Воронежский Государственный Университет
Факультет прикладной математики, информатики и механики

Кафедра вычислительной математики и прикладных информационных технологий

Применение персистентных гомологий для задачи классификации изображений

Бакалаврская работа
Направление 01.03.02 Прикладная математика и информатика
Профиль Математическое моделирование и вычислительная математика

Обучающийся		Снопов П.М.
Руководитель	д.т.н., проф.	Леденева Т.М.

Содержание

Цель и задачи работы

Цель: Исследование подхода, основанного на топологическом анализе данных, для классификации изображений

Задачи:

- Изучение теоретических и практических основ топологического анализа данных
- Анализ подходов классификации изображений
- Формирование алгоритма на основе персистентных гомологий
- Проведение вычислительного эксперимента, выявление области применимости, плюсов и минусов данного подхода

Актуальность

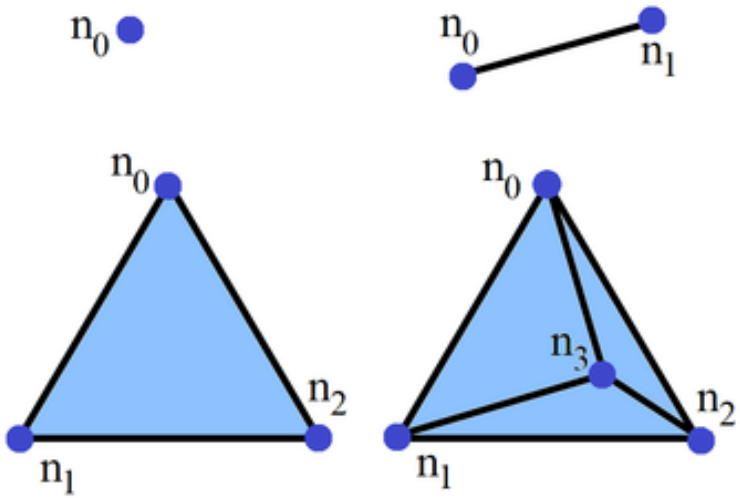
- Классификация изображений – фундаментальная задача анализа данных, одно из основных направлений ее развития, которое в последнее время имеет очень интенсивное развитие, связанное с достижениями нейронных сетей.
- Основным инструментом ТДА являются персистентные гомологии, о которых можно думать как об адаптации понятия гомологии к облаку точек. С помощью такого инструмента можно выявлять топологические характеристики исследуемого объекта.
- Задача классификации по сути является задачей определения характеристических свойств, которым удовлетворяют объекты одного класса. Поэтому персистентные гомологии могут быть полезны в данной задаче, так как зачастую такие свойства имеют геометрическую природу.

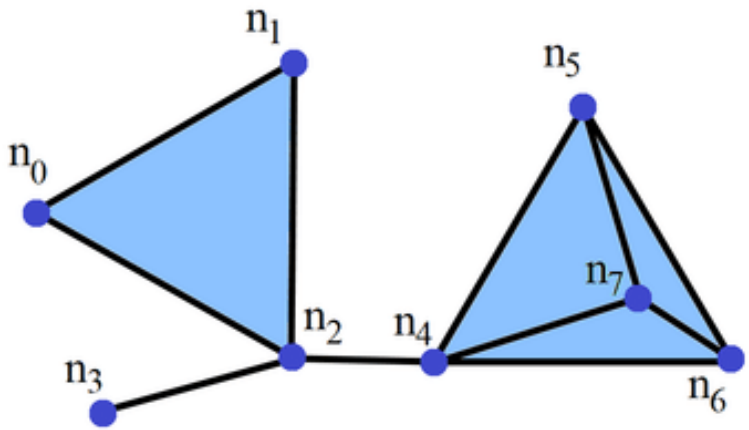
Симплициальные комплексы

Definition

Симплициальный комплекс K – это множество симплексов, т.е. выпуклых оболочек набора $n+1$ точек $\in \mathbb{R}^p$, таких, что векторы $x_1 - x_0, ..., x_n - x_0$ линейно независимы, при этом

- Для каждого симплекса из K его грани тоже лежат в K,
- Пересечение любых двух симплексов $\sigma, \tau \in K$ либо пусто, либо является гранью и σ, и τ.





