Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)

Метод разбиения категориальных данных на основе агломеративного подхода иерархической кластеризации

Студент: Ковалец Кирилл Эдуардович ИУ7-83Б
Научный руководитель: Новик Наталья Владимировна

МОСКВА, 2023

Актуальность

Недостаток иерархической кластеризации:

- попарное объединение обособленных равноудаленных кластеров.

Устранение этого недостатка позволит:

- улучшить компактность кластеров;
- улучшить разделенность групп.

Цель и задачи

Цель работы: разработка метода разбиения категориальных данных на основе агломеративного подхода иерархической кластеризации.

Задачи:

- провести аналитический обзор известных методов кластеризации данных;
- формализировать поставленную задачу;
- выбрать меры расстояний и критерии связи кластеров;
- разработать гибридный метод разбиения данных;
- разработать алгоритмы и структуру ПО;
- провести исследования разработанного метода.

Выбор метода разбиения для уточнения элементов в кластерах

Метод	Числовые признаки	Категориальные признаки	Тип алгоритма	Входные данные	Выходные данные
Иерархический	+	+	Иерархический	Матрица несходства	Бинарное дерево кластеров
K-средних	+	–	Центроидный	Массив объектов, число кластеров	Кластеры с центрами и номерами объектов
K-режимов	–	+	Центроидный	Массив объектов, число кластеров	Кластеры с центрами и номерами объектов
K-прототипов	+	+	Центроидный	Массив объектов, число кластеров	Кластеры с центрами и номерами объектов
C-средних	+	–	Центроидный	Массив объектов, число кластеров	Центры кластеров, матрица принадлежности
DBSCAN	+	+	На основе плотности	Матрица несходства, радиус, кол-во соседей	Набор кластеров и точки, не вошедшие ни в одну группу
МПД	+	+	На основе графов	Матрица несходства, пороговое значение расстояний	Древовидная структура кластеров

Метод разбиения на основе иерархической кластеризации

Выбор критерия связи кластеров

Критерий связи	**Расстояние между двумя кластерами**
Одиночная связь	кратчайшее расстояние между двумя точками в каждом кластере
Полная связь	самое длинное расстояние между двумя точками в каждом кластере
Средняя связь	среднее расстояние между каждой точкой в одном кластере до каждой точки в другом кластере

Сравнение мер расстояний

- К1 — возможность находить расстояние между числовыми данными;
- К2 — возможность находить расстояние между категориальными данными.

Мера расстояния	К1	К2
Евклидово расстояние	+	–
Квадрат Евклидово расстояние	+	–
Расстояние городских кварталов	+	–
Расстояние Чебышева	+	–
Расстояние Минковского	+	–
Степенное расстояние	+	–
Расстояние Хэмминга	–	+
Расстояние Говера	+	+

Вычисление расстояния Говера

Для двух векторов $\vec{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})$ и $\vec{x_j} = (x_{j1}, x_{j2}, \ldots, x_{jp})$ формула расстояния Говера:

$$d(\vec{x_i}, \vec{x_j}) = \frac{\sum_{k=1}^{p} s_{ijk}}{p},$$

где p — количество признаков в векторах;

s_{ijk} — оценка сходства двух признаков.

- для числовых переменных:

