Монотонные классификаторы для задач медицинской диагностики

М. Ю. Швец, А. В. Зухба, К. В. Воронцов

Московский физико-технический институт

Конференция ММРО-17

19-25 сентября 2015

Задача построения монотонного классификатора

Дано

$X = (x_i, y_i)_{i=1}^{\ell}$ – обучающая выборка
$x_i = (f_1(x_i), \ldots, f_t(x_i))$ – признаковое описание объекта x_i
$y_i = \{-1, +1\}$.

Построить

Монотонный классификатор
$a : X \to Y \qquad x < x' \to a(x) \leq a(x')$

Критерии

- Минимизация числа дефектных пар путем удаления минимального числа объектов $|D| = 0$, $|X| \to \max$
- Минимизация числа используемых признаков при сохранении качества классификации $F \subseteq \mathbb{F} = \{f_1, \ldots f_t\}$

Отбор признаков

Сортирующий критерий $g : \mathbb{F} \to \mathbb{R}$

Выбор k лучших признаков по $g(f_j)$

Средняя частота и встречаемость признака j в классе y

$$P_{jy} = \frac{1}{\ell_y} \sum_i f_j(x_i)\,[y_i = y]; \quad B_{jy} = \frac{1}{\ell_y} \sum_i [f_j(x_i) \geq \theta]\,[y_i = y]$$

Используемые критерии

Bayes	$g(f_j) = \ln P_{j+} - \ln P_{j-}$
Freq	$g(f_j) = P_{j+}$
FreqDiff	$g(f_j) = P_{j+} - P_{j-}$
FreqDiffAbs	$g(f_j) = \lvert P_{j+} - P_{j-} \rvert$
Occur	$g(f_j) = B_{j+}$
OccurDiff	$g(f_j) = B_{j+} - B_{j-}$
OccurDiffAbs	$g(f_j) = \lvert B_{j+} - B_{j-} \rvert$

Варианты монотонной классификации

1. Монотонная классификация ближайшего соседа с предварительной монотонизацией выборки (Воронцов, 2000)
2. Классификатор ближайшего соседа с M-функцией расстояния, не требующий предварительной монотонизации (Махина, 2012)

---

1. Воронцов К. В. Оптимизационные методы линейной и монотонной коррекции в алгебраическом подходе к проблеме распознавания // ЖВМ и МФ. — 2000. — Т. 40, N° 1. — С. 166–176.

2. Махина Г.А. О восстановлении монотонных булевых функций методом ближайшего соседа // Международная конференция «Интеллектуализация обработки информации» (ИОИ-9), Черногория, г.Будва, 16–22 сентября 2012 г. С.78-81.

Отбор объектов (монотонизация выборки)

Количество дефектных пар, в которых участвует объект

$L_i = \{x_k \in X : \ y_i \neq y_k, \ (x_i, x_k) - \text{дефектная пара}\}.$

Удаление минимального числа объектов

Задача получения подвыборки максимальной мощности, в которой отсутствуют дефектные пары, является полиномиальной по количеству объектов.

α-монотонизация

1. Упорядочим объекты класса $y = -1$: $x_{i_1}, x_{i_2} \ldots x_{i_s} \ldots,$ по убыванию $|L_i|$: $|L_{i_1}| > \cdots > |L_{i_s}| > 0 = |L_{i_{s+1}}| = \ldots;$
2. удалим из выборки первые s' объектов $\{x_{i_1} \ldots x_{i_{s'}}\}$, где s' выбрано из условия $s' = \lfloor \alpha s \rfloor$;
3. удалим эти объекты также из всех множеств L_i;
4. удалим из выборки все объекты x_i класса $y = +1$, у которых $|L_i| > 0$.

