**5. ВЫЯВЛЕНИЕ МНОЖЕСТВ**

1. **Описание задачи**

Имеются данные о совместных продажах различных товаров. Необходимо выявить какие товары наиболее часто покупаются совместно.

Речь фактически идет о проведении специфического кластерного анализа, при котором товары, наиболее часто встречающиеся при совместных покупках относятся к одному множеству товаров (к одному кластеру).

1. **Пример**

На сетевом диске имеется файл "Множества.xls". Скопируйте его в свою рабочую папку.

В нем на листе “Продажи” приводятся сведения о примерно 6000 продаж. При этом указывается какие товары были куплены, в каждой покупке (1 – товар присутствовал в покупке, 0 – товар отсутствовал).

Расшифровка кодов товаров прведена на листе “Товары”.

Для определения количества совместных покупок каждого товара с каждым создается следующая таблица (рекомендуется придерживаться указанной в таблице адресации):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | AB | AC | AD | AE | AF | AG | AH | AI | AK | AL | AM | AN | AO | AP | AQ | AR | AS | AT | AU | AV | AW | AX | AY | AZ | BA | BB |
| 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 5 |  |  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |  |
| 6 |  | 1 | 2043 | 1262 | 1252 | 1169 | 1075 | 1011 | 1154 | 600 | 1238 | 470 | 478 | 411 | 322 | 307 | 475 | 586 | 201 | 227 | 296 | 605 | 432 | 255 | 237 |  |
| 7 |  | 2 | 1262 | 2019 | 1242 | 1150 | 1036 | 960 | 1145 | 592 | 1197 | 465 | 504 | 420 | 311 | 255 | 479 | 606 | 204 | 223 | 293 | 573 | 420 | 239 | 227 |  |
| 8 |  | 3 | 1252 | 1242 | 2016 | 1117 | 1015 | 964 | 1137 | 593 | 1218 | 490 | 507 | 439 | 310 | 288 | 483 | 604 | 206 | 236 | 279 | 598 | 450 | 270 | 255 |  |
| 9 |  | 4 | 1169 | 1150 | 1117 | 1929 | 972 | 925 | 1081 | 568 | 1164 | 452 | 500 | 410 | 294 | 273 | 478 | 548 | 199 | 228 | 287 | 580 | 393 | 252 | 262 |  |
| 10 |  | 5 | 1075 | 1036 | 1015 | 972 | 1835 | 868 | 1014 | 533 | 1070 | 433 | 438 | 375 | 268 | 258 | 456 | 558 | 210 | 218 | 281 | 530 | 382 | 247 | 214 |  |
| 11 |  | 6 | 1011 | 960 | 964 | 925 | 868 | 1784 | 954 | 574 | 1024 | 410 | 441 | 386 | 265 | 269 | 402 | 519 | 198 | 221 | 270 | 538 | 388 | 267 | 242 |  |
| 12 |  | 7 | 1154 | 1145 | 1137 | 1081 | 1014 | 954 | 1971 | 589 | 1154 | 458 | 494 | 397 | 288 | 272 | 461 | 584 | 208 | 237 | 298 | 616 | 392 | 272 | 251 |  |
| 13 |  | 8 | 600 | 592 | 593 | 568 | 533 | 574 | 589 | 2281 | 845 | 894 | 909 | 850 | 723 | 695 | 312 | 394 | 197 | 216 | 204 | 948 | 561 | 1072 | 869 |  |
| 14 |  | 9 | 1238 | 1197 | 1218 | 1164 | 1070 | 1024 | 1154 | 845 | 2287 | 579 | 647 | 580 | 493 | 466 | 442 | 594 | 233 | 253 | 308 | 742 | 483 | 421 | 337 |  |
| 15 |  | 10 | 470 | 465 | 490 | 452 | 433 | 410 | 458 | 894 | 579 | 1592 | 809 | 613 | 458 | 456 | 200 | 272 | 128 | 146 | 154 | 643 | 410 | 849 | 755 |  |
| 16 |  | 11 | 478 | 504 | 507 | 500 | 438 | 441 | 494 | 909 | 647 | 809 | 1619 | 643 | 466 | 450 | 195 | 283 | 119 | 116 | 134 | 630 | 359 | 826 | 700 |  |
| 17 |  | 12 | 411 | 420 | 439 | 410 | 375 | 386 | 397 | 850 | 580 | 613 | 643 | 1493 | 461 | 464 | 232 | 296 | 156 | 146 | 140 | 701 | 454 | 712 | 594 |  |
| 18 |  | 13 | 322 | 311 | 310 | 294 | 268 | 265 | 288 | 723 | 493 | 458 | 466 | 461 | 1177 | 422 | 153 | 224 | 113 | 123 | 119 | 618 | 381 | 535 | 426 |  |
| 19 |  | 14 | 307 | 255 | 288 | 273 | 258 | 269 | 272 | 695 | 466 | 456 | 450 | 464 | 422 | 1153 | 152 | 216 | 103 | 117 | 111 | 596 | 361 | 540 | 433 |  |
| 20 |  | 15 | 475 | 479 | 483 | 478 | 456 | 402 | 461 | 312 | 442 | 200 | 195 | 232 | 153 | 152 | 1477 | 716 | 507 | 557 | 635 | 390 | 479 | 187 | 185 |  |
| 21 |  | 16 | 586 | 606 | 604 | 548 | 558 | 519 | 584 | 394 | 594 | 272 | 283 | 296 | 224 | 216 | 716 | 1683 | 538 | 578 | 647 | 458 | 535 | 211 | 214 |  |
| 22 |  | 17 | 201 | 204 | 206 | 199 | 210 | 198 | 208 | 197 | 233 | 128 | 119 | 156 | 113 | 103 | 507 | 538 | 1071 | 510 | 584 | 265 | 439 | 157 | 154 |  |
| 23 |  | 18 | 227 | 223 | 236 | 228 | 218 | 221 | 237 | 216 | 253 | 146 | 116 | 146 | 123 | 117 | 557 | 578 | 510 | 1145 | 605 | 294 | 442 | 164 | 170 |  |
| 24 |  | 19 | 296 | 293 | 279 | 287 | 281 | 270 | 298 | 204 | 308 | 154 | 134 | 140 | 119 | 111 | 635 | 647 | 584 | 605 | 1266 | 307 | 476 | 151 | 161 |  |
| 25 |  | 20 | 605 | 573 | 598 | 580 | 530 | 538 | 616 | 948 | 742 | 643 | 630 | 701 | 618 | 596 | 390 | 458 | 265 | 294 | 307 | 2065 | 642 | 835 | 687 |  |
| 26 |  | 21 | 432 | 420 | 450 | 393 | 382 | 388 | 392 | 561 | 483 | 410 | 359 | 454 | 381 | 361 | 479 | 535 | 439 | 442 | 476 | 642 | 1630 | 526 | 435 |  |
| 27 |  | 22 | 255 | 239 | 270 | 252 | 247 | 267 | 272 | 1072 | 421 | 849 | 826 | 712 | 535 | 540 | 187 | 211 | 157 | 164 | 151 | 835 | 526 | 1656 | 906 |  |
| 28 |  | 23 | 237 | 227 | 255 | 262 | 214 | 242 | 251 | 869 | 337 | 755 | 700 | 594 | 426 | 433 | 185 | 214 | 154 | 170 | 161 | 687 | 435 | 906 | 1463 |  |
| 29 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

