Статистические методы улучшения классификации в задаче прогнозирования послеоперационных кардиологических осложнений

Комлева Дарья Михайловна, гр. 522

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., доц. Алексеева Н.П.
Рецензент: к.ф.-м.н., доц. Коробейников А.И.

Проблема анализа кардиологических данных

- Снижение риска возникновения осложнений после операции АКШ (Аорто-коронарное шунтирование)
- 112 индивидов в раннем послеоперационном периоде
- 11 количественных и 43 категориальные признака, характеризующие предоперационный и интраоперационный период
- Механизм возникновения СМВ трудно предсказуем с клинической точки зрения

Цель — Классификация индивидов и прогнозирование послеоперационного осложнения СМВ (Синдром Малого сердечного Выброса)

Байесовская процедура классификации в случае двух популяций

$x = (x_1, \ldots, x_p)$ — реализация случайного вектора признаков,
q_1, q_2 — априорные вероятности популяций W_1, W_2,
μ_1, μ_2 — вектора средних,
$\hat{\Sigma} = \hat{\Sigma}_1 = \hat{\Sigma}_2$ — ковариационные матрицы,
$\beta = \hat{\Sigma}^{-1}(\mu_1 - \mu_2)$ — коэффициенты дискриминантной функции $f(x) = x^T\beta$.
Правило классификации: $x \in W_1$, если

$$\frac{\beta^T\mu_1 - \beta^T\mu_2}{2} + \ln(\frac{q_1}{q_2}) \leqslant x^T\beta.$$

Ограничения метода:

- $\hat{\Sigma} = \hat{\Sigma}_1 = \hat{\Sigma}_2$
- Проблема включения в анализ категориальных признаков
- Сложность интерпретации дискриминантных функций в случае большого количества признаков

Другие методы классификации

Методы	Основные преимущества
Пошаговый дискриминантный анализ	Редукция размерности
Регуляризованный дискриминантный анализ (RDA) [Фридман, 1989]	$p >> n$
Дискриминантный анализ с разрежением (SDA) [L.Clemmensen, 2011]	Редукция размерности, $p >> n$
Метод опорных векторов (SVM) [Вапник, 1963]	Нелинейность
Случайный лес (Random Forest) [Breiman, 2001]	Рейтинговый подход, случайное дерево
Стратификационный дискриминантный анализ [Алексеева, 2012]	Рейтинговый подход, разделяемые подвыборки

Методы можно использовать как изолированно, так и в сочетании.

Алгоритм стратификационного дискриминантного анализа в случае популяций W_0, W_1

- Переменные: классифицирующие $X = (X_1, \ldots, X_p)$, расслаивающие бинарные $Y = (Y_1, \ldots, Y_k)$. Итоговая характеристика $Z = -1$, если $X \in W_0$, $Z = 1$, если $X \in W_1$.
- N_c — пороговое значение объемов подвыборок, P_c — граница правильной классификации, c — граничное значение апостериорной вероятности или дискриминантной функции
- Расслоение популяций на (W_0^{i0}, W_1^{i0}) при $Y_i = 0$ и (W_0^{i1}, W_1^{i1}) при $Y_i = 1$, $i = 1, \ldots, k$. Объемы выборок n_0^{il}, n_1^{il}.
- Применение $LDA(SLDA, RDA, SDA)$ с переменными $X = (X_1, \ldots, X_p)$ для (W_0^{il}, W_1^{il}), $l = 0, 1$, $i = 1, \ldots, k$ при условии $n_0^{il} \geq N_c, n_1^{il} \geq N_c$.

Выходные данные алгоритма $LDAwS$

- Вероятность правильной классификации $P_{il}(X)$. Апостериорная вероятность или дискриминантная функция $d_{il}(X)$.
- Стратифицирующее множество подвыборок $\mathcal{L}(X) = \{(i,l)|\ \ n_0^{il} \geq N_c, n_1^{il} \geq N_c, P_{il}(X) > P_c\}$
- Средняя дискриминантная функция $d_m(X) = \sum\limits_{(i,l)\in\mathcal{L}} d_{il}(X)$
- Множество правильно классифицирующих подвыборок $\mathcal{L}_*(X) = \{(i,l)|\ \ (i,l) \in \mathcal{L}(X), \mathrm{sgn}(d_{il} - c)Z = 1\}$
- Индекс классификации

$$I(X) = \frac{\mathrm{card}(\mathcal{L}_*(X))}{\mathrm{card}(\mathcal{L}(X))}.$$

Расширение множества дихотомических признаков

Конечно-линейная стратификация.

