Проектная смена СИРИУС 2016

МЕДИЦИНСКАЯ ДИАГНОСТИКА ПО ЭЛЕКТРОКАРДИОГРАММЕ

Извлекаем пользу из Big Data

РУКОВОДИТЕЛЬ:
д.ф.-м.н., профессор РАН
Воронцов Константин Вячеславович
АССИСТЕНТ
Темирчев Павел Георгиевич

УЧАСТНИКИ:
Максим Деб Натх
Святослав Дженжер
Никита Ермаков
Роман Москаленко
Анна Нестеренко
Адьян Очиров
Николай Сафонов
София Семенова-Звенигородская
Андрей Шлапко

ПРОБЛЕМА

Совершенствование технологии информационного анализа электрокардиосигналов для скрининговой диагностики

АКТУАЛЬНОСТЬ

Здравоохранению нужна система *ранней диагностики* широкого спектра заболеваний внутренних органов человека

СУЩЕСТВУЮЩИЕ РЕШЕНИЯ

- Обычно ЭКГ используют для диагностики заболеваний сердца.
- Диагностическая система «Скринфакс»: 15 лет эксплуатации, накоплено 15 тыс. записей с диагнозами по 40 заболеваниям (но диагностические правила до сих пор строятся вручную)

НОВЫЕ РЕШЕНИЯ

Применение современных методов машинного обучения для построения диагностических правил по обучающим выборкам

ЦЕЛЬ ПРОЕКТА

Повышение точности диагностики заболеваний внутренних органов человека с помощью машинного обучения и технологии информационного анализа электрокардиосигналов

ЗАДАЧИ ПРОЕКТА

1. Придумать и проверить новые способы символьного кодирования ЭКГ
2. Адаптировать и применить методы поиска закономерностей в символьных последовательностях, применяемые в биоинформатике
3. Реализовать «наивный» линейный классификатор с отбором признаков
4. Сравнить с методами машинного обучения из Python scikit-learn
5. Придумать и проверить несколько композитных решений

ФАЗЫ ПРОЕКТА

1. Соревнование алгоритмов на платформе Kaggle in Class
2. Соревнование композитных решений

МЕТОДЫ

Сбор данных

- Данные были собраны заранее с помощью системы Скринфакс
- Диагнозы ставились в клиниках Москвы и Санкт-Петербурга

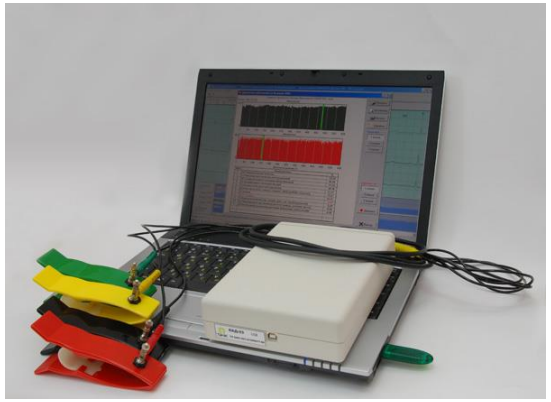
Обработка и анализ данных

- Методы дискретизации сигналов (символьная динамика)
- Методы машинного обучения, кросс-валидация

МАТЕРИАЛЫ И ОБОРУДОВАНИЕ

- Цифровой электрокардиограф Скринфакс:
  - усиленная помехозащищённость
  - полоса пропускания: 0,1—500 Гц
  - частота дискретизации: 1кГц

- Исходные данные:
  - 6 классов: здоровые + 5 болезней
  - обучающая выборка: 2515 записей
  - тестовая выборка: 595 записей

