Статистические обоснования информационного анализа электрокардиосигналов для диагностики заболеваний внутренних органов

Целых Влада

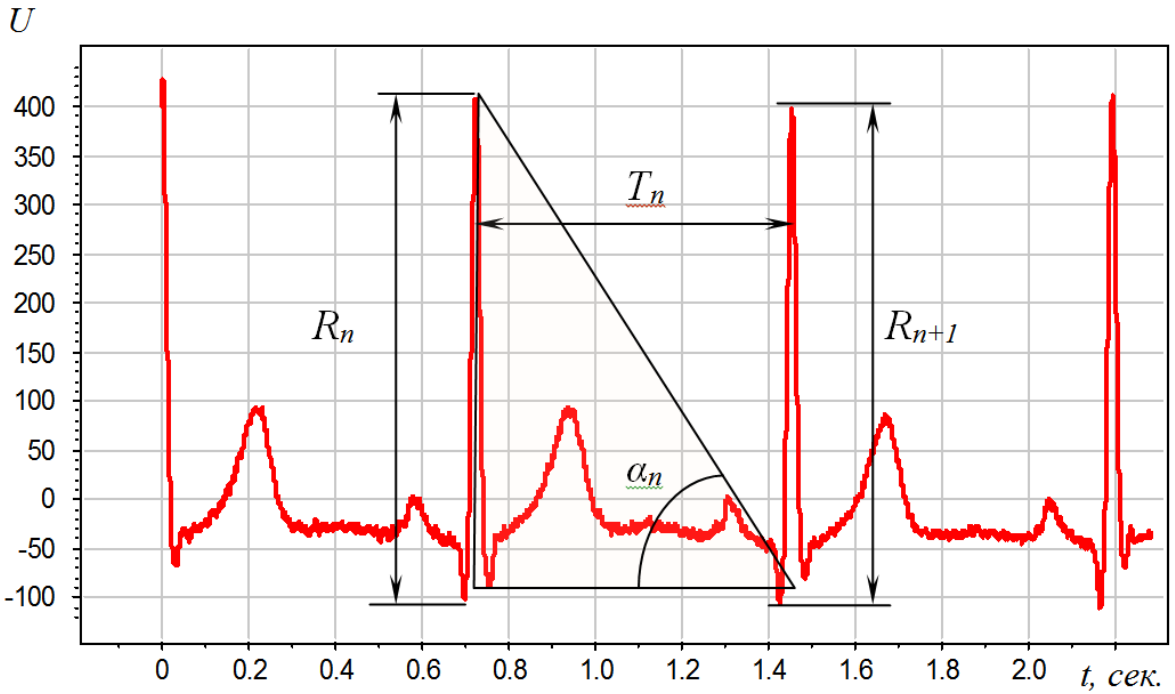
Московский физико-технический институт

Научный руководитель:
ст.н.с. ВЦ РАН, д.ф.-м.н.
Воронцов Константин Вячеславович

24 июня 2015 года

Технология информационного анализа ЭКГ по В.М.Успенскому

Диагностика болезней по знакам приращений амплитуд $R_{n+1} - R_n$, интервалов $T_{n+1} - T_n$ и углов $\alpha_{n+1} - \alpha_n$.



$$\alpha_n = \text{arctg}\,\frac{R_n}{T_n}$$

Технология информационного анализа ЭКГ по В.М.Успенскому

Этапы предварительной обработки ЭКГ-сигнала:

1. *Демодуляция* — вычисление амплитуд, интервалов и углов по кардиограмме длиной 600 кардиоциклов
2. *Дискретизация* — перевод в *кодограмму* — 599-символьную строку в 6-буквенном алфавите
3. *Векторизация* — перевод в вектор 6^k частот k-грамм, $k = 3$

Цель работы

Статистически обосновать отдельные этапы технологии информационного анализа ЭКГ, исследовать возможность диагностики заболеваний внутренних органов по ЭКГ.

