Математические структуры и моделирование
2023. №3 (67). С. 94–103

УДК 004.93’1
DOI 10.24147/2222-8772.2023.3.94-103

ИСПОЛЬЗОВАНИЕ ПЕРСИСТЕНТНОЙ ЭНТРОПИИ ДЛЯ ТОПОЛОГИЧЕСКОГО АНАЛИЗА ДАННЫХ

С.Н. Чуканов[1]
д.т.н., профессор, ведущий научный сотрудник, e-mail: ch_sn@mail.ru
И.С. Чуканов[2]
студент, e-mail: chukanov022@gmail.com
С.В. Лейхтер[3]
старший преподаватель, e-mail: leykhter@mail.ru

[1]Институт математики им. С.Л. Соболева СО РАН, Омский филиал, Омск, Россия
[2]Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, Екатеринбург, Россия

[3]Омский государственный университет им. Ф.М. Достоевского, Омск, Россия

Аннотация. Персистентная гомология и персистентная энтропия в последнее время стали полезными инструментами для распознавания образов. В работе найдены требования, при которых персистентная энтропия устойчива к малым возмущениям входных данных и инвариантна к масштабу. Описаны устойчивые суммирующие функции, сочетающие персистентную энтропию и кривую Бетти.

Ключевые слова: топологический анализ данных, персистентная гомология, персистентная энтропия, суммирующие функции.

1. Введение

Топологический анализ данных (TDA) использует инструменты вычислительной топологии для изучения наборов данных [1]. Интуитивно топологические признаки, такие как гомологии, можно рассматривать как качественные геометрические свойства, связанные с понятиями близости и непрерывности, следовательно, они могут быть полезными инструментами для распознавания образов. TDA стал областью исследований с персистентной гомологией в качестве ключевого инструмента.

Стандартный рабочий процесс TDA выглядит следующим образом:

– Начало с набора данных, снабжённого некоторым понятием близости (обычно метрикой).

– Построение симплициального комплекса и фильтрующей функции. Вычисление вложенной последовательности возрастающих подкомплексов, используя функцию фильтра.

– Вычисление гомологии каждого подкомплекса (интуитивно гомология захватывает «дыры» лежащего в основе пространства) и изучение того, как он развивается в последовательности, что приводит к ключевой концепции персистентной гомологии.

Персистентная гомология может быть компактно представлена с использованием персистентных бар-кодов [2], диаграмм [3] и ландшафтов [4, 5]. Эти представления устойчивы к малым возмущениям заданных данных. Существует множество программных пакетов для расчёта персистентной гомологии и её представлений.

Хотя бар-коды, диаграммы и ландшафты персистентности представляют собой метрические пространства, используемые для сравнения персистентной гомологии наборов данных, бар-коды и диаграммы персистентности не работают должным образом для статистического анализа; например, они не могут иметь уникальное среднее значение. Полезнее суммировать информацию, содержащуюся в персистентной гомологии, используя только числа. Это становится особенно целесообразным, когда доступны только небольшие выборки, поскольку в этих случаях требуются одномерные непараметрические тесты.

Персистентная энтропия является кандидатом для суммирования персистентной гомологии с использованием только чисел. В частности, персистентная энтропия – это энтропия Шеннона распределения вероятности, полученного из персистентной гомологии. Некоторые успешные приложения персистентной энтропии были разработаны для распознавания образов сигналов [6], сложных систем [7] и кластеризации [8]. Теоретический подход позволяет использовать персистентную энтропию, чтобы отличить топологические признаки от шума [9, 10]. Персистентная энтропия уже реализована как метод в библиотеке Gudhi, библиотеке scikit-TDA и библиотеке Giotto.

