



Л е к ц и я № 12

Снижение размерности

Разработал профессор П. М. Васильев
Кафедра фармакологии и биоинформатики

Для студентов, обучающихся по направлению 06.03.01 «Биология»
профили Биохимия, Генетика
при изучении дисциплины «Цифровые технологии в биологии»

П л а н л е к ц и и

- Методы снижения размерности
- Выбор признаков
- Свертка
- Многомерное шкалирование
- Метод главных компонент

Снижение размерности

Dimensionality Reduction

Преобразование данных, состоящее в уменьшении числа переменных.

Наличие в модели избыточных, неинформативных или слабо информативных признаков снижает эффективность модели.

Методы снижения размерности

- Выбор признаков
- Свертка
- Многомерное шкалирование
- Метод главных компонент

Выбор признаков

Feature Selection

Эти методы оставляют некоторое подмножество исходного набора признаков, избавляясь от признаков избыточных и слабо информативных.

Методы выбора признаков

➤ Фильтры, *Filter methods*

Коэффициент ранговой корреляции Спирмена

$$R_{Sp} = 1 - \frac{\sum_{i=1}^N (\text{rank}(x_i) - \text{rank}(y_i))^2}{N(N^2 - 1)}$$

➤ Правила обрезки

По числу признаков, по порогу

Методы выбора признаков

➤ Оберточные методы, *Wrapper methods*

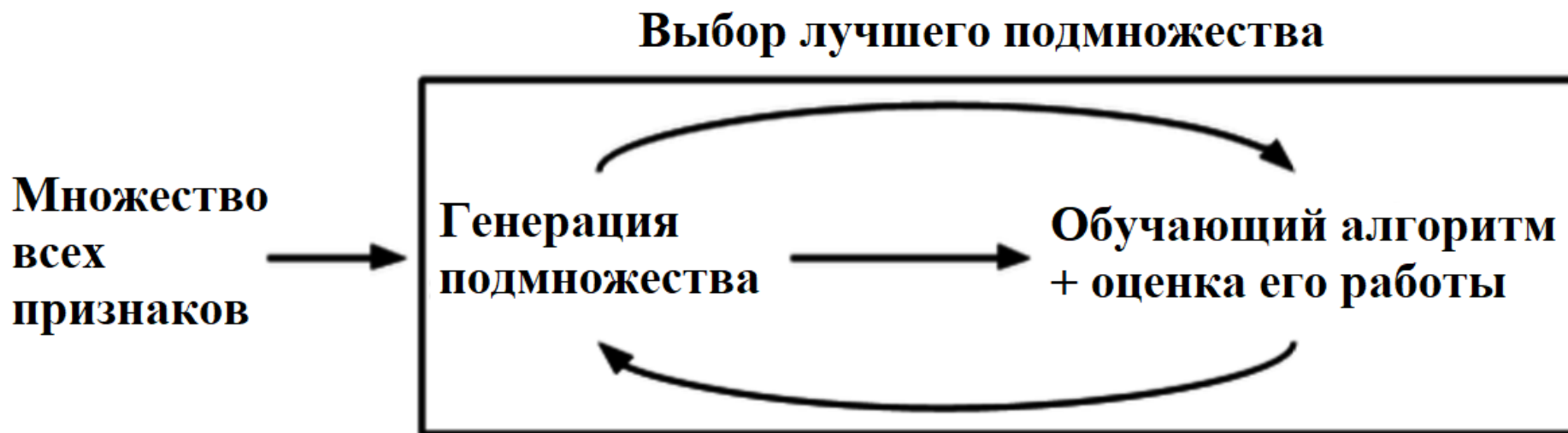
Последовательный прямой выбор (*SFS, Sequential Forward Selection*)



