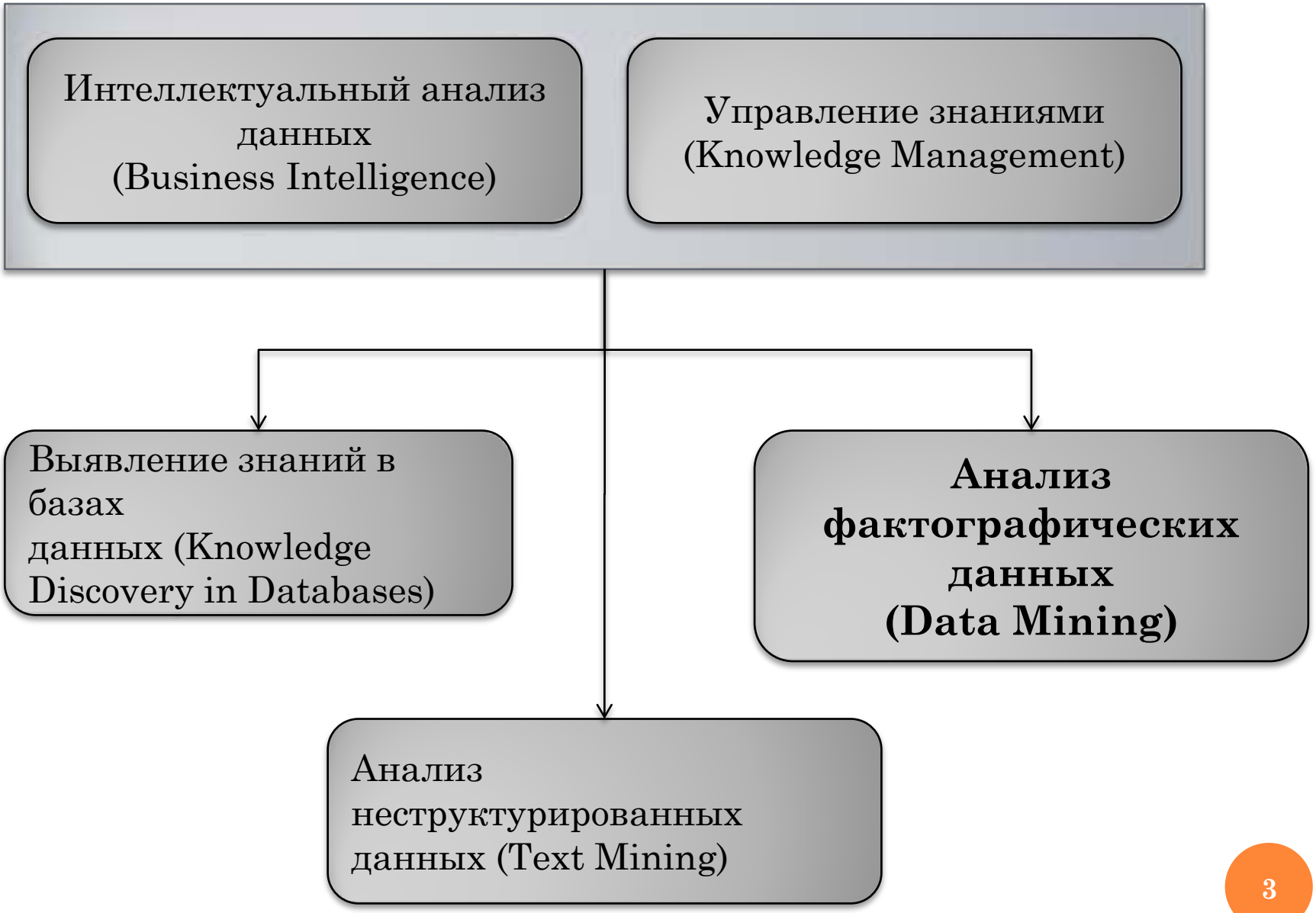
Интеллектуальный анализ данных

Нейский И.М.

Интеллектуальный анализ данных (ИАД) — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.


Интеллектуальный анализ данных (Business Intelligence)
Управление знаниями (Knowledge Management)
Выявление знаний в базах данных (Knowledge Discovery in Databases)
Анализ фактографических данных (Data Mining)
Анализ неструктурированных данных (Text Mining)

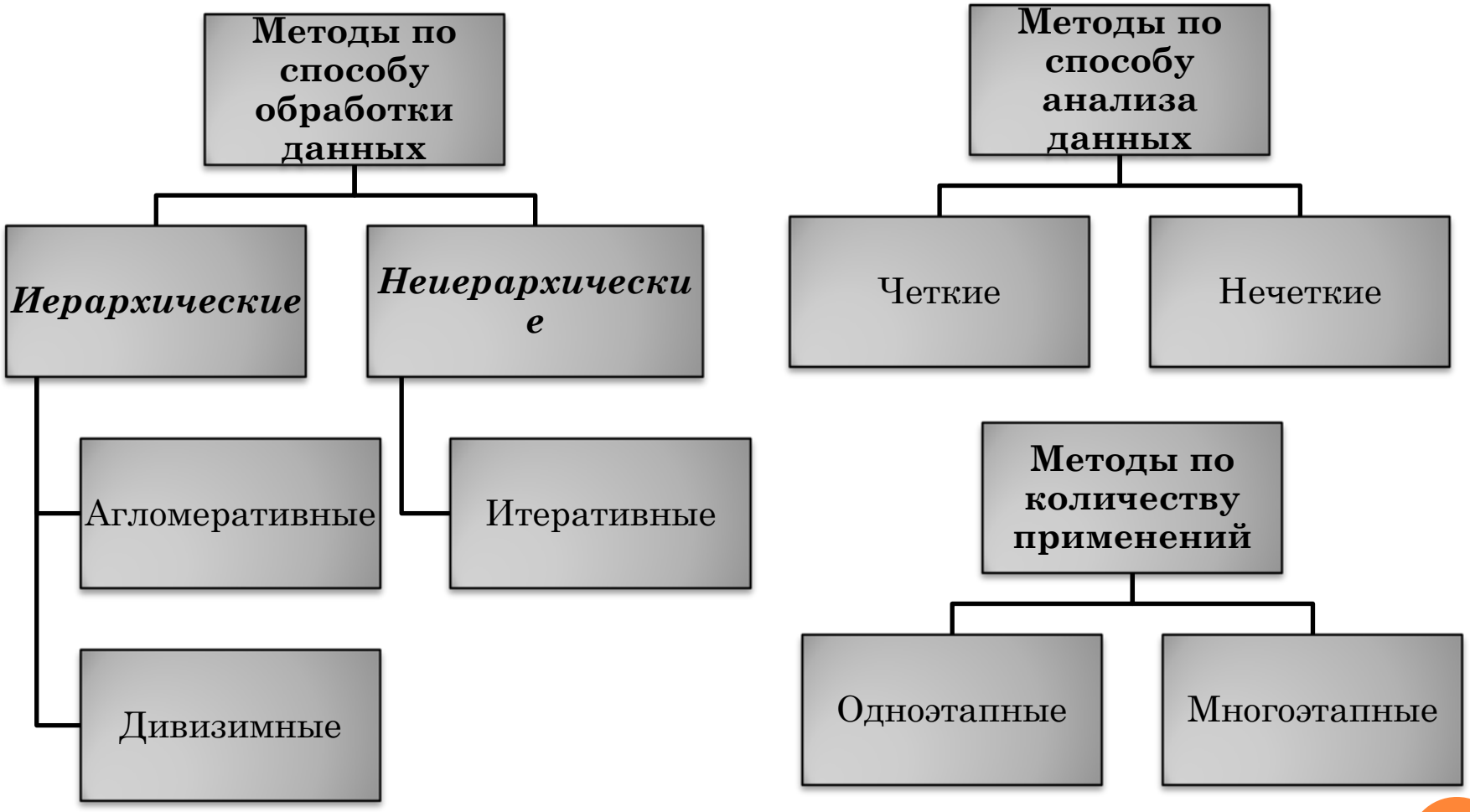
Стадии ИАД

1. Выявление закономерностей
2. Использование выявленных закономерностей для предсказания неизвестных значений
3. Анализ исключений

