МАШИННОЕ ОБУЧЕНИЕ
• распознавание языка текста •
• диагностика по электрокардиограмме •

Воронцов Константин Вячеславович
ФУПМ МФТИ • ВМК МГУ • Яндекс • FORECSYS

Презентация проекта • 2 июля 2016
Сочи, Сириус • Проектная смена • 1–24 июля 2016

Содержание

Декларация прав человека. На каких языках?

Статья 1. Все люди рождаются свободными и равными в своем достоинстве и правах. Они наделены разумом и совестью и должны поступать в отношении друг друга в духе братства.

Стаття 1. Всі люди народжуються вільними і рівними у своїй гідності та правах. Вони наділені розумом і совістю і повинні діяти у відношенні один до одного в дусі братерства.

Article 1. All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

Article 1. Tous les êtres humains naissent libres et égaux en dignité et en droits. Ils sont doués de raison et de conscience et doivent agir les uns envers les autres dans un esprit de fraternité.

Декларация прав человека. На каких языках?

rus: Russian

Статья 1. Все люди рождаются свободными и равными в своем достоинстве и правах. Они наделены разумом и совестью и должны поступать в отношении друг друга в духе братства.

ukr: Ukrainian

Стаття 1. Всі люди народжуються вільними і рівними у своїй гідності та правах. Вони наділені розумом і совістю і повинні діяти у відношенні один до одного в дусі братерства.

eng: English

Article 1. All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

frn: French

Article 1. Tous les êtres humains naissent libres et égaux en dignité et en droits. Ils sont doués de raison et de conscience et doivent agir les uns envers les autres dans un esprit de fraternité.

Декларация прав человека. На каких языках?

Artikel 1. Alle Menschen sind frei und gleich an Würde und Rechten geboren. Sie sind mit Vernunft und Gewissen begabt und sollen einander im Geiste der Brüderlichkeit begegnen.

Artikel 1. Alle menslike wesens word vry, met gelyke waardigheid en regte, gebore. Hulle het rede en gewete en behoort in die gees van broederskap teenoor mekaar op te tree.

Artículo 1. Todos los seres humanos nacen libres e iguales en dignidad y derechos y. Dotados como están de razón y conciencia, deben comportarse fraternalmente los unos con los otros.

Artigo 1. Todos os seres humanos nascem livres e iguais em dignidade e em direitos. Dotados de razão e de consciência, devem agir uns para com os outros em espírito de fraternidade.

Декларация прав человека. На каких языках?

ger: German

Artikel 1. Alle Menschen sind frei und gleich an Würde und Rechten geboren. Sie sind mit Vernunft und Gewissen begabt und sollen einander im Geiste der Brüderlichkeit begegnen.

afk: Afrikaans

Artikel 1. Alle menslike wesens word vry, met gelyke waardigheid en regte, gebore. Hulle het rede en gewete en behoort in die gees van broederskap teenoor mekaar op te tree.

spn: Spanish

Artículo 1. Todos los seres humanos nacen libres e iguales en dignidad y derechos y. Dotados como están de razón y conciencia, deben comportarse fraternalmente los unos con los otros.

por: Portuguese

Artigo 1. Todos os seres humanos nascem livres e iguais em dignidade e em direitos. Dotados de razão e de consciência, devem agir uns para com os outros em espírito de fraternidade.

Декларация прав человека. На каких языках?

Artikla 1. Kaikki ihmiset syntyvät vapaina ja tasavertaisina arvoltaan ja oikeuksiltaan. Heille on annettu järki ja omatunto, ja heidän on toimittava toisiaan kohtaan veljeyden hengessä.

Artikkel 1. Kõik inimesed sünnivad vabadena ja võrdsetena oma väärikuselt ja õigustelt. Neile on antud mõistus ja südametunnistus ja nende suhtumist üksteisesse peab kandma vendluse vaim.

Artikel 1. Alla människor är födda fria och lika i värde och rättigheter. De har utrustats med förnuft och samvete och bör handla gentemot varandra i en anda av gemenskap.

Artikkel 1. Alle menneske er fødde til fridom og med same menneskeverd og menneskerettar. Dei har fått fornuft og samvit og skal leve med kvarandre som brør.

