ряды могут быть как дискретными, так и непрерывными. Предлагается дискретизовать непрерывные ряды с помощью перевода в символьное представление [2], а дискретные оставить без изменений, после чего разбить исходные данные на участки однородности и кластеризовать их. Конечное представление исходных данных получается путем замены участков однородности на метки соответствующих кластеров.

Данный метод можно применять к любым входным данным, которые удовлетворяют следующим предположениям, использованным в данном подходе:

- ряды описывают один и тот же процесс, происходящий с различными объектами, например: показания датчиков в течение различных полетов одного и того же самолета;
- время измерения и число временных рядов, описывающих объекты, велико;
- объекты можно разбить на участки однородности;
- аномалии – маловероятные события как внутри объектов, так и на множестве объектов.

Главными преимуществами предлагаемого алгоритма является значительное уменьшение размерности входных данных и локализация аномалий внутри участков однородности.

2. Обзор литературы

Задача поиска аномалий в сдномерных временных рядах уже рассматривалась ранее, однако в большинстве случаев, вводилась метрика между объектами и строилась матрица попарных расстояний между объектами. После этого эти объекты кластеризовались. В большинстве работ объекты представлялись дискретными последовательностями [3–5].

Для многомерных рядов известно гораздо меньше алгоритмов. Опишем их основные идеи.

- Алгоритм MKAD [6]. Используется обобщение на многомерный случай метрики между последовательностями, основанной на $nLCS$ – наибольшей общей подпоследовательности. Строится матрица попарных расстояний, после чего объекты кластеризуются.

- Данные – многомерные бинарные временные ряды. Делается кластеризация каждого объекта по времени методом фон Мизеса-Фишера [7]. Таким образом сводят каждый объект к одномерному дискретному ряду.

- Данные – многомерные непрерывные. Настраивается модель прогнозирования временных рядов VARIMA [8] и моменты, когда у нее большие остатки считаются аномальными.

- Данные – многомерные непрерывные. Используется метод независимых компонент (ICA) [9]. Аномальные моменты оказываются в первой компоненте.

С разнотипными многомерными рядами работает только один из этих алгоритмов – MKAD. Он кластеризует объекты, с помощью одноклассового метода опорных векторов.

Задача понижения размерности была рассмотрена только в [7], где данными являлись многомерные бинарные ряды.

3. Постановка задачи

В данной работе рассматривается задача поиска аномалий в многомерных временных рядах. Исходными данными этой задачи является множество объектов $X = \{X^i_{jt}\}$, $i = 1, \ldots, N$, где каждый объект X^i описывается множеством временных рядов $j = 1, \ldots, J$ в течение некоторого времени $t = 1, \ldots, T_i$.Эти временные ряды можно разбить на два типа: непрерывные J_c и дискретные J_d, то есть:

$$X^i_{jt} \in \mathbb{R}, \quad \text{где } j \in J_c;$$
$$X^i_{jt} \in \Sigma_j, \quad \text{где } j \in J_d.$$

Здесь Σ_j – конечный алфавит j-го временного ряда.
Опуская некоторые из индексов, будем получать нужные срезы, например: X_j – совокупность j-х временных рядов всех объектов.

В работе использовалось несколько предположений:

- ряды описывают один и тот же процесс, происходящий с различными объектами, например: показания датчиков в течение различных полетов одного и того же самолета;
- время измерения и число временных рядов описывающих объекты велико;
- каждый объект можно разбить на последовательные участки однородности – сегменты;
- аномалии – маловероятные события как внутри объектов, так и на множестве объектов.

Требуется предложить преобразование исходных данных, уменьшающее размерность пространства (N, J, T), не теряя информации об аномалиях.

Поскольку уменьшать число объектов не имеет большого смысла, данную задачу можно разбить на две подзадачи:

1. понижение размерности по времени;
2. понижение размерности по компонентам многомерного ряда.

4. Понижение размерности по времени

4.1. Дискретизация временных рядов

Хочется работать с одним типом рядов, поэтому предлагается дискретизовать непрерывные временные ряды. Для этого используется аналог символьного агрегированного представления (SAX), описанного в [2, 10]. В этом алгоритме строится эмпирическая функция распределения, её область определения разбивается на интервалы, концами которых являются выборочные квантили. Значения, лежащие между соседними квантилями, заменяются на соответствующие буквы из алфавита, как показано на рисунке 1.

Алгоритм 1 Дискретизация непрерывного временного ряда.

Вход: X_j, где $j \in J_c$; – совокупность значений j-го временного ряда всех объектов

n – размер алфавита в который будет преобразован ряд;

$\{p_k\}_{k=0}^{n}$ – вероятности для квантилей, $p_0 = 0$, $p_n = 1$.

Выход: множество преобразованных рядов $\{X_j^i\}_{i=1}^{N}$: $X_{jt}^i \in (a_1, \ldots, a_n)$.

1: вычислить эмпирическую функцию распределения F_j для $\{X_j^i\}_{i=1}^{N}$;

2: вычислить выборочные квантили q_k: $P(X_j < q_k) = F_j(q_k) = p_k$;

3: в соответствии с полученными квантилями преобразовать исходные данные: если $X_j \in [q_{k-1}, q_k]$, то заменить это значение на символ a_k.