Когда нет необходимости находить существенные различия в данных, но нужна задача классификации, обычный подход заключается в замене статистических тестов методами машинного обучения. В этом случае суммирование персистентных гомологий в числах может быть слишком ограничительным, поскольку мы проецируем бесконечномерное пространство (постоянство бар-кодов) только на одно измерение (персистентная энтропия). Одним из решений может быть использование вместо этого суммирующих функций. Общие подходы к обобщению бар-кодов персистентности включают функции ядра, такие как многомасштабное ядро персистентности [11], взвешенное гауссовское ядро персистентности [12], а также векторизации диаграммы персистентности, такие как уже упомянутый ландшафт персистентности, силуэты персистентности [13], характеристические кривые Эйлера [14], топологические отображения интенсивности [15] и кривые Бетти [16].

2. Обзор TDA

Чтобы применить инструменты алгебраической топологии к анализу данных, мы должны обобщить информацию, предоставленную данными, в комбинаторной структуре; наиболее часто используется симплициальная структура. Напомним, что n-симплекс – это выпуклая оболочка $(n + 1)$ аффинно независимых точек. 0-симплекс – это точка, 1-симплекс – это отрезок, 2-симплекс – это треугольник, 3-симплекс – это тетраэдр и т. д.

Симплициальный комплекс – это множество симплексов, склеенных определённым образом. Абстрактный симплициальный комплекс можно рассматривать как способ хранения комбинаторной структуры симплициального комплекса.

Пусть X – конечное множество.

Семейство K подмножеств X называется абстрактным симплициальным комплексом, если для любых подмножеств $\sigma \in K; \sigma' \in X$ имеем, что $\sigma' \subset \sigma$ влечёт $\sigma' \in K$ (т. е. непустые пересечения симплексов в K также являются симплексами K). Подмножество в K из $(m+1)$ элемента X называется m-симплексом.

Когда конечное множество X представляет данные, геометрическая структура связанного с ним симплициального комплекса может предоставить информацию о том, как связаны данные. Обычно эти отношения не являются одинаково значимыми, поэтому обычно определяют порядок их симплексов, чтобы представить их важность. Это можно сделать неявно, используя функцию фильтра.

Функция фильтра на симплициальном комплексе K является монотонной функцией $f : K \to \mathbb{R}$; $\sigma' \subset \sigma$ подразумевает $f(\sigma') \leqslant f(\sigma)$. Фильтрацией на K, полученной из f, называется последовательность подкомплексов $(K_t)_{t\in\mathbb{R}}$, где $K_t = f^{-1}(-\infty, t]$. Заметим, что из-за монотонности f множество K_t является симплициальным комплексом для всех t и из $t_1 < t_2$ следует, что $K_{t_1} \subseteq K_{t_2}$. Параметр t будем называть временем, хотя его физический смысл может быть совершенно другим.

Пусть X – конечное множество точек, наделённых расстоянием dX. Фильтрацией Виеториса–Рипса X называется последовательность $(Rips(X,t))_{t\in\mathbb{R}}$, полученная из функции фильтра $f([x_0,...,x_m]) = \max\limits_{0\leqslant i,j\leqslant m} d_X(x_i,x_j)$, где для каждого $t \in \mathbb{R}$, симплексы симплициального комплекса Виеториса–Рипса $Rips(X,t)$ определяются как: $\sigma = \langle x_0,\ldots,x_m\rangle \in Rips(X,t) \Leftrightarrow f([x_0,...,x_m]) \leqslant t$.

Группы гомологий симплициальных комплексов дают формальную интерпретацию того, что такое n-мерная «дыра». Интуитивно понятно, что 0-мерное отверстие – это компонент связности, 1-мерное отверстие – это петля, 2-мерное отверстие – это полость и т. д. Для симплициального комплекса K m-цепь c является формальной суммой m-симплексов K. То есть $c = \sum\limits_{i=1}^{k} a_i\sigma_i$, где при $1 \leqslant i \leqslant k$, σ_i является m-симплексом K, a_i – коэффициент в унитальном кольце R.

