Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»
при Вычислительном центре им. А. А. Дородницына РАН

Бунаков Василий Андреевич

Методы нечёткого кодирования в информационном анализе электрокардиосигналов

010990 — Интеллектуальный анализ данных

Магистерская диссертация

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

Москва
2015 г.

Содержание

1 Введение

Существуют электрофизиологические методы исследования сердца, из которых важнейшую роль играет электрокардиография. Являясь основным методом современной кардиологии, научной и практической медицины, электрокардиография позволяет достаточно глубоко оценить состояние миокарда и функций сердца.

В опытах по изучению вариабельности сердечного ритма (ВСР) показано, что электрокардиоимпульсы также могут быть носителями информации о состоянии системы регуляции основных функций организма, в норме и при различных заболеваниях [1, 2]. На основе этих наблюдений профессором В. М. Успенским была предложена *теория информационной функции сердца* и обоснована роль сердца как информационного органа [3].

В отличие от исследований ВСР, в методе В. М. Успенского исследуется вариабельность не только R-R-интервалов, но и R-амплитуд. Предполагается, что в организме существуют механизмы передачи сигналов, аналогичные амплитудной и частотной модуляции в теории сигналов и связи. Для демодуляции этих сигналов и дешифровки содержащейся в них информации разработана *технология информационного анализа электрокардиосигнало*в [3, 4, 5, 6, 7, 8, 9, 10, 11, 12], реализованная в диагностической системе «Скринфакс» [3]. Система позволяет диагностировать по одной электрокардиограмме более 30 различных болезней, включая, помимо сердечно-сосудистых заболеваний, мочекаменную болезнь, сахарный диабет, рак, некроз головки бедренной кости и другие заболевания внутренних органов.

В методе информационного анализа ЭКГ каждый сигнал представляется сначала последовательностью целочисленных амплитуд и интервалов $(T_n, R_n)_{n=1}^N$, затем преобразуется в последовательность символов — *кодограмму*, а затем в числовой вектор признаков, что позволяет решать задачу диагностики методами машинного обучения.

Метод кодирования сигнала обладает следующей особенностью: символ кодограммы c_n порождается значениями амплитуд и интервалов пары последовательных кардиоциклов $T_n, R_n, T_{n+1}, R_{n+1}$, причём для кодирования достаточно лишь знаков приращений амплитуд, интервалов и их отношений: $T_{n+1} - T_n$, $R_{n+1} - R_n$, $R_{n+1}/T_{n+1} - R_n/T_n$. Значения T_n, R_n имеют погрешности, которые складываются из погрешностей измерения и погрешностей округления. По этой причине знаки приращений не всегда определяются верно: в 5% случаев по крайней мере одно из при-

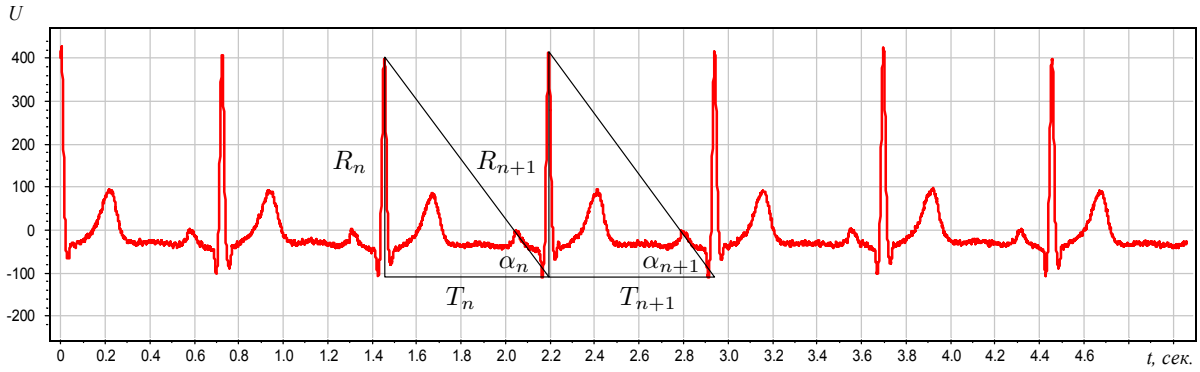

Рис. 1: Пример электрокардиограммы. Два последовательных кардиоцикла с амплитудами R_n, R_{n+1}, интервалами T_n, T_{n+1} и «фазовыми углами» α_n, α_{n+1}.

ращений амплитуд или интервалов равно нулю, а в 14% случаевЁ— по модулю не превосходит единицы, то есть заведомо находится в пределах погрешности измерений. В результате происходит искажение символьной последовательности и вектора признаков, которое может отрицательно сказываться на качестве диагностики.

В данной работе исследуются методы сглаживания шумов и неопределённостей при кодировании и векторизации сигналов. Общие методы, рассмотренные в работе, не зависят от способов кодирования и дискретизации, но являются ресурсоёмкими. Для технологии информационного анализа электрокардиосигналов предлагается ускоренный аналитический метод сглаживания.

Целью работы является повышение качества диагностики заболеваний внутренних органов с помощью технологии информационного анализа электрокардиосигналов при использовании данных методов.

2 Технология информационного анализа электрокардиосигналов

В этом разделе описывается технология диагностики заболеваний по одной электрокардиограмме, включающая предобработку сигнала и классификацию методами машинного обучения. Обучающая выборка представлена признаковыми описаниями электрокардиосигналов и метками классов, соответствующими здоровым и больным пациентам. Поскольку у одного человека может быть много заболеваний, задача построения диагностического правила ставится как задача классификации с пересекающимися классами.

2.1 Предобработка сигналов

Технология предобработки сигнала включает следующие этапы: вычисление амплитуд и интервалов, дискретизация и векторизация [3].

2.1.1 Вычисление амплитуд и интервалов

На первом этапе предобработки, электрокардиограмма преобразуется в последовательность пар $(T_n, R_n)_{n=1}^N$, где T_n — интервал, R_n — амплитуда n-го кардиоцикла, рис. 1.

Последовательность $(T_n)_{n=1}^N$ называется *кардиоинтервалограммой*, а $(R_n)_{n=1}^N$ — *кардиоамплитудограммой* [2]. Число кардиоциклов N, согласно методике измерений, имеет порядок нескольких сотен, обычно $N = 600$. Также вводится арктангенс отношения амплитуд и интервалов $\alpha_n = \operatorname{arctg} R_n/T_n$ как аналог фазового угла в гармонических сигналах.

2.1.2 Дискретизация

Предполагается, что знаки приращений интервалов $T_{n+1} - T_n$, амплитуд $R_{n+1} - R_n$ и углов $\alpha_{n+1} - \alpha_n$ в последовательных кардиоциклах обладают значительно большей диагностической ценностью, чем сами эти значения, подверженные влиянию множества факторов.

Возможны только 6 сочетаний знаков приращений этих трёх величин. Эти сочетания предлагается кодировать буквами 6-символьного алфавита $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$. В таблице «+» означает положительное приращение, «−» — отрицательное:

Рис. 2: Пример кодограммы.

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

Рис. 3: Векторное представление $n_w(C)$ кодограммы C, приведённой на рис. 2. Показаны только 64 из 216 триграмм, имеющих число вхождений $n_w(C) \geqslant 2$.

