УДК 004.93'1 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait/1995-5499/2022/3/115-126
Поступила в редакцию 01.06.2022
Подписана в печать 30.09.2022

ФОРМИРОВАНИЕ ПРИЗНАКОВ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ ТОПОЛОГИЧЕСКОГО АНАЛИЗА ДАННЫХ

© 2022 С. Н. Чуканов[1]✉, И. С. Чуканов[2]

[1]Институт математики им. Соболева С.Л. СО РАН
ул. Певцова, 13, 644043 Омск, Российская Федерация
[2]Уральский федеральный университет им. первого президента России Б. Н. Ельцина
ул. Мира, 32, 620078 Екатеринбург, Российская Федерация

Аннотация. В настоящее время возрос интерес к использованию методов алгебраической топологии для топологического анализа данных и применению топологического анализа данных в различных областях знаний. Целью топологического анализа данных является определение информативных топологических свойств и использование их в качестве дескрипторов при машинном обучении. Применение методов машинного обучения для сложных систем большой размерности затруднено из-за методов адекватного представления функций.

Метод персистентной гомологии из вычислительной топологии обеспечивает баланс между уменьшением размерности данных и характеристикой внутренней структуры объекта. Совмещению персистентной гомологии и машинного обучения препятствуют топологические представления данных, метрики расстояния и представление объектов данных. В работе используется метод персистентной гомологии, основанный применении фильтрации для присвоения каждому топологическому признаку геометрической размерности. Процесс фильтрации генерирует серии симплициальных комплексов, кодируемых со структурной информацией различных масштабов. Персистентная гомология может быть представлена персистентным баркодом или персистентной диаграммой.

В работе рассматриваются математические модели и функции представления объектов персистентного ландшафта на основе метода персистентной гомологии. Рассмотрены персистентные функции Бетти и функции персистентного ландшафта. Функции персистентного ландшафта позволяют отображать персистентные диаграммы и персистентные баркоды в гильбертово пространство. Рассмотрены представления топологических характеристик в различных моделях машинного обучения. Рассмотрена структура ядра для анализа персистентных диаграмм и персистентное взвешенное ядро Гаусса. Метод персистентного взвешенного ядра позволяет контролировать персистентность при анализе данных. Расстояния между персистентными ландшафтами определяются с помощью нормы пространства L^p. Приведены примеры нахождения расстояния между изображениями. В приложениях приведены основные понятия алгебраической топологии и метод воспроизводящего ядра гильбертова пространства для целей машинного обучения.

Ключевые слова: симплициальный комплекс, персистентные гомологии, персистентный ландшафт, машинное обучение, RKHS, гильбертово пространство.

ВВЕДЕНИЕ

В последние годы возрос интерес к использованию методов алгебраической топологии для топологического анализа данных (TDA — topological data analysis) [1] и применению в различных областях знаний. Целью TDA является определение информативных топологических свойств и использование их в качестве дескрипторов.

Ключевым математическим инструментом в топологическом анализе данных является метод персистентных гомологий (PH —

✉ Чуканов Сергей Николаевич
e-mail: ch_sn@mail.ru

persistent homology), который используется для извлечения топологической информации из данных. Рассмотрим способ формирования PH из точек данных в евклидовом пространстве. Целью является получение топологии из конечных данных. Рассмотрим *r*-шары (радиуса *r*) для реконструкции топологии. Ожидается, что модель *r*-шаров может представлять основные топологические структуры. Если r мал, то объединение всех *r*-шаров состоит из непересекающихся *r*-шаров. Если слишком большие, то объединение становится одним пространственным компонентом. Персистентная гомология [2, 3] рассматривает все значения r одновременно и обеспечивает выражение для топологических свойств.

Персистентная гомология может быть визуализирована персистентной диаграммой (PD — persistence diagram) $D = \{(b_i, d_i) \in \mathbb{R}^2 \mid i \in I, b_i \le d_i\}$. Каждая точка $(b_i, d_i) \in D$, которая называется генератором персистентной гомологии, представляет топологическое свойство, которое представляет топологическое свойство, появляющееся при X_{b_i} и исчезающее при X_{d_i} в модели *r*-шаров. Топологическое свойство с высокой персистентностью $d_i - b_i$ может рассматриваться как надежная структура, в то время как топологическое свойство с низкой персистентностью может рассматриваться как шум. Персистентной диаграммы кодируют топологическую и геометрическую информацию о точках данных.

Применение методов машинного обучения для сложных систем большой размерности затруднено из-за методов адекватного представления функций [4]. Геометрический анализ характеризует локальную структуру, но приводит к сложности представления данных. Элементы, полученные из топологических моделей определяют глобальную внутреннюю структурную информацию, но редуцируют много локальной структурной информации [5].

Методы персистентной гомологии разработаны для многомасштабного представления топологических признаков [1, 2, 6]. Метод персистентной гомологии обеспечивает мост между топологическими и геометрическими методами. Основная идея метода PH — применение фильтрации для присвоения каждому топологическому признаку геометрической размерности. Процесс фильтрации генерирует серии симплициальных комплексов, кодируемых со структурной информацией различных масштабов. Персистентной гомология может быть представлена персистентным баркодом (PB — persistent barcode) или персистентной диаграммой (PD — persistent diagram).

Машинное обучение может быть использовано для анализа топологических данных. Существуют подходы к машинному обучению с использованием персистентных диаграмм. Однако, стандартные меры для PD (например, расстояние Вассерштейна) не подходят для машинного обучения. Один из подходов к машинному обучению является отображение PD в гильбертово пространство формированием персистентных функций ландшафта (landscape functions) [7].

Из-за важной роли метода ядер в моделях машинного обучения были предложены различные ядра персистентной гомологии. Эти ядра, формируемые из PD/PB, могут быть интегрированы методы машинного обучения, основанные на ядрах. Топологические характеристики могут быть извлечены из PD/PB. Простейший путь получения свойств, основанных на PD/PB, — это собрать их статистические свойства. Специфические топологические характеристики, такие как числа Betti для определенного значения фильтрации, могут рассматриваться как признаки для машинного обучения. Более систематическим методом получения вектора топологических признаков их PD/PB является binning метод, в котором PD/PB расщепляются на различные компоненты, которые затем комбинируются в вектор признаков [8].