Декларация прав человека. На каких языках?

fin: Finnish

Artikla 1. Kaikki ihmiset syntyvät vapaina ja tasavertaisina arvoltaan ja oikeuksiltaan. Heille on annettu järki ja omatunto, ja heidän on toimittava toisiaan kohtaan veljeyden hengessä.

est: Estonian

Artikkel 1. Kõik inimesed sünnivad vabadena ja võrdsetena oma väärikuselt ja õigustelt. Neile on antud mõistus ja südametunnistus ja nende suhtumist üksteisesse peab kandma vendluse vaim.

swd: Swedish

Artikel 1. Alla människor är födda fria och lika i värde och rättigheter. De har utrustats med förnuft och samvete och bör handla gentemot varandra i en anda av gemenskap.

nrn: Norwegian

Artikkel 1. Alle menneske er fødde til fridom og med same menneskeverd og menneskerettar. Dei har fått fornuft og samvit og skal leve med kvarandre som brør.

Задача «Language Identification»

Как обучить машину определять язык текста автоматически?

Зачем это нужно:

- ~~Это просто прикольно~~
- Поисковые системы
- Системы агрегации контента
- Системы автоматического перевода

Постановка задачи

Дано:
обучающая выборка текстов с известными *классификациями*:

$$\langle \text{текст}_1, \text{язык}_1 \rangle, \langle \text{текст}_2, \text{язык}_2 \rangle, \ldots, \langle \text{текст}_\ell, \text{язык}_\ell \rangle$$

Найти:
Правило (функцию, алгоритм) классификации любого текста

$$\text{текст} \xrightarrow{?} \text{язык}$$

Критерий качества решения:
Алгоритм должен как можно реже ошибаться.

Задача поставлена, когда у неё есть Д.Н.К.

Математическая модель классификации текстов

Каждый язык имеет уникальное распределение частот N-грамм.

Все люди рождаются свободными и равными в своем достоинстве и...

л, ю, д, и — униграммы

лю, юд, ди — биграммы

люд, юди — триграммы

Линейная модель классификации:

Оценка близости текста к языку по всем N-граммам x:

$$\text{оценка}(\text{текст}, \text{язык}) := \sum_x \text{вес}[x, \text{язык}] \cdot \text{частота}[x, \text{текст}]$$

Правило классификации: отнести текст к тому языку, для которого оценка(текст, язык) наибольшая.

Машинное обучение настраивает оптимальные веса N-грамм в языках по обучающей выборке.

10 самых частых триграмм в 7 языках

В этом эксперименте

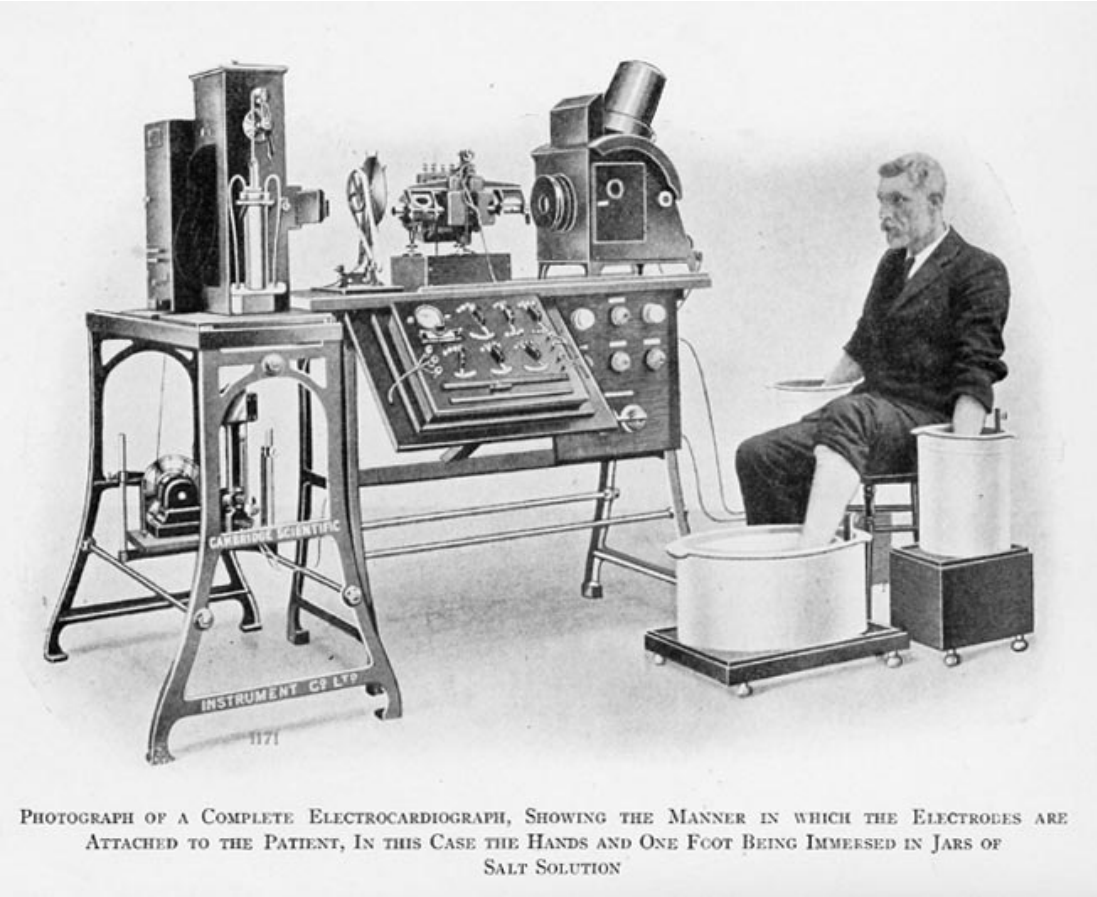
- использовались только тексты Декларации,
- использовались только языки на основе латиницы,
- все диакритические знаки и пробел были заменены на «-»