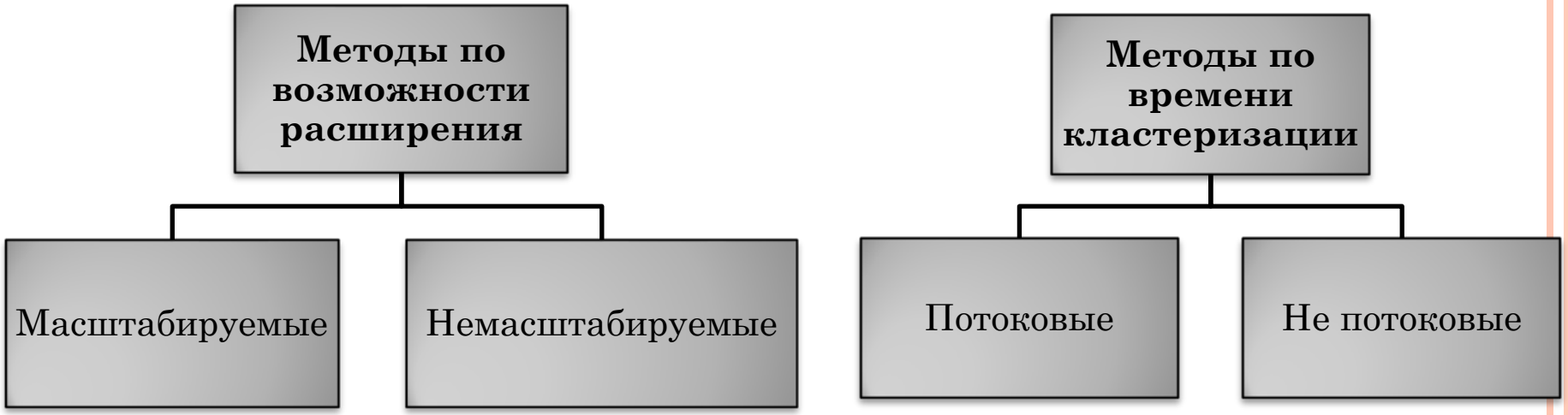
Задачи ИАД

1) Классификация
2) ***Кластеризация***
3) Выявление ассоциаций
4) Выявление последовательностей
5) Регрессия
6) Прогнозирование

Классификация методов кластеризации



Классификация методов кластеризации



Примеры методов кластеризации

- CURE (Clustering Using Representatives)
- CHAMELEON
- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)
- k-средних
- LargeItem
- Самоорганизующиеся карты Кохонена
- Fuzzy C-means и др.

Базовые понятия

Кластер – совокупность объектов, выделенная по формальному критерию их близости.

Характеристики кластера:

Центр кластера – это среднее геометрическое место точек в пространстве характеристик объектов, образующих кластер.

Радиус кластера – это максимальное расстояние до объекта, входящего в кластер, от центра кластера.

Размер кластера – определяется либо по радиусу кластера, либо по среднеквадратичному отклонению объектов для этого кластера.

Определение меры расстояния / сходства

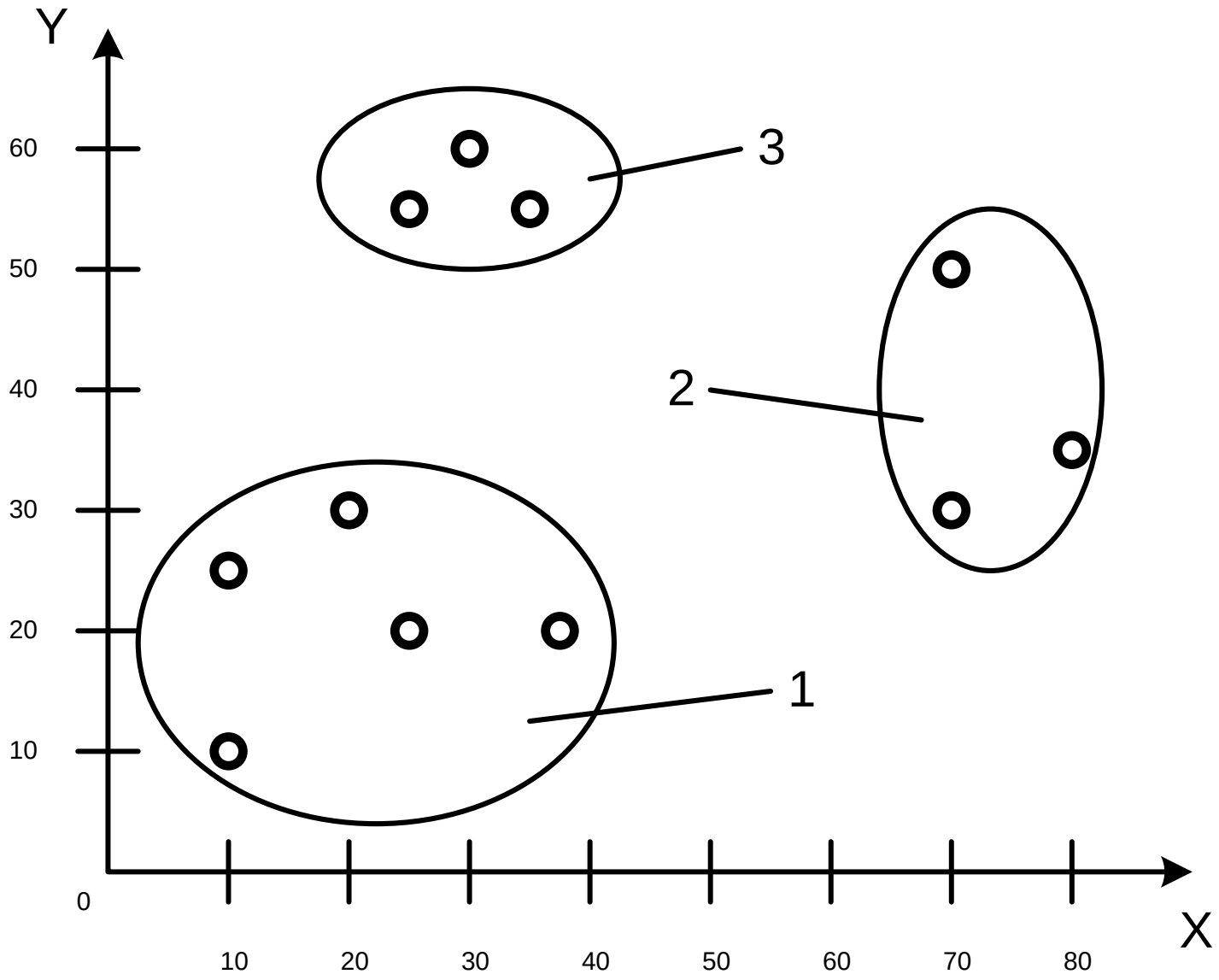
Евклидово расстояние $$D = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2 + (c_1 - c_2)^2 + \ldots}$$

Квадрат Евклидова расстояния: $$D = (a_1 - a_2)^2 + (b_1 - b_2)^2 + (c_1 - c_2)^2 + \ldots$$

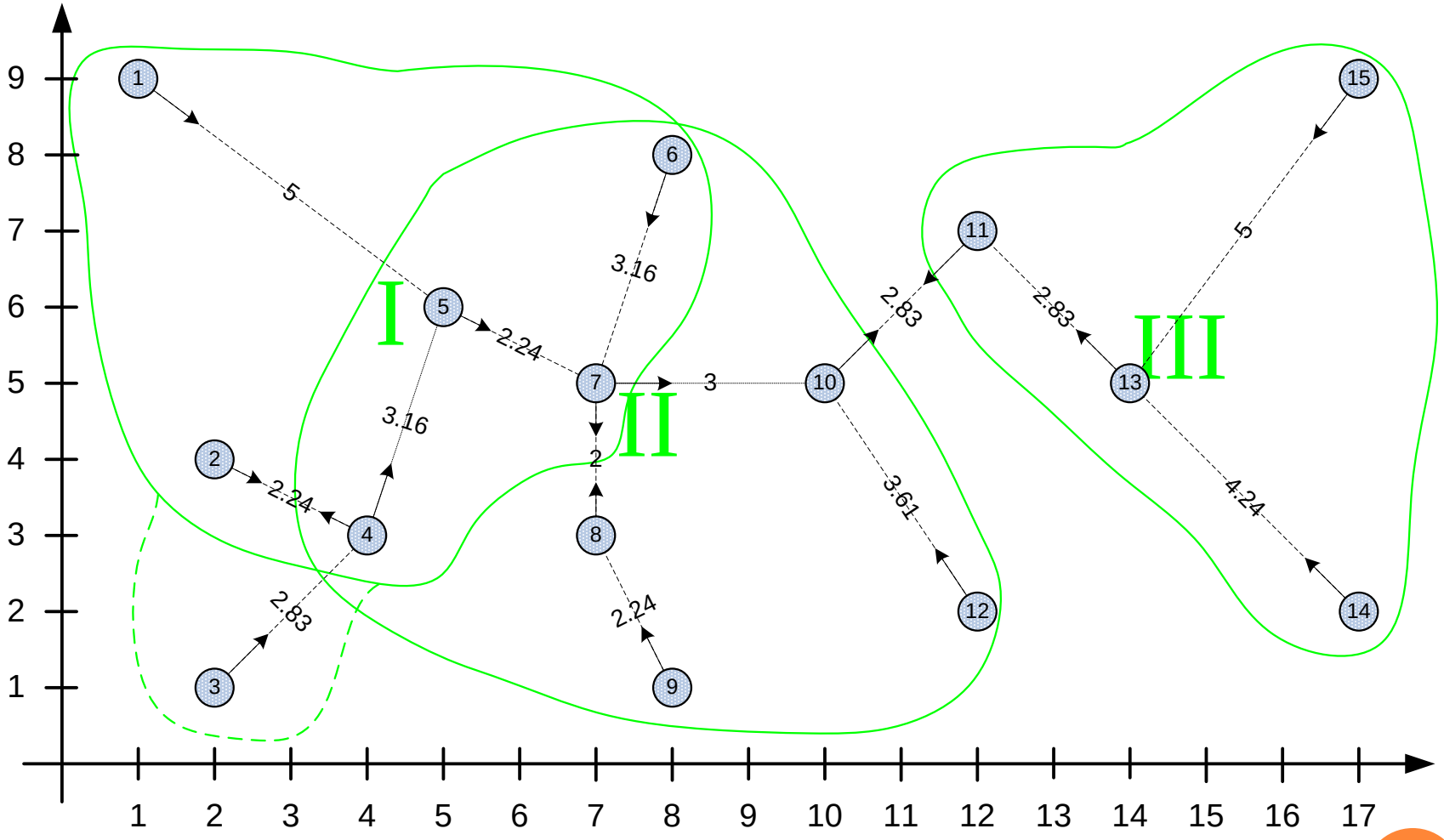
Манхэттенское расстояние: $$D = \frac{(a_1 - a_2) + (b_1 - b_2) + (c_1 - c_2) + \ldots}{n}$$

