



ANALÝZA A KLASIFIKACE BIOMEDICÍNSKÝCH DAT



prof. Ing. Jiří Holčík, CSc.

VII. VOLBA A VÝBĚR PŘÍZNAKŮ



ZAČÍNÁME

☑ kolik a jaké příznaky ?

→ málo příznaků – možná chyba klasifikace;

→ moc příznaků – možná nepřiměřená pracnost, vysoké náklady;



KOMPROMIS

(potřebujeme kritérium)

ZAČÍNÁME

KOMPROMIS

(potřebujeme kritérium)

- ☑ přípustná míra spolehlivosti klasifikace (např. pravděpodobnost chybné klasifikace, odchylka obrazu vytvořeného z vybraných příznaků vůči určitému referenčnímu);
- ☑ určit ty příznakové proměnné, jejichž hodnoty nesou nejvíce informace z hlediska řešené úlohy, tj. ty proměnné, kterou jsou nejefektivnější pro vytvoření co nejoddělenějších klasifikačních tříd;

ZAČÍNÁME

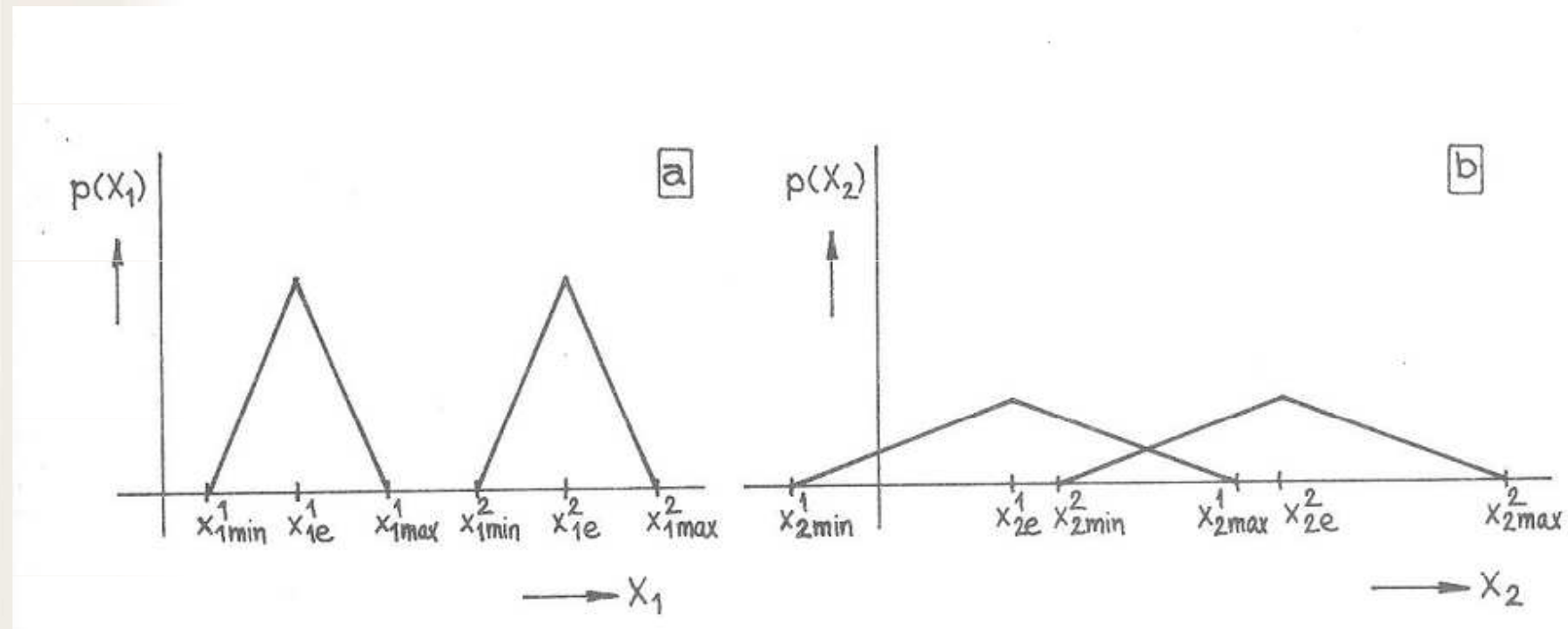
- ☑ algoritmus pro určení příznakových veličin nesoucích nejvíce informace pro klasifikátor není dosud teoreticky formalizován - pouze dílčí suboptimální řešení spočívající:
 - ve výběru nezbytného množství veličin z předem zvolené množiny;
 - vyjádření původních veličin pomocí menšího počtu skrytých nezávislých veličin, které zpravidla nelze přímo měřit, ale mohou nebo také nemusí mít určitou věcnou interpretaci

VOLBA PŘÍZNAKŮ

- ☑ počáteční volba příznakových veličin je z velké části empirická, vychází ze zkušeností získaných při empirické klasifikaci člověkem a závisí, kromě rozboru podstaty problému i na technických (ekonomických) možnostech a schopnostech hodnoty veličin určit