	A	B	C	D	E	F
$T_{n+1} - T_n$	$+$	$-$	$-$	$+$	$+$	$-$
$R_{n+1} - R_n$	$+$	$-$	$+$	$-$	$+$	$-$
$\alpha_{n+1} - \alpha_n$	$+$	$+$	$+$	$-$	$-$	$-$

Результатом дискретизации амплитудограмм и интервалограмм является построение кодограммы — символьной последовательности $C = (c_n)_{n=1}^{N-1}$, состоящей из символов алфавита $\mathcal{A}$, рис. 2. Каждый символ является сжатой характеристикой типа взаимосвязи между двумя соседними кардиоциклами. Таким образом, в кодограмме отражается наиболее важная для диагностики информация из исходного электрокардиосигнала. Выражение этой информации в символьной форме позволяет применять методы анализа символьных последовательностей и машинного обучения, аналогичные используемым в вычислительной лингвистике [14], и биоинформатике [15].

2.1.3 Векторизация

Векторизацией называется заключительный этап предобработки, превращающий кодограмму в числовое признаковое описание для задачи машинного обуче-

ния. Назовём *k-граммой* слово $c_n, \ldots, c_{n+k-1}$, образованное k последовательными буквами кодограммы C. Множество всех возможных k-грамм $W = \mathcal{A}^k$, составленных из букв 6-буквенного алфавита содержит 6^k элементов. *Частота* k-граммы $w = (w_0, \ldots, w_{k-1})$ определяется как отношение её *числа вхождений* $n_w(C)$ в кодограмму C к общему числу k-грамм в кодограмме, равному $N - k$:

$$n_w(C) = \sum_{n=1}^{N-k} \prod_{j=0}^{k-1} [C_{n+j} = w_j]; \qquad f_w(C) = \frac{n_w(C)}{N-k}.$$

На рис. 3 показан пример представления кодограммы в виде вектора частот триграмм, $k = 3$.

В методе В. М. Успенского рассматриваются наборы k-грамм, совместная встречаемость которых свидетельствует о наличии в организме *информационной сущности* или *программы* определённого заболевания. Её наличие может говорить о предрасположенности к заболеванию, она также проявляется у человека на любой стадии заболевания. Эта особенность позволяет использовать метод для ранней диагностики заболеваний, в том числе задолго до возникновения симптомов и перехода заболевания в активную фазу.

2.2 Используемые методы машинного обучения

Основные подходы и алгоритмы машинного обучения, используемые для анализа векторизованных сигналов ЭКГ с целью диагностики заболеваний, изложены в [13]. Ниже приводится краткий обзор алгоритмов, используемых в настоящей работе.

2.2.1 Линейные модели классификации

Для простоты рассмотрим задачу классификации пациентов на два класса: здоровые и больные. Обучающая выборка

$$X = \{\mathbf{x}_l, y_l\}, \quad \mathbf{x}_l = (T_n, R_n)_{n=1}^{N}, \quad l = 1 \ldots L,$$

содержит электрокардиосигналы пациентов, заданные последовательностями амплитуд и интервалов, и метки классов $y_l \in \{0, 1\}$. Существует процедура порождения признакового описания сигнала в виде вектора встречаемостей k-грамм:

$$F(\mathbf{x}_l) = \|f_w(\mathbf{x}_l)\|, \quad w = 1 \ldots 6^k.$$

Синдромный алгоритм представляет собой модификацию наивного байесовского классификатора. Используются бинаризованные признаки $[f_w > \theta]$, где θ — порог бинаризации. Как и для всякого линейного классификатора обучение сводится к нахождению весов γ_w признаков:

$$a(\mathbf{x}) = \big[b(\mathbf{x}) \geqslant \beta\big], \qquad b(\mathbf{x}) = \sum_{w \in W} \gamma_w \big[f_w(\mathbf{x}) > \theta\big].$$

Веса настраиваются по обучающей выборке следующим образом. Пусть X_m — множество объектов класса $m \in \{0, 1\}$. Обозначим $B_w(X_m, \theta)$ встречаемость k-граммы w в классе X_m:

$$B_w(X_m, \theta) = \sum_{\mathbf{x} \in X_m} \big[f_w(\mathbf{x}) > \theta\big].$$

Величина $B_w(X_m, \theta)$ характеризует информативность k-граммы w в классе X_m. В каждом классе отбирается K k-грамм с наибольшей информативностью. Веса признаков вычисляются по формуле:

$$\gamma_w = B_w(X_1, \theta) - B_w(X_0, \theta).$$

Параметр K настраивается по обучающей выборке.

Логистическая регрессия также является линейным классификатором с пороговым решающим правилом. Рассматриваются два вида признаков: вещественные частоты триграмм $f_w(S)$ и бинарные встречаемости $\big[f_w(\mathbf{x}) \geqslant \theta\big]$ с параметром θ. Для снижения размерности используется метод главных компонент. Число компонент K настраивается по обучающей выборке.

2.2.2 Случайный лес

Каждое решающее дерево в ансамбле строится по случайным подвыборкам, полученным в результате сэмплирований с возвращениями объектов обучающей выборки. Для классификации объектов используется простое голосование: каждое дерево относит объект к тому из классов, за который проголосует более βD деревьев, где D — общее количество деревьев, а β — порог принятия решения. В данной работе количество деревьев фиксировано, $D = 100$.

2.2.3 Оценивание качества диагностики

Все вышеперечисленные алгоритмы на выходе имеют пороговое решающее правило. Универсальным критерием качества является AUC (Area Under Curve) — пло-

щадь под ROC-кривой, отображающей зависимость доли верных положительных классификаций (TPR) от доли ложных положительных классификаций (FPR):

$$\mathrm{TPR} = \frac{1}{|X_1|}\sum_{\mathbf{x}\in X_1}\big[a(\mathbf{x})=1\big], \quad \mathrm{FPR} = \frac{1}{|X_0|}\sum_{\mathbf{x}\in X_0}\big[a(\mathbf{x})=1\big].$$

Также AUC можно определять как долю правильно упорядоченных пар прецедентов:

$$\mathrm{AUC} = \frac{1}{|X_0|\cdot|X_1|}\sum_{\mathbf{x}\in X_0}\sum_{\mathbf{x}'\in X_1}\big[b(\mathbf{x}) < b(\mathbf{x}')\big].$$

3 Общие методы нечёткого кодирования

В этом разделе описываются принципы построения признакового описания электрокардиосигналов, сглаженного относительно шумов и неопределённостей. Предлагаемые методы универсальны и не накладывают никаких ограничений на процедуру дискретизации сигнала.

3.1 Обозначения

Обучающая выборка описана в разделе 2.2.1. Введём обозначения:

- $\mathbf{x} = (T_n, R_n)_{n=1}^{N}$ — последовательность амплитуд и интервалов;
- $C(\mathbf{x})$ — кодограмма длины $N-1$;
- $W(\mathbf{x})$ — вектор длины $N-k$, состоящий из последовательных k-грамм;
- $F(\mathbf{x})$ — вектор встечаемостей k-грамм сигнала $\mathbf{x}$.

Гистограммная свёртка. Пусть матрица A размером $m \times n$ состоит из элементов конечного множества $\mathbb{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_h\}$. Построим матрицу B размером $|\mathbb{A}| \times n$ таким образом, чтобы её столбцы содержали доли элементов множества $\mathbb{A}$ в соответствующих столбцах матрицы A. Полученную матрицу B назовём *гистограммной свёрткой* матрицы A и обозначим

$$B = Hist\, A.$$

Пример. Рассматривая столбец последовательных k-грамм W^T как матрицу A, где $\mathbb{A}$ — множество всевозможных k-грамм, для столбца встречаемостей k-грамм F^T верно соотношение:

$$F^T = Hist\, W^T.$$

3.2 Модель измерений

Для простоты будем рассматривать действительные значения T_n, R_n и считать, что погрешности являются случайными и состоят только из ошибок измерения. По определению, погрешность измерения — это отклонение результата измерения от истинного (действительного) значения измеряемой величины [16]:

$$\varepsilon_T(\sigma_T) = T_n - \tilde{T}_n, \quad \varepsilon_R(\sigma_R) = R_n - \tilde{R}_n$$

Случайные величины ε_T, ε_R предполагаются независимыми. Их матожидания при отсутствии систематических погрешностей равны нулю, а дисперсии равны σ_T^2, σ_R^2 соответственно.

