Применение гиперболических вложений данных в машинном обучении
Часть 2. Применение метода развёртки в машинном обучении

Александра Бернадотт

НИУ ВШЭ, ФКН, Лаборатория алгебраической топологии и её приложений, Москва
МИСиС, Кафедра инженерной кибернетики, Москва
ООО Нейроспутник, Москва

Москва, 2020

Цель и задачи

Цель:

- оптимизировать точность классификации состояний мозга

Задачи:

- максимизировать точность классификации патологических состояний (алкоголизм, нейродеградация) и состояний здорового мозга;
- максимизировать точность классификации состояний, соответствующих (квази-)движениям.

Часть 1. Применение гиперболических вложений данных в машинном обучении

Почему выбрали методы гиперболической геометрии?

Часть 1. Применение гиперболических вложений данных в машинном обучении

В евклидовом пространстве существуют различные псевдосферы, имеющие конечную площадь постоянной отрицательной гауссовой кривизны.

Theorem 1 (Гильберта, 1901)

Никакая полная регулярная поверхность постоянной отрицательной гауссовой кривизны не погружается в $\mathbb{R}^3$

Наши данные тип 1: 3Д вектор значений из Q. Для задачи оценки морфо-функционального состояния мозга.

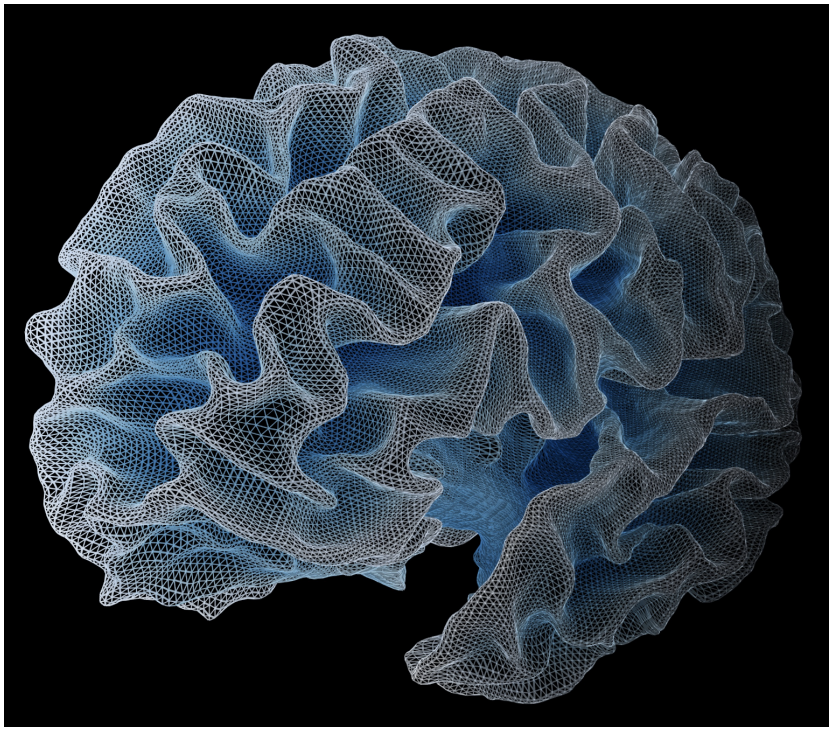
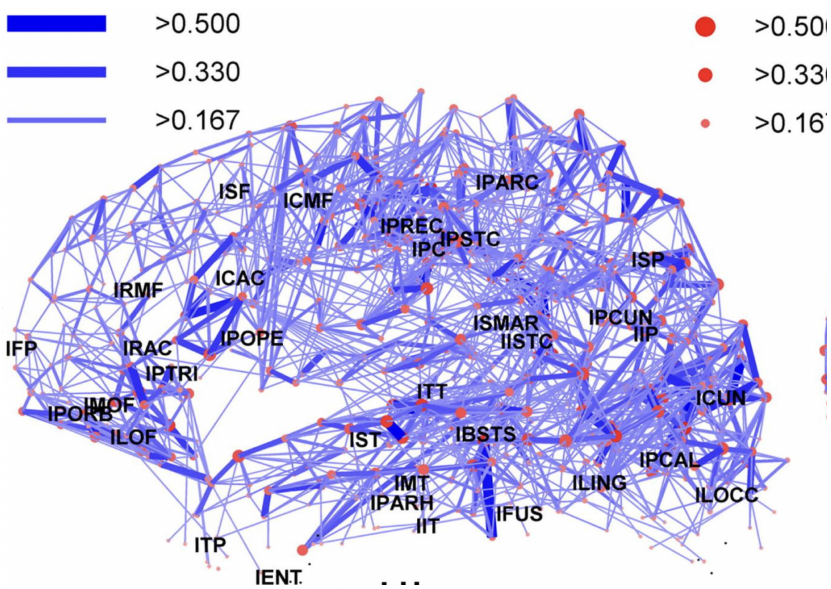


Рис.: Коннектом коры на микроуровне.

Часть 1. Применение гиперболических вложений в МО

Наши данные тип 2: коннектом на макроуровне в виде нагруженного ориентированного графа с координатами в выбранном метрическом пространстве. $G = (V, E, \varphi, \psi, \omega)$, V – где вершины графа $v \in V$, E – рёбра графа, φ – функция инцидентности, сопоставляющая каждому ребру $e \in E$ пару его вершин $v_i, v_j \in V$, ψ – функция лакализации, сопоставляющая каждой вершине координаты в соответсвующем метрическом пространстве, ω – функция нагруженности вершины (из R) от времени.