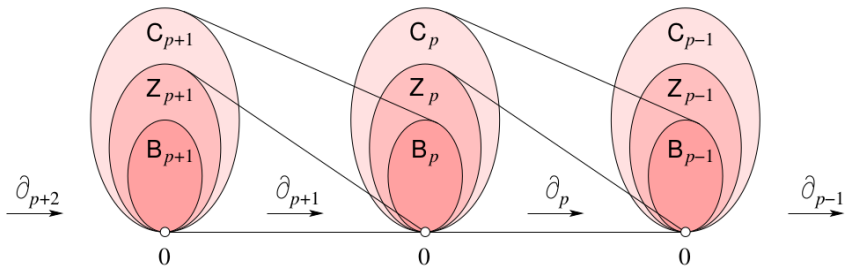
Симплициальные гомологии

Рассмотрим свободную абелеву группу $C_k(K)$ k-цепей симплициального комплекса K и гомоморфизм $\partial_k : C_k(K) \to C_{k-1}(K)$ называют *граничным оператором*. Он удовлетворяет следующему свойству:

$$\partial_{k-1} \circ \partial_k = 0.$$

T.e. $\operatorname{im} \partial_{k+1} \leq \ker \partial_k \leq C_k(K)$.
Последовательность $C_k(K)$ и ∂_k называется *цепным комплексом*

$$\ldots \xrightarrow{\partial_{k+2}} C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} \ldots \xrightarrow{\partial_1} C_0.$$



Симплициальные гомологии

Definition

k-ой группой гомологий симплициального комплекса K называют следующую фактор-группу:

$$H_k(K) = \ker \partial_k \,/\, \mathrm{im}\, \partial_{k+1}.$$

Тогда k-ое число Бетти – размерность k-ой группы гомологий: $\beta_k(K) = \dim H_k(K)$.

При $k = 0$ число Бетти описывает количество компонент связности данного пространства. При $k = 1$ – количество циклов. При $k = 2$ число Бетти описывает количество "полостей".

