

КЛАСТЕРНЫЙ АНАЛИЗ

Кластерный анализ позволяет разбивать некоторые объекты на кластеры или группы, в соответствии с некоторыми признаками. Существует много методов кластерного анализа, в данной работе рассмотрим один из основных: метод k-средних. Рассмотрим его на примере.

ПРИМЕР. Проводится маркетинговое исследование 25 различных товаров. Были собраны данные об объемах продаж (в % от всего объема проданных аналогичных товаров). Причем данные собирались в двух категориях: x – это процент продаж в городских территориях, и y – процент продаж в сельской местности. Исходные данные по объемам продаж каждого товара – в таблице.

Товар	1	2	3	4	5	6	7	8	9	10	11	12	13
x	90	99	82	61	28	53	23	97	46	78	56	77	80
y	96	11	12	57	58	34	24	42	41	56	29	30	79
Товар	14	15	16	17	18	19	20	21	22	23	24	25	
x	76	40	22	18	78	31	99	26	63	32	67	90	
y	55	33	83	81	6	7	20	57	8	39	9	43	

Ставится задача определить три кластера товаров. Первый кластер - это товары, которые плохо продаются в городе, но хорошо в сельской местности, центр этого кластера определяем, как $x_c=30\%$, $y_c=70\%$. Второй кластер - это товары, которые умеренно продаются как в городе, так и в сельской местности, центр этого кластера определяем, как $x_c=50\%$, $y_c=50\%$. Третий кластер - это товары, которые хорошо продаются в городе, но плохо в сельской местности, центр этого кластера определяем, как $x_c=70\%$, $y_c=30\%$.

Открываем лист Excel и вводим координаты центров кластеров в ячейки B2-D3. Также оформляем лист для проведения дальнейших вычислений, как это показано на рисунке.

	A	B	C	D	E	F	G	H	I	J
1	Центр	Кластер1	Кластер2	Кластер3						
2	x	30	50	70						
3	y	70	50	30						
4	Данные	x	y	Расстояние	Кластер1	Кластер2	Кластер3	Минимум	Кластер	Новый центр

Продолжение рисунка по горизонтали

	K	L	M	N	O	P	Q	R	S	T
1										
2										
3					Кластер 1		Кластер 2		Кластер 3	
4	Кластер1	Кластер2	Кластер3		x	y	x	y	x	y

Вводим данные из таблицы. Эту таблицу целиком в вертикальном виде вводим в ячейки A5-C29. То есть номер товара вводим в столбец A5-A29, при этом рекомендуем воспользоваться автозаполнением. Значения показателя x в диапазоне B5-B29, а показатель y в диапазон C5-C29.

Далее находим расстояния от каждого элемента до центра каждого кластера. При этом воспользуемся формулой квадрата Евклидова расстояния, когда мера расстояния от какого-то объекта (товара) с показателями (x ; y) до какого-то центра с координатами (x_c ; y_c) определяется формулой: $(x - x_c)^2 + (y - y_c)^2$. Для этого в ячейку E5 вводим формулу $= (B\$2 - \$B\$5)^2 + (B\$3 - \$C\$5)^2$ и с помощью автозаполнения распространяем ее на диапазон E5-G29.

Товар относится к тому кластеру, расстояние до которого минимальное. Рассчитываем минимальное из этих трех расстояний. В ячейку H5 вводим формулу $= \text{МИН}(E5:G5)$ и автозаполняем ее на диапазон H5-H29. Далее, определяем номер кластера, к которому относится этот товар. Для этого воспользуемся логической функцией Excel. Вводим в ячейку I5 формулу $= \text{ЕСЛИ}(E5=H5;1;\text{ЕСЛИ}(F5=H5;2;3))$ и автозаполняем ее на диапазон I5-I29.

На следующем этапе определяем новые координаты центров кластеров, которые равны среднеарифметическому соответствующих координат объектов, которые входят в этот кластер. При расчете нужно взять сумму каждой координаты для элементов из данного кластера (с использованием функции СУММЕСЛИ) и поделить на количество суммируемых элементов (с использованием функции СЧЁТЕСЛИ). Производим расчет новых координат в ячейках K5-M6, вводя формулы в соответствии с таблицей:

Ячейка	Формула
K5	$= \text{СУММЕСЛИ}(I5:I29;"=1";B5:B29) / \text{СЧЁТЕСЛИ}(I5:I29;"=1")$
K6	$= \text{СУММЕСЛИ}(I5:I29;"=1";C5:C29) / \text{СЧЁТЕСЛИ}(I5:I29;"=1")$
L5	$= \text{СУММЕСЛИ}(I5:I29;"=2";B5:B29) / \text{СЧЁТЕСЛИ}(I5:I29;"=2")$
L6	$= \text{СУММЕСЛИ}(I5:I29;"=2";C5:C29) / \text{СЧЁТЕСЛИ}(I5:I29;"=2")$
M5	$= \text{СУММЕСЛИ}(I5:I29;"=3";B5:B29) / \text{СЧЁТЕСЛИ}(I5:I29;"=3")$
M6	$= \text{СУММЕСЛИ}(I5:I29;"=3";C5:C29) / \text{СЧЁТЕСЛИ}(I5:I29;"=3")$

Вычислены новые координаты центров кластеров. Перенесем новые координаты в качестве исходных. Для этого выделяем область ячеек K5-M6, копируем данные (Ctrl+C), затем ставим курсор в ячейку B2 и вставляем данные в виде значений. Для этого нажимаем правую кнопку мыши и в поле параметры вставки выбираем вторую справа иконку. Автоматически просчитались новые координаты центров. Вновь копируем их в исходные координаты, и так до тех пор, пока результаты копирования не приведет к

тому, что исходные координаты и рассчитанные будут совпадать. В данном примере потребовалось три итерации.

