Московский Физико-Технический Институт
(Государственный Университет)

Факультет Управления и Прикладной Математики
Кафедра «Интеллектуальные Системы»

Вдовина Евгения Александровна

Отбор признаков для многоклассовой классификации символьных последовательсностей

Выпускная квалификационная работа магистра

Научный руководитель:
д.ф-м.н. Воронцов Константин Вячеславович

Москва, 2016

Аннотация

В работе проблема отбора признаков исследуется на примере прикладной задачи — диагностики заболеваний внутренних органов по ЭКГ. Диагностика основана на информационном анализе кардиограммы, в процессе которого она переводится в символьную последовательность. Недостатком данного метода является низкое качество дифференциальной диагностики. В работе проверяется, можно ли улучшить его, изменив метод отбора признаков.

Содержание

Введение

Задача классификации символьных последовательностей возникает в символьной динамике, категоризации текстов [1] и других случаях, в том числе и в диагностике заболеваний по ЭКГ, основанной на информационном анализе кардиограммы [2, 3, 13, 14, 4, 5, 6, 7, 8, 9].

В настоящее время технология информационного анализа электрокардиосигналов реализована в диагностической системе «Скринфакс» [2]. Она позволяет диагностировать по одной электрокардиограмме более 30 различных заболеваний внутренних органов, не ограничиваясь заболеваниями сердечно-сосудистой системы. За 10 лет врачебного применения накоплена обучающая выборка — более 20 тысяч прецедентов записей электрокардиограмм и соответствующих им диагнозов.

Кардиограмма изначально представляет собой временной ряд. Однако, технология информационного анализа ЭКГ включает преобразование этого временного ряда в символьную последовательность — кодограмму. Статистическое обоснование данной технологии приведено в [10]. Алфавит кодограммы содержит 6 символов. Таким образом, диагностика заболеваний по ЭКГ сводится к многоклассовой классификации символьных последовательностей с пересекающимися классами, так как у одного человека одновременно могут быть несколько болезней.

Ранее в основном ставилась задача «абсолютно здоровые — страдающие болезнью А», и качество классификации при такой постановке значительно лучше, чем в задаче «страдающие болезнью А — страдающие болезнью Б». Причин этому может быть несколько:

- отсутствие диагноза в обучающей выборке может трактоваться как отсутсвие информации о наличии болезни (диагноза нет, потому что эту болезнь не искали);
- простые модели, успешно отличающие абсолютное здоровье от заболевания, не подходят для дифференцированной диагностики, модель нужно модифицировать.

Целью данной работы является проверить второе предположение. Модификация производится за счет изменения метода отбора признаков.

Глава 1

Постановка задачи

В задаче диагностики заболеваний по ЭКГ в качестве объекта выступает кодограмма кардиограммы пациента, а в качестве класса — болезнь (или абсолютное здоровье), от которой пациент страдает в момент снятия кардиограммы. Один пациент может иметь одновременно несколько заболеваний, но если он абсолютно здоров, то у него не может быть болезней. Следовательно, абсолютное здоровье — особый класс, единственный, гарантированно не имеющий пересечений со всеми остальными классами. Собственно, выборка абсолютно здоровых так и подбиралась — это ЭКГ, снятое у людей, не имеющих заболеваний.

Для работы с кодограммой используется выделение n-грамм — слов из n подряд идущих символов. В данном исследовании использовались триграммы. Всего в алфавите из 6 символов существует $6^3 = 216$ триграмм.

Таким образом, исходные данные выглядят так: N записей кардиограмм, для которых указано наличие хотя бы одной из L болезней (в том числе и абсолютное здоровье). Для каждой такой записи известны:

- вектор отметок о болезнях (d_{ic}), где i – номер записи ЭКГ, $i = 1, \ldots, N$; c — номер болезни (или абсоютного здоровья); $c = 1, \ldots, L$;

$$d_{ic} = \begin{cases} 0, \text{ болезни нет или ее наличие не проверялось;} \\ 1, \text{ болезнь есть (диагностирована врачом);} \\ 2, \text{ болезнь есть, эталонный случай (особо достоверный диагноз).} \end{cases}$$

- вектор частот триграмм в соответствующей кодограмме (n_{ik}), где i — номер записи, $i = 1, \ldots, N$; k – номер триграммы, $k = \overline{1, M}, M = 216$. n_{ik} – сколько раз триграмма k встретилась в кодограмме i-ой записи ЭКГ.

Ранее было выявлено, что следующая бинаризация данных улучшает качество классификации базовым алгоритмом:

$$x_{ik} = \begin{cases} 1, \ n_{ik} \geq 2 \\ 0, \ \text{иначе} \end{cases}$$

В данной работе не делалось различий между обычным и эталонным случаями, поэтому имело место следующее:

$$b_{ic} = \begin{cases} 0, \text{ если } d_{ic} = 0 \\ 1, \text{ иначе} \end{cases}$$

Во всех экспериментах использовались одни и те же данные следующего размера: $N = 15183$, $L = 32$. Список диагнозов и их условных обозначений представлен в таблице 1.1.

Таблица 1.1: Для каждого заболевания: название, аббревиатура, номер.

абсолютное здоровье	ЗД	1
вегетососудистая дистония	ВД	2
гипертоническая болезнь	ГБ	3
желчнокаменная болезнь	ЖК	4
ишемическая болезнь сердца	ИБ	5
мочекаменная болезнь	МК	6
миома матки	ММ	7
сахарный диабет	СД	8
узловой (диффузный) зоб щитовидный железы	УЩ	9
хронический гастрит (гастродуоденит) гипоацидный	ХГ	10
холецистит хронический	ХХ	11
анемия	А	12
аденома простаты	АП	13
аднексит хронический	АХ	14
язвенная болезнь	ЯБ	15
некроз головки бедренной кости	ГБК	16
хронический гастрит (гастродуоденит) гиперацидный	ГДЭ	17
дискинезия желчевыводящих путей	ДЖЭ	18
рак общий (онкопатология различной локализации)	РОЭ	19
рак (онкопатология различной локализации)	Cr	20
хронический энтероколит	БК	21
гепатоз	ГПЗ	22
фиброзно-кистозная мастопатия	МП	23
полип желудка	ПГ	24
полип желчного пузыря	ПЖ	25
полип кишки	ПК	26
простатит	ПС	27
полип(оз) эндомертия матки	ПУ	28
рак молочной железы	РМ	29
хронический бронхит	ХБ	30
эндометриоз	ЭМ	31
язва желудка	ЯЖ	32

Глава 2

Критерий качества и базовый метод классификации

Раннее к этой задаче применялись различные методы [11], в том числе наивный байесовский классификатор, случайный лес (random forest), тематическое моделирование [12]. Базовым выбран двухклассовый наивный байесовский классификатор, так как в данной случае он линеен и по качеству классификации не уступает случайному лесу.

Итак, принцип максимума апостериорной вероятности для случая двух классов:

$$a(\mathbf{x}) = \arg\max_{\{-1;+1\}} \{p(-1|\mathbf{x});\, p(+1|\mathbf{x})\} = \arg\max_{\{-1;+1\}} \left\{\frac{p(-1,\mathbf{x})}{p(\mathbf{x})};\, \frac{p(+1,\mathbf{x})}{p(\mathbf{x})}\right\} =$$

$$= \arg\max_{\{-1;+1\}} \{p(-1,\mathbf{x});\, p(+1,\mathbf{x})\} = \arg\max_{\{-1;+1\}} \{P(-1)p(\mathbf{x}|-1);\, P(+1)p(\mathbf{x}|+1)\} =$$

$$= sign\left(\log\frac{P(+1)p(\mathbf{x}|+1)}{P(-1)p(\mathbf{x}|-1)}\right) = sign\left(\log\frac{p(\mathbf{x}|+1)}{p(\mathbf{x}|-1)} + \log\frac{P(+1)}{P(-1)}\right) =$$

$$= sign\left(\log\frac{p(\mathbf{x}|+1)}{p(\mathbf{x}|-1)} + w\right)$$

Предположим, что признаки независимы, тогда:

$$p(\mathbf{x}|y) = p(x_1|y)\cdot\ldots\cdot p(x_M|y),\; y\in\{-1;+1\},\; \mathbf{x} = (x_1,\ldots,x_M)$$

Так как признаки бинарные, $x_k \in \{0;1\}$, то МП-оценка вероятности $p(x_k = v|y)$ имеет следующий вид:

$$p(x_k = v|y) = \frac{\sum_{i=1}^{N}[y_i = y][x_{ik} = v]}{\sum_{i=1}^{N}[y_i = y]},$$

где $y_i \in \{+1,-1\}$ — метка класса,

$$[expression] = \begin{cases} 1, \text{ если } expression \text{ принимает значение } true \\ 0, \text{ если } expression \text{ принимает значение } false \end{cases}$$

Тогда принцип максимума апостериорной вероятности принимает вид:

$$a(\mathbf{x}) = sign\left(\log\frac{p(\mathbf{x}|+1)}{p(\mathbf{x}|-1)} + w\right) = sign\left(\log\frac{\prod_{k=1}^{M} p(x_k|+1)}{\prod_{k=1}^{M} p(x_k|-1)} + w\right) =$$

$$= sign\left(\log\prod_{k=1}^{M}\frac{p(x_k|+1)}{p(x_k|-1)} + w\right) = sign\left(\sum_{k=1}^{M}\log\frac{p(x_k|+1)}{p(x_k|-1)} + w\right);$$

$$p(x_k|y) = p(x_k = 0|y)^{1-x_k} p(x_k = 1|y)^{x_k};$$

$$\log\frac{p(x_k|+1)}{p(x_k|-1)} = \log\frac{p(x_k = 0|+1)^{1-x_k} p(x_k = 1|+1)^{x_k}}{p(x_k = 0|-1)^{1-x_k} p(x_k = 1|-1)^{x_k}} =$$

$$= \log\frac{p(x_k = 0|+1)}{p(x_k = 0|-1)}\frac{p(x_k = 1|+1)^{x_k} p(x_k = 0|-1)^{x_k}}{p(x_k = 1|-1)^{x_k} p(x_k = 0|+1)^{x_k}} =$$

$$= \log\frac{p(x_k = 0|+1)}{p(x_k = 0|-1)} + x_k \log\frac{p(x_k = 1|+1)p(x_k = 0|-1)}{p(x_k = 1|-1)p(x_k = 0|+1)};$$

В итоге наивный байесовский классификатор оказывается линейным:

$$a(\mathbf{x}) = sign\left(\sum_{k=1}^{M}\log\frac{p(x_k|+1)}{p(x_k|-1)} + w\right) = sign\left(\sum_{k=1}^{M} w_k x_k + w_0\right) = sign(f(\mathbf{x}, \mathbf{w}) + w_0),$$

$$\text{где } w_k = \log\frac{p(x_k = 1|+1)p(x_k = 0|-1)}{p(x_k = 1|-1)p(x_k = 0|+1)} =$$

$$= \log\frac{\sum_{i=1}^{N}[y_i = +1][x_{ik} = 1]\cdot\sum_{i=1}^{N}[y_i = -1][x_{ik} = 0]}{\sum_{i=1}^{N}[y_i = -1][x_{ik} = 1]\cdot\sum_{i=1}^{N}[y_i = +1][x_{ik} = 0]} \text{ — вес признака.}$$

