

ВЫБОР ПАРАМЕТРОВ ВЛОЖЕНИЯ ВРЕМЕННЫХ РЯДОВ ДЛЯ ВЫЯВЛЕНИЯ АНОМАЛИЙ ДАННЫХ В СИСТЕМАХ ПОВЕДЕНЧЕСКОЙ АНАЛИТИКИ

Нашивочников Н. В.¹, Меженева И. О.², Чатоян С. К.³

DOI: 10.21681/2311-3456-2025-2-20-28

Цель статьи: исследование влияния выбора алгоритма определения параметров вложения скалярных временных рядов на результативность обнаружения аномалий в системах поведенческой аналитики и выработка рекомендаций по их выбору.

Метод: исследование базируется на теоремах Такенса-Мане о реконструкции аттракторов динамических систем по скалярным наблюдаемым, методах топологического анализа данных и классификации (методах машинного обучения).

Полученный результат: экспериментально установлено, что результативность решения задач обнаружения аномалий в системах поведенческой аналитики методами машинного обучения и топологического анализа данных существенно зависит от качества построения реконструированного пространства по скалярной наблюдаемой, выработаны рекомендации по выбору параметров вложения – размерности реконструируемого пространства и временной задержки – в зависимости от используемого метода вложения.

Научная новизна: в отличие от известных исследований по тематике оптимального выбора параметров вложения временных рядов в задачах прогнозирования поведения (функционирования) динамических систем, это первая научная работа, в которой в прямой постановке исследовалось применение популярных алгоритмов выбора параметров вложения временных рядов для повышения точности выявления аномалий системами поведенческой аналитики с использованием топологических дескрипторов.

Ключевые слова: скалярные временные ряды, размерность вложения, задержка вложения, топологические дескрипторы, машинное обучение, классификация, HAI Security Dataset, обнаружение аномалий.

Введение

В системах поведенческой аналитики (User and Entity Behavior Analytics (UEBA)) одной из ключевых задач является задача формирования (вычисления) дескрипторов, описывающих как базовый, так и текущий (наблюдаемый) профили поведения пользователя/сущности. Исходной информацией для решения этой задачи, как правило, выступают облака данных – неупорядоченные наборы, не привязанные к какой-либо шкале измерений⁴ [1]. Одной из разновидностей таких наборов данных, наиболее часто встречающейся на практике, являются временные ряды [2].

Не ограничивая общности, в дальнейшем будем считать, что для формирования дескрипторов, отвечающих различным профилям поведения наблюдаемых объектов, используются подходы, основанные на топологическом анализе данных (Topological Data Analysis (TDA)) [4]. В работах [4–6] показано, что указанная технология в целом не предназначена для работы непосредственно с временными

рядами – она применяется к так называемым пространствам вложения⁵, которые специальным образом сопоставляются соответствующим наборам данных. В этой связи возникает закономерный вопрос, как параметры вложения влияют на качество решения задачи выявления аномалий (паттернов) в поведении пользователей/сущностей с использованием, например, таких топологических дескрипторов, как кривые Бетти⁶ [7,8].

1. Методы и алгоритмы выбора параметров вложения скалярных временных рядов в системах поведенческой аналитики

Источниками данных для систем поведенческой аналитики, как правило, являются различные источники, генерирующие данные в виде коллекций временных рядов. Каждый временной ряд из такой коллекции соответствует тому или иному аспекту⁷ поведения наблюдаемого объекта, а вся коллекция характеризует его поведение в целом. Причем,

1. Нашивочников Николай Васильевич, CISSP, заместитель генерального директора – технический директор ООО «Газинформсервис», г. Санкт-Петербург, Россия. E-mail: cto@gaz-is.ru

2. Меженева Ирина Олеговна, ведущий инженер аналитик лаборатории технологий искусственного интеллекта аналитического центра кибербезопасности ООО «Газинформсервис», г. Санкт-Петербург, Россия. E-mail: mezheneva-i@gaz-is.ru

3. Чатоян Сергей Камалович, кандидат технических наук, старший научный сотрудник, главный специалист лаборатории технологий искусственного интеллекта аналитического центра кибербезопасности ООО «Газинформсервис», г. Санкт-Петербург, Россия. E-mail: chatoyan-s@gaz-is.ru

4. Jiawei Han, Micheline Kamber, Jian Pei, Data mining: concepts and techniques. – Morgan kaufmann. – 3rd ed. – 740 S. ISBN 978-0-12-381479-1.

5. Псевдофазовым, реконструированным пространствам [3].

6. В этом смысле данная статья является естественным продолжением работы [6].

7. Предполагается, что каждому аспекту поведения соответствует свои, присущие только ему, показатели.

как показывают результаты предварительной обработки таких временных рядов, они в большинстве своем являются персистентными. Исходя из этого наблюдения, для вложения временных рядов в псевдофазовое (реконструируемое) пространство целесообразно использовать методы нелинейной динамики, в частности, теоремы Такенса-Мане [9,10]. Известно, что в этом случае необходимо решать задачу по определению оптимальных параметров вложения: размерности пространства вложения d и временной задержки τ [11, 12]. Следует подчеркнуть, что каких-либо однозначных критериев оптимального выбора указанных параметров в настоящее время не существует [13]. Более того, в работе⁸ указывается на неоднозначность в определении относительной важности между размерностью d и задержкой вложения τ , причем была даже высказана гипотеза о том, что эти параметры являются зависимыми. В этой связи в качестве параметра при оптимизации процедур вложения предлагается использовать окно реконструкции (вложения) $w_{d,\tau}$, которое определяется следующим образом:

$$w_{d,\tau} = (d - 1) \cdot \tau.$$

Неформальную рекомендацию по его выбору можно сформулировать следующим образом: оптимальное окно вложения $w_{d,\tau}$ должен быть таким, чтобы оно было не слишком коротким (возникает большое количество ложных соседей) и не слишком длинным (наблюдается эффект ложного увеличения размерности). Данная рекомендация справедлива для персистентных процессов, характерных для задач поведенческой аналитики [4].