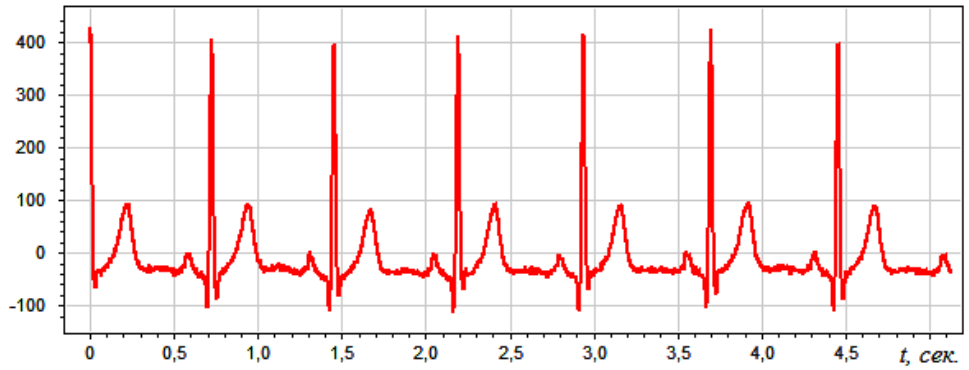


Скринфакс -- система скрининговой диагностики

д.м.н., проф. В.М.Успенский, автор теории информационной функции сердца и технологии информационного анализа электрокардиосигналов

ИСХОДНЫЕ ДАННЫЕ ДЛЯ ОБУЧЕНИЯ АЛГОРИТМОВ КЛАССИФИКАЦИИ

Объекты – записи ЭКГ по 600 кардиоциклов



6 классов

АЗ	«абсолютно здоровые»
ВД	вегетососудистая дистония
ИБ	ишемическая болезнь сердца
СД	сахарный диабет
УЩ	узловой зоб щитовидной железы
ЯБ	язвенная болезнь

Число объектов обучающей выборки

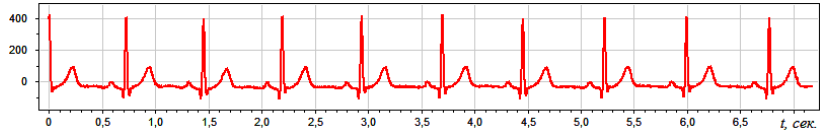
	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ	268	0	0	0	0	0
ВД		447	6	3	17	4
ИБ			783	38	59	37
СД				216	8	8
УЩ					620	15
ЯБ						373

Число объектов тестовой выборки

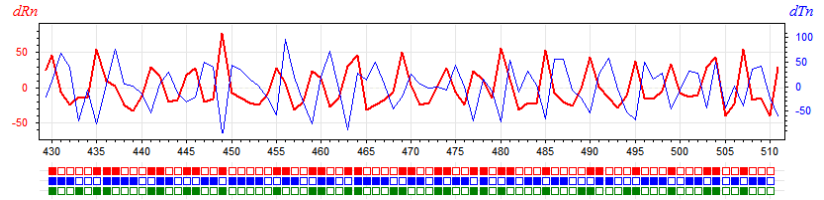
	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ	100	0	0	0	0	0
ВД		106	2	3	7	3
ИБ			135	23	22	9
СД				111	8	10
УЩ					125	11
ЯБ						112

