Персистентные гомологии и анализ гистологических данных

Георгий Каданцев, Александр Синицын (Санкт-Петербург, Россия)

Аннотация

Работа посвящена изучению гистологических изображений (WSI — Whole–Slide Imaging) при помощи методов топологического анализа данных. В частности, изображений рака толстой кишки. Основной характеристикой изображения для нас является персистентная энтропия, которая извлекается из нулевых симплициальных персистентентных гомологий изображения. Наша цель — показать, что персистентная энтропия может быть полезна для компьютерной диагностики различных видов рака, в том числе колоректального. В этой работе нами реализован алгоритм вычисления персистентной энтропии, проведен анализ набора патчей WSI–изображений здоровой ткани и колоректального рака. В энтропии изображений здоровой ткани и рака были найдены существенные различия. Данные наблюдения могут стать основой нового метода диагностики рака.

Содержание

1 Введение **2**

2 Основные определения **3**

2.1 (Дискретный) персистентный модуль 3

2.2 Интервальный дискретный персистентный модуль 4

2.3 Персистентная диаграмма 4

2.4 Энтропия 5

3 Анализ изображений **6**

4 Главные результаты **7**

1 Введение

Топологический анализ данных — область прикладной математики, в которой топологические идеи применяются для изучения различных видов данных [1], [2], [3]. В нашей работе мы применяем методы топологического анализа данных для изучения гистологических WSI–изображений (histological whole-slide images). Основной характеристикой изображения для нас является персистентная энтропия изображения, которая извлекается из нулевых персистентных гомологий фильтрованного пространства изображения. Эта характеристика уже показала себя как удобный инструмент при изучении идиотипической имунной сети (idiotypic immune network) (см. [7]). Грубо говоря, персистентная энтропия изображения – это численная мера хаотичности изображения, выраженная на языке топологического анализа данных.

Рак толстой кишки или колоректальный рак — злокачественная опухоль толстой кишки и ее придатка — червеобразного отростка. Данный вид рака второй наиболее часто диагностируемый среди мужчин и третий — у женщин. Наиболее распространенный форма этого рака — аденокарцинома (найден в 95% случаев рака толстой кишки), которая развивается в железистых клетках кишечника.

В обычном процессе диагностики патолог анализирует кусочки ткани под микроскопом и ищет определенные свойства строения клеток и то, как они варьируюстя. Однако это может быть сложно при большой трудовой нагрузке, поэтому с недавнего времени внимание особенное обращено на компьютерные методы анализа гистологических данных (цифровых копий препаратов в сверх-высоком разрешении).

Наша работа идейно близка к работе [6]. В ней разрабатывается алгоритм диагностики рака и алгоритм сегментации раковых клеток на изображении с использованием топологического анализа данных и сверточных нейронных сетей. Однако в алгоритме используются только простейшие топологические идеи (числа Бетти и их распределение, так называемые persistent homology profiles).