Определение

Пусть $X = (X_1, \ldots, X_m)^T$ вектор дихотомических признаков с компонентами, принимающими значения 0 и 1, $\tau = (t_1, \ldots, t_k) \in (1, 2, \ldots, m)$. Симптом k ранга — линейная комбинация вида $X_\tau = A_\tau X(\mod 2)$, где $A_\tau = (a_1, \ldots, a_m)$ — вектор-строка с компонентами

$$a_j = \begin{cases} 1, & j \in \tau \\ 0, & j \notin \tau. \end{cases}$$

В задаче прогнозирования СМВ достаточно использование симптомов ранга $k = 2, 3$ по $m = 43$ бинарным признакам.

Дискриминантный анализ с разрежением (SDA)

Пусть X — матрица наблюдений,
Y — $n \times K$ матрица фиктивных переменных для K классов,
θ_k — K-вектор корректирующих коэффициентов для классов,
λ, γ — неотрицательные параметры,
Ω — положительно определенная матрица.

Параметры (θ_k, β_k) являются решением задачи:

$$\min_{\beta_k, \theta_k} \left\{ ||Y\theta_k - X\beta_k||^2 + \gamma {\beta_k}^T \Omega \beta_k + \lambda ||\beta_k||_1 \right\},$$

$$\frac{1}{n} {\theta_k}^T Y^T Y \theta_k = 1,$$

$${\theta_k}^T Y^T Y \theta_l = 0 \ \forall l < k.$$

Соотношение между коэффициентами SDA и LDA в случае двух классов

Утверждение (1)

1. *Если* $X^T X = I$, *то коэффициенты* $\hat{\beta}_{SDA}$ *имеют вид:*

$$\hat{\beta}_{SDA} = \left(|(I + \gamma\Omega)X^T Y\theta| - \lambda/2\right)_+ \operatorname{sgn}\left((I + \gamma\Omega)X^T Y\theta\right),$$

где $z_+ = z$, *если* $z > 0$, *иначе* $z_+ = 0$.
В случае центрированных данных $\hat{\beta}_{LDA}$ *и* $\hat{\beta}_{SDA}$ *соотносятся как:*

$$\hat{\beta}_{SDA} = \frac{n\sqrt{q_1 q_2}}{(n-2)}\left(\hat{\beta}_{LDA} + \frac{n-2}{n}\hat{\Sigma}_b^{-1}(\hat{\mu_2} - \hat{\mu_1}) - \lambda/2\right)_+ \operatorname{sgn}\left(\hat{\beta}_{LDA} + \frac{n-2}{n}\hat{\Sigma}_b^{-1}(\hat{\mu_2} - \hat{\mu_1})\right).$$

2. *Если* $X^T X \neq I$ *и* $\lambda = 0$, *то* $\hat{\beta}_{SDA}$ *и* $\hat{\beta}_{LDA}$ *соотносятся как:*

$$\hat{\beta}_{SDA} = \frac{n\sqrt{q_1 q_2}}{(n-2)}\left(\hat{\beta}_{LDA} + \frac{n-2}{n}\hat{\Sigma}_b^{-1}(\hat{\mu_2} - \hat{\mu_1})\right)$$

Регуляризованный дискриминантный анализ

Во-первых с помощью параметра регуляризации α, где $0 < \alpha < 1$, вычисляется комбинация:

$$\hat{\Sigma}_i(\alpha) = (1 - \alpha)\hat{\Sigma}_i + \alpha\hat{\Sigma}.$$

Затем, используя параметр регуляризации γ, где $0 < \gamma < 1$, строится следующая оценка:

$$\hat{\Sigma}_i(\alpha, \gamma) = (1 - \gamma)\hat{\Sigma}_i(\alpha) + \gamma\frac{1}{d}\operatorname{tr}[\hat{\Sigma}_i(\alpha)]I,$$

где $\frac{1}{d}\operatorname{tr}[\hat{\Sigma}_i(\alpha)]$ — среднее значение диагональных элементов матрицы $\hat{\Sigma}_i(\alpha)$,
$\hat{\Sigma}_i$ — ковариационные матрицы для каждого класса,
$\hat{\Sigma}$ — общая ковариационная матрица.