В заключении, построим график кластеров и их центров. Для этого перечислим координаты каждого объекта (товара) в каждом кластере, если объект не относится к кластеру, то его координаты нулевые. Для этого в ячейку O5 вводим формулу =ЕСЛИ(\$I5=1;B5;0) и автозаполняем на диапазон O5-P29, в ячейку Q5 вводим =ЕСЛИ(\$I5=2;B5;0) и автозаполняем на Q5-R29, в ячейку S5 вводим =ЕСЛИ(\$I5=3;B5;0), автозаполняем на S5-T29.

Для построения графика ставим курсор в свободное место ниже расчетов и вызываем «Вставка», там выбираем точечную диаграмму (по центру в третьем ряду, в открывшемся окне, выбираем самую первую. График пока пустой, заполняем его.

Щелкаем правой кнопкой мыши по пустой диаграмме и выбираем «Выбрать данные». В поле «Элементы легенды (ряды)» нажимаем «Добавить». В окне «Изменение ряда» в поле «Имя ряда» вводим текст «Кластер 1», в поле «Значение X» вводим ссылку на диапазон O5-O29, обводя его мышью, в поле «Значения Y» удаляем имеющуюся там информацию и делаем ссылку на диапазон P5-P29, нажимаем «Ок».

Вновь нажимаем «Добавить». В окне «Изменение ряда» в поле «Имя ряда» вводим текст «Кластер 2», в поле «Значение X» вводим ссылку на диапазон Q5-Q29, в поле «Значения Y» делаем ссылку на диапазон R5-R29, нажимаем «Ок».

В третий раз нажимаем «Добавить». В поле «Имя ряда» вводим текст «Кластер 3», в поле «Значение X» вводим ссылку на диапазон S5-S29, в поле «Значения Y» делаем ссылку на диапазон T5-T29, нажимаем «Ок».

Теперь вводим центры кластеров, нажимаем «Добавить». В поле «Имя ряда» вводим текст «Центры кластеров», в поле «Значение X» вводим ссылку на диапазон B2-D2, в поле «Значения Y» делаем ссылку на диапазон B3-D3, нажимаем «Ок» и еще раз «Ок».

Чтобы выделить центры кластеров щелкаем левой кнопкой по любой точке – центру кластера на графике, потом по ней же правой кнопкой и выбираем «Формат ряда данных». На первой вкладке выбираем «Маркер», цвет делаем черным, ширина 2, закрываем окно формата ряда данных. Название диаграммы «Кластеры товаров». Работа выполнена.

Следует отметить, что оба показателя x и y измеряются по одной шкале. Если бы шкалы были разными, то необходимо было бы проводить нормализацию.

Задания для самостоятельного решения

ЗАДАНИЕ 1

Проводится анализ 22 сайтов в плане посещений их пользователями. Имеются 2 показателя: X – среднесуточное количество посещений в рабочие дни (тыс.) и Y – среднесуточное количество посещений в выходные дни (тыс.). Разделить сайты на 3 кластера: K1 – сайты с малым посещением, как в рабочие, так и в выходные дни, K2 – сайты со средним посещением и K3 – сайты со большим количеством посещений. Исходные данные в таблице:

Номер сайта	1	2	3	4	5	6	7	8	9	10	11
X	12,1	18,2	3,9	25,3	21,9	5,6	23,7	15,8	27,4	13	23
Y	8,3	29,7	14,7	22,5	8	7,8	27,5	15,4	17,7	22,9	20,3
Номер сайта	12	13	14	15	16	17	18	19	20	21	22
X	10,2	25,3	8,2	26,3	21,5	10	26,7	26,1	25,5	9,5	18,4
Y	12,4	10,2	23,4	21,9	16	24,5	16,7	27,9	19,1	24,6	12,4

В качестве центров рекомендовано взять: K1(10; 10), K2(17; 17), K3(24; 24).

Построить график.

ЗАДАНИЕ 2

Решим задачу кластеризации для одного признака. Исследуются 30 городов с точки зрения востребованности некоторой услуги. Объем указанных услуг в каждом городе за день приведены в таблице. Необходимо разделить города на 4 кластера по объему указанных услуг. В качестве центров кластеров взять числа 100, 200, 300 и 400. График строить не надо.

Город	1	2	3	4	5	6	7	8	9	10
Объем услуг	211	263	347	330	289	221	288	230	452	284
Город	11	12	13	14	15	16	17	18	19	20
Объем услуг	81	417	366	428	264	155	436	373	144	401
Город	21	22	23	24	25	26	27	28	29	30
Объем услуг	450	349	403	346	396	198	82	372	142	423

Указание. В качестве расстояний между объектами взять модуль разницы между координатой объекта и координатой центра