ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Январь–март 2022. Т. 29, № 1. С. 18–32

УДК 519.87+519.854 DOI 10.33048/daio.2022.29.713

ВЫЧИСЛИТЕЛЬНАЯ СЛОЖНОСТЬ ДВУХ ЗАДАЧ КОГНИТИВНОГО АНАЛИЗА ДАННЫХ

О. А. Кутненко[1, 2]

[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: olga@math.nsc.ru

Аннотация. Доказана NP-трудность в сильном смысле двух задач когнитивного анализа данных: задачи таксономии (кластеризации) — разбиения неклассифицированной выборки объектов на непересекающиеся подмножества — и задачи выбора подмножества типичных представителей классифицированной выборки, состоящей из объектов двух образов. Первую задачу можно рассматривать как частный случай второй задачи при условии, что один из образов состоит из одного объекта. Для количественной оценки качества множества выбранных типичных представителей выборки используется функция конкурентного сходства (FRiS-функция), с помощью которой оценивается сходство объекта с ближайшим типичным объектом. Ил. 1, библиогр. 18.

Ключевые слова: NP-трудность, таксономия (кластеризация), выбор типичных объектов (прототипов), функция конкурентного сходства.

Введение

Задача таксономии (кластеризации) — это задача разбиения неклассифицированной выборки объектов на непересекающиеся подмножества (кластеры) таким образом, чтобы каждый кластер состоял из близких (похожих) по некоторому критерию объектов, непохожих на объекты других кластеров. В анализе данных эта проблема относится к классу задач обучения без учителя. В качестве количественной оценки разбиения используется функция конкурентного сходства — FRiS-функция

Исследование выполнено в рамках государственного задания ИМ СО РАН (проект № FWNF–2022–0015).

(function of rival similarity) [1], с помощью которой оценивается сходство объекта с ближайшим типичным представителем выборки.

Проблема выбора типичных представителей выборки состоит в нахождении подмножества объектов, на котором достигается оптимальное значение функционала качества решения поставленной задачи. В литературе подобные объекты называются по-разному: столпами [2], прототипами [3], релевантными или типичными объектами [4], прецедентами или эталонными объектами [5] и т. п. Описание данных с помощью типичных представителей выборки (или прототипов) позволяет лучше понять структуру анализируемой выборки. Для решения прикладных задач используются различные эвристические методы для поиска прототипов [2, 5–13].

Задача выбора подмножества прототипов выборки, состоящей из объектов двух образов, типична в анализе данных. Предполагается, что разбиение на два класса задано и каждый класс может описываться несколькими прототипами. Задачу таксономии можно рассматривать как частный случай данной задачи при условии, что один из образов состоит из одного объекта. В [14] рассматривается задача выбора набора прототипов минимальной мощности для выборки, состоящей из объектов двух классов, при этом мощность одного из классов равна единице; критерием отбора прототипов является минимум частоты ошибок на всех остальных объектах выборки, в качестве расстояния берётся расстояние до ближайших прототипов классов. Показано, что в данной постановке задача выбора прототипов NP-трудна.

В рассматриваемой постановке в качестве прототипов выборки, состоящей из объектов двух классов, выбираются объекты, на которые максимально похожи объекты из того же класса и не похожи объекты другого класса. В качестве меры сходства объектов в некотором фиксированном признаковом пространстве используется функция конкурентного сходства, успешно применяемая в когнитивном анализе данных при решении различных прикладных задач [10, 15, 16]. В [17] показано, что данная задача NP-трудна. В настоящей работе предложено другое решение поставленной проблемы.

1. Таксономия (кластеризация) неклассифицированной выборки с помощью функции конкурентного сходства

1.1. FRiS-функция и качество кластеризации выборки. Особенностью задачи таксономии неклассифицированной выборки является то, что априори неизвестны как принадлежность объектов выборки к тому или иному образу (классу), так и число таких образов. Для решения задачи используется редуцированная функция конкурентного сходства [11]

$$F^*(z, \mathbf{A}) = \frac{\tau^* - \tau(z, \mathbf{A})}{\tau^* + \tau(z, \mathbf{A})}, \tag{1}$$

где τ^* — константа, интерпретируемая как расстояние от каждого объекта $z \in \mathbf{A}$ до виртуального образа (или образа-конкурента), все объекты которого являются прототипами, и расстояние от любого объекта выборки $\mathbf{A}$ до ближайшего объекта образа-конкурента равно τ^*; $\tau(z, \mathbf{A}) = \min\limits_{x \in \mathbf{A} \setminus \{z\}} \tau(z, x)$ — расстояние от объекта z до множества $\mathbf{A} \setminus \{z\}$.

