

Функции расстояния. Меры сходства

- Функция расстояния
- Виды функции расстояния
- Матрица сходства
- Меры сходства

Для того чтобы решить задачу кластерного анализа, построить алгоритм классификации, необходимо количественно определить функции расстояния, понятия сходства (близости) и разнородности (отдаленности) объектов, кластеров между собой.

Что означает утверждение “два объекта S_i, S_j различны”?

Это означает, что если объекты S_i, S_j попадают в один и тот же класс (при построении кластеров различными алгоритмами), когда расстояние (отдаленность) между соответствующими объектами было бы “достаточно малым” в определенном смысле, и, наоборот, попадают в разные классы, если расстояние между S_i, S_j будет “достаточно большим”. Таким образом, определяем понятие расстояния между объектами S_i, S_j из исходного множества допустимых объектов S .

Если объекты S_i, S_j являются точками n -мерного пространства, то исходное множество объектов S будет подмножеством евклидова пространства E_p .

Определение 2.1. Неотрицательная вещественнозначная функция $\rho(S_i, S_j)$ называется функцией расстояния (метрикой), если выполняемы аксиомы:

1. $\rho(S_i, S_j) \geq 0$ для всех S_i, S_j из S ;
2. $\rho(S_i, S_j) = 0$ тогда и только тогда, когда $S_i = S_j$;
3. $\rho(S_j, S_i) = \rho(S_i, S_j)$;
4. $\rho(S_i, S_j) \leq \rho(S_i, S_k) + \rho(S_k, S_j)$, где S_i, S_j и S_k - любые вектора из E_p .

Значение $\rho(S_i, S_j)$ для заданных S_i, S_j называется расстоянием между объектами S_i, S_j .

Аксиомы 1,2 определяют положительность метрики, аксиома 3 - определяет свойство симметричности, аксиома 4 - неравенство треугольника.

Евклидова метрика очень популярна и наиболее употребительна. Если $S_i = (i_1, i_2, \dots, i_n)$ и $S_j = (j_1, j_2, \dots, j_n)$, то евклидова метрика представляется как

$$\rho(S_i, S_j) = \sqrt{(i_1 - j_1)^2 + (i_2 - j_2)^2 + \dots + (i_n - j_n)^2} \quad (2.1)$$

При использовании этого расстояния будем считать, что компоненты (признаки) S_i взаимно независимы, однородны по своему физическому смыслу и все они одинаково важны с точки зрения вопроса об отнесении объекта к тому или иному классу, а также признаковое пространство совпадает с геометрическим пространством и понятие близости объектов соответственно совпадает с понятием геометрической близости в этом пространстве.

“Взвешенное” евклидово расстояние. Обычно применяется в ситуациях, когда некоторые признаки содержательно важны в смысле формирования кластеров, поэтому удается приписать каждому из признаков некоторый неотрицательный “вес” ω_i , “пропорциональный” степени его важности

$$\rho(S_i, S_j) = \sqrt{\omega_1(\alpha_1 - \beta_1)^2 + \omega_2(\alpha_2 - \beta_2)^2 + \dots + \omega_n(\alpha_n - \beta_n)^2} \quad (2.2)$$

Конечно, определение весов $\omega_i, i=1, n$ связано с дополнительным исследованием, таким как организация опроса экспертов и обработки их мнений, исследования обучающей выборки, использования специальных процедур предварительной обработки.

Хемингово расстояние. Используется как мера различия объектов, задаваемых дихотомическими признаками. Хемингово расстояние задается с помощью формулы

$$\rho_H(S_i, S_j) = \sum_{k=1}^n |\alpha_{ik} - \beta_{jk}|, \quad (2.3)$$

откуда видно что оно равно числу ij несовпадений значений соответствующих признаков в рассматриваемых S_i -м и S_j -м объектах.

Также можно использовать в качестве меры близости объектов S_i и S_j величину $(S_i, S_j) = ij/n$, где n - число признаков.

Также используются меры типа l – норма, наиболее простая с вычислительной точки зрения

$$\rho(S_i, S_j) = \sum_{k=1}^n |\alpha_{ik} - \beta_{jk}| \quad (2.3.1)$$

для произвольных значений признаков.

