Статистический анализ экспрессии генов: систематизация и классификация

Зиннатулина Белла Раифовна

Санкт-Петербургский Государственный Университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель — к.ф.-м.н. доцент Н.П. Алексеева
Рецензент — д.ф.-м.н. проф. М.С. Ермаков

Санкт-Петербург
2015г.

Структура данных

- Данные — экспрессии генов стволовых клеток (Северо-Западный федеральный медицинский исследовательский центр).
- Выборка ($N = 42$) больных сердечной недостаточностью ($\mathbf{HF}$).
- Признаки ($p = 36$) — экспрессии генов, характеризующие $\mathbf{HF}$.

Проблема: Безошибочная классификация данных по наличию диабета и ожирения по полному набору признаков линейным дискриминантным анализом. Возможно наличие неинформативных признаков и переобучение.

Цель: Выявить экспрессии генов, влияющих на $\mathbf{HF}$ опосредованно через факторы диабета ($\mathbf{D}$) и ожирения ($\mathbf{O}$).

Задачи

- Проверка действия классификаторов
  - Логистическая регрессия
  - Случайный лес
  - k ближайших соседей
  - Линейный дискриминантный анализ
- Отбор информативных признаков методами
  - Дискриминантный анализ с разрежением
  - Специальный метод
- Систематизация признаков при помощи кластерного анализа с учетом влияния на факторы $\mathbf{D}$ и $\mathbf{O}$.

Постановка задачи классификации

- $X \subset \mathbb{R}^p$ — множество наблюдений (индивидов).
- $Y = \{0, 1\}$ — метки классов в случае бинарной классификации.
- $(x_1, y_1), \ldots, (x_n, y_n)$, $y_i \in Y$, $x_i \in X$ — обучающая выборка.
- Целевая функция $f : X \to Y$, $f(x_i) = y_i$, $i = \overline{1 : n}$.
- Задача: аппроксимация f на все пространство X.

Метод: логистическая регрессия

- $f(z) = \dfrac{1}{1 + \mathrm{e}^{-z}}$ — логистическая (сигмоидная) функция.
- $\mathbb{P}(y = 1 \mid x) = f(\theta^{\mathrm{T}} x) = f(\theta_1 x_1 + ... + \theta_p x_p)$, $x \in X$,

  $\mathbb{P}(y = 0 \mid x) = 1 - f(\theta_1 x_1 + ... + \theta x_p)$.
- $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ — вектор значений вещественных параметров. Подбирается по методу максимального правдоподобия.

Применение метода в задаче классификации: Класс $y = 1$ или $y = 0$ объекта x определяется как $\mathbb{P}(y = 1 \mid x) > 0.5$ или $\mathbb{P}(y = 1 \mid x) \leq 0.5$ соответственно.

Метод: случайный лес (RF)

Случайный лес основан на построении ансамбля деревьев принятия решений. Параметр метода — количество деревьев N.

Дерево принятия решений

- $X \subset \mathbb{R}^p$.
- Узлы помечены предикатом $\Pi_i : X \to \{\text{True}, \text{False}\}, i = 0, 1, \ldots,$ который является критерием разбиения данных.
- Разбиение данных происходит в каждом узле по одному признаку.
- Листья — $y_i \in Y = \{0, 1\}$.

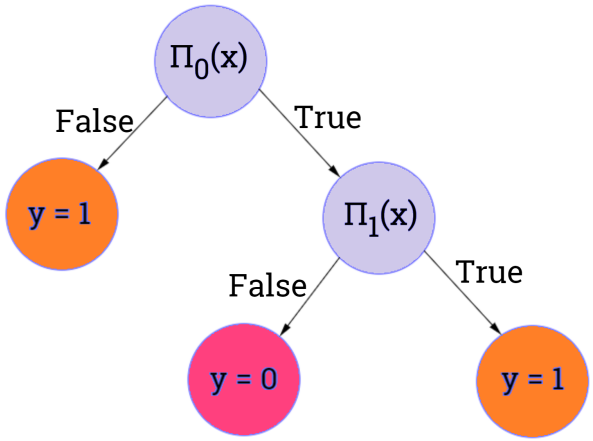

Рис.: Абстрактное дерево принятий решений.

Строится N деревьев принятия решений по случайным подвыборкам X_i, $i = \overline{1:N}$. Принадлежность элемента классу определяется путем голосования деревьев.

Метод: k ближайших соседей (KNN)

- Параметр метода — k.
- Метод основан на оценивании сходства объектов.
- $\rho : X \times X \to \mathbb{N}$ — функция расстояния, евклидова метрика .
- $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, $x_i \in X$, $y_i \in Y$ — обучающая выборка.
- $\widetilde{x}$ — элемент тестовой выборки:
  $\rho(\widetilde{x}, x^{[1]}) \leq \rho(\widetilde{x}, x^{[2]}) \leq \cdots \leq \rho(\widetilde{x}, x^{[n]})$.
- Выбор класса $\widetilde{x}$ определяется $\widetilde{y} = \left\lceil (y^{[1]} + .. + y^{[k]})/k - 0.5 \right\rceil$.

