НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2024 Том 24 № 2 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2024 Vol. 24 No 2 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2024-24-2-230-240
УДК 004.8

Гарантированное обнаружение структурных аномалий в потоковых данных с использованием метода RRCF: выбор параметров обнаружителя и его стабилизация в условиях аддитивных шумов

Андрей Владимирович Тимофеев✉

ТОО «Эквалайзум», Астана, 010000, Казахстан
timofeev.andrey@gmail.com✉, https://orcid.org/0000-0001-7212-5230

Аннотация

Введение. Предложены метод стабилизации обнаружения структурных аномалий в условиях аддитивных шумов, а также алгоритм формального выбора параметров решающего правила в обнаружителе структурных аномалий на основе метода Robust Random Cut Forest (RRCF). **Метод.** В рамках разработанного метода, для стабилизации процесса обнаружения структурных аномалий в условиях воздействия аддитивных шумов, предложено подавать на вход RRCF-обнаружителя поток данных, который предварительно обработан одним из методов цифровой фильтрации. При этом правило принятия решения об обнаружении аномалии строго формализовано и прозрачно интерпретируется. **Основные результаты.** Формализован выбор параметров стабилизированного методами предварительной фильтрации данных входного потока обнаружителя аномалий на базе RRCF. Параметр обнаружителя, выбранный в рамках предложенный схемы, гарантирует априорно заданную верхнюю границу для вероятности ложной тревоги при принятии решения об обнаружении структурной аномалии. Это свойство строго доказано и оформлено в виде теоремы. Эффективность работы стабилизированного RRCF-обнаружителя аномалий исследована численным методом. Достигнутые результаты подтверждают работоспособность рассмотренного подхода при условии выбора порога обнаружения предложенным способом. Приведен пример практического использования предложенного RRCF-обнаружителя. **Обсуждение.** Разработанный подход перспективен для обнаружения структурных аномалий в условиях зашумления наблюдений аддитивной помехой, в случае, когда важно гарантировать верхнюю границу для вероятности ложной тревоги. В частности, подход может найти применение при контроле технологических режимов прокачки жидкости в трубопроводных системах или в системах обнаружения предотказных состояний технологического оборудования.

Ключевые слова

Robust Random Cut Forest, обнаружение структурных аномалий, потоковая обработка данных, гарантированное обнаружение аномалий

Ссылка для цитирования: Тимофеев А.В. Гарантированное обнаружение структурных аномалий в потоковых данных с использованием метода RRCF: выбор параметров обнаружителя и его стабилизация в условиях аддитивных шумов // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 2. С. 230–240. doi: 10.17586/2226-1494-2024-24-2-230-240

Guarantee structural anomaly detection in streaming data using the RRCF model: selection of detector parameters and its stabilization under additive noise conditions

Andrey V. Timofeev✉

LLP “EqualiZoom”, Astana, 010000, Kazakhstan
timofeev.andrey@gmail.com✉, https://orcid.org/0000-0001-7212-5230

Abstract

A method for stabilizing structural anomaly detection under additive noise conditions as well as an algorithm for formal selection of the parameters of the solver rule in the structural anomaly detector based on the Robust Random Cut Forest

(RRCF) method are proposed. In the framework of the developed approach, in order to stabilize the process of structural anomaly detection under the influence of additive noise, it is proposed to feed to the input of the RRCF-detector a data stream which is pre-processed by one of the digital filtering methods. In this case, the decision rule for anomaly detection is strictly formalized and transparently interpreted. The selection of parameters of the RRCF-based anomaly detector stabilized by pre-filtering methods of the input data stream is formalized. The RRCF-detector parameters choice within the proposed scheme guarantees a predetermined upper bound for the false alarm probability when deciding to detect a structural anomaly. This property is rigorously proved and formalized as a theorem. The performance of the stabilized RRCF-detector is investigated numerically. The achieved results confirm the performance of the proposed approach provided that the detection threshold is selected in the way proposed in this paper. An example of practical application of the proposed method is presented. The developed approach is promising for the detection of structural anomalies in conditions of observation additive noise, in a situation where it is important to guarantee an upper bound for the probability of false alarm. In particular, the approach can find application in monitoring technological regimes of liquid pumping in pipeline systems or in systems for detecting pre-failure states of technological equipment.

Keywords

Robust Random Cut Forest, structural anomaly detection, streaming data processing, guaranteed anomaly detection

For citation: Timofeev A.V. Guarantee structural anomaly detection in streaming data using the RRCF model: selection of detector parameters and its stabilization under additive noise conditions. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 2, pp. 230–240 (in Russian). doi: 10.17586/2226-1494-2024-24-2-230-240

Введение

Проблема оперативного обнаружения аномалий в непрерывном потоке данных часто встречается на практике, например, в системах контроля технологических процессов, при обнаружении мошенничества (fraud) в банковских транзакциях, для обеспечения безопасности телекоммуникационных сетей, а также в других прикладных областях [1–12]. В ряде случаев, практически приемлемый уровень показателей эффективности решения данной задачи обеспечивают классические методы, основанные на обнаружении разладки случайных процессов (change point detection), а также методы, основанные на использовании машинного обучения («one class SVM» и др.). Всем этим методам характерны как достоинства, так и недостатки. Основным недостатком этих методов является сравнительно низкая чувствительность к малоамплитудным структурным аномалиям, когда по амплитудно-частотным характеристикам аномалия отличается от нормы незначительно. Частично эти недостатки способен компенсировать сравнительно новый метод **ансамблевого обнаружения аномалий**, который называется Robust Random Cut Forest (RRCF) [13]. В настоящей работе исследовано несколько важных свойств данного метода, сформирована методика определения его параметров, а также изучены способы стабилизации процесса обнаружения аномалий в условиях аддитивных шумов, отличных от классического метода стабилизации «bagging» [14]. Под стабилизацией понимается внесение в метод RRCF определенных алгоритмических дополнений, которые обеспечивают сохранение способности данного метода к эффективному обнаружению структурных аномалий при наличии аддитивного центрированного шума наблюдений с конечной дисперсией.

Определения и постановка задачи

Пусть в моменты времени $T = (t_0, t_1, \ldots)$ выполнены измерения случайного процесса $z(t)$, $\forall t \in T$: $z(t) \in Z$, где множество Z — априорно задано. Существуют априорно неизвестные величины $\tau_1, \tau_2 \in T$ такие, что для некоторых функций ρ, g ($\rho \neq g$) допустима запись:

$$z(t) = \begin{cases} g(t) + \xi(t), t \notin [\tau_1, \tau_2] \\ \rho(t) + \xi(t), t \in [\tau_1, \tau_2] \end{cases}, t \in T,$$

где функции ρ, g — неизвестны; $\xi(t)$ — шумовой случайный процесс с неизвестным распределением; $\mathbf{E}\xi(t) = 0$, $\mathbf{E}\xi^2(t) = \sigma^2 < \infty$; $\underset{t \neq k}{\forall} \mathbf{E}\xi(t)\xi(k) = 0$; σ — величина неизвестна. Здесь и далее $\mathbf{E}(x)$ — математическое ожидание величины x, а $\mathbf{P}(\omega)$ — вероятность события ω.

