ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

IV - pokračování
KLASIFIKACE PODLE MINIMÁLNÍ VZDÁLENOSTI

METRIKY PRO URČENÍ VZDÁLENOSTI MEZI DVĚMA OBRAZY S KVALITATIVNÍMI PŘÍZNAKY

KONTINGENČNÍ MATICE

- vycházejí z pojmu kontingenční matice (tabulka);
- předpokládejme, že hodnoty uvažovaných vektorů patří do konečné k-prvkové množiny F kategoriálních, nebo případně diskrétně kvantitativních hodnot. Dále předpokládejme, že máme vektory $\mathbf{x},\mathbf{y} \in F^n$, kde n je jejich délka a nechť $\mathbf{A}(\mathbf{x},\mathbf{y}) = \{a_{ij}\}$, $i,j \in F$, je matice o rozměru $k \times k$, a její prvky a_{ij} jsou určeny počtem případů, kdy se hodnota i nachází na určité pozici ve vektoru **x** a současně se na téže pozici nachází hodnota j ve vektoru **y**. Matici **A** nazýváme ***kontingenční tabulkou*** (***maticí***). Pokud je kontingenční tabulka rozměru 2 x 2, tj. k = 2, nazýváme ji ***čtyřpolní tabulkou***, slouží ke srovnání dichotomických znaků.

KONTINGENČNÍ MATICE - PŘÍKLAD

- předpokládejme, že množina F obsahuje symboly {0, 1, 2}, tj. k = 3 a vektory **x** a **y** jsou
- $\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$ a
- $\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$, n = 6. Potom kontingenční matice **A**(**x**,**y**) je

$$\mathbf{A}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

- součet hodnot všech prvků matice **A**(**x**,**y**) je roven délce n obou vektorů, tj. v našem případě

$$\sum_{i=0}^{2}\sum_{j=0}^{2} a_{ij} = 6$$

HAMMINGOVA METRIKA

- je definována počtem pozic, v nichž se oba vektory liší

$$\rho_{HQ}(\mathbf{x},\mathbf{y}) = \sum_{i=0}^{k-1} \sum_{\substack{j=0 \\ i \neq j}}^{k-1} a_{ij}$$

- tj. je dána součtem všech prvků matice **A**, které leží mimo hlavní diagonálu.

HAMMINGOVA METRIKA

- pro k = 2, kdy jsou hodnoty obou vektorů binární, se definiční vztah Hammingovy vzdálenosti transformuje na

$$\rho_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}(x_i + y_i - 2x_i y_i)$$

kde třetí člen v závorce kompenzuje případ, kdy jsou hodnoty x_i i y_i rovny jedné a součet prvních členů v závorce je tím pádem roven dvěma, nicméně nastává shoda hodnot, která k celkové vzdálenosti nemůže přispět.

- protože x_i a y_i nabývají hodnot pouze 0 a 1, můžeme také psát

$$\rho_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}(x_i^2 + y_i^2 - 2x_i y_i) = \sum_{i=1}^{n}(x_i - y_i)^2$$

- díky speciálnímu případu hodnot x_i a y_i je možná i nejjednodušší forma

$$\rho_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}\left|x_i - y_i\right|$$

HAMMINGOVA METRIKA

- v případě bipolárních vektorů, kdy jednotlivé složky vektorů nabývají hodnot +1 a -1, je Hammingova vzdálenost určena vztahem

$$\rho_{HQP}(\mathbf{x},\mathbf{y}) = \frac{\left(n - \sum_{i=1}^{n} x_i y_i\right)}{2}$$

HAMMINGOVA METRIKA – PŘÍKLAD 1

Určete Hammingovu vzdálenost vektorů z předchozího příkladu, tj.

$\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$ a

$\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$.

Vzájemným porovnáním obou vektorů lze určit, že oba vektory se liší v první, druhé a páté souřadnici, to znamená, že se oba vektory liší ve třech pozicích, což definuje hodnotu Hammingovy vzdálenosti obou vektorů, tj .

$$d_{HQ}(\mathbf{x},\mathbf{y}) = 3.$$

HAMMINGOVA METRIKA – PŘÍKLAD 1

Určete Hammingovu vzdálenost vektorů z předchozího příkladu, tj.

$\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$ a

$\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$.

$$\mathbf{A}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Z kontingenční matice $\mathbf{A}(\mathbf{x},\mathbf{y})$ je vzdálenost určena součtem všech prvků matice $\mathbf{A}(\mathbf{x},\mathbf{y})$ mimo hlavní diagonálu.

Tedy $d_{HQ}(\mathbf{x},\mathbf{y}) = a_{12} + a_{21} + a_{31} = 1 + 1 + 1 = 3$.

