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

Швец Михаил Юрьевич

Монотонные классификаторы для задач медицинской диагностики

010900 — Прикладные математика и физика

Бакалаврская диссертация

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

Москва
2015 г.

Содержание

1 Введение

Теория информационной функции сердца, развитая в [1], основывается на предположении о том, что сигнал электрокардиограммы несет информацию о функционировании всех систем организма и каждое заболевание имеет уникальное влияние на этот сигнал. Рассматриваемый метод дает возможность на ранней диагностики, поскольку информация о заболевании может проявляться на любой стадии. Статистическое обоснование метода проведено в работе [2]. За время эксплуатации диагностической системы, работающей на этом методе, была накоплена выборка из более 20 тысяч прецедентов с данными о более чем 40 заболеваниях.

В данной работе для каждой болезни, представленной в выборке, рассматривается задача двухклассовой классификации. Каждый объект имеет признаковое описание, полученное в результате анализа электрокардиограммы в соответствии с теорией информационной функции сердца. Среди объясняющих признаков есть такие, по которым целевой признак является монотонным. Это означает, что чем выше значение соответствующих признаков, тем больше вероятность того, что рассматриваемый объект относится к классу больных. Таким образом, возникает задача отбора признаков. Вопросы вычислительной сложности этой задачи, а также задачи отбора объектов, были исследованы в работе [3]. Естественные требования, которые возникают в процессе решения, порождают NP-трудные задачи, поэтому в нашей работе предлагаются жадные критерии отбора признаков, называемые сортирующими критериями. Каждому признаку сопоставляется число, после чего происходит сортировка признаков и отбрасывание всех, кроме определенного числа признаков с наибольшими сопоставленными значениями.

Во многих работах для решения задачи классификации с монотонными ограничениями рассматриваются линейные модели. Однако множество монотонных моделей значительно шире. В работе рассматривается монотонный классификатор ближайшего соседа. Этот классификатор был предложен в [4], [5] в качестве корректирующей операции в алгебраическом подходе [6]. Монотонный классификатор ближайшего использовался для задачи ранжирования [7]. В работе [8] содержится теоретическое обоснование обобщающей способности рассматриваемого нами метода на основании оценок полного скользящего контроля.

При построение монотонного классификатора ближайшего соседа мы будем требовать монотонности выборки. Это предположение, которое в общем случае неверно.

Монотонизация на практике проводится с помощью переопределения меток классов [9] или отбрасывания объектов обучающей выборки. В этой работе монотонизация проводится с помощью отбрасывания объектов. Сравниваются различные типы монотонизации.

Также в работе экспериментально исследуется классификатор ближайшего соседа с введенной в [10] функцией расстояния. Отличительная особенность этой модели в том, что она не требует предположения о монотонности обучающей выборки. При этом результирующий классификатор является монотонным.

Цель работы.

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

Положения, выносимые на защиту:

- 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.

2 Постановка задачи

Дано конечное множество объектов $\mathbb{X} = \{x_1, \ldots, x_m\}$, называемое обучающей выборкой, для которых известна истинная классификация $y : \mathbb{X} \to Y$. Будем обозначать $y_i = y(x_i)$. Рассматривается задача с двумя классами, то есть $Y = \{-1, +1\}$. Далее для краткости будем использовать обозначение $Y = \{-, +\}$. Также дано конечное множество признаков $P = \{p_1, \ldots, p_t\}$ – отображений вида $p_j : \mathbb{X} \to E_j$, где E_j – линейно-упорядоченное множество. В данной работе $\forall j : E_j = \{0, \ldots, n-1\}$.

Каждый объект отождествляется с вектором из пространства $W = E_1 \times E_2 \times \cdots \times E_t$, которое является пространством t-мерных n-значных векторов.

2.1 Понятие монотонности

Для объектов введем отношение предшествования. Будем говорить, что $\alpha \in W$ предшествует (строго предшествует) $\beta \in W$ по множеству признаков $Q \subseteq P$, если $\forall p_j \in Q$ выполнено $p_j(\alpha) \leqslant p_j(\beta)$ $(p_j(\alpha) < p_j(\beta))$. В таком случае будем писать $\alpha \preceq \beta$ $(\alpha \prec \beta)$. Если два объекта не находятся в отношении предшествования, будем говорить, что они несравнимы.

Определение 2.1. *Обучающая выборка $\mathbb{X}$ называется **монотонной** по множеству признаков $Q \subseteq P$, если для всех $x_i, x_k \in \mathbb{X}$ из $x_i \preceq x_k$ по Q следует $y_i \leqslant y_k$.*

Нам также понадобятся отношения на парах объектов разных классов из обучающей выборки.

Определение 2.2. *Рассмотрим два объекта $x_i, x_k \in \mathbb{X}$ $(x_i \neq x_k)$, такие что $y_i = -1$ и $y_k = +1$. Будем говорить, что пара объектов является **монотонной** по множеству признаков $Q \subseteq P$, если $x_i \preceq x_j$ по Q, и **дефектной**, если $x_j \preceq x_i$ по Q.*

Отметим, что монотонная выборка – это выборка, в которой количество дефектных пар равно нулю.

Далее рассмотрим ограничение W_Q пространства W на множество признаков $Q = \{p_{j_1}, \ldots p_{j_{|Q|}}\} \subseteq P$:

$$W_Q = E_{j_1} \times E_{j_2} \times \cdots \times E_{j_{|Q|}}. \tag{2.1}$$

Напомним, что $\forall j : E_j = \{0, \ldots, n-1\}$. Таким образом, W_Q является пространством $|Q|$-мерных n-значных векторов.

В работе требуется выбрать подмножество признаков $Q \subset P$ и определить монотонную на W_Q функцию, называемую монотонным классификатором, которая сопоставляет объекту некоторый ответ.

Определение 2.3. ***Монотонный классификатор*** *– функция $f : W_Q \mapsto Y$, удовлетворяющая условиям монотонности*

$$\forall u, v \in W_Q : \quad u \preceq v \text{ по множеству } Q \longrightarrow f(u) \leqslant f(v)$$

В работе [11] доказана лемма о существовании такой функции. Ниже приводится лишь формулировка леммы.

