Сравнительный анализ разных методов классификации с приложением в кардиологии

Мунхтогоо Норжинсурэн, гр. 21.M03-мм

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

Научный руководитель: к. ф.-м. н., Алексеева Н.П.
Рецензент: к. ф.-м. н., Ананьевская П.В.

Санкт-Петербург
2023г.

Особенность медико-биологических данных

- Невозможно достичь полноты данных
- Число признаков больше числа индивидов
- Потеря информативности при удалении пропусков
- Заполнение пропусков проблематично при совместной работе с экспериментатором

Приложение в кардиологии: посткардиотомный синдром (ПКТС). Данные содержат 55 признаков, 458 индивидов:

- x1 . . . x16 – 16 количественные,
- y1. . . y37 – 37 категориальные.

Постановка задачи

Цель – решение проблемы с пропусками применяя метод частичной классификации и улучшения точности прогноза с использованием метода классификации по расслоению.

Задачи:

- Получить оценку полного прогноза по частичным для задачи классификации и применить для прогнозирования ПКТС.
- На основе симптомно-синдромального анализа сделать классификацию по расслоению и построить алгоритмическое дерево.

Обычные методы классификации машинного обучения

Пропорция класса 0 и 1 классифицирующего фактора **0.4**.
Прологарифмировали признаки: x1, x2, x7, x8, x10, x11, x14, x15, x16.

Методы	Вся выборка	Обучающая выборка	Тестовая выборка	CV
LDA	0.63	0.603	0.614	0.59
QDA	0.68	0.69	0.554	0.53
Лог.регрессия	0.64	0.615	0.626	0.64
Случайный лес	1	1	0.662	-

Таблица: Точность классификации разных методов

При использовании обычных методов машинного обучения потеряются **133 индивидов из 458 (30% данных)**.

Основные понятия метода частичной классификации

Задача дискриминантного анализа состоит в построении дискриминантной функции $z = \alpha^T X$ по независимым переменным $X_1, \ldots, X_n$ в случае двух популяций P_1 и P_2.

- $\alpha = (\alpha_1, \ldots, \alpha_p)^T$ – вектор коэффициентов.
- μ_1, μ_2 – вектора средних в двух популяции.
- $\Sigma = \{\sigma_{ij}\}_{i,j=1}^n$ – ковариационная матрица с дисперсией σ^2. Будем считать, что $EX_i = 0$, $i = 1, \ldots, n$.

Наилучший классификатор (Фишер, 1936): $z = (\mu_1 - \mu_2)^T \Sigma^{-1} X$

$$f_0 = F(X) = \sum_{j=1}^{n} \alpha_j X_j, \quad \text{где } \alpha_j = \frac{1}{|\Sigma|} \sum_{i=1}^{n} (\mu_1^{(i)} - \mu_2^{(i)}) |\Sigma_{ji}|, \qquad (1)$$

$|\Sigma|$ – определитель матрицы Σ, $|\Sigma_{ij}|$ – алгебраические дополнения по i-строке и j-столбцу матрицы Σ.

Построение частичных классификаторов

Определение 1. Полный классификатор – наилучшая классификация по всем независимым переменным $X_1 \dots X_n$.

Определение 2. Частичный классификатор – наилучшая классификация по какому-нибудь подмножеству независимых переменных $X_\tau = (X_{t_1}, \dots, X_{t_p})$, где $\tau = (t_1, \dots, t_p) \subseteq (1, 2, \dots, n)$.

Предложение 1.

Пусть Z – вектор с компонентами $Z_i = \frac{X_i}{\sqrt{\sigma_{ii}}}$ и μ – вектор с компонентами $\mu_i = \frac{\mu_1^{(i)} - \mu_2^{(i)}}{\sqrt{\sigma_{ii}}}$, $i = 1, 2, \dots, n$ и $\Sigma = EXX^T$.

