ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

V.
KLASIFIKACE PODLE MINIMÁLNÍ VZDÁLENOSTI

PRINCIPY KLASIFIKACE

- pomocí diskriminačních funkcí – funkcí, které určují míru příslušnosti k dané klasifikační třídě;
- pomocí definice hranic mezi jednotlivými třídami a logických pravidel;
- pomocí vzdálenosti od reprezentativních obrazů (etalonů) klasifikačních tříd;
- pomocí ztotožnění s etalony;

PRINCIPY KLASIFIKACE

- pomocí diskriminačních funkcí – funkcí, které určují míru příslušnosti k dané klasifikační třídě;
- pomocí definice hranic mezi jednotlivými třídami a logických pravidel;
- pomocí vzdálenosti od reprezentativních obrazů (etalonů) klasifikačních tříd;
- pomocí ztotožnění s etalony;

METRIKA - VZDÁLENOST

Metrika ρ na X je funkce $\rho: X \times X \rightarrow R$, kde R je množina reálných čísel, taková, že:

$$\exists \rho_0 \in R: -\infty < \rho_0 \leq \rho(x,y) < +\infty, \forall x,y \in X$$

$$\rho(x,x) = \rho_0, \forall x \in X$$

a

$$\rho(x,y) = \rho(y,x), \forall x,y \in X. \textit{(symetrie)}$$

Když dále

$$\rho(x, y) = \rho_0 \text{ když a jen když } x = y \textit{ (totožnost)}$$

a $$\rho(x, z) \leq \rho(x, y) + \rho(y, z), \forall x,y,z \in X. \textit{(}\Delta\textit{ nerovnost)}$$

Prostor X, ve kterém metrika ρ definována, nazýváme metrickým prostorem.

Vzdálenost je hodnota určená podle metriky.

METRIKA PODOBNOSTI - PODOBNOST

Metrická míra podobnosti s na X je funkce $s: X \times X \rightarrow R$, kde R je množina reálných čísel, taková, že:

$$\exists s_0 \in R: -\infty < s(x,y) \leq s_0 < +\infty, \forall x,y \in X$$

$$s(x,x) = s_0, \forall x \in X$$

a

$$s(x,y) = s(y,x), \forall x,y \in X. \textit{(symetrie)}$$

Když dále

$$s(x,y) = s_0 \text{ když a jen když } x = y \textit{ (totožnost)}$$

a $$s(x,y).s(y,z) \leq [s(x,y) + s(y,z)].s(x,z), \forall x,y,z \in X.$$

MÍRY PODOBNOSTI VS. NEPODOBNOSTI

Vzdálenostní míry (míry nepodobnosti) mohou být transformovány na podobnostní míry různými transformacemi, např.

$$s_{ij} = 1/\rho_{ij}$$

$$s_{ij} = 1/(1+ \rho_{ij})$$

$$s_{ij} = c - \rho_{ij},\ c \geq \max \rho_{ij},\ \forall i,j$$

MÍRY PODOBNOSTI VS. NEPODOBNOSTI

Vzdálenostní míry (míry nepodobnosti) mohou být transformovány na podobnostní míry různými transformacemi, např.

$$s_{ij} = 1/\rho_{ij}$$

$$s(x,y).s(y,z) \leq [s(x,y) + s(y,z)].s(x,z), \forall x,y,z \in X$$

$$s_{ij} = 1/(1+ \rho_{ij})$$

$$s_{ij} = c - \rho_{ij}, c \geq \max \rho_{ij}, \forall i,j$$

MÍRY PODOBNOSTI VS. NEPODOBNOSTI

Vzdálenostní míry (míry nepodobnosti) mohou být transformovány na podobnostní míry různými transformacemi, např.

$$s_{ij} = 1/\rho_{ij}$$

$$s(x,y).s(y,z) \leq [s(x,y) + s(y,z)].s(x,z), \forall x,y,z \in X$$

$$s_{ij} = 1/(1+ \rho_{ij})$$

$$s_{ij} = c - \rho_{ij},\ c \geq \max \rho_{ij},\ \forall i,j$$

MÍRY PODOBNOSTI VS. NEPODOBNOSTI

Vzdálenostní míry (míry nepodobnosti) mohou být transformovány na podobnostní míry různými transformacemi, např.