Для векторизации персистентных диаграмм используется концепция вложения ядерных мер в RKHS (reproducing kernel Hilbert spaces) [9, 10]. Для вложения персистентных диаграмм в RKHS предлагается класс положительно определенных ядер PWGK (persistence weighted Gaussian kernel) [3]. Преимущества использования PWGK заключаются в следующем:

(i) способность контролировать эффект персистентности;

(ii) расстояние, определяемое нормой RKHS для PWGK, удовлетворяет свойству персистентности, которое обеспечивает непрерывность от данных к векторному представлению PD;

(iii) PWGK позволяет проводить расчеты с использованием функций Фурье и применима к PD с большим количеством генераторов.

Проводилось тестирование методов топологического анализа данных (на основе персистентной гомологии) по отношению к методам традиционной алгебраической топологии для определения расстоянии между изображениями цифр 6 и 9; при этом методы методов топологического анализа данных показали возможность определения такого расстояния, в то время как методы традиционной алгебраической топологии определяют такое расстояние как равное нулю.

1. ПЕРСИСТЕНТНЫЕ ГОМОЛОГИИ

Персистентные гомологии можно представить в виде топологических генераторов (баркодов) — пар появления (BT — birth time) и исчезновения (DT — death time) баркодов, которые можно обозначить как $l_j^k = \{b_j^k, d_j^k\}$, $j \in \{1,2,...,N_k\}$, где N_k — общее число k-мерных топологических генераторов [11]. Определим множество баркодов k-го измерения: $L_k = \left\{ l_j^k = \{b_j^k, d_j^k\}_{j\in\{1,2,...,N_k\}} \right\}$. Топологическая персистентность может быть представлена персистентным баркодом (каждый l_j^k рассматривается как баркод) или персистентной диаграммой (каждый l_j^k рассматривается как двумерная точка с координатой $l_j^k = (b_j^k, d_j^k)$) [12].

Пусть $X = \{x_1,...,x_n\}$ — конечное подмножество в метрическом пространстве (M, d_M). Чтобы проанализировать топологические свойства X, рассмотрим модель $X_r = \bigcup_{i=1}^{n} B(x_i; r)$, состоящую из шаров $B(x_i; r) = \{x \in M \mid d_M(x_i, x) \le r\}$ с радиусом r, и используем гомологии $H_q(X_r)$ для описания топологии X_r. Здесь для топологического пространства S его q-я гомология $H_q(S)$; $q = 0,1,\ldots$ определяется как векторное пространство. Так как $X_r \subset X_s$ для $r \le s$ множество $X = \{X_r \mid r \ge 0\}$ становится фильтрацией. При изменении радиуса новый генератор $a_i \in H_q(X_r)$ появляется на каком-то радиусе $r = b_i$ (birth) и исчезает на радиусе $r = d_i$ (death) большем чем b_i. Собирая все a_i $(i \in I)$ в фильтрации X, получаем множество пар $\underline{D}_q(X) = \{(b_i, d_i) \in \mathbb{R}^2 \mid i \in I\}$ в виде мультимножества. Персистентная диаграмма $D_q(X)$ определяется несвязным объединением $\underline{D}_q(X)$ и диагонального множества $\Delta = \{(a,a) \mid a \in R\}$, учитываемого с бесконечной кратностью. Точку $x = (b,d) \in D_q(X)$ называют генератором персистентной диаграммы. Персистентность точки x равна: $\mathrm{pers}(x) = d - b$.

Желательно, чтобы персистентные диаграммы были устойчивыми при возмущении данных. Мерой для изучения сходства между двумя персистентными диаграммами D и E является расстояние bottleneck $d_B(D,E) = \inf_{\gamma} \sup_{x\in D} \left\| x - \gamma(x) \right\|_\infty$, где γ это различные биекции от D до E: $(x \in D) \to (\gamma(x) \in E)$. В качестве расстояния между конечными множествами X, Y в метрическом пространстве M можно использовать расстояние Хаусдорфа, определяемое формулой:

$$d_H(X,Y) = \max\left\{ \sup_{x\in X} \inf_{y\in Y} d_M(x,y), \sup_{y\in Y} \inf_{x\in X} d_M(x,y) \right\}. \quad (1)$$

Пусть X, Y конечные подмножества в метрическом пространстве (M, d_M). Тогда расстояние bottleneck между персистентными диаграммами удовлетворяют неравенству $d_B(D_q(X), D_q(Y)) \le d_H(X,Y)$.

Персистентные функции. Для представления топологической информации были предложены различные функции, основанные на результатах PB/PD.

Непрерывная персистентная функция Бетти определяется как [13]:

$$f(x; L_k) = \sum_j \exp\left(-\left(x - \frac{b_j^k + d_j^k}{2} \right) \left(w_j \left(d_j^k - b_j^k \right) \right)^{-1} \right), \quad (2)$$

где w_j — значения весов.

Для каждого отдельного баркода можно определить кусочно-линейную функцию $f(x, l_j^k)$ [7]:

$$f\left(x, l_j^k\right) = \begin{cases} 0 & \text{if } x \notin \left(b_j^k, d_j^k\right), \\ x - b_j^k & \text{if } x \in \left(b_j^k, \dfrac{b_j^k + d_j^k}{2}\right], \\ -x + d_j^k & \text{if } x \in \left[\dfrac{b_j^k + d_j^k}{2}, d_j^k\right). \end{cases} \quad (3)$$

1.1. Основные понятия алгебраической топологии

Основные понятия алгебраической топологии изложены в монографии [5].