Монотонный классификатор ближайшего соседа

Верхняя и нижняя тень объекта $x_i \in X$

$M_i^+ = \{u \in \mathbb{X} : x_i \preceq u\} \quad M_i^- = \{u \in \mathbb{X} : u \preceq x_i\}$

Расстояние до тени

$\rho(u, M_i) = \min\limits_{v \in M_i} \rho(u, v)$, где $\rho(u, v)$ – манхэттенское расстояние

Вычисление расстояния до тени

Расстояния от объекта $u \in \mathbb{X}$ до верхней и нижней теней объекта $x_i \in X$ вычисляются по следующим формулам:

$$\rho(u, M_i^-) = \sum_{f_j \in F} [f_j(u) - f_j(x_i)]_+,$$

$$\rho(u, M_i^+) = \sum_{f_j \in F} [f_j(x_i) - f_j(u)]_+.$$

Дискриминантная функция

Монотонный классификатор ближайшего соседа

$x_k = \arg\min_{x_i \in X} \rho(u, M_i)$ – ближайший к объекту $u \in \mathbb{X}$

$$a(u) = y_k$$

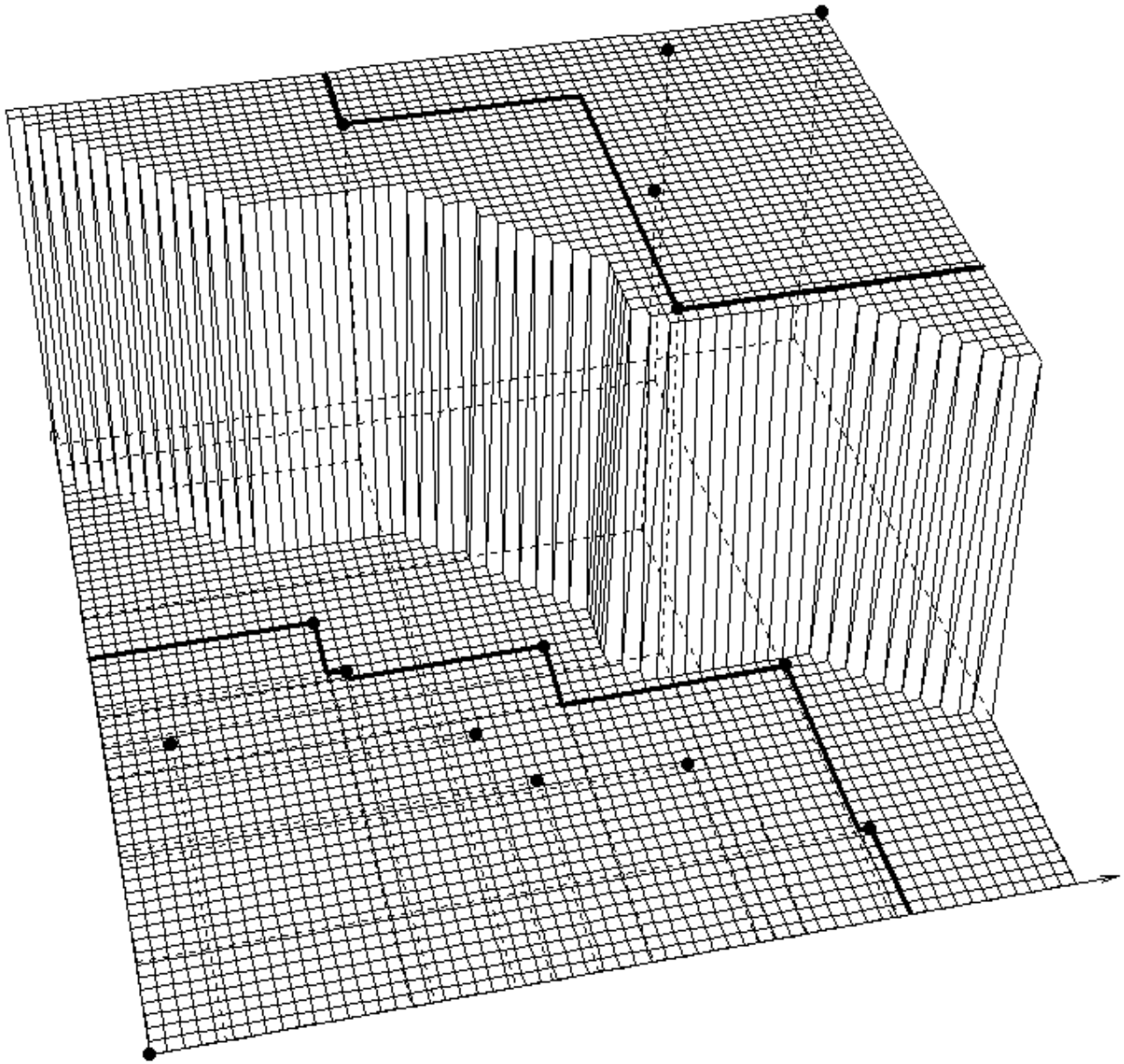
Непрерывная дискриминантная функция

$$\tilde{a}(u) = \begin{cases} a(u), & \rho_- = 0 \text{ или } \rho_+ = 0 \\ \frac{\rho_- - \rho_+}{\rho_- + \rho_+}, & \text{иначе.} \end{cases}$$

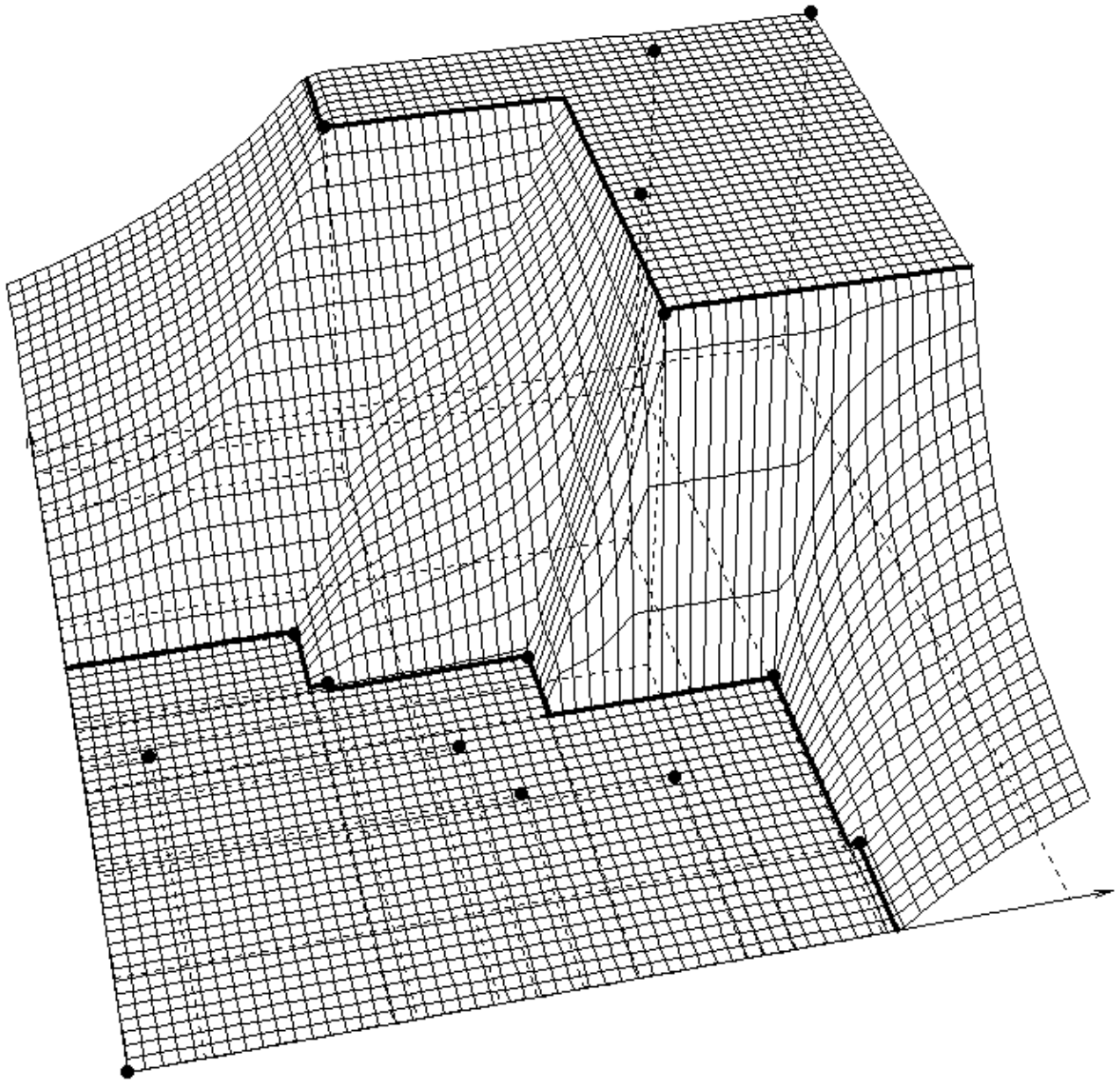
Здесь $\rho_y = \min_{y_i = y} \rho(u, M_i)$. Такая дискриминантная функция нужна для вычисления значения AUC.