HAMMINGOVA METRIKA – PŘÍKLAD 2

Určete Hammingovu vzdálenost binárních vektorů

$\mathbf{x} = (0, 1, 1, 0, 1)^T$ a

$\mathbf{y} = (1, 0, 0, 0, 1)^T$.

Podle definičního principu je vzdálenost obou vektorů dána počtem pozic, ve kterých se oba vektory liší, tj.

$$d_{HQB}(\mathbf{x},\mathbf{y}) = 3.$$

HAMMINGOVA METRIKA – PŘÍKLAD 2

Určete Hammingovu vzdálenost binárních vektorů

$\mathbf{x} = (0, 1, 1, 0, 1)^T$ a

$\mathbf{y} = (1, 0, 0, 0, 1)^T$.

Použijeme-li vztah

$$d_{HQB}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (x_i + y_i - 2x_i y_i) =$$

$$= (0+1-2.0.1) + (1+0-2.1.0) + (1+0-2.1.0) + (0+0-2.0.0) + (1+1-2.1.1) = 3.$$

HAMMINGOVA METRIKA – PŘÍKLAD 2

Určete Hammingovu vzdálenost binárních vektorů

$\mathbf{x} = (0, 1, 1, 0, 1)^T$ a

$\mathbf{y} = (1, 0, 0, 0, 1)^T$.

Podle vztahu

$$d_{HQB}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (x_i - y_i)^2 =$$

$$= (0-1)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (1-1)^2 = 1+1+1+0+0 = 3.$$

HAMMINGOVA METRIKA – PŘÍKLAD 2

Určete Hammingovu vzdálenost binárních vektorů

$\mathbf{x} = (0, 1, 1, 0, 1)^T$ a

$\mathbf{y} = (1, 0, 0, 0, 1)^T$.

Konečně, podle vztahu

$$d_{HQB}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| =$$

$$= |0-1| + |1-0| + |1-0| + |0-0| + |1-1| =$$

$$= 1+1+1+0+0 = 3.$$

HAMMINGOVA METRIKA – PŘÍKLAD 3

Určete Hammingovu vzdálenost bipolárních vektorů

$\mathbf{x} = (1, 1, 1, -1, 1)^T$ a

$\mathbf{y} = (1, -1, 1, -1, -1)^T$.

Podle definičního principu se oba vektory liší ve dvou pozicích, tj.

$$d_{HQP}(\mathbf{x},\mathbf{y}) = 2.$$

HAMMINGOVA METRIKA – PŘÍKLAD 3

Určete Hammingovu vzdálenost bipolárních vektorů

$\mathbf{x} = (1, 1, 1, -1, 1)^T$ a

$\mathbf{y} = (1, -1, 1, -1, -1)^T$.

Z kontingenční matice, která je pro tento případ rovna

$$\mathbf{A}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} 2 & 2 \\ 0 & 1 \end{bmatrix}$$

je $d_{HQP}(\mathbf{x},\mathbf{y})$ rovna součtu hodnot prvků ležících mimo hlavní diagonálu, tj. $d_{HQP}(\mathbf{x},\mathbf{y}) = 2$.

HAMMINGOVA METRIKA – PŘÍKLAD 3

Určete Hammingovu vzdálenost bipolárních vektorů

$\mathbf{x} = (1, 1, 1, -1, 1)^T$ a

$\mathbf{y} = (1, -1, 1, -1, -1)^T$.

Pomocí vztahu

$$d_{HQP}(\mathbf{x},\mathbf{y}) = \frac{\left(n - \sum_{i=1}^{n} x_i y_i\right)}{2} =$$

$$= \frac{5 - \left((1.1) + (1.(-1)) + (1.1) + ((-1).(-1)) + (1.(-1))\right)}{2} =$$

$$= \frac{5 - (1 - 1 + 1 + 1 - 1)}{2} = \frac{5 - 1}{2} = 2.$$

METRIKY PRO URČENÍ PODOBNOSTI MEZI DVĚMA OBRAZY S KVALITATIVNÍMI PŘÍZNAKY

- ☑ případy obecné
- ☑ případy s dichotomickými příznaky, pro které je definována celá řady tzv. ***asociačních koeficientů***.

  (Asociační koeficienty až na výjimky nabývají hodnot z intervalu ⟨0, 1⟩, hodnoty 1 v případě shody vektorů, 0 pro případ nepodobnosti.)

