УДК:519.7:004.852

Вычислительная сложность отбора объектов и признаков для задач классификации с ограничениями монотонности

©2015 Зухба А.В.¹

Московский физико-технический институт (государственный университет)
Россия, 141700, Московская область, г. Долгопрудный, Институтский пер., 9

Аннотация. В медицинской биологии часто используются решающие правила с ограничениями монотонности, разработанные в информатике. Одновременный отбор информативных признаков и эталонных объектов может существенно влиять на степень монотонности выборки и, как следствие, на качество классификации. В данной работе предлагается систематизация дискретных оптимизационных задач, возникающих при одновременном отборе информативных признаков и эталонных объектов, устанавливается вычислительная сложность этих задач.

Ключевые слова: машинное обучение, отбор признаков, отбор эталонов, монотонный классификатор, дискретная оптимизация, вычислительная сложность.

ВВЕДЕНИЕ

Во многих задачах классификации в области медицинской биологии используются векторные признаковые описания объектов и предполагается наличие монотонных зависимостей между значениями признаков и вероятностями классов. Например, «чем ниже уровень магния в крови, тем выше риск пролапса митрального клапана» [5], или «чем выше частота мотива A----GKS, тем выше вероятность, что белок обладает свойством связывания АТФ» [26]. Важность явного учёта априорных предположений о монотонном характере зависимости неоднократно отмечалась в литературе [24, 23, 3, 19].

На практике в таких случаях часто пользуются линейными моделями классификации с неотрицательными коэффициентами. Преимущество линейной модели — в её простоте и наличии готовых реализаций. Однако для многих задач линейная модель представляется слишком жёсткой. Множество монотонных функций существенно шире, чем множество линейных функций с неотрицательными коэффициентами. Поэтому нелинейные монотонные модели классификации имеют преимущества в задачах со сложной разделяющей поверхностью, что было показано на примере задач медицинской диагностики [21, 22], ранжирования поисковой выдачи [25], категоризации текстов [9], при построении композиций классификаторов [6, 7].

Известно много методов построения монотонных классификаторов [15, 22, 18, 19]. Для монотонного метода ближайшего соседа [2] имеются высокоточные комбинаторные оценки полного скользящего контроля [4, 11], из которых следует, что данный метод

¹ a__l@mail.ru

обладает высокой обобщающей способностью благодаря максимизации ширины зазора между классами. Данный факт позволяет провести аналогии между монотонными классификаторами и методом опорных векторов SVM [16], который является одним из лучших методов классификации именно благодаря принципу максимизации зазора.

Большинство методов монотонной классификации требуют предварительной монотонизации выборки, которая выполняется с помощью жадных эвристических алгоритмов [1] или методов изотонной регрессии [14]. Монотонизация сводится либо к отбрасыванию объектов обучающей выборки, нарушающих условие монотонности, либо к исправлению меток классов этих объектов. Кроме того, в процессе построения монотонного классификатора в каждом классе определяется подмножество недоминируемых объектов (эталонов, опорных объектов), которые и определяют границу зазора между классами. Данный факт также позволяет говорить об аналогии между монотонными классификаторами и методом опорных векторов SVM.

Многие задачи классификации, возникающие в медицинской диагностике и биоинформатике, являются задачами с большим числом классов. Для каждого класса требуется найти своё информативное подпространство признаков, связанных монотонной зависимостью с вероятностью данного класса. Например, найти все признаки, образующие синдром данного заболевания, или найти все мотивы, связанные с данной функцией белка, и т. д.

В информационном анализе электрокардиосигналов [13] решается задача диагностики заболеваний внутренних органов человека по электрокардиограмме, закодированной в последовательность символов [12] или векторов [27], каждый элемент которой соответствует одному кардиоциклу. В качестве признаков используются частоты встречаемости коротких кодовых фрагментов — локальных паттернов variability сердечного ритма. При этом оказывается, что для каждого заболевания характерен свой набор паттернов, и чем выше частота паттерна в ЭКГ обследуемого, тем выше вероятность данного заболевания. Естественным способом формализации данного наблюдения является предположение о монотонности классификатора в искомом подпространстве признаков, специфичных для данного заболевания.

Проблема отбора признаков является центральной для таких задач, причём решать её приходится многократно. Тем не менее, вопросы вычислительной сложности отбора признаков в монотонных классификаторах относительно мало исследованы. Известно, что задача выбора минимального подмножества признаков, позволяющего приблизить целевую функцию с заданной точностью, является NP-полной [20]. Отсюда следует, что более общая задача отбора признаков также является NP-полной. Остаётся открытым вопрос, не позволяет ли ограничение монотонности быстрее находить подпространство информативных признаков, в котором задача решается наиболее точно.

В данной работе исследуется связь задачи монотонизации выборки с отбором эталонных объектов и отбором признаков. Предлагается систематизация возникающих при этом задач дискретной оптимизации, устанавливается статус их вычислительной сложности, предлагаются способы решения.

ОПРЕДЕЛЕНИЯ И ОБОЗНАЧЕНИЯ

Дано конечное множество объектов $\mathbb{X} = \{x_1, \dots, x_\ell\}$, называемое обучающей выборкой, и конечное множество признаков $\mathbb{F} = \{f_1, \dots, f_t\}$ — отображений вида $f_j: \mathbb{X} \rightarrow E_j$, где E_j — линейно упорядоченное множество. Каждый объект x_i относится к одному из двух классов $y_i = y^*(x_i) \in \{0, 1\}$. Обозначим множество объектов обучающей выборки класса 1 через $\mathbb{A}$, а класса 0 — через $\mathbb{B}$.

Любое непустое подмножество множества признаков $F \subseteq \mathbb{F}$ индуцирует отношение частичного порядка на $\mathbb{X}$: $x \leq x'$ тогда и только тогда, когда $f(x) \leq f(x')$ для всех $f \in F$.

Определение 1. Пара объектов $(a, b) \in \mathbb{A} \times \mathbb{B}$ называется монотонной, если $a > b$. Множество всех монотонных пар обозначается через M .

Определение 2. Пара объектов $(a, b) \in \mathbb{A} \times \mathbb{B}$ называется дефектной, если $a < b$. Множество всех дефектных пар обозначается через D .

Множество пар, монотонных по признаку f , будем обозначать через M_f , монотонных по совокупности признаков F — через M_F . Множество пар, дефектных по признаку f , будем обозначать через D_f , дефектных по совокупности признаков F — через D_F .