Методы выбора признаков

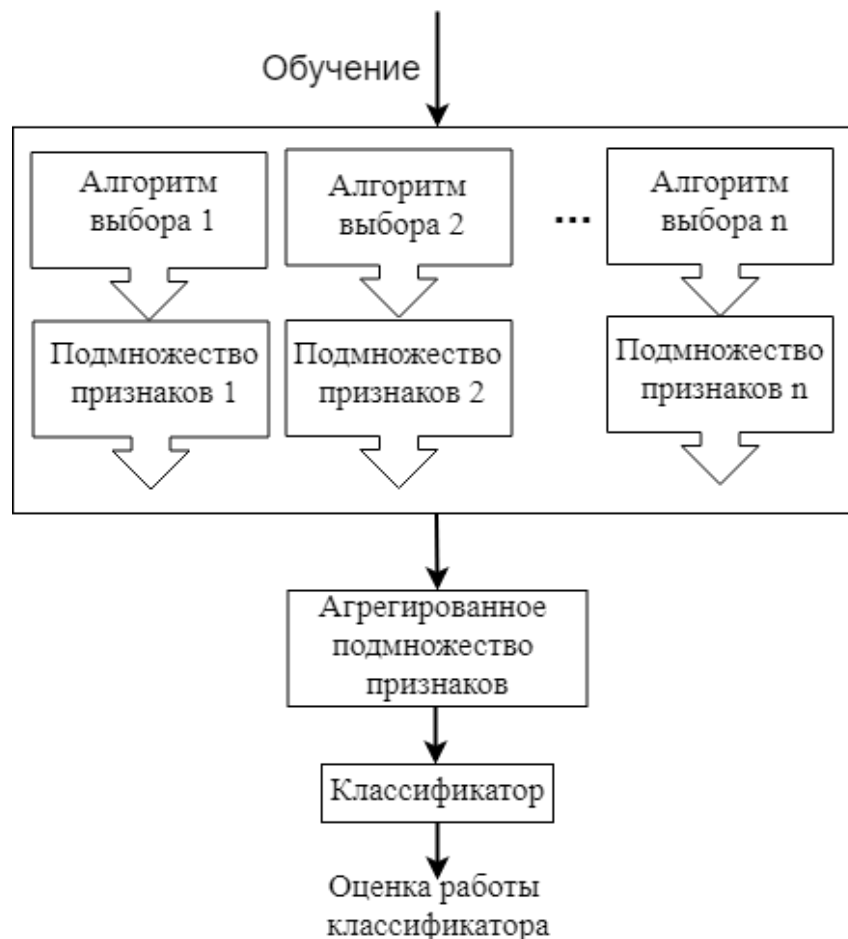
- Встроенные методы, *Embedded methods*

Случайный лес



Методы выбора признаков

➤ Гибридные и ансамблевые методы



Свертка

Convolution

Операция над парой матриц **A** (размера $n_x \times n_y$) и **B** (размера $m_x \times m_y$), результатом которой является матрица **C=A*B** размера $(n_x - m_x + 1) \times (n_y - m_y + 1)$.

Каждый элемент результата вычисляется как скалярное произведение матрицы B и некоторой подматрицы A такого же размера (подматрица определяется положением элемента в результате).

Ядра свертки изображений

Операция	Ядро
Тождественное отображение	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
Выделение границ	$\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$
Резкость	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$
Размытие по Гауссу (3 x 3)	$\frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$
Тиснение	$\begin{bmatrix} -2 & -1 & 0 \\ -1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix}$

Пример свертки изображения

Входное изображение

12	14	41
43	84	24
2	1	43



Матрица

0,5	0,75	0,5
0,75	1,0	0,75
0,5	0,75	0,5

—
—

Результат

—
—

$$\begin{pmatrix} 12 * 0,5 + 14 * 0,75 + 41 * 0,5 + \\ 43 * 0,75 + 84 * 1,0 + 24 * 0,75 + \\ 2 * 0,5 + 1 * 0,75 + 43 * 0,5 \end{pmatrix}$$



$$\frac{1}{\text{div}}$$

—
—

32,41667

div = 6

Результаты свертки изображения

Размытие



Результаты свертки изображения

Повышение четкости



Результаты свертки изображения

Нахождение границ



Пример свертки данных

Скользящее среднее

Медианная свертка

[illegible]