Необходимо создать решающее правило $\Psi(t, \theta | \{z(t) | t \in T\})$ такое, что

$$[\Psi(t, \theta | \{z(t) | t \in T\}) = true] \Rightarrow [t \in o_\varepsilon([\tau_1, \tau_2])],$$

$$[\Psi(t, \theta | \{z(t) | t \in T\}) = false] \Rightarrow [t \notin o_\varepsilon([\tau_1, \tau_2])],$$

а для априорно заданной величины $\alpha \in]0, 1[$ имеет место следующее неравенство:

$$\mathbf{P}(\Psi(t, \theta | \{z(t) | t \in T\}) = true | t \notin o_\varepsilon([\tau_1, \tau_2])) \leq \alpha, \quad (1)$$

где $o_\varepsilon([\tau_1, \tau_2])$ — ε-окрестность замкнутого интервала $[\tau_1, \tau_2]$, где $o_\varepsilon([\tau_1, \tau_2]) = o_\varepsilon([\tau_1 - \varepsilon, \tau_2 + \varepsilon])$ для некоторой, достаточно малой величины $\varepsilon > 0$; θ — порог принятия решения, величина которого зависит от функции g и априорно неизвестной константы σ.

Процесс $\Phi_T = \{z(t) | t \notin [\tau_1, \tau_2]\}$ назовем **базовым** (фоновым) процессом, а процесс $A_T = \{z(t) | t \in [\tau_1, \tau_2]\}$ — **аномалией**.

Robust Random Cut Forest для обнаружения аномалий

В основе метода RRCF лежит идея оперативного контроля сложности анализируемого фрагмента потока наблюдений, которая является новой для задач данного класса [13]. В последние годы метод RRCF часто используется на практике [15, 16]. Допустим, что аномалия представляет собой редкое событие, тогда

можно считать, что в основу функционирования RRCF положена следующая последовательность шагов.

1. Формируется лес F (F-ансамбль) из бинарных деревьев. Число деревьев и листьев в каждом дереве являются настроечными параметрами, которые адаптируются под анализируемый процесс.
2. При поступлении партии новых наблюдений (точек) формируется сдвигающееся окно, состоящее из точек, которые вставляются в каждое дерево из F-ансамбля с использованием формального метода — «вставка в бинарное дерево». Напомним, что каждый узел в бинарном дереве представляет собой «признак разделения», который является способом разделения пространства наблюдений на два подмножества. В случае бинарного дерева каждый узел может быть «левым» или «правым», что соответствует двум возможным значениям признака разделения. Когда реализуется операция вставки новой точки в бинарное дерево, всегда выбирается признак разделения для этой точки на основе некоторой стратегии. При использовании метода RRCF, признак разделения выбирается случайным образом из всех возможных вариантов. Такой метод выбора служит для обеспечения устойчивости к изменениям в данных, в рамках bagging-идеологии. Таким образом, операция «вставка в бинарное дерево» осуществляется для случайно выбранного признака разделения. В итоге все деревья из F-ансамбля модифицируются различным способом. Постепенно в структуре F-ансамбля отображается фоновая (нормальная) модель процесса Φ_T. Через некоторое время, в зависимости от темпа поступления входных данных, фоновая модель будет обучена инкрементальным методом. Другими словами, F-ансамбль будет настроен на норму Φ_T, причем сложность нормы известна и определена структурой инкрементально обученного F-ансамбля.
3. Окно из точек сдвигается на шаг по времени и производится оценка того, насколько изменилась сложность модели после добавления группы точек из окна? Если оцененная сложность модели превысила некоторый порог, считается, что в окне содержатся точки, соответствующие аномалии. Таким образом, сложность анализируемой порции точек (измерений) значимо отлична от сложности F-ансамбля и в результате можно сделать вывод о наличии аномалии.

В качестве функции, измеряющей сложность модели, используется некоторая функция, определенная для каждой вершины дерева и зависящая от его глубины, которая называется коллизионным перемещением (Collusive Displacement, *CoDisp*). В случае, если функция *CoDisp* определяется для группы точек, образующих сдвиговое окно, величина *CoDisp* сначала вычисляется для каждой точки (по всему F-ансамблю), а в качестве итога рассчитывается среднее значение по всем точкам. Фактически функция *CoDisp* представляет собой меру ранжирования вершин внутри дерева, которое в зависимости от величины некоторого порога θ, позволяет отделить аномальные измерения от нормальных. В рамках метода RRCF, при выборе порогового значения θ для величины *CoDisp* практически определяется то, какие вершины будут считаться аномалиями, в частности, к множеству аномальных относятся все вершины, для которых функция $CoDisp > \theta$. Чем ниже порог θ, тем больше вершин считаются аномалиями, и наоборот. В ряде научных работ, где рассмотрен метод RRCF, **не приводится** методика выбора порога θ, поэтому данная методика представлена в настоящей работе.

Напомним, что операция insert_point (вставка новой точки в бинарное дерево *Tr*) описывается следующей последовательностью действий.

1. Процедура начинается с корня дерева.
2. Значение вставляемой точки сравнивается с текущим узлом. Если значение меньше, точка сдвигается влево (в направлении «левого» поддерева), если больше — вправо (в направление «правого» поддерева).
3. Процесс продолжается до тех пор, пока не будет найден пустой узел (либо узел, в котором нет ни одного из потомков), куда и вставляется новая точка.

На практике операция insert_point означает создание нового узла и добавление его в дерево в соответствии с правилами бинарного дерева. Как следует из [13], в отличие от стандартного бинарного дерева поиска, RRCF использует механизм случайного выбора признака, по которому будет производиться разделение. При этом остальные признаки все равно участвуют в разделении, определяя какой узел будет являться родительским для новой точки. Такой подход делает метод RRCF более устойчивым к выбросам и менее чувствительным к выбору признаков. Когда новую точку данных вставляют в дерево из F-ансамбля, функция *CoDisp* вычисляет: насколько сильно включение этой точки меняет структуру дерева. Если включение новой точки значительно увеличивает сложность модели (увеличивает битовую глубину дерева), то аномальность этой точки считается более вероятной. Заметим, что для включения в бинарное дерево аномальной точки обязательно потребуется использование *большей* битовой глубины. И наоборот: для включения в бинарное дерево «нормальной» точки будет использована битовая глубина, характерная для инкрементально обученного дерева. Важно отметить и то, что вставляемые точки, которые находятся ближе к корню дерева, скорее всего не будут считаться выбросами. Это обусловлено тем, что точки ближние к корню обычно имеют больше общего с остальной частью данных, и поэтому менее вероятно то, что эти точки принадлежат аномальному процессу A_T.

Обратим внимание, что деревья в RRCF не обучаются в «традиционном» смысле: они обучаются на потоке данных по мере их поступления при условии, что поступающие данные являются нормальными (фоновыми). Иначе говоря, эти данные должны быть элементами процесса Φ_T. При этом каждая вставленная точка изменяет структуру дерева в F-ансамбле, что в свою очередь влияет на способность дерева классифицировать новые точки данных. Таким образом, вставка точки данных в дерево в RRCF является частью процесса *инкрементального обучения* модели. Из изложенного следует, что сложность модели можно представить как

сумму битовых глубин всех узлов дерева. При этом аномалия определяется как точка (группа точек), которая значительно увеличивает сложность модели при ее включении в дерево. Количественная оценка изменения сложности модели в методе RRCF может быть выражена как ожидаемое изменение битовой глубины всех листьев в дереве из F-ансамбля при удалении точки z. Такое изменение обозначим *Disp* (d-смещение). Изменение *Disp* является ключевым аспектом определения аномалий в RRCF и, согласно работе [3], принимая допущение о равновероятности деревьев $Tr \in \mathrm{F}$, определяется в виде:

$$Disp(z, Z) = \sum_{Tr, y \in Z - z} (f(y, Z, Tr) - f(y, Z - z, Tr)) \cdot |\mathrm{F}|^{-1}, \quad (2)$$

где $|\mathrm{F}|$ — мощность F-ансамбля; Z — множество измерений (точек); $f(y, Z, Tr)$ — глубина точки $y \in Z$ в бинарном дереве Tr.