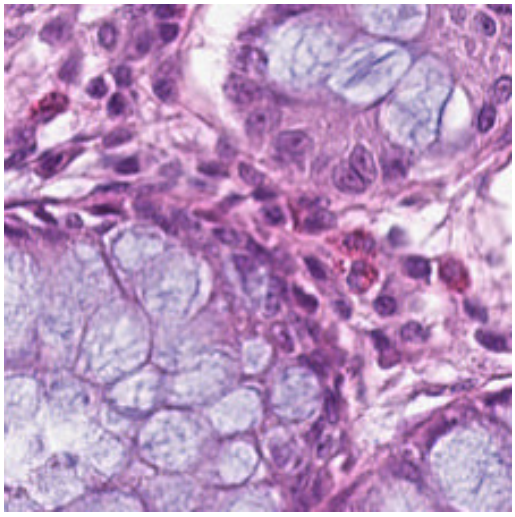

Рис. 1: Здоровая ткань

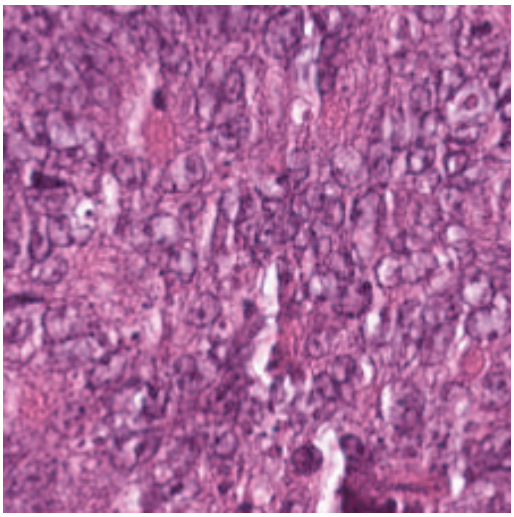

Рис. 2: Рак

Наша глобальная цель — решить задачу определения: представляет ли образец здоро-

вую ткань или опухоль. В отличии от работы [6], мы стараемся избежать использования искусственных нейронных сетей. Их недостаток заключается в том, что искусственная нейронная сеть — это "черный ящик", про который мы не вполне понимаем, как он работает. Использование таких "черных ящиков" имеет ряд очевидных минусов. Мы стремимся избавиться от них, используя более глубокие идеи топологического анализа данных. Цель этой работы — показать, что для решения данной задачи может быть полезно использование персистентной энтропии.

При исследовании в начале большое по размеру WSI–изображение делится на патчи (patches, маленькие кусочки исходного WSI–изображения), и каждый патч исследуется отдельно. На рисунках 1 и 2 мы приводим примеры патча здоровой ткани и патча с раковой тканью. Можно заметить, что правый патч выглядит более хаотично, в то время как на левом патче более ясно прослеживается структура. Персистентная энтропия численно учитывает эту разницу.

Мы предлагаем оригинальный алгоритм вычисления интервального разложения (и персистентной диаграммы) персистентных нулевых гомологий фильтрованного топологического пространства $X_0 \subseteq X_1 \subseteq \cdots \subseteq X_N$, основанный на естественном изоморфизме

$$H_0(X, F) \cong F^{(\pi_0(X))},$$

где F — произвольное поле, а $\pi_0(X)$ — множество компонент связности X. Мы рассматриваем отображения $\pi_0(X_n) \to \pi_0(X_m)$ для $n \leqslant m$ и обозначаем через $r_{n,m}$ количество элементов в его образе

$$r_{n,m} = \Big|\mathrm{Im}(\pi_0(X_n) \to \pi_0(X_m))\Big|.$$

Тогда число интервальных модулей $I(n, m)$ в интевальном разложении нулевых персистентных гомологий $H_0(X_*, F)$ равно

$$s_{n,m} = r_{n,m} - r_{n-1,m} - r_{n,m+1} + r_{n-1,m+1},$$

где $r_{n_0,m_0} = 0$, если $n_0 = n - 1 < 0$ или $m_0 = m + 1 > N$. (см. лемму 2). Судя по всему, эта лемма известна специалистам по топологическому анализу данных (см. [9]), но нам не удалось найти точную ссылку, поэтому далее мы приведем ее доказательство.

2 Основные определения

Далее мы подразумеваем, что у нас фиксировано поле F и все векторные пространства и линейные отображения рассматриваются над ним.

2.1 (Дискретный) персистентный модуль

(Дискретный) персистентный модуль длины k — это последовательность конечномерных векторных пространств и линейных отображений (без каких либо условий на них)

$$M: \quad M_1 \xrightarrow{f_1} M_2 \xrightarrow{f_2} \dots \xrightarrow{f_{k-1}} M_k.$$

Морфизм персистентных модулей длины k — это набор линейных отображений $\varphi_i : M_i \to N_i$ таких, что диаграмма коммутативна:

$$\begin{array}{ccc} M & M_1 \xrightarrow{f_1^M} M_2 \xrightarrow{f_2^M} \dots \xrightarrow{f_{k-1}^M} M_k \\ \downarrow \varphi & : \quad \downarrow \varphi_1 \qquad \downarrow \varphi_2 \qquad\qquad \downarrow \varphi_k \\ N & N_1 \xrightarrow{f_1^N} N_2 \xrightarrow{f_2^N} \dots \xrightarrow{f_{k-1}^N} N_k \end{array}$$

Изоморфизм персистентных модулей — это морфизм, компоненты которого — изоморфизмы.

Прямая сумма персистентных модулей $M \oplus N$ определяется покомпонентно:

$$M_1 \oplus N_1 \xrightarrow{f_1^M \oplus f_1^N} M_2 \oplus N_2 \xrightarrow{f_2^M \oplus f_2^N} \dots \xrightarrow{f_{k-1}^M \oplus f_{k-1}^N} M_k \oplus N_k.$$

2.2 Интервальный дискретный персистентный модуль

Пусть $0 \leqslant n \leqslant m \leqslant k$, тогда **интервальный дискретный персистентный модуль** $I(n, m)$ определяется как

$$I(n,m)_i = \begin{cases} F & i \in \{n, n+1, \dots, m\} \\ 0 & i \notin \{n, n+1, \dots, m\} \end{cases},$$

причем $f_i = \mathsf{id}$, при $n \leqslant i < m$.