Лемма 2.1. *Монотонный классификатор, для которого выполнено условие*

$$f(x_i) = y_i, \text{ для всех } x_i \in \mathbb{X},$$

то есть не допускающий ошибок на выборке $\mathbb{X}$*, существует тогда и только тогда, когда выборка* $\mathbb{X}$ *является монотонной.*

2.2 Модель

Для построения искомого монотонного классификатора предлагается использовать модель, состоящую из нескольких блоков. Модель включает в себя отбор признаков, монотонизацию выборки, построение монотонного классификатора ближайшего соседа.

2.3 О вычислительной сложности некоторых задач

Естественными требованиями при работе с выборкой являются следующие:

- число дефектных пар должно быть как можно меньше;
- число монотонных пар должно быть как можно больше;
- число отобранных признаков – как можно меньше;
- число отобранных объектов – как можно больше.

Однако, эти четыре условия являются зависимыми и частично противоречат друг другу. Вопросы вычислительной сложности отбора признаков и объектов были исследованы в работе [3]. Ниже приводятся формулировки теорем о вычислительной сложности некоторых корректно поставленных задач.

Теорема 2.1. *Задача выбора признаков так, чтобы монотонных пар было не менее* m*, а дефектных не более* d*, является NP-трудной.*

Теорема 2.2. *Задача выбора признаков так, чтобы монотонных пар было не менее* m*, а признаков не менее* q*, является NP-трудной.*

Теорема 2.3. *Задача выбора признаков так, чтобы дефектных пар было не более d, а признаков не более q, является NP-трудной.*

Теорема 2.4. *Задача выбора объектов так, чтобы монотонных пар было не менее m, а дефектных не более d, является NP-трудной.*

Таким образом, большинство задач, которые возникают естественным образом, являются NP-трудными, вследствие чего данная работа основывается на жадных методах отбора объектов и признаков.

2.4 Жадный отбор признаков

При построении модели проводится жадный отбор признаков по некоторому критерию. Напомним, что каждый объект выборки $x_i \in \mathbb{X}$ описывается t числовыми признаками. На произвольном множестве признаков P введем **сортирующий признаки критерий** – функцию

$$g : P \to \mathbb{R},$$

которая каждому признаку ставит в соответствие число на основании обучающей выборки. Упорядочим признаки $p_{j_1} \dots p_{j_t}$ так, что $g(p_{j_1}) \geqslant g(p_{j_2}) \geqslant \dots \geqslant g(p_{j_t})$. Фиксируя некоторое число k, проведем отбор k признаков по правилу

$$Q = \{p_{j_1} \dots p_{j_k}\}.$$

2.5 Монотонизация

Если обучающая выборка не является монотонной по множеству отобранных признаков Q, то для построения монотонного классификатора, который будет описан ниже, в работе необходимо решить задачу монотонизации. Для этого из выборки удаляются объекты, нарушающие монотонность.

Таким образом, возникает задача нахождения монотонной по Q подвыборки обучающей выборки $\widetilde{X} \subseteqq \mathbb{X}$.

2.6 Классификатор монотонного ближайшего соседа

Для каждого объекта выборки введем понятия верхней и нижней тени на множестве признаков $Q \subseteq P$. Для этого нам понадобится определенное в (2.1) пространство W_Q.

Определение 2.4. *Верхней (нижней) тенью объекта $x_i \in \mathbb{X}$ на Q называется множество $M_i^+ = \{a \in W_Q : x_i \preceq a\}$ ($M_i^- = \{x \in W_Q : x_i \succeq a\}$).*

Для каждого объекта выборки введем понятие тени, в котором будет учтена метка класса объекта.

Определение 2.5. *Тенью объекта $x_i \in \mathbb{X}$ называется нижняя тень, если $y_i = -1$, и верхняя тень, если $y_i = +1$, то есть $M_i = M_i^{y_i}$, что означает*

$$M_i = \begin{cases} M_i^+, & y_i = +1 \\ M_i^-, & y_i = -1 \end{cases}. \tag{2.2}$$

В терминах манхэттенского расстояния

$$\rho(u, v) = \sum_{p_j \in Q} |p_j(u) - p_j(v)|, \quad u, v \in W_Q \tag{2.3}$$

введем понятия расстояния от произвольного объекта $u \in W_Q$ до тени объекта из выборки $x_i \in \mathbb{X}$:

$$\rho(u, M_i) = \min_{a \in M_i} \rho(u, a) \tag{2.4}$$

В работе [10] доказана лемма о вычислении расстояний до теней. Приведем здесь формулировку этой леммы.

Лемма 2.2. *Расстояния от объекта $u \in W_Q$ до верхней и нижней теней объекта $x_i \in \mathbb{X}$ по множеству признаков Q вычисляются по следующим формулам:*

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

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

Здесь использовано обозначение усеченной разности

$$[a - b]_+ = \begin{cases} a - b, & a \geqslant b, \\ 0, & a < b. \end{cases}$$

Таким образом, для построения монотонного классификатора ближайшего соседа, нужно найти объект из обучающей выборки, к тени которого ближе всего лежит

поданный на вход классификатора объект, после чего метку класса этого объекта назначить в качестве ответа. Во введенных обозначениях получаем:

$$x_k = \arg\min_{x_i \in \mathbb{X}} \rho(u, M_i), \tag{2.5}$$

$$f(u) = y_k. \tag{2.6}$$

2.7 Отбор эталонных объектов

Можно заметить, что для построения монотонного классификатора ближайшего соседа (2.5, 2.6) исходная обучающая выборка становится избыточной. А именно, любой объект $x_k \in \mathbb{X}$, что $\exists x_i \in \mathbb{X} : x_k \in M_i$, то есть лежащий в тени некоторого объекта x_i, может быть удален из выборки, так как любой объект $u \in W$, который может быть классифицирован по x_k, может также быть классифицирован по x_i, поскольку $\rho(u, M_i) \leqslant \rho(u, M_k)$. Таким образом, нужно отобрать из обучающей выборки эталонные объекты.

Определение 2.6. *Объект $x_k \in \mathbb{X}$ называется* ***эталонным****, если $\forall x_i \in \mathbb{X} : x_i \neq x_k$ верно $x_k \notin M_i$*

Окончательно, отбор эталонных объектов имеет вид:

$$\widetilde{X} = \{x_i \in \mathbb{X} : \ x_i - \text{эталонный}\}. \tag{2.7}$$

2.8 Классификатор ближайшего соседа с M-функцией расстояния

Функция, введенная в (2.5, 2.6), требует монотонности обучающей выборки. В работе [10] была предложена M-функция расстояния, использование которой не предполагает монотонности выборки.