Дискретизация ЭКГ-сигнала

Вход: последовательность интервалов и амплитуд $(T_n, R_n)_{n=1}^{N}$

Правила кодирования:

если	$R_n < R_{n+1}$,	$T_n < T_{n+1}$,	$\alpha_n < \alpha_{n+1}$	то	$s_n = \mathtt{A}$
если	$R_n \geqslant R_{n+1}$,	$T_n \geqslant T_{n+1}$,	$\alpha_n < \alpha_{n+1}$	то	$s_n = \mathtt{B}$
если	$R_n < R_{n+1}$,	$T_n \geqslant T_{n+1}$,	$\alpha_n < \alpha_{n+1}$	то	$s_n = \mathtt{C}$
если	$R_n \geqslant R_{n+1}$,	$T_n < T_{n+1}$,	$\alpha_n \geqslant \alpha_{n+1}$	то	$s_n = \mathtt{D}$
если	$R_n < R_{n+1}$,	$T_n < T_{n+1}$,	$\alpha_n \geqslant \alpha_{n+1}$	то	$s_n = \mathtt{E}$
если	$R_n \geqslant R_{n+1}$,	$T_n \geqslant T_{n+1}$,	$\alpha_n \geqslant \alpha_{n+1}$	то	$s_n = \mathtt{F}$

Выход: *кодограмма* $S = (s_n)_{n=1}^{N-1}$ — последовательность символов алфавита $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$:

```
DBFEACFDAAFBABDDAADFAAFFEACFEACFBAEFFAABFFAAFFAAFFAAFFAAEBFAEBFEAAFCAAFFAAD
FCAFFAADFCADFCCDFDACFFACDFAEFFACFFEADFCAFBCADFFECFFAAFFAAFFAAFFAEFFCACFCAEFFCAD
DAADBFAAFFAEBFAABFACDFFAAFBAADFAADFDAAFCECFCEDFCEEFCAEFBECBBBAADBAACFFAAFFA
CFFCECFDAABDAEFFAAFFCEDBFAAFFAEFFAEFBACFBAEDFEAAFFCAFFDAAFFAEBDAADBBADFDAFF
EABFCCAFDEEBDECFFACFFAABFAADFBAAFFACFFFAEFFACFFACFFCECFBAAFFFAAFFFAAFFAADFB
AABFACDFDAEFFAADBAAEFFEAFBCECFDECCFBAAFFAADFDACDFAAFFAADFCAADFAEFBAAFFCADFE
AFFCECFCECFFAAFFABCFDAAAFFADBFCAEFFAABFACBFAAEBFAEBFCAFFBAAFFAAFFDACFDAABFB
CAFFAECFFACFFACDFCADFDAABFAAEDDABBFCACDBAAFFAAFFCADFAADFDACFFAEDFCACFCAEBCE
```

Векторизация кодограммы ЭКГ-сигнала

Вход: *кодограмма* $S = (s_1, \ldots, s_{N-1})$ как текстовая строка

Выход: вектор частот k-грамм $p_w(S)$, $w \in \mathcal{A}^k$ (*частота k-граммы* — отношение числа ее вхождений в кодограмму к общему числу k-грамм в кодограмме).

Обозначения

y_0 — класс здоровых

$y_1, \dots y_M$ — классы больных

X_m — выборка кодограмм класса y_m, $m = 0, 1, \dots M$

Средняя частота k-граммы w в классе y_m:

$$F_w(X_m) = \frac{1}{|X_m|} \sum_{S \in X_m} p_w(S)$$

Средняя встречаемость k-граммы w в классе y_m:

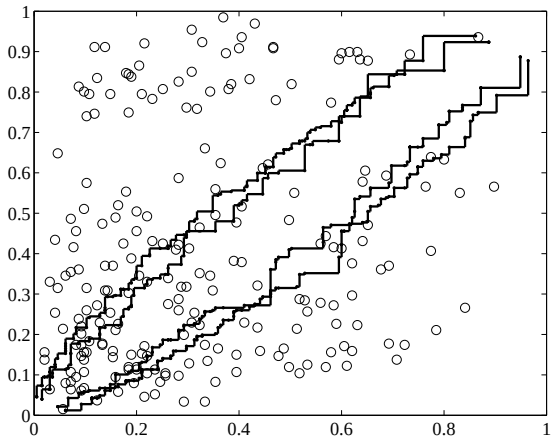
$$B_w(X_m) = \frac{1}{|X_m|} \sum_{S \in X_m} \left[p_w(S) \geqslant \theta \right]$$

Перестановочный тест для поиска информативных триграмм

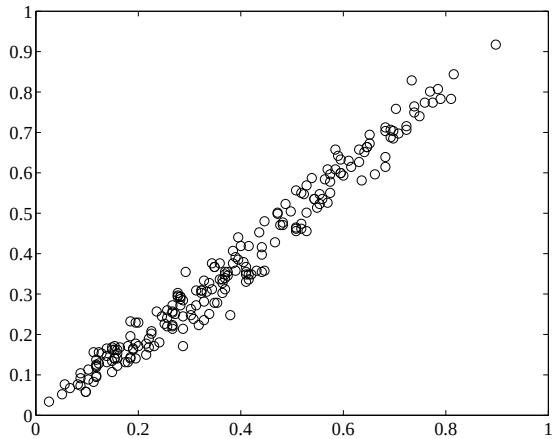
Точки на графиках — это триграммы

ось X: $B_w(X_0)$ — доля здоровых с частой триграммой w
ось Y: $B_w(X_m)$ — доля больных с частой триграммой w

