МЕРКИ НА РАЗЛИЧИЕ И ПОДОБИЕ ПРИ АНАЛИЗА НА ДАННИ

Доц. д-р Здравко Славов

Мерките на различие и подобие са фундаментално важни в анализа на данни. Тези функции намират приложение в много научни области – информатика, статистика, мениджмънт, икономика, финанси, метеорология, биология, медицина и психология. В тази работа са въведени някои математически идеи и геометрични методи в анализа на данни, при които се използват мерки на различие и подобие.

1. Въведение

В природата и човешкото общество рядко съществуват две абсолютно еднакви неща. До голяма степен нещата са както различни, така и подобни (сходни). Някои неща са по-близки помежду си, а други са по-различни. Следователно в познавателната си дейност ние се нуждаем от възможността да сравняваме нещата и да оценяваме количествено колко те са различни и колко – подобни.

От друга страна, нещата имат много характеристики (признаци, свойства, особености, черти, features), които определят тяхното различие и подобие в сравнение с останалите. Теоретично броят на тези характеристики е безкраен, но практически във всяко научно изследване работим с краен брой характеристики, които определят едно крайномерно пространство. Кои характеристики ще участват в изследването, зависи от самото изследване и мнението на изследователите.

Най-общо казано, измерителите на различие и подобие представляват числови характеристики, които позволяват да се формулират изводи за относителното разположение на всеки елемент от изследваното множество в едно крайномерно пространство спрямо останалите елементи от множеството [16].

2. Математически модел

Нека изследваното множество M има $n \geq 3$ елемента и нека елементите му са номерирани с числата $i = 1,2,...n$. Освен това нека в изследването участват $m \geq 3$ числови случайни величини Y_j, $j = 1,2,...m$. В резултат на извършените наблюдения получаваме матрицата $A = \lfloor a_{ij} \rfloor$ с емпирични данни при $i = 1,2,...,n$ и $j = 1,2,...,m$. Имаме, че числото a_{ij} е наблюдението на i-тия елемент от множеството M спрямо j-тата величина, т.е. имаме $a_{ij} = Y_j$ за i-тия елемент от множеството M.

Тук имаме две възможности:

(1) Първата възможност е да измерваме различието или подобието между елементите на множеството M. В този случай ще работим с редовете на матрицата A. Всеки елемент от M ще се характеризира с m координати.

(2) Втората възможност е да измерваме различието или подобието между случайните величини $\{Y_j\}_{j=1}^{m}$. В този случай ще работим със стълбовете на матрицата A. Всяка случайна величина Y_j ще се характеризира с n координати.

Принципно различие между двете възможности няма. Без ограничение можем да се насочим към първата възможност. В този случай имаме следния математически модел: Всеки ред на матрицата A разглеждаме като точка или вектори в R^m.

Следователно разполагаме с точките $x_1, x_2, ..., x_n \in R^m$, които са съответните редове на матрицата A. Тези точки можем да разглеждаме и като вектори. За удобство няма да използваме различно означение за точка и вектор. Практически координатите на всяка една от точки $x_1, x_2, ..., x_n \in R^m$ имат долна и горна граница, следователно съществува ограничено числово множество $X \subset R^m$ – такова, че $x_i \in X$ при $i = 1,2,...,n$. Разбира се, теоретично е възможно да имаме $X = R^m$. За следващите разглеждания това няма значение. Съществен е фактът, че $X \subset R^m$, като е възможно и равенство.

3. Дефиниране на метрика

Ще стартираме нашето изложение със случая, когато елементите на матрицата A са числа. По-късно ще направим обобщение и за случаите, когато елементите на матрицата A може да не са само числа.

3.1. Разстояние между две точки

От геометрията идва идеята за разстояние между точките в реалното пространство. По този начин точките се различават на базата на разстоянието между тях. Тази идея е обобщена в произволно крайномерно пространство.

Разглеждаме функцията $d : R^m \times R^m \to R$, удовлетворяваща следните аксиоми при $x, y, z \in R^m$:

(m1) $d(x, y) \geq 0$;

(m2) $d(x, y) = d(y, x)$ (симетричност);

(m3) $d(x, y) + d(y, z) \geq d(x, z)$ (неравенство на триъгълника);

(m4) Ако $x = y$, то $d(x, y) = 0$;

(m5) Ако $d(x, y) = 0$, то $x = y$.

Ясно е, че горните аксиоми се удовлетворяват в реалния смисъл, който влагаме в понятието “разстояние”.

Ако са изпълнени аксиомите (m1), (m2), (m3), (m4) и (m5), то казваме, че функцията d е метрика (distance measure). Двойката (X, d) се нарича метрично пространство.

Ако са изпълнени аксиомите (m1), (m2), (m3) и (m4), то казваме, че функцията d е псевдометрика (pseudodistance measure). Двойката (X, d) се нарича псевдометрично пространство. Разликата е в аксиома (m5).

Забелязваме, че функцията d е ограничена отдолу, но не е ограничена отгоре.

В топологията псевдометричните пространства се различават в известен смисъл много малко от метричните пространства, но все пак се различават. Аксиома (m5) се оказва несъществена за много дефиниции и твърдения. Разбира се, съществуват редица дефиниции и твърдения, при които аксиома (m5) се оказва съществена [4] [7]. Това е наложило подробното анализиране на ролята на тази аксиома за нуждите на топологията.

Пример 1. Необходимостта от аксиома (m5) може да се види в следния елементарен пример. Нека $d(x, y) = 0$ при всички $x, y \in X$. Лесно се установява, че тази функция удовлетворява аксиоми (m1), (m2), (m3) и (m4), но не удовлетворява аксиома (m5). Следователно тази функция дефинира псевдометрично пространство и от вида й следва, че няма да върши някаква конкретна работа.

Пример 2. За $x, y \in X$ нека $d(x,y) = 0$ при $x = y$ и $d(x,y) = 1$ при $x \neq y$. Лесно се установява, че тази функция удовлетворява аксиоми (m1), (m2), (m3), (m4) и (m5). Следователно тази функция дефинира метрично пространство и е много малко вероятно да може да се използва.

Горните примери са елементарни, но показателни за ролята на аксиома (m5).

Не е трудно да се провери, че ако d_1 е метрика, а d_2 е псевдометрика, то $d = d_1 + d_2$ е метрика.

Нека d_1 е метрика и $d(x,y) = \min(1, d_1(x,y))$ при $x, y \in R^m$. Лесно се установява, че d също е метрика, удовлетворяваща допълнителните условия при $x, y, z \in R^m$:

(1) $0 \le d(x,y) \le 1$;

(2) Ако $d_1(x,y) = d_1(x,z)$, то $d(x,y) = d(x,z)$;

(3) Ако $d_1(x,y) < d_1(x,z)$, то $d(x,y) \le d(x,z)$;

(4) Пораждащите топологии са еквиваленти [4].

Нека d_1 е метрика и $d(x,y) = \dfrac{d_1(x,y)}{1 + d_1(x,y)}$ при $x, y \in R^m$. Лесно се установява, че d също е метрика, удовлетворяваща допълнителните условия при $x, y, z \in R^m$:

(1) $0 \le d(x,y) < 1$;

(2) Ако $d_1(x,y) = d_1(x,z)$, то $d(x,y) = d(x,z)$;

(3) Ако $d(x,y) = d(x,z)$, то $d_1(x,y) = d_1(x,z)$;

(4) Ако $d_1(x,y) < d_1(x,z)$, то $d(x,y) < d(x,z)$;

(5) Ако $d(x,y) < d(x,z)$, то $d_1(x,y) < d_1(x,z)$;

(6) Пораждащите топологии са еквиваленти [4].

3.2. Норма и метрика

Разглеждаме функцията норма $\|\cdot\| : R^m \to R$, удовлетворяваща следните аксиоми при $x, y \in R^m$ и $a \in R$:

(n1) $\|x\| \ge 0$;

(n2) $\|a.x\| = |a|.\|x\|$;

(n3) $\|x\| + \|y\| \ge \|x + y\|$;

(n4) Ако $x = 0$, то $\|x\| = 0$;

(n5) Ако $\|x\| = 0$, то $x = 0$.

Най-популярните норми при $x(x_1, x_2, ..., x_m) \in R^m$ са:

(1) Норма L_1 – има вида $\|x\| = \sum_{i=1}^{m} |x_i|$;

(2) Норма L_2 – има вида $\|x\| = \sqrt{\sum_{i=1}^{m} x_i^2}$;

(3) Норма L_p, $p \ge 1$ – има вида $\|x\| = \left(\sum_{i=1}^{m} |x_i|^p \right)^{1/p}$;

(4) Норма L_∞ – има вида $\|x\| = \max_i |x_i|$.