В работе [3] рассмотрена важная концепция определения d-смещения, которая учитывает так называемые «дубликаты» или «близкие дубликаты» («colluders») измерений, существование которых способно маскировать наличие выбросов. Определения этих важных понятий будут даны далее по тексту. Данная концепция состоит в том, что если существует только один аномальный выброс (назовем его первым), то проблем с определением d-смещения, согласно формуле (2), нет: величина $Disp(z, Z)$ будет значительна. Проблема возникает в том случае, когда рядом с первым аномальным выбросом существует второй, близкий к первому. В этом случае d-смещение при удалении второго, в присутствии первого, будет сравнительно малым, так как при удалении из дерева второй выброс сдвинет первый внутри дерева. Такое поведение может привести к тому, что второй выброс будет *маскировать* наличие первого выброса, делая его **менее заметным** для метода RRCF. В этом случае первая и вторая аномалии называются «дубликатами» или «близкими дубликатами», для которых на английском языке используют термин «colluders». Чтобы нивелировать данную проблему, в [3] предложен концепт «Duplicate Resilience», в рамках которого осуществлена модификация формулы (2). При модификации также вычислено d-смещение (функция *CoDisp*), которое реализовано при одновременном удалении целого набора «colluders», представляющего собой множество $C_z \subseteq Z$ и находящихся рядом с целевой точкой $z \in Z$. В работе [1] функция *CoDisp* определена следующим образом:

$$CoDisp(z|Z|S|) = \\ = \mathop{\mathbf{E}}_{S \subseteq Z, T} \left(\max_{C_z \subseteq S} \frac{1}{|S|} \sum_{y \in S - C_z} (f(y, S, Tr) - f(y, S - C_z, Tr)) \right),$$

где $f(y, S, Tr)$ — глубина точки $y \in S \subseteq Z$ в бинарном дереве Tr, для некоторого (достаточно большого) $S \subseteq Z$; C_z — множество «colluders», соответствующих точке $z \in S \subseteq Z$. Отметим, что при условии концепта «Duplicate Resilience», элементы «colluders» соответствуют таким элементам данных, которые имеют схожую структуру или поведение, и поэтому могут быть рассмотрены в качестве *дубликатов* (или *близких дубликатов*) друг друга.

Определение функции $CoDisp(z|Z|S|)$ расширяет понятие модификации модели бинарного дерева Tr с учетом дубликатов удаляемой (добавляемой) точки, а также близких дубликатов, которые могут «замаскировать» наличие выбросов. При этом функция *CoDisp* вычисляется как ожидаемое изменение глубины точек в бинарном дереве Tr из F-ансамбля, когда набор точек C_z, содержащий интересующую нас точку z, удаляется из бинарного дерева Tr. Элементы «colluders» в этом контексте — элементы данных, которые удаляются вместе с элементом z.

Параметры и некоторые особенности метода RRCF

Основными параметрами метода RRCF являются:
- мощность F-ансамбля: $|\mathrm{F}|$;
- верхняя граница размера деревьев $Tr \in \mathrm{F}$: tree_size;
- длина сдвигового окна: shingle_size;
- порог принятия решения: θ.

Чем больше величина $|\mathrm{F}|$, тем устойчивее результат и выше вычислительные затраты. В ином случае, чем больше tree_size, тем устойчивее результат и выше вычислительные затраты. В свою очередь, чем больше shingle_size, тем больше чувствительность метода к слабовыраженным аномалиям, но при этом увеличивается величина $|o_\varepsilon([\tau_1, \tau_2])| - |\tau_1 - \tau_2| = 2\varepsilon$, т. е. падает точность оценивания интервала $[\tau_1, \tau_2]$. Выбор параметров $|\mathrm{F}|$, tree_size и shingle_size в основном определяется величинами ρ, g и σ, которые, как правило, на практике априорно неизвестны.

Однако важно то, что сам принцип построения величины *CoDisp* свидетельствует о том, что для $t \notin [\tau_1, \tau_2]$ распределение величин $CoDisp(z(t)|\cdot)$, для сравнительно небольших значений параметра σ, в общем случае должны иметь *почти стационарный* характер. Многочисленные вычислительные эксперименты подтверждают эту гипотезу.

Рассмотрим следующую модель для $CoDisp(z(t)|\cdot)$:

$$\forall t \notin [\tau_1, \tau_2]: CoDisp(z(t)|\cdot) = m(\rho, g, \sigma) + \varsigma(t). \quad (3)$$

Здесь для набора ρ, g, σ величина $m(\rho, g, \sigma) = \mathrm{const}(\rho, g, \sigma)$, а для величин $\{\varsigma(t)\}$ верно: $\mathbf{E}\varsigma(t) = 0$, $\mathbf{E}\varsigma^2(t) = \mathrm{const} < \infty$, $\forall_{k,l} \mathbf{E}\varsigma(k)\varsigma(l) = 0$. Случайная величина $\varsigma(t)$ зависит от $\{\xi(t)\}$ и ρ.

Для некоторого $\tau < \tau_1$ обозначим: $CoDisp(\tau) = \sum_{t \le \tau} CoDisp(z(t)|\cdot)\tau^{-1}$.

Теорема. Пусть:
1. допустимо представление (3);
2. для некоторого $P_c \in]0, 1[$: $\theta_\tau = CoDisp(\tau) + (1 + \tau^{-0,5}) \times \times \left(\frac{2\mathbf{E}\varsigma^2(\tau)}{(1 - P_c)} \right)^{0,5}$.

Тогда $\mathbf{P}(CoDisp(\tau) < \theta_\tau) \ge P_c$.

Доказательство. Рассмотрим очевидное представление:

$$CoDisp(\tau) = m(\rho, g, \sigma) + \sum_{t \le \tau} \varsigma(t) \cdot \tau^{-1} = m(\rho, g, \sigma) + \varsigma(\tau).$$

На основании неравенства Чебышева имеет место следующее неравенство:

$$\mathbf{P}\left(m(\rho, g, \sigma) \le CoDisp(\tau) + \left(\frac{\mathbf{E}\varsigma^2(\tau)}{\tau \cdot (1 - P_c)} \right)^{0,5} \right) \ge P_c.$$

Из которого следует неравенство:

$$\mathbf{P}\left(m(\rho, g, \sigma) \le CoDisp(\tau) + \left(\frac{2\mathbf{E}\varsigma^2(\tau)}{\tau \cdot (1 - P_c)} \right)^{0,5} \right) \ge P_c.$$

Рассмотрим события w_m, $\overline{w_m}$, w_ς, $\overline{w_\varsigma}$ и w_θ, определенные следующим образом:

$$w_m: \left\{ m(\rho, g, \sigma) \le CoDisp(\tau) + \left(\frac{2\mathbf{E}\varsigma^2(\tau)}{\tau \cdot (1 - P_c)} \right)^{0,5} \right\},$$

$$\overline{w_m}: \left\{ m(\rho, g, \sigma) > CoDisp(\tau) + \left(\frac{2\mathbf{E}\varsigma^2(\tau)}{\tau \cdot (1 - P_c)} \right)^{0,5} \right\},$$

$$w_\varsigma: \left\{ |\varsigma(\tau)| \le \left(\frac{\mathbf{E}\varsigma^2(\tau)}{(1 - P_c)} \right)^{0,5} \right\},$$

$$\overline{w_\varsigma}: \left\{ |\varsigma(\tau)| > \left(\frac{2\mathbf{E}\varsigma^2(\tau)}{(1 - P_c)} \right)^{0,5} \right\},$$

$$w_\theta: \{ CoDisp(\tau) < \theta_\tau \}.$$

На основании неравенства Чебышева имеют место неравенства:

$$\mathbf{P}(\overline{w_m}) \le (1 - P_c)/2,\ \mathbf{P}(\overline{w_\varsigma}) \le (1 - P_c)/2. \qquad (4)$$