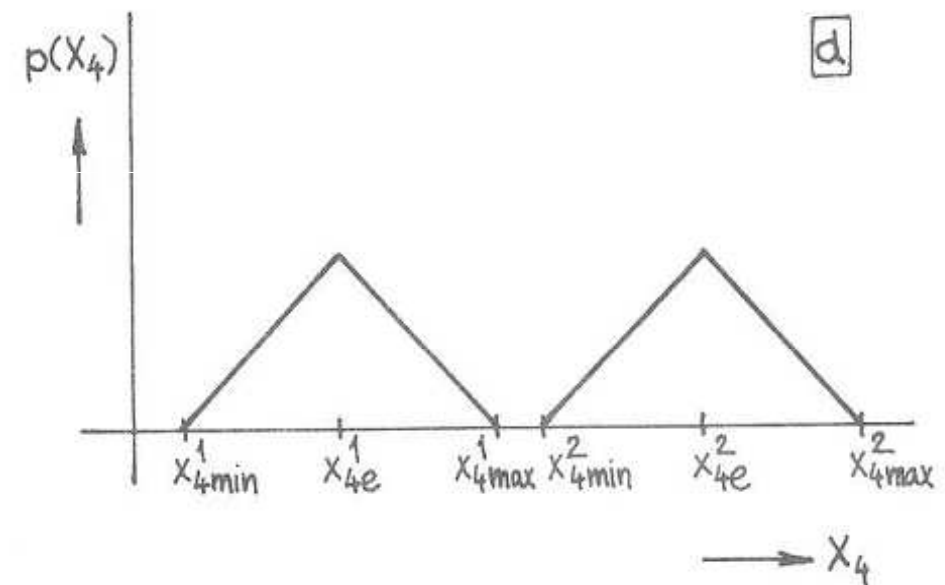
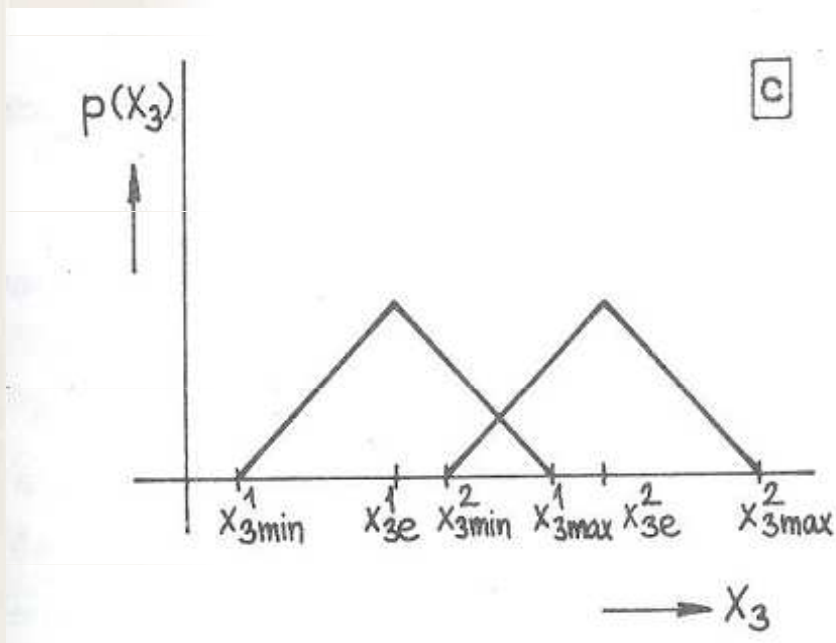
ZÁSADY PRO VOLBU PŘÍZNAKŮ

