ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

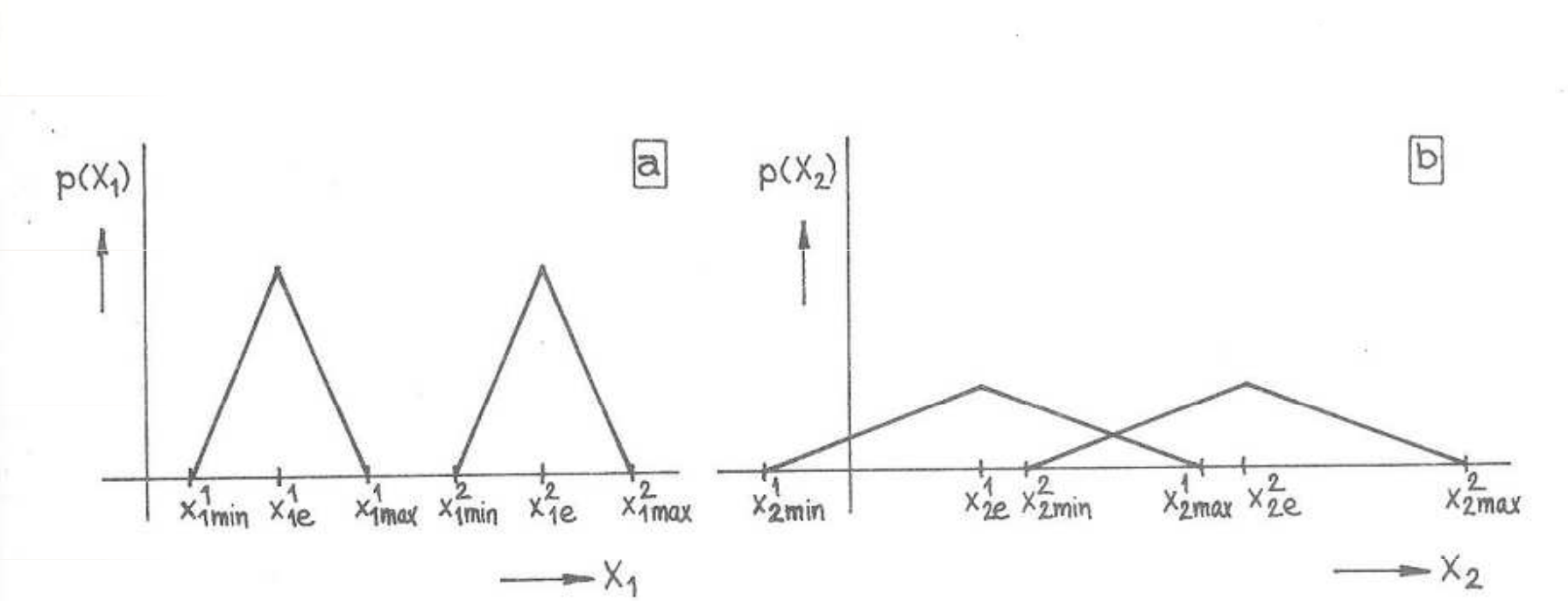
- algoritmus pro určení příznakových veličin nesoucích nejvíce informace pro klasifikátor není dosud teoreticky formalizován - pouze dílčí suboptimální řešení spočívající:
  - ve výběru nezbytného množství veličin z předem zvolené množiny;
  - vyjádření původních veličin pomocí menšího počtu skrytých 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