Используя совместно неравенства Буля и (4), получим:

$$\begin{aligned} &\mathbf{P}(w_m w_\varsigma) \ge 1 - (\mathbf{P}(\overline{w_m}) + \mathbf{P}(\overline{w_\varsigma})) \ge \\ &\ge 1 - ((1 - P_c)/2 + (1 - P_c)/2) \ge P_c. \end{aligned} \qquad (5)$$

Очевидна импликация:

$$w_m w_\varsigma \Rightarrow w_\theta. \qquad (6)$$

Из выражений (5) и (6) следует доказываемое утверждение.□

Использование θ_τ в качестве параметра «порог принятия решения», согласно сделанным предположениям и доказательству Теоремы, гарантирует, что

$$\begin{aligned} &\mathbf{P}(\Psi(t, \theta | \{z(t) | t \in T\}) = true | t \notin o_\varepsilon([\tau_1, \tau_2])) = \\ &= \mathbf{P}(CoDisp(\tau) \ge \theta_\tau | \tau < \tau_1) \le 1 - P_c. \end{aligned}$$

Если для заданного значения α выбрать $P_c = 1 - \alpha$, то

$$\mathbf{P}(\Psi(t, \theta | \{z(t) | t \in T\}) = true | t \notin o_\varepsilon([\tau_1, \tau_2])) \le \alpha.$$

В результате выполнено условие (1) постановки задачи.

При этом правило принятия решения $\Psi(\cdot)$ имеет вид:

$$\begin{cases} (\Psi(t, \theta | \{z(t) | t \in T\}) = true) \text{ если } (CoDisp(\tau) \ge \theta_\tau) \\ (\Psi(t, \theta | \{z(t) | t \in T\}) = false) \text{ если } (CoDisp(\tau) < \theta_\tau) \end{cases}.$$

Таким образом, требования постановки задачи выполнены: правило принятия решения сформулировано, а следование этому правилу гарантирует заданную верхнюю границу для вероятности ложной тревоги.

Так как $\mathbf{E}\varsigma^2(\tau)$ априорно неизвестно, эту величину следует оценить по доступным наблюдениям Φ_T. Для этих целей целесообразно использовать обычную выборочную, несмещенную оценку величины $\mathbf{E}\varsigma^2(\tau)$, которая эффективна для больших величин τ:

$$\begin{aligned} &\mathbf{Var}_\tau[CoDisp(\tau)] = \\ &= \sum_{t \le \tau} (CoDisp(\tau) - CoDisp(z(t) | \cdot))^2 (\tau - 1)^{-1}. \end{aligned}$$

В этом случае выражение для порога имеет следующий вид:

$$\tilde{\theta}_\tau = CoDisp(\tau) + (1 + \tau^{-0,5}) \left(\frac{2\mathbf{Var}_\tau[CoDisp(\tau)]}{(1 - P_c)} \right)^{0,5}. \qquad (7)$$

Для устранения негативного воздействия помех $\{\xi(t)\}$ и стабилизации метода RRCF, под которой подразумевается сохранение способности к обнаружению структурных аномалий в условиях искажения наблюдений аддитивным центрированным шумом с конечной дисперсией, предлагается применить цифровую фильтрацию к входному потоку данных. При этом, перед использованием метода RRCF, наблюдения $\{z(t)\}$ подвергаются обработке одним из заданного множества цифровых фильтров.

В качестве показателя, который характеризует стабильность RRCF при фиксированной дисперсии σ^2 аддитивного центрированного шума и использовании фильтра f предлагается использовать следующую метрику:

$$S^{(f)}(\sigma | P_\sigma^{(f)}, P_\sigma^{(0)}) = P_\sigma^{(f)} (\log_{10} 4) \log_{10} \left(\frac{2(1 + P_\sigma^{(f)})}{1 + P_\sigma^{(0)}} \right),$$

где $P_\sigma^{(f)}$ и $P_\sigma^{(0)}$ — вероятности обнаружения структурной аномалии методом RRCF в условиях искажения наблюдений аддитивным центрированным шумом с дисперсией σ^2 при использовании фильтра типа f и без использования фильтрации зашумленных наблюдений.

В дальнейшем, в том случае, когда это не вызывает неоднозначностей, вместо $S^{(f)}(\sigma | P_\sigma^{(f)}, P_\sigma^{(0)})$ будем использовать сокращенный вариант обозначения: $S^{(f)}(\sigma)$. С учетом того, что $P_\sigma^{(f)}, P_\sigma^{(0)} \in [0, 1]$ легко видеть, что $\forall: S^{(f)}(\sigma) \in [0, 1]$. Эта функция линейно зависит от $P_\sigma^{(f)}$ и ее величина пропорциональна логарифму величины $(1 + P_\sigma^{(f)})(1 + P_\sigma^{(0)})^{-1}$. Таким образом, величина $S^{(f)}(\sigma)$ тем больше, чем больше величина $P_\sigma^{(f)}$ превосходит $P_\sigma^{(0)}$, и наоборот. Другими словами, если вероятность обнаружения структурной аномалии с использованием фильтра f велика, и она выше, чем вероятность обнаружения этой аномалии без использования фильтра, то

показатель $S^{(f)}(\sigma)$ возрастает. Максимальное значение величины $S^{(f)}(\sigma)$ — единица. И наоборот, если вероятность обнаружения структурной аномалии с использованием фильтра f низка, и она ниже, чем вероятность обнаружения этой аномалии без использования фильтра, то показатель $S^{(f)}(\sigma)$ уменьшается. Минимальное значение величины $S^{(f)}(\sigma)$ — нуль. Таким образом, $S^{(f)}(\sigma)$ интерпретируемо характеризует эффективность использования фильтра f при стабилизации метода RRCF. Условимся называть эту метрику «обобщенным показателем стабилизации». В разделе «Численные исследования» показана иллюстрация использования различных типов фильтров f, а также предварительный анализ их эффективности.

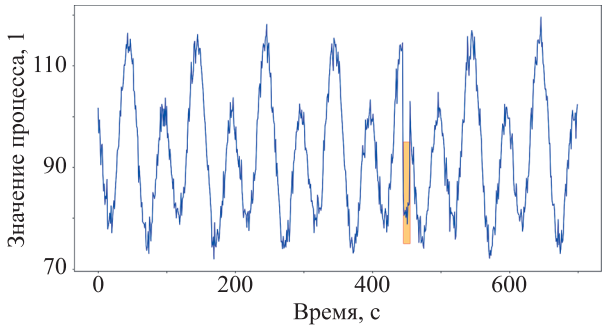

Рис. 1. Модельный процесс (при σ = 2) со слабо выраженной аномалией, которая выделена цветным прямоугольником

Fig. 1. Model process with a weak anomaly (highlighted by colored rectangle)

Численные исследования

Выбор конкретного фильтра зависит от специфики наблюдаемого процесса $\{z(t)\}$. Рассмотрим результаты использования технологии метода RRCF с параметрами: |F| = 120, tree_size = 140, shingle_size = 5. Выполним предварительную фильтрацию для следующего процесса:

— $\rho(t) = A\sin^2(Tt - T\varphi) + 0{,}5A\cos^2(T_1 t - T\varphi) + C + \xi(t)$, $A = 30$, $C = 70$, $\varphi = 20$, $T = 2\pi/100$, $T_1 = T/2$;

— $g(t) = G\cos^2(T_2 t) + \xi(t)$, $\tau_1 = 445$, $\tau_2 = 455$, $G = 90$, $T_2 = \pi/500$.

Здесь A, G — амплитудные параметры модели; t — время; C — константа уровня; T, T_1, T_2 и φ — параметры модели, определяющие ее частотно-фазовые характеристики.

Величины $\xi(t)$ распределены по нормальному закону с нулевым средним и среднеквадратическим отклонением σ, σ ∈ {0, 1, 2, 3, 4, 5, 6}. На рис. 1 представлен график этого процесса при σ = 2.