Цепи. Определим пространство $C_k(K)$ k-цепей на симплициальном комплексе K как векторное пространство линейных комбинаций ориентированных k-симплексов. C_k это свободная абелева группа, имеющая структуру векторного пространства вещественных функций. Элемент (k-цепь) $c \in C_k$ можно представить в виде суммы k-симплексов: $c = \sum_{\sigma \in S_k} w_\sigma \sigma$, где $w_\sigma \in \mathbb{R}$ — вес каждого k-симплекса. Линейные граничные отображения между последовательными цепными пространствами могут быть представлены как: $\partial_k c_k = 0$. k-й граничный оператор является линейным отображением: $\partial_k : C_k \to C_{k-1}$, которое определяется его операцией над базисными элементами C_k [16]:

$$\partial_k \left[v_{i_0}, \ldots, v_{i_k}\right] = \sum_{j=0}^{k} (-1)^j \left[v_{i_0}, \ldots, v_{i_{j-1}}, v_{i_{j+1}}, \ldots, v_{i_k}\right].$$

Если построить циклическую цепь $c_k \in C_k$, начинающуюся и заканчивающуюся в одном и том же симплексе, то $\partial_k c_k = 0$ и k-цепь $c_k \in \ker(\partial_k)$ называется k-циклом.

Группа гомологии. Граничные операторы это линейные отображения между конечномерными векторными пространствами. После выбора ориентации каждый из этих операторов может быть представлен матрицей, что позволяет нам выполнять вычисления. Матричное представление граничных операторов ∂_k будем обозначать через B_k. k-циклы это циклы в ядре граничного оператора, т. е. элементы $Z_k = \ker(\partial_k)$ отображения $\partial_k : C_k \to C_{k-1}$. k-границами называются циклы, образующие границы $(k+1)$-симплекса, т. е. элементы $B_k = \mathrm{im}(\partial_{k-1})$ отображения $\partial_{k+1} : C_{k+1} \to C_k$. Определим следующее подпространство k-цепей (k-ю группу гомологии): $H_k = Z_k / B_k = \ker(\partial_k) / \mathrm{im}(\partial_{k+1})$. Размерность H_k ($\dim H_k$) называется k-м числом Бетти.

Ориентация k-симплекса $\sigma^k = \{v_0, v_1, \cdots, v_k\}$ определяется классом эквивалентности порядка вершин σ^k, где $(v_0, v_1, \cdots, v_k) \sim$ $\sim (v_{\tau(0)}, v_{\tau(1)}, \cdots, v_{\tau(k)})$ — эквивалентные порядки, если перестановка τ четная. Обозначим ориентированный симплекс как $[\sigma^k] =$ $= \left[v_0, v_1, \cdots, v_k\right]$.

Фильтрацией симплициального комплекса K называется последовательность подкомплексов: $\varnothing \subset K^0 \subseteq K^1 \subseteq \cdots \subseteq K^n = K$. Персистентные гомологии (PH) определяются фильтрацией, в ходе которой формируются топологические пространства разного масштаба. Пусть K^l это фильтрация симплициального комплекса K и $Z_k^l = Z_k(K^l)$, $B_k^l = B_k(K^l)$; k-я группа гомологии K^l: $H_k^l = Z_k^l / B_k^l$. Персистентные гомологии обеспечивают геометрическое измерение топологическому инварианту.

Пусть X — это множество точек в евклидовом пространстве $\mathbb{R}^d$ и U хорошее покрытие X, т. е. $X \subseteq \bigcup_{i \in I} U_i$. Нерв N покрытия U определяется условиями: 1) $\varnothing \in N$; 2) $\left(\bigcap_{j \in J} U_j \neq 0 \Big|_{J \subseteq I} \right) \Rightarrow (J \in N)$. Комплекс Vietoris-Rips с параметром ε — это такое множество всех $\sigma \subseteq X$, что наибольшее евклидово расстояние между любыми его точками не превосходит 2ε.

2. ПЕРСИСТЕНТНЫЕ ЛАНДШАФТЫ (PL — PERSISTENT LANDSCAPES)

Персистентный ландшафт k-мерного баркода L_k — это последовательность функций: $\lambda_m^{PL} : \mathbb{R} \to \left[0, \infty\right]$, $m = 1, 2, 3, \ldots$, где $\lambda_m(x)$ — m-е наибольшее значение $\{f(x, l_j^k)\}_{j=1}^{N_k}$. Для баркодов $B = \{I_j\}$ можно определить PL как:

$$\lambda(k,t) =$$
$$\sup\left(h \geq 0 \middle| [t-h, t+h] \subset I_j, \text{ для } \geq k \text{ различных } j\right).$$

Определим функцию для PD

$$D = \left\{\left(b_i, d_i\right)\right\}, b_i < d_i:$$

$$f_{(b,d)}(t) = \max\left(0, \min\left(b+t, d-t\right)\right);$$

тогда $\lambda(k,t) = \text{kmax}\left\{f(b_i,d_i)(t)\right\}_{i\in I}$, где kmax обозначает k-й наибольший элемент.

Пусть задано множество S. Функция $F: S \to H$, где H — гильбертово пространство, называется функцией отображения признаков. Ядро на S является таким симметричным отображением $K: S \times S \to \mathbb{R}$, что для любого n и всех $x_1, ..., x_n \in S$, $a_1, ..., a_n \in \mathbb{R}$: $\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) \geq 0$.

RKHS (Reproducing kernel Hilbert space) на множестве S — это гильбертово пространство функций на S, где точечная оценка — непрерывный линейный функционал. Для заданного отображения характеристик существует ассоциированное ядро, определяемое формулой $K(x,y) = \left\langle F(x), F(y)\right\rangle_H$.

С ядром K связано гильбертово пространство RKHS H_k, которое является пополнением множества функций $K_x: S \to \mathbb{R}$, заданных формулой:

$$K_x(y) = K(x,y), \forall x \in S,$$

относительно скалярного произведения: $\left\langle K_x, K_y\right\rangle = K(x,y)$.