Чтобы связать m-цепи данного симплициального комплекса K с его m-мерными дырами, нам понадобится граничный оператор ∂_m: если $\langle x_0,\ldots,x_m\rangle$ есть m-симплекс поля K, то

$$\partial_m(\langle x_0,\ldots,x_m\rangle) = \sum_{i=0}^{m}(1)^i\langle x_0,\ldots,x_{i-1},x_{i+1},\ldots,x_m\rangle.$$

Мы можем распространить это определение на любую m-цепь по линейности. Так как граница границы равна нулю, то $\partial_{m-1}\circ\partial_m = 0$. m-мерные дыры K обнаруживаются по m-цепочкам, граница которых равна нулю, но сами не являются «границами». Более конкретно, m-мерная группа гомологий K определяется как факторгруппа $H_m(K) = \frac{\mathrm{Ker}\partial_m}{\mathrm{Im}\,\partial_{m+1}}$ и его m-мерное число Бетти как $\beta_m = \mathrm{rank}H_m(K)$. Интуитивно понятно, что β_0 подсчитывает количество независимых компонентов связности K, β_1 – количество независимых петель и т. д.

Пусть $\mathcal{H}_m$ – m-я персистентная гомология фильтрации $\mathcal{F}$. Для $a < b$ и $m \in \mathbb{Z}$

определим:

$$\mu_m^{a,b} = \left(rank\left(\operatorname{Im} v_m^{a,b-1}\right) - rank\left(\operatorname{Im} v_m^{a,b}\right)\right) - \left(rank\left(\operatorname{Im} v_m^{a-1,b-1}\right) - rank\left(\operatorname{Im} v_m^{a-1,b}\right)\right.,$$

что можно интерпретировать как число m-мерных классов гомологии, которые «рождаются» в момент времени a и «умирают» в момент времени b. Тогда $\mathcal{H}_m$ может быть представлено мультимножеством интервалов $\{[x_i^a, y_i^a)\}_{1 \leqslant i \leqslant n}$, называемым m-м персистентным бар-кодом или диаграммой , где каждый интервал $[x_i, y_i)$ появляется $\mu_m^{x_i,y_i}$ раз. При вычислении над полем группы гомологии представляют собой векторное пространство. Этот факт позволяет использовать постоянные гомологии для изучения фильтраций.

Пусть $\mathcal{F} = (K_t)_{t \in \mathbb{R}}$ – фильтрация. Предположим, что основное кольцо R является полем и $\forall t \in \mathbb{R}, m \in \mathbb{Z}$: m-мерная группа гомологий $H_m(K_t)$ является векторным пространством. Для $\forall a < b$ и m рассмотрим линейные отображения: $v_m^{a,b} : H_m(K_a) \to H_m(K_b)$, индуцированные включением $K_a \to K_b$. m-е персистентные группы гомологии являются образами линейных отображений $v_m^{a,b}$, обозначаемых через $\operatorname{Im} v_m^{a,b}$. Множество $\left\{\operatorname{Im} v_m^{a,b}\right\}_{a<b}$ называется m-й постоянной гомологией фильтрации $\mathcal{F}$ и обозначается $\mathcal{H}_m$.

Мы предполагаем, что ранг $H_m(K_t)$ конечен $\forall t \in \mathbb{R}, m \in \mathbb{Z}$. В этом случае персистентную гомологию можно компактно представить с помощью бар-кодов (или диаграмм) персистентности.

В работе мы предполагаем, что бар-коды имеют конечное число элементов.

Пусть $\mathcal{B}$ обозначает набор стойких бар-кодов. Для бар-кода персистентности $A \in \mathcal{B}$ его n_a интервалов будут обозначаться $[x_i^a, y_i^a), 1 \leqslant i \leqslant n_a$. Длину $[x_i^a, y_i^a)$ будем обозначать через $\ell_i^a = y_i^a - x_i^a$. L_a будет обозначать сумму: $L_a = \sum\limits_{i=1}^{n_a} \ell_i^a$. Кроме того, для двух бар-кодов персистентности A, B обозначим $\max\{n_a, n_b\}$ через $n_{\max}$ и $\max\{L_a, L_b\}$ через $L_{\max}$.

Определим следующие подмножества $\mathcal{B}$.