Многомерное шкалирование

Multidimensional Scaling

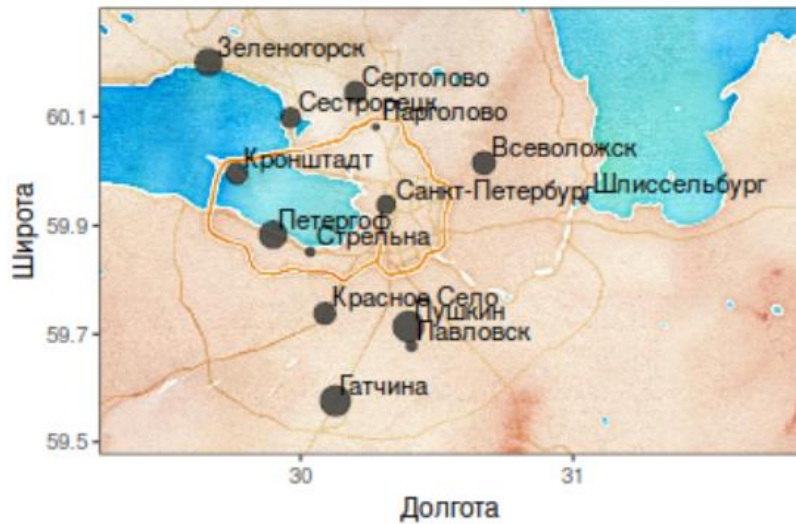
Методы многомерного шкалирования размещают объекты в пространстве заданной размерности и проверяют, насколько точно полученная конфигурация сохраняет расстояния между объектами.

Пример многомерного шкалирования

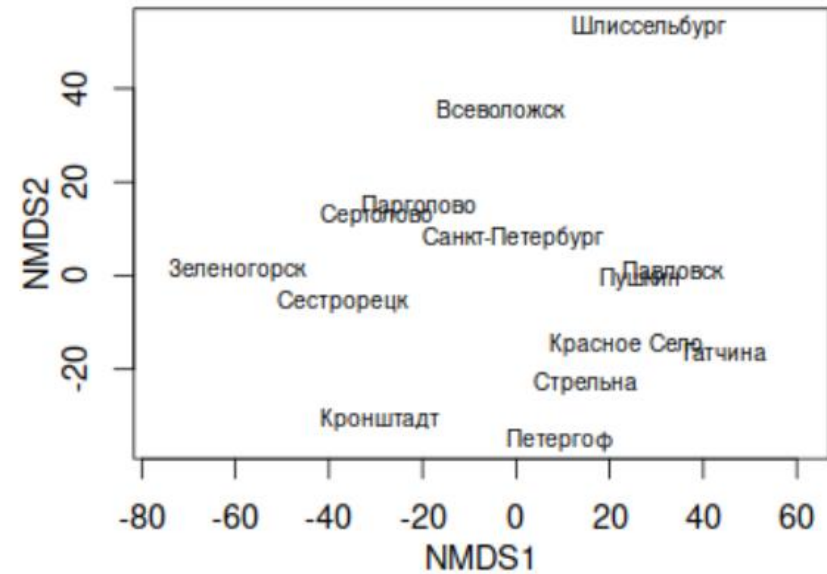
Расстояния по автодорогам					
name	Всеволожск	Гатчина	Зеленогорск	Красное Село	Кронштадт
Всеволожск	0	71	71	64	70
Гатчина	71	0	97	20	76
Зеленогорск	71	97	0	81	52
Красное Село	64	20	81	0	56
Кронштадт	70	76	52	56	0
Павловск	57	30	87	24	75
Парголово	31	70	44	53	42
Петергоф	73	47	77	28	37
Пушкин	52	29	82	21	70
Санкт-Петербург	24	47	53	31	51