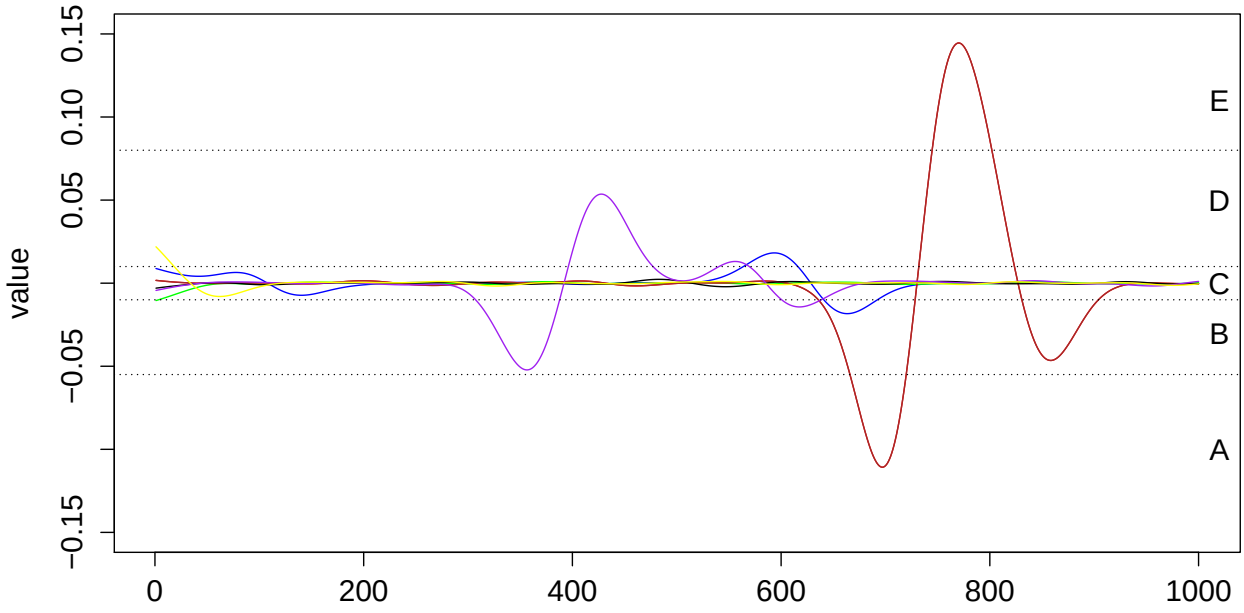


Рис. 1. Пример дискретизации временного ряда, разными цветами обозначен один и тот же временной ряд но для различных объектов. Пунктирные линии – квантили.

4.2. Сегментация

Теперь каждый объект описывается многомерным дискретным рядом. Для понижения размерности по времени предлагается разбить объекты на однородные участки – сегменты. Тогда объект можно будет представить как последовательность сегментов. Для этого необходимо уметь сравнивать вектора значений многомерного дискретного ряда в различные моменты времени.

Рассмотрим вектор значений многомерного временного ряда X^i в моменты времени t_1 и t_2. Расстояние между значениями, соответствующими изначально дискретным рядам будем считать по метрике Хэмминга:

$$\rho_{Ham}(a_k, a_l) = |a_k \neq a_l|.$$

Так как расстояние между значениями дискретизованных рядов неравномерно — расстояние между символами A и E значительно больше расстояния между символами A и B — введем метрику, которая учитывает упорядоченность этих символов:

$$\rho_c(a_k, a_l) = \left\| \frac{p_k + p_{k-1}}{2} - \frac{p_l + p_{l-1}}{2} \right\|,$$

где p_k, p_{k-1} – вероятности, которые подавались на вход алгоритма 1. Эти вероятности соответствуют квантилям q_k, q_{k-1}, ограничивающим интервал исходных значений, заменяемых на символ a_k в алгоритме 1.

Тогда расстояние между векторами значений многомерного дискретного временного ряда X^i в моменты времени t_1 и t_2 записывается следующей формулой:

$$\rho(X^i_{t_1}, X^i_{t_2}) = \sum_{j \in J_c} \rho_c(X^i_{jt_1}, X^i_{jt_2}) + 0.5 \sum_{j \in J_d} \rho_{Ham}(X^i_{jt_1}, X^i_{jt_2}), \quad (1)$$

$$\rho_{Ham}(X^i_{jt_1}, X^i_{jt_2}) = |X^i_{jt_1} \neq X^i_{jt_2}|.$$

Если два момента времени схожи по метрике (1), то их можно отнести к одному участку однородности.

Для определения границ участков однородности предлагается использовать скользящее окно. Окном ширины w с концом в моменте времени t многомерного временного ряда X^i размера $J \times T_i$, где J – число одномерных временных рядов, T_i –время измерения, назовем последовательность отсчётов $X^i_{t-w+1} \dots X^i_{t-1} X^i_t$. Каждому окну соответствует w моментов времени.

Проходя скользящим окном по многомерному временному ряду, будем вычислять внутри каждого окна среднее расстояние (1) от последнего момента времени в окне, до всех предыдущих в окне (см. рис. 2). Таким образом, для каждого момента времени $t = \overline{w, T}$ мы получим по одному значению $f(t)$.

