Статистические тесты для проверки однородности и воспроизводимости электрокардиосигналов

Успенский Вячеслав Максимилианович[1]
Воронцов Константин Вячеславович[2,3]
Бунакова Влада Руслановна[3]
Жариков Илья Николаевич[3]
Ишкина Шаура Хабировна[2]

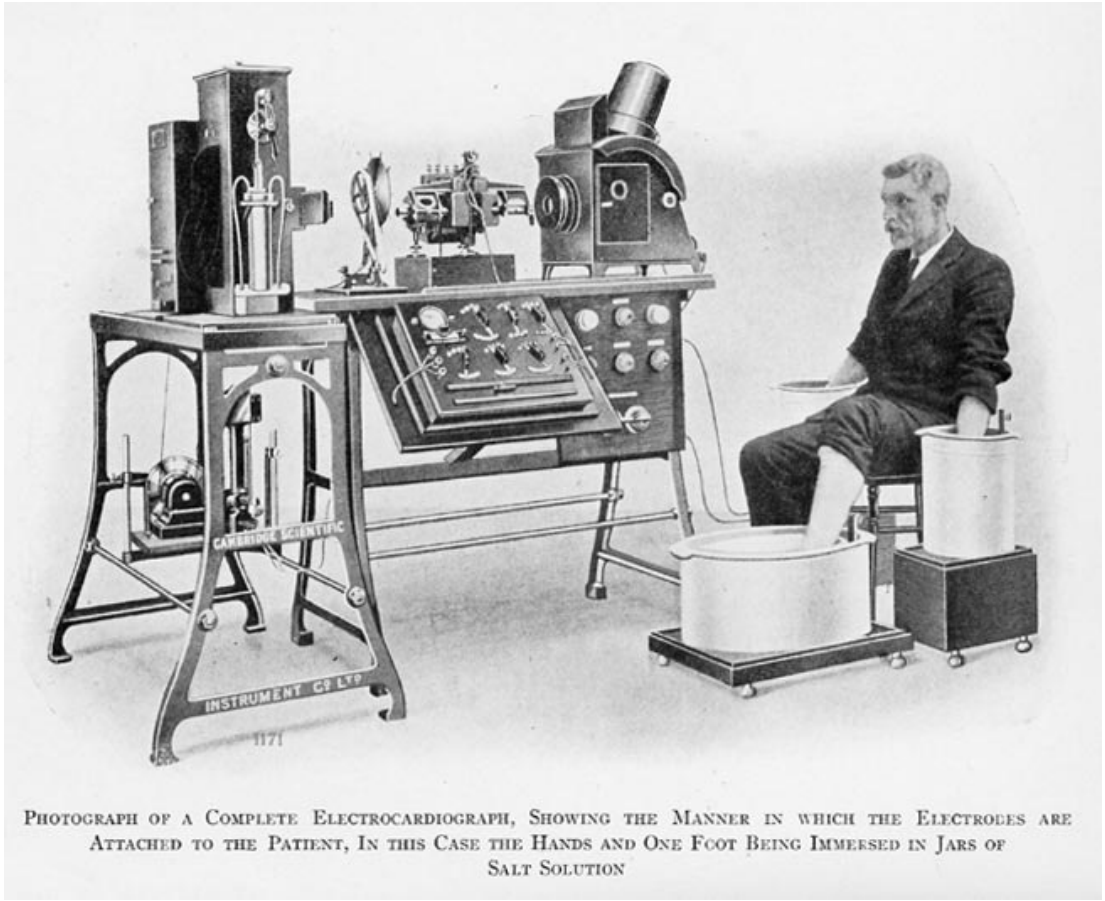
([1] *ФКУ ЦВКГ им. П.В.Мандрыка*, [2] *ФИЦ ИУ РАН*, [3] *МФТИ*)

Международная научно-практическая конференция
«175 лет ВНИИМ им. Д. И. Менделеева и Национальной системе обеспечения единства измерений»

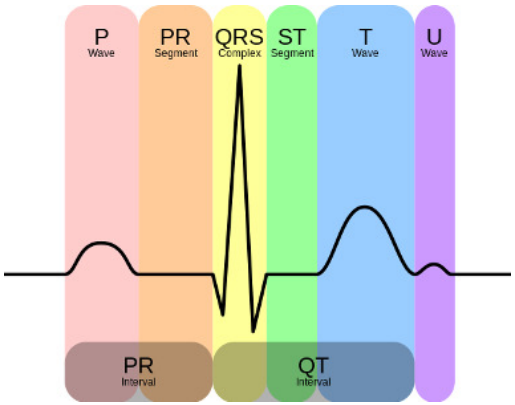
Санкт-Петербург 14–15 июня 2017

Содержание

Электрокардиография



Photograph of a Complete Electrocardiograph, Showing the Manner in which the Electrodes are Attached to the Patient, In this Case the Hands and One Foot Being Immersed in Jars of Salt Solution



1872 — первые записи электрической активности сердца

1911 — коммерческий электрокардиограф (фото)

1924 — нобелевская премия по медицине, Виллем Эйнтховен

Теория информационной функции сердца

Возможна ли диагностика несердечных заболеваний по ЭКГ?

Предпосылки:

- Китайская традиционная медицина: *пульсовая диагностика*
- Анализ вариабельности сердечного ритма [Р. М. Баевский]
- Цифровая электрокардиография высокого разрешения

Предположения:

- ЭКГ несёт информацию о работе всех систем организма
- Каждая болезнь по-своему «модулирует» ЭКГ-сигнал
- «Модуляция» происходит уже на ранней стадии болезни

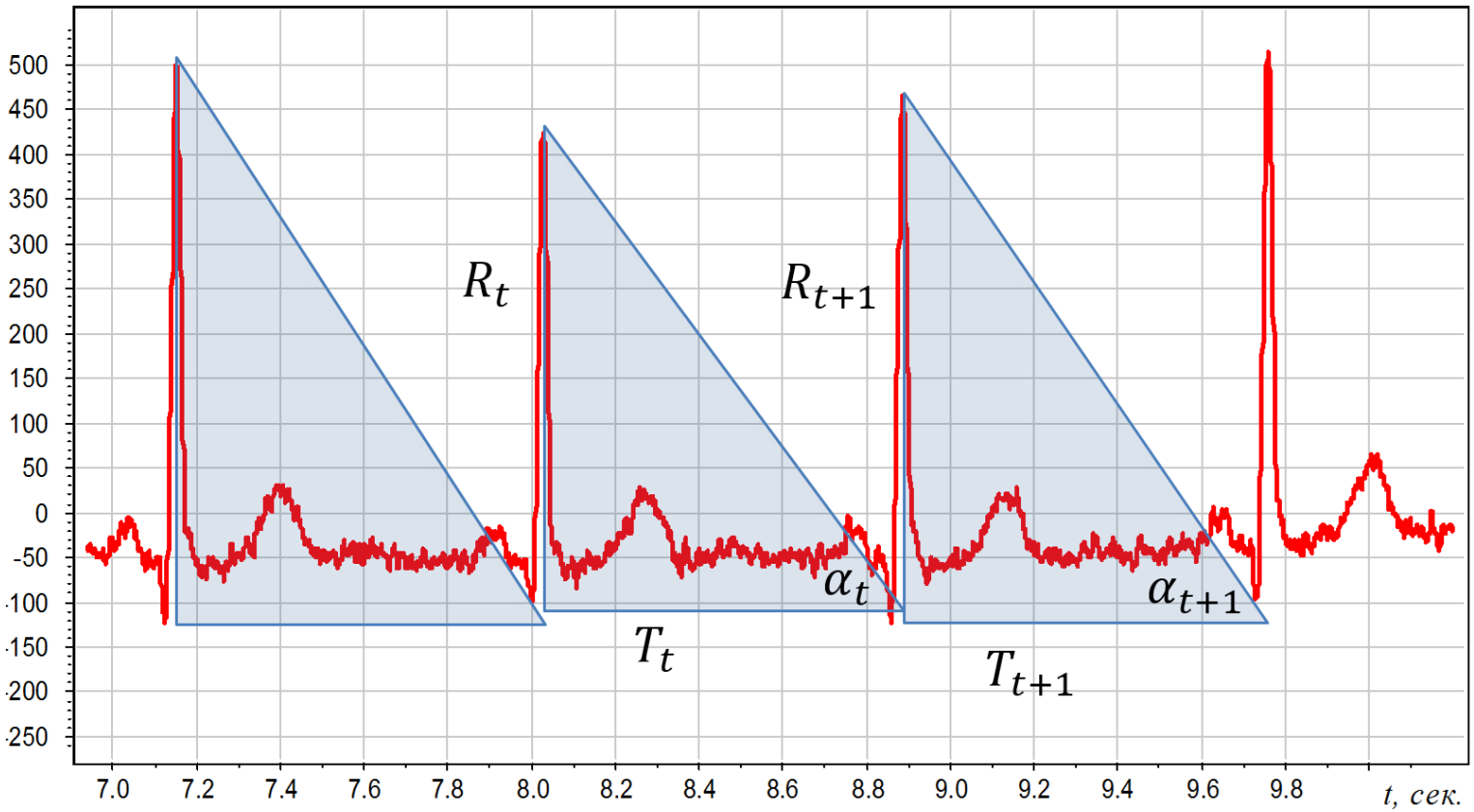
В.М.Успенский. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов. М.: Экономика и информатика, 2008.