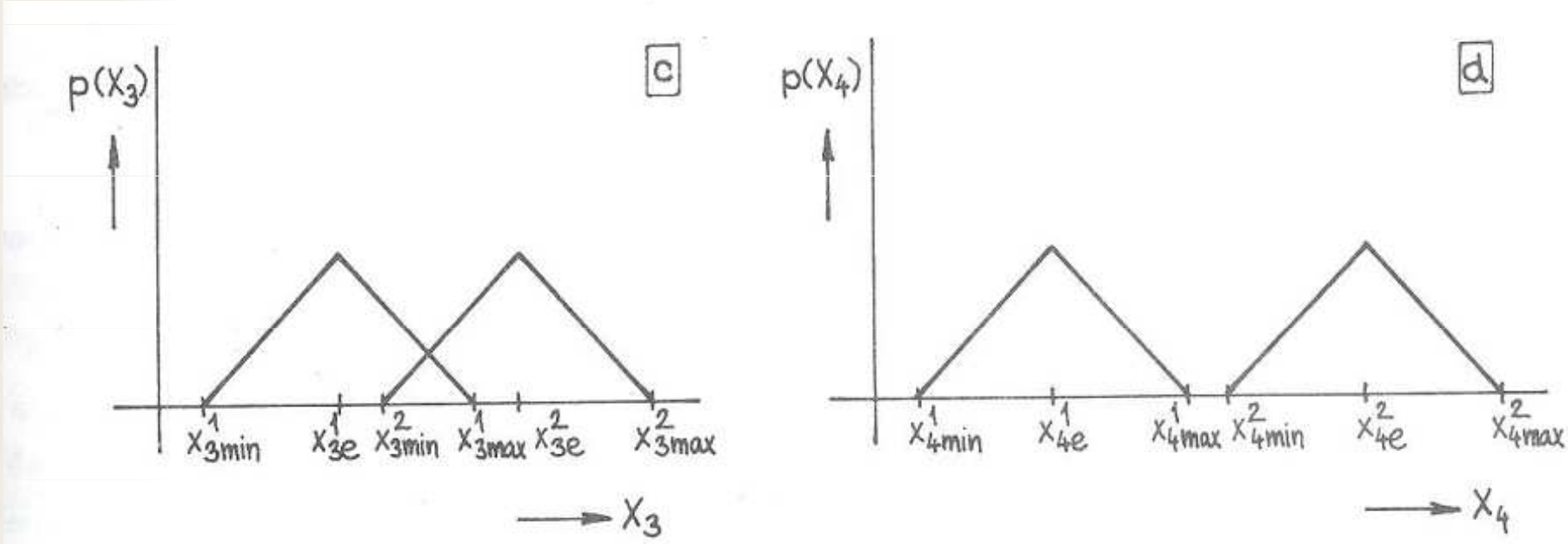
ZÁSADY PRO VOLBU PŘÍZNAKŮ

- ☑ výběr veličin s minimálním rozptylem uvnitř tříd



ZÁSADY PRO VOLBU PŘÍZNAKŮ

- ☑ výběr veličin s maximální vzdáleností mezi třídami

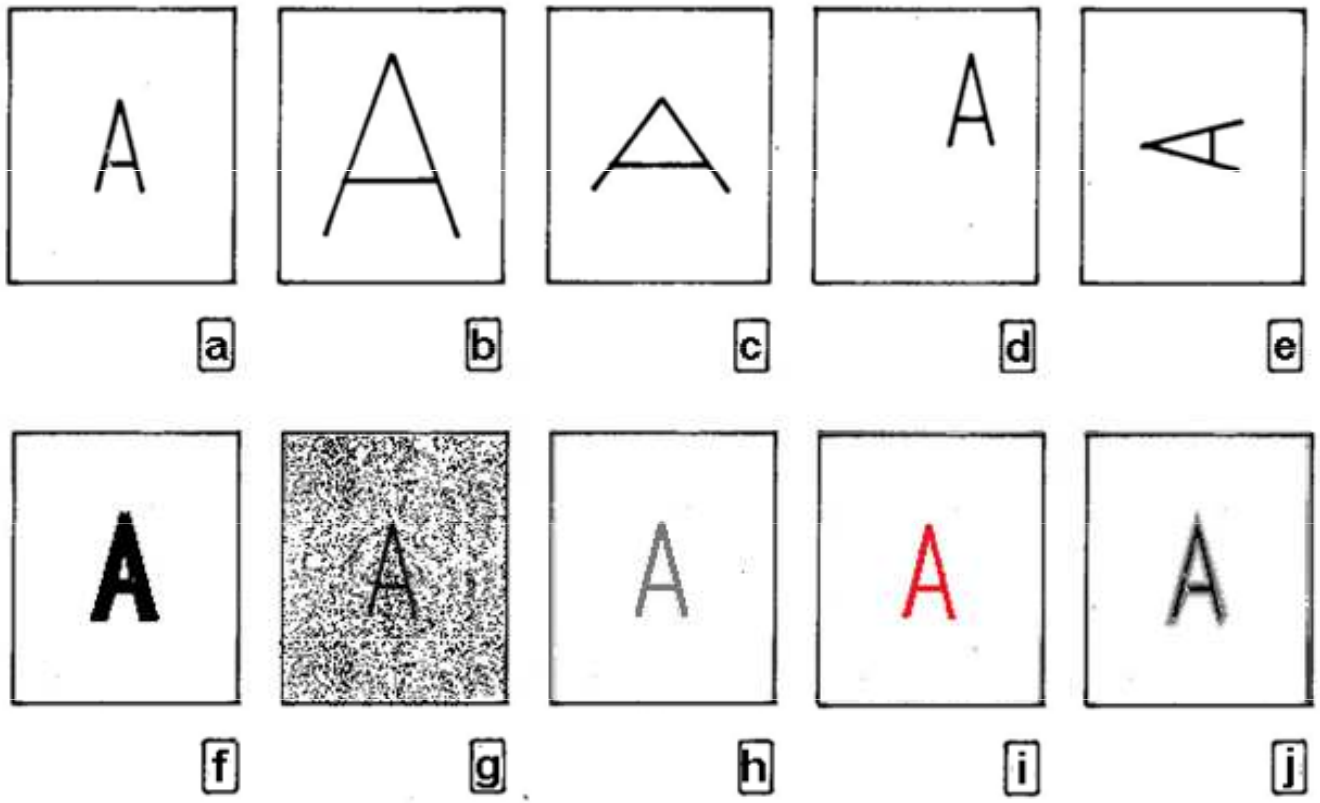


ZÁSADY PRO VOLBU PŘÍZNAKŮ

- ☑ 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

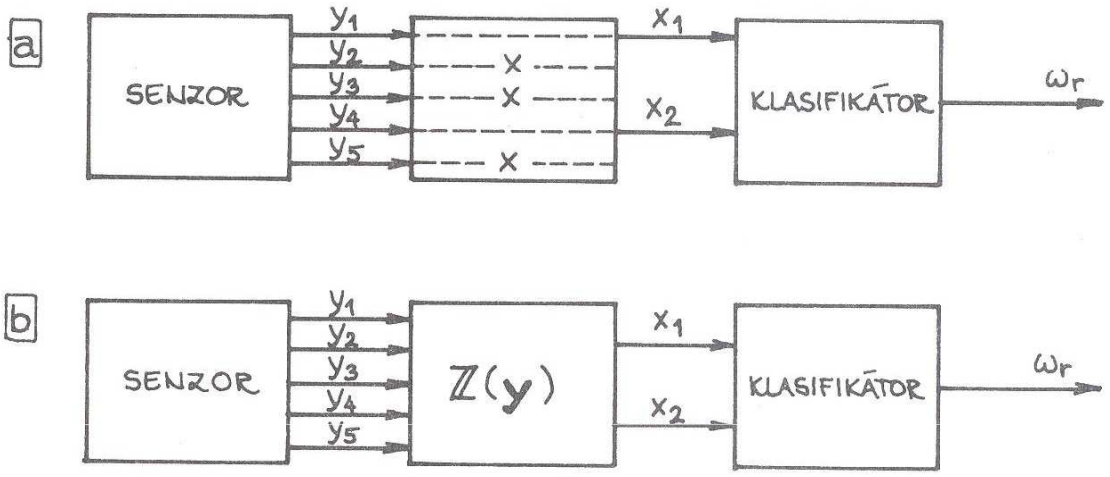
- formální popis objektu původně reprezentovaný m rozměrným vektorem se snažíme vyjádřit vektorem n rozměrným tak, aby množství diskriminační informace obsažené v původním vektoru bylo v co největší míře zachováno

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

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 kriteriální funkce, jejíž pomocí vybíráme příznakové veličiny.

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

$$\gtrsim(\mathbf{y}) = \underset{\forall \chi}{\text{extr}}\, J(\chi)$$

Problémy k řešení:

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

VÝBĚR PŘÍZNAKŮ
PODMÍNKY OPTIMALITY

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

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

$$\succsim(\mathbf{y}) = \underset{\forall \zeta}{\operatorname{extr}}\, J(\zeta)$$