Болезнь: некроз головки бедренной кости



истинные y_i

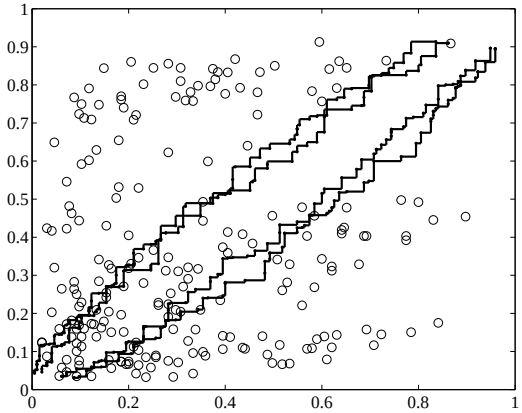


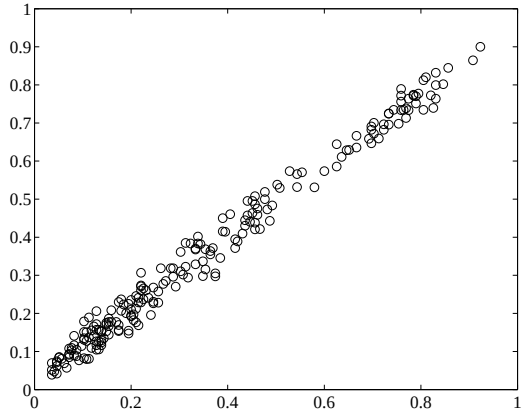
случайно перепутанные y_i

Значимые триграммы — вне 90% (99.8%) доверительной области, при 20 (1000) случайных перемешиваний меток y_i

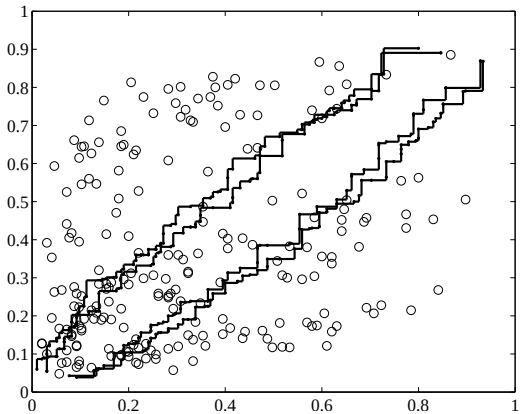
Перестановочный тест для поиска информативных триграмм

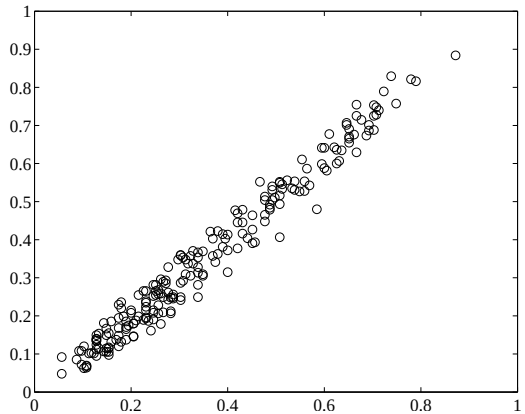
Болезнь: ишемия сердца





Болезнь: узловой зоб щитовидной железы





Модели классификации

$\{S_i\}_{i=1}^{\ell}$ — обучающая выборка кодограмм,
$\{y_i\}_{i=1}^{\ell}$ — диагнозы: 0 = здоровый, 1 = больной
$f_w(S_i)$ — числовой признак, монотонно зависящий от частоты k-граммы w в кодограмме S_i

Варианты определения признаков

- $f_w(S) = p_w(S)$ — частота k-граммы w
- $f_w(S) = [p_w(S) \geqslant \theta]$ — встречаемость k-граммы w

1. Линейные модели классификации

$$a(S) = \big[b(S) \geqslant \beta\big], \qquad b(S) = \sum_{w \in W} \gamma_w f_w(S),$$

где γ_w — вес k-граммы w.