Поскольку функция PL является отображением характеристик из множества PD в $L^2(\mathbb{N}\times\mathbb{R})$, то с ней ассоциируется ядро PL:

$$K\left(D^{(1)}, D^{(2)}\right) = \left\langle \lambda^{(1)}, \lambda^{(2)}\right\rangle = \\ = \sum_{k=1}^{\infty}\int_{-\infty}^{\infty} \lambda_k^{(1)}(t)\lambda_k^{(2)}(t)\,dt. \quad (4)$$

Для PL формируем p-норму:

$$\left\|\lambda\right\|_p = \sum_{k=1}^{\infty}\left[\int_{-\infty}^{\infty}\left(\lambda_k(t)\right)^p dt\right]^{\frac{1}{p}}, \text{ if } 1 \leq p < \infty,$$

и

$$\left\|\lambda\right\|_\infty = \sup_{k,t}\lambda_k(t), \text{ if } p = \infty.$$

Ядро можно рассматривать как ассоциированное отображение признаков: $D \to \sum_{k=1}^{\infty}\lambda_k(D)$, которое формирует отображение в гильбертово пространство со скалярным произведением:

$$\left\langle f, g\right\rangle = \sum_{k=1}^{\infty}\int_{-\infty}^{\infty} f_k(t)\,g_k(t)\,dt. \quad (5)$$

Расстояния между персистентными ландшафтами можно определить с помощью нормы L^∞:

$$\left\|\lambda^{PL} - \lambda'^{PL}\right\|_\infty = \sup_{k,t}\left|\lambda_k^{PL}(t) - \lambda_k'^{PL}(t)\right|,$$

или нормы [14, 15]:

$$\left\|\lambda^{PL} - \lambda'^{PL}\right\|_p = \\ \left[\sum_{k}\int_{-\infty}^{\infty}\left|\lambda_k^{PL}(t) - \lambda_k'^{PL}(t)\right|^p dt\right]^{\frac{1}{p}}, 1 \leq p < \infty. \quad (6)$$

3. МЕТОДЫ ЯДРА ДЛЯ ПЕРСИСТЕНТНЫХ ДИАГАММ

Рассмотрим ядро для персистентных диаграмм, называемое персистентным взвешенным ядром Гаусса (persistent weighted Gaussian kernel — PWGK) [3]. Пусть $k^w(x,y) = w(x)w(x)k(x,y)$ — взвешенное ядро весовой функцией $w(\cdot)$; рассмотрим отображение:

$$E_{k^w}: \mu_D \to \sum_{x\in L} w(x)\,w(\cdot)\,k(\cdot, x) \in \mathcal{H}_{k^w}.$$

Для практических целей выбираем ядро Гаусса $k_G(x,y) = \exp\left(-\frac{\left\|x-y\right\|^2}{2\sigma^2}\right)$; $\sigma > 0$ для k и $w_{\text{arc}}(x) = \arctan(C(b_x - a_x)^p)$; $C > 0$, $p > 0$ для весовой функции.

Персистентное взвешенное ядро Гаусса (PWGK) определяется следующим образом:

$$\kappa_{PWGK}\left(L_k, L_k', \sigma\right) = \\ \sum_{l_j^k \in L_k, l_{j'}^k \in L_{k'}} w_{\text{arc}}\left(l_j^k\right) w_{\text{arc}}\left(l_{j'}^k\right)\exp\left(-\frac{\left\|l_j^k - l_{j'}^k\right\|^2}{2\sigma^2}\right), \quad (7)$$

$$w_{\text{arc}}\left(l_j^k\right) = \arctan\left(C\left(d_j^k - b_j^k\right)^p\right); C, p > 0.$$

Коэффициент $w_{\text{arc}}(x)$ является возрастающей функцией по отношению к персистентности x. Следовательно, генератор x дает малое значение $w_{\text{arc}}(x)$ при малых x. Изменяя параметры C, p, мы можем контролировать эффект персистентности.

Методы ядра на RKHS. По расстоянию между множествами точек персистентных диаграмм L_k и L'_k можно оценить расстояния между соответствующими изображениями. Если персистентные диаграммы представлены векторами в RKHS, можно применять к этим векторам методы ядра для определения расстояния между L_k и L'_k. Самый простой выбор — рассмотреть линейное ядро на RKHS:

$$k_L(D,E) = \sum_{x\in L_k}\sum_{y\in L'_k} w_{\text{arc}}(x)\, w_{\text{arc}}(y)\, k_G(x,y). \tag{8}$$

Также можно рассмотреть нелинейное ядро на RKHS, такое как ядро Гаусса:

$$k_G(L_k, L'_k) = \exp\left(-\frac{d_{k_G}^{w_{\text{arc}}}(L_k, L'_k)^2}{2\tau^2}\right), \tau > 0, \tag{9}$$

где

$$\begin{aligned} d_{k_G}^{w_{arc}}(L_k, L'_k)^2 = {} & \sum_{x\in L_k}\sum_{x'\in L_k} w_{\text{arc}}(x)\, w_{\text{arc}}(x')\, k_G(x,x') + \\ & \sum_{y\in L'_k}\sum_{y'\in L'_k} w_{\text{arc}}(y)\, w_{\text{arc}}(y')\, k_G(y,y') - \\ & 2\sum_{x\in L_k}\sum_{y\in L'_k} w_{\text{arc}}(x)\, w_{\text{arc}}(y)\, k_G(x,y). \end{aligned} \tag{10}$$

3.2. Ядра в машинном обучении

В задачах машинного обучения нас интересует классификация данных во входном пространстве с помощью разделения гиперплоскостью. Однако использование линейного разделения ограничивает эффективность такого подхода. Можно использовать нелинейное разделение во входном пространстве и метод RKHS обеспечивает основу для достижения этого.

Рассмотрим симметричную меру подобия, называемую ядром: $\kappa : \Omega\times\Omega \to \mathbb{R}$, $(x,x') \to \kappa(x,x')$, $x, x' \in \Omega \in \mathbb{R}^d$. Так как $\Omega \in \mathbb{R}^d$, то существует возможность рассмотреть евклидово скалярное произведение для вычисления мер подобия: $\kappa(x,y) = x^T y$.