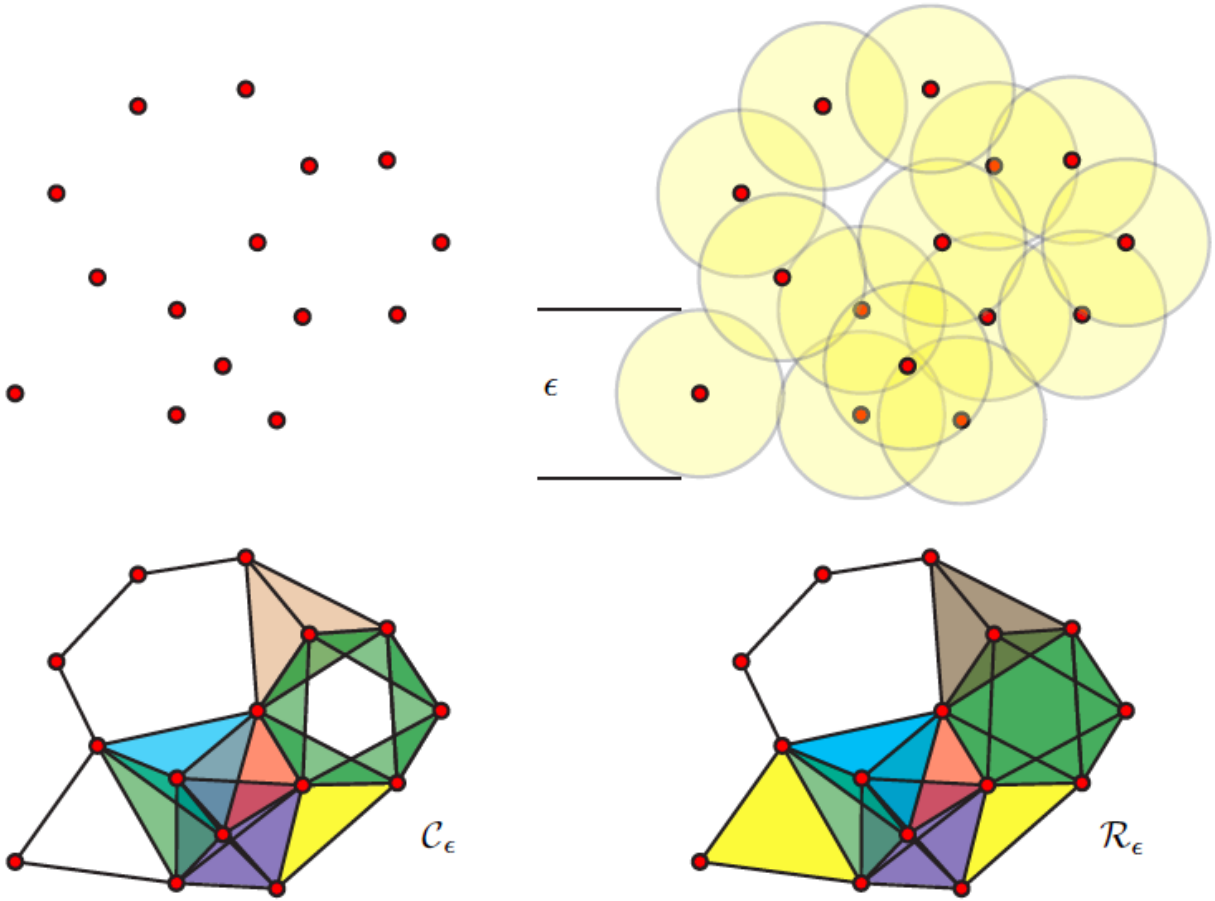
Симплициальные гомологии

Пространство	β_0	β_1	β_2
Pt	1	0	0
D^2	1	0	0
Треугольник	1	0	0
Граница треугольника	1	1	0
S^1	1	1	0
S^2	1	0	1
$\mathbb{T}^2 = S^1 \times S^1$	1	2	1

Таблица: Первые числа Бетти для некоторых пространств

Как построить симплициальный комплекс?

Имея облако точек $X \subset \mathbb{R}^n$, можно построить симплициальный комплекс по нему. Наиболее популярные – симплициальные комплексы *Чеха* и *Вьеториса—Рипса*.

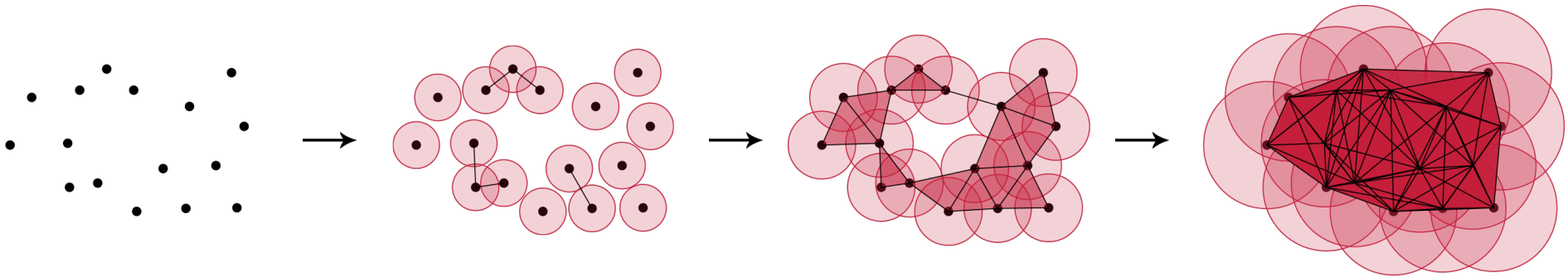


Фильтрации и устойчивые гомологии

Фильтрацией симплициального комплекса K называют вложенное семейство подкомплексов $(K_\tau)_{\tau \in T}$, такое, что если $\tau < \tau^{'}$, то $K_\tau \subseteq K_{\tau'}$.

Definition

n-ыми устойчивыми гомологиями фильтрованного комплекса $(K_\tau)_{\tau \in T}$ называют проиндексированное семейство абелевых групп и гомоморфизмов между ними $H_n(T) = \{(H_n(K_\tau))_{\tau \in T}, (H_n(K_\tau) \to H_n(K_\tau^{'}))_{\tau \leq \tau'}\}$.