В системах поведенческой аналитики, использующих алгоритмы TDA, зачастую важно обеспечить минимальную размерность реконструируемого псевдофазового пространства Такенса⁹. Тогда задача выбора окна вложения может быть сведена к задаче выбора оптимальной задержки вложения τ^* и проверке достоверности (корректности) вложения. В свою очередь методы, используемые для выбора оптимального параметра τ , отражают дихотомию в понятии качества вложения и их можно разделить на две категории: первая группа методов основана на динамических и информационно-теоретических подходах, а вторая – на топологических.

Методы, основанные на динамическом подходе¹⁰, сосредоточены на изучении механизмов генерации данных и взаимосвязей измерений¹¹. В этом случае

простейшим вариантом является равномерное вложение, когда при фиксированном значении параметра d задержка τ также выбирается постоянной. При равномерном вложении используются такие методы, как методы автокорреляционных функций и их нелинейных обобщений¹², функций минимальной взаимной информации $I(\tau, \cdot)$ ¹³, статистики непрерывности¹⁴, L -статистики¹⁵ и т.д. Необходимо отметить, что равномерное вложение ограничивает возможность выделять особенности в нескольких несопоставимых временных масштабах [14], а качество вложения может варьироваться в различных областях реконструированного аттрактора^{10, 17}.

Неравномерное вложение было предложено как естественное расширение с целью устранения некоторых недостатков равномерного вложения. Один из самых ранних методов выбора неравномерных задержек, предложенный в работе¹⁶, представляет собой вариацию метода ближайших соседей¹⁵. В рамках данного метода задержка выбирается рекурсивно с помощью так называемой N -статистики¹⁷ в течение нескольких циклов:

$$N(\tau) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(d_{2,\tau}^{(i)} > \alpha),$$

где N – длина временного ряда; $\mathbb{I}$ – индикаторная функция; α – некоторый порог¹⁸; $d_{1,\tau}$ – евклидово расстояние между парами ближайших соседей в реконструированном пространстве; $d_{2,\tau}$ – евклидово расстояние между парами ближайших соседей в реконструированном пространстве состояний на один шаг вперед по времени. Временной лаг, соответствующий первому минимуму функции $N(\tau)$, принимается за лаг вложения. В работе¹⁹ использовалась статистика непрерывности как способ построения неравномерных векторов задержки, основанный на идее функциональной независимости между их координатами. Функциональная зависимость между

8 Casdagli M. et al. State space reconstruction in the presence of noise // Physica D: Nonlinear Phenomena. – 1991. – Т. 51. – №. 1–3. – С. 52–98. DOI: 10.1016/0167-2789(91)90222-U.

9 Это связано с оптимизацией вычислительной сложности используемых алгоритмов.

10 Точнее – на методах нелинейной динамики.

11 К данной категории обычно относят также информационно-теоретические и статистические методы, например, такие как автокорреляционный метод, метод минимальной взаимной информации и другие¹⁶.

12 Fraser A. M., Swinney H. L. Independent coordinates for strange attractors from mutual information // Physical review A. – 1986. – Т. 33. – №. 2. – С. 1134. DOI: 10.1103/PhysRevA.33.1134.

13 Garcia S. P., Almeida J. S. Nearest neighbor embedding with different time delays // Physical Review E. – 2005. – Т. 71. – №. 3. – С. 037204. DOI: 10.1103/PhysRevE.71.037204.

14 Pecora L. M. et al. A unified approach to attractor reconstruction // Chaos: An Interdisciplinary Journal of Nonlinear Science. – 2007. – Т. 17. – №. 1. DOI: 10.1063/1.2430294.

15 Uzal L. C., Grinblat G. L., Verdes P. F. Optimal reconstruction of dynamical systems: A noise amplification approach // Physical Review E. – 2011. – Т. 84. – №. 1. – С. 016223. DOI: 10.1103/physreve.84.016223.

16 Kennel M. B., Brown R., Abarbanel H. D. I. Determining embedding dimension for phase-space reconstruction using a geometrical construction // Physical review A. – 1992. – Т. 45. – №. 6. – С. 3403. DOI: 10.1103/PhysRevA.45.3403.

17 N -статистика определяется как доля точек реконструированного аттрактора, у которых отношение евклидовых расстояний между парами ближайших соседей в реконструированном аттракторе.

18 Параметр α выбирается эвристически. В работе¹⁷ он принимается равным 10.

19 Pecora L. M. et al. A unified approach to attractor reconstruction // Chaos: An Interdisciplinary Journal of Nonlinear Science. – 2007. – Т. 17. – №. 1. DOI: 10.1063/1.2430294.

координатами вектора подразумевает, что существует произвольная гладкая функция $F(\cdot)$:

$$x(t - \tau_m) = F(x_m(t)), \quad (1)$$

где $x_m(t) \triangleq (x(t), x(t - \tau_1), \dots, x(t - \tau_m))^T \in \mathbb{R}^m$; $x(t - \tau_m) \in \mathbb{R}^1$.