Как видно из рис. 1, аномалия **слабо выражена** и замаскирована аддитивным шумом. Исследуем распределение $CoDisp(z(t)|\cdot)$, $t \notin [\tau_1, \tau_2]$ для различных вариантов реализации цифрового фильтра. В данном эксперименте использованы следующие типы фильтров (табл. 1).

На рис. 2 представлены плотности распределения величины $CoDisp(\cdot)$ для всех вариантов фильтрации (табл. 1). Каждая секция полученных зависимостей содержит изображение плотностей распределения $CoDisp(\cdot)$, соответствующих конкретному фильтру f и σ ∈ {0, 1, 2, 3, 4, 5, 6}.

Из рис. 2 видно, что плотности распределения величины $CoDisp(\cdot)$ всегда имеют унимодальный тип, с выраженной правой асимметрией (right-skewed distribution, positive skewness). С увеличением дисперсии помехи асимметрия данного типа увеличивается. Зависимости показали, что различные фильтры влияют на плотность распределения $CoDisp(\cdot)$ различным образом, с тенденцией к обострению функции распределения в области максимума. Иначе говоря: чем лучше работает фильтр, тем уже становится распределение вероятности.

Визуально, результаты работы фильтров median и order_filter выглядят предпочтительнее. Что полностью подтверждается результатами моделирования, которые показаны на рис. 3 и в табл. 2. Для получения сравнительных данных по эффективности стабилизации метода RRCF для различных фильтров выполнена серия вычислительных экспериментов, при которых для фильтров f осуществлена оценка величин $P_\sigma^{(f)}$, $P_\sigma^{(0)}$, соответствующих σ ∈ {0, 1, 2, 3, 4, 5, 6}, а также вычислены значения $S^{(f)}(\sigma)$. Мощность каждой серии экспериментов, проводимой для уникальных f и σ, равна

Таблица 1. Используемые фильтры

Table 1. Digital filters used

Обозначение фильтра	Общие характеристики фильтра
symiirorder	сглаживающий IIR-фильтр (рекурсивный фильтр, БИХ-фильтр) с зеркально-симметричными граничными условиями с помощью каскада секций первого порядка. Параметры фильтра: C0: 2, Z1:0,01 [17]
lfilter	фильтр с конечной импульсной характеристикой (фильтр скользящего среднего). Размер гауссового окна: 4 [17]
order_filter	порядковый фильтр 4-го ранга. Маска фильтра: [–1, –1, –1, 0,1, 1,1] [17]
median	обычный медианный фильтр. Размер окна фильтра: 5 [17]
savgol	фильтр Savitzky-Golay, который применяется для сглаживания данных и устранения шума. Основан на использовании локальной полиномиальной аппроксимации. Параметры фильтра: длина окна — 12; порядок полинома — 10; режим расширения — nearest [18]
non	без фильтрации

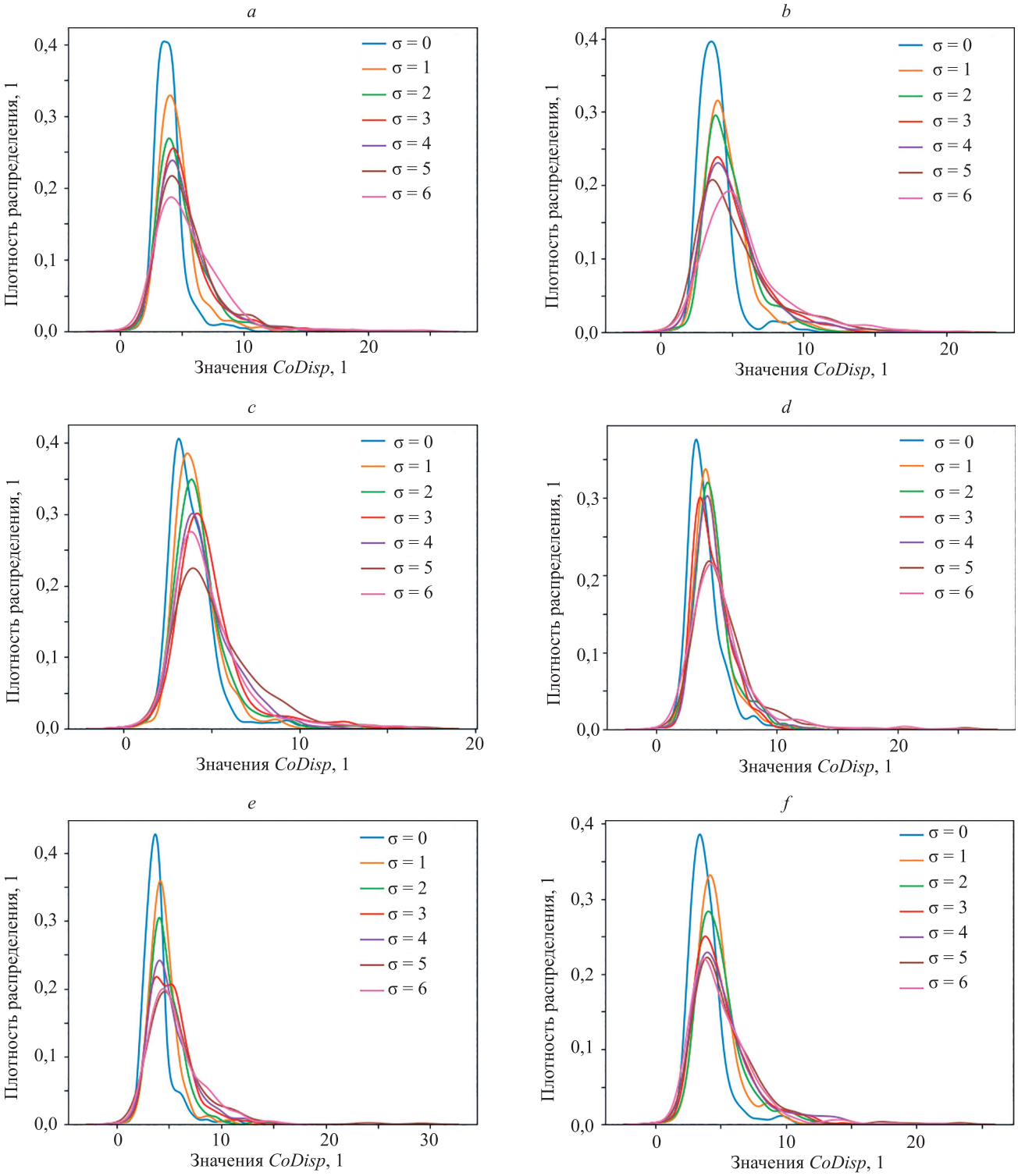

Рис. 2. Плотности распределения функции *CoDisp* при использовании различных фильтров для различных интенсивностей аддитивного шума, определяемых величиной σ: non (*a*); symiirorder (*b*); lfilter (*c*); order_filter (*d*); median (*e*); savgol (*f*)

Fig. 2. *CoDisp* distribution densities with different filters for various additive noise intensities defined by σ. Diagrams: no filters (*a*); symiirorder filter (*b*); lfilter filter (*c*); order_filter filter (*d*); median filter (*e*); savgol filter (*f*)

50. В табл. 2 представлены значения $P_{\sigma}^{(f)}$ для множества фильтров Fs = {non, savgol, lfilter, median, symiirorder, order_filter}, где non соответствует случаю отсутствия предварительной фильтрации. Зададим величину допустимой нижней границы $P_{per} \in [0, 1]$ для $P_{\sigma}^{(f)}$ на уровне 0,9, т. е. $P_{per} = 0{,}9$. В табл. 2 жирным шрифтом выделены значения $P_{\sigma}^{(f)}$, которые превышают P_{per}. Отметим, что данное представление результатов — одна из возможных форм представления факта обнаружения аномалии методом RRCF для разных вариантов реализации предобработки (фильтрации) входного потока измерений $\{z(t)\}$, соответствующих использованию

