1 Кластерный анализ

1.1 Описанире данных

Пусть из некоторого множества π (*генеральной совокупности*) отобрано n объектов $I = (I_1, \ldots, I_n)$. У каждого из этих объектов измеряется несколько характеристик $C = (C_1, \ldots, C_p)^T$. Мы предполагаем, что эти характеристики являются (*количественными.* Пусть далее x_{ij} есть результат измерения i-й характеристики у объекта I_j. Тогда $X_j = (x_{1j}, \ldots, x_{pj})^T$ есть измерения всех его характеристик у объекта I_j. В итоге мы имеем матрицу измерений:

$$X = (X_1, \ldots, X_n) \ .$$

Далее каждый набор X_j рассматривается как вектор в R^p.

1.2 Задача кластерного анализа

Пусть $m < n$. Требуется на основе измерений X разбить множество объектов I на m классов (*кластеров*) $\pi_1, \ldots, \pi_m$ так, чтобы:

1) каждый объект I_j принадлежалодному и только одному классу;

2) объекты внутри одного класса были бы в некотором смысле *сходными*;

3) объекты из разных классов были бы *несходными.*

Для решения этой задачи используется некоторая *целевая функция.* Она учитывает как число классов, так и качество группировки. Интуитивно ясно, объекты I_j и I_k нужно объединять в один класс, если расстояние между X_j и X_k будет *достаточно малым*, а для точек из разных классов оно должно быть *достаточно большим.*

1.3 Меры сходства

В этом разделе мы покажем как измерить близость выбранных объектов. Пусть в R^p задано некоторое расстояние (метрика) d. Рассмотрим несколько типичных примеров.

1) *Евклидово расстояние.*

$$d_2(X_k, X_j) := n[\sum_{i=1}^{p}(x_{ik} - x_{ij})^2]^{1/2} \ .$$

2) l_1*-норма.*

$$d_1(X_k, X_j) := n[\sum_{i=1}^{p}|x_{ik} - x_{ij}| \ .$$

3) *Максимальная норма.*

$$d_\infty(X_k, X_j) := \sup_i |x_{ik} - x_{ij}| \ .$$

4) *Расстояние Махаланобиса.*

$$D^2(X_k, X_j) := (X_k - X_j)^T W^{-1} (X_k - X_j) \ ,$$

где W есть так называемая матрица рассеяния, построенная по измерениям X (смотри ниже).

Важной особенностью этого расстояния является то, что оно не меняется при линейных преобразованиях: пусть $Y = B \cdot X$, тогда:

$$D^2(Y_k, Y_j) = D^2(X_k, X_j) \ .$$

Всюду далее через D будем обозначать матрицу попарных расстояний d_{kj} между векторами X_k X_j.

Определение 1 . *Неотрицательная функция $s(X_k, X_j) =: s_{kj}$ называется* **мерой сходства**, *если:*
1) $0 \le s(X_k, X_j) < 1$, *если* $X_k \neq X_j$;
2) $s(X_k, X_k) = 1$;
3) $s(X_k, X_j) = s(X_j, X_k)$.

Обозначим $S = (s_{kj})$.

Замечание 1 .
1) Величина s_{kj} называется **мерой сходства** *измерений X_k и X_j.*
2) Если X состоит из нулей и единиц, то s_{kj} называется **коэффициентом ассоциации** *или* **парным коэффициентом сопряженности**.
3) В статистике в качестве s_{kj} используют выборочные коэффициенты корреляции r_{kj}.

Рассмотрим теперь меры рассеяния и разнородности множества объектов $I = (I_1, \ldots, I_n)$.

Определение 2 . *Пусть X есть матрица измерений. Величина*

$$S_d := \frac{1}{2} \sum_{k,j=1}^{n} d(X_k, X_j)$$

называется **общим** *рассеянием множества объектов I, соответствующим данному расстоянию d.*

Определение 3 . *Величина* $\bar{s}_d := s_d/N_d$, *где* $N_d = (n^2 - n)/2$, *называется* **средним** *рассеянием множества объектов* I.

В статистике частоиспользуют следующие понятия.

Определение 4 . $p \times p$-*матрица*

$$S_X := \sum_{j=1}^{n} (X_j - \bar{X})(X_j - \bar{X})^T \tag{1}$$

называется **матрицей рассеяния** *множества* X, *где*

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j \ .$$

Определение 5 . *След матрицы* S_X *называется* **статистическим рассеянием** *множества* X *и обозначается*

$$s_t := trS_X = \sum_{j=1}^{n} \sum_{i=1}^{p} (X_{ij} - \bar{X}_i)^2 = \sum_{j=1}^{n} (X_j - \bar{X})^T (X_j - \bar{X}) =$$

$$\frac{1}{n} \sum_{j<k} d^2(X_j, X_k) \ .$$

Определение 6 . *Определитель* $|S_X|$ *матрицы* S_X *называется* **статистическим рассеиванием, соответствующим определителю**, *и обозначается* s_D.