Множество конечных бар-кодов персистентности определяется как:

$$\mathcal{B}_F = \{A \in \mathcal{B} : y_i^a < \infty, \forall [x_i^a, y_i^a) \in A\}.$$

Множество бар-кодов персистентности, все интервалы которых начинаются с 0, обозначается как $\mathcal{B}_0$: $\mathcal{B}_0 = \{A \in \mathcal{B} : x_i^a = 0, \forall [x_i^a, y_i^a) \in A\}$. Множество нормализованных бар-кодов персистентности определяется как:

$$\mathcal{B}_N = \left\{ A \in \mathcal{B} : \sum_{i=1}^{n_a} \ell_i^a = 1 \right\}.$$

Будем считать, что $n_a > 1$ для всех $A \in \mathcal{B}_F$, чтобы избежать вырожденных случаев. Существует соответствие между бар-кодами персистентности в $\mathcal{B}_F$ и бар-кодами персистентности в $\mathcal{B}_0 \cap \mathcal{B}_N$. Пусть $\psi : \mathcal{B}_F \to \mathcal{B}_0 \cap \mathcal{B}_N$ – проекция, которая определяется как композиция: $\psi = \phi \circ \pi$, где ϕ и π определяются следующим образом:

$$\phi : \mathcal{B}_F \to \mathcal{B}_N,$$

где

$$A = \{[x_i^a, y_i^a)\}_{1 \leqslant i \leqslant n_a} \to \varphi(A) = \left\{ \left[\frac{x_i^a}{L_a}, \frac{y_i^a}{L_a} \right) \right\}_{1 \leqslant i \leqslant n_a};$$

и

$$\pi : \mathcal{B}_F \to \mathcal{B}_0,$$

где

$$A = \{[x_i^a, y_i^a)\}_{1 \leqslant i \leqslant n_a} \to \pi(A) = \{[0, \ell_i^a)\}_{1 \leqslant i \leqslant n_a}.$$

Следующие метрики могут быть определены на $\mathcal{B}$. Пусть $A, B \in \mathcal{B}_F$ и $1 \leqslant p \leqslant \infty$. Определим p-е расстояние Вассерштейна как:

$$d_p(A, B) = \left(\min_{\gamma} \sum_{i=1}^{n_\gamma} \max \left\{ \left| x_i^a - x_{\gamma(i)}^b \right|^p, \left| y_i^a - y_{\gamma(i)}^b \right|^p \right\} \right)^{1/p},$$

где γ – любая биекция между мультимножествами (множества, элементы которых могут повторяться) $A = \{[x_i^a, y_i^a)\}_{1 \leqslant i \leqslant n_a}$ и $A = \left\{ \left[x_i^b, y_i^b \right) \right\}_{1 \leqslant i \leqslant n_b}$, а n_γ – кардинальное число γ. В случае $p = \infty$ это расстояние называется bottleneck расстоянием:

$$d_\infty(A, B) = \min_{\gamma} \min_{i} \max \left\{ \left| x_i^a - x_{\gamma(i)}^b \right|, \left| y_i^a - y_{\gamma(i)}^b \right| \right\}.$$

Пусть $f, g : X \to \mathbb{R}$ – две ручные функции Липшица на метрическом пространстве X, триангуляции которых растут полиномиально с постоянным показателем $j \geqslant 1$. Тогда существуют константы $c \geqslant 1, k \geqslant j$ такие, что p-е расстояние Вассерштейна между их соответствующими бар-кодами персистентности A, B удовлетворяет условию: $d_p(A, B) \leqslant c \|fg\|_\infty^{1-k/p}, \forall p \geqslant k$ [17]. Пусть K – симплициальный комплекс и $f, g : K \to \mathbb{R}$ – две монотонные функции. Если A, B – соответствующие бар-коды персистентности, полученные из f, g, то $d_\infty(A, B) \leqslant \|fg\|_\infty$ [18].

3. Устойчивость персистентной энтропии

В этом разделе показано, при каких условиях персистентная энтропия устойчива, т. е. она равномерно непрерывна или существует граница, которая «управляет» возмущением, вызванным шумом во входных данных.