- Синдромный алгоритм
- Логистическая регрессия

2. Случайный лес

Синдромный алгоритм

Различные сочетания типа используемых признаков, формулы весов и критерия информативности:

1. тип признаков: вещественные частоты k-грамм $p_w(S)$;
   - формула весов признаков:
     1) $\gamma_{mw} = 1$;
     2) $\gamma_{mw} = F_w(X_m)$;
     3) $\gamma_{mw} = F_w(X_m) - F_w(X_0)$;
     4) $\gamma_{mw} = \ln F_w(X_m) - \ln F_w(X_0)$;
     5) $\gamma_{mw} = DF_w(X_m)$;
   - критерий отбора K признаков с наибольшими значениями:
     1) $F_w(X_m)$; 2) $F_w(X_m)\big[w \notin T_0\big]$;
     3) $F_w(X_m) - F_w(X_0)$;
     4) $\ln F_w(X_m) - \ln F_w(X_0)$;
     5) $\big|\ln F_w(X_m) - \ln F_w(X_0)\big|$;
     6) $DF_w(X_m)$; 7) $\big|DF_w(X_m)\big|$.

Синдромный алгоритм

2. тип признаков: бинарные встречаемости $\left[p_w(S) \geqslant \theta\right]$;
   - формула весов признаков:
     1) $\gamma_{mw} = 1$; 2) $\gamma_{mw} = B_w(X_m)$;
     3) $\gamma_{mw} = B_w(X_m) - B_w(X_0)$;
     4) $\gamma_{mw} = \ln B_w(X_m) - \ln B_w(X_0)$;
     5) $\gamma_{mw} = \ln B_w(X_m)(1 - B_w(X_0)) - \ln B_w(X_0)(1 - B_w(X_m))$;
     6) $\gamma_{mw} = DB_w(X_m)$.
   - критерий отбора K признаков с наибольшими значениями:
     1) $B_w(X_m)$; 2) $B_w(X_m)\left[w \notin T_0\right]$;
     3) $B_w(X_m) - B_w(X_0)$;
     4) $\ln B_w(X_m) - \ln B_w(X_0)$;
     5) $\left|\ln B_w(X_m) - \ln B_w(X_0)\right|$;
     6) $\ln B_w(X_m)(1 - B_w(X_0)) - \ln B_w(X_0)(1 - B_w(X_m))$;
     7) $\left|\ln B_w(X_m)(1 - B_w(X_0)) - \ln B_w(X_0)(1 - B_w(X_m))\right|$;
     8) $DB_w(X_m)$; 9) $\left|DB_w(X_m)\right|$.

Логистическая регрессия

Методы понижения размерности:

- отбор K признаков с наибольшими значениями критерия информативности:
  1) $B_w(X_m, \theta)$;
  2) $F_w(X_m)$;
  3) $DB_w(X_m, \theta)$;
  4) $DF_w(X_m)$;
- L_1-регуляризация;
- метод главных компонент.

Оценивание качества диагностики

Доля правильно классифицируемых больных:

$$\text{чувствительность} = \frac{1}{|X_m|} \sum_{S \in X_m} \left[a_m(S) = 1 \right]$$

Доля правильно классифицируемых здоровых:

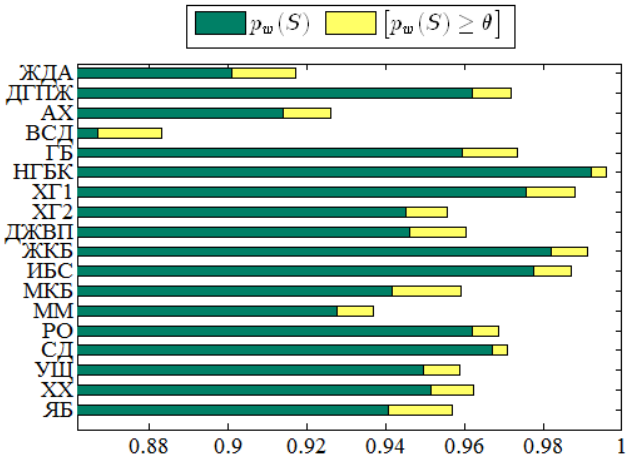
$$\text{специфичность} = \frac{1}{|X_0|} \sum_{S \in X_0} \left[a_m(S) = 0 \right]$$

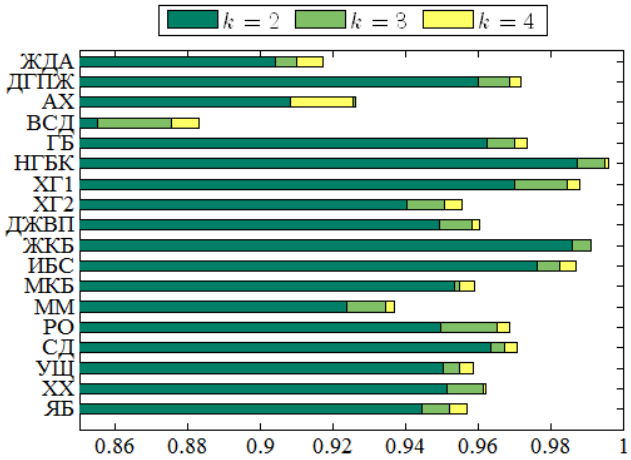
Area Under Curve — площадь под ROC-кривой, отображающей зависимость чувствительности от специфичности.