Определение 2.7. ***M-функцией расстояния*** *называется функция двух аргументов $r : W \times \mathbb{X} \to E_N$, где $N = (nt)^2 + nt + 1$, задаваемая правилом*

$$r(u, x_i) = nt\rho(u, M_i) + (nk - \rho(u, x)), \tag{2.8}$$

где $u \in W$, $x_i \in \mathbb{X}$.

В той же работе доказаны теоремы, формулировки которых приведены ниже.

Теорема 2.5. *Если* $\alpha,\ \beta \in W:\ \alpha \preceq \beta$, *то выполняются соотношения*

$$\forall x_i \in \mathbb{X} : y_i = -1\ \rightarrow\ r(\alpha, u) \leqslant r(\beta, u);$$

$$\forall x_i \in \mathbb{X} : y_i = +1\ \rightarrow\ r(\alpha, u) \geqslant r(\beta, u).$$

Теорема 2.6. *При использовании метода ближайшего соседа с функцией расстояния (2.8) при классификации объектов на основе выборки* $\mathbb{X}$ *получается монотонная функция.*

2.9 Функционал качества

Качество классификации объектов построенным классификатором в работе оценивается с помощью функционала AUC (Area Under Curve).

В терминах выбранного функционала качества в работе предлагается помимо функции $f(u)$, введенной в (2.6, 2.5), рассматривать дискриминантную функцию $\tilde{f}$, которая позволяет более точно вычислять значение AUC и после соответствующей нормировки может быть интерпретирована как вероятность принадлежности объекта к классу больных.

$$\tilde{f}(u) = \begin{cases} f(u), & \rho_- = 0 \text{ или } \rho_+ = 0 \\ \dfrac{\rho_- - \rho_+}{\rho_- + \rho_+}, & \text{иначе.} \end{cases} \tag{2.9}$$

Здесь используется обозначение

$$\rho_y = \min_{y_i = y} \rho(u, M_i)$$

для кратчайшего расстояния от объекта u до тени класса $y \in \{-, +\}$.

Докажем монотонность функции $\tilde{f}$.

Утверждение 2.1. *Функция* $\tilde{f}$, *определенная в формуле 2.9, является монотонной.*

Доказательство.

Пусть $\alpha \preceq \beta$. Тогда $\rho_-(\alpha) \leqslant \rho_-(\beta)$ и $\rho_+(\alpha) \geqslant \rho_+(\beta)$.

Выпишем числитель разности $\tilde{f}(\alpha) - \tilde{f}(\beta)$, приведенной к общему знаменателю:

$$(\rho_-(\alpha) - \rho_+(\alpha))(\rho_-(\beta) + \rho_+(\beta)) - (\rho_-(\beta) - \rho_+(\beta))(\rho_-(\alpha) + \rho_+(\alpha)) =$$

$$= 2\left(\rho_-(\alpha)\,\rho_+(\beta) - \rho_+(\alpha)\,\rho_-(\beta)\right).$$

Используя имеющиеся неравенства, имеем

$$\rho_-(\alpha)\,\rho_+(\beta) - \rho_+(\alpha)\,\rho_-(\beta) \leqslant \rho_-(\alpha)\,\rho_+(\alpha) - \rho_+(\alpha)\,\rho_-(\alpha) = 0.$$

Получили

$$\tilde{f}(\alpha) \leqslant \tilde{f}(\beta).$$

■

3 Решение

Для решения задачи строится набор моделей, состоящих из нескольких блоков.

3.1 Средняя частота и встречаемость

Далее нам понадобятся понятия *средней частоты* признака p_j в классе y обучающей выборки

$$F_j(y) = \frac{\sum\limits_{x_i \in \mathbb{X}} p_j(x_i)\,[y_i = y]}{\sum\limits_{x_i \in \mathbb{X}} [y_i = y]}$$

а также *встречаемости* признака p_j в классе y обучающей выборки

$$B_j(y,\theta) = \frac{\sum\limits_{x_i \in \mathbb{X}} [p_j(x_i) \geqslant \theta]\,[y_i = y]}{\sum\limits_{x_i \in \mathbb{X}} [y_i = y]},$$

то есть доли объектов класса y, в которых значение признака p_j превышает пороговое значение θ.

3.2 Предварительный отбор признаков

По каждому признаку, которые принимаются к рассмотрению, должна быть хорошая монотонность. Предлагается на первом этапе оставить только те признаки p_j, по которым средние встречаемости для обоих классов $\{+,-\}$, значимо различаются.

$$Q_0 = \{p_j \in P :\ |F_j(+) - F_j(-)| > \theta_F\},$$

где θ_F – некоторый заданный порог.

3.3 Отбор признаков по сортирующему критерию

На основании средней частоты признаков в каждом из классов множества Y в работе вводится три сортирующих критерия:

$$g_{F+}(p_j) = F_j(+), \tag{3.1}$$

$$g_{diff(F)}(p_j) = F_j(+) - F_j(-), \tag{3.2}$$

$$g_{|dif(F)|}(p_j) = |F_j(+) - F_j(-)|\,. \tag{3.3}$$

Аналогично вводятся три критерия на основании встречаемости признаков в каждом из классов множества Y:

$$g_{B+}(p_j) = B_j(+), \tag{3.4}$$

$$g_{dif(B)}(p_j) = B_j(+) - B_j(-), \tag{3.5}$$

$$g_{|dif(B)|}(p_j) = |B_j(+) - B_j(-)|\,. \tag{3.6}$$

Веса линейного классификатора Также сортирующий критерий можно естественным образом ввести, используя веса признаков некоторого линейного классификатора.

Линейный классификатор – это решающее правило, построенное на основании взвешенной суммы признаков.

$$A(x) = \left[\sum_{p_j \in Q} b_j p_j(x) > \beta\right]$$

Построим наивный байесовский классификатор, который является линейным в случае бинарных признаков. Признаки объекта x_i становятся бинарными после естественного преобразования по правилу $p_j(x_i) \to [p_j(x_i) > \theta_{feat}]$, $x_i \in \mathbb{X}$, где θ_{feat} – порог бинаризации.

Определим сортирующий критерий

$$g_{NB}(p_j) = b_j, \tag{3.7}$$

где b_j – веса признаков построенного классификатора.

