

NP-ТРУДНОСТЬ НЕКОТОРОЙ ЗАДАЧИ ЦЕНЗУРИРОВАНИЯ ДАННЫХ

О. А. Кутненко^{1,2, a}, А. В. Плясунов^{1,2, b}

¹ Институт математики им. С. Л. Соболева,
пр. Акад. Коптюга, 4, 630090 Новосибирск, Россия

² Новосибирский гос. университет,
ул. Пирогова, 2, 630090 Новосибирск, Россия

E-mail: ^aolga@math.nsc.ru, ^bapljas@math.nsc.ru

Аннотация. Доказана NP-трудность рассматриваемой в работе постановки задачи цензурирования данных. К решению такой задачи сводится одна из проблем анализа данных. В качестве количественной оценки компактности образа используется функция конкурентного сходства (FRiS-функция), с помощью которой оценивается локальное сходство объектов со своими ближайшими соседями. Ил. 1, библиогр. 23.

Ключевые слова: NP-трудность, цензурирование объектов, компактность образов, функция конкурентного сходства.

Введение

Развитие технологий в современном мире приводит к экспоненциальной скорости роста объёма информации в самых разных областях, что даёт, с одной стороны, новые возможности для решения различных прикладных задач, но, с другой стороны, повышает риск появления ошибок в анализируемых данных. Поэтому в настоящее время проблема цензурирования данных (data filtering, data cleaning) приобретает всё большую актуальность при решении самых разных задач [1–3]. В большинстве стандартных пакетов данных используются различные алгоритмы фильтрации данных [4–15].

В работе рассматривается задача очистки обучающей выборки, представленной объектами двух классов, от шумовых объектов только одного класса, т. е. при этом повышается качество описания одного образа при фиксированном втором образе. Такие задачи возникают, в частности,

Исследование выполнено в рамках государственного задания ИМ СО РАН (проекты № 0314–2019–0015, 0314–2019–0014).

при анализе биомедицинских данных, требующем полного сохранения данных одного из образов.

Исключение из обучающей выборки неверно классифицированных объектов (или объектов-выбросов) осуществляется на основе анализа локального окружения объектов. Данный подход опирается на гипотезу локальной компактности [16]. Количественная характеристика локальной компактности образа оценивается с помощью функции конкурентного сходства, успешно используемой в когнитивном анализе данных при решении различных прикладных задач [17–20].

1. Цензурирование объектов-выбросов с помощью функции конкурентного сходства

1.1. FRiS-компактность и качество описания выборки. Для получения количественной оценки компактности образов в фиксированном признаковом пространстве предлагается использовать FRiS-функцию, с помощью которой формализуется представление о компактности как о «высоком» сходстве объектов одного образа друг с другом и «низком» сходстве с объектами других образов. Пусть даны два образа: \mathbf{A} и \mathbf{B} . Для вычисления конкурентного сходства объекта z с объектом $a \in \mathbf{A}$ в конкуренции с объектом $b \in \mathbf{B}$ с опорой на некоторую метрику τ , определяющую расстояние между этими объектами, используется тернарная относительная мера, которая называется функцией конкурентного сходства или FRiS-функцией (function of rival similarity) [19]:

$$F(z, a|b) = \frac{\tau(z, b) - \tau(z, a)}{\tau(z, b) + \tau(z, a)}.$$

Конкурентное сходство объектов с образами будем определять по тому же принципу, что и конкурентное сходство объектов с объектами:

$$F(z, \mathbf{A}|\mathbf{B}) = \frac{\tau(z, \mathbf{B}) - \tau(z, \mathbf{A})}{\tau(z, \mathbf{B}) + \tau(z, \mathbf{A})}. \quad (1)$$

Для произвольного объекта $z \in \mathbf{A}$ мера конкурентного сходства этого объекта со своим образом в конкуренции с образом \mathbf{B} показывает, насколько этот объект похож на представителей своего образа и не похож на представителей образа \mathbf{B} . Эту характеристику можно определить для каждого объекта в отдельности и тем самым оценить вклад этого объекта в компактность своего образа [17]:

$$F_{\mathbf{A}|\mathbf{B}}(\mathbf{A}) = \frac{1}{|\mathbf{A}|} \sum_{a \in \mathbf{A}} F(a, \mathbf{A}|\mathbf{B}), \quad (2)$$

где $|\mathbf{A}|$ — число объектов образа \mathbf{A} . Далее эта величина будет называться FRiS-компактностью образа \mathbf{A} .