Утверждение 1. Для произвольного подмножества признаков $F \subseteq \mathbb{F}$

$$M_F = \bigcap_{f \in F} M_f, \quad D_F = \bigcap_{f \in F} D_f.$$

Следствие 1. Для любых подмножеств признаков F, G

$$F \subseteq G \Rightarrow |M_G| \leq |M_F|, \quad |D_G| \leq |D_F|. \quad (1)$$

Задача построения монотонного классификатора заключается в аппроксимации неизвестной функции $y^*: \mathbb{X} \rightarrow \{0, 1\}$, заданной в точках обучающей выборки, монотонной функцией $y: \mathbb{X} \rightarrow \{0, 1\}$. По определению монотонной функции для любых двух объектов $x, x' \in \mathbb{X}$ из $x < x'$ следует $y(x) \leq y(x')$. На практике обучающая выборка может не удовлетворять условию монотонности, и в таких случаях ставится задача её предварительной монотонизации.

Задача монотонизации обучающей выборки состоит в том, чтобы выбрать подмножество объектов $X \subseteq \mathbb{X}$ и подмножество признаков $F \subseteq \mathbb{F}$ так, чтобы среди пар объектов (a, b) из $(\mathbb{A} \cap X) \times (\mathbb{B} \cap X)$ оказалось как можно меньше дефектных пар и как можно больше монотонных: $|D_F| \rightarrow \min$, $|M_F| \rightarrow \max$. Согласно (1), первое условие можно заменить косвенным требованием $|F| \rightarrow \max$ или $|X| \rightarrow \min$, а второе — косвенным требованием $|F| \rightarrow \min$ или $|X| \rightarrow \max$.

Заметим, что минимизация $|X|$ при сохранении качества классификации соответствует задаче отбора эталонных объектов (prototype selection, PS) [8, 29, 17], а максимизация $|X|$ — задаче отсева выбросов (outlier detection) [28]. Требование максимизации числа признаков $|F|$ возникает, когда априори известно, что все признаки информативны, а требование минимизации $|F|$ соответствует задаче отбора признаков (feature selection, FS), когда требуется отбросить шумовые признаки и уменьшить переобучение [8]. Таким образом, дополнительные условия оптимальности могут частично противоречить друг другу. Для смягчения этих противоречий требования максимизации или минимизации могут заменяться ограничениями-неравенствами.

Систематизация возникающих при этом постановок задач показана на рис. 1. Стрелками соединены те задачи, которые получаются друг из друга изменением одного из условий оптимальности. Задачи отбора объектов (PS) и отбора признаков (FS) в данной работе рассматриваются отдельно.

Для выяснения статуса вычислительной сложности этих задач и доказательства теорем о сводимости нам понадобятся следующие известные задачи дискретной оптимизации [10].

Задача о покрытии множества подмножествами. *Покрытием* конечного множества U семейством его подмножеств S называется такое множество подмножеств $C = \{C_1, \dots, C_k\} \subseteq S$, что $\bigcup_{i=1}^k C_i = U$. *Размером* покрытия C называют число $k = |C|$.

Рис. 1. Постановки задачи монотонизации: FS — отбор признаков, PS — отбор объектов.

Задача 1. Найти покрытие $C \subseteq S$ множества U минимального размера k .

Данная задача является NP-полной.

Обычно предполагают, что покрытие существует, то есть $\bigcup_{C_i \in S} C_i = U$. Задача имеет еще одну формулировку, также являющуюся NP-полной: существует ли покрытие $C \subseteq S$ множества U размера $k \leq k_0$, где k_0 — заданное число.

Задача о рюкзаке. Дан набор из n предметов. Каждый предмет имеет два параметра: вес $z_i \in \mathbb{N}$ и ценность $c_i \in \mathbb{N}$, $i = 1, \dots, n$. Имеется рюкзак вместимости Z , $Z \in \mathbb{N}$. Положим x_i принимает два значения: $x_i = 1$, если груз упакован, и $x_i = 0$ в противном случае, для всех $i = 1, \dots, n$.

Задача 2. Максимизировать $\sum_{i=1}^n c_i x_i$ при наличии ограничения $\sum_{i=1}^n z_i x_i \leq Z$ на вместимость рюкзака.

Данная задача является NP-полной.

Задача о биклике. Полным двудольным графом или бикликой называется двудольный граф, у которого любая вершина первой доли соединена со всеми вершинами второй доли вершин.

Задача 3. Дан двудольный граф. Найти полный двудольный подграф $K_{i,i}$ с максимальным числом рёбер i^2 .

Данная задача является NP-полной.

Задача о вершинном покрытии в двудольном графе. Вершинное покрытие неориентированного графа $G = (V, E)$ — это множество его вершин S , такое, что у каждого ребра графа хотя бы один из концов принадлежит S . Паросочетанием называют множество попарно несмежных ребер, то есть рёбер, не имеющих общих вершин. Максимальное паросочетание — это такое паросочетание в графе, которое не содержится ни в каком другом паросочетании этого графа.

Задача 4. Дан двудольный граф. Найти его вершинное покрытие.

Точное решение данной задачи вычисляется за полиномиальное время.

Теорема 1. (Кёниг) Мощность максимального паросочетания в двудольном графе равна мощности его минимального вершинного покрытия.

Доказательство этой теоремы конструктивно: строится алгоритм, находящий вершинное покрытие двудольного графа за время $O(|E|\sqrt{|L|+|R|})$, где L, R — правая и левая доли соответственно.

ЗАДАЧИ ОТБОРА ПРИЗНАКОВ

Введем следующее представление множества объектов обучающей выборки $\mathbb{X}$: каждому признаку f поставим в соответствие матрицу T^f , строки которой соответствуют элементам множества $\mathbb{A}$, столбцы — элементам $\mathbb{B}$. Клетки матрицы соответствуют парам объектов из разных классов. В клетке запишем 1, если объект класса $\mathbb{A}$, соответствующий строке, больше объекта класса $\mathbb{B}$, соответствующего столбцу, и 0 в противном случае, то есть $T_{ab}^f = [f(a) > f(b)]$ для всех $(a, b) \in \mathbb{A} \times \mathbb{B}$.

Частичному порядку, индуцированному подмножеством признаков $F \subseteq \mathbb{F}$, соответствует матрица T^F , в каждой клетке которой записывается число признаков из F , по которым объект a , соответствующий строке, больше объекта b , соответствующего столбцу: для всех $(a, b) \in \mathbb{A} \times \mathbb{B}$

$$T_{ab}^F = \sum_{f \in F} [f(a) > f(b)].$$

Пример 1. Двумерная выборка и соответствующая ей матрица $T^{\{f_1, f_2\}}$ показаны на рис. 2.

Рис. 2. Пример выборки и матрицы: a_i — черные, b_j — белые.

Для дефектных пар в соответствующей клетке матрицы находится 0; для монотонных пар — число признаков $|F|$, то есть

$$\begin{aligned} (a, b) \in D_F &\Leftrightarrow T_{ab}^F = 0; \\ (a, b) \in M_F &\Leftrightarrow T_{ab}^F = |F|. \end{aligned}$$

Задача 5. Выбрать признаки так, чтобы при минимальном количестве дефектных пар получить максимальное возможное при этом количество монотонных пар: $FS(|D| \rightarrow \min: |M| \rightarrow \max)$.