Выборка делилась на обучающую и контрольную в отношении $4:1$. Настройка параметров проводилась на обучающей выборке по результатам 1×10 кросс-валидации.

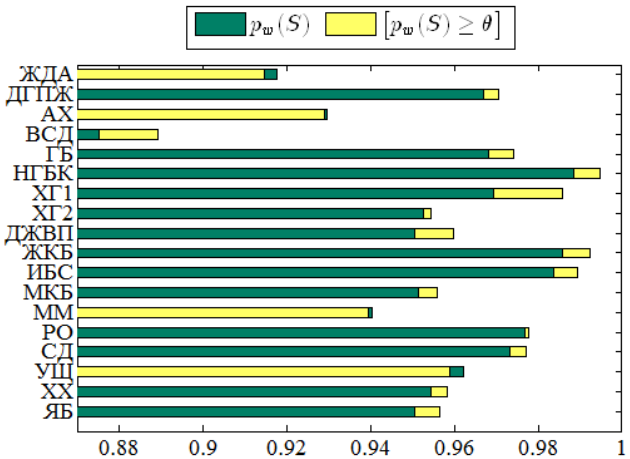
Линейные модели классификации

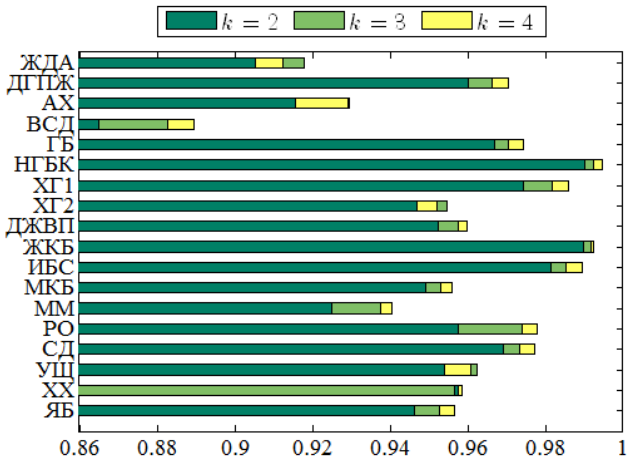
Синдромный алгоритм:





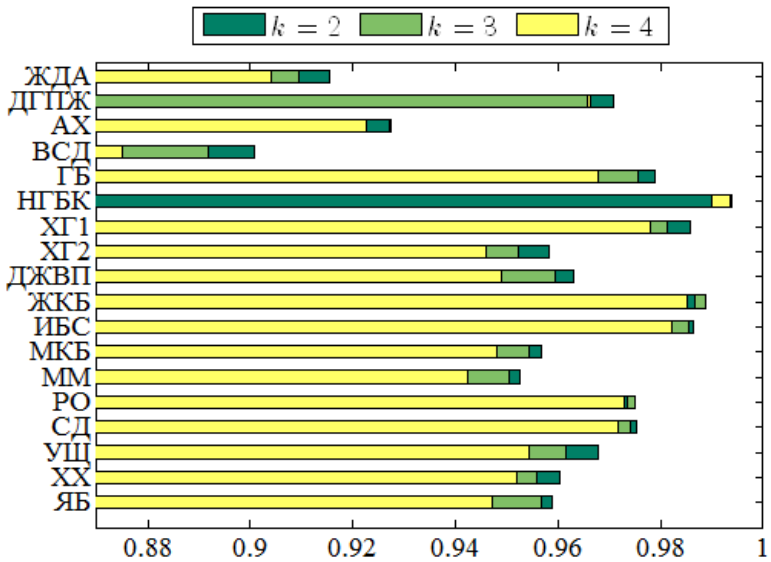
Логистическая регрессия:





Случайный лес

Значения AUC при $k = 2, 3, 4$:



Лучшие версии

Синдромный алгоритм (SA)

- тип признаков: бинарные, $k = 4$
- критерий информативности: $B_w(X_m)$
- формула весов:
$$\gamma_{mw} = \ln B_w(X_m)(1 - B_w(X_0)) - \ln B_w(X_0)(1 - B_w(X_m))$$