Предположим, что κ это положительно определенное ядро с действительными значениями и Ω непустое множество. Если $\mathbb{R}^\Omega = \{f : \Omega \to \mathbb{R}\}$, то отображение признаков — это такая функция, что $\Phi : \Omega \to \mathbb{R}^\Omega$, $x = \kappa(x,\cdot)$.

Φ отображает образы в функции на $\mathbb{R}^\Omega$. Это позволяет нам встраивать данные в векторное пространство признаков:

$$\mathcal{F} = \left\{ \sum_{i=1}^{n} \alpha_i \kappa(x_i,\cdot) \,\middle|\, n\in\mathbb{N}, x_i \in \Omega, \alpha_i \in \mathbb{R}, i = 1,\ldots,n \right\}.$$

Используя эту концепцию, мы можем построить предгильбертово пространство. Пусть $f, g \in \mathcal{F}$ ассоциируются с образами $x_i, x_j \in \Omega$; $i = 1,\ldots,n; j = 1,\ldots,n'$: $f = \sum_{i=1}^{n} \alpha_i \kappa(x_i,\cdot)$, $g = \sum_{i=1}^{n'} \beta_j \kappa(x'_j,\cdot); \alpha_i, \beta_j \in \mathbb{R}$. Определим внутреннее (скалярное) произведение:

$$\begin{aligned} \langle f, g\rangle &= \sum_{j=1}^{n'}\sum_{i=1}^{n} \alpha_i \beta_j \kappa(x_i, x'_j) = \\ &\sum_{j=1}^{n'} \beta_j f(x'_j) = \sum_{i=1}^{n} \alpha_i g(x_i). \end{aligned} \tag{11}$$

Функция κ, определенная на $\Omega\times\Omega$, является воспроизводящим ядром если и только если существуют гильбертово пространство H и отображение $\Phi : \Omega \to H$, такие что $\kappa(x,x') = \langle \Phi_x, \Phi_{x'}\rangle_{\mathcal{H}}$; $\forall x, x' \in \Omega$. С точки зрения внутреннего произведения пространства: $\kappa(x,x') = \langle \kappa(\cdot,x), \kappa(\cdot,x')\rangle_{\mathcal{H}_\kappa(\Omega)}$. Дополнительно $\langle f, g\rangle = \langle g, f\rangle$ и

$$\langle f, f\rangle = \sum_{i,j=1}^{n} \alpha_i \alpha_j \kappa(x_i, x_j) \geq 0,$$

$$\left\langle \sum_{i=1}^{n} c_i f_i, \sum_{j=1}^{n} c_j f_j \right\rangle \geq 0.$$

Это означает, что $\langle\cdot,\cdot\rangle$ является положительно определенным ядром в пространстве признаков.

Равенство $\langle f, f\rangle = 0$ подразумевает $f = 0$ и:

$$|f(x)|^2 = |(\kappa(x,\cdot), f)|^2 \leq \kappa(x,x)\cdot\langle f, f\rangle.$$

Таким образом, $\langle\cdot,\cdot\rangle$ это хорошо определенное скалярное произведение.

Вспоминая воспроизводящее свойство положительно определенных ядер, мы видим, что для всех функций из $\mathcal{F}$ имеем $\langle \kappa(x,\cdot), f\rangle = f(x)$, и в частности (kernel trick):

$$\left\langle \Phi(x), \Phi(x') \right\rangle = k(x, x'). \qquad (12)$$

4. ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ

В этом разделе рассмотрим примеры, иллюстрирующие применения методов, изложенных в разделах 2 и 3.

Пример 1. Рассмотрим пример, иллюстрирующий применение метода определения расстояния между персистентными ландшафтами (см. раздел 2). Рассмотрим изображение House из пяти точек $[-1,0;\ 1,0;\ 1,2;\ -1,2;\ 0,3]$. Определим баркоды размерности 0: $2[0\ \ 1,4142)$, $2[0\ \ 2)$, $[0\ \ \infty)$; размерности 1: $[2\ \ 2,82825)$; см. табл. 1.

Таблица 1. Баркоды изображения House
[Table 1. The barcodes of the image «House»]

barcode	dim	birth	peak	death
bar1,2	0	(0,0)	(0.707, 0.707)	(1.41,0)
bar3,4	0	(0,0)	(1,1)	(2,0)
bar5	1	(2,0)	(2.414,0.414)	(2.828,0)

Функции персистентных ландшафтов $\lambda(k,t)$ для размерности 0:

$$\lambda^{House}(1,t) = t \cdot st\left(t,(0\ldots1]\right) + (2-t)\cdot st\left(t,(1\ldots2]\right),$$

$$\lambda^{House}(2,t) = t \cdot st\left(t,(0\ldots0.707]\right) + (1.414-t)\cdot st\left(t,(0.707\ldots1.414]\right),$$

где

$$st\left(t,(a\ldots b]\right) = \begin{cases} 1 & \text{if } t \in (a\ldots b], \\ 0 & \text{if } t \neq (a\ldots b]. \end{cases}$$

Рассмотрим теперь изображение House1 из пяти точек $[-1,0;\ 1,0;\ 1,2;\ -1,2;\ 0,4]$.

Определим баркоды размерности 0: 3[0, 2.0), [0, 2.233), $[0,\infty)$; размерности 1: [2.0, 2.828); см. табл. 2.

Функции персистентных ландшафтов $\lambda(k,t)$ для размерности 0:

$$\lambda^{House1}(1,t) = t \cdot st\left(t,(0\ldots1.116]\right) + (2.233-t)\cdot st\left(t,(1.116\ldots2.233]\right),$$

Таблица 2. Баркоды изображения House1
[Table 2. The barcodes of the image «House1»]

barcode	dim	birth	peak	death
bar1,2,3	0	(0,0)	(1.0, 1.0)	(2.0,0)
bar4	0	(0,0)	(1.116, 1.116)	(2.233,0)
bar5	1	(2,0)	(2.414,1.298)	(2.828,0)

$$\lambda^{House1}(2,t) = t \cdot st\left(t,(0\ldots1]\right) + (2-t)\cdot st\left(t,(1\ldots2]\right).$$

Для нахождения расстояния между изображениями House и House1 используем соотношение (6): $\left\|\lambda^{House} - \lambda^{House1}\right\|_2 = 0.5451$.