ТРИ ЭТАПА ПЕРДВАРИТЕЛЬНОЙ ОБРАБОТКИ ДАННЫХ

Исходные данные – электрокардиограмма



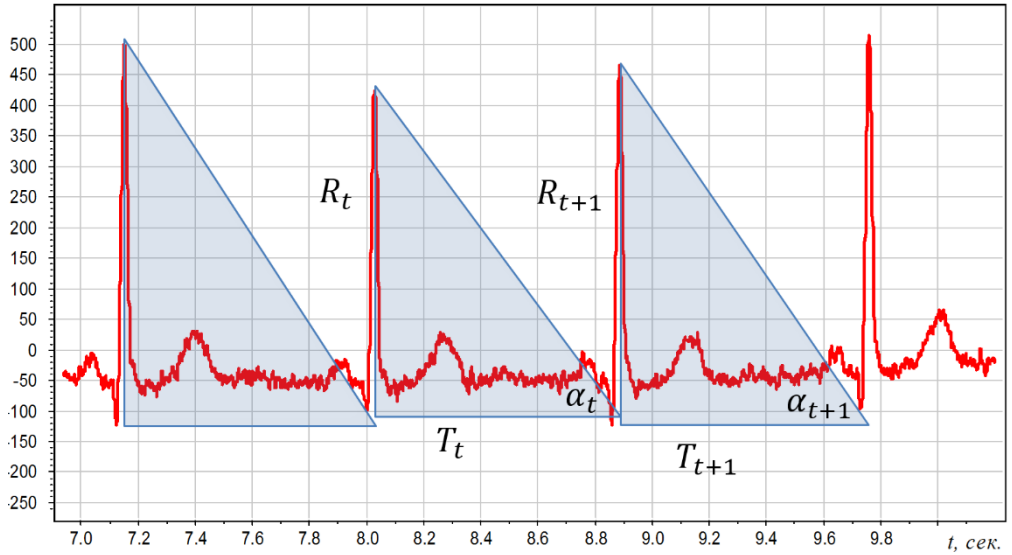
1. Приращения интервалов и амплитуд



2. Кодограмма – символьная последовательность

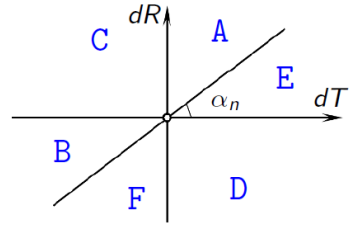
3. Векторы частот *n*-грамм

Интервалы и амплитуды кардиоциклов



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



ЭВРИСТИЧЕСКИЙ ПОИСК ДИАГНОСТИЧЕСКИХ ПРАВИЛ В ТЕКУЩЕЙ ВЕРСИИ «СКРИНФАКС»

Признаки – частоты триграмм, $n = 6^3 = 216$

Диагностический эталон – множество триграмм, совместно встречающихся у многих больных и ни у одного здорового

Обучение – построение множества диагностических эталонов для каждой болезни

Классификатор: score – доля из $|E|$ самых частых триграмм в кодограмме, которые входят в диагностический эталон E.

	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ		52,8	59,4	61,1	54,8	55,1
ВД	66,3		46,9	46,7	52,2	51,7
ИБ	93,3	67,9		63,7	60,8	60,8
СД	86,9	63,0	44,1		52,3	63,6
УЩ	77,4	60,9	47,4	51,9		55,2
ЯБ	80,1	63,1	47,2	52,9	51,2	

Средний AUC = 60,1%

Недостатки:

- Слишком простая многоклассовая стратегия «каждая болезнь против класса здоровых».
- Отбор диагностических эталонов автоматизирован лишь частично, необходимо привлечение эксперта.

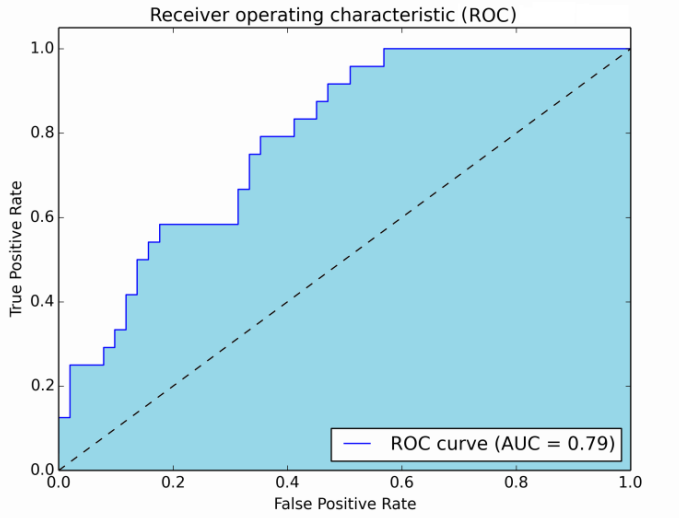
Пути устранения недостатков:

- Многоклассовая стратегия «каждый против всех»
- Машинное обучение.

КРИТЕРИЙ КАЧЕСТВА МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ

Стратегия «к каждый против вс всех остальных ешаем 6 двухклассовых задач классификации

	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ						
ВД						
ИБ						
СД						
УЩ						
ЯБ						



Для каждого класса вычисляется AUC – площадь под ROC-кривой. Затем усредняется по классам

Задача классификации с пересекающимися классами:

$$(x_i, Y_i)_{i=1}^{\ell}, \quad x_i \in \mathbb{R}^n, \quad Y_i \subset Y = \{0, 1, \ldots, 5\}.$$

Линейная модель многоклассовой классификации:

$$a_y(x) = \left[\text{score}(x, w_y) > w_{0y}\right], \quad y \in Y.$$

Средний AUC по классам при стратегии «каждый против всех»

$$\text{AUC} = \frac{1}{|Y|} \sum_{y \in Y} \frac{1}{|D|} \sum_{(i,j) \in D} \left[\text{score}(x_i, w_y) > \text{score}(x_j, w_y)\right],$$

где $D = \{(i,j) \mid y \in Y_i \text{ и } y \notin Y_j\}$ — множество пар объектов.