Наши данные тип 3: сементически размеченные временные ряды.
В задаче распознавания квази-движений мы смотрим на сигнал мозга в виде временных рядов как на иерархическую модель, представленную графовым древом.

Рис.: ЭЭГ сигнал мозга с визуализацией источников сигнала.

Часть 1. Применение гиперболических вложений в МО

Распознавание квази-движений (состояний мозга).

- Подзадача 1: представить временные ряды электро-магнитного сигнала мозга, соответствующего функциям квазидвижений и движений, в виде иерархической структуры и графа;
- Подзадача 2: предложить алгоритмы гиперболических вложений с учётом семантических связей для анализа электро-магнитного сигнала мозга, соответствующего функциям квазидвижений и движений, позволяющий повысить точность и производительность алгоритмов классификации.
- Подзадача 3: продемонстрироваьт результаты на экспериментальных данных. Получить дополнительную информацию о структуре семантических связей квазидвижений и движений, представленных электро-магнитным сигналом мозга.

Распознавание квази-движений (состояний мозга).

Рис.: Графовое представление данных, согласно семантике.

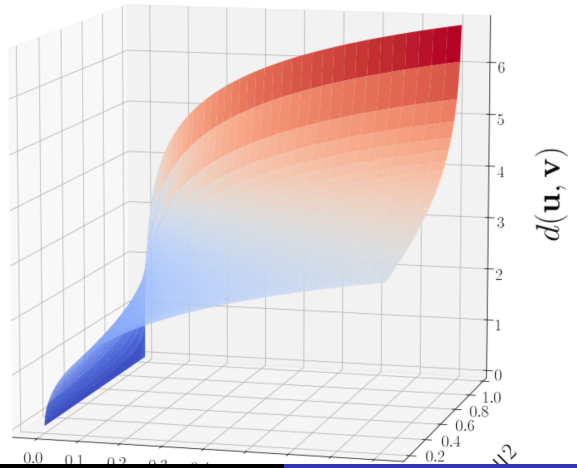
Эмбединг с использованием шара Пуанкаре.

Модель Пуанкаре гиперболического пространства соответствует риманову многообразию (B^n, g_x), т. е. открытому единичному шару, снабженному римановым метрическим тензором:

$$\delta(u, v) = 2\frac{\|u - v\|^2}{(1 - \|u\|^2)(1 - \|v\|^2)} \tag{1}$$

где u и v – вектора в $\mathbb{R}^n$ с евклидовой нормой меньше единицы $\|u\| < 1, \|v\| < 1$. Функция расстояния:

$$d(u, v) = arcosh(1 + \delta(u, v)) \tag{2}$$

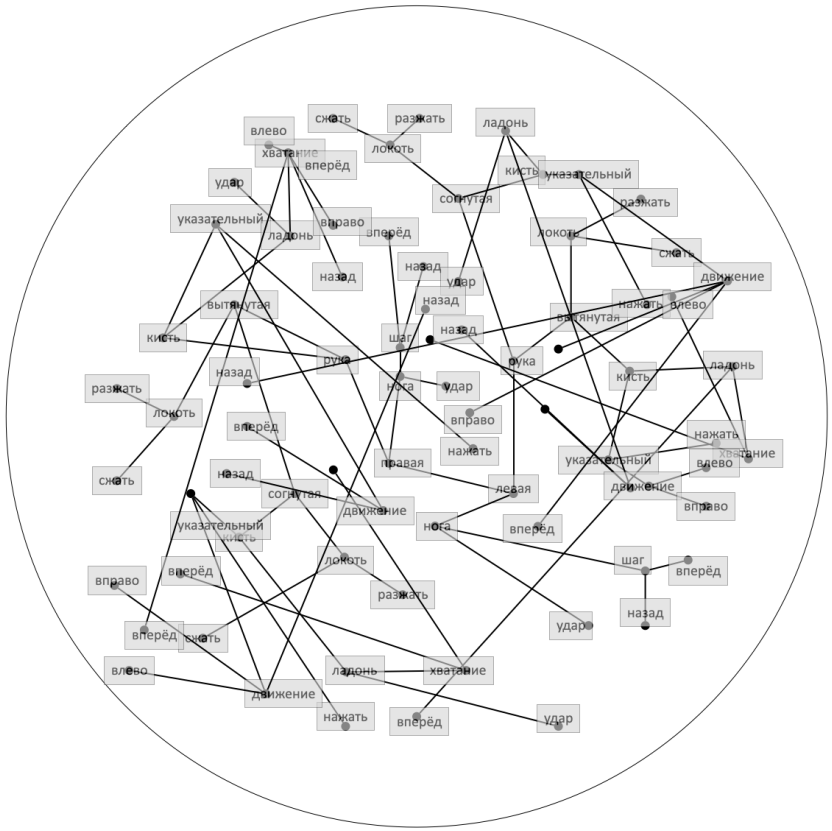


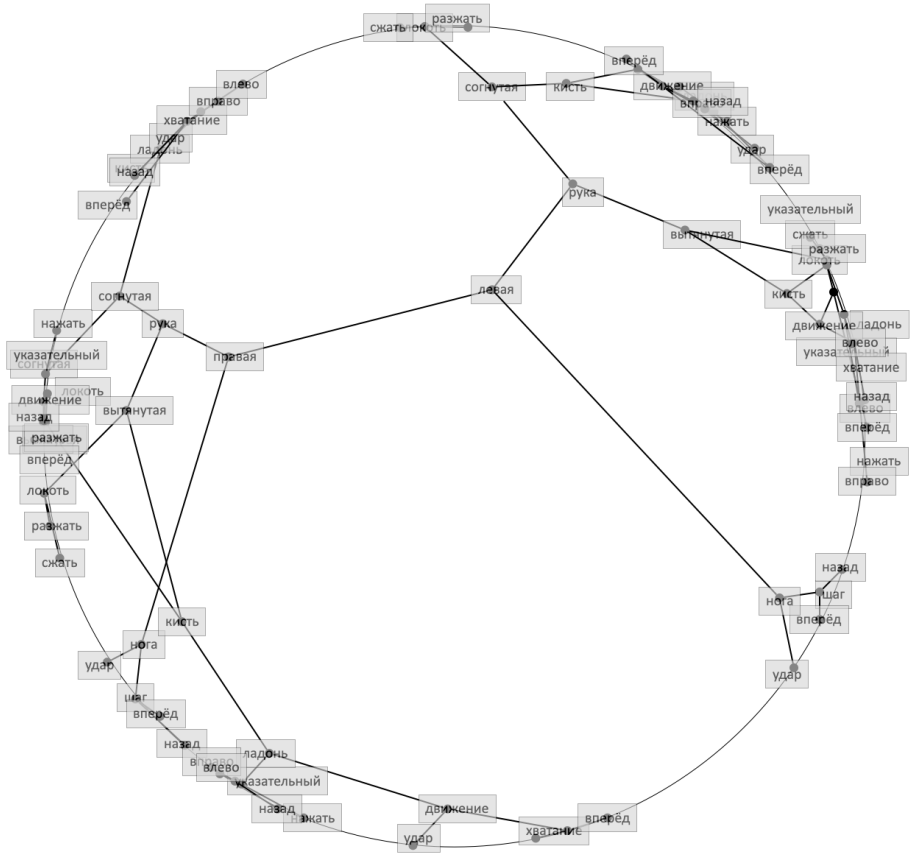
Мы минимизировали функцию потерь:

$$L(E) = \sum_{(u,v)\in S} log \frac{e^{-d(u,v)}}{\sum_{v' \notin S} e^{-d(u,v')}} \tag{3}$$

где S – семантическая близость векторов, E – гиперболическое вложение.
Для задач оптимизации гиперболических вложений разработаны RSGD, RSVRG, RAdaGrad, RAdam и RAMSGrad.

Часть 1. Применение гиперболических вложений в МО





Часть 2.Метод многомерной развертки для алгобро-топологического анализа данных морфо-функциональных структур мозга

Цель: распознавание состояний мозга.

- Подзадача 1: предложить метод классификации состояний мозга с учетом алгобро-топологического представления данных с использованием метода многомерной развёртки и метода цленаправленного проекцирования;
- Подзадача 2: предоставить алгоритм классификации;
- Подзадача 3: продемонстрировать на экспериментальных данных ЭЭГ (алкоголики и трезвинники).

Часть 2.Метод многомерной развертки для МО

Метод многомерной развёртки предложен В.М. Бухштабером в 1994 году

- Времянной ряд $f = (f_1, ..., f_N)$ – это последовательность значений функции $f(t)$, с фиксированным временным интервалом Δt.
- n-мерная развертка $X_f = X_f(N, n)$ временного ряда $f = (f_1, ..., f_N)$ в n-мерном евклидовом пространстве – это кусочно-линейная кривая, полученная последовательным соединением векторов $X_1, X_2, ..., X_p$, где $X_i^T = (f_i, f_{i+1}, ..., f_{i+n-1})$, $p = N - (n-1)$.
- ранг кривой K не превосходит r, если существует такая r-мерная плоскость $L \subset \mathbb{R}^n$, что $X_q \in L$ для всех $q = 1, ..., p$.
- ε-ранг кривой K не превосходит r, если существует такая r-мерная плоскость $L \subset \mathbb{R}^n$, что $\sum_{q=1} \delta(X_q, L)^2 < \varepsilon, \delta(X_q, L) = min_{z \in L} \|X_q - Z\|$,

- Матрицей рассеяния $T(K)$ кривой K бедем называть матрицу рассеяния совокупности её узлов $X_1, ..., X_p$, то есть $T(K) = \sum_{q=1}^{p}(X_q - E)(X_q - E)$, где $E = 1/p \sum_{q=1}^{p} X_q$

Theorem 2 (Бухштабер, 1994)

Ранг кривой K не превосходит $r(rk(K) \leq r)$ тогда и только тогда, когда $\lambda_s = 0$ для исех $s > r$, где $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq 0$ суть собственные значения матрицы рассеяния $T(K)$. Соответсвуенно, ε-ранг кривой K не превосходит $r(rk(K) \leq r)$ тогда и только тогда, когда $\sum_{s=r+1}^{n} \lambda_s \leq \varepsilon$.