$$s_{ij} = 1/\rho_{ij}$$

$$s(x,y).s(y,z) \leq [s(x,y) + s(y,z)].s(x,z), \forall x,y,z \in X$$

$$s_{ij} = 1/(1+ \rho_{ij})$$

$$s(x,y).s(y,z) \leq [s(x,y) + s(y,z) - s(x,y).s(y,z)].s(x,z), \forall x,y,z \in X$$

$$s_{ij} = c - \rho_{ij},\ c \geq \max \rho_{ij},\ \forall i,j$$

MÍRY PODOBNOSTI VS. NEPODOBNOSTI

Vzdálenostní míry (míry nepodobnosti) mohou být transformovány na podobnostní míry různými transformacemi, např.

$$s_{ij} = 1/\rho_{ij}$$

$$s(x,y).s(y,z) \leq [s(x,y) + s(y,z)].s(x,z), \forall x,y,z \in X$$

$$s_{ij} = 1/(1+ \rho_{ij})$$

$$s(x,y).s(y,z) \leq [s(x,y) + s(y,z) - s(x,y).s(y,z)].s(x,z), \forall x,y,z \in X$$

$$s_{ij} = c - \rho_{ij},\ c \geq \max \rho_{ij}, \forall i,j$$

$$s(x,z) \geq s(x,y) + s(y,z) - c$$

TYPY MĚR VZDÁLENOSTI (PODOBNOSTI)

- dle typu příznaků (numerické hodnoty, nominální či ordinální hodnoty, binární hodnoty);
- dle objektů, jejichž vztah hodnotíme – obrazy (vektory), množiny obrazů (vektorů), rozdělení
- deterministické (nepravděpodobnostní) vs. pravděpodobnostní míry

MÍRY VZDÁLENOSTI

obecné poznámky:

- ☑ výběr konkrétní metriky závisí na použití

  kritéria:

  - ➔ optimální výsledky (klasifikační chyby, ztráta, …)
  - ➔ výpočetní nároky
  - ➔ charakter rozložení dat

- ☑ obecně nelze doporučit vhodnou metriku pro určité standardní situace

NUMERICKÉ PŘÍZNAKY

EUKLIDOVA METRIKA

metrika zřejmě s nejnázornější geometrickou interpretaci

$$\rho_E(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{i=1}^{n} (x_{1i} - x_{2i})^2 \right]^{1/2}$$

- ☑ geometrickým místem bodů s toutéž Euklidovou vzdáleností od daného bodu je hyperkoule (kruh ve dvourozměrném prostoru);
- ☑ dává větší důraz na větší rozdíly mezi souřadnicemi (žádoucí nebo nežádoucí? – volba i podle toho, jak chceme zdůrazňovat rozdíly mezi jednotlivými souřadnicemi)
- ☑ čtverec euklidovské vzdálenosti (lépe se počítá) je stále mírou nepodobnosti, ale není metrikou

EUKLIDOVA METRIKA

metrika zřejmě s nejnázornější geometrickou interpretaci

$$\rho_E(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{i=1}^{n} (x_{1i} - x_{2i})^2 \right]^{1/2}$$