Обозначим через $\mathbf{S_A}$ множество прототипов выборки $\mathbf{A}$. Для оценки качества этого множества используется усреднённая величина

$$H(\mathbf{A}, \mathbf{S_A}) = \frac{1}{|\mathbf{A}|} \sum_{z \in \mathbf{A} \setminus \mathbf{S_A}} F^*(z, \mathbf{S_A}). \tag{2}$$

Для решения рассматриваемой задачи требуется найти множество $\mathbf{S_A}$ прототипов выборки $\mathbf{A}$, на котором достигается максимум функционала H.

1.2. Постановка задачи. Дано множество объектов $\mathbf{A} = \{a_i\}_{i=\overline{1,M}}$. Задана матрица попарных расстояний между всеми объектами множества. В качестве расстояния от объекта до образа используется расстояние до ближайшего объекта образа. Требуется выбрать $1 \leqslant p \leqslant M$ прототипов из данного множества таким образом, чтобы сходство всех оставшихся объектов множества с прототипом своего класса было максимально, а с прототипами других классов — минимально, т. е. требуется найти множество $\mathbf{S_A} \subseteq \mathbf{A}$, $|\mathbf{S_A}| = p$, на котором достигается максимум $H(\mathbf{A}, \mathbf{S_A})$.

С учётом (1) запишем (2) в следующем виде:

$$H(\mathbf{A}, \mathbf{S_A}) = \frac{1}{|\mathbf{A}|} \sum_{a_i \in \mathbf{A} \setminus \mathbf{S_A}} \frac{\tau^* - \min\limits_{a_j \in \mathbf{S_A}} \tau(a_i, a_j)}{\tau^* + \min\limits_{a_j \in \mathbf{S_A}} \tau(a_i, a_j)} = \frac{1}{|\mathbf{A}|} \sum_{a_i \in \mathbf{A} \setminus \mathbf{S_A}} \max_{a_j \in \mathbf{S_A}} \frac{\tau^* - \tau(a_i, a_j)}{\tau^* + \tau(a_i, a_j)}.$$

Множество $\mathbf{A}$ можно записать как выборку $\{(a_i, y_i)\}_{i=\overline{1,M}}$, где $y_i = 1$, $i = \overline{1, M}$, — номинальный целевой признак.

Определим множество $\mathbf{T} = \{\mathbf{t} = \{t_i\}_{i=\overline{1,M}} \mid t_i \in \{0, 1\}\}$, где $t_i = 1$ означает, что i-й объект выборки принадлежит множеству прототипов: $\mathbf{S_A} = \{a_i \mid t_i = 1,\ i = \overline{1, M}\}$, соответственно $\mathbf{A} \setminus \mathbf{S_A} = \{a_i \mid t_i = 0,\ i = \overline{1, M}\}$ и $p = |\mathbf{S_A}| = \sum\limits_{i=1}^{M} t_i$.

Обозначим через $\tau_{ij} = \tau(a_i, a_j)$ расстояние между i-м и j-м объектами множества $\mathbf{A}$. Тогда для решения рассматриваемой смысловой задачи требуется решить экстремальную задачу — найти вектор $\mathbf{t}^*$, определяющий множество $\mathbf{S}^*_{\mathbf{A}}$, на котором достигается максимум $H(\mathbf{A}, \mathbf{S}_{\mathbf{A}})$:

$$\mathbf{t}^* = \arg\max_{\mathbf{t}\in\mathbf{T}} \sum_{i:\ t_i=0} \max_{j:\ t_j=1} \frac{\tau^* - \tau_{ij}}{\tau^* + \tau_{ij}}, \quad p = \sum_{i=1}^{M} t_i^*.$$

2. Вычислительная сложность задачи таксономии

Доказательство NP-трудности будет выполнено сведением известной NP-полной задачи о вершинном покрытии графа к задаче выбора подмножества, на котором значение функционала H максимально.

Задача ВП (вершинное покрытие) [18]**.** *Дан граф $G = (V, E)$ и положительное целое число $J \leqslant |V|$. Имеется ли в графе G вершинное покрытие не более чем из J элементов, т. е. такое подмножество $V' \subseteq V$, что $|V'| \leqslant J$ и для каждого ребра $\{u, v\} \in E$ хотя бы одна из вершин u или v принадлежит V'?*

Для доказательства потребуется

Утверждение 1. *Для любых $r \in \mathbb{Q}$ и $R \in \mathbb{Q}$, удовлетворяющих условию $0 < r < R$, существует r_1 такое, что*

$$0 < r < r_1 < R, \tag{3}$$

$$\frac{2(r_1 - r)}{r_1 + r} + \frac{r_1 - R}{r_1 + R} < 0. \tag{4}$$

Здесь и далее $\mathbb{Q}$ — множество рациональных чисел.