Часть 2.Метод многомерной развертки для МО

- n-мерные развертки временных рядов лежат в подпространтствах L, размерность которых не зависит от n, N.
- Эти временные ряды принадлежат классу функций $g(t+\tau) = \sum_{k=1}^{m} \varphi_k(t)\psi_k(\tau)$ для соответсвующих $\varphi_k(t)$, $\psi_k(\tau)$ и параметра m.

Theorem 3 (Бухштабер, 1994)

Пусть $g(t)$ – достаточное число раз дифференцируемая вещественнозначная функция. Тогда абсолютный ранг функции $g(t)$ не превосходит r тогда и только тогда, когда $g(t)$ является решением обыкновенного дифференциального уравнения $\sum_{q=0}^{r} b_q \frac{d^q}{dt^q} g(t) = C$ с постоянными ненулевыми коэффициентами $b_0, ..., b_r$.

Часть 2.Метод многомерной развертки для МО

Идея классификатора на основе многомерной развертки для анализа мозга:

- Пусть n-мерная развертка $\xi_f = \xi_f(N, n)$ принадлежит ε-окрестности r-мерной гиперплоскости $L \subset \mathbb{R}^n$, если $\Sigma_{i=1}^{p}(\|L - X_i\|)^2 < \varepsilon$.
- Классификация данных, предствленных временными рядами, на классы эквивалентности согласно принадлежности многомерной развёртке соответствующего временного ряда $f = (f_1, ..., f_N)$, представленного кусочно-линейной кривой ξ в $\mathbb{R}^n$ с узлами $X_1, ..., X_p$, ε-окрестности r-мерной гиперплоскости $L \subset \mathbb{R}^n$.
- ε-окрестность r-мерной гиперплоскости может соответсвовать доверительному интервалу. Таким образом выходом предложенного алгоритма классификации будет принадлежность временного ряда классу и вероятность отнесения временного ряда данному классу.

Часть 2.Метод многомерной развертки для МО

Algorithm 1 Алгоритм формирования классов эквивалентности для временных рядов в пространстве пониженной размерности методом многомерной развёртки

Require: На вход алгоритма подаётся k временных рядов $\{f_1, ..., f_k\}$, $f_i = (f_{i_1}, ..., f_{i_N})$, сгруппированных в соответсвующие классы эквивалентности $\{K_1, ..., K_m\}, k \leq m$.

(1) Фиксируем размерность развёртки n и проекции r.

(2) Применяем метод многомерной развертки к каждому временному ряду из $\{f_1, ..., f_k\}$. Получаем: k штук n-мерных разверток, сгруппированных в соответсвующие m классов эквиваленитности, $\{X_1, ..., X_m\} = \{\{X_{1_1}, ..., X_{1_t}\}, ..., \{X_{m_1}, ..., X_{m_{t'}}\}\}$.

(3) Применяем метод целенаправленного проецирования на пространство размерности r для каждого класса эквивалентности, объединяя для этого все точки развёрток класса в одно множество. Получаем: m штук r-мерных гиперплоскостей $\{L_1, ..., L_m\}$.

(4) Присваиваем каждой гиперплоскости ε_i таким образом, чтобы все развёртки одного класса лежали в ε_i-окрестности соответствующей гиперплоскости L_i для соответсвующего доверительного интервала $(\sigma, 2 \times \sigma.3 \times \sigma)$. Получаем: $\{\varepsilon_1, ..., \varepsilon_m\}$.

(5) Проверяем лежат ли $\{L_1, ..., L_m\}$ в ε-окрестностях друг друга.

for $L_i, L_i \in \{L_1, ..., L_m\}$ **do**
  for $L_j, L_j \in \{L_1, ..., L_m\}, j \neq i$ **do**
    if (6) $L_j \in \varepsilon_i$-L_i **then**
      (7) $\varepsilon_i \leftarrow \frac{||L_i - L_j||}{2}$. Возврат на (6) ▷ Возможны другие подходы. Например, (а) поточечное удаление из L_i или (б) удаление из L_i соответствующей развёртки и пересчет L_i и ε_i.
    end if
  end for
end forПолучаем: обновленные значения $\{\varepsilon_1, ..., \varepsilon_m\}$.

(8) Проверяем лежат ли соответсвующие многомерные развертки из X_i, $X_i \in \{X_1, ..., X_m\}$ в обновлённой ε_i-окрестности L_i.

for $X_i \in \{X_1, ..., X_m\}$ **do**
  for $X_{i_j} \in X_i, X_i = \{X_{i_1}, ..., X_{i_t}\}$ **do**
    if (9) развёртка $X_{i_j} \in \varepsilon_i$-$L_i$ **then**
      (10) временной ряд $f_{i_j} \in K_i$.
    else if $X_{i_j} \notin \varepsilon_i$-$L_i$ **then**
      (11) $K_i^{'} = K_i \setminus f_{i_j}$.
    end if
  end for
end for(12) Получаем обновленные классы: $\{K_1^{'}, ..., K_m^{'}\}$.

Выход: $\{K_1^{'}, ..., K_m^{'}\}$, $\{L_1, ..., L_m\}$, $\{\varepsilon_1, ..., \varepsilon_m\}$.