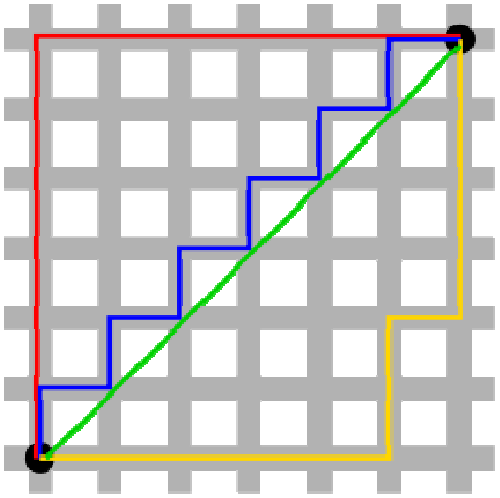
- **Sokalova metrika**

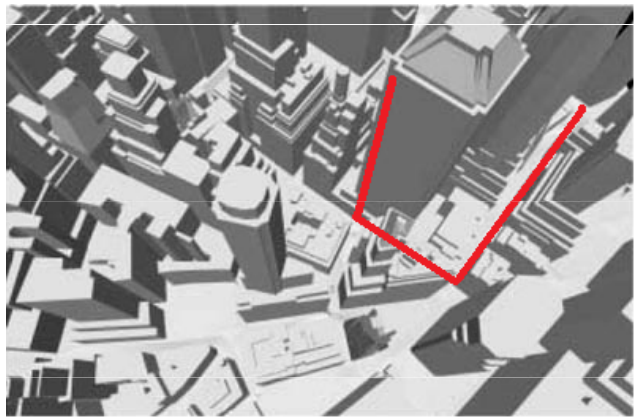
$$\rho_S(\mathbf{x}_1, \mathbf{x}_2) = \left\{ \frac{\rho_E^2(\mathbf{x}_1, \mathbf{x}_2)}{n} \right\}^{1/2}$$

HAMMINGOVA METRIKA

(metrika Manhattan, city-block m., taxi driver m.)

$$\rho_H(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{n} \left| x_{1i} - x_{2i} \right|$$





HAMMINGOVA METRIKA

(metrika Manhattan, city-block m., taxi driver m.)

$$\rho_H(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{n} \left| x_{1i} - x_{2i} \right|$$

- ☑ geometrickým místem bodů ve dvou rozměrném prostoru je kosočtverec;
- ☑ nižší výpočetní nároky než E.m. $\Rightarrow$ použití v úlohách s vysokou výpočetní pracností

MINKOVSKÉHO METRIKA

$$\rho_M(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{i=1}^{n} \left| x_{1i} - x_{2i} \right|^m \right]^{1/m}$$

- zobecnění Euklidovy a Hammingovy metriky;
- volba m záleží na míře důrazu – čím větší m, tím větší váha na velké rozdíly mezi příznaky,

  pro $m \to \infty$ metrika konverguje k Čebyševově metrice

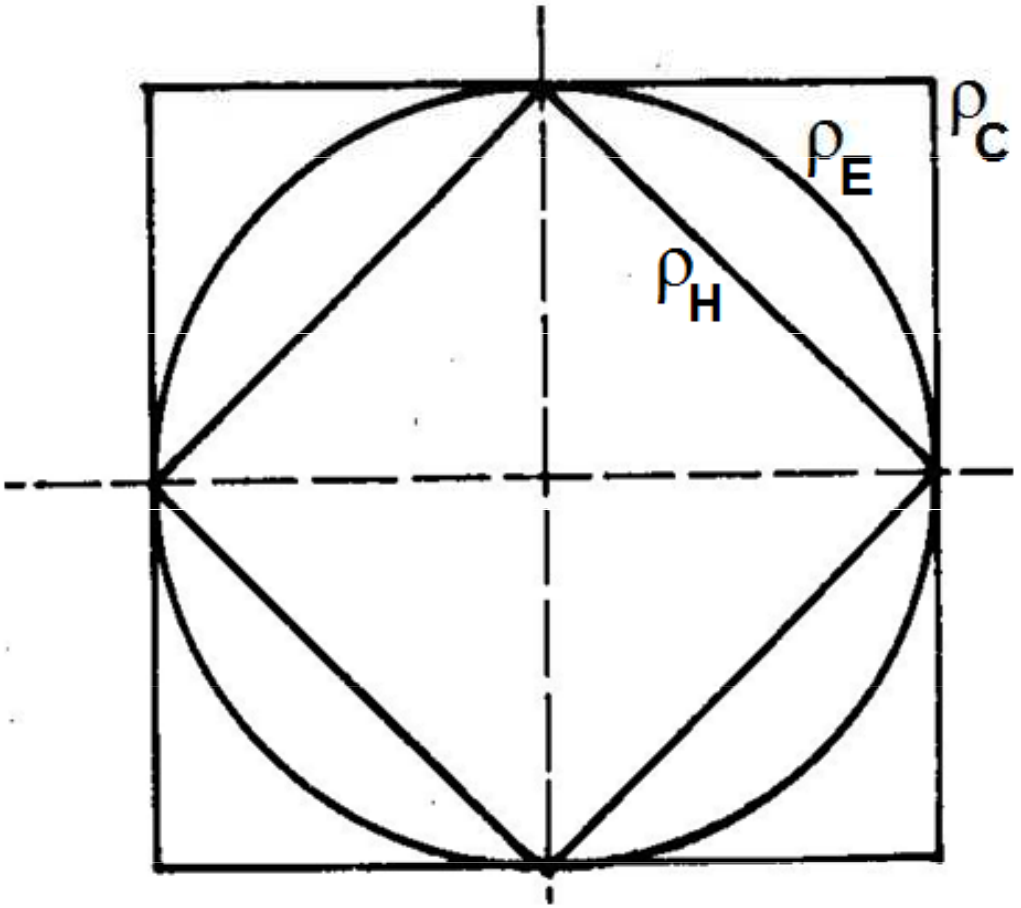
$$\rho_C(\mathbf{x}_1, \mathbf{x}_2) = \lim_{m \to \infty} \rho_M(\mathbf{x}_1, \mathbf{x}_2)$$

