Aplikace teorie neuronových sítí

Doc. RNDr. Iveta Mrázová, CSc.

Katedra teoretické informatiky
Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

Aplikace teorie neuronových sítí

- tradiční přístupy -

Doc. RNDr. Iveta Mrázová, CSc.

Katedra teoretické informatiky
Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

Bayesovo rozhodovací pravidlo (kritérium minimální chyby)

- ***c*** tříd: $\omega_1, \ldots, \omega_c$
- Vzor patří ke třídě ω_i s apriorní pravděpodobností $P_i ;\ P_i \geq 0 ;\ \sum_{i=1}^{c} P_i = 1$
- Vzor $\vec{x}$: náhodný vektor v ***n***-rozměrném příznakovém prostoru Ω s podmíněnou hustotou pravděpodobnosti $p(\vec{x}|\omega_i)$, patří-li $\vec{x}$ ke třídě ω_i ; $i = 1, \ldots, c$

Bayesovo rozhodovací pravidlo (2)

Klasifikace $\vec{x}$ z neznámé třídy ω:

- Pravděpodobnost, že $\vec{x}$ patří do třídy ω_j : $P(\omega_j|\vec{x})$

  (~ aposteriorní pravděpodobnost třídy ω_j)

- **Bayesův vzorec:** $$P(\omega_j|\vec{x}) = \frac{p(\vec{x}|\omega_j)P_j}{p(\vec{x})}$$

- Hustota pravděpodobnosti výskytu vzoru $\vec{x}$:

$$p(\vec{x}) = \sum_{j=1}^{c} p(\vec{x}|\omega_j) P_j$$

Bayesovo rozhodovací pravidlo (3)

- Rozhodnutí zařadit $\vec{x}$ do třídy ω_j : $\hat{\omega}(\vec{x}) = \omega_i$
  přinese ztrátu $\lambda(\omega_i | \omega_j)$ s pravděpodobností $P(\omega_j | \vec{x})$
  (~ vzor náležející ke třídě ω_j je chybně zařazen do třídy ω_i ; $i \neq j$)

- Očekávaná ztráta pro chybné rozhodnutí $\hat{\omega}(\vec{x}) = \omega_i$:

$$l^i(\vec{x}) = \sum_{j=1}^{c} \lambda(\omega_i | \omega_j) P(\omega_j | \vec{x})$$

Bayesovo rozhodovací pravidlo (4)

Každé rozhodovací pravidlo má:

- Podmíněnou ztrátu:

$$l(\vec{x}) = \sum_{j=1}^{c} \lambda\left(\hat{\omega}(\vec{x}) \middle| \omega_j\right) P(\omega_j \mid \vec{x})$$

- a střední ztrátu:

$$L = \int_{\Omega} l(\vec{x})\, p(\vec{x})\, d\,\vec{x}$$

$$= \int_{\Omega} \sum_{j=1}^{c} \lambda\left(\hat{\omega}(\vec{x}) \middle| \omega_j\right) p\left(\vec{x} \middle| \omega_j\right) P_j \; d\,\vec{x}$$

Bayesovo rozhodovací pravidlo (5)

Cíl: takové rozhodovací pravidlo $\hat{\omega}$, které by minimalizovalo L

Bayesovo rozhodovací pravidlo $\hat{\omega}^*(\vec{x})$:

$$\hat{\omega}^*(\vec{x}) = \omega_i \text{ , jestliže } l^i(\vec{x}) \le l^j(\vec{x}) \text{ pro } j = 1,\ldots,c$$

- Nejmenší podmíněná ztráta:

$$l^*(\vec{x}) = \min_{i=1,\ldots,c} l^i(\vec{x}) = \min_{i=1,\ldots,c} \sum_{j=1}^{c} \lambda(\omega_i \mid \omega_j) P(\omega_j \mid \vec{x})$$

- i střední ztráta (Bayesovské riziko): $L = \int_{\Omega} l^*(\vec{x})\, p(\vec{x})\, d\vec{x}$

Bayesovo rozhodovací pravidlo (6)

Pro ztráty: $\lambda(\omega_i|\omega_i) = 0$ a $\lambda(\omega_i|\omega_j) = 1$ pro $i \neq j$

- je podmíněná Bayesovská ztráta:

$$e^*(\vec{x}) = 1 - \max_{j=1,\ldots,c} P(\omega_j \mid \vec{x})$$

- a Bayesovské riziko: $E^* = \int_{\Omega} e^*(\vec{x})\, p(\vec{x})\, d\vec{x}$

Bayesovské rozhodovací pravidlo :

$$\hat{\omega}^*(\vec{x}) = \omega_i \text{ , jestliže } P(\omega_i|\vec{x}) = \max_{j=1,\ldots,c} P(\omega_j|\vec{x})$$

Bayesovo rozhodovací pravidlo (7)

Odhad podmíněných rozdělení

- proces odhadu parametrů na základě trénovací množiny ~ **UČENÍ**
- KRITÉRIUM MINIMÁLNÍ CHYBY:
  - Diskriminační funkce:

$$g_r(\vec{x}) = \hat{p}(\vec{x} \mid \omega_r)\, \hat{P}(\omega_r); \quad r = 1, \ldots, c$$

- Trénovací množina $\boldsymbol{T}$: $T = \{[\vec{x}_1, \Omega_1], \ldots, [\vec{x}_R, \Omega_R]\}$
- Konstrukce odhadů $\hat{p}(\vec{x} \mid \omega_r)$

Bayesovo rozhodovací pravidlo (8)

Vlastnosti odhadů:

- Nestrannost: záruka, že v průměru se bude odhad pohybovat kolem neznámé hustoty $p(\vec{x})$
- Konzistence: záruka, že čím větší bude počet vzorů množiny $T_r \subseteq T$, tím více se bude odhad $\hat{p}(\vec{x})$ blížit k neznámé hustotě $p(\vec{x})$ rozložení vzorů z třídy ω_r
- Eficience: takový nestranný odhad, který mezi všemi nestrannými odhady má nejmenší disperzi

Bayesovo rozhodovací pravidlo (9)

Vlastnosti odhadů (pokračování):

- Induktivnost (→ cíl učení): požadavek nalézt po předložení spočetně mnoha dvojic $[\vec{x}_k, \Omega_k]$ parametr $\vec{q}^*$, který minimalizuje střední ztrátu přes celou množinu X; $\vec{x} \in X$
- Sekvenčnost (→ postup učení): odhad na základě postupného předkládání dvojic $[\vec{x}_k, \Omega_k]$