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

Problémy k řešení:

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

SELEKCE PŘÍZNAKŮ
KRITERIÁLNÍ FUNKCE

- pro bayesovské klasifikátory (to už jsme si říkali)

  je-li $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ možná n-tice příznaků, vybraných ze všech možných m hodnot y_i, $i=1,\ldots,m$, $n \le m$, pak pravděpodobnost chybného rozhodnutí P_{eme} je pro tento výběr rovna

$$P_{eme} = J(a^*) = \min_{\forall a} J(a) = \int_{\chi} \min_{\forall r} L_{\chi}(\omega_r)\,d\mathbf{x} =$$

$$= \int_{\chi} \min_{\forall r} \left[p(\mathbf{x}) - p(\mathbf{x}\mid\omega_r).P(\omega_r)\right]d\mathbf{x} = \int_{\chi} p(\mathbf{x})d\mathbf{x} - \int_{\chi} \max_{\forall r} p(\mathbf{x}\mid\omega_r).P(\omega_r)d\mathbf{x} =$$

$$= 1 - \int_{\chi} \max_{\forall r}\ p(\mathbf{x}\mid\omega_r).P(\omega_r)d\mathbf{x}$$

SELEKCE PŘÍZNAKŮ
PRAVDĚPODOBNOSTNÍ MÍRY

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

$$e = 1 - \int_{\chi} \left| p(\mathbf{x} \mid \omega_1).P(\omega_1) - p(\mathbf{x} \mid \omega_2).P(\omega_2) \right| d\mathbf{x}$$