Доказательство. Имеем

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

Учитывая, что $(r + r_1)(r_1 + R) > 0$, найдём r_1, для которого выполняется (4):

$$2r_1^2 - 2rr_1 + 2r_1R - 2rR + r_1^2 + rr_1 - r_1R - rR$$
$$= 3r_1^2 - r_1r + r_1R - 3rR < 0.$$

Обозначим через $r^*_{1,2}$ корни квадратного уравнения

$$3r_1^2 - r_1(r - R) - 3rR = 0,$$
$$r^*_{1,2} = \frac{r - R \pm \sqrt{r^2 - 2rR + R^2 + 36rR}}{6},$$

т. е.

$$r_1 \in \left(\frac{r-R-\sqrt{r^2+R^2+34rR}}{6}, \frac{r-R+\sqrt{r^2+R^2+34rR}}{6}\right).$$

Покажем, что для любого $R > r$ выполняется

$$\frac{r-R+\sqrt{r^2+R^2+34rR}}{6} > r:$$

$$\sqrt{r^2+R^2+34rR} > 5r+R,$$
$$r^2+R^2+34rR > 25r^2+10rR+R^2,$$

что справедливо для $R > r$.

Нетрудно показать, что для любого $R > r$ также выполняется

$$\frac{r-R+\sqrt{r^2+R^2+34rR}}{6} < R.$$

Таким образом, с учётом (3) получим множество значений r_1:

$$r_1 \in \left(r, \frac{r-R+\sqrt{r^2+R^2+34rR}}{6}\right). \tag{5}$$

Показано, что для любых $r \in \mathbb{Q}$ и $R \in \mathbb{Q}$ таких, что $0 < r < R$, существует r_1, определяемое (5), удовлетворяющее требуемым условиям. Утверждение 1 доказано.

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

При этом выборка X_G строится по G за полиномиальное время и имеет полиномиальное количество объектов относительно $|V| + |E|$.

Доказательство. По заданному графу $G = (V, E)$ построим множество объектов, задающих образ $\mathbf{A}$: каждой вершине A графа поставим в соответствие объект A, каждому ребру (A, B) поставим в соответствие объект AB. Добавим в выборку X_G объект O виртуального образа-конкурента, т. е. $X_G = \mathbf{A} \cup \{O\}$.

Зададим матрицу попарных расстояний между объектами выборки следующим образом: положим расстояние равным r как между объектами, соответствующими смежным вершинам графа, так и между объектом, соответствующим ребру графа, и объектом, соответствующим вершине графа, являющейся одним из концов данного ребра; положим равными r_1 расстояния от всех объектов образа $\mathbf{A}$ до объекта O. Все неоговоренные выше расстояния положим равными R.

На r, r_1, R наложим условие $0 < r < r_1 < R < 2r$, при этом r_1 задаётся (5). Данная цепочка неравенств обеспечивает выполнение неравенства треугольника. Введённая таким образом матрица попарных расстояний удовлетворяет всем аксиомам расстояния.

Итак, выборка X_G построена (рис. 1).

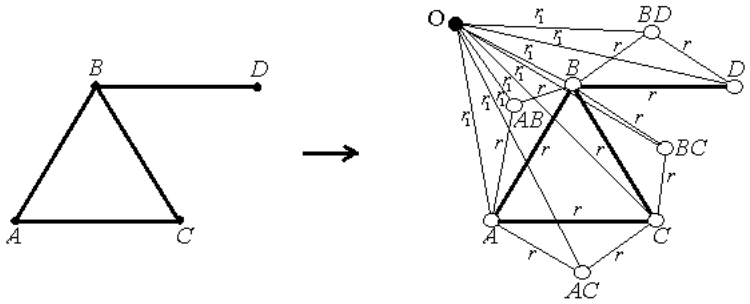

Рис. 1. Пример графа G, состоящего из четырёх вершин, и построение искусственной выборки X_G

Ставится задача: найти множество $\mathbf{S}^*_\mathbf{A}$, на котором достигается максимум функционала H, заданного (2). Далее входящий в $\mathbf{S}^*_\mathbf{A}$ объект будем называть *прототипом*. Заметим, что значение функционала зависит как от размерности множества $\mathbf{A} \setminus \mathbf{S}^*_\mathbf{A}$, так и от вклада каждого объекта из данного множества.

Рассмотрим произвольную группу объектов AB, A и B, соответствующих ребру (A, B) и его вершинам A и B, и проанализируем вклад каждого из этих объектов в H.

Пусть ни один из этих объектов не является прототипом, т. е. не входит в $\mathbf{S}^*_\mathbf{A}$. Вклад AB в данном случае всегда равен $\frac{r_1-R}{r_1+R}$. Вклад объекта A (B) равен $\frac{r_1-r}{r_1+r}$ или $\frac{r_1-R}{r_1+R}$. Тогда вклад данных объектов в H будет не больше $\frac{2(r_1-r)}{r_1+r} + \frac{r_1-R}{r_1+R}$. Равенство достигается, если для каждого объекта, соответствующего вершине, в $\mathbf{S}^*_\mathbf{A}$ входит объект, соответствующий смежной вершине, и/или объект, соответствующий ребру, инцидентному вершине. Согласно утверждению 1 при заданных условиях на r, r_1, R эта величина отрицательная.