  → $\exists \{f_k\}_{k=1}^{\infty}$ posloupnost funkcí ;

$$\hat{p}^{(k+1)}(\vec{x}) = f_{k+1}\left(\hat{p}^{(k)}(\vec{x}), \vec{x}(k+1)\right) \quad ; \quad k = 1, \ldots$$

$\hat{p}^{(k)}(\vec{x})$ … ***k*** –tý odhad hustoty $p(\vec{x})$

Bayesovo rozhodovací pravidlo (10)

Obecný algoritmus odhadu parametrů učením:

START: Zjisti c, počáteční odhad $\hat{p}^{(0)}(\vec{x}|\omega_r)$, anuluj $\hat{P}(\omega_r)$, $k := 0$;

ITERACE: $k := k + 1$

Přečti $\vec{x}$ a $\Omega = \omega_r$ z T ;

Zkoriguj $\hat{p}(\vec{x}|\omega_r)$; $\hat{P}(\omega_r) := \hat{P}(\omega_r) + 1$;

PODMÍNKA: Je na vstupu další dvojice z T ?

ANO: Proveď iteraci

NE: STOP

STOP: $\hat{P}(\omega_r) := \hat{P}(\omega_r)/K$; $r = 1,\ldots,c$

Vypiš $\hat{p}(\vec{x}|\omega_r)$; $\hat{P}(\omega_r)$; $r = 1,\ldots,c$

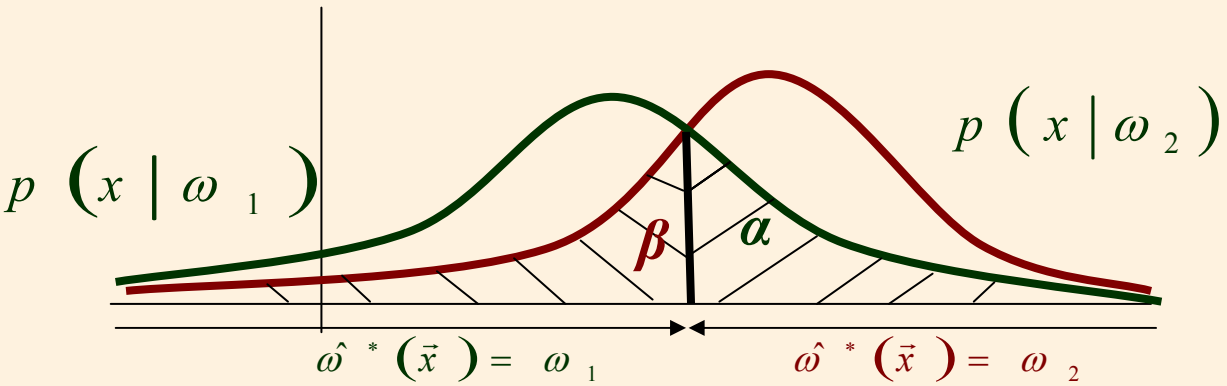
Bayesovo rozhodovací pravidlo (11)

Odhad apriorní pravděpodobnosti P_r :

$$\hat{P}(\omega_r) = \frac{K_r}{K}$$

K_r … počet vzorů ve třídě r

K … celkový počet vzorů z T



Střední ztráta je úměrná plochám α, β

Rozhodovací pravidlo nejbližšího souseda

- Dána množina $S_n = \{(\vec{x}_1, \vartheta_1), \ldots, (\vec{x}_n, \vartheta_n)\}$ nezávislých náhodných veličin

  ϑ … některé z c tříd $\omega_1, \ldots, \omega_c$

  ϑ_i … třída, ke které vzor $\vec{x}_i$ skutečně patří

- Klasifikace nového vzoru $\vec{x} \notin S_n$:
  - Určit jeho nejbližšího souseda $\vec{x}' \in S_n$
  - Zařadit $\vec{x}$ do třídy ϑ' odpovídající $\vec{x}'$

Rozhodovací pravidlo nejbližšího souseda (2)

- Odhad $\hat{\omega}$ neznámé třídy ω vzoru $\vec{x}$:

  $\hat{\omega} = \vartheta'$, jestliže $\delta(\vec{x}, \vec{x}') = \min_{i=1,\ldots,n} \delta(\vec{x}, \vec{x}'_n)$

  δ … metrika v příznakovém prostoru

- *1*-NNR, *k*-NNR

Konvergence nejbližšího souseda $\vec{x}'_n$ k $\vec{x}$

- $|S_n|$ dost velká, $p(\vec{x})$ spojitéa $p(\vec{x}) > 0 \Rightarrow \vec{x}'_n \xrightarrow{n} \vec{x}$ podle pravděpodobnosti
- $p(\vec{x} \mid \omega_i)$ spojité $\forall i \Rightarrow P(\omega \mid \vec{x}'_n) \xrightarrow{n} P(\omega \mid \vec{x}'_n)$ podle pravděpodobnosti

Rozhodovací pravidlo nejbližšího souseda (3)

Dále předpoklad:

Podmínky konvergence splněny, $\vec{x}, \vec{x}'_1, \ldots, \vec{x}'_n$ nezávislé

Libovolně velká S_n $(n \to \infty)$, $P(\omega_i \mid \vec{x}) \approx \eta_i(\vec{x})$

Podmíněná pravděpodobnost chyby *1*-NNR:

$$e_1(\vec{x} \mid \vec{x}'_n) = \sum_i P\{\omega = \omega_i, \vartheta' \neq \omega_i \mid \vec{x}, \vec{x}'_n\} =$$

$$= \sum_i P\{\omega = \omega_i \mid \vec{x}\} \cdot P\{\vartheta' \neq \omega_i \mid \vec{x}'_n\} =$$

$$= \sum_i \eta_i(\vec{x}) \cdot [1 - \eta_i(\vec{x}'_n)]$$

Rozhodovací pravidlo nejbližšího souseda (4)

$$n \to \infty : \quad e_1(\vec{x}) = \lim_{n\to\infty} e_1(\vec{x}, \vec{x}'_n) = 1 - \sum_i \eta_i(\vec{x})^2$$