Утверждение

Функция $\tilde{a}$ является монотонной.



Монотонный классификатор ближайшего соседа



Непрерывная дискриминантная функция

Классификатор ближайшего соседа с M-функцией расстояния, не требующий монотонизации

M-расстояние

$r : \mathbb{X} \times X \to E_N$, где $N = (nt)^2 + nt + 1$, задаваемая правилом $r(u, x_i) = nt\rho(u, M_i) + (nt - \rho(u, x))$.

Утверждение (о сохранении монотонности M-функции)

Для любых $u, v \in \mathbb{X}$, таких что $u \preceq v$, выполнено

$\forall x_i \in X \ (y_i = +1) : \quad r(u, x_i) \geq r(v, x_i)$

$\forall x_i \in X \ (y_i = -1) : \quad r(u, x_i) \leq r(v, x_i)$

Теорема

При использовании метода ближайшего соседа с M-функцией расстояния получаемая функция является монотонной.

Эксперимент

Информационный анализ ЭКГ сигналов (Успенский, 2008)

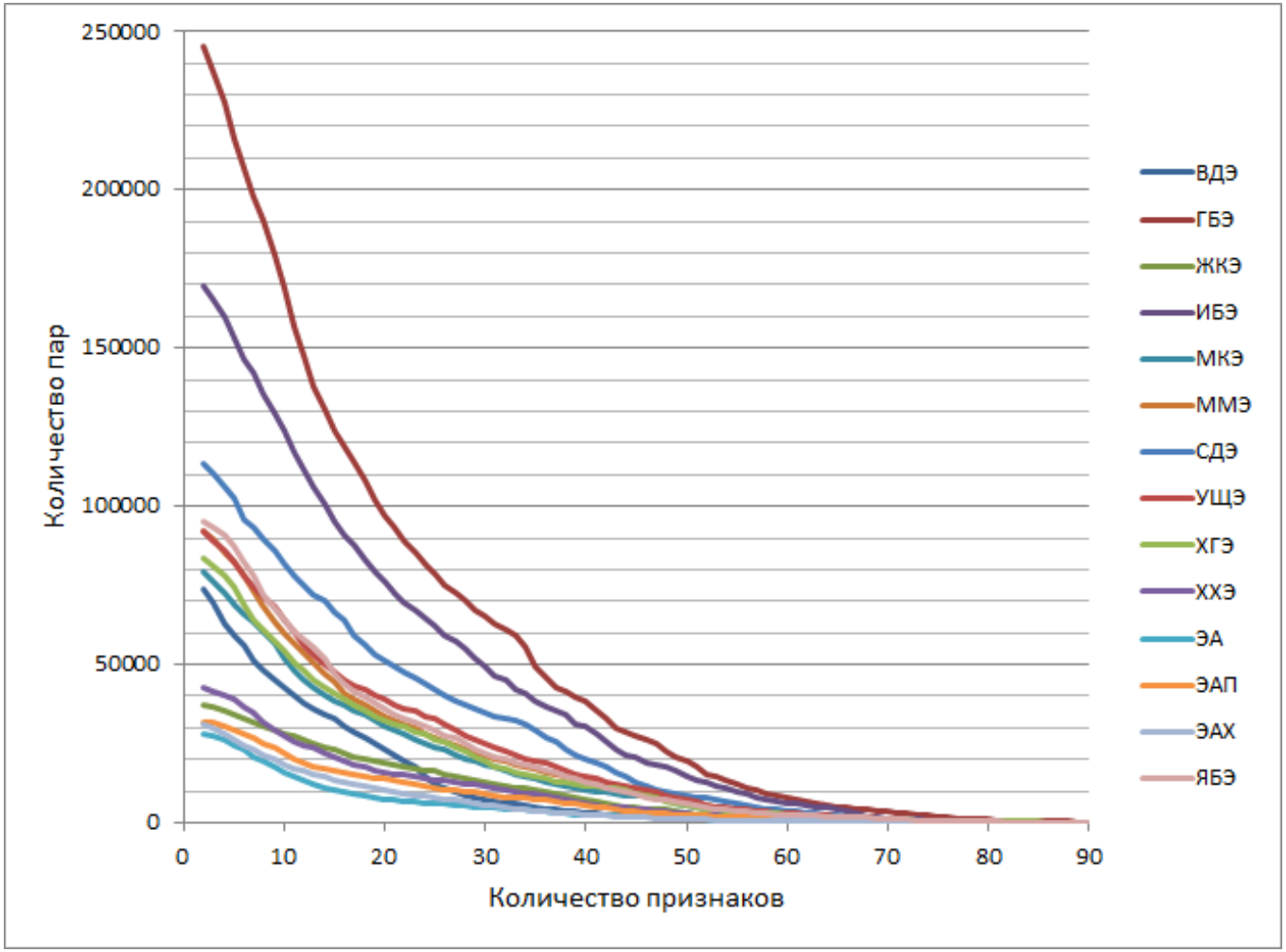
- Объем выборки – 10921
- Число болезней – 14
- Число здоровых – 198
- Признаки – частоты 216 триграмм по кодограмме в 6-буквенном алфавите

Сравнение моделей

- Функционал качества – AUC.
- Скользящий контроль по 10 блокам, 40 запусков.

Успенский В.М. Информационная функция сердца // Клиническая медицина, - 2008. – Т. 86. - №5. – С. 4-13