При этом в ней должно быть 23 строки и 23 столбца.

Непосредственно для расчетов в ячейку AD6 вводится формула:

=СЧЁТЕСЛИМН($D5:$D$5814;$AD$5;D$5:D$5814;$AD$5), которая затем копируется во все ячейки таблицы.

Если присмотреться к полученным результатам, то можно отметить следующее:

* Элементы на главной диагонали представляют собой общее количество покупок i-ого товара. В этом нетрудно убедиться, если подсчитать суммы по каждому товару в исходной таблице;
* Данные по частоте совместных покупок симметричны относительно главной диагонали. Такая симметрия позволяет просматривать таблицу как постолбцам, так и по строкам (исходя из соображений удобства);
* Специалисты, знакомые с основами кластерного анализа, сразу увидят, что таблица очень напоминает матрицу расстояний, используемую при стандартном кластерном анализе. Там элементами таблицы является расстояния между объектами и чем меньше эти расстояния, тем с большей вероятностью объекты можно отнести к одному кластеру.

Здесь же вместо расстояний находятся частоты покупок и чем эти частоты больше, тем с большей вероятностью объекты можно отнести к одному кластеру.

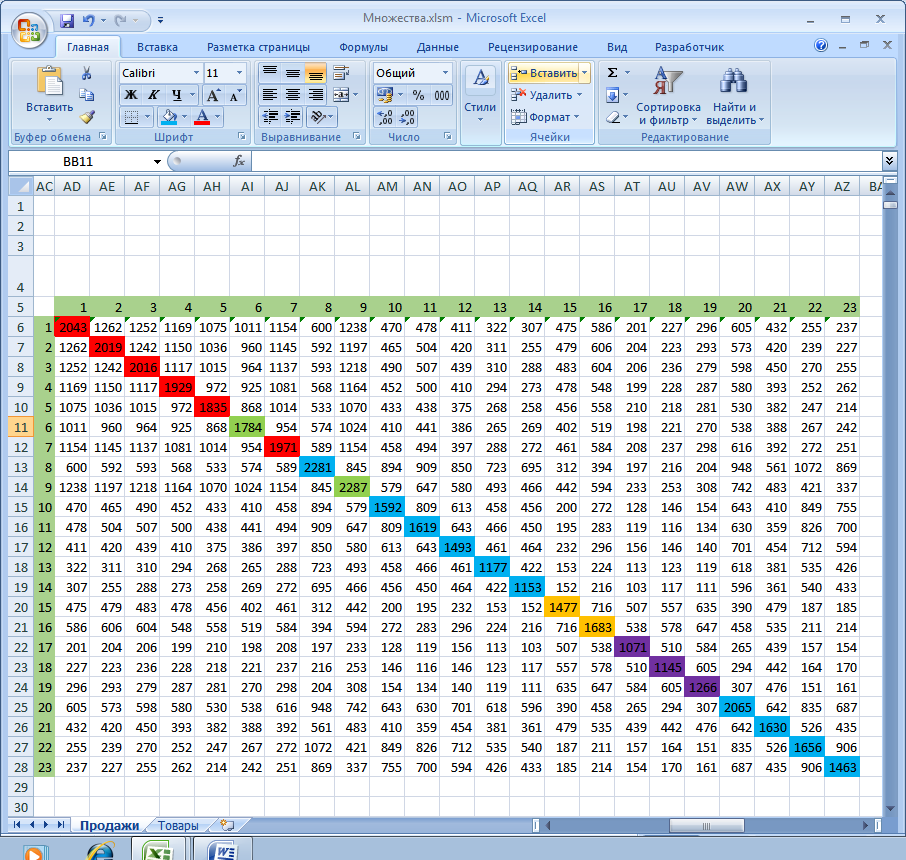
Последнее обстоятельство позволяет произвести кластерзацию вручную по следующей схеме:

1. Берется первый объект (в первом столбце) и заносится в первый кластер. Чтобы затем не ошибиться ячейку AD6 можно залить каким-то цветом (например, красным).
2. Переходим ко второму столбцу (второму объекту) и ищем в нем ячейку с максимальной частотой (при этом ячейка, находящаяся на главной диагонали не учитывается). Такой ячейкой является ячейка AE5, находящаяся в первой строке таблицы. Поскольку максимальная частота появления рассматриваемого второго объекта наблюдается совместно с первым объектом, то второй объект включается в первый кластер, где уже находится первый объект.

Поэтому закрашиваем ячейку AE7 цветом первого кластера (т.е. красным).

1. Переходим к третьему столбцу. Здесь максимальная частота также находится в первой строке. Т. е. третий объект чаще всего встречается вместе с первым. Поэтому и третий объект относим к первому кластеру и закрашиваем ячейку AF8 красным цветом.
2. Переходим к четвертому столбцу. После аналогичных рассуждений четвертый объект будет отнесен к первому кластеру и ячейка AG9 будет закрашена красным цветом.
3. Аналогично к первому кластеру будет отнесен и пятый объект.
4. В шестом столбце максимальная частота наблюдается в девятой строке. Поскольку ни шестой ни девятый объект еще “свободны” и максимально близки по частоте, о естественно предположить, что они представляют собой новый отдельный кластер. Закрасим их, например, зеленым цветом.
5. Седьмой объект имеет максимальную совместную частоту с первым объектом, а потому относим его к первому кластеру.
6. Восьмой объект имеет максимальную совместную частоту вместе с одинадцатым. Оба они пока еще “свободны”, а потому относим их к новому кластеру и соответствующие ячейки закрашиваем, например синим цветом.
7. Девятый столбец пропускаем, поскольку он уже включен во второй кластер.
8. Десятый объект имеет максимальную частоту в одинадцатой строке. Одинадцатый объект входит в третий кластер. Поэтому десятый объект также включаем в третий кластер.
9. И т.д.

В результате должно получиться, примерно, следующее:



Если перейти от кодов товаров к их названиям, то результат анализа будет выглядеть следующим образом:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Кластер 1 | Кластер 2 | Кластер 3 | Кластер 4 | Кластер 5 |
| Молоко | Хлеб белый | Алкоголь | Печенье | Зефир |
| Сметана | Йогурт | Хлеб черный | Конфеты | Шоколад |
| Творог |  | Мясо |  | Мармелад |
| Кефир |  | Колбаса |  |  |
| Сыр |  | Рыба |  |  |
| Ряженка |  | Курица |  |  |
|  |  | Индейка |  |  |
|  |  | Овощи |  |  |
|  |  | Фрукты |  |  |
|  |  | Табак |  |  |

При анализе реальных данных количество товаров может достигать нескольких сотен. Поэтому вместо ручных расчетов лучше использовать соответствующие программные средства.

В данном случае в редакторе Visual Basic файла «Множества.xls» приведен макрос, который воспроизводит вышеприведенный алгоритм выделения кластеров.

Перейдите в редактор Visual Basic и запустите макрос Raschet (при активном листе «Продажи). Результат будет напечатан в столбцах BD:BH.

1. **Варианты заданий**

|  |  |
| --- | --- |
| № варианта | Использовать исходные данные продаж, удалив столбец |
| 1 | 1 |
| 2 | 3 |
| 3 | 5 |
| 4 | 6 |
| 5 | 7 |
| 6 | 9 |
| 7 | 11 |
| 8 | 13 |
| 9 | 15 |
| 10 | 17 |
| 11 | 19 |
| 12 | 21 |
| 13 | 23 |