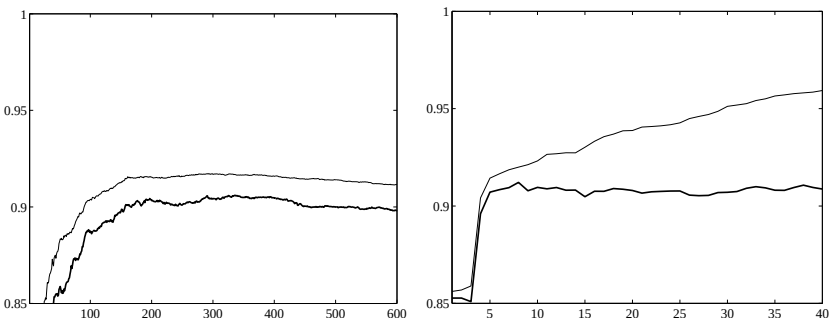
Логистическая регрессия (LR)

- тип признаков: бинарные, $k = 4$
- отбор признаков: метод главных компонент

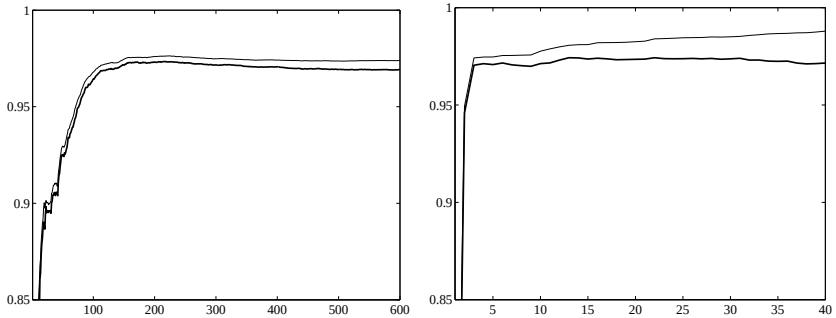
Случайный лес (RF)

- тип признаков: $k = 2$
- число деревьев: 300

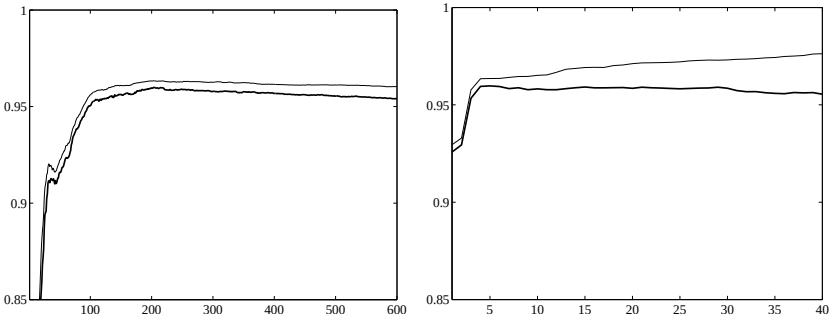
Зависимость AUC от числа признаков для SA и LR



