МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное общеобразовательное учреждение
высшего образования
«Курганский государственный университет»

Кафедра «Программное обеспечение
автоматизированных систем»

Анализ данных. Часть 3. Кластеризация. Методы машинного обучения.

Методические указания к выполнению лабораторных работ
для бакалавров направлений
09.03.03 «Прикладная информатика»,
09.03.04 «Программная инженерия»

Курган 2026

Кафедра: «Программное обеспечение автоматизированных систем»

Дисциплины: «Интеллектуальный анализ данных» (09.03.03), «Методы и алгоритмы анализа данных» (09.03.04)

Составитель: старший преподаватель Ю. В. Адаменко

Печатается в соответствии с планом издания, утвержденным методическим советом университета «13» декабря 2025 г.

Утверждены на заседании кафедры «5» декабря 2025 г.

1 Постановка задачи кластеризации

Кластерный анализ (**Data clustering**) – задача разбиения заданной выборки объектов (ситуаций) на непересекающиеся подмножества (называемые кластерами) так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Задача кластеризации относится к широкому классу задач обучения без учителя.

Типы входных данных:

• признаковое описание объектов. Каждый объект описывается набором своих характеристик, называемых ***признаками***. Признаки могут быть числовыми или нечисловыми.

• матрица расстояний между объектами. Каждый объект описывается расстояниями до всех остальных объектов обучающей выборки.

Матрица расстояний может быть вычислена по матрице признаковых описаний объектов бесконечным числом способов, в зависимости от того, как ввести функцию расстояния (метрику) между признаковыми описаниями. Часто используется евклидова метрика, однако этот выбор в большинстве случаев является эвристикой и обусловлен лишь соображениями удобства.

Алгоритмов кластерного анализа достаточно много. Все их можно подразделить на ***иерархические*** и ***неиерархические***. Иерархические (древовидные) процедуры – наиболее распространенные алгоритмы кластерного анализа по их реализации на ЭВМ. Различают агломеративные (от слова ***agglomerate*** – собирать) и итеративные дивизивные (от слова ***division*** – разделять) процедуры.

Принцип работы иерархических агломеративных процедур состоит в последовательном объединении групп элементов, сначала самых близких, а затем всё более отдаленных друг от друга. Принцип работы иерархических дивизивных процедур, наоборот, состоит в последовательном разделении групп элементов, сначала самых далеких, а затем всё более близких друг от друга. Большинство этих алгоритмов исходит из матрицы расстояний (сходства).

Постановка задачи. Пусть имеется выборка **U** из **n** объектов наблюдения, каждый из которых характеризуется **m** числовыми признаками $\mathbf{X_j}$. Известно, что каждый из этих объектов на самом деле относится к одному из **k** классов, причём признаки объектов из одного класса не слишком сильно различаются, а признаки объектов из разных классов различаются более существенно. Мера различия между признаками объектов может быть определена как расстояние **ρ(x, y)** между векторами признаков в соответствующем признаковом пространстве. Требуется разбить множество объектов наблюдения **U** на **k** подмножеств $\mathbf{U_I}$, так чтобы:

1) множества не пересекались, то есть

$$\forall\, \boldsymbol{i}, \boldsymbol{j} \in [\mathbf{1}; \boldsymbol{k}] \cap \boldsymbol{Z}: \boldsymbol{i} \neq \boldsymbol{j} \Rightarrow \boldsymbol{U_i} \cap \boldsymbol{U_j} = \emptyset;$$

2) каждый объект относился к одному из множеств, то есть

$$\bigcup_{l=1}^{k} \boldsymbol{U_l} = \boldsymbol{U};$$

3) при этом некоторый показатель ошибки кластеризации **J** был минимален.

Сами подмножества $\mathbf{U_l}$ и называются *кластерами*. Показатель ошибки кластеризации может выбираться по-разному. Допустим, он представляет собой сумму квадратов расстояний от каждого объекта до центра соответствующего кластера:

$$\boldsymbol{J} = \sum_{\boldsymbol{l=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{x \in U_k}} \boldsymbol{\rho^2}(\boldsymbol{x}, \bar{\boldsymbol{x}}_{\boldsymbol{l}}), \tag{1}$$

где $\bar{\boldsymbol{x}}_{\boldsymbol{l}}$– это центр $\boldsymbol{l}$-го кластера:

$$\bar{\boldsymbol{x}}_{\boldsymbol{l}} = \frac{\mathbf{1}}{|\boldsymbol{U_l}|} \sum_{\boldsymbol{x \in U_l}} \boldsymbol{x}\,. \tag{2}$$

Здесь под $|\boldsymbol{U_l}|$понимается количество объектов, отнесенных к $\boldsymbol{l}$-у кластеру.

При этом сама метрика **ρ(x, y)** также может быть выбрана по-разному. Некоторые примеры метрик приведены ниже.

1 Манхэттенское расстояние (расстояние городских кварталов):

$$\rho_1(x, y) = \sum_{j=1}^{m} |x_j - y_j|. \tag{3}$$

2 Евклидово расстояние:

$$\rho_2(x, y) = \sqrt{\sum_{j=1}^{m} (x_j - y_j)^2}. \tag{4}$$

3 Максимальное различие по координатам:

$$\rho_\infty(x, y) = \max_{j \in [1;m] \cap Z} |x_j - y_j|. \tag{5}$$

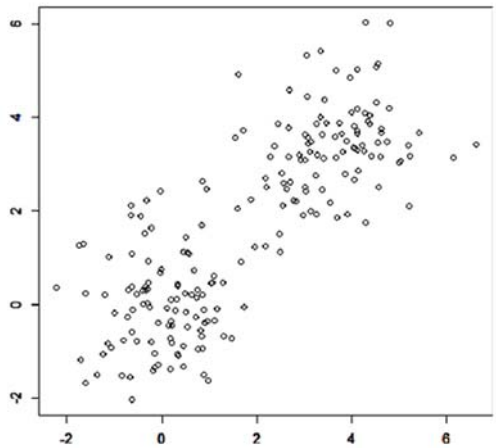


Рисунок 1 – Задача кластеризации на плоскости.

На рисунке 1 приведен пример задачи кластеризации точек на плоскости.

В действительности представленные точки относятся к одному из двух гауссовских распределений с одинаковыми корреляционными матрицами, но различными математическими ожиданиями. На рисунке видно, как точки концентрируются вокруг каждого из двух центров.

В такой постановке эта задача отличается высокой вычислительной сложностью. Можно показать справедливость следующих утверждений относительно класса её вычислительной сложности.

1 Задача кластеризации является **NP**-сложной в евклидовом пространстве даже для двух кластеров.

2 Задача кластеризации является **NP**-сложной для произвольного числа кластеров даже на плоскости.

3 Задача кластеризации может быть точно решена с помощью алгоритма, вычислительная сложность которого $\mathbf{O(n^{mk+1})}$.

Тем не менее, существует ряд эвристических алгоритмов, решающих эту задачу с некоторыми допущениями. Самый популярный из них – алгоритм **k-means**.

2 Алгоритм k-means

Идея алгоритма **k-средних** (англ. **k-means**) заключается в последовательном пересчёте средних по формуле 2.

Пусть вначале для каждого из **k** кластеров имеется некоторое начальное среднее $\bar{\boldsymbol{x}}_{\boldsymbol{l}}^{(\mathbf{0})}$. Если внутригрупповые средние заданы, то каждый кластер $\mathbf{U_l}$ очевидным образом определяется как множество объектов наблюдения, вектора признаков для которых ближе к центру $\bar{\boldsymbol{x}}_{\boldsymbol{l}}^{(\mathbf{0})}$, чем к центрам других кластеров.

После этого внутригрупповое среднее можно пересчитать по формуле (2), после чего снова пересчитать кластеры, пока средние не перестанут меняться.

Таким образом, любую **s**-ю итерацию алгоритма можно описать тремя шагами:

1 распределить объекты наблюдения по кластерам:

$$U_l^{(s)} = \left\{x \in U | \forall j \in [1;k] \cap Z : \rho\left(x, \bar{x}_l^{(s-1)}\right) \leq \rho\left(x, \bar{x}_j^{(s-1)}\right), l \leq j\right\}.$$

При этом объекты просто относятся к тому кластеру, до центра которого расстояние от этого объекта меньше, а в случае равенства наименьших расстояний – в кластер с меньшим номером.

2 пересчитать центры кластеров по формуле (2). Они считаются, как обычные выборочные средние.

3 Если ничего не поменялось с прошлой итерации, то есть

$$\forall \boldsymbol{l} \in [\mathbf{1};\boldsymbol{k}] \cap \boldsymbol{Z} : \bar{\boldsymbol{x}}_{\boldsymbol{l}}^{(\boldsymbol{s})} = \bar{\boldsymbol{x}}_{\boldsymbol{l}}^{(\boldsymbol{s-1})},$$

то закончить выполнение алгоритма, иначе перейти к шагу 1 на следующей итерации (**s + 1**).

Этот алгоритм всегда сходится за конечное количество итераций, поскольку значение ошибки кластеризации (1) не увеличивается в процессе работы алгоритма, а число возможных разбиений конечного множества на подмножества конечно. Тем не менее, этот алгоритм имеет экспоненциальную

сложность в худшем случае, хотя на практике обычно сходится довольно быстро. Кроме того, этот алгоритм может сойтись не к глобальному, а к локальному минимуму функции (1).

Еще одной особенностью является тот факт, что количество кластеров **k** должно быть известно заранее. В некоторых прикладных задачах оно действительно известно. Если же это не так можно перебирать это количество, оценивая результат в каждом случае. Тем не менее, нужно понимать, что этот алгоритм не предназначен для случая, когда количество кластеров не известно.

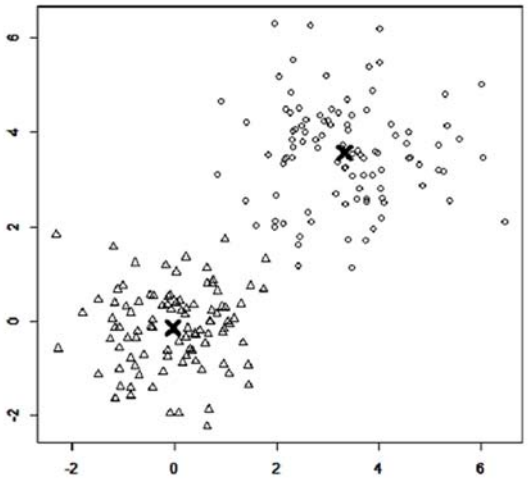


Рисунок 2 – Результат кластеризации с помощью алгоритма **k-means**

На рисунке 2 приведен пример результата кластеризации некоторого набора точек на плоскости с помощью алгоритма **k-means**. Кругами и треугольниками отмечены точки, отнесенные к двум различным кластерам. Крестами помечены окончательные центры кластеров.

Описанный алгоритм иногда называют алгоритмом **Ллойда** (англ. **Lloyd**) или **Ллойда-Форджи** (англ. **Lloyd-Forgy).** Существуют и альтернативные реализации описанного подхода. Так реализация **Маккуина** (англ. **MacQueen**) по-другому подходит к пересчёту центров кластеров. Она пересчитывает центры всякий раз, как очередной объект наблюдения меняет свой кластер, прямо в процессе перераспределения объектов по кластерам.

На практике очень часто используется реализация **Хартигана-Вонга** (англ. **Hartigan-Wong**). В ней для каждого объекта наблюдения вычисляется не только ближайший центр кластеров, но и второй по удаленности от него. На каждой итерации алгоритма дополнительно выполняется стадия быстрого перехода, на которой каждая точка перераспределяется во второй по удалённости кластер, если это уменьшит сумму всех расстояний между парами точек, лежащих в одном и том же кластере.

Многое зависит от выбора начальных центров $\bar{x}_l^{(0)}$. Как видно, единственное требование к ним – ни один кластер изначально не должен быть пустым. Один из простых способов выбора начальных центров, который гарантирует выполнение этого условия, – взять в качестве них любые **k** различных объектов наблюдения из множества **U**:

$$\bar{x}_l^{(0)} \in U \text{ и } \forall i,j \in [1;k] \cap Z: i \neq j \Rightarrow \bar{x}_i^{(0)} \neq \bar{x}_j^{(0)}.$$

Тогда, по крайней мере, сами объекты, выбранные в качестве центров, изначально попадут в соответствующие кластеры. Кроме того, существуют некоторые более эффективные способы автоматического выбора изначальных центров кластеров.

3 Автоматический выбор начальных центров кластеров

Две основные проблемы, возникающие при неудачном выборе в качестве центров кластеров нескольких случайных объектов наблюдения, состоят в большом количестве итераций алгоритма **k-means** и в достижении локального минимума функции (1) вместо глобального минимума. Использование специфического алгоритма автоматического выбора начальных центров кластеров (англ. **k-means**++) может решить эти проблемы.

Классический алгоритм автоматического выбора начальных центров кластеров состоит из **k** итераций, на каждой из которых определяется очередной начальный центр. Этот алгоритм можно описать в виде последовательности шагов следующим образом.

1 Первый начальный центр $\bar{x}_l^{(0)}$ выбирается равновероятно среди имеющихся в выборке U объектов наблюдения: $\bar{x}_l^{(0)} \in U$.

2 Пусть ранее уже определено **s** начальных центров кластеров из **k**. Объекты наблюдения распределяются по имеющимся кластерам, в соответствии с текущими центрами, как это делается на шаге 1 алгоритма **k-means**. Каждый объект наблюдения относится к тому кластеру, к центру которого он ближе. Для каждого объекта сохраняется расстояние до центра кластера, к которому он относится.

3 Очередной центр $\bar{x}_{s+1}^{(0)}$ выбирается случайно среди всех объектов наблюдения. При этом каждый объект x, относящийся к кластеру U_l, выбирается с вероятностью

$$\rho(x) = \frac{\rho^2(x,\bar{x}_l^{(0)})}{\sum_{j=1}^{s}\sum_{y\in U_j}\rho^2(y,\bar{x}_j^{(0)})}.$$

то есть вероятность выбора конкретного объекта наблюдения в качестве очередного начального центра кластера прямо пропорциональна квадрату расстояния от этого объекта наблюдения до ближайшего к нему ранее выбранного начального центра кластера.

4 Если все **k** начальных центров кластеров уже выбраны, закончить выполнение алгоритма, иначе перейти к шагу 2.

4 Метод k-медиан

Метод k-медиан – применяемая в статистике и машинном обучении вариация метода **k-means** для задач кластеризации, где для определения центроида кластера вместо среднего вычисляется медиана.

Задача определения **k-медиан** состоит в поиске таких **k** центров, что сформированные по ним кластеры будут наиболее «компактными».

Формально, при заданных точках данных $\mathbf{x_i}$, **k** центров $\mathbf{c_j}$ должны быть выбраны так, чтобы минимизировать сумму расстояний от каждой $\mathbf{x_i}$ до ближайшего $\mathbf{c_j}$.

Метод **k-медиан** иногда работает лучше, чем метод **k-means**, где минимизируется сумма квадратов расстояний. Критерий суммы расстояний широко используется для транспортных задач.

5 Метрики качества кластеризации

Задача оценки качества кластеризации является более сложной по сравнению с оценкой качества классификации. Во-первых, такие оценки не должны зависеть от самих значений меток, а только от самого разбиения выборки. Во-вторых, не всегда известны истинные метки объектов, поэтому также нужны оценки, позволяющие оценить качество кластеризации, используя только неразмеченную выборку.

Выделяют внешние и внутренние метрики качества. Внешние используют информацию об истинном разбиении на кластеры, в то время как внутренние метрики не используют никакой внешней информации и оценивают качество кластеризации, основываясь только на наборе данных. Оптимальное число кластеров обычно определяют с использованием внутренних метрик.

Силуэт. Данный коэффициент не предполагает знания истинных меток объектов и позволяет оценить качество кластеризации, используя только саму (неразмеченную) выборку и результат кластеризации. Сначала силуэт определяется отдельно для каждого объекта.

Обозначим через $\boldsymbol{a}$ – среднее расстояние от данного объекта до объектов из того же кластера, через $\boldsymbol{b}$ – среднее расстояние от данного объекта до объектов из ближайшего кластера (отличного от того, в котором лежит сам объект). Тогда силуэтом данного объекта называется величина:

$$s = \frac{b - a}{\max(a, b)}.$$

Силуэтом выборки называется средняя величина силуэта объектов данной выборки. Таким образом, силуэт показывает, насколько среднее расстояние до объектов своего кластера отличается от среднего расстояния до объектов других кластеров. Данная величина лежит в диапазоне [-1,1]. Значения, близкие к -1, соответствуют плохим (разрозненным) кластеризациям; значения, близкие к нулю, говорят о том, что кластеры пересекаются и накладываются друг на друга; значения, близкие к 1, соответствуют "плотным", четко выделенным кластерам. Таким образом, чем больше силуэт, тем более четко выделены кластеры, и они представляют собой компактные, плотно сгруппированные облака точек.

С помощью силуэта можно выбирать оптимальное число кластеров k (если оно заранее неизвестно) – число кластеров, максимизирующее значение силуэта. В отличие от предыдущих метрик, силуэт зависит от формы кластеров и

достигает больших значений на более выпуклых кластерах, получаемых с помощью алгоритмов, основанных на восстановлении плотности распределения.

6 Практические задания

6.1 Сегментирование клиентской базы компании средствами электронных таблиц

Используя **Excel** или **Calc** проведите сегментирование клиентской базы компании (кластеризацию методом **k-средних** и **k-медиан**) для проведения таргетированных рассылок о предложениях компании. Исходные данные находятся в файле «**WineKMC.xml**».

В нем содержатся данные о сделках Оптовой Винной Компании Джоуи Бэг О'Донатса за год:

• метаданные по каждому предложению (заказу) сохранены в электронной таблице, включая сорт, минимальное количество вина в заказе, скидку на розничную продажу, информацию о том, пройден ли ценовой максимум и о стране происхождения. Эти данные размещены во вкладке под названием **OfferInformation** как показано на рисунке 3;

• данные о заказах клиентов представлены в формате имя-№ заказа (рисунок 4).