Сюпремум норма также легко вычисляется, включает себе процедуру упорядочения, и может использоваться в качестве меры близости

$$(2.4)$$

Наиболее часто используемыми (двумерный случай) в задачах классификации изображений и анализа сцен являются метрика абсолютного значения $d_A(S_1, S_2)$ и метрика максимального значения $d_M(S_1, S_2)$:

$$d_A(S_1, S_2) = |\alpha_1 - \beta_1| + |\alpha_2 - \beta_2|,$$

$$d_M(S_1, S_2) = \max \{ |\alpha_1 - \beta_1|, |\alpha_2 - \beta_2| \}.$$

Выбор соответствующей метрики имеет под собой более глубокие основания, чем может показаться на первый взгляд. Это связано с формой представления в ЭВМ информации о графическом изображении и его геометрических параметрах. Так, например, любые точки изображения в цифровом представлении определяются их целочисленными координатами в поле кадра. Результат вычисления евклидова расстояния между парой точек в общем случае не является целым. Если использовать здесь округление, то, как легко убедиться на примере, можно нарушить свойство 4 определения метрики. В то же время, используя метрику абсолютного значения, мы всегда получаем целые расстояния при целочисленных координатах точек. Можно, отметить, что метрика d_A более проста в вычислительном отношении.

Несимметрическая функция сходства основания на косинус угла используется в тех случаях, когда кластеры обнаруживают тенденцию располагаться вдоль главных осей, как это показано на рис. 2.3.

Функция

$$K(S, Z) = \cos \theta = \frac{S'Z}{\|S\| \|Z\|} \quad (2.5)$$

представляет собой косинус угла, образованного вектором S и Z , и достигающая максимума, когда их направления совпадают. Этой мерой сходства удобно пользоваться, когда кластеры достаточно отстоят друг от друга и от начала координат.

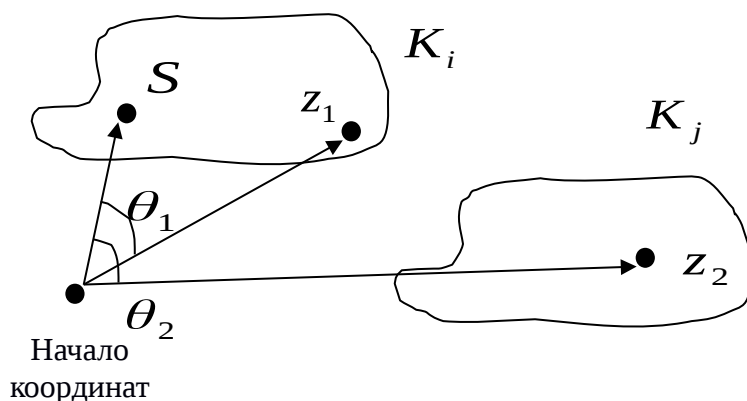


Рис. 2.3. Иллюстрация понятия меры сходства косинус угла

$$K(S, z_1) = \cos \theta_1 = \frac{S' z_1}{\|S\| \|z_1\|}, K(S, z_2) = \cos \theta_2 = \frac{S' z_2}{\|S\| \|z_2\|}.$$

Из рис. 2.3 видно, что образ z_1 обладает большим сходством с образом S , чем образ z_2 , поскольку значение $K(S, z_1)$ больше значения $K(S, z_2)$.

Когда рассматриваются двоичные образы и их элементы принимают значение из множества $[0,1]$, функции сходства (2.5) можно дать интересную негеометрическую интерпретацию.

Если $\alpha_i = 1$, считается, что двоичный объект S обладает i -м признаком. В таком случае член $S'z$ в формуле (2.5) просто характеризует число общих для образов S и z признаков, а $\|S\| \|z\| = \sqrt{(S'S)(z'z)}$ - среднее геометрическое число признаков, которыми обладает объект S , и число признаков, которыми обладает объект z . Понятно, что $K(S, z)$ есть мера общих признаков у объектов S и z .

Мера Танимото является двоичным вариантом формулы (2.5), который нашел широкое распространение в информационной классификации болезней и таксономии (классификация видов животных и растений)

$$K(S, z) = \frac{S'z}{S'S + z'z - S'z} \quad (2.5.1)$$

Расстояние **Махаланобиса** оказывается полезной мерой сходства в тех случаях, когда статистические характеристики объектов присутствуют в явном виде

$$D = (S - z)' C^{-1} (S - z),$$

где C - ковариационная матрица совокупности объектов; z - вектор средних значений; а S - объект.

Данная функция Махаланобиса успешно используется, когда известны плотности распределения объектов в классах, т.е. когда используются статистические свойства классов.