В разделах 2.2.2–2.2.3 описывались процедуры дискретизации и векторизации сигналов, заданных наблюдаемыми значениями амплитуд и интервалов. В действительности, каждое измеренное значение порождает множество возможных реализаций истинных значений. Задача *нечёткого кодирования* — произвести дискретизацию и векторизацию сигнала, опираясь на все возможные реализации истинных значений амплитуд и интервалов, а не только на измеренные значения.

3.3 Общие методы сглаживания

В этом разделе описываются универсальные методы, не зависящие от способов кодирования и дискретизации сигналов.

Рассмотрим истинные значения амплитуд и интервалов как вероятностные распределения:

$$\tilde{T}_n = T_n + \varepsilon_T(\sigma_T), \quad \tilde{R}_n = R_n + \varepsilon_R(\sigma_R).$$

Предполагая погрешности симметричными, для удобства возьмём ε_T, ε_R с противоположными знаками.

Сигнал также будем рассматривать как вероятностное распределение $\tilde{\mathbf{x}}$ над множеством возможных последовательностей амплитуд и интервалов. Аналогично, рассматриваем как вероятностные распределения кодограмму $C(\tilde{\mathbf{x}})$, последовательность k-грамм $W(\tilde{\mathbf{x}})$ и вектор встречаемостей $F(\tilde{\mathbf{x}})$. Задача состоит в получении оценки $\mathsf{M}F(\tilde{\mathbf{x}})$ значения вектора встречаемостей, усреднённого по возможным реализациям амплитуд и интервалов в рамках модели измерений.

3.3.1 Семплирование векторов встречаемостей

Для нахождения оценки среднего значения $\mathsf{M}F(\tilde{\mathbf{x}})$ рассмотрим S реализаций истинных значений амплитуд и интервалов, семплированных из модельных распределений ε_T, ε_R. То есть вместо исходного сигнала $\mathbf{x}$ рассматриваем совокупность

$$\chi = \{\tilde{\mathbf{x}}_s\}, \quad s = 1, \ldots, S$$

сигналов, каждый из которых представляет собой сумму исходного сигнала и последовательности семплированных значений ε_T, ε_R.

Самый простой способ оценки $\mathsf{M}F(\tilde{\mathbf{x}})$ состоит в вычислении векторов встречаемостей для всех сигналов $\mathbf{x}_s$ и усреднении по всем S реализациям:

$$\mathsf{M}F(\tilde{\mathbf{x}}) = \frac{1}{S}\sum_{s=1}^{S} P(\tilde{\mathbf{x}}_s).$$

Преимущество метода состоит в том, что в нём не используется никаких предположений относительно способов дискретизации и векторизации сигналов. Недостаток у всех методов, использующих семплирование, один и тот же: низкая скорость вычислений и высокий расход памяти. Вместо одного сигнала обработке подвергается S сигналов, что может существенно увеличить время работы для $S \sim 100$ и достаточно больших объёмов данных.

Ниже приводится ещё два метода, основанных на семплировании.

3.3.2 Семплирование k-грамм

Пусть каждый сигнал $\tilde{\mathbf{x}}_s$ преобразуется в последовательность k-грамм $W(\tilde{\mathbf{x}}_s)$, а совокупность χ — в матрицу $W(\chi)$, состоящую из строк $W(\tilde{\mathbf{x}}_s)$, рис. 4. Столбец $W_n(\chi)$ матрицы $W(\chi)$ содержит S k-грамм, которые могли реализоваться вместо n-ной k-граммы сигнала при S различных реализациях $\tilde{\mathbf{x}}_s$.

Гистограммная свёртка $Hist\, W_n(\chi)$ даёт эмпирическое распределение вероятностей возможных k-грамм в данном месте сигнала. Сумма таких гистограмм по всем столбцам и является оценкой средней встречаемости k-грамм $\mathsf{M}F(\tilde{\mathbf{x}})$:

$$\mathsf{M}F(\tilde{\mathbf{x}}) = \sum_{n=1}^{N-k} Hist\, W_n(\chi), \quad \chi = \{\tilde{\mathbf{x}}_s\}, \quad s = 1, \ldots, S$$

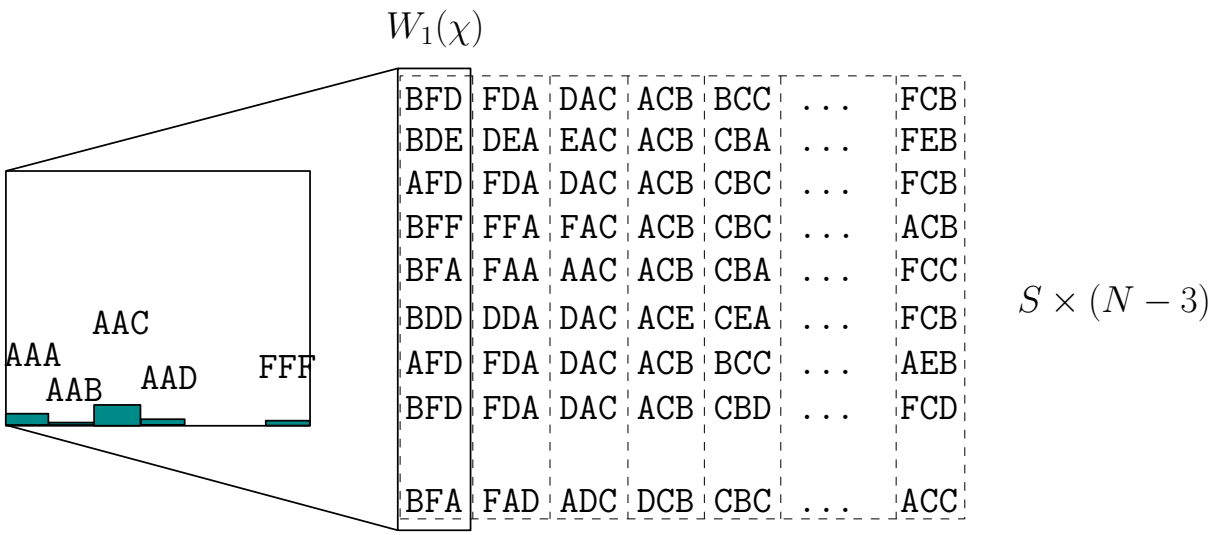

Рис. 4: Вычисление нечёткой k-граммы с помощью семплирования.