	A	B	C	D	E	F	G
1	Offer #	Campaign	Varietal	Minimum Qty (kg)	Discount (%)	Origin	Past Peak
2	1	January	Malbec	72	56	France	FALSE
3	2	January	Pinot Noir	72	17	France	FALSE
4	3	February	Espumante	144	32	Oregon	TRUE
5	4	February	Champagne	72	48	France	TRUE
6	5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE
7	6	March	Prosecco	144	86	Chile	FALSE
8	7	March	Prosecco	6	40	Australia	TRUE
9	8	March	Espumante	6	45	South Africa	FALSE
10	9	April	Chardonnay	144	57	Chile	FALSE
11	10	April	Prosecco	72	52	California	FALSE
12	11	May	Champagne	72	85	France	FALSE
13	12	May	Prosecco	72	83	Australia	FALSE
14	13	May	Merlot	6	43	Chile	FALSE
15	14	June	Merlot	72	64	Chile	FALSE
16	15	June	Cabernet Sauvignon	144	19	Italy	FALSE
17	16	June	Merlot	72	88	California	FALSE
18	17	July	Pinot Noir	12	47	Germany	FALSE
19	18	July	Espumante	6	50	Oregon	FALSE
20	19	July	Champagne	12	66	Germany	FALSE
21	20	August	Cabernet Sauvignon	72	82	Italy	FALSE
22	21	August	Champagne	12	50	California	FALSE
23	22	August	Champagne	72	63	France	FALSE
24	23	September	Chardonnay	144	39	South Africa	FALSE
25	24	September	Pinot Noir	6	34	Italy	FALSE
26	25	October	Cabernet Sauvignon	72	59	Oregon	TRUE
27	26	October	Pinot Noir	144	83	Australia	FALSE
28	27	October	Champagne	72	88	New Zealand	FALSE
29	28	November	Cabernet Sauvignon	12	56	France	TRUE
30	29	November	Pinot Grigio	6	87	France	FALSE
31	30	December	Malbec	6	54	France	FALSE
32	31	December	Champagne	72	89	France	FALSE
33	32	December	Cabernet Sauvignon	72	45	Germany	TRUE

Рисунок 3 – Детали последних 32 заказов

	A	B
1	Customer Last Name	Offer #
2	Smith	2
3	Smith	24
4	Johnson	17
5	Johnson	24
6	Johnson	26
7	Williams	18
8	Williams	22
9	Williams	31
10	Brown	7
11	Brown	29
12	Brown	30
13	Jones	8
14	Miller	6
15	Miller	10

Рисунок 4 – Список количества заказов по покупателям

Порядок выполнения работы

Для выполнения работы выполните следующие этапы.

1 Создайте матрицу сделок по покупателям (лист **Pivot**), используя Мастер создания сводных таблиц, В результате получается таблица, показанная на рисунке 5.

Count of Offer #	Column Labels																							
Row Labels	Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell	Bennett	Brooks	Brown	Butler	Campbell	Carter	Clark	Collins	Cook	Cooper	Cox	Cruz	Davis	Diaz	Edwards	Evans	Fisher
1											1													1
2							1						1					1						1
3									1															
4											1			1			1							
5																								
6																								
7					1	1						1									1			
8								1	1												1	1		
9			1																					
10						1	1																	
11									1					1	1									1
12																				1				
13														1										
14																								
15																								
16																		1						
17							1												1					
18		1																						
19						1																		
20																		1						
21								1																
22								1		1		1								1			1	1

Рисунок 5 – Сводная таблица «клиент-сделка»

2 Скопируйте лист **OfferInformation** и назовите его **Matrix**. В этот новый лист вставьте значения из сводной таблицы (не нужно копировать и вставлять номер сделки, потому что он уже содержится в информации о заказе), начиная со столбца H. В итоге у вас должна получиться расширенная версия матрицы, дополненная информацией о заказах, как на рисунке 6.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Offer #	Campaign	Varietal	Minimum Qty (kg)	Discount (%)	Origin	Past Peak	Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell	Bennett
3	2	January	Pinot Noir	72	17	France	FALSE							1	
4	3	February	Espumante	144	32	Oregon	TRUE								
5	4	February	Champagne	72	48	France	TRUE								
6	5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE								
7	6	March	Prosecco	144	86	Chile	FALSE								
8	7	March	Prosecco	6	40	Australia	TRUE					1	1		
9	8	March	Espumante	6	45	South Africa	FALSE								1
10	9	April	Chardonnay	144	57	Chile	FALSE		1						
11	10	April	Prosecco	72	52	California	FALSE						1	1	
12	11	May	Champagne	72	85	France	FALSE								
13	12	May	Prosecco	72	83	Australia	FALSE								
14	13	May	Merlot	6	43	Chile	FALSE								
15	14	June	Merlot	72	64	Chile	FALSE								
16	15	June	Cabernet Sauvignon	144	19	Italy	FALSE								
17	16	June	Merlot	72	88	California	FALSE								
18	17	July	Pinot Noir	12	47	Germany	FALSE							1	
19	18	July	Espumante	6	50	Oregon	FALSE	1							

Рисунок 6 – Описание сделок и данные о заказах, слитые в единую матрицу

3 Проведите кластерный анализ для 4-х кластеров (значение k=4).

3.1 Скопируйте данные из листа **Matrix**, в новый лист и назовите его **4MC**. Вставьте четыре столбца после ценового максимума в столбцы от **H** до **K**, которые будут кластерными центрами. Назовите эти кластеры от **Cluster** 1 до **Cluster** 4. Лист **4MC** появится будет выглядеть, как показано на рисунке 7.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Offer #	Campaign	Varietal	Minimum Qty	Discount (%)	Origin	Past Peak	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Adams	Allen	Anderson
2	1	January	Malbec	72	56	France	FALSE	0,028	0,012	0,043	0,275			
3	2	January	Pinot Noir	72	17	France	FALSE	0,234	0,022	0,108	0,115			
4	3	February	Espumante	144	32	Oregon	TRUE	0,017	0,023	0,054	0,160			
5	4	February	Champagne	72	48	France	TRUE	0,016	0,026	0,130	0,174			
6	5	February	Cabernet Sau	144	44	New Zealand	TRUE	0,023	0,022	0,054	0,057			
7	6	March	Prosecco	144	86	Chile	FALSE	0,028	0,010	0,095	0,297			
8	7	March	Prosecco	6	40	Australia	TRUE	0,034	0,541	0,123	0,086			
9	8	March	Espumante	6	45	South Africa	FALSE	0,007	0,430	0,155	0,122			
10	9	April	Chardonnay	144	57	Chile	FALSE	0,014	0,014	0,141	0,114		1	
11	10	April	Prosecco	72	52	California	FALSE	0,010	0,041	0,083	0,084			
12	11	May	Champagne	72	85	France	FALSE	0,032	0,025	0,128	0,280			
13	12	May	Prosecco	72	83	Australia	FALSE	0,016	0,041	0,073	0,088			
14	13	May	Merlot	6	43	Chile	FALSE	0,026	0,194	0,016	0,011			
15	14	June	Merlot	72	64	Chile	FALSE	0,045	0,034	0,061	0,163			
16	15	June	Cabernet Sau	144	19	Italy	FALSE	0,013	0,024	0,052	0,171			
17	16	June	Merlot	72	88	California	FALSE	0,030	0,009	0,063	0,027			
18	17	July	Pinot Noir	12	47	Germany	FALSE	0,608	0,032	0,005	0,061			
19	18	July	Espumante	6	50	Oregon	FALSE	0,027	0,419	0,074	0,054	1		
20	19	July	Champagne	12	66	Germany	FALSE	0,025	0,015	0,078	0,071			
21	20	August	Cabernet Sau	72	82	Italy	FALSE	0,024	0,015	0,069	0,062			

Рисунок 7 – Пустые кластерные центры, помещенные на лист 4MC

3.2 Рассчитайте евклидовы расстояния для каждого клиента до каждого кластерного центра. В ячейках **G34:G37** впишите названия «**Distance to Cluster 1**».. «**Distance to Cluster 4**».

Например, в ячейке **L34** под заказами Адамса можно вычислить разницу между вектором Адамса и кластерным центром, возвести ее в квадрат, сложить и затем извлечь корень, используя следующую формулу для массивов (отметьте абсолютные ссылки, позволяющие вам перетаскивать эту формулу вправо или вниз без изменения ссылки на кластерный центр):

{=КОРЕНЬ(СУММА(L$2:L$33-$H$2:$H$33)A2))}.

Формулу для массивов (введите формулу и нажмите **Ctrl+Shift+Enter**) нужно использовать, потому что ее часть **(L2:L33-H2:H33)^2** должна «знать», куда обращаться для вычисления разниц и возведения их в квадрат, шаг за шагом.

Аналогично посчитайте расстояния для каждого из 4 кластеров, а затем растяните получившиеся формулы на всех клиентов. Рассчитанные расстояния показаны на рисунке 8.

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Discount (%)	Origin	Past Peak	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell	Bennett	Brooks
23	63	France	FALSE	0,009	0,024	0,023	0,951							1		
24	39	South Africa	FALSE	0,029	0,023	0,036	0,062									
25	34	Italy	FALSE	0,941	0,043	0,017	0,035			1				1		
26	59	Oregon	TRUE	0,025	0,034	0,083	0,114									
27	83	Australia	FALSE	0,690	0,030	0,090	0,130			1				1		
28	88	New Zealand	FALSE	0,010	0,021	0,087	0,141		1							
29	56	France	TRUE	0,026	0,017	0,090	0,030									
30	87	France	FALSE	0,012	0,619	0,043	0,038	1							1	
31	54	France	FALSE	0,020	0,729	0,079	0,136	1				1				
32	89	France	FALSE	0,023	0,027	0,211	0,259						1	1		
33	45	Germany	TRUE	0,093	0,013	0,053	0,125									
34			Distance to Cluster 1					2,166	1,939	0,740	1,924	2,372	2,387	0,929	1,942	2
35			Distance to Cluster 2					1,044	1,886	1,865	1,042	2,092	2,311	2,318	1,235	2
36			Distance to Cluster 3					1,691	1,339	1,428	1,359	1,806	1,875	1,953	1,362	1
37			Distance to Cluster 4					2,012	1,731	1,781	1,749	2,122	1,667	2,196	1,785	1

Рисунок 8 – Расчет расстояний от каждого покупателя до всех кластерных центров

3.3 Проведите распределение по кластерам по кратчайшему расстоянию следующим образом:

- добавьте названия строк 38 и 39 в ячейки 38 и 39 столбца G «**Minimum Cluster Distance**» и «**Assigned Cluster**»;

- рассчитайте минимальные расстояния до кластерного центра по каждому клиенту (функция МИН);

- выведите номер кластера с минимальным расстоянием для каждого клиента (используем формулу ПОИСКПОЗ).

Изначально для всех клиентов это будет кластер 1 (рисунок 9).

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	Discount (%)	Origin	Past Peak	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell	Bennett	Bro
23	63	France	FALSE	0,009	0,024	0,023	0,951						1			
24	39	South Africa	FALSE	0,029	0,023	0,036	0,062									
25	34	Italy	FALSE	0,941	0,043	0,017	0,035			1				1		
26	59	Oregon	TRUE	0,025	0,034	0,083	0,114									
27	83	Australia	FALSE	0,690	0,030	0,090	0,130			1				1		
28	88	New Zealand	FALSE	0,010	0,021	0,087	0,141		1							
29	56	France	TRUE	0,026	0,017	0,090	0,030									
30	87	France	FALSE	0,012	0,619	0,043	0,038	1							1	
31	54	France	FALSE	0,020	0,729	0,079	0,136	1			1					
32	89	France	FALSE	0,023	0,027	0,211	0,259					1	1			
33	45	Germany	TRUE	0,093	0,013	0,053	0,125									
34			Distance to Cluster 1					2,166	1,939	0,740	1,924	2,372	2,387	0,929	1,942	
35			Distance to Cluster 2					1,044	1,886	1,865	1,042	2,092	2,311	2,318	1,235	
36			Distance to Cluster 3					1,691	1,339	1,428	1,359	1,806	1,875	1,953	1,362	
37			Distance to Cluster 4					2,012	1,731	1,781	1,749	2,122	1,667	2,196	1,785	
38			Minimum Cluster Distance					1,044	1,339	0,740	1,042	1,806	1,667	0,929	1,235	
39			**Assigned Cluster**					**2**	**3**	**1**	**2**	**3**	**4**	**1**	**2**	

Рисунок 9 – Добавление на лист привязки к кластерам

3.4 Осуществите поиск решений для кластерных центов. Чтобы установить наилучшее положение кластерных центров, нужно найти такие значения в столбцах от H до K, которые минимизируют общее расстояние между покупателями и кластерными центрами, к которым они привязаны. Оптимизация в **Excel** производится с помощью надстройки «Поиск решения» (вкладка «**Данные-Анализ**»). В **Calc** оптимизация проводится с помощью встроенной функции «Решатель» в меню «Сервис».

Для этого:

3.4.1 Задайте целевую функцию в ячейке A36 (сумма минимальных расстояний от клиентов до кластеров);

3.4.2 Осуществите постановку задачи для Поиска решения (рисунок 10):

– цель: минимизировать общие расстояния от покупателей к их кластерным центрам (A36);

– переменные: вектор каждой сделки относительно кластерного центра (H2:K33);

– условия: кластерные центры должны иметь значения в пределах 0 - 1.

Поскольку евклидово расстояние – нелинейная функция, используйте эволюционный алгоритм. Также установите параметры Эволюционного алгоритма (рекомендуемое время ожидания - 600 с).

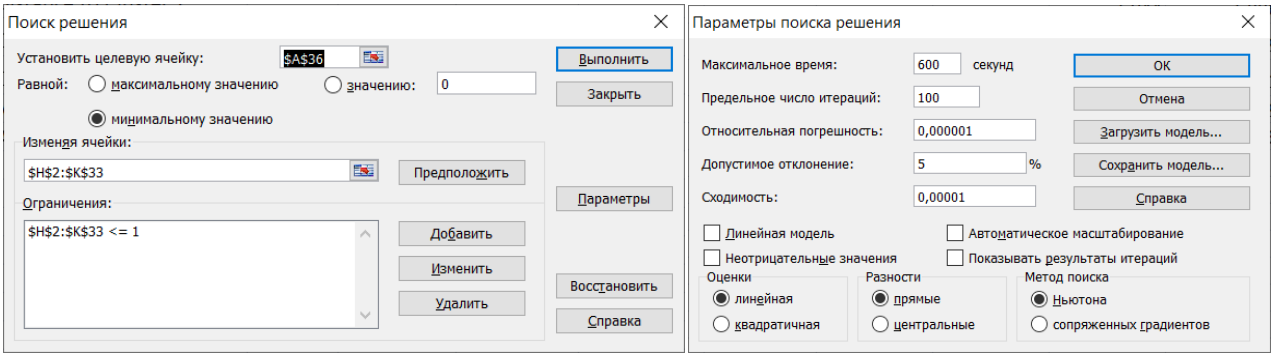


Рисунок 10 – Установки «Поиска решения» для 4-центровой кластеризации в Excel

3.5 Осуществите поиск решения (кнопка «Найти решение»). Проанализируйте результат и сделайте выводы.

3.6 Проведите рейтингование сделок по результатам кластеризации.

3.6.1 Скопируйте лист **OfferInformation**, копию назовите 4МС – **TopDealsByCluster**. Пронумеруйте столбцы от H до K на этом новом листе от 1 до 4 (как на рисунке 11).

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Offer date	Product	Minimum Qt	Discount	Origin	Past Peak	1	2	3	4
2	1	January	Malbec	72	56	France	ЛОЖЬ				
3	2	January	Pinot Noir	72	17	France	ЛОЖЬ				
4	3	February	Espumante	144	32	Oregon	ИСТИНА				
5	4	February	Champagne	72	48	France	ИСТИНА				
6	5	February	Cabernet Sa	144	44	New Zealanc	ИСТИНА				
7	6	March	Prosecco	144	86	Chile	ЛОЖЬ				
8	7	March	Prosecco	6	40	Australia	ИСТИНА				
9	8	March	Espumante	6	45	South Africa	ЛОЖЬ				
10	9	April	Chardonnay	144	57	Chile	ЛОЖЬ				
11	10	April	Prosecco	72	52	California	ЛОЖЬ				
12	11	May	Champagne	72	85	France	ЛОЖЬ				
13	12	May	Prosecco	72	83	Australia	ЛОЖЬ				
14	13	May	Merlot	6	43	Chile	ЛОЖЬ				

Рисунок 11 – Создание листа таблицы для подсчета популярности сделок с помощью кластеров

3.6.2 Посчитайте количество сделок по кластерам. Для этого, исходя из привязок клиентов по кластерам на листе 4МС, нужно сравнить названия столбцов от H до K на листе 4МС – **TopDealsByCluster** – с кластером на листе 4МС и затем сложить количество сделок в каждой строке. Используйте функцию СУММЕСЛИ. Примените условное форматирование (рисунок 12).

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	1	2	3	4
2	1	January	Malbec	72	56	France	ЛОЖЬ	0	0	4	6
3	2	January	Pinot Noir	72	17	France	ЛОЖЬ	4	0	4	2
4	3	February	Espumante	144	32	Oregon	ИСТИНА	0	0	2	4
5	4	February	Champagne	72	48	France	ИСТИНА	0	0	7	5
6	5	February	Cabernet Sa	144	44	New Zealanc	ИСТИНА	0	0	2	2
7	6	March	Prosecco	144	86	Chile	ЛОЖЬ	0	0	5	7
8	7	March	Prosecco	6	40	Australia	ИСТИНА	0	12	4	3
9	8	March	Espumante	6	45	South Africa	ЛОЖЬ	0	11	6	3
10	9	April	Chardonnay	144	57	Chile	ЛОЖЬ	0	0	7	3
11	10	April	Prosecco	72	52	California	ЛОЖЬ	0	0	5	2
12	11	May	Champagne	72	85	France	ЛОЖЬ	0	0	7	6
13	12	May	Prosecco	72	83	Australia	ЛОЖЬ	0	0	3	2
14	13	May	Merlot	6	43	Chile	ЛОЖЬ	0	6	0	0
15	14	June	Merlot	72	64	Chile	ЛОЖЬ	0	0	5	4
16	15	June	Cabernet Sa	144	19	Italy	ЛОЖЬ	0	0	2	4
17	16	June	Merlot	72	88	California	ЛОЖЬ	0	0	5	0
18	17	July	Pinot Noir	12	47	Germany	ЛОЖЬ	7	0	0	0
19	18	July	Espumante	6	50	Oregon	ЛОЖЬ	0	11	2	1
20	19	July	Champagne	12	66	Germany	ЛОЖЬ	0	0	2	3
21	20	August	Cabernet Sa	72	82	Italy	ЛОЖЬ	0	0	4	2
22	21	August	Champagne	12	50	California	ЛОЖЬ	0	0	2	2
23	22	August	Champagne	72	63	France	ЛОЖЬ	0	0	0	21

Рисунок 12 – Общее количество сделок по каждому предложению, разбитое по кластерам

3.6.3 Проанализируйте полученные данные по сделкам и сделайте предположения о клиентах каждого кластера B.

