Лабораторная работа № 7.

Решение слабо структуризованных задач. Кластерный анализ.

Цель работы:
1) Изучить основные понятия кластерного анализа;
2) изучить метод K средних, метод максимина.

Методы кластерного анализа предназначены для разделения множества анализируемых объектов и явлений на кластеры, т.е. на группы объектов, схожих друг с другом по каким-либо признакам.

Пример. Анализируется информация о девяти инвестиционных фондах (Ф1, Ф2,…,Ф9). Показатели, характеризующие деятельность фондов, приведены в табл.10.1.

Таблица 10.1 – Информация о девяти инвестиционных фондах (Ф1, Ф2,…,Ф9)

Фонд	Ф1	Ф2	Ф3	Ф4	Ф5	Ф6	Ф7	Ф8	Ф9
Прибыль за анализируемый период, тыс. ден.ед.	16476	17081	13827	13187	11793	16728	10386	15145	15596
Экспертная оценка риска, баллы	4	4	5	1	3	4	2	7	4

Примечание. Оценки риска заданы экспертом по десятибалльной шкале: 1 – минимальный риск, 10 – максимальный.

Требуется выделить группы фондов, имеющих сходные значения показателей.

В данной задаче имеется девять объектов (M=9). Для их описания используется два признака (N=2). Решение этой задачи различными методами рассматривается ниже.

Подготовка данных для кластерного анализа

1. Меры различия

Как отмечено выше, задача кластерного анализа состоит в разделении множества анализируемых объектов на группы объектов, сходных друг с другом по каким-либо признакам. При этом необходимо учитывать, что каждый объект, как правило, описывается несколькими признаками. Эти признаки обычно различаются по размерности (измеряются в разных единицах) и по диапазону значений (одни признаки выражаются большими числами, другие – малыми). Некоторые признаки могут указываться в виде балльных оценок (например, по 10- или 100-балльной шкале). В некоторых случаях объекты описываются качественными (словесными) признаками: для описания объектов используются оценки “отлично”, ”хорошо”, ”часто”, ”редко” и т.д. Такое разнообразие оценок затрудняет сопоставление объектов и делает невозможным получение оценки различия между объектами в виде одного числа. Поэтому, прежде чем применять какие-либо методы кластерного анализа, необходимо выполнить нормирование признаков объектов. В результате нормировки все значения признаков объектов должны быть безразмерными (т.е. не должны измеряться в каких-либо единицах) и нахо-

диться в некотором ограниченном диапазоне (например, от нуля до единицы). Существует несколько методов нормирования. Обычно применяются следующие методы:

1) **деление на максимальное значение:** значения признака для всех объектов делятся на максимальное значение этого признака. Результатом являются безразмерные величины, находящиеся в диапазоне от нуля до единицы;

2) **стандартизация:** из каждого значения признака вычитается среднее значение данного признака, полученная разность делится на стандартное отклонение данного признака. Результатом являются безразмерные величины, большинство из которых принимает значения в диапазоне от –3 до 3.

При небольшом количестве анализируемых объектов обычно применяется деление на максимальное значение, при большом количестве объектов – стандартизация.

Если для описания объектов используются качественные (словесные) оценки, то следует перейти от таких оценок к числовым величинам (например, балльным экспертным оценкам), а затем выполнить нормировку на основе одного из рассмотренных способов.

Рассмотрим пример, приведенный выше (анализ информации об инвестиционных фондах). В этом примере объекты описываются двумя признаками. Первый из этих признаков (прибыль) измеряется в денежных единицах, второй (оценка риска) – в баллах. Кроме того, первый признак может принимать большие значения (десятки тысяч), а второй – выражается числами от 1 до 10. На основе таких оценок невозможно получить какую-либо величину, характеризующую различие между объектами. Поэтому требуется нормировка значений признаков.

1) Выполним нормирование, используя **деление на максимальное значение**. Для признака “прибыль” максимальное значение равно 17 081, для признака “оценка риска” – 7 (см. табл.10.2). Для нормирования разделим каждое значение признака на соответствующее максимальное значение. Результаты приведены в табл.10.2.

