Отбор признаков для многоклассовой классификации символьных последовательсностей

Вдовина Е. А.

Московский физико-технический институт

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

2016

Технология информационного анализа ЭКГ

Этапы предварительной обработки ЭКГ-сигнала:

1. *Демодуляция* — вычисление амплитуд, интервалов и углов по кардиограмме
2. *Дискретизация* — перевод в *кодограмму* — символьную строку в 6-буквенном алфавите

Постановка задачи

Задача диагностики заболеваний внутренних органов по ЭКГ.

1. объект — кодограмма
2. класс — заболевание (или абсолютное здоровье)
3. классы могут пересекаться
4. если объект относится к классу «абсолютное здоровье», то он не может принадлежать ни одному другому классу

Признаковое описание объекта — вектор частот триграмм $(n_1, ..., n_M), M = 216$. Признаки бинаризуются:

$$x_k = \begin{cases} 1, & n_k \geq 2 \\ 0, & \text{иначе} \end{cases}$$

Кодограмме сопоставлен бинарный вектор принадлежности классу.

Условные обозначения заболеваний

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 зоб щитовидный железы	УЩ	9
хронический гастрит 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

Условные обозначения заболеваний (продолжение)

хронический гастрит гиперацидный	ГДЭ	17
дискинезия желчевыводящих путей	ДЖЭ	18
рак общий (различной локализации)	РОЭ	19
рак (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

Суть проблемы

Проблема: качество классификации в двухклассовой задаче «здоровые против больных» значительно лучше, чем в задаче «больные заболеванием А против больных заболеванием Б»

Цель работы

Проверить гипотезу, что с помощью модификации отбора признаков можно улучшить качество дифференциальной диагностики.

Базовый алгоритм двухклассовой классификации

Принцип максимума апостериорной вероятности для 2 классов:

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

Если признаки независимые и бинарные, то МП-оценка:

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

Тогда наивный байесовский классификатор является линейным:

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

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

Отбор признаков в базовом алгоритме

Отбираются K признаков с наибольшими значениями информативности ψ.
В базовом алгоритме $\psi_k = |w_k|$.

Эталон — множество информативных признаков.
Диагностический эталон — множество информативных признаков для отдельной болезни.

Критерий качества классификации

AUC — доля правильно упорядоченных пар $(\mathbf{x}_i, \mathbf{x}_j)$, обобщение на случай пересекающихся классов:

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

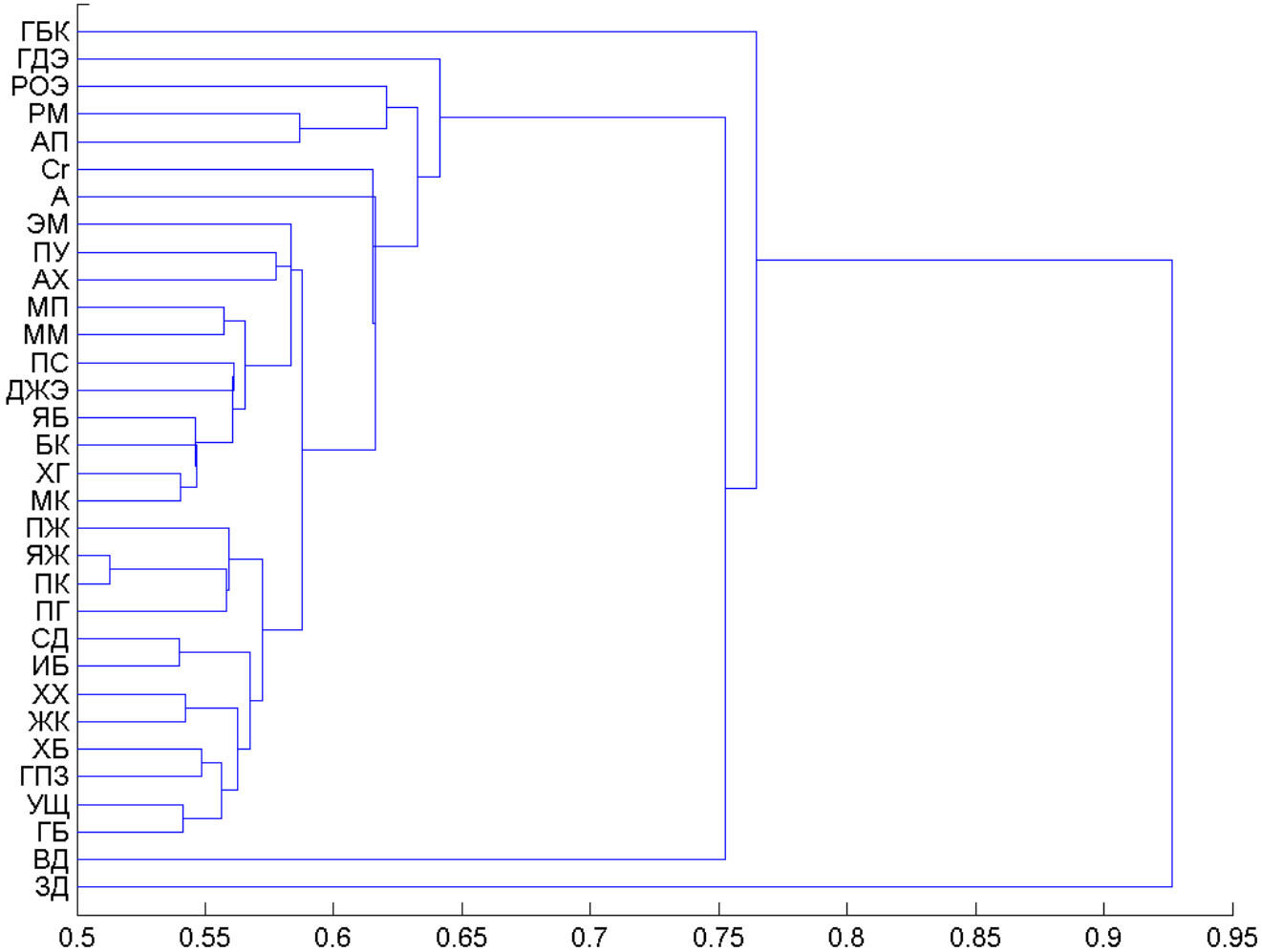
где

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