Используемый в данной работе критерий качества двухклассовой классификации — это AUC, площадь под ROC-кривой. Однако, AUC также и доля правильно упорядоченных пар ($\mathbf{x}_i$, $\mathbf{x}_j$):

$$AUC = AUC(X^N, \mathbf{f}(X^N)) = \frac{1}{N_+ N_-}\sum_{i=1}^{N}\sum_{j=1}^{N}[y_i \le y_j][f(\mathbf{x}_i, \mathbf{w}) < f(\mathbf{x}_j, \mathbf{w})], \qquad (2.1)$$

$$\mathbf{f}(X^N) = (f(\mathbf{x}_1, \mathbf{w}), ..., f(\mathbf{x}_N, \mathbf{w}))$$

где N_+ и N_- – количество объектов, принадлежащих классам «+1» и «−1» соответственно. Как отмечалось в предыдущем разделе, в данной задаче большинство классов могут иметь пересечения друг с другом. При вычислении AUC случай, когда объект принадлежит обоим классам, нужно учесть отдельно. Поэтому в формуле 2.1

$$y_i = \begin{cases} +1, \text{ если } \mathbf{x}_i \text{ принадлежит только классу } «+1» \\ -1, \text{ если } \mathbf{x}_i \text{ принадлежит только классу } «-1» \\ 0, \text{ если } \mathbf{x}_i \text{ принадлежит обоим классам} \end{cases}$$

и объекты, принадлежащие пересечению классов, учитываются и в N_+, и в N_-.

Алгоритм 2.0.1. Отбор признаков в наивном байесовском классификаторе.

Вход: обучающая выборка $X^N = \{\mathbf{x}_i\}_{i=1}^N$, вектора весов $\mathbf{w} = (w_1, ..., w_M)$ и инфомативности $\psi = (\psi_1, ..., \psi_M)$ признаков

Выход: мощность эталона K, эталон $\mathbf{k}$;

ПРОЦЕДУРА Selection$(X^N, \mathbf{w}, \psi)$:
отсортировать признаки $(1, ..., M)$ в порядке убывания критерия информативности: $(k_1, ..., k_M)$, $(\psi_{k_1}, ..., \psi_{k_M})$, $l < m \Leftrightarrow \psi_{k_l} \geq \psi_{k_m}$;
для всех $l = 1 \dots M$
    для всех $i = 1 \dots N$
        вычислить $\mathbf{f} = (f_1, ..., f_N)$, $f_i = \sum_{m=1}^{l} w_{k_m} x_{ik_m}$;
    вычислить $AUC_l = AUC(X^N, \mathbf{f})$;
$K = \arg\max_l \{AUC_l\}_{l=1}^M$;
$\mathbf{k} = (k_1, ..., k_K)$;
вернуть K, $\mathbf{k}$;

Для улучшения качества классификации проводится отбор признаков, основанный на критерии их информативности ψ для пары классов. Способ отбора описан в Алгоритме 2.0.1. В базовом методе используется $\psi = |w|$, где w – вес признака. **Множество информативных признаков далее называется эталоном. Диагностический эталон — эталон для отдельной болезни (а не для пары болезней).**

Для оценки качества по выборке проводится кросс-валидация, а именно контроль по Q блокам (Q-fold CV) со стратификацией классов. Усреднение AUC, весов признаков $\mathbf{w}$ и количества значимых признаков K проводится путем вычисления среднего арифметического, для K с округлением до ближайшего целого. В данной работе $Q = 10$. Обучение и тестирование описано в Алгоритме 2.0.2. Так как в базовом методе $\psi = |w|$, а вектор весов во время скользящего контроля считается отдельно для каждой из Q обучающих выборок, то получается, что и вектор информативности признаков также вычисляется для каждой выборки отдельно.

Алгоритм 2.0.2. Обучение и тестирование.

Вход: обучающие выборки X^{-1} и X^{+1} для классов «-1» и «$+1$», количество блоков Q, вектор информативности ψ;

Выход: усредненная мощность эталона K, усредненный вектор весов $\mathbf{w}$, усредненный AUC на контроле;

разбить выборки на Q примерно равных частей: $X^{-1} = \bigcup_{q=1}^{Q} X_q^{-1}$, $X^{+1} = \bigcup_{q=1}^{Q} X_q^{+1}$;
для всех $q = 1 \dots Q$
    $X_q = X_q^{-1} \bigcup X_q^{+1}$;
для всех $q = 1 \dots Q$
    $L_q = \bigcup_{t \neq q} X_t$;
    вычислить по выборке L_q вектор весов $\mathbf{w}_q$;
    $\{K_q, \mathbf{k}_q\}$ =Selection$(L_q, \mathbf{w}_q, \psi)$ (см. Алгоритм2.0.1);
    вычислить $\mathbf{f}_q = (f_{q1}, ..., f_{qN})$, где $f_{qi} = \sum_{l=1}^{K_q} w_{qk_l} x_{ik_l}$, $\mathbf{x_i} \in X_q$;
    $AUC_q = AUC(X_q, \mathbf{f}_q)$;
$AUC = \frac{1}{Q} \sum_{q=1}^{Q} AUC_q$;
$\mathbf{w} = \frac{1}{Q} \sum_{q=1}^{Q} \mathbf{w}_q$;
$K = [\frac{1}{Q} \sum_{q=1}^{Q} K_q]$, где $[\,]$ – округление до ближайшего целого;
вернуть $AUC, K, \mathbf{w}$;

Глава 3

Качество классификации базовым методом

Наивный байесовский классификатор показывает относительно хорошие результаты ($AUC > 0.85$) в двухклассовой задаче «абсолютно здоровые — страдающие болезнью А», где один класс — это абсолютное здоровье, а другой — какая-нибудь болезнь. Но в двуклассовой задаче «страдающие болезнью А — страдающие болезнью Б» качество заметно падает (см. рис. 3.1 и 3.2). Как отмечалось выше, классы могут пересекаться. На рис. 3.3 показаны мощности попарных пересечений классов, на диагонали — мощности самих классов. Некоторые пересечения по мощности сравнимы с классами. Однако, как видно на рис. 3.4, удаление пересечений классов на общую картину влияет незначительно.

	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
ЗД	50	87	94	95	95	93	93	95	93	91	95	89	95	93	92	98	96	93	94	90	92	95	90	93	93	93	94	92	96	95	91	93
ВД	87	50	73	76	78	69	68	76	71	66	73	63	78	64	67	84	69	67	79	70	68	77	67	77	72	76	70	62	78	74	61	75
ГБ	94	73	50	60	60	57	60	57	54	56	56	65	60	64	58	76	66	59	62	62	56	55	62	60	63	57	59	63	63	56	63	57
ЖК	95	76	60	50	67	63	65	59	58	65	54	71	60	67	62	72	65	62	66	67	63	60	65	63	65	63	63	65	59	59	68	62
ИБ	95	78	60	67	50	63	65	54	58	62	56	71	58	68	63	75	69	63	59	66	61	58	67	60	66	60	64	69	65	58	68	62
МК	93	69	57	63	63	50	57	63	58	54	59	59	63	61	55	80	62	55	68	65	55	60	59	63	62	62	56	61	63	59	60	58
ММ	93	68	60	65	65	57	50	65	58	57	60	59	66	58	56	79	66	59	69	61	58	62	56	69	60	62	59	58	64	61	59	59
СД	95	76	57	59	54	63	65	50	60	64	58	71	56	67	63	77	70	64	59	65	61	59	67	63	64	60	63	69	61	56	68	60
УЩ	93	71	54	58	58	58	58	60	50	57	56	64	61	62	56	75	64	60	66	61	56	60	58	59	57	56	60	60	62	57	63	56
ХГ	91	66	56	65	62	54	57	64	57	50	59	57	64	61	55	80	66	57	67	63	57	60	58	65	59	61	60	57	66	60	58	60
ХХ	95	73	56	54	56	59	60	58	56	59	50	67	60	60	56	73	63	57	63	64	57	57	62	64	64	61	57	65	63	56	63	62
А	89	63	65	71	71	59	59	71	64	57	67	50	72	62	61	84	69	62	72	67	60	67	60	69	63	66	65	60	72	68	59	65
АП	95	78	60	60	58	63	66	56	61	64	60	72	50	71	63	78	69	65	61	66	64	57	67	61	66	55	63	68	59	56	71	59
АХ	93	64	64	67	68	61	58	67	62	61	60	62	71	50	60	76	68	61	72	66	59	65	62	73	63	69	62	58	68	63	61	67
ЯБ	92	67	58	62	63	55	56	63	56	55	56	61	63	60	50	79	64	56	69	65	56	61	57	65	59	59	56	57	63	58	59	62
ГБК	98	84	76	72	75	80	79	77	75	80	73	84	78	76	79	50	74	76	78	79	79	78	79	82	81	83	78	77	79	77	79	79
ГДЭ	96	69	66	65	69	62	66	70	64	66	63	69	69	68	64	74	50	61	72	74	64	69	69	71	70	72	64	61	69	66	63	68
ДЖЭ	93	67	59	62	63	55	59	64	60	57	57	62	65	61	56	76	61	50	69	69	60	59	61	68	65	65	56	61	65	61	60	63
РОЭ	94	79	62	66	59	68	69	59	66	67	63	72	61	72	69	78	72	69	50	66	64	61	72	66	70	65	68	71	65	62	71	65
Cr	90	70	62	67	66	65	61	65	61	63	64	67	66	66	65	79	74	69	66	50	60	66	65	68	58	59	67	63	68	62	68	59
БК	92	68	56	63	61	55	58	61	56	57	57	60	64	59	56	79	64	60	64	60	50	58	59	65	60	61	60	59	64	58	59	61
ГПЗ	95	77	55	60	58	60	62	59	60	60	57	67	57	65	61	78	69	59	61	66	58	50	65	59	63	57	59	68	63	55	67	60
МП	90	67	62	65	67	59	56	67	58	58	62	60	67	62	57	79	69	61	72	65	59	65	50	69	57	61	62	62	66	63	61	62
ПГ	93	77	60	63	60	63	69	63	59	65	64	69	61	73	65	82	71	68	66	68	65	59	69	50	63	55	68	67	67	64	72	58
ПЖ	93	72	63	65	66	62	60	64	57	59	64	63	66	63	59	81	70	65	70	58	60	63	57	63	50	55	65	61	66	64	65	57
ПК	93	76	57	63	60	62	62	60	56	61	61	66	55	69	59	83	72	65	65	59	61	57	61	55	55	50	62	64	62	59	69	51
ПС	94	70	59	63	64	56	59	63	60	60	57	65	63	62	56	78	64	56	68	67	60	59	62	68	65	62	50	61	63	57	63	62
ПУ	92	62	63	65	69	61	58	69	60	57	65	60	68	58	57	77	61	61	71	63	59	68	62	67	61	64	61	50	70	65	58	62
РМ	96	78	63	59	65	63	64	61	62	66	63	72	59	68	63	79	69	65	65	68	64	63	66	67	66	62	63	70	50	59	69	59
ХБ	95	74	56	59	58	59	61	56	57	60	56	68	56	63	58	77	66	61	62	62	58	55	63	64	64	59	57	65	59	50	66	57
ЭМ	91	61	63	68	68	60	59	68	63	58	63	59	71	61	59	79	63	60	71	68	59	67	61	72	65	69	63	58	69	66	50	66
ЯЖ	93	75	57	62	62	58	59	60	56	60	62	65	59	67	62	79	68	63	65	59	61	60	62	58	57	51	62	62	59	57	66	50