Съществува директна връзка между норма и разстояние, изразена чрез равенствата: $d(x,y) = \|x - y\|$ и $\|x\| = d(x,0)$.

Важно свойство при разстоянието между две точки е, че прибавянето на някоя константа към определена координата и за двете точки не оказва значение за разстоянието. Това не важи за нормата на един вектор.

Пример 3. Най-популярните метрики при $x(x_1, x_2, ..., x_m), y(y_1, y_2, ..., y_m) \in R^m$ са:

(1) При използване на норма L_1 имаме

$d(x,y) = \sum_{i=1}^{m} |x_i - y_i|$.

Горната метрика се нарича абсолютна, линейна или абсолютно-линейна (Manhattan distance, city-block, taxicab).

(2) При използване на норма L_2 имаме

$d(x,y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$.

Горната метрика се нарича Евклидова (Euclidean distance).

(3) При използване на норма L_p, $p \geq 1$, имаме

$$d(x,y) = \left(\sum_{i=1}^{m} |x_i - y_i|^p \right)^{1/p}.$$

Горната метрика се нарича разстояние на Минковски (Minkowski distance). Ясно е, че предните две метрики са частни случаи на тази метрика съответно при $p = 1$ и $p = 2$.

(4) При използване на норма L_∞ имаме

$d(x,y) = \max_i |x_i - y_i|$.

Горната метрика се нарича разстояние на Чебишов (Chebyshev distance).

Сега да разгледаме векторите $x(x_1, x_2, ..., x_m), y(y_1, y_2, ..., y_m) \in R^m$. Скаларното произведение на тези два вектора се дефинира по следния начин:

$\langle x, y \rangle = \sum_{i=1}^{m} x_i . y_i$.

Естествено, при $x = y$ имаме

$\langle x, x \rangle = \sum_{i=1}^{m} x_i . x_i = \sum_{i=1}^{m} x_i^2 = \|x\|$ (норма L_2).

Следователно за Евклидово разстояние получаваме:

$d(x,y) = \langle x - y, x - y \rangle$.

3.3. Свойства на разстоянието

Сега нека имаме функцията $h : R_+ \to R_+$ със следните свойства при $a, b \in R_+$:

(1) $h(0) = 0$;

(2) Ако $a < b$, то $h(a) < h(b)$ (строго монотонно растяща);

(3) $h(a) + h(b) \geq h(a + b)$.

Получаваме, че ако d е метрика, то $h \circ d$ също е метрика [7]. Така можем да генерираме нови метрики. Желателно е функцията h да е непрекъсната. Аналогично, ако d е псевдометрика, то $h \circ d$ също е псевдометрика.

Функцията $h: R_+ \to R_+$ може да се дефинира и по друг начин, а именно, ако при $a, b \in R_+$ има свойствата:

(1) $h(0) = 0$;

(2) Ако $h(a) = 0$, то $a = 0$;

(3) Ако $a < b$, то $h(a) \le h(b)$ (монотонно растяща);

(4) $h(a) + h(b) \ge h(a + b)$.

Пример 4. За всяка една от метриките, посочени в *пример 3* е изпълнено

$$d(a.x + b.e, a.y + b.e) = |a|.d(x, y)$$

където $a, b \in R$, $x, y \in R^m$ и $e = (1,1,...,1)$.

Сега да разгледаме правата l, определена от точка $x(x_1, x_2, ..., x_m) \in R^m$ и вектор $s(s_1, s_2, ..., s_m) \in R^m$. Нека точките $y, z \in l$ и са от едната страна на точка x в следната последователност: x, y, z. От това следва, че има две числа с еднакви знаци $a, b \in R$ – такива, че $y = x + a.s$ и $z = y + b.s$, следователно $z = x + (a + b).s$. Имаме $d(x, y) = |a|.d(0, s)$, $d(y, z) = |b|.d(0, s)$ и $d(x, z) = |a + b|.d(0, s)$. Окончателно получаваме

$$d(x, y) + d(y, z) = d(x, z).$$

Получихме един частен случай на неравенството на триъгълника, когато трите точки, лежат на една права. Това се покрива с реалната ни представа за разстояние между три точки лежащи на една права.

В предходния пример срещнахме равенството $d(x, y) + d(y, z) = d(x, z)$. Когато е изпълнено това равенство, казваме, че точката y е между точките x и z.

Пример 5. Трима ученици се характеризират с оценките по математика, физика и история, както следва: $x(5,4,6)$, $y(6,4,5)$ и $z(6,5,4)$. Видно е, че всеки от тях има една четворка, една петица и една шестица.

Първо да намерим разстоянието (различието) между учениците, използвайки абсолютна метрика. Така получаваме:

$$d_1(x, y) = |5-6| + |4-4| + |6-5| = 2;$$

$$d_1(x, z) = |5-6| + |4-5| + |6-4| = 4;$$

$$d_1(y, z) = |6-6| + |4-5| + |5-4| = 2.$$

Сега да използваме Евклидова метрика:

$$d_2(x, y) = \sqrt{(5-6)^2 + (4-4)^2 + (6-5)^2} = \sqrt{2};$$

$$d_2(x, z) = \sqrt{(5-6)^2 + (4-5)^2 + (6-4)^2} = \sqrt{6};$$

$$d_2(y, z) = \sqrt{(6-6)^2 + (4-5)^2 + (5-4)^2} = \sqrt{2}.$$

Да намерим разстоянието между учениците, използвайки разстоянието на Минковски при $p = 3$:

$$d_3(x, y) = \sqrt[3]{|5-6|^3 + |4-4|^3 + |6-5|^3} = \sqrt[3]{2};$$

$$d_3(x, z) = \sqrt[3]{|5-6|^3 + |4-5|^3 + |6-4|^3} = \sqrt[3]{10};$$

$$d_3(y, z) = \sqrt[3]{|6-6|^3 + |4-5|^3 + |5-4|^3} = \sqrt[3]{2}.$$

Накрая да използваме разстоянието на Чебишов:

$d_4(x,y) = \max(|5-6|,|4-4|,|6-5|) = 1;$

$d_4(x,z) = \max(|5-6|,|4-5|,|6-4|) = 2;$

$d_4(y,z) = \max(|6-6|,|4-5|,|5-4|) = 1.$

Получаваме различни резултати, но подредбата е винаги една и съща. Разстоянията (различията) между x и y и между y и z са най-малки и тези разстояния (различия) са равни. Най-голямо е разстоянието (различието) между x и z.

3.4. Метрика с тегла

Данните в статистическия анализ могат да бъдат от различен мащаб и да имат различна тежест. Това налага използването на тегла [1] [9]. Нека имаме тегла $\{w_i\}_{i=1}^m$, $w_i > 0$ при $i = 1,2,...,m$. Разглеждаме метриките, дефинирани в *пример 3*, но с тегла:

(1) Абсолютната метрика с тегла има вида

$d(x,y) = \sum_{i=1}^m w_i.|x_i - y_i|.$

(2) Евклидовата метрика с тегла има вида

$d(x,y) = \sqrt{\sum_{i=1}^m w_i.(x_i - y_i)^2}.$

(3) Метриката на Минковски с тегла има вида

$d(x,y) = \left(\sum_{i=1}^m w_i.|x_i - y_i|^p \right)^{1/p}.$

(4) Метриката на Чебишов с тегла има вида

$d(x,y) = \max_i w_i.|x_i - y_i|.$

Пример 6. Имаме трима кандидат-студенти, които се характеризират с оценките по математика, физика, история и български език, както следва: $x(6,5,4,5)$, $y(5,6,5,4)$ и $z(4,5,6,5)$. Видно е, че всеки от тях има една четворка, две петици и една шестица.

Ще разгледаме две ситуации за измерване на разстоянието (различието) между кандидат-студентите при използване на абсолютно разстояние.