Персистентная гомология может быть представлена с использованием персистентных бар-кодов. Тем не менее иногда мы можем предпочесть использовать только число для суммирования персистентной гомологии (например, персистентная энтропия), даже если при этом теряется информация.

Персистентная энтропия $E(A)$ персистентного бар-кода $A = \{[x_i^a, y_i^a)\}_{1 \leqslant i \leqslant n_a}$ в $\mathcal{B}_F$ определяется как:

$$E(A) = -\sum_{i=1}^{n_a} \frac{\ell_i^a}{L_a} \log \left(\frac{\ell_i^a}{L_a} \right).$$

Для вычисления персистентной энтропии необходимо учитывать только длину ℓ_i^a каждого интервала $[x_i^a, y_i^a)$. Если $A \in \mathcal{B}_F$, то $E(\psi(A)) = E(A)$.

Пусть $A, B \in \mathcal{B}_F$ и $1 \leqslant p \leqslant \infty$. Относительная ошибка $r_p(A, B)$ определяется как:

$$r_p(A, B) = \frac{2(n_p)^{1-1/p}}{L_{\max}} d_p(A, B).$$

Выполняется неравенство

$$d_p(\pi(A), \pi(B)) \leqslant \frac{L_{\max}}{(n_p)^{1-1/p}} r_p(A, B).$$

Обобщим результат непрерывности персистентной энтропии относительно bottleneck расстояния на расстояние Вассерштейна [9, 10].

Пусть $A, B \in \mathcal{B}_F$ и пусть d_p – p-е расстоянием Вассерштейна ($1 \leqslant p \leqslant \infty$). Если мы зафиксируем максимальное количество интервалов и минимальную сумму длин интервалов в бар-коде персистентности, то персистентная энтропия E непрерывна на $(\mathcal{B}_F, d_p)$: $\forall\varepsilon\exists\delta : d_p(A, B) \leqslant \delta \Rightarrow |E(A) - E(B)| \leqslant \varepsilon$.

Пусть $P, Q \in \mathbb{R}^u$ – два конечных распределения вероятностей, $E_S(P), E_S(Q)$ – их энтропии Шеннона. Если $\|P - Q\|_1 \leqslant 1/2$, то $|E_S(P) - E_S(Q)| \leqslant \|P - Q\|_1 (\log(u) - \log(\|P - Q\|_1))$[Cover]. Поскольку пространство $\mathcal{B}_0 \cap \mathcal{B}_N$ можно интерпретировать как конечное распределение вероятностей, мы можем сначала спроецировать бар-коды персистентности $\mathcal{B}_F$ на $\mathcal{B}_0 \cap \mathcal{B}_N$, а затем применить предыдущую теорему для получения желаемого результата устойчивости.

Пусть $A, B \in \mathcal{B}_F$. Предположим, что $r_p(A, B) \leqslant 1/4$. Тогда: $|E(A) - E(B)| \leqslant 2r_p(A, B)(\log(n_a + n_b) - \log(2r_p(A, B)))$.

Хотя $|E(A) - E(B)|$ может стремиться к ∞ при сколь угодно большом $n = n_a + n_b$, относительное значение $\frac{|E(A)-E(B)|}{\log n}$ ограничено, так как $r_p(A, B) \leqslant 1/4$:

$$\lim_{n\to\infty} \sup_{\mathcal{B}_F} \frac{|E(A) - E(B)|}{\log n} = 2r_p(A, B).$$

Чтобы расширить определение персистентной энтропии на бар-коды персистентности с интервалами бесконечной длины, обычно определяют проекцию $\mathcal{B} \to \mathcal{B}_F$, которая преобразует интервалы бесконечной длины в интервалы конечной длины. Есть много способов сделать это, и в зависимости от выбора персистентная энтропия может больше не быть устойчивой или масштабно-инвариантной. Рассмотрим некоторые проекции и их свойства.