1. Наилучший классификатор *(1)* имеет вид $F(X) = (\mu_1 - \mu_2)^T \Sigma^{-1} X = \mu^T \Lambda^{-1} Z$.
2. $F(AX) = F(X)$, где A квадратная матрица полного ранга n.

Строятся частичные классификаторы $f = (f_1, \dots f_p)$, $f_i = F(X_{\tau_i})$:

$$F(X_\tau) = (\mu_1 - \mu_2)^T \Sigma_p^{-1} X_\tau$$

Математическое ожидание определителя матрицы Λ_n (Алексеева Н. П.)

Λ_n – корреляционная матрица частичных классификаторов $f_1, \ldots, f_n$.
Модель корреляционной матрицы классификаторов:

$$\Lambda_n = \begin{bmatrix} 1 & r + x_{12} & \ldots & r + x_{1n} \\ r + x_{21} & 1 & \ldots & r + x_{2n} \\ \vdots & \vdots & \ldots & \vdots \\ r + x_{n1} & r + x_{n2} & \ldots & 1 \end{bmatrix} \tag{2}$$

где $\mathrm{E}x_{ij} = 0$, $\mathrm{D}x_{ij} = \sigma_0^2$, $x_{ij} = x_{ji}$, $0 < r < 1$.

$$\mathrm{E}\Lambda_n = rnJ_{n-1}(\sigma_0, r) + J_n(\sigma_0, r) \quad \text{и} \quad \mathrm{E}\Lambda_{n,kj} = -rJ_{n-2}(\sigma_0, r), \ k \neq j. \tag{3}$$

где
$$J_n(\sigma_0, r) = \sum_{k=0}^{[n/2]} C_n^{2k}(-1)^k \psi_k \sigma_0^{2k}(1-r)^{n-2k}, \quad n = 1, 2, 3, \ldots$$

$\psi_k = 1 \times 3 \times 5 \times \ldots \times (2k-1) = \frac{(2k)!}{2^k k!}$ – нечетный факториал.

Оценка полного прогноза по частичным

Утверждение 1.

Наилучший классификатор имеет вид

$$F(\mathrm{f}) = \mu^T \Lambda_n^{-1} Z = \sum_{j=1}^{n} \gamma_j z_j, \quad \text{где} \quad \gamma_j = \frac{1}{|\Lambda_n|} \sum_{i=1}^{n} \mu_i |\Lambda_{n,ij}| \tag{4}$$

компоненты вектора $\gamma = \mu^T \Lambda^{-1}$, $\delta_j = 1 - \frac{\bar{\mu}}{\mu_j}$, $f_j = \mu_j z_j$. Тогда для оценки полного прогноза $f_b(r, n, \sigma_0, \mu)$ по частичным классификаторам справедливо выражение

