Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАТИКИ
И РАДИОЭЛЕКТРОНИКИ

Кафедра информационных технологий автоматизированных систем

Отчет по лабораторной работе №7
«АНАЛИЗ И ПРИНЯТИЕ РЕШЕНИЙ НА ОСНОВЕ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА»
Вариант №5

Выполнил:
Ст. Гр. 820601
Шведов А.Р.

Проверила:
Протченко Е.В.

Минск 2020

Задача

Анализируются сведения о продукции девяти предприятий (П1, П2,...,П9). Имеются следующие показатели.

Предприятие	П1	П2	П3	П4	П5	П6	П7	П8	П9
Доля продукции, поставляемой на экспорт, %	42	17	38	54	14	11	50	24	62
Доля высокотехнологичной продукции, %	30	20	24	60	10	7	30	20	42

Требуется выделить группы предприятий, имеющих сходные значения показателей.

При решении задачи с использованием метода K средних выделить следующие группы: 1) предприятия с высокой долей экспортной продукции и высокотехнологичной продукции; 2) предприятия с низкой долей экспортной продукции и высокотехнологичной продукции; 3) предприятия со средними значениями обоих показателей.

Решение

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 и по диапазону значений. Некоторые признаки могут указываться в виде балльных оценок. В некоторых случаях объекты описываются качественными (словесными) признаками.

Существует несколько методов нормировки. Обычно применяются следующие методы:

- 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;
- 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.

Выполним нормировку, используя **деление на максимальное значение**.

	П1	П2	П3	П4	П5	П6	П7	П8	П9
доля пр-ии на экспорт	42	17	38	54	14	11	50	24	62
доля доля пр-ии на экспорт	30	20	24	60	10	7	30	20	42
Средняя доля пр-ии на экспорт	0,67742	0,27419	0,6129	0,87097	0,22581	0,17742	0,80645	0,3871	1
Средняя доля высокотех. пр-ии	0,5	0,33333	0,4	1	0,16667	0,11667	0,5	0,33333	0,7

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:

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

0	0,43631	0,11901	0,53615	0,56131	0,63004	0,12903	0,33476	0,37955
	0	0,34521	0,89475	0,17355	0,2373	0,55774	0,1129	0,81317
		0	0,65314	0,45198	0,51954	0,21786	0,23544	0,48974
			0	1,05389	1,12307	0,50415	0,82376	0,32657
				0	0,06958	0,66952	0,23193	0,94012
					0	0,73663	0,30151	1,00842
						0	0,45126	0,27832
							0	0,71421

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.

Метод К средних

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. Для каждого кластера выбирается объект-*прототип* – 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.

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.

метод К средних									
3 кластера									
высокие доли пр-ии на экспорт и высокотех.									
низкие доли пр-ии на экспорт и высокотех.									
средние доли пр-ии на экспорт и высокотех.									
0 итерация									
объект	x1	x2	x3	x4	x5	x6	x7	x8	x9
P01 (x9)	0,37955	0,81317	0,48974	0,32657	0,94012	1,00842	0,27832	0,71421	0
P02 (x6)	0,63004	0,2373	0,51954	1,12307	0,06958	0	0,73663	0,30151	1,00842
P03 (x1)	0	0,43631	0,11901	0,53615	0,56131	0,63004	0,12903	0,33476	0,37955
кластер	3	2	3	1	2	2	3	2	1
1 итерация	P11	P12	P13						
	0,93548	0,26613	0,69892						
	0,85	0,2375	0,46667						
объект	x1	x2	x3	x4	x5	x6	x7	x8	x9
P11	0,43485	0,8392	0,55368	0,16329	0,98518	1,05472	0,37303	0,75344	0,16329
P12	0,48792	0,09617	0,38296	0,97326	0,08151	0,1499	0,60071	0,15433	0,86745
P13	0,03967	0,44517	0,10883	0,5604	0,56022	0,62807	0,11258	0,33914	0,38091
кластер	3	2	3	1	2	2	3	2	1

Таким образом, результаты разделения предприятий на группы (кластеры) оказались следующими. К первой группе (высокие доли пр-ии на экспорт и высокотехн.) относятся предприятия П4, П9. Ко второй группе (низкие доли пр-ии на экспорт и высокотехн.) можно отнести П2, П5, П6, П8. В третью группу (средние доли пр-ии на экспорт и

высокотехн.) входят П1, П3, П7.

Метод максимин

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.

1. 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.

2. 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.

3. 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.

4. 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.

5. 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).

6. 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.

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

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

8. Выполняется возвращение к шагу 4.

метод максимина											
объект	x1	x2	x3	x4	x5	x6	x7	x8	x9		
P1=X1	0	0,43631	0,11901	0,53615	0,56131	0,63004	0,12903	0,33476	0,37955		
											порог
объект	x1	x2	x3	x4	x5	x6	x7	x8	x9		0,31501764
P1=X1	0	0,43631	0,11901	0,53615	0,56131	0,63004	0,12903	0,33476	0,37955		
P2=X6	0,63004	0,2373	0,51954	1,12307	0,06958	0	0,73663	0,30151	1,00842		
кластер	1	2	1	1	2	2	1	2	1		новый порог
											0,38154326
объект	x1	x2	x3	x4	x5	x6	x7	x8	x9		
P1=X1	0	0,43631	0,11901	0,53615	0,56131	0,63004	0,12903	0,33476	0,37955		
P2=X6	0,63004	0,2373	0,51954	1,12307	0,06958	0	0,73663	0,30151	1,00842		
P3=X4	0,53615	0,89475	0,65314	0	1,05389	1,12307	0,50415	0,82376	0,32657		
кластер	1	2	1	3	2	2	1	2	3		

Таким образом, результаты разделения предприятий на группы (кластеры) оказались следующими. К первому кластеру относятся П1, П3, П7, ко второму – П2, П5, П6, П8, к третьему – П4, П9. Интерпретация этих результатов возможна только на основе анализа, выполняемого специалистами в соответствующей предметной области.

Проанализировав показатели, можно предложить следующую интерпретацию полученного разбиения. Первый кластер включает предприятия со средними долями пр-ии на экспорт и высокотехн. Второй кластер соответствует низким долям пр-ии на экспорт и высокотехн. Третий кластер - высокие доли пр-ии на экспорт и высокотехн