OBECNÉ METRIKY

HAMMINGOVA METRIKA

pro nedichotomické příznaky

$$\sigma_{HQ}(\mathbf{x},\mathbf{y}) = b_{max} - \rho_{HQ}(\mathbf{x},\mathbf{y}).$$

TANIMOTOVA METRIKA

Předpokládejme, že máme dvě množiny X a Y a n_X, n_Y a $n_{X\cap Y}$ jsou kardinality (počty prvků) množin X, Y a $X \cap Y$. V tom případě je Tanimotova míra podobnosti dvou množin určena podle vztahu

$$\sigma_T(X,Y) = \frac{n_{X\cap Y}}{n_X + n_Y - n_{X\cap Y}} .$$

- jinými slovy, Tanimotova podobnost dvou množin je určena počtem společných prvků obou množin vztaženým k počtu všech rozdílných prvků.

TANIMOTOVA METRIKA

Pro výpočet Tanimotovy podobnosti dvou vektorů s kvalitativními příznaky jsou použity všechny páry složek srovnávaných vektorů, kromě těch, jejichž hodnoty jsou obě nulové.

Definujme pro porovnávané vektory **x** a **y** hodnoty

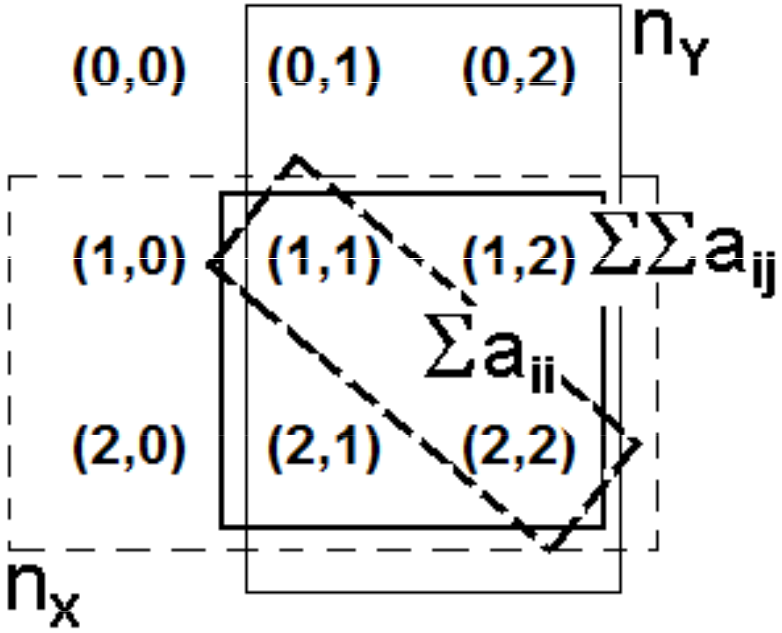
$$n_x = \sum_{i=1}^{k-1}\sum_{j=0}^{k-1} a_{ij} \quad \text{a} \quad n_y = \sum_{i=0}^{k-1}\sum_{j=1}^{k-1} a_{ij}$$

kde k je počet hodnot souřadnic obou vektorů a a_{ij} jsou prvky kontingenční matice **A**(**x**,**y**), tzn. že n_x, resp. n_y udává počet nenulových položek vektoru **x**, resp. **y**.

TANIMOTOVA METRIKA

$$n_x = \sum_{i=1}^{k-1}\sum_{j=0}^{k-1} a_{ij} \qquad n_y = \sum_{i=0}^{k-1}\sum_{j=1}^{k-1} a_{ij}$$

$$\sigma_{TQ}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{n_x + n_y - \sum_{i=1}^{k-1}\sum_{i=1}^{k-1} a_{ij}}$$



TANIMOTOVA METRIKA - PŘÍKLAD

Určete hodnoty Tanimotových podobností $s_{TQ}(\mathbf{x},\mathbf{x})$, $s_{TQ}(\mathbf{x},\mathbf{y})$ a $s_{TQ}(\mathbf{x},\mathbf{z})$, jsou-li vektory

$\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$ a

$\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$ a

$\mathbf{z} = (2, 0, 0, 0, 0, 2)^T$.

Ze zadání je množina symbolů F = {0, 1, 2}, k = 3, n = 6.

Kontingenční tabulky jsou

$$\mathbf{A}(\mathbf{x},\mathbf{x}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}; \mathbf{A}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}; \quad \mathbf{A}(\mathbf{x},\mathbf{z}) = \begin{bmatrix} 0 & 0 & 1 \\ 2 & 0 & 1 \\ 2 & 0 & 0 \end{bmatrix}.$$

TANIMOTOVA METRIKA - PŘÍKLAD

V prvním případě při maximální podobnosti jsou nenulové prvky kontingenční tabulky pouze na hlavní diagonále, v případě nejmenší podobnosti jsou naopak na hlavní diagonále jen nulové prvky.