Пример многомерного шкалирования

Карта

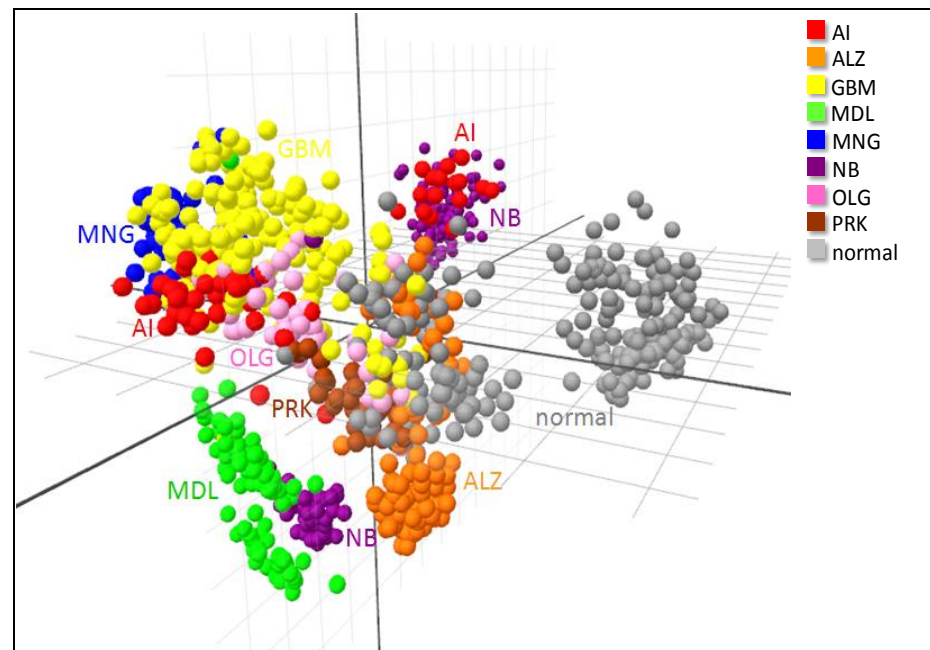
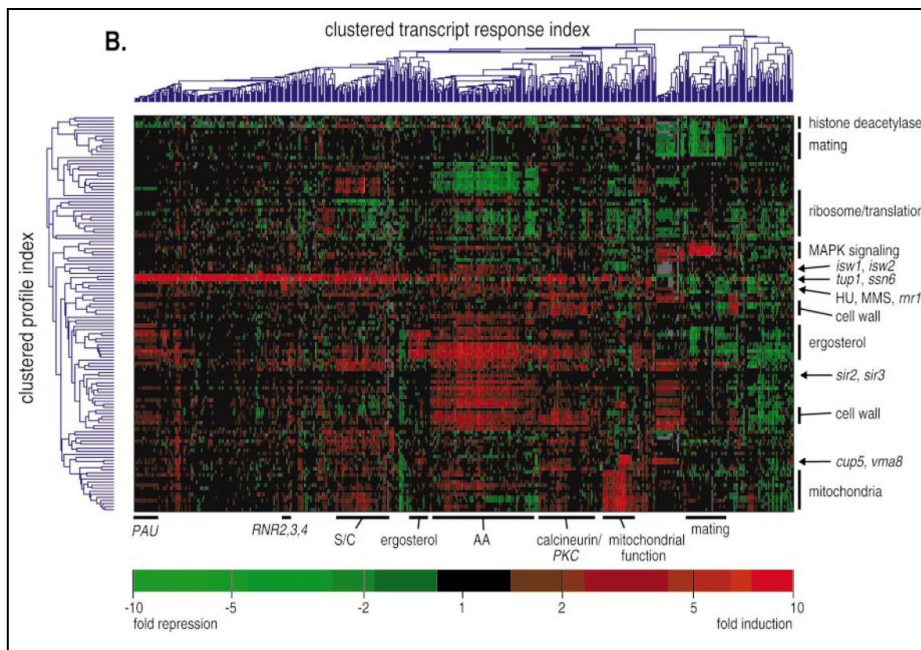