Пусть $c \in \mathbb{R}$. Определим проекцию $\xi_c : \mathcal{B} \to \mathcal{B}_F$ таким образом, что для

$$A = \{[x_i^a, z_i^a)\} \in \mathcal{B} : \xi_c(A) = \{[x_i^a, z_i^a)\},$$

где

$$z_i^a = \begin{cases} c, & y_i^a = \infty; \\ y_i^a, & \text{otherwise} \end{cases}.$$

Следующий результат подтверждает устойчивость проекции. Пусть $A, B \in \mathcal{B}$. Тогда проекция ξ_c удовлетворяет неравенству: $d_p(\xi_c(A), \xi_c(B)) \leqslant d_p(A, B)$.

Несмотря на устойчивость, ξ_c не является масштабно-инвариантной. По определению проекция $f : \mathcal{B} \to \mathcal{B}_F$ масштабно-инвариантна, если $f(\lambda A) = \lambda f(A)$; величина λA является скалярным произведением каждого из интервалов ($\lambda \cdot \infty = \infty$).

Определим устойчивые и масштабно-инвариантные проекции $\mathcal{B} \to \mathcal{B}_F$. Пусть $\lambda \geqslant 0; 1 \leqslant p \leqslant \infty$ и $A = \{[x_i^a, y_i^a)\} \in \mathcal{B}$.

Запишем выражения для проекций $\tau_\lambda, \mu_\lambda$:

$$\mu_\lambda(A) = \{[x_i^a, z_i^a)\},$$

где

$$z_i^a = \begin{cases} z_i^a + \lambda \ell_i^a, & y_i^a = \infty; \\ y_i^a, & \text{otherwise}; \end{cases}$$

$\ell_{\max}^a$ – максимальное конечное значение для $\ell_i^a = y_i^a - x_i^a$. $\nu_{\lambda,p}(A) = \{[x_i^a, z_i^a)\}$,

$$z_i^a = \begin{cases} z_i^a + \lambda L_{a,p}, & y_i^a = \infty; \\ y_i^a, & \text{otherwise}; \end{cases}$$

$$L_{a,p} = \left(\sum_{i \in I} (\ell_i^a)^p \right)^{1/p}; I = \{i : 1 \leqslant i \leqslant n_a\}, \ell_i^a < \infty.$$

Для двух персистентных бар-кодов с одинаковым числом m интервалов бесконечной длины, имеем:

$$d_p(\mu_\lambda(A), \mu_\lambda(B)) \leqslant (1 + m 2^p \lambda^p)^{1/p} d_p(A, B);$$

$$d_p(\nu_{\lambda,p}(A), \nu_{\lambda,p}(B)) \leqslant (1 + m 2^p \lambda^p)^{1/p} d_p(A, B).$$

4. Суммирующие функции на основе энтропии

Суммирующие функции (такие как уже упомянутые силуэты персистентности, характеристические кривые Эйлера, топологические карты интенсивности или ландшафты персистентности) использовались для получения статистической информации из бар-кодов персистентности. Например, простым способом обобщения бар-кода персистентности является кривая Бетти, определяемая следующим образом: если $A = \{[x_i^a, y_i^a)\} \in \mathcal{B}$, то $\beta(A)(t) = \text{card}\{[x_i^a, y_i^a) : x_i^a \leqslant t \leqslant y_i^a\}$. То есть $\beta(A)(t)$ – это количество интервалов в A, которые «живы» в момент времени t.

Определим новую суммирующую кусочно-постоянную функцию. Она похожа на кривую Бетти, но использует постоянную энтропию вместо чисел Бетти. Нормализация этой функции является устойчивой.

Определим новую функцию, которая связывает бар-код персистентности $A \in \mathcal{B}_F$ с вещественной кусочно-персистентной функцией. Эта новая функция суммирует информацию о количестве интервалов данного бар-кода персистентности и их однородности и является устойчивой по отношению к bottleneck расстоянию.