Главным утверждением теории персистентных модулей является следующая теорема об интервальном разложении.

Теорема 1. *Пусть M — персистентный модуль длины k. Тогда существуют такие $n_1, \dots, n_t, m_1, \dots, m_t \in \{0, \dots, k\}$, что M изоморфен прямой сумме интервальных персистентных модулей.*

$$M \cong I(n_1, m_1) \oplus I(n_2, m_2) \oplus \dots \oplus I(n_t, m_t),$$

причем набор $(n_1, m_1), \dots, (n_t, m_t)$ определен однозначно с точностью до перестановки.

Такой изоморфизм называется **интервальным разложением персистентного модуля**.

2.3 Персистентная диаграмма

Персистентная диаграмма персистентного модуля — это мультимножество (множество с учетом кратностей), состоящее из пар (n, m), где $I(n, m)$ — интервальный модуль из интервального разложения M. Более строго можно сказать, что мультимножество — это пара $(X, s : X \to \mathbb{N}_0)$, где X — обычное множество и s — отображение в натуральные числа с нулем. Для любого пресистентного модуля есть изоморфизм $M \cong \bigoplus_{n \leqslant m} I(n, m)^{s_{n,m}}$, где $s_{n,m} \in \mathbb{N}_0$. Тогда **персистентная диаграмма** M — это пара $(X, s : X \to \mathbb{N}_0)$, где $X = \{(n, m) \mid 0 \leqslant n \leqslant m \leqslant k\}$ и $s(n, m) = s_{n,m}$.

2.4 Энтропия

Энтропия чисел $0 \leqslant p_1, \ldots, p_t \leqslant 1$ таких, что $p_1 + \cdots + p_t = 1$, определяется как

$$H(p_1, \ldots, p_t) = -\sum_i p_i \log_2(p_i).$$

Персистентная энтропия. Пусть $l_1, \ldots, l_t$ — длины интервалов в интервальном разложении персистентного модуля M. Обозначим $L = \sum l_i$ и отнормируем эти длины $p_i := \dfrac{l_i}{L}$. Персистентная энтропия M определяется как энтропия чисел p_i. Это интересный инвариант персистентного модуля, который показал себя полезным в [4].

Фильтрованное топологическое пространство X_* — это топологическое пространство X вместе с цепочкой подпространств

$$X_0 \subseteq X_1 \subseteq \cdots \subseteq X_k = X.$$

Нулевые персистентные гомологии фильтрованного пространства X_* — это персистеный модуль M, для которого

$$M_n = H_0(X_n, F)$$

и отображения $M_n \to M_{n+1}$ индуцируются вложениями $X_n \hookrightarrow X_{n+1}$.

Фильтрованное пространство изображения. Для черно–белого изображения размера $p \times q$ и числа $0 \leqslant n \leqslant 255$ мы определим подмножество X_n квадрата $[0, p] \times [0, q] \subseteq \mathbb{R}^2$ как объединение замкнутых квадратов $[i, i+1] \times [j, j+1]$ таких, что интенсивность пикселя с координатами (i, j) меньше n. Тогда множества X_n образуют фильтрованное пространство

$$X_0 \subseteq X_1 \subseteq \cdots \subseteq X_{255} \subseteq \mathbb{R}^2.$$

Персистетная энтропия изображения — это персистентная энтропия нулевых персистентных гомологий фильтрованного пространства изображения.

Одним из главных преимуществ персистентной гомологии перед другими средствами анализа данных является строго обоснованная устойчивость к шуму (отсюда слово *персистентный*), т. е. при небольших изменениях исходных данных штрих-код особо не поменяется.

Судя по всему, следующая лемма хорошо известна специалистам, но мы не нашли точной ссылки, поэтому приводим ее с доказательством.

Лемма 1. *Пусть M — персистентный модуль длины k и для $0 \leqslant n \leqslant m \leqslant k$ мы обозначим*

$$r_{n,m} = \operatorname{rank}(M_n \to M_m).$$

Тогда количество $s_{n,m}$ интервальных модулей $I(n, m)$ в интервальном разложении M вычисляется по формуле

$$s_{n,m} = r_{n,m} - r_{n-1,m} - r_{n,m+1} + r_{n-1,m+1}$$

(здесь мы подразумеваем $r_{-1,m} = r_{n,k+1} = 0$).