3.4 Монотонизация

Для каждого объекта обучающей выборки $x_i \in \mathbb{X}$ определим объекты противоположного класса, с которыми рассматриваемый объект образует дефектные пары:

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

Задача получения подвыборки максимальной мощности, в которой отсутствуют дефектные пары, является полиномиальной по количеству объектов. Эта задача сводится к задаче минимального вершинного покрытия двудольного графа. Однако, в случае несбалансированных классов при таком отборе из выборки будут удаляться объекты класса меньшей мощности.

В работе предлагаются жадные критерии удаления объектов из выборки. α-монотонизация позволяет сдвигать разделяющую поверхность в сторону одного из классов, варьирую отношение между чувствительностью и специфичностью. Для этого из выборки удаляется заданный процент объектов класса $-$, образующих наибольшее количество дефектных пар. Если из выборки удаляется какой-либо объект x_k, то этот объект также удаляется и из всех множеств $L_i : \ x_k \in L_i$. На последнем шаге монотонизации происходит удаление всех объектов класса $+$, которые все еще участвуют в образовании дефектных пар.

α-монотонизация

Вход: $\mathbb{X}$ – немонотонная выборка;

Q – множество признаков;

α – параметр монотонизации;

Выход: $\widetilde{X}$ – монотонная по Q выборка;

1: $\widetilde{X} = \mathbb{X}$

2: определить объекты класса -1 с $|L_i| > 0$ и их количество s:
$s = |X_s|$, где $X_s = \{x_i \in \mathbb{X} : \ y_i = -1, \ |L_i| > 0\}$;

3: упорядочить объекты X_s: $x_{i_1} \dots x_{i_s}$,
по убыванию $|L_i|$: $|L_{i_1}| > \dots > |L_{i_s}|$;

4: удалить из выборки первые s' объектов $\{x_{i_1} \dots x_{i_{s'}}\}$,
где s' выбрано из условия $s' \leqslant \alpha s < s' + 1$:

5: **пока** $s' \leqslant \alpha s$

6: $\quad \widetilde{X} = \widetilde{X} \setminus \{x_{i_{s'}}\}$;

7: $\quad$ удалить $x_{i_{s'}}$ из всех множеств L_i:

8: **для всех** $x_i \in \widetilde{X}:\ y_i = +1$

9: $L_i = L_i \setminus \{x_{i_{s'}}\}$;

10: удалить из выборки все объекты x_i класса $y = +1$, у которых $|L_i| > 0$:
$\widetilde{X} = \widetilde{X} \setminus \{x_i \in \widetilde{X}:\ y_i = +1,\ |L_i| > 0\}$.

3.5 Алгоритм

Приведем полную версию построенного алгоритма в виде псевдокода.

Обучение монотонного классификатора

Вход: $\mathbb{X}$ – обучающая выборка;

P – множество признаков;

g – сортирующий критерий;

α – параметр монотонизации;

θ – параметр;

k – количество признаков;

Выход: X – отобранные эталонные объекты;

Q – отобранные признаки;

предварительный отбор признаков и смена знаков:
$Q_1 = \emptyset$ – набор признаков, по которым наблюдается монотонность;
для всех $p_j \in P$
  рассчитать среднюю частоту признака j в каждом классе: $F_j(+)$ и $F_j(-)$;
  если $F_j(+) - F_j(-) > \theta$ **то**
    $Q_1 = Q_1 \cup \{p_j\}$;
  иначе если $F_j(-) - F_j(+) > \theta$ **то**
    определить p'_j: $p'_j(x) = -p_j(x)$, $\forall x \in \mathbb{X}$;
    $Q_1 = Q_1 \cup \{p'_j\}$;
применить сортирующий критерий:
$Q_2 = \{p_{j_1} \dots p_{j_k}\}$,
где $g(p_{j_1}) \geqslant g(p_{j_2}) \geqslant \dots \geqslant g(p_{j_{|Q_1|}})$;
провести монотонизацию выюорки $\mathbb{X}$ по множеству Q_2 с параметром α;
получить монотонную выборку X;
отбор эталонных объектов:
для всех $x_i \in X$
  если $\exists x_t \in X:\ y_i = y_t$ и $x_i \in M_t$ **то**

15: $X = X \setminus \{x_i\}$;

3.6 Композиция монотонных классификаторов

В работе предлагается строить композицию моделей, полученных на предыдущих шагах:

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

где корректирующая операция F является взвешенным голосованием моделей

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

Мы будем рассматривать веса, все равные единице, что соответствует простому голосованию алгоритмов, а также веса, равные последовательным членам геометрической прогрессии со знаменателем q, то есть $w_k = q^{k-1}$.

Из следующего утверждения следует, что построенный классификатор также лежит в классе монотонных функций.

Утверждение 3.1. *Если функция $b_1(u)$ монотонна на множестве признаков Q_1, а функция $b_2(u)$ монотонна на множестве признаков Q_2, то их линейная комбинация $w_1 b_1(u) + w_2 b_2(u)$ с неотрицательными весами $w_1, w_2 > 0$ монотонна на множестве признаков $Q_1 \cup Q_2$.*

Доказательство.

$$u \preceq v \text{ на } Q_1 \cup Q_2 \Rightarrow u \preceq v \text{ на } Q_1 \Rightarrow b_1(u) \leqslant b_1(v)$$

Проделывая то же для b_2, получаем, что b_1 и b_2 монотонны на множестве признаков $Q_1 \cup Q_2$, а значит и их взвешенная сумма с неотрицательными весами монотонна на этом множестве. ■

Классификаторы, над которыми производится операция голосования, в данной работе строятся с использованием разных сортирующих критериев. Классификаторы обучаются последовательно. Объекты, оказавшиеся эталонными для одного классификатора, исключаются из обучающей выборки при обучении следующих классификаторов. Ниже приведен псевдокод для построения композиции алгоритмов.