Вариабельность интервалов и амплитуд кардиоциклов

приращение амплитуд: $dR_t = R_{t+1} - R_t$

приращение интервалов: $dT_t = T_{t+1} - T_t$

приращение углов: $d\alpha_t = \alpha_{t+1} - \alpha_t, \quad \alpha_t = \operatorname{arctg} \frac{R_t}{T_t}$



Технология информационного анализа ЭКГ-сигналов

Этап I. Методы символьной динамики

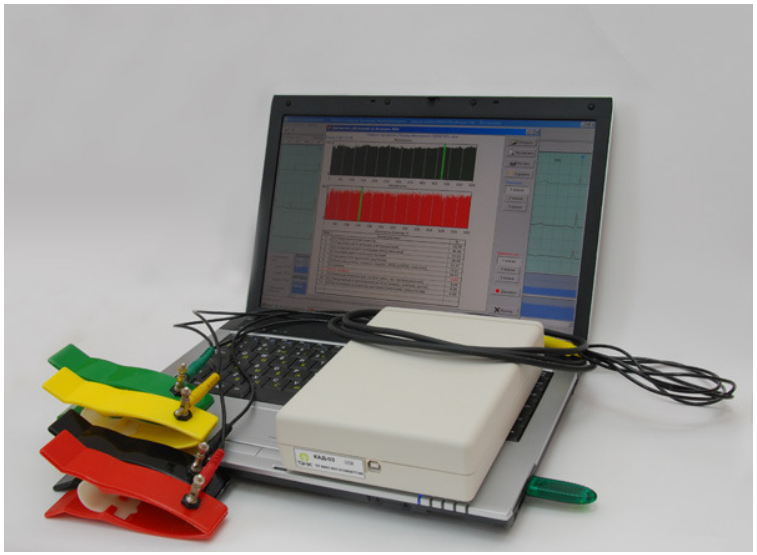
1. *Демодуляция* — вычисление амплитуд, интервалов и углов по кардиограмме длиной 600 кардиоциклов
2. *Дискретизация* — перевод в *кодограмму* — 599-символьную строку в 6-буквенном алфавите
3. *Векторизация* — перевод в вектор $6^3{=}216$ частот триграмм

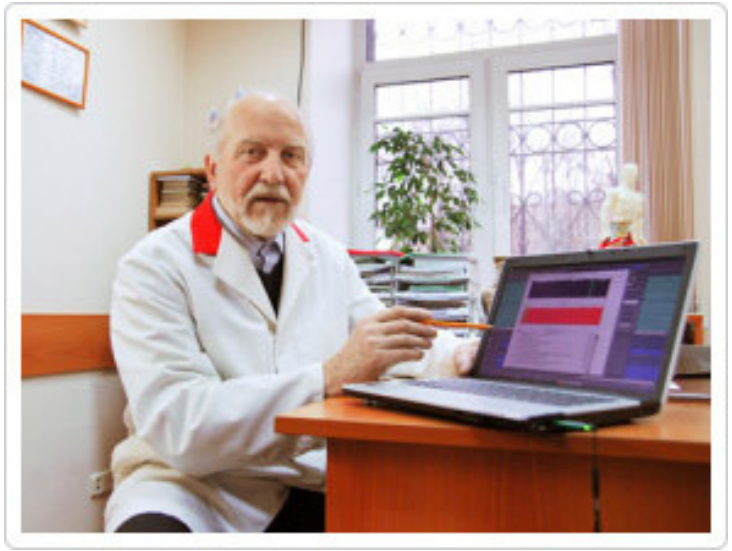
Этап II. Методы машинного обучения

1. Формирование обучающих выборок здоровых и больных
2. Построение моделей классификации
3. Обучение моделей классификации
4. Оценивание качества диагностики

Диагностическая система «Скринфакс»

Цифровой электрокардиограф с улучшенной помехозащищённостью и расширенной полосой пропускания.





- более 15 лет исследований и накопления данных
- более 20 тысяч прецедентов (кардиограмма + диагнозы)
- более 40 заболеваний

Объём исходных данных (по заболеваниям)

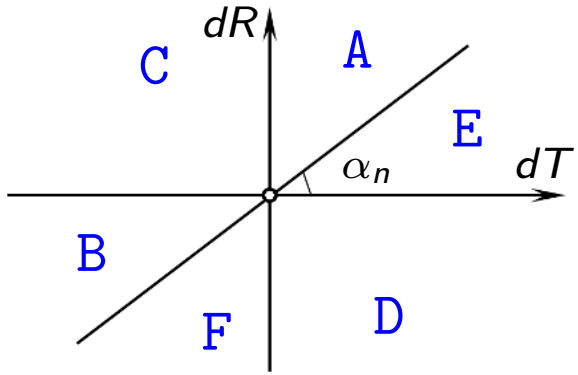
«абсолютно здоровые»	АЗ	193
гипертоническая болезнь	ГБ	1894
ишемическая болезнь сердца	ИБС	1265
сахарный диабет (СД1 и СД2)	СД	871
язвенная болезнь	ЯБ	785
миома матки	ММ	781
узловой (диффузный) зоб щитовидный железы	УЩ	748
дискинезия желчевыводящих путей	ДЖВП	717
хронический гастрит (гастродуоденит) гипоацидный	ХГ2	700
вегетососудистая дистония	ВСД	694
мочекаменная болезнь	МКБ	654
рак общий (онкопатология различной локализации)	РО	530
холецистит хронический	ХХ	340
асептический некроз головки бедренной кости	НГБК	324
хронический гастрит (гастродуоденит) гиперацидный	ХГ1	324
желчнокаменная болезнь	ЖКБ	278
аднексит хронический	АХ	276
аденома простаты	ДГПЖ	260
анемия железодефицитная	ЖДА	260

Дискретизация ЭКГ-сигнала

Вход: интервалограмма $(T_t)_{t=1}^{N+1}$ и амплитудограмма $(R_t)_{t=1}^{N+1}$

Правила кодирования:

$dR_t = R_{t+1} - R_t$	+	−	+	−	+	−
$dT_t = T_{t+1} - T_t$	+	−	−	+	+	−
$d\alpha_t = \alpha_{t+1} - \alpha_t$	+	+	+	−	−	−
s_t	A	B	C	D	E	F



Выход: *кодограмма* $x = (s_t)_{t=1}^{N}$ — последовательность символов алфавита $\{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$:

DBFEACFDAAFBABDDAADFAAFFEACFEACFBAEFFAABFFAAFFAAFFAAFFAAEBFAEBFEAAFCAAFFAAD
FCAFFAADFCADFCCDFDACFFACDFAEFFACFFEADFCAFBCADFFECFFAAFFAAFFAEFFCACFCAEFFCAD
DAADBFAAFFAEBFAABFACDFFAAFBAADFAADFDAAFCECFCEDFCEEFCAEFBECBBBAADBAACFFAAFFA
CFFCECFDAABDAEFFAAFFCEDBFAAFFAEFFAEFBACFBAEDFEAAFFCAFFDAAFFAEBDAADBBADFDAFF
EABFCCAFDEEBDECFFACFFAABFAADFBAAFFACFFFAEFFACFFACFFCECFBAAFFFAAFFFAAFFAADFB
AABFACDFDAEFFAADBAAEFFEAFBCECFDECCFBAAFFAADFDACDFAAFFAADFCAADFAEFBAAFFCADFE
AFFCECFCECFFAAFFABCFDAAAFFADBFCAEFFAABFACBFAAEBFAEBFCAFFBAAFFAAFFDACFDAABFB
CAFFAECFFACFFACDFCADFDAABFAAEDDABBFCACDBAAFFAAFFCADFAADFDACFFAEDFCACFCAEBCE

Векторизация ЭКГ-сигнала

x — *кодограмма*, последовательность символов $\{\mathtt{A},\mathtt{B},\mathtt{C},\mathtt{D},\mathtt{E},\mathtt{F}\}$:

$\left(f_j(x)\right)_{j=1}^{n}$ — вектор частот триграмм, $n=6^3=216$:

1. FFA - 42	17. EFF - 10	33. CEC - 6	49. EAC - 3
2. FAA - 33	18. DAA - 10	34. ADB - 5	50. DDA - 3
3. AFF - 32	19. ECF - 9	35. FFE - 5	51. CAC - 3
4. AAF - 30	20. FFC - 9	36. EBF - 5	52. EDF - 3
5. ADF - 18	21. FEA - 9	37. CFD - 5	53. EFB - 3
6. FCA - 18	22. DFC - 8	38. AFB - 4	54. DBA - 3
7. ACF - 17	23. ABF - 8	39. AAE - 4	55. FCC - 2
8. AAD - 15	24. AAB - 8	40. CFC - 4	56. AFC - 2
9. CFF - 14	25. FCE - 8	41. CAE - 4	57. EAA - 2
10. AEF - 13	26. AEB - 7	42. DAC - 4	58. CED - 2
11. FDA - 13	27. DFD - 7	43. DBF - 4	59. CAA - 2
12. FAE - 12	28. ACD - 6	44. BFC - 4	60. BCA - 2
13. FAC - 12	29. CDF - 6	45. CFB - 4	61. BBA - 2
14. FBA - 11	30. DFA - 6	46. AED - 3	62. DFF - 2
15. BFA - 11	31. CAF - 6	47. FFF - 3	63. BDA - 2
16. BAA - 11	32. CAD - 6	48. FBC - 3	64. DAE - 2

Линейная модель классификации с двумя классами

$\{x_i\}_{i=1}^{\ell}$ — обучающая выборка кодограмм
y_i — класс объекта x_i: больной $y_i = 1$, здоровый $y_i = 0$

Основная эмпирическая гипотеза:

- у больных одни триграммы частые, у здоровых — другие

Линейная модель классификации:

$$\langle x, w \rangle = \sum_{j=1}^{n} w_j f_j(x), \qquad a(x) = \begin{cases} 1, & \langle x, w \rangle \geqslant w_0 \\ 0, & \langle x, w \rangle < w_0 \end{cases}$$

где w_j — вес j-й триграммы:

- $w_j > 0$, если триграмма более характерна для больных
- $w_j < 0$, если триграмма более характерна для здоровых
- $w_j = 0$, если триграмма не информативна для этой болезни

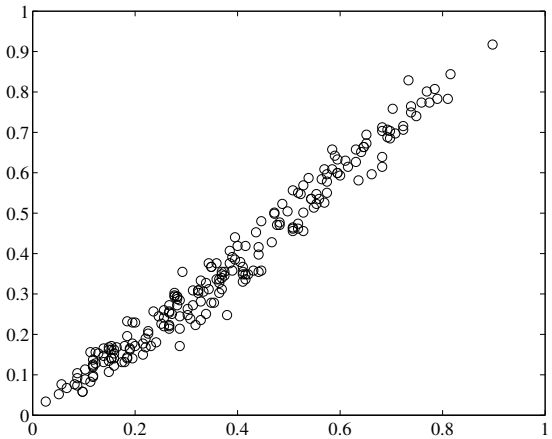
Методы машинного обучения

- **Наивный байесовский классификатор**
  - ☺ простой интерпретируемый линейный классификатор
  - ☹ качество классификации невысокое
- **Наивный байесовский классификатор + отбор признаков**
  - 😐 качество классификации получше
  - 😐 находит один диагностический эталон каждой болезни
- **Метод главных компонент + логистическая регрессия**
  - ☺ качество классификации высокое
  - ☹ не определяет диагностические эталоны болезней
- **SVM, нейронные сети, случайный лес**
  - ☺ качество классификации высокое
  - ☹ неоправданно сложное, неинтерпретируемое решение
- **Тематические модели классификации**
  - ☺ автоматически находит все диагностические эталоны
  - 😐 качество классификации среднее

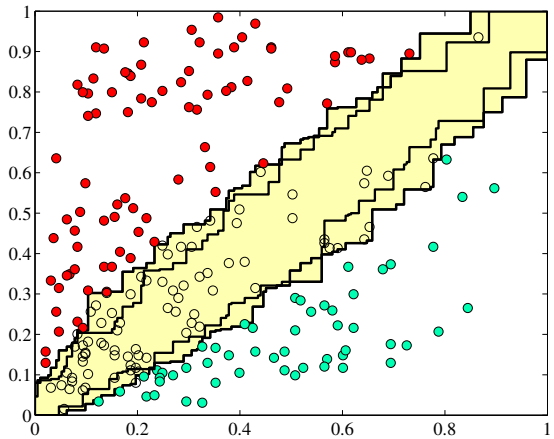
Существуют сочетания триграмм, специфичные для болезней