Заметим, что достаточно наличия в $\mathbf{S}^*_\mathbf{A}$ хотя бы одного объекта из рассматриваемой группы объектов, чтобы вклад оставшихся объектов был положительным. Независимо от того, какой объект из рассматриваемой группы входит в $\mathbf{S}^*_\mathbf{A}$, вклад оставшихся двух объектов равен $\frac{2(r_1-r)}{r_1+r} > 0$. Однако при вхождении в $\mathbf{S}^*_\mathbf{A}$ объекта, соответствующего вершине, он

обеспечивает положительный вклад в H также объектам, соответствующим смежным вершинам, и объектам, соответствующим рёбрам, инцидентным данной вершине. Поэтому без потери качества функционала объекты, входящие в $\mathbf{S}^*_\mathbf{A}$, являются объектами типа A, каждому из которых соответствует вершина графа G.

Поскольку входящие в $\mathbf{S}^*_\mathbf{A}$ объекты дают нулевой вклад в H, сокращение числа объектов типа A, входящих в $\mathbf{S}^*_\mathbf{A}$, увеличит соответственно значение функционала H:

$$H(\mathbf{A}, \mathbf{S}^*_\mathbf{A}) = \frac{|V| - |V'| + |E|}{|V| + |E|} \cdot \frac{r_1 - r}{r_1 + r}.$$

Таким образом, максимальное значение функционала H достигается, когда в $\mathbf{S}^*_\mathbf{A}$ входит минимальное число $|V'| \leqslant |V|$ таких объектов, где $V' \subseteq V$.

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

Пусть данные вершины не образуют вершинного покрытия, т. е. существует объект AB, соответствующий ребру (A, B), такой, что в $\mathbf{S}^*_\mathbf{A}$ не входят объекты A и B. Тогда вклад AB в $H(\mathbf{A}, \mathbf{S}^*_\mathbf{A})$, равный $F(AB, \mathbf{A}^*) = \frac{r_1 - R}{r_1 + R} < 0$, уменьшает значение функционала H, что противоречит оптимальности $H(\mathbf{A}, \mathbf{S}^*_\mathbf{A})$.

Пусть образованное множеством вершин V' покрытие не является минимальным, т. е. существует множество $\mathbf{S}^{**}_\mathbf{A} \subset \mathbf{A}$, содержащее $|V'_1|$ объектов типа A, и $|V'_1| < |V'|$. С учётом того, что вклад в H объектов типа A, входящих в $\mathbf{S}^*_\mathbf{A}$, нулевой, получим $H(\mathbf{A}, \mathbf{S}^{**}_\mathbf{A}) > H(\mathbf{A}, \mathbf{S}^*_\mathbf{A})$, что противоречит предположению об оптимальности $H(\mathbf{A}, \mathbf{S}^*_\mathbf{A})$.

Таким образом доказано, что для достижения максимального значение H необходимо и достаточно вхождения в $\mathbf{S}^*_\mathbf{A}$ объектов типа A, соответствующих минимальному вершинному покрытию $V' \subseteq V$ графа G.

Показан способ построения минимального вершинного покрытия по известному множеству $\mathbf{S}^*_\mathbf{A}$, т. е. задача поиска минимального вершинного покрытия произвольного графа G сведена к задаче выбора из некоторой искусственной выборки X_G множества $\mathbf{S}^*_\mathbf{A}$, $|\mathbf{S}^*_\mathbf{A}| = |V'| = p$, прототипов выборки $\mathbf{A}$, на котором достигается максимум критерия H. При этом время построения выборки X_G и время преобразования множества $\mathbf{S}^*_\mathbf{A}$ в вершинное покрытие имеют порядок $O(|V|+|E|)$. Теорема 1 доказана.

Замечание 1. Из доказательства теоремы 1 следует, что задача таксономии NP-трудна в сильном смысле.

3. Задача выбора прототипов выборки, представленной объектами двух классов

3.1. Критерий качества системы прототипов. В качестве прототипов классифицированной выборки, состоящей из $M = |\mathbf{A} \cup \mathbf{B}|$ объектов двух образов, выбираются объекты, на которые максимально похожи объекты из того же класса и не похожи объекты другого класса. Задана матрица попарных расстояний между всеми объектами множества. В качестве расстояния от объекта до образа используется расстояние до ближайшего объекта образа. В качестве меры сходства объектов в некотором фиксированном признаковом пространстве используется функция конкурентного сходства.