Рис. 3.1: Матрица значений 100×AUC в двухклассовой задаче (розовый — минамальное значение, белый — максимальное)

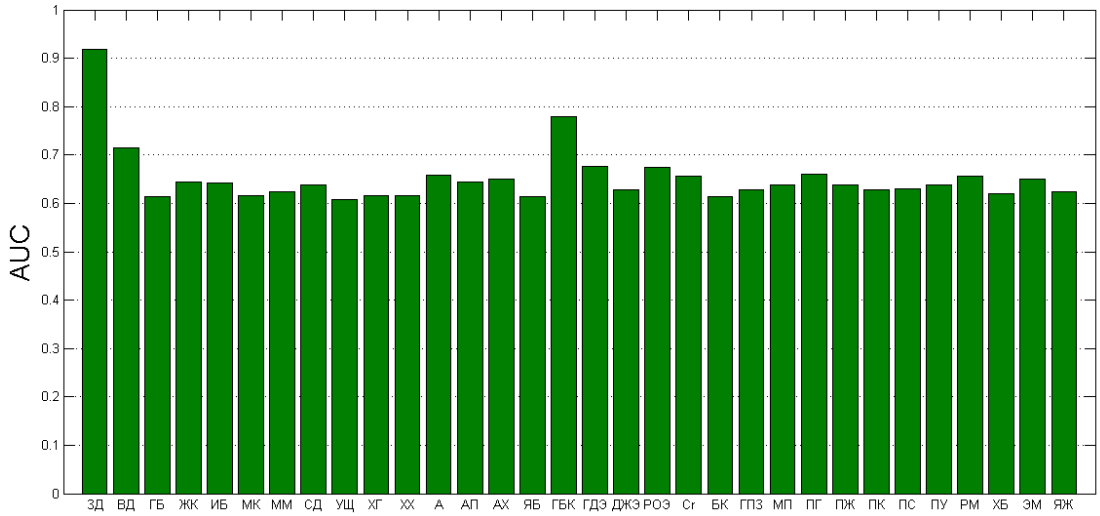


Рис. 3.2: Диаграмма значений AUC, усредненных по двухклассовым задачам для каждого класса (см. формулу 7.1).

	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
ЗД	536	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ВД	0	1260	37	4	28	46	49	18	71	59	53	37	7	81	33	0	66	70	0	1	51	6	73	7	0	5	79	16	0	27	80	0
ГБ	0	37	2764	172	1480	613	323	513	370	316	510	75	313	145	254	0	130	459	78	4	331	296	221	34	95	56	275	27	36	218	128	29
ЖК	0	4	172	620	188	93	84	116	75	50	59	14	68	33	66	0	0	7	9	3	55	49	60	8	6	12	29	5	17	55	31	11
ИБ	0	28	1480	188	2770	458	221	494	339	309	463	63	379	86	272	0	61	238	77	3	327	236	165	46	53	66	229	20	41	259	120	40
МК	0	46	613	93	458	1770	229	221	182	121	266	105	175	98	219	0	51	140	19	13	256	131	152	20	18	49	172	13	34	181	136	17
ММ	0	49	323	84	221	229	1777	103	297	133	171	151	1	214	114	0	12	75	14	8	160	69	339	32	61	19	0	38	52	152	322	11
СД	0	18	513	116	494	221	103	1267	166	71	261	25	123	102	127	0	5	41	104	0	171	183	63	13	70	23	79	0	24	122	68	12
УЩ	0	71	370	75	339	182	297	166	1641	120	198	63	93	89	84	0	16	83	17	0	178	57	223	24	22	34	38	18	33	98	135	9
ХГ	0	59	316	50	309	121	133	71	120	1396	255	164	63	64	14	0	0	94	0	0	244	49	101	60	15	51	48	14	15	61	78	4
ХХ	0	53	510	59	463	266	171	261	198	255	1558	88	140	156	166	0	56	242	47	4	274	130	190	17	83	26	126	13	29	126	116	35
А	0	37	75	14	63	105	151	25	63	164	88	747	30	85	25	0	2	35	22	11	62	9	58	26	17	19	19	6	14	26	92	0
АП	0	7	313	68	379	175	1	123	93	63	140	30	866	0	83	0	33	71	23	13	28	53	1	10	8	40	24	0	0	72	1	23
АХ	0	81	145	33	86	98	214	102	89	64	156	85	0	770	38	0	17	48	3	0	150	33	151	4	63	12	0	16	31	82	219	6
ЯБ	0	33	254	66	272	219	114	127	84	14	166	25	83	38	1212	0	3	54	14	0	69	62	86	3	13	15	113	7	10	116	90	50
ГБК	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	307	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ГДЭ	0	66	130	0	61	51	12	5	16	0	56	2	33	17	3	0	306	108	0	0	53	8	21	0	4	2	90	9	0	20	19	0
ДЖЭ	0	70	459	7	238	140	75	41	83	94	242	35	71	48	54	0	108	675	4	4	148	57	62	5	12	11	106	6	0	43	47	5
РОЭ	0	0	78	9	77	19	14	104	17	0	47	22	23	3	14	0	0	4	509	57	3	0	2	0	12	0	12	0	62	19	0	0
Cr	0	1	4	3	3	13	8	0	0	0	4	11	13	0	0	0	0	4	57	255	0	0	0	0	0	0	3	0	0	12	0	0
БК	0	51	331	55	327	256	160	171	178	244	274	62	28	150	69	0	53	148	3	0	1197	82	146	18	55	36	128	16	10	125	126	14
ГПЗ	0	6	296	49	236	131	69	183	57	49	130	9	53	33	62	0	8	57	0	0	82	599	37	16	30	11	89	0	3	47	40	7
МП	0	73	221	60	165	152	339	63	223	101	190	58	1	151	86	0	21	62	2	0	146	37	1488	18	48	7	0	32	1	112	194	18
ПГ	0	7	34	8	46	20	32	13	24	60	17	26	10	4	3	0	0	5	0	0	18	16	18	179	10	10	3	0	0	0	0	2
ПЖ	0	0	95	6	53	18	61	70	22	15	83	17	8	63	13	0	4	12	12	0	55	30	48	10	379	0	12	14	12	12	16	0
ПК	0	5	56	12	66	49	19	23	34	51	26	19	40	12	15	0	2	11	0	0	36	11	7	10	0	320	15	0	0	13	12	3
ПС	0	79	275	29	229	172	0	79	38	48	126	19	24	0	113	0	90	106	12	3	128	89	0	3	12	15	983	0	0	86	0	17
ПУ	0	16	27	5	20	13	38	0	18	14	13	6	0	16	7	0	9	6	0	0	16	0	32	0	14	0	0	214	1	2	35	0
РМ	0	0	36	17	41	34	52	24	33	15	29	14	0	31	10	0	0	0	62	0	10	3	1	0	12	0	0	1	492	25	35	0
ХБ	0	27	218	55	259	181	152	122	98	61	126	26	72	82	116	0	20	43	19	12	125	47	112	0	12	13	86	2	25	987	81	4
ЭМ	0	80	128	31	120	136	322	68	135	78	116	92	1	219	90	0	19	47	0	0	126	40	194	0	16	12	0	35	35	81	1018	5
ЯЖ	0	0	29	11	40	17	11	12	9	4	35	0	23	6	50	0	0	5	0	0	14	7	18	2	0	3	17	0	0	4	5	196

Рис. 3.3: Матрица мощностей попарных пересечений классов, по диагонали — мощность класса (белый — минамальное значение, серый — максимальное).

Рис. 3.4: Сравнение качества классификации в случае удаления пересечений классов со случаем, когда пересечения не удалялись. Вертикальная ось — AUC для c-ого класса, усредненный по всем двухклассовым задачам с участием c-ого класса (см. формулу 7.1).

Глава 4

Кластеризация болезней

В предыдущей главе отмечалось, что качество классификации в задаче «страдающие болезнью А — страдающие болезнью Б» значительно ниже, чем в «абсолютно здоровые — страдающие болезнью А». Всвязи с этим была проведена кластеризация болезней двумя разными способами. Целью кластеризации было проверить, какие заболевания окажутся наиболее схожими и можно ли это сходство объяснить с точки зрения медицины.

AUC в качестве расстояния между кластерами. AUC принимает значения от 0.5 до 1. Чем ближе AUC к 0.5, тем менее различны классы для данного алгоритма, то есть AUC можно рассматривать как меру близости (функцию расстояния) между кластерами.

Проводится классификация по стратегии «каждый против каждого». Таким образом, для всех неупорядоченных пар болезней становится известен AUC. После слияния пары заболеваний с наименьшим AUC в один кластер этот кластер считается «агрегированной болезнью» и для каждой пары «новый кластер — болезнь (возможно, «агрегированная»)» проводится классификация и подсчет AUC. Так продолжается до тех пор, пока не останется всего 1 кластер. Результат в виде дендрограммы показан на рис. 4.1. По дендрограмме видно, что к моменту, когда минимальное расстояние между кластерами (то есть AUC) стало больше 0.7, осталось всего 4 кластера, причем три из них содержат по одному классу. Трудно с точки зрения медицины объяснить такое сильное сходство между заболеваниями.

В данной задаче классы могут пересекаться, и это может влиять на качестве классификации. Поэтому кластеризация с AUC в качестве расстояния была повторена для случая отсутствия пересечений. В этом случае для каждой пары (агрегированных) болезней выборки создаются каждый раз заново. То есть из исходных данных выбираются все записи, для которых отмечено наличие 1-ой болезни (1-ого набора болезней), и все записи, с отметкой о 2-ой болезни (2-ом наборе болезней). Затем из каждого множества удаляются элементы, входящие в их пересечение. Далее так же, как и без удаления пересечений, применяется Алгоритм 2.0.2. Как видно на рис. 4.2, результат принципиально не изменился.