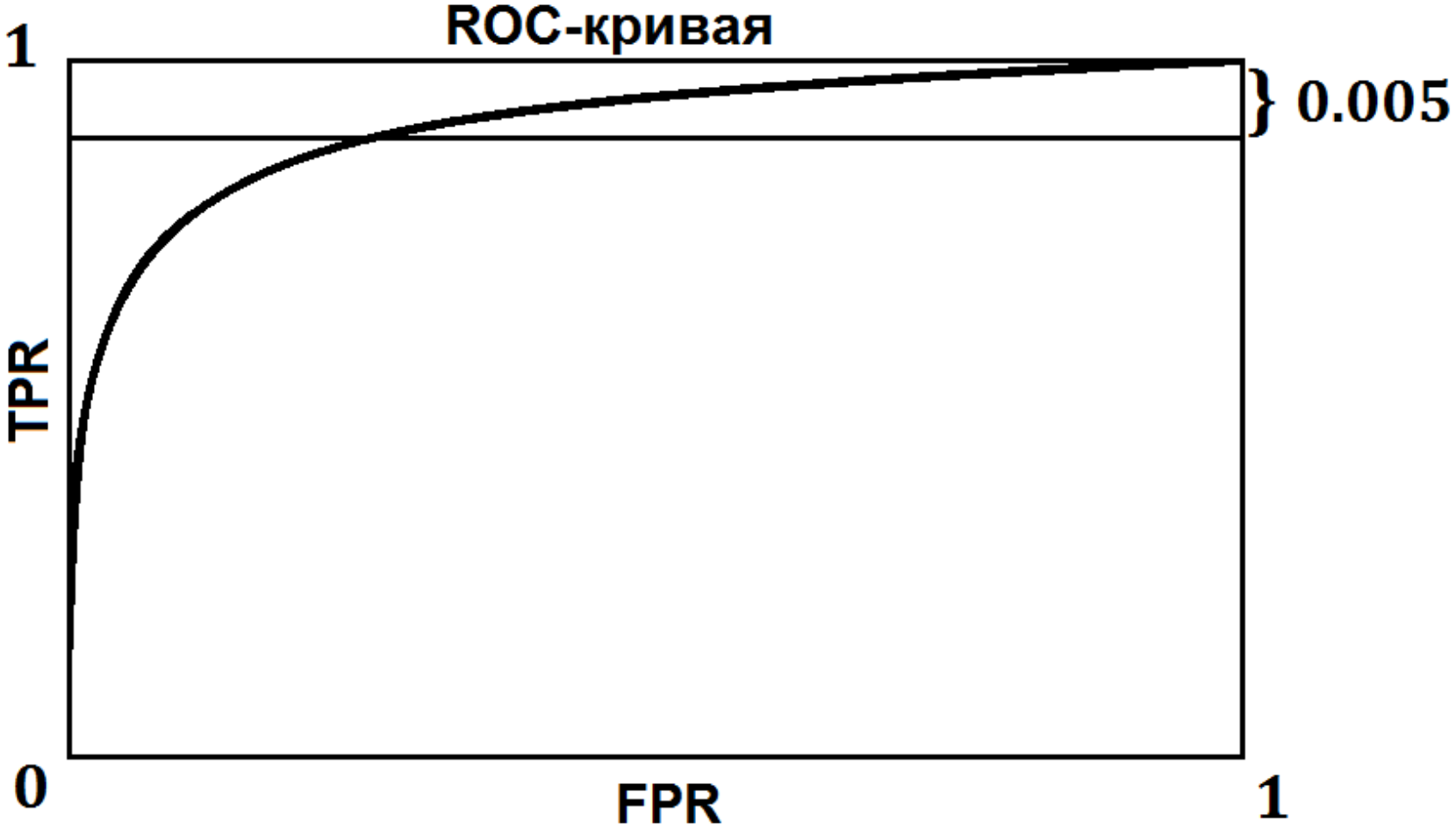
Настройка параметров проводилась на обучающей выборке по результатам $1 \times Q$ кросс-валидации, $Q = 10$.