Таблица 10.2 – Нормирование с использованием **деления на максимальное значение**

Фонд	Ф1	Ф2	Ф3	Ф4	Ф5	Ф6	Ф7	Ф8	Ф9
Прибыль за анализируемый период	0,96	1,00	0,81	0,77	0,69	0,98	0,61	0,89	0,91
Экспертная оценка риска	0,57	0,57	0,71	0,14	0,43	0,57	0,29	1,00	0,57

2) Выполним нормирование на основе **стандартизации**. Для этого необходимо сначала найти среднее значение и стандартное отклонение каждого признака. Обозначим средние значения признаков как $\overline{A_i}$, а стандартные отклонения – как σ_i, i=1,…,M (где M – количество признаков), N – количество объектов. Эти величины находятся по следующим формулам:

$$\overline{A_i} = \frac{1}{N}\sum_{j=1}^{N} X_{ij}\ , \qquad (10.1)$$

$$\sigma_i = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(X_{ij} - \overline{A_i})^2}\ . \qquad (10.2)$$

Для данного примера $\overline{A_1}$=14468,78; $\overline{A_2}$=3,78; σ_1=2333,68; σ_2=1,72.

Как указано выше, нормирование выполняется следующим образом: из каждого значения признака (табл.10.2) вычитается среднее значение данного признака, полученная разность делится на стандартное отклонение. Результаты нормирования приведены в табл.10.3.

Таблица 10.3 – Нормирование на основе **стандартизации**

Фонд	Ф1	Ф2	Ф3	Ф4	Ф5	Ф6	Ф7	Ф8	Ф9
Прибыль за анализируемый период	0,86	1,12	-0,28	-0,55	-1,15	0,97	-1,75	0,29	0,48
Экспертная оценка риска	0,13	0,13	0,71	-1,62	-0,45	0,13	-1,04	1,88	0,13

Здесь, например, значение признака “прибыль за анализируемый период” для фонда Ф1 получено следующим образом: (16476-14468,78)/2333,68=0,86.

Во всех последующих расчетах будут использоваться только нормированные значения признаков.

Чтобы принять решение о том, можно ли считать некоторые объекты достаточно сходными и отнести их к одному кластеру, необходимо использовать некоторую числовую *меру различия* между объектами. Обычно в качестве такой меры различия используется *евклидово расстояние*. Значение евклидова расстояния между некоторыми объектами X_j и X_k определяется по следующей формуле:

$$D(X_j, X_k) = \sqrt{\sum_{i=1}^{M} (X_{ij} - X_{ik})^2} \,. \qquad (10.3)$$

Найдем, например, евклидово расстояние между объектами X_1 и X_2 из рассматриваемого примера, т.е. меру различия между фондами Ф1 и Ф2. Для расчета евклидова расстояния будем использовать нормированные значения признаков, полученные путем деления на максимальное значение (табл.10.3):

$$D(X_1, X_2) = \sqrt{(0{,}96-1)^2 + (0{,}57-0{,}57)^2} = 0{,}04.$$

Если признаки различны по важности (т.е. различия по одним признакам необходимо учитывать в большей степени, по другим – в меньшей), то в качестве меры различия используется взвешенное евклидово расстояние:

$$D(X_j, X_k) = \sqrt{\sum_{i=1}^{M} W_i \cdot (X_{ij} - X_{ik})^2} \,, \qquad (10.4)$$

где Wi, i=1,…,M – веса признаков (чем важнее признак, тем больше его вес). Они могут определяться, например, на основе методов экспертного анализа. Обычно используются значения весов, удовлетворяющие следующему условию: $W_1+W_2+\ldots+W_M=1$.

В некоторых случаях применяются также следующие меры различия:

- **расстояние city-block (манхэттенское):**

$$D(X_j, X_k) = \sum_{i=1}^{M} \left| X_{ij} - X_{ik} \right| ; \qquad (10.5)$$

Это расстояние рассчитывается как среднее разностей по координатам. В большинстве случаев эта мера расстояния приводит к результатам, подобным расчетам расстояния евклида. Однако, для этой меры влияние отдельных выбросов меньше, чем при

использовании *евклидова расстояния*, поскольку здесь координаты не возводятся в квадрат.

- **чебышевское расстояние:**

$$D(X_j, X_k) = \max_i \left| X_{ij} - X_{ik} \right|. \qquad (10.6)$$

Это расстояние стоит использовать, когда необходимо определить два объекта как "различные", если они отличаются по какому-то одному измерению.

Смысл всех мер различия следующий: чем больше различаются значения признаков, описывающих объекты, тем большее значение принимают меры различия. Объекты с небольшими значениями мер различия должны относиться к одному кластеру, с большими – к разным.

2. Разделение объектов на заданное число кластеров

Разделение объектов на заданное число кластеров может производиться с использованием **метода K средних, метода максимина.**

Метод K средних (K-Means Cluster)

Метод предназначен для разделения объектов на заданное число кластеров.