Расстояние между множествами значимых признаков в качестве расстояния между кластерами. Рассматриваются только пары «абсолютно здоровые

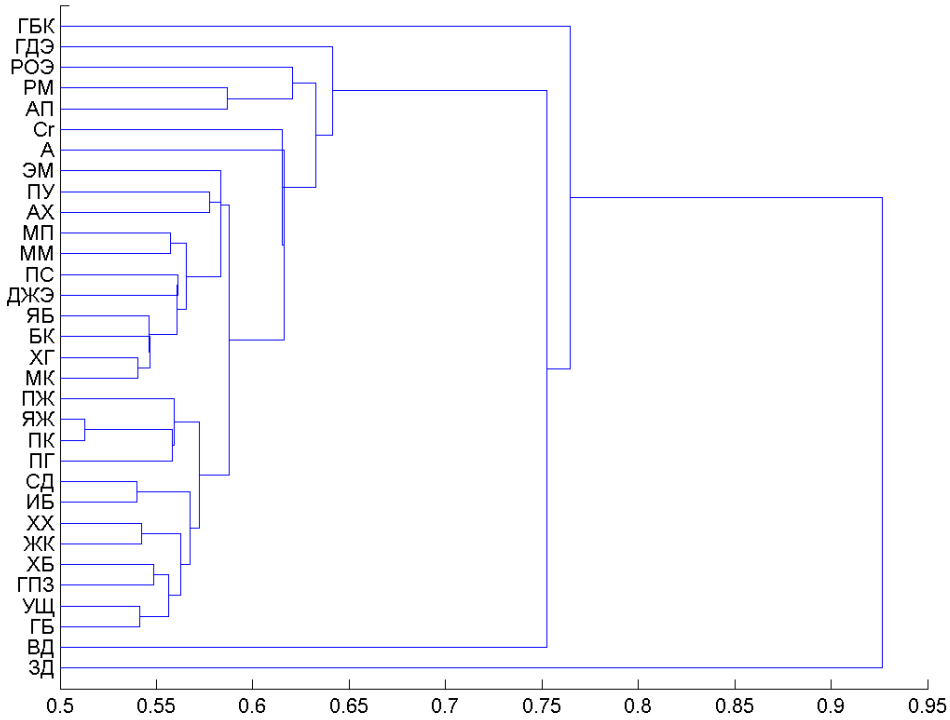


Рис. 4.1: Кластеризация болезней с использованием AUC в качестве расстояния между кластерами (без удаления пересечений кластеров). Горизонтальная ось — значения AUC, вертикальная ось — болезни.

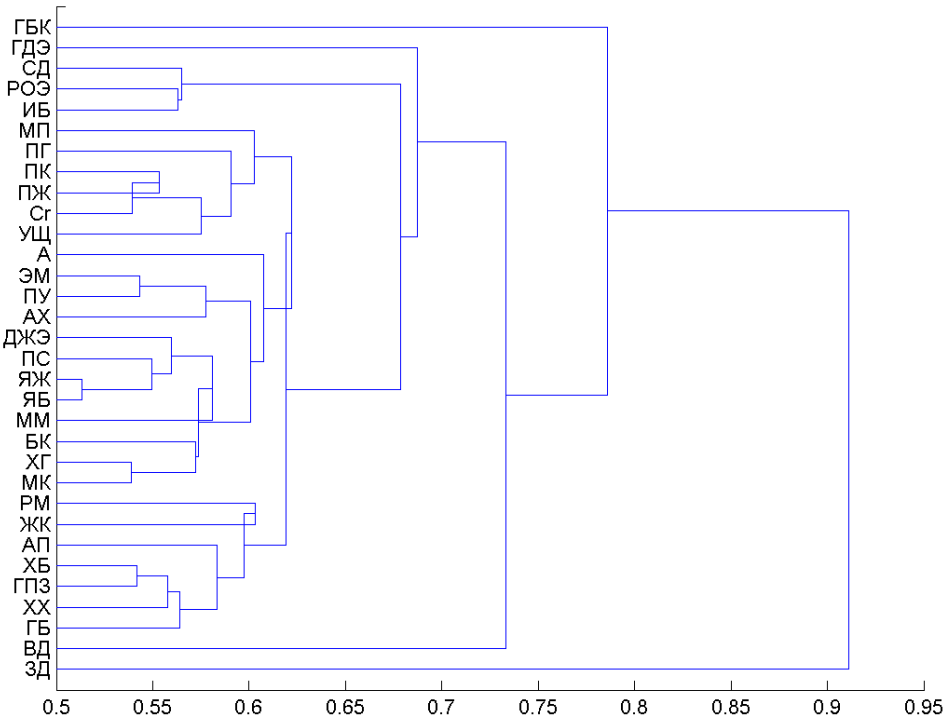


Рис. 4.2: Кластеризация болезней с использованием AUC в качестве расстояния между кластерами (без пересечений кластеров). Горизонтальная ось — значения AUC, вертикальная ось — болезни.

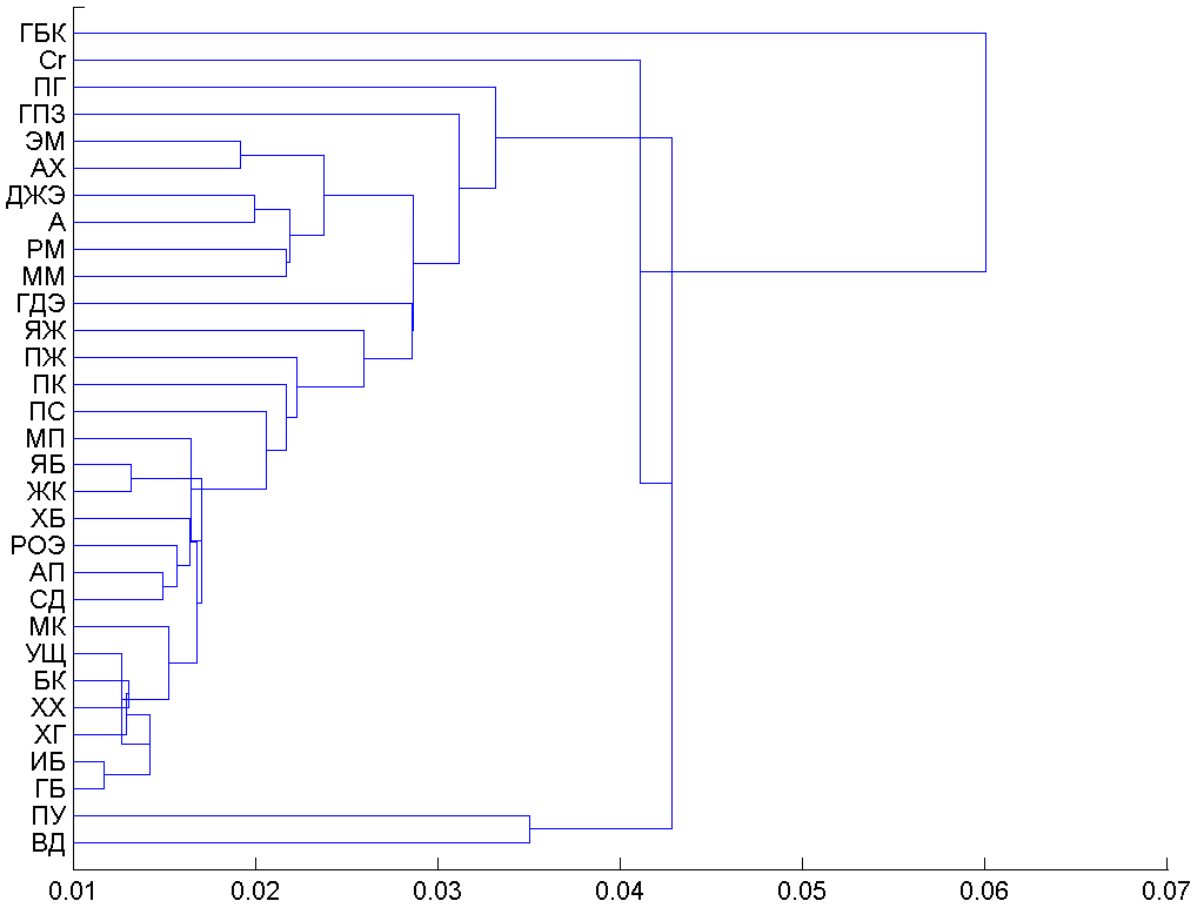


Рис. 4.3: Кластеризация болезней с использованием расстояния между множествами значимых признаков в качестве расстояния между кластерами. Горизонтальная ось — значения $f(d_1, d_2)$, расстояния между кластерами; вертикальная ось — болезни.

— страдающие болезнью (набором болезней) A». Так как здоровье исключает наличие заболеваний, то проблема пересечения классов не стоит. После проведения такой классификации у каждой болезни есть множество значимых признаков T и вектор весов признаков $\mathbf{w}$. Расстояние между кластерами определяется через веса общих значимых признаков:

$$f(d_1, d_2) = \frac{1}{2n} \sum_{k \in T_1 \bigcap T_2} |\frac{w_{1k}}{w_1^{max}} - \frac{w_{2k}}{w_2^{max}}|, 0 \leq f(d_1, d_2) \leq 1$$

где w_{ck} — вес k-ого признака для c-ого кластера, являющегося значимым для обоих кластеров ($k \in T_1 \bigcap T_2$); $w_c^{max} = \max_{k=\overline{1,M}} |w_{ck}|$ — модуль наибольшего по модулю веса признака для c-ого кластера; d_c — номер болезни (набор номеров болезней), относящейся к c-ому кластеру; $n = |T_1 \bigcap T_2|$ — количество общих значимых признаков; T_c — множество значимых признаков c-ого кластера.

Естественно, что здоровье в данном случае в кластеризации не участвует. Этот класс используется только для получения вектора весов и отбора признаков для вновь образовавшихся кластеров. Результат в виде дендрограммы показан на рис. 4.3.

Глава 5

Другие методы отбора признаков

Как уже отмечалось ранее, качество классификации в задаче «страдающие заболеванием А – страдающие заболеванием Б» довольно низкое — AUC в среднем по парам болезней (без учета абсолютного здоровья) составляет примерно 0.63. AUC, усредненный по парам классов, — это

$$\overline{AUC} = \frac{1}{L(L-1)} \sum_{c_{-1}<c_{+1}} AUC_{c_{-1}c_{+1}}, \text{ где } c_{-1}, c_{+1} = \overline{1, L}$$

где $AUC_{c_{-1}c_{+1}}$ – значение AUC для двухклассовой задачи с классами «-1» и «+1» — соответственно c_{-1}-м и c_{+1}-м классами.

При этом если рассматривать зависимость AUC от количества учитываемых признаков, то характерным для большинства пар классов будет следующее: с определенного момента AUC почти не меняется. В случае байесовского классификатора это значит, что есть большая свобода в выборе безызбыточного эталона. Иллюстрация этого факта представлена на рис. 5.1. На нем показана диаграмма количества признаков K и K_δ, где

$$K_\delta = \arg\max_{l=\overline{1,M}} [AUC(l) \leq AUC - \delta] \cdot AUC(l),$$

$$AUC(l) = \frac{1}{Q} \sum_{q=1}^{Q} AUC_q(l),\ AUC_q(l) = AUC(X_q, \mathbf{f}_q(l)),$$

$$\mathbf{f}_q(l) = (f_{q1}(l), ..., f_{qN}(l)),\ f_{qi}(l) = \sum_{s=1}^{l} w_{qk_s} x_{ik_l},\ \mathbf{x}_i \in X_q$$

(смотреть Алгоритмом 2.0.2).