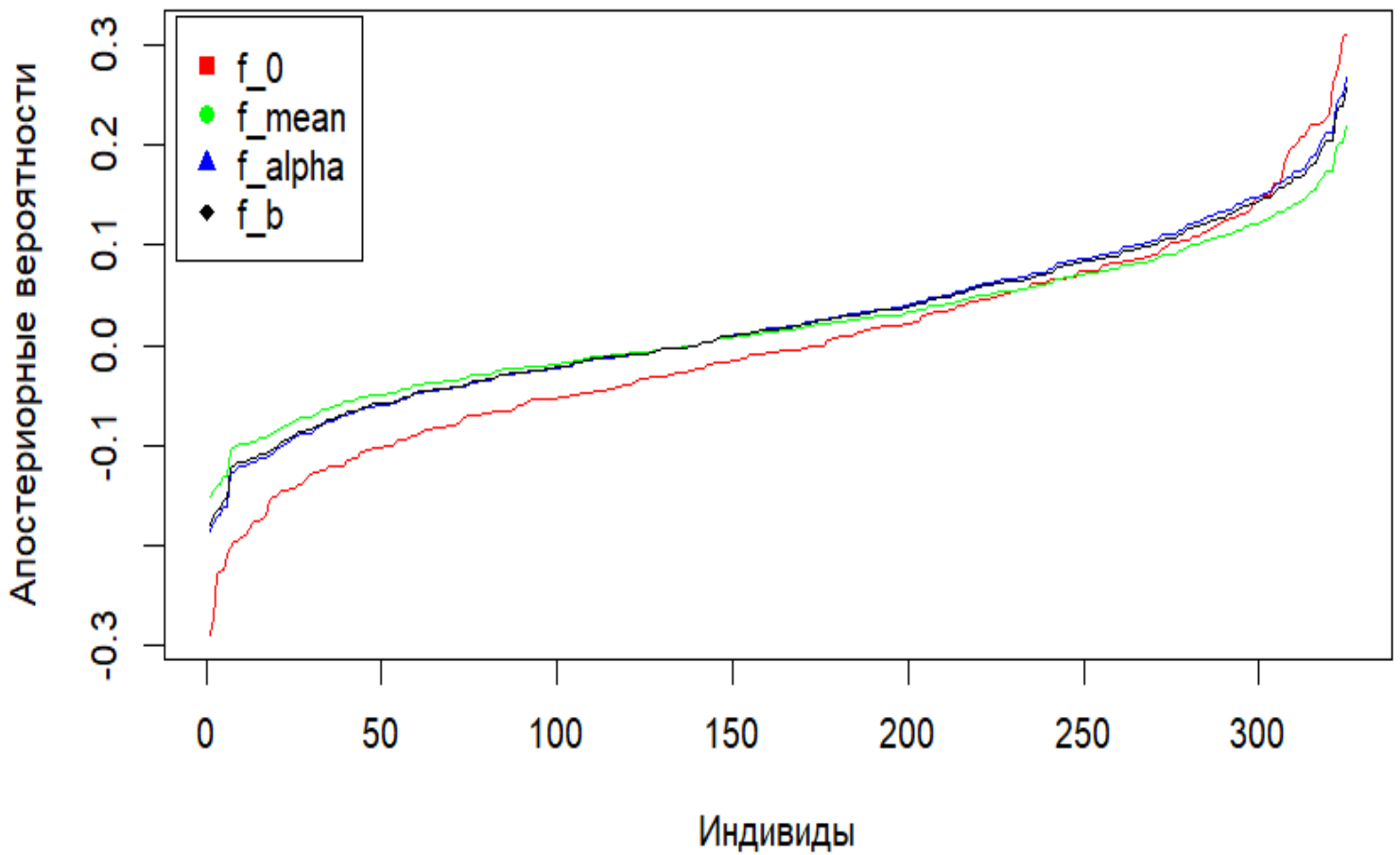
$$F(f) = f_b = \sum_{j=1}^{n} c_j f_j, \quad \text{где } c_j = \frac{J_{n-1}(\sigma_0, r) + rnJ_{n-2}(\sigma_0, r)\delta_j}{J_n(\sigma_0, r) + rnJ_{n-1}(\sigma_0, r)} \tag{5}$$

Следствие 1. Пусть имеется оценка f_b (5), то когда $\mu_i = \mu_0$ и $\sigma_0 = 0$ получим оценку $f_\alpha = f_b(r, n, 0, \mu_0 1^n)$:

$$f_\alpha = C_\alpha \bar{f}, \quad \text{где} \quad C_\alpha = \frac{n}{1 + r(n-1)} \tag{6}$$

Результаты применения частичной классификации

462 частичных классификаторов по 5 признакам $\Rightarrow$ были центрированы $\Rightarrow$ выбраны самые лучшие по точности классификаторы $\Rightarrow$ построена Λ_n вычислен $r = 0.8157 \Rightarrow$ вычислены оценки f_{mean}, f_α, f_b.



Расслоение популяции

Вероятности ошибочной классификации в популяции W_1, W_2:
$\mathcal{P} = q_1 P(2|1) + q_2 P(1|2)$.