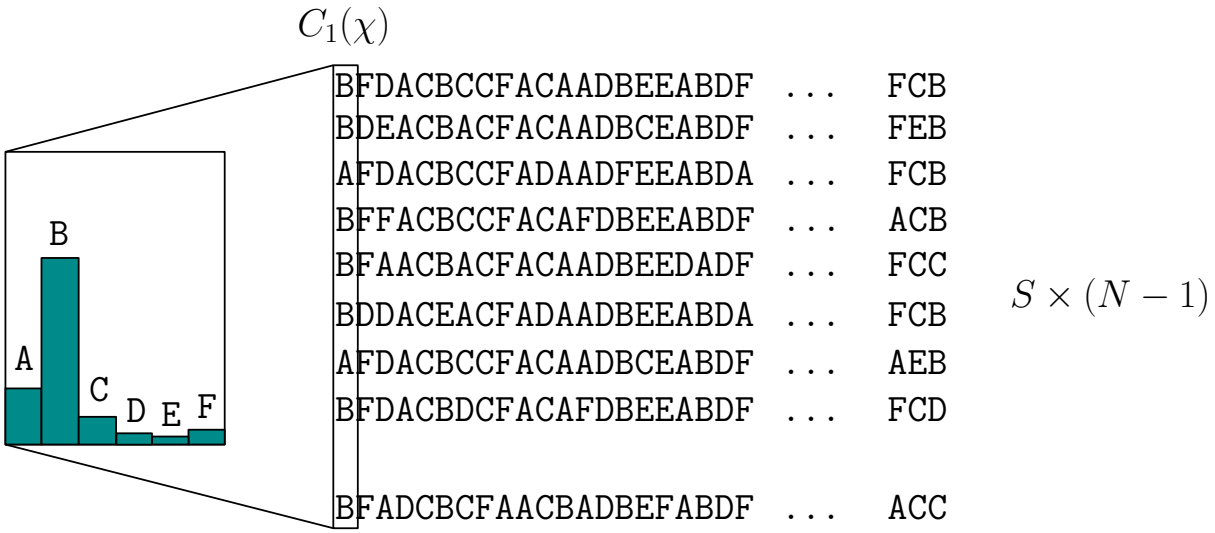

Рис. 5: Вычисление нечёткой кодограммы с помощью семплирования.

3.3.3 Семплирование униграмм, гипотеза независимости

Нечёткие униграммы или *символы* получаются аналогично нечётким k-граммам. Разница заключается в том, что для каждой из S реализаций сигнала вычисляется кодограмма $C(\tilde{\mathbf{x}}_s)$, рис. 5. Таким образом, нечёткая униграмма представляет собой вектор-столбец $Hist\, C_n(\chi)$.

Нечёткая кодограмма есть результат вычисления нечётких униграмм для каждого столбца матрицы $C(\chi)$. Другими словами, нечёткой кодограммой называется гистограммная свёртка матрицы кодограмм: $Hist\, C(\chi)$. Столбцы нечёткой кодограммы отражают «степень присутствия» каждого символа в данном месте сигнала, рис. 6. На этой идее основаны дальнейшие рассуждения, приводящие к вычислению нечёт-

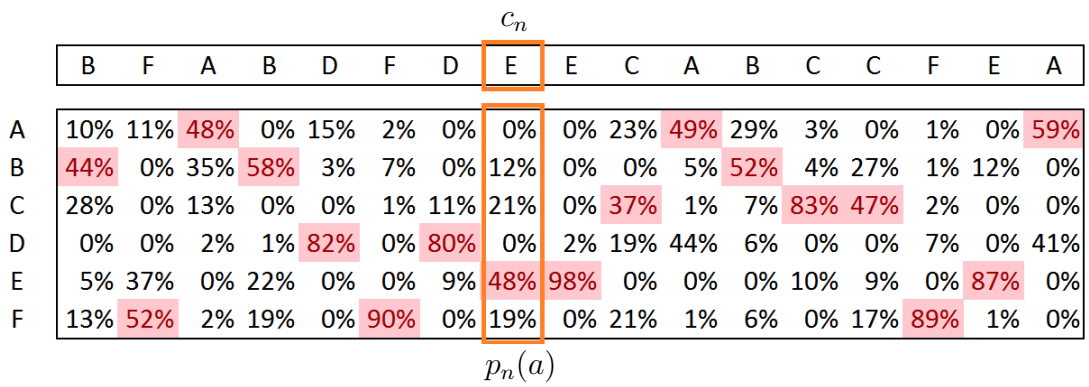

Рис. 6: Нечёткая кодограмма.

ких k-грамм и получению среднего вектора встречаемостей аналогично предыдущему методу.

Гипотеза независимости. Векторизация нечёткой кодограммы основывается на следующих принципах:

1. Нечёткие униграммы рассматриваются как распределения вероятностей $p_n(a)$ над алфавитом символов $a \in \mathcal{A}$.

2. *Гипотеза независимости:* при рассмотрении коротких отрезков длины k, распределения $p_n, p_{n+1}, \ldots, p_{n+k-1}$ предполагаются независимыми.

3. Нечёткая k-грамма представляет собой совместное распределение независимых нечётких униграмм $p_n, p_{n+1}, \ldots, p_{n+k-1}$.

4. Средний вектор встречаемостей k-грамм $\mathsf{M}F(\tilde{\mathbf{x}})$ рассчитывается как сумма нечётких k-грамм.

Исходя из третьего принципа и определения нечёткой униграммы, нечёткую k-грамму можно рассчитать как совокупность произведений всевозможных компонентов столбцов $Hist\, C_n(\chi), Hist\, C_{n+1}(\chi), \ldots, Hist\, C_{n+k}(\chi)$.

В следующем разделе на базе этого метода предлагается процедура аналитического вычисления векторов нечёткой кодограммы без использования семплирования.

4 Нечёткое кодирование в технологии информационного анализа электрокардиосигналов

В этом разделе описывается метод аналитического вычисления нечётких кодограмм. Метод позволяет существенно ускорить вычисления по сравнению с семплированием, так как не требуется совершать преобразований большого числа S копий сигнала.

Метод основывается на процедуре дискретизации сигнала, предложенной В. М. Успенским и использует гипотезу независимости, изложенную в предыдущем разделе. Таким образом, данный метод является наименее общим из всех предложенных методов нечёткого кодирования.

4.1 Графическая интерпретация

Процесс дискретизации, используемый в технологии информационного анализа электрокардиосигналов, описан в разделе 2.1.2.

Каждой паре последовательных кардиоциклов ставится в соответствие символ 6-буквенного алфавита. Символ соответствует какому-то одному из возможных сочетаний знаков приращений интервалов $T_{n+1}-T_n$, амплитуд $R_{n+1}-R_n$ и углов $\alpha_{n+1}-\alpha_n$. Этот процесс имеет важную графическую интерпретацию, которая послужит переходным звеном от дискретного кодирования к нечёткому. Как и прежде, рассматривается сигнал

$$\mathbf{x} = (T_n, R_n)_{n=1}^{N}.$$

Рассмотрим двумерную координатную плоскость. Изобразим на плоскости точки $(\Delta T_n,\, \Delta R_n)$, соответствующие приращениям $\Delta T_n = T_{n+1} - T_n$ и $\Delta R_n = R_{n+1} - R_n$. Точки, соответствующие нулевым приращениям $\Delta T_n = 0$, $\Delta R_n = 0$, лежат на координатных прямых. Нетрудно заметить, что условию $\Delta\alpha_n = 0$ соответствуют точки, расположенные на прямой $y = \arctan\alpha_n \cdot x$.

Таким образом, прямые $y = 0$, $x = 0$ и $y = \arctan\alpha_n \cdot x$ разбивают плоскость на шесть векторов, каждый из которых соответствует одной из комбинаций знаков приращений, рис. 7.

Модель измерений полностью аналогична введённой в предыдущем разделе. Рассмотрим пару последовательных кардиоциклов: $T_n, R_n, T_n + \Delta T_n, R_n + \Delta R_n$.