ČEBYŠEVOVA METRIKA

$$\rho_C(\mathbf{x}_1, \mathbf{x}_2) = \max_{\forall i} \left\{ \left| x_{1i} - x_{2i} \right| \right\}$$

- ☑ používá se ve výpočetně kriticky náročných případech, kdy je pracnost výpočtu dle euklidovsky orientovaných metrik nepřijatelná;
- ☑ geometrickým místem bodů s toutéž Čebyševovou vzdáleností od daného bodu je hyperkrychle (čtverec ve dvourozměrném prostoru)

SROVNÁNÍ GEOMETRICKÝCH MÍST



ČEBYŠEVOVA METRIKA

- pokud je třeba použít „euklidovskou“ metriku, ale s nižší výpočetní pracností, používá se v první řadě Hammingova nebo Čebyševova metrika;
- lepším přiblížením je kombinace obou metrik

$$\rho_A(\mathbf{x}_1, \mathbf{x}_2) = \max(2\rho_H / 3; \rho_C)$$

(ve dvourozměrném prostoru tvoří geometrické místo bodů o téže vzdálenosti osmiúhelník)

KVADRATICKÁ VZDÁLENOST

$$\rho_Q^2(\mathbf{x}_1,\mathbf{x}_2)=(\mathbf{x}_1-\mathbf{x}_2)^T.\mathbf{Q}.(\mathbf{x}_1-\mathbf{x}_2)$$

- vhodný výběr matice Q je inverzní matice kovariance uvnitř množiny obrazů;
- pak se to jmenuje Mahalanobisova metrika

$$\rho_{MA}(\mathbf{x}_1,\mathbf{x}_2)=\left[(\mathbf{x}_1-\mathbf{x}_2)^T.\mathbf{K}^{-1}.(\mathbf{x}_1-\mathbf{x}_2)\right]^{1/2}$$

METRIKA CANBERRA

$$\rho_{CA}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{n} \frac{|x_{1i} - x_{2i}|}{x_{1i} + x_{2i}}$$

- je vhodná pro proměnné s nezápornými hodnotami
  - pokud jsou obě hodnoty x_{1i} a x_{2i} nulové, potom předpokládáme, že hodnota zlomku je nulová;
  - je-li jenom jedna hodnta nulová, pak je zlomek roven jedné, nezávile na velikosti druhé hodnoty;
  - někdy se nulové hodnoty nahrazují malým kladným číslem (menším, než nejmenší naměřené hodnoty);

NELINEÁRNÍ VZDÁLENOST

$$\rho_N(\mathbf{x}_1,\mathbf{x}_2)=\begin{cases}0 & \text{když } \rho_E(\mathbf{x}_1,\mathbf{x}_2)<D\\ H & \text{když } \rho_E(\mathbf{x}_1,\mathbf{x}_2)\geq D\end{cases}$$

kde D je prahová hodnota a H je nějaká konstanta. Uvádí se, že dobrý výběr hodnot H a D by měl splňovat vztah

$$H=\frac{\Gamma(n/2)}{D^n\sqrt{\pi^n}}$$

když D splňuje nestrannost a konzistenční podmínku Parzenova odhadu, především $D^nN\to\infty$ a $D\to 0$, když $N\to\infty$ (N je počet obrazů v množině)

ÚHLOVÁ VZDÁLENOST

$$\frac{\sum_{i=1}^{n} x_{1i}x_{2i}}{\sum_{i=1}^{n} x_{1i}^2 \sum_{i=1}^{n} x_{2i}^2}$$

- Úhlová vzdálenost (je to spíš míra podobnosti, než nepodobnosti) určuje úhel mezi jednotkovými vektory, které mají směr obou zkoumaných vektorů.
- Vhodná v případě, pokud je informativní pouze relativní hodnota příznaků.