Из результатов данного примера можно сделать вывод о возможности нахождения расстояния между изображениями. □

Пример 2. Рассмотрим пример, аналогичный примеру 1, в котором находится расстояние между изображениями стеклянных бутылок.

Аппроксимируем четырнадцатью точками контур 2D изображения бутылки молока (в нотации Matlab):

```
qmilk_x=
[0,-1,-1.75,-1.75,-0.75,-1,-
1,1,1,0.75,1.75,1.75,1,0];
qmilk_y=
[0,0,1,6.5,9.25,9.25,10,10,9.25,9.25,6.5,
1,0,0] ;
plot(qmilk_x,qmilk_y).
```

и бутылки шампанского (в нотации Matlab):

```
qchamp_x=
[0,-1.25,-1.75,-1.75,-0.4,-0.5,
-0.5,0.5,0.5,0.4,1.75,1.75,1.25, 0];
qchamp_y=
[0,0,0.5,4,9.5,9.75,10,10,9.75,9.5,4,0.5,
0,0];
plot(qchamp_x,qchamp_y).
```

По полученным баркодам сформируем функции персистентных ландшафтов $\lambda(k,t)$ изображения бутылки молока для размерности 0:

$$\lambda^{milk}(1,t) = t \cdot st\left(t,(0\ldots2.75]\right) + (5.5-t)\cdot st\left(t,(2.75\ldots5.5)\right);$$

$$\lambda^{milk}(2,t) = t \cdot st\left(t,(0\ldots1.45]\right) + (2.9-t)\cdot st\left(t,(1.45\ldots2.9)\right);$$

$$\lambda^{milk}(3,t) = t \cdot st\left(t,(0\ldots0.75]\right) + (1.5-t)\cdot st\left(t,(0.75\ldots1.5)\right);$$

$$\lambda^{milk}(4,t) = t \cdot st\left(t,(0\ldots0.615]\right) + (1.23-t)\cdot st\left(t,(0.615\ldots1.23)\right);$$
$$\lambda^{milk}(5,t) = t \cdot st\left(t,(0\ldots0.5]\right) + (1.0-t)\cdot st\left(t,(0.5\ldots1.0)\right);$$

изображения для бутылки шампанского:

$$\lambda^{champ}(1,t) = t \cdot st\left(t,(0\ldots2.935]\right) + (5.87-t)\cdot st\left(t,(2.935\ldots5.87)\right);$$
$$\lambda^{champ}(2,t) = t \cdot st\left(t,(0\ldots1.75]\right) + (3.5-t)\cdot st\left(t,(1.75\ldots3.5)\right);$$
$$\lambda^{champ}(3,t) = t \cdot st\left(t,(0\ldots0.615]\right) + (1.23-t)\cdot st\left(t,(0.615\ldots1.23)\right);$$
$$\lambda^{champ}(4,t) = t \cdot st\left(t,(0\ldots0.5]\right) + (1.0-t)\cdot st\left(t,(0.5\ldots1.0)\right);$$

Для нахождения расстояния между контурами 2D изображений бутылки молока и бутылки шампанского используем соотношение (6):

$$\left\|\lambda^{milk}-\lambda^{champ}\right\|_2 = \sqrt{\sum_k \int_{-\infty}^{\infty} \left|\lambda^{milk}(k,t)-\lambda^{champ}(k,t)\right|^2 dt}.$$

В результате получим расстояние между аппроксимированными контурами 2D изображений бутылки молока и бутылки шампанского:

$$\left\|\lambda^{milk}-\lambda^{champ}\right\|_2 = 1.0012,$$

что указывает на возможность сравнения аппроксимированных контуров 2D изображений и распознавание различий между этими изображениями. □

Пример 3. Рассмотрим пример, иллюстрирующий применение метода ядра для персистентных баркодов и персистентных диаграмм (см. раздел 3). Рассмотрим изображение House из пяти точек $[-1,0;\ 1,0;\ 1,2;\ 1,2;\ -1,2;\ 0,4]$ с баркодами в размерности 0: $2[0\ \ 1,4142)$, $2[0\ \ 2)$, $[0\ \ \infty)$ и изображение House1 из пяти точек $[-1,0;\ 1,0;\ 1,2;\ -1,2;\ 0,4]$ с баркодами в размерности 0: 3[0, 2.0), [0, 2.233), $[0,\infty)$. Определим расстояние $d_{k_G}(House, House_1)^2$ на основе соотношения (9) при $w_{\text{arc}}^{House} = w_{\text{arc}}^{House1} = 1,\ \ 2\tau^2 = 1$: $\sum_{x\in L_k}\sum_{x'\in L_k} k_G(x,x') = 4.838$; $\sum_{y\in L'_k}\sum_{y'\in L'_k} k_G(y,y') = 5.841$; $\sum_{x\in L_k}\sum_{y\in L'_k} k_G(x,y) = 4.098$; $d_{k_G}^{w_{arc}}(L_k, L'_k)^2 = 2.482$. Для оценивания расстояния между изображениями используем формулу (8): $k_G(L_k, L'_k) = 0.0836$.

Из результатов данного примера можно сделать вывод о возможности нахождения расстояния между изображениями на основе метода RKHS. □

Из результатов определения расстояний в примерах 1 и 2 следует, что:

1) определить расстояния между изображениями на основе методов алгебраической топологии (используя информацию о числах Бетти) не представляется возможным;

2) определены расстояния между изображениями на основе методов топологического анализа данных;

3) для повышения точности определения расстояний можно использовать персистентные гомологии и фильтрацию не только на основе r-шаров, но и на основе других параметров (например, сканирования слева-направо или снизу-вверх), что повышает разнообразие информации и, следовательно, точность определения расстояний.