(1) Ако тези кандидат-студенти кандидатстват за специалност “Математика”, практиката е да се изчислява средният успех с различна тежест на оценките. В този случай най-важна е оценката по математика, след това оценката по физика и сравнително малко участват оценките по история и български език. Например теглата нека са съответно 4, 2, 1 и 1. Така получаваме съответно притегления среден успех на тримата кандидат-студенти:

$\bar{x} = \frac{4.6 + 2.5 + 1.4 + 1.5}{4 + 2 + 1 + 1} = 5{,}375;$

$\bar{y} = \frac{4.5 + 2.6 + 1.5 + 1.4}{4 + 2 + 1 + 1} = 5{,}125;$

$\bar{z} = \frac{4.4 + 2.5 + 1.6 + 1.5}{4 + 2 + 1 + 1} = 4{,}625.$

Естествено, можем да приемем тези тегла да участват при определяне на разстоянието (различието) между кандидат-студентите. Така получаваме:

$d(x,y) = 4.|6-5| + 2.|5-6| + 1.|4-5| + 1.|5-4| = 8;$

$d(x,z) = 4.|6-4| + 2.|5-5| + 1.|4-6| + 1.|5-5| = 10;$

$d(y,z) = 4.|5-4| + 2.|6-5| + 1.|5-6| + 1.|4-5| = 8$.

Ако приемем теглата да са $w_1 = 7$, $w_2 = 1$, $w_3 = 1$ и $w_4 = 1$, получаваме:

$d(x,y) = 7.|6-5| + 1.|5-6| + 1.|4-5| + 1.|5-4| = 10$;

$d(x,z) = 7.|6-4| + 1.|5-5| + 1.|4-6| + 1.|5-5| = 18$;

$d(y,z) = 7.|5-4| + 1.|6-5| + 1.|5-6| + 1.|4-5| = 10$.

(2) Ако тези кандидат-студенти кандидатстват за специалност "Право", естествено е да използваме други тегла. Например можем да използваме теглата $w_1 = 1$, $w_2 = 1$, $w_3 = 4$ и $w_4 = 4$. В този случай получаваме съответно претегления среден успех на тримата кандидат-студенти:

$$\bar{x} = \frac{1.6 + 1.5 + 4.4 + 4.5}{1+1+4+4} = 4{,}70;$$

$$\bar{y} = \frac{1.5 + 1.6 + 4.5 + 4.4}{1+1+4+4} = 4{,}70;$$

$$\bar{z} = \frac{1.4 + 1.5 + 4.6 + 4.5}{1+1+4+4} = 5{,}30.$$

За разстоянията (различията) получаваме:

$d(x,y) = 1.|6-5| + 1.|5-6| + 4.|4-5| + 4.|5-4| = 10$;

$d(x,z) = 1.|6-4| + 1.|5-5| + 4.|4-6| + 4.|5-5| = 10$;

$d(y,z) = 1.|5-4| + 1.|6-5| + 4.|5-6| + 4.|4-5| = 10$.

Горният пример поставя въпроса как да избираме теглата. Това е въпрос, който няма еднозначен отговор и трябва да се решава за всеки отделен случай.

Има случаи, при които координатите са от различно естество и мащаб. По този начин влиянието на отделните координати при формиране на разстоянието ще бъде с различна тежест. Сега възниква задачата за изравняване тежестта на координатите. Най-често се използват техниките на нормиране и стандартизиране [1] [13]. При тях се стремим стойностите на различните координати да са в един и същ диапазон.

Пример 7. За нуждите на едно криминално разследване се анализират четири характеристики на четирима мъже: възраст, височина, тегло и номер на обувките. Разполага се със следните емпирични данни:

$x(28{,}174{,}71{,}42)$;

$y(26{,}181{,}68{,}42)$;

$z(32{,}178{,}69{,}43)$;

$v(24{,}177{,}68{,}41)$.

Виждаме, че координатите имат различен мащаб и ще оказват различно влияние при определяне на разстоянието (различието) между мъжете. Ако променим мерните единици на теглото от килограми в грамове, получаваме:

$x(28{,}174{,}71000{,}42)$;

$y(26{,}181{,}68000{,}42)$;

$z(32{,}178{,}69000{,}43)$;

$v(24{,}177{,}68000{,}41)$.

Ясно е, че сега участието на третата координата е много определящо за разстоянието между мъжете.

4. Дефиниране на функция на различие

В топологията аксиома (m3) има важно място, например тя е свързана с непрекъснатостта на функцията за разстояние [4]. Геометричният смисъл на тази аксиома е, че най-краткото разстояние между две точки е по правата линия, определена от тях. Ако се откажем от аксиома (m3), то това ни отдалечава от интуитивната ни представа за геометрично разстояние. Прието е при дискретни данни тази аксиома евентуално да се пропуска.

Разглеждаме функцията $d : X \times X \to R$, удовлетворяваща следните аксиоми при $x, y, z \in X$:

(d1) $d(x, y) \geq 0$;

(d2) $d(x, y) = d(y, x)$ (симетричност);

(d3) Ако $x = y$, то $d(x, y) = 0$;

(d4) Ако $d(x, y) = 0$, то $x = y$.

Ако са изпълнени аксиомите (d1), (d2), (d3) и (d4), то казваме, че функцията d е функция на различие (мярка на различие, dissimilarity measure) [11] [12] [13].

Ясно е, че всяка метрика се явява функция на различие. Обратното твърдение не е вярно, защото тук липсва аксиомата за триъгълника.

Ако са изпълнени аксиомите (d1), (d2) и (d3), то казваме, че функцията d е функция на псевдоразличие (мярка на псевдоразличие, pseudodissimilarity measure).

Функцията d е ограничена отдолу, но не е ограничена отгоре.

Практическото тълкуване на функцията на различие е, че с нарастване на нейната стойност $d(x, y)$ нараства и различието между елементите x и y. Аксиома (d3) ни осигурява, че при $x = y$ имаме $d(x, y) = 0$ и това е най-малката стойност на тази функция. Обратното твърдение е спорната аксиома (d4).

Пример 8. Най-популярната функцията на различие при $x(x_1, x_2, ..., x_m), y(y_1, y_2, ..., y_m) \in X$ се дефинира като

$$d(x, y) = \left(\sum_{i=1}^{m} |x_i - y_i|^p \right)^{1/q},$$

където $p \geq 1$ и $q \geq 1$ са параметри. При специални стойности на параметрите p и q получаваме:

(1) При $p = q = 1$ – линейна метрика.

(2) При $p = q = 2$ – Евклидова метрика.

(3) При $p = q$ – разстояние на Минковски.

(4) При $q = 1$ и $p \to \infty$ – разстояние на Чебишов.

Видяхме, че горните четири функции се явяват метрики в R^m.

(5) При $p = 2$ и $q = 1$ имаме

$d(x, y) = \sum_{i=1}^{m} (x_i - y_i)^2$.

Не е трудно да се провери, че аксиоми (d1), (d2), (d3) и (d4) са в сила.

Сега да разгледаме точките $x(-1,2,0)$, $y(0,0,0)$ и $z(2,-3,0)$. Те образуват тъпоъгълен триъгълник при Евклидова метрика. Пресмятаме $d(x, y) = 5$, $d(x, z) = 13$ и $d(y, z) = 34$. Получаваме, че

$d(x, y) + d(y, z) = 5 + 13 < 34 = d(x, z)$.

Следователно не в сила аксиома (m3). Така установяваме, че тази функция на различие не е метрика. Тя се нарича квадрат на Евклидово разстояние (squared Euclidean distance).

Този път да разгледаме точките $x(1,1,1)$, $y(3,3,3)$ и $z(8,8,8)$. Те лежат на една права и y е между x и z. Пресмятаме $d(x,y)=12$, $d(x,z)=147$ и $d(y,z)=75$. Окончателно получаваме $d(x,y)+d(y,z)=12+75\neq 127=d(x,z)$. Отново резултат, който противоречи на реалните ни представи.

Пример 9. За $x,y\in X$ нека $d(x,y)=\left(\sum_{i=1}^{m}(x_i-y_i)^2\right)^2$. Лесно се проверява, че се удовлетворяват аксиоми (d1), (d2), (d3) и (d4).

Сега да разгледаме точките $x(0,4,0)$, $y(0,0,0)$ и $z(3,0,0)$. Факт е, че те образуват правоъгълен триъгълник при Евклидова метрика. В нашата метрика (четвърта степен на Евклидово разстояние) пресмятаме $d(x,y)=256$, $d(x,z)=625$ и $d(y,z)=81$. Получаваме, че

$d(x,y)+d(y,z)=256+81<625=d(x,z)$.

Следователно не в сила аксиома (m3). Така установяваме, че тази функция на различие не е метрика.

Целта на горните два примера е да се изостри вниманието при използването на функции на различие, които не са метрики. Тук не трябва да се осланяме на геометрична интуиция, а на научно обосновани резултати.