ÚHLOVÁ VZDÁLENOST

$$\frac{\sum_{i=1}^{n} x_{1i}x_{2i}}{\sum_{i=1}^{n} x_{1i}^2 \sum_{i=1}^{n} x_{2i}^2}$$

co takhle korelační koeficient ?

- Úhlová vzdálenost (je to spíš míra podobnosti, než nepodobnosti) určuje úhel mezi jednotkovými vektory, které mají směr obou zkoumaných vektorů.
- Vhodná v případě, pokud je informativní pouze relativní hodnota příznaků.

NEPRAVDĚPODOBNOSTNÍ METRIKY

nevýhody:

- ☑ fyzikální nesmyslnost vytvářet kombinaci veličin s různým fyzikálním rozměrem
- ☑ jsou-li příznakové veličiny zahrnovány do výsledné vzdálenosti se stejnými vahami, zvyšuje se vliv korelovaných veličin

NEPRAVDĚPODOBNOSTNÍ METRIKY

možné odstranění (potlačení) nevýhod:

vztažením k nějakému vyrovnávacímu faktoru, např. střední hodnotě, směrodatné odchylce, normě daného obrazu $x=(x_1, x_2, \ldots, x_n)$

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2},$$

rozpětí

$$\Delta_i = \max_j x_{ij} - \min_j x_{ij},$$

resp. standardizací podle vztahu

$$u_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, \quad i = 1,\ldots,n;\ j = 1,\ldots,K$$

NEPRAVDĚPODOBNOSTNÍ METRIKY

možné odstranění (potlačení) nevýhod:

- lze i subjektivně či na základě nějaké apriorní informace o úloze přiřadit každé příznakové proměnné váhový koeficient, např. váhovaná Minkovského metrika má tvar

$$\rho_{WM}(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{i=1}^{n} a_i . \left| x_{1i} - x_{2i} \right|^m \right]^{1/m}$$

NEPRAVDĚPODOBNOSTNÍ METRIKY

možné odstranění (potlačení) nevýhod:

- ☑ váhování příznaků lze zapsat maticově

$$\mathbf{u}_i = {}^{T}\mathbf{C}.\mathbf{x}_i,$$

kde prvky transformační matice **C** jsou definovány jako

$$c_{ii} = a_i, \text{ pro } i = 1, \ldots, n$$

$$c_{ij} = 0, \text{ pro } i \neq j$$

Za tohoto formalismu je Euklidova metrika definována vztahem

$$\rho_{EW}(\mathbf{x}_1, \mathbf{x}_2) = \left[(\mathbf{x}_1 - \mathbf{x}_2)^T.\mathbf{C}.\mathbf{C}^T.(\mathbf{x}_1 - \mathbf{x}_2)\right]^{1/2}$$

NEPRAVDĚPODOBNOSTNÍ METRIKY

možné odstranění (potlačení) nevýhod:

- pokud jsou složky transformovaného obrazu dány lineární kombinací více složek původního obrazu, není ani matice **C**, ani matice $\mathbf{C}.^{T}\mathbf{C}$ čistě diagonální. Použijeme-li místo matice $\mathbf{C}.^{T}\mathbf{C}$ inverzní kovarianční matice $\mathbf{K}^{-1}$, pak definiční vztah pro váhovanou Euklidovu metriku je definičním vztahem pro **Mahalanobisovu metriku**

$$\rho_E(\mathbf{u}_1,\mathbf{u}_2) = \rho_{MA}(\mathbf{x}_1,\mathbf{x}_2) = \left[(\mathbf{x}_1-\mathbf{x}_2)^T.\mathbf{K}^{-1}.(\mathbf{x}_1-\mathbf{x}_2)\right]^{1/2}$$

NEPRAVDĚPODOBNOSTNÍ METRIKY

možné odstranění (potlačení) nevýhod:

- pokud jsou složky transformovaného obrazu dány lineární kombinací více složek původního obrazu, není ani matice **C**, ani matice $\mathbf{C}.^{T}\mathbf{C}$ čistě diagonální. Použijeme-li místo matice $\mathbf{C}.^{T}\mathbf{C}$ inverzní kovarianční matice $\mathbf{K}^{-1}$, pak definiční vztah pro váhovanou Euklidovu metriku je definičním vztahem pro **Mahalanobisovu metriku**

$$\rho_E(\mathbf{u}_1,\mathbf{u}_2) = \rho_{MA}(\mathbf{x}_1,\mathbf{x}_2) = \left[(\mathbf{x}_1-\mathbf{x}_2)^T.\mathbf{K}^{-1}.(\mathbf{x}_1-\mathbf{x}_2)\right]^{1/2}$$