Принцип работы метода следующий. На основе имеющейся информации о предметной области **задается количество кластеров (K)**. При этом указывается также содержательный смысл каждого кластера (например, объекты с высоким значением некоторого признака, объекты со средним значением некоторых признаков и т.д.). Для каждого кластера выбирается объект-*прототип* (представитель), т.е. объект, наиболее подходящий для данного класса по значениям признаков. Находится первоначальный вариант разделения объектов на кластеры: каждый объект относится к кластеру, представляемому *ближайшим* объектом-прототипом. Затем в каждом кластере находится новый прототип со средними (для данного кластера) значениями признаков. Снова выполняется отнесение каждого объекта к кластеру, представляемому ближайшим прототипом. Процедура повторяется до получения окончательного разбиения, т.е. до тех пор, пока на двух последовательных итерациях метода будет получено одинаковое разбиение.

Приведем пошаговый **алгоритм** разбиения объектов на заданное число кластеров на основе метода K средних.

1. Номер итерации алгоритма принимается равным нулю: s = 0.
2. Задается количество кластеров (K). Для каждого кластера выбирается первоначальный объект-прототип: P_k^0, k=1,…, K.
3. Выполняется переход к очередной итерации алгоритма: s = s + 1.
4. Находятся расстояния от каждого из анализируемых объектов до каждого из объектов-прототипов. Выполняется отнесение каждого объекта к ближайшему кластеру, т.е. к кластеру, для которого расстояние между этим объектом и прототипом кластера минимально.
5. В каждом кластере определяется новый объект-прототип: P_k^s, k = 1,…,K. Значение каждого признака этого объекта-прототипа определяется как среднее арифметическое значений этого признака для всех объектов, входящих на текущей итерации в данный кластер.
6. Если объекты-прототипы всех кластеров на данной и на предыдущей итерации совпадают (т.е. выполняется условие $P_k^s = P_k^{s-1}$, k=1,…,K), то алгоритм завершается. Если на данной итерации получено разбиение объектов, отличное от предыдущего, то выполняется возврат к шагу 3.

Рассмотрим реализацию приведенного алгоритма на примере (анализ информации об инвестиционных фондах). Во всех расчетах будут использоваться нормированные значения признаков объектов, полученные путем деления на максимальное значение признака (табл.10.3).

Пусть предполагается, что инвестиционные фонды можно в основном разделить на три группы: 1) фонды с низким риском и низкой прибылью; 2) фонды с высоким риском и высокой прибылью; 3) фонды со средними значениями обоих показателей. Таким образом, предполагается, что число кластеров равно трем (K=3). Пошаговый алгоритм разбиения объектов на заданное число кластеров на основе метода K средних выполняется в следующем порядке.

1. Номер итерации принимается равным нулю: s = 0.
2. Задается количество кластеров: K=3. Для каждого кластера требуется выбрать прототип. Пусть на основе анализа показателей (табл.10.2) эксперт указал, что наиболее характерным представителем первого кластера (фонды с низким риском и низкой прибылью) является фонд, обозначенный как Ф4. В качестве характерного представителя второго кластера (фонды с высоким риском и высокой прибылью) экспертом указан фонд Ф8, а в качестве представителя третьего кластера (фонды со средними значениями обоих показателей) – фонд Ф9. Таким образом, $P_1^0 = X_4 = (0{,}77;\ 0{,}14)$, $P_2^0 = X_8 = (0{,}89;\ 1)$, $P_3^0 = X_9 = = (0{,}91;\ 0{,}57)$.
3. Выполняется **переход к очередной итерации: s = 1.**
4. Находятся расстояния от каждого из анализируемых объектов до каждого из прототипов по формуле (10.3). Эти расстояния приведены в табл.10.4.

Например, расстояние между объектом X_1 и прототипами кластеров найдено следующим образом:

$$D(X_1, P_1^0) = \sqrt{(0{,}96-0{,}77)^2 + (0{,}57-0{,}14)^2} = 0{,}47,$$

$$D(X_1, P_2^0) = \sqrt{(0{,}96-0{,}89)^2 + (0{,}57-1)^2} = 0{,}44,$$

$$D(X_1, P_3^0) = \sqrt{(0{,}96-0{,}91)^2 + (0{,}57-0{,}57)^2} = 0{,}05.$$

По найденным расстояниям выполняется отнесение каждого объекта к кластеру, представленному ближайшим прототипом. Например, для объекта X_1 расстояние до прототипа P_1^0 составляет 0,47, до прототипа P_2^0 - 0,44, до P_3^0 - 0,05. Таким образом, ближайшим к объекту X_1 оказался прототип третьего кластера, поэтому X_1 относится к этому кластеру. Результаты разделения объектов на кластеры приведены в табл.10.4 (последняя строка).