V případě první podobnosti $s_{TQ}(\mathbf{x},\mathbf{x})$ je $n_x = 5$, $n_Y = 5$, součet prvků na hlavní diagonále Σa_{ii} také 5 a konečně součet $\Sigma\Sigma a_{ij}$ opět 5. Hodnota podobnosti pak po dosazení je

$$s_{TQ}(\mathbf{x},\mathbf{x}) = \frac{5}{5+5-5} = 1.$$

TANIMOTOVA METRIKA - PŘÍKLAD

Pro podobnost $s_{TQ}(\mathbf{x},\mathbf{y})$ je $n_x = 5$, $n_y = 4$, součet prvků na hlavní diagonále $\Sigma a_{ii} = 3$ a konečně součet $\Sigma\Sigma a_{ij} = 3$. Hodnota podobnosti pak po dosazení je

$$s_{TQ}(\mathbf{x},\mathbf{y}) = \frac{3}{5+4-3} = 0{,}5.$$

Konečně, pro podobnost $s_{TQ}(\mathbf{x},\mathbf{z})$, což představuje srovnání dvou nejméně podobných vektorů, je $n_x = 5$, $n_y = 2$, součet prvků na hlavní diagonále $\Sigma a_{ii} = 0$ a konečně součet $\Sigma\Sigma a_{ij} = 1$.

Hodnota podobnosti pak po dosazení je

$$s_{TQ}(\mathbf{x},\mathbf{z}) = \frac{0}{5+2-1} = 0.$$

DALŠÍ OBECNÉ METRIKY

Další míry podobnosti vektorů $\mathbf{x},\mathbf{y} \in F^n$ jsou definovány pomocí různých prvků kontingenční matice $\mathbf{A}(\mathbf{x},\mathbf{y})$. Některé z nich používají pouze počet shodných pozic v obou vektorech (ovšem s nenulovými hodnotami), jiné míry používají i shodu s nulovými hodnotami. Příkladem metriky podobnosti z první uvedené kategorie může být např. metrika definovaná vztahem

$$\sigma_1(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{n}$$

nebo i metrika

$$\sigma_2(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{n - a_{00}}$$

Příkladem metriky druhé uvedené skupiny je např.

$$\sigma_3(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=0}^{k-1} a_{ii}}{n}$$

ASOCIAČNÍ KOEFICIENTY

		$\mathbf{x}_j$	
		false/0	true/1
$\mathbf{x}_i$	false/0	D	C
	true/1	B	A

A. hodnota k-té souřadnice obou vektorů signalizuje, že u obou obrazů sledovaný jev nastal (oba odpovídající si příznaky mají hodnotu true, resp.1) – ***pozitivní shoda***;
B. ve vektoru $\mathbf{x}_i$ jev nastal (x_{ik} = true), zatímco ve vektoru $\mathbf{x}_j$ nikoliv (x_{jk} = false, resp.0);
C. u obrazu $\mathbf{x}_i$ jev nenastal (k-tá souřadnice má hodnotu x_{ik} = false), zatímco u obrazu $\mathbf{x}_j$ ano (x_{jk} = true);
D. sledovaný jev nenastal (oba odpovídající si příznaky mají hodnotu false) – ***negativní shoda***.

Při výpočtu podobnosti dvou vektorů sledujeme kolikrát pro všechny souřadnice obou vektorů $\mathbf{x}_j$ a $\mathbf{x}_j$ nastaly případy shody či neshody – A+D určuje celkový počet shod, B+C celkový počet neshod a A+B+C+D = n, tj. celkový počet souřadnic obou vektorů (obrazů).

JACCARDŮV – TANIMOTŮV ASOCIAČNÍ KOEFICIENT

$$\sigma_{JT}(\mathbf{x},\mathbf{y}) = \frac{A}{A+B+C}$$

což je díky zjednodušení i dichotomická varianta metriky podle vztahu

$$\sigma_{TQ}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{n_x + n_y - \sum_{i=1}^{k-1}\sum_{i=1}^{k-1} a_{ij}}$$

Tento vztah se dominantně používá v ekologických studiích.

RUSSELŮV – RAOŮV ASOCIAČNÍ KOEFICIENT

$$\sigma_{RR}(\mathbf{x},\mathbf{y}) = \frac{A}{A+B+C+D}$$

je to dichotomická varianta metriky podle vztahu

$$\sigma_1(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{n}$$

SOKALŮV – MICHENERŮV ASOCIAČNÍ KOEFICIENT

$$\sigma_{SM}(\mathbf{x},\mathbf{y}) = \frac{A+D}{A+B+C+D}$$

je dichotomická varianta vztahu

$$\sigma_3(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=0}^{k-1} a_{ii}}{n}$$

DICEŮV (CZEKANOWSKÉHO) ASOCIAČNÍ KOEFICIENT

$$\sigma_{DC}(\mathbf{x},\mathbf{y}) = \frac{2A}{2A+B+C} = \frac{2A}{(A+B)+(A+C)}$$

V případě Jaccardova a Diceova koeficientu je třeba vyřešit (pokud jsou používány v situacích, kdy může nastat úplná negativní shoda) jejich hodnotu, když A = B = C =0. Pak zpravidla předpokládáme, že $\sigma_{JT}(\mathbf{x},\mathbf{y}) = \sigma_{DC}(\mathbf{x},\mathbf{y}) = 1$.