Míra chyby ***1***-NNR, E_1: $E_1 = \lim_{n\to\infty} E\{e_1(\vec{x}, \vec{x}'_n)\}$

$e_1(\vec{x}, \vec{x}'_n)$ omezena: $E_1 = E\left\{\lim_{n\to\infty} e_1(\vec{x}, \vec{x}'_n)\right\}$

$$E_1 = \int \sum_{i<j}^{c} 2\,\eta_i(\vec{x})\eta_j(\vec{x})\,p(\vec{x})\,d\vec{x}$$

Rozhodovací pravidlo nejbližšího souseda (5)

Porovnání míry chyby ***k***-NNR, E_k, s Bayesovskou, E^*:

- Je-li poslední nejbližší soused sudého řádu, nepřispívá žádnou dodatečnou informací ke klasifikaci: $E_{2k'-1} = E_{2k'}$
- $k \to \infty$; $E^* \leq E_{k-1}$
- Posloupnost horních mezí pro E^*:

  $E^* \leq \ldots \leq E_{2k'+2} = E_{2k'+1} \leq E_{2k'} = E_{2k'-1} \leq \ldots \leq E_2 = E_1 \leq 2E^*$
- Dolní mez pro E^*: $E_k \leq E^* + E_1 / \sqrt{k\,\pi}$

  => konvergence E_k k E^*

Rozhodovací pravidlo nejbližšího souseda (6)

Redukce počtu vzorů v S_n :

- Výběrem reprezentativní podmnožiny
- Klasifikace pomocí redukované množiny je s pravděpodobností blízkou 1 stejně dobrá jako pomocí celé množiny

Rozhodovací pravidlo nejbližšího souseda (7)

Editační algoritmus:

2 třídy, k liché, S_n rozdělena do *2* nezávislých podmnožin $S_{n'}$ a $S_{n''}$; $n' + n'' = n$ a $0 << n'/n'' << \infty$

Krok 1: Klasifikace vzorů z $S_{n'}$ pomocí k –NNR a trénovací množiny $S_{n''}$

Krok 2: Vyřazení všech vzorů z $S_{n'}$, které byly v Kroku 1 klasifikovány chybně. Nově vzniklá podmnožina bude označena $S_{n'}$

Následná klasifikace pomocí 1 –NNR a $S_{n'}$ (kvasi Bayesovsky optimální pro malou pravděpodobnost chyby)

Rozhodovací pravidlo nejbližšího souseda (8)

Opakované editování:

Krok 1: Difuze ~ náhodně rozděl S do N podmnožin $S_1,\ldots, S_N$; $N \geq 3$

Krok 2: Klasifikace ~ klasifikuj vzory z S_i pomocí 1–NNR a trénovací množiny $S_{S(i+1) \bmod N}$; $i = 1,\ldots,N$

Krok 3: Editování ~ vyřaď všechny vzory chybně klasifikované v Kroku 2

Krok 4: Konfuze ~ ze zbylých vzorů vytvoř novou S

Rozhodovací pravidlo nejbližšího souseda (9)

Opakované editování (pokračování):

Krok 5: Ukončení

- Pokud během posledních ***I*** iterací nebyl vyřazen žádný vzor → STOP
- Jinak přejdi ke Kroku 1

- Následná klasifikace nových vzorů pomocí ***1***–NNR a výsledné trénovací množiny ***S***
  - Statistická nezávislost vzorů
  - Asymptoticky Bayesovsky optimální

Rozhodovací pravidlo nejbližšího souseda (10)

Eliminace vzorů, které nepřispívají k definici dělicí nadplochy:

~ vybrat co nejmenší podmnožinu trénovací množiny tak, aby 1-NNR pomocí vybrané podmnožiny správně klasifikoval zbylé vzory z trénovací množiny

2 zásobníky: **SKLAD** a **PYTEL**

první vzor nechť je v zásobníku **SKLAD**

n_P ~ aktuální počet vzorů v zásobníku **PYTEL** při každém vstupu algoritmu na Krok 1

Rozhodovací pravidlo nejbližšího souseda (11)

Kondenzační algoritmus:

Krok 1: Klasifikuj ***i***-tý vzor ze zásobníku **PYTEL** pomocí ***1***-NNR a aktuálního zásobníku **SKLAD**

- Je-li vzor klasifikován správně, vrať ho do zásobníku **PYTEL**
- Jinak ho ulož do zásobníku **SKLAD**

Opakuj operace pro $i = 1,\ldots,n_P$

Krok 2: Pokud byl Krok 1 proveden bez přesunu ze zásobníku **PYTEL** do zásobníku **SKLAD** nebo je **PYTEL** prázdný → STOP

Jinak přejdi ke Kroku 1

Dynamické shlukování

Problém:

Rozdělení trénovací množiny $Y = \{\vec{y}_i; i = 1, \ldots, n\}$ do c shluků Γ_k ; $k = 1, \ldots, c$ reprezentovaných střední hodnotou $\vec{m}_k$

n_k ~ počet prvků k-tého shluku

$$\vec{m}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \vec{y}_i \; ; \quad \vec{y}_i \in \Gamma_k$$

- míra podobnosti mezi vzorem $\vec{y} \in \Gamma_k$ a představitelem shluku Γ_k :

$$\delta_E(\vec{y}, \vec{m}_k) < \delta_E(\vec{y}, \vec{m}_l) \quad \forall\, l \neq k$$

Dynamické shlukování (2)

Cíl: Nalézt mezi všemi možnými rozděleními

$$\Gamma = \left\{ \Gamma_j ;\ j = 1, \ldots , c ; \bigcap_{j=1}^{c} \Gamma_j = \{\ \}; \bigcup_{j=1}^{c} \Gamma_j = Y \right\}$$

množiny ***Y*** do ***c*** shluků takové, které minimalizuje funkci shlukovacího kritéria ***J(Γ)***

- $\Gamma^* = \{Y_j ;\ j = 1, \ldots , c\}$ optimální $\Leftrightarrow$ $J(\Gamma^*) = \min_{\Gamma} J(\Gamma)$
- $J(\Gamma) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \delta_E^2 (\vec{y}_i , \vec{m}_j) = \sum_{j=1}^{c} J(\Gamma_j); \quad \vec{y}_i \in \Gamma_j$

Dynamické shlukování (3)