- ☑ výběr veličin s minimálním rozptylem uvnitř tříd



ZÁSADY PRO VOLBU PŘÍZNAKŮ

- ☑ výběr veličin s maximální vzdáleností mezi třídami

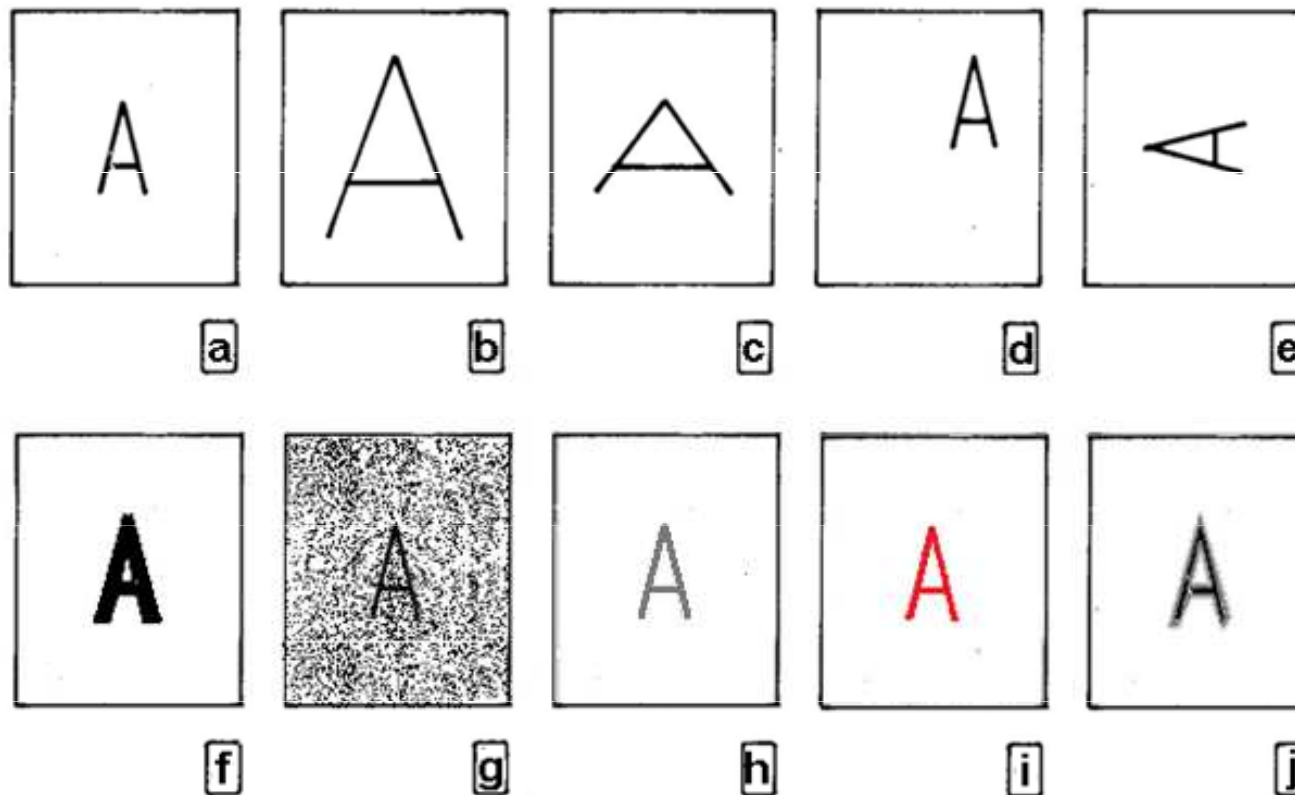


ZÁSADY PRO VOLBU PŘÍZNAKŮ

- ☑ výběr vzájemně nekorelovaných veličin
 - pokud jsou hodnoty jedné příznakové veličiny závislé na příznacích druhé veličiny, pak použití obou těchto veličin nepřináší žádnou další informaci pro správnou klasifikaci – stačí jedna z nich, jedno která

ZÁSADY PRO VOLBU PŘÍZNAKŮ

- ☑ výběr veličin invariantních vůči deformacím
 - volba elementů formálního popisu závisí na vlastnostech původních i předzpracovaných dat a může ovlivňovat způsob předzpracování



VÝBĚR PŘÍZNAKŮ

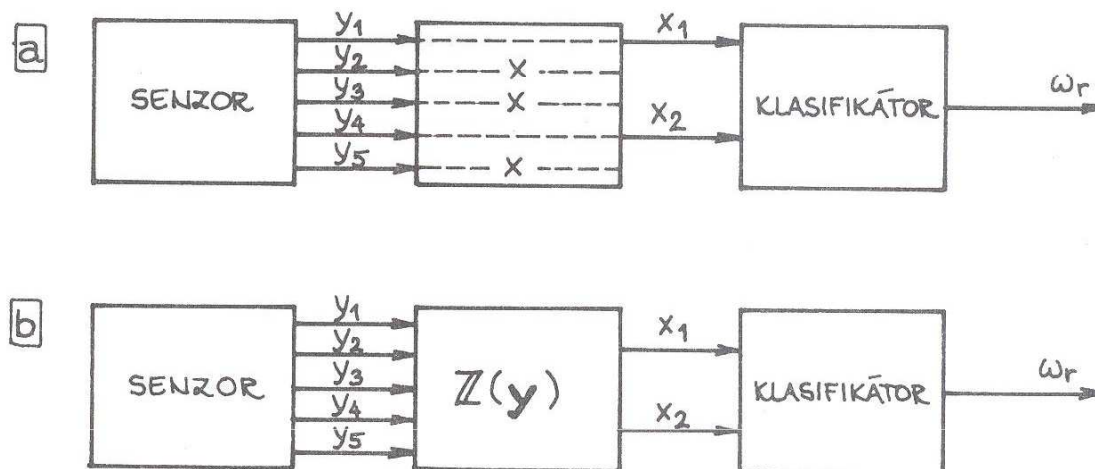
- ☑ formální popis objektu původně reprezentovaný m rozměrným vektorem se snažíme vyjádřit vektorem n rozměrným tak, aby množství diskriminační informace obsažené v původním vektoru bylo v co největší míře zachováno

$$Z: \mathcal{Y}^m \rightarrow \mathcal{X}^n$$

VÝBĚR PŘÍZNAKŮ

dva principiálně různé způsoby:

- ☑ **selekce** – nalezení a odstranění těch příznakových funkcí, které přispívají k separabilitě klasifikačních tříd nejméně;
- ☑ **extrakce** – transformace původních příznakových proměnných na menší počet jiných příznakových proměnných



VÝBĚR PŘÍZNAKŮ

dva principiálně různé způsoby:

- ☑ **selekce** – nalezení a odstranění těch příznakových funkcí, které přispívají k separabilitě klasifikačních tříd nejméně;
- ☑ **extrakce** – transformace původních příznakových proměnných na menší počet jiných příznakových proměnných

Abychom dokázali realizovat libovolný z obou způsobů výběru, je třeba definovat a splnit určité podmínky optimality.

VÝBĚR PŘÍZNAKŮ

PODMÍNKY OPTIMALITY

Nechť J je kritériální funkce, jejíž pomocí vybíráme příznakové veličiny.

V případě **selekce** vybíráme vektor $x = {}^T(x_1, \dots, x_n)$ ze všech možných n -tic χ příznaků y_i , $i = 1, 2, \dots, m$.

Optimalizaci selekce příznaků formálně zapíšeme jako

$$\underset{\forall \chi}{\text{extr}} J(\chi)$$

Problémy k řešení:

- stanovení kritériální funkce;
- stanovení nového rozměru kritériální funkce;
- stanovení optimalizačního postupu

VÝBĚR PŘÍZNAKŮ

PODMÍNKY OPTIMALITY

Nechť J je kritériální funkce, jejíž pomocí vybíráme příznakové veličiny.