На диаграмме $\delta = 0.005$.

Поиск диагностического эталона.Возможно, что плохое качество классификации является результатом избыточного набора информативных признаков. К тому же, было бы идеально, если бы для каждой болезни существовало такое множество признаков, которое отличало ее от всех остальных заболеваний — диагностический эталон. Поэтому родилась следующая идея. В предыдущих экспериментах было построено по 31 бинарному классификатору для каждой болезни, который отличает

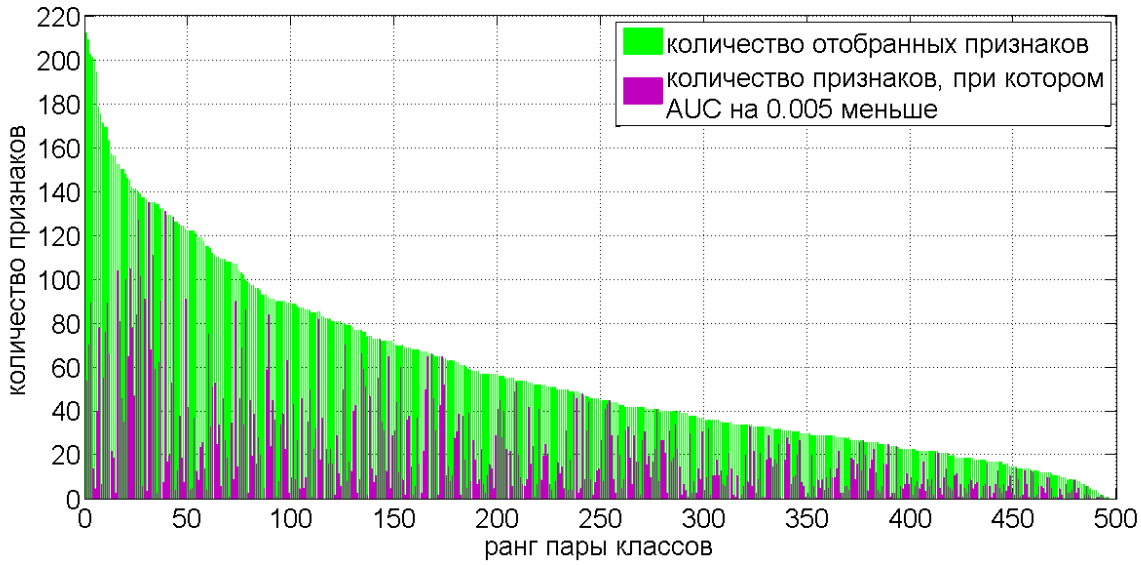


Рис. 5.1: Зеленый – K, розовый – K_δ, $\delta = 0.005$. Как видно на диаграмме, K намного обольше, чем K_δ.

эту болезнь от фиксированной другой. Идея состояла в том, чтобы из этого 31 эталнона сделать один, который бы отличал болезнь от всех остальных, а потом его редуцировать. Однако, как можно видеть на рис. 5.2, тривиальное решение не работает: пересечение всех эталонов для одной болезни почти у всех пустое. Кроме того, применение классификатора «абсолютно здоровые – страдающие заболеванием А» к случаю «страдающие заболеванием Б – страдающие заболеванием А» показало, что эталон, отличающий болезнь от здоровья, плохо отличает болезнь от болезни, а значит, не является диагностическим. Это видно на рис. 5.4 и 5.5. Обучение по стратегии «один против всех» также не дало хорошего результата, так как отобранных эталонов не достаточно, как показано на рис. 5.3.

	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
мощность объединения	105	190	216	216	216	216	215	216	216	210	216	201	211	203	216	212	214	214	210	210	214	216	215	214	209	215	216	213	210	215	215	214
мощность пересечения	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Рис. 5.2: Мощности пересечений и объединений всех эталонов (31 эталона) для каждой болезни

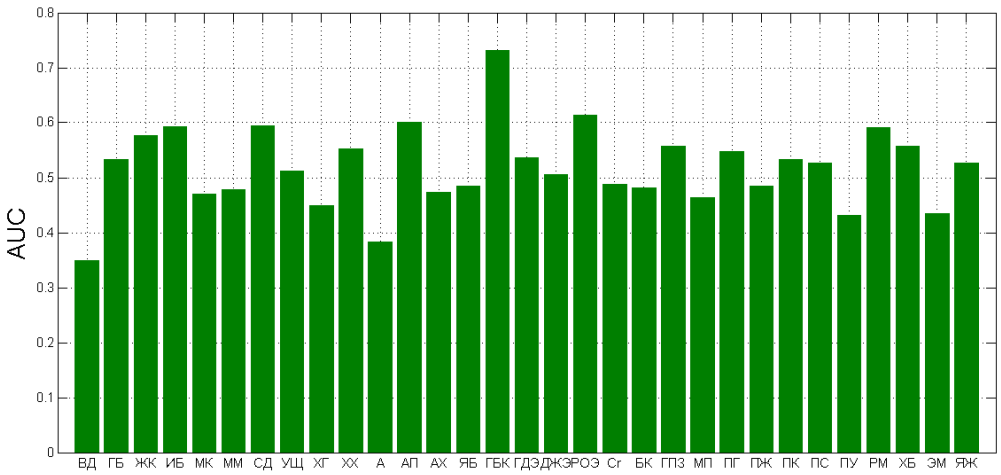
	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
AUC на обучении	0,91	0,74	0,81	0,85	0,81	0,83	0,82	0,82	0,79	0,82	0,85	0,83	0,83	0,89	0,80	0,79	0,87	0,91	0,79	0,66	0,83	0,90	0,82	0,87	0,85	0,82	0,83	0,84	0,75	0,85	0,81	0,83
K	38	1	1	1	1	1	1	1	1	1	1	1	1	1	1	74	1	1	1	29	1	1	1	1	1	1	1	1	1	1	1	1

Рис. 5.3: Результат обучения классификатора «страдающие заболеванием А – страдающие остальными заболеваниями». K – мощность найденного эталона.

Другие критерии информативности. Так как модуль веса в качестве критерия информативности признака приводит к избыточному набору информативных

	остальные болезни																													
ВД	33	30	28	38	37	29	35	41	30	47	28	38	37	23	34	36	29	42	37	32	41	38	39	36	33	41	29	31	41	36
ГБ	71	47	44	56	56	45	53	58	49	65	44	59	55	41	56	55	43	57	55	49	59	53	56	52	52	62	46	48	62	52
ЖК	74	55	49	61	61	50	56	63	53	69	49	62	59	43	58	58	48	62	60	54	63	58	61	57	56	66	51	53	66	57
ИБ	76	56	52	62	62	50	58	64	54	71	50	65	61	46	62	61	48	62	61	55	64	57	62	58	59	68	52	54	68	58
МК	66	44	40	39	50	39	46	52	42	59	38	53	49	36	49	48	38	52	49	42	52	46	50	45	45	57	39	42	56	47
ММ	66	45	42	39	51	39	47	53	43	60	38	53	50	36	51	50	39	52	50	44	53	48	51	47	47	57	40	42	56	47
СД	75	56	54	50	62	62	59	64	55	70	50	65	61	48	62	61	48	62	61	55	64	58	62	58	59	68	51	54	68	58
УЩ	70	48	44	42	55	54	43	56	47	63	42	57	53	39	53	53	42	55	53	47	56	50	54	49	50	60	43	46	60	50
ХГ	63	42	40	37	48	47	37	44	40	57	36	50	47	34	47	46	36	50	47	41	50	45	48	44	43	53	37	40	53	44
ХХ	72	52	49	46	58	58	47	55	60	67	46	60	57	41	57	56	45	60	57	51	61	56	59	55	54	65	47	50	63	55
А	57	36	33	30	41	41	31	38	44	34	30	44	40	28	40	39	31	44	41	34	44	39	41	37	37	47	31	33	47	39
АП	76	57	54	51	63	63	51	59	65	55	72	66	62	48	62	61	49	63	62	56	65	59	63	59	59	69	52	54	68	58
АХ	63	45	41	39	51	49	40	47	53	42	59	40	49	32	47	48	40	53	49	44	53	50	51	48	46	55	41	42	54	48
ЯБ	66	46	43	40	52	51	40	48	54	44	61	39	54	35	49	49	40	53	51	45	54	50	52	48	47	57	40	43	57	48
ГБК	83	71	66	66	76	75	67	72	77	68	82	69	74	75	69	72	65	77	75	72	77	76	78	77	73	78	68	70	77	73
ГДЭ	70	51	46	45	57	57	46	53	59	48	66	46	57	55	35	53	45	60	56	51	59	56	59	56	51	61	46	49	60	55
ДЖЭ	68	48	44	42	54	53	43	49	56	45	62	43	55	52	37	50	42	56	53	46	56	51	54	50	48	60	43	46	59	52
РОЭ	76	58	56	53	64	64	53	61	65	57	71	53	66	63	50	64	63	64	63	57	66	61	64	61	61	69	55	56	68	60
Cr	66	46	44	41	52	51	41	48	54	45	60	40	54	51	40	53	52	41	51	45	53	47	50	47	49	56	42	44	57	47
БК	66	45	43	39	51	51	40	48	53	44	60	39	53	50	37	51	50	38	52	44	54	48	51	47	47	57	41	43	57	47
ГПЗ	75	52	49	47	59	59	46	55	60	51	67	46	62	58	46	60	57	44	59	57	61	54	58	53	55	66	48	50	65	54
МП	65	44	41	38	50	49	38	45	51	42	58	37	52	48	37	50	48	38	50	49	42	45	48	44	46	54	38	41	55	45
ПГ	74	51	48	45	58	59	46	53	59	50	66	45	62	57	44	59	57	44	57	57	50	60	57	52	54	64	48	50	64	54
ПЖ	67	45	42	39	52	52	41	47	53	43	61	40	53	50	36	51	50	40	53	51	44	54	47	47	47	58	41	44	57	49
ПК	72	50	46	44	56	56	44	52	58	49	65	43	60	55	44	58	55	43	56	55	48	58	51	55	53	63	45	48	63	52
ПС	70	50	46	44	56	55	45	52	58	48	65	44	57	54	40	53	53	44	58	55	48	58	54	57	52	62	45	47	61	53
ПУ	60	41	38	35	47	45	35	43	49	38	56	34	48	45	31	44	44	35	49	46	40	49	45	47	43	42	35	38	51	42
РМ	75	56	53	50	62	62	50	58	64	54	71	50	64	61	46	60	60	49	63	61	55	64	59	62	59	58	67	54	67	58
ХБ	72	53	50	47	59	58	47	55	61	51	67	47	61	57	44	58	57	46	59	57	52	61	56	59	55	54	64	48	64	55
ЭМ	60	41	38	35	47	46	36	43	49	38	56	36	47	45	28	43	44	36	50	46	40	50	46	47	45	42	51	37	39	45
ЯЖ	70	50	48	44	56	55	44	52	58	49	64	43	58	55	42	56	55	43	56	55	48	58	51	55	51	52	61	45	47	61

Рис. 5.4: $100 \times AUC$. Результат применения классификатора «абсолютно здоровые – страдающие заболеванием А» к случаю «страдающие заболеванием Б – страдающие заболеванием А». Розовый — минимальное значение, белый — максимальное.