Сега нека имаме функцията $h:R_+\to R_+$ със следните свойства:

(1) $h(0)=0$;

(2) При $a,b\in R_+$ ако $a<b$, то $h(a)<h(b)$ (строго монотонно растяща).

Получаваме, че ако d е функция на различие, то $h\circ d$ също е функция на различие. Така можем да генерираме нови функции на различие. Желателно е функцията h да е непрекъсната. Аналогично, ако d е функция на псевдоразличие, то $h\circ d$ също е функция на псевдоразличие.

Различните функции на различие се използват при различни методи за обработка на данни. Например при клъстерния анализ избирането на съответната функция на различие определя метода за клъстеризация на данните [12] [14].

Пример 10. Нека $x\in R^m$, $C\in R$ и $k\in N$. Разглеждаме k-ти момент относно C:

$m_k(x,C)=\frac{1}{m}\sum_{i=1}^{m}(x_i-C)^k$.

От *пример 8* следва, че при четно k имаме функцията на различие

$d(x,y)=\sum_{i=1}^{m}(x_i-y_i)^k$.

При $h(a)=\frac{1}{m}a$ получаваме, че

$d_1(x,y)=h(d(x,y))=\frac{1}{m}\sum_{i=1}^{m}(x_i-y_i)^k$

също е функция на различие.

При $y=(C,C,...,C)$ имаме, че k-ти момент относно C $m_k(x,C)=d_1(x,y)$ се явява специфична функция на различие.

Пример 11. Нека $x, y \in R^m$ и $C \in R$. Разглеждаме функцията на различие $d(x,y) = \sum_{i=1}^{m} |x_i - y_i|^p$.

При $p = 1$ и $y = (C, C, ..., C)$ имаме $d(x,y) = \sum_{i=1}^{m} |x_i - C|$. Търсим такова C, че $d(x,y) = \min\{d(x,z) : z \in R^m\}$. Известно е, че решението на този оптимизационен проблем е $C = median(x)$.

При $p = 2$ и $y = (C, C, ..., C)$ имаме $d(x,y) = \sum_{i=1}^{m} (x_i - C)^2$. Търсим такова C, че $d(x,y) = \min\{d(x,z) : z \in R^m\}$. Известно е, че решението на този оптимизационен проблем е $C = mean(x) = \frac{1}{m} \sum_{i=1}^{m} x_i$.

5. Дефиниране на функция на подобие

Идеята отново можем да намерим в геометрията, например при подобните триъгълници. Да разгледаме всички триъгълници с ъгли 40^o, 60^o и 80^o. Знаем, че те са подобни и страните им са пропорционални. Ако вземем отношението между страните на малкия към големия триъгълник, ще получим коефициент, известен като коефициент на подобие. Този коефициент има стойности от 0 до 1 и при стойност 1 триъгълниците са еднакви. Ако този коефициент не е 1, но е близко до 1, то триъгълниците ще имат почти равни страни.

Разглеждаме функцията $s : X \times X \to R$, удовлетворяваща следните аксиоми при $x, y \in X$:

(s1) $0 \le s(x,y) \le 1$;

(s2) $s(x,y) = s(y,x)$;

(s3) Ако $x = y$, то $s(x,y) = 1$;

(s4) Ако $s(x,y) = 1$, то $x = y$.

Ако са изпълнени аксиомите (s1), (s2), (s3) и (s4), то казваме, че функцията s е функция на подобие (мярка на подобие, similarity measure). Ако са изпълнени аксиомите (s1), (s2) и (s3), то казваме, че функцията s е функция на псевдоподобие (мярка на псевдоподобие) [9] [10] [15].

Забелязваме, че функцията s е ограничена отдолу и отгоре.

Практическото тълкуване на функцията на подобие е, че с нарастване на нейната стойност $s(x,y)$ нараства и подобието между елементите x и y. Аксиома (s3) ни осигурява, че при $x = y$ имаме $s(x,y) = 1$ и това е най-голямата стойност на тази функция. Обратното твърдение е спорната аксиома (s4). От аксиоматиката не следва, че се достига най-малката стойност на функцията.

Пример 12. Да разгледаме ненулевите вектори $x(x_1, x_2, ..., x_m), y(y_1, y_2, ..., y_m) \in R^m$ и нека q е ъгълът между тях. Известно е, че $\cos q = \frac{\sum_{i=1}^{n} x_i . y_i}{\sqrt{\sum_{i=1}^{m} x_i^2} . \sqrt{\sum_{i=1}^{m} y_i^2}}$. Лесно се установява, че $s(x,y) = |\cos q| = \frac{\left|\sum_{i=1}^{n} x_i . y_i\right|}{\sqrt{\sum_{i=1}^{m} x_i^2} . \sqrt{\sum_{i=1}^{m} y_i^2}}$ се явява функция на псевдоподобие при $X = R^m \setminus \{0\}$, т.е. удовлетворява аксиомите (s1), (s2) и (s3), но не удовлетворява аксиома (s4). Например при $x(1,0,...,0)$ и

$y(-1,0,...,0)$ имаме $s(x,y)=1$, а $x \neq y$. Лесно се установява, че при $x,y \in R^m \setminus \{0\}$ и $a,b \in R \setminus \{0\}$ имаме $s(a.x,b.y)=s(x,y)$. Сега, ако ограничим множеството X, например при $X=\{z \in R_+^m : \sum_{i=1}^m z_i = 1\}$ – част от единичната сфера, то s се явява функция на подобие.

От горния пример можем да направим извода, че не само видът на функцията s е важен, а също така и допустимото множество X.

Да отбележим, че ако функцията s удовлетворява аксиоми (s1), (s2), (s3) и (s4), то функцията $d(x,y)=1-s(x,y)$ удовлетворява аксиоми (d1), (d2), (d3) и (d4).

Също така, ако функцията d удовлетворява аксиоми (d1), (d2), (d3) и (d4), то функцията $s(x,y)=\dfrac{1}{1+d(x,y)}$ удовлетворява аксиоми (s1), (s2), (s3) и (s4).

Аналогично, ако функцията s удовлетворява аксиоми (s1), (s2) и (s3), то функцията $d(x,y)=1-s(x,y)$ удовлетворява аксиоми (d1), (d2) и (d3).

Също така, ако функцията d удовлетворява аксиоми (d1), (d2) и (d3), то функцията $s(x,y)=\dfrac{1}{1+d(x,y)}$ удовлетворява аксиоми (s1), (s2) и (s3).

От горното можем да направим извода, че съществува пряка връзка между функцията на различие и функцията на подобие, както и между функцията на псевдоразличие и функцията на псевдоподобие.

Пример 13. Ако d е функция на различие, то лесно се проверява, че $s(x,y)=e^{-d(x,y)}$ е функция на подобие.

Нека имаме функцията $h: R_+ \to R_+$ със следните свойства:

(1) $h(0)=0$ и $h(1)=1$;

(2) При $a,b \in R_+$, ако $a<b$, то $h(a)<h(b)$ (строго монотонно растяща).

Получаваме, че ако d е функция на подобие, то $h \circ d$ също е функция на подобие. Така можем да генерираме нови функции на подобие. Желателно е функцията h да е непрекъсната. Аналогично, ако d е функция на псевдоподобие, то $h \circ d$ също е функция на псевдоподобие.

6. Коефициент на корелация

Коефициентите на корелация са индекси, които описват до каква степен два елемента или две променливи са преди всичко линейно свързани. Те са известни още като мярка за взаимовръзка. При тях липсва единна аксиоматика, което им дава по-голяма свобода при дефинирането им [2] [3] [8].

Популярните коефициенти са от няколко типа.

(1) Първият тип се дефинират чрез функцията $r: X \times X \to R$, удовлетворяваща следните аксиоми при $x,y \in X$:

(r1) $-1 \le r(x,y) \le 1$;

(r2) $r(x,y)=r(y,x)$;

(r3) Ако $x=y$, то $r(x,y)=1$;

(r4) $r(x,-y)=-r(x,y)$.

(2) Вторият тип коефициенти на корелация се дефинират чрез функцията $r: X \times X \to R$, удовлетворяваща аксиомите (r1), (r2) и (r3).

(3) Третият тип коефициенти на корелация се дефинират чрез функцията $r : X \times X \to R$, удовлетворяваща аксиомите (s1), (s2) и (s3). Следователно те могат да се разглеждат и като функции на псевдоподобие.