ЗАКЛЮЧЕНИЕ

Для повышения производительности моделей машинного обучения необходимо ввести функциональные возможности, способные сохранить внутреннюю информацию данных и уменьшить размерность данных. Метод PH из вычислительной топологии обеспечивает баланс между редукцией размерности данных и характеристиками внутренней структуры. Однако сочетанию машинного обучения и постоянных гомологий препятствуют топологические представления данных, метрики расстояния и представление объектов данных. В статье рассматриваются математические модели персистентных гомологий и функции персистентных ландшафтов представления признаков в машинном обучении. Функции персистентных ландшафтов отображают диаграммы пер-

систентности в гильбертово пространство. Рассмотрены представления топологических характеристик объектов в моделях машинного обучения. Представлена структура ядра для анализа персистентных диаграмм и метод персистентного взвешенного ядра Гаусса (PWGS — persistent weighted Gaussian kernel). Преимущество метода заключается в том, что метод PWGS позволяет контролировать персистентность при анализе данных.

Проводилось тестирование методов топологического анализа данных (на основе персистентной гомологии) по отношению к методам традиционной алгебраической топологии для определения расстоянии между изображениями цифр 6 и 9; при этом методы методов топологического анализа данных показали возможность определения такого расстояния, в то время как методы традиционной алгебраической топологии определяют такое расстояние как равное нулю.

По сравнению с методом формирования дескрипторов гистограмм ориентированного градиента (HOG [17]) методы формирования признаков на основе TDA не требуют нахождения направлений градиентов, инвариантны к евклидовым преобразованиям; они могут быть организованы по нескольким параметрам, что повышает разнообразие и надежность формируемой информации.

Методы формирования признаков машинного обучения на основе TDA могут быть использованы не только для исследования изображений объектов, но и в других приложениях; например, для идентификации признаков отказов сложных систем управления.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке Программы фундаментальных исследований СО РАН № I.5.1., проект № 0314-2019-0020 и Российского научного фонда, грант № 22-21-00035.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Carlsson G.* Topology and data. Bulletin of the American Mathematical Society. – 2009. – Vol. 46, No 2. – P. 255–308. DOI: 10.1090/S0273-0979-09-01249-X

2. *Edelsbrunner H., Harer J.* Computational topology: an introduction. – American Mathematical Soc., 2010.

3. *Kusano G., Hiraoka Y., Fukumizu K.* Persistence weighted Gaussian kernel for topological data analysis // International Conference on Machine Learning. – PMLR, 2016. – P. 2004–2013.

4. *Hofer C., Kwitt R., Niethammer M., Uhl A.* Deep learning with topological signatures. In Advances in Neural Information Processing Systems. – 2017. – P. 1634–1644.

5. *Hatcher A.* Algebraic Topology. – Cambridge UP. – 2005.

6. *Zomorodian A. J.* Topology for computing. – Cambridge UP. – 2005. – Vol. 16.

7. *Bubenik P.* The persistence landscape and some of its properties // Topological Data Analysis. – Springer, Cham. – 2020. – P. 97–117. DOI: 10.1007/978-3-030-43408-3_4

8. *Pun C. S., Xia K., Lee S. X.* Persistent-Homology-based Machine Learning and its Applications – A Survey // arXiv preprint arXiv:1811.00252. – 2018. DOI: 10.48550/arXiv.1811.00252

9. *Kwitt R., Huber S., Niethammer, M., Lin W., Bauer U.* Statistical topological data analysis – a kernel perspective. In Advances in Neural Information Processing Systems 28. Curran Associates, Inc. – 2015. – P. 3052–3060.

10. *Sriperumbudur B. K., Fukumizu K., Lanckriet G. R. G.* Universality, Characteristic Kernels and RKHS Embedding of Measures // Journal of Machine Learning Research. – 2011. – Vol. 12, No 7 – P. 2389–2410. DOI: 10.48550/arXiv.1003.0887

11. *Ghrist R.* Barcodes: the persistent topology of data // Bulletin of the American Mathematical Society. – 2008. – Vol. 45, No 1. – P. 61–75. DOI: 10.1090/S0273-0979-07-01191-3

12. *Mischaikow K., Nanda V.* Morse theory for filtrations and efficient computation of persistent homology // Discrete & Computational Geometry. – 2013. – Vol. 50, No 2. – P. 330–353. DOI: 10.1007/s00454-013-9529-6

13. *Xia K.* A quantitative structure comparison with persistent similarity // arXiv preprint arXiv:1707.03572. – 2017. DOI: 10.48550/arXiv.1707.03572

14. *Chukanov S. N.* Comparison of objects' images based on computational topology methods // Informatics and Automation. – 2019. – Vol. 18, No 5. – P. 1043–1065.

15. *Chukanov S. N.* The Comparison of Diffeomorphic Images based on the Construction of Persistent Homology // Automatic Control and Computer Sciences. – 2020. – Vol. 54, No 7. – P. 758–771. DOI: 10.3103/S0146411620070056

16. *Barbarossa S., Sardellitti S.* Topological signal processing over simplicial complexes // IEEE Transactions on Signal Processing. – 2020. – Vol. 68. – P. 2992–3007. DOI: 10.1109/TSP.2020.2981920

17. *Dalal N., Triggs B.* Histograms of oriented gradients for human detection // Comp. Vis. and Patt. Rec. – 2005. – Vol. 1. – P. 886–893. DOI: 10.1109/CVPR.2005.177

Чуканов Сергей Николаевич — д-р техн. наук, ведущий научный сотрудник Института математики им. С. Л. Соболева СО РАН (Омский филиал), проф.,
E-mail: ch_sn@mail.ru
ORCID iD: https://orcid.org/0000-0002-8106-9813

Чуканов Илья Станиславович — студент, Уральский федеральный университет им. первого президента России Б. Н. Ельцина.
E-mail: chukanov022@gmail.com
ORCID iD: https://orcid.org/0000-0001-9946-7484

DOI: https://doi.org/10.17308/sait/1995-5499/2022/3/115-126 ISSN 1995-5499
Received 01.06.2022
Accepted 30.09.2022