Так как с ростом числа исключённых объектов неизбежно повышается переобученность метода, для учёта этого эффекта используется нормирующий коэффициент $|\mathbf{A}^*|/|\mathbf{A}|$, где $|\mathbf{A}^*| = |\mathbf{A} \setminus \mathbf{A}'|$ — число объектов, оставшихся после удаления $|\mathbf{A}'|$ объектов, принадлежащих образу \mathbf{A} . Тем самым компактность «очищенного» образа задаётся формулой

$$H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*) = \frac{|\mathbf{A}^*|}{|\mathbf{A}|} F_{\mathbf{A}^*|\mathbf{B}}(\mathbf{A}^*) = \frac{1}{|\mathbf{A}|} \sum_{a \in \mathbf{A}^*} F(a, \mathbf{A}^*|\mathbf{B}). \quad (3)$$

Для решения рассматриваемой задачи необходимо найти множество \mathbf{A}' удалённых объектов или множество $\mathbf{A}^* = \mathbf{A} \setminus \mathbf{A}'$ оставшихся объектов, на котором достигается максимум $H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*)$.

1.2. Постановка задачи. Даны два непересекающихся конечных множества объектов \mathbf{A} и \mathbf{B} , $|\mathbf{A} \cup \mathbf{B}| = M$. Задана матрица попарных расстояний между всеми объектами множества $\mathbf{A} \cup \mathbf{B}$; в качестве расстояния от объекта до образа используется среднее расстояние до $k \geq 1$ ближайших объектов образа при условии, что $|\mathbf{A}| \geq k + 1$, $|\mathbf{B}| \geq k + 1$. Требуется найти множество объектов $\mathbf{A}' \subset \mathbf{A}$, удаление которых обеспечивает максимум функции $H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*)$.

Множество $\mathbf{A} \cup \mathbf{B}$ можно записать как выборку $\{(x_i, y_i)\}_{i=\overline{1, M}}$, где $y_i \in \{-1, 1\}$ — номинальный целевой признак. Образ \mathbf{A} запишем как множество объектов $\{x_i \mid y_i = 1, i = \overline{1, M}\}$, \mathbf{B} — как $\{x_i \mid y_i = -1, i = \overline{1, M}\}$; $\mathbf{y} = \{y_i\}_{i=\overline{1, M}}$. Тогда (2) запишем в следующем виде:

$$F_{\mathbf{A}|\mathbf{B}}(\mathbf{A}) = \frac{1}{|\{x_i \mid y_i = 1, i = \overline{1, M}\}|} \times \sum_{i: y_i=1} F(x_i, \{x_j \mid y_j = 1, j = \overline{1, M}\} \setminus \{x_i\} \mid \{x_j \mid y_j = -1, j = \overline{1, M}\}). \quad (4)$$

Для заданного вектора \mathbf{y} определим множество

$$\mathbf{T} = \{\mathbf{t} = \{t_i\}_{i=\overline{1, M}} \mid t_i \in \{-1, 0, 1\}; t_i = y_i, \text{ если } y_i = -1\}.$$

Здесь $t_i = 0$ означает, что i -й объект выборки исключён, а второе условие гарантирует, что удаляться могут только объекты образа \mathbf{A} . При решении задачи поиска объектов-выбросов образа \mathbf{A} все изменения касаются только целевого признака объектов образа \mathbf{A} , поэтому целью является нахождение вектора значений целевого признака $\mathbf{t} \in \mathbf{T}$. Тогда $\mathbf{A}' = \{x_i \mid t_i = 0, i = \overline{1, M}\}$, $\mathbf{A}^* = \{x_i \mid t_i = 1, i = \overline{1, M}\}$.

Таким образом, для решения рассматриваемой смысловой задачи требуется решить экстремальную задачу — найти вектор \mathbf{t}^* , определяющий

множество \mathbf{A}^* , на котором достигается максимум $H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*)$:

$$\mathbf{t}^* = \arg \max_{\mathbf{t} \in \mathbf{T}} \sum_{i: t_i=1} F(x_i, \{x_j \mid t_j = 1, j = \overline{1, M}\} \mid \{x_j \mid t_j = -1, j = \overline{1, M}\}).$$

2. Вычислительная сложность задачи

Доказательство NP-трудности будет выполнено сведением известной NP-полной задачи о вершинном покрытии графа к задаче выбора подмножества, на котором компактность образа будет максимальна.

Задача ВП (вершинное покрытие) [21]. Даны граф $G = (V, E)$ и положительное целое число $J \leq |V|$. Имеется ли в графе G вершинное покрытие не более чем из J элементов, т. е. такое подмножество $V' \subseteq V$, что $|V'| \leq J$ и для каждого ребра $\{u, v\} \in E$ хотя бы одна из вершин u или v принадлежит V' ?