V případě **extrakce** transformujeme příznakový prostor na základě výběru zobrazení \mathcal{Z} z množiny všech možných zobrazení ζ prostoru \mathcal{Y}^m do \mathcal{X}^n , tj.

$$\mathcal{Z}(\mathbf{y}) = \underset{\forall \zeta}{\text{extr}} J(\zeta)$$

Příznakový prostor je pomocí optimálního zobrazení \mathcal{Z} dán vztahem $\mathbf{x} = \mathcal{Z}(\mathbf{y})$

Problémy k řešení:

- stanovení kritériální funkce;
- stanovení nového rozměru kritériální funkce;
- zvolení požadavků na vlastnosti zobrazení;
- stanovení optimalizačního postupu

SELEKCE PŘÍZNAKŮ PRAVDĚPODOBNOSTNÍ MÍRY

☑ pro bayesovské klasifikátory

je-li $\chi = (\chi_1, \chi_2, \dots, \chi_n)$ možná n -tice příznaků, vybraných ze všech možných m hodnot y_i , $i=1, \dots, m$, $n \leq m$, pak pravděpodobnost chybného rozhodnutí P_{eme} je pro tento výběr rovna

$$\begin{aligned} P_{\text{eme}} &= J(a^*) = \min_{\forall a} J(a) = \int_{\chi} \min_{\forall r} L_{\chi}(\omega_r) d\chi = \\ &= \int_{\chi} \min_{\forall r} [p(\chi) - p(\chi | \omega_r) \cdot P(\omega_r)] d\chi = \int_{\chi} p(\chi) d\chi - \int_{\chi} \max_{\forall r} p(\chi | \omega_r) \cdot P(\omega_r) d\chi = \\ &= 1 - \int_{\chi} \max_{\forall r} p(\chi | \omega_r) \cdot P(\omega_r) d\chi \end{aligned}$$

SELEKCE PŘÍZNAKŮ PRAVDĚPODOBNOSTNÍ MÍRY

- ☑ pro dichotomický bayesovský klasifikátor ($R=2$) je celková pravděpodobnost chybného rozhodnutí

$$e = 1 - \int_{\mathcal{X}} |p(\chi | \omega_1)P(\omega_1) - p(\chi | \omega_2)P(\omega_2)| d\chi$$

- ☑ pravděpodobnost chyby bude maximální, když integrál bude nulový – obě váhované hustoty pravděpodobnosti budou stejné, pravděpodobnost chyby bude minimální, když se obě hustoty nebudou překrývat.
- ☑ Čím větší vzdálenost mezi klasifikačními třídami, tím menší pravděpodobnost chyby



**Integrál může být považován za vyjádření
„pravděpodobnostní vzdálenosti“**

SELEKCE PŘÍZNAKŮ PRAVDĚPODOBNOSTNÍ MÍRY

- ☑ zobecnění

$$J(\chi) = \int_{\chi} f[p(\chi | \omega_i)P(\omega_i), i = 1,2]d\chi$$

- ☑ $J(\chi) \geq 0$;
- ☑ $J(\chi) = 0$, když jsou hustoty pravděpodobnosti totožné;
- ☑ $J(\chi)$ je maximální, když se hustoty nepřekrývají

SELEKCE PŘÍZNAKŮ

PRAVDĚPODOBNOSTNÍ MÍRY

Chernoffova vzdálenost

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

Bhattacharyyova vzdálenost

$$J_B = - \ln \int [p(x|\omega_1) \cdot p(x|\omega_2)]^{\frac{1}{2}} dx ;$$

Divergence

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

Patrikova - Fisherova vzdálenost

$$J_P = \left\{ \int [p(x|\omega_1) - p(x|\omega_2)]^2 dx \right\}^{\frac{1}{2}},$$

SELEKCE PŘÍZNAKŮ PRAVDĚPODOBNOSTNÍ MÍRY

zprůměrněná Chernoffova vzdálenost

$$J_C = - \ln \int [p(x|\omega_1)P(\omega_1)]^s \cdot [p(x|\omega_2)P(\omega_2)]^{s-1} dx, \quad s \in \langle 0, 1 \rangle;$$

zprůměrněná Bhattacharyyova vzdálenost

$$J_B = - \ln \int [p(x|\omega_1) \cdot P(\omega_1) \cdot p(x|\omega_2) \cdot P(\omega_2)]^{\frac{1}{2}} dx ;$$

zprůměrněná divergence

$$J_D = \int [p(x|\omega_1)P(\omega_1) - p(x|\omega_2)P(\omega_2)] \cdot \ln \left(\frac{p(x|\omega_1)P(\omega_1)}{p(x|\omega_2)P(\omega_2)} \right) dx ;$$

zprůměrněná Patrikova - Fisherova vzdálenost

$$J_P = \left\{ \int [p(x|\omega_1)P(\omega_1) - p(x|\omega_2)P(\omega_2)]^2 dx \right\}^{\frac{1}{2}} .$$

SELEKCE PŘÍZNAKŮ PRAVDĚPODOBNOSTNÍ MÍRY

- ☑ pro více klasifikačních tříd tzv. bayesovská vzdálenost

$$J_{BA} = \int_{\chi} \left(\sum_{r=1}^R P^2(\omega_r | \chi) \right) \cdot p(\chi) d\chi$$

SELEKCE PŘÍZNAKŮ POMĚR ROZPTYLŮ

- ☑ rozptyl uvnitř třídy pomocí disperzní matice

$$D(\chi) = \sum_{r=1}^R P(\omega_r) \int_{\chi} (\chi - \mu_r)^T (\chi - \mu_r) p(\chi | \omega_r) d\chi,$$

kde

$$\mu_r = \int_{\chi} p(\chi | \omega_r) d\chi$$

SELEKCE PŘÍZNAKŮ POMĚR ROZPTYLŮ

- ☑ rozptyl mezi třídami může být dán

$$B(\chi) = \sum_{r=1}^{R-1} \sum_{s=r+1}^R P(\omega_r) \cdot P(\omega_s) \cdot \mu_{rs} \cdot \mu_{rs}^T,$$

$$\text{kde } \mu_{rs} = \mu_r - \mu_s$$

- ☑ pokud $\mu_0 = \sum_{r=1}^R P(\omega_r) \cdot \mu_r = \int \chi \cdot p(\chi) d\chi$

Ize také psát

$$B(\chi) = \sum_{r=1}^R P(\omega_r) \cdot (\mu_r - \mu_0) \cdot (\mu_r - \mu_0)^T,$$