Фильтрации и устойчивые гомологии

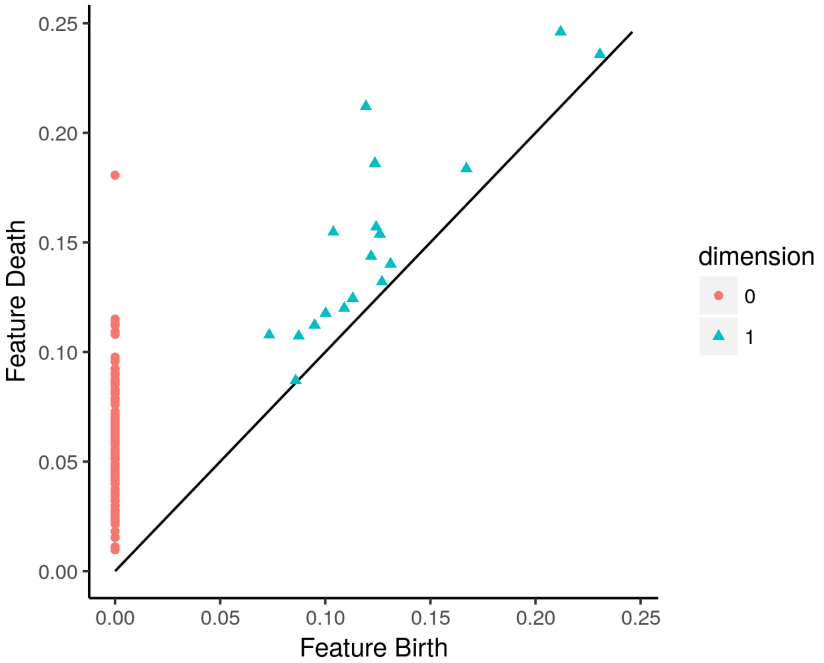
Устойчивые гомологии – это пример конечнопорожденного персистентного модуля. Главная теорема для конечнопорожденного персистентного модуля – это структурная теорема:

Theorem (Структурная теорема для конечнопорожденного персистентного модуля)

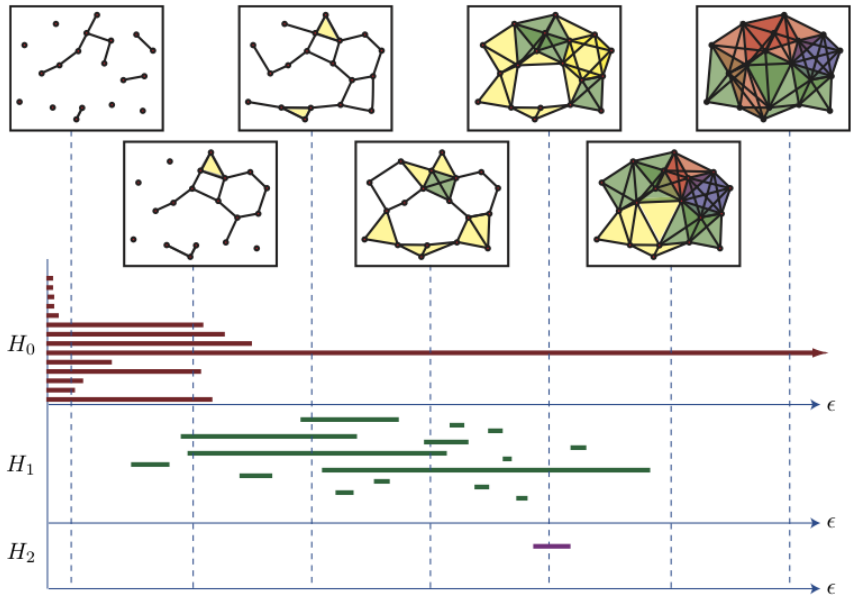
Пусть R – произвольное поле, а (A_*, x) – конечнопорожденный персистентный R-модуль. Тогда (A_*, x) имеет единственное с точностью до перестановки слагаемых представление в виде прямой суммы конечного числа интервальный модулей:

$$(A_*, x) \simeq \left(\bigoplus_k I_{[j_k, s_k]}\right) \oplus \left(\bigoplus_k I_{[r_k, \infty)}\right)$$