Голосование монотонных классификаторов

Вход: $\mathbb{X}$ – обучающая выборка;

Q – множество признаков;

T – количество классификаторов в композиции;

$\{k_1, \ldots, k_T\}$ – количество признаков, отбираемых для каждого классификатора;

$\{g_1, \ldots, g_T\}$ – сортирующие критерии для каждого классификатора;

Выход: A – результирующая композиция классификаторов;

1: $X = \mathbb{X}$

2: Последовательное обучение классификаторов:

3: **для всех** $t \in \{1 \ldots T\}$

4: обучить классификатор A_t на выборке X на множестве признаков Q с отбором k_t признаков по критерию g_t;

5: получить множество эталонных объектов классификатора A_t: X_t;

6: исключить из выборки объекты X_t:
$X = X \setminus X_t$;

7: задать веса классификаторов w_i;

8: построить результирующий классификатор $A = \sum_{t=1}^{T} w_t A_t$

---

4 Структура реальных данных

Эксперимент проводится на данных электрокардиограмм пациентов и соответствующих им диагнозам. Технология информационного анализа ЭКГ-сигналов развита в работе [1]. Исходный ЭКГ-сигнал преобразовывается в кодограмму в алфавите из 6 символов, после чего для каждой возможной комбинации из 3 букв (3-граммы) вычисляется число вхождений комбинации в кодограмму. Всего в таком алфавите возможно составить $6^3 = 216$ триграмм. Таким образом, в нашей работе исходное пространство признаков имеет размерность $t = 216$.

В выборке накоплены данные по пациентам с различными заболеваниями. Также в выборке присутствуют данные по здоровым людям (не являющимся больными ни одним из рассматриваемых заболеваний). Далее в тексте используются следующие аббревиатуры: ВДЭ (вегетососудистая дистония), ГБЭ (гипертоническая болезнь), ЖКЭ (желчнокаменная болезнь), ИБЭ (ишемическая болезнь сердца), МКЭ (мочекаменная болезнь), ММЭ (миома матки), СДЭ (сахарный диабет), УЩЭ (узловой зоб щитовидной железы), ХГЭ (хронический гастрит гипоацидный), ХХЭ (холеци-

стит хронический), ЭА (анемия железодефицитная), ЭАП (аденома простаты), ЭАХ (аднексит хронический), ЯБЭ (язвенная болезнь).

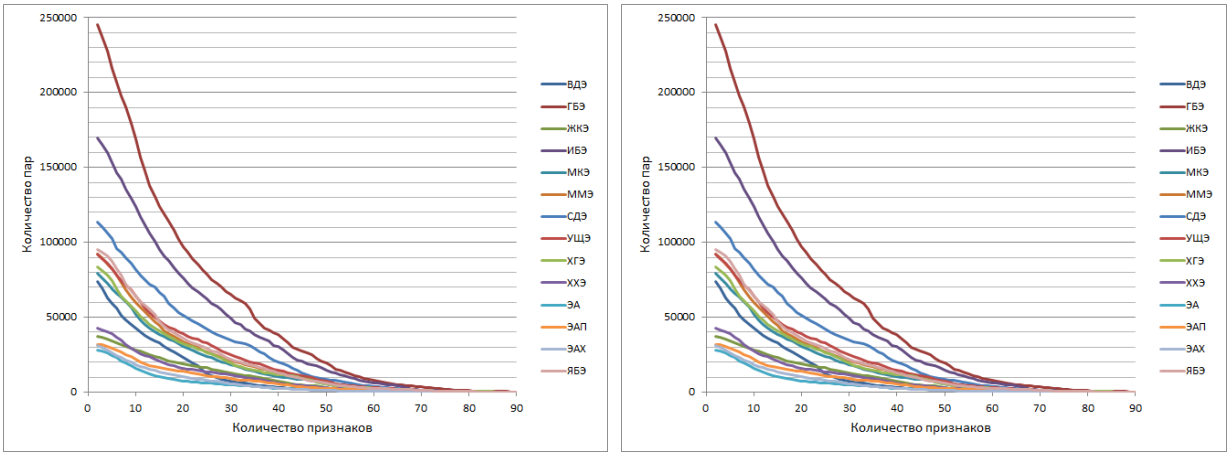
Разработанный алгоритм тестируется на каждой болезни, представленной в выборке. Таким образом, рассматривается задача бинарной классификации, где в качестве класса с меткой -1 выступают здоровые люди, а метку $+1$ имеют больные.

5 Эксперимент

Вычислительный эксперимент реализован с использованием средств языка программирования C++. Программная реализация доступна по ссылке [12]. Числа, приводимые в таблицах для указания качества классификации, являются средними значениями функционала AUC на контрольной выборке при 40-кратном проведении скользящего контроля с разбиением выборки на 10 блоков.

5.1 Монотонные и дефектные пары

На рис. 1 построена зависимость количества монотонных и дефектных пар от количества признаков, отобранных по критерию $g_{diff(B)}$ для всех рассматриваемых болезней. Количество монотонных пар на малом числе признаков на порядок превышает количество дефектных пар. Полученные графики зависимостей свидетельствуют о выполнении предположения о хорошей монотонности выборки, а также показывают быстрое убывание монотонных и дефектных пар с ростом размерности пространства.



(a) Монотонные пары

(b) Дефектные пары

Рис. 1: Количество монотонных и дефектных пар от количества признаков

5.2 Сравнение сортирующих критериев

Рассмотрим модель, содержащую блоки предварительного отбора признаков, отбора признаков по сортирующему критерию, монотонизации и классификации с помощью монотонного классификатора ближайшего соседа. Фиксируем параметр $\alpha = 0.5$. Для всех болезней, представленных в выборке, рассмотрим сортирующие критерии, описанные в разделе "Решение".

Сравнение критериев при отборе $k = 5$ признаков приведено в таблице 1. Максимальное значение функционала качества в каждой строке выделено. Среди предложенных сортирующих критериев, лучшим на большинстве болезней является критерий $g_{|dif(B)|}$. На трех болезнях максимальное значение функционала качества соответствует использованию критерия $g_{dif(B)}$, на двух – критерия $g_{|dif(F)|}$ и на одной болезни – критерия g_{NB}.

Аналогично, в таблицах 2, 3 приведено сравнение критериев при отборе $k = 20$ и $k = 50$ признаков. Лучшими по значению функционала качества для $k = 20$ признаков являются $g_{dif(F)}$ и $g_{dif(B)}$, а также критерий $g_{|dif(B)|}$. Для $k = 50$ большинство болезней имеют максимальное значение AUC после применения критериев $g_{dif(F)}$ и $g_{dif(B)}$.

Таким образом, на разных размерностях лучшее значение функционала качества достигается при применении разных сортирующих критериев.

На рис. 2 по всем болезням показаны графики зависимостей AUC от количества признаков.

5.3 Классификатор ближайшего соседа с M-функцией расстояния

На рис. 3 по всем болезням показаны графики зависимости функционала качества для классификатора ближайшего соседа с M-функцией расстояния, определенной в (2.8).