Обозначим через $\mathbf{S_A}$ и $\mathbf{S_B}$ множества прототипов образа $\mathbf{A}$ и образа $\mathbf{B}$ соответственно. Для оценки качества выбранных прототипов используется усреднённая величина

$$H(\mathbf{A}, \mathbf{S_A}, \mathbf{B}, \mathbf{S_B}) = \frac{1}{|\mathbf{A} \cup \mathbf{B}|} \Bigg(\sum_{x \in \mathbf{A} \setminus \mathbf{S_A}} \frac{\tau(x, \mathbf{S_B}) - \tau(x, \mathbf{S_A})}{\tau(x, \mathbf{S_B}) + \tau(x, \mathbf{S_A})} + \sum_{x \in \mathbf{B} \setminus \mathbf{S_B}} \frac{\tau(x, \mathbf{S_A}) - \tau(x, \mathbf{S_B})}{\tau(x, \mathbf{S_A}) + \tau(x, \mathbf{S_B})} \Bigg). \quad (6)$$

Для решения рассматриваемой задачи требуется найти множества $\mathbf{S_A}$ и $\mathbf{S_B}$ прототипов выборки, на которых достигается максимум функционала H.

3.2. Постановка задачи. Дана выборка, состоящая из объектов двух образов $\mathbf{A}$ и $\mathbf{B}$, $\mathbf{A} \cap \mathbf{B} = \varnothing$, $|\mathbf{A} \cup \mathbf{B}| = M$, и задана матрица попарных расстояний между всеми объектами множества. Требуется выбрать $2 \leqslant p \leqslant M$ типичных объектов выборки таким образом, чтобы сходство всех оставшихся объектов множества с прототипом своего класса было максимально, а с прототипами других классов минимально:

$$H(\mathbf{A}, \mathbf{S_A}, \mathbf{B}, \mathbf{S_B}) \to \max_{\substack{\mathbf{S_A} \subseteq \mathbf{A},\, \mathbf{S_A} \neq \varnothing, \\ \mathbf{S_B} \subseteq \mathbf{B},\, \mathbf{S_B} \neq \varnothing, \\ |\mathbf{S_A}| + |\mathbf{S_B}| = p}}.$$

Перепишем (6) в следующем виде:

$$H(\mathbf{A}, \mathbf{S_A}, \mathbf{B}, \mathbf{S_B}) = \frac{1}{|\mathbf{A} \cup \mathbf{B}|} \times \Bigg(\sum_{a_i \in \mathbf{A} \setminus \mathbf{S_A}} \frac{\min\limits_{b_j \in \mathbf{S_B}} \tau(a_i, b_j) - \min\limits_{a_k \in \mathbf{S_A}} \tau(a_i, a_k)}{\min\limits_{b_j \in \mathbf{S_B}} \tau(a_i, b_j) + \min\limits_{a_k \in \mathbf{S_A}} \tau(a_i, a_k)}$$

$$+\sum_{b_i\in\mathbf{B}\setminus\mathbf{S_B}}\frac{\min\limits_{a_j\in\mathbf{S_A}}\tau(b_i,a_j)-\min\limits_{b_k\in\mathbf{S_B}}\tau(b_i,b_k)}{\min\limits_{a_j\in\mathbf{S_A}}\tau(b_i,a_j)+\min\limits_{b_k\in\mathbf{S_B}}\tau(b_i,b_k)}\Big).$$

Множество $\mathbf{A}\cup\mathbf{B}$ представим как выборку $\{(x_i,y_i)\}_{i=\overline{1,M}}$, в которой $y_i\in\{-2,2\}$, $i=\overline{1,M}$, — номинальный целевой признак. Тогда образ $\mathbf{A}$ можно записать как множество объектов $\{x_i \mid y_i=-2,\ i=\overline{1,M}\}$, $\mathbf{B}$ — как $\{x_i \mid y_i=2,\ i=\overline{1,M}\}$.

Для заданного вектора $\mathbf{y}$ определим множество

$$\mathbf{T}=\{\mathbf{t}=\{t_i\}_{i=\overline{1,M}} \mid t_i\in\{-2,-1,1,2\};\\ t_i\in\{-1,-2\},\text{ если } y_i=-2;\ t_i\in\{1,2\},\text{ если } y_i=2\},$$

где $t_i=-1$ означает, что i-й объект выборки принадлежит множеству прототипов образа $\mathbf{A}$: $\mathbf{S_A}=\{x_i \mid t_i=-1,\ i=\overline{1,M}\}$, соответственно $\mathbf{S_B}=\{x_i \mid t_i=1,\ i=\overline{1,M}\}$.