- $q = P(W_1)$ – априорные вероятности, тогда $1 - q = P(W_2)$.
- Будем делать расслоение популяции с вероятностями $s = P(S^1),\ 1 - s = P(S^2)$.
- Если обозначить $x = P(W_1^1)$, тогда $P(W_2^1) = s - x,\ P(W_1^2) = q - x,$ $P(W_2^2) = 1 - s - q + x$.

Вычисляем вероятности ошибочной классификации:

$$\mathcal{P}_0 = q\Phi(u_0) + (1 - q)\Phi(v_0),$$
$$s\mathcal{P}_1 = x\Phi(u_1) + (s - x)\Phi(v_1),$$
$$(1 - s)\mathcal{P}_2 = (q - x)\Phi(u_2) + (1 - s - q + x)\Phi(v_2).$$

где $u_i = \frac{K_i}{\Delta_i} - \frac{\Delta_i}{2},\ v_i = -\frac{K_i}{\Delta_i} - \frac{\Delta_i}{2},\ K_0 = ln\left(\frac{1}{q} - 1\right),\ K_1 = ln\left(\frac{s}{x} - 1\right),\ K_2 = ln\left(\frac{1-s}{q-x} - 1\right)$. $i = 0$ нерасслоенная и $i = 1, 2$ расслоенные выборки.

Вероятность ошибочной классификации

Эффективность расслоения будет определятся через разность

$$P(x) = \mathcal{P}_0 - (s\mathcal{P}_1 + (1-s)\mathcal{P}_2)$$

Предложение 2.

Если расстояние Махаланобиса инвариантны относительно расслоения популяции, то вероятность случайной классификации при расслоении не увеличится.
Если признаки, определяющие расслоение общей популяции на $W_1^1, W_2^1, W_1^2, W_2^2$, независимы, то $x = qs$, следовательно, $K_0 = K_1 = K_2$, $u_0 = u_1 = u_2$, $v_0 = v_1 = v_2$, $P(sq) = 0, P'(sq) = 0, P''(sq) \geq 0$, т.е. в точке $x = qs$ функция P(x) имеет минимум.

Доказательство. Аналитическое выражение о том, что $P''(sq) \geq 0$:

$$P''(qs) = \frac{1}{\sqrt{2\pi}} e^{\frac{-v_1^2}{2}} \frac{1}{\Delta q^2 s(1-q)} + \frac{1}{\sqrt{2\pi}} e^{\frac{-v_2^2}{2}} \frac{1}{\Delta q^2 (1-s)(1-q)} \geq 0. \quad \blacksquare$$

Построение мультипликативного синдрома

Определение симптома. Пусть вектор $A = (a_1, \ldots, a_m)^T$, где элементы $a_i \in \{0, 1\}$ и набор $\tau = \{t : a_t = 1\}$, k – длина τ. Линейная комбинация

$$X_\tau = \sum_{t=1}^{m} a_t X_t (mod 2) = A^T \boldsymbol{X}_m (mod 2)$$

называется симптомом ранга k.

Определение синдрома. Пусть имеется $k + 1 > 0$ симптомов $X_0, \ldots, X_k$. Совокупность $2^{k+1} - 1$ симптомов вида $\beta_1 X_0 + \ldots + \beta_k X_k (mod 2)$, где коэффициенты $\beta_i \in F_2$ не равны нулю одновременно, называется синдромом k-ого порядка S_k.

$$S(X_1) = X_1, S(X_1, X_2) = S(X_1), X_2, S(X_1) + X_2 \, (mod 2)), \ldots,$$
$$S(\boldsymbol{X}_m = (S(\boldsymbol{X}_{m-1}), X_m, S(\boldsymbol{X}_{m-1} + X_m \, (mod 2)) \qquad (7)$$
$$\text{где } X_m \notin S(\boldsymbol{X}_{m-1}), m > 2.$$

Алгоритм применения симптомно-синдромального анализа

$y1 \dots y37$ – бинарные признаки выборки

$\Downarrow$

строим всевозможные комбинации для пары (y_i, y_j), $i, j = 1 \dots 37$, $i \neq j$