$$s_{ijk} = \frac{|x_{ik} - x_{jk}|}{R_k},$$

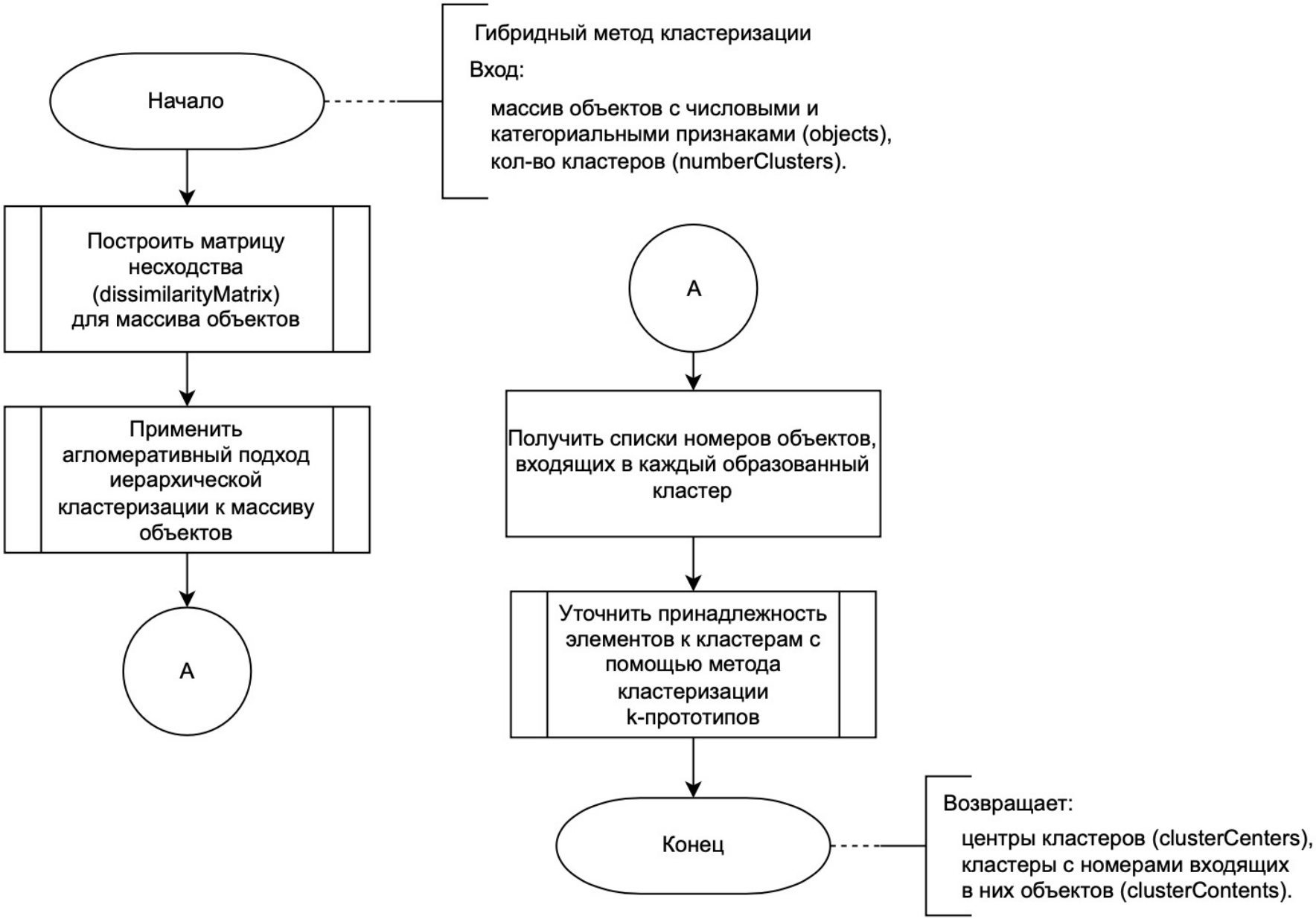
где R_k — диапазон значений признака k;

x_{ik}, x_{jk} — значения k-ых признаков векторов $\vec{x_i}$, $\vec{x_j}$ соответственно.

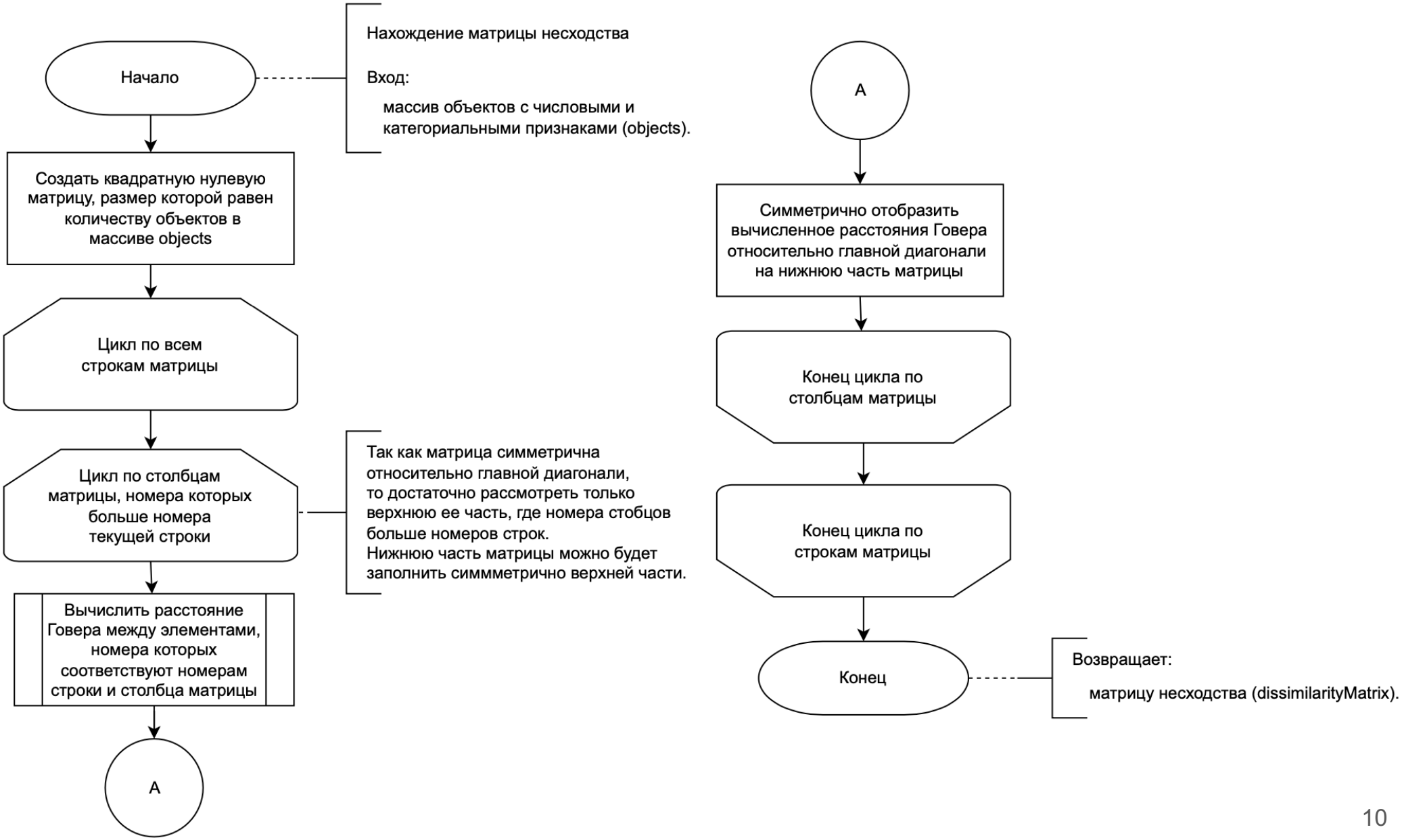
- для категориальных переменных:

$$s_{ijk} = \begin{cases} 0, & \text{если } x_{ik} = x_{jk}; \\ 1, & \text{иначе.} \end{cases}$$

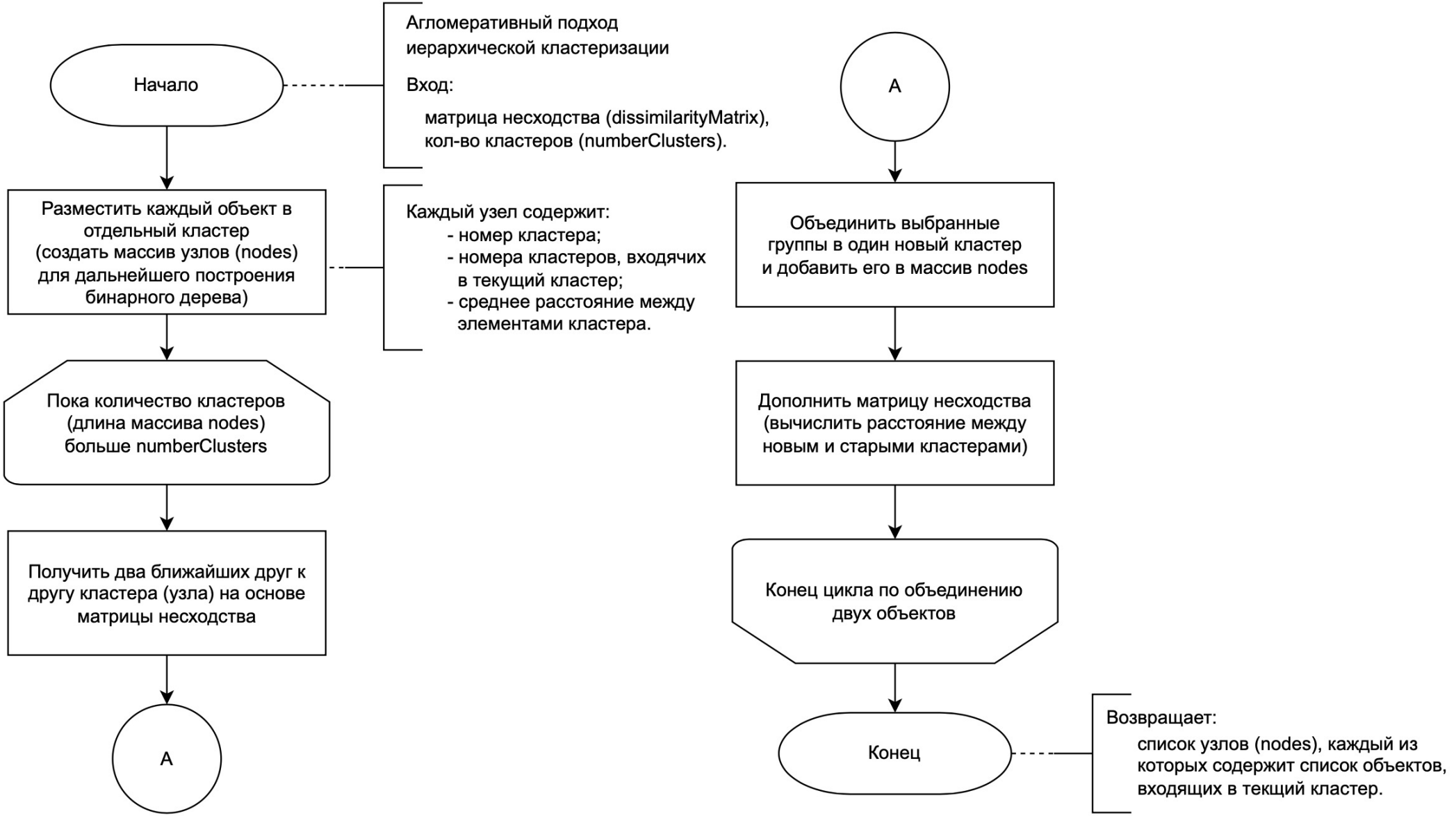
Гибридный метод кластеризации



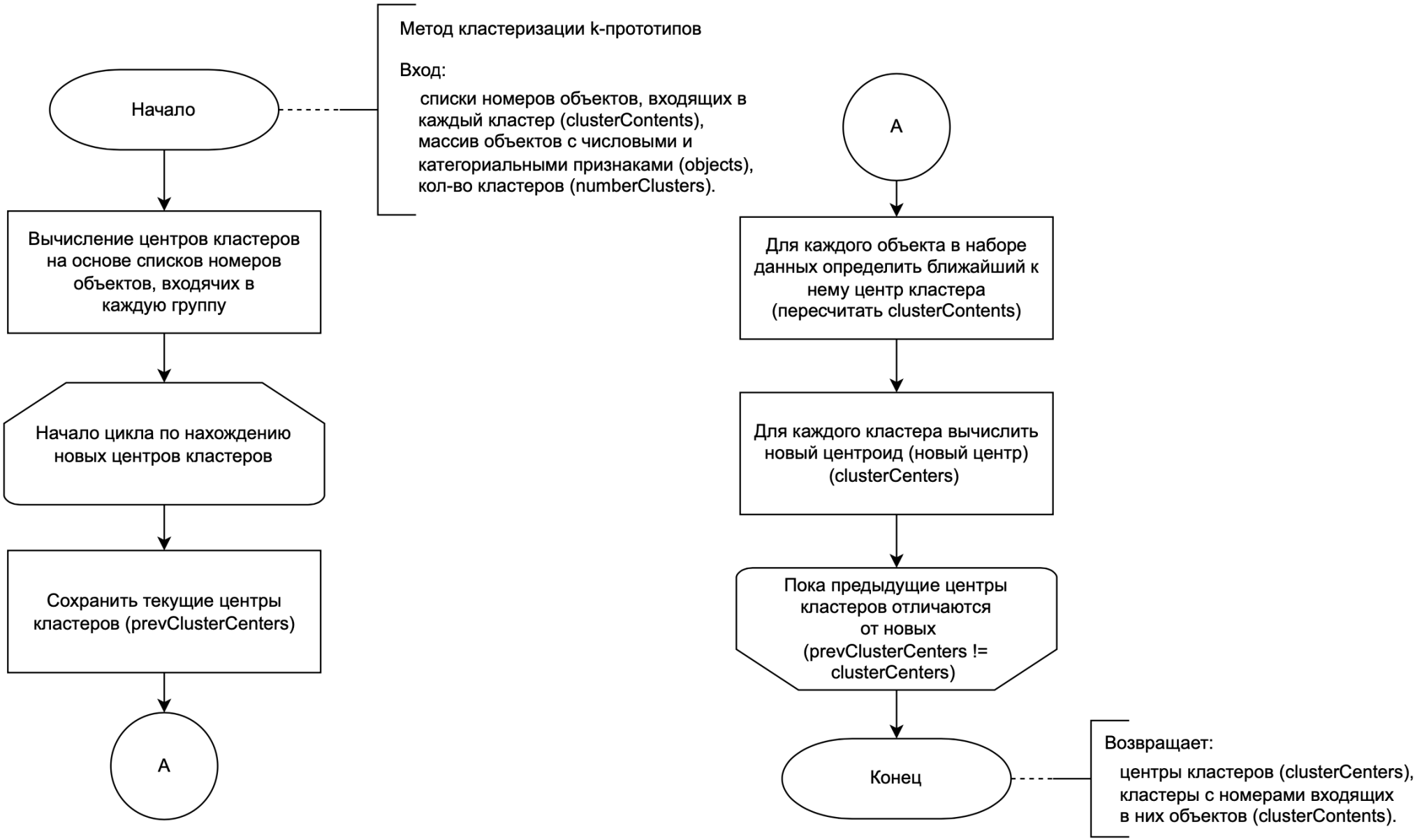
Нахождение матрицы несходства



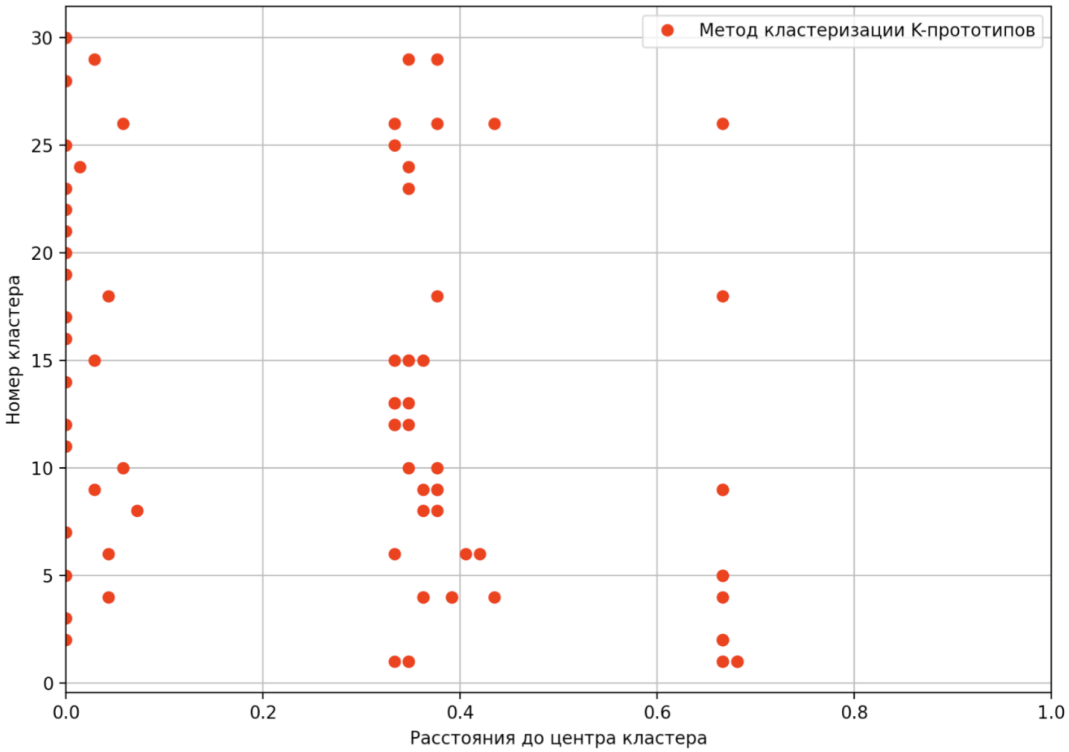
