Монотонные классификаторы для задач медицинской диагностики

Швец Михаил

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель ст.н.с ВЦ РАН, д.ф.-м.н.
К. В. Воронцов

Группа 174, 1 июля 2015

Цели

- Предложить вычислительно эффективные методы отбора признаков и объектов для монотонных классификаторов
- Применить разработанные методы к задаче диагностики заболеваний по электрокардиосигналу

Определения и обозначения

$\mathbb{X} = \{x_1, \dots, x_m\}$, $y : \mathbb{X} \to Y = \{-, +\}$.

Множество признаков $P = \{p_1, \dots, p_t\}$, где
$p_j : \mathbb{X} \to E_j$. $\forall j : E_j = \{0, \dots, n-1\}$.

Пространство $W = E_1 \times E_2 \times \dots \times E_t$.

$\alpha \in W$ предшествует $\beta \in W$ $(\alpha \preceq \beta)$ по множеству $Q \subseteq P$, если $\forall p_j \in Q$ выполнено $p_j(\alpha) \le p_j(\beta)$.

Пара $x_i, x_k \in \mathbb{X}$, $y_i = -1$ и $y_k = +1$, является **монотонной** по $Q \subseteq P$, если $x_i \preceq x_k$, и **дефектной**, если $x_k \preceq x_i$.

$\mathbb{X}$ – **монотонная** по $Q \subseteq P$, если в $\mathbb{X}$ нет дефектных пар.

Монотонный классификатор – функция $f : W \mapsto Y$, удовлетворяющая условиям монотонности
$\forall u, v \in W : u \preceq v$ по множеству $Q \subseteq P \longrightarrow f(u) \le f(v)$

О вычислительной сложности задач (А. Зухба)

Теорема 1

Задача выбора признаков так, чтобы монотонных пар было не менее m, а дефектных не более d, является NP-трудной.

Теорема 2

Задача выбора признаков так, чтобы монотонных пар было не менее m, а признаков не менее q, является NP-трудной.

Теорема 3

Задача выбора признаков так, чтобы дефектных пар было не более d, а признаков не более q, является NP-трудной.

Теорема 4

Задача выбора объектов так, чтобы монотонных пар было не менее m, а дефектных не более d, является NP-трудной.

Отбор признаков

Определение критерия

Сортирующий критерий – функция $g : P \to \mathbb{R}$. Признаки упорядочиваются $p_{j_1} \dots p_{j_t}$ так, что $g(p_{j_1}) > g(p_{j_2}) > \dots > g(p_{j_t})$. При фиксированном k, выбор признаков проводится по правилу $Q = \{p_{j_1} \dots p_{j_k}\}$.

Средняя частота и встречаемость признака

$$F_j(y) = \frac{\sum\limits_{x_i \in \mathbb{X}} p_j(x_i)[y_i = y]}{\sum\limits_{x_i \in \mathbb{X}} [y_i = y]}; \quad B_j(y, \theta) = \frac{\sum\limits_{x_i \in \mathbb{X}} \big[p_j(x_i) \geq \theta\big][y_i = y]}{\sum\limits_{x_i \in \mathbb{X}} [y_i = y]}$$

Используемые критерии

$$g_{F+}(p_j) = F_j(+) \qquad g_{B+}(p_j) = B_j(+)$$
$$g_{dif(F)}(p_j) = F_j(+) - F_j(-) \qquad g_{dif(B)}(p_j) = B_j(+) - B_j(-)$$
$$g_{|dif(F)|}(p_j) = |F_j(+) - F_j(-)| \qquad g_{|dif(B)|}(p_j) = |B_j(+) - B_j(-)|$$

а также $g_{NB}(p_j) = b_j$, b_j – веса линейного классификатора.

Отбор объектов (монотонизация выборки)

Количество дефектных пар, в которых участвует объект

$L_i = \{x_k \in \mathbb{X} :\ y_i \neq y_k,\ (x_i, x_k)$ – 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 $y = -1$: $x_{i_1}, x_{i_2} \dots x_s \dots,$ по убыванию $|L_i|$: $|L_{i_1}| > \dots > |L_{i_s}| > 0 = |L_{i_{s+1}}| = \dots;$
2. удалим из выборки первые s' объектов $\{x_{i_1} \dots x_{i_{s'}}\}$, где s' выбрано из условия $s' \leq \alpha s < s' + 1$;
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 $x_i \in \mathbb{X}$

$$M_i^+ = \{a \in W : x_i \preceq a\} \quad M_i^- = \{a \in W : a \preceq x_i\}$$

Расстояние до тени

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

Лемма (о вычислении расстояния до тени)

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

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

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

Результирующая функция

Монотонный классификатор ближайшего соседа

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

$$f(u) = y_k$$

Непрерывный аналог результирующей функции

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

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

Утверждение

Функция $\tilde{f}$ является монотонной.