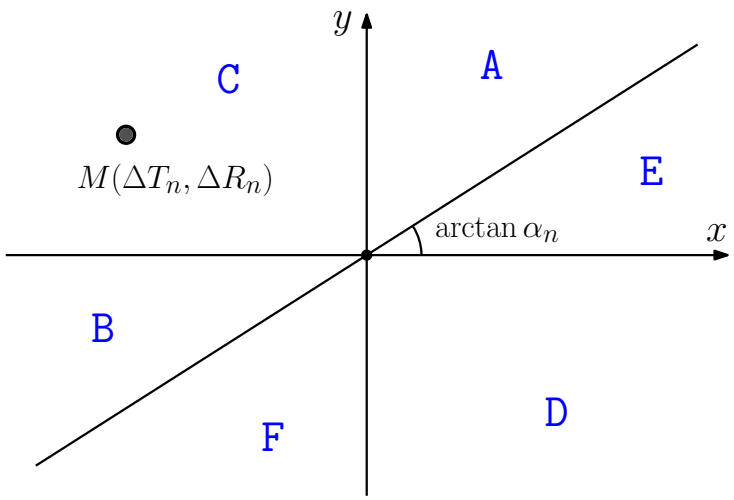

Рис. 7: Графическая интерпретация.

Основное предположение состоит в том, что погрешности измерений целиком заложены в приращениях, а сами значения T_n и R_n, и, следовательно, угол $\alpha_n = R_n/T_n$ считаются зафиксированными:

$$\Delta\tilde{T}_n = \Delta T_n + \varepsilon_T^*(\sigma_T^*), \quad \Delta\tilde{R}_n = \Delta R_n + \varepsilon_R^*(\sigma_R^*),$$

Также считаем, что ε_T^* и ε_R^* — независимые случайные величины с нулевыми матожиданиями и дисперсиями σ_T^{*2}, σ_R^{*2}.

Модель измерения эквивалентна следующей: в точке $M(\Delta T_n, \Delta R_n)$ находится двумерное вероятностное распределение с нулевым матожиданием и плотностью

$$\varphi(x, y, \sigma_T^*, \sigma_R^*) = \varphi_T(x, \sigma_T^*)\varphi_R(y, \sigma_R^*).$$

Проинтегрировав плотность $\varphi(x, y, \sigma_T^*, \sigma_R^*)$ по каждому сектору, получаем вероятности появления соответствующих символов. Вычислив эти вероятности для каждой пары последовательных циклов, получаем аналог нечёткой кодограммы, полученной в разделе 3.3.4 при условии $\sigma_T^{*2} = 2\sigma_T^2$, $\sigma_R^{*2} = 2\sigma_R^2$.

4.2 Аналитическое вычисление нечёткой кодограммы

В некоторых случаях можно обойтись без семплирования и ускорить вычисления, рассчитав интегралы аналитически.

Например, для распределения Лапласа, плотность φ равна

$$\varphi(x, y, \sigma_T^*, \sigma_R^*) = \frac{1}{\sigma_T^*\sqrt{2}} \exp\left(\frac{-|x|\sqrt{2}}{\sigma_T^*}\right) \cdot \frac{1}{\sigma_R^*\sqrt{2}} \exp\left(\frac{-|y|\sqrt{2}}{\sigma_R^*}\right).$$

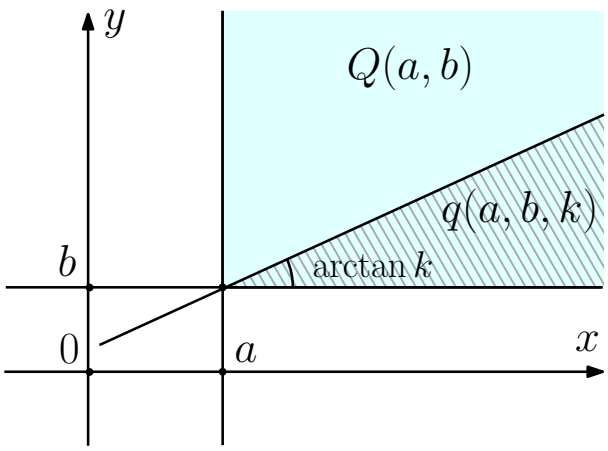

Рис. 8: Базовые сектора.

Введём вспомогательные функции $Q(a,b)$ и $q(a,b,k)$ — интегралы обезразмеренной плотности $\varphi(x,y,1,1)$ по двум базовым секторам (см. Рис. 8):

$$Q(a,b) = \iint\limits_{x>a,\ y>b} \frac{1}{2} e^{-\sqrt{2}(|x|+|y|)} dxdy,$$
$$q(a,b,k) = \iint\limits_{\substack{x>a,\\ b<y<kx}} \frac{1}{2} e^{-\sqrt{2}(|x|+|y|)} dxdy.$$

Нетрудно показать, что площади каждого из 6 секторов выражаются через базовые функции следующим образом:

$$p_{\mathtt{A}} = Q\left(-\frac{\Delta T_n}{\sigma_T}, -\frac{\Delta R_n}{\sigma_R}\right) - q\left(-\frac{\Delta T_n}{\sigma_T}, -\frac{\Delta R_n}{\sigma_R}, \frac{k\sigma_T}{\sigma_R}\right)$$
$$p_{\mathtt{B}} = Q\left(\frac{\Delta T_n}{\sigma_T}, -\frac{\Delta R_n}{\sigma_R}\right)$$
$$p_{\mathtt{C}} = q\left(\frac{\Delta T_n}{\sigma_T}, \frac{\Delta R_n}{\sigma_R}, \frac{k\sigma_T}{\sigma_R}\right)$$
$$p_{\mathtt{D}} = Q\left(-\frac{\Delta T_n}{\sigma_T}, \frac{\Delta R_n}{\sigma_R}\right)$$
$$p_{\mathtt{E}} = q\left(\frac{\Delta T_n}{\sigma_T}, \frac{\Delta R_n}{\sigma_R}, \frac{k\sigma_T}{\sigma_R}\right)$$
$$p_{\mathtt{F}} = Q\left(\frac{\Delta T_n}{\sigma_T}, \frac{\Delta R_n}{\sigma_R}\right) - q\left(\frac{\Delta T_n}{\sigma_T}, \frac{\Delta R_n}{\sigma_R}, \frac{k\sigma_T}{\sigma_R}\right).$$

Аналитические выражения для функций $Q(a,b)$ и $q(a,b,k)$:

$$Q(a,b)=\begin{cases}\frac{1}{4}e^{-a-b}, & \text{при } a\geqslant 0, b\geqslant 0;\\ \frac{1}{2}e^{-b}-\frac{1}{4}e^{a-b}, & \text{при } a<0, b\geqslant 0;\\ 1-\frac{1}{2}e^{a}-\frac{1}{2}e^{b}+\frac{1}{4}e^{a+b}, & \text{при } a<0, b<0;\\ \frac{1}{2}e^{-a}-\frac{1}{4}e^{b-a}, & \text{при } a\geqslant 0, b<0.\end{cases}$$

$$q(a,b,k\neq 1)=\begin{cases}\frac{k}{4(k+1)}e^{-a-b}, & \text{при } a\geqslant 0, b\geqslant 0;\\ \frac{1}{2}e^{-b}+\frac{k}{4(1-k)}e^{a-b}-\frac{1}{2(1-k^2)}e^{ak-b}, & \text{при } a<0, b\geqslant 0;\\ \frac{k^2}{2(k^2-1)}e^{b/k-a}-\frac{1}{2}e^{b}-\frac{1}{2(k^2-1)}e^{b-ak}+\frac{k}{4(k+1)}e^{a+b}, & \text{при } a<0, b<0, ak\geqslant b;\\ 1-\frac{1}{2}e^{b}+\frac{k}{4(k+1)}e^{a+b}+\frac{k^2}{2(1-k^2)}e^{a-b/k}-\frac{1}{2(1-k^2)}e^{ak-b}, & \text{при } a<0, b<0, ak<b;\\ \frac{k^2}{2(k^2-1)}e^{b/k-a}-\frac{k}{4(k-1)}e^{b-a}, & \text{при } a\geqslant 0, b<0.\end{cases}$$

$$q(a,b,k=1)=\begin{cases}\frac{1}{8}e^{-a-b}, & \text{при } a\geqslant 0, b\geqslant 0;\\ \frac{1}{2}e^{-b}+\frac{2a-3}{8}e^{a-b}, & \text{при } a<0, b\geqslant 0;\\ -\frac{1}{2}e^{b}+\frac{1}{8}e^{a+b}+\frac{a-b+2}{4}e^{b-a}, & \text{при } a<0, b<0, a\geqslant b;\\ 1-\frac{1}{2}e^{b}+\frac{1}{8}e^{a+b}+\frac{a-b+2}{4}e^{b-a}, & \text{при } a<0, b<0, a<b;\\ \frac{1-2b}{8}e^{b-a}, & \text{при } a\geqslant 0, b<0.\end{cases}$$

5 Вычислительный эксперимент

В этом разделе описывается применение методов нечёткого кодирования для технологии информационного анализа ЭКГ и исследуется возможность увеличения качества диагностики с помощью этих методов.