Таблица 2. Значения величины $P_{\sigma}^{(f)}$ для разных фильтров f и значений σ

Table 2. Values of $P_{\sigma}^{(f)}$ for various filters f and values of σ

Значение σ	Тип фильтра					
	non	savgol	lfilter	median	symiirorder	order_filter
0	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
1	0,81	**0,98**	**0,98**	**0,98**	**0,99**	**1,00**
2	0,32	**0,95**	**0,90**	**0,97**	**0,98**	**0,99**
3	0,00	0,30	0,20	0,35	0,40	**0,98**
4	0,00	0,10	0,05	0,03	0,20	**0,96**
5	0,00	0,00	0,00	0,00	0,05	0,35
6	0,00	0,00	0,00	0,00	0,00	0,00

различных типов фильтров $f \in Fs$. Полученные данные свидетельствуют: результаты обнаружения аномалии данного типа для рассмотренных вариантов предобработки $f \in Fs$, при повышении интенсивности шума (величины σ), далеки от идеального. В первую очередь это обусловлено тем, что смоделированная структурная аномалия довольно слаба на фоне воздействия аддитивного шума $\{\xi(t)\}$. Тем не менее, использование фильтра order_filter позволило устойчиво обнаружить аномалию для всех σ ∈ {0, 1, 2, 3, 4}. Ожидаемо наихудшие результаты соответствуют варианту non (полное отсутствие предобработки).

На рис. 3 представлены графики зависимостей $S^{(f)}(\sigma)$ от значений σ для фильтров $f \in Fs$. Серой пунктирной линией обозначена сглаженная допустимая нижняя граница величины обобщенного показателя стабилизации при $P_{per} = 0{,}9$ в виде $S_{per}^{(f)}(\sigma) = S^{(f)}(\sigma|\, P_{per}, P_{\sigma}^{(0)})$. Если $S^{(f)}(\sigma) \le S_{per}^{(f)}(\sigma)$, считается: при данном значении σ стабилизированный при помощи $f \in Fs$ метод RRCF — неэффективен. Из рис. 3 следует: наибольшую эффективность обеспечивает RRCF-обнаружитель, стабилизированный фильтром order_filter. В данном случае при $P_{per} = 0{,}9$ обеспечивается обнаружение структурной аномалии для всех σ ∈ {0, 1, 2, 3, 4}. Фильтры savgol, lfilter, median и symiirorder показали приблизительно равную эффективность стабилизации метода, обеспечивая обнаружение аномалии для всех σ ∈ {0, 1, 2}.

В процессе выполненных расчетов порог принятия решения об обнаружении аномалии выбирался согласно выражению (7). На рис. 4 представлены варианты реализации случайной величины $CoDisp(\cdot)$, соответствующие различным фильтрам $f \in Fs$ и σ помехового процесса. Зеленая пунктирная линия обозначает порог принятия решения θ. Область реализации аномалии выделена розовым цветом.

Представленные результаты доказывают: использование фильтрации в качестве предварительной обработки данных стабилизирует метод RRCF в условиях воздействия аддитивной помехи высокой интенсивности.

Пример практического использования

Стабилизированный метод RRCF применен для обнаружения аномальных вибраций трубопроводной конструкции в системе отвода шахтных вод в криолитозоне. Давление рассола, который отводится через эту

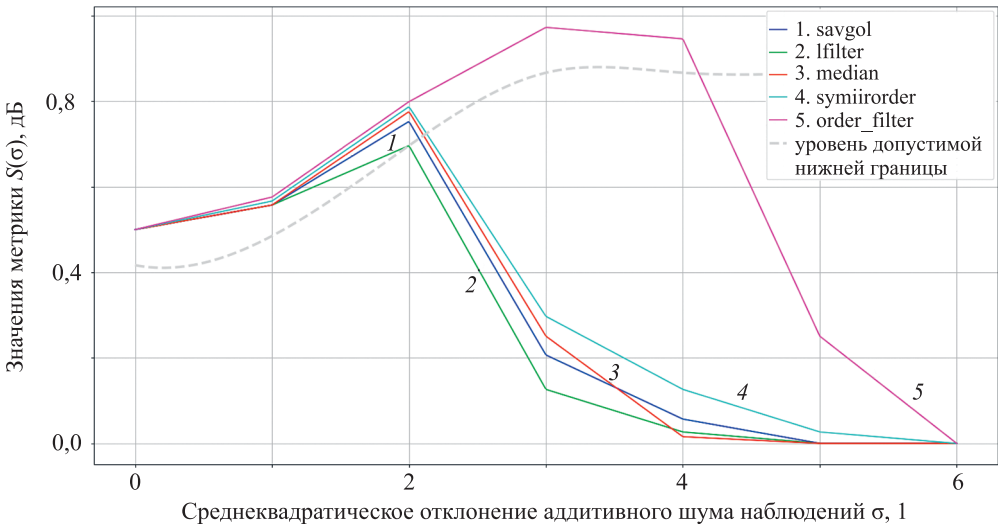

Рис. 3. Зависимости величины обобщенного показателя стабилизации $S^{(f)}(\sigma)$ от значений σ для фильтров $f \in Fs$