Метод: дискриминантный анализ с разрежением (sparseLDA)

- $X \subset \mathbb{R}^p$ — множество наблюдений, принадлежащих K классам.
- $H = \{H_{ij} = \mathbb{1}_{\{x_i \to y_j\}},\ i \in \overline{1:n},\ j \in \overline{1:k}\}$.
- Ω — положительно определенная матрица.
- θ_k — корректирующий вектор весов.
- λ и γ — неотрицательные настраиваемые параметры.
- $\beta_k \in \mathbb{R}^p$ — k-й дискриминантный вектор.
- (θ_k, β_k) является k-м решением задачи

$$\min_{\beta_k, \theta_k} \{\| H\theta_k - X\beta_k \|^2 + \gamma\beta_k^{\mathrm{T}}\Omega\beta_k + \lambda \| \beta_k \|_1\},$$

$$\frac{1}{n}\theta_k^{\mathrm{T}} H^{\mathrm{T}} H\theta_k = 1, \quad \theta_k^{\mathrm{T}} H^{\mathrm{T}} H\theta_l = 0, \quad \forall l < k.$$

- Классификация применением стандартного линейного дискриминантного анализа к матрице $(X\beta_1 ... X\beta_q)$, $q < K$.

Результаты классификации

Таблица: Доля правильной классификации выборки

Метод	Тренировочная выборка		Тестовая выборка	
	Диабет	Ожирение	Диабет	Ожирение
Random Forest	0.7	0.72	0.6	0.66
KNN	0.78	0.76	0.75	0.75
sparseLDA	0.83	0.75	0.8	0.7
Logistic Regression	1	1	1	1
LDA	1	1	1	1

Выделение информативных признаков: этап 1

- Матрица наблюдений $X(n,p) = [X_1 \vdots \ldots \vdots X_p]$, где n — число индивидов, p — число признаков. $Y = \{\mathbf{O}, \mathbf{D}\}$.
- Усеченная матрица наблюдений $X_\tau(n,\theta) = [X_{\tau_1} \vdots \ldots \vdots X_{\tau_\theta}]$, где $\tau = (\tau_1, \ldots, \tau_\theta) \subset (1, 2, \ldots, p) = \mathcal{N}_p$.
- Множество $\Theta(p,\theta)$, $|\Theta(p,\theta)| = C_p^\theta$ всех θ-подмножеств из $\mathcal{N}_p$.
- Классификатор $f^Y : X_\tau \to [0,1]$.
- Информативные θ-наборы по классификаторам f_r^Y, $r \in \overline{1:m}$ с уровнем вероятности P

$$\Sigma_r = \Sigma_r(\theta, p, f_r^Y, P, Y) = \{i | i \in \Theta(p,\theta), f_r^Y(X_\tau) \geq P\}.$$

Выделение информативных признаков: этап2

- Значения линейных дискриминантных функций
  $DF_j(X_\tau, Y) = \alpha_0 + X_\tau \alpha^{\mathrm{T}}, \ \tau \in \bigcup_{r=1}^{m} \Sigma_r, \ j = \overline{1:n}.$
- Персональные наборы $\widetilde{\tau}_j = \arg\max_{\tau} |DF_j(X_\tau)|$.
- Финальный θ-набор — мода $(\widetilde{\tau}_1, \ldots, \widetilde{\tau}_n)$.
- Рейтинги в P-значимой f_r^Y-классификации
  $Z_{ir}^Y = |\{\tau \in \Sigma_r | i \in \tau\}|, i \in \overline{1:p}, r \in \overline{1:m}.$