	1	2	3	4	5	6	7	8	9	10
английский	-an	and	nd-	the	-th	he-	ion	of-	-of	tio
французский	-de	es-	de-	le-	et-	ion	nt-	tio	-et	te-
немецкий	en-	ein	er-	der	ine	nd-	cht	ung	-un	ich
африкаанс	ie-	die	-di	en-	ing	ng-	an-	et-	-re	reg
финский	ise	sta	an-	en-	ta-	ais	aan	la-	ell	ist
испанский	os-	-de	-la	de-	la-	-y-	es-	-a-	ent	ien
португальский	de-	-de	os-	-e-	em-	o-d	to-	-a-	-di	dir

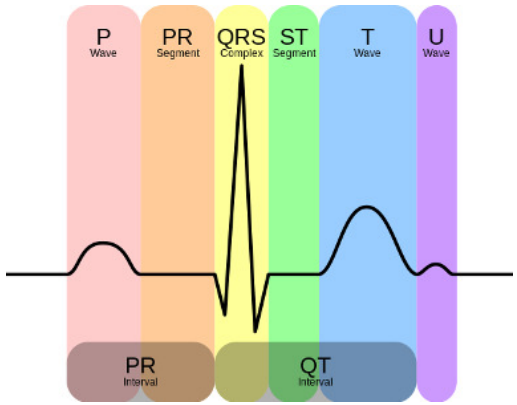
Что показывают эксперименты

- язык текста можно распознавать автоматически,
- с очень высокой надёжностью,
- используя частоты триграмм или биграмм,
- точность распознавания быстро увеличивается с ростом длины текста — сотни символов хватает для распознавания даже близких языков
- для профессионального решения задачи дополнительно используют словари слов и аббревиатур