Иерархическая часть гибридного метода



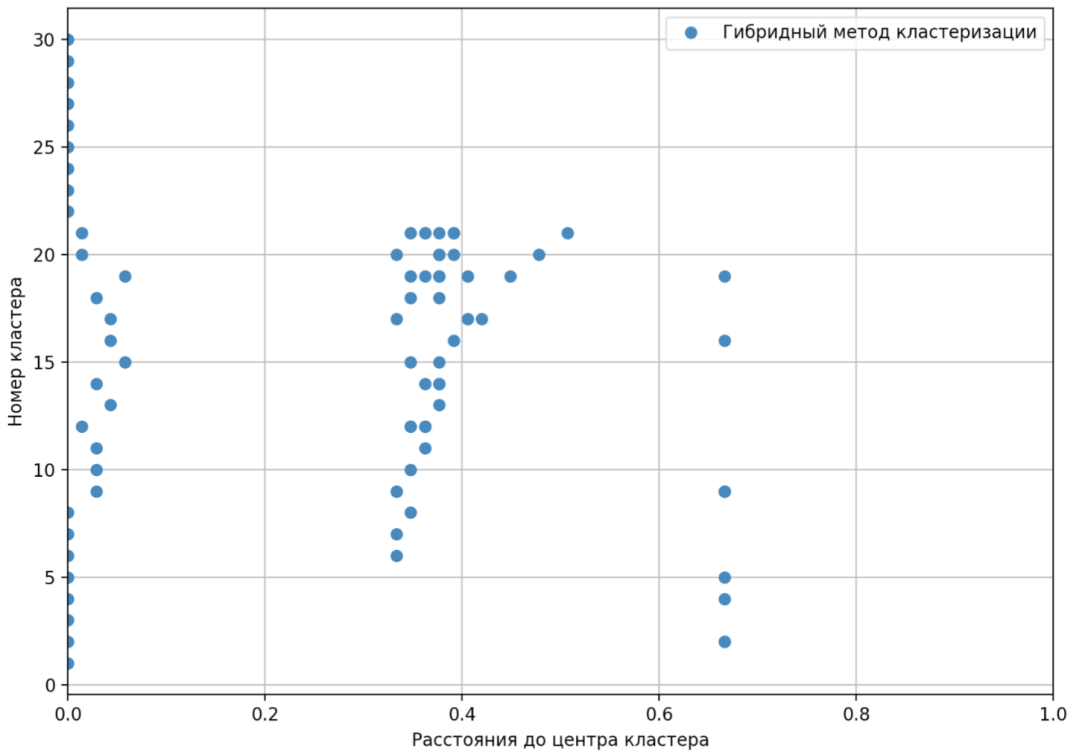
Центроидная часть гибридного метода



Результаты разбиения для 30 кластеров

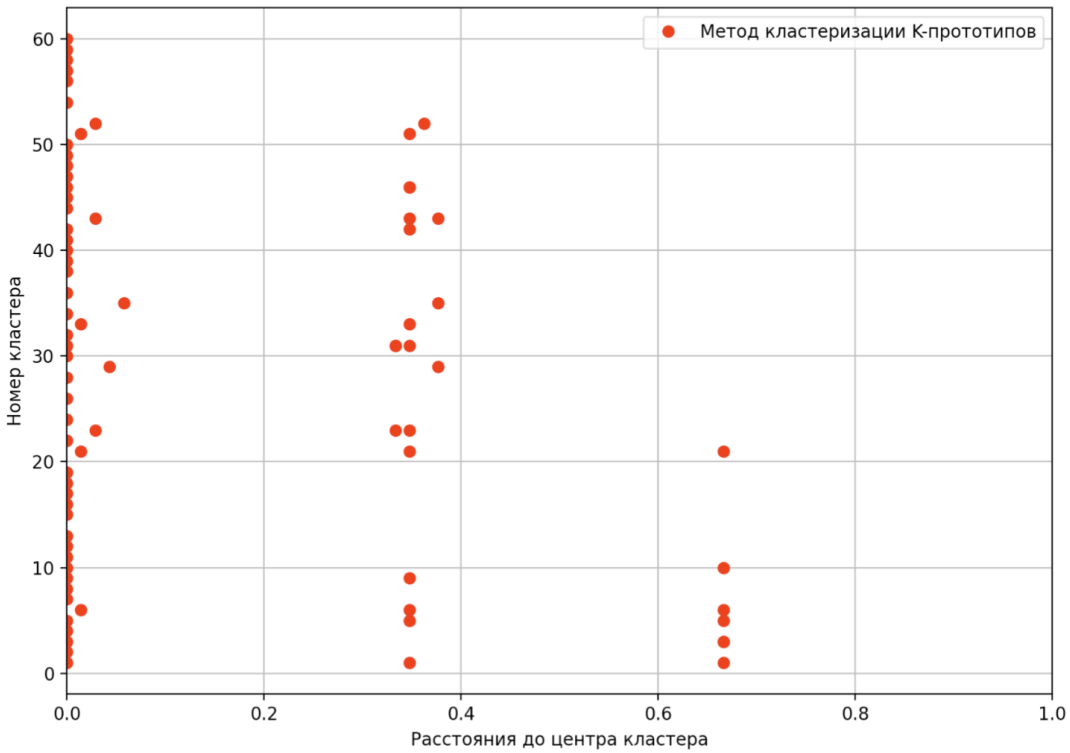


Метод K-прототипов

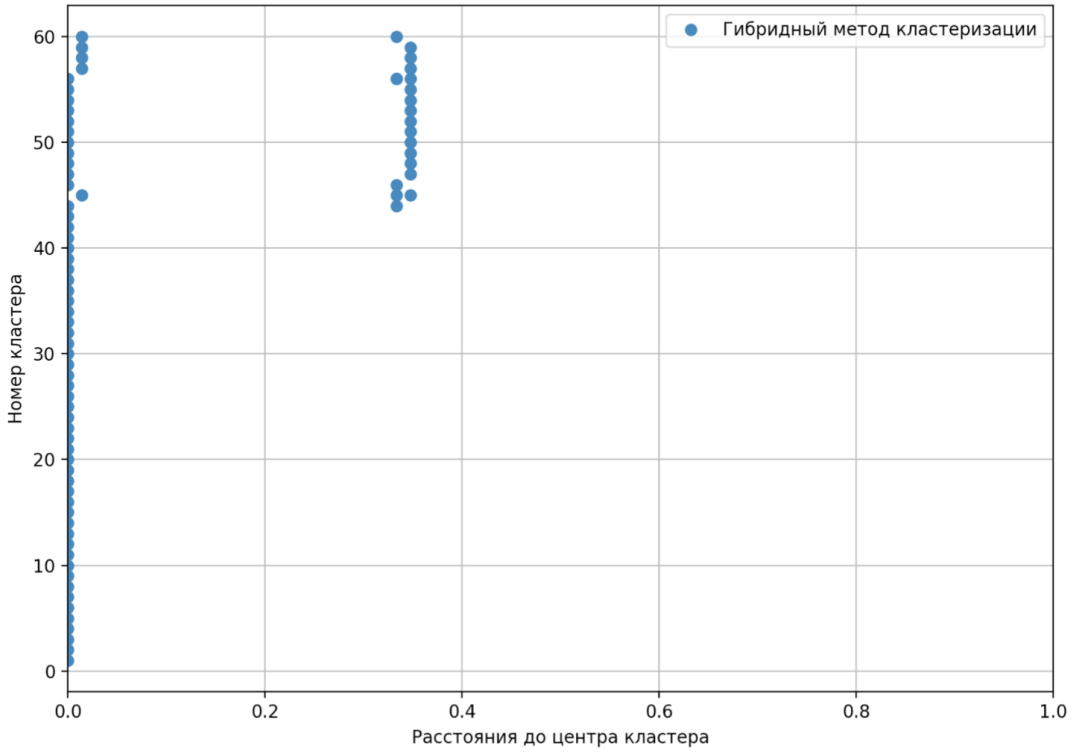