Расстояние Чебышева: $$D = Max\{|a_1 - a_2|, |b_1 - b_2|, |c_1 - c_2|, \ldots\}$$

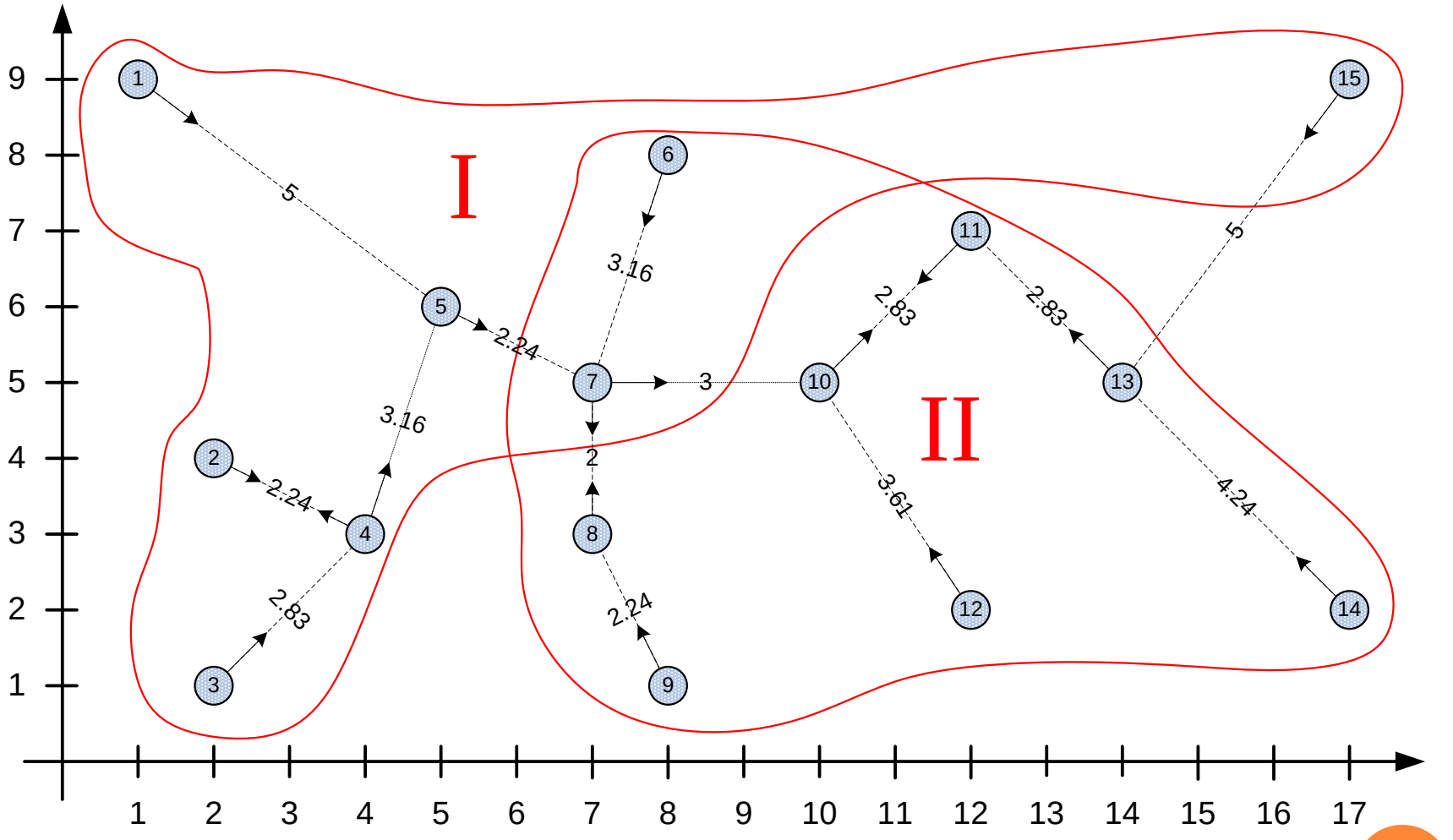
Примеры кластеризации



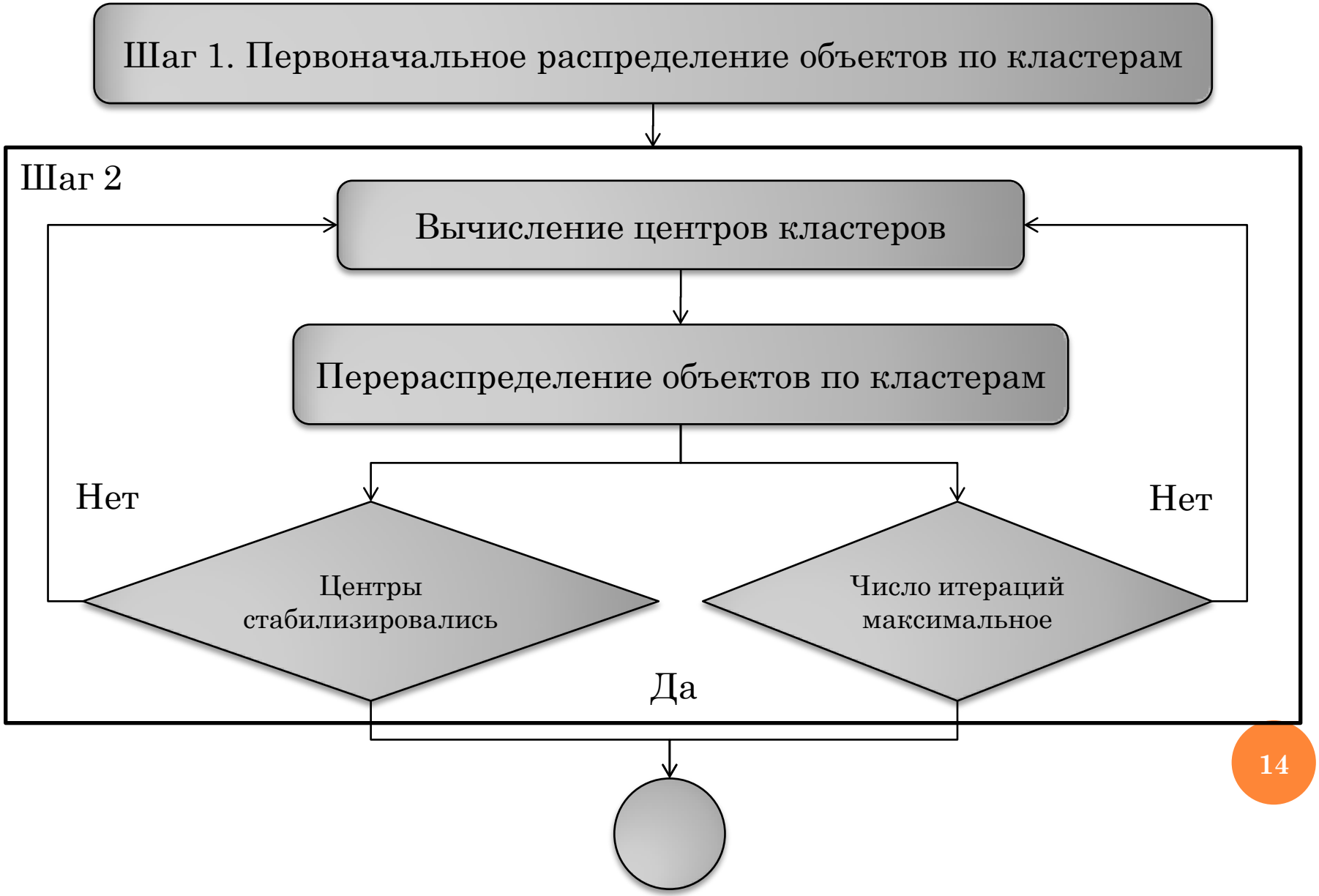
Примеры кластеризации



Примеры кластеризации



K-СРЕДНИЕ (K-MEANS)