В качестве метрики τ в (1) используется среднее расстояние до $k \geq 1$ ближайших объектов образа. Для доказательства потребуется

Утверждение 1. Для любого $k \in \mathbb{N}$ и любого положительного $r_1 \in \mathbb{Q}$ существуют числа r и r_2 такие, что

$$0 < r < r_1 < r_2 < 2r, \quad (5)$$

$$\frac{r - r_1}{r + r_1} + \frac{r_2 - r_1}{2kr_2 - r_2 + r_1} > 0. \quad (6)$$

Здесь и далее \mathbb{N} — множество натуральных чисел, \mathbb{Q} — множество рациональных чисел.

ДОКАЗАТЕЛЬСТВО. Имеем

$$\frac{r - r_1}{r + r_1} + \frac{r_2 - r_1}{2kr_2 - r_2 + r_1} = \frac{(r - r_1)(2kr_2 - r_2 + r_1) + (r_2 - r_1)(r + r_1)}{(r + r_1)(2kr_2 - r_2 + r_1)}.$$

Учитывая, что $(r + r_1)(2kr_2 - r_2 + r_1) > 0$, найдём r_2 , для которых выполняется (6):

$$\begin{aligned} (r - r_1)(2kr_2 - r_2 + r_1) + (r_2 - r_1)(r + r_1) \\ = 2krr_2 - 2kr_1r_2 + 2r_1r_2 - 2r_1^2 > 0, \end{aligned}$$

т. е. $r_2(kr - kr_1 + r_1) > r_1^2$. Отсюда $kr - kr_1 + r_1 > 0$, что верно для всех $k \in \mathbb{N}$ при

$$r > \frac{k-1}{k} r_1. \quad (7)$$

Итак,

$$r_2 > \frac{r_1^2}{kr - kr_1 + r_1}. \quad (8)$$

Далее найдём r , для которых

$$\frac{r_1^2}{kr - kr_1 + r_1} < 2r, \quad (9)$$

т. е. $r_1^2 < 2kr^2 - 2krr_1 + 2rr_1$. Обозначим через $r_{1,2}^*$ корни квадратного уравнения

$$2kr^2 - (2kr_1 - 2r_1)r - r_1^2 = 0, \\ r_{1,2}^* = \frac{(2kr_1 - 2r_1) \pm \sqrt{4k^2r_1^2 - 8kr_1^2 + 4r_1^2 + 8kr_1^2}}{4k} = r_1 \frac{k - 1 \pm \sqrt{k^2 + 1}}{2k}.$$

Так как $k - 1 < \sqrt{k^2 + 1}$, при $r > 0$ имеем

$$r > r_1 \frac{k - 1 + \sqrt{k^2 + 1}}{2k}.$$

Поскольку $\frac{k-1+\sqrt{k^2+1}}{2k} > \frac{k-1}{k}$ для всех $k \in \mathbb{N}$, выполняется (7). Нетрудно показать, что

$$\frac{k - 1 + \sqrt{k^2 + 1}}{2k} < 1.$$

Таким образом, учитывая (5), получим множество значений r :

$$r \in \left(\frac{k - 1 + \sqrt{k^2 + 1}}{2k} r_1, r_1 \right). \quad (10)$$

Определим множество значений r_2 : $r_1 < r_2 < 2r$. Очевидно, что для всех $k \in \mathbb{N}$ при $r < r_1$ справедливо

$$r_1 < \frac{r_1^2}{kr - kr_1 + r_1}. \quad (11)$$

Из (8), (9) и (11) следует, что

$$r_2 \in \left(\frac{r_1^2}{kr - kr_1 + r_1}, 2r \right). \quad (12)$$

Показано, что для любого $k \in \mathbb{N}$ и любого положительного $r_1 \in \mathbb{Q}$ существуют r и r_2 , определяемые согласно (10) и (12) и удовлетворяющие требуемым условиям. Утверждение 1 доказано.

Для удобства изложения и визуализации \mathbf{A} назван образом белых объектов, а \mathbf{B} — образом чёрных объектов.

Теорема 1. Задача поиска наименьшего вершинного покрытия произвольного графа $G = (V, E)$ сводится к задаче выбора из некоторой искусственной выборки X_G множества объектов $\mathbf{A}^* \subseteq \mathbf{A}$, на котором достигается максимум функционала H .