$\Downarrow$

получим $C_{37}^2 = \frac{37!}{2!(37-2)!} = 666$ пары

$\Downarrow$

для каждой пары (y_i, y_j) строим $C_{37}^2 = 666 \times 7 = 4662$ синдромы

$\Downarrow$

Перечислим все логические сочетания:

$$s_1 = y_i,$$
$$s_2 = y_j,$$
$$s_3 = (y_i + y_j)\, mod 2,$$
$$s_4 = y_i\, y_j,$$
$$s_5 = (y_i + y_i\, y_j)\, mod 2,$$
$$s_6 = (y_j + y_i\, y_j)\, mod 2,$$
$$s_7 = (y_i + y_j + y_i\, y_j)\, mod 2.$$

Таблица: Интерпретация симптомов:

y_i	y_j	s_1	s_2	s_3	s_4	s_5	s_6	s_7
0	0	0	0	0	0	0	0	0
0	1	0	1	1	0	0	1	1
1	0	1	0	1	0	1	0	1
1	1	1	1	0	1	0	0	1

Результаты классификации по расслоению на полных данных

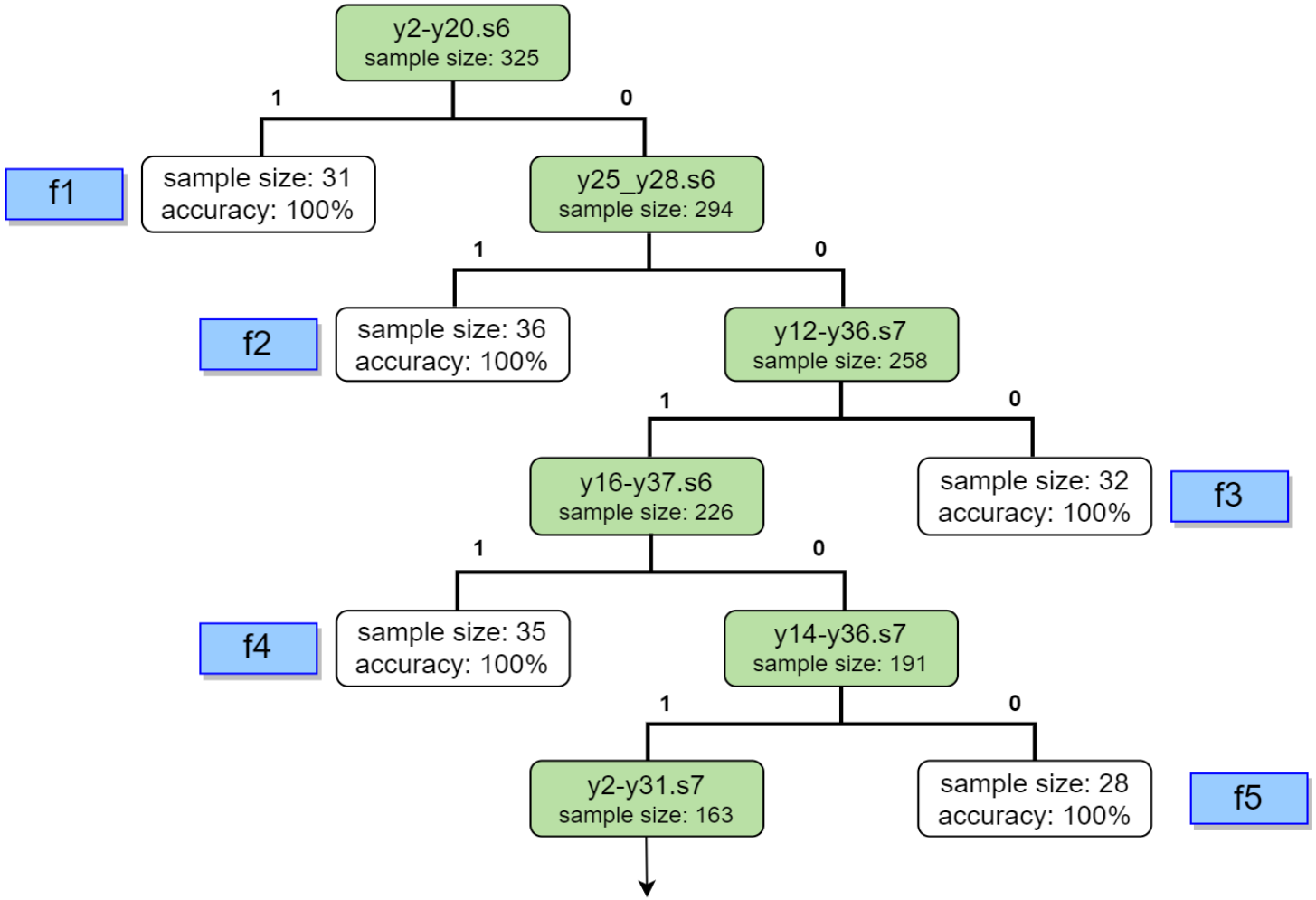


Рис.: График итерационной классификации по расслоению в виде дерева

Результаты классификации по расслоению на полных данных

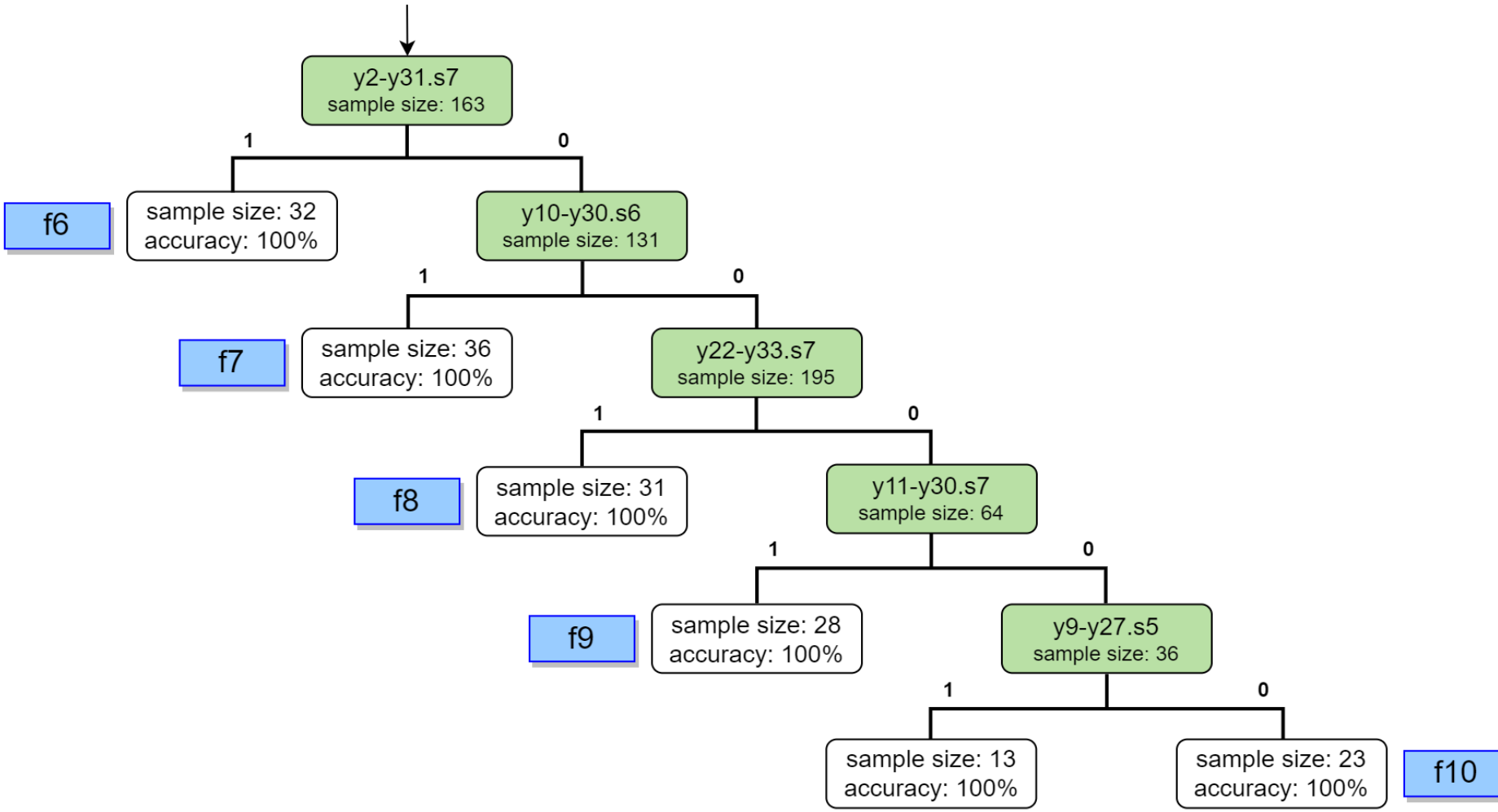