«НАИВНЫЙ» ЛИНЕЙНЫЙ КЛАССИФИКАТОР

Признаки x^j – частоты триграмм, $n = 6^3 = 216$

Линейная модель многоклассовой классификации:

$$a_y(x) = \left[\text{score}(x, w_y) > w_{0y}\right], \quad \text{score}(x, w_y) = \sum_{j=1}^{n} w_{yj} x^j$$

Веса признаков вычисляются независимо друг от друга.
Варианты эвристических формул для весов w_{yj}:

$$w_{yj} = \frac{N^j_{y1}}{N^j_{\bar{y}1}} \qquad w_{yj} = \log \frac{N^j_{y1}}{N^j_{\bar{y}1}} \qquad w_{yj} = \sqrt{N^j_{y1}} - \sqrt{N^j_{\bar{y}1}}$$

N^j_{yz} — доля объектов класса y, у которых $z = \left[x^j_i \geqslant A\right]$.

	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ		81,8	94,6	92,1	91,1	87,6
ВД	48,5		71,6	73,5	71,7	69,9
ИБ	94,0	78,2		51,1	59,1	65,0
СД	91,6	76,3	49,6		58,9	64,1
УЩ	76,7	69,3	52,5	51,6		53,1
ЯБ	66,2	56,6	59,5	58,5	60,7	

Средний AUC = 69,5%

Результат: Качество гораздо выше, хотя никак не учитываются зависимости между признаками.

«НАИВНЫЙ» ЛИНЕЙНЫЙ КЛАССИФИКАТОР С ОТБОРОМ ПРИЗНАКОВ

Признаки x^j – частоты триграмм, $n = 6^3 = 216$.
Линейная модель многоклассовой классификации.
Жадный отбор K признаков с максимальным весом $|w_{yj}|$.

Перебирались три параметра:

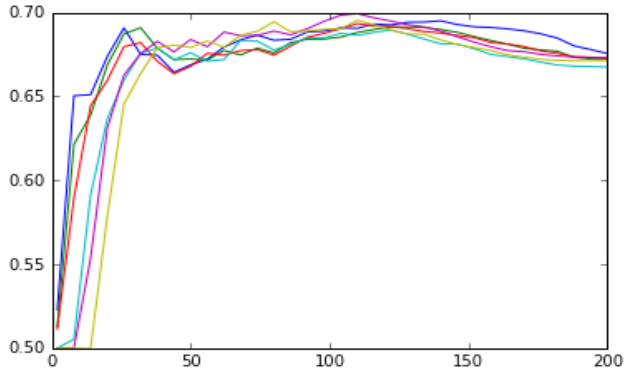
- A – порог частоты триграммы;
- K – число признаков;
- 6 вариантов формулы весов w_{yj} .

	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ		82,0	97,1	96,8	94,5	92,4
ВД	53,4		78,2	68,4	80,0	74,4
ИБ	91,2	80,5		55,3	65,8	61,8
СД	87,7	79,0	52,2		64,4	62,4
УЩ	77,7	69,6	41,4	56,8		49,1
ЯБ	78,3	61,3	43,6	44,6	58,2	

Средний AUC = 69,9%

Результаты:

- Отбор признаков немного улучшает качество, устойчивый оптимум при $A = 5..9$ и $K = 70..100$
- Формула весов $w_{yj} = N^j_{y1} \;/\; N^j_{\bar{y}1}$



Зависимость AUC от числа признаков K при различных значениях параметра A

«НАИВНЫЙ» ЛИНЕЙНЫЙ КЛАССИФИКАТОР С ОТБОРОМ ВЫСОКОЧАСТОТНЫХ ПРИЗНАКОВ

Признаки x^j – частоты триграмм, $n = 6^3 = 216$.
Линейная модель многоклассовой классификации.