Теорема 2. Решение задачи о минимальном покрытии множества подмножествами сводится к решению задачи 5 на некоторой искусственной обучающей выборке $\mathbb{X}$.

Доказательство. Пусть U — конечное множество, $|U| = n$, $S = \{C_1, \dots, C_{|S|}\}$ — семейство его подмножеств.

Каждому элементу $u_i \in U$ поставим в соответствие объект $a_i \in \mathbb{A}$. Каждому подмножеству C_t поставим в соответствие объект $b_{t+1} \in \mathbb{B}$ и признак f_{t+1} . Добавим два вспомогательных объекта a^M и a^* класса $\mathbb{A}$, один вспомогательный объект b_1 класса $\mathbb{B}$ и один вспомогательный признак f_1 .

Первый признак f_1 задает следующий порядок на множестве объектов:

$$f_1(a^M) > f_1(b_1) > f_1(a_1) > \dots > f_1(a_n) > f_1(a^*) > f_1(b_2) > \dots > f_1(b_{|S|+1}).$$

Матрица T^{f_1} будет выглядеть следующим образом: столбцы упорядочены по номерам элементов класса $\mathbb{B}$, строки — по номерам элементов класса $\mathbb{A}$, две нижние строки соответствуют объектам a^* и a^M . Последняя строка состоит из единиц, первый столбец, кроме последнего элемента — нули, остальные — единицы, рис. 3.

	b_1	b_2		$b_{ S +1}$
a_1	0	1	1	1
a_2	0	1	1	1
a_n	0	1	1	1
a^*	0	1	1	1
a^M	1	1	1	1

Рис. 3. Матрица для первого признака.

Остальные признаки строятся следующим образом: признак с номером $t + 1$ соответствует подмножеству $C_t = \{u_{t1}, \dots, u_{t|C_t|}\}$ из семейства S . Он устанавливает следующий порядок на множестве объектов обучающей выборки:

$$\begin{aligned} & \left\{ f_{t+1}(a_r) \mid a_r \in \{a_{t1}, \dots, a_{t|C_t|}\} \right\} > f_{t+1}(b_1) > \\ & > \left\{ f_{t+1}(a_r) \mid a_r \in \{a_1, \dots, a_n\} \setminus \{a_{t1}, \dots, a_{t|C_t|}\} \right\} > f_{t+1}(b_{t+1}) > \\ & > f_{t+1}(a^*) > \left\{ f_{t+1}(b_r) \mid b_r \in \mathbb{B} \setminus \{b_1, b_{t+1}\} \right\} > f_{t+1}(a^M). \quad (2) \end{aligned}$$

В матрице $T^{f_{t+1}}$ последняя строка, соответствующая a^M , состоит из нулей; в строке, соответствующей a^* , нули стоят в 1-м и в $(t + 1)$ -м столбце, остальные элементы — единицы; в первом столбце единицы стоят только в строках $\{a_{t1}, \dots, a_{t|C_t|}\}$, остальные элементы — нули; остальные элементы матрицы — единицы, рис. 4.

	b_1	b_2		b_{t+1}		$b_{ S +1}$
a_1	0	1	1	1	1	1
a_2	0	1	1	1	1	1
a_{ti}	1	1	1	1	1	1
a_n	0	1	1	1	1	1
a^*	0	1	1	1	0	1
a^M	0	0	0	0	0	0

Рис. 4. Матрица для признака $t + 1$.

Итак, построены $|S| + 1$ признака. Первый признак — вспомогательный, каждый признак с номером $t + 1$ ($t > 0$) взаимно однозначно соответствует подмножеству $C_t \subset S$. Соответствие устанавливается по первому столбцу матрицы: единицы стоят в тех и только в тех строках, которые соответствуют элементам C_t .

Далее для доказательства теоремы нам понадобится несколько вспомогательных утверждений.

Утверждение 2. *Оптимальное множество признаков F содержит первый признак f_1 .*

Доказательство. Предположим, было выбрано оптимальное подмножество признаков F . Пусть первый признак ему не принадлежит. Рассмотрим матрицу T^F . Она равна поэлементной сумме матриц всех признаков. При этом нули соответствуют дефектным парам объектов, а элементы со значением $|F|$ — монотонным парам. Поскольку первый признак не принадлежит F , последняя строка T^F будет состоять из нулей. Добавим к множеству F первый признак $F' = F \cup \{f_1\}$. Рассмотрим матрицу $T^{F'}$. Матрица $T^{F'}$ равна сумме матрицы T^F и матрицы T^{f_1} первого признака. Следовательно, все элементы $T^{F'}$ не меньше, чем в матрице T^F . Все элементы T^F , равные $|F|$, в $T^{F'}$ равны $|F'|$, а в последней строке стоят не нули, а единицы. Таким образом, нулей в матрице $T^{F'}$ строго меньше, чем в T^F , то есть дефектных пар по подмножеству признаков F' строго меньше, чем по подмножеству F , а монотонных пар не меньше, что противоречит оптимальности F . Таким образом, признак f_1 обязательно принадлежит оптимальному подмножеству признаков. Утверждение 2 доказано.

Утверждение 3. *Матрица T^F оптимального подмножества признаков F содержит ровно один нулевой элемент $T_{a^*b_1}^F = 0$.*

Доказательство. Матрица любого подмножества признаков является суммой матриц каждого из признаков. Во всех матрицах признаков в первом столбце в строке a^* находится нуль. Следовательно, для любого подмножества признаков F элемент $T_{a^*b_1}^F$ равен нулю. Рассмотрим матрицу всех признаков $T^{\mathbb{F}}$. Из предположения о существовании покрытия $\bigcup_{C_i \in S} C_i = U$ следует, что для каждого элемента первого столбца $T_{ab_1}^{\mathbb{F}}$, кроме $T_{a^*b_1}^{\mathbb{F}}$, найдется такой признак f_i , что $T_{ab_1}^{f_i} = 1$. По построению, $T_{a^M b_1}^{f_1} = 1$. Таким образом, в матрице $T^{\mathbb{F}}$ ровно один нулевой элемент, $T_{a^*b_1}^{\mathbb{F}} = 0$. Следовательно, по условию оптимальности множества признаков F , в матрице T^F ровно один нулевой элемент $T_{a^*b_1}^F = 0$. Утверждение 3 доказано.