FORMATION OF FEATURES OF MACHINE LEARNING ON THE BASIS OF TOPOLOGICAL DATA ANALYSIS

© 2022 S. N. Chukanov[1]✉, I. S. Chukanov[2]

[1]Sobolev Institute of Mathematics of the Siberian Branch of Russian Academy of Sciences
13, Pevtsova Street, 644043 Omsk, Russian Federation
[2]Ural Federal University named after the First President of Russia B. N. Yeltsin
32, Mira Street, 620078 Yekaterinburg, Russian Federation

Annotation. At the present time, interest has increased in the use of algebraic topology methods for topological data analysis and the application of topological data analysis in various fields of knowledge. The goal of topological data analysis is to identify informative topological properties and use them as descriptors in machine learning. The application of machine learning methods for complex systems of large dimensions is difficult due to the methods of adequate representation of functions.

The persistent homology method from computational topology provides a balance between reducing the data dimension and characterizing the internal structure of an object. The combination of persistent homology and machine learning is hampered by topological representations of data, distance metrics, and representation of data objects. The paper uses the method of persistent homology, based on the use of filtering to assign a geometric dimension to each topological feature. The filtering process generates a series of simplicial complexes encoded with structural information of various scales. Persistent homology can be represented by a persistent barcode or a persistent diagram.

✉ Чуканов Сергей Николаевич
e-mail: ch_sn@mail.ru

The paper considers mathematical models and functions for representing persistent landscape objects based on the persistent homology method. Betty's persistent functions and persistent landscape functions are considered. The persistent landscape functions allow you to map persistent diagrams and persistent barcodes into Hilbert space. The representations of topological characteristics in various machine learning models are considered. The structure of the kernel for the analysis of persistent diagrams and the persistent weighted Gaussian kernel are considered. The persistent weighted kernel method allows you to control the persistence in data analysis. Distances between persistent landscapes are defined using the norm of the space L^p. Examples of finding the distance between images are given. The appendices present the basic concepts of algebraic topology and the Hilbert space reproducing kernel method for the purposes of machine learning.
Keywords: simplicial complex, persistent homology, persistent landscape, machine learning, RKHS, Hilbert space.

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Carlsson G.* (2009) Topology and data. *Bulletin of the American Mathematical Society*. Vol. 46, No 2. P. 255–308. DOI: 10.1090/S0273-0979-09-01249-X

2. *Edelsbrunner H., Harer J.* (2010) Computational topology: an introduction. *American Mathematical Soc*.

3. *Kusano G., Hiraoka Y., Fukumizu K.* (2016) Persistence weighted Gaussian kernel for topological data analysis. *International Conference on Machine Learning. PMLR, 2016*. P. 2004–2013.

4. *Hofer C., Kwitt R., Niethammer M., Uhl A*. (2017) Deep learning with topological signatures. *In Advances in Neural Information Processing Systems*. P. 1634–1644.

5. *Hatcher A*. (2005) Algebraic Topology. *Cambridge UP*.

6. *Zomorodian A. J.* (2005) Topology for computing. *Cambridge UP*. Vol. 16.

7. *Bubenik P.* (2020) The persistence landscape and some of its properties. *Topological Data Analysis. Springer, Cham*. P. 97–117. DOI: 10.1007/978-3-030-43408-3_4

8. *Pun C. S., Xia K., Lee S. X*. (2018) Persistent-Homology-based Machine Learning and its Applications – A Survey. arXiv preprint arXiv:1811.00252. DOI: 10.48550/arXiv.1811.00252

9. *Kwitt R., Huber S., Niethammer M., Lin W., Bauer U.* (2015) Statistical topological data analysis – a kernel perspective. *In Advances in Neural Information Processing Systems 28. Curran Associates, Inc*. P. 3052–3060.

10. *Sriperumbudur B. K., Fukumizu K., Lanckriet G. R. G.* (2011) Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*. Vol. 12, No 7 – P. 2389–2410. DOI: 10.48550/arXiv.1003.0887

11. *Ghrist R.* (2008) Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*. Vol. 45, No 1. P. 61–75. DOI: 10.1090/S0273-0979-07-01191-3

12. *Mischaikow K., Nanda V.* (2013) Morse theory for filtrations and efficient computation of persistent homology. *Discrete & Computational Geometry*. Vol. 50, No 2. P. 330–353. DOI: 10.1007/s00454-013-9529-6

13. *Xia K.* (2017) A quantitative structure comparison with persistent similarity. arXiv preprint arXiv:1707.03572. DOI: 10.48550/arXiv.1707.03572

14. *Chukanov S. N.* (2019) Comparison of objects' images based on computational topology methods. *Informatics and Automation*. Vol. 18, No 5. P. 1043–1065.

15. *Chukanov S. N.* (2020) The Comparison of Diffeomorphic Images based on the Construction of Persistent Homology. *Automatic Control and Computer Sciences*. Vol. 54, No 7. P. 758–771. DOI: 10.3103/S0146411620070056

16. *Barbarossa S., Sardellitti S.* (2020) Topological signal processing over simplicial complexes. *IEEE Transactions on Signal Processing*. Vol. 68. P. 2992–3007. DOI: 10.1109/TSP.2020.2981920

17. *Dalal N., Triggs B.* (2005) Histograms of oriented gradients for human detection. *Comp. Vis. and Patt. Rec*. Vol. 1. P. 886–893. DOI: 10.1109/CVPR.2005.177

Chukanov Sergey Nikolayevich — leading researcher at Sobolev Institute of Mathematics of the Siberian Branch of Russian Academy of Sciences (Omsk branch), professor, doctor of technical sciences.
E-mail: ch_sn@mail.ru
ORCID iD: https://orcid.org/0000-0002-8106-9813

Chukanov Ilya Stanislavovich — student at Ural Federal University named after the First President of Russia B. N. Yeltsin.
E-mail: chukanov022@gmail.com
ORCID iD: https://orcid.org/0000-0001-9946-7484