Классификатор ближайшего соседа с M-функцией расстояния (Г. Махина)

M-расстояние

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

Утверждение (о сохранении монотонности M-функции)

Для любых $u, v \in W$, таких что $u \preceq v$, выполнено

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

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

Теорема

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

Композиция моделей

Взвешенное голосование

$$a(x) = F\left(b_1(x), \ldots, b_K(x)\right),$$

$$F\left(b_1(x), \ldots, b_K(x)\right) = \sum_{k=1}^{K} w_k b_k(x)$$

Выбор весов

$w_k = 1$ – простое голосование

$w_k = q^{k-1}$ – члены геометрической прогрессии

Утверждение

Если функция $b_1(x)$ монотонна на множестве признаков Q_1, а функция $b_2(x)$ монотонна на множестве признаков Q_2, то их линейная комбинация с неотрицательными весами w_1 и w_2, $w_1 b_1(x) + w_2 b_2(x)$, монотонна на множестве признаков $Q_1 \cup Q_2$.

Эксперимент

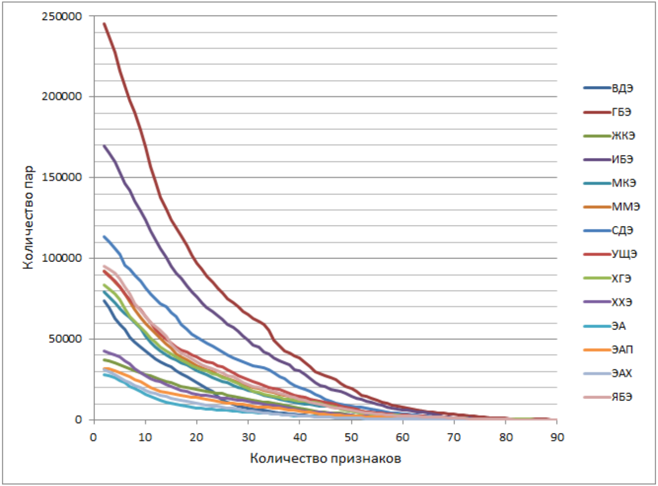
Данные [Успенский]

- Данные о 14 болезнях.
- Признаковое описание получено с помощью технологии информационного анализа электрокардиосигналов.

Сравнение моделей

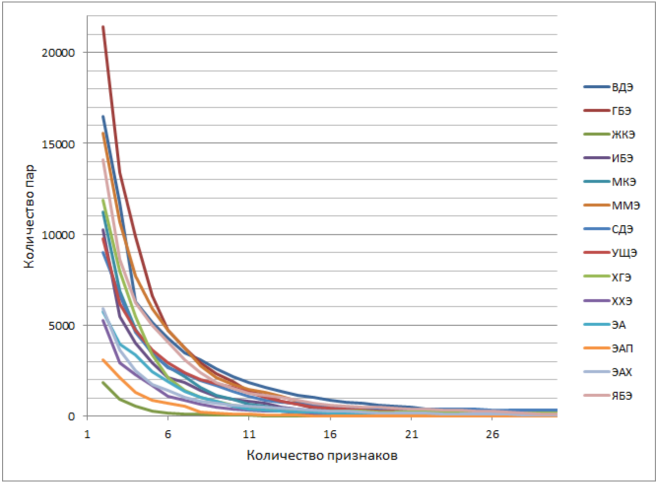
- Функционал качества – AUC.
- Скользящий контроль по 10 блокам, 40 запусков.