Точки на графиках соответствуют триграммам, $j = 1, \ldots, 216$
— ось X: доля здоровых с частой триграммой $f_j(x_i) \geqslant 2$
— ось Y: доля больных с частой триграммой $f_j(x_i) \geqslant 2$

НГБК (асептический некроз головки бедренной кости)



случайно перемешанные y_i

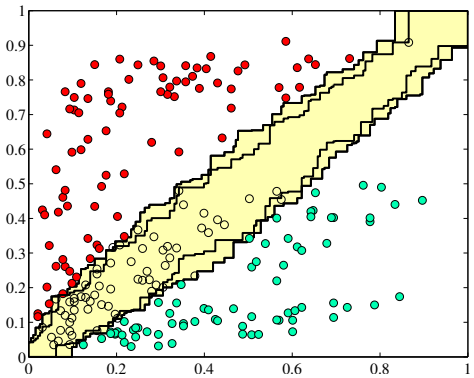


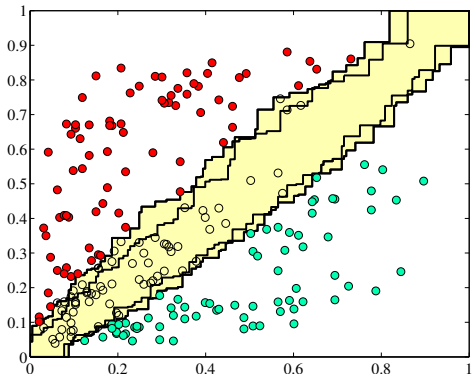
наблюдаемые y_i

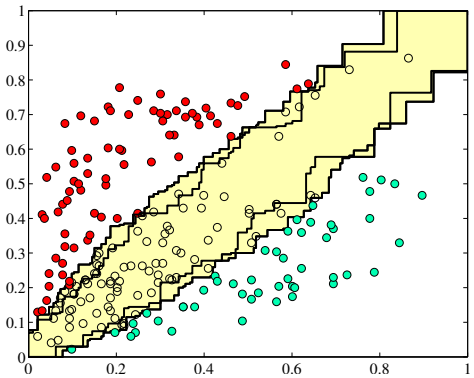
Слева: как распределятся точки, если объектам x_i назначить случайно переставленные метки классов y_i.
Жёлтая область: если случайно перемешать 20 раз, 1000 раз.

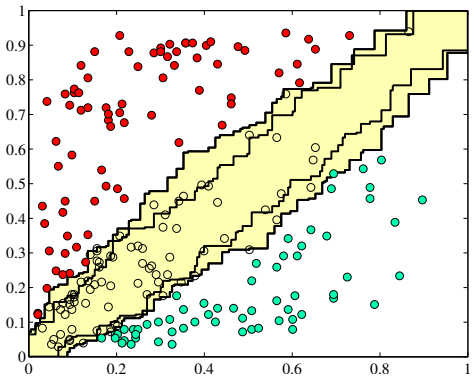
Существуют сочетания триграмм, специфичные для болезней

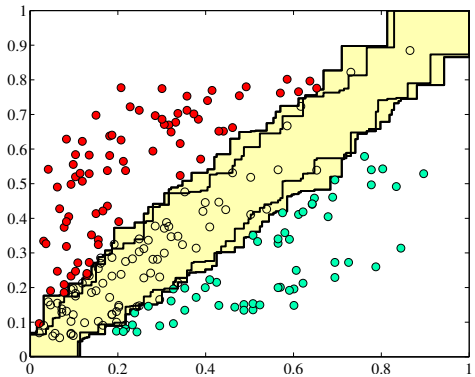
Для каждой болезни есть свои неслучайно частые триграммы

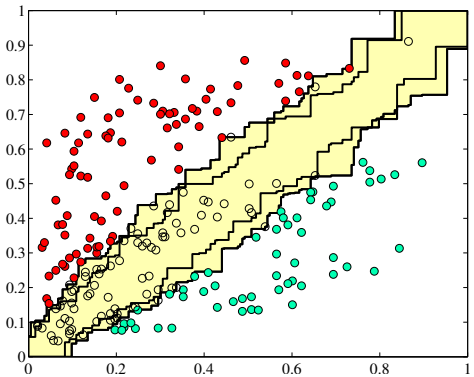

ишемия сердца


гипертония


рак


желчнокаменная болезнь


миома матки


язвенная болезнь

Терминология диагностики

Положительный диагноз — алгоритм предсказывает болезнь

Доля больных с верным положительным диагнозом:

$$\text{чувствительность} = \frac{\sum_{i=1}^{\ell}[y_i = 1][a(x_i) = 1]}{\sum_{i=1}^{\ell}[y_i = 1]}$$

Доля здоровых с верным отрицательным диагнозом:

$$\text{специфичность} = \frac{\sum_{i=1}^{\ell}[y_i = 0][a(x_i) = 0]}{\sum_{i=1}^{\ell}[y_i = 0]}$$

Максимизируем чувствительность и специфичность

- они не зависят от соотношения мощностей классов
- они подходят для несбалансированных выборок

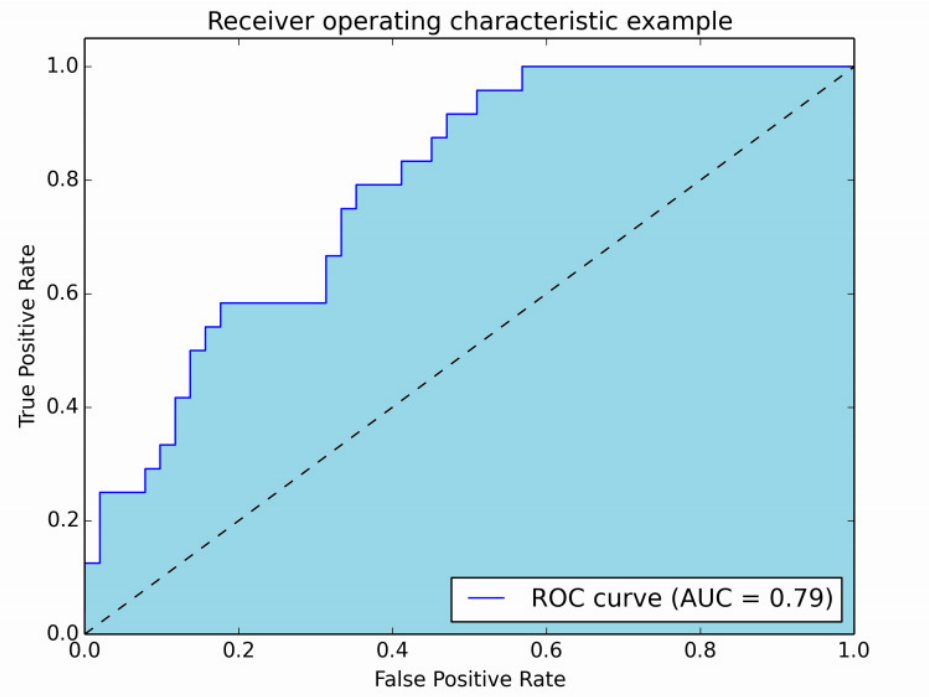
ROC-кривой и AUC — площадь под ROC-кривой

Модель классификации: $a(x) = [\langle x, w \rangle > w_0]$

по оси X: 1 – специфичность = FPR, False Positive Rate,

по оси Y: чувствительность = TPR, True Positive Rate

Каждая точка ROC-кривой соответствует значению порога w_0
(ROC — «receiver operating characteristic»),

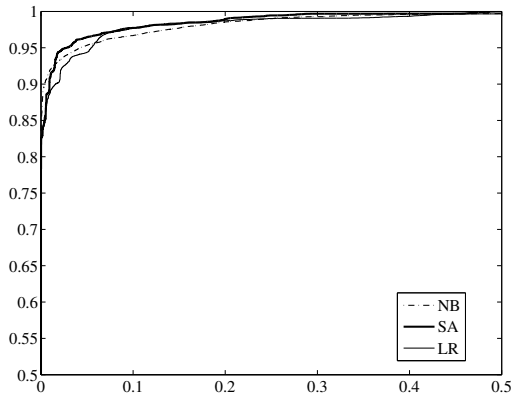


Результаты кросс-валидации

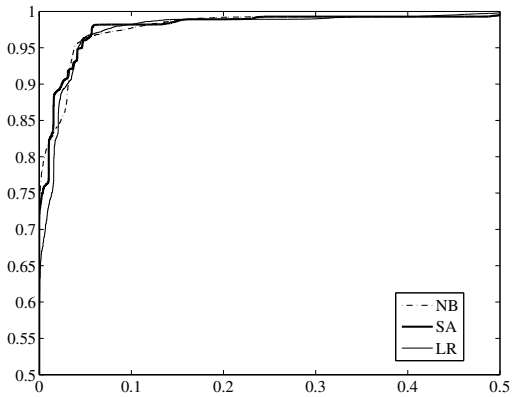
Обучающая выборка: оптимизация параметров модели
Тестовая выборка: Чувствительность, Специфичность, AUC
40×10-fold cross-validation — для доверительного оценивания

болезнь	выборка	AUC, %	С% при Ч=95%
некроз головки бедренной кости	327	99.19 ± 0.10	96.6 ± 1.76
желчнокаменная болезнь	277	98.98 ± 0.23	94.4 ± 1.54
ишемическая болезнь сердца	1262	97.98 ± 0.14	91.1 ± 1.86
гастрит	321	97.76 ± 0.11	88.3 ± 2.64
гипертоническая болезнь	1891	96.76 ± 0.09	84.7 ± 1.99
сахарный диабет	868	96.75 ± 0.19	85.3 ± 2.18
аденома простаты	257	96.49 ± 0.13	80.1 ± 3.19
рак	525	96.49 ± 0.28	82.2 ± 2.38
узловой зоб щитовидной железы	750	95.57 ± 0.16	73.5 ± 3.41
холецистит хронический	336	95.35 ± 0.12	74.8 ± 2.46
дискинезия ЖВП	714	94.99 ± 0.16	70.3 ± 4.67
мочекаменная болезнь	649	94.99 ± 0.11	69.3 ± 2.14
язвенная болезнь	779	94.62 ± 0.10	63.6 ± 2.55