Перебирались три параметра:

- A – порог частоты триграммы;
- K – число признаков;
- 6 вариантов формулы весов w_{yj} .

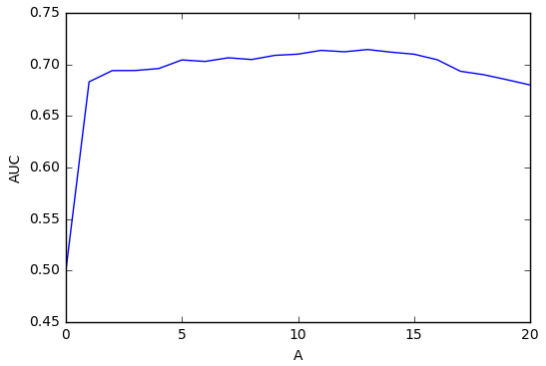
Идея: повышение порога A >10 оставляет только самые информативные признаки

Результаты:

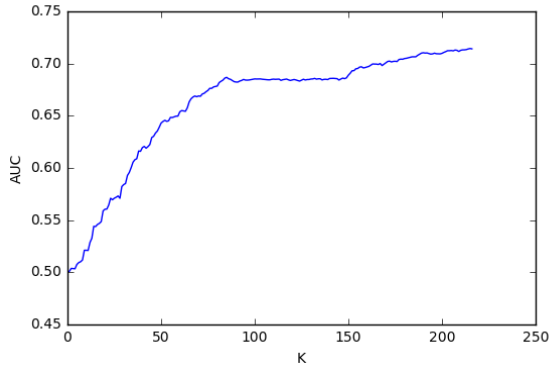
- Устойчивый оптимум при $A^* = 13$ и $K^* = 215$ (почти все признаки)
- Формула весов $w_{yj} = \ln N^j_{y1} - \ln N^j_{\bar{y}1}$

	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ		83,3	95,1	92,1	92,5	88,6
ВД	55,5		73,9	77,6	74,1	71,6
ИБ	96,5	80,6		52,4	59,4	65,0
СД	93,4	80,5	53,7		63,4	68,1
УЩ	88,5	76,3	51,7	54,1		56,4
ЯБ	85,1	56,7	47,5	48,2	51,5	

Средний AUC = 71,4%



Зависимость AUC от A при K^*



Зависимость AUC от K при A^*

ЛИНЕЙНЫЙ КЛАССИФИКАТОР – ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Признаки – частоты триграмм, $n = 6^3 = 216$.
Линейная модель многоклассовой классификации.

Стандартная реализация из Python scikit-learn

	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ		83,5	95,4	92,1	92,0	87,3
ВД	65,6		78,9	80,8	75,7	70,5
ИБ	90,8	77,0		53,1	63,0	64,3
СД	86,8	72,4	54,2		65,0	65,7
УЩ	70,4	64,9	55,4	59,1		59,5
ЯБ	66,1	57,4	53,3	48,9	57,6	

Средний AUC = 70,4%

Дополнительно сделан отбор признаков,
оптимум при $K = 111$ из 216 триграмм.

	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ		84,5	96,6	93,3	94,2	90,2
ВД	63,2		78,8	79,0	76,6	71,5
ИБ	93,2	78,5		53,5	63,0	68,5
СД	87,5	72,0	59,0		67,3	69,7
УЩ	76,8	66,1	53,6	57,5		56,2
ЯБ	70,8	60,2	57,8	56,2	59,4	

Средний AUC = 71,9%

Результат: Качество улучшается незначительно по сравнению с «наивным» линейным классификатором.

ЛИНЕЙНЫЙ КЛАССИФИКАТОР – ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ + РЕГУЛЯРИЗАЦИЯ

Признаки – частоты триграмм, $n = 6^3 = 216$.
Линейная модель многоклассовой классификации.

Стандартная реализация из Python scikit-learn,
L_2-регуляризация,
определение константы регуляризации по кросс-валидации.

Отбора признаков нет.

	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ		96,6	94,2	89,5	93,2	85,2
ВД	94,7		63,0	69,7	53,1	81,6
ИБ	80,0	56,8		58,7	60,9	68,2
СД	72,5	54,0	55,4		51,8	55,9
УЩ	91,4	55,1	69,6	71,8		80,6
ЯБ	64,7	80,6	76,9	72,3	81,5	

Средний AUC = 72,8%

Результат:

- учёт зависимости между признаками + регуляризация для компенсации переобучения

работает на 1,4% лучше, чем

- предположение о независимости признаков + отбор признаков.