Algorithm 2 Алгоритм классификации для временных рядов в пространстве пониженной размерности

Require: На вход подаётся временной ряд $f_i = (f_{i_1}, ..., f_{i_N})$ (или множество временных рядов), гиперплоскости $\{L_1, ..., L_m\}$, $\{\varepsilon_1, ..., \varepsilon_m\}$-значения, $\{K_1^{'}, ..., K_m^{'}\}$, полученные с помощью Алгоритма 1.

Применяем метод многомерной развертки ко временному ряду $f_i = (f_{i_1}, ..., f_{i_N})$. Получаем n-мерную развертку $X = (f_i, f_{i+1}, ..., f_{i+n-1})$, $p = N - (n-1)$.

Проверяем попадание развертки в ε-окресность соотвествующей гиперплоскости из $\{L_1, ..., L_m\}$.

for $L_i \in \{L_1, ..., L_m\}$ **do**

  if $\Sigma_{i=1}^{p}(||L_i - X||)^2 < \varepsilon$ **then**

    $f_i \in K_i^{'}$. Выход из алгоритма со значением i

  end if

end for

if $\Sigma_{i=1}^{p}(||L_i - X||)^2 \geq \varepsilon$ для любого L_i из $\{L_1, ..., L_m\}$ **then**

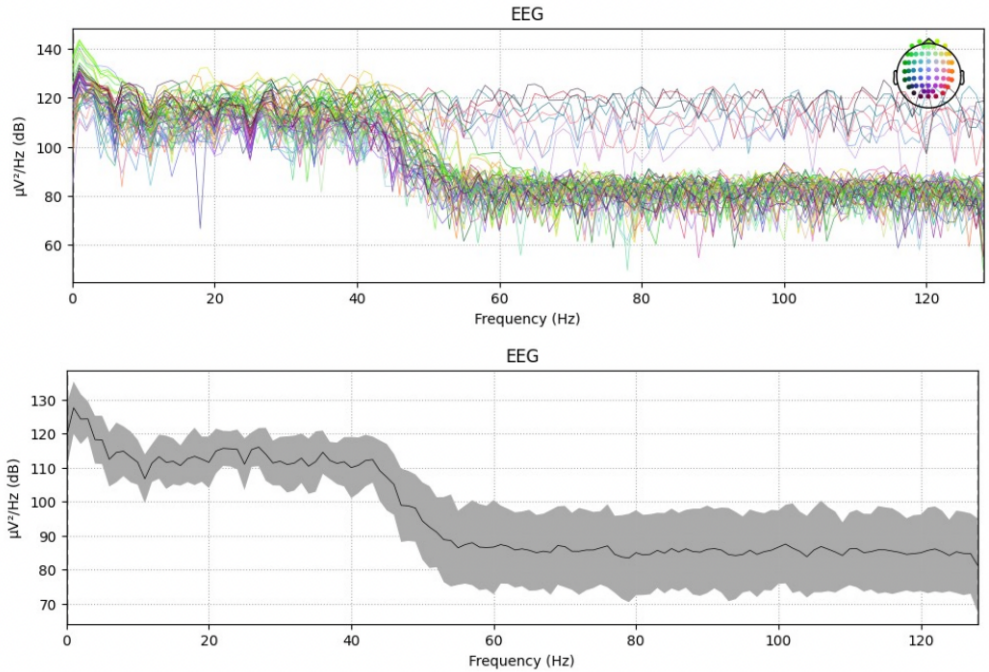
  Выход: i соответсвующее L_i с минимальным расстоянием X и L_i из $\{L_1, ..., L_m\}$.

end if

Часть 2.Метод многомерной развертки для МО

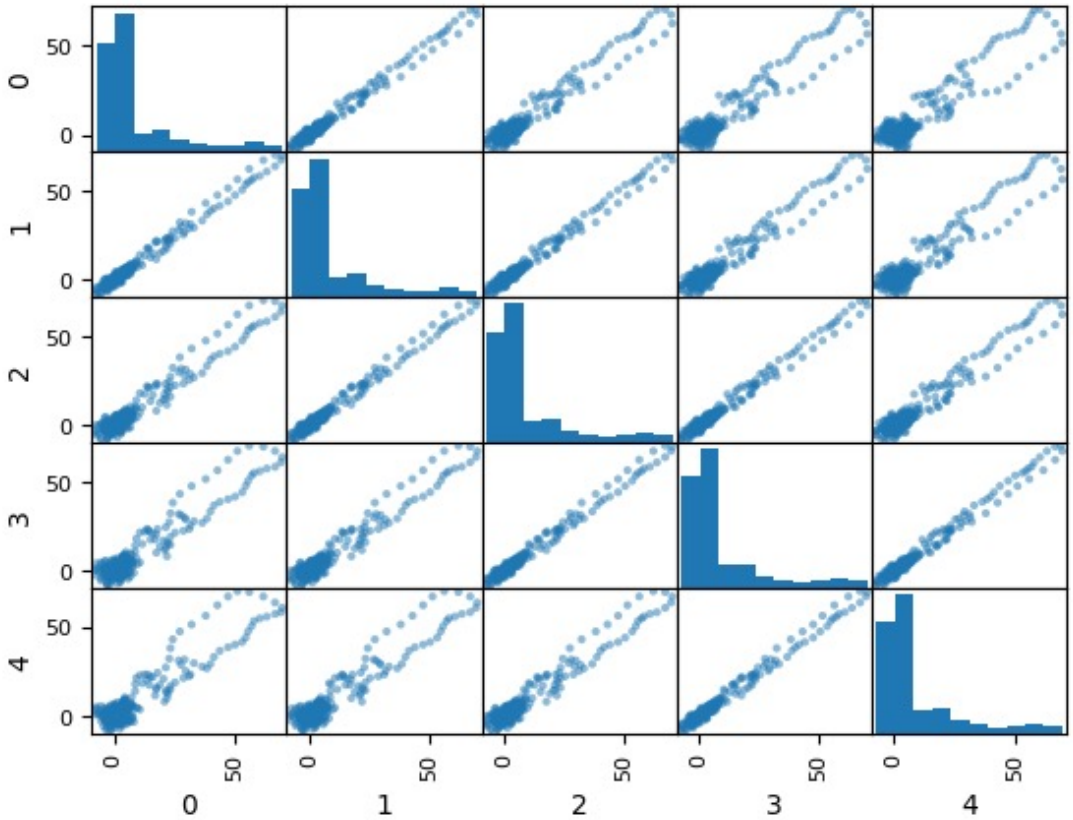
- Данные ЭЭГ записаны на устройстве из 64 электрода. Данные были записаны на 122 испытуемых, разделенных на две группы: с диагностированным хроническим алкоголизмом и группа, не соответствующая критерием алкоголизма. От каждого испытуемого получали по 120 записей длины 1 секунда.
- Каждый испытуемый подвергался воздействию двух стимулов, которые представляли собой изображения объектов, выбранных из набора изображений Снодграсса и Вандерварта.
- Оцифрованные данные представляли собой временные ряды длиной 1 секунда, с частотой дискретизации сигнала в 256 Гц, то есть временной ряд предславлен был 256-ю значениями.