Кластеризация методом к-средних

Исходные данные ***K = 3***

Vehicle	Top_speed	Colour	Air_resistance	Weight	Type
V1	220	red	0.30	1 300	Sport
V2	230	black	0.32	1 400	Sport
V3	260	red	0.29	1 500	Sport
V4	140	grey	0.35	800	Medium market
V5	155	blue	0.33	950	Medium market
V6	130	white	0.40	600	Medium market
V7	100	black	0.50	3 000	Lorry
V8	105	red	0.60	2 500	Lorry
V9	110	grey	0.55	3 500	Lorry

Кластеризация методом к-средних

Итерация 1

	V1	V2	V3	V4	V5	V6	V7	V8	V9
V1	0.00	100.50	203.96	506.36	355.98	705.76	1 704.23	1 205.50	2 202.75
V2	100.50	0.00	104.40	606.71	456.21	806.23	1 605.27	1 107.08	2 103.43
V3	203.96	104.40	0.00	710.21	559.93	909.34	1 508.51	1 011.94	2 005.62
V4	506.36	606.71	710.21	0.00	150.75	200.25	2 200.36	1 700.36	2 700.17
V5	355.98	456.21	559.93	150.75	0.00	350.89	2 050.74	1 550.81	2 550.40
V6	705.76	806.23	909.34	200.25	350.89	0.00	2 400.19	1 900.16	2 900.07
V7	1 704.23	1 605.27	1 508.51	2 200.36	2 050.74	2 400.19	0.00	500.03	500.10
V8	1 205.50	1 107.08	1 011.94	1 700.36	1 550.81	1 900.16	500.03	0.00	1 000.01
V9	2 202.75	2 103.43	2 005.62	2 700.17	2 550.40	2 900.07	500.10	1 000.01	0.00

Кластер1: V1, V2, V3, Кластер 2: V5, V4, V6, Кластер 3: V8,V7, V9

Кластеризация методом к-средних

Новые центры

$$C_i = AVG\left[x_j\right], x_j \in i$$

	Top_speed	Air_resistance	Weight
C1	237	0.30	1 400
C2	142	0.36	783
C3	105	0.55	3 000

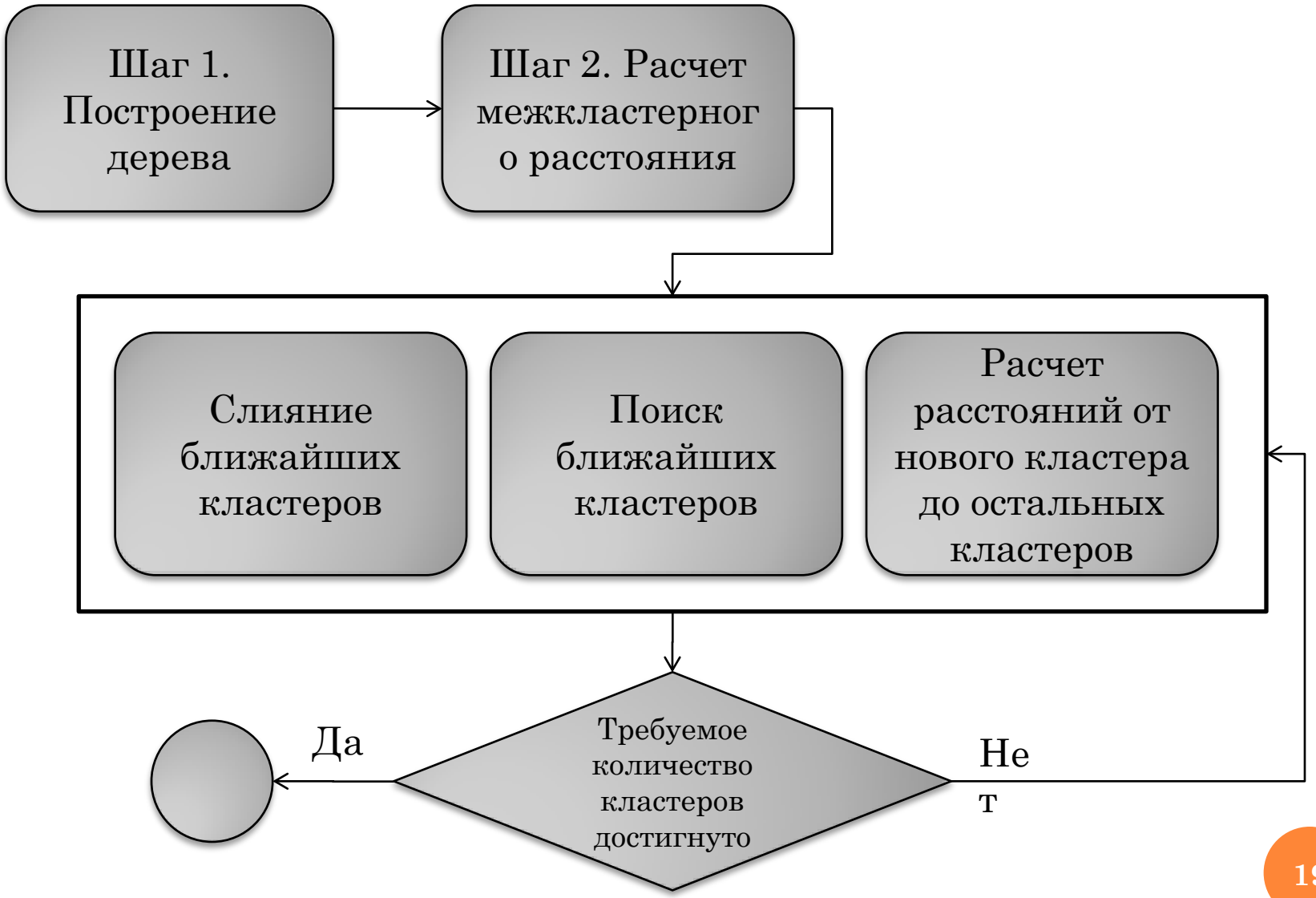
Кластеризация методом к-средних

Итерация 2

	C1	C2	C3
V1	101.38	522.57	1 703.89
V2	6.67	622.96	1 604.88
V3	102.69	726.37	1 507.99
V4	607.74	16.75	2 200.28
V5	457.35	167.20	2 050.61
V6	807.08	183.70	2 400.13
V7	1 605.83	2 217.06	5.00
V8	1 107.85	1 717.06	500.00
V9	2 103.82	2 716.85	500.02