Фильтрации и устойчивые гомологии



(a) Диаграмма персистентности



(b) Баркод

Рис.: Способы кодирования информации об персистентных гомологиях

Метрическое пространство всех диаграмм персистентности

На множестве $\mathcal{D}$ диаграмм персистентности можно ввести структуру метрического пространства. Например, можно ввести т.н. *p−метрику Васерштейна* W_p, где $1 \leq p < \infty$:

$$W_p(B, B^{'}) = \inf_{\gamma: B \to B^{'}} \left(\sum_{u \in B} \|u - \gamma(u)\|_{\infty}^{p} \right)^{\frac{1}{p}},$$

где $B, B^{'}$ – диаграммы персистентности.
Другой естественной метрикой является т.н. *bottleneck distance* W_∞:

$$W_\infty(B, B^{'}) = \inf_{\gamma: B \to B^{'}} \sup_{u \in B} \|u = \gamma(u)\|_{\infty}$$

Векторизация диаграмм персистентности

Векторизацией для пространства диаграмм персистентности называют отображение $\varphi : \mathcal{D} \to V$, где V – векторное пространство.
Амплитудой на $\mathcal{D}$ называют отображение $A : \mathcal{D} \to \mathbb{R}$, для которого $\exists \varphi : \mathcal{D} \to V$, где V – нормированное пространство, такое, что $\forall B \in \mathcal{D} : \quad A(B) = \|\varphi(B)\|$.
Персистентной энтропией $E(B)$ диаграммы $B = \{(b_i, d_i)\}_{i \in I}$ называют меру энтропии по Шеннону точек на диаграмме персистентности:

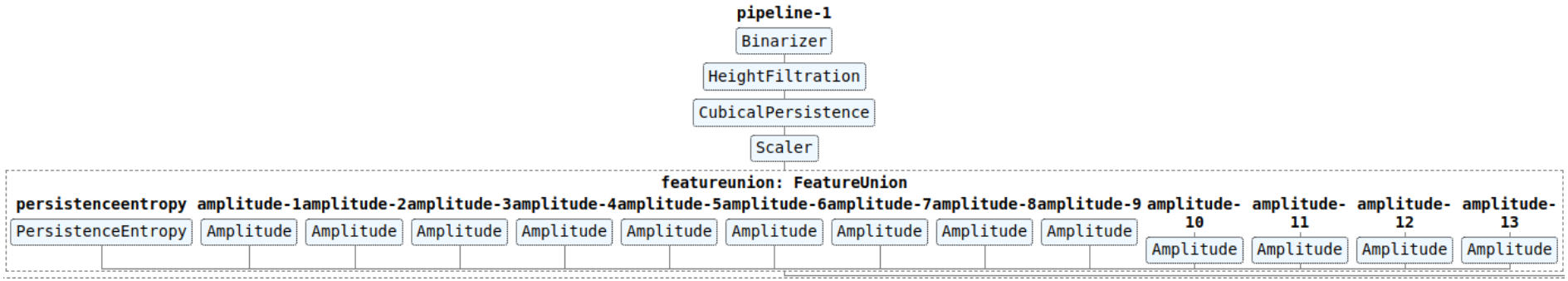
$$E(B) = -\sum_{i \in I} p_i \log(p_i), \text{ где } p_i = \frac{d_i - b_i}{\sum\limits_{i \in I} d_i - b_i}.$$