Утверждение 4. *Подмножества C_i , соответствующие признакам оптимального множества F , образуют покрытие множества U .*

Доказательство. По построению, каждому элементу $u_i \in U$ соответствует одна из строк матрицы T^F с соответствующим номером $i = 1, \dots, n$. Согласно утверждению 3, в матрице T^F ровно один нулевой элемент, причем он находится в первом столбце в строке a^* . Следовательно, для каждой строки с номером $i = 1, \dots, n$ существует признак f_{j+1} такой, что $T_{a_i b_1}^{f_{j+1}} = 1$. Тогда в соответствующем множестве C_j содержится элемент u_i . То есть для каждого элемента $u_i \in U$ найдется такое подмножество C_j , соответствующее признаку $f_{j+1} \in F$, что $u_i \in C_j$. Таким образом, подмножества C_i , соответствующие признакам оптимального множества F , образуют покрытие множества U . Утверждение 4 доказано.

Итак, мы доказали, что оптимальное множество признаков F обязательно содержит первый признак, и что множество подмножеств, соответствующих признакам из F , являются покрытием множества U . Теперь рассмотрим вопрос минимальности данного покрытия. Для этого рассмотрим матрицу $T^{f_{t+1}}$ признака $t+1$, соответствующего подмножеству C_t . В этой матрице на пересечении строки a^* и столбца b_{t+1} стоит ноль, $T_{a^* b_{t+1}}^{f_{t+1}} = 0$. Рассмотрим матрицу T^F : первый столбец, кроме элемента $T_{a^* 1}^F = 0$, состоит из элементов, больших 0 и меньших $|F|$, то есть соответствующих парам несравнимых объектов. Элемент $T_{a^* 1}^F = 0$ соответствует единственной дефектной паре. Строка a^M состоит из единиц, значит соответствуют парам несравнимых объектов. В пересечении

всех строк с номерами $i = 1, \dots, n$ и столбцов с номерами $j = 2, \dots, |S|+1$ стоят элементы, равные $|F|$, то есть соответствующие монотонным парам. Рассмотрим единственную оставшуюся строку a^* : первый элемент этой строки равен нулю; остальные элементы этой строки равны $|F|$ или $|F| - 1$, то есть соответствуют или монотонной или несравнимой паре. По построению, элемент $T_{a^*b_{t+1}}^F$ равен $|F| - 1$ тогда и только тогда, когда признак f_{t+1} , соответствующий подмножеству C_t , принадлежит F , то есть тогда и только тогда, когда $f_{t+1} \in F$. Таким образом, в строке a^* ровно один нулевой элемент, соответствующий дефектной паре; ровно $|F| - 1$ элемент, равный $|F| - 1$ и соответствующий несравнимой паре; остальные элементы равны $|F|$ и соответствуют монотонным парам.

Утверждение 5. Подмножества C_{i-1} , соответствующие признакам $f_i \in F \setminus \{f_1\}$ оптимального подмножества признаков F , образуют минимальное покрытие S множества U .

Доказательство. То, что соответствующие подмножества C_i образуют покрытие S множества U , мы доказали в утверждении 4. Покажем теперь, что покрытие S минимально. Пусть F — оптимальное подмножество признаков. Предположим, что соответствующее покрытие U не является минимальным, то есть существует покрытие S' множества U подмножествами $C_i \in S$, содержащее меньше $|F| - 1$ элементов. Построим множество признаков F' следующим образом: пусть оно состоит из первого признака f_1 и всех признаков, соответствующих подмножествам C_i из минимального покрытия.

Поскольку S' является покрытием, в матрице $T^{F'}$ первый столбец кроме строки a^* состоит из элементов, больших 0 и меньших $|F'|$, то есть соответствующих парам несравнимых объектов. Элемент, находящийся на пересечении первого столбца и строки a^* , является нулем и соответствует единственной дефектной паре. Строка a^M состоит из единиц, а единицы больше 0 и меньше $|F'|$, значит, соответствуют парам несравнимых объектов. В пересечении всех строк с номерами $i = 1, \dots, n$ и столбцов с номерами $j = 2, \dots, |S|+1$ стоят элементы, равные $|F'|$, то есть соответствующие монотонным парам. По построению, в строке a^* ровно один нулевой элемент, соответствующий дефектной паре; ровно $|F'| - 1$ элемент, равный $|F'| - 1$ и соответствующий несравнимой паре; остальные элементы равны $|F'|$ и соответствуют монотонным парам. По предположению, $|F'| < |F|$. Тогда множество признаков F' индуцирует на обучающей выборке порядок, при котором дефектная пара одна, а несравнимых пар меньше, чем в порядке, индуцируемом F . То есть дефектных пар столько же, а монотонных пар больше, что противоречит оптимальности F . Следовательно, S является минимальным покрытием множества U подмножествами. Утверждение 5 доказано.

Таким образом мы получили конструктивное доказательство теоремы.

Следствие 2. Задача $FS(|D| \rightarrow \min: |M| \rightarrow \max)$ является NP-трудной.

Задача 6. Выбрать множество признаков F так, чтобы монотонных пар было не менее m , а дефектных не более d : $FS(|D_F| \leq d, |M| \geq m)$.

Теорема 3. Задача $FS(|D_F| \leq d, |M| \geq m)$ является NP-трудной.

Доказательство. Предположим, что для задачи 6 имеется полиномиальный алгоритм. Построим тогда алгоритм, решающий задачу при всех возможных параметрах d и m . Поскольку d и m могут принимать целые значения от 0 до $|A||B|$, потребуется перебрать не более $(|A||B| + 1)^2$ вариантов. То есть получаемый алгоритм также является полиномиальным, но решает задачу $FS(|D| \rightarrow \min: |M| \rightarrow \max)$. Получили противоречие. Значит, задача 6 является NP-трудной. Теорема доказана.

Задача 7. Получить не менее t монотонных пар, выбрав максимальное возможное при этом количество признаков: $FS(|M| \geq t, |F| \rightarrow \max)$.

Теорема 4. Решение задачи о рюкзаке сводится к решению задачи 7 на некоторой искусственной обучающей выборке $\mathbb{X}$.

Доказательство. Пусть дан набор из n предметов. Каждый предмет имеет два параметра: вес $z_i \in \mathbb{N}$ и ценность $c_i \in \mathbb{N}$, $i = 1, \dots, n$, и имеется рюкзак определенной вместимости Z , $Z \in \mathbb{Z}$.

Построим обучающую выборку $\mathbb{X}$, состоящую из n непересекающихся множеств объектов X_i , $i = 1, \dots, n$. Построим множество признаков $\mathbb{F}$, состоящее из n непересекающихся множеств признаков F_i , $i = 1, \dots, n$. Каждому предмету i из рюкзака взаимно однозначно поставим в соответствие множество объектов X_i и множество признаков F_i . Признаки будем строить как линейные порядки на множестве объектов.