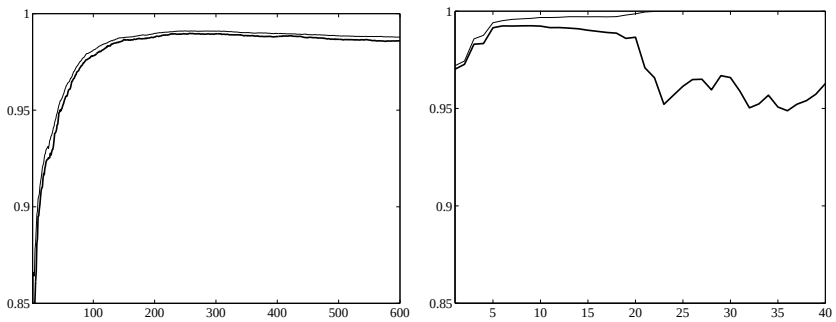
анемия



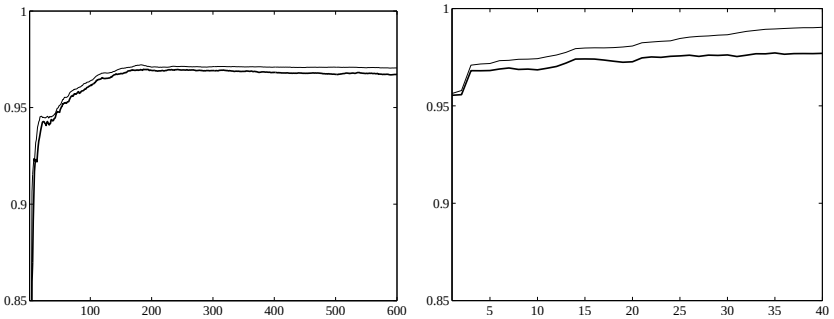
гипертония



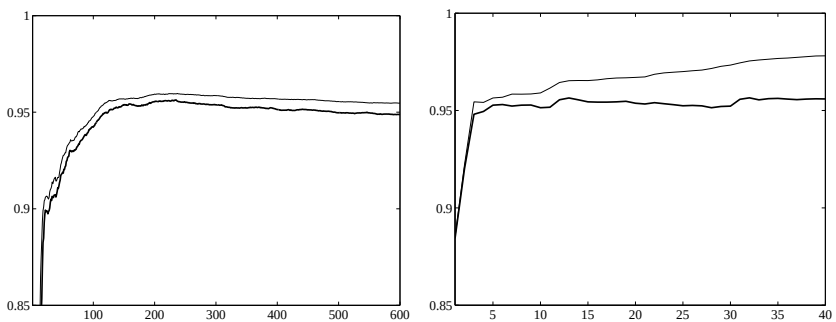
дискинезия ЖВП



желчнокаменная болезнь



сахарный диабет

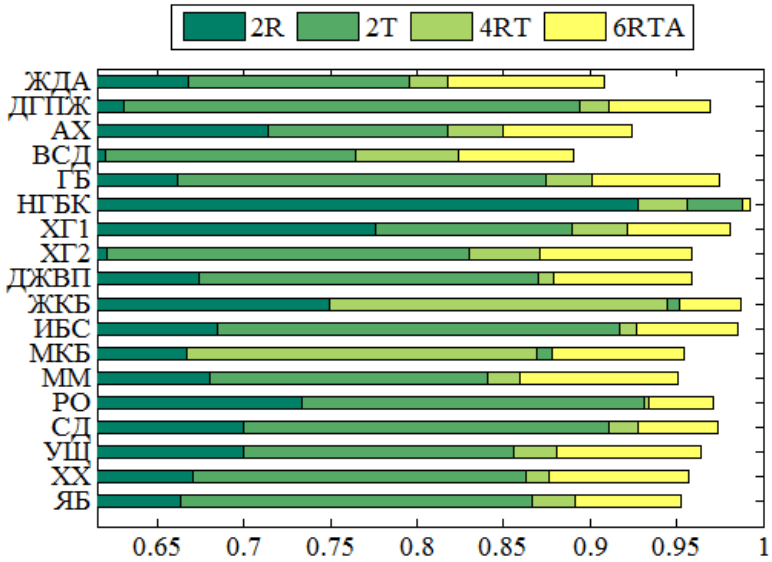


язвенная болезнь

Значения AUC, контрольная выборка (мощность выборки здоровых 193)