Обозначим через τ_{ij} расстояние между i-м и j-м объектами множества $\mathbf{A}\cup\mathbf{B}$. Таким образом, для решения рассматриваемой смысловой задачи требуется решить экстремальную задачу — найти вектор $\mathbf{t}^*$, определяющий множества $\mathbf{S}^*_\mathbf{A}$ и $\mathbf{S}^*_\mathbf{B}$, на которых достигается максимум функционала $H(\mathbf{A},\mathbf{S_A},\mathbf{B},\mathbf{S_B})$:

$$\mathbf{t}^*=\arg\max_{\mathbf{t}\in\mathbf{T}}\Big(\sum_{i:\ t_i=-2}\frac{\min\limits_{j:\ t_j=1}\tau_{ij}-\min\limits_{k:\ t_k=-1}\tau_{ik}}{\min\limits_{j:\ t_j=1}\tau_{ij}+\min\limits_{k:\ t_k=-1}\tau_{ik}}\\ +\sum_{i:\ t_i=2}\frac{\min\limits_{j:\ t_j=-1}\tau_{ij}-\min\limits_{k:\ t_k=1}\tau_{ik}}{\min\limits_{j:\ t_j=-1}\tau_{ij}+\min\limits_{k:\ t_k=1}\tau_{ik}}\Big),$$

$$p=|\mathbf{S}^*_\mathbf{A}|+|\mathbf{S}^*_\mathbf{B}|=\sum_{i:\ |t_i|=1}|t_i|.$$

4. Вычислительная сложность задачи выбора прототипов выборки, представленной объектами двух классов

NP-трудность данной задачи следует из теоремы 1.

Следствие 1. *Задача поиска множества прототипов выборки, представленной объектами двух классов,* NP*-трудна.*

Доказательство. По данным графам $G_1=(V_1,E_1)$ и $G_2=(V_2,E_2)$ построим выборку $X_{G_1G_2}$.

Каждой вершине A_1 графа G_1 поставим в соответствие объект A_1, каждому ребру (A_1,B_1) — объект A_1B_1. Построенные объекты задают

множество $\mathbf{A}$. Аналогично каждой вершине A_2 графа G_2 поставим в соответствие объект A_2, каждому ребру (A_2, B_2) — объект A_2B_2. Данные объекты задают множество $\mathbf{B}$, т. е. $X_{G_1G_2} = \mathbf{A} \cup \mathbf{B}$.

Зададим матрицу попарных расстояний между объектами выборки следующим образом. Для объектов множества $\mathbf{A}$, соответствующих графу $G_1 = (V_1, E_1)$, положим расстояние равным r как между объектами, соответствующими смежным вершинам данного графа, так и между объектом, соответствующим ребру этого графа, и объектом, соответствующим вершине графа, являющейся одним из концов данного ребра. Аналогично зададим расстояния между объектами множества $\mathbf{B}$, соответствующими графу $G_2 = (V_2, E_2)$. Положим равными r_1 расстояния от всех объектов множества $\mathbf{A}$ до всех объектов множества $\mathbf{B}$. Все неоговоренные выше расстояния положим равными R.

На r, r_1, R наложим условие $0 < r < r_1 < R < 2r$, при этом r_1 задаётся (5). Данная цепочка неравенств обеспечивает выполнение неравенства треугольника. Заданная таким образом матрица попарных расстояний удовлетворяет всем аксиомам расстояния. Итак, выборка $X_{G_1G_2}$ построена.

Ставится задача: найти множества $\mathbf{S}^*_{\mathbf{A}}$ и $\mathbf{S}^*_{\mathbf{B}}$, на которых достигается максимум функционала H, заданного (6). Из теоремы 1 и (2) следует, что максимум данного функционала достигается при $\mathbf{S}^*_{\mathbf{A}}$ и $\mathbf{S}^*_{\mathbf{B}}$, состоящих из объектов, соответствующих минимальным вершинным покрытиям V_1' и V_2' графов G_1 и G_2:

$$H(\mathbf{A}, \mathbf{S}^*_{\mathbf{A}}, \mathbf{B}, \mathbf{S}^*_{\mathbf{B}}) = \frac{|V_1| - |V_1'| + |E_1| + |V_2| - |V_2'| + |E_2|}{|V_1| + |E_1| + |V_2| + |E_2|} \cdot \frac{r_1 - r}{r_1 + r},$$

где $|V_1'| = |\mathbf{S}^*_{\mathbf{A}}|$, $|V_2'| = |\mathbf{S}^*_{\mathbf{B}}|$ — мощности минимальных вершинных покрытий графов G_1 и G_2 соответственно.

При этом время построения выборки $X_{G_1G_2}$ и время преобразования множеств $\mathbf{S}^*_{\mathbf{A}}$ и $\mathbf{S}^*_{\mathbf{B}}$ в вершинные покрытия соответствующих графов имеют порядок $O(|V_1| + |E_1| + |V_2| + |E_2|)$. Следствие 1 доказано.

Замечание 2. Из доказательства следствия 1 следует, что задача выбора прототипов выборки, представленной объектами двух классов, NP-трудна в сильном смысле.

NP-трудность рассмотренных задач когнитивного анализа данных обосновывает применение различных эвристических алгоритмов для решения задач выбора прототипов классифицированных, неклассифицированных и смешанных выборок, в которых для количественной оценки качества выбранных типичных объектов образов используется функция конкурентного сходства [2, 10–12].