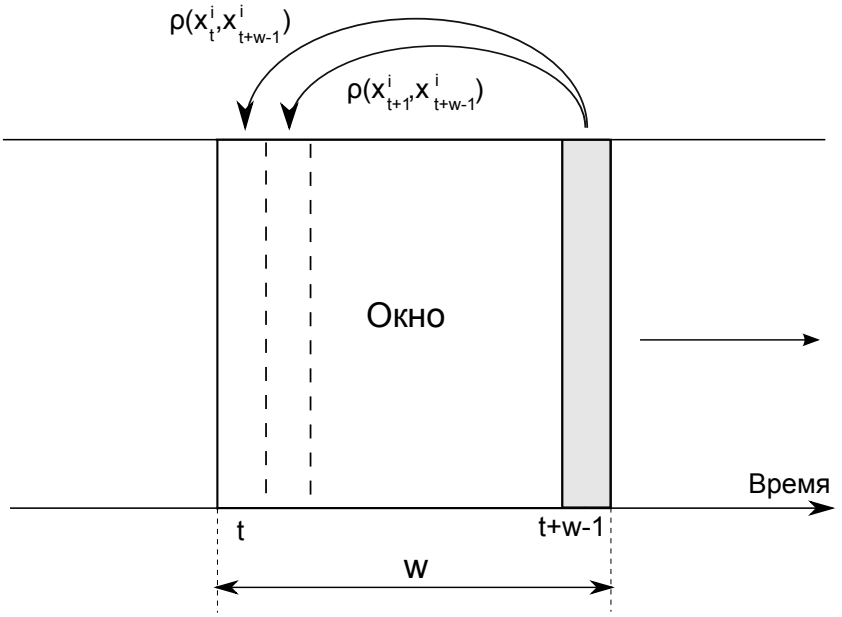


Рис. 2. Иллюстрация окна ширины w

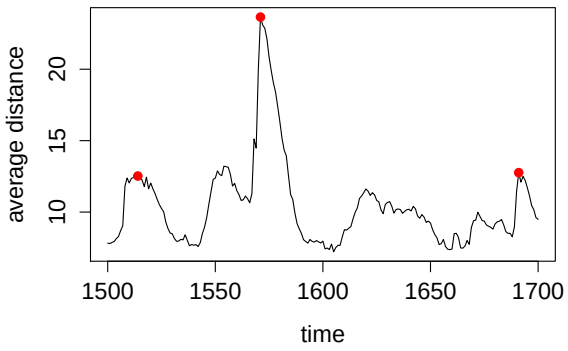


Рис. 3. Пример графика среднего расстояния, красные точки - главные локальные максимумы

Смысл этих значений в том, что они описывают однородность внутри объекта. Большие значения расстояния соответствуют структурным изменениям во временном ряде, а малые значения говорят об однородности временного ряда внутри данного окна.

Найдя локальные максимумы полученных значений, получаем разбиение на участки однородности – сегменты; например, из рисунка 3 видно, что получится три сегмента.

На реальных данных выбор локальных максимумов затруднен – выбирается большое количество “ложных” максимумов (см. рис. 4), которые возникают из-за шума, в связи с чем получается слишком сильное дробление на сегменты. Эту проблему можно решить прореживая найденные максимумы следующим образом:

- выбираем все локальные максимумы полученных значений;
- среди уже найденных локальных максимумов ещё раз выбираем локальные максимумы;
- проделываем эту процедуру до тех пор, пока визуально не будет видно, что выделены все главные максимумы.

4.3. Кластеризация сегментов

Теперь предлагается кластеризовать полученные сегменты всех объектов. Заменяя каждый сегмент на метку соответствующего кластера, получим представление объекта одномерным дискретным рядом. То есть уменьшается размерность по времени. Опишем подробно процедуру кластеризации.

Алгоритм 2 Сегментация объекта представленного многомерным дискретным рядом.

Вход: $\{X^i_{jt}\}, j = \overline{1,J}, t = \overline{1,T_i}$ – объект

w – ширина окна.

n – количество итераций выбора локальных максимумов

Выход: сегменты $\{S^i_m\}$, $m = 1, \ldots, M_i$.

1: **для всех** $t \in \{w, \ldots, T_i\}$

2: берем окно с концом в точке t и считаем среднее расстояние (1) от последнего момента времени до всех предыдущих:
$f(t) = \frac{1}{w-1} \sum\limits_{\tau=t-w+1}^{t-1} \rho(X^i_\tau, X^i_t)$.

3: ищем локальные максимумы: $t^1_{max} = \varnothing$.

4: **для всех** $t \in \{w, \ldots, T_i\}$

5: **Если** $f(t) > \max(f(t-1), f(t+1))$ **то**

6: $t^1_{max} = t^1_{max} \cup t$

7: **для всех** $k \in \{2, \ldots, n\}$

8: среди уже найденных локальных максимумов ищем новые: $t^k_{max} = \varnothing$

9: **для всех** $t \in t^{k-1}_{max}$

10: **Если** $f(t) > \max(f(t-1), f(t+1))$ **то**

11: $t^k_{max} = t^k_{max} \cup t$

12: t^n_{max} – границы сегментов. Тогда определим:
$S^i_1 = X^i_1 X^i_2 \ldots X^i_{t_{(1)}}$, $S^i_2 = X^i_{t_{(1)}+1} \ldots X^i_{t_{(2)}}$,и так далее, где
$t_{(1)}, t_{(2)} \in t^n_{max}$ – отсортированные по возрастанию локальные максимумы,
$|t^n_{max}| = M_i - 1$.