Выделяя столбцы от A до K и применяя автофильтрацию, вы можете сортировать полученные данные. Например, отсортировав от наибольшего к наименьшему столбец H, мы видим, какие сделки наиболее популярны в кластере 1 (рисунок 13). Можно сделать предположение, что клиенты 1 кластера предпочитают сорт Пино Нуар.

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	1	2	3	4
2	24	September	Pinot Noir	6	34	Italy	ЛОЖЬ	12	0	0	0
3	26	October	Pinot Noir	144	83	Australia	ЛОЖЬ	8	0	5	2
4	17	July	Pinot Noir	12	47	Germany	ЛОЖЬ	7	0	0	0
5	2	January	Pinot Noir	72	17	France	ЛОЖЬ	4	0	4	2
6	1	January	Malbec	72	56	France	ЛОЖЬ	0	0	4	6

Рисунок 13 – Сортировка кластера 1. Любители Пино

3.7 Оцените качество кластеризации по 4 кластерам. Рассчитайте силуэт.

3.7.1 Рассчитайте матрицу расстояний, для этого:

- создайте новый лист «Distances»;

- вставьте список клиентов по горизонтали и вертикали, пронумеруйте клиентов по строкам и столбцам от 0 до 99 (расположите нумерацию, соответственно, в первой строке и первом столбце);

- рассчитайте евклидовы расстояния между всеми клиентами (рисунок 14). Для удобства используйте функцию СМЕЩ. Не забывайте про формулы массивов!

Пример формулы для клетки C3:

={КОРЕНЬ(СУММ((СМЕЩ(Matrix!$H$2:$H$33;0;Distances!C$1)-СМЕЩ(Matrix!$H$2:$H$33;0;Distances!$A3))^2))}

Пример формулы для клетки E9:

={КОРЕНЬ(СУММ((СМЕЩ(Matrix!$H$2:$H$33;0;Distances!E$1)-СМЕЩ(Matrix!$H$2:$H$33;0;Distances!$A9))^2))}

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1			0	1	2	3	4	5	6	7	8	9	10	11
2			Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell	Bennett	Brooks	Brown	Butler	Campbell
3	0	Adams	0,000	2,236	2,236	1,732	2,646	2,646	2,646	1,732	2,646	1,414	2,449	2,449
4	1	Allen	2,236	0,000	2,000	2,000	2,449	2,449	2,449	2,000	2,449	2,236	2,646	2,236
5	2	Anderson	2,236	2,000	0,000	2,000	2,449	2,449	1,414	2,000	2,449	2,236	2,646	1,000
6	3	Bailey	1,732	2,000	2,000	0,000	2,000	2,449	2,449	2,000	2,449	1,000	2,236	2,236
7	4	Baker	2,646	2,449	2,449	2,000	0,000	2,000	2,828	2,449	2,828	2,236	3,000	2,646
8	5	Barnes	2,646	2,449	2,449	2,449	2,000	0,000	2,828	2,449	2,449	2,646	2,646	2,646
9	6	Bell	2,646	2,449	1,414	2,449	2,828	2,828	0,000	2,449	2,828	2,646	3,000	1,000
10	7	Bennett	1,732	2,000	2,000	2,000	2,449	2,449	2,449	0,000	2,000	1,732	2,646	2,236
11	8	Brooks	2,646	2,449	2,449	2,449	2,828	2,449	2,828	2,000	0,000	2,646	2,646	2,646
12	9	Brown	1,414	2,236	2,236	1,000	2,236	2,646	2,646	1,732	2,646	0,000	2,449	2,449
13	10	Butler	2,449	2,646	2,646	2,236	3,000	2,646	3,000	2,646	2,646	2,449	0,000	2,828
14	11	Campbell	2,449	2,236	1,000	2,236	2,646	2,646	1,000	2,236	2,646	2,449	2,828	0,000
15	12	Carter	1,732	2,449	2,449	1,414	2,449	2,828	2,828	2,000	2,828	1,000	2,646	2,646
16	13	Clark	2,646	2,449	2,449	2,449	2,449	2,449	2,828	2,449	2,449	2,646	2,236	2,646
17	14	Collins	1,732	2,000	2,000	1,414	2,449	2,449	2,449	2,000	2,000	1,732	2,236	2,236
18	15	Cook	2,236	2,000	0,000	2,000	2,449	2,449	1,414	2,000	2,449	2,236	2,646	1,000
19	16	Cooper	2,646	2,449	2,449	2,449	2,828	2,828	2,828	2,449	2,828	2,646	2,646	2,646
20	17	Cox	2,646	2,449	1,414	2,449	2,828	2,828	0,000	2,449	2,828	2,646	3,000	1,000
21	18	Cruz	1,000	2,000	2,000	1,414	2,449	2,449	2,449	1,414	2,449	1,000	2,236	2,236
22	19	Davis	2,449	2,236	2,236	2,236	2,646	2,236	2,646	2,236	2,236	2,449	2,449	2,449

Рисунок 14 – Матрица расстояний

3.7.2 Рассчитайте силуэт, для этого:

3.7.2.1 Создайте новый лист 4MC Silhouette. Скопируйте с листа 4MC имена клиентов в столбец A, а привязки к кластерам (значения) в столбец B. Озаглавьте столбцы с C до F «Distance from people in 1» … «Distance from people in 4».

3.7.2.2 Рассчитайте средние расстояния для каждого клиента между ним и другими клиентами, входящими в конкретный кластер. Используйте функцию СРЗНАЧЕСЛИ.

Например, для Адамса, формулы будут иметь вид:

=СРЗНАЧЕСЛИ('4MC'!$L$39:$DG$39;1;Distances!$C3:$CX3)
=СРЗНАЧЕСЛИ('4MC'!$L$39:$DG$39;2;Distances!$C3:$CX3)
=СРЗНАЧЕСЛИ('4MC'!$L$39:$DG$39;3;Distances!$C3:$CX3)
=СРЗНАЧЕСЛИ('4MC'!$L$39:$DG$39;4;Distances!$C3:$CX3).

3.7.2.3 В столбце G (Заголовок столбца – Closest) найдите ближайшую группу покупателей. Используйте функцию МИН.

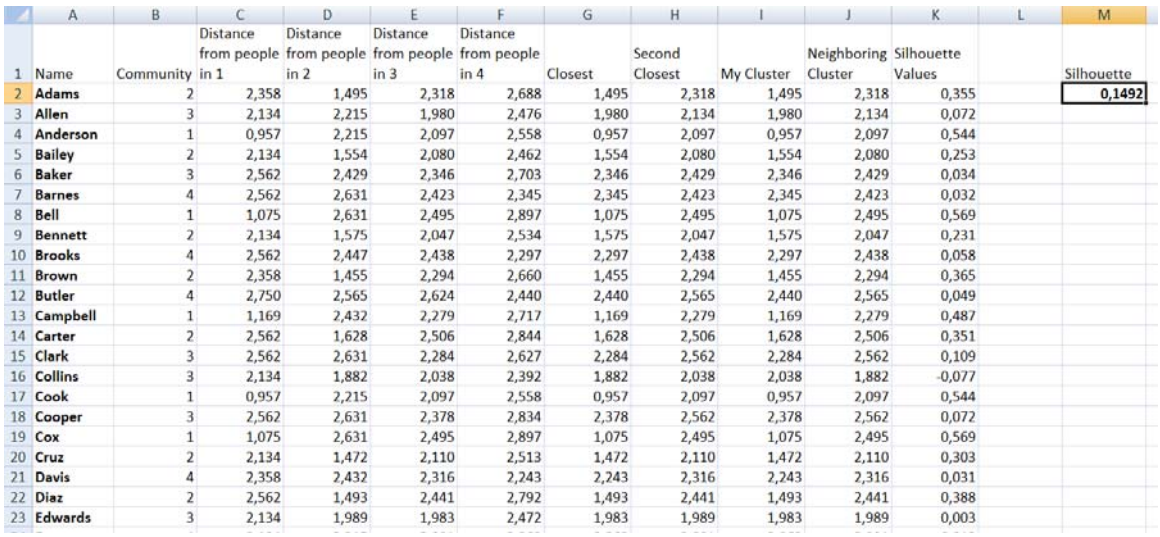
3.7.2.4 В столбце H (Заголовок столбца – Second Closest) найдите вторую по близости группу покупателей. Используйте функцию НАИМЕНЬШИЙ.

3.7.2.5 В столбце I (Заголовок столбца – My Cluster) найдите расстояние до членов собственного кластера. Используйте функцию ИНДЕКС.

3.7.2.6 В столбце J (Заголовок столбца – Neighboring Cluster) найдите расстояние до ближайшего группы покупателей, находящихся не в вашем кластере. Если собственное кластерное расстояние равно расстоянию ближайшего кластера, то ответ в столбце H (Second Closest), если нет – в столбце G (Closest). Используйте функцию ЕСЛИ.

3.7.2.7 Рассчитайте силуэты по каждому клиенту по формуле $s=(b-a)/\max(a,b)$, где a – My Cluster, B – Neighboring Cluster. Проанализируйте получившиеся значения.

3.7.2.8 Рассчитайте Итоговый силуэт как среднее значение всех силуэтов по клиентам (рисунок 15). Проанализируйте получившееся значение.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Name	Community	Distance from people in 1	Distance from people in 2	Distance from people in 3	Distance from people in 4	Closest	Second Closest	My Cluster	Neighboring Cluster	Silhouette Values		Silhouette
2	Adams	2	2,358	1,495	2,318	2,688	1,495	2,318	1,495	2,318	0,355		0,1492
3	Allen	3	2,134	2,215	1,980	2,476	1,980	2,134	1,980	2,134	0,072		
4	Anderson	1	0,957	2,215	2,097	2,558	0,957	2,097	0,957	2,097	0,544		
5	Bailey	2	2,134	1,554	2,080	2,462	1,554	2,080	1,554	2,080	0,253		
6	Baker	3	2,562	2,429	2,346	2,703	2,346	2,429	2,346	2,429	0,034		
7	Barnes	4	2,562	2,631	2,423	2,345	2,345	2,423	2,345	2,423	0,032		
8	Bell	1	1,075	2,631	2,495	2,897	1,075	2,495	1,075	2,495	0,569		
9	Bennett	2	2,134	1,575	2,047	2,534	1,575	2,047	1,575	2,047	0,231		
10	Brooks	4	2,562	2,447	2,438	2,297	2,297	2,438	2,297	2,438	0,058		
11	Brown	2	2,358	1,455	2,294	2,660	1,455	2,294	1,455	2,294	0,365		
12	Butler	4	2,750	2,565	2,624	2,440	2,440	2,565	2,440	2,565	0,049		
13	Campbell	1	1,169	2,432	2,279	2,717	1,169	2,279	1,169	2,279	0,487		
14	Carter	2	2,562	1,628	2,506	2,844	1,628	2,506	1,628	2,506	0,351		
15	Clark	3	2,562	2,631	2,284	2,627	2,284	2,562	2,284	2,562	0,109		
16	Collins	3	2,134	1,882	2,038	2,392	1,882	2,038	2,038	1,882	-0,077		
17	Cook	1	0,957	2,215	2,097	2,558	0,957	2,097	0,957	2,097	0,544		
18	Cooper	3	2,562	2,631	2,378	2,834	2,378	2,562	2,378	2,562	0,072		
19	Cox	1	1,075	2,631	2,495	2,897	1,075	2,495	1,075	2,495	0,569		
20	Cruz	2	2,134	1,472	2,110	2,513	1,472	2,110	1,472	2,110	0,303		
21	Davis	4	2,358	2,432	2,316	2,243	2,243	2,316	2,243	2,316	0,031		
22	Diaz	2	2,562	1,493	2,441	2,792	1,493	2,441	1,493	2,441	0,388		
23	Edwards	3	2,134	1,989	1,983	2,472	1,983	1,989	1,983	1,989	0,003		

Рисунок 15 – Средние расстояния до клиентов из всех кластеров и силуэты

4 Проведите кластерный анализ для 5 кластеров (значение k=5). Этапы выполнения аналогичны. Результаты должны быть представлены в листах, соответственно, 5MC, 5MC –TopDealsByCluster, 5MC Silhouette.

5 Проведите кластерный анализ методом k-медиан, используя расстояние по косинусу.

5.1 Скопируйте лист 5MC и переименуйте его в 5MedC и удалите данные, рассчитанные с помощью Поиска решения (кластерные центры).

5.2 Рассчитайте расстояние по косинусу в строках 34-38. Близость по косинусу: косинус угла между двумя бинарными заказами — это число совпадений заказов в двух векторах, разделенное на произведение квадратных корней количества заказов первого и второго векторов. Расстояние по косинусу = 1–Близость по косинусу.

Чтобы найти значение совпадающих заказов у клиента и кластера, используйте функцию СУММПРОИЗВ.

Вставьте в формулу проверку на ошибку, если кластерный центр окажется равным 0. В этом случае присвойте ему значение 1.

Таким образом, формула для расчёта расстояния по косинусу от Адамса до кластера 1 имеет вид:

=ЕСЛИОШИБКА(1-СУММПРОИЗВ(M$2:M$33;$H$2:$H$33)/
/(КОРЕНЬ(СУММ(M$2:M$33))*КОРЕНЬ(СУММ($H$2:$H$33)));1).

5.3 Осуществите поиск решения, для этого замените условие <=1 на бинарное (рисунок 16). Проанализируйте результат и сделайте выводы.

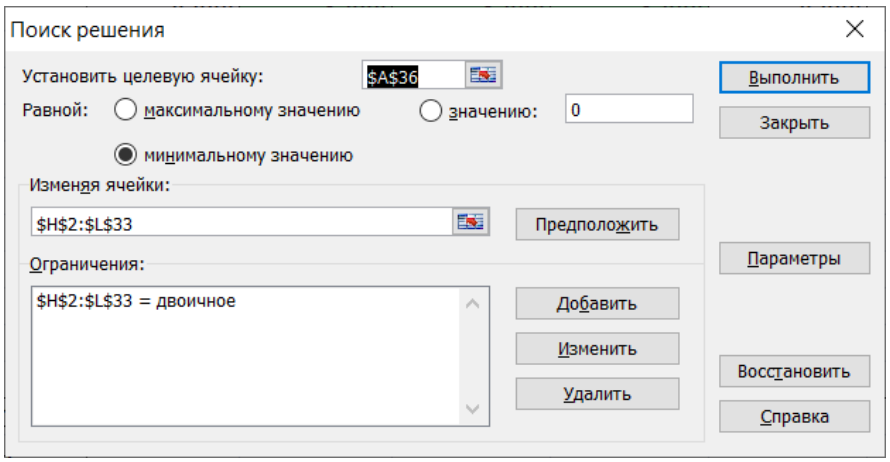


Рисунок 16 – Постановка задачи поиска решения

5.4 Рассчитайте рейтинг сделок для 5-медианных кластеров. Для этого скопируйте лист 5C – TopDealsByCluster и переименуйте его в 5MedC – TopDealsByCluster. Измените ссылки на листы в формулах.

5.5 Проанализируйте полученные данные по сделкам и сделайте предположения о клиентах каждого кластера.

5.6 Сравните результаты кластеризации по k-средним и k-медианам.

6.2 Реализация алгоритма k-средних в языке R

В языке **R** для кластеризации методом **k** внутригрупповых средних используется функция **kmeans** из пакета **stats**. Она позволяет производить кластеризацию данных различными методами. Ниже приведена сигнатура этой функции.

kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"), trace=FALSE),

где **x** – матрица данных, каждая строка которой является вектором признаков очередного объекта наблюдения.

centers – количество кластеров **k** или набор начальных центров кластеров. Если набор начальных центров не задан, то в качестве начальных центров выбираются **k** случайных объектов наблюдения.

iter.max – максимальное количество итераций алгоритма.

nstart – количество наборов случайных начальных центров кластеров в случае, если центры выбираются случайно.

algorithm – реализация алгоритма **k** внутригрупповых средних, с помощью которой будет производиться кластеризация. Допустимые значения: «**Lloyd**» или «**Forgy**» – алгоритм Ллойда-Форджи, «**MacQueen**» – алгоритм Маккуина, «**Hartigan-Wong**» – алгоритм Хартигана-Вонга.

trace – уровень протоколирования для алгоритма Хартигана-Вонга. Чем выше значение, тем больше сопроводительной информации выводится.

Функция возвращает объект класса **kmeans**, который имеет методы **print** и **fitted**. Этот объект представляет собой список, содержащий следующие элементы.

cluster – вектор из n целых положительных чисел, обозначающих номер кластера соответствующего объекта наблюдения.
centers – матрица, строки которой представляют собой центры соответствующих кластеров.
totss – сумма квадратов расстояний от объектов наблюдения до соответствующих центров кластеров.
withinss – вектор внутрикластерных сумм квадратов расстояний между объектами одного кластера для каждого кластера.
tot.**withinss** – общая сумма внутрикластерных сумм квадратов расстояний между объектами, то есть сумма элементов вектора **withinss**.
betweenss – междукластерная сумма квадратов расстояний, то есть разница между **totss** и **tot**.**withinss**.
size – количество объектов наблюдения, попавших в каждый из кластеров.
iter – количество внешних итераций, совершённых в ходе работы алгоритма.
ifault – код ошибки, возникшей в ходе работы алгоритма.

Доступ к этим значениям можно получить путём записи их названия в двойных квадратных скобках справа от переменной, содержащей модель, либо через знак доллара. Например, «**a[["cluster"]]**» или «**a$cluster**».