В рамках данного метода построение неравномерного вложения с задержкой требует итерационного построения вектора $\tau^* = \{\tau_1, \dots, \tau_m\}^T \in \mathbb{R}^m$, который минимизирует вероятность функциональной зависимости между компонентами вектора задержки. На каждой итерации перспективная задержка τ_i проверяется на функциональную зависимость с существующими компонентами вектора τ . Если «существенной» функциональной зависимости не обнаружено, то задержка τ_i может быть добавлена в коллекцию задержек. Другими словами, рассмотрим m мерные векторы $x_m(t) \in \mathbb{R}^m$, построенные с использованием задержек $\{\tau_1, \dots, \tau_m\}$, и потенциальную новую задержку τ_{m+1} . Чтобы проверить выполнение равенства (1) в пространстве вложения, выбирается опорная точка $x_m(t_0) \in \mathbb{R}^m$. Если равенство (1) с гладкой функцией $F(\cdot)$ выполняется, то из непрерывности функции $F(\cdot)$ следует, что любые векторы $x_{m,i}(t_0) \in O_{x_m(t_0)} \subset \mathbb{R}^m$, где

$$O_{x_m(t_0)} = \{x_{m,i} \in \mathbb{R}^m \mid \|x_{m,i}(t_0) - x_m(t_0)\| < \delta, \delta \in \mathbb{R}_+^1\},$$

должны иметь компоненту τ_{m+1} :

$$|x_{m,i}(t_0 - \tau_{m+1}) - x_m(t_0 - \tau_{m+1})| < \varepsilon, \varepsilon \in \mathbb{R}_+^1.$$

Доля p точек $x_{m,i} \in \mathbb{R}^m$, удовлетворяющих данному условию, сравнивается с нулевой гипотезой, согласно которой взаимосвязь между этими наборами является случайной. Большие значения показателя p указывают на сильную связь между m мерной реконструкцией и новой компонентой задержки τ_{m+1} . Чтобы проверить, существует ли функциональная зависимость относительно некоторого выбранного параметра $\varepsilon > 0$, в работе²⁰ предлагается использовать биномиальное распределение с критическим значением $p^* = 0,5^{[20]}$. Для проверяемого значения τ_{m+1} и данной выборки точек ряда тест на непрерывность применяется с уменьшающимися значениями ε до тех пор, пока нулевая гипотеза не будет отвергнута. Наименьшее возможное значение для отклонения нулевой гипотезы задается как ε^* . Это значение усредняется по совокупности контрольных точек, отобранных из данных, чтобы вычислить статистику непрерывности $\langle \varepsilon^* \rangle(\tau_{m+1})$. На каждой итерации при выборе τ_{m+1} для существующей коллекции задержек $\tau = \{\tau_1, \dots, \tau_m\}$ вычисляется профиль статистики непрерывности $\langle \varepsilon^* \rangle(\tau_{m+1})$. В качестве новой задержки τ_{m+1} берется задержка, соответствующая относительно

максимуму профиля статистики непрерывности. Эта процедура повторяется до тех пор, пока не будет достигнута желаемая размерность реконструированного евклидового пространства.

Метод, предложенный в работе²¹ (он получил название метод L -статистики) базируется на концепции усиления шума при реконструкции. Данная концепция была сформулирована в работе²² в попытке количественно оценить качество вложения. Она основана на предположении о том, что «хорошее» вложение должно давать и «хороший» прогноз. Кроме того, такое вложение должно работать относительно «хорошо» даже в присутствии шума в данных. Количественно усиление шума (для данного вектора реконструкции $x(t)$) определяется как ошибка прогноза на T шагов при наличии шума. В общем случае она определяется следующим образом:

$$\sigma(T, x) = \lim_{\varepsilon \rightarrow 0} \sigma_\varepsilon(T, x),$$

$$\sigma_\varepsilon(T, x) = \frac{1}{\varepsilon} \sqrt{\text{Var}[x(T) \mid y = B_\varepsilon(x)]},$$

где $\text{Var}[x(T) \mid y = B_\varepsilon(x)]$ – условная дисперсия при прогнозе на T шагов от начального состояния x в пространстве вложений $\mathbb{R}^m$, «загрязненном» добавленным «малым» шумом наблюдений – ε , $y = B_\varepsilon(x)$ – «восстановленный» вектор состояния в исходном аттракторе при заданном уровне шума ε . Считается, что прогнозы не имеют модельных ошибок. Это предположение выполняется за счет выбора ближайших соседей во вложении $\mathbb{R}^m$ в качестве косвенного признака зашумленных начальных условий. Чтобы вычислить усиление шума σ_Σ , ошибка $\sigma(T, x)$ усредняется по набору опорных точек, отобранных по всему временному ряду. Вложения с высоким усилением шума σ_Σ означают, что ближайшие соседи в пространстве вложений $\mathbb{R}^m$ имеют тенденцию к расхождению траекторий в будущем (наблюдается рост старшего показателя Ляпунова), поскольку они не соответствуют реальным соседям в истинном аттракторе. Таким образом, влияние шума значительно усиливается, поскольку небольшие возмущения в реконструированном пространстве $\mathbb{R}^m$ приводят к большим неопределенностям в истинном состоянии системы. Одним из недостатков концепции усиления шума является необходимость определения горизонта прогнозирования, для которого и рассчитывается усиление шума. В работе²³ было предложено модифицировать выражение для $\sigma(T, x)$ следующим образом:

21 Fraser A.M., Swinney H.L. Independent coordinates for strange attractors from mutual information //Physical review A. – 1986. – Т. 33. – №. 2. – С. 1134. DOI: 10.1103/PHYSREVA.33.1134.

22 Casdagli M. et al. State space reconstruction in the presence of noise // Physica D: Nonlinear Phenomena. – 1991. – Т. 51. – №. 1-3. – С. 52-98. DOI: 10.1016/0167-2789(91)90222-U.

23 Uzal L.C., Grinblat G.L., Verdes P.F. Optimal reconstruction of dynamical systems: A noise amplification approach //Physical Review E. – 2011. – Т. 84. – №. 1. – С. 016223. DOI: 10.1103/physreve.84.016223.

20 Это распределение предлагается использовать благодаря его простоте и устойчивости к шуму.

Таблица 1.