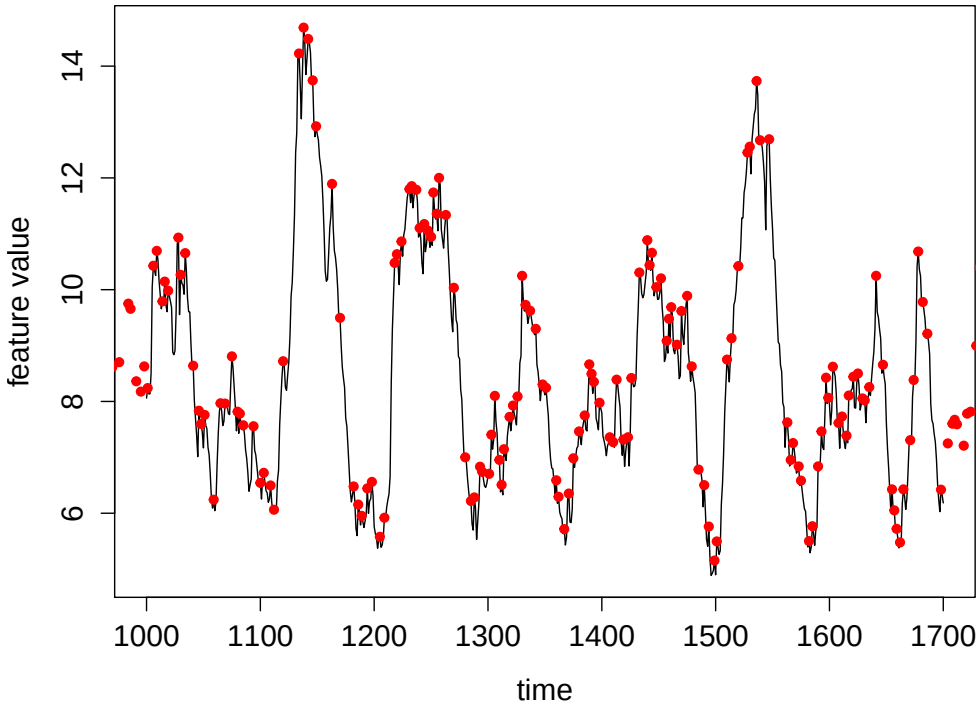


Рис. 4. Первичный выбор локальных максимумов. Локальные максимумы обозначены красными точками.

В соответствии с тем как мы разбивали на сегменты — внутри сегментов временные ряды однородны — перейдём к частотному описанию этих сегментов, практически не теряя информации об аномалиях. Каждый сегмент представляет собой матрицу размера $J \times T$. Рассмотрим j-ю компоненту этого ряда, то есть j-й временной ряд. Множество значений этого временного ряда есть дискретный алфавит $\Sigma_j, |\Sigma_j| = n$.
Тогда поставим в соответствие каждому временному ряду j вектор частот ν_j, $|\nu_j| = n$, $\sum_{k=1}^{n} \nu_{kj} = 1$.

Введем среднюю метрику Хеллингера между вышеопределенными частотными представлениями сегментов:

$$\rho_H(S^1, S^2) = \frac{1}{J\sqrt{2}} \sum_{j=1}^{J} \sqrt{\sum_{k=1}^{n} \left(\sqrt{\nu_{kj}^1} - \sqrt{\nu_{kj}^2}\right)^2} \tag{2}$$

Все сегменты всех объектов предлагается кластеризовать иерархически методом Уорда. Число кластеров C выбирается из предположения о том, что все объекты в большинстве своем схожи, то есть схожие по времени сегменты будут объединяться в один кластер. Тогда число кластеров ограничено сверху минимальным количеством сегментов в объ-

екте.

После кластеризации каждый объект представим одномерным дискретным рядом: заменяя каждый сегмент S_m^i на метку $\{1, \ldots, C\}$ соответствующего кластера, получим символьное представление объекта. Таким образом мы свели исходную задачу к задаче поиска аномалий в одномерных дискретных последовательностях.

5. Понижение размерности по компонентам многомерного ряда

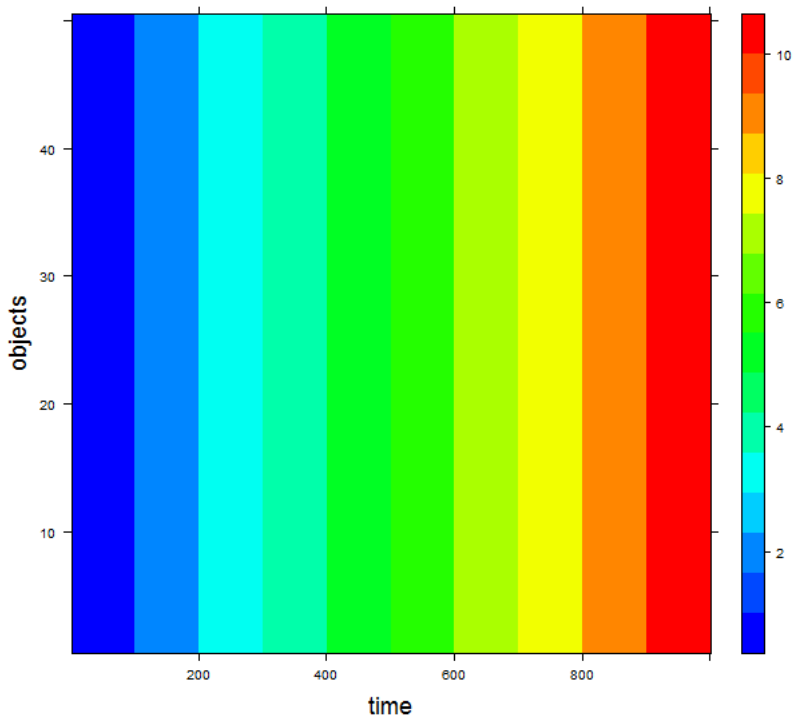


Рис. 5. “Идеальная” кластеризация сегментов

Исходя из предположений о свойствах объектов — ряды описывают один и тот же процесс, происходящий с различными объектами — ожидается, что схожие по времени сегменты различных объектов будут объединяться в один кластер (см. рис. 5). Кластеризация сегментов может не дать желаемого результата, если среди временных рядов описывающих объект много шумовых. Отберем ряды, позволяющие кластеризовать наши объекты наилучшим образом, с помощью жадного алгоритма. Требуется предложить функционал качества кластеризации, который будет учитывать порядок следования сегментов в объектах. Предлагается выбрать в качестве такого функционала среднее попарное сходство между объектами на основе наибольшей общей подпоследовательности:ы

$$Q(\rho) = \frac{2}{N^2} \sum_{i_1=1}^{N-1} \sum_{i_2=i_1+1}^{N} nLCS(X_\rho^{i_1}, X_\rho^{i_2}) \to \max, \tag{3}$$