Кластер1: V1, V2, V3, Кластер 2: V5, V4, V6, Кластер 3: V8,V7, V9

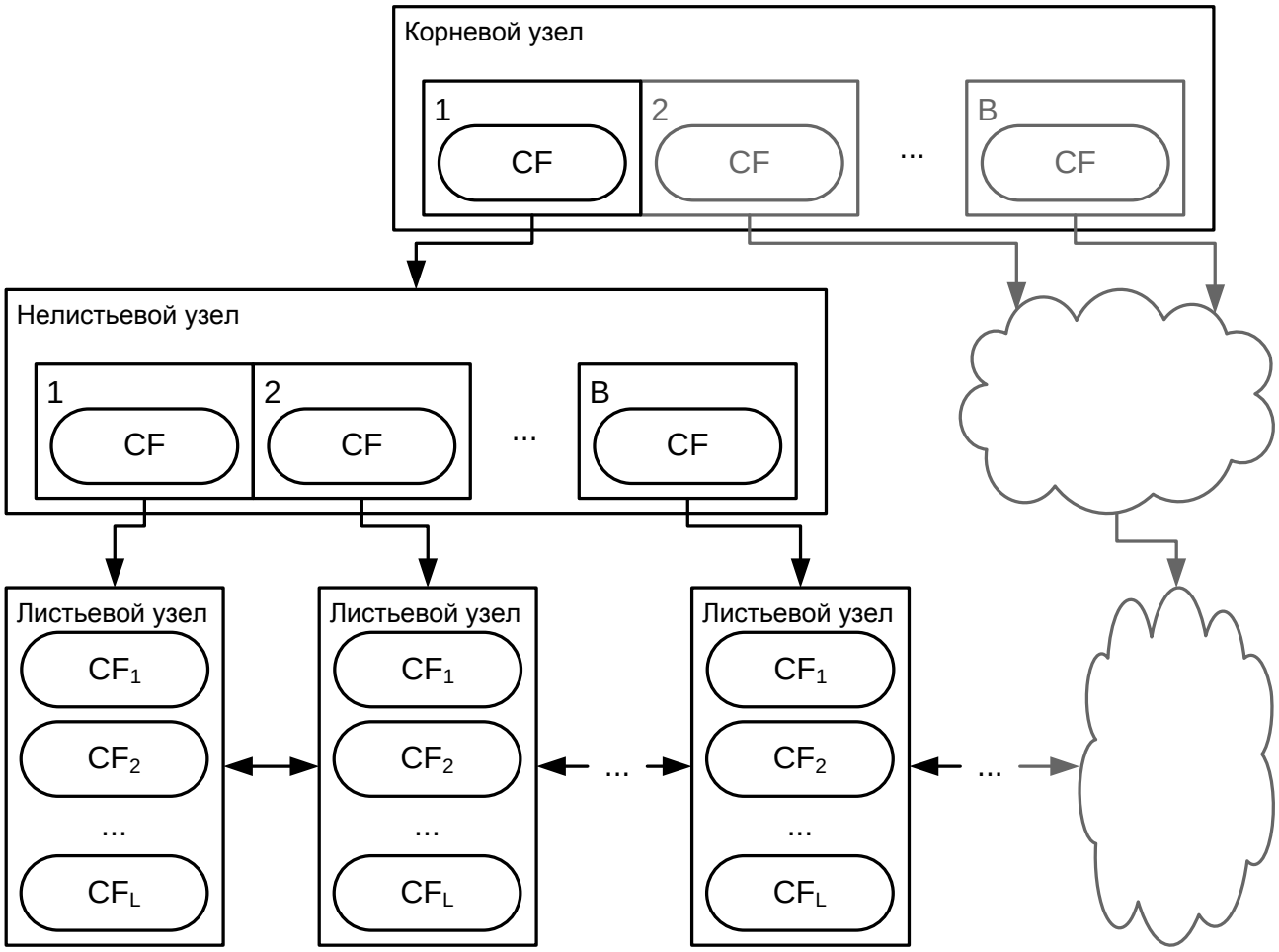
CURE (Clustering Using REpresentatives)



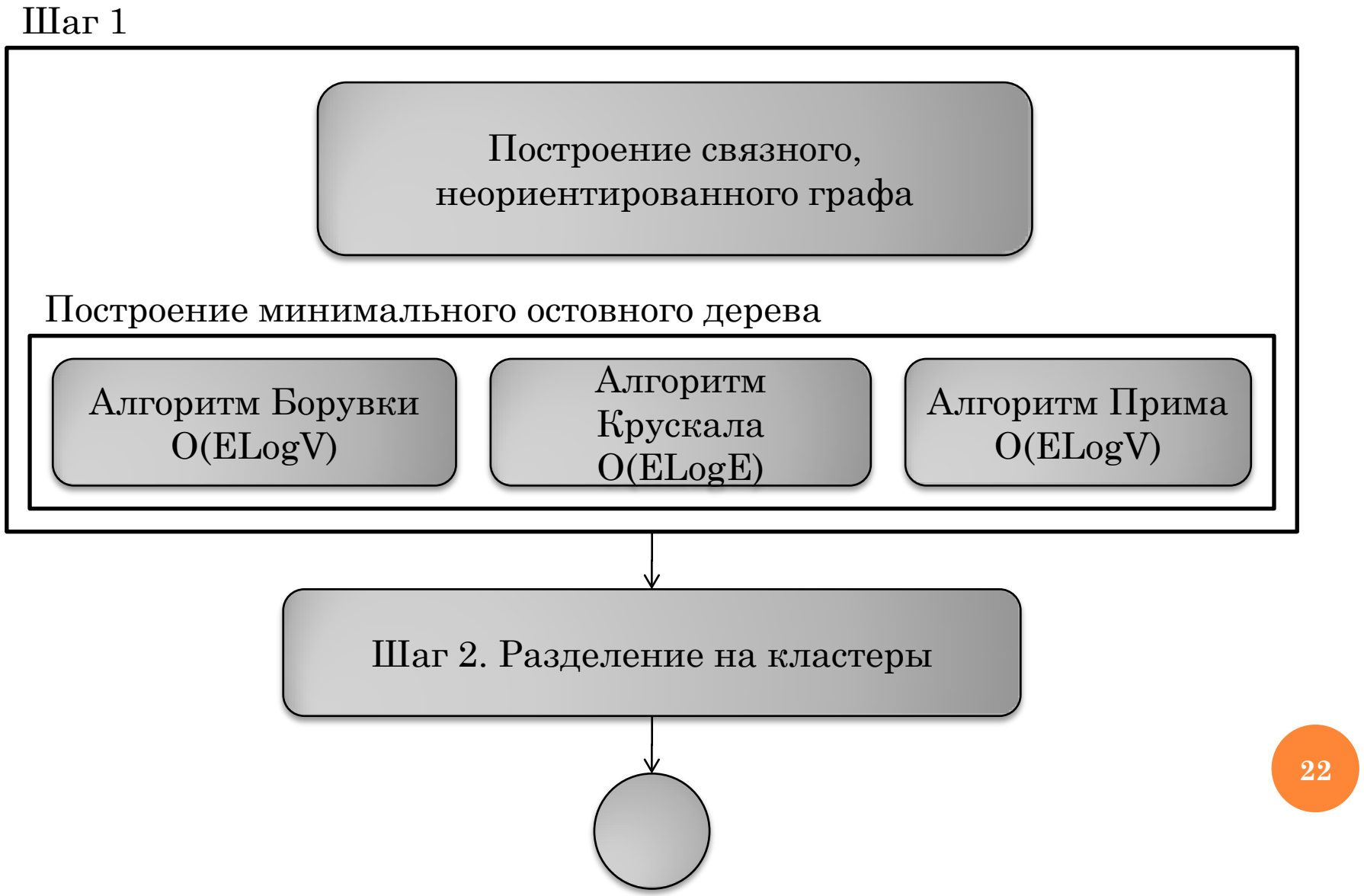
BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