Суммирующая функция энтропии (ES-функция) персистентного бар-кода $A = \{[x_i^a, z_i^a)\} \in \mathcal{B}_F$ представляет собой кусочно-линейную знакопостоянную вещественную функцию:

$$S(A)[t] = -\sum_{i=1}^{n_a} w_i^a(t) \frac{\ell_i^a}{L_a} \log\left(\frac{\ell_i^a}{L_a}\right),$$

где

$$w_i^a(t) = \begin{cases} 1, & x_i^a \leqslant t \leqslant y_i^a; \\ 0, & \text{otherwise} \end{cases}.$$

Другими словами, ES-функция связывает бар-код персистентности $A = \{[x_i^a, z_i^a)\}$ и момент времени t с частичной суммой $E(A)$, соответствующей интервалам $[x_i^a, z_i^a) \in A$, которые «живы» в этот момент t, т. е. $x_i^a \leqslant t \leqslant y_i^a$. Обратите внимание на то, что $S(A)$: $\mathbb{R} \to \mathbb{R}$ и $S : \mathcal{B}_F \to \mathcal{C}$, являющееся $\mathcal{C}$ пространством кусочно-постоянных вещественных функций.

Пусть S будет ES-функцией, d_∞ – bottleneck расстояние, $A, B \in \mathcal{B}_F$ – бар-коды персистентности. Пусть n_∞ – кардинальное число биекции, обозначаемой как γ_∞, когда $d_\infty(A, B)$ достигается.

Если $r_\infty(A, B) \leqslant \frac{2}{3e}$, то

$$\|S(A) - S(B)\|_1 \leqslant r_\infty(A, B) L_{\max}\left(\frac{\log n_{\max}}{n_{\max}} - \frac{3}{2}\log\left(\frac{3}{2} r_\infty(A, B)\right)\right).$$

ES-функция основана на персистентной энтропии, тогда как кривая Бетти состоит из подсчёта количества «живых» интервалов. Обе функции (ES-функция и кривая Бетти) непрерывны относительно bottleneck расстояния, если фиксировано максимальное число интервалов. ES-функция работает лучше, чем кривая Бетти, в шумном контексте, поскольку персистентная энтропия является устойчивой, а подсчёт количества интервалов – нет.

Одной из основных целей персистентной гомологии является представление формы входных данных. В некоторых приложениях, таких как анализ изображений, может быть важно обнаруживать некоторые повторяющиеся закономерности независимо от размера входного набора данных. Возможным инструментом для этого является нормализованная версия суммирующей функции, чтобы попытаться зафиксировать форму пространства, а не размер.

Нормализованная суммирующая функция энтропии (NES-функция) бар-кода персистентности $A = \{[x_i^a, z_i^a)\} \in \mathcal{B}_F$ определяется как:

$$NES(A)[t] = \frac{S(A)[t]}{\|S(A)\|_1}.$$

Как и ES-функция, эта функция также является устойчивой. Если $r_\infty(A, B) \leqslant \frac{2}{3e}$, то

$$\|NES(A) - NES(B)\|_1 \leqslant \frac{r_\infty(A, B) L_{\max}\left(\frac{\log n_{\max}}{n_{\max}} - \frac{3}{2}\log\left(\frac{3}{2} r_\infty(A, B)\right)\right)}{\min\{\|S(A)\|_1, \|S(B)\|_1\}}.$$

5. Заключение

В работе рассматривается устойчивость персистентной энтропии. Персистентная энтропия использовалась для формирования устойчивой суммирующей функции (ES-функции) и её нормализованной версии (NES-функции). В целом они работают лучше, чем кривая Бетти, в шумном контексте и могут быть полезны для задач машинного обучения. Несколько типов кривых персистентности, были также определены Y.M Chung и A. Lawson в [20].

6. Благодарности

Работа выполнена в рамках государственного задания ИМ СО РАН, проект FWNF–2022–0016, и при поддержке Российского научного фонда, грант № 22-21-00035.