признаков, то логично попробовать другие критерии. Были проверены еще три критерия — r^{min}, r^{ave} и r^{ln}. Пусть $B = \|b_{ic}\|$ — матрица $N \times L$ принадлежности к классу, где

$$b_{ic} = \begin{cases} 1, \text{ если } \mathbf{x}_i \text{ принадлежит классу } c \\ 0, \text{ иначе} \end{cases}$$

Для r^{min} и r^{ave}:

$$F_{ck} = \frac{\sum_{i=1}^{N} b_{ic} x_{ik}}{\sum_{i=1}^{N} b_{ic}}$$

$$D_{c_1 c_2 k} = |F_{c_1 k} - F_{c_2 k}|$$

$$R^{min}_{ck} = \min_{c' \neq c} D_{cc'k}$$

$$R^{ave}_{ck} = \frac{1}{L-2} \sum_{c' \neq c, c' \neq 1} |D_{cc'k}|,$$

где $c = 1$ – номер класса абсолютно здоровых. Он не учитывается, так как этот класс сильно отличается от всех остальных.

Для r^{ln}:

$$D_{c_1 c_2 k} = |\ln F_{c_1 k} - \ln F_{c_2 k}|$$

$$R^{\ln}_{ck} = \min_{c' \neq c} D_{cc'k}$$

Таким образом вычисляется вектор информативности признаков R^v_c, $v \in \{min, ave, ln\}$ для класса c.

Для двухклассового байесовского классификатора нужна информативность признака для пары классов:

$$r^v_{c_1 c_2 k} = \max\{R^v_{c_1 k}; R^v_{c_2 k}\},\ v \in \{min, ave, ln\}.$$

Результаты экспериментов с критериями $\psi = r^v$, $v \in \{min, ave, ln\}$ представлены в главе «Вычислительные эксперименты». В них и в экспериментах с комбинацией двух подходов критерий информативности, в отличие от вектора весов, вычислен по всей выборке до начала кросс-валидации.

Комбинация двух подходов. Сочетание критерия информативности $\psi = r^{min}$ со стратегией обучения «один против всех» дает примерно такой же результат, что и сочетание этой же стратегии и базового критерия информативности (результат на рис. 5.6).

Так как при обучении по стратегии «один против всех» отбирается слишком мало признаков, была проверена еще одна стратегия – «усредненный один против всех», описанная в Алгоритме 5.0.3. Она так же, как и «один против всех», отбирает один эталон для каждого класса (а не для пары классов), только он получается как результат максимизации среднего арифмитического значений AUC по парам, составленным с каждым из других классов. Однако, такая стратегия приводит к отбору почти всех признаков, как видно на рис. 5.7.

	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
AUC на обучении	0,90	0,71	0,69	0,74	0,69	0,71	0,69	0,74	0,69	0,69	0,72	0,73	0,75	0,76	0,70	0,75	0,76	0,75	0,68	0,61	0,72	0,76	0,70	0,74	0,70	0,72	0,72	0,71	0,64	0,73	0,75	0,74
K	49	1	1	1	1	1	1	1	1	1	1	1	1	1	1	97	1	1	1	213	1	1	1	1	1	1	1	1	28	1	1	1

Рис. 5.6: Результат обучения по стратегии «один против всех». Критерий информативности $\psi = r^{min}$. K – мощность найденного эталона.

	ЗД	ВД	ГБ	ЖК	ИБ	МК	ММ	СД	УЩ	ХГ	ХХ	А	АП	АХ	ЯБ	ГБК	ГДЭ	ДЖЭ	РОЭ	Cr	БК	ГПЗ	МП	ПГ	ПЖ	ПК	ПС	ПУ	РМ	ХБ	ЭМ	ЯЖ
AUC на обучении	0,93	0,73	0,64	0,68	0,67	0,65	0,66	0,67	0,64	0,65	0,65	0,68	0,68	0,68	0,65	0,81	0,72	0,67	0,71	0,70	0,65	0,66	0,67	0,72	0,69	0,68	0,66	0,68	0,69	0,65	0,67	0,70
K	49	215	216	202	212	214	214	216	215	216	215	216	107	201	210	154	216	216	208	213	209	210	216	123	194	158	202	212	114	215	206	69

Рис. 5.7: Результаты на обучении по стратегии «усредненный один против всех». Критерий информативности признаков $\psi = r^{min}$. K – мощность найденного эталона.

Алгоритм 5.0.3. Стратегия обучения «усредненный один против всех».

Вход: обучающие выборки X, $c = \overline{1, L}$ для всех классов;

Выход: вектор усредненных мощностей эталонов $\mathbf{K}$;

1: **для всех** $c = 1 \dots L$
2: **для всех** $l = 1 \dots M$
3: **для всех** $c' = 1 \dots L$, $c' \neq c$
4: вычислить для двухклассовой задачи «класс c – класс c'» усредненный по кросс-валидации $AUC_{cc'l}$ на обучении;
5: усреднить $AUC_{cc'l}$ по болезням: $AUC_{cl} = \frac{\sum_{c'} AUC_{cc'l}}{(L-1)}$
6: найти максимум AUC_{cl} по количесву признаков: $AUC_c = \max_l AUC_{cl}$, $K_c = \arg\max_l AUC_{cl}$
7: **вернуть** $\mathbf{K} = (K_1, ..., K_L)$;

Другой порог бинаризации признаков. Плохое качество классификации может быть связано с тем, что признаки зашумлены. Чтобы убрать общую для всех классов компоненту и сделать признаковое описание классов более различным, можно применить в качестве порога бинаризации среднее значение признака. $\|n_{ik}\|_{N \times M}$ – небинаризованная матрица объект-признак без объектов, принадлежащих классу «здоровье», так как этот класс сильно отличается от всех остальных. Бинаризация проходит следующим образом:

$$x_{ik} = \begin{cases} 1, & n_{ik} \geq \frac{1}{N} \sum_{i=1}^{N} n_{ik}; \\ 0, & \text{иначе.} \end{cases}$$

Результат эксперимента приведен в главе «Вычислительные эксперименты».

Глава 6

Неустойчивость весов признаков

При скользящем контроле по Q блокам веса признаков вычисляются Q раз, каждый раз на другой выборке, состоящей из $Q-1$ блоков. То есть для двухклассовой задачи вычисляется Q векторов $\mathbf{w}_q$, $q = \overline{1,Q}$, см. Алгоритм 2.0.2. Оказалось, что значения весов сильно зависит от выборки.

Среднее значение вектора весов:

$$\overline{\mathbf{w}} = \frac{1}{Q}\sum_{q=1}^{Q}\mathbf{w}_q$$

Среднеквадратичное отклонение:

$$\sigma_k = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}(w_{qk} - \overline{w}_{qk})^2}$$

Если $|\frac{\sigma_k}{\overline{w}_{qk}}| > 0.1$, то вес признака считается неустойчивым. Как видно на рис. 6.2, признаков с неустойчивыми весами довольно много. На рис. **??** показана зависимость среднеквадратичного отклонения от среднего значения для бинарных признаков для болезни ХГ. Такой вид зависимости характерен для всех классов. Характерным является то, что точки, каждая из которых соответствует одному признаку, равномерно расположены на кривой и нет четкого разделения на устойчивые и неустойчивые признаки. Форма кривой определятеся тем, что бинарные признаки имеют распределение Бернулли, а у него дисперсия является квадратичной функцией от математического ожидания. Как проводилась оценка математического ожидания и квадратного корня из дисперсии, описано в Алгоритме 6.0.4.

Устойчивость весов можно использовать при отборе признаков, исключая признаки с неустойчивыми весами. Это было сделано двумя способами — мягким и жестким. Мягкий способ заключается в том, что информативность признака с неустойчивым весом считается равной нулю, но сам признак в отборе участвует. При жестком учете устойчивости признаки с неустойчивыми весами не участвуют в отборе признаков и в эталон не включаются никогда. Результаты соответсвующих экспериментов представлены в главе «Вычислительные эксперименты»

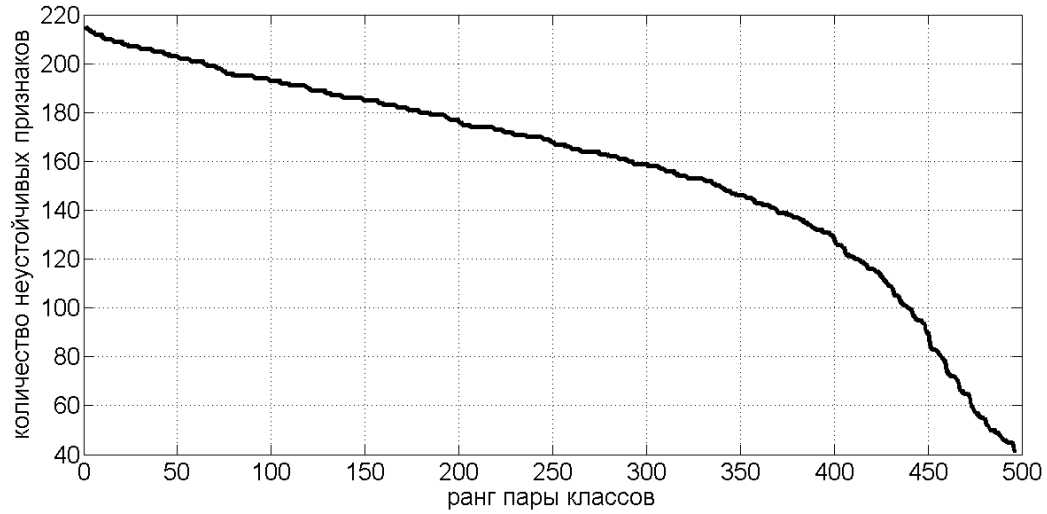


Рис. 6.1: Количество признаков с неустойчивыми весами, то есть такими весами, что $|\frac{\sigma_k}{w_{qk}}| > 0.1$. Горизонтальная ось — ранг пары классов, вертикальная ось — количество признаков.