Заключение

Показана NP-трудность в сильном смысле экстремальных задач поиска множества типичных объектов анализируемых данных, на котором достигается согласно заданным критериям максимум оценки качества выбранных прототипов. Таким образом, показана труднорешаемость соответствующих проблем анализа данных. Отметим, что алгоритмов поиска типичных объектов данных с гарантированными оценками точности для решения рассмотренных задач в настоящее время неизвестно.

ЛИТЕРАТУРА

1. **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. V. 18, No. 1. P. 1–6.
2. **Борисова И. А., Дюбанов В. В., Загоруйко Н. Г., Кутненко О. А.** Сходство и компактность // Докл. 14-й Всерос. конф. «Математические методы распознавания образов» (Суздаль, Россия, 21–25 сентября 2009 г.). М.: Макс Пресс, 2009. С. 89–92.
3. **Burges C. J. C.** A tutorial on support vector machines for pattern recognition // Data Mining Knowl. Discov. 1998. V. 2, No. 2. P. 121–167.
4. **Tipping M. E.** The relevance vector machine // Advances in Neural Information Processing Systems 12. Proc. 1999 Conf. (Denver, CO, USA, Nov. 29–Dec. 4, 1999). Cambridge, MA: MIT Press, 2000. P. 652–658.
5. **Загоруйко Н. Г.** Прикладные методы анализа данных и знаний. Новосибирск: Изд-во Ин-та математики 1999. 268 с.
6. **Воронцов К. В., Колосков А. О.** Профили компактности и выделение опорных объектов в метрических алгоритмах классификации // Искусств. интеллект. 2006. № 2. С. 30–33.
7. **Иванов М. Н., Воронцов К. В.** Отбор эталонов, основанный на минимизации функционала полного скользящего контроля // Докл. 14-й Всерос. конф. «Математические методы распознавания образов» (Суздаль, Россия, 21–25 сентября 2009 г.). М.: Макс Пресс, 2009. С. 119–122.
8. **Bermejo S., Cabestany J.** Learning with nearest neighbor classifiers // Neural Proc. Lett. 2001. V. 13, No. 2. P. 159–181.
9. **Вапник В. Н.** Задача обучения распознаванию образов. М.: Знание, 1971. 60 с.
10. **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.
11. **Борисова И. А.** Алгоритм таксономии FRiS-Tax // Науч. вестн. НГТУ. 2007. № 3. С. 3–12.
12. **Борисова И. А., Загоруйко Н. Г.** Алгоритм FRiS-TDR для решения обобщённой задачи таксономии и распознавания // Матер. 2-й Всерос. конф. «Знания — Онтологии — Теории» (Новосибирск, Россия, 20–22 октября 2009 г.). Т. 1. Новосибирск: ИМ СО РАН, 2009. С. 93–102.

13. MacQueen J. B. Some methods for classification and analysis of multivariate observations // Proc. 5th Berkeley Symp. Math. Stat. Probab. (Berkeley, USA, June 21–July 18, 1965; Dec. 27, 1965–Jan. 7, 1966) V. 1. Berkeley: Univ. California Press, 1967. P. 281–297.

14. Zukhba A. V. NP-completeness of the problem of prototype selection in the nearest neighbor method // Pattern Recognit. Image Anal. 2010. V. 20, No. 4. P. 484–494.

15. 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.; V. 6581).

16. Загоруйко Н. Г., Борисова И. А., Кутненко О. А., Дюбанов В. В. Построение сжатого описания данных с использованием функции конкурентного сходства // Сиб. журн. индустр. математики. 2013. Т. 16, № 1. С. 29–41.

17. Борисова И. А. Вычислительная сложность задачи выбора типичных представителей в 2-разбиении конечного множества точек метрического пространства // Дискрет. анализ и исслед. операций. 2020. Т. 27, № 2. С. 5–16.

18. Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи. М.: Мир, 1982. 416 с.

Кутненко Ольга Андреевна

Статья поступила
26 апреля 2021 г.
После доработки —
2 декабря 2021 г.
Принята к публикации
3 декабря 2021 г.

DISKRETNYI ANALIZ I ISSLEDOVANIE OPERATSII
/DISCRETE ANALYSIS AND OPERATIONS RESEARCH/
January–March 2022. Vol. 29, No. 1. P. 18–32

UDC 519.87+519.854 DOI 10.33048/daio.2022.29.713

COMPUTATIONAL COMPLEXITY OF TWO PROBLEMS OF COGNITIVE DATA ANALYSIS

O. A. Kutnenko[1,2]

[1] Sobolev Institute of Mathematics,
4 Acad. Koptyug Avenue, 630090 Novosibirsk, Russia
[2] Novosibirsk State University,
2 Pirogov Street, 630090 Novosibirsk, Russia
E-mail: olga@math.nsc.ru

Abstract. The NP-hardness in the strong sense is proved for two problems of cognitive data analysis. One of them is the problem of taxonomy (clustering), i. e. splitting an unclassified sample of objects into disjoint subsets. The other is the problem of sampling a subset of typical representatives of a classified sample which consists of objects of two images. The first problem can be considered as a special case of the second problem, provided that one of the images consists of one object. To obtain a quantitative quality estimate for the set of selected typical representatives of the sample, the function of rival similarity (FRiS function) is used, which assesses the similarity of an object with the closest typical object. Illustr. 1, bibliogr. 18.