ROGERSŮV – TANIMOTŮV ASOCIAČNÍ KOEFICIENT

$$\sigma_{RT}(\mathbf{x},\mathbf{y}) = \frac{A+D}{A+D+2.(B+C)} = \frac{A+D}{(B+C)+(A+B+C+D)}$$

oba posledně uvedené koeficienty zvyšují význam shod v datech – Diceův koeficient zvýšením váhy počtu pozitivních shod v čitateli i jmenovateli, v druhém případě zvýšením váhy počtu neshod ve jmenovateli.

HAMANŮV ASOCIAČNÍ KOEFICIENT

$$\sigma_{HA}(\mathbf{x},\mathbf{y}) = \frac{A + D - (B + C)}{A + B + C + D}$$

nabývá na rozdíl od všech dříve uvedených koeficientů hodnot z intervalu ⟨-1, 1⟩. Hodnoty -1 nabývá, pokud se příznaky pouze neshodují, je roven nule, když je počet shod a neshod v rovnováze a +1 v případě úplné shody všech příznaků.

- Z asociačních koeficientů, které vyjadřují míru podobnosti, lze jednoduše odvodit i míry nepodobnosti (vzdálenosti) pomocí formule

$$\rho_X(\mathbf{x},\mathbf{y}) = 1 - \sigma_X(\mathbf{x},\mathbf{y}).$$

Na základě četností A až D lze pro případ binárních příznaků vytvářet i zajímavé vztahy pro již dříve uvedené míry:

Hammingova metrika $\rho_H(\mathbf{x},\mathbf{y}) = B + C\,;$

Euklidova metrika $\rho_H(\mathbf{x},\mathbf{y}) = \sqrt{B + C}\,;$

Pearsonův korelační koeficient

$$\sigma_{PC}(\mathbf{x},\mathbf{y}) = \frac{A.D - B.C}{\sqrt{(A+B).(C+D).(A+C).(B+D)}}$$

DETERMINISTICKÉ METRIKY PRO URČENÍ VZDÁLENOSTI MEZI DVĚMA MNOŽINAMI OBRAZŮ

PODOBNOST MEZI TŘÍDAMI

- „podobnost“ jednoho obrazu s více obrazy jedné třídy (skupin, množin, shluků);
- „podobnost“ obrazů dvou tříd (skupin, množin, shluků);
- zavedeme funkci, která ke každé dvojici skupin obrazů (C_i, C_j) přiřazuje číslo D(C_i, C_j), které podobně jako míry podobnosti či nepodobnosti (metriky) jednotlivých obrazů musí splňovat minimálně podmínky:

PODOBNOST MEZI TŘÍDAMI

PODMÍNKY

- ☑ (S1) $D(C_i, C_j) \geq 0$
- ☑ (S2) $D(C_i, C_j) = D(C_j, C_i)$
- ☑ (S3) $D(C_i, C_i) = \max_{i,j} D(C_i, C_j)$

(pro míry podobnosti)

- ☑ (S3') $D(C_i, C_i) = 0$ pro všechna i

(pro míry vzdálenosti)

METODA NEJBLIŽŠÍHO SOUSEDA

- je-li d libovolná míra nepodobnosti (vzdálenosti) dvou obrazů a C_i a C_j jsou libovolné skupiny množiny obrazů $\{x_i\}$, i=1,…,K, potom metoda nejbližšího souseda definuje mezi skupinami C_i a C_j vzdálenost

$$D_{NN}(C_i, C_i) = \min_{\substack{x_p \in C_i \\ x_q \in C_j}} d(x_p, x_q)$$

Pozn.:

Při použití této metody se mohou vyskytovat v jednom shluku často i poměrně vzdálené obrazy. Tzn. metoda nejbližšího souseda může generovat shluky protáhlého tvaru.

METODA K NEJBLIŽŠÍCH SOUSEDŮ

Je zobecněním metody nejbližšího souseda.

Je definována vztahem

$$D_{NNk}(C_i, C_i) = \min_{\substack{x_p \in C_i \\ x_q \in C_j}} \sum^{k} d(x_p, x_q),$$

tj. vzdálenost dvou shluků je definována součtem k nejkratších vzdáleností mezi obrazy dvou skupin obrazů.