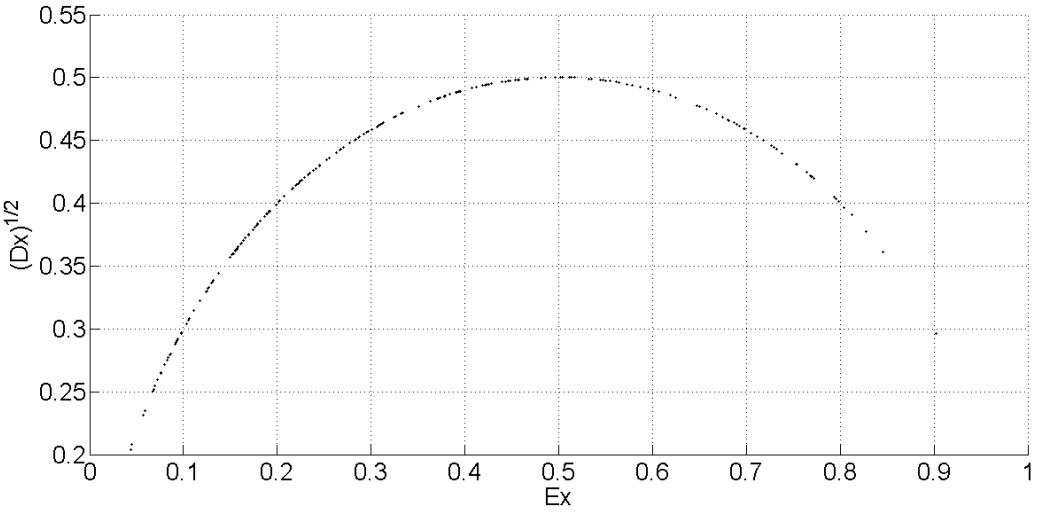


Рис. 6.2: Характерный вид зависимости среднеквадратичного отклонения от среднего значения для бинарного признака (бинаризация по порогу 2). Горизонтальная ось — среднее значение, вертикальная ось — среднеквадратичное отклонение. Каждая точка соответствует одному признаку. Как видно на графике, точки равномерно расположены на кривой и четкого разделения на устойчивые и неустойчивые нет.

Алгоритм 6.0.4. Оценка матожидания и квадратного корня из дисперсии признаков для одного класса.

Вход: обучающая выборка X для одного класса, количество блоков Q;

Выход: вектор оценок матожидания $\overline{\mathbf{x}}$ и квадратного корня из дисперсии σ для всех признаков;

для всех $q = 1 \dots Q$
  разбить выборку на Q примерно равных частей: $X = \bigcup_{q=1}^{Q} X_q$;
  для всех $t = 1 \dots 100$
    случайным образом выбрать из X_q подвыборку X_q^t, содержащую примерно 70% от X_q;
    вычислить по X_q^t вектора средних значений $\overline{\mathbf{x}}_q^t$ и среднеквадратичных отклонений σ_q^t;
  $\sigma_q = \frac{1}{100} \sum_{t=1}^{100} \sigma_q^t$;
  $\overline{\mathbf{x}}_q = \frac{1}{100} \sum_{t=1}^{100} \overline{\mathbf{x}}_q^t$;
$\overline{\mathbf{x}} = \frac{1}{Q} \sum_{q=1}^{Q} \overline{\mathbf{x}}_q$;
$\sigma = \frac{1}{Q} \sum_{q=1}^{Q} \sigma_q$;
вернуть $\overline{\mathbf{x}}$, σ;

Глава 7

Вычислительные эксперименты

Далее по вертикальным осям графиков часто будет откладываться AUC для c-ого класса, усредненный по всем двухклассовым задачам с участием c-ого класса:

$$\overline{AUC}_c = \frac{1}{L-1}\sum_{c} AUC_{cc'}, \text{ где } c' = \overline{1,L},\ c' \neq c, \tag{7.1}$$

$AUC_{cc'}$ – значение AUC для двухклассовой задачи с c-м и c'-м классами. **Вектор весов признаков во время кросс-валидации вычисляется отдельно на каждой обучающей выборке, в точности как в Алгоритме 2.0.2.** На рис. 7.1, 7.2 и 7.3 представлены результаты применения критериев информативности $\psi = r^v$, $v \in \{min, ave, ln\}$ соответственно по сравнению с базовым критерием $\psi = |w|$. Во всех случаях порог бинаризации равен 2.

В таблице 7.1 для каждого эксперимента приведены 5 классов, для которых прирост AUC оказался наибольшим. На рис. 7.7 показаны все результаты для случая, когда веса признаков вычислялись во время кросс-валидации, в одной системе координат, а также показан общий тренд.

Вектор весов признаков вычисляется по всей выборке до начала кросс-валидации. В остальном все как в Алгоритме 2.0.2. Далее во всех случаях используется критерий информативности $\psi = |w|$.

На рис. 7.4 и 7.5представлены результаты мягкого и жесткого учета устойчивости весов признаков при отборе в сравнении со случаем, когда устойчивость веса не учитывается вообще (порог бинаризации везде равен 2). При мягком учете информативность признака с неустойчивым весом считается равной 0, но признак участвует в отборе и может попасть в эталон. При жестком учете признак с неустойчивым весом в эталон попасть не может.

На рис. 7.6 представлен результат применения бинаризации признаков по порогу, равному среднему значению признака, в сравнении с бинаризацией по порогу 2 (устойчивость весов признаков в обоих случаях не учитывается).

В таблице 7.2 для каждого эксперимента приведены 5 классов, для которых прирост AUC оказался наибольшим. На рис. 7.8 показаны все результаты для случая, когда веса признаков вычислялись до кросс-валидации, в одной системе координат, а

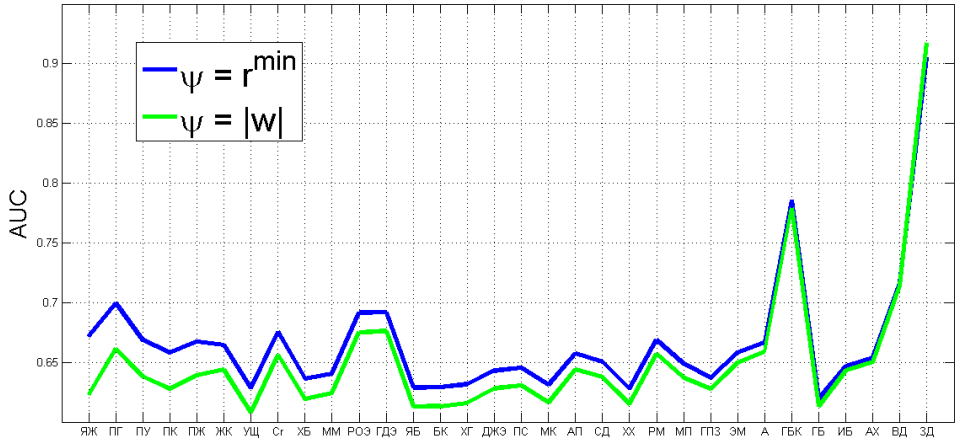


Рис. 7.1: Сравнение критериев информативности $\psi = r^{min}$ и $\psi = |w|$. Вертикальная ось — AUC для c-ого класса, усредненный по всем двухклассовым задачам с участием c-ого класса.

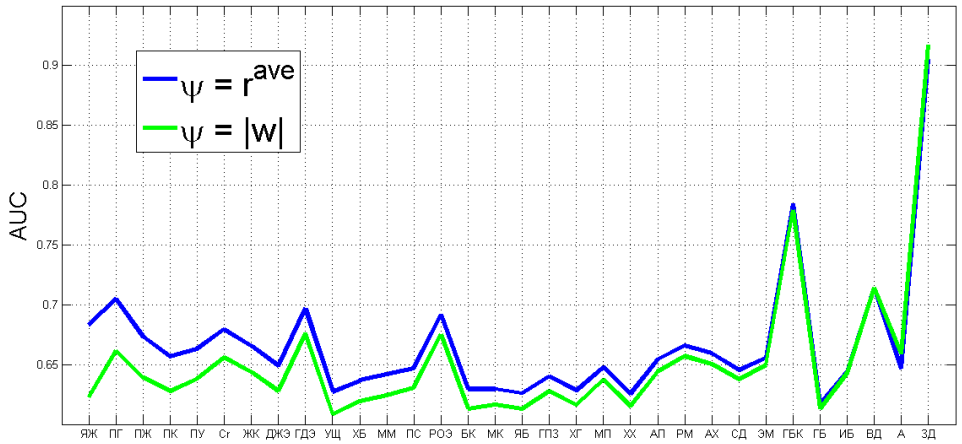


Рис. 7.2: Сравнение критериев информативности $\psi = r^{ave}$ и $\psi = |w|$. Вертикальная ось — AUC для c-ого класса, усредненный по всем двухклассовым задачам с участием c-ого класса.

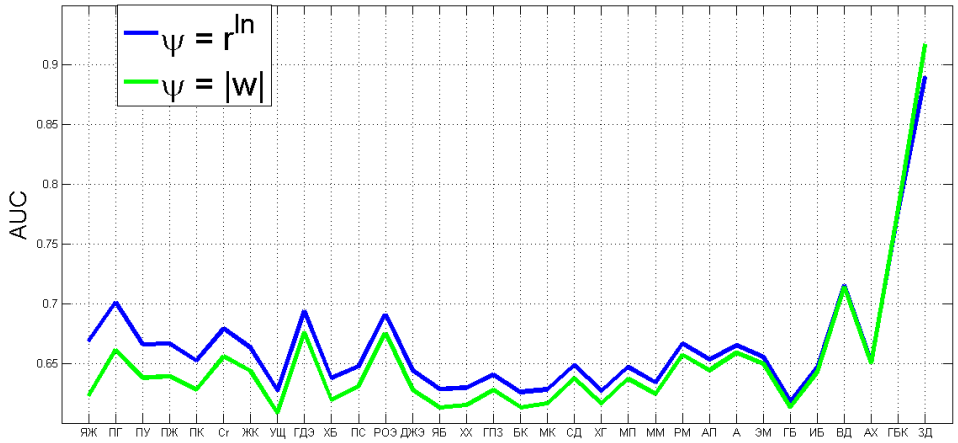


Рис. 7.3: Сравнение критериев информативности $\psi = r^{ln}$ и $\psi = |w|$. Вертикальная ось — AUC для c-ого класса, усредненный по всем двухклассовым задачам с участием c-ого класса.

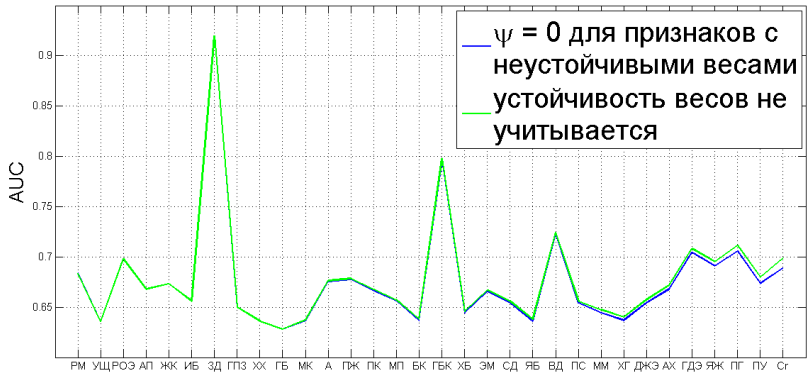


Рис. 7.4: Сравнение случая мягкого учета устойчивости весов признаков при отборе и случая, когда устойчивость веса не учитывается вообще. Вертикальная ось — AUC для c-ого класса, усредненный по всем двухклассовым задачам с участием c-ого класса.