Ясно е, че ако r е коефициент на корелация от първи или втори тип, то $|r|$ или r^2 са коефициенти на корелация от трети тип, т.е. функции на псевдоподобие. Оттук следва, че $1-|r|$ или $1-r^2$ са функции на псевдоразличие.

За коефициенти на корелация от първи и втори тип имаме $-1 \le r(x,y) \le 1$. Понякога знакът им има някакво адекватно тълкуване, а понякога – не. Това зависи както от самия коефициент, така и от характера на емпиричните данни.

Пример 14. Ако се върнем към *пример 12*, то имаме $r(x,y) = \cos q = \dfrac{\sum_{i=1}^{n} x_i . y_i}{\sqrt{\sum_{i=1}^{m} x_i^2} . \sqrt{\sum_{i=1}^{m} y_i^2}}$. Лесно се установява, че аксиоми (r1), (r2), (r3) и (r4) се удовлетворяват, следователно този корелационен коефициент е от първи тип. При $a, b \in R_+ \setminus \{0\}$ имаме $r(a.x, b.y) = r(x,y)$.

Пример 15. Ако в *пример 14* заменим ъгъл q с ъгъл j, който е ъгълът между векторите $m_x(x_1 - \bar{x}, x_2 - \bar{x}, ..., x_m - \bar{x})$ и $m_y(y_1 - \bar{y}, y_2 - \bar{y}, ..., y_m - \bar{y})$, получаваме коефициента на Пирсън

$$r(x,y) = \cos j = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m} (x_i - \bar{x})^2} . \sqrt{\sum_{i=1}^{m} (y_i - \bar{y})^2}} .$$

Тук имаме ограничението, че векторите $\vec{m}_x$ и $\vec{m}_y$ трябва да са ненулеви.

Корелационният коефициент на Пирсън е от първи тип и от него получаваме функции на псевдоподобие

$$s(x,y) = |r(x,y)| = \frac{\left|\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right|}{\sqrt{\sum_{i=1}^{m} (x_i - \bar{x})^2} . \sqrt{\sum_{i=1}^{m} (y_i - \bar{y})^2}} .$$

Нека разгледаме вектора $e(1,1,...,1) \in R^m$ и правата $l_0 = \{a.e \in R^m : a \in R\}$. При $a, b, c, d \in R \setminus \{0\}$ имаме $r(a.x + b.e, c.y + d.e) = r(x,y)$.

Сега, ако разгледаме функцията $h(a) = a^2$, получаваме нова функция на псевдоподобие

$$s_1(x,y) = h(s(x,y)) = \big(s(x,y)\big)^2 = \frac{\left(\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{m} (x_i - \bar{x})^2 . \sum_{i=1}^{m} (y_i - \bar{y})^2} .$$

Пример 16. Нека r е коефициентът на Пирсън. Да разгледаме коефициента $r_1 = r^3$, който също удовлетворява аксиоми (r1), (r2), (r3) и (r4). Следователно r_1 е корелационен коефициент, но е съмнително да изразява линейна зависимост.

Коефициентът на корелация на Пирсън има централно място в корелационния анализ. Той дава количествена оценка за линейната зависимост между две величини. Това е причината, когато говорим за корелационни коефициенти, да ги свързваме само с линейна зависимост, но видяхме, че това твърдение не е вярно за всички коефициенти.

7. Дефиниране на функция на различие между две множества с данни

Последователно ще разгледаме две конструкции, чрез които ще дефинираме съответно функция на различие и функция на псевдоразличие.

7.1. Дефиниране на функция на различие

Нека d е метрика. Да разгледаме крайните непразни подмножества на X. Целта е да дефинираме функция на различие за тези подмножества d, удовлетворяваща аксиоми (d1), (d2), (d3) и (d4). Ясно е, че d ще зависи от d и при едноелементи подмножества двете функции трябва да съвпадат.

Нека имаме две крайни непразни подмножества $A, B \subset X$. Първо ще дефинираме разстояние от точка $x \in A$ до множеството B чрез

$$d(x, B) = \min\{d(x, y) : y \in B\}.$$

Горното разстояние е най-малкото от разстоянията между x и елементите на B. Оттук дефинираме функция на различие на двете подмножества

$$d^*(A, B) = \max\{d(x, B) : x \in A\}.$$

Горната функция не е симетрична, т.е. съществуват две подмножества, за които $d^*(A, B) \neq d^*(B, A)$.

Съществуват два варианта на функцията d. Първият вариант е известен като разстояние на Хаусдорф (Hausdorff distance) и се дава чрез

$$d(A, B) = \max\{d^*(A, B), d^*(B, A)\}.$$

Проверява се, че горната функция удовлетворява аксиоми (d1), (d2), (d3) и (d4) [13].

Вторият вариант на функцията d има вида

$$d(A, B) = d^*(A, B) + d^*(B, A).$$

Проверява се, че горната функция също удовлетворява аксиоми (d1), (d2), (d3) и (d4).

7.2. Дефиниране на функция на псевдоразличие

В този случай конструкцията на функцията ще е по-проста, защото отпада аксиома (d4). Отново нека d е метрика. Да разгледаме крайните непразни подмножества на X. Целта е да дефинираме функция на псевдоразличие за тези подмножества d, удовлетворяваща аксиоми (d1), (d2) и (d3). Ясно е, че d ще зависи от d и при едноелементи подмножества двете функции трябва да съвпадат.

Нека имаме две крайни непразни подмножества $A, B \subset X$ и нека

$$d(A, B) = \min\{d(x, y) : x \in A, y \in B\}.$$

Горната функция удовлетворява аксиоми (d1), (d2) и (d3). Лесно може да се посочи пример, при който не се удовлетворява аксиома (d4).

8. Различие и подобие при бинарни данни

Досега разгледахме случая при работа с непрекъснати данни, а сега да преминем към дискретни. Първо да разгледаме случая, когато елементи на матрицата A могат да бъдат само числата 0 или 1. Така получаваме задачата да измерваме различието и подобието на бинарни данни [5] [6].

Нека разгледаме бинарните вектори $x, y \in \{0,1\}^m \subset R^m$. Сравнявайки съответните координати на тези вектори, получаваме четири случая:

(1) числото m_{01} е броят на координатите, при които координатата на x е 0 и координатата на y е 1;

(2) числото m_{10} е броят на координатите, при които координатата на x е 1 и координатата на y е 0;

(3) числото m_{00} е броят на координатите, при които координатата на x е 0 и координатата на y е 0;

(4) числото m_{11} е броят на координатите, при които координатата на x е 1 и координатата на y е 1;

Ясно е, че имаме $m_{01} + m_{10} + m_{00} + m_{11} = m$.

Пример 17. Ако използваме линейната метрика $d(x,y) = \sum_{i=1}^{m} |x_i - y_i|$, получаваме бинарната метрика (линейна бинарна метрика, binary squared Euclidean measure)

$d(x,y) = m_{01} + m_{10}$.

Ако използваме Евклидовата метрика $d(x,y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$, получаваме друга бинарна метрика (Евклидова бинарна метрика, binary Euclidean measure)

$d(x,y) = \sqrt{m_{01} + m_{10}}$.

Горните две бинарни метрики са известни още като метрики на Хеминг (Hamming distance).

Пример 18. От *пример 17* имаме бинарната метрика $d(x,y) = m_{01} + m_{10}$ и нека $h(a) = \frac{1}{m} a$, следователно

$$d_1(x,y) = h(d(x,y)) = \frac{1}{m} d(x,y) = \frac{m_{01} + m_{10}}{m_{00} + m_{01} + m_{10} + m_{11}}$$

също е бинарна метрика. Ясно е, че специфичното за тази метрика е, че имаме $0 \le d_1(x,y) \le 1$. Също така получаваме, че

$$s(x,y) = 1 - d_1(x,y) = \frac{m_{11} + m_{00}}{m_{00} + m_{01} + m_{10} + m_{11}}$$

е функция на подобие. Тя е известна с названието елементарен измерител на подобие (simple matching similarity measure).

При $h(a) = \frac{1}{4m} a$ получаваме нова бинарна метрика

$$d_2(x,y) = h(d(x,y)) = \frac{1}{4m} d(x,y) = \frac{m_{01} + m_{10}}{4(m_{00} + m_{01} + m_{10} + m_{11})},$$

известна като вариационен измерител на различие (variance dissimilarity measure) .