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

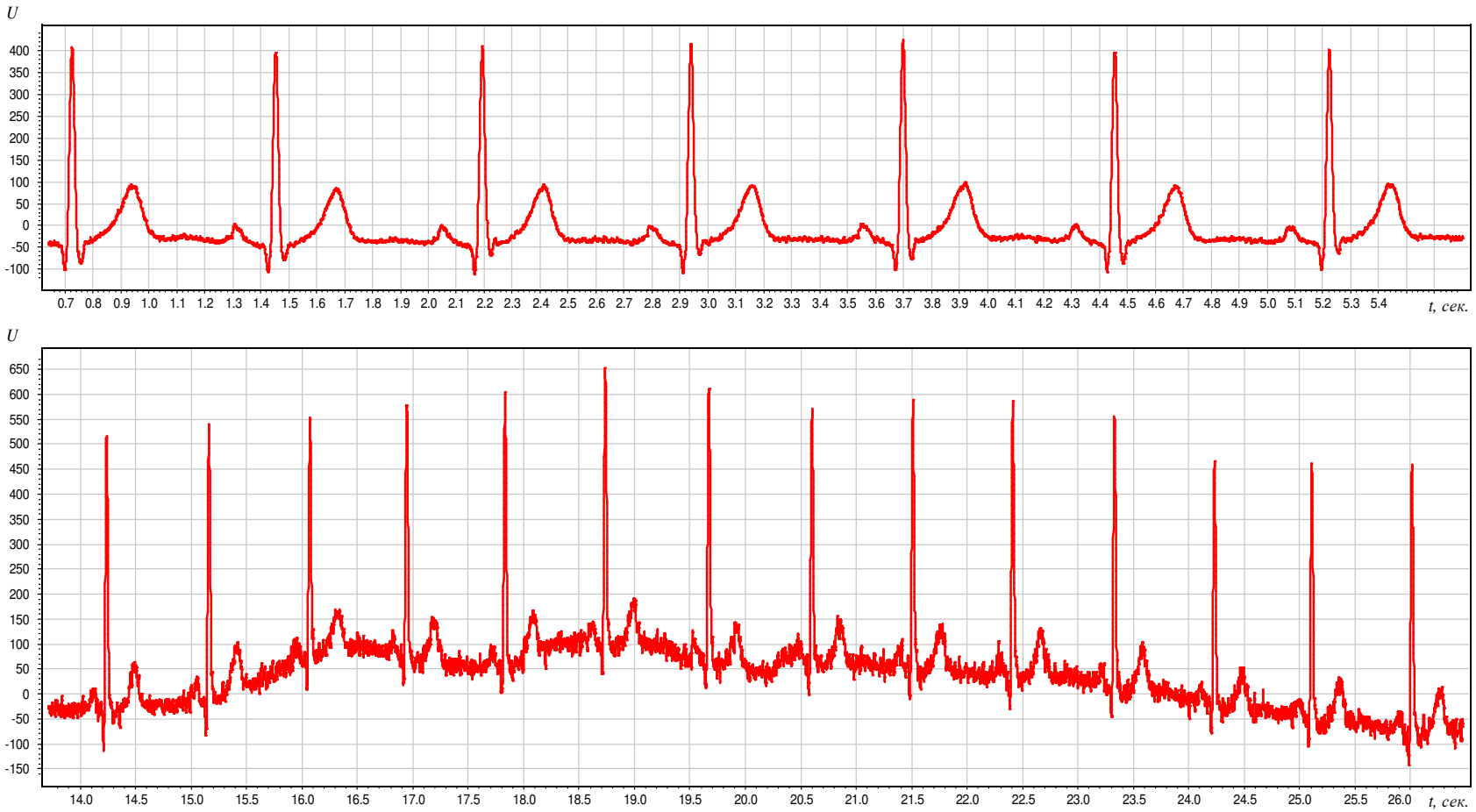


1872 — первые записи электрической активности сердца

1911 — коммерческий электрокардиограф (фото)

1924 — нобелевская премия по медицине, Виллем Эйнтховен

Примеры электрокардиограмм



В основе диагностики заболеваний сердца — многочисленные наблюдения за особенностями PQRST-комплекса

Теория информационной функции сердца [В.М.Успенский]

Предпосылки:

- Китайская традиционная медицина: *пульсовая диагностика*
- Р. М. Баевский: использование вариабельности сердечного ритма (*интервалов кардиоциклов*) в целях диагностики
- Появление цифровой электрокардиографии

Предположения:

- ЭКГ-сигнал несёт информацию о функционировании всех систем организма, не только сердца
- Каждое заболевание по-своему изменяет ЭКГ-сигнал
- Информация о заболевании может проявляться на любой его стадии, поэтому возможна *ранняя* диагностика