Несмотря на то, что предложенный классификатор мало изучен и неустойчив к выбросам, качество классификации на небольших размерностях пространств оказывается сравнимо со значениями предыдущего раздела.

Таблица 1: Сравнение сортирующих критериев для $k = 5$ признаков

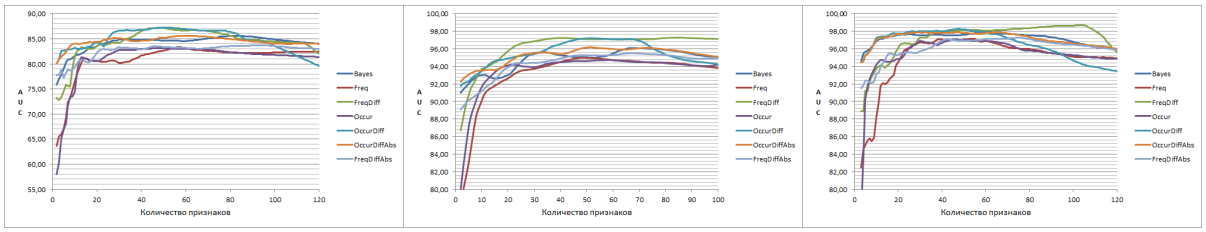
	g_{NB}	g_{F+}	$g_{dif(F)}$	g_{B+}	$g_{dif(B)}$	$g_{\|dif(B)\|}$	$g_{\|dif(F)\|}$
ВДЭ	78,96	67,05	74,49	66,88	**82,63**	81,86	77,19
ГБЭ	92,00	82,93	90,68	86,98	92,30	**93,05**	90,10
ЖКЭ	**95,68**	85,01	91,25	90,14	95,23	95,24	92,39
ИБЭ	93,90	66,92	88,38	77,42	93,55	**95,27**	90,44
МКЭ	89,38	82,83	89,60	81,10	90,57	**90,63**	88,22
ММЭ	87,00	80,55	86,75	81,90	87,00	**88,06**	86,17
СДЭ	93,02	88,76	92,79	91,58	**94,05**	93,71	92,00
УЩЭ	91,06	80,85	90,39	86,09	90,95	**92,36**	90,35
ХГЭ	91,08	83,34	88,24	80,15	**91,48**	91,35	89,07
ХХЭ	92,01	78,04	90,18	86,14	91,54	**92,07**	90,48
ЭА	85,50	75,21	84,72	69,70	85,00	82,08	**87,14**
ЭАП	93,65	84,45	92,10	87,41	93,52	**94,64**	92,26
ЭАХ	85,32	76,36	84,55	76,77	87,86	85,71	**88,42**
ЯБЭ	89,72	80,89	89,61	84,75	90,90	**91,19**	87,01

Таблица 2: Сравнение сортирующих критериев для $k = 20$ признаков

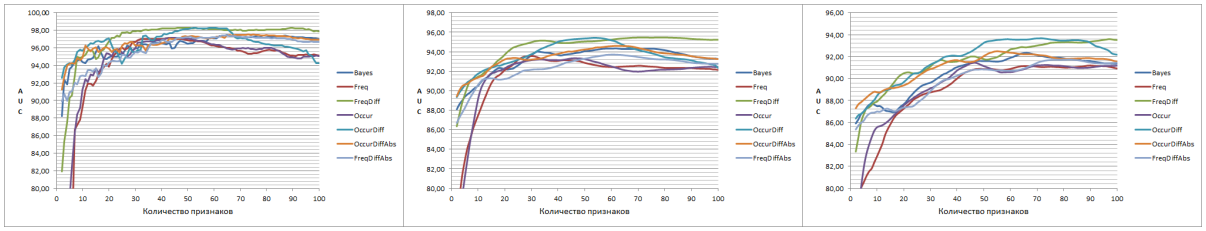
	g_{NB}	g_{F+}	$g_{dif(F)}$	g_{B+}	$g_{dif(B)}$	$g_{\|dif(B)\|}$	$g_{\|dif(F)\|}$
ВДЭ	83,63	80,62	83,56	80,55	84,34	**84,43**	82,29
ГБЭ	93,02	92,63	**95,74**	94,11	94,93	94,44	93,89
ЖКЭ	97,58	94,43	96,05	94,84	**97,73**	97,70	95,35
ИБЭ	96,24	94,54	96,42	95,70	96,07	**96,61**	95,71
МКЭ	92,21	92,11	**93,71**	92,29	92,76	93,21	91,22
ММЭ	87,66	87,26	**90,49**	87,56	89,68	89,38	87,36
СДЭ	94,20	92,44	94,93	93,12	**95,69**	94,96	93,89
УЩЭ	93,71	90,09	**94,12**	91,23	93,38	93,23	93,40
ХГЭ	92,53	91,81	**94,05**	90,40	93,21	92,42	91,48
ХХЭ	92,46	89,48	94,13	91,14	**94,61**	93,20	92,93
ЭА	88,24	85,10	**89,50**	85,56	88,78	87,65	86,11
ЭАП	94,75	92,26	95,09	91,78	**95,40**	95,39	94,04
ЭАХ	90,55	88,04	**91,19**	88,54	91,03	91,18	87,16
ЯБЭ	91,70	88,31	92,06	89,21	**92,62**	92,07	89,06