Часто задача кластеризации, описанная в разделе 1, возникает в несколько более сложной постановке: точное число кластеров **k** может быть не известно.

В этом случае использование алгоритма **k** внутригрупповых средних не представляется возможным. Конечно, можно просто перебирать значения **k** и для каждого из них использовать этот алгоритм, но сравнение среднеквадратических расстояний до центров кластеров (1) для различных количеств кластеров лишено смысла.

Вместо этого используются алгоритмы иерархической кластеризации (англ. **hierarchical clustering**), которые разбивают множество объектов наблюдения на кластеры, которые в свою очередь также разбиваются на кластеры и так далее. Получившийся граф, вершинами которого являются кластеры, а дуги направлены от кластеров более высокого уровня к соответствующим кластерам более низкого уровня, называется ***дендрограммой***.

Построенная дендрограмма содержит информацию о близости между отдельными подмножествами объектов наблюдения, что позволяет подбирать число кластеров и производить более тщательный кластерный анализ в отдельных случаях.

Иерархическая кластеризация, как и обыкновенная, производится на основе расстояний **ρ(x, y)** между объектами наблюдения из заданной выборки. Эти расстояния также могут быть определены по-разному. На их основе строятся более сложные показатели качества кластеризации, которые допускают сравнение для различного числа кластеров. Далее с помощью различных методов, основанных на последовательном объединении и разбиении имеющихся кластеров, строится дендрограмма.

Существует ряд различных методов иерархической кластеризации. Ниже приведены названия некоторых из этих методов.

1 Метод Уорда (англ. **Ward’s method**).

2 Метод одиночной связи (англ. **single linkage**).
3 Метод полной связи (англ. **complete linkage**).
4 Метод средней связи (англ. **pair-group method using arithmetic averages**).
5 Метод МакКуитти (англ. **McQuitty's method**).
6 Метод медиан.
7 Центроидный метод (англ. **pair-group method using the centroid average**).

В языке R для вычисления расстояния между всеми парами объектов наблюдения из заданного блока данных используется функция **dist** из пакета **stats**. Эта функция возвращает нижнюю треугольную матрицу расстояний между всеми соответствующими парами объектов. Сигнатура этой функции описана ниже.

dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)

x – числовая матрица или блок данных.
method – метрика, используемая для вычисления расстояния. Может принимать одно из значений, указанных в таблице 1.

Таблица 1 – Метрики, используемые для вычисления расстояний в R

Название	Название в R	Формула
Евклидово расстояние	`eucliden`	$\rho_2(x,y)=\sqrt{\sum_{j=1}^{m}(x_j-y_j)^2}$
Максимум разницы компонент	`maximum`	$\rho_\infty(x,y)=\max_{j\in[1;m]\cap\mathbf{Z}}\lvert x_j-y_j\rvert$
Манхэттенское расстояние	`manhattan`	$\rho_1(x,y)=\sum_{j=1}^{m}\lvert x_j-y_j\rvert$
Для неотрицательных значений	`canberra`	$\rho_c(x,y)=\sum_{j=1}^{m}\frac{\lvert x_j-y_j\rvert}{\lvert x_j+y_j\rvert}$
Бинарное расстояние	`binary`	$\rho_b(x,y)=\frac{\lvert\{j\in[1,m]\cap\mathbf{Z}\mid x_j\wedge y_j\}\rvert}{\lvert\{j\in[1,m]\cap\mathbf{Z}\mid x_j\vee y_j\}\rvert}$
Расстояние Минковского	`minkowski`	$\rho_p(x,y)=\left(\sum_{j=1}^{m}(x_j-y_j)^p\right)^{1/p}$

diag – логическое значение, определяющее, выводить ли диагональ матрицы при выводе матрицы на экран.
upper – логическое значение, определяющее, выводить ли верхний треугольник матрицы при выводе матрицы на экран.
p – степень в расстоянии Минковского.

Для иерархической кластеризации в **R** предназначена функция **hclust** из пакета **stats**. Она принимает блок данных и возвращает объект класса **hclust**, содержащий информацию о результатах кластеризации. Сигнатура этой функции приведена ниже:

hclust(d, method = "complete", members = NULL),

где **d** – матрица расстояний, полученная с помощью функции dist.

method – используемый метод иерархической кластеризации. Может принимать одно из значений «**ward**.**D**» – метод Уорда, «**ward**.**D2**» – другой метод Уорда, «**single**» – метод одиночной связи, «**complete**» – метод полной связи, «**average**» средней связи, «**mcquitty**» – метод МакКуитти, «**median**» медиан или «**centroid**» – центроидный метод.

members – **NULL** или вектор того же размера, что и d, позволяющий отдельно начать иерархическую кластеризацию конкретного участка дендрограммы.

Для графического изображения дендрограммы можно просто передать результат, возвращаемый функцией **hclust** в функцию **plot**. Например, для кластеризации штатов США по уровню преступности можно использовать команду

plot(hclust(dist(USArrests))).

На рисунке 17 приведена дендрограмма, получающаяся в результате выполнения такой команды.

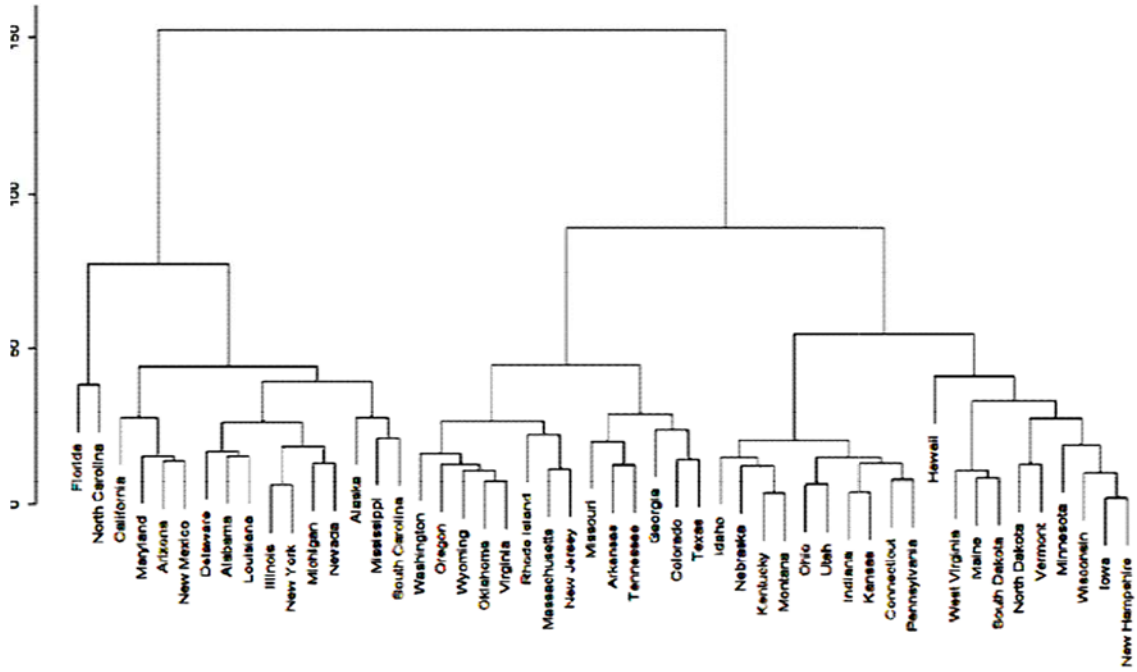


Рисунок 17 – Иерархическая кластеризация штатов США по уровню преступности

Рассмотрим пример.

Входные данные: **n** объектов, каждый из которых характеризуется двумя числовыми признаками $\{x_i\}_{i=1}^n$ и $\{y_i\}_{i=1}^n$, а также номером класса $\{c_i\}_{i=1}^n$.

Требуется исследовать работу алгоритмов кластеризации объектов наблюдения по двум признакам. Для каждого набора данных требуется выполнить следующие задания.

1 Провести кластеризацию объектов наблюдения с помощью алгоритма k внутригрупповых средних.

2 Графически изобразить на плоскости разбиения объектов наблюдения в соответствии с кластерами и в соответствии с классами c_i. Также отметить центры каждого кластера. Количество кластеров должно соответствовать количеству классов.
3 Для разбиения на кластеры вычислить сумму квадратов расстояний (1) от каждого объекта наблюдения до центра соответствующего кластера.

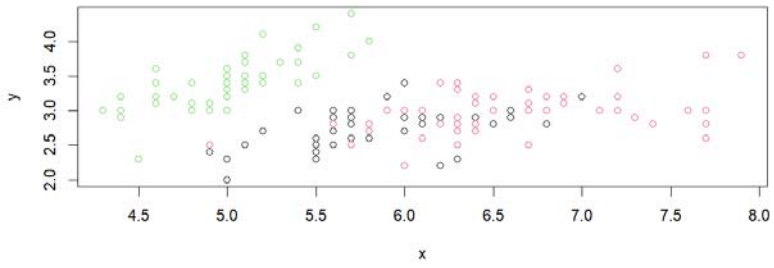
Все описанные задания требуется выполнить для двух наборов данных.
1 Смоделированные независимые случайные векторы (X, Y), n_1 из которых относятся к первому классу, а n_2 – ко второму классу. Векторы, относящиеся к первому классу, распределены по гауссовскому закону с математическим ожиданием a_1 и корреляционной матрицей R_1, а векторы, относящиеся ко второму классу, – по гауссовскому закону с математическим ожиданием a_2 и корреляционной матрицей R_2.
2 Реальные статистические данные из заданного набора (выдаются преподавателем).

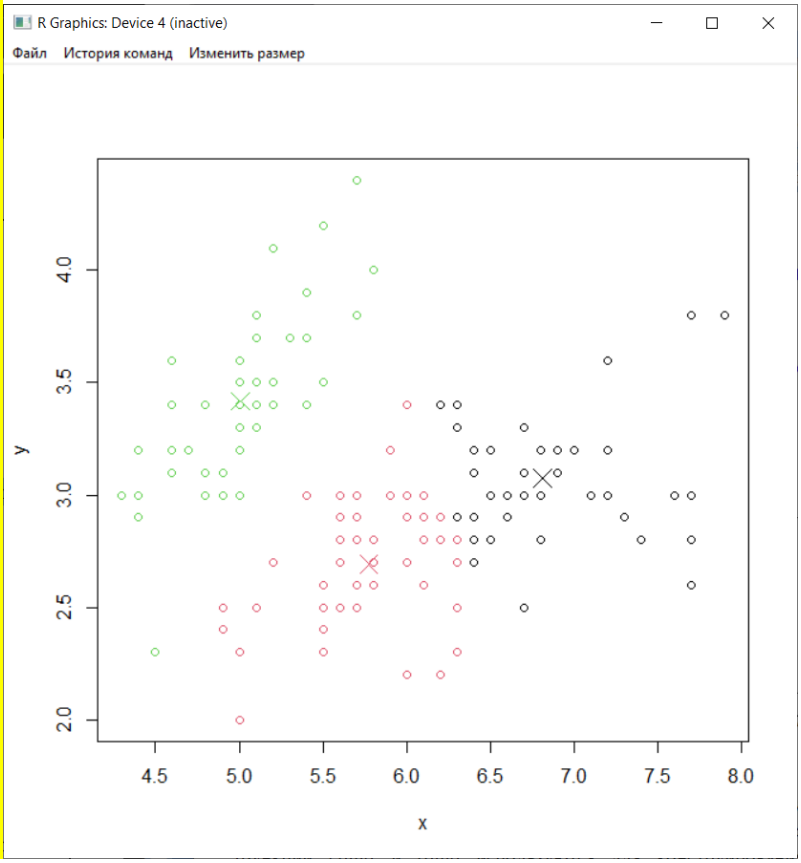
Отчет кроме прочих обязательных элементов должен включать:
1) изображения данных в виде точек на плоскости, причем данные из разных классов должны изображаться отличающимися друг от друга;
2) изображения результатов кластеризации данных в виде точек на плоскости, причем данные из разных кластеров должны изображаться отличающимися друг от друга, кроме того, требуется отметить центры кластеров;
3) значения суммы квадратов расстояний (1) от каждого объекта наблюдения до центра соответствующего кластера.

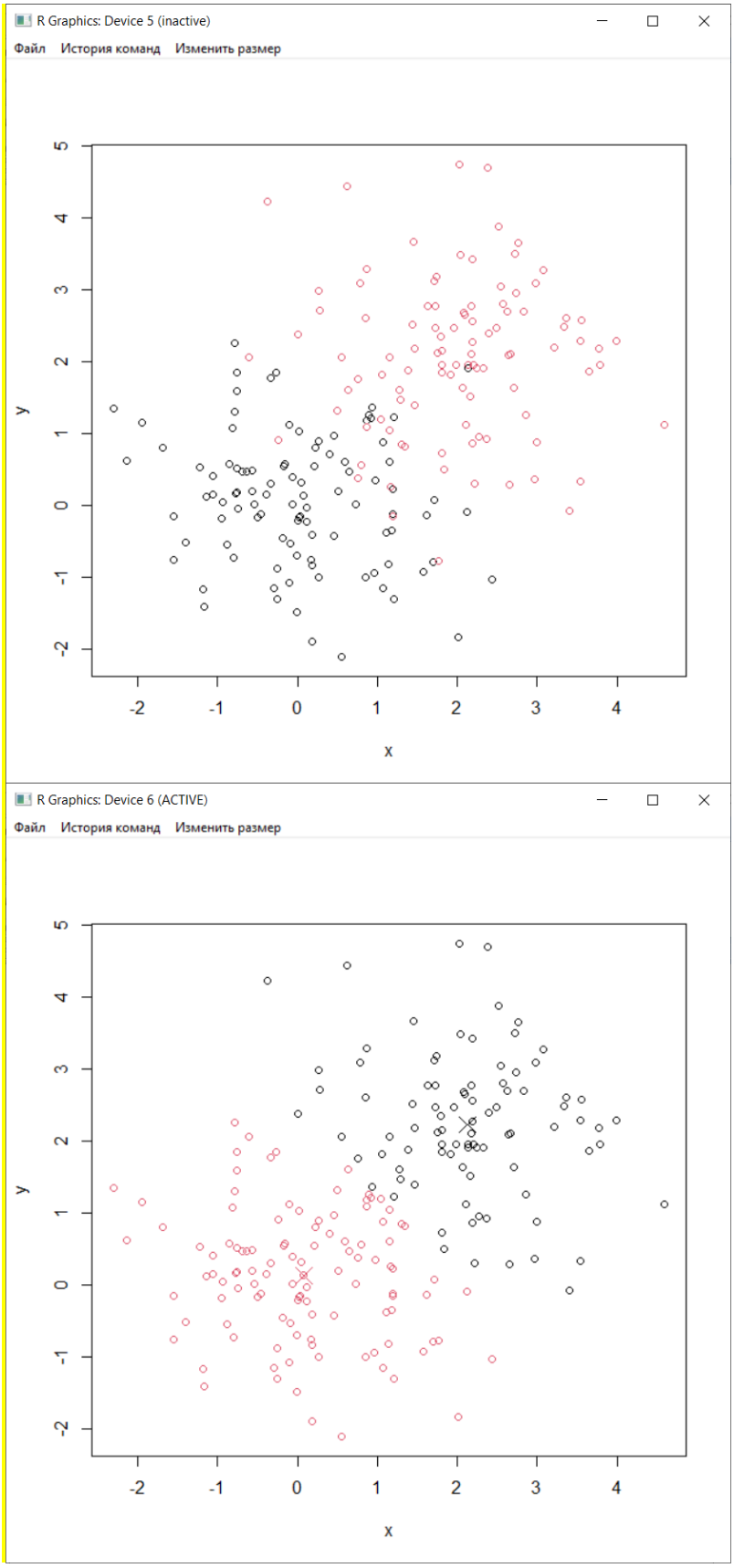
Пример реализации кластерного анализа.
В листинге 1 приведён пример выполнения нулевого варианта на языке R.

```
> require(MASS)
Загрузка требуемого пакета: MASS
>
> analyse_clust <- function(x, y, clazz) {
+     k <- length(unique(clazz))
+     clust <- kmeans(cbind(x, y), k)
+     print(clust$totss)
+     dev.new()
+     plot(x, y, col=as.factor(clazz))
+     dev.new()
+     plot(x, y, col=as.factor(clust$cluster))
+     points(clust$centers, col=1:length(clust$centers), pch=4, ce
x=2)
+ }
>
> dat <- read.csv("00-iris.txt",sep=';')
> analyse_clust(dat$Sepal.Length, dat$Sepal.Width,as.factor(dat$Sp
ecies))
[1] 130.1809
> n1 <- 100
> a1 <- c(0, 0)
> r1 <- cbind(c(1, 0), c(0, 1))
> n2 <- 100
> a2 <- c(2, 2)
> r2 <- cbind(c(1, 0), c(0, 1))
> dat <- rbind(mvrnorm(n1, a1, r1), mvrnorm(n2, a2, r2))
> analyse_clust(dat[,1], dat[,2], c(rep(1, n1), rep(2,
+                                                   n2)))
[1] 775.5319
```

Листинг 1 – Пример кластерного анализа данных на языке R


y
2.0 2.5 3.0 3.5 4.0
4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0
x


R Graphics: Device 4 (inactive)
Файл История команд Изменить размер
y
2.0 2.5 3.0 3.5 4.0
4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0
x



Для определения количества различных классов используется функция **unique**, возвращающая вектор из уникальных значений, содержащихся во входном векторе. Функции **cbind** и **rbind** используются для конструирования больших матриц из подматриц: **cbind** объединяет матрицы по столбцам, **rbind** – по строкам.

При рисовании графиков с помощью функции **plot** можно передавать именованный параметр **col**, отвечающий за цвет каждой отображаемой точки. Функция **points** используется для наложения центров кластеров на уже имеющийся график. Она принимает, кроме всего прочего, форму точек **pch** (4 означает кресты) и модификатор размера **cex**.

Синтаксическая конструкция **a:b** в **R** возвращает вектор из последовательных целых чисел от **a** до **b** включительно. Векторы из нескольких одинаковых элементов получаются с помощью функции **rep**. Функция **as.factor** представляет объект в виде фактора, то есть занумерованного набора данных. Если он состоял из строк, то эти строки нумеруются в том порядке, в котором они встречались в наборе, причём одинаковые строки получают одинаковый номер.