Литература

1. Чуканов С.Н., Чуканов И.С., Лейхтер С.В. Формирование признаков машинного обучения на основе методов вычислительной топологии // Математические структуры и моделирование. 2022. № 4 (64). С. 89–99.
2. Carlson G., Zomorodian A., Collins A., Guibas L. Persistence barcodes for shapes // International Journal of Shape Modeling. 2005. Vol. 11, No. 2. P. 149–187.
3. Edelsbrunner H., Letscher D., Zomorodian A. Topological persistence and simplification // Discrete & Computational Geometry. 2002. Vol. 28, No. 4. P. 511–533.
4. Bubenik P. Statistical topology using persistence landscapes // Journal of Machine Learning Research. 2015. Vol. 16. P. 77–102.
5. Лейхтер С.В., Чуканов С.Н., Чуканов И.С., Широков И.В. Анализ данных. Омск : ОмГУ, 2022. 108 с.
6. Rucco M. et al. A new topological entropy-based approach for measuring similarities among piecewise linear functions // Signal Processing. 2017. Vol. 134. P. 130–138.
7. Binchi J. et al. Jholes: A tool for understanding biological complex networks via clique weight rank persistent homology // Electronic Notes in Theoretical Computer Science. 2014. Vol. 306. P. 5–18.
8. Wang X. et al. Scale space clustering evolution for salient region detection on 3d deformable shapes // Pattern Recognition. 2017. Vol. 71. P. 414–427.
9. Atienza N., Gonzalez-Diaz R., Rucco M. Persistent entropy for separating topological features from noise in Vietoris-Rips complexes // Journal of Intelligent Information Systems. 2019. Vol. 52. P. 637–655.
10. Atienza N., Gonzalez-Diaz R., Soriano-Trigueros M. On the stability of persistent entropy and new summary functions for topological data analysis // Pattern Recognition. 2020. Vol. 107. P. 107509.
11. Reininghaus J. et al. A stable multi-scale kernel for topological machine learning // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. P. 4741–4748.
12. Kusano G., Hiraoka Y., Fukumizu K. Persistence weighted Gaussian kernel for topological data analysis // International conference on machine learning. PMLR, 2016. P. 2004–2013.

13. Chazal F. et al. Stochastic convergence of persistence landscapes and silhouettes // Proceedings of the thirtieth annual symposium on Computational geometry. 2014. P. 474–483.
14. Richardson E., Werman M. Efficient classification using the Euler characteristic // Pattern Recognition Letters. 2014. Vol. 49. P. 99–106.
15. Pranav P. et al. The topology of the cosmic web in terms of persistent Betti numbers // Monthly Notices of the Royal Astronomical Society. 2017. Vol. 465, No. 4. P. 4281–4310.
16. Umeda Y. Time series classification via topological data analysis // Information and Media Technologies. 2017. Vol. 12. P. 228–239.
17. Cohen-Steiner D. et al. Lipschitz functions have L p-stable persistence // Foundations of computational mathematics. 2010. Vol. 10, No. 2. P. 127–139.
18. Edelsbrunner H., Harer J. Computational Topology: An Introduction. American Mathematical Society, 2010. 241 p.
19. Cover T.M. Elements of information theory. John Wiley & Sons, 2006. 784 p.
20. Chung Y.M. et al. Topological approaches to skin disease image analysis // 2018 IEEE International Conference on Big Data. IEEE, 2018. P. 100–105.

USING PERSISTENT ENTROPY FOR TOPOLOGICAL DATA ANALYSIS

S.N. Chukanov[1]
Dr.Sc. (Techn.), Professor, Leading Scientist Researcher, e-mail: a@a.ru
I.S. Chukanov[2]
Student, e-mail: chukanov022@gmail.com
S.V. leykhter[3]
Assistant Professor, e-mail: leykhter@mail.ru

[1]Sobolev Institute of Mathematics, Omsk branch, Omsk, Russia
[2]Ural Federal University named after the first President of Russia B.N. Yeltsin, Ekaterinburg, Russia
[3]Dostoevsky Omsk State University, Omsk, Russia

Abstract. Persistent homology and persistent entropy have recently become useful tools for pattern recognition. In the paper, requirements are found under which the persistent entropy is stable to small perturbations of the input data and is scale invariant. Stable summary functions are described that combine the persistent entropy and the Betti curve.

Keywords: topological data analysis, persistent homology, persistent entropy, summary functions.

Дата поступления в редакцию: 10.07.2023