5.1 Обучающая выборка

Выборка X содержит электрокардиограммы здоровых и больных пациентов. Больные имеют одну или несколько болезней из списка 18 заболеваний, диагностируемых системой. Обозначим через y_0 класс здоровых людей, через $y_1, \ldots, y_M$ — классы $M = 18$ заболеваний. К классу y_m относятся все пациенты, имеющие заболевание под номером m, не зависимо от того, есть ли у них другие заболевания. Обозначим X_m множество пациентов, класса m.

Пациент представляется признаковым описанием — усреднённым вектором частот k-грамм $\mathsf{M}F$. Для каждого заболевания формируется двухклассовая подвыборка векторов частот k-грамм, полученных одним из методов нечёткого кодирования.

5.2 Оптимизация параметров нечёткого кодирования

Параметры модели нечёткого кодирования оптимизируются по всей выборке пациентов при $y_1, \ldots, y_M$. Оптимизируются следующие параметры:

- Вид распределений ε_T, ε_R
  - нормальное распределение (N)
  - распределение Лапласа (L)
- Дисперсии распределений — параметры σ_T, σ_R
- Способ векторизации:
  - семплирование векторов встречаемостей (S_1)
  - сэмплирование триграмм (S_2)
  - сэмплирование униграмм + гипотеза независимости (S_3)
  - аналитическая оценка + гипотеза независимости (A)

Качество классификации оценивается по критерию AUC (см. раздел 2.2.3) с использованием 1×10 кросс-валидации. Параметры σ_T, σ_R оптимизируются на двумерной сетке и усредняются по всем классам y_m:

$$(\sigma_T, \sigma_R) = \frac{1}{M} \sum_{m=1}^{M} \arg\max_{\mathbb{R}^2_+} \mathrm{AUC}\big(\{X_0, X_m\}, \sigma_T, \sigma_R\big).$$

Ниже приведены двумерные графики, изображающие значения AUC при различных параметрах σ_T, σ_R, взятых на двумерных сетках $\sigma_T = 0, \ldots, 15$, $\sigma_R = 0, \ldots, 8$. Значения AUC приведены для различных моделей нечёткого кодирования, использующих биграммы и триграммы, и вычислены с помощью различных алгоритмов классификации, описанных в разделах 2.2.2–2.2.3.

Все графики имеют характерный вид распределения с выраженным максимумом, соответствующим оптимальным значениям σ_T, σ_R. Важным полученным результатом является тот факт, что форма распределений и положение максимума схожи для всех рассмотренных методов нечёткого кодирования. Также, отсутствует зависимость от типа модельных распределений ε_T, ε_R.