Таблица 10.4 – Результаты разделения объектов на кластеры методом K средних

Объект	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
P_1^0	0,47	0,49	0,57	0,00	0,30	0,48	0,22	0,87	0,45
P_2^0	0,44	0,44	0,30	0,87	0,60	0,44	0,76	0,00	0,43
P_3^0	0,05	0,09	0,17	0,45	0,26	0,07	0,41	0,43	0,00
Кластер	3	3	3	1	3	3	1	2	3

Таким образом, получено следующее разбиение объектов на кластеры: к первому кластеру относятся объекты X_4 и X_7, ко второму – только объект X_8, к третьему – объекты X_1, X_2, X_3, X_5, X_6, X_9.

5. В каждом кластере определяется новый объект-прототип. Значение каждого признака этого объекта-прототипа определятся как среднее арифметическое значений этого признака для всех объектов, входящих на текущей итерации в данный кластер. Для данного примера значения признаков объекта-прототипа первого кластера находятся как средние арифметические значений признаков объектов X_4 и X_7:

$$P_1^1 = \left(\frac{0,77 + 0,61}{2}; \frac{0,14 + 0,29}{2} \right) = (0,69;\ 0,22).$$

Во втором кластере находится только один объект (X_8), поэтому он и становится прототипом этого класса: $P_2^1 = (0,89;\ 1,00)$.

Значения признаков объекта-прототипа третьего кластера находятся как средние арифметические значений признаков объектов X_1, X_2, X_3, X_5, X_6, X_9:

$$P_3^1 = \left(\frac{0,96+1+0,81+0,69+0,98+0,91}{6}; \frac{0,57+0,57+0,71+0,43+0,57+0,57}{6} \right) = (0,89;\ 0,57)$$

6. Выполняется сравнение прототипов, полученных на данной и на предыдущей итерациях. На данной итерации получены прототипы $P_1^1 = (0,69;\ 0,22)$, $P_2^1 = (0,89;\ 1,00)$, $P_3^1 = (0,89;\ 0,57)$. На предыдущей итерации прототипы были следующими: $P_1^0 = (0,77;\ 0,14)$, $P_2^0 = (0,89;\ 1)$, $P_3^0 = (0,91;\ 0,57)$. Прототипы не совпадают, поэтому требуется следующая итерация. Выполняется возврат к шагу 3.

3. Выполняется **переход к очередной итерации: s = 2.**

4. Находятся расстояния от каждого из анализируемых объектов до каждого из прототипов по формуле (10.3). По найденным расстояниям выполняется отнесение каждого объекта к кластеру, представленному ближайшим прототипом. Расстояния, а также результаты разделения объектов на кластеры приведены в табл.10.5.

Например, расстояние между объектом X_1 и прототипами кластеров найдено следующим образом:

$$D(X_1, P_1^1) = \sqrt{(0,96 - 0,69)^2 + (0,57 - 0,22)^2} = 0,45,$$

$$D(X_1, P_2^1) = \sqrt{(0,96 - 0,89)^2 + (0,57 - 1)^2} = 0,44,$$

$$D(X_1, P_3^1) = \sqrt{(0,96 - 0,89)^2 + (0,57 - 0,57)^2} = 0,07.$$

Таким образом, объект X_1 относится к третьему кластеру.

Таблица 10.5 – Результаты разделения объектов на кластеры

Объект	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
P_1^1	0,45	0,47	0,51	0,11	0,21	0,46	0,11	0,81	0,42
P_2^1	0,44	0,44	0,30	0,87	0,60	0,44	0,76	0,00	0,43
P_3^1	0,07	0,11	0,16	0,45	0,25	0,09	0,40	0,43	0,02
Кластер	3	3	3	1	1	3	1	2	3

Таким образом, получено следующее разбиение объектов на кластеры: к первому кластеру относятся объекты X_4, X_5 и X_7, ко второму – только объект X_8, к третьему – объекты X_1, X_2, X_3, X_6, X_9.