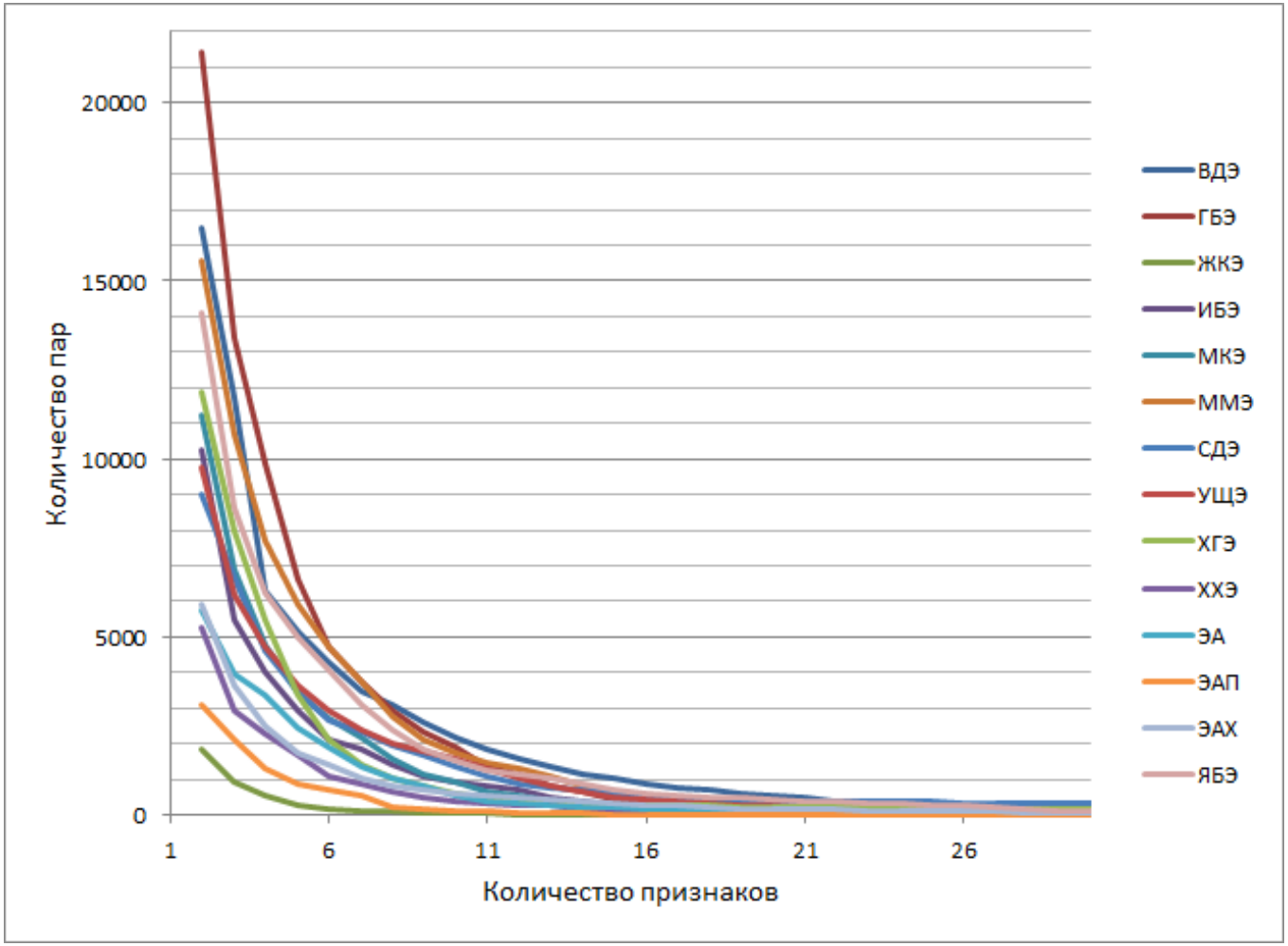
Монотонные пары



Предположение о хорошей монотонности выборки выполняется.

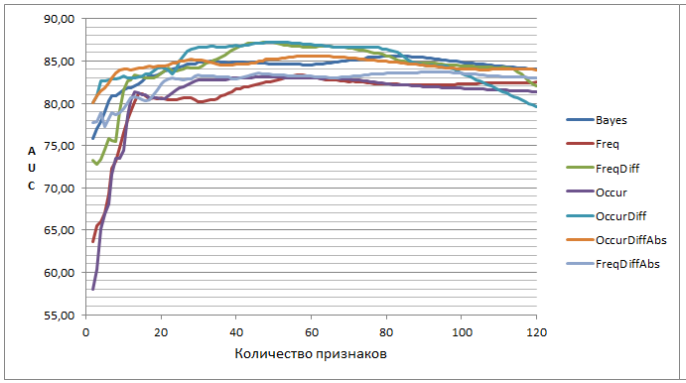
С ростом размерности количество пар быстро убывает.