SELEKCE PŘÍZNAKŮ POMĚR ROZPTYLŮ

- ☑ vyjádření vztahu obou rozptylů

$$J_{r1}(\chi) = \text{tr}(D^{-1}(\chi) \cdot B(\chi))$$

$$J_{r2}(\chi) = \text{tr}(B(\chi)) / \text{tr}(D(\chi))$$

$$J_{r3}(\chi) = |D^{-1}(\chi) \cdot B(\chi)| = |B(\chi)| / |D(\chi)|$$

$$J_{r4}(\chi) = \ln(J_{r3}(\chi))$$

NEPRAVDĚPODOBNOSTNÍ METRIKY

- ✓ **Euklidovská metrika** (s nejnázornější geometrickou interpretací)

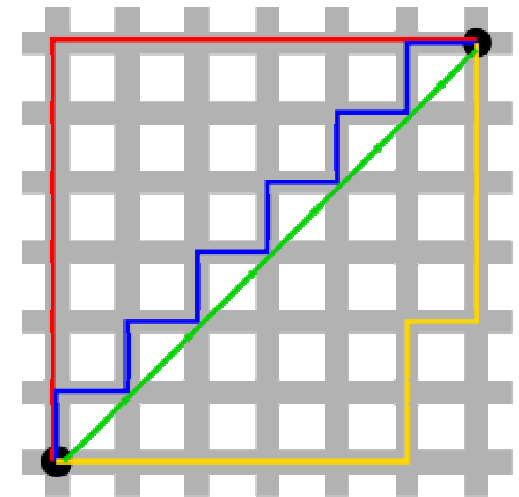
$$\rho_E(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{i=1}^n (x_{1i} - x_{2i})^2 \right]^{1/2}$$

- ✓ **Hammingova metrika (Manhattan m.)**

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

- ✓ **Minkovského metrika**

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



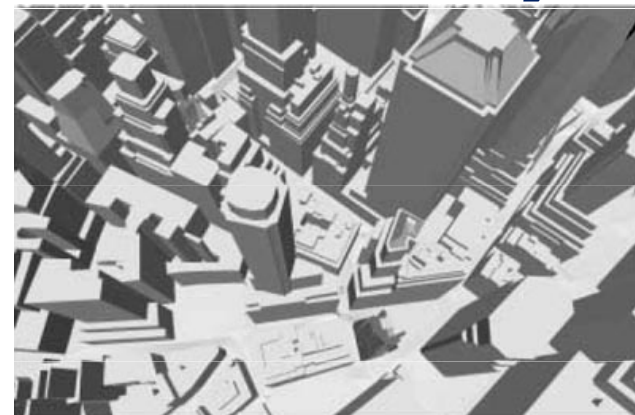
NEPRAVDĚPODOBNOSTNÍ METRIKY

- ✓ **Euklidovská metrika** (s nejnázornější geometrickou interpretací)

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

- ✓ **Hammingova metrika (Manhattan m.)**

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



- ✓ **Minkovského metrika**

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

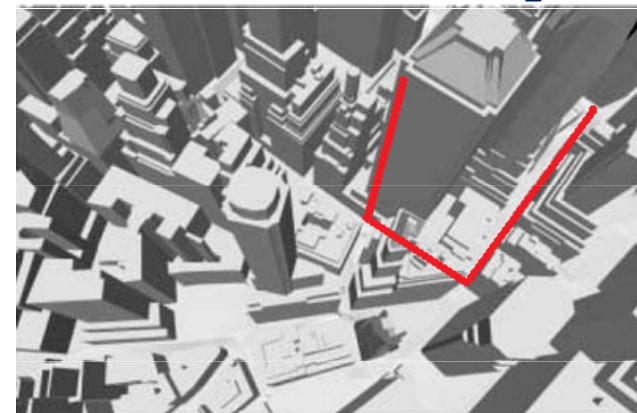
NEPRAVDĚPODOBNOSTNÍ METRIKY

- ✓ **Euklidovská metrika** (s nejnázornější geometrickou interpretací)

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

- ✓ **Hammingova metrika (Manhattan m.)**

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



- ✓ **Minkovského metrika**

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

NEPRAVDĚPODOBNOSTNÍ METRIKY

☑ Čebyševova metrika

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

Čebyševovu metriku lze vyjádřit jako limitu

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

☑ Sokalova metrika

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

NEPRAVDĚPODOBNOSTNÍ METRIKY

☑ Čebyševova metrika

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

Čebyševovu metriku lze vyjádřit jako limitu

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

☑ Sokalova metrika

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

NEPRAVDĚPODOBNOSTNÍ METRIKY

☑ Čebyševova metrika

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

Čebyševovu metriku lze vyjádřit jako limitu

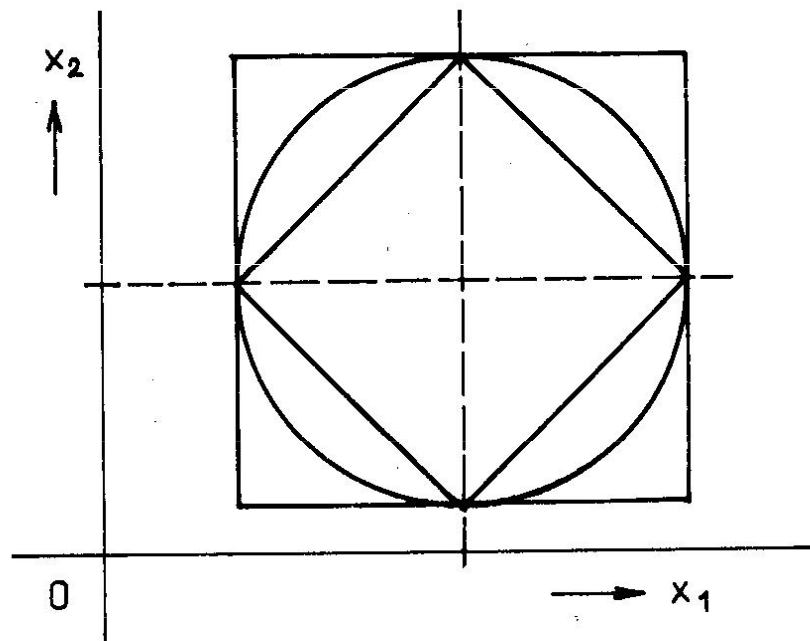
$$\rho_C(\mathbf{x}_1, \mathbf{x}_2) = \lim_{m \rightarrow \infty} \rho_M(\mathbf{x}_1, \mathbf{x}_2)$$