Доказательство. Заметим, что $r_{n,m}$ — это количество интервальных модулей $I(n_0, m_0)$ в интервальном разложении M для которых $n_0 \leqslant n$ и $m \leqslant m_0$. Поэтому, чтобы найти $s_{n,m}$ мы пользуемся формулой "включений–исключений":

$$s_{n,m} = r_{n,m} - r_{n-1,m} - r_{n,m+1} + r_{n-1,m+1}.$$

Из числа интервалов, начинающихся до (и включая) n и заканчивающихся после (и включая) m вычитается количество интервалов, которые начинаются строго до n и заканчиваются после m, а также начинаются до n и заканчиваются строго после m. Также необходимо добавить количество интервалов, которые начинаются строго до n и заканчиваются строго после m, потому что они были исключены два раза.

Лемма 2. *Пусть X_* — фильтрованное топологическое пространство и*

$$r_{n,m} = \Big|\mathrm{Im}(\pi_0(X_n) \to \pi_0(X_m))\Big|.$$

Тогда число интервальных модулей $I(n, m)$ в интевальном разложении нулевых персистентных гомологий $H_0(X_*, F)$ равно

$$s_{n,m} = r_{n,m} - r_{n-1,m} - r_{n,m+1} + r_{n-1,m+1},$$

где $r_{n_0,m_0} = 0$, если $n_0 < 0$ или $m_0 > k$.

3 Анализ изображений

Нами исследуются изображения в формате RGB 224×224 пикселей здоровой эпителиальной ткани кишечника и опухоли из набора (https://clck.ru/LqX7n).

В анализе изображений мы сталкиваемся с проблемой вариативности условия получения изображений ткани, разнообразия в интенсивности изображений. Чтобы привести все изображения к одному виду, мы пользуемся методом нормализации гистологичскиxx изображений, описанный в [8]. Пользуясь этим методом, мы выделяем два красителя — эозин и гематоксилин — и исследуем их по отдельности. Мы создаем матрицу 256×256 в которую будем записывать $r_{n,m}$, $n \leqslant m$. Каждое такое X_n мы получаем следующим образом:

Каждый пиксель становится:

- черным, если сумма его цветов $\dfrac{r+g+b}{3} < n$;
- белым, если сумма его цветов $\dfrac{r+g+b}{3} \geqslant n$.

После такой фильтрации изображения, картинка становится черно–белой и можно реализовать обход графа в глубину. Так мы получаем множество компонент связности, с которыми в дальнейшем мы будем работать.

Мы вычисляем нулевые гомологии, подсчитывая компоненты связности для каждого $0 \leqslant t \leqslant 255$, множество которых мы обозначаем за $\pi(B_t(X))$. Можно рассмотреть

отображение $\pi(B_n(X)) \to \pi(B_m(X))$ и вычислить $r_{n,m} = \left|\mathrm{Im}(\pi(B_n(X)) \to \pi(B_m(X)))\right|$, где $0 \leqslant n \leqslant m \leqslant 255$. Затем $s_{n,m}$ вычисляется по своеобразной формуле "включений–исключений":

$$s_{n,m} = r_{n,m} - r_{n-1,m} - r_{n,m+1} + r_{n-1,m+1}.$$

В случае, если индексы $n-1$ или $m+1$ выходят за границы отрезка $[0, 255]$, соответствующее значение r полагается равным 0.

Кроме этого, мы еще вычисляем персистентные гомологии инвертированной картинки. Так, образно говоря, то, что в фильтрованном пространстве являлось компонентой связности становится дыркой и наоборот.

4 Главные результаты

Мы обнаружили, что для различия здоровых тканей и зараженных раком наиболее полезна энтропия нулевой персистентной гомологии инвертированного изображения по каналу эозина. Видно, что среднее арифметическое значение энтропии различается: у здоровых $\overline{n} = 8.51$, в то время как среднее значение энтропии изображений рака — $\overline{t} = 8.96$. Стандартное статистическое отклонение примерно одинаковое, с $\mu_n = 0.26$ у здоровых тканей и с $\mu_t = 0.24$ у рака.