Гибридный метод

Результаты разбиения для 60 кластеров



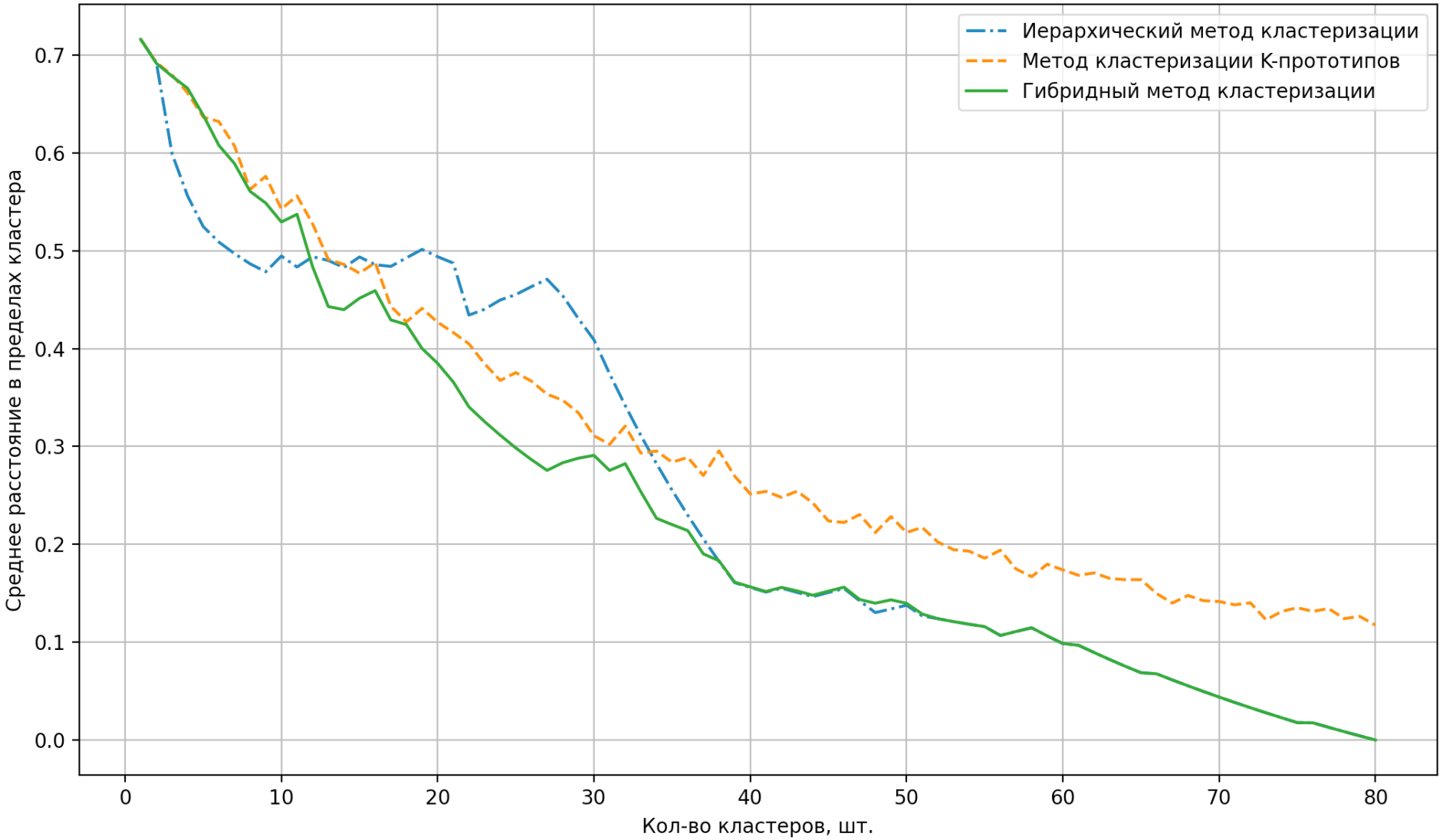
Метод K-прототипов



Гибридный метод

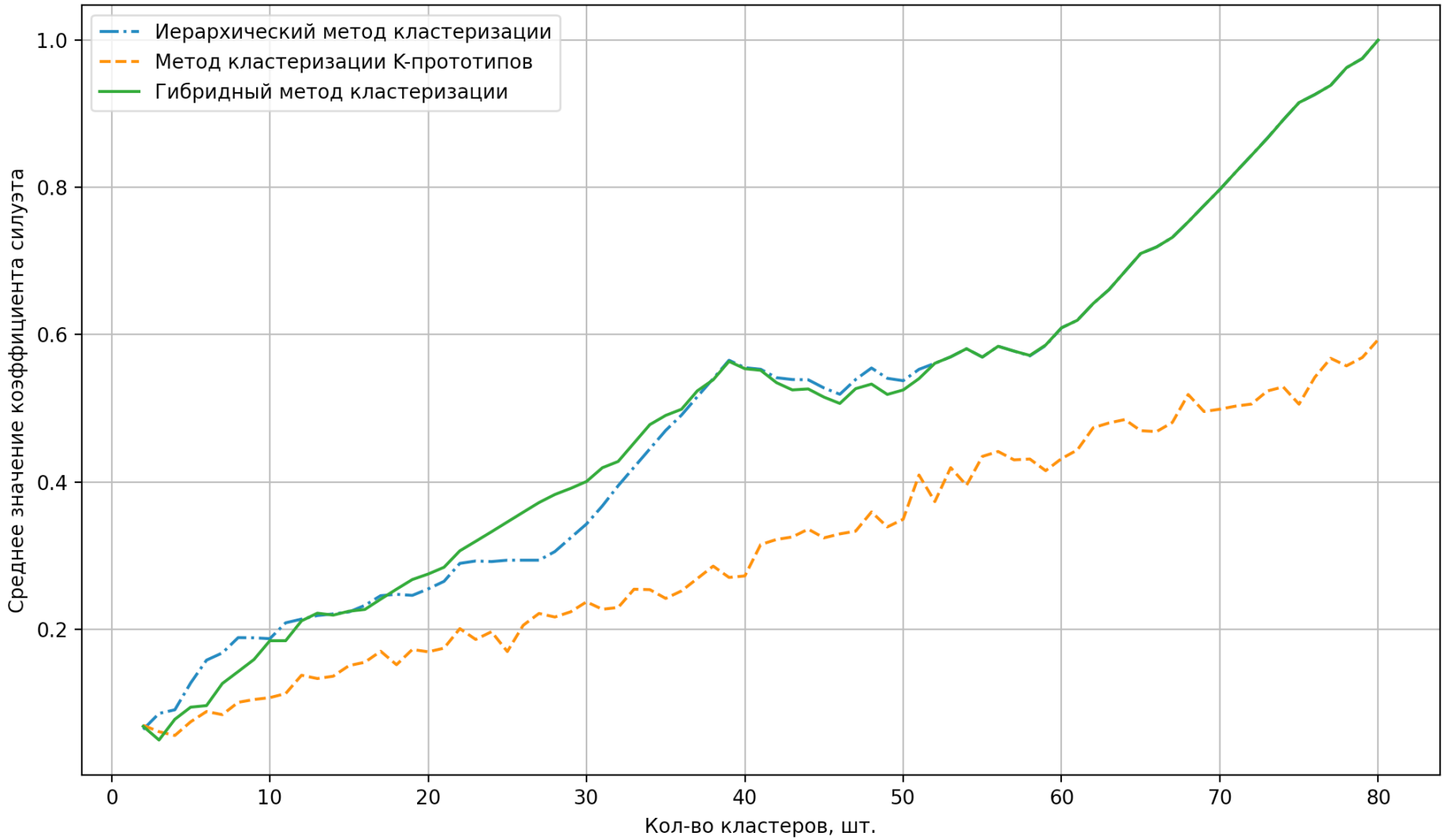
Оценка качества кластеризации (метод локтя)

- Показывает среднее расстояние между объектами внутри групп.



Оценка качества кластеризации (метод оценки силуэтов)

- Показывает, насколько близко каждая точка внутри одной группы расположена к точкам ближайшего соседнего кластера.



Заключение

В ходе выполнения работы были выполнены все поставленные задачи:

- проведен аналитический обзор известных методов кластеризации данных;
- формализирована поставленная задача;
- выбраны меры расстояний и критерии связи кластеров;
- разработан гибридный метод разбиения данных;
- разработаны алгоритмы и структура ПО;
- проведены исследования разработанного метода.

Направление дальнейшего развития

- Разработать алгоритмы управления кластеризацией в зависимости от особенностей исходных структур.
- Добавить возможность выбора метода вычисления расстояний в матрице несходства.