Таблица 3: Сравнение сортирующих критериев для $k = 50$ признаков

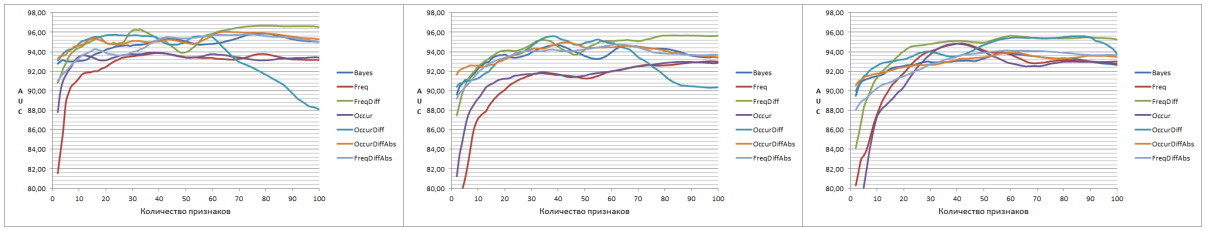
	g_{NB}	g_{F+}	$g_{dif(F)}$	g_{B+}	$g_{dif(B)}$	$g_{\|dif(B)\|}$	$g_{\|dif(F)\|}$
ВДЭ	84,63	82,61	**87,19**	83,15	87,18	85,25	83,38
ГБЭ	95,05	94,96	97,10	94,61	**97,20**	96,16	95,28
ЖКЭ	97,61	97,06	98,10	97,00	**98,22**	97,83	96,92
ИБЭ	96,39	95,74	**97,87**	95,67	97,36	96,55	96,87
МКЭ	94,06	92,86	95,04	93,25	**95,37**	94,23	93,29
ММЭ	91,60	90,86	91,78	91,14	**93,26**	92,06	90,84
СДЭ	95,06	93,44	93,95	93,40	95,06	94,93	**95,38**
УЩЭ	93,56	91,26	94,20	91,54	**94,93**	94,43	94,24
ХГЭ	93,31	94,14	**94,95**	94,02	94,70	93,44	93,90
ХХЭ	94,22	91,83	**95,57**	91,62	95,17	94,33	93,12
ЭА	87,00	85,97	**89,22**	85,85	87,87	87,45	87,79
ЭАП	95,10	93,83	95,31	94,14	**95,40**	95,11	94,38
ЭАХ	89,09	87,17	**90,73**	86,85	89,31	89,22	88,31
ЯБЭ	92,75	91,18	94,19	91,63	**94,54**	92,09	90,95



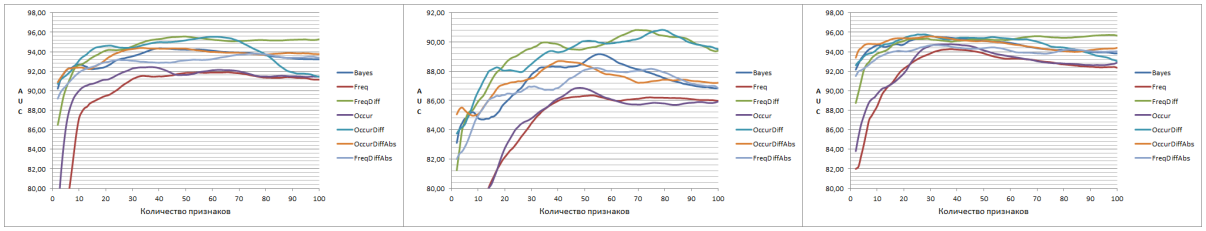
(a) ВДЭ (b) ГБЭ (c) ЖКЭ



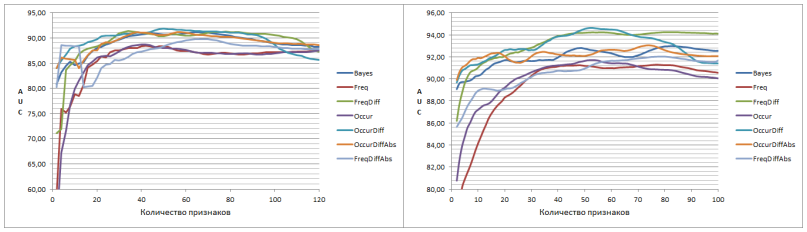
(d) ИБЭ (e) МКЭ (f) ММЭ



(g) СДЭ (h) УЩЭ (i) ХГЭ

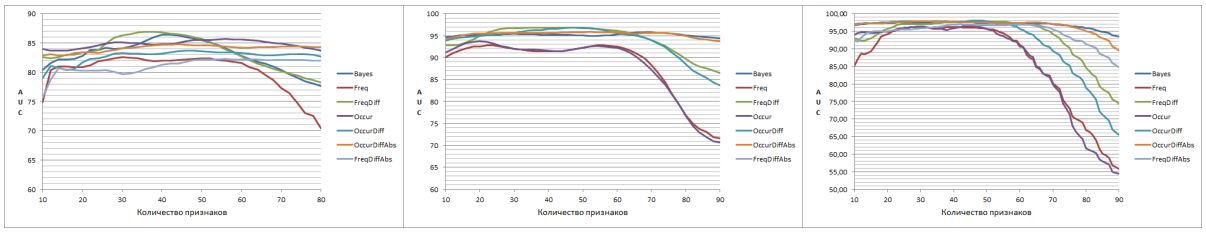


(j) ХХЭ (k) ЭА (l) ЭАП

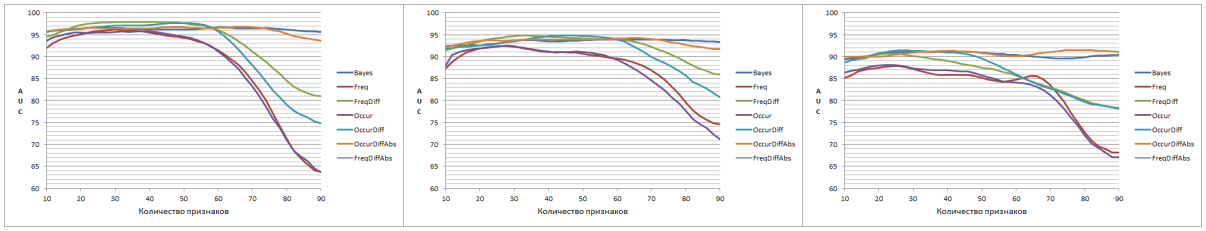


(m) ЭАХ (n) ЯБЭ

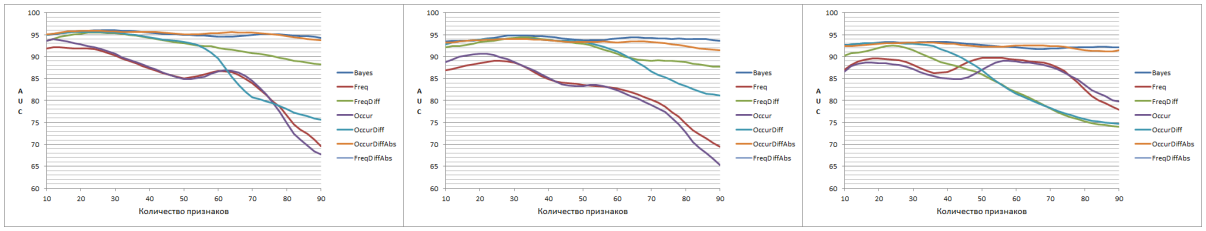
Рис. 2: Зависимость AUC от количества признаков