Кластеризация болезней

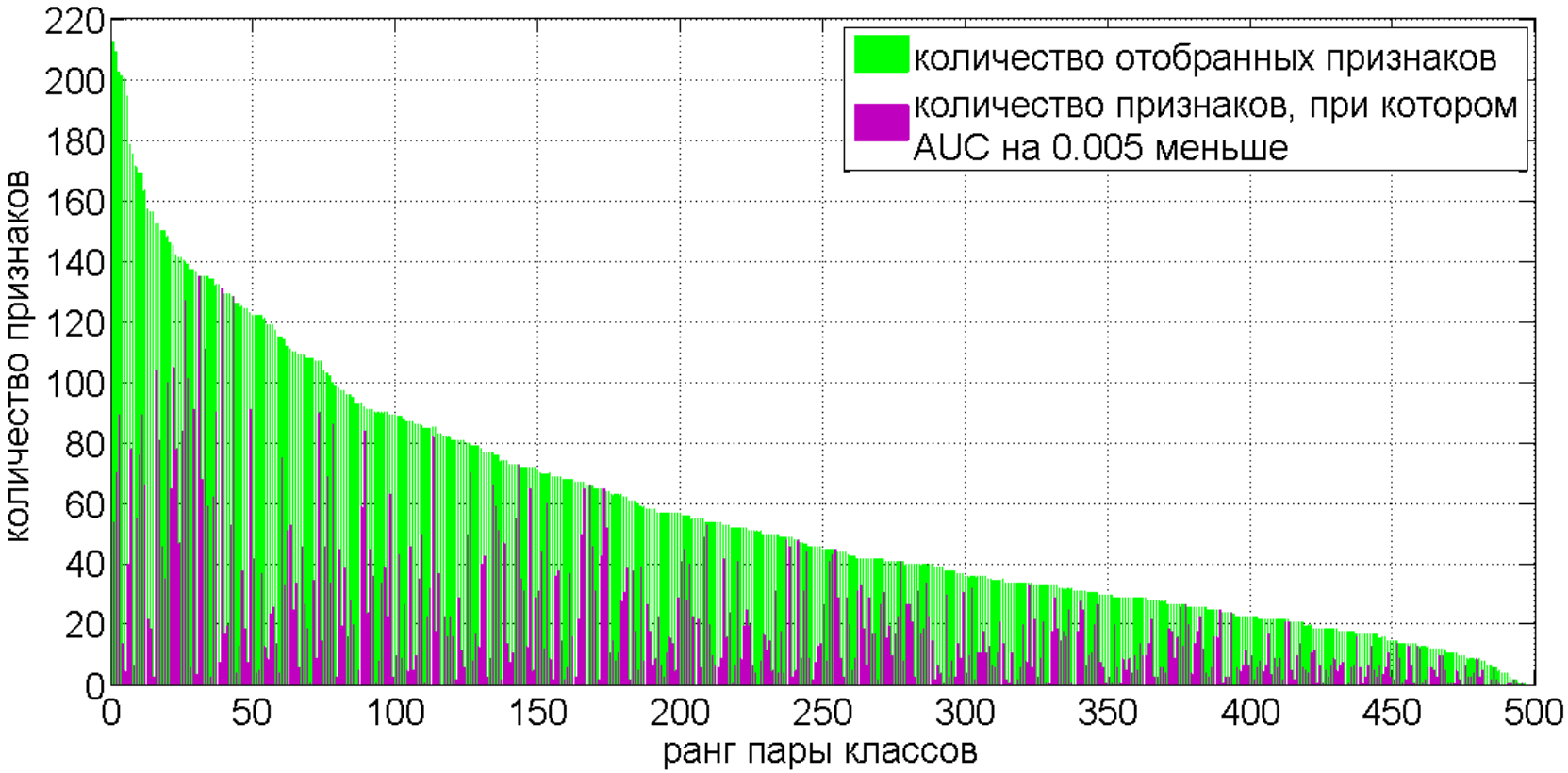
По горизонтальной оси — AUC, расстояние между кластерами.



Зависимость AUC от количества признаков



Зависимость AUC от количества признаков



Существует ли диагностический эталон?

Диагностический эталон должен отличать болезнь от всех остальных. Тривиальное решение — вырожденное:

- пересечение эталонов, полученных для c-ой болезни при решении всевозможных двухклассовых задач – пустое

	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

- стратегия обучения «каждый против всех» – отбирается 1 признак

	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

Другие критерии информативности признаков

$c = 1$ – номер класса «абсолютное здоровье»; $\overline{x}_k(c)$ – среднее значение признака для класса c. Для r^{min} и r^{ave}:

$$D_k(c_1, c_2) = |\overline{x}_k(c_1) - \overline{x}_k(c_2)|$$

$$R_k^{min}(c) = \min_{c' \neq c} D_k(c, c')$$

$$R_k^{ave}(c) = \frac{1}{L-2} \sum_{c' \neq c, c' \neq 1} |D_k(c, c')|$$

Для r^{ln}:

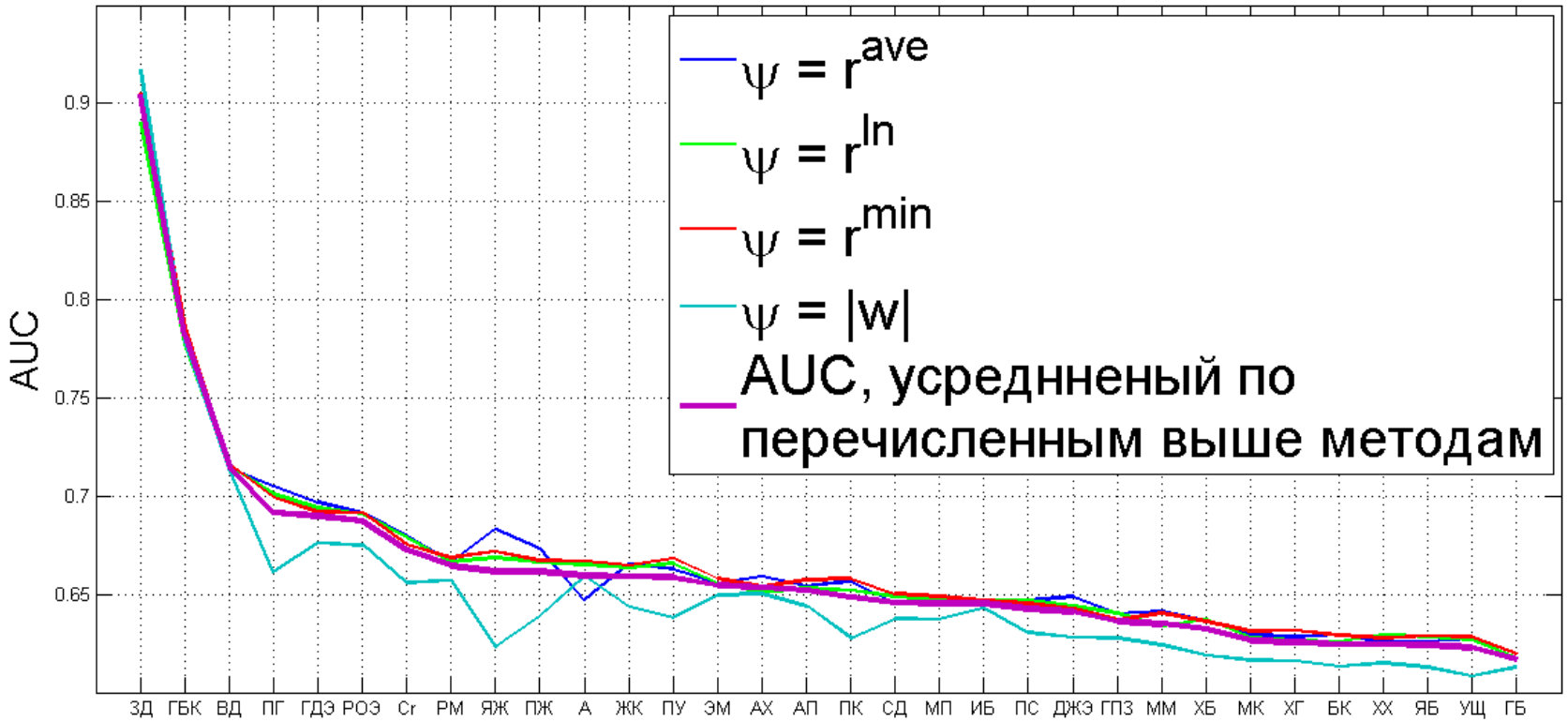
$$D_k(c_1, c_2) = |\ln \overline{x}_k(c_1) - \ln \overline{x}_k(c_2)|$$

$$R_k^{\ln}(c) = \min_{c' \neq c} D_k(c, c')$$

Информативность признака для пары классов:

$$r_k^v(c_1, c_2) = \max_{c \in \{c_1; c_2\}} R_k^v(c),\ v \in \{min,\ ave,\ ln\}$$