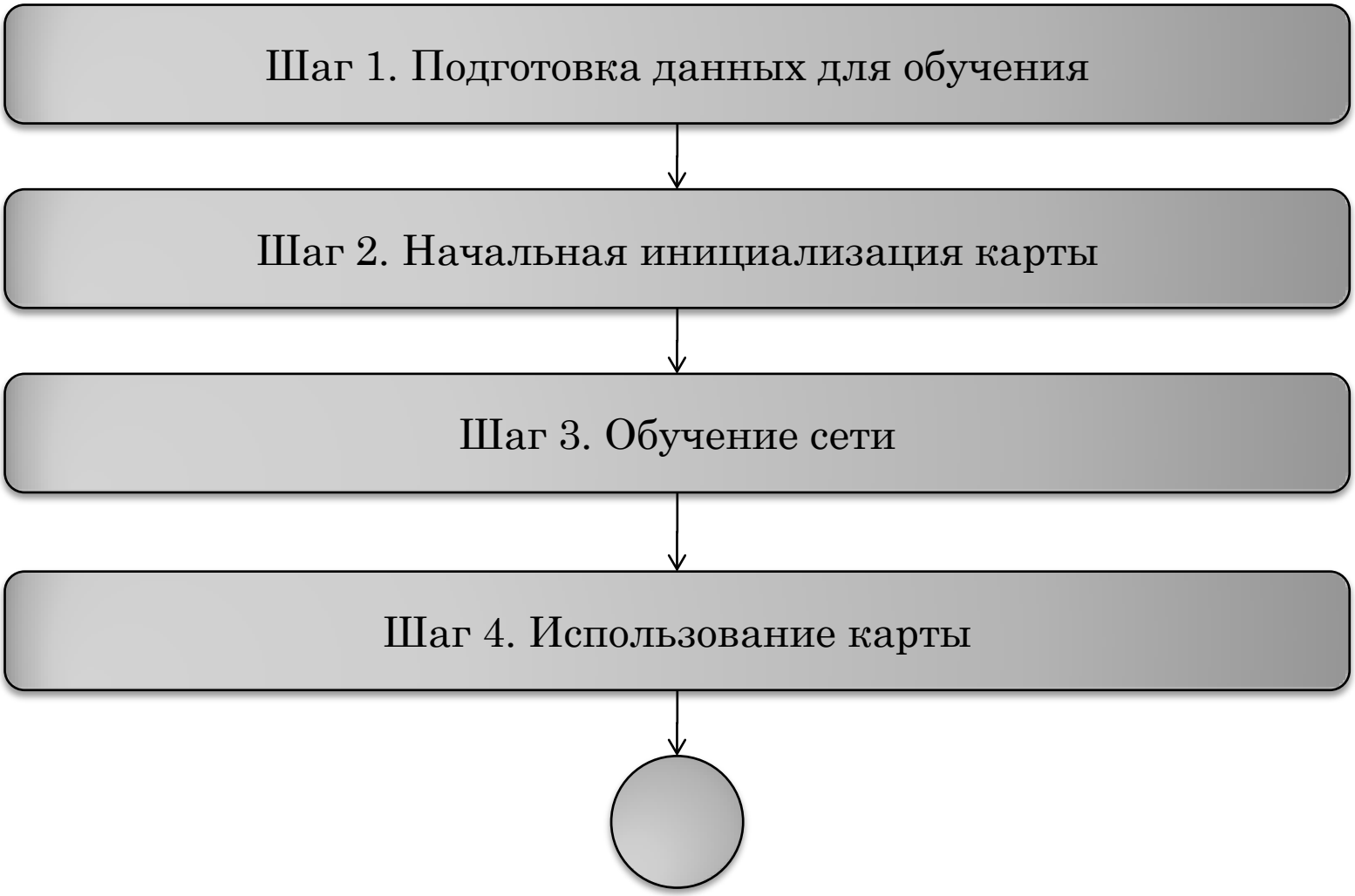
Кластерное дерево



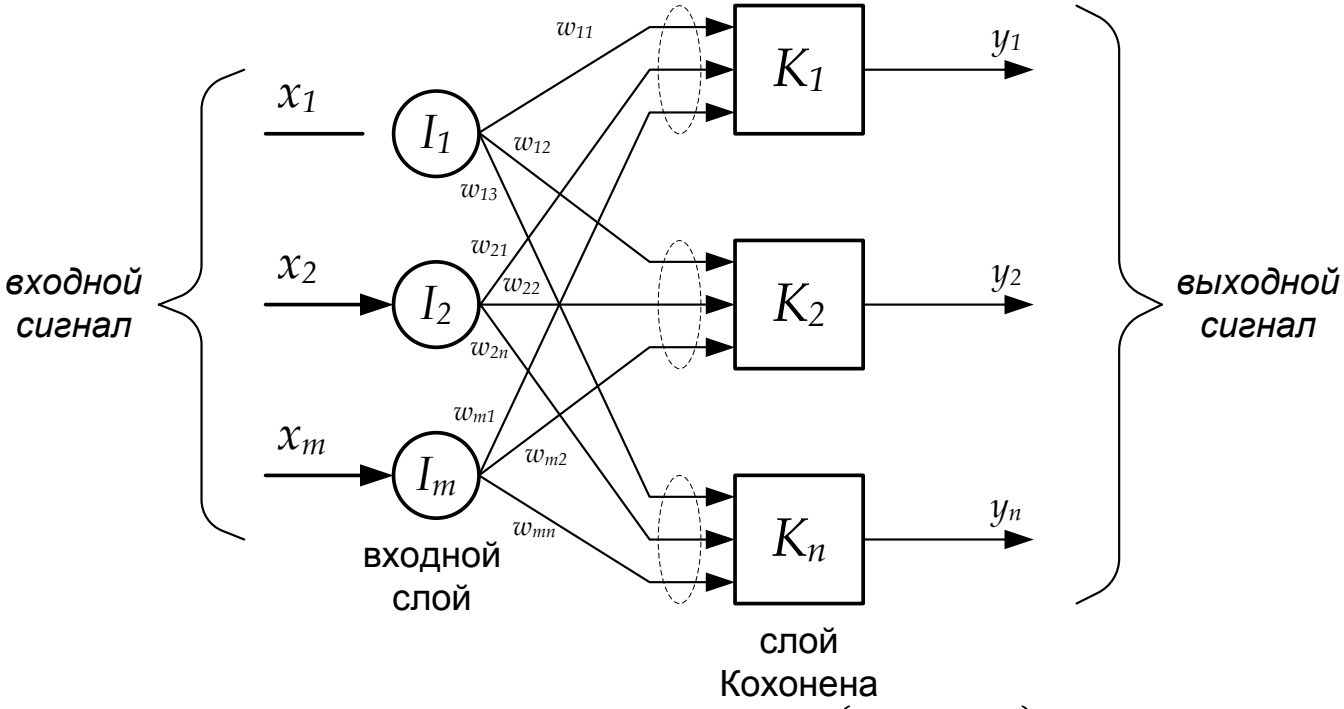
MST (Algorithm based on Minimum Spanning Trees)



Самоорганизующиеся карты Кохонена



Самоорганизующиеся карты Кохонена



Выбор нейрона-победителя: $\|x - w_c\| = \min_i \{\|x - w_i\|\}$

Модификация весов: $w_i(t+1) = w_i(t) + h_{ci}(t) * [x(t) - w(t)]$

$$h(t) = h(\|r_c - r_i\|, t) * a(t)$$

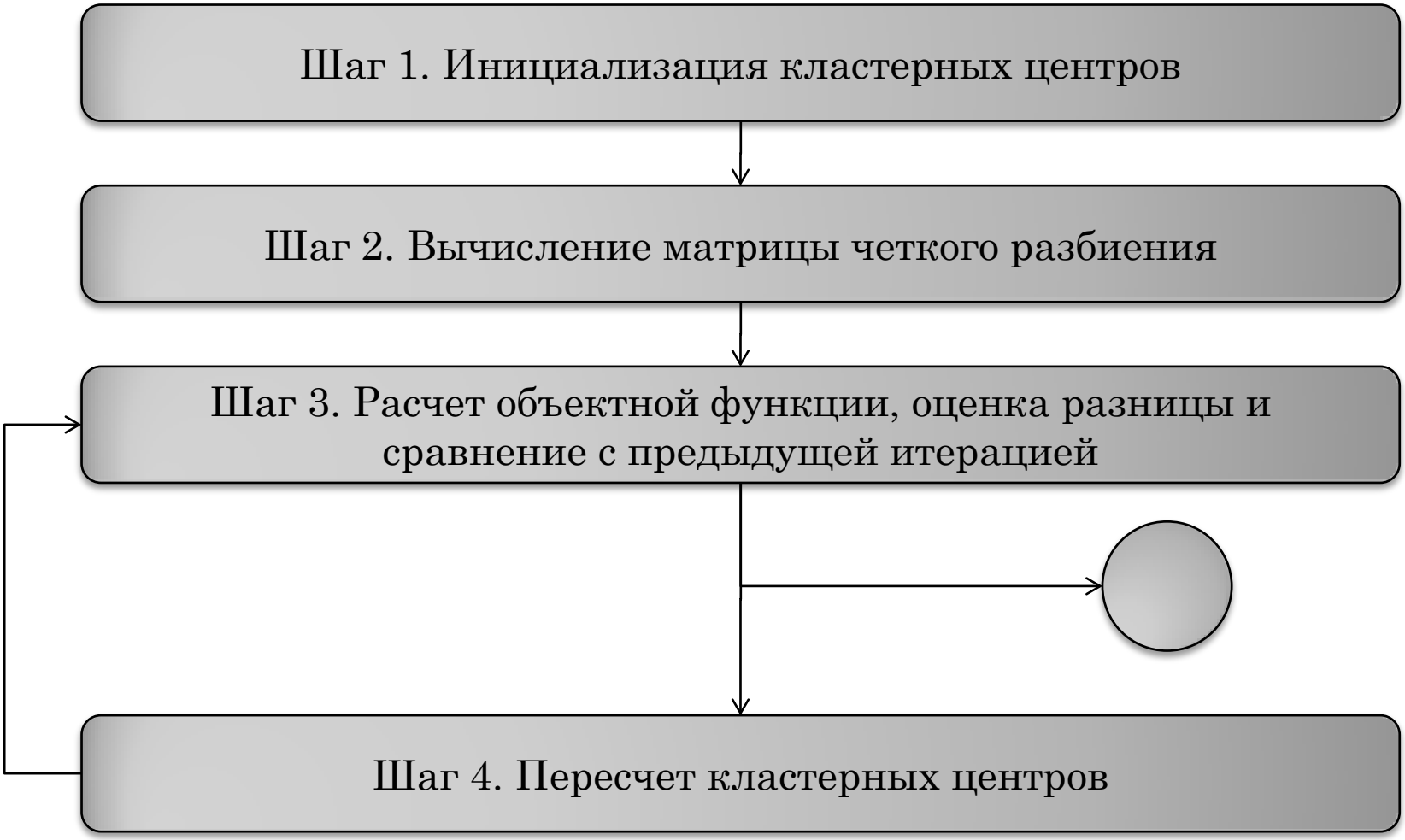
Функции расстояния:

Постоянная: $h(d,t) = \begin{cases} const, d \le \delta(t) \\ 0, d > \delta(t) \end{cases}$

Гауссова функция: $h(d,t) = e^{-\frac{d^2}{2*\delta^2(t)}}$

Функция скорости обучения сети: $a(t) = \frac{A}{t+B}$

HCM (Hard C – Means)



HCM (Hard C – Means)

Матрица четкого разбиения:

$$m_{ik} = \begin{cases} 1, \|u_k - c_i\|^2 \le \|u_k - c_j\|^2 \\ 0, \text{в остальных случаях} \end{cases} \quad i \neq j, \; i = \overline{1,c}, \; k = \overline{1,K}$$

Свойства матрицы:

$$\sum_{i=1}^{c} m_{ij} = 1, \sum_{j=1}^{K} m_{ij} = K$$

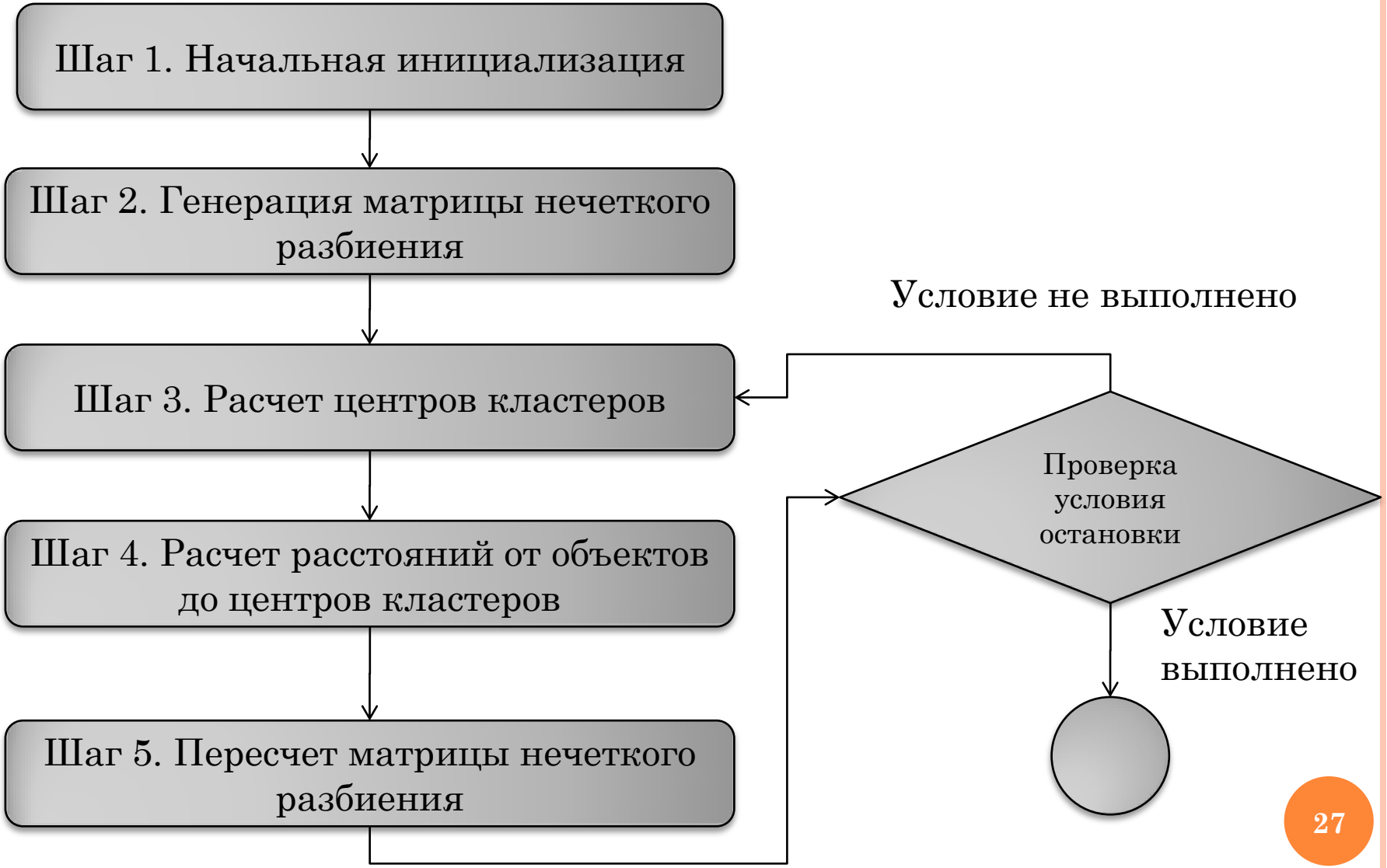
Объектная функция:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, u_k \in C_i} \|u_k - c_i\|^2 \right)$$

Пересчет кластерных центров:

$$c_i = \frac{1}{|C_i|} * \sum_{k, u_k \in C_i} u_k$$

Fuzzy C-means



Кластеризация методом Fuzzy C-means

Исходные данные

Vehicle	Top_speed	Colour	Air_resistance	Weight	Type
V1	220	red	0.30	1 300	Sport
V2	230	black	0.32	1 400	Sport
V3	260	red	0.29	1 500	Sport
V4	140	grey	0.35	800	Medium market
V5	155	blue	0.33	950	Medium market
V6	130	white	0.40	600	Medium market
V7	100	black	0.50	3 000	Lorry
V8	105	red	0.60	2 500	Lorry
V9	110	grey	0.55	3 500	Lorry

Кластеризация методом Fuzzy C-means

Шаг 1. Начальная инициализация

c = 3, m = 2, ξ – граничное условие

Шаг 2. Генерация матрицы нечеткого разбиения

	V1	V2	V3	V4	V5	V6	V7	V8	V9	Проверка строки
C1	0.70	0.80	0.60	0.30	0.00	0.00	0.20	0.00	0.20	2.80
C2	0.20	0.20	0.20	0.50	0.50	1.00	0.30	0.10	0.20	3.20
C3	0.10	0.00	0.20	0.20	0.50	0.00	0.50	0.90	0.60	3.00
Проверка столбца	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	

Кластеризация методом Fuzzy C-means

Шаг 3. Расчет центров кластеров

$$V_i = \frac{\sum_{j=\overline{1,n}} \mu_{ji}^{\,m} \times x_j}{\sum_{j=\overline{1,n}} \mu_{ji}^{\,m}}$$

	Top_speed	Air_resistance	Weight
C1	222.65	0.32	1448.80
C2	140.11	0.39	2487.22
C3	117.39	0.52	2487.22

Кластеризация методом Fuzzy C-means

Шаг 4. Расчет расстояний от объектов до центров кластеров

	C1	C2	C3
V1	148.82	1 189.90	1 191.64
V2	49.35	1 090.93	1 093.03
V3	63.38	994.47	997.46
V4	654.04	1 687.22	1 687.37
V5	503.36	1 537.29	1 537.68
V6	853.84	1 887.24	1 887.26
V7	1 556.05	514.35	513.08
V8	1 057.77	37.37	17.80
V9	2 054.30	1 013.23	1 012.81

Кластеризация методом Fuzzy C-means

Шаг 5. Пересчет матрицы нечеткого разбиения

$$\mu_{ji} = \frac{1}{\left(D_{ij}^{\ 2} * \sum_{k=1}^{c} \frac{1}{D_{kj}^{\ 2}} \right)^{1/(m-1)}}, i = \overline{1,c},\ j = \overline{1,n}$$

	V1	V2	V3	V4	V5	V6	V7	V8	V9	Проверка строки
C1	0.97	1.00	0.99	0.77	0.82	0.71	0.05	0.00	0.11	5.42
C2	0.02	0.00	0.00	0.12	0.09	0.15	0.47	0.18	0.45	1.47
C3	0.02	0.00	0.00	0.12	0.09	0.15	0.48	0.81	0.45	2.11
Проверка столбца	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	