Пусть каждое множество объектов X_i состоит из одного объекта a_i класса 1 и z_i объектов $\{b_{i,1}, \dots, b_{i,z_i}\}$ класса 0. Зададим i -е множество признаков F_i как $c_i = |F_i|$ признаков, каждый из которых устанавливает следующий порядок на множестве объектов $\mathbb{X}$:

$$\left\{ f_i(a_r) \mid a_r \in \mathbb{A} \setminus \{a_i\} \right\} > \left\{ f_i(b_r) \mid b_r \in \{b_{i,1}, \dots, b_{i,z_i}\} \right\} > f_i(a_i) > \left\{ f_i(b_r) \mid b_r \in \mathbb{B} \setminus \{b_{i,1}, \dots, b_{i,z_i}\} \right\}. \quad (3)$$

Порядок внутри $\mathbb{A} \setminus \{a_i\}$, $\{b_{i,1}, \dots, b_{i,z_i}\}$, $\mathbb{B} \setminus \{b_{i,1}, \dots, b_{i,z_i}\}$ не влияет на монотонность, потому может быть произвольным для каждого из этих признаков.

Пронумеруем объекты класса 1 от 1 до $|\mathbb{A}|$, а объекты класса 0 от 1 до $|\mathbb{B}|$ произвольным образом. Построим два вспомогательных признака f_1^* и f_2^* , которые устанавливают на обучающей выборке $\mathbb{X}$ два линейных порядка, соответственно,

$$\begin{aligned} f_1^*(a_1) &> \dots > f_1^*(a_{|\mathbb{A}|}) > f_1^*(b_1) > \dots > f_1^*(b_{|\mathbb{B}|}); \\ f_2^*(a_{|\mathbb{A}|}) &> \dots > f_2^*(a_1) > f_2^*(b_{|\mathbb{B}|}) > \dots > f_2^*(b_1). \end{aligned}$$

Таким образом, мы построили обучающую выборку $\mathbb{X}$, состоящую из $n + \sum_{i=1}^n z_i$ объектов, и множество $\mathbb{F}$, состоящее из $2 + \sum_{i=1}^n c_i$ признаков.

Договоримся, что переменная $x_i \in \{0, 1\}$ принимает значение 1, если i -е множество признаков присутствует в оптимальном множестве F , и значение 0 в противном случае, для всех $i = 1, \dots, n$. Положим $t = n \sum_{i=1}^n z_i - Z$.

Предположим теперь, что мы получили не менее t монотонных пар, выбрав максимальное количество признаков.

Утверждение 6. Признаки f_1^* и f_2^* входят в оптимальное множество признаков F .

Доказательство. Если предположить обратное, то их можно добавить. При этом количество монотонных пар не уменьшится, а количество признаков станет больше, что противоречит оптимальности F .

Утверждение 7. Каждое множество признаков F_i , соответствующее определенному предмету i , либо присутствует в F целиком, либо не пересекается с F .

Доказательство. Действительно, пусть один из признаков множества F_i присутствует в F . Тогда добавление к F остальных признаков F_i не сократит количество монотонных пар объектов, но увеличит количество признаков, что противоречит оптимальности F .

Утверждение 8. *Рассмотрим те множества признаков, которые присутствуют в F . Соответствующие им объекты и будут решением задачи о рюкзаке.*

Доказательство. Число монотонных пар $n \sum_{i=1}^n z_i - \sum_{i=1}^n z_i x_i \geq n \sum_{i=1}^n z_i - Z = m$, следовательно $\sum_{i=1}^n z_i x_i \leq Z$, то есть выполняется ограничение на размер рюкзака.

Количество выбранных признаков равно $2 + \sum_{i=1}^n c_i x_i$ и максимально по предположению. Тогда ценность соответствующих предметов тоже максимальна.

Предположим обратное: пусть существует набор предметов, который при заданном ограничении на вес имеет бóльшую ценность. Переменная x_i^* принимает два значения: $x_i^* = 1$, если предмет присутствует в оптимальном наборе предметов, и $x_i^* = 0$ в противном случае, для всех $i = 1, \dots, n$. Выберем новое множество F^* признаков, включающее в себя f_1^* и f_2^* , а также все множества признаков, соответствующие выбранным предметам. Тогда выполнение условия $M \geq m$ будет следовать из выполнения условия $\sum_{i=1}^n z_i x_i^* \leq Z$, при этом число признаков будет равно $2 + \sum_{i=1}^n c_i x_i^*$.

По предположению, $\sum_{i=1}^n c_i x_i^* \geq \sum_{i=1}^n c_i x_i$, но тогда $2 + \sum_{i=1}^n c_i x_i^* \geq 2 + \sum_{i=1}^n c_i x_i$, что противоречит оптимальности множества F .

Таким образом, мы свели задачу о рюкзаке к решению задачи 7.

Теорема доказана.

Следствие 3. *Задача $FS(|M| \geq m, |F| \rightarrow \max)$ является NP-трудной.*

Задача 8. *Получить не менее m монотонных пар, выбрав не менее q признаков: $FS(|M| \geq m, |F| \geq q)$.*

Теорема 5. *Задача $FS(|M| \geq m, |F| \geq q)$ является NP-трудной.*

Доказательство. Предположим, что для задачи 8 имеется полиномиальный алгоритм решения. Минимальное количество признаков q может принимать целые значения от 0 до $|F|$. Тогда построим новый алгоритм, решающий задачу 8 для каждого возможного значения q . Он будет решать задачу $FS(|M| \geq m, |F| \rightarrow \max)$ за полиномиальное время, что противоречит её NP-трудности. Таким образом, задача 8 является NP-трудной. Теорема доказана.

Задача 9. *Получить минимальное количество дефектных пар, выбрав не более q признаков: $FS(|D_F| \rightarrow \min, |F| \leq q)$.*

Теорема 6. *Решение задачи о минимальном покрытии множества подмножествами сводится к решению задачи 9 на некоторой искусственной обучающей выборке $\mathbb{X}$.*

Доказательство. Пусть U — конечное множество, $|U| = n$, S — семейство его подмножеств.

Поставим в соответствие каждому элементу u_j множества U клетку $T_{a_j b_j}$ диагонали матрицы T . Для краткости главную диагональ $T_{a_j b_j}$ матрицы T будем называть просто *диагональю*. Каждому элементу C_i множества S поставим в соответствие признак f_i , матрица T^{f_i} которого имеет единицы выше диагонали, нули ниже диагонали,

а на диагонали единицы в тех и только в тех клетках $T_{a_j b_j}^{f_i}$, которые соответствуют элементам u_j множества C_i .

	b_1	b_2	b_{i_1}	b_{i_k}	b_n
a_1	0	1	1	1	1
a_2	0	0	1	1	1
a_{i_1}	0	0	1	1	1
a_{i_k}	0	0	0	1	1
a_n	0	0	0	0	0

Рис. 5. Матрица для i -го признака.

Далее нам понадобится несколько вспомогательных утверждений.

Утверждение 9. Все элементы матрицы T^F ниже диагонали равны 0, а выше диагонали равны $|F|$.