Keywords: NP-hardness, taxonomy (clustering), typical object (prototypes) selection, function of rival similarity.

REFERENCES

1. **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).
2. **I. A. Borisova, V. V. Dyubanov, N. G. Zagoruiko,** and **O. A. Kutnenko,** Similarity and compactness, in *Proc. 14th All-Russian Conf. "Mathematical Methods for Pattern Recognition", Suzdal, Russia, Sept. 21–25, 2009* (Maks Press, Moscow, 2009), pp. 89–92 [Russian].

The study is carried out within the framework of the state contract of the Sobolev Institute of Mathematics (Project FWNF–2022–0015).

English version: Journal of Applied and Industrial Mathematics **16** (1) (2022).

3. **C. J. C. Burges,** A tutorial on support vector machines for pattern recognition, *Data Mining Knowl. Discov.* **2** (2), 121–167 (1998).
4. **M. E. Tipping,** The relevance vector machine, in *Advances in Neural Information Processing Systems 12* (Proc. 1999 Conf., Denver, CO, USA, Nov. 29–Dec. 4, 1999) (MIT Press, Cambridge, MA, 2000), pp. 652–658.
5. **N. G. Zagoruiko,** *Applied Methods of Data and Knowledge Analysis* (Izd. Inst. Mat., Novosibirsk, 1999) [Russian].
6. **K. V. Vorontsov** and **A. O. Koloskov,** Compactness profiles and prototype object selection in metric classification algorithms, *Iskusstv. Intell.,* No. 2, 30–33 (2006) [Russian].
7. **M. N. Ivanov** and **K. V. Vorontsov,** Prototypes selection based on minimization of a complete follow-control functional, in *Proc. 14th All-Russian Conf. "Mathematical Methods for Pattern Recognition", Suzdal, Russia, Sept. 21–25, 2009* (Maks Press, Moscow, 2009), pp. 119-–122 [Russian].
8. **S. Bermejo** and **J. Cabestany,** Learning with nearest neighbor classifiers, *Neural Proc. Lett.* **13** (2), 159-–181 (2001).
9. **V. N. Vapnik,** *The Task of Learning Pattern Recognition* (Znanie, Moscow, 1971) [Russian].
10. **N. G. Zagoruiko, I. A. Borisova, V. V. Dyubanov,** and **O. A. Kutnenko,** A quantitative measure of compactness and similarity in a competitive space, *Sib. J. Ind. Math.* **13** (1), 59–71 (2010) [Russian] [*J. Appl. Ind. Math.* **5** (1), 144–154 (2011)].
11. **I. A. Borisova,** A taxonomy algorithm FRiS-Tax, *Nauchn. Vestn. NGTU,* No. 3, 3–12 (2007) [Russian].
12. **I. A. Borisova** and **N. G. Zagoruiko,** A FRiS-TDR algorithm for solving a generalized taxonomy and recognition problem, in *Proc. 2nd All-Russian Conf. "Knowledge–Ontology–Theory", Novosibirsk, Russia, Oct. 22–24, 2009,* Vol. 1 (Inst. Mat., Novosibirsk, 2009), pp. 93–102 [Russian].
13. **J. B. MacQueen,** Some methods for classification and analysis of multivariate observations, in *Proc. 5th Berkley Symp. Math. Stat. Prob., Berkley, USA, June 21–July 18, 1965; Dec. 27, 1965–Jan. 7, 1966,* Vol. 1 (Univ. California Press, Berkley, 1967), pp. 281—297.
14. **A. V. Zukhba,** NP-completeness of the problem of prototype selection in the nearest neighbor method, *Pattern Recognit. Image Anal.* **20** (4), 484–494 (2010).
15. **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).
16. **N. G. Zagoruiko, I. A. Borisova, O. A. Kutnenko,** and **V. V. Dyubanov,** A construction of a compressed description of data using a function of rival similarity, *Sib. J. Ind. Math.* **16** (1), 29–41 (2013) [Russian] [*J. Appl. Ind. Math.* **7** (2), 275–286 (2013)].

17. I. A. Borisova, Computational complexity of the problem of choosing typical representatives in a 2-clustering of a finite set of points in a metric space, *Discrete Anal. Oper. Res.* **27** (2), 5–16 (2020) [Russian] [*J. Appl. Ind. Math.* **14** (2), 242–248 (2020)].

18. 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]).

Olga A. Kutnenko

Received April 26, 2021
Revised December 2, 2021
Accepted December 3, 2021