Другие критерии информативности признаков



Неустойчивость весов признаков

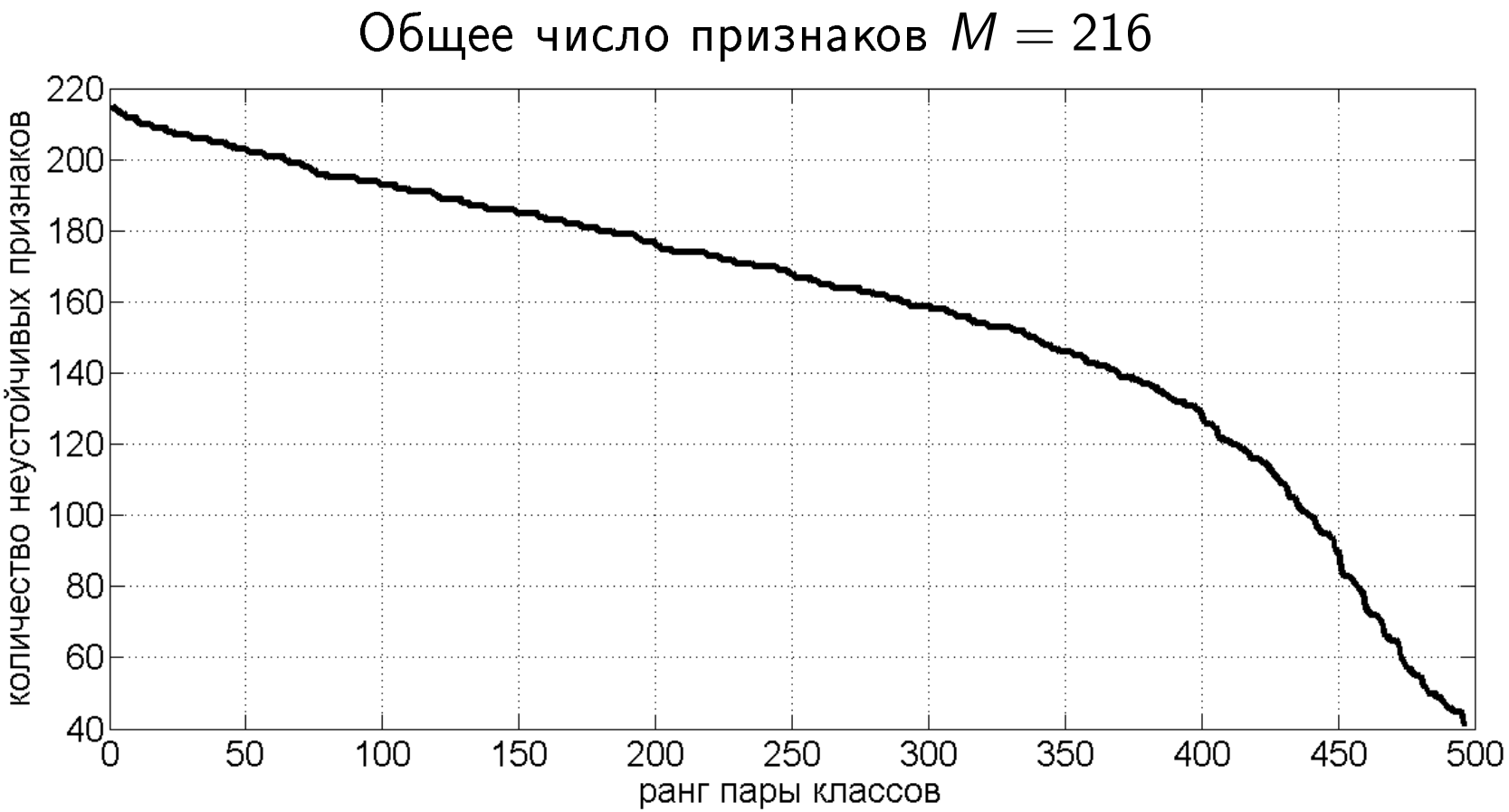
До этого вектор весов вычислялся во время скользящего контроля по Q блокам, то есть Q раз, каждый раз на другой выборке, состоящей из $Q-1$ блоков: $\mathbf{w}_q$, $q = \overline{1,Q}$.

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

Если $|\frac{\sigma_k}{\overline{w}_{qk}}| > 0.1$, то вес признака считается неустойчивым.
ψ – информативность признака.
Учет устойчивости веса признака