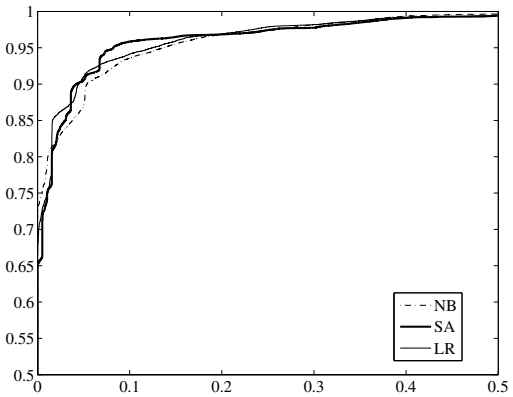
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



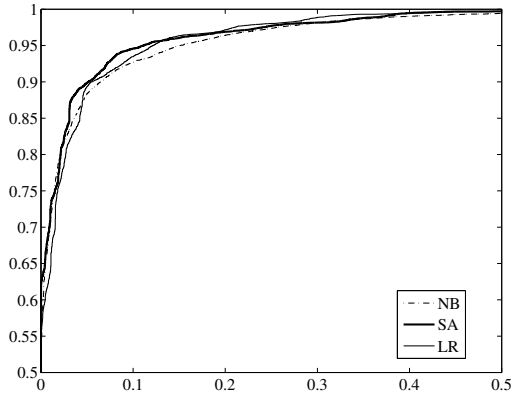
асептический некроз ГБК



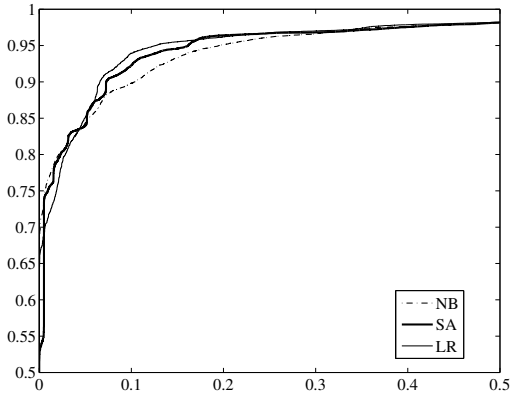
желчнокаменная болезнь



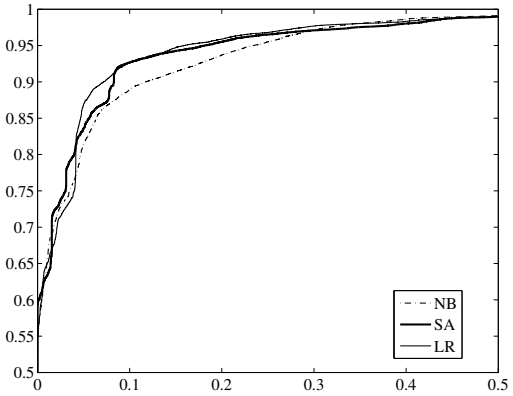
ишемическая болезнь



хронический гастрит 1



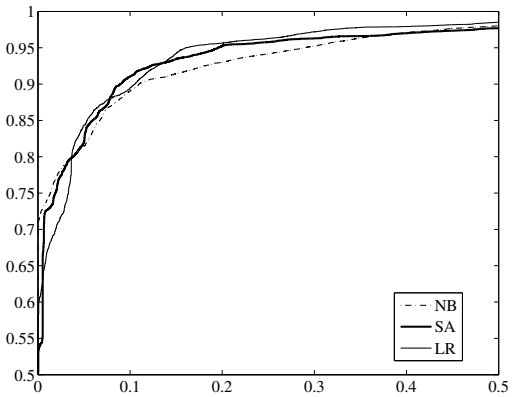
сахарный диабет



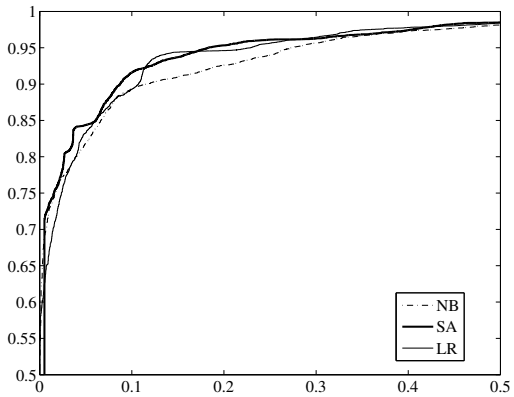
гипертония

NB — Naïve Bayes, SA — Syndrome Algorithm, LR — Logistic Regression

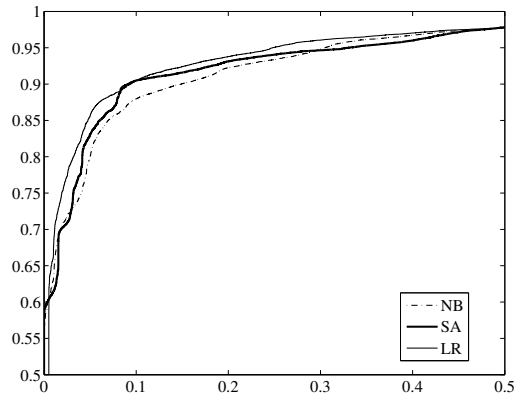
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



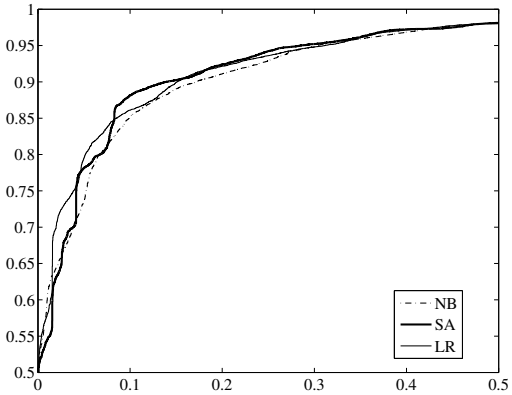
рак общий



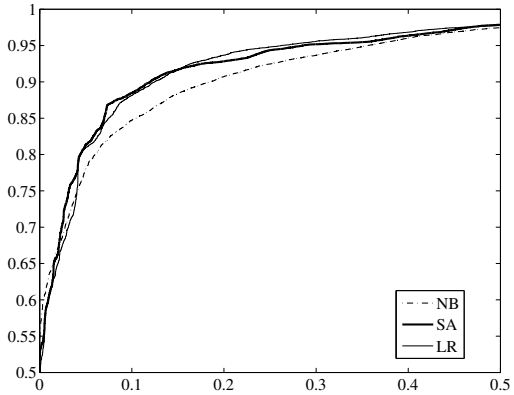
аденома простаты



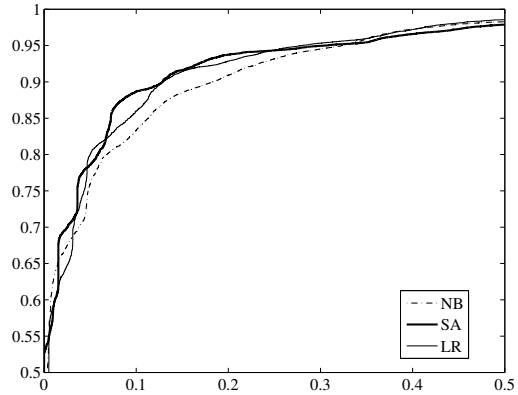
зоб щитовидной железы



хронический гастрит 2



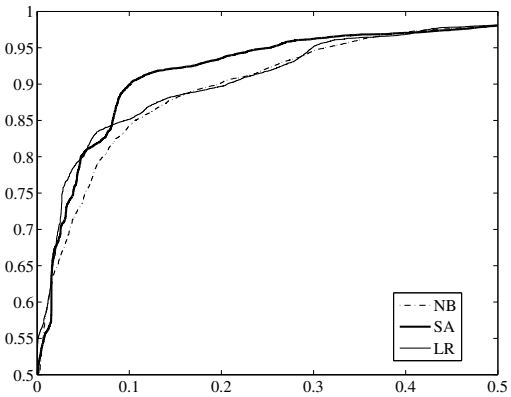
дискинезия ЖВП



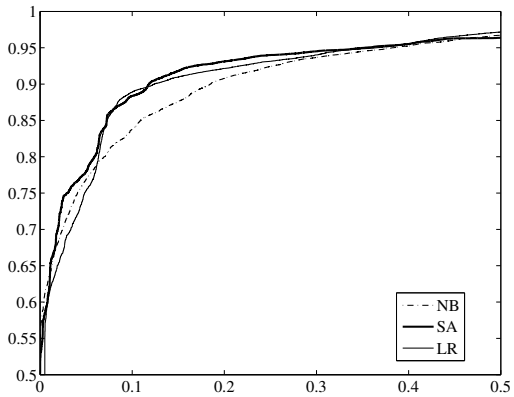
мочекаменная болезнь

NB — Naïve Bayes, SA — Syndrome Algorithm, LR — Logistic Regression

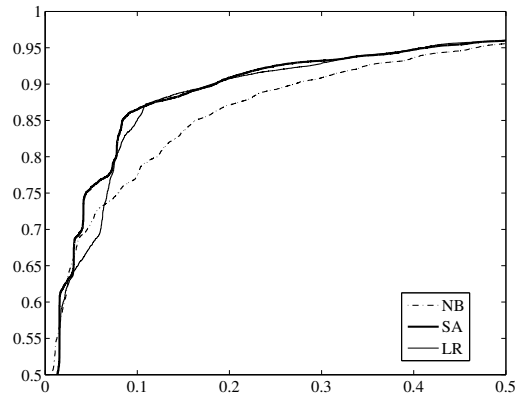
ROC-кривые в осях X:(1−специфичность), Y:чувствительность