Сравнение методов на данных по Синдрому Малого Выброса (CMB)

Таблица 1: Доли правильной классификации

Метод	Обучающая выборка	Контрольная выборка	Вся выборка
LDA	0.611	0.636	0.616
RDA	0.64	0.545	0.625
SDA	0.644	0.5	0.616
LDA Greedy Wilks	0.622	0.682	0.553
SVM	1	0.636	0.928
Random Forest	1	0.545	0.911
LDAwS	0.9	0.91	0.955
RDAwS	0.9	0.86	0.973
SDAwS	0.9	0.91	0.955

LDAwS, RDAwS, SDAwS на всех данных

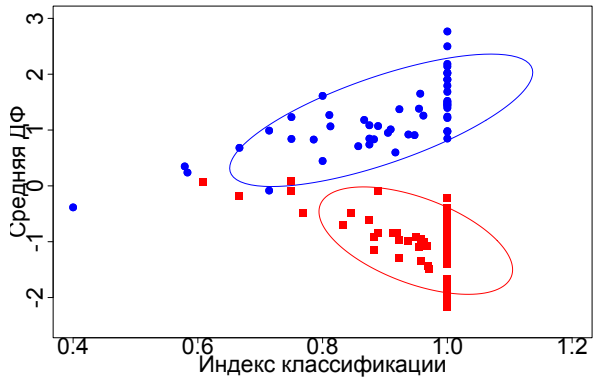

Рис. 1: Результаты LDAwS

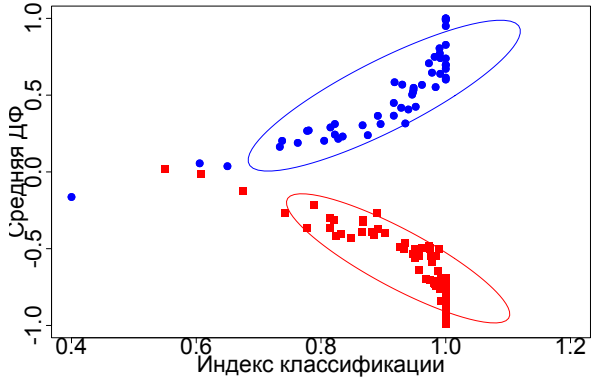

Рис. 2: Результаты RDAwS

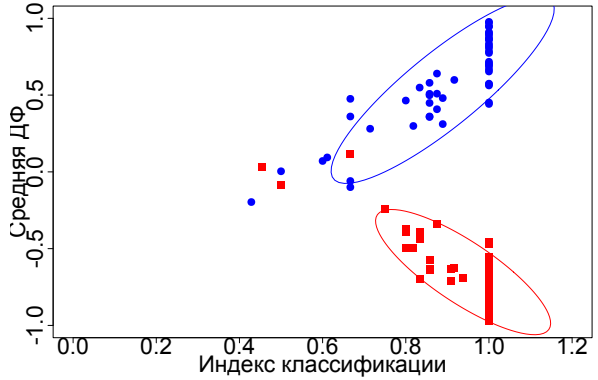

Рис. 3: Результаты SDAwS

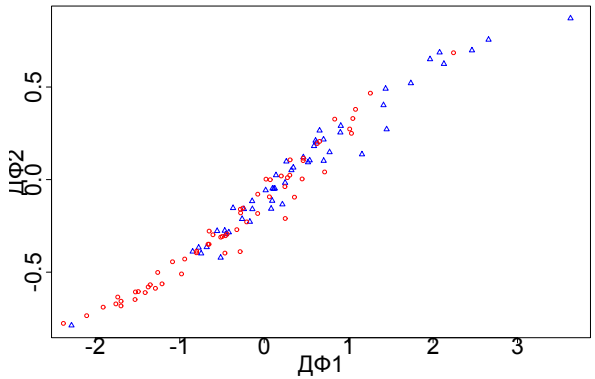

Рис. 4: Без стратификации

LDAwS, RDAwS, SDAwS с обучением

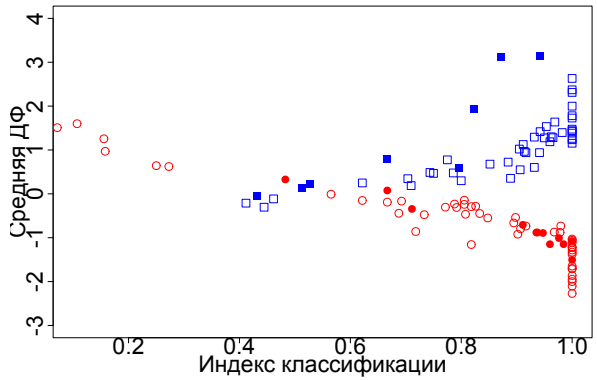

Рис. 5: Результаты LDAwS