Общая идея алгоритма

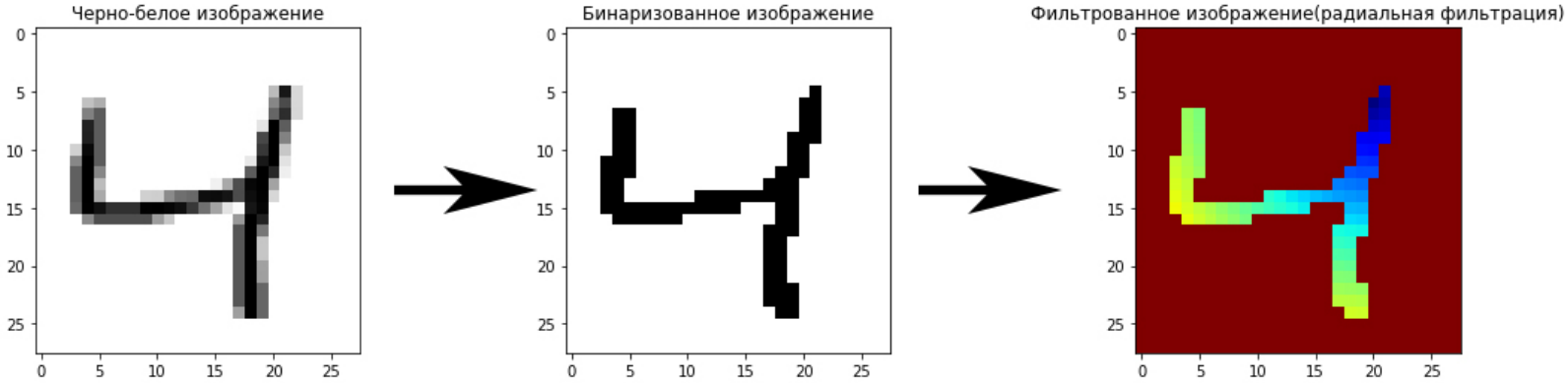
- По изображению строим фильтрацию;
- По построенной фильтрации находим кубический комплекс, персистентные гомологии и строим диаграмму устойчивости;
- Векторизуем диаграмму устойчивости, получаем векторное представление, которое можно использовать в моделях машинного обучения.

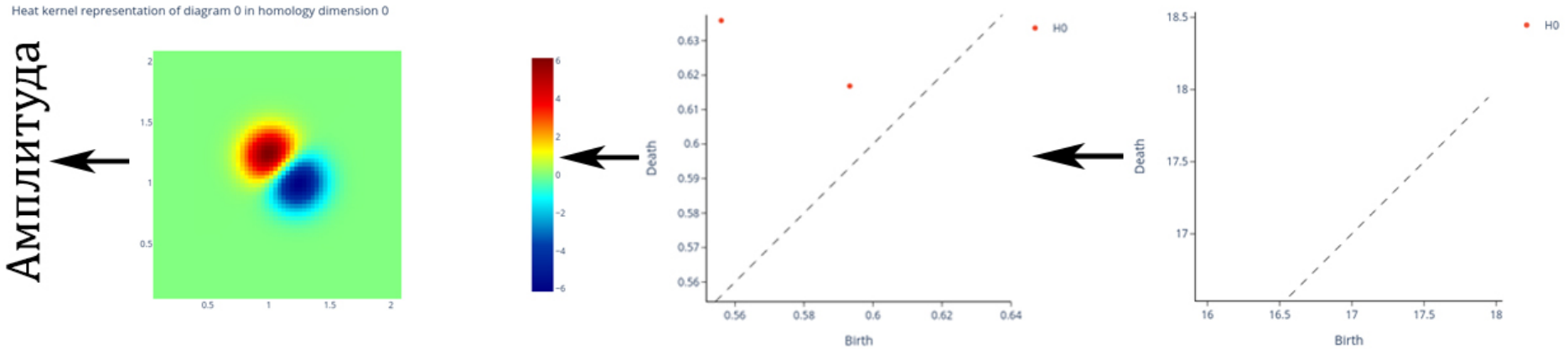
Пайплайн используемого метода построения векторного представления

Построить фильтрацию для изображения можно различными методами. В настоящей работе было построено 17 разнообразных фильтрацией, для каждой из которых были получены 0 и 1 персистентные гомологии, из диаграмм которых было получено 14 признаков. Таким образом, для одной картинки было получено $17 \times 2 \times 14 = 476$ признаков.



Процесс получения векторного представления





Сравнение результатов различных методов машинного обучения

Модель	Точность на тестовой выборке
Случайный лес	0.88
Логистическая регрессия	0.9
Метод опорных векторов	0.88
LightGBM	0.86
CatBoost	0.88
XGBoost	0.88

Таблица: Сравнение результатов различных методов машинного обучения

Сравнение результатов различных методов машинного обучения

Модель	Точность на тестовой выборке
Случайный лес	0.88
Логистическая регрессия	0.9
Метод опорных векторов	0.9
LightGBM	0.88
CatBoost	0.89
XGBoost	0.89

Таблица: Сравнение результатов различных методов машинного обучения с подобранными гиперпараметрами