- **Kovarianční matice** dvou (náhodných) vektorů $\mathbf{x}=^{T}(x_1,\ldots,x_m)$ a $\mathbf{y}=^{T}(y_1,\ldots,y_n)$ je dána vztahem
  - $\mathbf{K}(\mathbf{x},\mathbf{y})=E((\mathbf{x}-E\mathbf{x}).(\mathbf{y}-E\mathbf{y})^T) = [\mathrm{cov}(x_i,y_j)]_{m,n}$

NOMINÁLNÍ A ORDINÁLNÍ PROMĚNNÉ

NOMINÁLNÍ A ORDINÁLNÍ PROMĚNNÉ

- **Nominální** proměnná je taková, o jejíž dvou hodnotách můžeme pouze říci, zda jsou stejné či různé (škola, fakulta, obor). Hodnotami mohou být texty (písmena), případně i číselné kódy. Lze u nich zjišťovat jen rozdělení četností, nemůžeme provádět aritmetické operace (sčítat apod.), výjimkou jsou binární proměnné (viz dále).
- **Ordinální** (**pořadová**), u jejíž dvou hodnot můžeme navíc určit pořadí (úroveň spokojenosti, vzdělání). Jako hodnoty lze použít text, datum, číslo. Pro statistické analýzy (s výjimkou zjišťování četností) je třeba texty převést na čísla. S typem datum lze provádět jen některé výpočty, a to pouze v některých programových systémech.

NOMINÁLNÍ A ORDINÁLNÍ PROMĚNNÉ

Nominální proměnné jsou často reprezentovány binárně kódem jedna z m. Vzdálenost mezi takovými obrazy je dána součtem příspěvků od jednotlivých proměnných

V případě ordinálních proměnných vzdálenost mezi dvěma vektory nezávisí jednoduše na hodnotách proměnných.

Pokud proměnná v jednom vektoru nabývá hodnoty m a v druhém hodnoty k (m<k), pak požadujeme, aby

$$\delta_{mk} \geq \delta_{ms} \text{ pro } s<k$$

$$\delta_{mk} \geq \delta_{sk} \text{ pro } s>m$$

Hodnota δ_{mk} je velice závislá na řešeném problému.

Př. rostlina s krátkými, dlouhými a velmi dlouhými plody. Samozřejmě chceme, aby vzdálenost mezi velmi dlouhým a krátkým plodem byla větší než mezi dlohým a krátkým plodem. To splňuje kódování 1,2,3, ale také 1,10,100.

BINÁRNÍ PROMĚNNÉ

BINÁRNÍ PROMĚNNÉ

KOEFICIENTY ASOCIACE

KOEFICIENTY ASOCIACE

- **Koeficienty asociace** jsou míry podobnosti mezi obrazy obsahujícími logické (binární, dichotomické) příznakové veličiny.
- Ke zjištění podobnosti je třeba sledovat shodu či neshodu hodnot odpovídajících si příznaků ⇒ čtyři možné situace

KOEFICIENTY ASOCIACE

A. u obou obrazů sledovaný jev nastal (oba odpovídající si příznaky mají hodnotu true) – **pozitivní shoda**;

B. u obrazu $\mathbf{x}_i$ jev nastal (x_{ik} = true), zatímco u obrazu $\mathbf{x}_j$ nikoliv (x_{jk} = false);

C. u obrazu $\mathbf{x}_i$ jev nenastal (x_{ik} = false), zatímco u obrazu $\mathbf{x}_j$ ano (x_{jk} = true);

D. u obou obrazů sledovaný jev nenastal (oba odpovídající si příznaky mají hodnotu false) – **negativní shoda**;

KOEFICIENTY ASOCIACE

		x_j	
		true	false
x_i	true	A	B
	false	C	D

KOEFICIENTY ASOCIACE

- sledujeme, kolikrát pro všechny příznaky obrazů x_i a x_j nastaly případy shody a neshody
  - A+D celkový počet shod příznaků;
  - B+C celkový počet neshod příznaků;
  - A+B+C+D =n tj. počet příznaků obou obrazů

Na základě počtu zjištěných shod a neshod jsou definovány různé koeficienty asociace.