Характеристики набора данных HAI

Версия набора данных	Количество характеристик (датчиков), шт.	Скорость измерений, точек/с	Данные нормального поведения (имя файла; интервал, размер)	Данные, содержащие атаки (имя файла; интервал; размер; количество атак)
HAI 23.05	84	86	hai-train3.csv; 35 часов; 69.4 Мбайт	hai-test1.csv; 15 часов; 29.8 Мбайт; 14 атак

$$\sigma(T_M, x) = \frac{1}{T_M} \int_0^{T_M} \sigma_\varepsilon^2(T, x) dT,$$

где T_M – максимальный горизонт прогнозирования.

Оценка $\sigma(T_M, x)$ оказывается относительно устойчивой для достаточно больших T_M . Алгоритм, используемый для вычисления σ_Σ , основан на использовании k ближайших соседей для опорной точки x_i в качестве прокси. В зависимости от распределения точек это может привести к тому, что эффективный уровень шума для каждой точки будет разным. Поэтому в работе для учета данной вариации было предложено использовать следующую постоянную нормализации:

$$\alpha_k^2 = \left[\sum_i \varepsilon_k^2(x_i) \right]^{-1}.$$

Объединив эти две идеи, в работе выдвинута гипотеза, что усиление шума σ_Σ измеряет некоторое понятие избыточности, а α_k – измеряет некоторое понятие нерелевантности. Исходя из этого предположения, L -статистика трактуется как функция затрат, а параметры реконструкции определяются в результате решения следующей задачи оптимизации:

$$L = \log(\sigma_\Sigma \cdot \alpha_k) = \log(\sigma_\Sigma) + \log(\alpha_k) \rightarrow \min.$$

Недостаток метода статистики непрерывности заключается в неоднозначности выбора оптимальной задержки τ на каждой итерации вложения. Кроме того, метод не учитывает эффекты от выбора различных метрик, используемых для определения близких точек в реконструкции. Наконец, статистика недостаточной выборки, первоначально предложенная в качестве условия завершения алгоритма вложения, требует больших вычислительных затрат и не дает информации о том, какую из перспективных задержек следует выбрать. В работе [15] был предложен метод PECUZAL, объединяющий статистику непрерывности с L -статистикой. Алгоритм использует статистику непрерывности для прямого перебора множества временных задержек и определения потенциальных задержек, соответствующих локальным максимумам профиля статистики непрерывности $\langle \varepsilon^* \rangle(\tau_{m+1})$. Затем L -статистика используется для выбора перспективной задержки в каждом цикле вложения, обеспечивая наибольшее снижение L -статистики.

Метод максимальных производных по проекции (MDOP) – это геометрическая альтернатива статистическим и информационно теоретическим подходам, предложенная в работе²⁴. В рамках данного метода оптимизируются параметры вложения, основываясь на критерии, что реконструированный аттрактор должен быть максимально развернут, с одной стороны, и минимизировалась избыточность в компонентах задержки, с другой стороны. Исходя из этого, вектор задержки строится рекурсивно через циклы вложения с использованием производной по направлению от некоторой функции, заданной на реконструированном пространстве состояний. Процесс циклически повторяется до достижения желаемой размерности вложения, выбираемой, например, с использованием метода ложных ближайших соседей.

2. Экспериментальное исследование методов и алгоритмов построения пространств вложения

Предварительно заметим, что для выбора того или иного метода построения пространств вложения необходимо на их множестве задать критерий (соответственно, отношение предпочтения), в соответствии с которым будет осуществляться окончательный выбор. При этом надо иметь в виду, что выбор такого критерия (критериев) представляет собой самостоятельную задачу, решение которой, при прочих равных условиях, зависит как от используемых топологических дескрипторов, так и от анализируемого набора данных. В дальнейшем в качестве топологических дескрипторов используются кривые Бетти, а в качестве набора данных – общедоступный набор данных HAI²⁵, собранный на испытательном стенде, который имитирует работу электростанций. Набор данных включает данные с датчиков при обычных операциях и при атаках²⁶. В таблице 1 приведены основные характеристики набора данных.

На рисунке 1 представлены примеры временных рядов нормального поведения и с атаками. Каждая

²⁴ Nichkawde C. Optimal state-space reconstruction using derivatives on projected manifold // Physical Review E. – 2013. – Т. 87. – №. 2. – С. 022905. DOI: 10.1103/PhysRevE.87.022905.

²⁵ HAI Security Dataset // Kaggle URL: <https://www.kaggle.com/datasets/icsdataset/hai-security-dataset/data> (дата обращения: 22.05.2024).

²⁶ При атаках злоумышленник вызывает постоянные ошибки в физических измерениях, манипулируя пакетами.

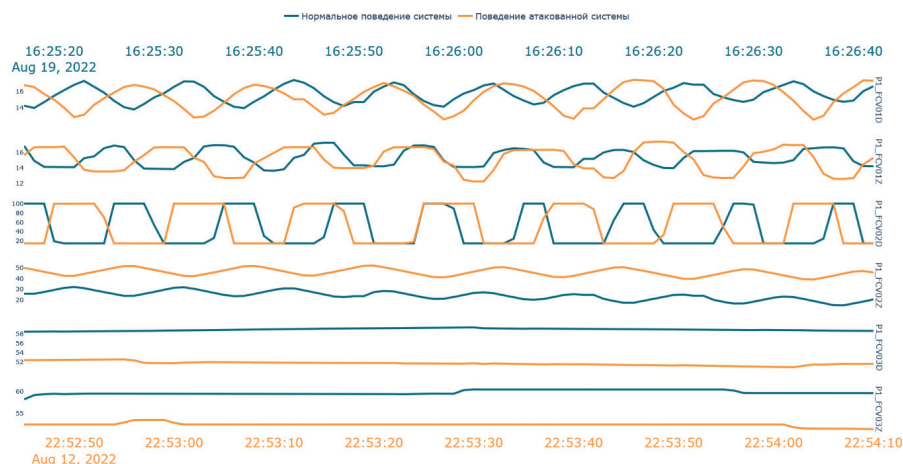


Рис. 1. Временные ряды поведения системы

точка на графике соответствует одному измерению (86 точек/с).