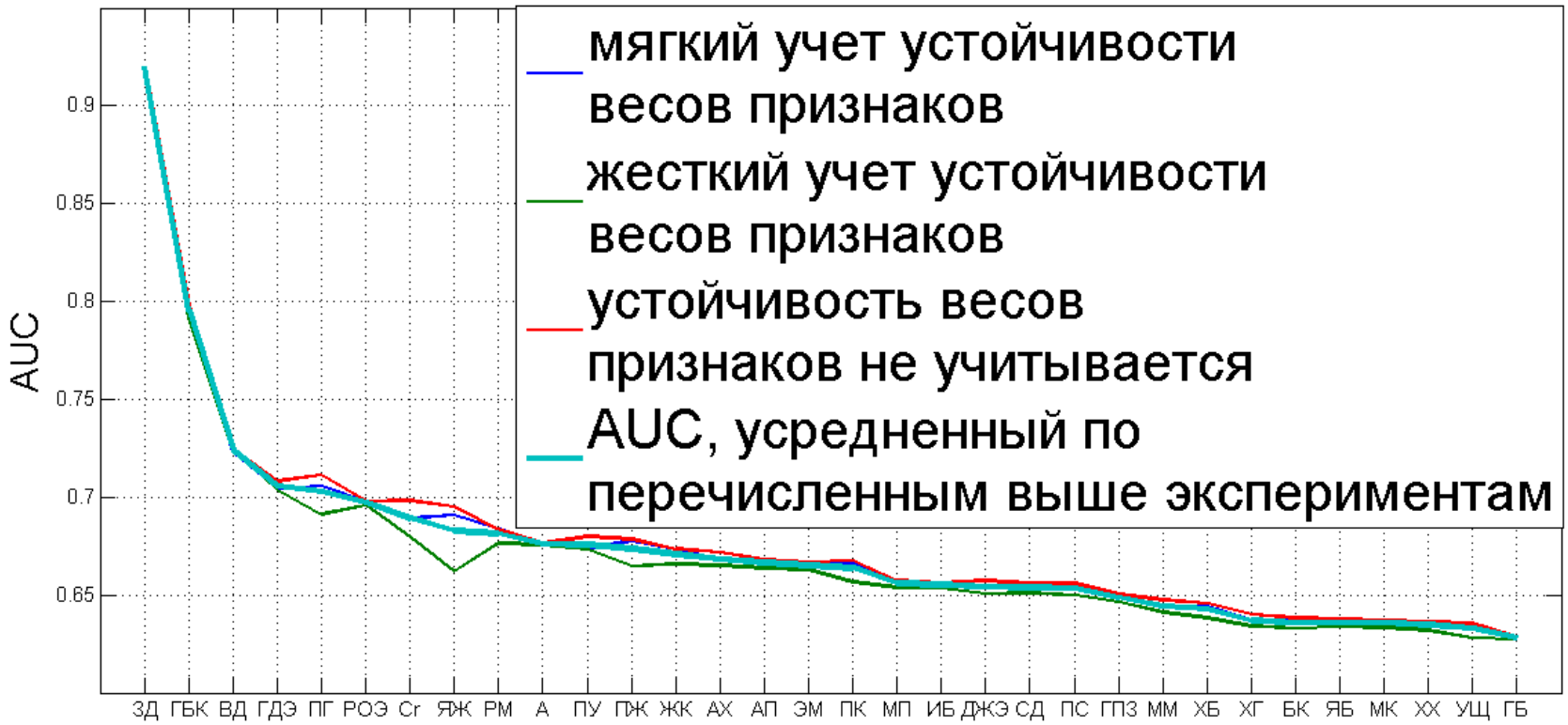
- мягкий: вес неустойчивый $\Rightarrow \psi = 0$, но признак может попасть в эталон;
- жесткий: вес неустойчивый $\Rightarrow$ признак никогда в эталон не попадет.

Неустойчивость весов признаков



Учет устойчивости весов признаков

w вычисляется до кросс-валидации.