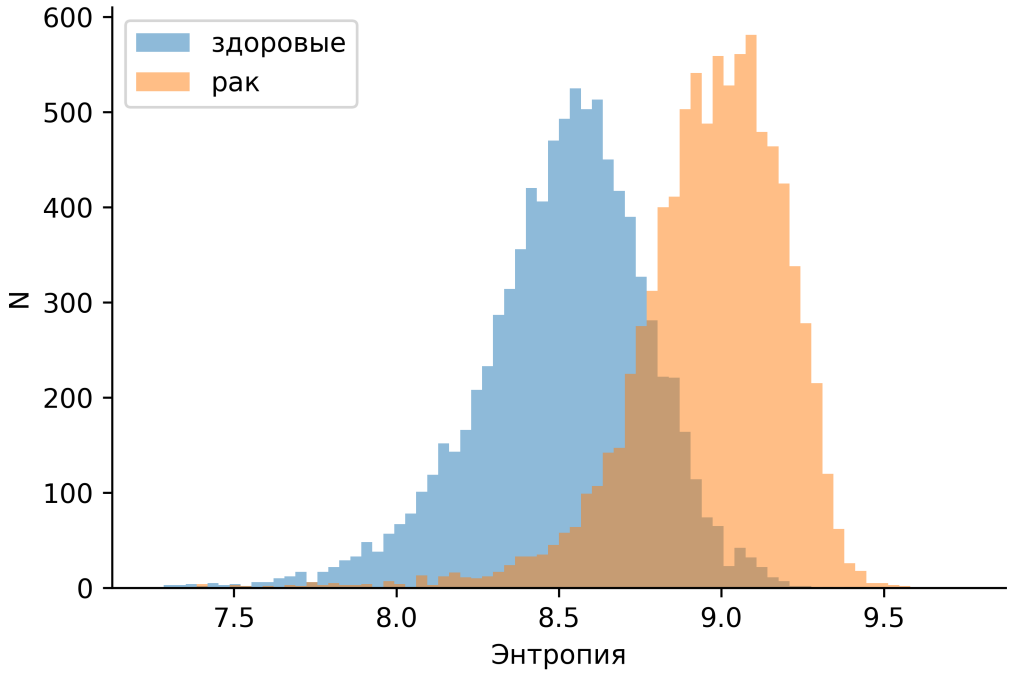

Рис. 3: Распределение значений энтропии

Мы проанализировали 9 тысяч картинок здоровой ткани и 23 тысячи изображений

рака. На графике 3 представлено распределение энтропии 9 тысяч "здоровых" изображений и 9 тысяч изображений "рака".

Авторы статьи выражают благодарность (к.ф.-м.н., старший научный сотрудник лаборатории "Современная Алгебра и Приложения" с СПбГУ.) Сергею Олеговичу Иванову за проделанную работу и полезные обсуждения.

Приложение

- Данные для анализа можно найти по ссылке: Kather, Jakob Nikolas, Halama, Niels, & Marx, Alexander. (2018). 100,000 histological images of human colorectal cancer and healthy tissue (Version v0.1) [Data set].
  Zenodo. http://doi.org/10.5281/zenodo.1214456
- Код программы находится в общем доступе на github:
  https://github.com/Sannitsa/Persistent-homology-and-histological-data/

Список литературы

[1] G. Carlsson: Topology and data. Bullentin of the American Mathematical Society 46, 2 (2009), 255–308.

[2] S. Y. Oudot: Persistence Theory: From Quiver Representations to Data Analysis, vol. 209 of Mathematical Surveys and Monographs. American Mathematical Society (2015).

[3] H. Edelsbrunner, D. Morozov: Persistent homology: theory and practice. In Proceedings of the European Congress of Mathematics, pages 31–50, (2012).

[4] N. Atienza, R. Gonzalez Diaz, M. Soriano Trigueros: On the stability of persistent entropy and new summary functions for Topological Data Analysis. arXiv:1803.08304v6 (2019)

[5] N. Atienza, L. M. Escudero, M. J. Jimenez, M. Soriano-Trigueros: Persistent entropy: a scale-invariant topological statistic for analyzing cell arrangements. arXiv:1803.08304v6 (2019)

[6] T. Qaiser, Y.-W. Tsang, and N. Rajpoot et al.: Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. Mèd Image Anal., 55:1-14, (2019).

[7] M. Rucco, F. Castiglione, E. Merelli, and M. Pettini.: Characterisation of the idiotypic immune network through persistent entropy. In Proceedings of ECCS 2014, 117–128. Springer, (2016).

[8] Macenko, Marc & Niethammer, Marc & Marron, J. & Borland, David & Woosley, John & Guan, Xiaojun & Schmitt, Charles & Thomas, Nancy. (2009). A Method for Normalizing Histology Slides for Quantitative Analysis. Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009. 9. 1107-1110. 10.1109/ISBI.2009.5193250.

[9] Gunnar Carlsson, Afra Zomorodian: The Theory of Multidimensional Persistence

НЦ Лаборатория Непрерывного Математического Образования

12 января 2020