- Přemístění vzoru $\vec{y}'$ ze shluku Γ_k do Γ_j:
  pokles hodnoty kritéria $J(\Gamma)$ <=>

$$\frac{n_j}{n_j + 1} \delta_E^2 \left(\vec{y}', \vec{m}_j \right) < \frac{n_k}{n_k - 1} \delta_E^2 \left(\vec{y}', \vec{m}_k \right)$$

~ je-li $\vec{y}'$ blíž k $\vec{m}_j$ než k $\vec{m}_k$

Dynamické shlukování (4)

c-průměrová metoda (k-means algoritmus):

Inicializace:

- Trénovací množinu Y libovolně rozděl do c shluků Γ_j ; $1 \le j \le c$
- Urči střední hodnoty $\vec{m}_j$ příslušných shluků Γ_j ; $1 \le j \le c$

Krok 1: $\forall \vec{y} \in Y$; zařaď $\vec{y}$ do Γ_j, jestliže

$$\delta_E(\vec{y}, \vec{m}_j) = \min_k \delta_E(\vec{y}, \vec{m}_k)$$

Krok 2: Aktualizuj $\vec{m}_j$; $1 \le j \le c$

nastala změna středních hodnot?

ANO: přejdi na Krok 1

NE: STOP

Dynamické shlukování (5)

Informace o struktuře shluku:

shluk Γ_j reprezentován jádrem $K_j = (\vec{y}, V_j)$

$V_j \sim$ množina parametrů definujících K_j

$\Delta(\vec{y}, K_j) \sim$ míra podobnosti vektoru $\vec{y}$ a shluku Γ_j (reprezentovaného jádrem K_j)

=> KRITÉRIUM: $J(\Gamma) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \Delta(\vec{y}_i, K_j)$

Dynamické shlukování (6)

Dynamický shlukovací algoritmus:

Inicializace:

- Trénovací množinu Y libovolně rozděl do c shluků Γ_j ; $1 \le j \le c$
- Urči jádra K_j příslušných shluků Γ_j ; $1 \le j \le c$

Krok 1: $\forall \vec{y} \in Y$; zařaď $\vec{y}$ do Γ_j, jestliže

$$\Delta\left(\vec{y}, K_j\right) = \min_k \Delta\left(\vec{y}, K_k\right)$$

Krok 2: Aktualizuj K_j ; $1 \le j \le c$

došlo ke změně K_j ?

ANO: přejdi ke Kroku 1

NE: STOP

Dynamické shlukování (7)

Dynamický shlukovací algoritmus (pokračování):

Konvergence algoritmu za podmínky:

$J(\Gamma', K') \leq J(\Gamma, K')$, jestliže $J(\Gamma, K') \leq J(\Gamma, K)$

- $J(\Gamma, K)$ … hodnota funkce kritéria $J(\Gamma)$ odpovídající množině jader $K = \{ K_j ; j = 1,\ldots,c \}$
- Γ , Γ' … rozdělení trénovací množiny Y získané při klasifikaci prvků z Y pomocí množiny jader K , resp. K'
- K' … množina aktualizovaných jader $K' = \{ K_j' ; j=1,\ldots,c \}$

Dynamické shlukování (8)

Dynamické shlukovací metody jsou obecně výpočetně velmi výkonné a atraktivní pro uživatele

- Nedostatky: zvolený model jen zřídka odpovídá skutečné stochastické struktuře dat
  → nežádoucí shlukování

  → analýza dat pomocí co největšího počtu libovolně zvolených jader a následná volba optimálního řešení úlohy

Metody shlukové analýzy jsou vhodné především jako prostředek poskytující nezávislou informaci o struktuře dat

Hierarchické shlukování

- Na každém stupni shlukovacího procesu splynutí dvou navzájem si nejpodobnějších shluků
- Na začátku je každý vzor z trénovací množiny považován za samostatný shluk
- Na každém dalším stupni vytvoří dva navzájem si nejpodobnější shluky jeden shluk
- Proces plývání pokračuje i v následných stupních shlukové analýzy – přitom se v každém stupni počet shluků sníží o 1
- Shlukovací procedura končí, když jsou všechny vzory zařazeny do jednoho shluku

Hierarchické shlukování (2)

- Míra podobnosti $\Delta(\Gamma_i, \Gamma_j)$ umožňuje vyjádřit vzájemný vztah shluků
- Nejvíc užívané míry podobnosti pro hierarchické shlukování jsou definovány vzdálenostmi mezi body v obrazovém prostoru:
  - Nejbližší soused: $\Delta(\Gamma_i, \Gamma_j) = \min_{\vec{y} \in \Gamma_i, \vec{y}' \in \Gamma_j} \delta(\vec{y}, \vec{y}')$
  - Nejvzdálenější soused: $\Delta(\Gamma_i, \Gamma_j) = \max_{\vec{y} \in \Gamma_i, \vec{y}' \in \Gamma_j} \delta(\vec{y}, \vec{y}')$
  - Střed: $\Delta(\Gamma_i, \Gamma_j) = \delta(\vec{m}_i, \vec{m}_j)$

$\delta(\vec{y}, \vec{v})$ je nějaká metrika

Hierarchické shlukování (3)

Hierarchický shlukovací algoritmus:

Inicializace:

- Polož $\Gamma_j = \vec{y}_j$, $\forall \vec{y} \in I$, kde $I = \{j; j = 1, \ldots, n\}$

Krok 1: Najdi mezi $\{\Gamma_j; j \in I\}$ takovou dvojici shluků Γ_i, Γ_k, aby:

$$\Delta(\Gamma_i, \Gamma_k) = \min_{\forall j,l \in I} \Delta(\Gamma_j, \Gamma_l)$$

Krok 2: Spoj Γ_i s Γ_k a vymaž Γ_i.

Krok 3: Odstraň i z množiny indexů I.
Je-li kardinalita I dvě, ukonči algoritmus.
V opačném případě přejdi ke Kroku 2.