☑ Sokalova metrika

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

NEPRAVDĚPODOBNOSTNÍ METRIKY

Volba podle toho jak potřebujeme posílit vliv proměnných, u nichž je pro dané obrazy x_1 a x_2 velký rozdíl.



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 $\mathbf{x} = (x_1, x_2, \dots, x_n)$

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

rozpětí

$$\Delta_j = \max_i x_{ij} - \min_i x_{ij},$$

resp. standardizací podle vztahu

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

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

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

NEPRAVDĚPODOBNOSTNÍ METRIKY

možné odstranění (potlačení) nevýhod:

- ☑ váhování příznaků lze zapsat maticově

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

kde prvky transformační matice \mathbf{C} jsou definovány jako

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

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

Za tohoto formalismu je Euklidova metrika definována vztahem

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

NEPRAVDĚPODOBNOSTNÍ METRIKY

možné odstranění (potlačení) nevýhod:

- ☑ pokud jsou složky transformovaného obrazu dány lineární kombinací více složek původního obrazu, není ani matice \mathbf{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**

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

NEPRAVDĚPODOBNOSTNÍ METRIKY

možné odstranění (potlačení) nevýhod:

- ☑ pokud jsou složky transformovaného obrazu dány lineární kombinací více složek původního obrazu, není ani matice \mathbf{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**

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

- ☑ **Kovarianční matice** dvou (náhodných) vektorů $\mathbf{x} = \mathbf{1}^T(x_1, \dots, x_m)$ a $\mathbf{y} = \mathbf{1}^T(y_1, \dots, y_n)$ je dána vztahem
 - ☑ $\mathbf{K}(\mathbf{x}, \mathbf{y}) = E((\mathbf{x} - E\mathbf{x}) \cdot \mathbf{1}^T(\mathbf{y} - E\mathbf{y})) = [\text{cov}(x_i, y_j)]_{m,n}$

KOEFICIENTY KORELACE

KOEFICIENTY KORELACE

TO SNAD OPAKOVAT NEMUSÍME

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

☑ Jaccardův (Tanimotův) koeficient

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

(není definován pro dvojice obrazů, které vykazují negativní shodu ve všech příznacích);

☑ Soklův- Michenerův koeficient

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

☑ Diceův koeficient

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

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

☑ Rogersův-Tanimotův koeficient

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

☑ 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 $\langle -1, 1 \rangle$, 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:

☑ **Pearsonův korelační koeficient**

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

☑ **kritérium shody χ^2**

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

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

Z koeficientů asociace, které vyjadřují míru podobnosti lze 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

NÁZEV	ROZSAH	VZTAH
Cosine	0.0, 1.0	$\frac{c}{\sqrt{(a+c)*(b+c)}}$
Dice	0.0, 1.0	$\frac{2.0*c}{(a+c)+(b+c)}$
Euclid	0.0, 1.0	$\sqrt{\frac{c+d}{a+b+c+d}}$
Forbes	0.0, ∞	$\frac{c*(a+b+c+d)}{(a+c)*(b+c)}$
Hamman	-1.0, 1.0	$\frac{(c+d)-(a+b)}{a+b+c+d}$
Jaccard	0.0, 1.0	$\frac{c}{a+b+c}$
Kulczynski	0.0, 1.0	$0.5*\left(\frac{c}{a+c} + \frac{c}{b+c}\right)$
Manhattan	1.0, 0.0	$\frac{(a+b)}{(a+b+c+d)}$

KOEFICIENTY ASOCIACE

NÁZEV	ROZSAH	VZTAH
Sokal-Michener	0.0,1.0	$\frac{c+d}{a+b+c+d}$
Pearson	-1.0,1.0	$\frac{(c*d) - (a*b)}{\sqrt{(a+c)*(b+c)*(a+d)*(b+d)}}$
Rogers-Tanimoto	0.0,1.0	$\frac{c+d}{(a+b) + (a+b+c+d)}$
Russell-Rao	0.0,1.0	$\frac{c}{a+b+c+d}$
Simpson	0.0,1.0	$\frac{c}{\min((a+c), (b+c))}$
Yule	-1.0,1.0	$\frac{(c*d) - (a*b)}{(c*d) + (a*b)}$

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 $a \in C_i$ a $b \in C_j$ jsou libovolné skupiny množiny obrazů $\{x_i\}$, $i=1, \dots, K$, potom metoda nejbližšího souseda definuje mezi skupinami C_i a C_j vzdálenost

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

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

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

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

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

Pozn.:

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

- ☑ je možné i zobecnění pro více nejbližších sousedů

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

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.