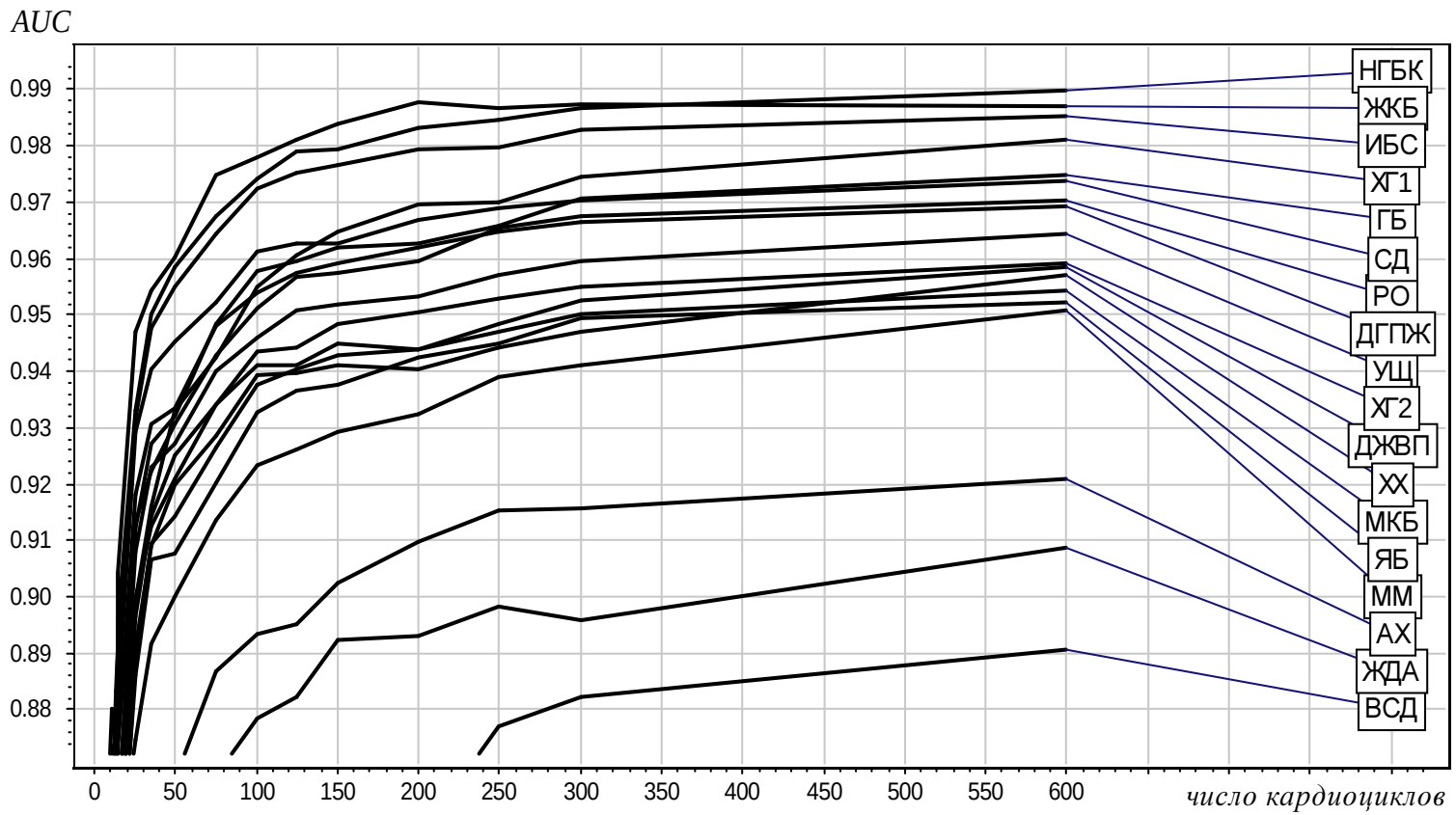
болезнь	$\lvert X_m \rvert$	HM, %	SA, %	LR, %	RF, %
некроз головки бедренной кости	324	98,84	98,86	99,29	99,41
желчнокаменная болезнь	278	95,42	98,40	97,39	98,63
ишемическая болезнь сердца	1265	94,57	96,12	96,65	98,14
сахарный диабет	871	95,92	96,29	96,89	97,68
гастродуоденит гиперацидный	324	92,23	95,78	95,46	97,59
аденома простаты	260	94,38	95,66	96,99	97,39
гипертоническая болезнь	1894	93,19	95,16	95,71	96,95
гастродуоденит гипоацидный	700	86,90	93,50	94,51	95,81
узловой зоб щитовидный железы	748	86,39	93,62	94,15	95,80
мочекаменная болезнь	654	80,25	93,89	93,23	95,54
рак общий	530	–	93,57	95,38	95,38
холецистит хронический	340	85,90	92,46	91,40	94,78
миома матки	781	86,69	93,67	92,44	94,55
дискинезия ЖВП	717	84,17	92,11	92,04	93,66
язвенная болезнь	785	90,83	91,88	92,80	93,22
аднексит хронический	276	–	90,26	89,74	90,16
анемия железодефицитная	260	79,14	83,97	85,21	89,00
вегетососудистая дистония	694	73,38	85,32	84,84	87,21

Выбор способа кодирования



- 2R — только амплитуды, 2-буквенный алфавит
- 2T — только интервалы, 2-буквенный алфавит
- 4RT — интервалы и амплитуды, 4-буквенный алфавит
- 6RTA — интервалы, амплитуды и углы, 6-буквенный алфавит

Зависимость AUC от длины кардиосигнала



Выводы

- проверены статистические гипотезы о неслучайном характере вариаций интервалов и амплитуд кардиоциклов и о взаимосвязи этих вариаций с заболеваниями;
- разработан легко интерпретируемый, практически не подверженный переобучению синдромный алгоритм;
- достигнуты высокие уровни чувствительности и специфичности для выбранных заболеваний;
- обоснован учет совместной вариабельности интервалов, амплитуд и их отношений при кодировании ЭКГ-сигнала;
- показано, что длины кардиосигнала, равной 300, достаточно для получения высокого качества классификации.