Рис.: График итерационной классификации по расслоению в виде дерева

Обьединение факторов риска

Таблица: Коэффициенты дискриминантных функций

	$f1$	$f2$	$f3$	$f4$	$f5$	$f6$	$f7$	$f8$	$f9$	$f10$
x1	1.62	0.46	0.11	-0.54	-0.5	-0.69	0.29	-2.15	2.09	1.08
x2	1.08	-0.05	-0.26	-0.27	0.15	0.09	-0.85	-0.78	-1.23	-0.71
x3	0.58	-0.72	0.57	-0.47	1.5	1.34	-1.01	1.42	-2.75	0.76
x4	-3.51	1.33	0.33	-0.27	-0.38	-2.31	-0.58	-0.74	1.49	0.76
x5	-0.22	0.29	0.48	0.63	0.2	0.18	0.76	0.26	-0.48	0.28
x6	0.38	-0.02	0.14	-0.14	2.69	0.98	-1.67	3.29	-4.76	-0.79
x7	0.29	-0.10	-0.51	-0.15	0.02	1.38	0.73	-0.21	0.64	0.21
x8	3.01	0.53	0.5	0.87	-1.21	-1.2	0.45	4.17	-3.27	2.1
x9	2.14	0.55	0.13	-5.35	-0.08	-0.13	-0.37	2.81	-0.97	2.26
x10	1.56	0.53	1.09	0.35	-0.005	-0.07	0.68	-3.69	0.39	0.37
x11	3.54	-1.75	-1.84	-1.05	-1.27	-2.39	0.07	-1.37	-10.4	0.36
x12	-3.9	1.69	0.92	-0.03	0.11	0.78	0.34	0.21	4.18	0.24
x13	1.84	-0.75	-0.45	0.53	-0.09	1.76	-1.1	-4.05	-3.31	0.58
x14	-1.17	-0.95	-0.69	-0.22	0.67	-1.08	-0.4	3.44	-2.12	0.65
x15	0.34	-0.62	0.63	1.14	0.59	2.1	-0.47	0.87	6.81	-2.76
x16	2.33	-0.66	0.49	1.5	-1.29	-0.65	-0.44	-1.0	1.15	0.71

Заключение

- Получена оценка полного прогноза по частичным для задачи классификации и мною были доказаны **Утверждение 1**, **Следствие 1** и **Предложение 1, 2**.
- Применила метод частичной классификации для прогнозирования ПКТС и получила вывод, что обычное усреднение частичных классификаторов достаточно для классификации.
- Для улучшения точности классификации применила расслоение на основе мультипликативного синдрома и получила алгоритмическое дерево.
- В результате расслоения получили классификацию с удовлетворительной точностью.

Спасибо за внимание!