Fig. 3. Dependence $S^{(f)}(\sigma)$ vs. σ values for different $f \in Fs$

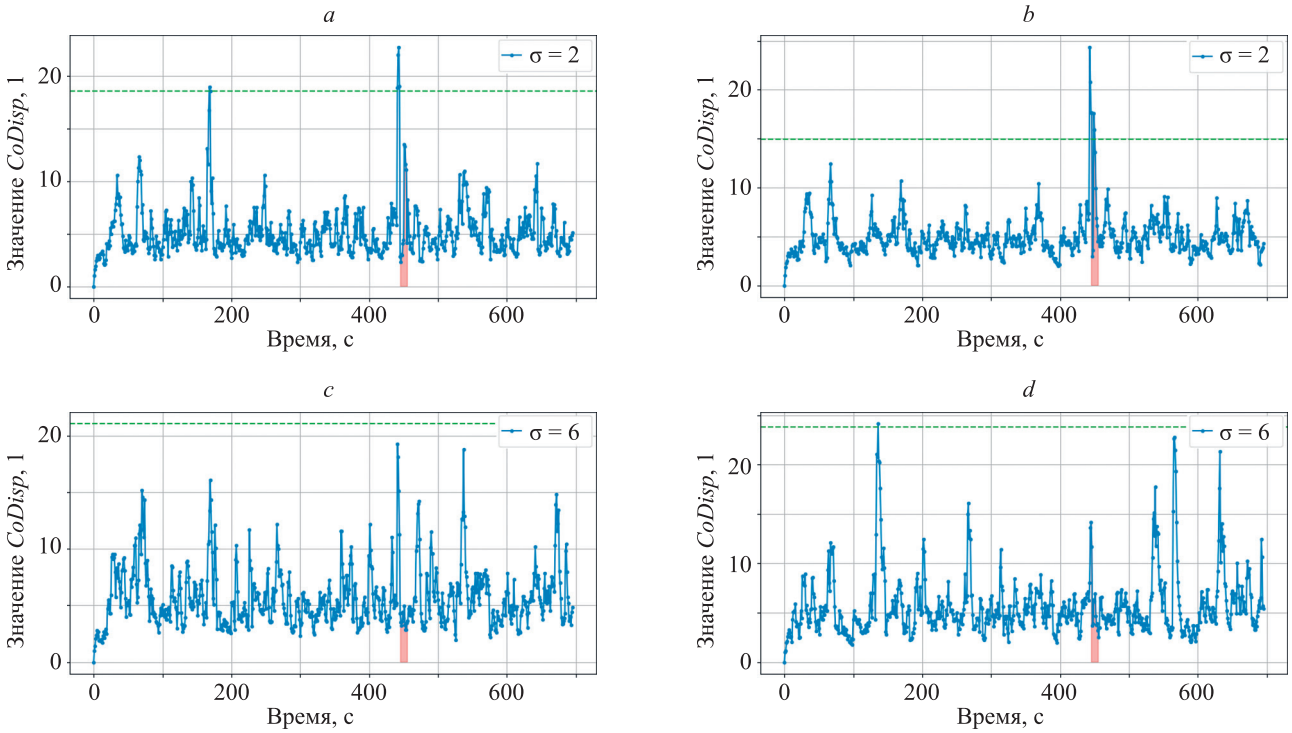

Рис. 4. Примеры реализации функции *CoDisp*, соответствующие различным значениям σ помехового процесса для фильтров: non, σ = 2 (*a*); order_filter, σ = 2 (*b*); median, σ = 6 (*c*); order_filter, σ = 6 (*d*)

Fig. 4. Examples of *CoDisp* realizations corresponding to different filters and different σ of the noise process: no filtering, σ = 2 (*a*); ordinal filter, σ = 2 (*b*); median filter, σ = 6 (*c*); ordinal filter, σ = 6 (*d*)

систему, в зимний период достигает 18 бар и более, а элементы трубопроводной конструкции часто расположены на неровной поверхности, с ярко выраженными спусками и подъемами. Общее напряженно-деформированное состояние конструкции изменяется в зависимости от состояния ее опор, степени изношенности элементов трубопроводной системы, технологических режимов перекачки, а также вследствие влияния иных факторов. О состоянии напряженно-деформированного состояния конструкции, согласно ГОСТ 57727-2007[1], можно объективно судить по характеру ее вибрации. В процессе оптоволоконного мониторинга вибрации трубопроводной конструкции [19] существует необходимость обнаруживать моменты смена режимов вибрации (МСРВ), которые происходят, например, из-за смены технологического режима прокачки рассола, в момент начала неуправляемого разрушения элемента конструкции или при изменении напряженно-деформированного статуса элемента конструкции во время проседания опоры на слабом грунте. Обнаружение МСРВ крайне важно для результатов мониторинга, поэтому эта задача выделяется в отдельный информационный процесс. При этом сами МСРВ, в зависимости от их причины, могут быть как достаточно частыми (регулярными), так и крайне редкими (проседание опоры или лавинообразный процесс разрушения конструкции). По этой причине, в базе данных наблюдений за динамикой вибрации конструкции трубопровода в основном присутствуют регулярные МСРВ. Именно для обнаружения этого типа МСРВ и был применен стабилизированный метод RRCF с конфигурацией |F| = 200 , tree_size = 150, shingle_size = 30, order_filter. Используются два независимых обнаружителя, периоды адаптации которых равны 30 мин, но сдвинуты друг относительно друга на 15 мин. Согласно используемым определениям, в процессе адаптации строится модель Φ_T. После окончания периода адаптации производится сброс настроек к начальным и процесс адаптации начинается вновь. Данная схема показала высокую практическую эффективность в условиях, когда вибрационные образы технологических режимов отличались сравнительно высокой нестабильностью вследствие специфики работы насосного оборудования и искажений, возникающих в измерительном канале. В результате многочисленных экспериментов было выяснено, что стабилизированный метод RRCF с вероятностью близкой к 100 % обнаруживает регулярные МСРВ, обеспечивая задержку принятия решения не хуже 10–15 с. Предложенный метод обеспечивал не более одного ложного срабатывания в сутки. Достигнутые показатели приемлемы практически и были достигнуты потому, что регулярные МСРВ, в отличие от ранее рассмотренного примера, соответствуют достаточно контрастным, скачкообразным изменениям множества параметров, характеризующих вибрацию конструкции и отражающихся в реализации наблюдаемого процесса $\{z(t)\}$.

[1] ГОСТ 57727-2007 Техническая диагностика. Акустико-эмиссионная диагностика. Общие требования. Введен 01.10.2007. М.: Издательство стандартов, 2007. 11 с.

Обсуждение

Как показали проведенные исследования, стабилизированный при помощи предварительной фильтрации метод RRCF представляет собой мощный метод для обнаружения аномалий в потоке данных, обладая способностью обнаруживать слабо выраженные аномалии структурного типа в потоке данных, искаженном аддитивной помехой высокой интенсивности. В рамках данного исследования был строго обоснован выбор порога принятия решения θ. Возможно, порог θ, вычисляемый в рамках предложенной процедуры, является чрезмерно осторожным, так как алгоритм выбора этого параметра основан на использовании неравенства Чебышева. Предположительно, чтобы выбрать порог θ более оптимально, необходимо вместо неравенства Чебышева использовать P_c-квантиль, построенный по выборочному распределению Φ_T. С другой стороны, стабилизация RRCF на базе фильтрации, несмотря на ее кажущуюся очевидность, должна быть изучена более глубоко: оптимально было бы сформулировать более формальные правила настройки фильтров в зависимости от статистических свойств Φ_T. Эти вопросы являются предметом дальнейших исследований.

Заключение

В работе исследована стабилизация и выбор параметров метода Robust Random Cut Forest (RRCF) для обнаружения аномальностей в потоке зашумленных данных. Предложенные методы позволяют сделать RRCF-обнаружитель более устойчивым к воздействию аддитивных помех, а также формализовать процедуру определения порога принятия решения, обеспечивающую верхнюю границу для вероятности ложной тревоги. Предложенная модификация метода RRCF была апробирована при решении реальной задачи, результаты апробации подтвердили ее практическую эффективность.

Литература

1. Gomes H.M., Read J., Bifet A. Streaming random patches for evolving data stream classification // Proc. of the IEEE International Conference on Data Mining (ICDM). 2019. P. 240–249. https://doi.org/10.1109/ICDM.2019.00034
2. Pang Z., Cen J., Yi M. Unsupervised concept drift detection method based on robust random cut forest // International Journal of Machine Learning and Cybernetics. 2023. V. 14. N 12. P. 4207–4222. https://doi.org/10.1007/s13042-023-01890-x
3. Zheng M., Geng L., Zuo B., Nakata T. A dynamic thresholds based anomaly detection algorithm in energy consumption process of industrial equipment // Proc. of the 2023 7th International Conference on Big Data and Internet of Things. 2023. P. 201–209. https://doi.org/10.1145/3617695.3617706
4. Marathe A. LRZ convolution: An algorithm for automatic anomaly detection in time-series data // Proc. of the 32nd International Conference on Scientific and Statistical Database Management. 2020. P. 1–12. https://doi.org/10.1145/3400903.3400904
5. Bohlke-Schneider M., Kapoor S., Januschowski T. Resilient neural forecasting systems // Proc. of the Fourth International Workshop on Data Management for End-to-End Machine Learning. 2022. P. 1–5. https://doi.org/10.1145/3399579.3399869
6. Тимофеев А.В. Обнаружение сигналов случайной формы при непараметрической априорной неопределенности относительно распределения наблюдений // Известия вузов. Радиоэлектроника. 1991. № 7. С. 64–68.
7. Timofeev A.V., Denisov V.M. Multimodal heterogeneous monitoring of super-extended objects: modern view. recent advances in systems safety and security // Studies in Systems, Decision and Control. 2016. V. 62. P. 97–116. https://doi.org/10.1007/978-3-319-32525-5_6
8. Gomes H., Read J., Bifet A., Barddal J., Gama J. Machine learning for streaming data: state of the art, challenges, and opportunities // ACM SIGKDD Explorations Newsletter. 2019. V. 21. N 2. P. 6–22. https://doi.org/10.1145/3373464.3373470
9. Tatbul N., Lee T., Zdonik S., Alam M., Gottschlich J. Precision and recall for time series // Advances in Neural Information Processing Systems. 2018. V. 31. P. 1924–1934.
10. Siddiqui M., Fern A., Dietterich T., Wright R., Theriault A., Archer D. Feedback-guided anomaly discovery via online optimization // Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018. P. 2200–2209. https://doi.org/10.1145/3219819.3220083
11. Hariri S., Kind M. Batch and online anomaly detection for scientific applications in a Kubernetes environment // Proc. of the 9th Workshop on Scientific Cloud Computing. 2018. P. 1–7. https://doi.org/10.1145/3217880.3217883
12. Salehi M., Rashidi L. A survey on anomaly detection in evolving data // ACM SIGKDD Explorations Newsletter. 2018. V. 20. N 1. P. 13–23. https://doi.org/10.1145/3229329.3229332