Часть 2.Метод многомерной развертки для МО



Часть 2.Метод многомерной развертки для МО

Построили развертку для каждого канала, $n = 5, 25, 100$, $N = 256$. Посчитали матрицу рассеяния для каждого канала на развертке.



Посчитали собственные значения, полученные на матрице рассеяния для 64-х 5,25 и 100-мерных разверток.

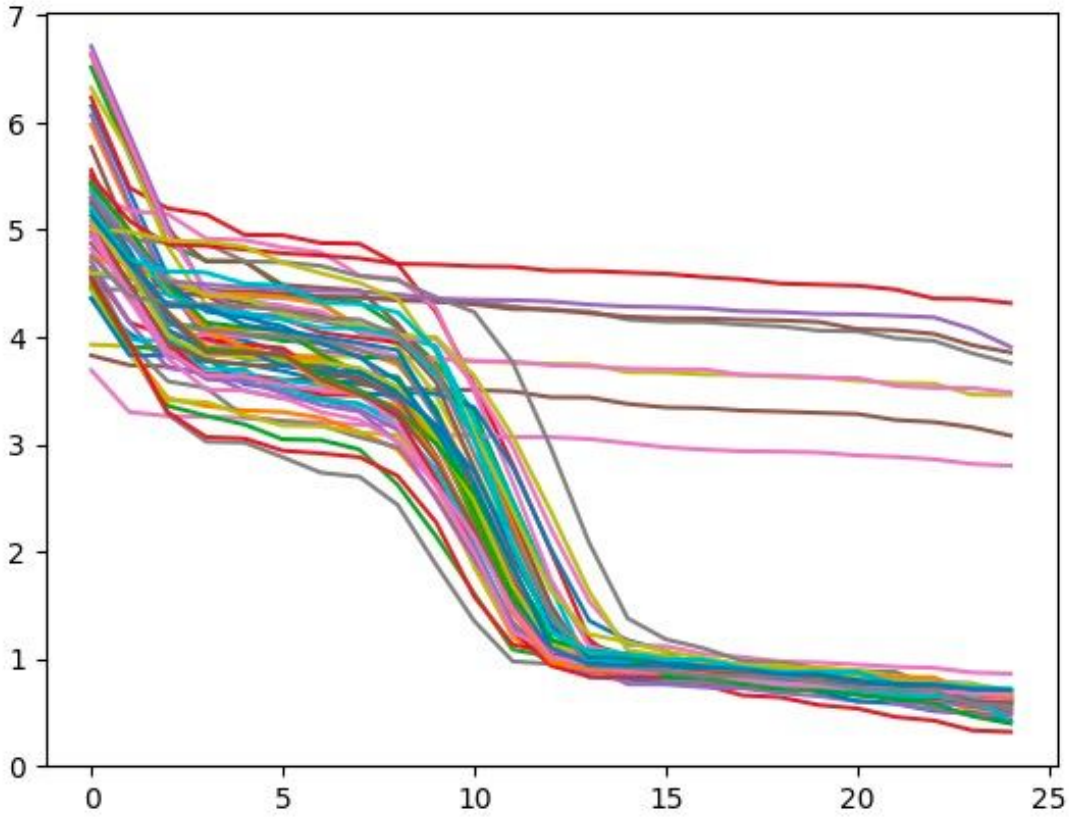
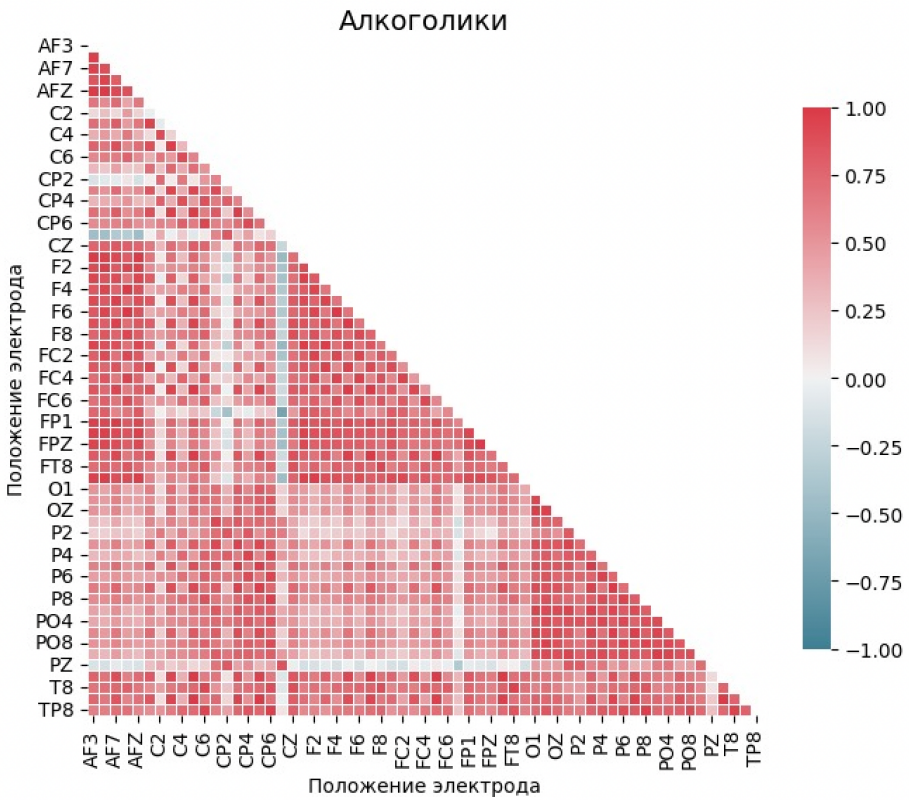
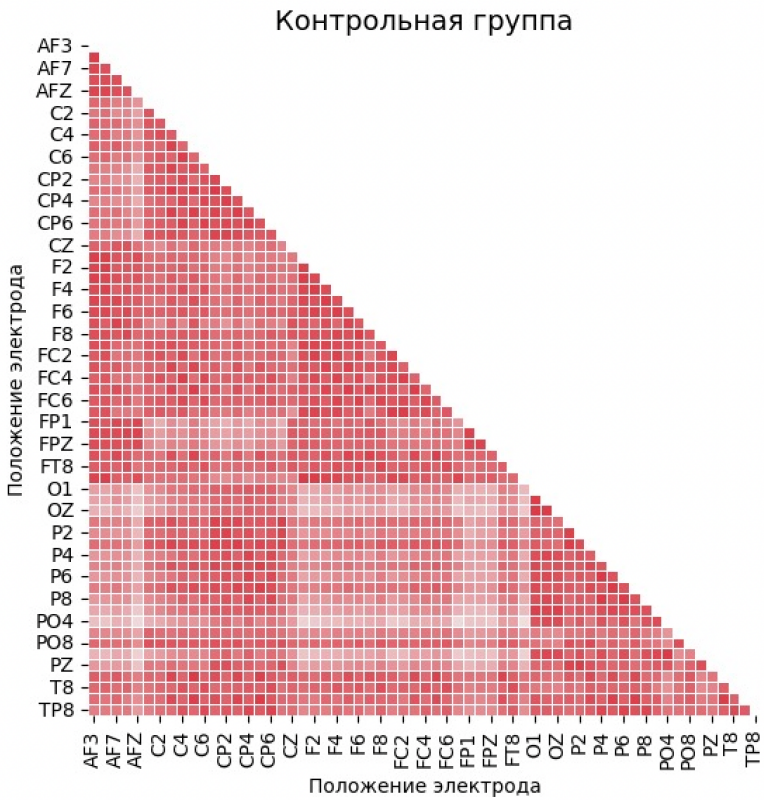


Рис.: Собственные значения, полученные на матрице рассеяния для 64-х 25-мерных разверток.

Часть 2.Метод многомерной развертки для МО

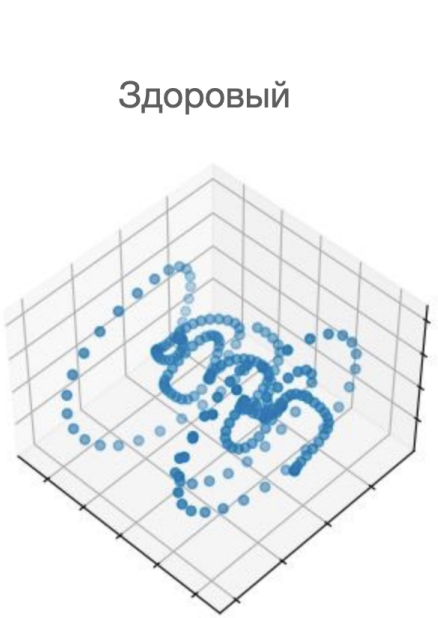
Посчитали корреляцию между каналами с выравниваем.