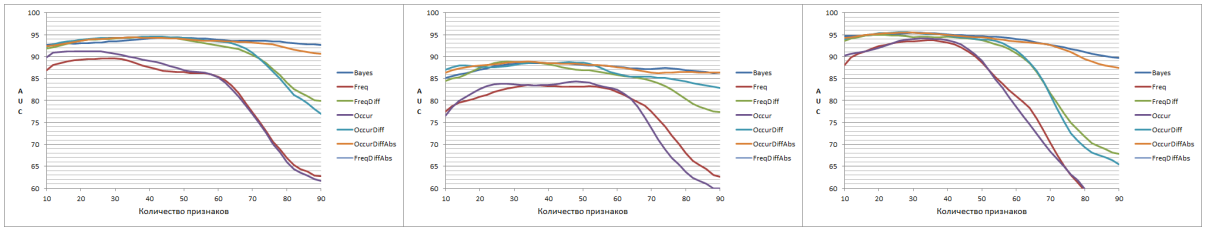
(a) ВДЭ (b) ГБЭ (c) ЖКЭ



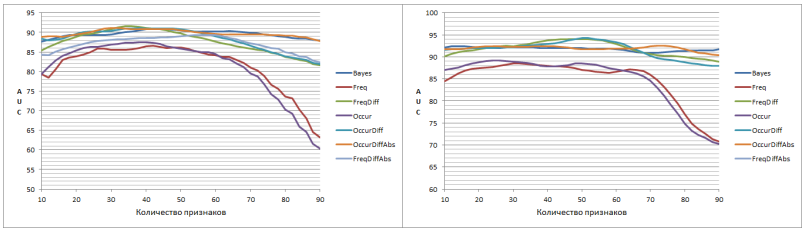
(d) ИБЭ (e) МКЭ (f) ММЭ



(g) СДЭ (h) УЩЭ (i) ХГЭ



(j) ХХЭ (k) ЭА (l) ЭАП



(m) ЭАХ (n) ЯБЭ

Рис. 3: Зависимость AUC от количества признаков для классификатора ближайшего соседа с M-функцией расстояния

5.4 α-монотонизация

В этом разделе фиксируем все уровни, кроме уровня монотонизации. Отбирается $k = 8$ признаков по критерию $g_{dif(B)}$. В таблице 4 указаны значения чувствительности и специфичности, получаемые на контрольной выборке при каждом уровне α. Увеличение параметра α приводит к увеличению специфичности и уменьшению чувствительности, то означает сдвиг разделяющей поверхности в сторону класса $y = +1$ больных.

5.5 Сравнение результатов

В таблице 5 сравниваются результаты лучших моделей монотонного классификатора для всех болезней с результатами, полученными в работе [2].

6 Заключение

В данной работе были исследованы различные способы отбора объектов и признаков для монотонного классификатора ближайшего соседа, предложены жадные методы решения NP-трудных задач. Была выполнена программная реализация и проведены численные эксперименты, показывающие применимость исследуемых моделей к задаче медицинской диагностики. Также, на практике был изучен алгоритм ближайшего соседа с M-функцией расстояния, описанный в [10]. Полученные результаты оказались сравнимыми с результатами, которые были получены ранее в работе [2] на той же выборке болезней.

Список литературы

[1] В. М. Успенский. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов. *Экономика и информация*, 1:116, 2008.

[2] В. Р. Целых В. М. Успенский, К. В. Воронцов. Статистическое обоснование информационного анализа электрокардиосигналов для диагностики заболеваний внутренних органов. *Математическая биология и биоинформатика*, 2014.

Таблица 4: Изменение параметра монотонизации α

		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
ХГЭ	чувствительность	94,16	89,27	85,18	82,74	80,29
	специфичность	62,21	74,19	80,65	83,54	85,86
ХХЭ	чувствительность	93,16	87,47	83,25	80,62	77,50
	специфичность	66,17	79,17	84,19	87,08	89,58
ЭА	чувствительность	93,95	83,79	77,80	72,77	68,51
	специфичность	37,53	72,37	79,86	83,83	87,30
ЭАП	чувствительность	93,00	89,45	86,74	83,84	80,76
	специфичность	72,85	81,67	86,21	88,93	90,83
ЭАХ	чувствительность	94,40	88,24	82,94	78,79	75,24
	специфичность	51,15	70,36	80,76	85,22	88,76

Таблица 5: 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

[3] А. В. Зухба. Вычислительная сложность отбора объектов и признаков для задач классификации с ограничениями монотонности. 2014.

[4] K. V. Rudakov and K. V. Vorontsov. Methods of optimization and monotone correction in the algebraic approach to the recognition problem. *Doklady Mathematics*, 60:139, 1999.

[5] К. В. Воронцов. Оптимизационные методы линейной и монотонной коррекции в алгебраическом подходе к проблеме распознавания. *Ж. вычисл. матем. и матем. физ.*, 40:166–176, 2000.

[6] Ю. И. Журавлев. Об алгебраическом подходе к решению задач распознавания или классификации. *Проблемы кибернетики*, 33:5–68, 1978.

[7] Nikita Spirin and Konstantin Vorontsov. Learning to rank with nonlinear monotonic ensemble. In Carlo Sansone, Josef Kittler, and Fabio Roli, editors, *Multiple Classifier Systems*, volume 6713 of *Lecture Notes in Computer Science*, pages 16–25. Springer Berlin Heidelberg, 2011.

[8] Г. А. Воронцов, К. В. Махина. Принцип максимизации зазора для монотонного классификатора ближайшего соседа. volume 0.25, pages 117–121. Доклады Всероссийской конференции «Математические методы распознавания образов» (ММПО-15), 2011.

[9] A. Feelders. Monotone relabeling in ordinal classification. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 803–808, Dec 2010.

[10] Г. А. Махина. Построение монотонного классификатора по методу ближайшего соседа. 2014.

[11] К. В. Воронцов. Монотонная непрерывная интерполяция. 2012.

[12] М. Ю. Швец. Реализация монотонных классификаторов для задач медицинской диагностики, 2015. Доступна по ссылке https://svn.code.sf.net/p/mlalgorithms/code/Group174/Shvets2015MonotonicClassification/code/.