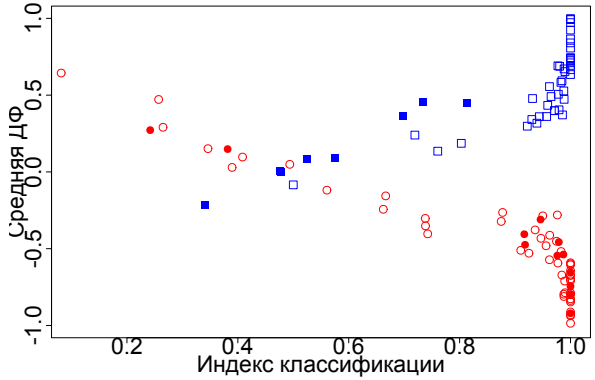

Рис. 6: Результаты RDAwS

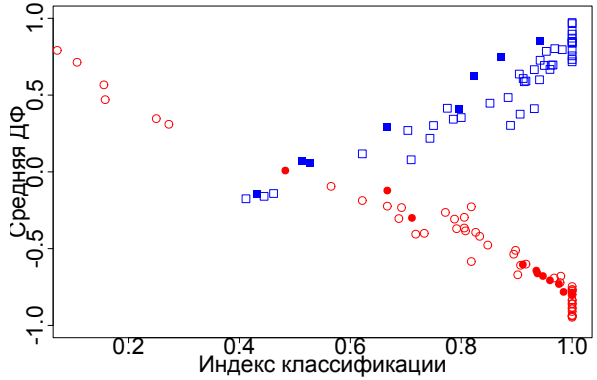

Рис. 7: Результаты SDAwS

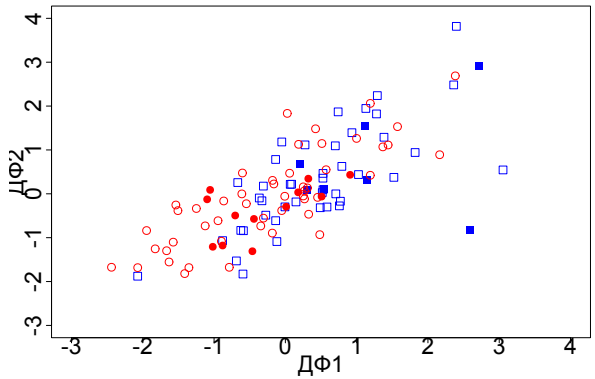

Рис. 8: Без стратификации

Основные результаты

- Алгоритмы LDAwS, RDAwS, SDAwS реализованы в виде программного кода на языке R.
- Для прогнозирования CMB использовались разнообразные методы классификации с разбиением выборки на обучающую и контрольную.
- Наилучший результат показывают методы RDAwS и SDAwS, 86-92% правильной классификации.

- Стратификационный дискриминантный анализ позволяет отделить людей с наиболее вероятным возникновением послеоперационного осложнения с небольшой перестраховкой.
- Для больных без осложнений характерен более высокий индекс правильной классификации.