Доказательство. По построению, все элементы матрицы каждого признака выше диагонали равны единице, а ниже диагонали равны нулю. Тогда из равенства $T_{ab}^F = \sum_{f \in F} T_{ab}^f$ следует, что все элементы матрицы T^F ниже диагонали равны 0, а выше диагонали равны $|F|$. Утверждение 9 доказано.

Утверждение 10. Все элементы диагонали матрицы T^F не равны нулю тогда и только тогда, когда множества C_i , соответствующие признакам $f_i \in F$, образуют покрытие множества U , то есть $\bigcup_{i: f_i \in F} C_i = U$.

Доказательство. Поскольку $T_{ab}^F = \sum_{f \in F} T_{ab}^f$, все элементы диагонали не равны нулю тогда и только тогда, когда для каждого $j = 1, \dots, n$ существует признак $f_i \in F$ такой, что $T_{a_j b_j}^{f_i} = 1$. В то же время $T_{a_j b_j}^{f_i} = 1$, тогда и только тогда, когда $u_j \in C_i$. Следовательно, все элементы диагонали не равны нулю тогда и только тогда, когда для каждого $j = 1, \dots, n$ существует $f_i \in F$ такой, что $u_j \in C_i$. Тогда $\bigcup_{\{i: f_i \in F\}} C_i = U$.

Утверждение 10 доказано.

Следствие 4. Пусть F — множество, состоящее не более чем из q признаков, такое, что все элементы диагонали T^F не равны нулю. Тогда множество подмножеств C_i , соответствующих признакам $f_i \in F$, является решением задачи о покрытии множества подмножествами.

Осталось показать, что построенные нами матрицы действительно могут соответствовать признакам. Пусть элементы, соответствующие столбцам и строкам, упорядочены: $a_1 > a_2 > \dots > a_n$, $b_1 > b_2 > \dots > b_n$. Пусть в соответствующей матрице единицы на диагонали стоят в клетках $i_1, \dots, i_k$. Тогда $a_j > b_j$ для всех $j \in \{i_1, \dots, i_k\}$ и $a_j < b_j$ для всех $j \notin \{i_1, \dots, i_k\}$. Причем $a_j > a_{j+1}$, $b_j > b_{j+1}$, $a_j > b_{j+1}$, $b_j > a_{j+1}$ для всех $j = 1, \dots, n-1$. Таким образом, матрица действительно задает отношение порядка на элементах и является матрицей признака.

Итак, мы свели задачу о минимальном покрытии множества подмножествами к задаче $FS(|D_F| \rightarrow \min, |F| \leq q)$.

Теорема доказана.

Следствие 5. Задача $FS(|D_F| \rightarrow \min, |F| \leq q)$ является NP-трудной.

Задача 10. Получить не более d дефектных пар, выбрав не более q признаков: $FS(|F| \leq q, |D_F| \leq d)$.

Теорема 7. Задача $FS(|F| \leq q, |D_F| \leq d)$ является NP-трудной.

Доказательство. Предположим, что для задачи 10 имеется полиномиальный алгоритм решения. Число d может принимать целые значения от 0 до $|\mathbb{A}||\mathbb{B}|$. Тогда построим новый алгоритм, решающий задачу 10 для каждого возможного значения d . Он будет решать задачу $FS(|D_F| \rightarrow \min, |F| \leq q)$ за полиномиальное время, что противоречит её NP-трудности. Таким образом, задача отбора признаков 10 является NP-трудной. Теорема доказана.

ЗАДАЧИ ОТБОРА ОБЪЕКТОВ

Задача 11. Выбрать объекты так, чтобы при числе дефектных пар не более d получить максимальное число монотонных пар: $PS(|D| \leq d, |M| \rightarrow \max)$.

Теорема 8. Решение задачи о рюкзаке сводится к решению задачи 11 на некоторой искусственной обучающей выборке $\mathbb{X}$.

Доказательство. Пусть дан набор из n предметов. Каждый предмет имеет два параметра: вес $z_i \in \mathbb{N}$ и ценность $c_i \in \mathbb{N}$, $i = 1, \dots, n$, и имеется рюкзак, определенной вместимости $Z \in \mathbb{N}$.

Положим $d = Z$. Построим обучающую выборку $\mathbb{X}$, состоящую из n множеств объектов X_i , каждое из которых соответствует одному из предметов. Внутри каждого множества X_i находится $c_i + z_i$ объектов класса 1 и $c_i + z_i + 1$ объект класса 0, упорядоченных следующим образом:

$$b_{i,c_i+z_i} < \dots < b_{i,2} < b_{i,1} < a_{i,c_i+z_i} < \dots < a_{i,c_i+1} < b_i^* < a_{i,c_i} < \dots < a_{i,2} < a_{i,1}.$$

Для построения такого порядка на обучающей выборке $\mathbb{X}$ достаточно двух признаков. Пусть каждый из них задает указанный линейный порядок внутри множеств X_i , но первый признак упорядочивает множества X_i по возрастанию i , а второй — по убыванию.

Пусть имеется полиномиальный алгоритм, решающий задачу 11. Докажем от противного, что в оптимальном множестве X каждое множество X_i или присутствует целиком, или присутствует без объекта b_i^* . Допустим, что для некоторого множества X_j не выполняется $X_j \setminus \{b_j^*\} \supseteq X$. Тогда возможны два случая.

1. Пусть в оптимальном множестве X отсутствует b_j^* . По предположению, существуют еще объекты множества X_j , которых нет в оптимальном множестве X . Тогда количество дефектных пар в данном множестве X_j равно 0, а количество монотонных пар меньше $(c_j + m_j)^2$. Добавим к оптимальному множеству X все недостающие объекты множества X_j кроме b_j^* . Количество дефектных пар при этом не изменится, а количество монотонных пар увеличится, что противоречит предположению об оптимальности множества X .

2. Пусть в оптимальном множестве X присутствует b_j^* . Тогда в X присутствуют все объекты из подмножества $\{a_{j,1}, \dots, a_{j,c_j}\}$ и $\{b_{j,1}, \dots, b_{j,c_j+z_j}\}$, в противном случае можно было бы их добавить, не изменив таким образом число дефектных пар, но увеличив количество монотонных пар. Тогда по предположению во множестве X отсутствует хотя бы один объект a из объектов подмножества $\{a_{j,c_j+1}, \dots, a_{j,c_j+z_j}\}$. Добавим a к множеству объектов X и удалим b_j^* из него. Тогда количество дефектных пар не увеличится, а количество монотонных пар увеличится на $(c_j + z_j) - c_j$, что противоречит оптимальности множества X .

Таким образом, мы доказали, что в оптимальном множестве X каждое множество X_i или присутствует целиком, или присутствует без объекта b_i^* .