Рис. 7.5: Сравнение случая жесткого учета устойчивости весов признаков при отборе и случая, когда устойчивость веса не учитывается вообще. Вертикальная ось — AUC для c-ого класса, усредненный по всем двухклассовым задачам с участием c-ого класса.

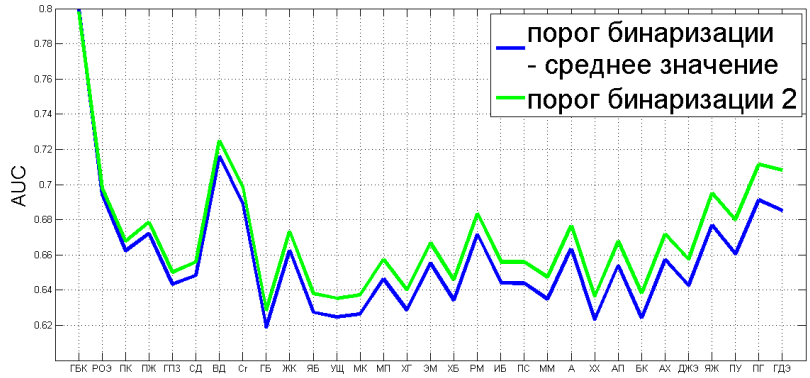


Рис. 7.6: Сравнение бинаризации по порогу, равному среднему значению признака, и по порогу 2. Вертикальная ось — AUC для c-ого класса, усредненный по всем двухклассовым задачам с участием c-ого класса.

также показан общий тренд. Как видно из графиков и из таблиц, мягкий учет устойчивости веса признака дает лучший результат, чем жесткий. Отсюда можно сделать вывод, что признаки с неустойчивыми весами могут быть информативными.

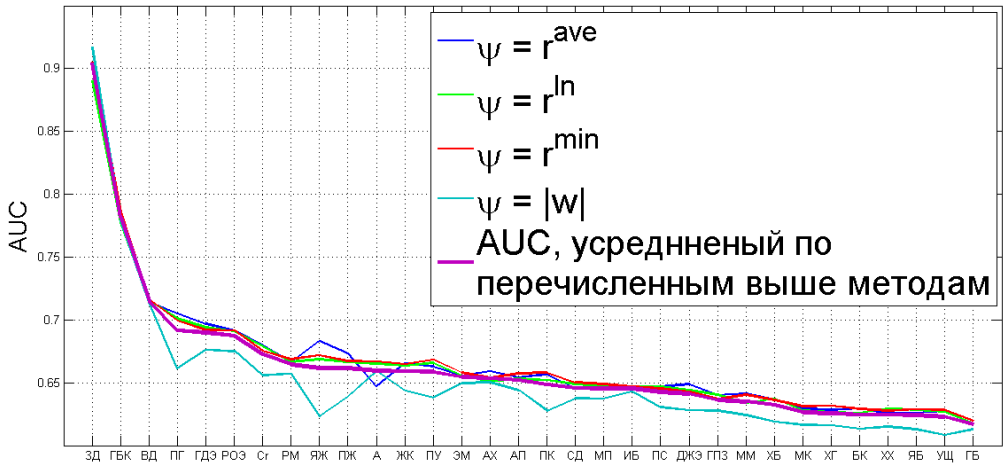


Рис. 7.7: Общий тренд разных методов. Веса признаков вычислялись во время кросс-валидации. Вертикальная ось — AUC для c-ого класса, усредненный по всем двухклассовым задачам с участием c-ого класса.

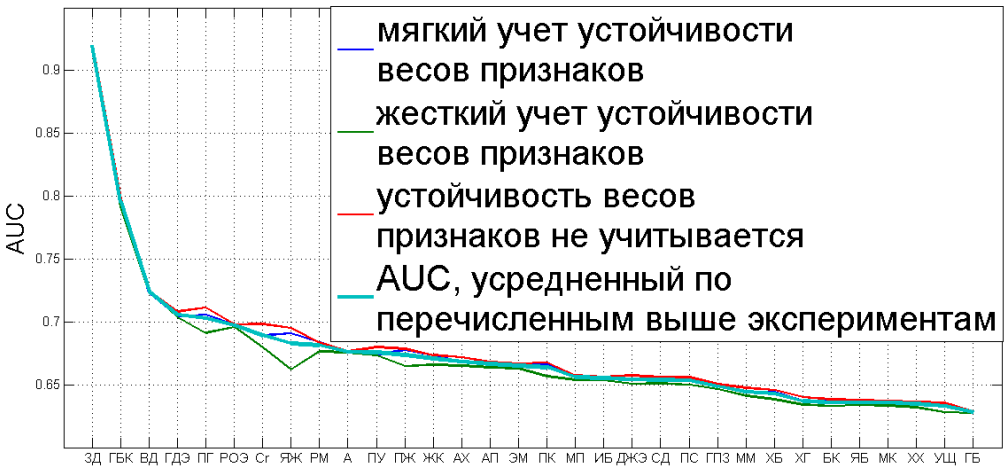


Рис. 7.8: Общий тренд разных методов. Веса признаков вычислялись до кросс-валидации. Вертикальная ось — AUC для c-ого класса, усредненный по всем двухклассовым задачам с участием c-ого класса.

Итог. В таблице 7.3 приведен средний по парам классов (без учета пар, в которых есть класс «абсолютное здоровье») AUC для всех экспериментов, описанных в этой главе.

Таблица 7.1: Веса признаков вычислялись во время кросс-валидации. Для каждого из трех экспериментов: критерий информативности и 5 классов, на которых достигается наибольший прирост AUC (в скобках значение прироста).

$\psi = r^{min}$	ЯЖ(+0.05)	ПГ(+0.04)	ПУ(+0.03)	ПК(+0.03)	ПЖ(+0.03)
$\psi = r^{ave}$	ЯЖ(+0.06)	ПГ(+0.04)	ПЖ(+0.03)	ПК(+0.03)	ПУ(+0.03)
$\psi = r^{ln}$	ЯЖ(+0.05)	ПГ(+0.04)	ПУ(+0.03)	ПЖ(+0.03)	ПК(+0.02)

Таблица 7.2: Веса признаков вычислялись до кросс-валидации. Для каждого из трех экспериментов: номер и 5 классов, на которых достигается наибольший прирост AUC (в скобках значение прироста). Нумерация экспериментов: 1 – мягкий учет устойчивости, 2 – жесткий учет устойчивости, 3 – бинаризация по порогу, равному среднему значению.

1	РМ($+4.5 \cdot 10^{-4}$)	УЩ($+3.2 \cdot 10^{-5}$)	РОЭ($+3.5 \cdot 10^{-8}$)	АП($-5.2 \cdot 10^{-5}$)	ЖК($-6.2 \cdot 10^{-5}$)
2	ЗД($-1.7 \cdot 10^{-4}$)	А($-8.6 \cdot 10^{-4}$)	ВД($-9.9 \cdot 10^{-4}$)	ГБ(-0.0012)	РОЭ(-0.0023)
3	ГБК($+0.0029$)	РОЭ(-0.0038)	ПК(-0.0056)	ПЖ(-0.0065)	ГПЗ(-0.0070)

Таблица 7.3: Для каждого эксперимента: краткое описание и AUC, средний по парам классов без учета пар, в которых есть класс «абсолютное здоровье».

эксперимент	AUC
базовый: $\psi = \|w\|$	0.63
$\psi = r^{min}$	0.65
$\psi = r^{ave}$	0.65
$\psi = r^{ln}$	0.65
базовый: без учета устойчивости весов, порог бинаризации – 2	0.66
мягкий учет устойчивости весов	0.66
жесткий учет устойчивости весов	0.65
бинаризация по порогу – среднему значению признака	0.66

Заключение

В работе исследованы различные методы отбора признаков, такие как многоклассовые стратегии обучения и критерии информативности признаков. С каждым из исследованных методов проведен вычислительный эксперимент на реальных данных. По результатам экспериментов видно, что изменение стратегии обучения приводит к крайним случаям: либо отбираются почти все признаки, либо всего один-два из 216. Применение различных критериев информативности значительно на качество классификации не влияет. В данной задаче качество классификации больше зависит от класса, чем от метода.

Литература

[1] *Sebastiani, Fabrizio* Machine learning in automated text categorization. // ACM Computing Surveys, 2002, vol. 34, pp. 1-47.

[2] *Успенский В. М.* Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов. – М.: Экономика и информатика, 2008. 116 с.

[3] *Успенский В. М.* Информационная функция сердца. *Клиническая медицина.* 2008. Т. 86. № 5. С. 4–13.

[4] *Uspenskiy V. M.* Diagnostic System Based on the Information Analysis of Electrocardiogram. In: *Proceedings of MECO 2012. Advances and Challenges in Embedded Computing.* Bar, Montenegro, June 19–21, 2012. Pp. 74–76.

[5] *Успенский В. М., Кравченко Ю. Г., Павловский К. П., Авербах Ю. И.* Устройство экспресс-диагностики заболеваний внутренних органов и онкопатологии. Патент на изобретение № 2159574 от 27 ноября 2000 г.

[6] *Успенский В. М.* Способ диагностики болезней неинфекционной этиологии. Патент на изобретение № 2157093 от 10 октября 2000 г.

[7] *Успенский В. М.* Способ диагностики заболеваний внутренних органов неинфекционной природы на любой стадии их развития. Патент на изобретение № 2163088 от 20 февраля 2001 г.

[8] *Успенский В. М.* Способ суточного кардиомониторирования для определения наличия и активности заболеваний человека неинфекционной природы. Патент на изобретение № 2211658 от 10 сентября 2003 г.

[9] *Успенский В. М.* Способ диагностики заболеваний внутренних органов. Патент на изобретение № 2407431 от 27 декабря 2010 г.

[10] *Успенский В. М., Воронцов К. В., Целых В. Р.* Статистические обоснования информационного анализа электрокардиосигналов для диагностики заболеваний внутренних органов. // Математическая биология и информатика. – 2014.

[11] *Целых В. Р.* Статистические обоснования информационного анализа электрокардиосигналов для диагностики заболеваний внутренних органов. – 2015. http://www.machinelearning.ru/wiki/images/8/8d/Tselykh2015Diploma.pdf

[12] *Цыганова С. В.* Применение тематической модели классификации в информационном анализе электрокардиосигналов. – 2015.

[13] *Uspenskiy V. M.* Information Function of the Heart. Biophysical substantiation of technical requirements for electrocardioblock registration and measurement of electrocardiosignals parameters acceptable for information analysis to diagnose internal diseases. In: *Joint International IMEKO TC1+TC7+TC13 Symposium.* August 31–September 2, 2011, Jena, Germany.

[14] *Uspenskiy V. M.* Information Function of the Heart. A Measurement Model. In: *Measurement 2011: 8-th International Conference.* Smolenice, Slovakia, April 27–30, 2011. Pp. 383–386.