Основные недостатки методов

1) необходимость задания пороговых значений;
2) необходимость задания количества кластеров;
3) работа только с данными одного типа (числовые или текстовые);
4) выделение кластеров определенной формы;
5) чувствительность к аномалиям в наборе данных;
6) существующие методы являются контекстно-зависимыми;
7) возможно возникновение неопределенностей;
8) медленная работа на больших объемах данных;
9) отсутствие гарантии в нахождении оптимального решения;
10) нелинейность времени работы алгоритма в зависимости от объема входных данных;
11) вычислительная сложность;

Основные проблемы кластерного анализа

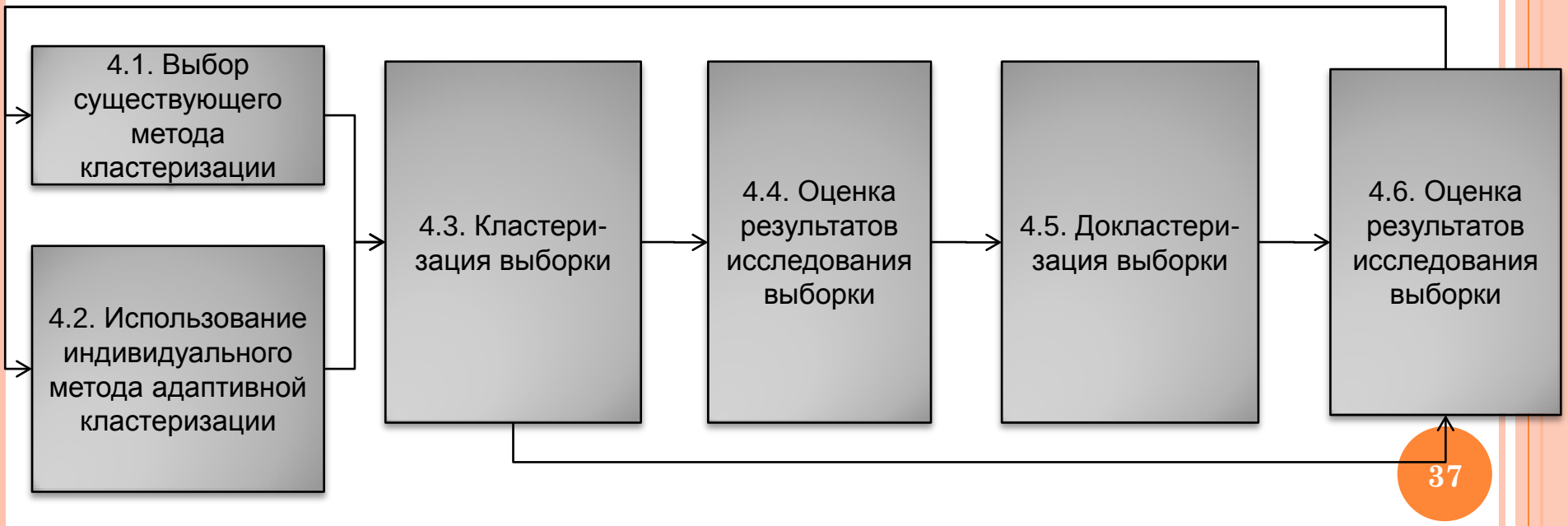
1. Выбор метода исследования
2. Оценка качества полученного разбиения
3. Выбор значения параметра «Количество кластеров»
4. Постоянно растущие объемы данных

Докластеризация

Адаптивная кластеризация

Методика адаптивной кластеризации

Выбор метода кластерного анализа



Выбор существующего метода кластеризации

Шаг 1. Выбор метода кластерного анализа

1. На основе существующих рекомендаций по исследованию предметных областей и задач.
2. На основе критериев
3. Общий алгоритм

Шаг 2. Настройка параметров выбранного метода кластерного анализа

— объем обучающего множества;
— объем валидационного множества;
— объем тестового множества;
— количество атрибутов входного набора данных;
— тип атрибутов входного набора данных;
— используемость атрибутов входного набора данных.

Характеристические параметры

— количество кластеров;
— алгоритм выполнения дополнительной кластеризации;
— пороговое значение остановки работы алгоритма;
— способ выбора начальных центров;
— максимальное количество итераций;
— количество одновременно обрабатываемых данных;
— количество предварительных разделов;
— коэффициент удаленности.

Итерационные параметры

— способ определения расстояния между кластерами;
— метод оценки качества кластеризации;
— пороговое значение для метода оценки качества кластеризации;
— начальное пороговое значение алгоритма;
— процент аномалий (выбросов) в полном объеме;
— разделяющая функция;
— скорость обучения сети.

Экспертные параметры

Шаг 3. Анализ массива фактографических данных и оценка разбиения

Кластеризация массива

Аналитическая оценка

Индекс оценки$_1$

Индекс оценки$_2$

…

Индекс оценки$_к$

Качество?

Да

Результат

Нет

Выбор метода на основе рекомендаций

Метод	Рекомендации	Противопоказания
CURE	Выявляет кластеры произвольной формы, метод менее чувствителен к выбросам, чем MST. Время работы алгоритма незначительное	Не использовать для исследования объектов с большим количеством атрибутов, требует задания пороговых значений
BIRCH	Метод предназначен для очень больших наборов данных. Работает с произвольным количеством оперативной памяти. Получаемое разбиение обладает высоким качеством	Не использовать для получения несферических форм кластеров, требует задания пороговых значений
MST	Лучше всего подходит для выделения кластеров произвольной формы	Очень чувствителен к выбросам и может медленно работать на больших массивах данных
k-средних	Показывает хорошие результаты при работе с данными, которые распределены по компактным группам сферической формы	Очень чувствителен к выбросам и может медленно работать на больших массивах данных
PAM	Показывает хорошие результаты при работе с данными, которые распределены по компактным группам сферической формы	Чувствителен к выбросам и может медленно работать на больших массивах данных
SOM	Поиск и анализ закономерностей в данных	Требуется минимизация размеров карты, проблема с аналитическим обоснованием результатов исследования
HCM	Показывает высокие результаты при работе с данными, которые распределены по компактным группам сферической формы	Чувствителен к выбросам и может медленно работать на больших массивах данных
Fuzzy C-Means	Относит объект к разным кластерам на основе степени принадлежности элемента к кластерам, выделяет кластеры сферической формы	Высокие требования к вычислительной мощности используемого аппаратного обеспечения, не работает с объектами, которые удалены от всех кластеров, и с вложенными кластерами

Критерии оценки качества разбиения

Индекс «Хие-Бени»

$$\chi = \frac{\sum_{i=1,c}\sum_{j=1,M} \mu_{ij}^{\ m} * \left\| X_j - V_i \right\|^2}{M * \min_{i \neq j}\left(\left\| X_j - V_i \right\|^2 \right)}$$

Индекс истинности разбиения

$$O = \frac{r}{n} * \begin{cases} q/k, q \leq k \\ k/q, q > k \end{cases}$$

Коэффициент разбиения

$$PC = \frac{\sum_{i=1}^{Q}\sum_{j=1}^{K} u_{ij}^2}{Q}, PC \in \left[\frac{1}{K}, 1 \right]$$

Индекс четкости

$$Cl = \frac{K * PC - 1}{K - 1}, Cl \in [0,1]$$

Показатель компактности и изолированности

$$CS = \frac{\sum_{q=1}^{Q}\sum_{k=1}^{K} u_{qk}^2 * d^2\left(x_q, c_k\right)}{Q * \min\left\{ d^2\left(c_i, c_j\right) \middle| i, j \in \overline{1, K}, i \neq j \right\}}$$

Индекс эффективности (совокупность межкластерных и внутрикластерных отличий)

$$PI = \sum_{q=1}^{Q}\sum_{k=1}^{K} u_{qk}^2 * \left[d^2\left(c_k, \bar{x}\right) - d^2\left(x_q, c_k\right) \right]$$

Рекомендуемые источники

1. Jain, Dubes «Algorithms for clustering data»;
2. Материалы, представленные в библиотеке на сайте http://www.basegroup.ru;
3. Баргесян А.А., Куприянов М.С., Степаненко В.В. и Холод И.И. «Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP»;
4. Книги Дюка по Data Mining;
5. Чубукова И.А. «Data Mining».

СПАСИБО ЗА ВНИМАНИЕ!

Форум по теме: www.philippovich.ru в разделе «Семинары НОК CLAIM» ветка «Методика адаптивной кластеризации»