5. В каждом кластере определяется новый объект-прототип:

$$P_1^2 = \left(\frac{0,77+0,69+0,61}{3}; \frac{0,14+0,43+0,29}{3} \right) = (0,69\,;\,0,29)\,,$$

$$P_2^2 = (0,89;\ 1,00),$$

$$P_3^2 = \left(\frac{0,96+1+0,81+0,98+0,91}{5}; \frac{0,57+0,57+0,71+0,57+0,57}{5} \right) = (0,93;\ 0,60)\,.$$

6. Выполняется сравнение прототипов, полученных на данной и на предыдущей итерациях. На данной итерации получены прототипы $P_1^2 = (0,69;\ 0,29)$, $P_2^2 = (0,89;\ 1,00)$, $P_3^2 = (0,93;\ 0,63)$, на предыдущей – прототипы $P_1^1 = (0,69;\ 0,22)$, $P_2^1 = (0,89;\ 1,00)$, $P_3^1 = (0,89;\ 0,57)$. Таким образом, прототипы не совпадают. Требуется следующая итерация. Выполняется возврат к шагу 3.

3. Выполняется **переход к очередной итерации: s = 3.**

4. Находятся расстояния от каждого из анализируемых объектов до каждого из прототипов по формуле (10.3), и выполняется отнесение каждого объекта к кластеру, представленному ближайшим прототипом. Результаты приведены в табл.10.6.

Таблица 10.6 – Результаты разделения объектов на кластеры

Объект	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
P_1^2	0,39	0,42	0,44	0,17	0,14	0,41	0,08	0,74	0,36
P_2^2	0,44	0,44	0,30	0,87	0,60	0,44	0,76	0,00	0,43
P_3^2	0,04	0,07	0,17	0,49	0,29	0,06	0,45	0,40	0,04
Кластер	3	3	3	1	1	3	1	2	3

Таким образом, получено следующее разбиение объектов на кластеры: к первому кластеру относятся объекты X_4, X_5 и X_7, ко второму – только объект X_8, к третьему – объекты X_1, X_2, X_3, X_6, X_9.

5. В каждом кластере определяется новый объект-прототип (как и на предыдущих итерациях): $P_1^3 = (0,69;\ 0,29)$, $P_2^3 = (0,89;\ 1,00)$, $P_3^3 = (0,93;\ 0,60)$.

6. Выполняется сравнение прототипов, полученных на данной и на предыдущей итерациях. Прототипы совпадают: $P_1^3 = P_1^2$, $P_2^3 = P_2^2$,. $P_3^3 = P_3^2$. Это означает, что получено окончательное разбиение объектов на кластеры.

Примечание. В качестве признака окончания алгоритма (т.е. окончательного разбиения) можно использовать совпадение разбиения на двух последовательных итерациях.

Таким образом, результаты разделения инвестиционных фондов на группы (кластеры) оказались следующими. К первой группе (фонды с низким риском и низкой прибылью) относятся фонды Ф4, Ф5, Ф7. Ко второй группе (фонды с высоким риском и высокой прибылью) можно отнести только фонд Ф8. В третью группу (фонды со средними значениями обоих показателей) входят фонды Ф1, Ф2, Ф3, Ф6, Ф9.

Метод максимина

Метод предназначен для разделения объектов на кластеры, причем количество кластеров **заранее неизвестно**; оно определяется автоматически в процессе разбиения объектов.

Принцип работы метода следующий. Выбирается один из объектов (любой); он становится прототипом первого кластера. Находится объект, наиболее удаленный от выбранного; он становится прототипом второго кластера. Все объекты распределяются по двум кластерам; каждый объект относится к кластеру, представленному ближайшим прототипом. Затем в каждом из кластеров находится объект, *наиболее удаленный* от своего прототипа. Если расстояние между этим объектом и прототипом кластера оказывается значительным (превышающим некоторую предельную величину), то объект становится новым прототипом, т.е. образуется новый кластер. После этого распределение объектов по кластерам выполняется заново. Процесс продолжается, пока не будет получено такое разбиение на кластеры, при котором расстояние от каждого объекта до прототипа кластера не будет превышать заданную предельную величину.

Приведем пошаговый **алгоритм р**еализации метода максимина.

- Выбирается любой из объектов, например, первый в списке объектов (X_1). Он становится прототипом первого кластера: $P_1=X_1$. Количество кластеров принимается равным единице: K=1.

1. Определяются расстояния от объекта P_1 до всех остальных объектов: $D(P_1, X_j)$, j=1,…,N. Определяется объект, наиболее удаленный от P_1, т.е. объект X_f, для которого выполняется условие: $D(P_1, X_f) = \max_j D(P_1, X_j)$. Этот объект становится прототипом второго кластера: $P_2=X_f$. Количество кластеров принимается равным двум: K=2.