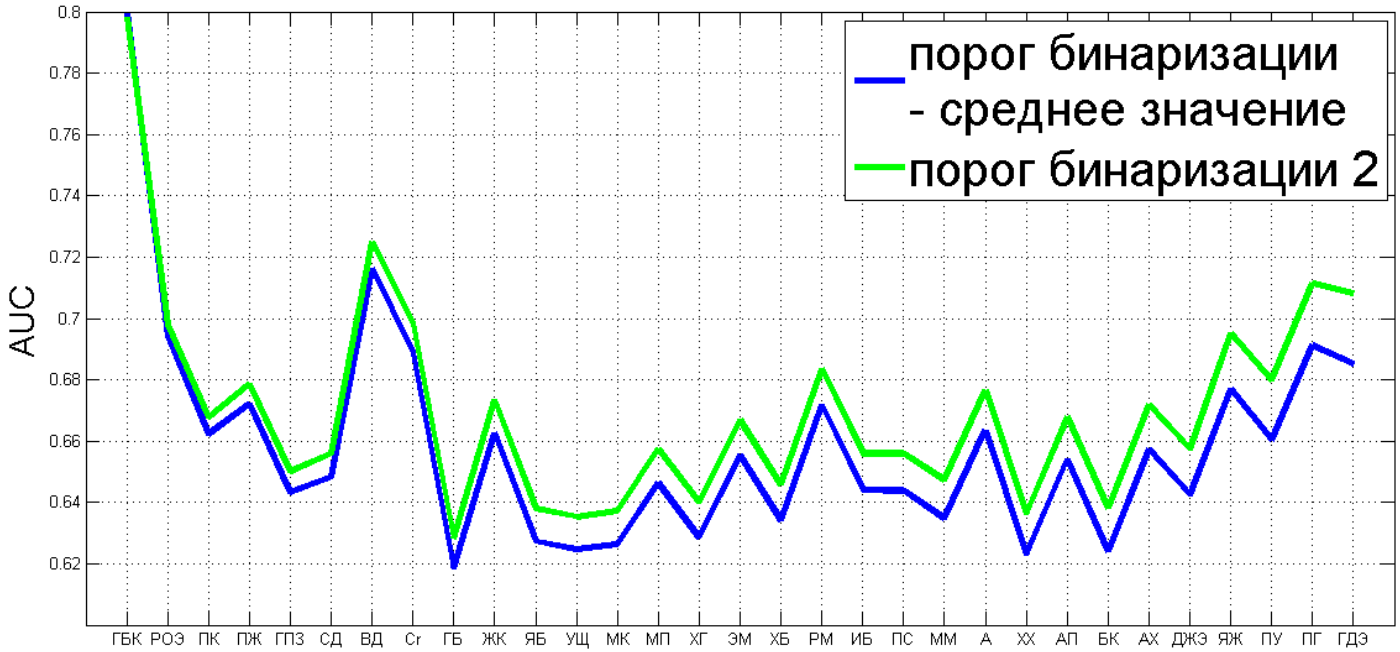
Другой порог бинаризации

Плохое качество классификации может быть связано с тем, что признаки зашумлены.
Чтобы убрать шум на стадии бинаризации признаков:
$\|n_{ik}\|_{N\times M}$ – небинаризованная матрица объект-признак без объектов, принадлежащих классу «абсолютное здоровье»;

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

w вычисляется по всей выборке перед началом кросс-валидации.

Другой порог бинаризации



Сравнение результатов

Для каждого эксперимента: 5 классов, на которых достигается наибольший прирост AUC (в скобках значение прироста).
Веса признаков вычислялись во время кросс-валидации.

$\psi = r^{min}$	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)

Сравнение результатов

Для каждого эксперимента: 5 классов, на которых достигается наибольший прирост AUC (в скобках значение прироста).
Веса признаков вычислялись до кросс-валидации.
Условные обозначения: 1 – мягкий учет устойчивости, 2 – жесткий учет устойчивости, 3 – бинаризация по порогу, равному среднему значению.

1	РМ($+5 \cdot 10^{-4}$)	УЩ($+3 \cdot 10^{-5}$)	РОЭ($+4 \cdot 10^{-8}$)	АП($-5 \cdot 10^{-5}$)	ЖК($-6 \cdot 10^{-5}$)
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)

Сравнение результатов

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

эксперимент	AUC
базовый: $\psi = \lvert w \rvert$	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

- Рассмотрен ряд модификаций отбора признаков.
- Проведено их сравнение друг с другом и с базовым методом.
- Исследованные методы влияют на качество классификации незначительно.
- Качество больше зависит от класса, чем от метода.