Набор данных не сбалансирован, с низкой долей количества аномалий, что затрудняет прогнозирование положительных классов. В этой связи в качестве критериев (метрик) как для оценки эффективности моделей выявления (распознавания) аномалий, так и для выбора метода построения пространств вложения были выбраны метрики Precision (точность), Recall (полнота), F_1 score (F -мера), False Positive Rate (уровень ложноположительных результатов, FPR) и False Negative Rate (уровень ложных отрицательных результатов, FNR) [16].

В исследовании каждая характеристика набора данных является отдельным временным рядом, который для анализа делится на сегменты по 86 точек, преобразующихся в облака точек методом вложений Такенса с различными алгоритмами выбора параметров. Описание методики анализа данных, основанной на топологических конструкциях²⁷, представлено в [4,5,7]. В качестве примера на рисунках 2–4 представлены соответственно временной ряд и полученные при исследовании алгоритма с параметрами вложения временных рядов с помощью метода L -статистики облако точек и кривые Бетти.

Полученные в ходе анализа метрики использовались для классификации с применением классических методов: К ближайших соседей (KNC), метод опорных векторов (SVC), классификатор гауссовских процессов (GPC), деревья решений (DTC), метод случайного леса (RFC), многослойный перцептрон (MLPC), гауссовский байесовский классификатор (GNB) [17]. Для каждого метода приводятся оценки качества моделей классификации.

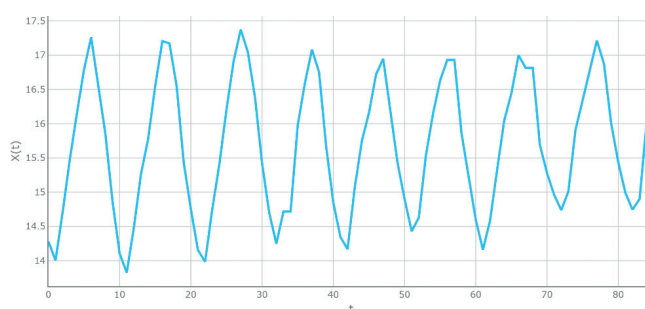


Рис. 2. Временной ряд

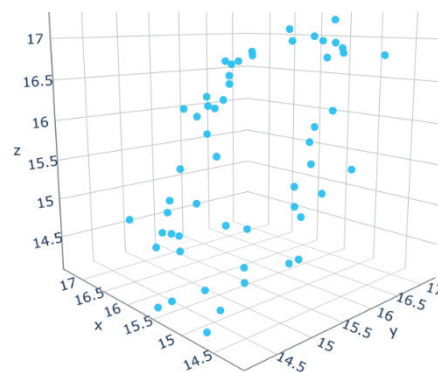


Рис. 3. Облако точек

В таблице 2 показаны результаты оценки качества с выбором параметров методом ложных ближайших соседей.

В таблице 3 представлены результаты оценки качества с выбором параметров методом статистики непрерывности.

В таблице 4 показаны результаты оценки качества с выбором параметров методом L -статистики.

В таблице 5 представлены результаты оценки качества с выбором параметров методом PECUZAL.

В таблице 6 приведены результаты оценки качества с выбором параметров методом MDOP.

²⁷ Математическое описание и интерпретация конструкций TDA, используемых в настоящем исследовании, в необходимом и достаточном для понимания объеме приведены в работе: Нашивочников, Н.В., Чатоян, С.К., Лукашин, А.А., Заборовский, В.С., Введение в топологический анализ данных: учебное пособие. – 2023. – СПб: ПОЛИТЕХ ПРЕСС – 244 с. DOI: 10.18720/SPBPU/2/id23-431.

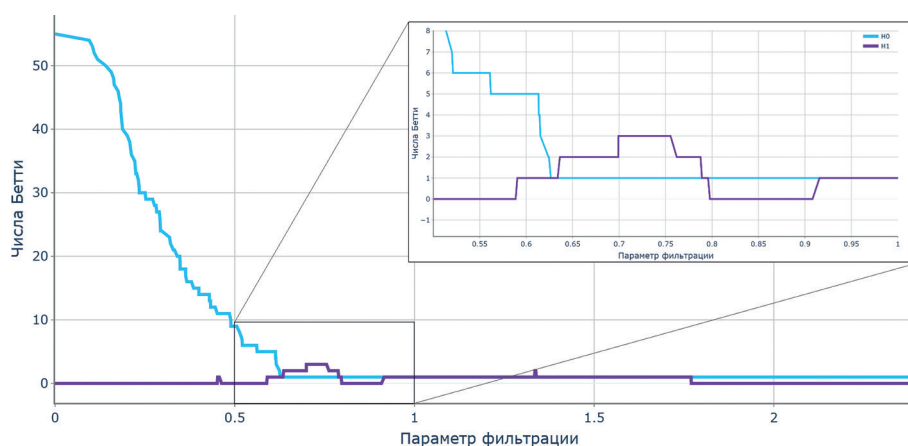


Рис. 4. Кривые Бетти

Таблица 2.

Результаты оценки качества с алгоритмом ложных ближайших соседей

Метод классификации	FNR	FPR	Accuracy	Precision	Recall	F_1 -score
KNC	0,364	0,0	0,862	1,0	0,636	0,778
SVC	0,364	0,0	0,862	1,0	0,636	0,778
GPC	0,375	0,071	0,818	0,833	0,625	0,714
DTC	0,125	0,059	0,92	0,875	0,875	0,875
RFC	0,375	0,0	0,864	1,0	0,625	0,769
MLPC	0,125	0,059	0,92	0,875	0,875	0,875
GNB	0,125	0,143	0,864	0,778	0,875	0,824

Таблица 3.