Koeficienty asociace jsou míry podobnosti

KOEFICIENTY ASOCIACE

- **Sokalův- Michenerův koeficient** (koeficient jednoduché vazby)

$$s_{SM}(\mathbf{x}_i, \mathbf{x}_j) = \frac{A + D}{A + B + C + D}$$

Problém je se hodnotou D – společná absence jevu – problém „double zero“.

To, že někde něco není často nevede k větší podobnosti (ekologie), nebo naopak společná absence se špatně určuje (detekce určitých prvků v signálu).

KOEFICIENTY ASOCIACE

☑ Jaccardův (Tanimotův) koeficient

$$s_J(\mathbf{x}_i, \mathbf{x}_j) = \frac{A}{A + B + C}$$

- ➔ vůbec neobsahuje člen D – masivní využití v ekologii
- ➔ není definován pro dvojice obrazů, které vykazují negativní shodu ve všech příznacích;

KOEFICIENTY ASOCIACE

Diceův koeficient (Czekanowského)

$$s_D(\mathbf{x}_i, \mathbf{x}_j) = \frac{2A}{2A + B + C} = \frac{2A}{(A + B) + (A + C)}$$

v podstatě totéž jako Jaccardův koeficient, pouze koincidence má dvojnásobnou váhu

KOEFICIENTY ASOCIACE

Russelův- Raoův koeficient

$$s_{RR}(\mathbf{x}_i, \mathbf{x}_j) = \frac{A}{A + B + C + D}$$

Asociační koeficienty zpravidla nabývají hodnot z intervalu $\langle 0,1 \rangle$. V případě R-R koeficientu je při srovnání dvou týchž obrazů hodnota $s_{RR} = 1$ pouze když došlo u všech příznaků jen k pozitivní shodě.

KOEFICIENTY ASOCIACE

Z koeficientů asociace, které vyjadřují míru podobnosti lze zpravidla odvodit koeficienty nepodobnosti

$$d_X(\mathbf{x}_i, \mathbf{x}_j) = 1 - s_x(\mathbf{x}_i, \mathbf{x}_j)$$

V případě Jaccardova a Dicova koeficientu nepodobnosti je dodefinována hodnota i pro případy úplné negativní shody tak, že

$$d_J(x_i,x_j) = d_D(x_i,x_j) = 0 \text{ pro } A = B = C = 0$$

KOEFICIENTY ASOCIACE

☑ Rogersův-Tanimotův koeficient

$$s_{RT}(\mathbf{x}_i, \mathbf{x}_j) = \frac{A + D}{A + D + 2.(B + C)} = \frac{A + D}{(B + C) + (A + B + C + D)}$$

☑ Hammanův koeficient

$$s_H(\mathbf{x}_i, \mathbf{x}_j) = \frac{A + D - (B + C)}{A + B + C + D}$$

Na rozdíl od všech předcházejících nabývá Hammanův koeficient hodnot z intervalu ⟨-1,1⟩, přičemž hodnoty -1 nabývá, pokud se příznaky neshodují ani jednou, 0 nabývá když je počet shod a neshod v rovnováze a +1 je v případě úplné shody mezi všemi příznaky.

KOEFICIENTY ASOCIACE

Na základě četností A až D lze vytvářet také dříve uvedené míry:

- **Hammingova vzdálenost**

$$\rho_H(\mathbf{x}_i, \mathbf{x}_j) = B + C$$

- **Euklidova vzdálenost**

$$\rho_E(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{B + C}$$

KOEFICIENTY ASOCIACE

Na základě četností A až D lze vytvářet také dříve uvedené míry:

- **Pearsonův korelační koeficient**

$$s_r(\mathbf{x}_i, \mathbf{x}_j) = \frac{A.D - B.C}{\sqrt{(A+B).(C+D).(A+C).(B+D)}}$$

- **kritérium shody χ^2**

$$s_\chi(\mathbf{x}_i, \mathbf{x}_j) = n.s_r^2(\mathbf{x}_i, \mathbf{x}_j)$$

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 podobnosti)

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.