Пример 19. Ако обединим резултатите от *пример 13* и *16*, получаваме функциите на подобие:

$s_1(x,y) = e^{-m_{01} - m_{10}}$;

$s_2(x,y) = e^{-\sqrt{m_{01} + m_{10}}}$.

Пример 20. От *пример 17* също можем да получим още две функции на подобие:

$$s_1(x,y) = \frac{1}{1 + m_{01} + m_{10}};$$

$$s_2(x,y)=\frac{1}{1+\sqrt{m_{01}+m_{10}}}.$$

Пример 21. Ако приложим коефициента на Пирсън за бинарни данни, получаваме

$$r(x,y)=\frac{m_{00}.m_{11}-m_{01}.m_{10}}{\sqrt{(m_{00}+m_{01})(m_{10}+m_{11})(m_{00}+m_{11})(m_{01}+m_{11})}}.$$

Тук знакът няма съществен реален смисъл, следователно можем да разгледаме неговата абсолютна стойност и получаваме функцията на псевдоподобие за бинарни данни

$$s(x,y)=\frac{\left|m_{00}.m_{11}-m_{01}.m_{10}\right|}{\sqrt{(m_{00}+m_{01})(m_{10}+m_{11})(m_{00}+m_{11})(m_{01}+m_{11})}},$$

известна като бинарен корелационен коефициент (fourfold point correlation).

Ако разгледаме функцията $h(a)=a^2$, получаваме нова функция на псевдоподобие

$$s_1(x,y)=h(s(x,y))=\frac{\left(m_{00}.m_{11}-m_{01}.m_{10}\right)^2}{(m_{00}+m_{01})(m_{10}+m_{11})(m_{00}+m_{11})(m_{01}+m_{11})}.$$

Пример 22. Нека означим:

(1) числото m_0^x е броят на координатите на x, които са 0;

(2) числото m_1^x е броят на координатите на x, които са 1;

(3) числото m_1^y е броят на координатите на y, които са 0;

(4) числото m_1^y е броят на координатите на y, които са 1.

Получаваме две функции на подобие на Танимото (Tanimoto similarity measure):

$$s_1(x,y)=\frac{m_{00}}{m_0^x+m_0^y-m_{00}};$$

$$s_2(x,y)=\frac{m_{11}}{m_1^x+m_1^y-m_{11}}.$$

Аналогично получаваме две функции на подобие на Дайк (Dice similarity measure):

$$s_3(x,y)=\frac{2.m_{00}}{m_0^x+m_0^y};$$

$$s_4(x,y)=\frac{2.m_{11}}{m_1^x+m_1^y}.$$

Ясно е, че имаме: $m_0^x=m_{00}+m_{01}$, $m_1^x=m_{10}+m_{11}$, $m_0^y=m_{00}+m_{10}$ и $m_1^y=m_{01}+m_{11}$. Оттук получаваме съответните функции на различие на Танимото (Tanimoto dissimilarity measure):

$$d_1(x,y)=1-s_1(x,y)=1-\frac{m_{00}}{m_0^x+m_0^y-m_{00}}=\frac{m_{01}+m_{10}}{m_{01}+m_{10}+m_{00}};$$

$$d_2(x,y) = 1 - s_2(x,y) = 1 - \frac{m_{11}}{m_1^x + m_1^y - m_{11}} = \frac{m_{01} + m_{10}}{m_{01} + m_{10} + m_{11}}.$$

Естествено, получаваме и съответните функции на различие на Дайк (Dice dissimilarity measure):

$$d_3(x,y) = 1 - s_3(x,y) = 1 - \frac{2.m_{00}}{m_0^x + m_0^y} = \frac{m_{01} + m_{10}}{m_{01} + m_{10} + 2.m_{00}};$$

$$d_4(x,y) = 1 - s_4(x,y) = 1 - \frac{2.m_{11}}{m_1^x + m_1^y} = \frac{m_{01} + m_{10}}{m_{01} + m_{10} + 2.m_{11}}.$$

Пример 23. Функциите на подобие на Танимото от *пример 22* може да се преработят и в новия си вид те са известни като функции на подобие на Джакард (Jaccard similarity measure):

$$s_1(x,y) = \frac{m_{00}}{m_0^x + m_0^y - m_{00}} = \frac{m_{00}}{m_{01} + m_{10} + m_{00}};$$

$$s_2(x,y) = \frac{m_{11}}{m_1^x + m_1^y - m_{11}} = \frac{m_{11}}{m_{01} + m_{10} + m_{11}}.$$

Нека разгледаме бинарните данни $x(1,0,1,1,0,1,1,1)$ и $y(0,1,0,1,1,1,1,1)$. В този случай получаваме:

$$s_1(x,y) = \frac{0}{2+2+0} = 0;$$

$$s_2(x,y) = \frac{4}{2+2+4} = 0{,}5.$$

Двата коефициента дават различни резултати и естествено е да си зададем въпроса на кой от двата да вярваме, т.е. да приемем за меродавен. От факта, че функциите на подобие на Джакард или Танимото намират голямо практическо приложение, следва да си зададем въпроса: В кой случай коя функция да използваме?

Да сравним функцията на подобие (елементарен измерител на подобие) от *пример 18* $s(x,y) = \frac{m_{11} + m_{00}}{m_{00} + m_{01} + m_{10} + m_{11}}$ с двете функции на Джакард $s_1(x,y) = \frac{m_{00}}{m_{01} + m_{10} + m_{00}}$ и $s_2(x,y) = \frac{m_{11}}{m_{01} + m_{10} + m_{11}}$. При функцията s_1 от числителя и знаменателя липсва m_{11}, а при s_2 от числителя и знаменателя липсва m_{00}. Следователно при функциите на Джакард умишлено не се отчита част от информацията. Например в биологията освен елементарният измерител на подобие се използва и първата функция на Джакард при необходимостта да се разгледа т. нар. негативна двойка, т.е. при едновременното отсъствие на дадена характеристика при нейно означаване с 0 [12]. За нашия пример имаме:

$$s(x,y) = \frac{m_{11} + m_{00}}{m_{00} + m_{01} + m_{10} + m_{11}} = \frac{4+0}{0+2+2+4} = 0{,}5.$$

Следователно s и s_2 имат стойност 0,5, а s_1 има стойност 0. Ако гледаме подобието на двата елемента x и y относно общите нули (липса на характеристика), то според s_1 е нулево. Наистина двата елемента нямат общи нули и оттук липсва подобие.

9. Различие и подобие при номинални данни

Досега разгледахме случая, когато стойностите на величините са числа. Величините, между стойностите на които не съществува някаква наредба, се наричат номинални. Сега да разгледаме случая, когато всяка координата може да бъде една от няколко възможни номинални стойности. Нека имаме крайно множество от номинални стойности (последователност от един или повече символи, стингове, string) $a=\{a_1,a_2,...,a_k\}$, $k\geq 2$. Нека координатите на x и y са някои от елементите на a, следователно $x,y\in a^m$. При $k=2$ можем да приемем, че символите са 0 или 1, т.е. получаваме бинарни данни. Следователно сега разглеждаме по-общ проблем.

Нека означим при $i=1,2,...,m$

$$d_i=\begin{cases}0, & x_i=y_i\\ 1, & x_i\neq y_i\end{cases}.$$

Разглеждаме функцията $d(x,y)=\sum_{i=1}^{m}d_i$. Ясно е, че $0\leq d(x,y)\leq m$. Тази функция е известна като метрика на Хеминг (Hamming distance) и удовлетворява аксиоми (m1), (m2), (m3), (m4) и (m5) [13].

При $h(a)=\sqrt{a}$ получаваме друга метрика на Хеминг

$$d_1(x,y)=h(d(x,y))=\sqrt{\sum_{i=1}^{m}d_i}\ .$$

При $a=\{0,1\}$ получаваме бинарните метрики на Хеминг.

Нека имаме тегла $\{w_i\}_{i=1}^{m}$, $w_i>0$ при $i=1,2,...,m$. Често се поставя допълнителното изискване $\sum_{i=1}^{m}w_i=1$ или $\sum_{i=1}^{m}w_i=m$, но то не е задължително. Доказва се, че функциите $d(x,y)=\sum_{i=1}^{m}w_i.d_i$ и $d_1(x,y)=\sqrt{\sum_{i=1}^{m}w_i.d_i}$ също са метрики [13].

Лесно се установява, че $s(x,y)=\dfrac{m-d(x,y)}{m}$ и $s_1(x,y)=\dfrac{m-d_1(x,y)}{m}$ са функции на подобие.