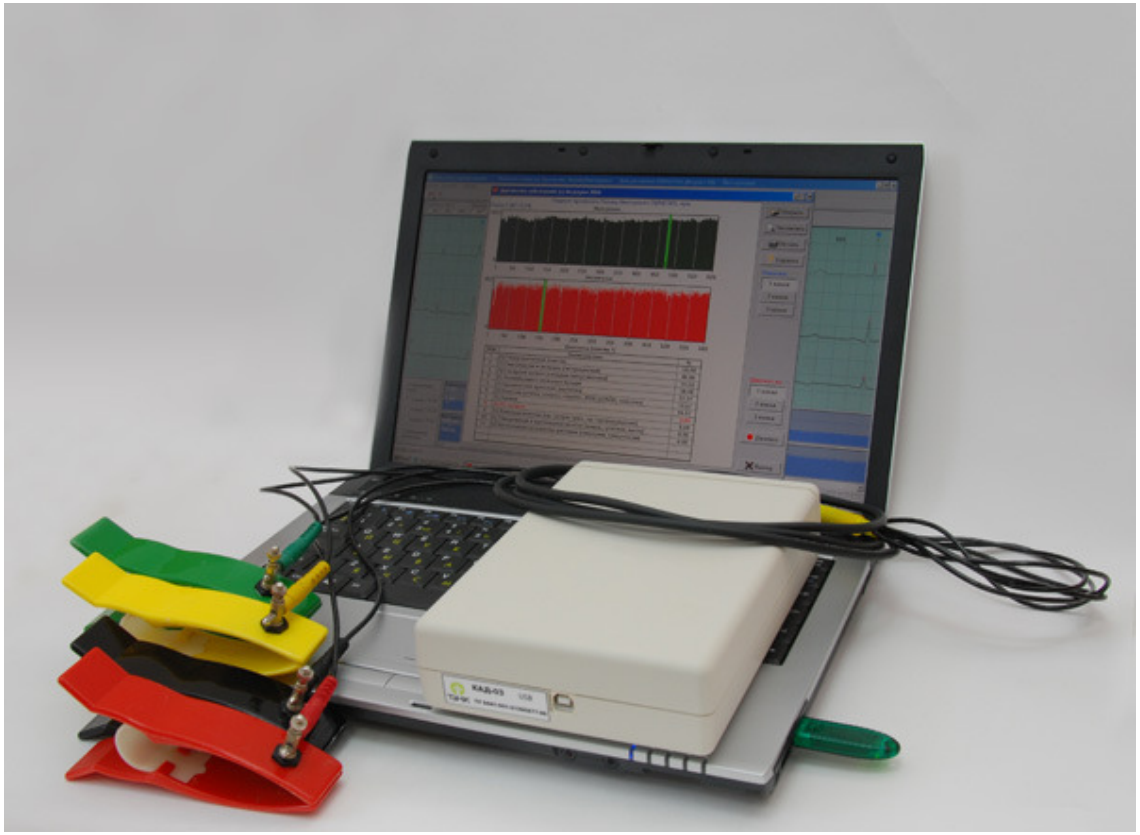
Дискретизация и векторизация ЭКГ-сигнала

По ЭКГ строится текстовая строка — *кодограмма*:

Частоты триграмм — число вхождений триграммы в кодограмму:

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

Диагностическая система «Скринфакс»



- более 10 лет эксплуатации
- более 20 тысяч прецедентов (кардиограмма + диагноз)
- более 40 заболеваний

Объём исходных данных (по заболеваниям)

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

Результаты кросс-валидации

Обучающая выборка — для оптимизации параметров модели
Тестовая выборка — для оценивания чувс., спец., 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

Выводы

- задача удивительно похожа на распознавание языка текста,
- многие болезни можно диагностировать по ЭКГ,
- с очень высокой надёжностью,
- используя частоты триграмм или биграмм,
- причём точность диагностики увеличивается с ростом времени регистрации ЭКГ, и нескольких сотен кардиоциклов уже хватает для распознавания.

Открытые проблемы и исследовательские задачи

- Болезни различаются между собой намного хуже, чем состояния болезни и здоровья
- Существуют ли лучшие способы кодирования?
- Существуют ли лучшие модели классификации?
- Возможно ли углубить аналогию между языком и ЭКГ?

Цели проекта

Для врачей и пациентов:

- Улучшить качество дифференциальной диагностики

Для разработчиков новых методов диагностики:

- Решить несколько исследовательских задач

Для себя:

- Овладеть основами профессии анализа данных

Наши планы

1. освоить программирование в Python
2. научиться работать с данными и «смотреть на данные»
3. попробовать несколько методов «из коробки»
4. сделать что-то своё и добиться улучшений
5. провести соревнование методов
6. побороть открытые проблемы
7. провести несколько «мозговых штурмов»
8. научиться оформлять результаты исследований
9. подготовить выступление на конференции

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov

Когда что-то не понятно,
не стесняйтесь подходить и спрашивать :)