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

- ☑ pak

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

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

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

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

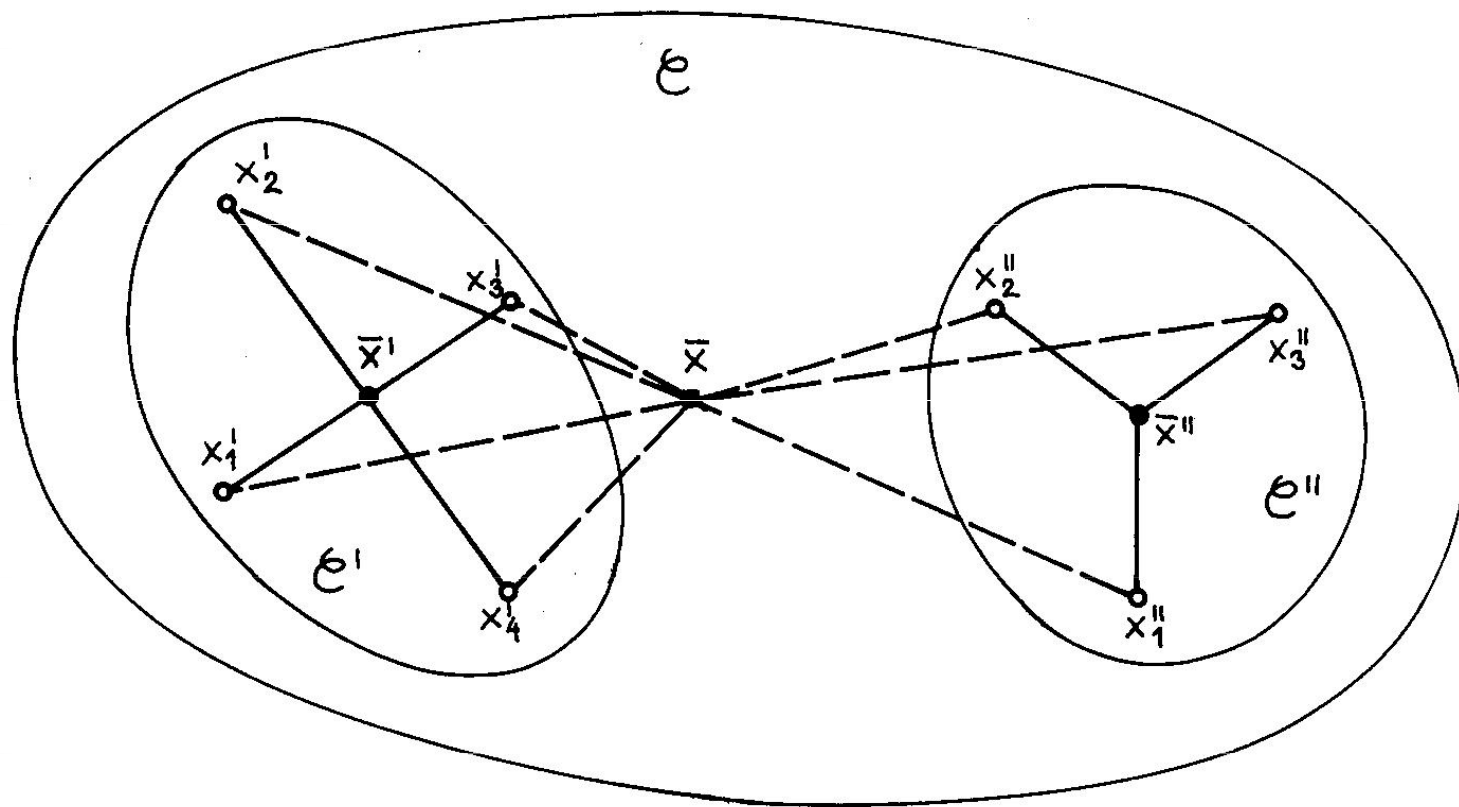
- ☑ jsou-li $\bar{\mathbf{x}}_i$ a $\bar{\mathbf{x}}_j$ těžiště tříd C_i a C_j a $\bar{\mathbf{x}}$ těžiště sjednocené množiny, pak Wardova vzdálenost obou shluků je definována výrazem

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

Pozn.:

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

WARDOVA METODA



ALGORITMY SELEKCE PŘÍZNAKŮ

- ☑ výběr optimální podmnožiny obsahující n ($n \leq m$) příznakových proměnných – kombinatorický problém ($m!/(m-n)!n!$ možných řešení)



hledáme jen kvazioptimální řešení

ALGORITMUS OHRANIČENÉHO VĚTVENÍ

předpoklad:

- ☑ monotónnost kritéria selekce - označíme-li X_j množinu obsahující j příznaků, pak monotónnost kritéria znamená, že podmnožiny