При этом выборка X_G строится по G за полиномиальное время и имеет полиномиальное число объектов относительно $|V| + |E|$, а в \mathbf{A}^* содержится не менее $k + 1$ объектов образа \mathbf{A} .

ДОКАЗАТЕЛЬСТВО. По заданному графу $G = (V, E)$ построим X_G следующим образом.

Пусть в X_G будет два класса: \mathbf{A} — белые объекты, \mathbf{B} — чёрные объекты, а удаляться могут только белые объекты. Каждой вершине A графа поставим в соответствие белый объект A , а каждому ребру AB поставим в соответствие группу из k белых объектов $\{(AB)_i\} = \{(AB)_i \mid i = 1, \dots, k\}$. Расстояние между объектами, соответствующими смежным вершинам графа, положим равным r_1 . Расстояние между каждым объектом группы, соответствующей ребру графа, и объектом, соответствующим вершине графа, являющейся одним из концов данного ребра, также положим равным r_1 .

Для каждой вершины A добавим группу из k чёрных объектов: $\{\bar{A}_i\} = \{\bar{A}_i \mid i = 1, \dots, k\}$, каждый из которых расположим на расстоянии r от объекта A и на расстоянии r_2 от каждого объекта из группы $\{(AB)_i\}$, соответствующей ребру AB . Расстояния между объектами внутри группы чёрных объектов положим равными r , внутри группы белых объектов — равными r_2 .

На r, r_1, r_2 наложим условие $0 < r < r_1 < r_2 < 2r$, при этом r и r_2 задаются согласно (10) и (12).

Все неговоренные выше расстояния положим равными $R > r_2$. Предполагается, что $R < 2r$, тогда заданная цепочка неравенств $r < r_1 < r_2 < R < 2r$ обеспечивает выполнение неравенства треугольника.

Итак, выборка X_G построена (рис. 1).

Из (1) следует, что вклады в $H_{\mathbf{A}|\mathbf{B}}(\mathbf{A})$ всех белых объектов, входящих в группы типа $\{(AB)_i\}$, одинаковы. Учитывая, что каждой вершине соответствует один белый объект A , а каждому ребру соответствует группа из k белых объектов $\{(AB)_i\}$, получим

$$H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}) = \frac{|V|F(A, \mathbf{A}|\mathbf{B}) + |E|kF((AB)_i, \mathbf{A}|\mathbf{B})}{|V| + k|E|}.$$

Ставится задача найти множество $\mathbf{A}^* = \mathbf{A} \setminus \mathbf{A}'$, на котором значение функционала $H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*)$ максимально.

Рассмотрим произвольную группу белых объектов, соответствующих ребру AB и его вершинам A и B , а также достроенные к ним чёрные объекты, и проанализируем вклад каждого из белых объектов в $H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*)$.

Вклад объекта A (B) не превосходит $(r - r_1)/(r + r_1)$; равенство достигается, если в \mathbf{A}^* входят k белых объектов образа \mathbf{A} , находящихся на расстоянии r_1 от $A(B)$. Согласно (5) этот вклад отрицательный.

