Статистические методы обработки экспериментальных данных

Лекция №8

Санкт-Петербург
2022

Кластерный анализ
Метод поиска сгущений

Метод поиска сгущений является еще одним итеративным методом кластерного анализа.

Основная идея метода заключается в построении гиперсферы заданного радиуса, которая перемещается в пространстве классификационных признаков в поисках локальных сгущений объектов.

Метод поиска сгущений **требует**, прежде всего, **вычисления матрицы расстояний** (или матрицы мер сходства) между объектами **и выбора первоначального центра сферы**.

Метод поиска сгущений

Как правило, на первом шаге центром сферы служит объект (точка), в ближайшей (заданной) окрестности которого расположено наибольшее число соседей. На основе заданного радиуса сферы *(R)* определяется совокупность точек внутри этой сферы, и для них вычисляются координаты центра (вектор средних для попавших в сферу значений признаков).

Когда очередной пересчет координат центра сферы приводит к такому же результату, как и на предыдущем шаге, перемещение сферы прекращается, а точки, попавшие в нее, образуют кластер, и из дальнейшего процесса кластеризации исключаются.

Метод поиска сгущений

Перечисленные процедуры повторяются для всех оставшихся точек. Работа алгоритма завершается за конечное число шагов, и все точки оказываются распределенными по кластерам. Число образовавшихся кластеров заранее неизвестно и сильно зависит от заданного радиуса сферы.

Для оценки устойчивости полученного разбиения целесообразно повторить процесс кластеризации несколько раз для различных значений радиуса сферы, изменяя каждый раз радиус на небольшую величину.

Метод поиска сгущений

Существуют различные способы выбора начального радиуса сферы. В частности, если обозначить через d_{ij} расстояние между i-м и j-м объектами, то в качестве нижней границы значения радиуса сферы можно выбрать минимальное из таких расстояний, а в качестве верхней границы – максимальное:

$$R_{\min} = \min_{i,j} d_{ij};\ R_{\max} = \max_{i,j} d_{ij}$$

Тогда, если начинать алгоритма работу с

$$R = R_{\min} + \delta;\ \delta > 0$$

и при каждом его повторении увеличивать значение δ на некоторую величину, то в конечном итоге можно найти значения радиусов, которые приводят к устойчивому разбиению на кластеры.

Метод поиска сгущений

Следует отметить следующие существенные при реализации метода поиска сгущений моменты:

1. В случае разномасштабности квалификационных признаков необходимо проведение их нормировки перед началом работы метода;
2. Возможны два варианта реализации метода. Один из них не предполагает изменения заданного значения радиуса сферы до завершения кластеризации, а другой – предполагает изменение этого радиуса в процессе кластеризации при начале построения очередной сферы;
3. В отличие от метода k-средних метод поиска сгущений не требует задания количества кластеров на которые предполагается разбить исходное множество объектов.

Метод поиска сгущений

4. Качество полученного в результате применения метода итогового разбиения на кластеры оценивается, как и методе k-средних, с помощью введенных на предыдущей лекции критериев качества разбиения F_1, F_2 и F_3;
5. Получение в результате кластеризации пересекающихся кластеров (наличие спорных объектов) в принципе является неудовлетворительным результатом. На практике в этом случае необходимо скорректировать процесс, либо выбрать другой метод кластеризации.

Метод поиска сгущений

Пример.
Пусть требуется произвести классификацию магазинов, квалификационные признаки, которых представлены в таблице:

Номер магазина	X_1	X_2	X_3	Номер магазина	X_1	X_2	X_3
1	100	160	25	6	85	200	35
2	130	200	30	7	60	170	28
3	80	180	20	8	110	150	18
4	40	100	22	9	55	110	15
5	150	90	15	10	110	100	12

Метод поиска сгущений

Рассчитаем расстояния между объектами по евклидовой метрике

$$d_{ij} = \sqrt{\sum_{k=1}^{3} \left(z_k^{(i)} - z_k^{(j)} \right)^2}$$

где:

$$z_k^{(i)} = \frac{X_k^{(i)} - \bar{X}_k}{\sigma_k}$$

$$Z = \begin{pmatrix} 0{,}243 & 0{,}345 & 0{,}426 \\ 1{,}156 & 1{,}338 & 1{,}136 \\ -0{,}365 & 0{,}838 & -0{,}284 \\ -1{,}582 & -1{,}134 & 0{,}000 \\ 1{,}764 & -1{,}381 & -0{,}994 \\ -0{,}273 & 1{,}332 & 1{,}847 \\ -0{,}973 & 0{,}592 & 0{,}852 \\ 0{,}547 & 0{,}099 & -0{,}568 \\ -1{,}125 & -0{,}888 & -0{,}994 \\ 0{,}547 & -1{,}134 & -1{,}420 \end{pmatrix}$$

Метод поиска сгущений

Получим матрицу расстояний:

$$D = \begin{bmatrix} 0 & 1{,}524 & 1{,}052 & 2{,}609 & 2{,}324 & 1{,}805 & 1{,}312 & 1{,}069 & 2{,}325 & 2{,}385 \\ & 0 & 2{,}140 & 3{,}860 & 3{,}507 & 1{,}596 & 2{,}274 & 2{,}193 & 3{,}833 & 3{,}608 \\ & & 0 & 2{,}335 & 3{,}156 & 2{,}189 & 1{,}312 & 1{,}208 & 2{,}015 & 2{,}452 \\ & & & 0 & 3{,}499 & 3{,}348 & 2{,}019 & 2{,}525 & 1{,}121 & 2{,}559 \\ & & & & 0 & 4{,}425 & 3{,}846 & 1{,}963 & 2{,}931 & 1{,}313 \\ & & & & & 0 & 1{,}424 & 2{,}833 & 3{,}705 & 4{,}175 \\ & & & & & & 0 & 2{,}138 & 2{,}371 & 3{,}233 \\ & & & & & & & 0 & 1{,}988 & 1{,}499 \\ & & & & & & & & 0 & 1{,}743 \\ & & & & & & & & & 0 \end{bmatrix}$$

Большая часть небольших расстояний находится в первой строке матрицы. Следовательно первый объект можно выбрать в качестве центра первой сферы.

Метод поиска сгущений

Зададим радиус сферы R=2.

Тогда в первый кластер попадают объекты (1,2,3,6,7,8).

Пересчитаем центр кластера. Получим точку

$$\overline{z}_* = (0,056;\ 0,757;\ 0,568)$$

Переносим центр кластера в эту точку и вновь проверяем объекты на попадание в первый кластер. Получим тот же список объектов (1,2,3,6,7,8). Поскольку центр кластера в этом случае не изменяется, **первый кластер считаем сформированным.**

Для формирования второго кластера нужно определиться с его начальным центром. Анализ матрицы расстояний позволяет в качестве этого центра выбрать объекты 9 или 10. Пусть это будет 9-й объект.

Метод поиска сгущений

Тогда при R=2 в этот кластер попадают объекты (4,8,9,10).

Если пересчитать центр кластера, то в кластер попадут объекты (1,3,4,8,9,10).

После нескольких итераций с пересчетом положения центра кластера и формированием нового списка попадающих в этот кластер объектов, получим, что во второй кластер войдут объекты (1,3,4,7,8,9,10).