2. Определяется пороговое расстояние. Оно принимается равным *половине* расстояния между прототипами P_1 и P_2: $T= D(P_1, P_2) / 2$. Эта величина будет использоваться для проверки условия окончания алгоритма.

3. Находятся расстояния от каждого из анализируемых объектов до каждого из имеющихся объектов-прототипов. Выполняется отнесение каждого объекта к ближайшему кластеру, т.е. кластеру, для которого расстояние между этим объектом и прототипом кластера минимально.

4. В каждом кластере определяется объект, наиболее удаленный от прототипа своего кластера. Обозначим эти объекты как Y_k, k=1,...,K (здесь k – номер кластера, K – количество кластеров).

5. Для каждого из наиболее удаленных объектов, найденных на шаге 5, проверяется условие: $D(P_k,Y_k)<T$, k=1,…,K. Если это условие выполняется *для всех кластеров*, то алгоритм *завершается*. Если для некоторого объекта Y_k это условие не выполняется, то он становится прототипом нового кластера, и количество кластеров увеличивается на единицу (K=K+1). В результате этого шага количество кластеров K увеличивается на число, равное количеству новых кластеров.

6. Находится новое пороговое расстояние. Оно определяется как половина среднего арифметического всех расстояний между прототипами:

$$T = \frac{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} D(P_i, P_j)}{K \cdot (K-1)}. \qquad (10.7)$$

7. Выполняется возврат к шагу 4.

Таким образом, окончательным является разбиение, для которого во всех кластерах расстояние от прототипа кластера до каждого из объектов, входящих в этот кластер (даже до самого удаленного), не превышает некоторой предельной величины (порогового расстояния).

Рассмотрим реализацию приведенного алгоритма на примере (анализ информации об инвестиционных фондах).

1. Первый из объектов принимается в качестве прототипа первого кластера: $P_1 = X_1 = (0,96; 0,57)$. Количество кластеров принимается равным единице: K = 1.

2. Определяются расстояния от прототипа P_1 до всех остальных объектов по формуле (10.3). Эти расстояния приведены в табл.10.7.

Таблица 10.7 – Расстояния от прототипа P_1 до всех остальных объектов

Объект	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
$P_1 = X_1$	0,00	0,04	0,21	0,47	0,30	0,02	0,45	0,44	0,05

Здесь, например, расстояние между прототипом P_1 и объектом X_2 найдено следующим образом: $D(P_1,X_2)=\sqrt{(0,96-1)^2+(0,57-0,57)^2}=0,04$.

Из таблицы 10.7 видно, что наиболее удаленным от прототипа P_1 является объект X_4. Он становится прототипом второго кластера: $P_2=X_4=(0,77; 0,14)$. Количество кластеров становится равным двум: K=2.

3. Определяется пороговое расстояние: $T= D(P_1, P_2) / 2 = 0,235$.

4. Находятся расстояния от каждого из анализируемых объектов до каждого из имеющихся объектов-прототипов. По найденным расстояниям выполняется отнесение каждого объекта к кластеру, представленному ближайшим прототипом. Расстояния, а также результаты разбиения объектов на кластеры приведены в табл.10.8.

Таблица 10.8 – Результаты разбиения объектов на кластеры методом максимина

Объект	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
$P_1 = X_1$	0,00	0,04	0,21	0,47	0,30	0,02	0,45	0,44	0,05
$P_2 = X_4$	0,47	0,49	0,57	0,00	0,30	0,48	0,22	0,87	0,45
Кластер	1	1	1	2	2	1	2	1	1

Примечание. Если вычислить расстояния от объекта X_5 до прототипов P_1 и P_2 с большей точностью (до трех знаков), то они окажутся следующими: $D(P_1,X_5)=0,304$, $D(P_2,X_5)=0,301$. Поэтому объект X_5 отнесен ко второму кластеру.

5. В каждом кластере определяется объект, наиболее удаленный от прототипа *своего* кластера. Из таблицы 10.8 видно, что в первом кластере таким объектом является X_8, во втором – X_5. Таким образом, $Y_1=X_8$, $Y_2=X_5$.