Ординация nMDS



Пример многомерного шкалирования

Анализ транскриптом, измененных в результате заболеваний мозга



Данные анализа на микрочипах парной коэкпрессии генов

Метод главных компонент

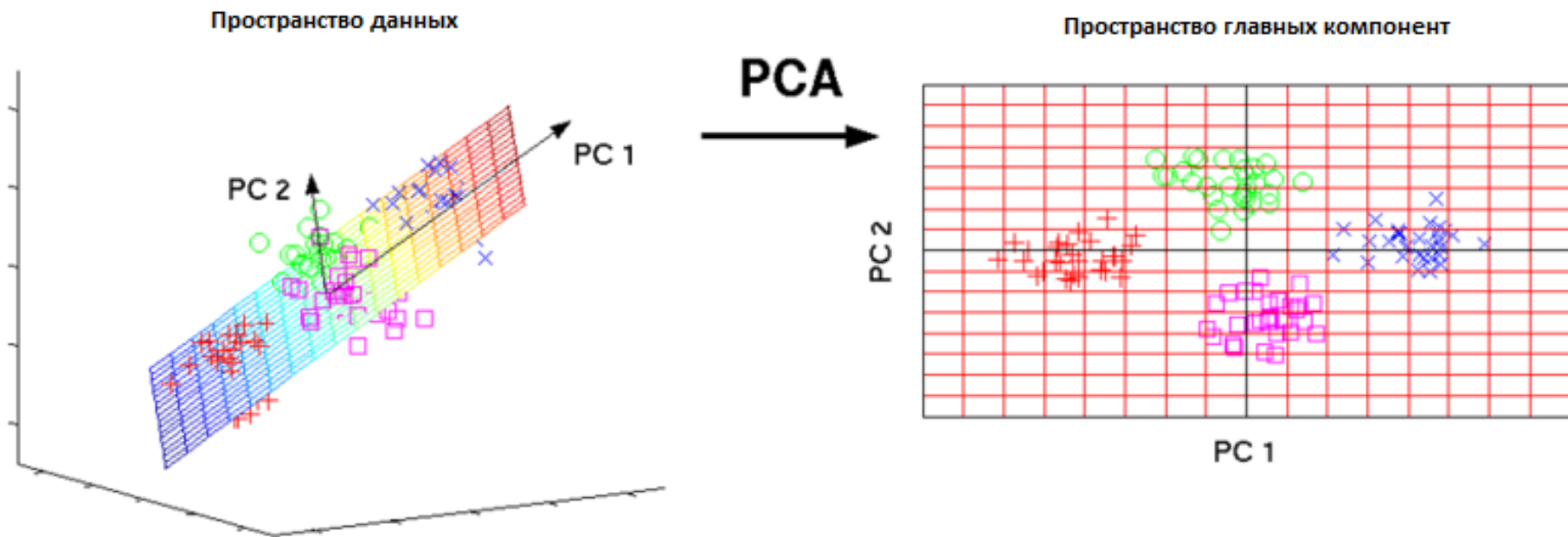
Principal Component Analysis, PCA

Поиск такой гиперплоскости, на которую при ортогональной проекции всех признаков максимизируется дисперсия.

Данное преобразование может быть произведено с помощью сингулярного разложения матриц и создает проекцию только на линейные многомерные плоскости.

Основан на вычислении собственных векторов и собственных значений ковариационной матрицы данных.

Метод главных компонент



To be continued ...