… и пока это наилучшее решение!

НОВАЯ КОДИРОВКА С 12-СИМВОЛЬНЫМ АЛФАВИТОМ

Идея: добавим приращения через два кардиоцикла:

$$\Delta\alpha_2 = \alpha_{t+1} - \alpha_{t-1}$$

Вместо 6-символьного алфавита получим 12-символьный.
Признаки – частоты биграмм, $n = 12^2 = 144$.

Линейная модель многоклассовой классификации.
Логистическая регрессия из Python scikit-learn,
L_2-регуляризация, кросс-валидация, отбора признаков нет.

	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ		96,4	93,8	89,0	93,0	84,5
ВД	95,1		60,9	68,6	53,6	81,0
ИБ	82,7	54,0		53,9	55,4	70,6
СД	76,6	54,9	57,5		56,5	52,9
УЩ	92,2	53,9	66,0	71,8		81,4
ЯБ	62,3	78,8	77,0	74,8	81,2	

Средний AUC = 72,4%

Результат:
Замена 3-граммы на 2-граммы сокращает число признаков с 216 до 144.
Тем не менее, каждый признак по-прежнему охватывает 4 кардиоцикла.

Небольшое ухудшение -0,4%. Идея кажется удачной, попробуем её развить.

НОВАЯ КОДИРОВКА С 12-СИМВОЛЬНЫМ АЛФАВИТОМ + СТАРАЯ КОДИРОВКА

Идея: Объединим признаки двух кодировок.
Признаки – частоты 2-грамм и 3-грамм, $n = 6^3 + 12^2 = 360$.

Линейная модель многоклассовой классификации.
Логистическая регрессия из Python scikit-learn,
L_2-регуляризация, кросс-валидация, отбора признаков нет.

	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ		96,6	94,4	89,2	93,1	85,2
ВД	95,4		62,3	69,8	53,2	82,5
ИБ	82,2	56,7		57,5	60,5	68,8
СД	73,0	53,4	59,3		55,5	56,2
УЩ	92,5	57,1	69,9	73,1		81,0
ЯБ	64,8	81,4	77,7	74,1	82,3	

Средний AUC = 73,4%

Результат:
Наилучший результат в проекте!

УПРОЩЁННАЯ КОДИРОВКА С 3-СИМВОЛЬНЫМ АЛФАВИТОМ

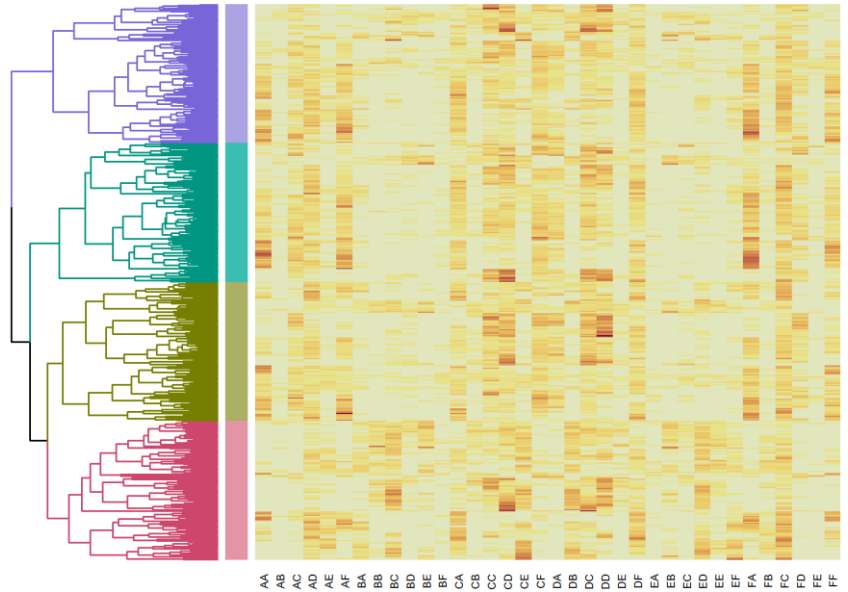
Визуализация: строки – объекты 4х классов (розовый–АЗ), столбцы – признаки в лексикографическом порядке.
Признаки – частоты биграмм, $n = 6^2 = 36$.

