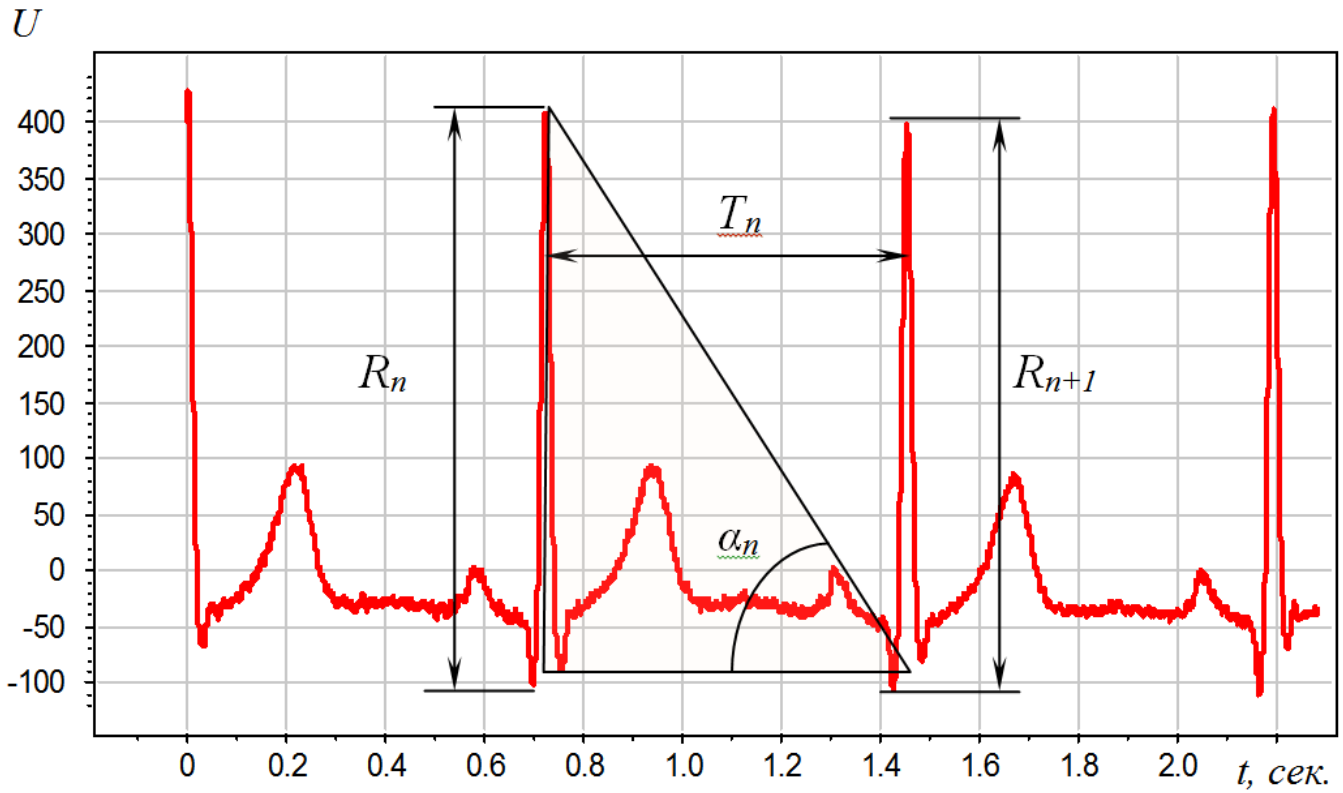
Задача диагностики заболеваний по дискретизированному ЭКГ-сигналу (практическое задание)

Воронцов Константин Вячеславович

ШАД Яндекс • 13 марта 2014



Открытие д.м.н. проф. В.М.Успенского:
возможна ранняя диагностика многих заболеваний по ЭКГ, причём достаточно использовать только знаки приращений амплитуд $R_{n+1} - R_n$, интервалов $T_{n+1} - T_n$ и углов $\alpha_{n+1} - \alpha_n$.

Диагностическая система «Скринфакс» (2-е поколение)

- более 10 лет эксплуатации
- более 20 тысяч прецедентов (кардиограмма + диагноз)
- более 50 заболеваний
- из них более 20 имеют отобранные эталонные выборки