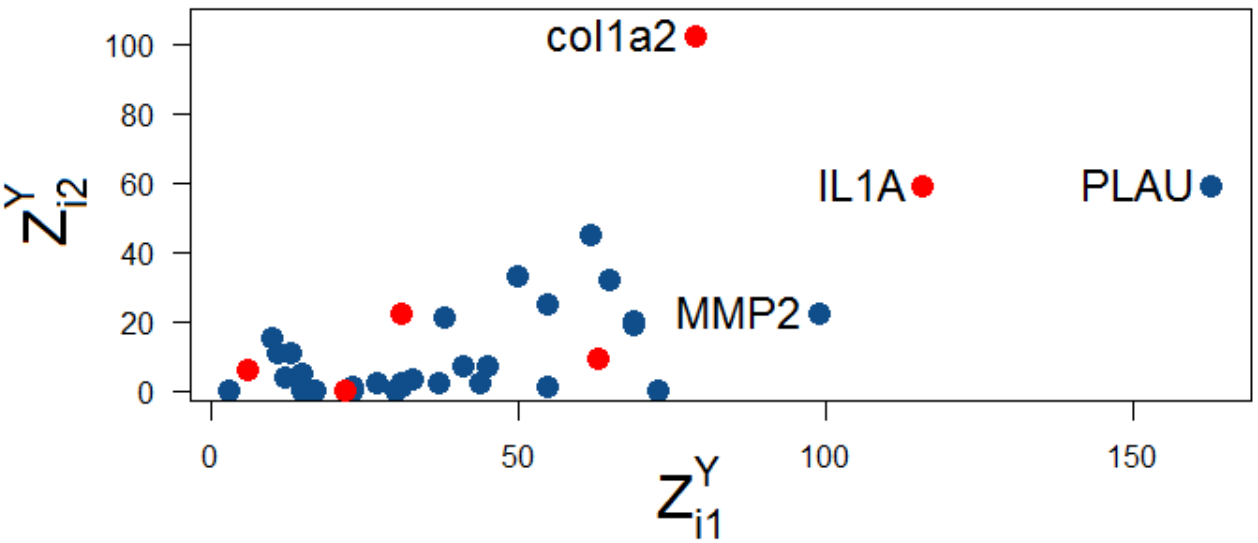

Рис.: Рейтинги признаков $f_1^Y = KNN$, $f_2^Y = RF$, $\theta = 4$, $Y = \mathbf{O}$.