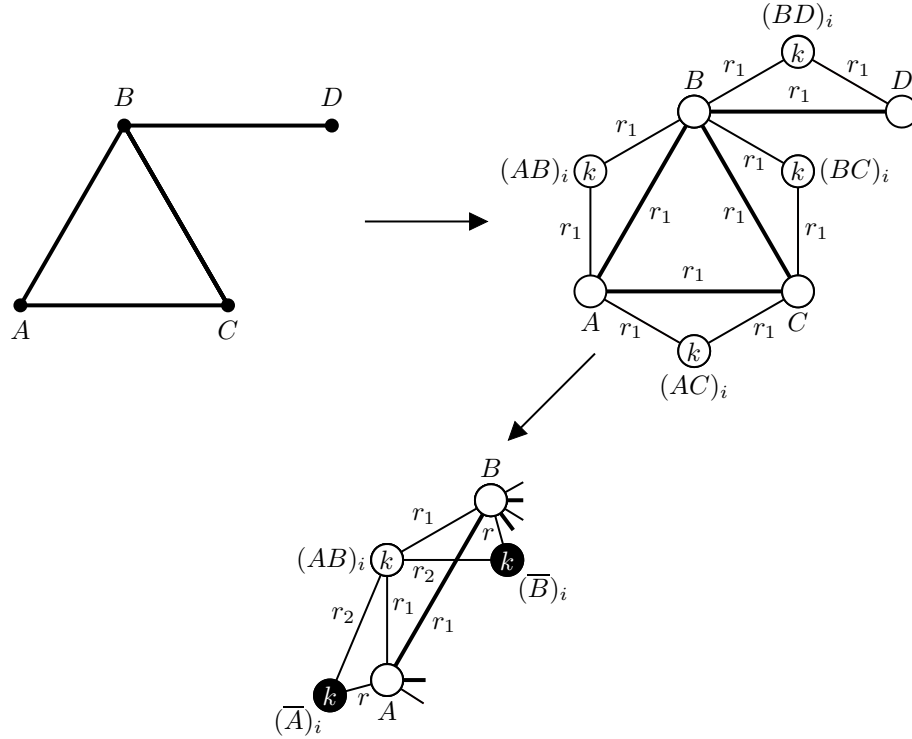


Рис. 1. Пример графа G , состоящего из четырёх вершин, и построение искусственной выборки X_G , $k \geq 1$

Если в \mathbf{A}^* не входят объекты типа A , то вклад объекта $(AB)_i$ максимален при вхождении в \mathbf{A}^* всех объектов группы $\{(AB)_i\}$ и будет равен

$$\frac{kr_2 - (k-1)r_2 - R}{kr_2 + (k-1)r_2 + R},$$

где $((k-1)r_2 + R)/k$ — среднее расстояние до k ближайших белых объектов образа \mathbf{A} , входящих в \mathbf{A}^* . В этом случае вклад отрицательный, так как $r_2 < R$.

Если в \mathbf{A}^* входит объект A и/или B и вся группа $\{(AB)_i\}$, то вклад каждого объекта $(AB)_i$ из этой группы равен

$$\frac{kr_2 - (k-1)r_2 - r_1}{kr_2 + (k-1)r_2 + r_1},$$

где $((k-1)r_2 + r_1)/k$ — среднее расстояние до k ближайших белых объектов образа \mathbf{A} , входящих в \mathbf{A}^* . В данном случае согласно (5) вклад положительный.

Заметим, что при заданных условиях на расстояния r, r_1, r_2 в силу утверждения 1 вклад в $H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*)$ объекта $(AB)_i$ и объекта типа A равен

$$\frac{r - r_1}{r + r_1} + \frac{r_2 - r_1}{2kr_2 - r_2 + r_1} > 0.$$

Таким образом, для того чтобы объект типа $(AB)_i$ вошёл в \mathbf{A}^* , необходимо и достаточно, чтобы в \mathbf{A}^* входил по крайней мере один объект типа A , соответствующий вершине A или B .

Поскольку объекты типа A входят с отрицательным весом, очевидно, что максимальное значение функционала H достигается, когда в \mathbf{A}^* входит минимальное число $|V'| \leq |V|$ таких объектов:

$$\begin{aligned} H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*) &= \frac{|V'|F(A, \mathbf{A}^*|\mathbf{B}) + |E|kF((AB)_i, \mathbf{A}^*|\mathbf{B})}{|V| + k|E|} \\ &= \frac{1}{|V| + k|E|} \left(|V'| \frac{r - r_1}{r + r_1} + |E|k \frac{r_2 - r_1}{2kr_2 - r_2 + r_1} \right). \end{aligned}$$

Далее покажем, что вершины, соответствующие объектам типа A , входящим в множество \mathbf{A}^* , на котором достигается максимальное значение функционала H , образуют минимальное вершинное покрытие графа $G(V, E)$.

Пусть данные вершины не образуют вершинного покрытия, т. е. существует объект $(AB)_i$, входящий в соответствующую ребру AB группу, такой, что в \mathbf{A}^* не входят объекты A и B . Тогда вклад $(AB)_i$ в $H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*)$, равный

$$F((AB)_i, \mathbf{A}^*|\mathbf{B}) = \frac{kr_2 - (k - 1)r_2 - R}{kr_2 + (k - 1)r_2 + R} < 0,$$

уменьшает значение функционала H , что противоречит оптимальности $H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*)$.

Пусть вершинное покрытие V' , образованное данными вершинами, не минимально, т. е. существует множество $\mathbf{A}_1^* \subset \mathbf{A}$, содержащее все объекты типа $(AB)_i$ и $|V'_1|$ объектов типа A , и $|V'_1| < |V'|$. Учитывая, что вклад объектов типа A , равный $(r - r_1)/(r + r_1)$, отрицателен, получим $H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*) < H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}_1^*)$, что противоречит предположению об оптимальности $H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*)$.

Таким образом, доказано, что для достижения максимального значения $H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*)$ необходимо и достаточно вхождения в \mathbf{A}^* всех объектов входящих в группы типа $\{(AB)_i\}$, соответствующие рёбрам графа G , и объектов типа A , соответствующих минимальному вершинному покрытию $V' \subseteq V$ графа G .