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

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

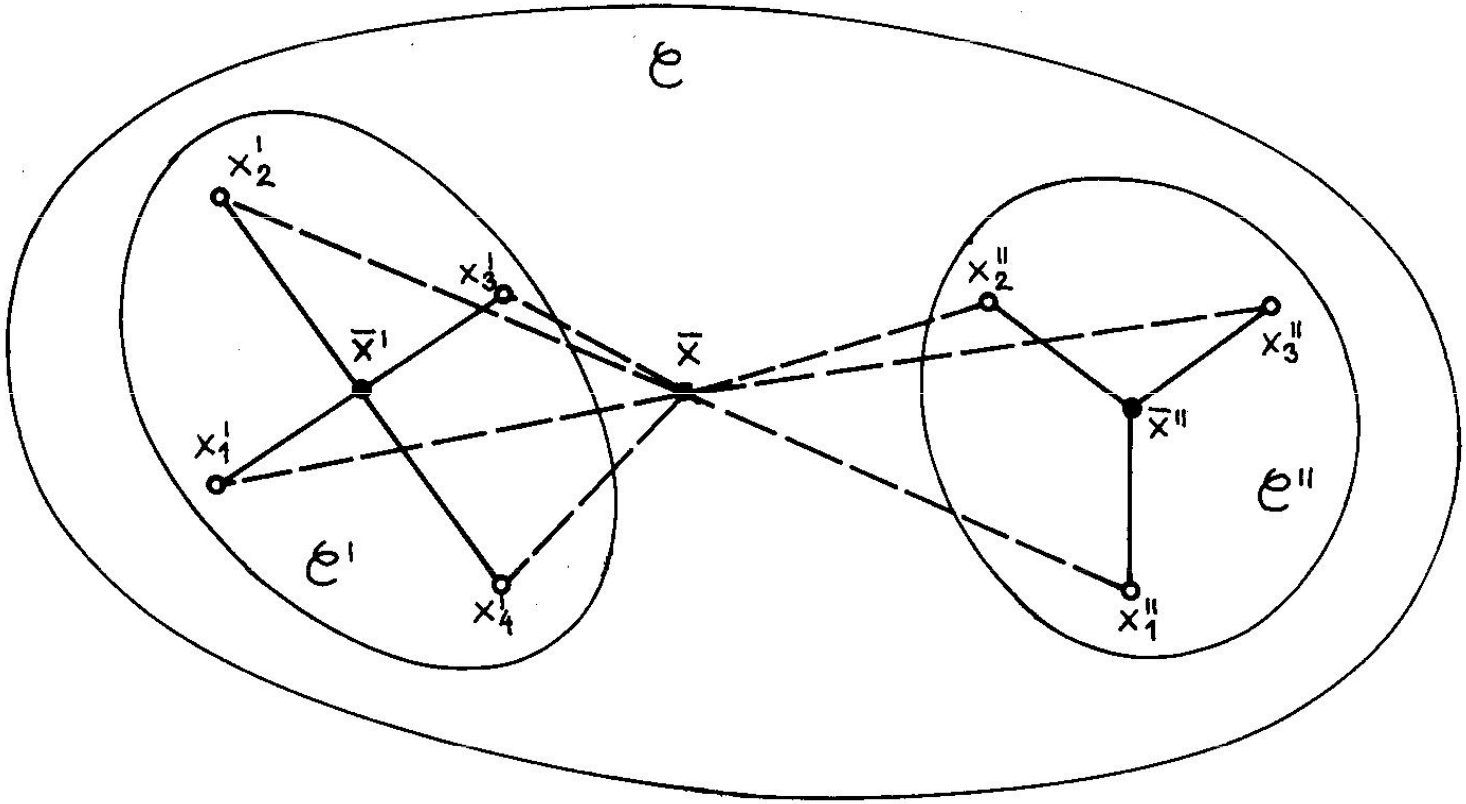
$$D_W(C_i, C_i) = \sum_{x_i \in C_i \cup C_j} \sum_{k=1}^{n} (x_{ik} - \bar{x}_k)^2 -$$

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

Pozn.:

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

WARDOVA METODA



Příprava nových učebních materiálů oboru Matematická biologie

je podporována projektem ESF

č. CZ.1.07/2.2.00/07.0318

„VÍCEOBOROVÁ INOVACE STUDIA MATEMATICKÉ BIOLOGIE“

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