Hierarchické shlukování (4)

Dendrogram algoritmus:

- Grafické znázornění hierarchické struktury shluků vytvořených hierarchickým shlukovacím algoritmem (ilustruje posloupnost splývání shluků a odpovídající hodnoty míry podobnosti)
- Pro libovolný práh míry podobnosti dostáváme shluky, jejichž podshluky mají míru podobnosti rovnu přinejmenším té prahové
- **Volba prahové hodnoty:** tak, aby rozptyl mezi shluky byl podstatně větší než rozptyl uvnitř shluků

Hierarchické shlukování (5)

Výsledky shlukové analýzy:

- **Vliv mnoha faktorů:**
  - Míra podobnosti:
    - **Nejbližší soused:**
      - spojování dvou různých kompaktních shluků
    - **Nejvzdálenější soused:**
      - citlivý na body vzdálené od centra shluku
      → neschopnost detekovat protáhlé shluky
  - Měřítko na osách, použitá metrika, shlukovací kritérium, počet vzorů v analyzované množině, …

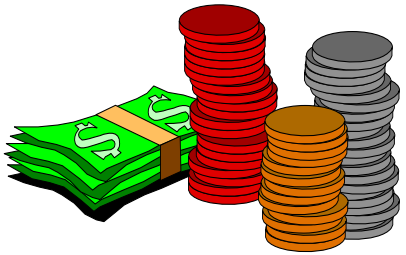
Automatická detekce shluků

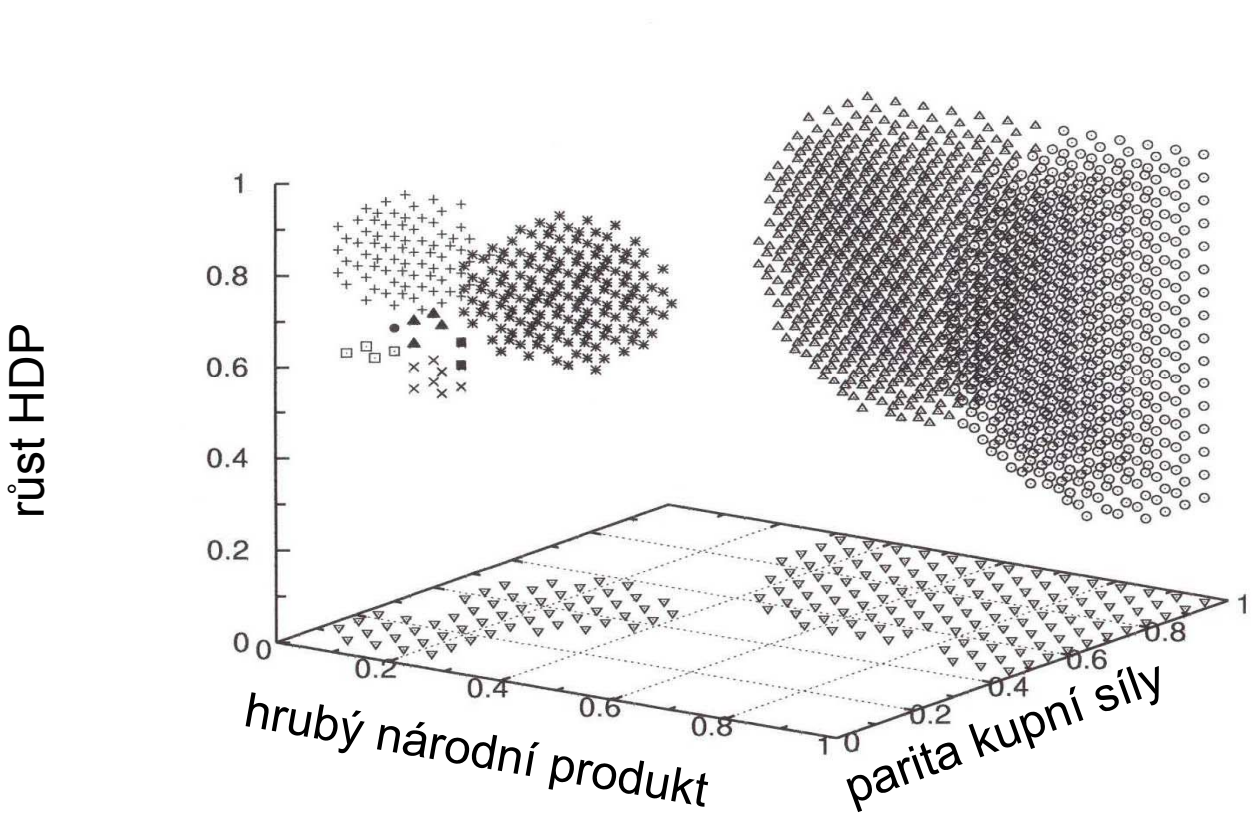
- Cíl:

Nalézt předem neznámé podobnosti v datech!

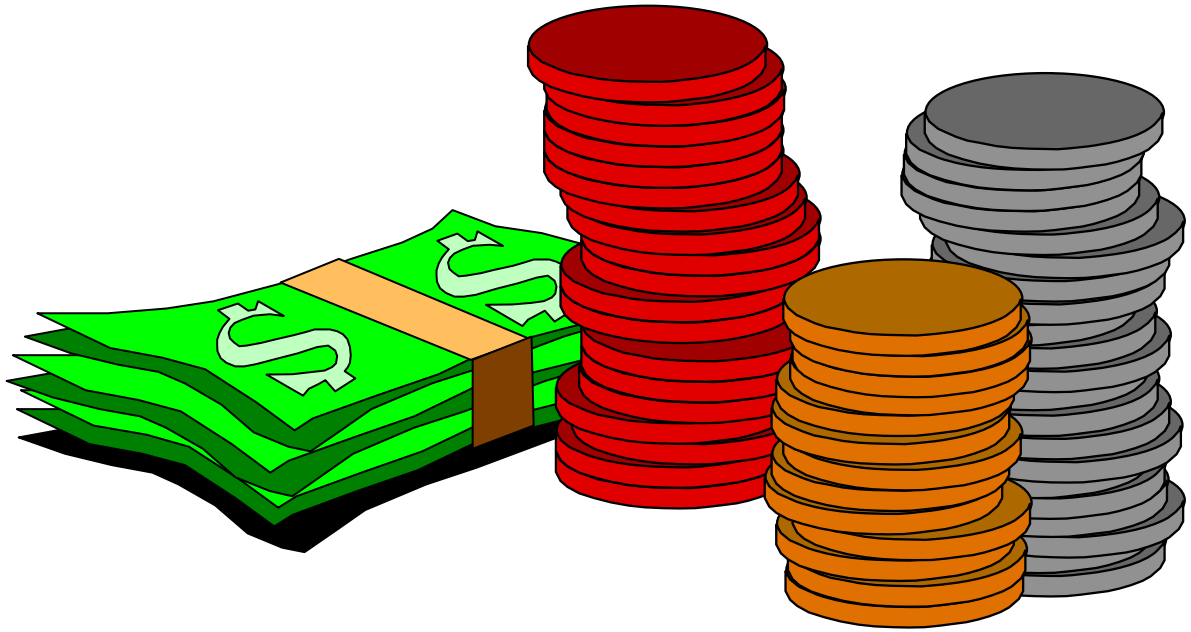
- Vytvořit modely rozpoznávající navzájem podobné vzory
- Vhodné pro počáteční analýzu dat
- Samoorganizace - “učení bez učitele”