Также показан способ построения минимального вершинного покрытия по известному множеству \mathbf{A}^* , т. е. задача поиска минимального вершинного покрытия произвольного графа G сведена к задаче выбора из некоторой искусственной выборки X_G множества объектов \mathbf{A}^* , на котором достигается максимум критерия H .

При этом время построения X_G и время преобразования множества \mathbf{A}^* в вершинное покрытие имеет порядок $O(|V| + k|E|)$. Так как в множество \mathbf{A}^* наряду с k объектами $(AB)_i$ входит объект A и/или B , то $|\mathbf{A}^*| \geq k + 1$, что позволяет использовать заданную выше метрику τ в качестве расстояния от объекта до образа. Теорема 1 доказана.

Поскольку $\mathbf{A}' = \mathbf{A} \setminus \mathbf{A}^*$, из доказанной теоремы вытекает

Следствие 1. Задача поиска множества $\mathbf{A}' = \mathbf{A} \setminus \mathbf{A}^*$ удаляемых объектов-выбросов образа \mathbf{A} NP-трудна.

NP-трудность рассмотренной постановки задачи цензурирования объясняет применение различных эвристических алгоритмов для решения задач цензурирования данных, в которых в качестве количественной оценки компактности образа используется функция конкурентного сходства [4, 22, 23].

Заключение

Показана NP-трудность экстремальной задачи поиска множества, на котором согласно заданному критерию достигается максимум оценки компактности образа. Таким образом, показана труднорешаемость соответствующей проблемы анализа данных. Отметим, что в настоящее время неизвестно алгоритмов цензурирования данных с гарантированными оценками точности для решения рассмотренной задачи.

ЛИТЕРАТУРА

1. Osborne J. W. Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data. Los Angeles: SAGE Publ., 2013. 296 p.
2. Farcomeni A., Greco L. Robust methods for data reduction. New York: CRC Press, 2015. 297 p.
3. Waal T. D., Pannekoek J., Scholtus S. Handbook of statistical data editing and imputation. Hoboken, NJ: Wiley, 2011. 456 p.
4. Борисова И. А., Кутненко О. А. Цензурирование ошибочно классифицированных объектов выборки // Машин. обучение и анализ данных. 2015. Т. 1, № 11. С. 1632–1641.
5. Aggarwal C. C. Data mining. Cham: Springer, 2015. 734 p.
6. Brighton H., Mellish C. Advances in instance selection for instance-based learning algorithms // Data Min. Knowl. Discov. 2002. Vol. 6, No. 2. P. 153–172.

7. **Delany S. J., Segata N., Mac Namee B.** Profiling instances in noise reduction // *Knowl.-B. Syst.* 2012. Vol. 31. P. 28–40.
8. **Frenay B., Verleysen M.** Classification in the presence of label noise: A survey // *IEEE Trans. Neural Netw. Learn. Syst.* 2014. Vol. 25, No. 5. P. 845–869.
9. **Jankowski N., Grochowski M.** Comparison of instances selection algorithms I. Algorithms survey // *Artificial Intelligence and Soft Computing — ICAISC 2004. Proc. 7th Int. Conf. (Zakopane, Poland, June 7–11, 2004)*. Heidelberg: Springer, 2004. P. 598–603. (Lect. Notes Comput. Sci.; Vol. 3070).
10. **Massie S., Craw S., Wiratunga N.** When similar problems don't have similar solutions // *Case-Based Reasoning Research and Development. Proc. 7th Int. Conf. (Belfast, NI, UK, Aug. 13–16, 2007)*. Heidelberg: Springer, 2007. P. 92–106. (Lect. Notes Comput. Sci.; Vol. 4626).
11. **Quinlan J. R.** Induction of decision trees // *Mach. Learn.* 1986. Vol. 1. P. 81–106.
12. **Segata N., Blanzieri E.** Noise reduction for instance-based learning with a local maximal margin approach // *J. Intel. Inf. Syst.* 2010. Vol. 35, No. 2. P. 301–331.
13. **Son S.-H., Kim J.-Y.** Data reduction for instance-based learning using entropy-based partitioning // *Computational Science and Its Applications — ICCSA 2006. Proc. Int. Conf. (Glasgow, UK, May 8–11, 2006)*. Pt. 3. Heidelberg: Springer, 2006. P. 590–599. (Lect. Notes Comput. Sci.; Vol. 3982).
14. **Teng C. M.** A comparison of noise handling techniques // *Proc. 14th Int. Florida Artificial Intelligence Res. Soc. Conf. (Key West, FL, USA, May 21–23, 2001)*. Menlo Park, CA: AAAI Press, 2001. P. 269–273.
15. **Wilson D. R., Martinez T. R.** Reduction techniques for instance-based learning algorithms // *Mach. Learn.* 2000. Vol. 38, No. 3. P. 257–286.
16. **Аркадьев А. Г., Браверман Э. М.** Обучение машины распознаванию образов. М.: Наука, 1964. 112 с.
17. **Загоруйко Н. Г.** Когнитивный анализ данных. Новосибирск: Акад. изд-во ГЕО, 2013, 186 с.
18. **Borisova I. A., Dyubanov V. V., Kutnenko O. A., Zagoruiko N. G.** Use of the FRiS-function for taxonomy, attribute selection and decision rule construction // *Knowledge Processing and Data Analysis. Rev. Sel. Pap. 1st Int. Conf. KONT 2007 (Novosibirsk, Russia, Sept. 14–16, 2007); 1st Int. Conf. KPP 2007 (Darmstadt, Germany, Sept. 28–30, 2007)*. Heidelberg: Springer, 2011. P. 256–270. (Lect. Notes Comput. Sci.; Vol. 6581).
19. **Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.** Methods of recognition based on the function of rival similarity // *Pattern Recognit. Image Anal.* 2008. Vol. 18, No. 1. P. 1–6.
20. **Загоруйко Н. Г., Борисова И. А., Дюбанов В. В., Кутненко О. А.** Количественная мера компактности и сходства в конкурентном пространстве // *Сиб. журн. индустр. математики.* 2010. Т. 13, № 1. С. 59–71.
21. **Гэри М., Джонсон Д.** Вычислительные машины и труднорешаемые задачи. М.: Мир, 1982. 416 с.