Варианты заданий

В таблице 2 для каждого варианта задания приведены значения количеств объектов наблюдения для каждого класса (n_1 и n_2), значения векторов математических ожиданий для каждого класса (a_1 и a_2) и корреляционные матрицы для каждого класса (R_1 и R_2) для моделируемой выборки из гауссовских случайных векторов.

Таблица 2 – Варианты задания для моделирования данных

№	n_1	a_1	R_1	n_2	a_2	R_2
0	100	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$	100	$\begin{pmatrix}2\\2\end{pmatrix}$	$\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$
1	100	$\begin{pmatrix}-1\\-1\end{pmatrix}$	$\begin{pmatrix}1 & 0{,}5\\0{,}5 & 2\end{pmatrix}$	200	$\begin{pmatrix}2\\2\end{pmatrix}$	$\begin{pmatrix}2 & 0{,}5\\0{,}5 & 1\end{pmatrix}$
2	1000	$\begin{pmatrix}-1\\0\end{pmatrix}$	$\begin{pmatrix}1 & 0{,}9\\0{,}9 & 2\end{pmatrix}$	2000	$\begin{pmatrix}2\\3\end{pmatrix}$	$\begin{pmatrix}2 & 0{,}9\\0{,}9 & 1\end{pmatrix}$
3	100	$\begin{pmatrix}-1\\1\end{pmatrix}$	$\begin{pmatrix}1 & 0{,}1\\0{,}1 & 2\end{pmatrix}$	50	$\begin{pmatrix}2\\4\end{pmatrix}$	$\begin{pmatrix}2 & 0{,}1\\0{,}1 & 1\end{pmatrix}$
4	1000	$\begin{pmatrix}0\\-1\end{pmatrix}$	$\begin{pmatrix}1 & -0{,}5\\-0{,}5 & 2\end{pmatrix}$	500	$\begin{pmatrix}3\\2\end{pmatrix}$	$\begin{pmatrix}2 & -0{,}5\\-0{,}5 & 1\end{pmatrix}$

Варианты реальных наборов данных.

0 Ирисы Фишера

Название файла: 00-iris.txt

Ссылка: https://archive.ics.uci.edu/ml/datasets/Iris

Первый признак: Sepal.Length (столбец № 2)

Второй признак: Sepal.Width (столбец № 3)

Класс: Species (столбец № 6)

1 Soybean
Название файла: 01-soybean.txt
Первый признак: protein (столбец № 10)
Второй признак: oil (столбец № 11)
Класс: Loc (столбец № 3)
2 Данные о новорожденных
Название файла: 02-birth.txt
Первый признак: Head (столбец № 6)
Второй признак: Chest (столбец № 7)
Класс: Sex (столбец № 2)
3 Потребление пищи в Дании в 1985 году
Название файла: 03-vitamina.txt
Первый признак: Avit (столбец № 9)
Второй признак: Cvit (столбец № 20)
Класс: Sex (столбец № 4)
4 Physical Growth of California Boys and Girls
Название файла: 04-physical-growth.txt
Первый признак: WT2 (столбец № 3)
Второй признак: WT9 (столбец № 5)
Класс: Sex (столбец № 2)

6.3 Кластеризация в DEDUCTOR

В качестве функции расстояния **k-means** в **Deductor** использует:

– для непрерывных числовых полей, а также упорядоченных категориальных признаков – евклидово расстояние. Использует для вычисления расстояний следующее правило:

$$d_E(x,y) = \sqrt{\sum_i (x_i - y_i)^2};$$

где $\mathbf{x = (x_1, x_2,..., x_m)}$, $\mathbf{y = (y_1, y_2,..., y_m)}$ – наборы (векторы) значений признаков двух записей.

Поскольку множество точек, равноудаленных от некоторого центра при использовании евклидовой метрики будет образовывать сферу (или круг в двумерном случае), то и кластеры, полученные с использованием евклидова расстояния, также будут иметь форму, близкую к сферической;

– для неупорядоченных категориальных признаков – функцию отличия:

$$d(x,y) = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases}.$$

В **Deductor Studio** для автоматизации этого процесса есть соответствующий инструмент – **Кластеризация**.

1) рассмотрим механизм кластеризации, реализованный на алгоритме

k-means, основываясь на данных роста численности населения по регионам.

Исходная таблица находится в файле **Population.txt**. Задача состоит в распределении регионов на функциональные группы по демографической картине в них и выявлении скрытых закономерностей.

Вначале необходимо осуществить импорт рассматриваемых данных из файла.

2) после этого выбираем и запускаем **Мастер обработки Кластеризация**.

При запуске **Мастера** необходимо настроить назначения столбцов, т. е. выбрать свойства, по которым будет происходить группировка объектов (рисунок 18). Укажем столбцам назначение соответственно рисунку.

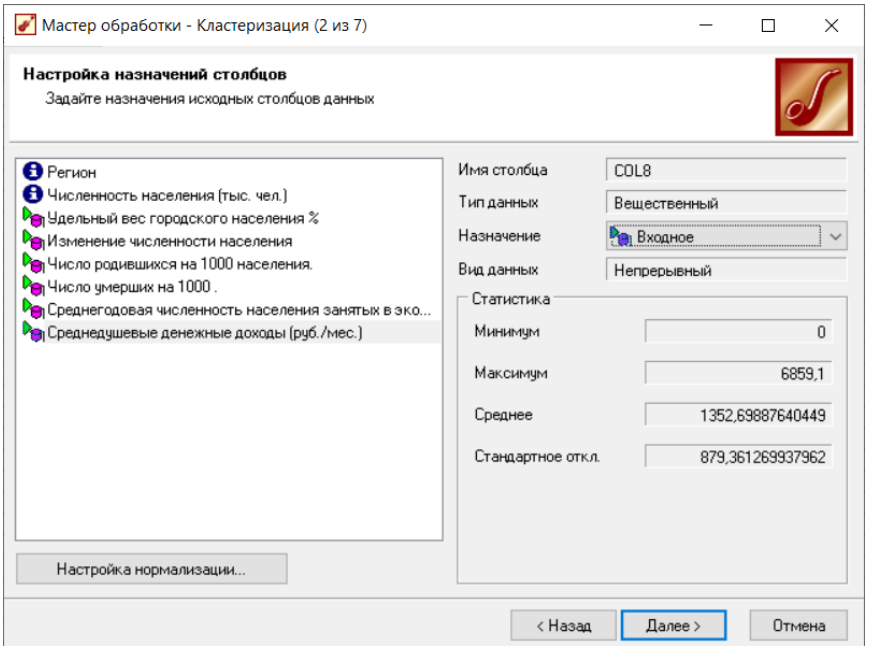


Рисунок 18 – Настройка назначения столбцов

3) на следующем шаге **Мастера** необходимо настроить способ разделения исходного множества данных на тестовое и обучающее, а также количество примеров в том и другом множестве. Укажем, что данные обоих множеств берутся случайным образом, и определим все множество как обучающее (рисунок 19).

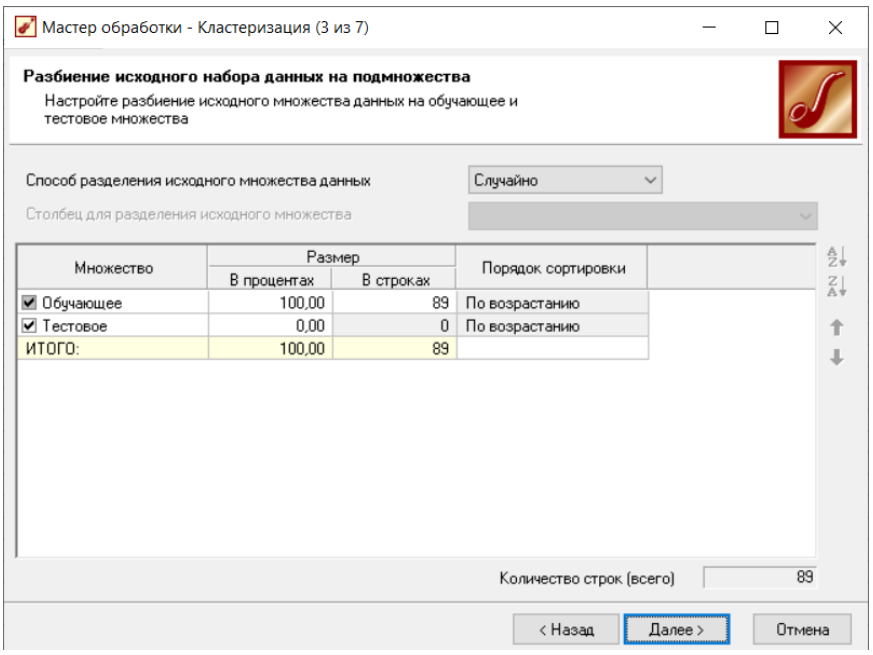


Рисунок 19 – Разбиение исходного набора данных на подмножества

4) следующий шаг предлагает настроить параметры кластеризации, определить, на какое количество кластеров будет распределяться исходное множество. По мнению экспертов в стране наблюдается четыре тенденции развития регионов, поэтому выберем фиксированное количество кластеров, равное четырем (рисунок 20).

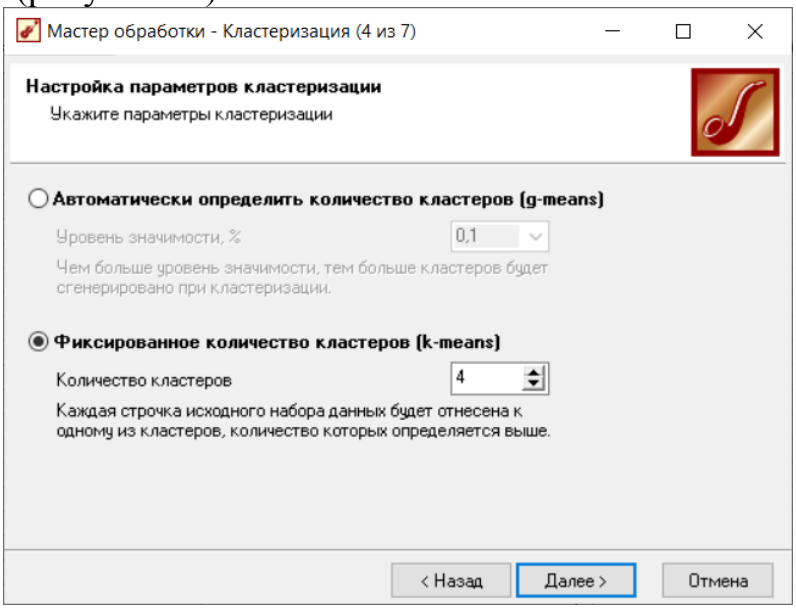


Рисунок 20 – Настройка параметров кластеризации

5) для отображения полученных групп кластеров выберем в обработчике Кластеризация из списка визуализаторов способы отображения данных: **Что-если** для решения задачи классификации, отнесение региона к одному из кластеров, Профили кластеров – для определения структуры формирования группы кластеров и **Куб** – для наглядного просмотра полученных результатов (рисунок 21).

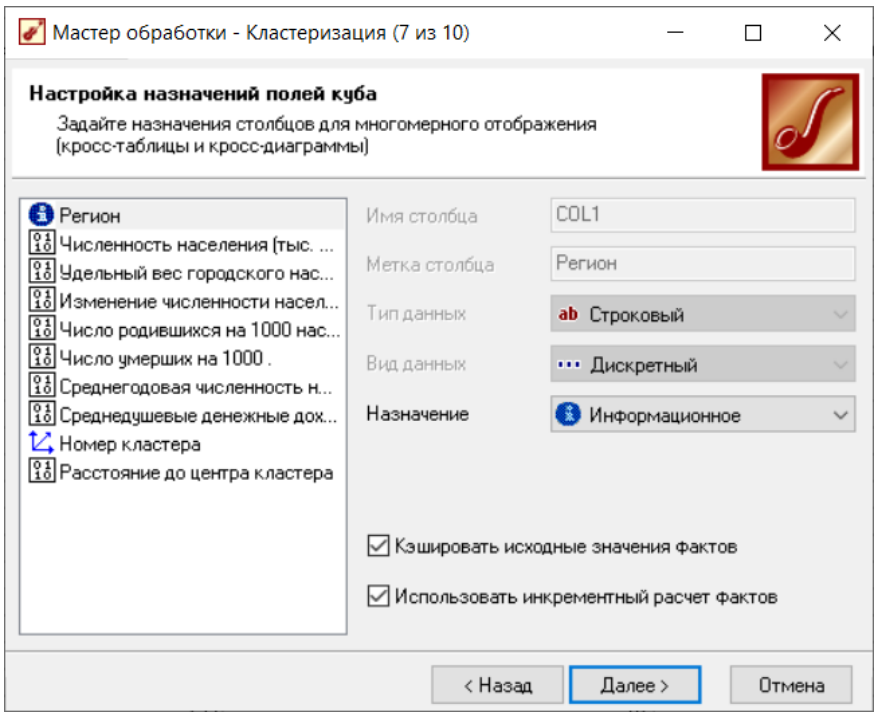


Рисунок 21– Настройка полей визуализатора Куб

Для настройки визуализатора Куб необходимо выбрать рассматриваемые свойства как факты, а Номер кластера – как измерение и Регионы – информационное.

Наиболее правильно в дальнейших настройках задать отображение всех фактов как среднее по рассматриваемое группе (рисунок 22).

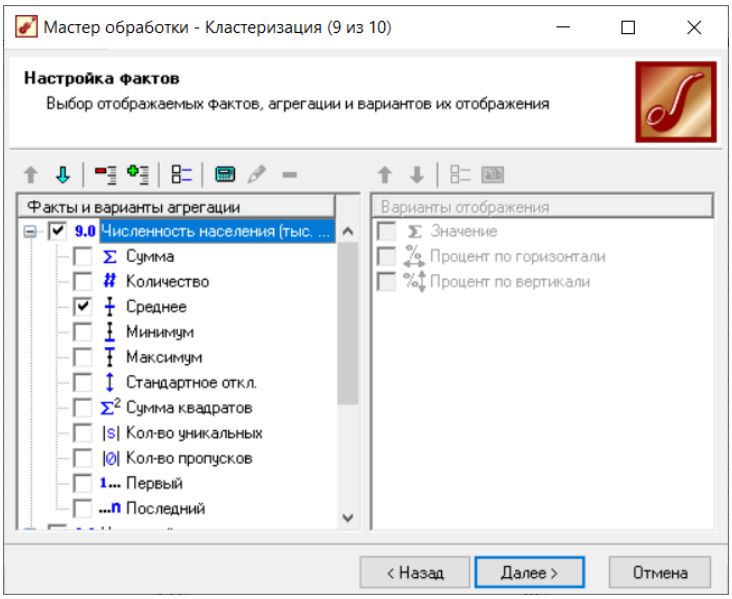


Рисунок 22 – Настройка отображаемых фактов

6) общую структуру сформированных алгоритмом кластеров можно просмотреть в визуализаторе Профили кластеров. Отсортировав кластеры по убыванию, получаем следующие профили кластеров (рисунок 23).

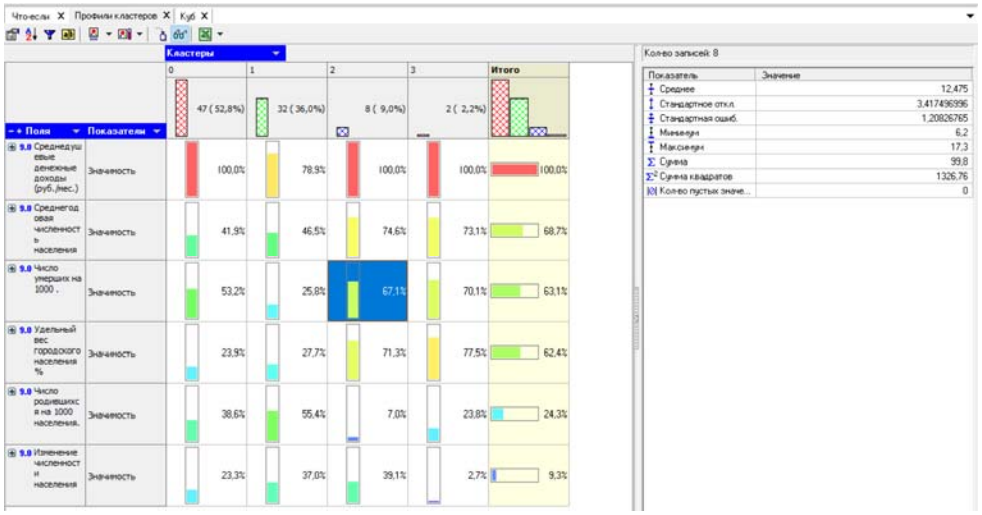


Рисунок 23 – Результат сортировки кластеров по убыванию

В визуализаторе представлены все рассматриваемые свойства вместе с характером влияния их на состав кластера. Основным, определяющим состав кластера фактором является значимость свойств, выраженная в процентах.

Алгоритм автоматически разбил регионы на четыре кластера с разной поддержкой и разными процентами значимости свойств. Нулевой кластер является показателем демографической обстановки страны, так как собрал в себя максимальное количество регионов. Наиболее ярко выраженными кластерами по заданным свойствам является нулевой и первый кластеры. Они максимально отличаются от остальных рассматриваемых групп значениями свойств.

7) определим кластеры, где самым значимым параметром является среднедушевой доход. Для этого нажмем кнопку настройка сортировки на панели инструментов , и зададим параметры сортировки. Выберем тип сортировки по значимости, направление по убыванию и поле, по которому будем производить сортировку, остальное оставим без изменения (рисунок 24).