6. Расстояния между наиболее удаленными объектами (найденными на шаге 5) и объектами-прототипами сравниваются с пороговым расстоянием. В данном примере наиболее удаленным объектом первого кластера оказался объект X_8. Расстояние между этим объектом и прототипом P_1 равно 0,44; оно превышает пороговое расстояние (0,235). Таким образом, можно считать, что объект X_8 существенно отличается от прототипа своего кластера. Поэтому он становится прототипом нового кластера (P_3), и количество кластеров увеличивается на единицу ($K=3$). Во втором кластере наиболее удаленным объектом является X_5. Расстояние между этим объектом и прототипом P_2 равно 0,3; оно превышает пороговое расстояние. Поэтому объект X_5 также становится прототипом нового кластера (P_4), и количество классов увеличивается еще на единицу ($K=4$).

7. Находится новое пороговое расстояние. Оно определяется как половина среднего арифметического всех расстояний между прототипами. В данном случае имеется четыре прототипа (X_1, X_4, X_8, X_5). Требуется найти расстояния между ними: $D(X_1,X_4)=0,47$, $D(X_1,X_8)=0,44$, $D(X_1,X_5)=0,3$, $D(X_4,X_8)=0,87$, $D(X_4,X_5)=0,3$, $D(X_8,X_5)=0,6$. Новое пороговое расстояние находится по формуле (10.7) следующим образом:

$$T = \frac{0,47 + 0,44 + 0,3 + 0,87 + 0,3 + 0,6}{4 \cdot 3} = 0,25 .$$

8. Выполняется возврат к шагу 4.

4. Находятся расстояния от каждого из анализируемых объектов до каждого из имеющихся объектов-прототипов. По найденным расстояниям выполняется отнесение каждого объекта к кластеру, представленному ближайшим прототипом. Результаты приведены в табл.10.9.

Таблица 10.9 – Результаты отнесения каждого объекта к кластеру, представленному ближайшим прототипом

Объект	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
$P_1 = X_1$	0,00	0,04	0,21	0,47	0,30	0,02	0,45	0,44	0,05
$P_2 = X_4$	0,47	0,49	0,57	0,00	0,30	0,48	0,22	0,87	0,45
$P_3 = X_8$	0,44	0,44	0,30	0,87	0,60	0,44	0,76	0,00	0,43
$P_4 = X_5$	0,30	0,34	0,30	0,30	0,00	0,32	0,16	0,60	0,26
Кластер	1	1	1	2	4	1	4	3	1

5. В каждом кластере определяется объект, наиболее удаленный от прототипа *своего* кластера. Из табл. 10.9 видно, что в первом кластере таким объектом является X_3, в четвертом – X_7. Во второй и третий кластеры входит по одному объекту (X_4 и X_8 соответственно). Таким образом, $Y_1=X_3$, $Y_2=X_4$, $Y_3=X_8$, $Y_4=X_7$.

6. Расстояния между наиболее удаленными объектами (найденными на шаге 5) и объектами-прототипами сравниваются с пороговым расстоянием. В данном примере $D(P_1,Y_1)=0{,}21$, $D(P_2,Y_2)=0$, $D(P_3,Y_3)=0$, $D(P_4,Y_4)=0{,}16$. Все эти расстояния меньше порогового (T=0,25). Таким образом, можно считать, что во всех кластерах объекты достаточно близки к своим прототипам, т.е. имеют сходные значения признаков. Алгоритм завершается.

Таким образом, результаты разделения инвестиционных фондов на группы (кластеры) оказались следующими. К первому кластеру относятся фонды Ф1, Ф2, Ф3, Ф6, Ф9, ко второму – фонд Ф4, к третьему – Ф8, к четвертому - Ф5 и Ф7. Интерпретация этих результатов возможна только на основе анализа, выполняемого специалистами в соответствующей предметной области (в данном случае – экспертами-экономистами). Проанализировав показатели фондов, приведенные в табл.10.2, можно предложить следующую интерпретацию полученного разбиения. Первый кластер включает фонды со средней и высокой прибылью и со средними оценками риска. Второй кластер, включающий только фонд Ф4, соответствует средней прибыли и самому низкому риску. Третий кластер (фонд Ф8) соответствует достаточно высокой прибыли и самому высокому риску. Четвертый кластер включает фонды с низкой прибылью и невысоким (ниже среднего) риском.

Задания

Варианты заданий 1-15 по кластерному анализу даны в таблице 1, а значения показателей производственно-хозяйственной деятельности предприятий машиностроения приведены в таблице 2.

Четные варианты выполняются с использованием метода K средних.

Нечетные варианты выполняются с использованием метода максимина.