где $nLCS(X_\rho^{i_1}, X_\rho^{i_2}) = \frac{|LCS(X_\rho^{i_1}, X_\rho^{i_2})|}{\sqrt{l(X_\rho^{i_1}) l(X_\rho^{i_2})}}$, N – число объектов,

$X_\rho^{i_1}$, $X_\rho^{i_2}$ – объекты, представленные одномерным дискретным рядом (после кластеризации сегментов с метрикой между ними ρ)

$l(X_\rho^{i_1})$, $l(X_\rho^{i_2})$ – длины объектов

$LCS(\cdot,\cdot)$ – наибольшая общая подпоследовательность

$nLCS(\cdot,\cdot)$ – длина $LCS(\cdot,\cdot)$ нормированная на длины последовательностей

Здесь общей подпоследовательностью двух последовательностей $\mathbf{y} = y_1, \ldots, y_n$ и $\mathbf{z} = z_1, \ldots, z_m$ является такая последовательность $\mathbf{x} = x_1, \ldots x_k, k \leqslant \min(m, n)$, которая может быть получена из y и из z путем выбрасывания любых элементов.

Например:

$$x = ABAACB, y = ACBAB$$

Их наибольшей общей подпоследовательностью будет $ABAB$.

Алгоритм 3 Отбор временных рядов.

Вход: $\{S_m^i\}$, где $i = \overline{1, N}$, $m = \overline{1, M_i}$ — множество сегментов всех объектов.

Выход: метрика ρ между сегментами

1: Инициализация: $n = 1$
2: **для всех** $j \in J$
3: кластеризовать сегменты с метрикой (2), где $J = j$.
Обозначим эту метрику ρ_j.
4: вычислить значение функционала (3)
5: Выбрать тот временной ряд, для которого значение функционала максимально:
$\rho^1 = arg\max_{j \in J} Q(\rho_j)$
6: **повторять**
7: **для всех** $j \in J$
8: **для всех** $\alpha = 0 : 0.01 : 1$
9: кластеризовать сегменты с метрикой $\rho^n(\alpha, j) = \alpha\rho^{n-1} + (1 - \alpha)\rho_j$
10: вычислить значение функционала (3)
11: обновление метрики: $\rho^n = arg\max_{j \in J, \alpha} Q(\rho^n(\alpha, j))$
12: **пока** $Q(\rho^n) <= Q(\rho^{n-1})$

Таким образом, жадный алгоритм 3 отберет те временные ряды, которые позволяют кластеризовать сегменты наилучшим образом. Исходя из предположения, что различные объекты должны быть похожи во времени, данный алгоритм позволит уменьшить размерность пространства по временным рядам.

6. Вычислительный эксперимент

Алгоритм, предложенный в данной работе, был применен к данным по полетам самолетов. Данные представляли собой показания 304 датчиков для круизной фазы 79 полетов одного типа самолета. Длина круизной фазы для различных полетов варьировалась от 3000 до 6000. В введенных ранее терминах: $N = 79$ – число объектов, $J = 304$ – число временных рядов описывающих объект, $T \in [3000, 6000]$ – время измерения временных рядов. Информация об аномалиях отсутствовала.

Количество непрерывных и дискретных датчиков равнялось 195 и 109 соответственно. Непрерывные датчики были дискретизованы, для всех датчиков был выбран пятибуквенный алфавит (A, B, C, D, E), вероятности для квантилей были выбраны равными $(0, 0.005, 0.335, 0.665, 0.995, 1)$. Таким образом, после дискретизации все маловероятные события будут сосредоточены в буквах A и E. Это делается для того, чтобы мы не теряли аномалии при дискретизации: если, например, рассмотреть равновероятный переход к пятибуквенному алфавиту — $(p_0, p_1, \ldots, p_5) = (0, 0.2, 0.4, 0.6, 0.8, 1)$ — то мы не сможем отличить маловероятное событие от любого другого. Из рисунков 8 и 10 видно, что хотя различия в полетах и остаются, они становятся значительно меньше.