Положим, что переменная x_i принимает два значения: $x_i = 1$, если b_i^* присутствует в оптимальном множестве X , и $x_i = 0$ в противном случае, для всех $i = 1, \dots, n$. Поскольку в оптимальном множестве X каждое множество X_i или присутствует целиком, или присутствует без объекта b_i^* , то решение задачи о поиске оптимального множества X в данной обучающей выборке $\mathbb{X}$ состоит в поиске оптимального набора x_i . Рассмотрим те множества X_i , которые присутствуют в X вместе с объектом b_i^* , то есть те, для которых $x_i = 1$: соответствующие им предметы являются решением задачи о рюкзаке. Требование $\sum_{i=1}^n z_i x_i \leq Z$ выполняется, т. к. $d = Z$, а количество дефектных пар равно $\sum_{i=1}^n z_i x_i$. При этом количество монотонных пар равно $\sum_{i=1}^n (c_i + z_i)^2 + \sum_{i=1}^n c_i x_i$, и оно максимально при заданном ограничении на дефектные пары. Но тогда $\sum_{i=1}^n c_i x_i$ максимально при $\sum_{i=1}^n z_i x_i$, поскольку $\sum_{i=1}^n (c_i + z_i)^2$ не зависит от набора x_i .

Таким образом, решение задачи о рюкзаке полиномиально сводится к решению задачи отбора объектов $PS(|D| \leq d, |M| \rightarrow \max)$.

Теорема доказана.

Задача 12. *Выбрать объекты так, чтобы монотонных пар было не менее t , а дефектных пар не более d : $PS(|D| \leq d, |M| \geq t)$.*

Теорема 9. *Задача $PS(|D| \leq d, |M| \geq t)$ является NP-трудной.*

Доказательство. Предположим, что для задачи 12 имеется полиномиальный алгоритм. Построим тогда алгоритм, решающий задачу при всех возможных параметрах t . Поскольку t может принимать целые значения от 0 до $|\mathbb{A}||\mathbb{B}|$, потребуется перебрать не более $|\mathbb{A}||\mathbb{B}| + 1$ вариантов. То есть получаемый алгоритм также является полиномиальным, но решает задачу $PS(|D| \leq d, |M| \rightarrow \max)$. Получили противоречие. Следовательно, задача 12 является NP-трудной. Теорема доказана.

Задача 13. *Выбрать не более x объектов так, чтобы число монотонных пар было не менее t : $PS(|M| \geq t, |X| \leq x)$.*

Теорема 10. *Задача $PS(|M| \geq t, |X| \leq x)$ является NP-трудной.*

Доказательство. Сведем задачу о поиске максимального полного двудольного подграфа к данной задаче.

Пусть дан двудольный граф G . Поставим в соответствие каждой вершине левой доли объект класса 0, а каждой вершине правой доли — объект класса 1. Пусть пары, связанные в G ребром, будут соответствовать монотонным парам.

Покажем, что такую подвыборку всегда можно построить, то есть построить признаки, индуцирующие соответствующий частичный порядок. Пронумеруем объекты класса 0 и класса 1. Построим два вспомогательных признака, индуцирующих два линейных порядка, соответственно:

$$\begin{aligned} a_1 &> a_2 > \dots > a_{|\mathbb{A}|} > b_1 > b_2 > \dots > b_{|\mathbb{B}|}; \\ b_1 &< b_2 < \dots < b_{|\mathbb{B}|} < a_1 < a_2 < \dots < a_{|\mathbb{A}|}. \end{aligned}$$

Тогда в частичном порядке, индуцированном этими двумя признаками, все объекты класса 0 будут меньше объектов класса 1, а объекты внутри классов будут несравнимы.

Теперь для каждого объекта a_i класса 1 и соответствующей вершины r_i графа рассмотрим те объекты класса 0, которым соответствуют вершины, соединенные с r_i

ребром. Построим признак q_i , принимающий значение 1 на объекте a_i , значение 3 на остальных объектах класса 1; значение 0 на объектах, которым соответствуют вершины, соединенные с r_i ребром; значение 2 на остальных объектах класса 0.

Получим искомую подвыборку с описанным выше частичным порядком. Она включает в себя $|V|$ объектов и не более $|V| + 2$ признаков. В ней каждой вершине графа взаимно однозначно соответствует объект, а каждому ребру — монотонная пара.

Предположим, что существует полиномиальный алгоритм, решающий задачу выбора не более x объектов так, чтобы монотонных пар было не менее m .

Тогда для каждого $i = 1, \dots, |V|$ построим соответствующую обучающую выборку и запустим алгоритм поиска не более $2i$ объектов так, чтобы монотонных пар было не менее i^2 . Найденные подвыборки будут соответствовать бикликам.

Тогда из предположения существования полиномиального алгоритма для нашей задачи следует существование полиномиального алгоритма для поиска биклики. Таким образом, задача 13 является NP-трудной.

Теорема доказана.

Задача 14. Устранить все дефектные пары, оставив в обучающей выборке как можно больше объектов: $PS(|D| = 0, |X| \rightarrow \max)$.

Эту задачу можно переформулировать так: устранить все дефектные пары, удалив из обучающей выборки как можно меньше объектов. Построим полиномиальный алгоритм решения задачи 14 и оценим его сложность.

Рассмотрим объекты, участвующие в дефектных парах. Представим их в виде двудольного графа: в одной доле объекты класса $\mathbb{A}$, в другой доле объекты класса $\mathbb{B}$, ребрами соединены объекты, образующие дефектные пары. Тогда задача об удалении минимального количества объектов так, чтобы дефектных пар не было, линейно сводится к задаче о поиске минимального вершинного покрытия для данного двудольного графа.

Пусть d — количество дефектных пар, a_d и b_d — число объектов класса $\mathbb{A}$ и $\mathbb{B}$ соответственно, задействованных в дефектных парах. Тогда время поиска удаляемых объектов составляет $O(d(a_d + b_d)^{0,5})$ по теореме Кёнига.

Задача 15. Устранить все дефектные пары, выбрав не менее x объектов обучающей выборки: $PS(|D| = 0, |X| \geq x)$.

Рассмотрим множество объектов $\bar{X}$, являющееся решением задачи 14 на обучающей выборке $\mathbb{X}$. Тогда при $x \leq |\bar{X}|$ множество $\bar{X}$ является также и решением задачи 15. Из максимальной $|\bar{X}|$ следует, что при $x > |\bar{X}|$ задача 15 не имеет решения.

Задача 16. Получить не более d дефектных пар, выбрав не менее x объектов обучающей выборки $\mathbb{X}$: $PS(|D| \leq d, |X| \geq x)$.

Теорема 11. Задача 16 является полиномиальной по числу объектов.

Доказательство. Рассмотрим объекты, участвующие в дефектных парах. Представим их в виде двудольного графа: в одной доле объекты класса $\mathbb{A}$, в другой доле объекты класса $\mathbb{B}$, ребрами соединены объекты, образующие дефектные пары. Удалим d ребер всеми возможными способами и сведем задачу к предыдущей. Это возможно сделать $C_{|D|}^d$ способами, то есть за время $O(|X|^{2d})$, полиномиальное по числу объектов в обучающей выборке, но экспоненциальное по d . Теорема доказана.