Pozn.:

Při shlukování metoda částečně potlačuje generování řetězcových struktur.

METODA NEJVZDÁLENĚJŠÍHO SOUSEDA

- opačný princip než nejbližší sousedi

$$D_{FN}(C_i, C_i) = \max_{\substack{x_p \in C_i \\ x_q \in C_j}} d(x_p, x_q)$$

Pozn.:

Generování protáhlých struktur tato metoda potlačuje, naopak vede ke tvorbě nevelkých kompaktních shluků.

- je možné i zobecnění pro více nejbližších sousedů

$$D_{FNk}(C_i, C_i) = \max_{\substack{x_p \in C_i \\ x_q \in C_j}} \sum^{k} d(x_p, x_q),$$

METODA CENTROIDNÍ

- vychází z geometrického modelu v euklidovském n rozměrném prostoru a určuje vzdálenost dvou tříd jako čtverec Euklidovy vzdálenosti těžišť obou tříd.

je-li těžiště třídy definováno jako střední hodnota z obrazů patřících do této třídy, tj.

$$\mathbf{x}_{rk} = \{x_{rk1}, x_{rk2}, \ldots, x_{rkn}\}, \bar{\mathbf{x}}_{ri} = \sum_{k=1}^{K} x_{rik}, i = 1, \ldots, n,$$

pak

$$D_C(C_i, C_i) = \rho_E^2(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j),$$

METODA PRŮMĚRNÉ VAZBY

- vzdálenost dvou tříd C_i a C_j je průměrná vzdálenost mezi všemi obrazy tříd C_i a C_j. Obsahuje-li shluk C_i P obrazů a C_j Q obrazů, pak jejich vzdálenost je definována vztahem

$$D_{GA}(C_i, C_i) = \frac{1}{PQ} \sum_{p=1}^{P} \sum_{q=1}^{Q} d(x_p, x_q).$$

Pozn.:

Metoda často vede k podobným výsledkům jako metoda nejvzdálenějšího souseda.

WARDOVA METODA

- vzdálenost mezi třídami (shluky) je definována přírůstkem součtu čtverců odchylek mezi těžištěm a obrazy shluku vytvořeného z obou uvažovaných shluků C_i a C_j oproti součtu čtverců odchylek mezi obrazy a těžišti v obou shlucích C_i a C_j.

WARDOVA METODA

- jsou-li $\bar{\mathbf{x}}_i$ a $\bar{\mathbf{x}}_j$ těžiště tříd C_i a C_j a $\bar{\mathbf{x}}$ těžiště sjednocené množiny, pak Wardova vzdálenost obou shluků je definována výrazem

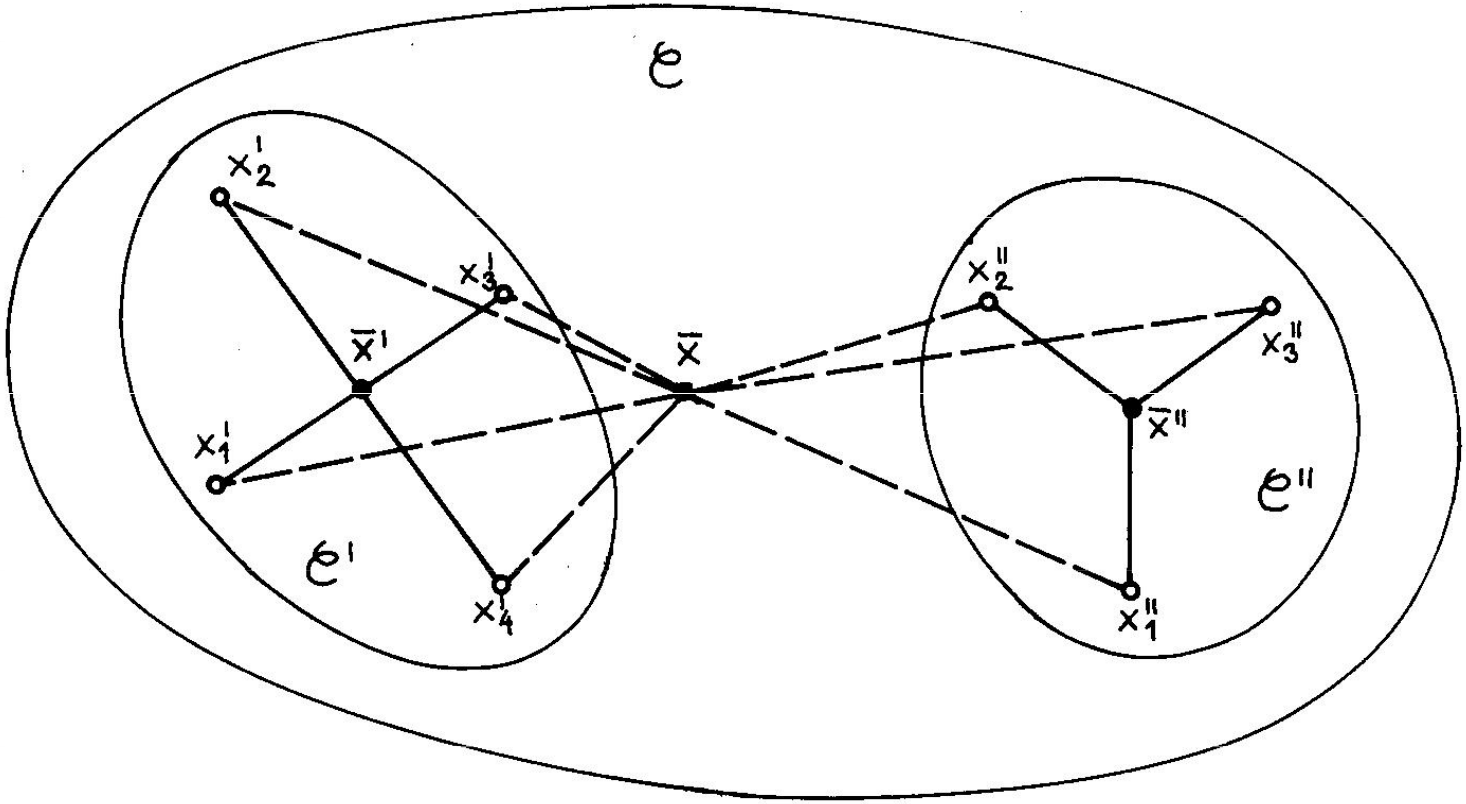
$$D_W(C_i, C_i) = \sum_{x_i \in C_i \cup C_j} \sum_{k=1}^{n} (x_{ik} - \bar{x})^2 -$$