References

1. Gomes H.M., Read J., Bifet A. Streaming random patches for evolving data stream classification. *Proc. of the IEEE International Conference on Data Mining (ICDM)*, 2019, pp. 240–249. https://doi.org/10.1109/ICDM.2019.00034
2. Pang Z., Cen J., Yi M. Unsupervised concept drift detection method based on robust random cut forest. *International Journal of Machine Learning and Cybernetics*, 2023, vol. 14, no. 12, pp. 4207–4222. https://doi.org/10.1007/s13042-023-01890-x
3. Zheng M., Geng L., Zuo B., Nakata T. A dynamic thresholds based anomaly detection algorithm in energy consumption process of industrial equipment. *Proc. of the 2023 7th International Conference on Big Data and Internet of Things*, 2023, pp. 201–209. https://doi.org/10.1145/3617695.3617706
4. Marathe A. LRZ convolution: An algorithm for automatic anomaly detection in time-series data. *Proc. of the 32nd International Conference on Scientific and Statistical Database Management*, 2020, pp. 1–12. https://doi.org/10.1145/3400903.3400904
5. Bohlke-Schneider M., Kapoor S., Januschowski T. Resilient neural forecasting systems. *Proc. of the Fourth International Workshop on Data Management for End-to-End Machine Learning*, 2022, pp. 1–5. https://doi.org/10.1145/3399579.3399869
6. Timofeev A.V. Detection of randomly shaped signals under nonparametric a priori uncertainty about the distribution of observations. *Izvestija vuzov. Radiojelektronika*, 1991, no. 7, pp. 64–68. (in Russian)
7. Timofeev A.V., Denisov V.M. Multimodal heterogeneous monitoring of super-extended objects: modern view. recent advances in systems safety and security. *Studies in Systems, Decision and Control*, 2016, vol. 62, pp. 97–116. https://doi.org/10.1007/978-3-319-32525-5_6
8. Gomes H., Read J., Bifet A., Barddal J., Gama J. Machine learning for streaming data: state of the art, challenges, and opportunities. *ACM SIGKDD Explorations Newsletter*, 2019, vol. 21, no. 2, pp. 6–22. https://doi.org/10.1145/3373464.3373470
9. Tatbul N., Lee T., Zdonik S., Alam M., Gottschlich J. Precision and recall for time series. *Advances in Neural Information Processing Systems*, 2018, vol. 31, pp. 1924–1934.
10. Siddiqui M., Fern A., Dietterich T., Wright R., Theriault A., Archer D. Feedback-guided anomaly discovery via online optimization. *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2200–2209. https://doi.org/10.1145/3219819.3220083
11. Hariri S., Kind M. Batch and online anomaly detection for scientific applications in a Kubernetes environment. *Proc. of the 9th Workshop on Scientific Cloud Computing*, 2018, pp. 1–7. https://doi.org/10.1145/3217880.3217883
12. Salehi M., Rashidi L. A survey on anomaly detection in evolving data. *ACM SIGKDD Explorations Newsletter*, 2018, vol. 20, no. 1, pp. 13–23. https://doi.org/10.1145/3229329.3229332

13. Guha S., Mishra N., Roy G., Schrijvers O. Robust random cut forest based anomaly detection on streams // Proceedings of Machine Learning Research. 2016. V. 46. P. 2712–2721.
14. Breiman L. Bagging predictors // Machine Learning. 1996. V. 24. N 2. P. 123–140. https://doi.org/10.1007/bf00058655
15. Putina A., Rossi D. Online anomaly detection leveraging stream-based clustering and real-time telemetry // IEEE Transactions on Network and Service Management. 2021. V. 18. N 1. P. 839–854. https://doi.org/10.1109/TNSM.2020.3037019
16. Vardhan H., Sztipanovits J. Reduced robust random cut forest for out-of-distribution detection in machine learning models // ArXiv. 2022. arXiv:2206.09247. https://doi.org/10.48550/arXiv.2206.09247
17. Arce G.R. Nonlinear Signal Processing: A Statistical Approach. Wiley, 2005. 480 p.
18. Savitzky A., Golay M.J.E. Smoothing and differentiation of data by simplified least squares procedures // Analytical Chemistry. 1964. V. 36. N 8. P. 1627–1639. https://doi.org/10.1021/ac60214a047
19. Тимофеев А.В., Максимов П.Н., Грознов Д.И. Применение оптоволоконной технологии для мониторинга трубопроводных систем отведения шахтных вод в криолитозоне // Гидротехника. 2023. № 3. С. 34–43. https://doi.org/10.55326/22278400_2023_3_34

13. Guha S., Mishra N., Roy G., Schrijvers O. Robust random cut forest based anomaly detection on streams. *Proceedings of Machine Learning Research*, 2016, vol. 46, pp. 2712–2721.
14. Breiman L. Bagging predictors. *Machine Learning*, 1996, vol. 24, no. 2, pp. 123–140. https://doi.org/10.1007/bf00058655
15. Putina A., Rossi D. Online anomaly detection leveraging stream-based clustering and real-time telemetry. *IEEE Transactions on Network and Service Management*, 2021, vol. 18, no. 1, pp. 839–854. https://doi.org/10.1109/TNSM.2020.3037019
16. Vardhan H., Sztipanovits J. Reduced robust random cut forest for out-of-distribution detection in machine learning models. *ArXiv*, 2022, arXiv:2206.09247. https://doi.org/10.48550/arXiv.2206.09247
17. Arce G.R. *Nonlinear Signal Processing: A Statistical Approach*. Wiley, 2005, 480 p.
18. Savitzky A., Golay M.J.E. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 1964, vol. 36, no. 8, pp. 1627–1639. https://doi.org/10.1021/ac60214a047
19. Timofeev A.V., Maksimov P.N., Groznov D.I. Application of fiber optic technology for monitoring the mine water drainage pipeline system in the permafrost zone. *The Hydrotechnika*, 2023, no. 3, pp. 34–43. (in Russian). https://doi.org/10.55326/22278400_2023_3_34

Автор

Тимофеев Андрей Владимирович — доктор технических наук, научный директор, ТОО «Эквалайзум», Астана, 010000, Казахстан, SC 56689367600, https://orcid.org/0000-0001-7212-5230, timofeev.andrey@gmail.com

Author

Andrey V. Timofeev — D.Sc., Chief Scientific Officer, LLP "EqualiZoom", Astana, 010000, Kazakhstan, SC 56689367600, https://orcid.org/0000-0001-7212-5230, timofeev.andrey@gmail.com

Статья поступила в редакцию 15.01.2024
Одобрена после рецензирования 04.02.2024
Принята к печати 14.03.2024

Received 15.01.2024
Approved after reviewing 04.02.2024
Accepted 14.03.2024