Наблюдение: на картинке имеется осевая симметрия! Символы образуют комплементарные пары A-F, B-E, C-D.

Линейная модель, логистическая регрессия из scikit-learn, L_2-регуляризация, кросс-валидация, отбора признаков нет.

Результат: трёхбуквенный алфавит не улучшает AUC:
3-граммы, $n = 33 = 27$, AUC = 68.6%
4-граммы, $n = 34 = 81$, AUC = 68.2%
5-граммы, $n = 35 = 243$, AUC = 68.5%

Идея: возьмём 6-буквенный алфавит и оставим только первую половину признаков ➔ AUC = 73,0%.



	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ		97,1	95,2	90,1	93,6	85,3
ВД	95,2		63,3	71,3	53,2	81,3
ИБ	81,4	56,4		56,5	58,6	72,4
СД	73,2	56,0	55,2		54,9	55,4
УЩ	91,2	56,8	70,1	71,5		78,0
ЯБ	62,8	79,4	77,5	72,0	80,7	

Средний AUC = 73,0%

ОТБРАСЫВАНИЕ ПОЛОВИНЫ ДУБЛИРУЮЩИХ ПРИЗНАКОВ – НЕ ХУЖЕ РЕГУЛЯРИЗАЦИИ

Признаки x^j – частоты триграмм.
Оставим только первую половину признаков
(в лексикографическом порядке), $n = 6^3/2 = 108$.

Наивный линейный классификатор без отбора признаков.

Перебирались два параметра:
- A – порог частоты триграммы;
- 10 вариантов формулы весов w_{yj} .

	АЗ	ВД	ИБ	СД	УЩ	ЯБ
АЗ		96,7	94,9	89,6	93,1	84,8
ВД	95,2		63,3	71,3	53,2	81,3
ИБ	83,2	50,5		54,8	51,8	72,7
СД	77,2	45,6	43,8		44,2	57,1
УЩ	91,3	56,9	70,2	71,5		77,9
ЯБ	62,4	79,4	77,5	71,9	80,8	

Средний AUC = 73,0%

Результат:
Отбрасывание парных сильно зависимых (почти дублирующих) признаков имеет тот же эффект, что регуляризация с кросс-валидацией, и даёт лучший AUC, чем жадный отбор признаков.

Это второй лучший результат в проекте.

ИТОГОВЫЙ РЕЗУЛЬТАТ

Качество диагностики улучшено несколькими методами в смысле среднего по классам значения AUC:

- 60,1% -- базовый метод для текущей версии «Скринфакс»

1. 69,5% -- «наивный » линейный классификатор
2. 69,9% -- он же с отбором признаков по весу
3. 71,4% -- он же с отбором высокочастотных признаков
4. 70,4% -- логистическая регрессия без отбора признаков
5. 71,9% -- логистическая регрессия с отбором признаков
6. 72,8% -- логистическая регрессия с регуляризацией
7. 73,0% -- отбрасывание половины зависимых признаков
8. 73,4% -- логистическая регрессия с 12-символьным кодом

Найдены новые способы 3- и 12-символьного кодирования ЭКГ-сигналов.

Найден новый способ отбора признаков по высокой частоте.

Maxim Deb Natkh	0.75569	44
Pavel Temirchev (MMP, MSU, Russia)	**0.75546**	**32**
Anton Popov	0.75487	28
LadyPython	0.75236	12
Adyan	0.74909	25
Simple Logistic Regression	0.74842	
petrie	0.74610	14
JustLogin	0.73642	4
Nickolay Safonov	0.72528	14
Simple XGBoost	0.69283	
sdjenjer	0.68952	98
Целых Влада (ШАД)	0.59323	2

Результаты соревнования на платформе Kaggle in Class: ники участников проекта, AUC на скрытой выборке, число загрузок.

ПРОЕКТНАЯ СМЕНА СИРИУС 2016

Ученик СУНЦ МГУ Максим Деб Натх рассказывает Президенту В.В.Путину о задаче информационного анализа электрокардиосигналов