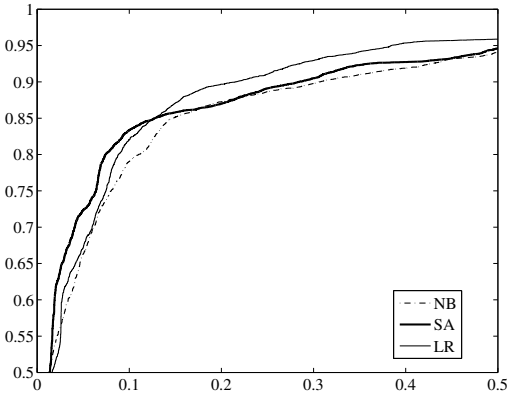
хронический холецистит



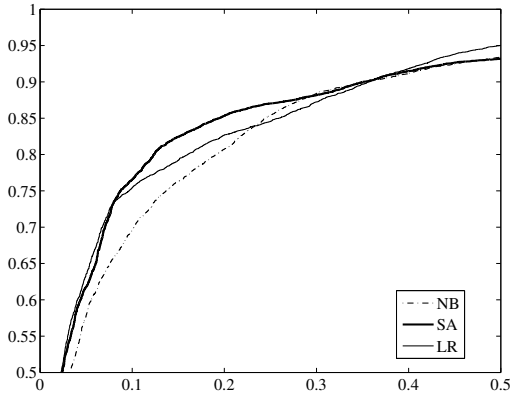
язвенная болезнь



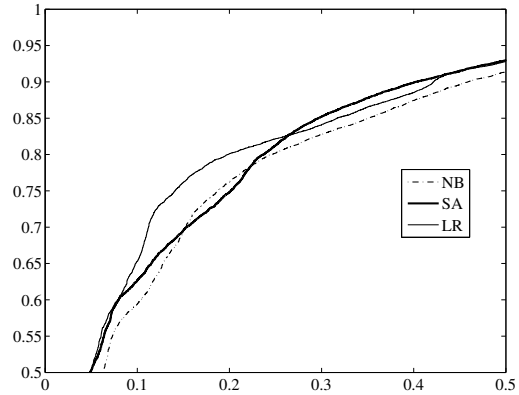
миома матки



хронический аднексит



анемия



вегетососудистая дистония

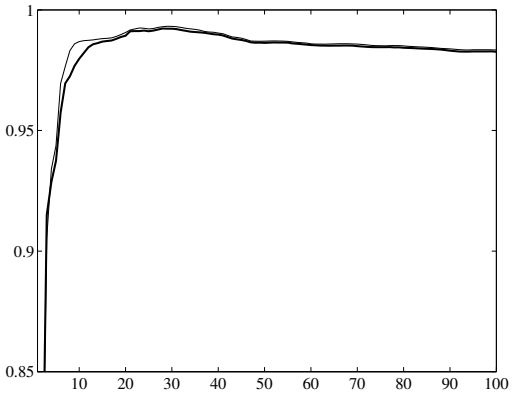
NB — Naïve Bayes, SA — Syndrome Algorithm, LR — Logistic Regression

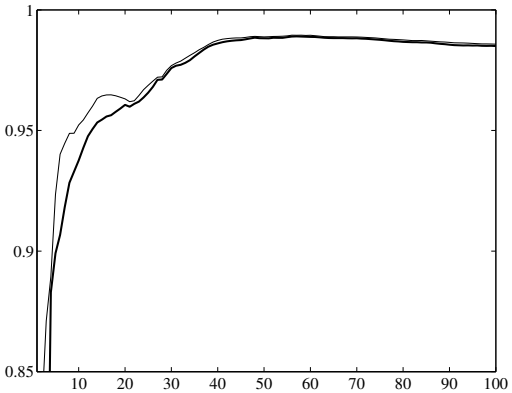
Эксперименты по выбору признаков и модели классификации

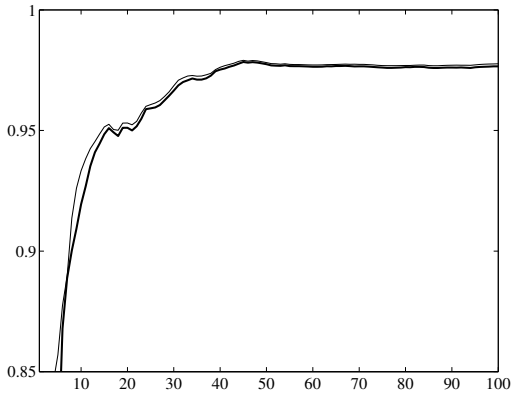
1. Выбор подпространства признаков
   - синдромный алгоритм: K наиболее значимых признаков
   - логистическая регрессия: K главных компонент
2. Выбор типа символьного кодирования
   - что использовать: интервалы, амплитуды, их отношения
   - ширина окна h-грамм
3. Выбор модели классификации
   - HM — полу-автоматический отбор триграмм
   - SA — синдромный алгоритм, NB с отбором признаков
   - LR — логистическая регрессия на главных компонентах
   - RF — случайный лес (Random Forest)
4. Оценка достаточной длительности регистрации ЭКГ
5. Тестирование на открытых данных по инфарктам миокарда

Зависимости AUC от числа используемых признаков K

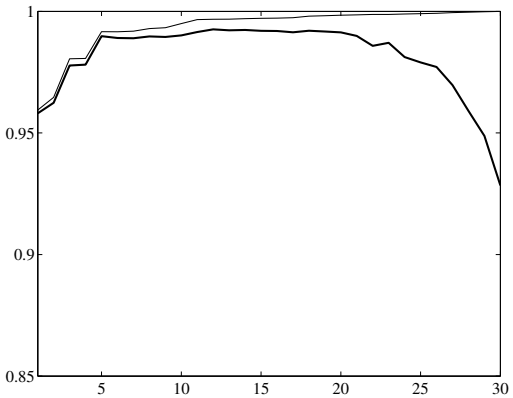
Синдромный алгоритм (наивный Байес на K признаках):

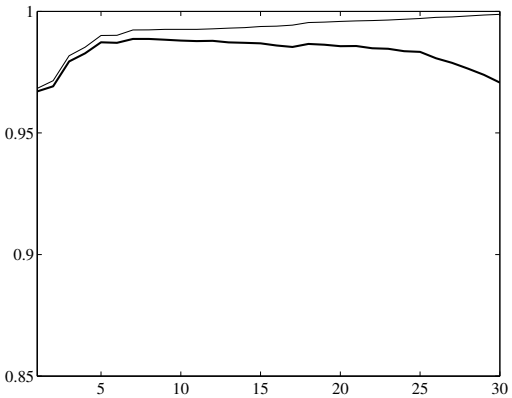


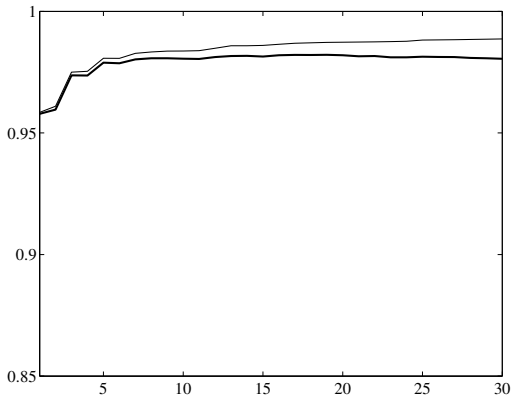




Логистическая регрессия (K — число главных компонент):





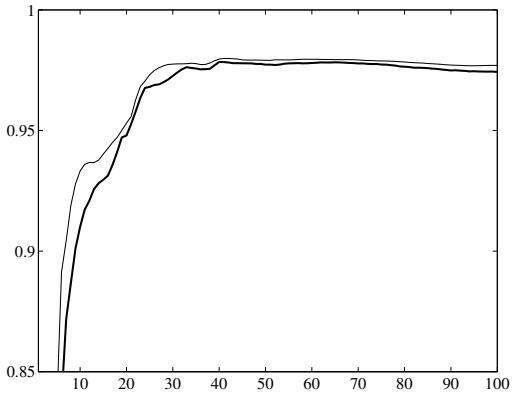


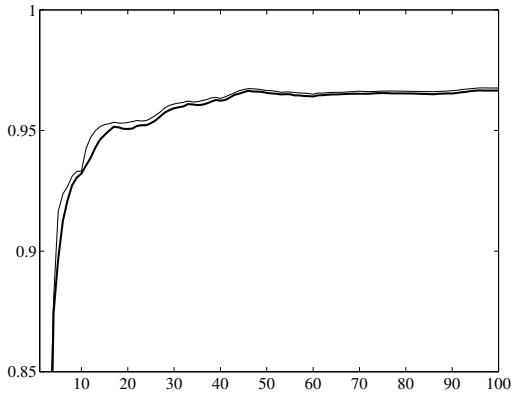
асептический некроз ГБК | желчнокаменная болезнь | ишемическая болезнь

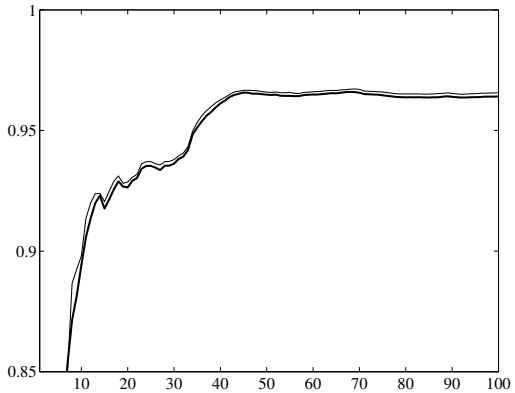
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

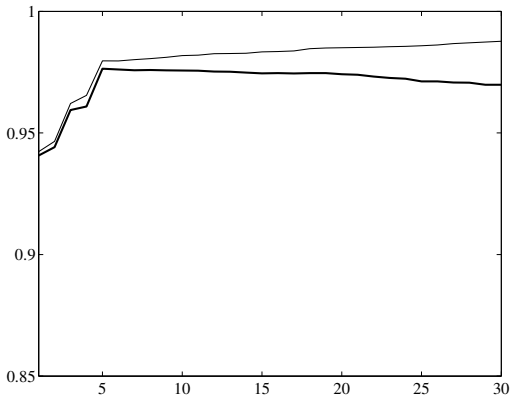
Синдромный алгоритм (K — число признаков):

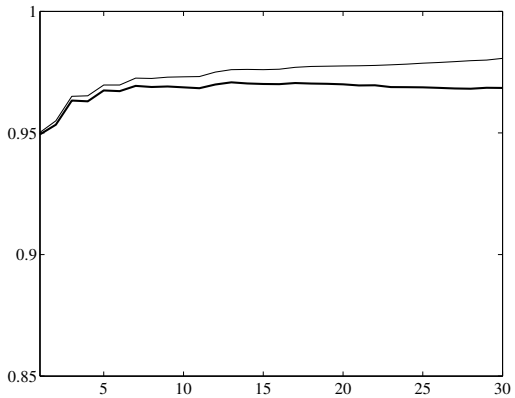


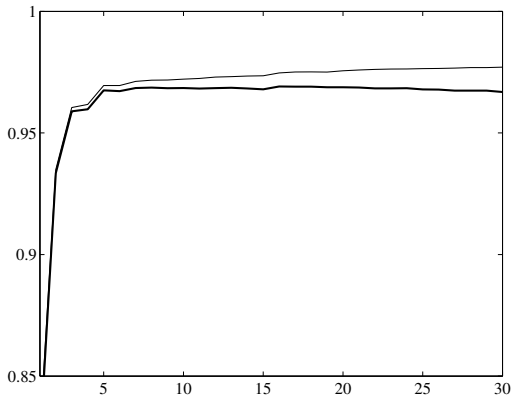




Логистическая регрессия (K — число главных компонент):