- 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

- pro více klasifikačních tříd tzv. bayesovská vzdálenost

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

SELEKCE PŘÍZNAKŮ
POMĚR ROZPTYLŮ

- rozptyl uvnitř třídy pomocí disperzní matice

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

kde

$$\mu_r = \int_{\chi} p(\mathbf{x} \mid \omega_r) d\mathbf{x}$$

SELEKCE PŘÍZNAKŮ
POMĚR ROZPTYLŮ

- rozptyl mezi třídami může být dán

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

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

- pokud $\mu_0 = \sum_{r=1}^{R} P(\omega_r).\mu_r = \int_{\chi} \mathbf{x}.p(\mathbf{x})d\mathbf{x}$

lze také psát

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

SELEKCE PŘÍZNAKŮ
POMĚR ROZPTYLŮ

- vyjádření vztahu obou rozptylů

$$J_{r1}(\mathbf{x})=tr(D^{-1}(\mathbf{x}).B(\mathbf{x}))$$

$$J_{r2}(\mathbf{x})=tr(B(\mathbf{x})/tr(D(\mathbf{x}))$$

$$J_{r3}(\mathbf{x})=|D^{-1}(\mathbf{x}).B(\mathbf{x})|= |B(\mathbf{x})|/|D(\mathbf{x})|$$

$$J_{r4}(\mathbf{x}) = \ln(J_{r3}(\mathbf{x}))$$

ALGORITMY SELEKCE PŘÍZNAKŮ

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

$$\Downarrow$$

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

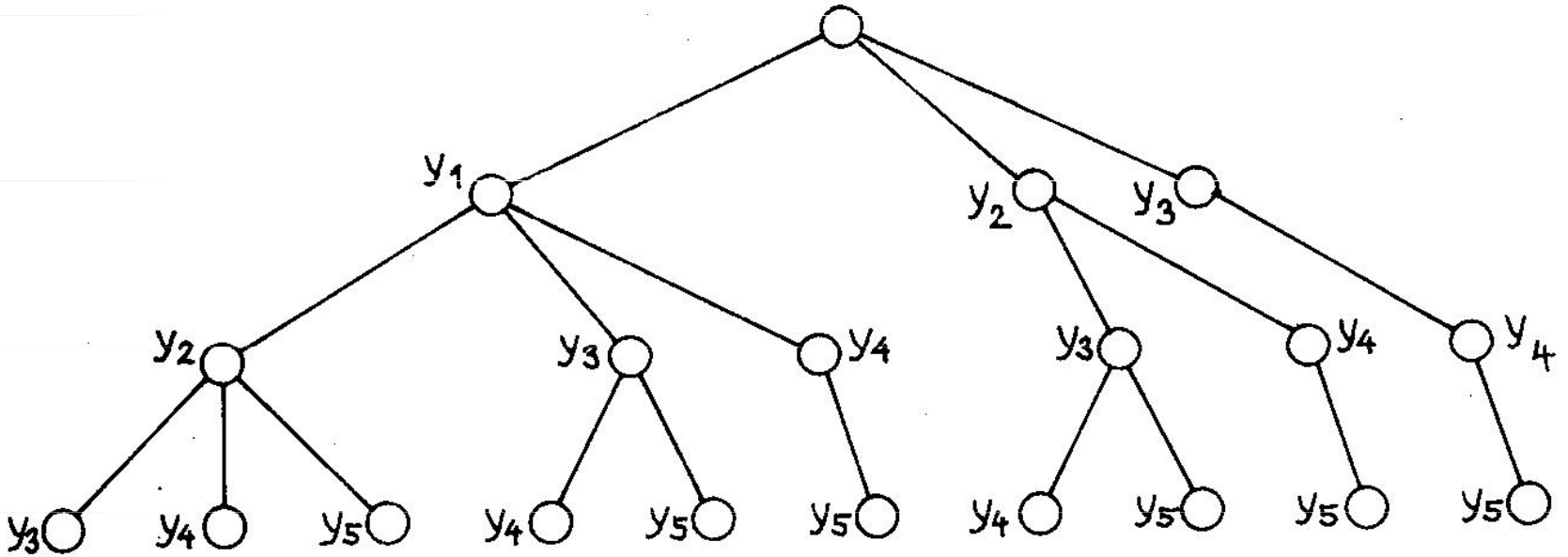
$$X_1 \subset X_2 \subset \ldots \subset X_j \subset \ldots \subset X_m$$

splňuje selekční kritérium vztah

$$J(X_1) \leq J(X_1) \leq \ldots \leq J(X_m)$$

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 kriteriální 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 kriteriální funkce pouze v jedno- a dvourozměrném příznakovém prostoru.

Předpokládejme, že bylo vybráno k příznakových veličin do množiny $\{X_k\}$ a zbývají veličiny z množiny $\{Y\text{-}X_k\}$. Výběr veličiny $y_j \in \{Y\text{-}X_k\}$ 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

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

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"

EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