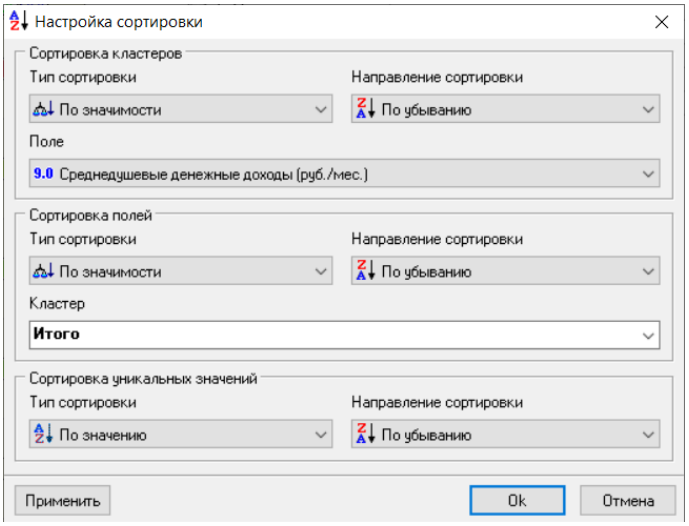


Рисунок 24 – Настройка сортировки

Кластеры поменялись местами в зависимости от значимости среднедушевого дохода в рассматриваемом наборе. Наиболее отличающиеся

кластеры по среднедушевому годовому отчету будут иметь максимальную значимость (рисунок 25).

Рисунок 25 – Результат сортировки кластеров

8) результаты по сформированным кластерам наиболее удобно рассматриваются с помощью визуализатора Куб, в котором встроена кроссдиаграмма, изображающая полученные кластеры в графическом виде, что существенно упрощает анализ (рисунок 26).

В трех из четырех кластеров наблюдается картина того, что численность населения очень сильно падает, число умерших в несколько раз больше числа родившихся. Эти кластеры показывают демографическую обстановку страны, так как в их состав входит большая часть регионов. Имеется только один кластер, где положение дел более-менее хорошее, это первый кластер.

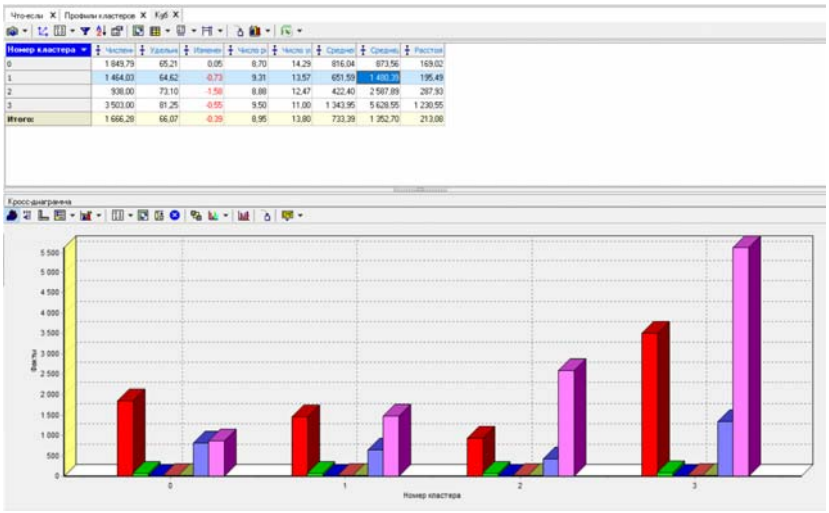


Рисунок 26 – Кластеризация отображена визуализатором Куб

Сохраните результат в файле L1.ded.

Рассмотренный пример проиллюстрировал применение кластеризации для группового анализа данных. С помощью задачи кластеризации все регионы

сгруппировались на кластеры по параметрам входных значений, интерпретация которых осуществляется с помощью кросс-диаграммы и куба. Но кажущаяся простота задачи кластеризации обманчива, она требует полной собранности аналитика при анализе полученных результатов и наличии чувства интуиции. Именно аналитик решает, на сколько кластеров необходимо разбить исследуемый набор данных и какие свойства будут основными при построении кластера, т. е. аналитик закладывает фундамент решения задачи. Но это не все проблемы, связанные с задачей кластеризации, – одной из особенностей применения k-means алгоритма, а также и многих других. Например, то, что при повторном построении задачи кластеризации можно не получить одинакового результата, это связано с тем, что данные очень разрозненные, и алгоритм выбирает случайным образом центры кластеров.

Сегментация клиентов телекоммуникационой компании с использованием карт Кохонена

В такой высокотехнологической отрасли как телекоммуникации, методы и подходы Data Mining получили широкое применение. Решаемые задачи, прежде всего, связаны с программами лояльности и удержанием существующей клиентской базы, а также с привлечением новых потребительских услуг.

В биллинговых системах телекоммуникационных компаний накапливаются большие объемы данных. В первую очередь это информация об абонентах и статистика использованных услуг. Анализ такой информации ручными и полуручными методами малоэффективны.

Постановка задачи

Руководство филиала региональной телекоммуникационной компании, представляющей услуги мобильной связи, поставило задачу сегментации абонентской базы. Ее целями являются:

– построение профилей абонента путем выявления их схожего поведения в плане частоты, длительности и времени звонков, а также ежемесячных расходов;

– оценка наиболее и наименее доходных сегментов.

Эта информация в дальнейшем может использоваться для разработки маркетинговых акций, направленных на определенные группы клиентов, на разработку новых тарифных планов, предотвращения оттока клиентов в другие компании.

В качестве исходных данных используются данные, взятые из биллинговой системы за последние несколько месяцев.

Решение задачи

Решение задачи следует разбить на два этапа:

1 кластеризация объектов алгоритмом Кохонена.

2 построение и интерпретация карты Кохонена.

В программе Deductor сети и карты Кохонена реализованы в обработчике Карта Кохонена, где содержится сам алгоритм Кохонена и специальный визуализатор карт Кохонена.

Алгоритм Кохонена применяется к сети Кохонена, состоящей из ячеек,

упорядоченных на плоскости. В выходном наборе данных алгоритм Кохонена формирует поля Номер ячейки и Расстояние до центра ячейки. Далее ячейки объединяются в кластеры при помощи алгоритма k-means или g-means. При этом каждый входной признак имеет весовой коэффициент в диапазоне от 0 до 100 %, который влияет на расчет евклидового расстояния между векторами.

1 Кластеризация при равном весе входных атрибутов

1) импортировать в Deductor набор данных из файлов mobile.txt.

2) запустить Мастер обработки и выбрать узел Карта Кохонена.

Установить все поля, кроме Код входными (рисунок 27).

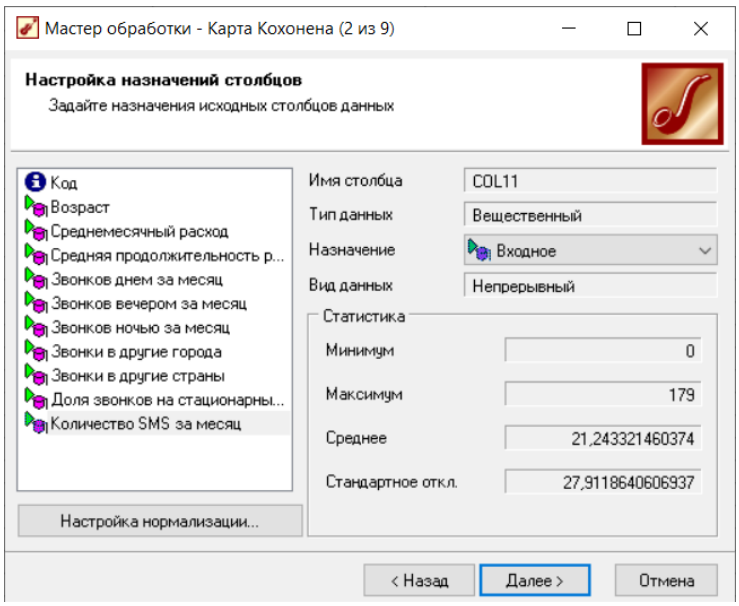


Рисунок 27– Настройка назначений столбцов

На этой же вкладке нажмите кнопку Настройка нормализации. При этом откроется окно, где можно будет задать значимость каждого входного поля. Для начала значимость всех полей надо оставить без изменения.

3) перейти ко второму шагу – разбиение исходного набора данных на подмножества. В обучающем множестве необходимо оставить 100 % записей, поскольку в алгоритме Кохонена необходимость в выделении отдельного тестового множества отсутствует (рисунок 28).

4) перейти к четвертому шагу – настройка параметров карты Кохонена. Выбираем размер сетки 24x18, общее число элементов составляет теперь 432 (рисунок 29).

5) на шаге Настройка параметров установки обучения изменить настройки обучения в соответствии с рисунком (рисунок 30).

Далее установить фиксированное число кластеров – 6 (рисунок 31).

6) выполнить построения карты Кохонена, нажав кнопку Пуск (рисунок 32).

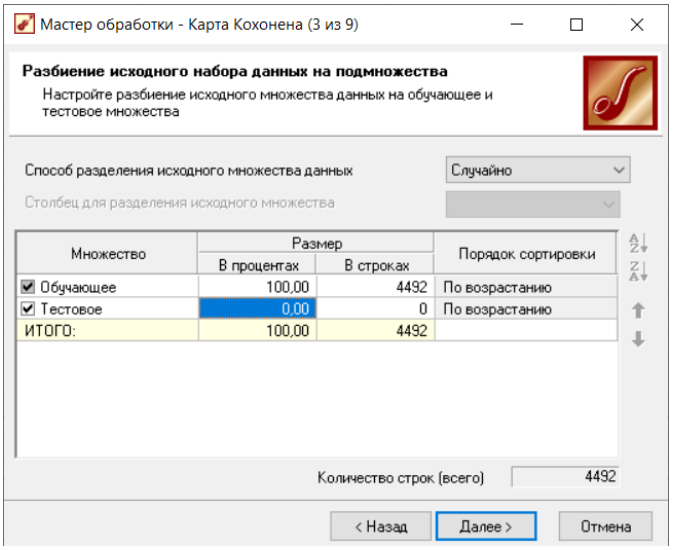


Рисунок 28 – Разбиение исходного набора данных на подмножества

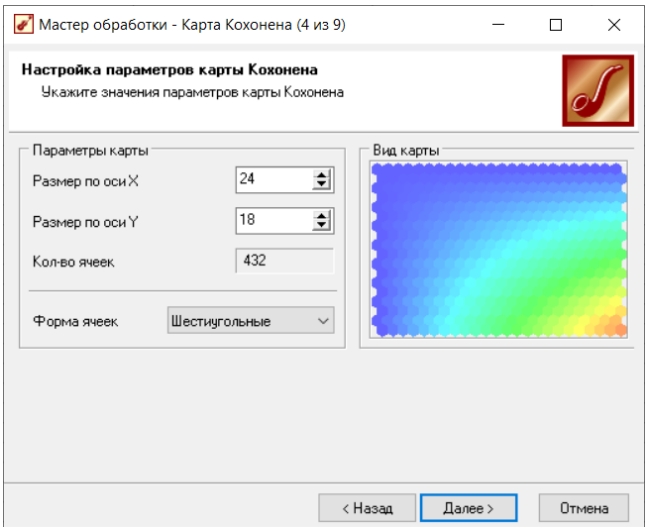


Рисунок 29 – Настройка параметров карты Кохонена

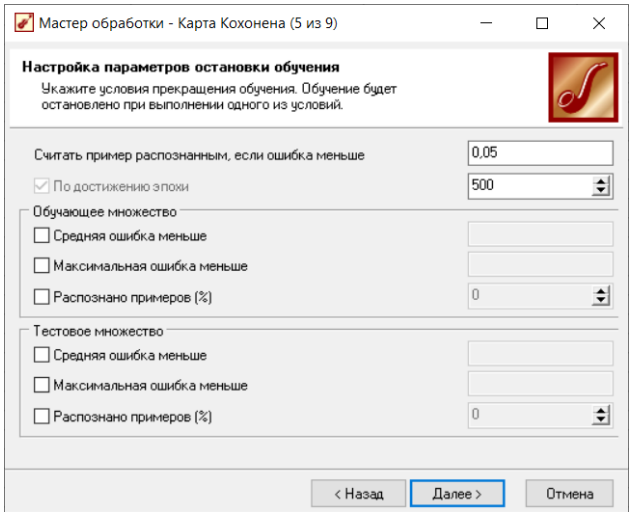


Рисунок 30 – Настройка параметров установки обучения

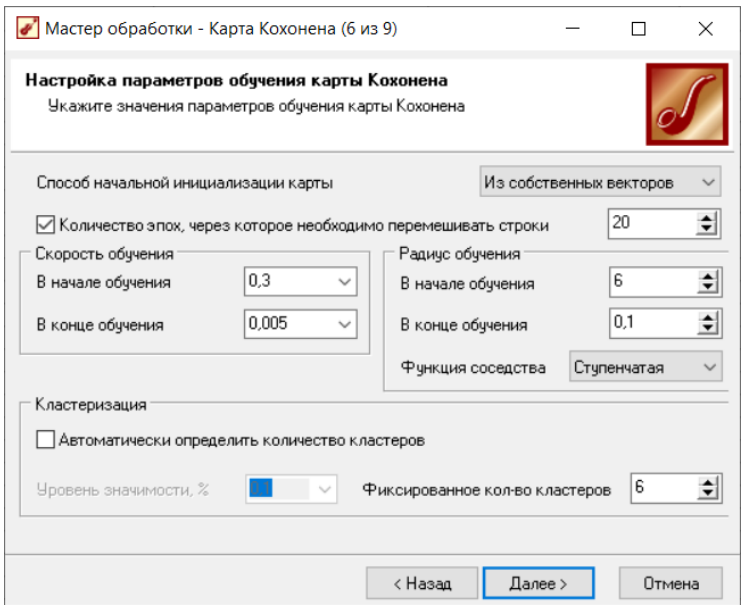


Рисунок 31 – Настройка параметров установки обучения

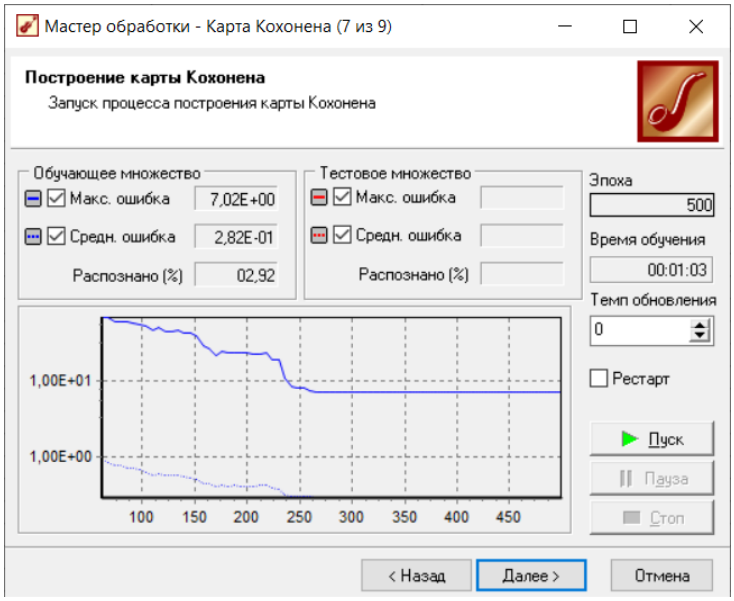


Рисунок 32 – Построение карты Кохонена

7) после того, как карта была получена, перейти к следующему шагу. В качестве способа визуализации выбрать Карта Кохонена и Профили кластеров.

! Процесс займет некоторое время.

8) на последнем шаге выполняется настройка отображений карты Кохонена (рисунок 33). Здесь следует выбрать все входные столбцы и некоторые специальные (Матрица ошибок квантования, Матрица плотности попадания и Кластеры).

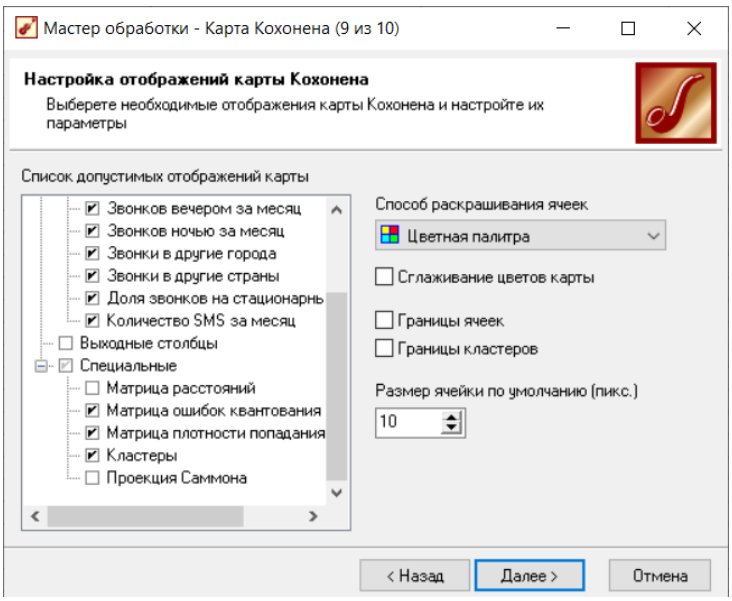


Рисунок 33 – Настройка отображений карты Кохонена

9) на следующем рисунке приведены карты Кохонена для всех выбранных столбцов. Каждая ячейка карты соответствующей характеристики окрашена в цвет, теплота которого пропорциональна среднему арифметическому значений этой характеристики для всех абонентов, которые были к ней отнесены. Темно-синий цвет соответствует минимальному значению из всех средних арифметических, а красный – максимальному (рисунок 34).

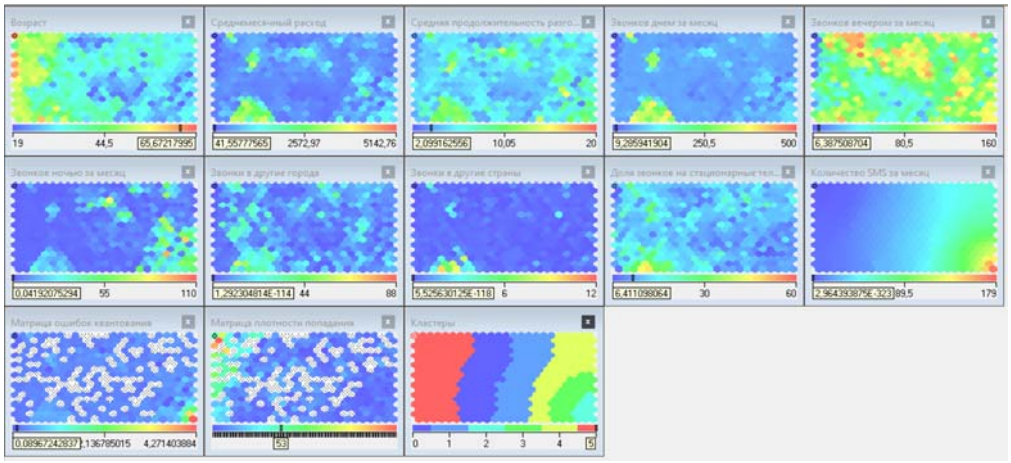


Рисунок 34 – Карты Кохонена для всех выбранных столбцов

Рассматривая карты, можно выделить группы абонентов, которые явно сгруппированы по каким-либо признакам, и сделать вывод об особенностях и предпочтениях этих людей в сфере услуг мобильной связи. Так, на карте «Возраст» можно выделить группу людей среднего и пожилого возраста и убедиться в том, что они практически не пользуются SMS-сообщениями, но при этом для них характерно среднее число вечерних звонков – около 70 за месяц.