Результаты: двумерная классификация

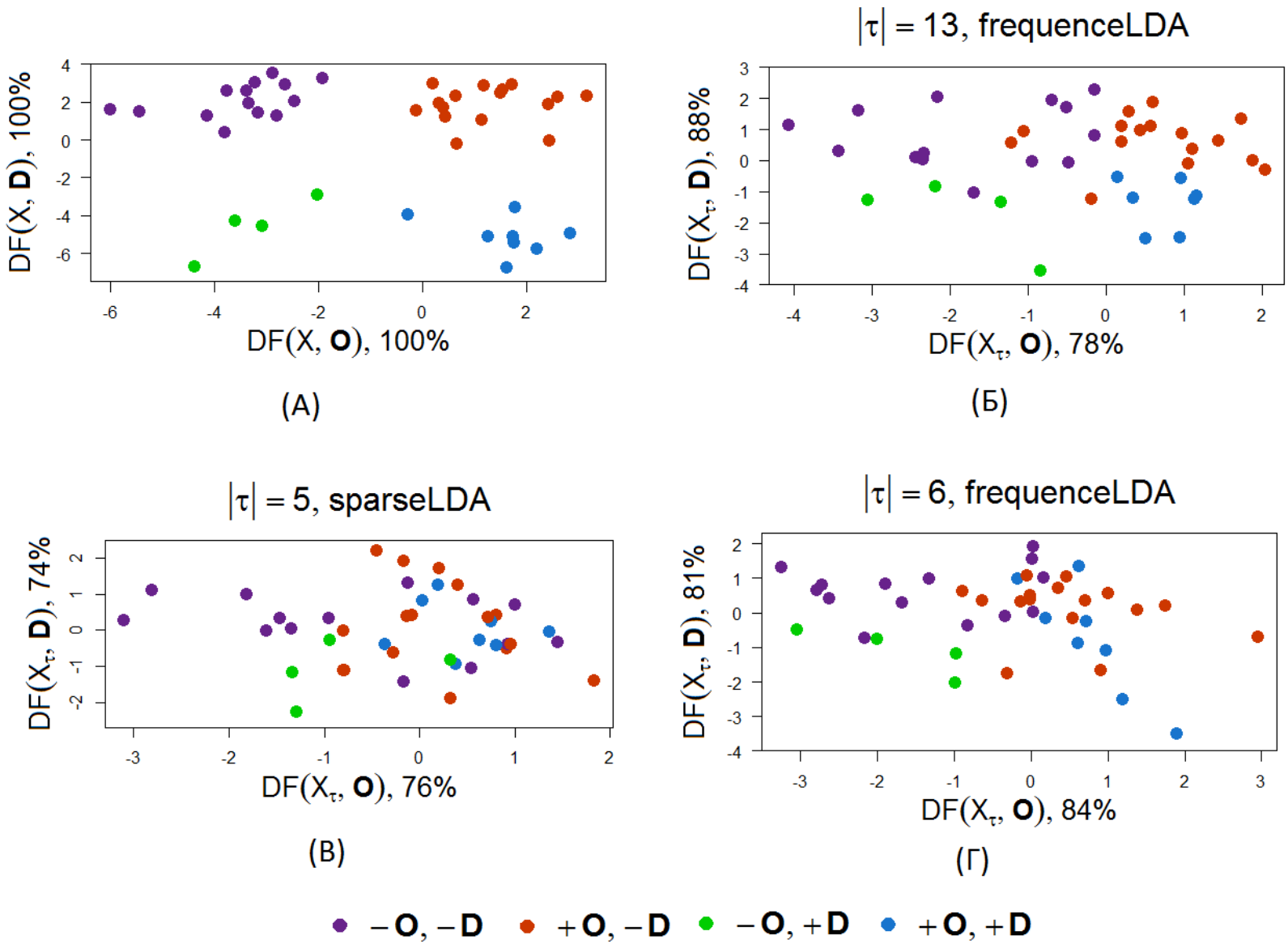

Рис.: Визуализация данных по значениям дискриминантных функций.

Систематизация. Кластерный анализ признаков

- Метрика — информационное разнообразие.
- Стратегия — информационный выигрыш от объединения групп.
- Объект кластеризации — рейтинги N_{jr}^{Y} в порядковой шкале признаков X_τ классификатора f_r и разделителя Y,
  $j \in \overline{1:n}$, $Y = \{\mathbf{O}, \mathbf{D}\}$, $r \in \overline{1:m}$, $\tau \in \bigcup_{r=1}^{m} \Sigma_r$.

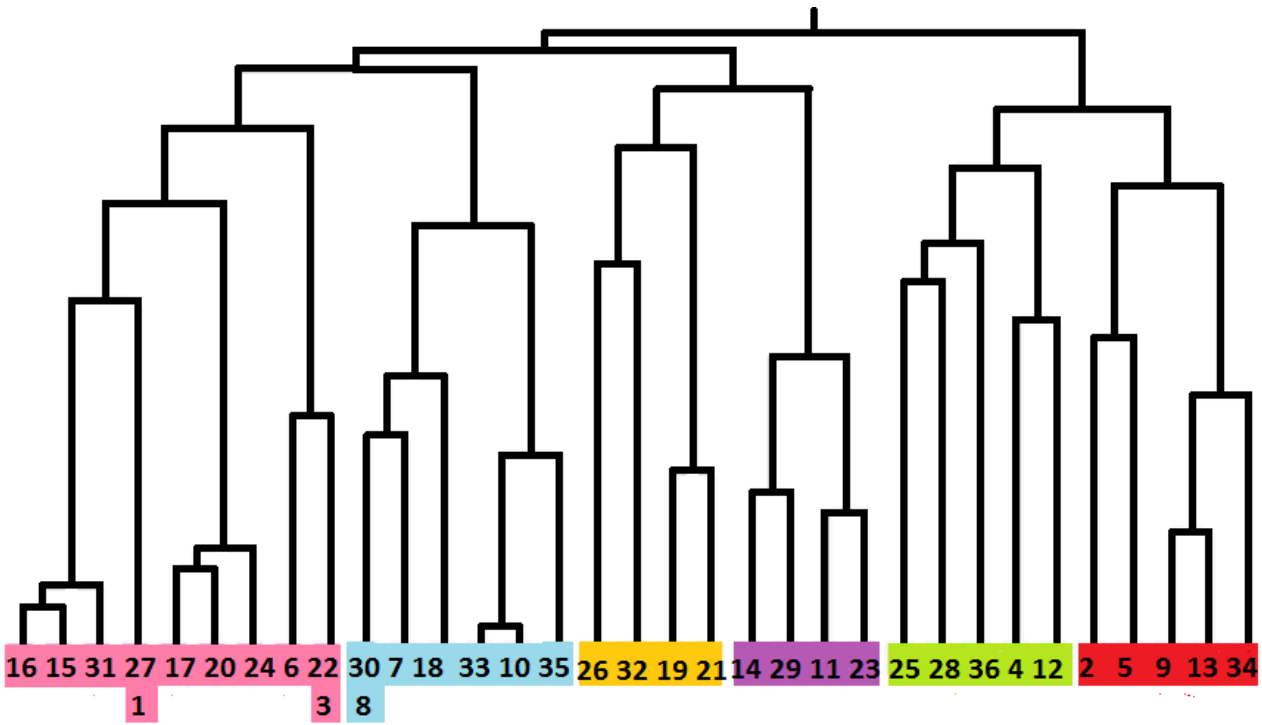

Рис.: Дерево кластеризации признаков.

Основные результаты

Результаты:

- Алгоритмы классификации и отбора признаков реализованы на языке R.
- Выделены группы признаков, дающих наилучшую классификацию по наличию диабета и ожирения с наименьшими потерями точности.Удалось сократить пространство признаков с 36 до 6.
- Признаки разделены на 6 групп по схожести роли, которую они играют в классификации.

Планы:

- Исследовать структуру данных относительно разделителей и классификаторов.