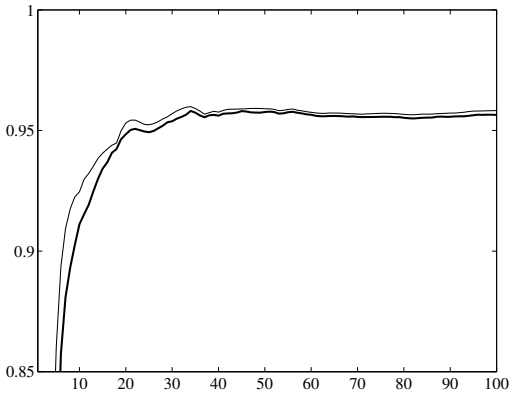


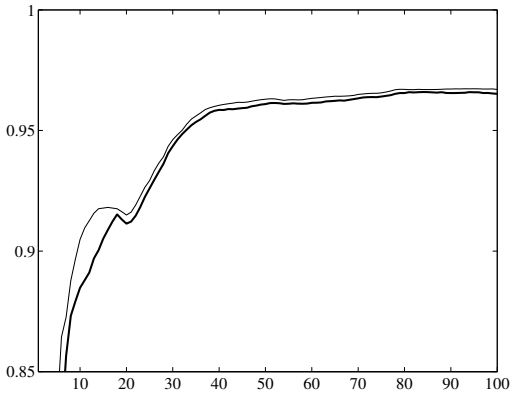
хронический гастрит 1 сахарный диабет гипертония

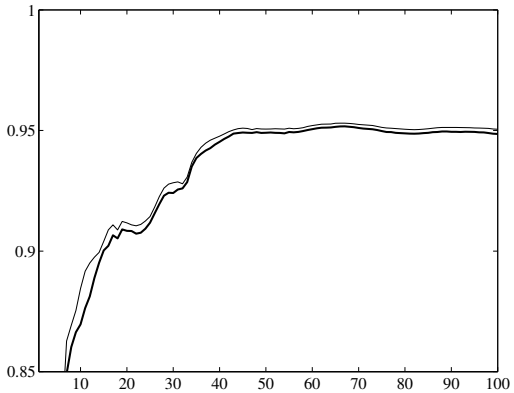
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

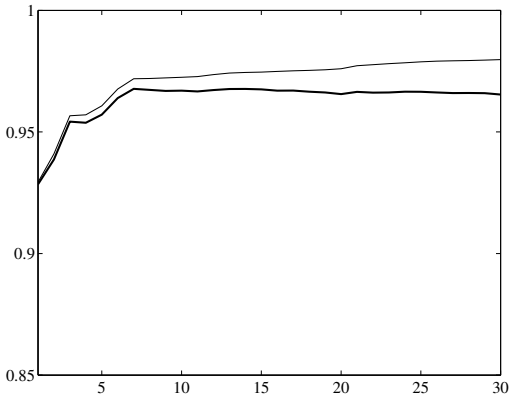
Синдромный алгоритм (K — число признаков):

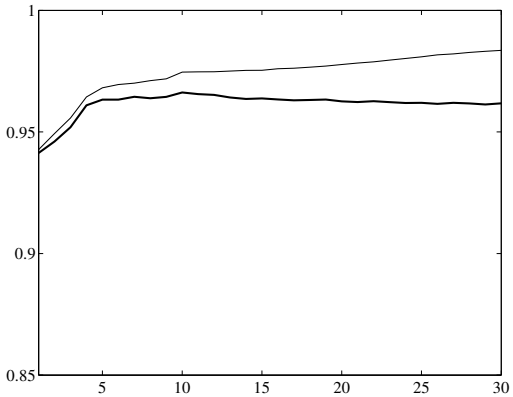


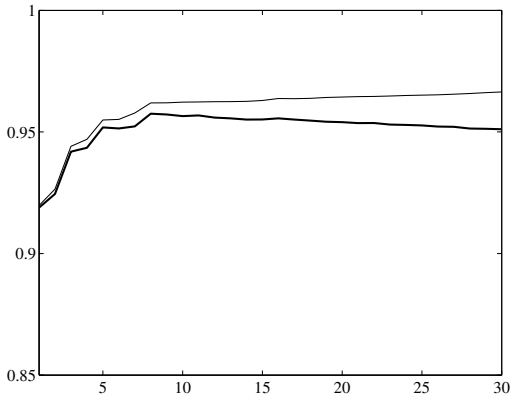




Логистическая регрессия (K — число главных компонент):





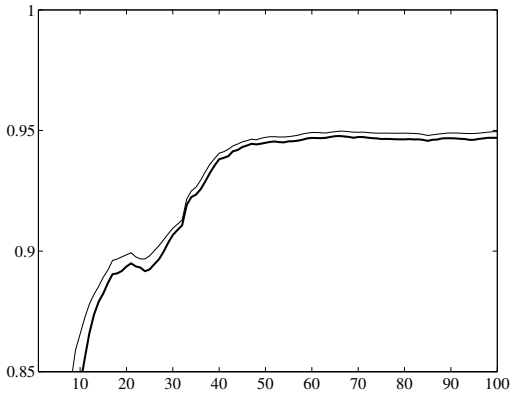


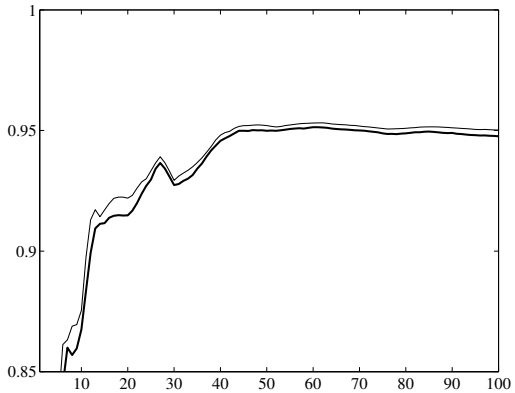
рак общий | аденома простаты | зоб щитовидной железы

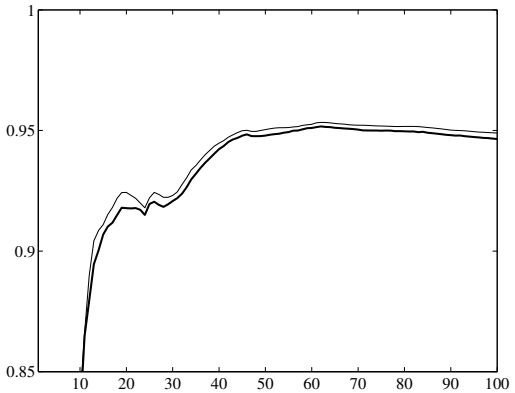
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

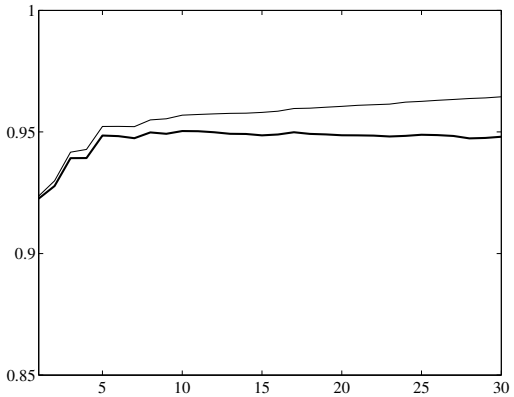
Синдромный алгоритм (K — число признаков):

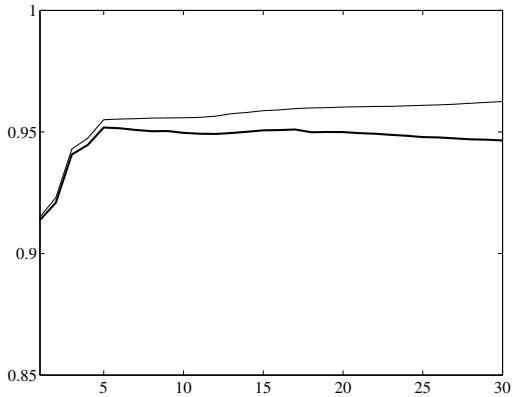


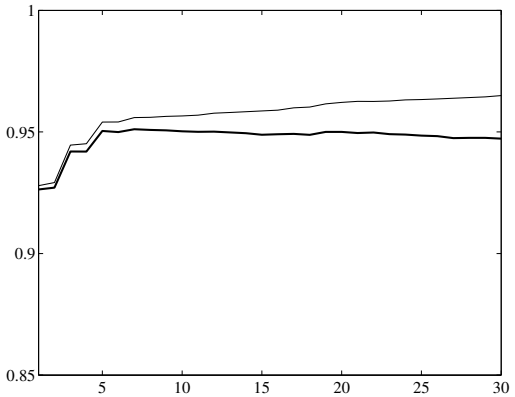




Логистическая регрессия (K — число главных компонент):





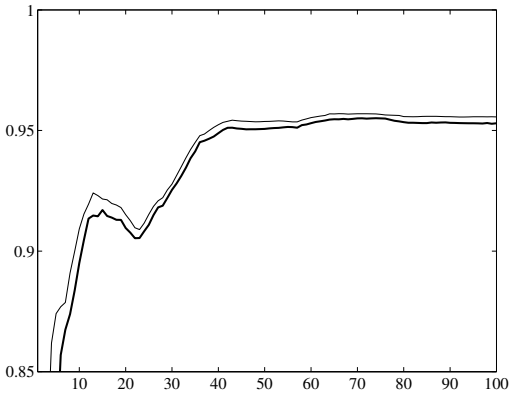


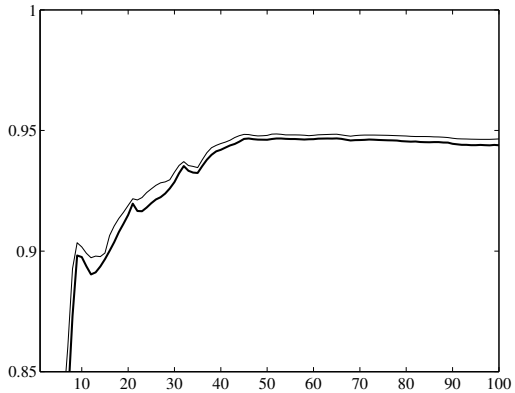
хронический гастрит 2 дискинезия ЖВП мочекаменная болезнь

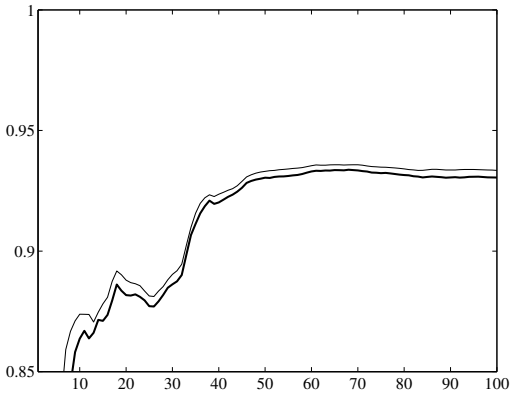
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