Рассматриваются следующие показатели:

Y1 - производительность труда;

Y2 - индекс снижения себестоимости продукции;

Y3- рентабельность;

X4 - трудоемкость единицы продукции;

X5 - удельный вес рабочих в составе ППП;

X6 - удельный вес покупных изделий;

X7- коэффициент сменности оборудования;
X8- премии и вознаграждения на одного работника;
X9 - удельный вес потерь от брака;
X10 - фондоотдача;
X11 - среднегодовая численность ППП;
X12 - среднегодовая стоимость ОПФ;
X13 - среднегодовой фонд заработной платы ППП;
X14- фондовооруженность труда;
X15- оборачиваемость нормируемых оборотных средств;
X16 - оборачиваемость ненормируемых оборотных средств;
X17 - непроизводственные расходы.

Таблица 1 – **Варианты заданий 1-15 по кластерному анализу**

№ варианта	Номера показателей, X
1	1, 6, 8, 12, 17
2	1, 8, 11, 13, 17
3	1, 6, 8, 14, 17
4	1, 8, 11, 14, 17
5	1, 7, 11, 12, 13
6	1, 8, 9, 13, 14
7	1, 5, 6, 7, 9
8	1, 5, 9, 11, 17
9	3, 8, 10, 15, 16
10	3, 5, 10, 15, 17
11	3, 5, 7, 11, 12
12	3, 8, 9, 10, 17
13	2, 4, 5, 6, 9
14	2, 4, 6, 8, 9
15	2, 4, 5, 8, 17

Таблица 2 – **Таблица исходных данных**

№ пред-при-ятия	Y1	Y2	Y3	X4	X5	X6	X7	X8	X9	X10
1	9,26	204,2	13,26	0,23	0,78	0,40	1,37	1,23	0,23	1,45
2	9,38	209,6	10,16	0,24	0,75	0,26	1,49	1,04	0,39	1,30
3	12,11	222,6	13,72	0,19	0,68	0,40	1,44	1,80	0,43	1,37
4	10,81	236,7	12,85	0,17	0,70	0,50	1,42	0,43	0,18	1,65
5	9,35	62,0	10,63	0,23	0,62	0,40	1,35	0,88	0,15	1,91
6	9,87	53,1	9,12	0,43	0,76	0,19	1,39	0,57	0,34	1,68
7	8,17	172,1	25,83	0,31	0,73	0,25	1,16	1,72	0,38	1,94
8	9,12	56,5	23,39	0,26	0,71	0,44	1,27	1,70	0,09	1,89
9	5,88	52,6	14,68	0,49	0,69	0,17	1,16	0,84	0,14	1,94
10	6,30	46,6	10,05	0,36	0,73	0,39	1,25	0,60	0,21	2,06
11	6,22	53,2	13,99	0,37	0,68	0,33	1,13	0,82	0,42	1,96
12	5,49	30,1	9,68	0,43	0,74	0,25	1,10	0,84	0,05	1,02
13	6,50	146,4	10,03	0,35	0,66	0,32	1,15	0,67	0,29	1,85
14	6,61	18,1	9,13	0,38	0,72	0,02	1,23	1,04	0,48	0,88
15	4,32	13,6	5,37	0,42	0,68	0,06	1,39	0,66	0,41	0,62

Продолжение таблицы 2.2

№ пред-приятия	X11	X12	X13	X14	X15	X16	X17
1	26006	167,69	47750	6,40	166,32	10,08	17,72
2	23935	186,10	50391	7,80	92,88	14,76	18,39
3	22589	220,45	43149	9,76	158,04	6.48	26,46
4	21220	169,30	41089	7,90	93,96	21.96	22,37
5	7394	39,53	14257	5,35	173,88	11,88	28,13
6	11586	40,41	22661	9,90	162,30	12,60	17,55
7	26609	102,96	52509	4,50	88,56	11,52	21,92
8	7801	37,02	14903	4,88	101,16	8,28	19,52
9	11587	45,74	25587	3,46	166,32	11,52	23,99
10	9475	40,07	16821	3,60	140,76	32,40	21,76
11	10811	45,44	19459	3,56	128,52	11,52	25,68
12	6371	41,08	12973	5,65	177,84	17,28	18,13
13	26761	136,14	50907	4,28	114,48	16,20	25,74
14	4210	42,39	6920	8,85	93,24	13,32	21,21
15	3557	37,39	5736	8.52	126,72	17,28	22,97

Содержание отчета

1 Тема и цель работы.
2 Текст программы.
3 Результаты выполнения программы.

Контрольные вопросы

1 Для чего предназначена кластеризация? Характеристики кластера.
2 Методы нормирования признаков объектов.
3 Какие бывают меры различия между объектами?
4 Сущность метода K средних и метода максимина.