$$- \left(\sum_{x_i \in C_i} \sum_{k=1}^{n} (x_{ik} - \bar{x}_{ik})^2 + \sum_{x_i \in C_j} \sum_{k=1}^{n} (x_{ik} - \bar{x}_{ik})^2 \right).$$

Pozn.:

Metoda má tendenci vytvářet shluky zhruba stejné velikosti, tedy odstraňovat shluky malé, resp. velké.

WARDOVA METODA



METRIKY PRO URČENÍ VZDÁLENOSTI MEZI DVĚMA MNOŽINAMI OBRAZŮ POUŽÍVAJÍCÍ JEJICH PRAVDĚPODOBNOSTNÍ CHARAKTERISTIKY

NA ÚVOD

Klasifikační třídy (množiny obrazů se společnými charakteristikami) nemusí být definovány jen výčtem obrazů, nýbrž vymezením obecnějších vlastností - definicí hranic oddělujících část obrazového prostoru, která náleží dané klasifikační třídě, diskriminační funkcí, pravděpodobnostními charakteristikami výskytu obrazů v dané třídě, atd.

NA ÚVOD

Pokud si na metriky klademe určité požadavky, i metriky pro stanovení vzdálenosti dvou množin, pro něž využíváme rozložení pravděpodobnosti výskytu obrazů, by měly vyhovovat standardním požadavkům. Logicky tyto metriky splňují následující vlastnosti (protože jejich výpočet je založen na poněkud jiném přístupu a protože i dále uvedené vlastnosti nesplňují vše, co od metrik očekáváme, bývá zvykem je značit jiným písmenem, zpravidla J):

1. J = 0, pokud jsou hustoty pravděpodobnosti obou množin identické, tj. když $p(\mathbf{x}|\omega_1) = p(\mathbf{x}|\omega_2)$;

2. $J \geq 0$;

3. J nabývá maxima, pokud jsou obě množiny disjunktní, tj. když

$$\int_{-\infty}^{\infty} p(\mathbf{x}|\omega_1).p(\mathbf{x}|\omega_2)d\mathbf{x} = 0$$

(Jak vidíme, není mezi vlastnostmi pravděpodobnostních metrik uvedena trojúhelníková nerovnost, jejíž splnění by se zajišťovalo vskutku jen velmi obtížně.)

NA ÚVOD

- Základní myšlenkou, na které jsou pravděpodobnostní metriky založeny, je využití pravděpodobnosti způsobené chyby. Čím více se hustoty pravděpodobnosti výskytu obrazů **x** v jednotlivých množinách překrývají, tím je větší pravděpodobnost chyby.