Пример 24. Можем да трансформираме точки, чиито координати имат номинални стойности, в бинарни. Целта се състои във факта, че бинарните данни се поддават на по-лесна и по-задълбочена обработка. Да разгледаме три елемента, които имат по пет координати. Елементите са: x=’праскова’, y=’портокал’ и z=’банан’. Последователно координатите са пет техни характеристики: цвят, форма, вкус, повърхност, структура. Разполагаме със следните данни:

x=(‘оранжев’,’кръгъл’,’сладък’,’мъхеста’,’костилка’);

y=(‘оранжев’,’кръгъл’,’сладък’,’гладка’,’семки’);

z=(‘жълт’,’дълъг’,’сладък’,’гладка’,’сегменти’).

Прилагайки метриката на Хеминг за номинални данни $d(x,y)=\sum_{i=1}^{5}d_i$, получаваме: $d(x,y)=2$, $d(y,z)=3$ и $d(x,z)=4$. Съответно за функцията на подобие $s(x,y)=\dfrac{m-d(x,y)}{m}$ получаваме: $s(x,y)=\dfrac{5-2}{5}=0{,}6$, $s(y,z)=\dfrac{5-3}{5}=0{,}4$ и $s(x,z)=\dfrac{5-4}{5}=0{,}2$.

Можем да бинаризираме горните три елемента, като всеки един от тях има десет координати $x' = (x_1, x_2, ..., x_{10})$, определени по следния начин: $x_1 = 1$ при оранжев цвят и $x_1 = 0$ при друг цвят, $x_2 = 1$ при жълт цвят и $x_2 = 0$ при друг цвят, $x_3 = 1$ при кръгла форма и $x_3 = 0$ при друга форма, $x_4 = 1$ при дълга форма и $x_4 = 0$ при друга форма, $x_5 = 1$ при сладък вкус и $x_5 = 0$ при друг вкус, $x_6 = 1$ при мъхеста повърхност и $x_6 = 0$ при друга повърхност, $x_7 = 1$ при гладка повърхност и $x_7 = 0$ при друга повърхност, $x_8 = 1$ при костилкова структура и $x_8 = 0$ при друга структура, $x_9 = 1$ при семкова структура и $x_9 = 0$ при друга структура, $x_{10} = 1$ при сегментна структура и $x_{10} = 0$ при друга структура. Окончателно получаваме

$x'(1,0,1,0,1,1,0,1,0,0)$;

$y'(1,0,1,0,1,0,1,0,1,0)$;

$z'(0,1,0,1,1,0,1,0,0,1)$.

Използвайки линейна метрика $d(x', y') = \sum_{i=1}^{10} |x_i' - y_i'|$, получаваме: $d(x', y') = 4$, $d(y', z') = 6$ и $d(x', z') = 8$.

При използване на функцията на подобие

$$s(x', y') = \frac{m_{11} + m_{00}}{m_{00} + m_{01} + m_{10} + m_{11}}$$

получаваме $s(x', y') = \frac{6}{10} = 0{,}6$, $s(y, z) = \frac{4}{10} = 0{,}4$ и $s(x', z') = \frac{2}{10} = 0{,}2$.

Сравнявайки получените резултати, виждаме, че те са пропорционални (при двете метрики коефициентът на пропорционалност е 2) или еднакви (при отчитане на подобие).

Пример 25. Нека изследваме четири едноклетъчни организма x, y, z и v, различаващи се по цвят, брой ядра и брой опашки. Първият организъм е светъл, с едно ядро и с една опашка, вторият – светъл, с две ядра и две опашки, третият – тъмен, с две ядра и с две опашки, а четвъртият – тъмен, с три ядра и с една опашка. Следователно всеки един от организмите ще има по три координати. Първата координата е номинална, а следващите две са числови. Това различие, естествено, създава проблеми. Имаме няколко възможности:

(1) Можем да кодираме първата координата, като при светъл цвят имаме код 1, а при тъмен – код 0. Така получаваме само числови координати, което е по-удобно за пресмятане. Окончателно получаваме: $x(1,1,1)$, $y(1,2,2)$, $z(0,2,2)$ и $v(0,3,1)$.

При използване на Евклидова метрика получаваме таблицата с разстоянията между точките:

	x	y	z	v
x	0	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{5}$
y	(2)	0	1	$\sqrt{3}$
z	(3)	(1)	0	$\sqrt{2}$
v	(4)	(3)	(2)	0

Сега нека кодираме първата координата не с 1 и 0, а с 10 и 0. Следователно имаме: $x(10,1,1)$, $y(10,2,2)$, $z(0,2,2)$ и $v(0,3,1)$. Отново при използване на Евклидова метрика получаваме таблицата с разстоянията между точките:

	x	y	z	v
x	0	$\sqrt{2}$	$\sqrt{102}$	$\sqrt{104}$
y	(1)	0	10	$\sqrt{102}$
z	(3)	(2)	0	$\sqrt{2}$
v	(4)	(3)	(1)	0

Сравнявайки резултатите от последните две таблици, виждаме, че резултатите са различни както по стойност, така и по подредба. При първата таблица най-малкото разстояние е $d(y,z)$, а при втората таблица е $d(x,y)$.

На горните координати, които са числа, можем да гледаме като на номинални стойности. В този случай можем да приложим метриката на Хеминг при номинални данни $d(x,y)=\sum_{i=1}^{m} d_i$ и получаваме таблицата с разстоянията между точките:

	x	y	z	v
x	0	2	3	2
y	(2)	0	1	3
z	(3)	(1)	0	2
v	(2)	(3)	(2)	0

Сега получихме резултати, които се различават от предходните два случая.

(2) Гледаме на всички характеристики на организмите като номинални и постъпваме както в *пример 24.* Първа координата 1 при светъл цвят и 0 при друг, втора координата 1 при тъмен цвят и 0 при друг, трета координата 1 при едно ядро и 0 при друг брой, четвърта координата 1 при две ядра и 0 при друг брой, пета координата 1 при три ядра и 0 при друг брой, шеста координата 1 при една опашка и 0 при друг брой, седма координата 1 при 2 опашки и 0 при друг брой. Така получаваме: $x(1,0,1,0,0,1,0)$, $y(1,0,0,1,0,0,1)$, $z(0,1,0,1,0,0,1)$ и $v(0,1,0,0,1,1,0)$. В този случай можем да приложим метриката на Хеминг при бинарни данни $d(x,y)=m_{01}+m_{10}$ и получаваме таблицата с разстоянията между точките:

	x	y	z	v
x	0	4	6	4
y	(2)	0	2	6
z	(3)	(1)	0	4
v	(2)	(3)	(2)	0

Получихме резултати, които са различни, сравнени с последните три случая. Забелязваме, че въпреки различието между третия и четвъртия случай разстоянията са пропорционални, а подредбата на разстоянията е еднаква. Резултатът е аналогичен на резултата от *пример 24.*

10. Определяне на типичен елемент

Да се върнем към изследваното множество M, което се състоеше от $n \geq 3$ елемента, дефинирани с m координати. Нека разгледаме негово непразно подмножество $M_0 \subset M$ с $n_0 \geq 2$ елемента. В частност може да имаме $M_0 = M$. Задачата е да намерим типичен елемент на множеството M_0, който да представлява всички елементи от M_0. Типичността на елемента се определя в основа на максимално подобие на този елемент с елементите от M_0 и минималното му подобие с елементите на $M \setminus M_0$ [1].

Нека s е функция на подобие и $x \in M_0$. Разглеждаме функцията

$$S_I(x) = \frac{1}{n_0} \sum_{y \in M_0} s(x, y),$$

която ни дава средното подобие на елемента x с всички елементи от M_0.

Имаме два случая: $M \setminus M_0 = \varnothing$ и $M \setminus M_0 \neq \varnothing$.

(1) Нека $M \setminus M_0 = \varnothing$, т.е. $M = M_0$.

Търсим такъв елемент $z \in M_0$, че да е в сила

$S_I(z) = \max\{S_I(x) : x \in M_0\}$.

Доказва се, че такъв елемент съществува, но той може да не е единствен. Този или тези елементи се наричат типични за множеството M_0, т.е. имаме $z = typ(M_0)$.

(2) Нека $M \setminus M_0 \neq \varnothing$, разглеждаме функцията

$$S_O(x) = \frac{1}{n - n_0} \sum_{y \in M \setminus M_0} s(x, y),$$

която ни дава средното подобие на елемента x с всички елементи вън от M_0.