Результаты оценки качества с алгоритмом статистики непрерывности

Метод классификации	FNR	FPR	Accuracy	Precision	Recall	F_1 -score
KNC	0,091	0,053	0,933	0,909	0,909	0,909
SVC	0,091	0,053	0,933	0,909	0,909	0,909
GPC	0,25	0,0	0,909	1,0	0,75	0,857
DTC	0,0	0,111	0,931	0,846	1,0	0,917
RFC	0,25	0,071	0,864	0,857	0,75	0,8
MLPC	0,091	0,053	0,933	0,909	0,909	0,909
GNB	0,125	0,143	0,864	0,778	0,875	0,824

Таблица 4.

Результаты оценки качества с алгоритмом L-статистики

Метод классификации	FNR	FPR	Accuracy	Precision	Recall	F_1 -score
KNC	0,125	0,0	0,955	1,0	0,875	0,933
SVC	0,091	0,0	0,967	1,0	0,909	0,952
GPC	0,545	0,0	0,8	1,0	0,455	0,625
DTC	0,091	0,053	0,933	0,909	0,909	0,909
RFC	0,091	0,053	0,933	0,909	0,909	0,909
MLPC	0,091	0,0	0,967	1,0	0,909	0,952
GNB	0,545	0,0	0,8	1,0	0,455	0,625

Таблица 5.

Результаты оценки качества с алгоритмом PECUZAL

Метод классификации	FNR	FPR	Accuracy	Precision	Recall	F_1 -score
KNC	0,125	0,143	0,864	0,778	0,875	0,824
SVC	0,091	0,056	0,931	0,909	0,909	0,909
GPC	0,727	0,053	0,7	0,75	0,273	0,4
DTC	0,0	0,111	0,931	0,846	1,0	0,917
RFC	0,636	0,105	0,7	0,667	0,364	0,471
MLPC	0,091	0,056	0,931	0,909	0,909	0,909
GNB	0,182	0,0	0,933	1,0	0,818	0,9

Таблица 6.

Результаты оценки качества с алгоритмом MDOP

Метод классификации	FNR	FPR	Accuracy	Precision	Recall	F_1 -score
KNC	0,125	0,0	0,955	1,0	0,875	0,933
SVC	0,125	0,0	0,955	1,0	0,875	0,933
GPC	0,727	0,0	0,733	1,0	0,273	0,429
DTC	0,091	0,056	0,931	0,909	0,909	0,909
RFC	0,125	0,0	0,955	1,0	0,875	0,933
MLPC	0,091	0,056	0,931	0,909	0,909	0,909
GNB	0,182	0,053	0,9	0,9	0,818	0,857

Таблица 7.

Сводная таблица результатов

	FNR	FPR	Accuracy	Precision	Recall	F_1 -score
Ложных ближайших соседей	0,125	0,059	0,92	0,875	0,875	0,875
Статистика непрерывности	0,0	0,111	0,931	0,846	1,0	0,917
L -статистика	0,091	0,0	0,967	1,0	0,909	0,952
PECUZAL	0,0	0,111	0,931	0,846	1,0	0,917
MDOP	0,125	0,0	0,955	1,0	0,875	0,933

В результате проведенных экспериментов мы выявили, что при решении задачи обнаружения аномалий для рассмотренного набора данных HAI, при использовании в качестве топологического дескриптора кривых Бетти, наилучшим методом выбора параметров вложения, с показателями F_1 -score = 0,952 и FPR = 0,0, является L -статистика. В таблице 7 представлены лучшие показатели обнаружения аномалий для пяти методов выбора параметров вложения временных рядов.

Таким образом, в общем случае выбор параметров вложений необходимо осуществлять для каждого набора данных индивидуально, исходя, например, из точности решения задач классификации. Почти все рассмотренные методы выбора задержки вложения фокусируются на динамике или на топологии

в определении хорошего вложения. Кроме того, неравномерное вложение может не давать полную картину относительной значимости каждой задержки, что, в свою очередь, приводит к неэффективному использованию вычислительных ресурсов. Поэтому необходима разработка новых методов выбора задержки вложения, которые будут учитывать, как динамику, так и топологию системы, а также будут эффективными в использовании вычислительных ресурсов [18,19].

Заключение

В результате проведенного исследования мы описали ключевые отличия и оценили влияние на точность обнаружения аномалий во временных рядах пяти популярных моделей выбора параметров вложения временных рядов – MDOP, PECUZAL, ложных

ближайших соседей, L -статистики и статистики непрерывности – в контексте обнаружения аномалий в технологических процессах АСУ ТП. Оценка этих методов на наборе данных HAI позволила нам сформулировать конкретные рекомендации для практиков, подчеркивая важность выбора подходящего метода в зависимости от данных и задачи – для предложенного подхода на основе топологического анализа данных и, в частности, кривых Бетти нами рекомендуется использование метода L -статистики.

Дальнейшие исследования должны быть направлены на расширение спектра применяемых дескрипторов и наборов данных для более всеобъемлющей оценки эффективности методов вложения.

Также актуально изучить влияние комбинированных подходов, которые могут объединять преимущества нескольких моделей, на повышение точности обнаружения аномалий. Кроме того, стоит рассмотреть адаптивные алгоритмы, способные динамически изменять свои параметры в зависимости от текущих условий работы систем, что может повысить их универсальность и эффективность в реальных киберфизических системах. Это позволит не только укрепить нашу теоретическую базу, но и сделать шаг ближе к разработке практически-ориентированных решений для обеспечения безопасности и надежности технологических процессов в современных инфраструктурах.