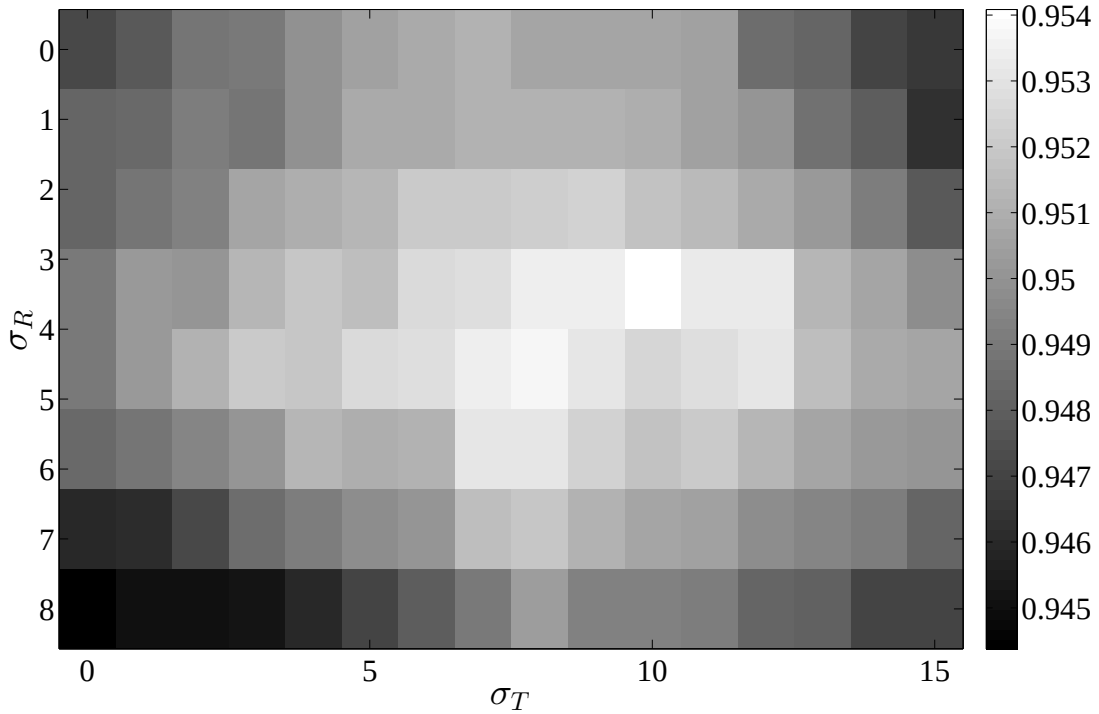

Рис. 9: Синдромный алгоритм, A-L, триграммы

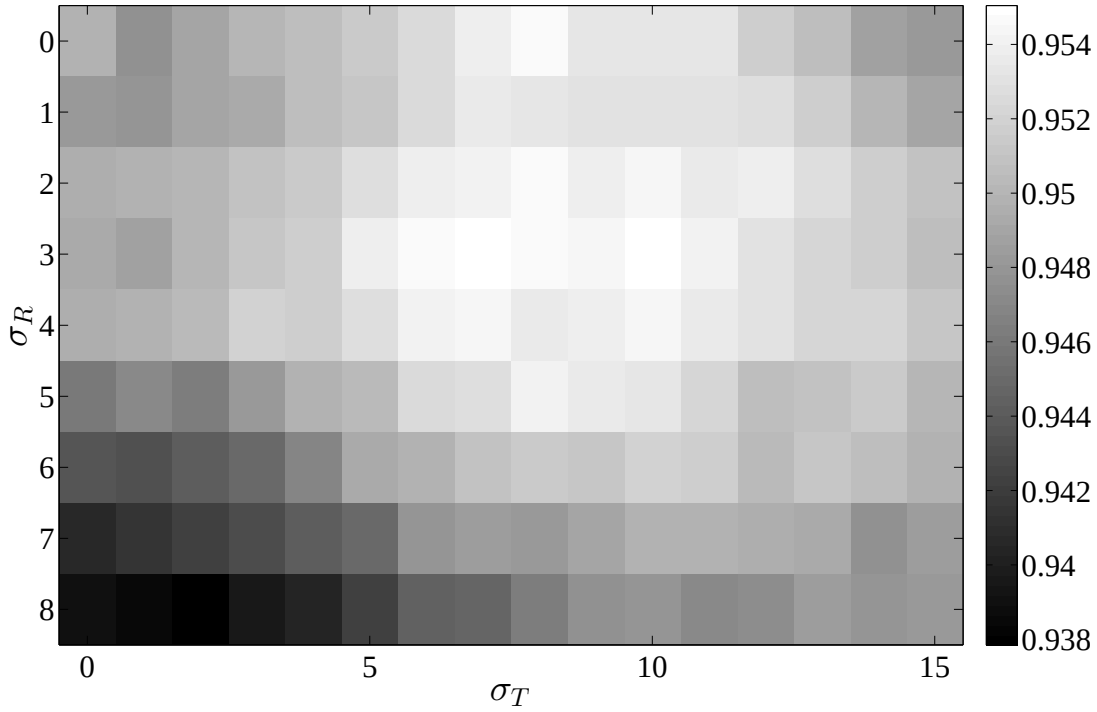

Рис. 10: Синдромный алгоритм, S_1-N, триграммы

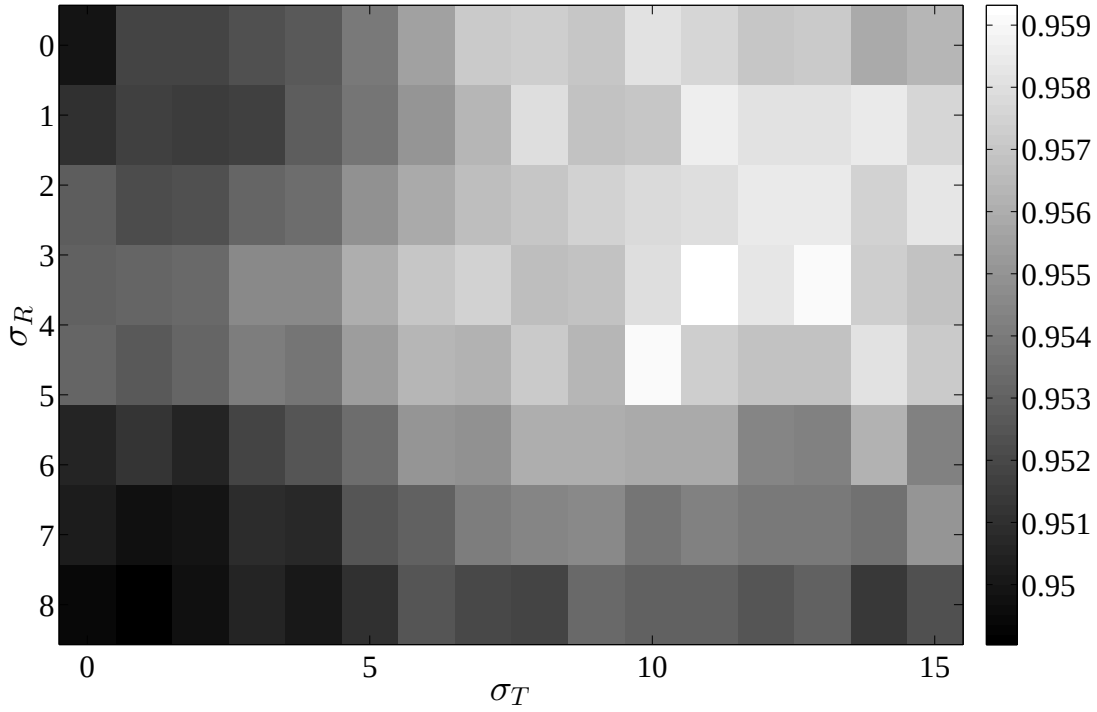

Рис. 11: Логистическая регрессия, S_2-N, триграммы

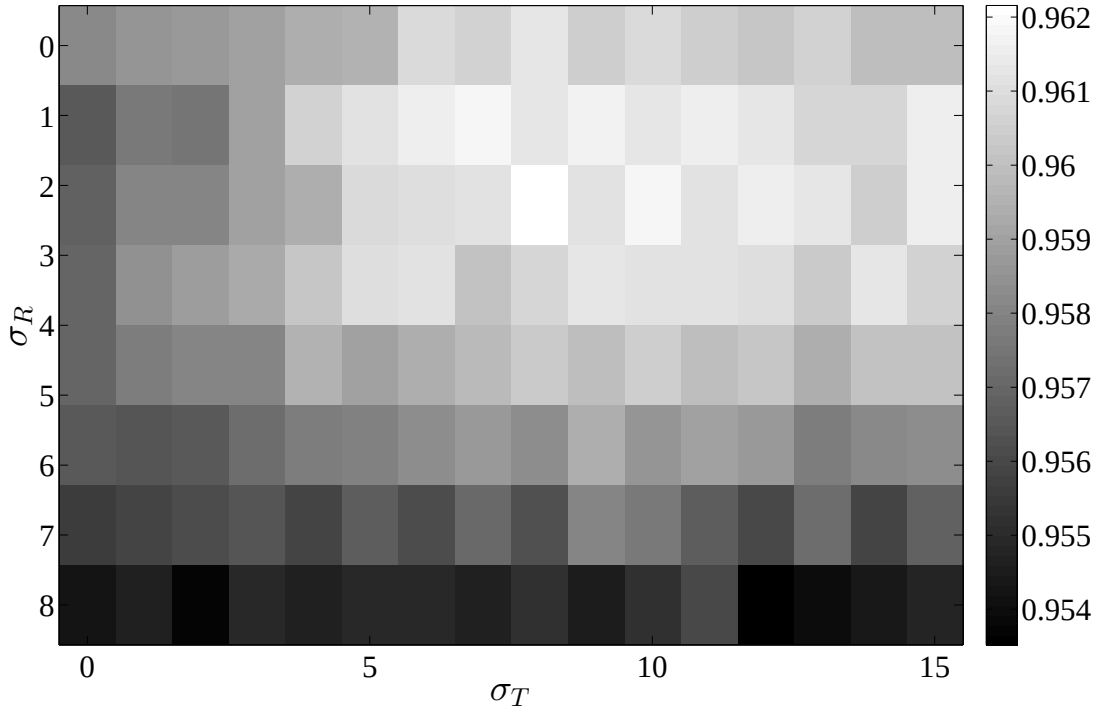

Рис. 12: Случайный лес, S_2-N, биграммы

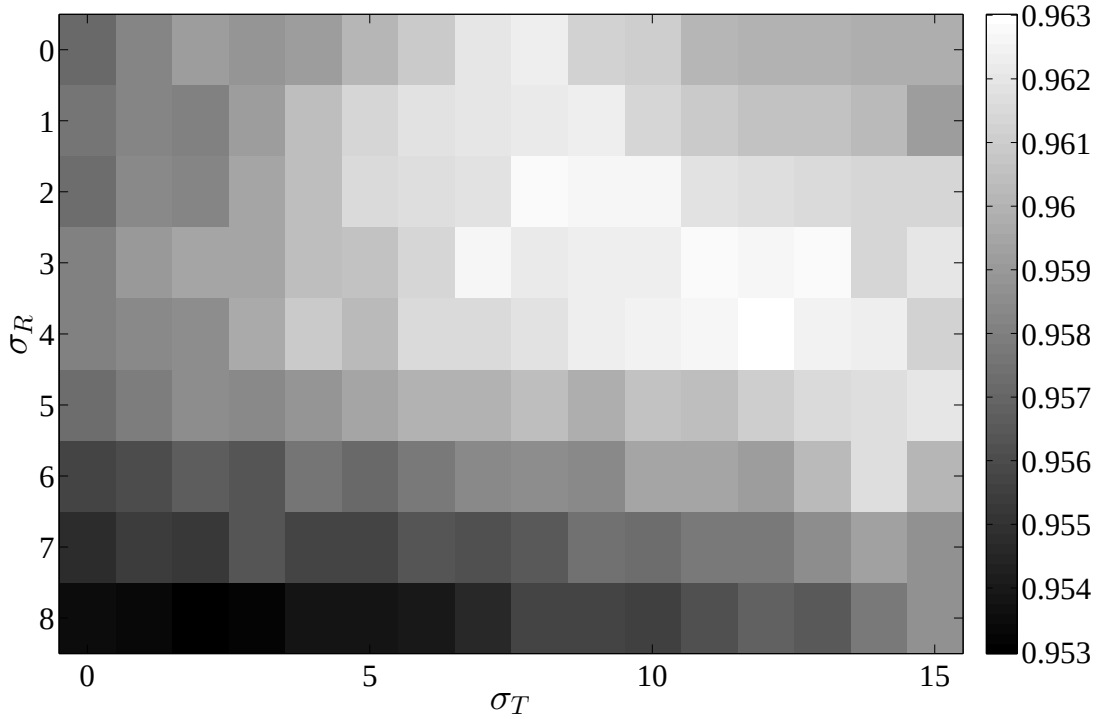

Рис. 13: Случайный лес, S_3-N, биграммы

5.3 Улучшение качества диагностики

В таблице указан прирост значений AUC в процентах при использовании различных методов нечёткого кодирования, для разных алгоритмов классификации. Для каждой болезни указан результат при средних оптимальных значениях σ_T, σ_R.

болезнь	S_1-N, SA	S_2-N, LR	S_3-N, RF	A-L, RF
оптимальные параметры σ_T, σ_R	10, 3	11, 3	11, 4	10, 2
анемия железодефицитная	0,8	0,8	1,3	1,9
мочекаменная болезнь	0,9	0,9	1,3	0,9
рак общий	0,4	1,2	0,9	0,8
гастродуоденит гипоацидный	1,1	0,7	0,8	0,7
холецистит хронический	0,9	1	0,5	0,5
дискинезия ЖВП	0,7	1,3	0,4	0,2
аденома простаты	0,4	1	0,7	0,5
узловой зоб щитовидный железы	0,5	0,4	1	0,4
миома матки	0,9	0,8	0,5	0,1
сахарный диабет	0,5	0,7	0,5	0,4
язвенная болезнь	0,4	0,6	0,4	0,6
гипертоническая болезнь	0,7	1	0,2	0,2
ишемическая болезнь сердца	0,7	0,7	0,1	0,2
желчнокаменная болезнь	0	0	0,2	0,3
аднексит хронический	-0,4	-0,2	0,3	0,3
гастродуоденит гиперацидный	0,4	0	-0,4	-0,2
вегетососудистая дистония	-0,1	0	-0,3	-0,1
некроз головки бедренной кости	-0,2	-0,4	-0,3	-0,2

Для подавляющего большинства заболеваний удалось достичь улучшения качества классификации при использовании всех описанных методов. Средний прирост составил около 0,5% вне зависимости от способа кодирования и алгоритма. Однако, по болезням он распределён неравномерно: для некоторых болезней достигает 1,9%, для двух болезней отсутствует, а ещё для двух отмечено ухудшение качества классификации. Данный факт в настоящее время остаётся невыясненным и требует дальнейшего исследования.

6 Заключение

Основные результаты работы

- Предложены методы предобработки сигналов ЭКГ, сглаживающие влияние шумов и неопределённостей при дискретизации в методе Успенского. Рассмотрены общие методы, использующие семплирование, не основанные ни на каких предположениях о методе кодирования, а также аналитический метод для частного случая метода Успенского.
- Показано, что метод аналитического интегрирования существенно быстрее и не снижает качество классификации, несмотря на предположения о независимости и параметрическом виде распределений.
- Найдены оптимальные параметры модели измерений и показано, что они не зависят от метода нечёткого кодирования и выбора модельного распределения.
- Достигнуто улучшение качества классификации до 1,9% для некоторых болезней

По результатам работы сделан доклад на международной конференции и опубликована статья [13].

Список литературы