- 22. Борисова И. А., Кутненко О. А.** Исправление диагностических ошибок в целевом признаке с помощью функции конкурентного сходства // Мат. биология и биоинформатика. 2018. Т. 13, № 1. С. 38–49.
- 23. Загоруйко Н. Г., Кутненко О. А.** Цензурирование обучающей выборки // Вестн. Томского гос. ун-та. Сер. Управление, вычисл. техника и информатика. 2013. № 22. С. 66–73.

Кутненко Ольга Андреевна
Плясунов Александр Владимирович

Статья поступила
10 июня 2020 г.
После доработки —
22 декабря 2020 г.
Принята к публикации
24 декабря 2020 г.

NP-HARDNESS OF SOME DATA CLEANING PROBLEM

O. A. Kutnenko^{1,2,a} and A. V. Plyasunov^{1,2,b}

¹ Sobolev Institute of Mathematics,
4 Acad. Koptug Avenue, 630090 Novosibirsk, Russia

² Novosibirsk State University,
2 Pirogov Street, 630090 Novosibirsk, Russia
E-mail: ^aolga@math.nsc.ru, ^bapljas@math.nsc.ru

Abstract. We prove the NP-hardness of the problem of outliers detection considered in this paper, to solving which a data analysis problem is reduced. As a quantitative assessment of the compactness of the image, the function of rival similarity (FRiS-function) is used, which evaluates the local similarity of objects with their closest neighbors. Illustr. 1, bibliogr. 23.

Keywords: NP-hardness, detecting outliers, image compactness, function of rival similarity.

REFERENCES

1. **J. W. Osborne**, *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data* (SAGE Publ., Los Angeles, 2013).
2. **A. Farcomeni** and **L. Greco**, *Robust Methods for Data Reduction* (CRC Press, New York, 2015).
3. **T. D. Waal**, **J. Pannekoek**, and **S. Scholtus**, *Handbook of Statistical Data Editing and Imputation* (Wiley, Hoboken, NJ, 2011).
4. **I. A. Borisova** and **O. A. Kutnenko**, Censoring misclassified sample items, *Mash. Obuch. Anal. Dannykh* **1** (11), 1632–1641 (2015) [Russian].
5. **C. C. Aggarwal**, *Data Mining* (Springer, Cham, 2015).
6. **H. Brighton** and **C. Mellish**, Advances in instance selection for instance-based learning algorithms, *Data Min. Knowl. Discov.* **6** (2), 153–172 (2002).

This research is carried out within the framework of the state contract of the Sobolev Institute of Mathematics (Projects 0314–2019–0015, 0314–2019–0014).

English version: Journal of Applied and Industrial Mathematics **15** (2), 285–291 (2021), DOI 10.1134/S1990478921020095.