Ekonomiky seskupené podle svých výsledků

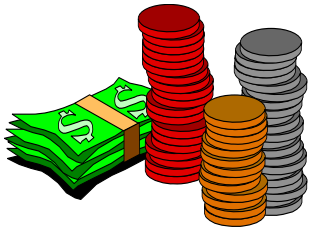




Analýza dat ze Světové banky:
klastrování podle fuzzy c-středů (fuzzy c-means clustering)

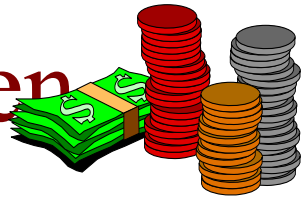


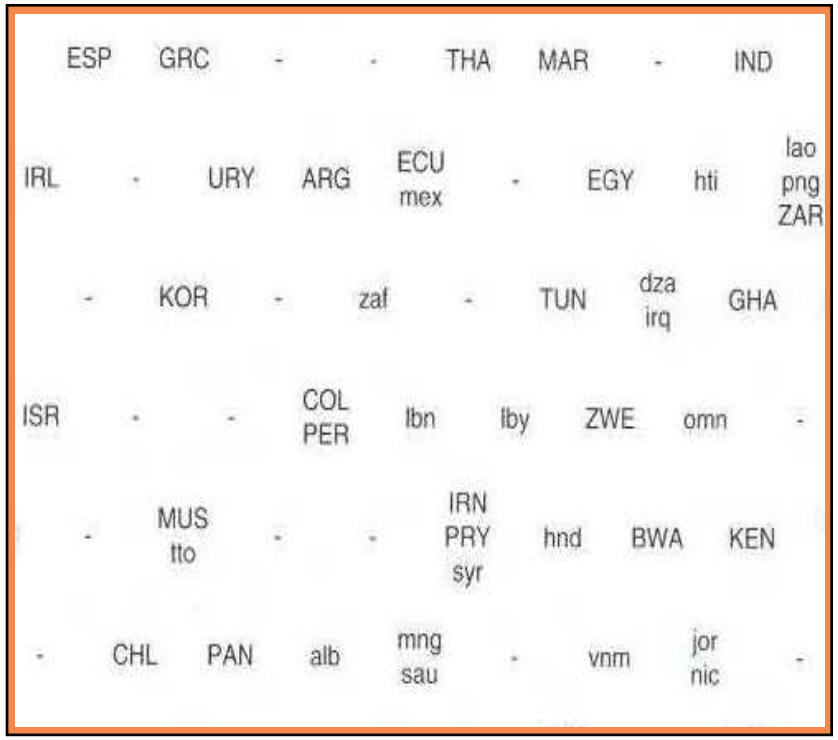
FCM-klastrování: úvod



- **WDI-indikátory (indikátory vývoje ve světě)**
  - každoročně zveřejňovány Světovou bankou
  - odrážejí stav vývoje a pokroku v jednotlivých zemích
  - neúplné a nepřesné údaje
- **(dříve) používané techniky**
  - regresní analýza - lineární závislosti
  - kategorizace států používaná v rozvinutých zemích(G. Ip, Wall Street Journal)
  - kategorizace zemí podle HDP (Světová banka)
  - Kohonenovy mapy (T. Kohonen, S. Kaski, G. Deboeck)

Mapa “chudoby” - T. Kohonen

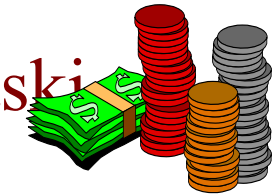


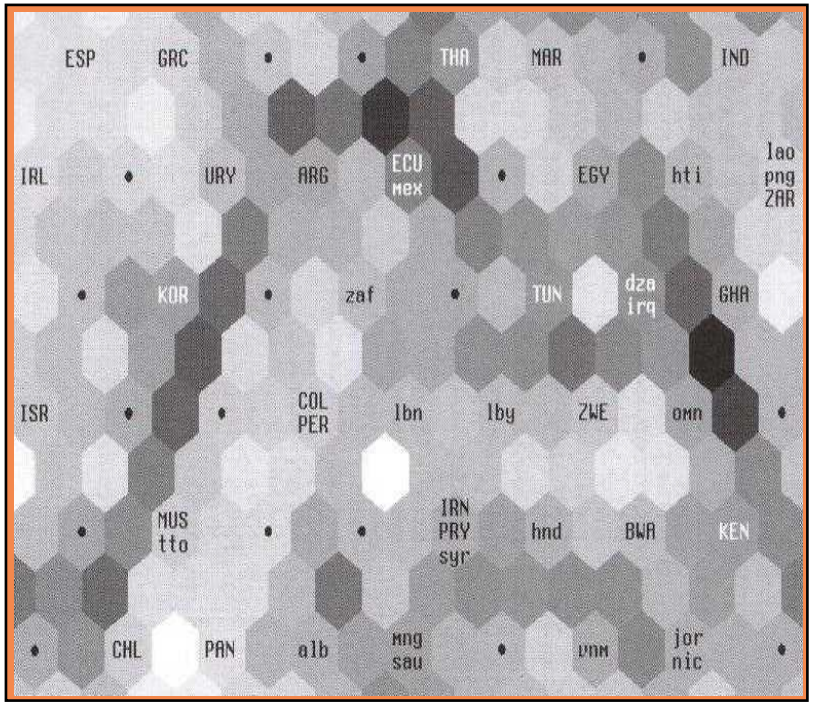


převzato z “T. Kohonen: *Self-Organizing Maps*, 3-rd Edition, Springer-Verlag, 2001”