NA ÚVOD

Pravděpodobnost P_e chybného zařazení je (VIZ Bayesův klasifikátor) rovna

$$P_e = J(\mathbf{a}^*) = \min_{\forall \mathbf{a}} J(\mathbf{a}) = \int_X \min_{\forall r} L_X(\mathbf{a}) d\mathbf{x} = \int_X \left[p(\mathbf{x}) - p(\mathbf{x}|\omega_r).P(\omega_r)\right] d\mathbf{x} =$$

$$= \int_X p(\mathbf{x}) d\mathbf{x} - \int_X \max_{\forall r} p(\mathbf{x}|\omega_r).P(\omega_r) d\mathbf{x} = 1 - \int_X \max_{\forall r} p(\mathbf{x}|\omega_r).P(\omega_r) d\mathbf{x} .$$

Pro dichotomický případ (R = 2) je celková pravděpodobnost chybného rozhodnutí určena vztahem

$$P_e = 1 - \int_X \left| p(\mathbf{x}|\omega_1).P(\omega_1) - p(\mathbf{x}|\omega_2).P(\omega_2) \right| d\mathbf{x},$$

což lze podle Bayesova vzorce upravit i do tvaru

$$P_e = 1 - \int_X \left| P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x}) \right| .p(\mathbf{x}).d\mathbf{x}.$$

Kolmogorovova variační vzdálenost

NA ÚVOD

Hodnota Kolmogorovovy variační vzdálenosti přímo souvisí s pravděpodobností chybného rozhodnutí. Ostatní dále uvedené pravděpodobnostní vzdálenosti odvozené z obecné formule

$$J(\mathbf{x}) = \int f\left[p(\mathbf{x}|\omega_i), P(\omega_i), i = 1,2\right] d\mathbf{x}$$

už tuto přímou souvislost nemají.

PRAVDĚPODOBNOSTNÍ METRIKY

- ***Chernoffova metrika***

$$J_C(\omega_1,\omega_2) = -\ln \int p^s(\mathbf{x}|\omega_1).p^{1-s}(\mathbf{x}|\omega_2).d\mathbf{x},\ s \in \langle 0;1\rangle;$$

- ***Bhattacharyyova metrika***

$$J_B(\omega_1,\omega_2) = -\ln \int [p(\mathbf{x}|\omega_1).p(\mathbf{x}|\omega_2)]^{0,5}.d\mathbf{x}.$$

(Jak lze snadno rozpoznat, Bhattacharyyova metrika je speciální případ Chernoffovy metriky pro s = 0,5).

- ***Divergence***

$$J_D(\omega_1,\omega_2) = \int [p(\mathbf{x}|\omega_1) - p(\mathbf{x}|\omega_2)].\ln\left(\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)}\right).d\mathbf{x}\ ;$$

- ***Patrickova -Fisherova metrika***

$$J_{PF}(\omega_1,\omega_2) = \left\{\int [p(\mathbf{x}|\omega_1) - p(\mathbf{x}|\omega_2)]^2.d\mathbf{x}\right\}^{0,5}.$$

ZPRŮMĚRNĚNÉ PRAVDĚPODOBNOSTNÍ METRIKY

- ***zprůměrněná Chernoffova metrika***

$$J_{AC}(\omega_1,\omega_2) = -\ln \int [p(\mathbf{x}|\omega_1).P(\omega_1)]^s.[p(\mathbf{x}|\omega_2).P(\omega_2)]^{1-s1}.d\mathbf{x},..\ s \in \langle 0;1\rangle;$$

- ***zprůměrněná Bhattacharyyova metrika***

$$J_{AB}(\omega_1,\omega_2) = -\ln \int [p(\mathbf{x}|\omega_1).P(\omega_1).p(\mathbf{x}|\omega_2).P(\omega_2)]^{0,5}.d\mathbf{x};$$

- ***zprůměrněná divergence***

$$J_{AD}(\omega_1,\omega_2) = \int [p(\mathbf{x}|\omega_1).P(\omega_1) - p(\mathbf{x}|\omega_2).P(\omega_2)].\ln\left(\frac{p(\mathbf{x}|\omega_1).P(\omega_1)}{p(\mathbf{x}|\omega_2).P(\omega_2)}\right).d\mathbf{x};$$

- ***zprůměrněná Patrickova -Fisherova metrika***

$$J_{PF}(\omega_1,\omega_2) = \left\{\int [p(\mathbf{x}|\omega_1).P(\omega_1) - p(\mathbf{x}|\omega_2).P(\omega_2)]^2.d\mathbf{x}\right\}^{0,5}.$$

Příprava nových učebních materiálů
oboru Matematická biologie

je podporována projektem ESF

č. CZ.1.07/2.2.00/28.0043

„INTERDISCIPLINÁRNÍ ROZVOJ STUDIJNÍHO OBORU MATEMATICKÁ BIOLOGIE“

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