Технология информационного анализа ЭКГ по В.М.Успенскому

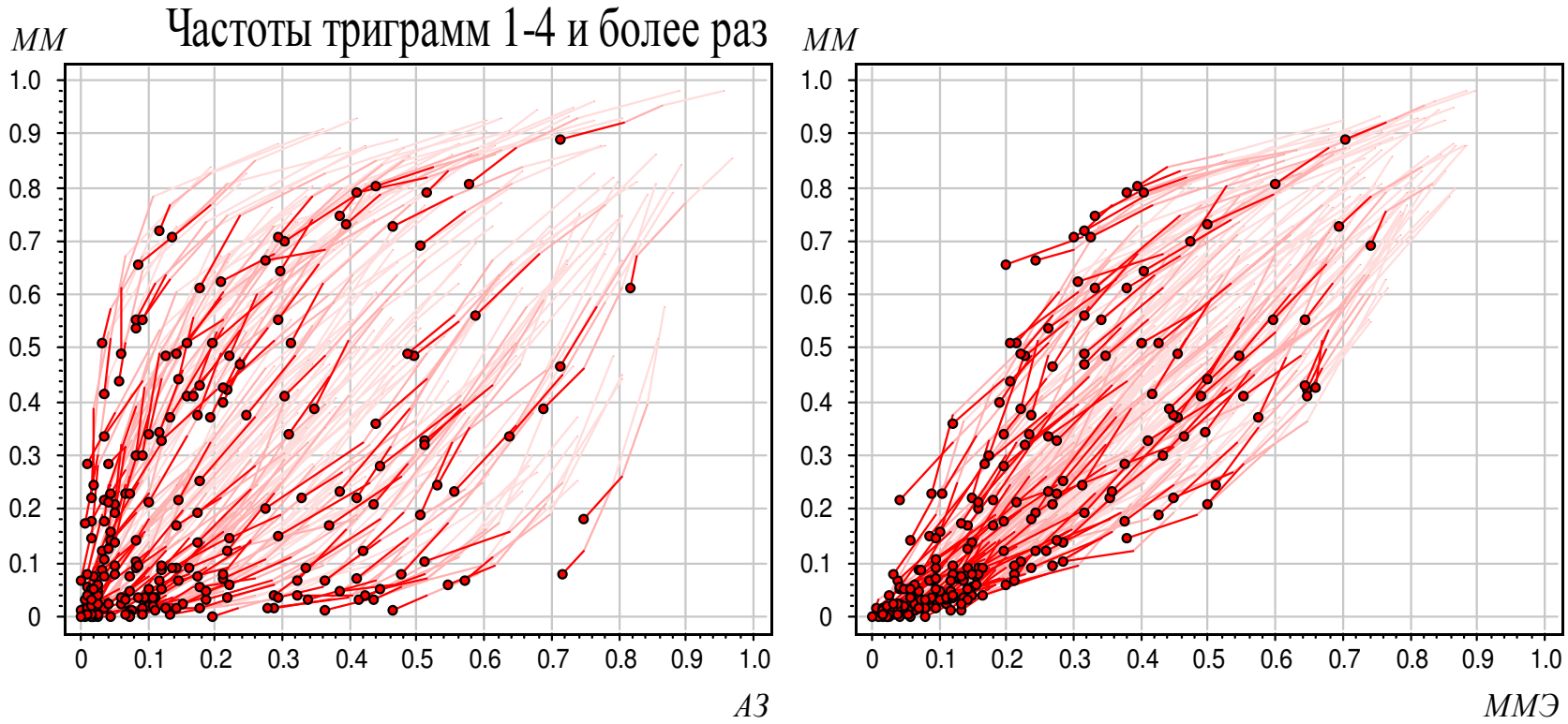
1. вычисление амплитуд, интервалов и углов
   по кардиограмме длиной 600 кардиоциклов
2. вычисление *кодограммы* — 599-символьной строки
   в 6-буквенном алфавите
3. вычисление $6^3 = 216$ признаков — частот триграмм
4. формирование эталонных выборок:
   1) абсолютно здоровых,
   2) больных (для каждого заболевания отдельная выборка)
5. поиск *диагностических эталонов* — наборов триграмм,
   совместно встречающихся у больных данным заболеванием
6. обучение алгоритма классификации
7. статистическая оценка точности диагностики
   по контрольным выборкам или скользящему контролю
8. применение алгоритма классификации для диагностики

Кодограмма и частоты триграмм

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

Поиск информативных признаков-триграмм

Слева: триграммы в осях «доля здоровых» — «доля больных».
Справа: триграммы в осях «доля больных» — «доля больных».



Вывод: болезнь имеет *диагностический эталон* — множество триграмм, часто встречающихся в кодограммах больных, и редко встречающихся в кодограммах здоровых людей.

Задача

Дано: матрицы «объекты–признаки» по 5 болезням,
первый столбец — метки классов (0–здоровый, 1–больной)

эталонные (здоровых, больных)	контрольные (здоровых, больных)
ВДЭ.txt (195, 697)	ВД.txt (356, 639)
ЖКЭ.txt (195, 281)	ЖК.txt (356, 393)
ИБЭ.txt (195, 1272)	ИБ.txt (356, 1611)
УЩЭ.txt (195, 754)	УЩ.txt (356, 999)
ЯБЭ.txt (195, 792)	ЯБ.txt (356, 474)

Найти: алгоритм классификации.

Критерий: чувствительность и специфичность на контроле

Описание задачи — на странице

http://www.MachineLearning.ru/wiki/index.php?title=User:Vokov

раздел «Диагностика заболеваний по ЭКГ»

http://www.MachineLearning.ru/wiki/images/e/e3/Voron-2014-task-ekg.pdf

http://www.MachineLearning.ru/wiki/images/3/37/Voron-2014-task-ekg-data.rar

Положительный диагноз — алгоритм предсказывает болезнь (хотя, казалось бы, что тут положительного...)

Определение

Чувствительность — доля больных с верным положительным диагнозом.
Доля ошибок 2-го рода $= 1 -$ чувствительность.

Определение

Специфичность — доля здоровых с верным отрицательным диагнозом.
Доля ошибок 1-го рода $= 1 -$ специфичность.

Чувствительность и специфичность надо максимизировать.
+ Они не зависят от соотношения мощностей классов.
+ Хорошо подходят для несбалансированных выборок.