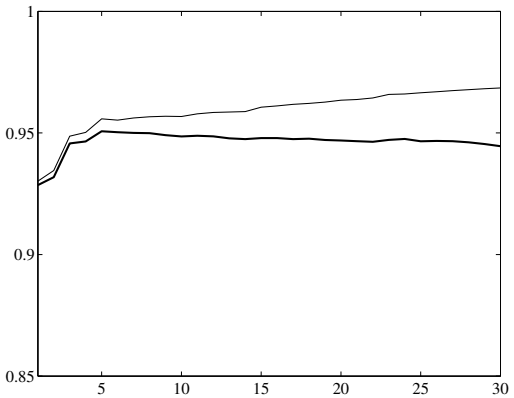
Синдромный алгоритм (K — число признаков):

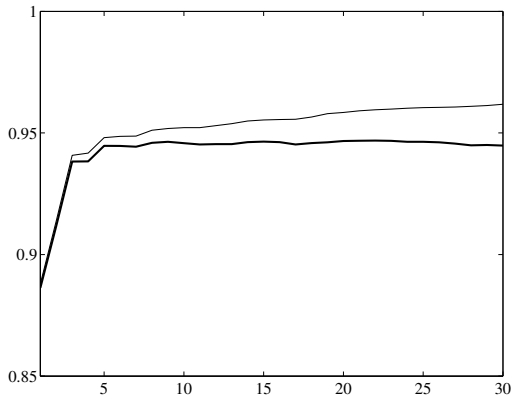


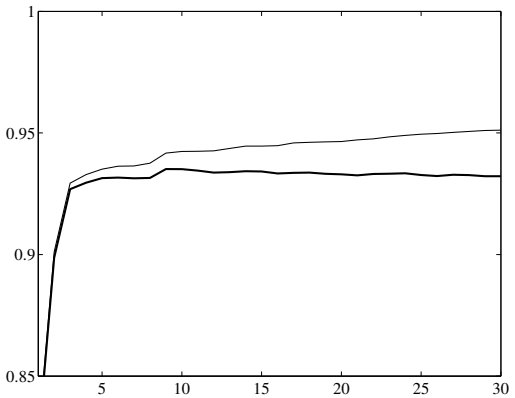




Логистическая регрессия (K — число главных компонент):





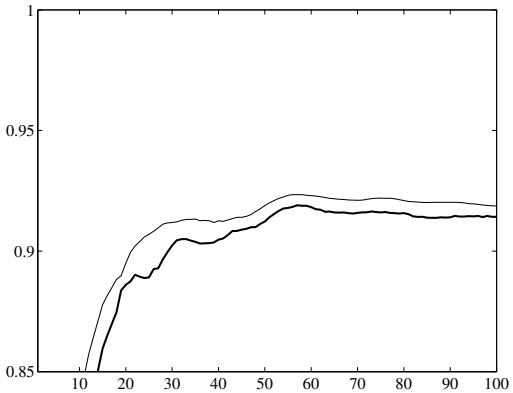


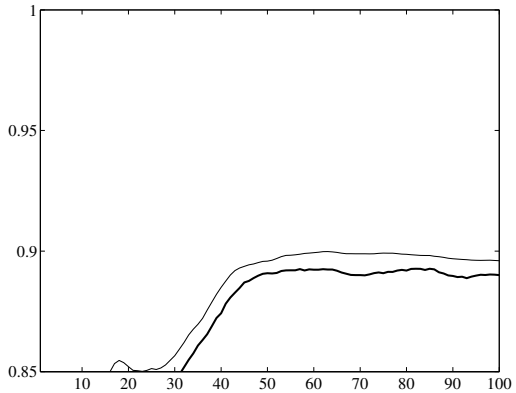
хронический холецистит | язвенная болезнь | миома матки

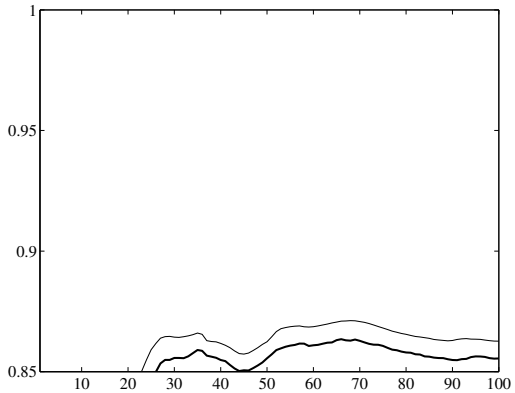
Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимости AUC от числа используемых признаков K

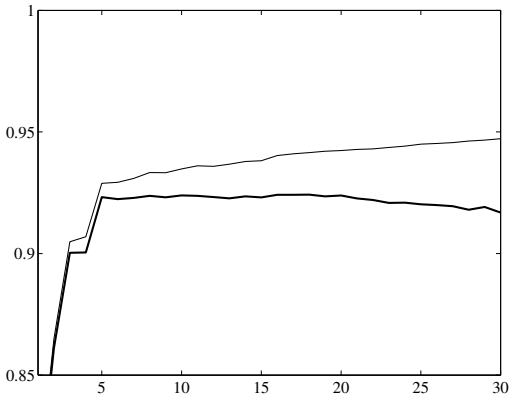
Синдромный алгоритм (K — число признаков):

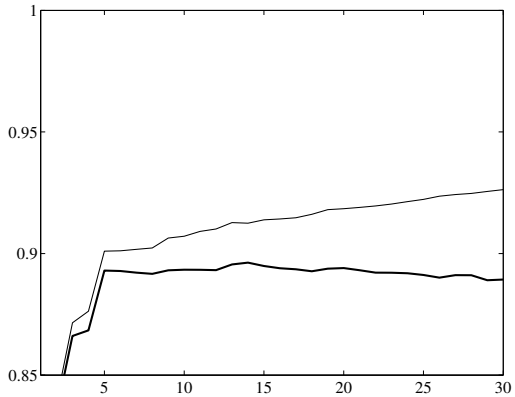


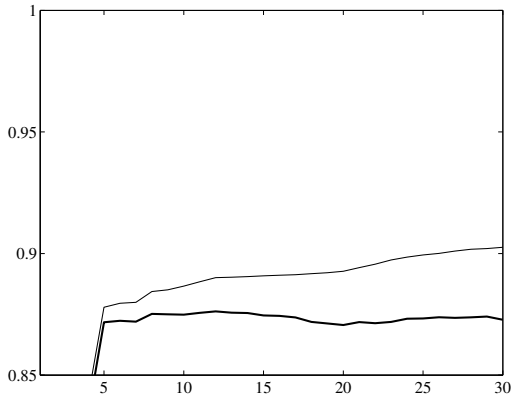




Логистическая регрессия (K — число главных компонент):



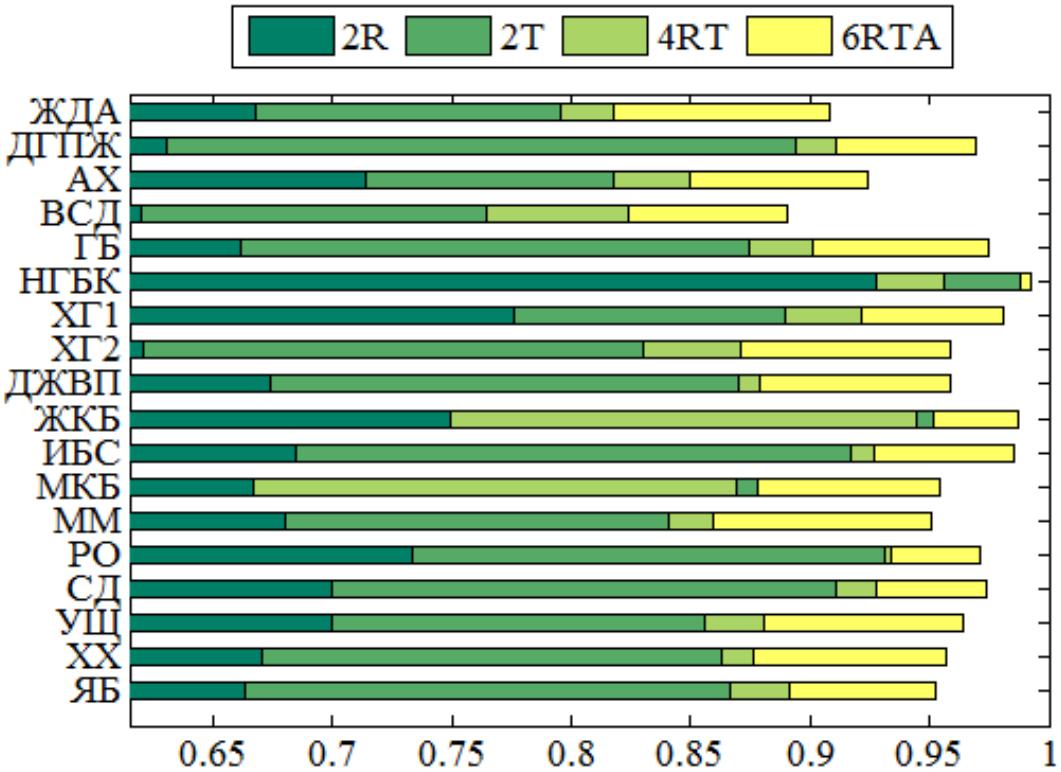




хронический аднексит | анемия | вегетососудистая дистония

Тонкая (верхняя) линия — на обучающей выборке
Толстая (нижняя) линия — на тестовой выборке