Подбор гиперпараметров и отбор признаков

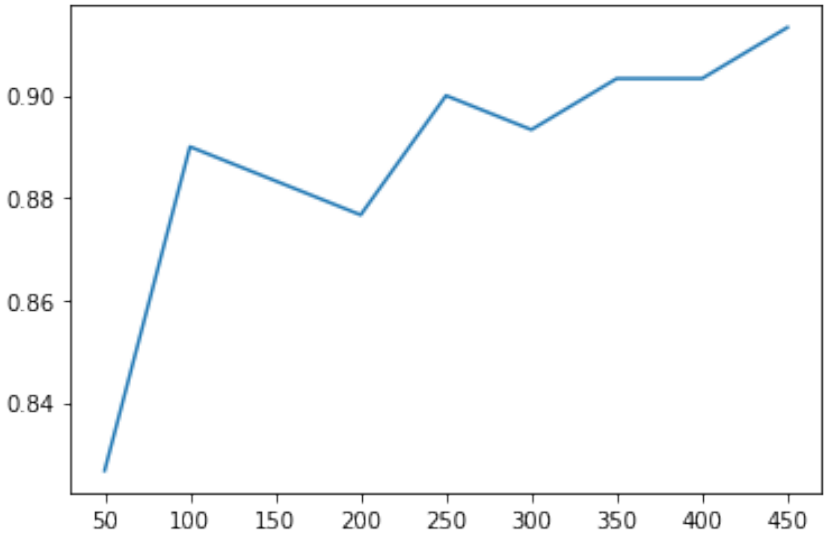


График зависимости точности на обучающей выборке и числа признаков

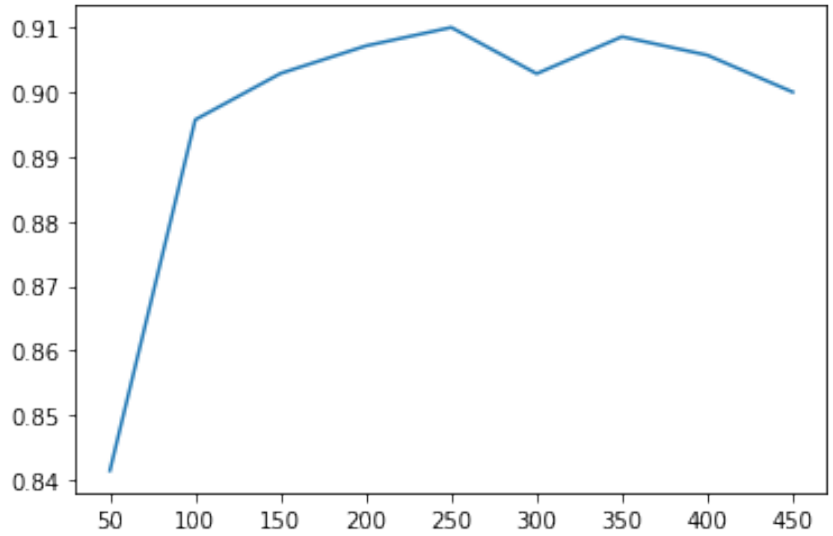


График зависимости точности на валидационной выборке и числа признаков

Что осталось сделать?

- Сравнить различные модели, которые обучались на топологических признаках, с теми же моделями, но которые обучаются прямо на картинке(т.е. в качестве признака берется каждый пиксель). Сравнить зависимости точности от количества признаков. Ожидаемый эффект: модели, обученные на топологических признаках, дают сравнимую точность при гораздо меньшем количестве признаков.
- Использовать топологический пайплайн как один из слоев нейросети, применить такую модель, сравнить результаты.
- Увеличить количество топологических признаков: использовать другие методы фильтрации и векторизации.
- Попробовать сделать отбор признаков на основе, например, коэффициента корреляции Пирсона.
- Попробовать обучить модели на полном датасете. Вероятно тут нужно будет использовать батчи для создания векторного представления. Можно будет попробовать связанные с этим трюки в духе batch normalization.

Воронежский Государственный Университет
Факультет прикладной математики, информатики и механики

Кафедра вычислительной математики и прикладных информационных технологий

Применение персистентных гомологий для задачи классификации изображений

Бакалаврская работа
Направление 01.03.02 Прикладная математика и информатика
Профиль Математическое моделирование и вычислительная математика

Обучающийся Снопов П.М.
Руководитель д.т.н., проф. Леденева Т.М.