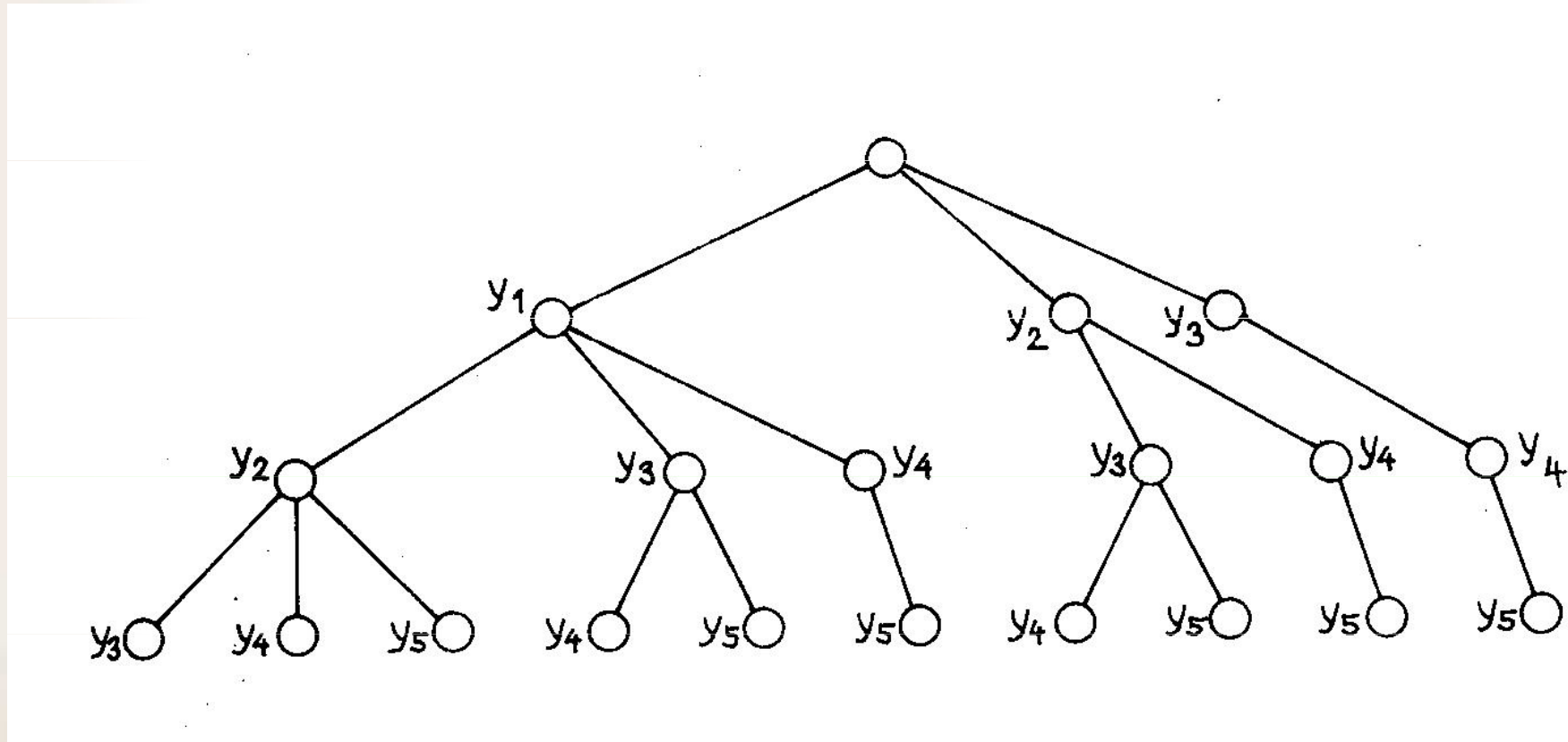
$$X_1 \subset X_2 \subset \dots \subset X_j \subset \dots \subset X_m$$

splňuje selekční kritérium vztah

$$J(X_1) \leq J(X_2) \leq \dots \leq J(X_m)$$

ALGORITMUS OHRANIČENÉHO VĚTVENÍ

uvažme případ selekce dvou příznaků z pěti



ALGORITMUS SEKVENČNÍ DOPŘEDNÉ SELEKCE

- ☑ algoritmus začíná s prázdnou množinou, do které se vloží proměnná s nejlepší hodnotou selekčního kritéria;
- ☑ v každém následujícím kroku se přidá ta proměnná, která s dříve vybranými veličinami dosáhla nejlepší hodnoty kritéria, tj.

$$J(\{X_{k+1}\}) = \max J(\{X_k \cup y_j\}), y_j \in \{Y - X_k\}$$

ALGORITMUS SEKVENČNÍ ZPĚTNÉ SELEKCE

- ☑ algoritmus začíná s množinou všech příznakových veličin;
- ☑ v každém následujícím kroku se eliminuje ta proměnná, která způsobuje nejmenší pokles kritériální funkce, tj. po $(k+1)$. kroku platí

$$J(\{X_{m-k-1}\}) = \max J(\{X_{m-k} - y_j\}), y_j \in \{X_{m-k}\}$$

ALGORITMY SEKVENČNÍ SELEKCE

SUBOPTIMALITA

Suboptimalita nalezeného řešení sekvenčních algoritmů je způsobena:

- ☑ dopředná selekce - tím, že nelze vyloučit ty veličiny, které se staly nadbytečné po přiřazení dalších veličin;
- ☑ zpětná selekce – neexistuje možnost opravy při neoptimálním vyloučení kterékoliv proměnné;

Dopředný algoritmus je výpočetně jednodušší, protože pracuje maximálně v n -rozměrném prostoru, naopak zpětný algoritmus umožňuje průběžně sledovat množství ztracené informace.

ALGORITMUS PLUS P MÍNUS Q

- ✓ po přidání p veličin se q veličin odstraní;
- ✓ proces probíhá, dokud se nedosáhne požadovaného počtu příznaků;
- ✓ je-li $p > q$, pracuje algoritmus od prázdné množiny;
- ✓ je-li $p < q$, varianta zpětného algoritmu

ALGORITMUS MIN - MAX

Heuristický algoritmus vybírající příznaky na základě výpočtu hodnot kritériální funkce pouze v jedno- a dvourozměrném příznakovém prostoru.

Předpokládejme, že bylo vybráno k příznakových veličin do množiny $\{X_k\}$ a zbývají veličiny z množiny $\{Y-X_k\}$. Výběr veličiny $y_j \in \{Y-X_k\}$ přináší novou informaci, kterou můžeme ocenit relativně k libovolné veličině $x_i \in X_k$ podle vztahu

$$\Delta J(y_j, x_i) = J(y_j, x_i) - J(x_i)$$

ALGORITMUS MIN - MAX

Informační přírůstek ΔJ musí být co největší, ale musí být dostatečný pro všechny veličiny již zahrnuté do množiny X_k .
Vybíráme tedy veličinu y_{k+1} , pro kterou platí

$$\Delta J(y_{k+1}, X_k) = \max_j \min_i \Delta J(y_j, x_i), x_i \in X_k$$