Дефектные пары

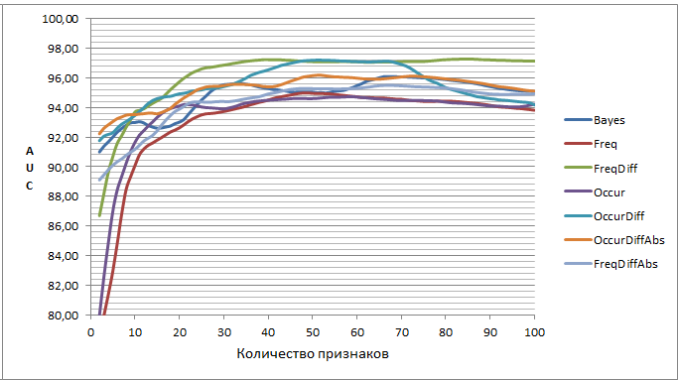


Дефектных пар существенно меньше, чем монотонных.

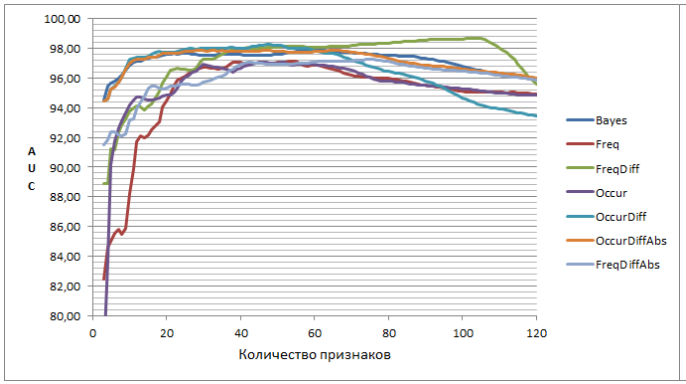
Сравнение сортирующих критериев



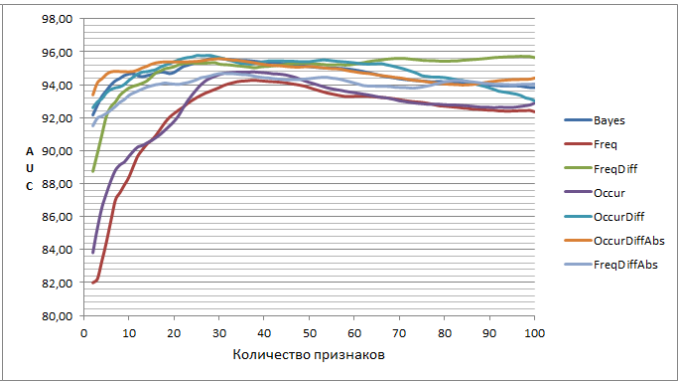
(a) ВДЭ



(b) ГБЭ



(c) ЖКЭ



(d) ЭАП

Высокое качество классификации на небольших размерностях.

Сравнение результатов

	ВДЭ	ГБЭ	ЖКЭ	ИБЭ	МКЭ	ММЭ	СДЭ
Monotonic	87,26	**97,29**	98,67	97,99	**95,47**	**93,67**	96,68
M_func	86,55	96,87	98,03	97,91	94,86	91,51	96,01
logReg	**87,62**	96,91	**99,00**	**98,21**	95,11	93,52	**97,08**
Syndr	86,35	96,60	98,90	97,84	95,17	93,37	96,66
	УЩЭ	ХГЭ	ХХЭ	ЭА	ЭАП	ЭАХ	ЯБЭ
Monotonic	95,67	**95,65**	**95,56**	**90,75**	95,78	91,82	94,63
M_func	94,84	93,38	94,59	88,87	95,59	91,59	94,22
logReg	**95,75**	95,22	95,07	90,04	**96,62**	**92,42**	**94,69**
Syndr	95,17	94,77	95,51	89,27	96,59	91,90	94,67

- Monotonic – монотонный классификатор ближайшего соседа с отбором объектов и признаков
- M_func – монотонный классификатор с M-функцией расстояния
- logReg – логистическая регрессия на главных компонентах
- Syndr – синдромный алгоритм

Заключение

- Предложены несколько способов одновременного отбора объектов и признаков для монотонного классификатора ближайшего соседа.
- Показано, что при диагностике заболеваний по ЭКГ качество классификации превосходит линейные модели для ряда заболеваний.