1.4 Расстояние между кластерами

Общая идея кластерного анализа состоит в том. что кластеры строятся последовательно. Начинаем с единичных объектом, рассматриваемых как кластеры. если мы уже построили некоторую систему кластеров, то далее мы объединяем некоторые кластеры, которые лежат близко друг к другу. Окончательное разбиение получается с помощью некоторой *целевой функции*. Для ее построения используют понятия *внутренней однородности* кластера и меры *разнородности* кластеров между собой.

Пусть мы имеем два набора объектов $I = (I_1, \ldots, I_{n_1})$ и $J = (J_1, \ldots, J_{n_2})$, которые приводят к двум множествам измерений $X = (X_1, \ldots, X_{n_1})$ и $Y = (Y_1, \ldots, Y_{n_2})$.

Определение 7 . *Величина*

$$D_1(I, J) := \min_{k,j} d(X_k, Y_j)$$

называется **минимальным локальным расстоянием** *между кластерами* I *и* J.

Определение 8 . *Величина*

$$D_2(I,J) := \max_{k,j} d(X_k, Y_j)$$

называется **максимальным локальным расстоянием** *между кластерами* I *и* J.

Определение 9 . *Величина*

$$D_3(I,J) := \sum_k \sum_j d(X_k, Y_j)/n_1 \cdot n_2$$

называется **средним расстоянием** *между кластерами* I *и* J.

Определение 10 . *Величина*

$$D_4(I,J) := \frac{n_1 n_2}{n_1 + n_2}(\bar{X} - \bar{Y})^T(\bar{X} - \bar{Y})$$

называется **статистическим расстоянием** *между кластерами* I *и* J.

Смысл последнего понятия проясняется ниже. Рассмотрим кластер $K = I \cup J$. По аналогии с (1) определим

$$S_K = \sum_{k=1}^{n_1}(X_k - M)(X_k - M)^T + \sum_{j=1}^{n_2}(Y_j - M)(Y_j - M)^T \ , \tag{2}$$

где

$$M = (\sum_k X_k + \sum_j Y_j)/(n_1 + n_2) = (n_1\bar{X} + n_2\bar{Y})/(n_1 + n_2) \ . \tag{3}$$

Нетрудно показать, что

$$S_K = S_I + S_J + \frac{n_1 n_2}{n_1 + n_2}(\bar{X} - \bar{Y})(\bar{X} - \bar{Y})^T \ . \tag{4}$$

Определение 11 . *Матрица*

$$\frac{n_1 n_2}{n_1 + n_2}(\bar{X} - \bar{Y})(\bar{X} - \bar{Y})^T \tag{5}$$

называется **матрицей межгруппового рассеивания.**

Определение 12 . *Величину*

$$tr\frac{n_1 n_2}{n_1 + n_2}(\bar{X} - \bar{Y})(\bar{X} - \bar{Y})^T = \frac{n_1 n_2}{n_1 + n_2}(\bar{X} - \bar{Y})^T(\bar{X} - \bar{Y}) \tag{6}$$

называют **статистическим расстоянием** *между кластерами* I *и* J *или* **межгрупповой суммой квадратов**

Тогда (3) можно интерпретировать следующим образом: общая сумма квадратов складывается из внутригрупповой суммы квадратов и межгрупповой суммы квадратов.

1.5 Последовательное построение кластеров

Обычно кластеры строят последовательно. Общая схема такого построения имеет следующий вид:

1) Сначала все объекты рассматривают как отдельные кластеры.

2) Выбирают два порога s и r.

3) Если все кластеры находятся на расстоянии более, чем s, то все заканчивается.

4) Если есть кластеры, которые ближе друг к другу, чем s, то находим два наиболее близких и объединяем их.

5) Находим расстояния внутри кластеров и расстояния между кластерами.

6) Процедура продолжается до тех пор, пока расстояния внутри всех кластеров не более r, а между кластерами не более s.

Существуют различные методы реализации этой схемы. Они программно реализованы в ППП по статистике. Подробное описание можно найти в руководствах по кластерному анализу.

Процедура может привести и к построению только одного кластера, т.е. ничего разделить не удалось.

Оценка качества работы процедура - это отдельная важная задача.

Список литературы

[1] Мандель И.Д. Кластерный анализ.

[2] Дюран Б., Оделл П. Кластерный анализ

[3] Факторный, дискриминантный и кластерный анализ (сборник статей).