ЗАКЛЮЧЕНИЕ

В работе предложена систематизация задач отбора объектов и признаков, возникающих при построении монотонных классификаторов. Исследованы вопросы вычислительной сложности этих задач. Показано, что лишь одна из них (задача отбора

объектов при ограничении числа дефектных пар сверху и числа отобранных объектов снизу) обладает полиномиальной вычислительной сложностью. Все остальные задачи являются NP-трудными, и для их решения оправдано применение приближённых эвристических алгоритмов. За рамками статьи остались задачи одновременного отбора как объектов, так и признаков. Практическое применение методов дискретной оптимизации для решения таких задач является направлением дальнейших исследований.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проекты № 14-07-31240 и № 14-07-00847.

СПИСОК ЛИТЕРАТУРЫ

1. Воронцов К.В. О проблемно-ориентированной оптимизации базисов задач распознавания. *ЖВМ и МФ*. 1998. Т. 38. № 5. С. 870–880.
2. Воронцов К.В. Оптимизационные методы линейной и монотонной коррекции в алгебраическом подходе к проблеме распознавания. *ЖВМ и МФ*. 2000. Т. 40. № 1. С. 166–176.
3. Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов. В: *Математические вопросы кибернетики*. Под ред. Лупанова. О. Б. М.: Физматлит, 2004. Т. 13. С. 5–36.
4. Воронцов К.В., Махина Г.А. Принцип максимизации зазора для монотонного классификатора ближайшего соседа. В: *15-я всероссийская конференция "Математические методы распознавания образов"*. М.: МАКС Пресс, 2011. С. 64–67.
5. Громова О.А., Торшин И.Ю., Рудаков К.В., Грустливая У.Е., Калачева А.Г., Юдина Н.В., Егорова Е.Ю., Лиманова О.А., Федотова Л.Э., Грачева О.Н. и др. Недостаточность магния — достоверный фактор риска коморбидных состояний: результаты крупномасштабного скрининга магниевого статуса в регионах России. *Фарматека*. 2013. № 6. С. 116–129.
6. Гуз И.С. Нелинейные монотонные композиции классификаторов. В: *13-я всероссийская конференция "Математические методы распознавания образов"*. М.: МАКС Пресс, 2007. С. 111–114.
7. Гуз И.С. Минимизация эмпирического риска при построении монотонных композиций классификаторов. *Труды МФТИ*. 2011. Т. 3. № 3(11). С. 115–121.
8. Загоруйко Н.Г. *Прикладные методы анализа данных и знаний*. Новосибирск: ИМ СО РАН, 1999.
9. Иванов М.Н., Воронцов К.В. Применение монотонного классификатора ближайшего соседа в задаче категоризации текстов. В: *Интеллектуализация обработки информации (ИОИ-9): Докл.* Москва: Торус Пресс, 2012. С. 621–624.
10. Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К. *Алгоритмы: построение и анализ, 2-е издание*. М.: Издательский дом "Вильямс" 2005.
11. Махина Г.А. О восстановлении монотонных булевых функций методом ближайшего соседа. В: *Интеллектуализация обработки информации (ИОИ-2012): Докл.* Москва: Торус Пресс, 2012. С. 78–81.
12. Успенский В.М. *Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов*. М.: Экономика и информатика, 2008. 116 с.

13. Успенский В.М. Информационная функция сердца. *Клиническая медицина*. 2008. Т. 86. № 5. С. 4–13.
14. Barlow R., Bartholomew D., Bremner J., Brunk H. *Statistical inference under order restrictions; the theory and application of isotonic regression*. New York: Wiley, 1972.
15. Chandrasekaran R., Ryu Y.U., Jacob V.S., Hong S. Isotonic separation. *INFORMS J. on Computing*. 2005. V. 17. № 4. P. 462–474.
16. Cortes C., Vapnik V. Support-vector networks. *Machine Learning*. 1995. V. 20. № 3. P. 273–297.
17. Ivanov M.N. Prototype sample selection based on minimization of the complete cross validation functional. *Pattern Recognition and Image Analysis*. 2010. V. 20. № 4. P. 427–437.
18. Kamp R., Feelders A., Barile N. Isotonic classification trees. In: *Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII*. Berlin, Heidelberg: Springer-Verlag, 2009. P. 405–416.
19. Malar B., Nadarajan R., Saisundarakrishnan G. Isotonic separation with an instance selection algorithm using softset: Theory and experiments. *WSEAS Transactions On Information Science And Applications*. 2012. V. 9. № 11. P. 350–367.
20. Natarajan B.K. Sparse approximate solutions to linear systems. *Society for Industrial and Applied Mathematics*. 1995. V. 24. P. 227–234.
21. Royston P. A useful monotonic non-linear model with applications in medicine and epidemiology. *Statistics in Medicine*. 2000. V. 19. № 15. P. 2053–2066.
22. Ryu Y.U., Chandrasekaran R., Jacob V.S. Breast cancer prediction using the isotonic separation technique. *European Journal of Operational Research*. 2007. V. 181. № 2. P. 842–854.
23. Sill J. Monotonic networks. In: *Advances in Neural Information Processing Systems*. Ed. Jordan M.I., Kearns M.J., Solla S.A. Cambridge: MIT Press, 1998. P. 661–667.
24. Sill J., Abu-Mostafa Y.S. Monotonicity hints. In: *Advances in Neural Information Processing Systems*. Ed. Mozer M.C., Jordan M.I., Petsche T. Cambridge: MIT Press, 1997. P. 634–640.
25. Spirin N.V., Vorontsov K.V. Learning to rank with nonlinear monotonic ensemble. In: *10th International Workshop on Multiple Classifier Systems (Naples, Italy, June 15–17, 2011)*: Lecture Notes in Computer Science. Ed. Sansone C., Kittler J., Roli F. Berlin: Springer-Verlag, 2011. P. 16–25.
26. Torshin I.Yu. The study of the solvability of the genome annotation problem on sets of elementary motifs. *Pattern Recognition and Image Analysis*. 2011. V. 21. P. 652–662.
27. 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. *Advances in Mathematical and Computational Tools in Metrology and Testing X, vol. 10*. 2015. V. 86. P. 377–384. (Series on Advances in Mathematics for Applied Sciences, World Scientific, Singapore).
28. Zhang J. Advancements of Outlier Detection: A Survey. *ICST Transactions on Scalable Information Systems*. 2013. V. 13.
29. Zukhba A.V. NP-completeness of the problem of prototype selection in the nearest neighbor method. *Pattern Recognition and Image Analysis*. 2010. V. 20. P. 484–494.

Материал поступил в редакцию 10.09.2015, опубликован 25.09.2015.