Зависимость AUC от типа символьного кодирования



2R: 2-символьная, только приращения амплитуд

2T: 2-символьная, только приращения интервалов

4RT: 4-символьная, приращения интервалов и амплитуд

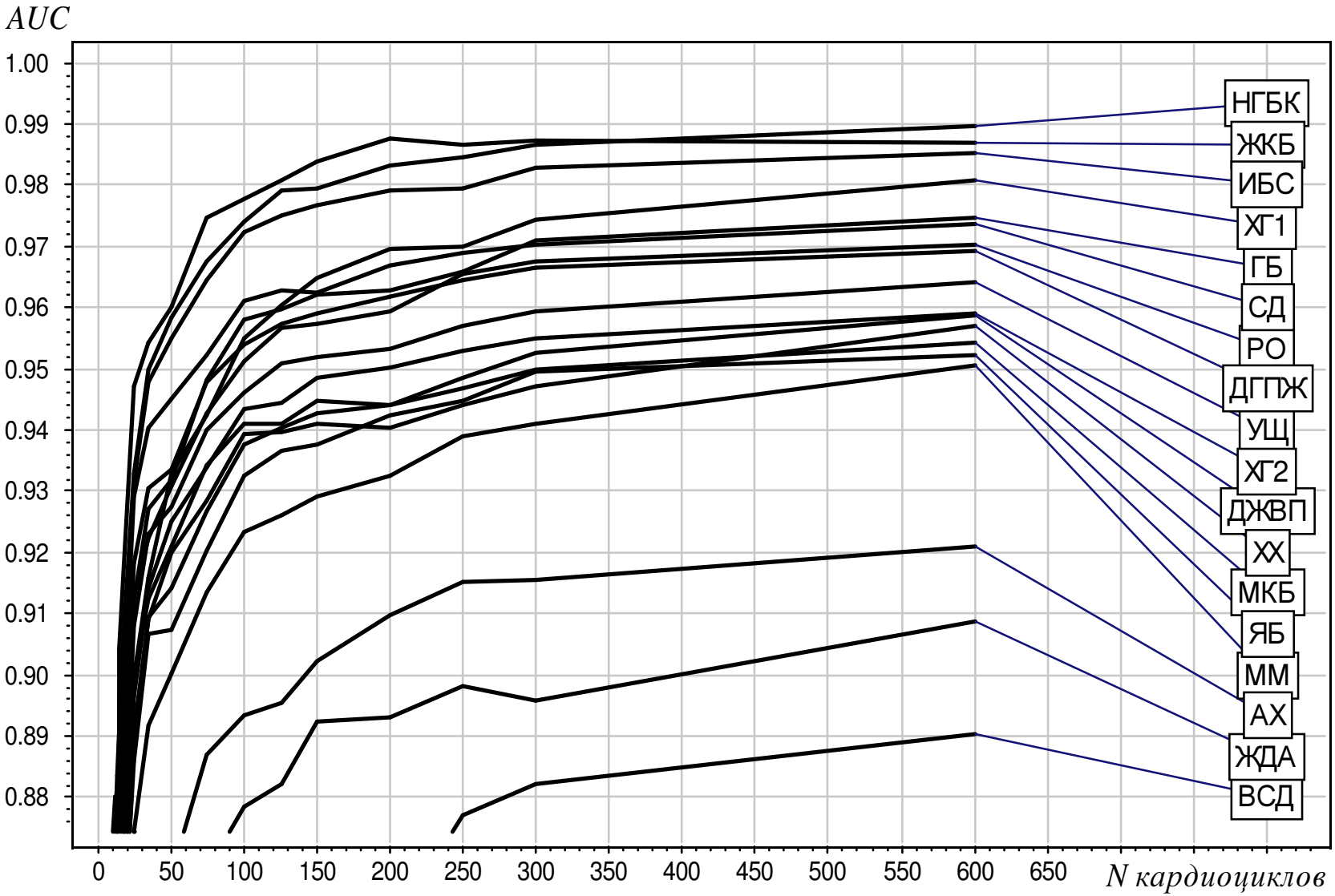
6RTA: 6-символьная, приращения интервалов, амплитуд и их отношений

Выбор модели классификации и ширины окна h-грамм

	HM:3	SA:2	SA:3	SA:4	LR:2	LR:3	LR:4	RF:2	RF:3	RF:4
НГБК	98,42	98,72	99,44	99,59	99,02	99,25	99,48	98,99	99,35	99,34
ЖКБ	97,19	98,45	99,03	98,98	98,99	99,19	99,25	98,67	98,87	98,53
ИБС	95,04	97,31	98,23	98,69	98,15	98,54	98,94	98,64	98,56	98,21
ХГ1	93,85	96,98	98,40	98,57	97,43	98,18	98,59	98,58	98,12	97,80
ГБ	89,80	95,39	96,94	97,34	96,68	97,05	97,43	97,90	97,57	96,78
СД	93,54	96,25	96,63	96,99	96,92	97,06	97,72	97,54	97,42	97,16
ДГПЖ	94,50	95,75	96,79	97,08	96,01	96,61	97,04	97,08	96,58	96,58
РО	—	94,42	96,49	96,78	95,64	97,22	97,78	97,35	97,49	97,29
УЩ	88,95	94,40	95,36	95,84	95,38	95,88	95,87	96,78	96,15	95,43
ХХ	90,61	95,12	96,13	96,23	95,76	95,66	95,84	96,02	95,58	95,20
ДЖВП	88,40	94,59	95,82	96,01	95,23	95,75	95,97	96,32	95,94	94,90
МКБ	75,64	94,43	95,46	95,81	94,91	95,28	95,58	95,68	95,42	94,81
ЯБ	91,54	94,06	95,18	95,64	94,60	95,25	95,65	95,88	95,68	94,72
ММ	79,45	91,81	93,47	93,29	92,47	93,71	93,96	95,25	95,04	94,25
АХ	—	90,84	92,31	92,18	91,53	92,91	92,91	92,70	92,73	92,26
ЖДА	77,89	89,40	91,01	90,61	90,51	91,46	91,21	91,55	90,95	90,42

- методы машинного обучения отличаются не существенно,
- но заметно превосходят полу-автоматический метод (HM)
- RF лидирует и лучше работает на биграммах

Зависимость AUC от длительности регистрации ЭКГ



Открытые данные по инфарктам миокарда: база данных PTB

Данные национального метрологического института Германии.

Число записей ЭКГ-сигналов: 320 больных, 74 здоровых.
Длительность регистрации ЭКГ: 100–200 кардиоциклов.

AUC при 6-символьном кодировании (6RTA) для трёх методов:
LR — логистическая регрессия,
RF — случайный лес,
SA — наивный Байес с отбором признаков

	LR	RF	SA
2-граммы	87.7	87.9	86.1
3-граммы	89.4	**89.6**	87.1
4-граммы	88.6	87.7	86.9

Bousseljot R., Kreiseler D., Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik. 1995.

Эксперименты по воспроизводимости кодограмм

Опр. Кодограммы *однородны*, если они имеют равные распределения вероятностей триграмм.

Воспроизводимость — однородность в повторных обследованиях.

- Данные
  - 7626 записей прибора **Скринфакс**, с диагнозами
  - 4918 записей прибора **CardioQvark**, без диагнозов
- Статистические тесты — критерии согласия:
  - FET: точный тест Фишера,
  - G-тест,
  - Z-тест
- Поправки на множественную проверку гипотез:
  - Holm: метод Холма
  - BH: Метод Бенджамини-Хохберга

Однородность кодограммы в одном обследовании

Доля кодограмм, однородных в пределах одного обследования:

			поправки	
		Pvalue	**Holm**	**BH**
Скринфакс	**FET**	0,744	0,950	0,861
	G-тест	0,973	0,993	0,991
	Z-тест	0,974	—	—
CardioQvark	**FET**	0,887	0,990	0,971
	G-тест	0,999	0,999	0,999
	Z-тест	0,997	—	—

Вывод:

- кодограммы в пределах одного обследования однородны.

Эксперименты по оцениванию мощности критериев:

- FET наиболее мощный против альтернативной гипотезы H_1: частоты m триграмм равны нулю

Однородность кодограмм в разных обследованиях

Доля пар кодограмм, однородных в разных обследованиях:

				поправки	
			Pvalue	**Holm**	**BH**
пары обследований одного человека	**Скринфакс**	**FET**	0,104	0,221	0,106
		G-тест	0,285	0,408	0,294
		Z-тест	0,275	—	—
	CardioQvark	**FET**	0,342	0,676	0,368
		G-тест	0,746	0,894	0,814
		Z-тест	0,712	—	—
пары обследований разных людей	**Скринфакс**	**FET**	0,002	0,010	0,002
		G-тест	0,051	0,095	0,052
		Z-тест	0,043	—	—
	CardioQvark	**FET**	0,060	0,207	0,061
		G-тест	0,349	0,629	0,381
		Z-тест	0,338	—	—

Выводы:

- кодограммы одного человека не всегда однородны
- кодограммы разных людей, как правило, неоднородны
- **Скринфакс** точнее: запись ЭКГ под наблюдением врача

Однородность кодограмм, снятых разными приборами

Данные: 2×23 синхронных записей **Скринфакс** и **CardioQvark**

Доля однородных пар кодограмм:

FET	0,96
G-тест	0,96
Z-тест	0,91

Выводы:

- кодограммы двух приборов можно смешивать при формировании обучающих и тестовых выборок

Проблемы:

- имеющаяся выборка синхронных записей слишком мала
- в данных **Скринфакс** диагнозы есть, в **CardioQvark** — нет
- возможно ли обучить классификатор по данным одного прибора и применять к данным другого прибора?

Однородность обучающих выборок от разных приборов

Схема эксперимента:

- вероятностные модели различий в данных двух приборов:
  - $\Delta(T_t)$: разность интервалограмм
  - $\Delta(R_t)$: разность амплитудограмм
  - $\Delta\big(f_j(x)\big)$: разность векторов частот триграмм
- тестирование адекватности вероятностной модели:
  H_0: различия случайны и не зависят от обследования
- эмуляция выборки **CardioQvark** по данным **Скринфакс**
- сравнение классификаторов, обученных по двум выборкам

Результат: доля пар обследований, в которых H_0 отвергается:

	p-value	**Holm**	**BH**
$\Delta(R_t)$	1,000	1,000	1,000
$\Delta(T_t)$	0,600	0,316	0,547
$\Delta\big(f_j(x)\big)$	0,058	0,000	0,000

Вывод: только разности частот подходят для эмуляции.

Однородность обучающих выборок от разных приборов

Алгоритмы классификации «здоровый-больной», 20 болезней:

- SA: синдромный алгоритм
- RF: случайный лес

Средний AUC (%) на контроле:

	обучение на данных Скринфакс		**обучение на эмуляции CardioQvark**	
	RF	SA	RF	SA
данные Скринфакс	95,34	94,38	94,92	94,30
эмуляция CardioQvark	94,97	93,94	94,73	93,94
наибольшая разность	1,15	1,13	0,54	0,98
средняя разность	0,38	0,44	0,22	0,36

- классификаторы, обученные по данным **Скринфакс** и эмулированным данным **CardioQvark**, отличаются мало
- можно объединять обучающие выборки двух приборов
- можно обучаться по данным одного прибора и использовать для диагностики другой прибор

Выводы

- Удивительно высокая точность диагностики
  - качество диагностики подтверждено кросс-валидацией
  - в том числе в экспериментах на открытых данных
- Даны статистические обоснования основных элементов *технологии информационного анализа ЭКГ-сигналов*:
  - выбор диагностических признаков болезней
  - выбор длительности регистрации сигнала (300–600)
  - выбор оптимального алфавита и типа кодирования
  - выбор методов машинного обучения
- Статистические тесты однородности:
  - интервалограммы и амплитудограммы не воспроизводимы
  - векторы частот триграмм воспроизводимы
  - можно объединять обучающие выборки двух приборов