- více neuronů než zemí
- význam mají jen lokální geometrické vztahy
- státům zobrazeným blízko u sebe odpovídá podobný stav vývoje

Mapa “chudoby” - T. Kohonen, S. Kaski



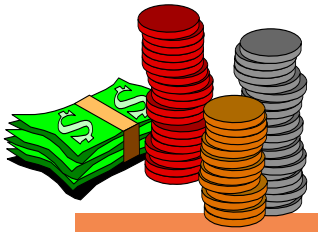


převzato z “T. Kohonen: *Self-Organizing Maps*, 3-rd Edition, Springer-Verlag, 2001”

U-matice:

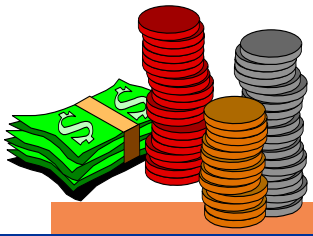
- znázornění “hranic” mezi shluky
- vyjadřuje průměrnou vzdálenost mezi sousedními neurony pomocí různých úrovní šedi
  - malá průměrná vzdálenost
    $\Rightarrow$ ***světlý odstín***
  - velká prům. vzdálenost
    $\Rightarrow$ ***tmavý odstín***

Náš cíl

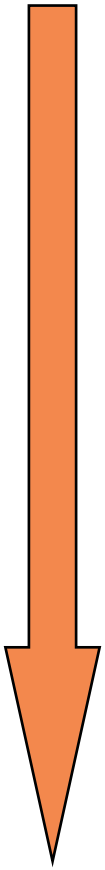


- Efektivně klastrovat nepřesná data
- Odhadnout počet shluků
- Vizualizovat výsledky
- Interpretovat výsledky

Náš cíl

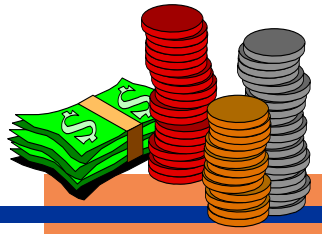


- Efektivně klastrovat nepřesná data
- Odhadnout počet shluků
- Vizualizovat výsledky
- Interpretovat výsledky



- **fuzzy *c* - means shlukování (FCM)**
- **indikátory pro validitu klastrů**
- **formou spread-sheetu**
- **nalezení “landmarks”**

Cílová funkce



- Odpovídá **vážené vzdálenosti** mezi vstupními vzory a centry shluků:

$$J_s(\mathbf{U},\mathbf{v}) = \sum_{p=1}^{P}\sum_{i=1}^{c}(u_{ip})^s\left[\sum_{j=1}^{n}(x_{pj} - v_{ij})^2\right]$$

střed shluku

fuzzyfikační parametr

stupeň členství

vstupní vzor

- stupně členství mezi 0 a 1: $0 \le u_{ip} \le 1$
- celkové členství vzoru odpovídá 1: $\forall p \left| \sum_{i=1}^{c} u_{ip} = 1 \right.$
- shluky nejsou prázdné ani plné: $\forall i \left| 0 < \sum_{p=1}^{P} u_{ip} < P \right.$

Fuzzy *c*-means shlukování (FCM)

- **Krok 1:** Inicializace c, s, ε a t. Zvolte náhodně $U^{(0)}$.
- **Krok 2:** Určete nová centra fuzzy shluků:

$$\vec{v}_i^{(t)} = \frac{1}{\sum_p (u_{ip}^{(t)})^s} \sum_p (u_{ip}^{(t)})^s \vec{x}_p$$

- **Krok 3:** Vypočítejte novou členskou matici $U^{(t+1)}$:

$$u_{ip}^{(t+1)} = \frac{(1 / \| \vec{x}_p - \vec{v}_i^{(t)} \|^2)^{1/s-1}}{\sum_{k=1}^{c} (1 / \| \vec{x}_p - \vec{v}_k^{(t)} \|^2)^{1/s-1}}$$

- **Krok 4:** Vyhodnoťte $\Delta = \| U^{(t+1)} - U^{(t)} \| = \max_{i,p} | u_{ip}^{(t+1)} - u_{ip}^{(t)} |$

  Jestliže $\Delta > \varepsilon$ potom nahraďte $t = t + 1$ a přejděte ke **Kroku 2**.
  Jestliže $\Delta \leq \varepsilon$ potom **Stop**.

- **KONEC FCM**

Kritéria pro validitu klastrů

- **Koeficient členské funkce:**

$$F\ (U\ ;c\ )\ =\ \frac{1}{P}\sum_{p\ =1}^{P}\ \sum_{i\ =1}^{c}\ (\ u_{\ ip}\ )^{\ 2}$$

shluky

stupeň členství

vzory

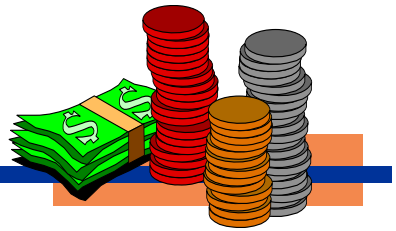
- **Entropie členské funkce:**

$$H\ (U\ ;c\ )\ =\ -\ \frac{1}{P}\sum_{p=1}^{P}\ \sum_{i=1}^{c}\ u_{\ ip}\ \ln(\ u_{\ ip}\ )\ ;\ \ u_{ip}\ln(u_{ip})=0\ \ \text{for}\ u_{ip}=0.$$

- **Windhamův proporční exponent:**

$$W(U;c)=-\sum_{p=1}^{P}\ln\left[\sum_{j=1}^{\lfloor\mu_p^{-1}\rfloor}(-1)^{j+1}\binom{c}{j}\left(1-\ j\cdot\mu_{\ p}\right)^{\ c-1}\right];\ \ \mu_p=\max_{1\le i\le c}\left\{u_{ip}\right\}$$

Kolik shluků?