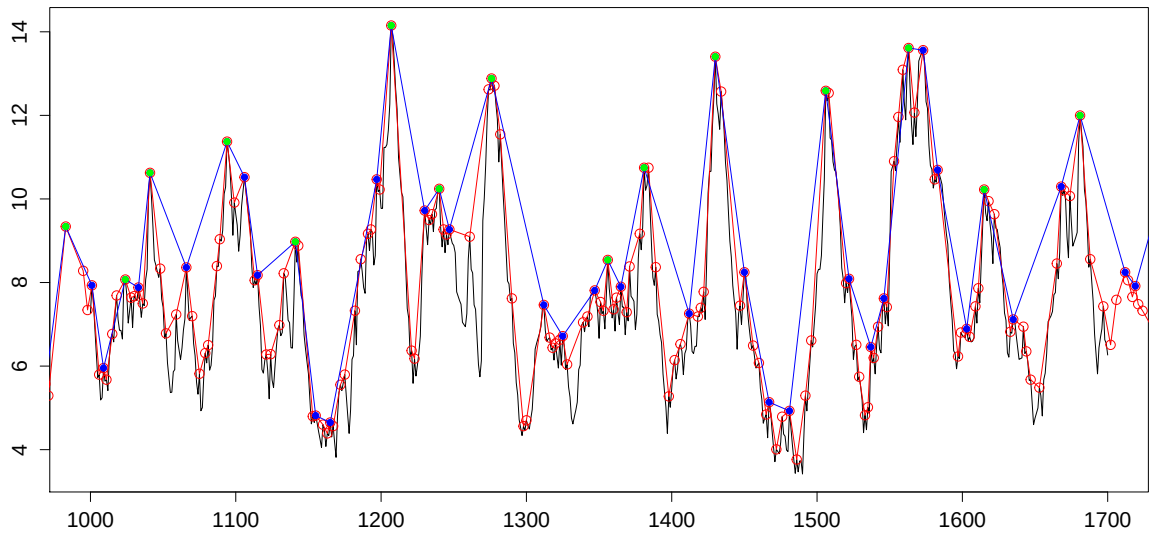


Рис. 6. Графики среднего расстояния в окне в зависимости от времени. Красные точки – первичные локальные максимумы, синие – вторичные, зеленые – итоговые.

При сегментации ширина окна w была выбрана равной 20, границы сегментов выбирались согласно алгоритму 2, с числом итераций $n = 3$.

Из графика 6 видно, что, после трёх проходов оказались выбраны почти все глав-

ные максимумы.

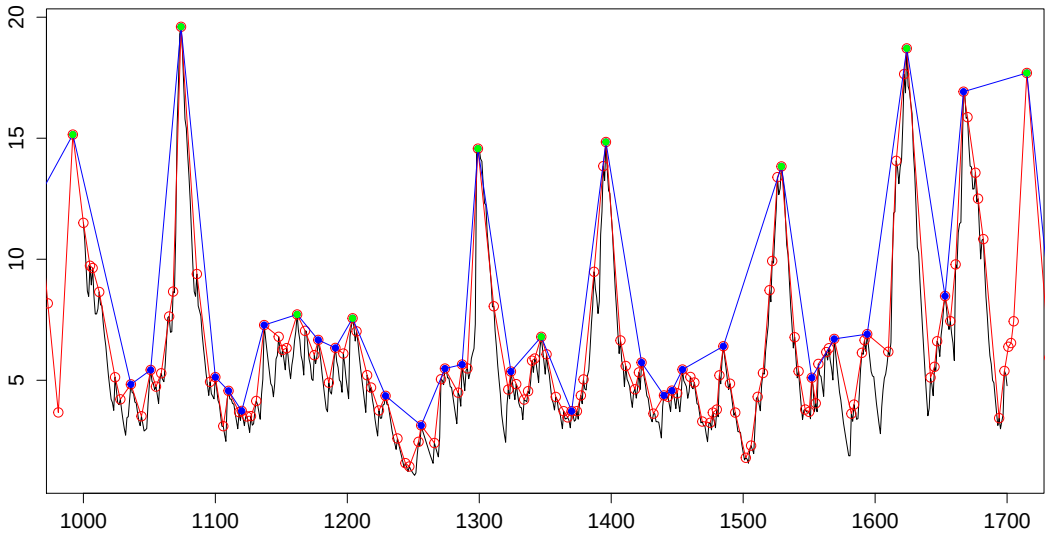


Рис. 7. График процесса выделения главных максимумов

В результате отбора временных рядов, отобралось 2 константных дискретных датчика. после чего функционал качества не улучшался. Несмотря на то, что данные датчики кластеризовали объекты в соответствии с нашими предположениями, очевидно, что такого малого количества датчиков не достаточно для данной задачи.

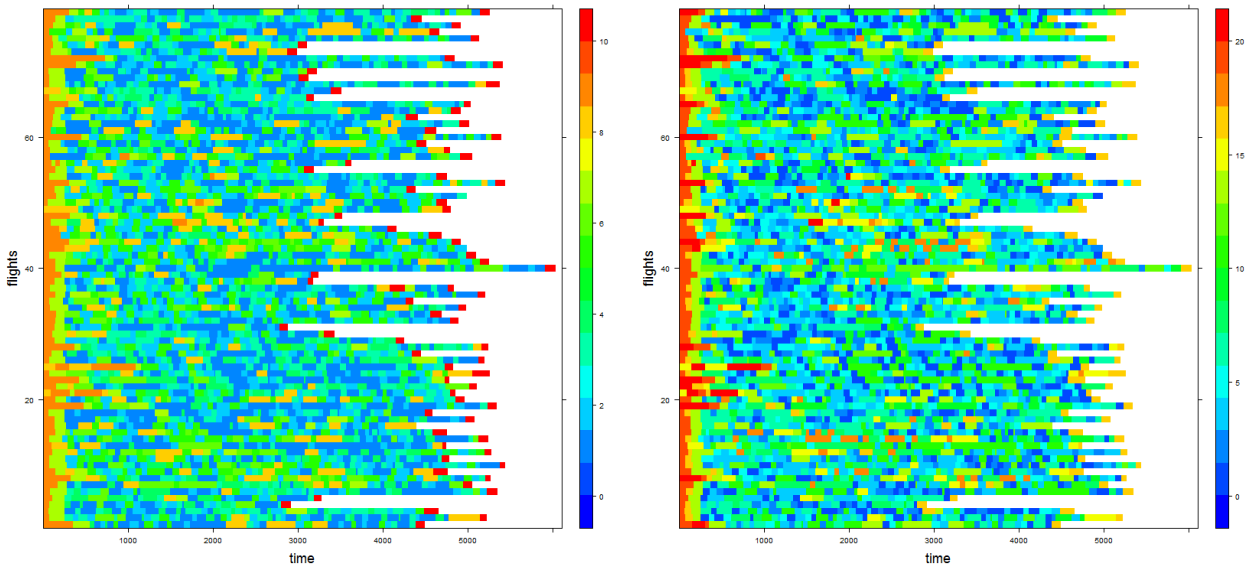
Была попытка изменить данный функционал, чтобы он не выбирал только дискретные датчики: было введено добавочное слагаемое, равное числу различных кластеров, представленных в объекте (полете), деленному на общее число сегментов в данном объекте. При таком задании функционала качества выбирались уже не только дискретные датчики, но и непрерывные, однако алгоритм остановился на 5 датчиках. Полученное значение функционала было меньше, чем если выбрать все датчики сразу, что свидетельствует о том, что нецелесообразно применять этот функционал для данной задачи.