Исходное множество объектов

$$S = S_1, S_2, \dots, S_m, S_i = (i_1, i_2, \dots, i_n)$$

может быть представлено в виде матрицы исходных данных размером $m \times n$:

$$S = [S_1, S_2, \dots, S_m] = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \dots & \dots & \dots & \dots \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{mn} \end{pmatrix}.$$

Аналогичным образом расстояния между парами объектов могут быть представлены в виде симметричной матрицы расстояний:

$$C = \hat{C}, \quad (2.6)$$

Если для элементов матрицы расстояний C , применить некоторый порог ρ , т. е.:

$$C_{ij} = \begin{cases} 1, & \text{если } \rho_{ij} < \rho \\ 0, & \text{иначе} \end{cases}, i, j = 1, 2, \dots, m \quad (2.7)$$

тогда получаем симметричную матрицу сходства $C = \|C_{ij}\|_{m \times m}$, где $C_{ii} = 1$, $C_{ij} = C_{ji}$, $C_{ij} \in [0, 1]$.

Элемент C_{ij} матрицы сходства C называется мерой сходства объектов S_i, S_j относительно выбранного порога ρ .

Воспользуемся теперь введенным понятием расстояния для вычисления меры рассеяния или разнородности множества объектов

$$S = S_1, S_2, \dots, S_m.$$

Определение 2.2. Пусть задано множество

$S = S_1, S_2, \dots, S_m$ допустимых объектов (множество наблюдений).

Величина $r_\rho = \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \rho(S_i, S_j)$ называется общим рассеянием, соответствующим данной функции расстояния $\rho(S_i, S_j)$.

Определение 2.3. Величина $r_\rho = \frac{r_\rho}{M_\rho}$, где $M_\rho = \left(\frac{m^2 - m}{2} \right)$ называется средним рассеянием множества S .

Обоснование определений 2.2, 2.3 следует из рассмотрения матрицы расстояний $C = \|_{ij} \|_{m \times m}$ с учетом того, что, во-первых, для всех i , $\rho_{ii} = \rho(S_i, S_i)$, а во-вторых, из $(S_i, S_j) = (S_j, S_i)$ следует $c_{ij} = c_{ji}$ для всех i, j , $i, j = 1, 2, \dots, m$.

Отсюда величина r представляет собой сумму m^2 расстояний, из которых m равны нулю, и $\frac{m^2 - m}{2}$ различны и неотрицательны. Поэтому r есть арифметическое среднее ненулевых расстояний между парами объектов из S . Также применяется другой вид рассеяния.

Определение 2.4. Матрица $R_3 = \sum_{i=1}^m (S_i - \bar{S}) \bar{S}$ размером $n \times n$ называется матрицей рассеяния множества S , причем $\bar{S} = \sum_{i=1}^m S_i / m$ - есть вектор арифметических средних размером n .

Понятие меры внутренней однородности и меры разнородности классов иногда является необходимым условием при построении некоторых алгоритмов кластерного анализа.

Пусть $K_1 = S_1, S_2, \dots, S_{m_1}$ и $K_2 = S_{m_1+1}, S_{m_1+2}, \dots, S_{m_2}$ обозначают два класса объектов. Ниже приводятся наиболее употребительные меры близости между классами объектов.

Определение 2.5. Величину

$$D_1(K_1, K_2) = \min_{\substack{i=1, 2, \dots, m_1, \\ j=m_1+1, m_1+2, \dots, m_2}} (S_i, S_j), \text{ для всех}$$

будем называть минимальным локальным расстоянием между кластерами K_1 и K_2 , соответствующим данной функции расстояния. Другими словами, расстояние измерено по принципу “ближний сосед” БС.

Определение 2.6. Величину

$$D_2(K_1, K_2) = \max_{\substack{i=1, 2, \dots, m_1, \\ j=m_1+1, m_1+2, \dots, m_2}} (S_i, S_j), \text{ для всех}$$

назовем максимальным локальным расстоянием между K_1 и K_2 или расстоянием, измеренным по принципу “дальний сосед” ДС.

Определение 2.7. Величина

$$D_3(K_1, K_2) = \sum_{i=1}^{m_2} \sum_{j=1}^{m_1} \frac{\rho(S_i, S_j)}{m_1 \cdot m_2}$$

называется средним расстоянием между K_1 и K_2 относительно $\rho(S_i, S_j)$ или расстоянием измеренным по принципу “средней связи” СС.