7. **S. J. Delany, N. Segata, and B. Mac Namee**, Profiling instances in noise reduction, *Knowl.-B. Syst.* **31**, 28–40 (2012).
8. **B. Frenay and M. Verleysen**, Classification in the presence of label noise: A survey, *IEEE Trans. Neural Netw. Learn. Syst.* **25** (5), 845–869 (2014).
9. **N. Jankowski and M. Grochowski**, Comparison of instances selection algorithms I. Algorithms survey, in *Artificial Intelligence and Soft Computing — ICAISC 2004* (Proc. 7th Int. Conf., Zakopane, Poland, June 7–11, 2004) (Springer, Heidelberg, 2004), pp. 598–603 (Lect. Notes Comput. Sci., Vol. 3070).
10. **S. Massie, S. Craw, and N. Wiratunga**, When similar problems don't have similar solutions, in *Case-Based Reasoning Research and Development* (Proc. 7th Int. Conf., Belfast, NI, UK, Aug. 13–16, 2007) (Springer, Heidelberg, 2007), pp. 92–106 (Lect. Notes Comput. Sci., Vol. 4626).
11. **J. R. Quinlan**, Induction of decision trees, *Mach. Learn.* **1**, 81–106 (1986).
12. **N. Segata and E. Blanzieri**, Noise reduction for instance-based learning with a local maximal margin approach, *J. Intel. Inf. Syst.* **35** (2), 301–331 (2010).
13. **S.-H. Son and J.-Y. Kim**, Data reduction for instance-based learning using entropy-based partitioning, in *Computational Science and Its Applications — ICCSA 2006* (Proc. Int. Conf., Glasgow, UK, May 8–11, 2006), Pt. 3, (Springer, Heidelberg, 2006), pp. 590–599 (Lect. Notes Comput. Sci., Vol. 3982).
14. **C. M. Teng**, A comparison of noise handling techniques, in *Proc. 14th Int. Florida Artificial Intelligence Res. Soc. Conf., Key West, FL, USA, May 21–23, 2001* (AAAI Press, Menlo Park, CA, 2001), pp. 269–273.
15. **D. R. Wilson and T. R. Martinez**, Reduction techniques for instance-based learning algorithms, *Mach. Learn.* **38** (3), 257–286 (2000).
16. **A. G. Arkadyev and Eh. M. Braverman**, *Machine Learning to Pattern Recognition* (Nauka, Moscow, 1964) [Russian].
17. **N. G. Zagoruiko**, *Cognitive Data Analysis* (Akad. Izd. GEO, Novosibirsk, 2013) [Russian].
18. **I. A. Borisova, V. V. Dyubanov, O. A. Kutnenko, and N. G. Zagoruiko**, Use of the FRiS-function for taxonomy, attribute selection and decision rule construction, in *Knowledge Processing and Data Analysis* (Rev. Sel. Pap. 1st Int. Conf. KONT 2007, Novosibirsk, Russia, Sept. 14–16, 2007; 1st Int. Conf. KPP 2007, Darmstadt, Germany, Sept. 28–30, 2007) (Springer, Heidelberg, 2011), pp. 256–270 (Lect. Notes Comput. Sci., Vol. 6581).
19. **N. G. Zagoruiko, I. A. Borisova, V. V. Dyubanov, and O. A. Kutnenko**, Methods of recognition based on the function of rival similarity, *Pattern Recognit. Image Anal.* **18** (1), 1–6 (2008).
20. **N. G. Zagoruiko, I. A. Borisova, V. V. Dyubanov, and O. A. Kutnenko**, A quantitative measure of compactness and similarity in competitive space, *Sib. Zh. Ind. Mat.* **13** (1), 59–71 (2010) [Russian] [*Sib. J. Ind. Math.* **5** (1), 144–154 (2011)].
21. **M. R. Garey and D. S. Johnson**, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, San Francisco, 1979; Mir, Moscow, 1982 [Russian]).

-
- 22. I. A. Borisova and O. A. Kutnenko**, Correction of diagnostic errors in the target attribute with the function of rival similarity, *Mat. Biol. Bioinform.* **13** (1), 38–49 (2018) [Russian].
- 23. N. G. Zagoruiko and O. A. Kutnenko**, Censoring of a train dataset, *Vestn. Tomsk. Gos. Univ., Ser. Upr. Vychisl. Tekh. Inform.*, No. 22, 66–73 (2013) [Russian].

Olga A. Kutnenko
Aleksandr V. Plyasunov

Received June 10, 2020
Revised December 22, 2020
Accepted December 24, 2020