В итоге отбор временных рядов в данной задаче не применялся.

На рисунке 9 можно увидеть, что после кластеризации сегменты различных объектов (полетов), схожие по времени, попадают в один кластер, как и подразумевалось.

Увеличение размера алфавита при дискретизации не оказывает существенного влияния на итоговый результат. Результаты кластеризации для одиннадцатибуквенного алфавита представлены на рисунке 11. Видно, что те же объекты сильно выделяются на фоне остальных.

Можно утверждать, что данные результаты хорошо описывают общую структу-



10 кластеров. 20 кластеров.

Рис. 8. Иллюстрация кластеризации для 10 и 20 кластеров, при равновероятном алфавите. Цвет – номер кластера.

ру объектов, и, как можно видеть из рисунков 10 и 9, объекты, которые могут быть аномальными, остаются таковыми при увеличении числа кластеров. Например, визуальный анализ позволяет выделить полеты под номерами 47 и 50 как наиболее отличающиеся от остальных по сегментной структуре.

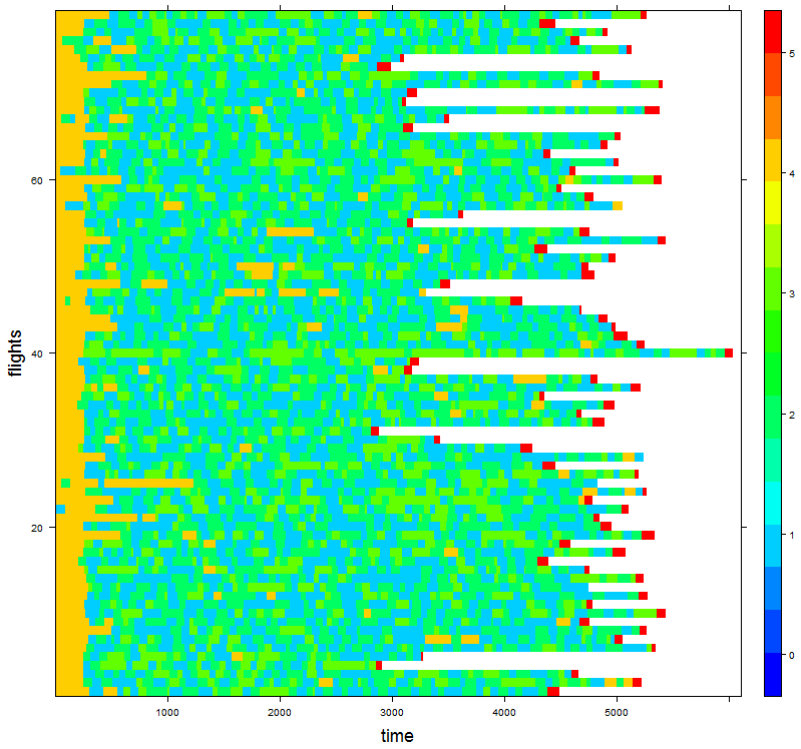


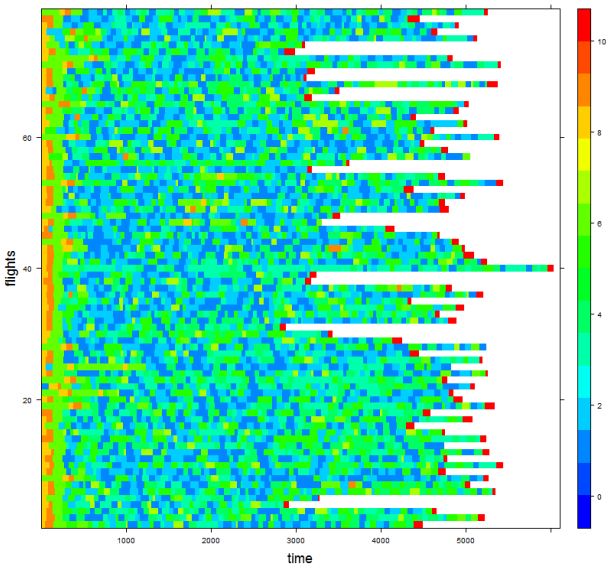
Рис. 9. Иллюстрация кластеризации сегментов для 5 кластеров. Цвет – номер кластера.