Если выделить конкретную ячейку, то по ней можно посмотреть детализацию – получить список тех абонентов, которые были к ней отнесены, и получить их общую статистику. Так, если выбрать правую нижнюю ячейку,

которой соответствует максимальное среднее число SMS-сообщений в месяц, то после установки параметра фильтра По ячейке в окне детализации , мы получим список из девяти абонентов с полной детализацией их характеристик (рисунок 35).

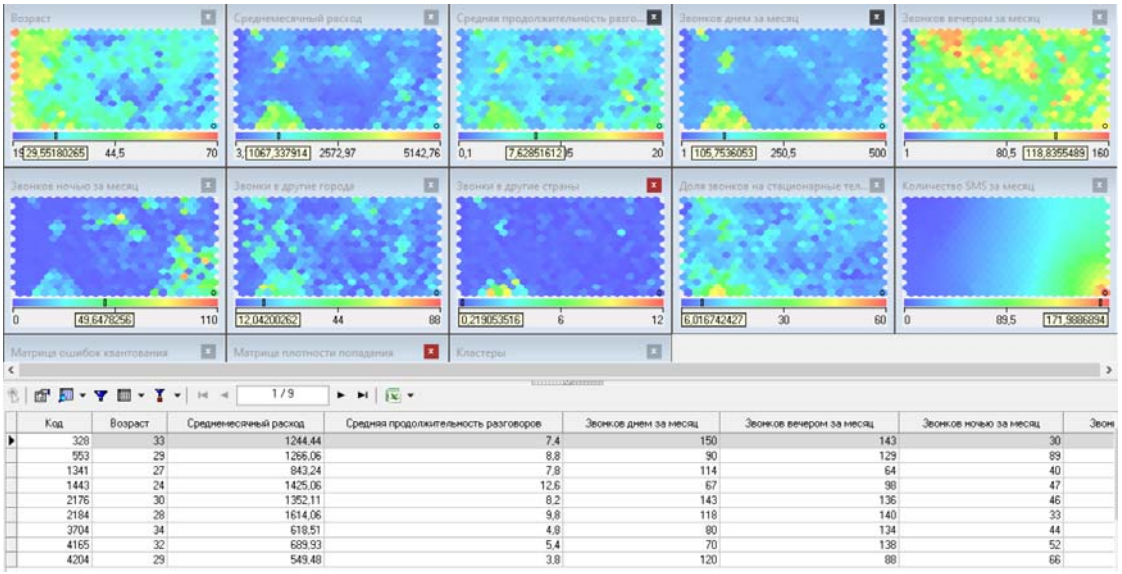


Рисунок 35 – Пример детализации по выбранной ячейке

Для получения статистики следует выбрать способ отображения Статистика в окне детализации и выполнить настройку ее параметров (из контекстного меню) – оставить гистограмму, среднее и стандартное отклонение. Остальные поля – убрать (рисунок 36).

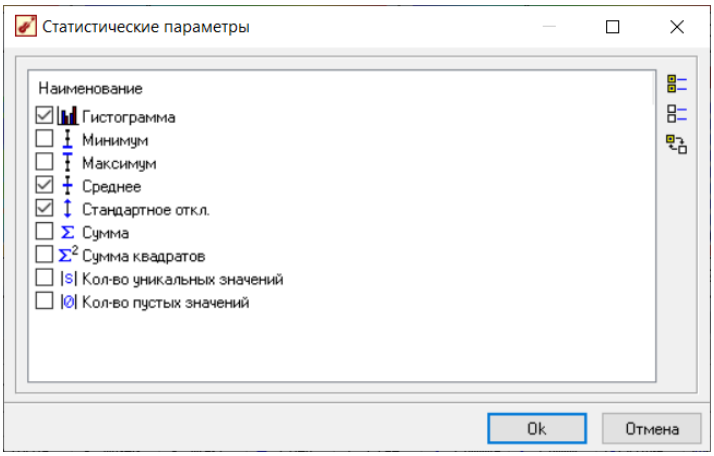


Рисунок 36 – Настройка параметров отображения Статистика

Также можно выполнить и настройку формата отображения данных (также из контекстного меню) (рисунок 37).

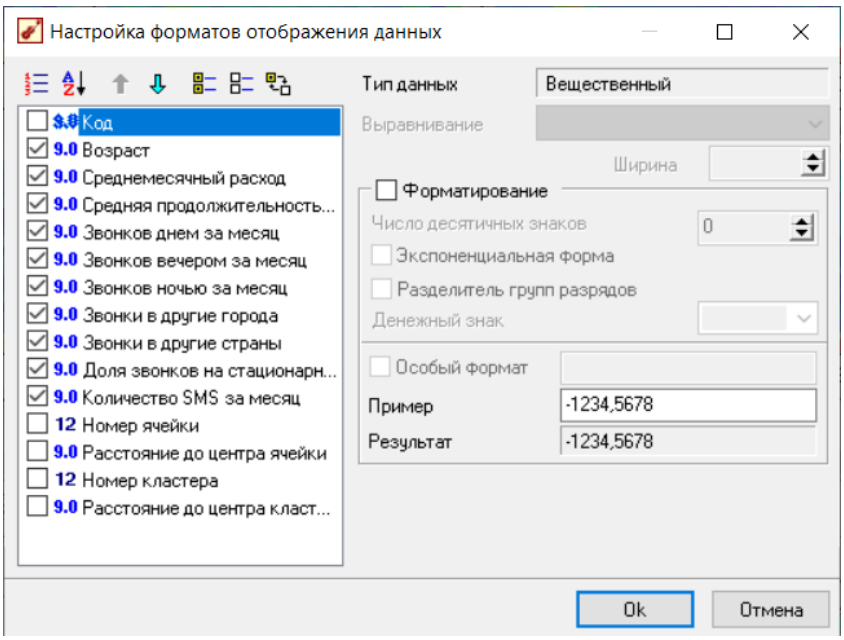


Рисунок 37 – Настройка формата отображения данных

Результат представлен на следующем рисунке (рисунок 38). Видно, что средний возраст этой группы абонентов составляет 30 лет, среднемесячный расход 1 067 руб., примерно одинаковое число звонков днем и вечером за месяц, в среднем в месяц они отправляют по 172 SMS-сообщения.

	Метка столбца	Статистика: Кол-во значений = 9		
		Гистогра...	Среднее	Стандартное откл.
1	Возраст		29,55555556	3,12694384
2	Среднемесячны...		1066,987778	393,8749339
3	Средняя продол...		7,622222222	2,711907897
4	Звонков днем з...		105,7777778	30,5318595
5	Звонков вечеро...		118,8888889	28,32597944
6	Звонков ночью ...		49,66666667	18,13146436
7	Звонки в другие...		12,11111111	13,03307758
8	Звонки в другие...		0,2222222222	0,6666666667
9	Доля звонков н...		6	4,062019202
10	Количество SM...		172	5,315072906

Рисунок 38 – Результат анализа статистики

10) попробуем самостоятельно выделить кластер и провести по нему анализ. Для этого нужно переключится в режим выделения  и указать ячейки на карте, которые должны быть отнесены к выделенной области. Выделим область, для которой характерна относительно большое число звонков в другие страны (рисунок 39).

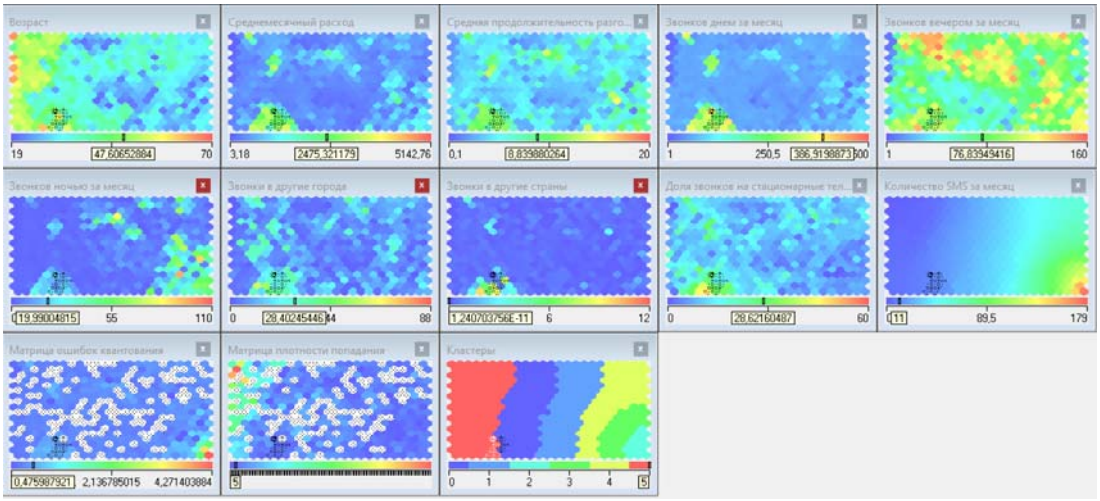


Рисунок 39 – Выделение области для анализа

Согласно статистике, средний возраст этой группы людей составляет 48 лет, для них характерен высокий среднемесячный расход – 2475 руб., они практически не пользуются SMS-сообщениями и в среднем за день совершают по 16 исходящих звонков продолжительностью по 9 минут (рисунок 40).

	Метка столбца	Статистика: Кол-во значений = 5		
		Гистогра...	Среднее	Стандартное откл.
1	9.0 Возраст		47,6	12,0124935
2	9.0 Среднемесячный расход		2474,642	318,589854
3	9.0 Средняя продолжительность разговоров		8,84	2,909982818
4	9.0 Звонков днем за месяц		386,8	131,6859901
5	9.0 Звонков вечером за месяц		76,8	17,79606698
6	9.0 Звонков ночью за месяц		20	12,88409873
7	9.0 Звонки в другие города		28,4	24,21363252
8	9.0 Звонки в другие страны		0	0
9	9.0 Доля звонков на стационарные телефо...		28,6	12,66096363
10	9.0 Количество SMS за месяц		11	0

Рисунок 40 – Результат статистики области для анализа

Детализация по выделенной группе абонентов приведена на следующем рисунке (рисунок 41).

Код	Возраст	Среднемесячный расход	Средняя продолжительность разговоров	Звонков днем за месяц	Звонков вечером за месяц
280	36	3577,77	13,8	346	
287	67	2389,13	11,4	184	
325	57	1970,05	7,8	311	
366	42	2086,81	10,4	252	
535	41	1574,47	6,6	265	
674	27	2215,23	8,4	296	
806	49	2596,22	10,6	286	
955	53	2541,63	7	474	
958	33	1196,8	5,6	271	
968	55	3139,5	13,8	217	
995	50	3224,24	14,8	246	
1072	25	3267,75	8,6	477	
1238	40	1819,53	7,4	272	
1323	28	2370,06	10,8	243	
1431	30	3078,37	14,6	267	
1552	29	2256,83	14,2	168	
1555	34	3577,77	13,2	322	

Рисунок 41 – Детализация статистики области для анализа

11) выделим другой кластер – по возрастной группе (рисунок 42).

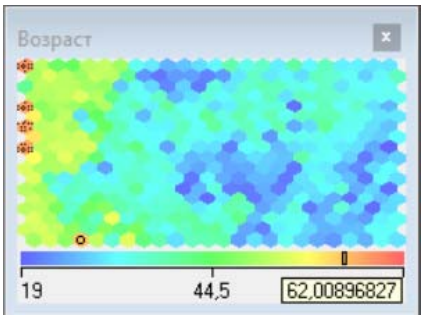


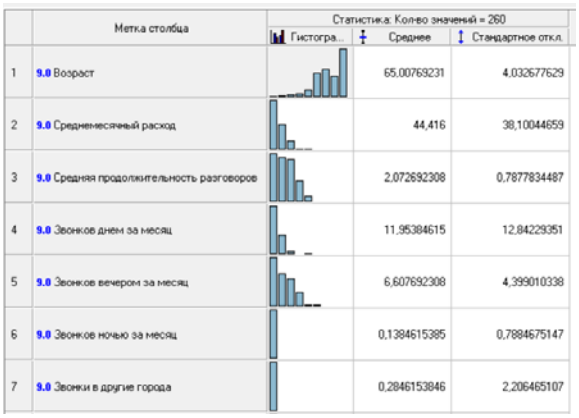
Рисунок 42 – Выбор кластера по возрастной группе

Для этого следует очистить предыдущее выделение и выделить ячейки, которым соответствует большое значение возраста абонентов. Детализация и статистика по выделенной группе приведена ниже. Видно, что средний возраст этой группы абонентов составляет 65 лет, среднемесячный расход составляет 44 руб., они совершают очень мало исходящих звонков со средней продолжительностью две минуты и не пользуются SMS-сообщениями (рисунок 43).

Код	Возраст	Среднемесячный расход	Средняя продолжительность разговоров	Звонков днем за месяц	Звонков вечером за месяц	Звонков ночью за месяц	Зв
7	66	14,52	1,1	10	1	0	
15	69	50,4	2,1	15	5	0	
20	59	74,88	2,4	9	17	0	
43	59	196	3,5	35	31	0	
46	68	39	2,5	9	4	0	
47	68	55,2	2,3	8	12	0	
50	65	73,92	2,2	16	12	0	
51	63	77,52	3,4	9	10	0	
66	69	25,92	1,2	9	9	0	
77	68	50,4	3,5	5	7	0	
78	67	14,4	1,5	3	5	0	
91	53	18,72	1,2	3	10	0	
101	64	7,2	1,2	4	1	0	
103	64	63,48	2,3	18	5	0	
105	66	71,76	2,3	17	9	0	
125	69	12,48	2,6	2	2	0	
134	65	48,6	2,7	6	9	0	
148	70	39	2,5	12	1	0	
181	69	36,48	1,6	11	8	0	
187	64	27,36	1,2	7	12	0	

Рисунок 43– Детализация статистики кластера по возрастной группе

Выполните оценку приведенной статистики (рисунок 44).



	Метка столбца	Статистика: Кол-во значений = 260		
		Гистогра...	Среднее	Стандартное откл.
1	9.0 Возраст		65,00769231	4,032677629
2	9.0 Среднемесячный расход		44,416	38,10044659
3	9.0 Средняя продолжительность разговоров		2,072692308	0,7877834487
4	9.0 Звонков днем за месяц		11,95384615	12,84229351
5	9.0 Звонков вечером за месяц		6,607692308	4,399010338
6	9.0 Звонков ночью за месяц		0,1384615385	0,7884675147
7	9.0 Звонки в другие города		0,2846153846	2,206465107

Рисунок 44 – Статистика кластера по возрастной группе

12) использование профилей кластеров для получения сводной оценки карт Кохонена.

Перейдя на соответствующую вкладку Профили кластеров вы можете переименовать кластеры и настроить их сортировку в соответствии с имеющимися приоритетами. Здесь была выполнена следующая настройка:

– переименование кластеров (акцент был сделан на возрастную характеристику) (рисунок 45);

– настройка отображения кластеров (выделены только именные кластеры) (рисунок 46);

– настройка сортировки кластеров (рисунок 47).

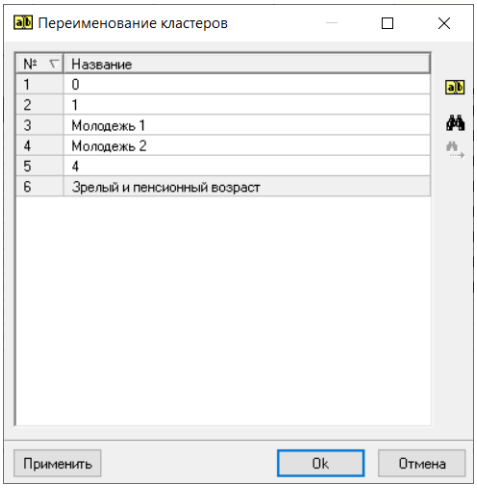


Рисунок 45 – Переименование кластеров

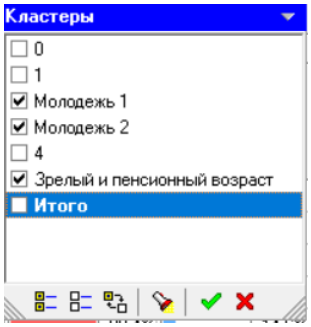


Рисунок 46 – Настройка отображения кластеров

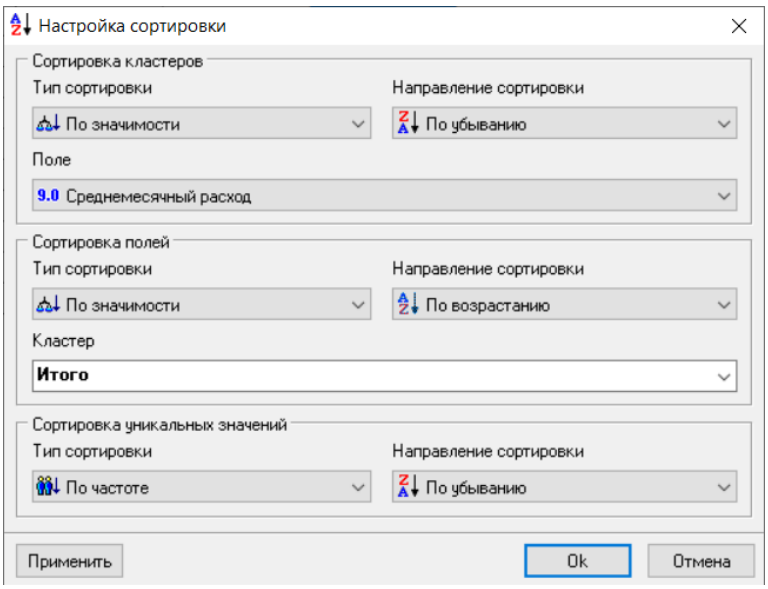


Рисунок 47– Настройка сортировки кластеров

Результат приведен на рисунке 48.