[1] Баевский Р. М., Иванов Г. Г. Вариабельность сердечного ритма: теоретические аспекты и возможности клинического применения. *Ультразвуковая и функциональная диагностика.* 2001. № 3. С. 108–127.

[2] Баевский Р. М., Иванов Г. Г., Чирейкин Л. В., Гаврилушкин А. П., Довгалевский П. Я., Кукушкин Ю. А., Миронова Т. Ф., Прилуцкий Д. А., Семенов Ю. Н., Федоров В. Ф., Флейшман А. Н., Медведев М. М. Анализ вариабельности сердечного ритма при использовании различных электрокардиографических систем (методические рекомендации). *Вестник аритмологии.* 2001. № 24. С. 65–87.

[3] Успенский В. М. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов. М.: Экономика и информатика, 2008. 116 с.

[4] Успенский В. М. Информационная функция сердца. *Клиническая медицина.* 2008. Т. 86. № 5. С. 4–13.

[5] Uspenskiy V. M. Information Function of the Heart. Biophysical substantiation of technical requirements for electrocardioblock registration and measurement of electrocardiosignals parameters acceptable for information analysis to diagnose internal diseases. In: *Joint International IMEKO TC1+TC7+TC13 Symposium.* August 31–September 2, 2011, Jena, Germany.

[6] Uspenskiy V. M. Information Function of the Heart. A Measurement Model. In: *Measurement 2011: 8-th International Conference*. Smolenice, Slovakia, April 27–30, 2011. Pp. 383–386.

[7] Uspenskiy V. M. Diagnostic System Based on the Information Analysis of Electrocardiogram. In: *Proceedings of MECO 2012. Advances and Challenges in Embedded Computing.* Bar, Montenegro, June 19–21, 2012. Pp. 74–76.

[8] Успенский В. М., Кравченко Ю. Г., Павловский К. П., Авербах Ю. И. Устройство экспресс-диагностики заболеваний внутренних органов и онкопатологии. Патент на изобретение № 2159574 от 27 ноября 2000 г.

[9] Успенский В. М. Способ диагностики болезней неинфекционной этиологии. Патент на изобретение № 2157093 от 10 октября 2000 г.

[10] Успенский В. М. Способ диагностики заболеваний внутренних органов неинфекционной природы на любой стадии их развития. Патент на изобретение № 2163088 от 20 февраля 2001 г.

[11] Успенский В. М. Способ суточного кардиомониторирования для определения наличия и активности заболеваний человека неинфекционной природы. Патент на изобретение № 2211658 от 10 сентября 2003 г.

[12] Успенский В. М. Способ диагностики заболеваний внутренних органов. Патент на изобретение № 2407431 от 27 декабря 2010 г.

[13] Uspenskiy V. M., Vorontsov K. V., Tselykh V. R., Bunakov V. A. Information Function of the Heart: Discrete and Fuzzy Encoding of the ECG-Signal for Multidisease Diagnostic System. In: *Advanced Mathematical and Computational Tools in Metrology — AMCTM 2014*.

[14] Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. М.: Вильямс, 2011.

[15] Torshin I. Yu. The study of the solvability of the genome annotation problem on sets of elementary motifs. *Pattern Recognition and Image Analysis*. 2011. Vol. 21, Issue 4. Pp. 652–662.

[16] Сергеев А. Г. Метрология и метрологическое обеспечение: учебник. М.: Высшее образование, 2008. 575 с.