Литература

1. Gómez Y. et al. A User and Entity Behavior Analysis for SIEM Systems: Preprocessing of The Computer Emergency and Response Team Dataset // *Journal of Soft Computing and Artificial Intelligence*. – 2023. – Т. 4. – № 1. – С. 1–6. DOI: 10.55195/jsc.ai.1213782.
2. Karan A., Kaygun A. Time series classification via topological data analysis // *Expert Systems with Applications*. – 2021. – Т. 183. – С. 115326. DOI: 10.1016/j.eswa.2021.115326.
3. Eroglu A., Eroglu H. U. Topological Data Analysis for Intelligent Systems and Applications // *Artificial Intelligence Applications in Intelligent System* – 2023. – ISRES Publishing – С. 27–60.
4. Нашивочников Н. В., Пустарнаков В. Ф. Топологические методы анализа в системах поведенческой аналитики // *Вопросы кибербезопасности*. – 2021. – № 2 (42). – С. 26–36. DOI: 10.21681/2311-3456-2021-2-26-36.
5. Нашивочников Н. В. Выявление отклонений в поведенческих паттернах пользователей корпоративных информационных ресурсов с использованием топологических признаков // *Вопросы кибербезопасности*. – 2023. – № 4. – С. 12–22. DOI: 10.21681/2311-3456-2023-4-12-22.
6. Chung Y. M., Huang W. K., Wu H. T. Topological data analysis assisted automated sleep stage scoring using airflow signals // *Biomedical Signal Processing and Control*. – 2024. – Т. 89. – С. 105760. DOI: 10.1016/j.bspc.2023.105760.
7. Меженева И. О., Лукашин А. А. Топологические дескрипторы для анализа данных в системах поведенческой аналитики // *Математические методы в технологиях и технике*. – 2023. – № 10. – С. 104–108. DOI: 10.52348/2712-8873_MMTT_2023_10_104.
8. Hensel F., Moor M., Rieck B. A survey of topological machine learning methods // *Frontiers in Artificial Intelligence*. – 2021. – Vol. 26, No. 4. – p. 681108. DOI: 10.3389/frai.2021.681108.
9. Datsis G., Parlitz U. Nonlinear dynamics: a concise introduction interlaced with code. – Springer Nature, 2022. DOI: 10.1007/978-3-030-91032-7.
10. Ravishanker N., Chen R. An introduction to persistent homology for time series // *Wiley Interdisciplinary Reviews: Computational Statistics*. – 2021. – Т. 13. – № 3. – С. e1548. DOI: 10.1002/wics.1548.
11. Tan E. et al. Selecting embedding delays: An overview of embedding techniques and a new method using persistent homology // *Chaos: An Interdisciplinary Journal of Nonlinear Science*. – 2023. – Т. 33. – № 3. DOI: 10.1063/5.0137223.
12. Marwan N., Kraemer K. H. Trends in recurrence analysis of dynamical systems // *The European Physical Journal Special Topics*. – 2023. – Т. 232. – С. 5–27. DOI: 10.1140/epjs/s11734-022-00739-8.
13. Ambika G., Hari Krishnan K. P. Methods of nonlinear time series analysis and applications: A review // *Dynamics and control of energy systems*. – 2020. – С. 9–27. DOI: 10.1007/978-981-15-0536-2_2.
14. Krakovská A. et al. State space reconstruction techniques and the accuracy of prediction // *Communications in Nonlinear Science and Numerical Simulation*. – 2022. – Т. 111. – С. 106422. DOI: 10.1016/j.cnsns.2022.106422.
15. Krämer K. H. et al. A unified and automated approach to attractor reconstruction // *New Journal of Physics*. – 2021. – Т. 23. – С. 033017. DOI: 10.1088/1367-2630/abe336.
16. Vujović Ž. et al. Classification model evaluation metrics // *International Journal of Advanced Computer Science and Applications*. – 2021. – Т. 12. – № 6. – С. 599–606. DOI: 10.14569/IJACSA.2021.0120670.
17. Douglass M. J. J. Book Review: Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow, by Aurélien Géron. – 2020 – O'Reilly Media – p. 1135–1136. DOI: 10.1007/s13246-020-00913-z.
18. Krämer K. H. et al. A unified and automated approach to attractor reconstruction // *New Journal of Physics*. – 2021. – Т. 23. – № 3. – С. 033017. DOI: 10.1088/1367-2630/abe336.
19. Axås J., Haller G. Model reduction for nonlinearizable dynamics via delay-embedded spectral submanifolds // *Nonlinear Dynamics*. – 2023. – Т. 111. – № 24. – С. 22079–22099. DOI: 10.1007/s11071-023-08705-2.

SELECTING TIME SERIES NESTING PARAMETERS TO DETECT DATA ANOMALIES IN BEHAVIORAL ANALYTICS SYSTEMS

Nashivochnikov N. V.²⁸, Mezhenева I. O.²⁹, Chatoyan S. K.³⁰

Keywords: UEBA, TDA, scalar time series, embedding dimension, embedding delay, topological descriptors, machine learning, classification, HAI Security Dataset, anomaly detection.

Purpose of the article: this article investigates the impact of embedding algorithm selection on the performance of anomaly detection in behavioral analytics systems for scalar time series. Recommendations for algorithm selection are provided.

Method: the research methodology is based on Takens-Mane theorems for reconstructing attractors of dynamical systems from scalar observables and methods of topological data analysis.

Result: the study experimentally establishes that the effectiveness of anomaly detection in behavioral analytics systems using machine learning and topological data analysis significantly depends on the quality of the reconstructed space from the scalar observable. Recommendations are developed for selecting embedding parameters – the dimension of the reconstructed space and the time delay – depending on the embedding method used.