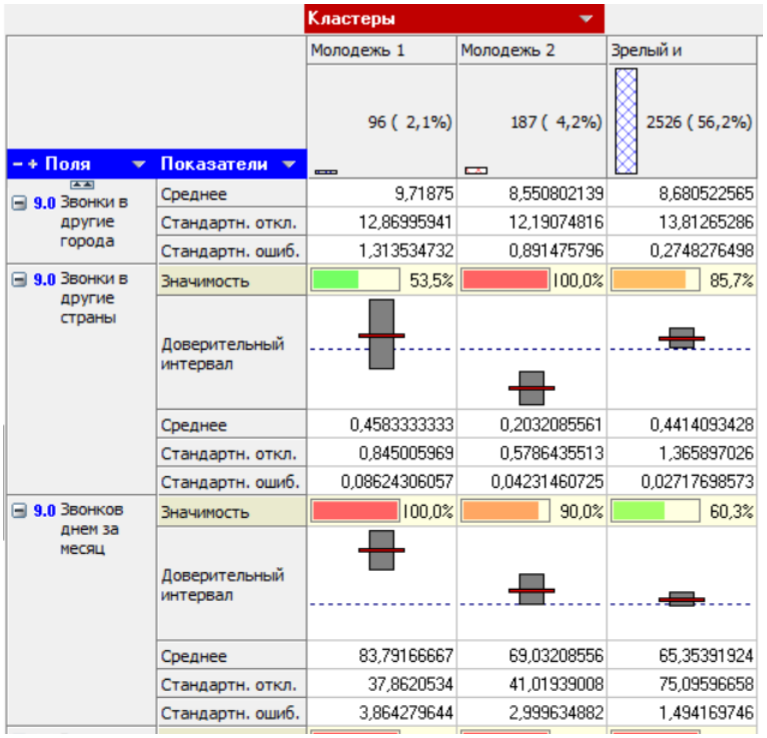


Рисунок 48 – Оценка приведенной статистики

Сохраните результат в файле L2.ded.

2 Кластеризация при различных весах входных атрибутов

Акцент на выделении кластера Активная молодежь

1) импортировать в Deductor набор данных из файлов mobile.txt.

2) запустить Мастер обработки и выбрать узел Карта Кохонена. Выполнить настройку нормализации входных столбцов (наиболее значимыми будут поля Возраст, Звонков ночью за месяц и Количество SMS за месяц) (рисунок 49).

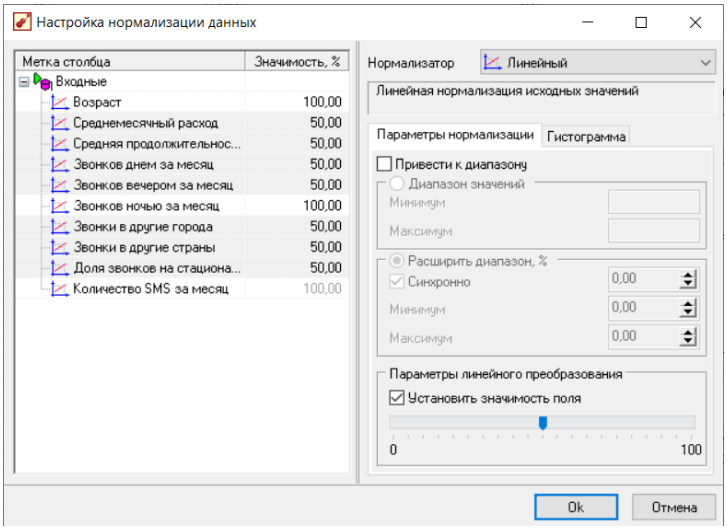


Рисунок 49 – Настройка нормализации данных

Большие значения в двух последних указанных полях как раз и должны характеризовать группу абонентов «Активная молодежь», поскольку для них характерны активное пользование SMS-сообщениями и звонки в ночное время.

Настроить карту Кохонена с параметрами из предыдущего примера.

3) проанализировав полученные карты, самостоятельно выделить кластер Активная молодежь (много ночных разговоров, много SMS, юный возраст) (рисунок 50).

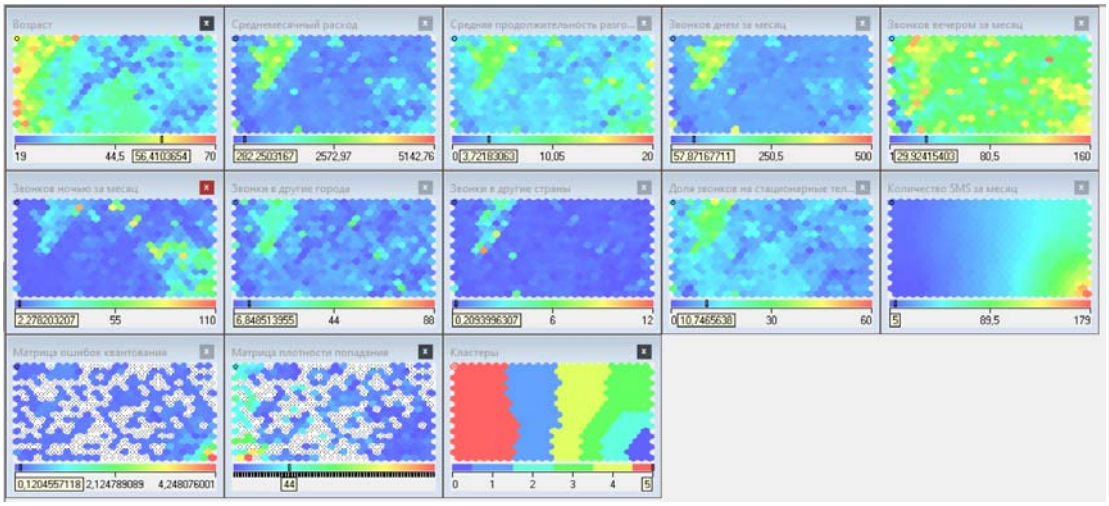


Рисунок 50 – Полученные карты Кохонена

Получить статистику по выделению группы активной молодежи.

4) выполнить автоматическую кластеризацию абонентов (рисунок 51), воспользовавшись кнопкой  , принудительно установив число кластеров, равным 3 (рисунок 52).

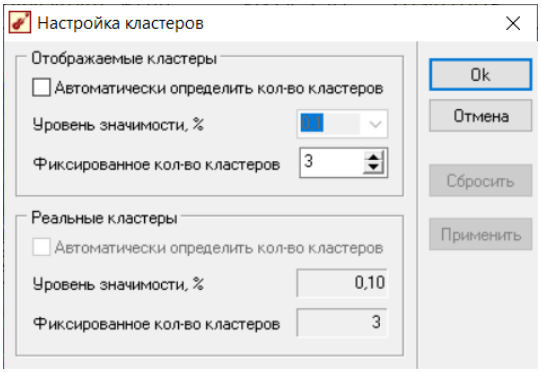


Рисунок 51 – Настройка кластеров

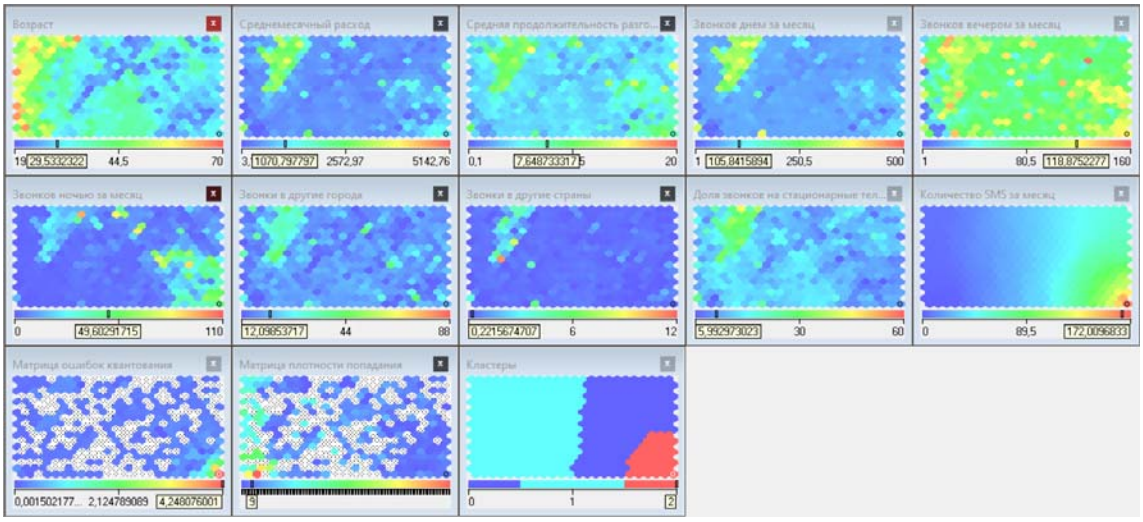


Рисунок 52 – Кластеризация абонентов

Обратите внимание, что один из кластеров явно соответствует группе Активная молодежь.

5) получить статистику по интересующему нас кластеру (рисунок 53), настроить и проанализировать профили кластеров (рисунок 54).

	Метка столбца	Статистика: Кол-во значений = 228		
		Гистогра...	Среднее	Стандар...
1	9.0 Возраст		27,74122807	5,909647327
2	9.0 Среднемесячны...		832,6456579	669,1454301
3	9.0 Средняя продол...		6,625438596	4,316280243
4	9.0 Звонков днем з...		76,31578947	39,38532471
5	9.0 Звонков вечеро...		97,80701754	36,73835989
6	9.0 Звонков ночью ...		36,89035088	29,04911978
7	9.0 Звонки в другие...		9,714912281	13,08469013
8	9.0 Звонки в другие...		0,3333333333	0,7290893078

Рисунок 53 – Статистика по кластеру

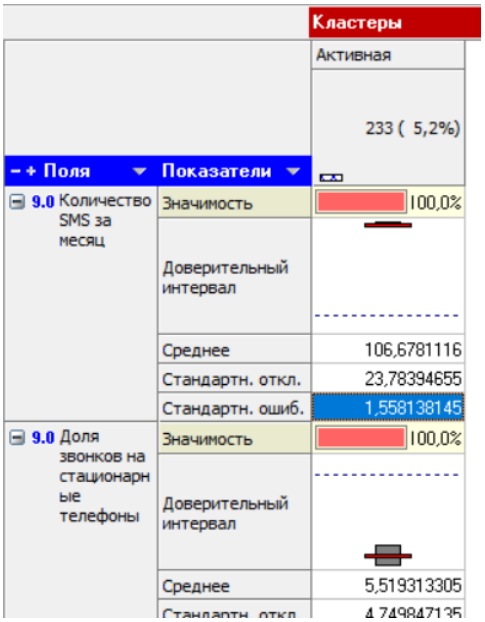


Рисунок 54 – Оценка статистики по заданному кластеру

6) получить карты Кохонена, делая акцент на формировании кластера VIP-клиенты – самые высокие расходы, продолжительные разговоры, частые международные звонки, много разговоров в рабочее время.

Сохраните результат в файле L2.ded.

3 Кластеризация при ограничении набора входных атрибутов

Акцент на выделении кластера Активная молодежь

1) импортировать в Deductor набор данных из файлов mobile.txt

2) запустить Мастер обработки и выбрать узел Карта Кохонена. Выбрать в качестве входных параметров поля Возраст, Звонков ночью за месяц и Количество SMS за месяц, остальные поля сделать выходными (рисунок 55).

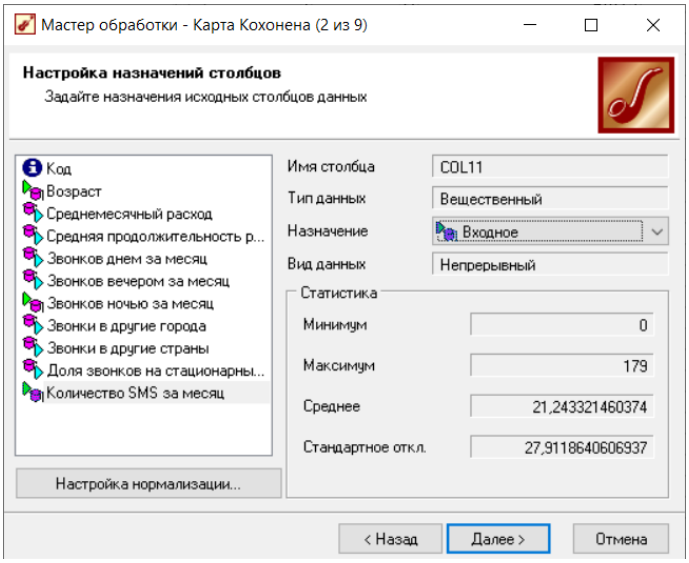


Рисунок 55 – Настройка назначений столбцов

3) настроить параметры обучения как в предыдущих примерах. И запустить процесс построения карты Кохонена.

4) проанализировав полученные карты, самостоятельно выделить кластер Активная молодежь (много ночных разговоров, много SMS, юный возраст) (рисунок 56). Получить статистику по полученному выделению. Обратите внимание на однородность заполнения ячеек выходных полей. Какой вывод можно из этого сделать? Предложить данной группе абонентов оптимальный тарифный план.

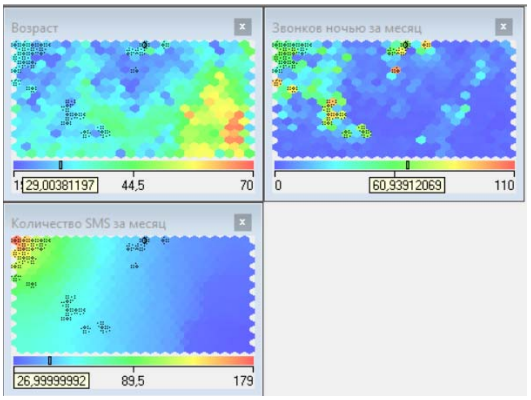


Рисунок 56 – Полученные карты Кохонена

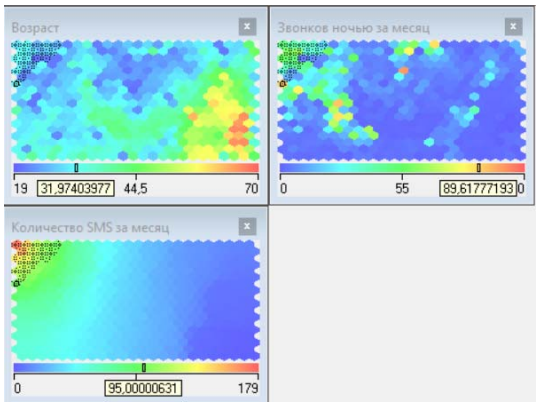


Рисунок 57 – Автоматическая кластеризация абонентов

5) выполнить автоматическую кластеризацию абонентов, принудительно установив число кластеров, равным 3 (рисунок 57). Обратите внимание, что один из кластеров близок к той группе, которую мы выделяли ранее.

6) получить статистику по интересующему нас кластеру (рисунок 58), настроить и проанализировать профили кластеров.

	Метка столбца	Статистика: Кол-во значений = 122		
		Гистогра...	Среднее	Станда...
1	9.0 Код		2370,352459	1370,763236
2	9.0 Возраст		28,39344262	5,647433963
3	9.0 Среднемесячны...		885,6281148	640,267564
4	9.0 Средняя продол...		6,990163934	3,990845099
5	9.0 Звонков днем з...		81,21311475	38,75377444
6	9.0 Звонков вечеро...		100,9836066	37,18014746

Рисунок 58 – Статистика по заданному кластеру

7) получить карты Кохонена, делая акцент на формировании кластера «VIP-клиенты» – самые высокие расходы, продолжительные разговоры, частые международные звонки, много разговоров в рабочее время.

Сохраните результат в файле L2.ded.

7 Контрольные вопросы

1 В чём состоит задача кластеризации данных?
2 Какие существуют различные способы определения расстояния между объектами наблюдения по их признакам?
3 К какому классу сложности относится задача кластеризации в классической постановке?
4 Как работает классическая реализация алгоритма k внутригрупповых средних?
5 Что можно сказать о сходимости алгоритма k внутригрупповых средних?
6 Какую функцию минимизирует алгоритм k внутригрупповых средних?
7 Какие существуют альтернативные варианты реализации алгоритма k внутригрупповых средних?
8 Какие существуют методы автоматического выбора начальных центров кластеров для алгоритма k внутригрупповых средних?
9 Как работает алгоритм автоматического выбора начальных центров кластеров kmeans++?
10 Что такое иерархическая кластеризация?

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1 Макшанов А. В. Технологии интеллектуального анализа данных : учебное пособие / А. В. Макшанов, А. Е. Журавлев. – 2-е изд., стер. – Санкт-Петербург : Лань, 2019. – 212 с.

2 Мицель А. А. Прикладная математическая статистика : учебное пособие / А. А. Мицель. – Томск : ТУСУР, 2019. – 113 с. – URL: https://edu.tusur.ru/publications/9151 (дата обращения: 01.10.2025).

3 Форман Д. Много цифр. Анализ больших данных при помощи Excel / Д. Форман; перевод А. Соколовой. – Москва : Альпина Паблишер, 2016. - 461 с. Электронно-библиотечная система «Лань». – URL: https://e.lanbook.com/book/87871 (дата обращения: 01.10.2025).

Адаменко Юлия Владимировна

Анализ данных. Часть 3. Кластеризация. Методы машинного обучения.
Методические указания к выполнению лабораторных работ
для бакалавров направлений
09.03.03 «Прикладная информатика»,
09.03.04 «Программная инженерия»

Редактор В. А. Лисина

БИЦ Курганского государственного университета.
640020, г. Курган, ул. Советская, 63/4.
Курганский государственный университет.