Монотонные пары



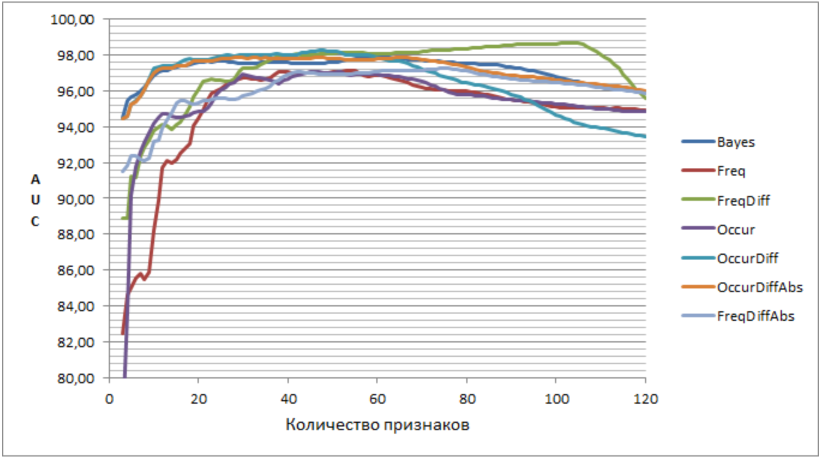
Предположение о хорошей монотонности выборки выполняется.
С ростом размерности количество пар быстро убывает.

Дефектные пары



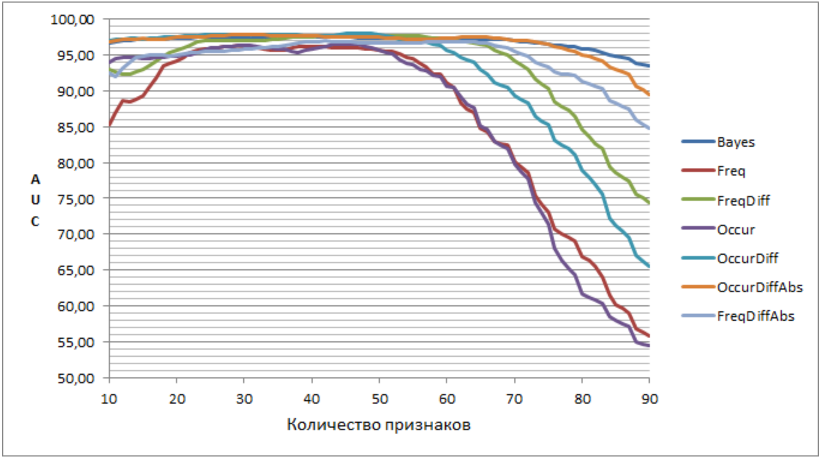
Дефектных пар существенно меньше, чем монотонных.

Сравнение сортирующих критериев



Высокое качество классификации на небольших размерностях.

Качество классификации. M-функция



Изменение параметра монотонизации

		α				
		0,1	0,3	0,5	0,7	0,9
ВДЭ	чувствительность	93,86	84,01	77,17	70,92	66,08
	специфичность	42,10	64,36	76,07	83,13	86,32
ГБЭ	чувствительность	97,49	93,50	90,32	88,38	87,07
	специфичность	51,99	72,03	81,84	85,43	87,11
ЖКЭ	чувствительность	95,81	93,67	92,92	91,90	89,65
	специфичность	80,83	87,81	88,96	90,72	92,84
ИБЭ	чувствительность	97,51	94,50	92,79	92,08	90,81
	специфичность	61,77	78,35	83,75	85,14	87,77
МКЭ	чувствительность	96,02	90,42	86,07	82,66	80,06
	специфичность	50,64	76,05	81,97	85,31	87,67
ММЭ	чувствительность	95,54	87,77	82,27	79,53	76,58
	специфичность	41,41	74,04	82,90	86,58	88,80
СДЭ	чувствительность	98,02	94,29	91,50	90,09	88,76
	специфичность	33,96	74,53	84,44	87,90	89,62
УЩЭ	чувствительность	97,34	91,72	88,34	85,73	83,56
	специфичность	35,30	74,61	82,92	88,03	89,96

С ростом α чувствительность падает, специфичность растет.

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

Заключение

- Исследованы различные способы отбора объектов и признаков для монотонного классификатора ближайшего соседа.
- Предложены жадные методы решения NP-трудных задач.
- Выполнена программная реализация и проведены численные эксперименты, показывающие применимость исследуемых моделей к задаче медицинской диагностики.

Направление развития работы:

- Предложение внутреннего критерия качества на основании количества монотонных и дефектных пар для стратегии пошагового добавления и удаления объектов и признаков.
- Развитие идеи композиции монотонных классификаторов, построение аналога синдромного алгоритма.