Scientific novelty: unlike previous studies on the optimal selection of time series embedding parameters in predicting the behavior (functioning) of dynamical systems, this is the first scientific work that directly investigates the application of popular algorithms for selecting time series embedding parameters to improve the accuracy of anomaly detection by behavioral analytics systems using topological descriptors.

References

1. Görmez Y. et al. A User and Entity Behavior Analysis for SIEM Systems: Preprocessing of The Computer Emergency and Response Team Dataset // Journal of Soft Computing and Artificial Intelligence. – 2023. – V. 4. – № 1. – P. 1–6. DOI: 10.55195/jscai.1213782.
2. Karan A., Kaygun A. Time series classification via topological data analysis // Expert Systems with Applications. – 2021. – V. 183. – P. 115326. DOI: 10.1016/j.eswa.2021.115326.
3. Eroglu A., Eroglu H. U. Topological Data Analysis for Intelligent Systems and Applications // Artificial Intelligence Applications in Intelligent System – 2023. – ISRES Publishing – P. 27–60.
4. Nashivochnikov, N. V., Chatoyan, S. K., Lukashin, A. A., Zaborovskij, V. S., Vvedenie v topologicheskij analiz dannyx: uchebnoe posobie. – 2023. – SPb: POLITEX PRESS – S. 244. DOI: 10.18720/SPBPU/2/id23-431.
5. Nashivochnikov N. V., Pustarnakov V. F. Topologicheskie metody' analiza v sistemax povedencheskoj analitiki // Voprosy' kiberbezopasnosti. – 2021. – № 2 (42). – S. 26–36. DOI: 10.21681/2311-3456-2021-2-26-36.
6. Chung Y. M., Huang W. K., Wu H. T. Topological data analysis assisted automated sleep stage scoring using airflow signals // Biomedical Signal Processing and Control. – 2024. – V. 89. – P. 105760. DOI: 10.1016/j.bspc.2023.105760.
7. Mezhenева I. O., Lukashin A. A. Topologicheskie deskripty' dlya analiza dannyx v sistemax povedencheskoj analitiki // Matematicheskie metody' v tekhnologiyax i tekhnike. – 2023. – № 10. – S. 104–108. DOI: 10.52348/2712-8873_MMTT_2023_10_104.
8. Hensel F., Moor M., Rieck B. A survey of topological machine learning methods // Frontiers in Artificial Intelligence. – 2021. – Vol. 26, No. 4. – p. 681108. DOI: 10.3389/frai.2021.681108.
9. Datsieris G., Parlitz U. Nonlinear dynamics: a concise introduction interlaced with code. – Springer Nature, 2022. DOI: 10.1007/978-3-030-91032-7.
10. Ravishanker N., Chen R. An introduction to persistent homology for time series // Wiley Interdisciplinary Reviews: Computational Statistics. – 2021. – V. 13. – № 3. – P. e1548. DOI: 10.1002/wics.1548.
11. Tan E. et al. Selecting embedding delays: An overview of embedding techniques and a new method using persistent homology // Chaos: An Interdisciplinary Journal of Nonlinear Science. – 2023. – V. 33. – № 3. DOI: 10.1063/5.0137223.
12. Marwan N., Kraemer K. H. Trends in recurrence analysis of dynamical systems // The European Physical Journal Special Topics. – 2023. – V. 232. – P. 5–27. DOI: 10.1140/epjs/s11734-022-00739-8.
13. Ambika G., Hari Krishnan K. P. Methods of nonlinear time series analysis and applications: A review // Dynamics and control of energy systems. – 2020. – P. 9–27. DOI: 10.1007/978-981-15-0536-2_2.
14. Krakovská A. et al. State space reconstruction techniques and the accuracy of prediction // Communications in Nonlinear Science and Numerical Simulation. – 2022. – V. 111. – P. 106422. DOI: 10.1016/j.cnsns.2022.106422.
15. Krämer K. H. et al. A unified and automated approach to attractor reconstruction // New Journal of Physics. – 2021. – V. 23. – P. 033017. DOI: 10.1088/1367-2630/abe336.
16. Vujović Ž. et al. Classification model evaluation metrics // International Journal of Advanced Computer Science and Applications. – 2021. – V. 12. – № 6. – P. 599–606. DOI: 0.14569/IJACSA.2021.0120670.
17. Douglass M. J. J. Book Review: Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow, by Aurélien Géron. – 2020 – O'Reilly Media – p. 1135–1136. DOI: 10.1007/s13246-020-00913-z.
18. Krämer K. H. et al. A unified and automated approach to attractor reconstruction // New Journal of Physics. – 2021. – V. 23. – № 3. – P. 033017. DOI: 10.1088/1367-2630/abe336.
19. Axås J., Haller G. Model reduction for nonlinearizable dynamics via delay-embedded spectral submanifolds // Nonlinear Dynamics. – 2023. – V. 111. – № 24. – P. 22079–22099. DOI: 10.1007/s11071-023-08705-2.

28 Nikolay V. Nashivochnikov, CISSP, Deputy General Director – Technical Director, Gazinformservice LLC, St. Petersburg, Russia. E-mail: cto@gaz-is.ru

29 Irina O. Mezhenева, Lead Engineer Analyst, Laboratory of Artificial Intelligence Technologies, Cyber Security Analytical Centre, Gazinformservice LLC, St. Petersburg, Russia. E-mail: mezhenева-i@gaz-is.ru

30 Sergey K. Chatoyan, Ph.D. in Technical Sciences, Senior Researcher, Chief Specialist, Laboratory of Artificial Intelligence Technologies, Cyber Security Analytical Centre, Gazinformservice LLC, St. Petersburg, Russia. E-mail: chatoyan-s@gaz-is.ru