7. Заключение

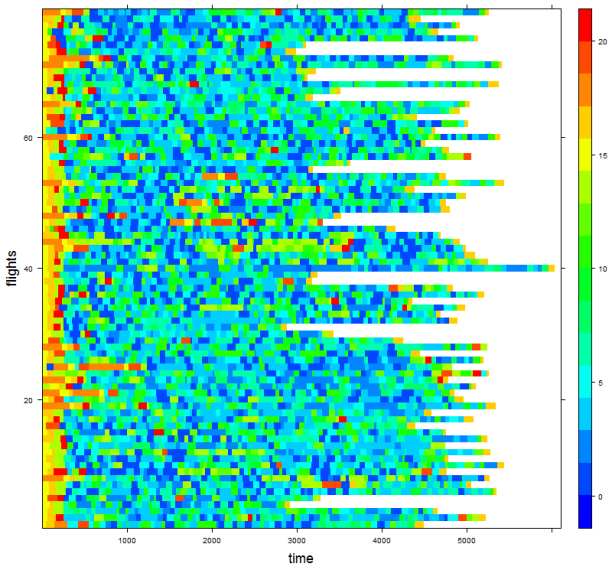
В работе рассмотрена задача поиска аномалий в объектах, которые описываются многомерными временными рядами. Предложен алгоритм, позволяющий уменьшить размерности пространства, при этом не теряя информации об аномалиях. Считалось что аномалия – достаточно продолжительный промежуток времени, тогда аномальным можно считать сегмент, что позволяет проводить дальнейший анализ в уже одномерных дискретных рядах.

Главные преимущества данного подхода:

- аномалии локализуются внутри сегментов;
- уменьшается размерность исходной задачи по времени и по количеству временных рядов;
- возможность использовать весь комплекс алгоритмов для задачи поиска аномалий в одномерных дискретных рядах;
- данный алгоритм можно использовать для предварительного анализа данных.

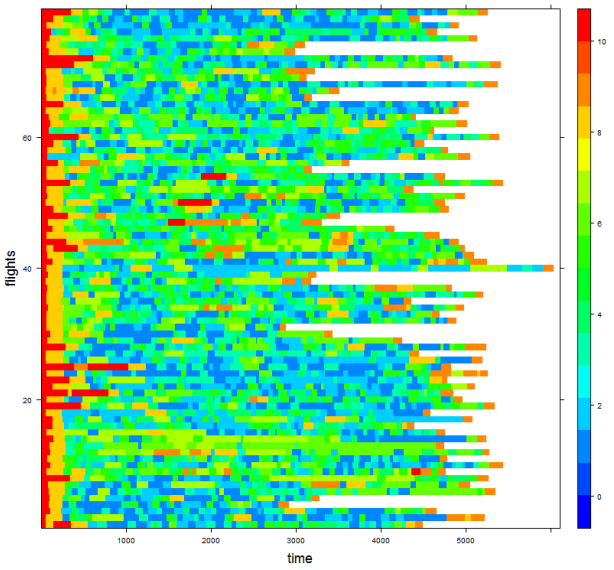


10 кластеров.

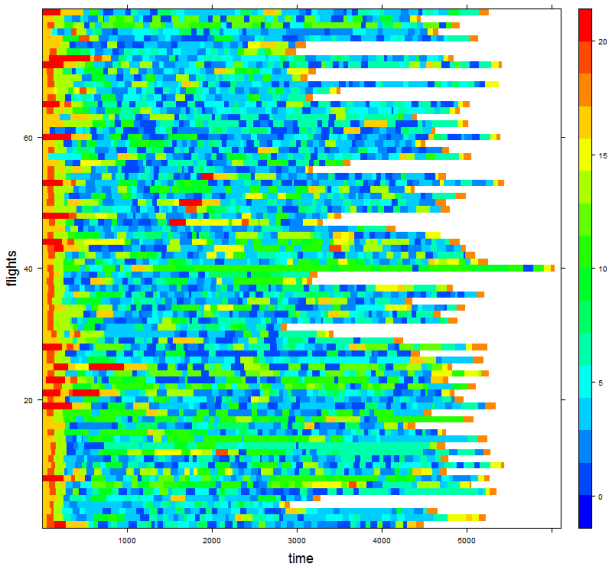


20 кластеров.

Рис. 10. Иллюстрация кластеризации для 10 и 20 кластеров. Цвет – номер кластера.



10 кластеров.



20 кластеров.

Рис. 11. Иллюстрация кластеризации для 10 и 20 кластеров при 11-буквенном алфавите. Цвет – номер кластера.

Список литературы

1. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.
2. J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pp. 2–11, ACM, 2003.
3. S. Budalakoti, A. N. Srivastava, and M. E. Otey, "Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 39, no. 1, pp. 101–113, 2009.
4. V. Chandola, *Anomaly detection for symbolic sequences and time series data*. PhD thesis, University of Minnesota, 2009.
5. A. Lazarevic, A. Banerjee, V. Chandola, V. Kumar, and J. Srivastava, "Data mining for anomaly detection," in *Tutorial at the European Conference on Principles and Practice of Knowledge Discovery in Databases, Antwerp, Belgium, September*, vol. 19, 2008.
6. S. Das, B. L. Matthews, A. N. Srivastava, and N. C. Oza, "Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 47–56, ACM, 2010.
7. A. N. Srivastava, "Discovering system health anomalies using data mining techniques," in *Proceedings of 2005 Joint Army Navy NASA Airforce Conference on Propulsion*, Citeseer, 2005.
8. R. S. Tsay, D. Peña, and A. E. Pankratz, "Outliers in multivariate time series," *Biometrika*, vol. 87, no. 4, pp. 789–804, 2000.
9. R. Baragona and F. Battaglia, "Outliers detection in multivariate time series by independent component analysis," *Neural computation*, vol. 19, no. 7, pp. 1962–1984, 2007.
10. J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing sax: a novel symbolic representation of time series," *Data Mining and Knowledge Discovery*, vol. 15, no. 2, pp. 107–144, 2007.