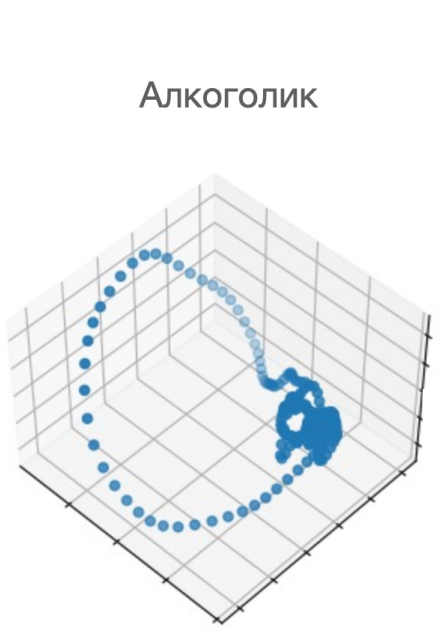


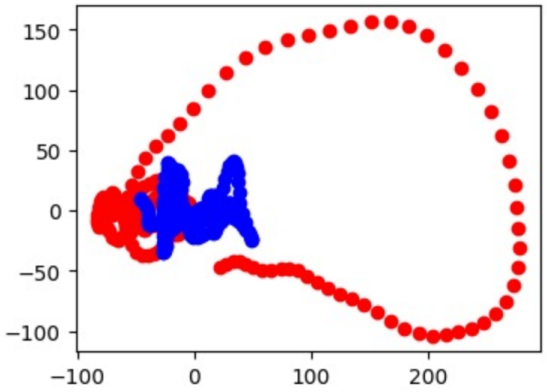


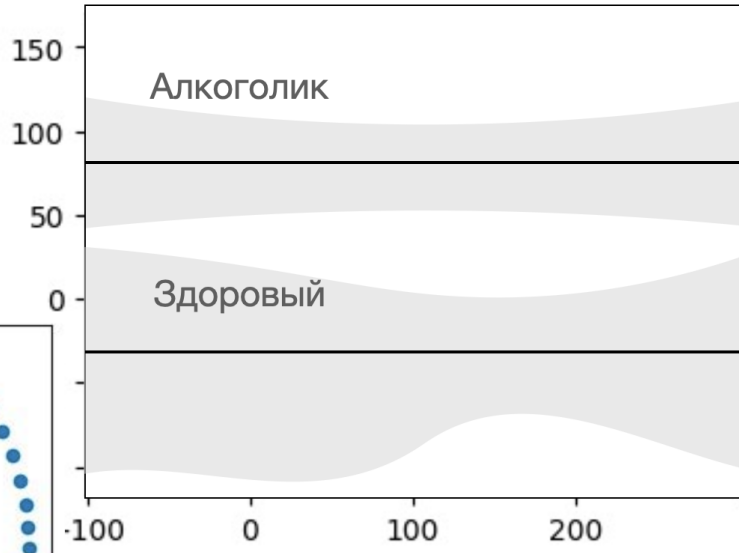
Часть 2.Метод многомерной развертки для МО

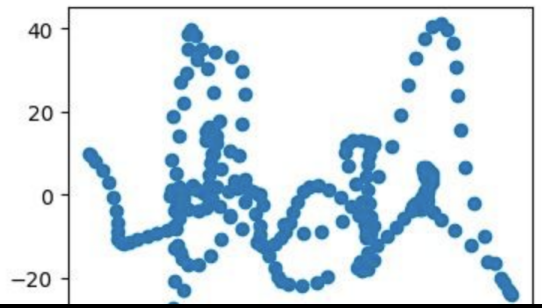
Строили гиперплоскости с доверительным интервалом для двух классов Алгоритмами 1 и 2.

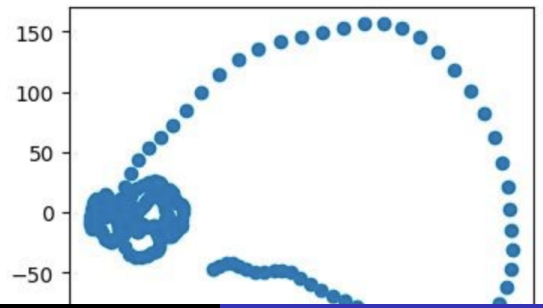












Часть 2.Метод многомерной развертки для МО

Сравнили точность классификации предложенного алгоритма по метрике $ACC = \frac{TP+TN}{TP+FP+TN+FN}$ с работами коллег на данном датасете.

Источник	Классифи-	Метрика (ACC)
Ekaputri et al. [16]	SVM	77.80
Ehlers et al. [8]	Дискриминантный анализ	88.00
Rieg et al. [17]	Random Forest	96.67
Sharma et al. [18]	LS-SVM	97.08
Malar et al. [19]	ELM	87.60
P. Dewi Purnamasari et al. [20]	BPNN	90.00
Faust et al. [28]	HOS	92.00
Kannathal et al. [29]	Дискриминантный анализ	90.00
Acharya et al. [9]	SVM	91.70
L. Farsi et al. [30]	PCA + ANN	71.00
	LSTM 1	91.00
	LSTM 2	93.00
M. Zubair et al. [31]	ANN	97.37
	KNN	98.22
	XGBoost	98.97
Предложенный метод	Алгоритм 1 и 2	100

Спасибо за внимание!