В този случай трябва да решим една двукритериална задача:

$S_I(x) \to \max$

$S_O(x) \to \min$.

Да означим решенията на първата задача $S_I(x) \to \max$ с $S_{I,\max}$, а на втората задача $S_O(x) \to \min$ с $S_{O,\min}$.

Имаме две ситуации: $S_{I,\max}$ **I** $S_{O,\min} \neq \varnothing$ и $S_{I,\max}$ **I** $S_{O,\min} = \varnothing$.

(а) Ако имаме идеалната ситуация $S_{I,\max}$ **I** $S_{O,\min} \neq \varnothing$, то всеки елемент от $S_{I,\max}$ **I** $S_{O,\min}$ ще бъде типичен. Така окончателно получаваме $typ(M_0) = S_{I,\max}$ **I** $S_{O,\min}$.

(б) Ако имаме ситуацията $S_{I,\max}$ **I** $S_{O,\min} = \varnothing$, то трябва да търсим решение, оптимално по Парето.

Например можем да разгледаме функцията $S(x) = S_I(x) - S_O(x)$ и да търсим такъв елемент $z \in M_0$, че да е в сила

$S(z) = \max\{S(x) : x \in M_0\}$.

Доказва се, че такъв елемент съществува, но той може да не е единствен. Този или тези елементи се наричат типични за множеството M_0, т.е. имаме $z = typ(M_0)$.

11. Интерпретация на основните статистически величини

Нека отново разгледаме вектора $e(1,1,...,1) \in R^m$ и правата $l_0 = \{a.e \in R^m : a \in R\}$. Ще разгледаме разстоянието или различието на точка $x \in R^m$ до правата l_0, а именно

$d(x, l_0) = \inf\{d(x, y) : y \in l_0\}$.

Имаме, че l_0 е затворено множество. Следователно получаваме, че

$d(x, l_0) = \min\{d(x, y) : y \in l_0\}$

и съществува точка $z \in l_0$ такава, че $d(x, l_0) = d(x, z)$ [11]. Възможно е точката z да не е единствена.

11.1. Стандартно отклонение

Известно е, че $\sqrt{\frac{\sum_{i=1}^m (x_i - C)^2}{m}} \geq \sqrt{\frac{\sum_{i=1}^m (x_i - \bar{x})^2}{m}}$ при $C \in R$, като равенството се достига единствено при $C = \bar{x}$.

Разглеждаме Евклидова метрика d и $h(a) = \frac{1}{\sqrt{m}} a$. Така получаваме нова метрика $d_1 = h \mathbf{o} d$ и окончателно имаме

$$d_1(x, l_0) = \min\{d_1(x, y) : y \in l_0\} = \min\{\sqrt{\frac{\sum_{i=1}^m (x_i - y_i)^2}{m}} : y \in l_0\} = s_x.$$

Точката $\bar{x}.e \in l_0$ е единствената точка, удовлетворяваща $d_1(x, l_0) = d_1(x, \bar{x}.e)$.

Известно е също така, че $s_x^2 = Var(x)$.

11.2. Средно линейно отклонение

Известно е, че $\frac{\sum_{i=1}^m |x_i - C|}{m} \geq \frac{\sum_{i=1}^m |x_i - Me|}{m}$ при $C \in R$.

Разглеждаме линейна метрика d и $h(a) = \frac{1}{m} a$. Така получаваме нова метрика $d_1 = h \mathbf{o} d$ и окончателно имаме

$$d_1(x, l_0) = \min\{d_1(x, y) : y \in l_0\} = \min\{\frac{\sum_{i=1}^m |x_i - y_i|}{m} : y \in l_0\} = d_x.$$

За разлика от математическото очакване $\bar{x}$ в 11.1 медианата Me може да има повече от една стойност. При нечетно m имаме единствена стойност на медианата, но при четно m това не е задължително.

11.3. Дисперсия (вариация, variance)

Известно е, че $\frac{\sum_{i=1}^m (x_i - C)^2}{m} \geq \frac{\sum_{i=1}^m (x_i - \bar{x})^2}{m}$ при $C \in R$, като равенството се достига единствено при $C = \bar{x}$.

Разглеждаме квадрат на Евклидово разстояние (функция на различие) d и $h(a) = \frac{1}{m} a$. Така получаваме нова функция на различие $d_1 = h \mathbf{o} d$ и окончателно имаме

$$d_1(x,l_0)=\min\{d_1(x,y):y\in l_0\}=\min\{\frac{\sum_{i=1}^{m}(x_i-y_i)^2}{m}:y\in l_0\}=Var(x).$$

Точката $\bar{x}.e\in l_0$ е единствената точка, удовлетворяваща $d_1(x,l_0)=d_1(x,\bar{x}.e)$ [11].

12. Теоретико-множествен подход за определяне на различие и подобие

Теоретико-множественият подход за определяне на различие и подобие се основава на знаменитата работа на Тверски [15], която е класика в съвременната теория за изучаване на подобие. Според Тверски подобието между два елемента x и y от изследваната съвкупност се определя от:

(1) Броя на характеристиките, общи и за двата елемента x и y. Да означим това множество с E.

(2) Броя на характеристиките, уникални за елемента x. Да означим това множество с B.

(3) Броя на характеристиките, уникални за елемента y. Да означим това множество с C.

Така лесно получаваме функциите на подобие:

$$s_1(x,y)=\frac{|E|}{|B|+|C|+|E|};$$

$$s_2(x,y)=\frac{2.|E|}{|B|+|C|+2.|E|}.$$

Ясно е, че при бинарни данни функцията на подобие s_1 съвпада с функцията на подобие на Танимото или Джакард, а функцията на подобие s_2 съвпада с функцията на подобие на Дайк.

Можем да въведем още едно множество – D, което е на липсващите характеристики и в двата елемента x и y. Получаваме трета функция на подобие:

$$s_3(x,y)=\frac{|E|}{|B|+|C|+|E|+|D|}.$$

Естествено, от функциите на подобие получаваме функциите на различие:

$$d_1(x,y)=1-s_1(x,y)=1-\frac{|E|}{|B|+|C|+|E|}=\frac{|B|+|C|}{|B|+|C|+|E|};$$

$$d_2(x,y)=1-s_2(x,y)=1-\frac{2.|E|}{|B|+|C|+2.|E|}=\frac{|B|+|C|}{|B|+|C|+2.|E|};$$

$$d_3(x,y)=1-s_3(x,y)=1-\frac{|E|}{|B|+|C|+|E|+|D|}=\frac{|B|+|C|+|D|}{|B|+|C|+|E|+|D|}.$$

Литература

1. **Агре, Г., З. Марков, Д. Дочев.** Увод в машинното самообучение. София, 2001.
2. **Димитров, Б., Н. Янев.** Вероятности и статистика. София, УИ ”Климент Охридски”, 1990.
3. **Калинов, К.** Статистически методи в поведенческите и социалните науки. София, НБУ, 2001.
4. **Кели, Д.** Обща топология. София, Наука и изкуство, 1971.
5. **Манов, А.** Статистика със SPSS. София, Тракия-М, 2001.
6. **Манов, А.** Многомерни статистически методи със SPSS. София, Стопанство, 2002.
7. **Рудин, У.** Основи на математическия анализ. София, Наука и изкуство, 1973.
8. **Стайков, Р., З. Славов.** Статистически методи в икономиката и управлението. Варна, ВСУ, 2008.
9. **Дидэ, Э.** Методы анализа данных. Москва, Финансы и статистика, 1985.
10. **Дубровский, С.** Прикладной многомерный статистческий анализ. Москва, Финансы и статистика, 1982.
11. Прикладной многомерный статистческий анализ. Москва, Наука, 1978.
12. **Ким, Д., Ч. Мьюллер, У. Клекка, М. Олдендерфер, Р. Блэшфилд.** Факторный, дискриминантный и кластерный анализ. Москва, Финансы и статистика, 1989.
13. **Фелингер, А.** Статистические алгоритмы в социологических исследованиях. Новосибирск, Наука, 1985.
14. **Slavov, Z.** Mathematical Ideas in Centroids-Based Partitioning Cluster Analysis, International Conference ISK 2008, Varna 26 – 28 June 2008: 336-345.
15. **Tversky, A.** Features of Similarity. Psychological Review 84 (1977): 327–352.
16. **Young, F., R. Hamer.** Theory and applications of multidimensional scaling. Hillsdale, NJ, Erlbaum, 1994.