- **Koeficient členské funkce :**

$$\max_{2 \le c \le P-1} \left\{ \max_{U} \left[F(U;c) \right] \right\}$$

partitions

clusters

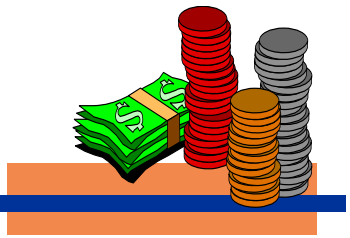
- **Entropie členské funkce:**

$$\min_{2 \le c \le P-1} \left\{ \min_{U} \left[H(U;c) \right] \right\}$$

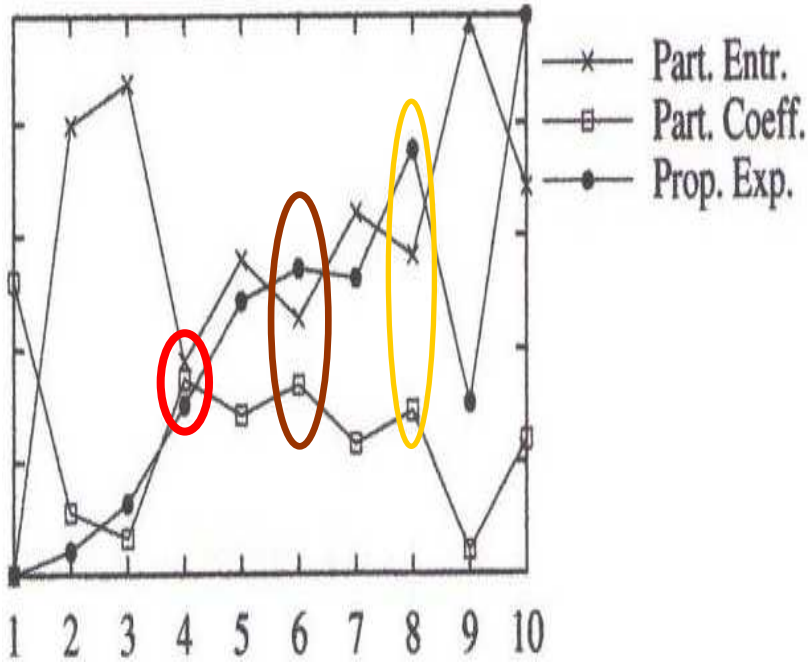
- **Windhamův proporční exponent:**

$$\max_{2 \le c \le P-1} \left\{ \max_{U} \left[W(U;c) \right] \right\}$$

Experimenty - “umělá” data

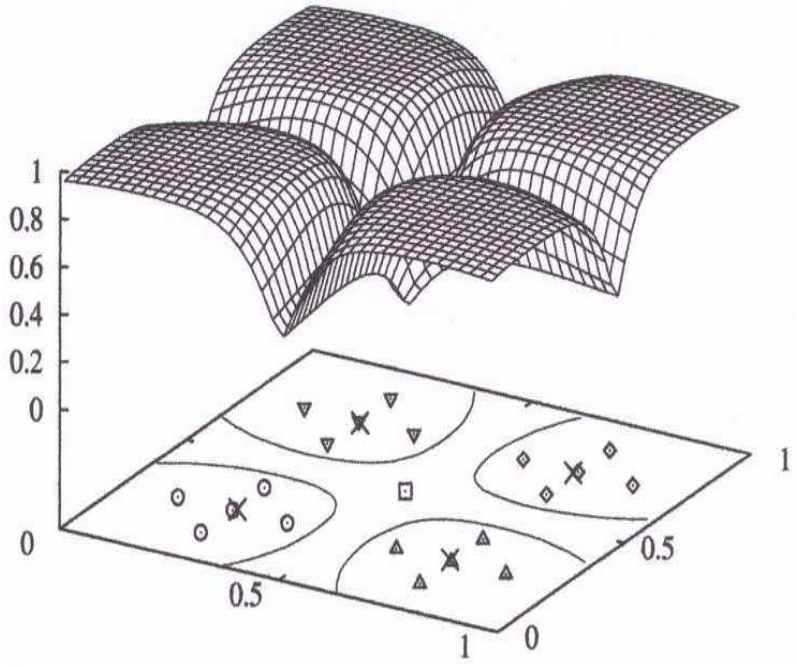


Indikátory validity shluků pro “umělá” data



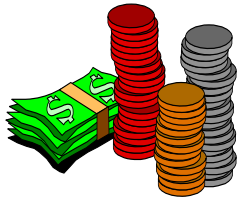
21 vstupních vzorů, $s = 1.4$, $\varepsilon = 0.05$

Fuzzy 4-rozčlenění vzorů

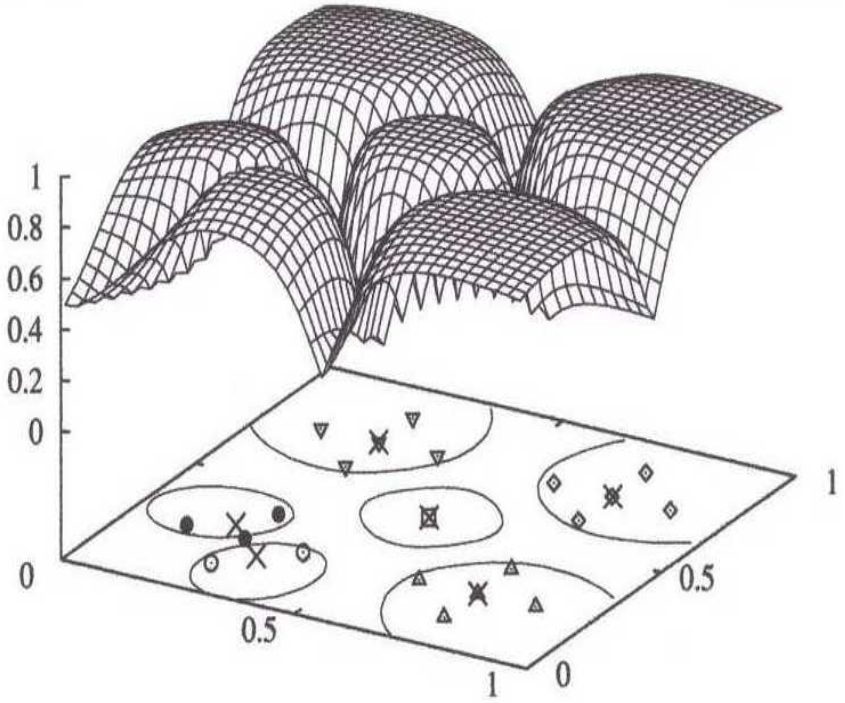


´×´ odpovídá centerům shluků, vzory ze stejných shluků jsou označeny stejně

Experimenty - “umělá” data

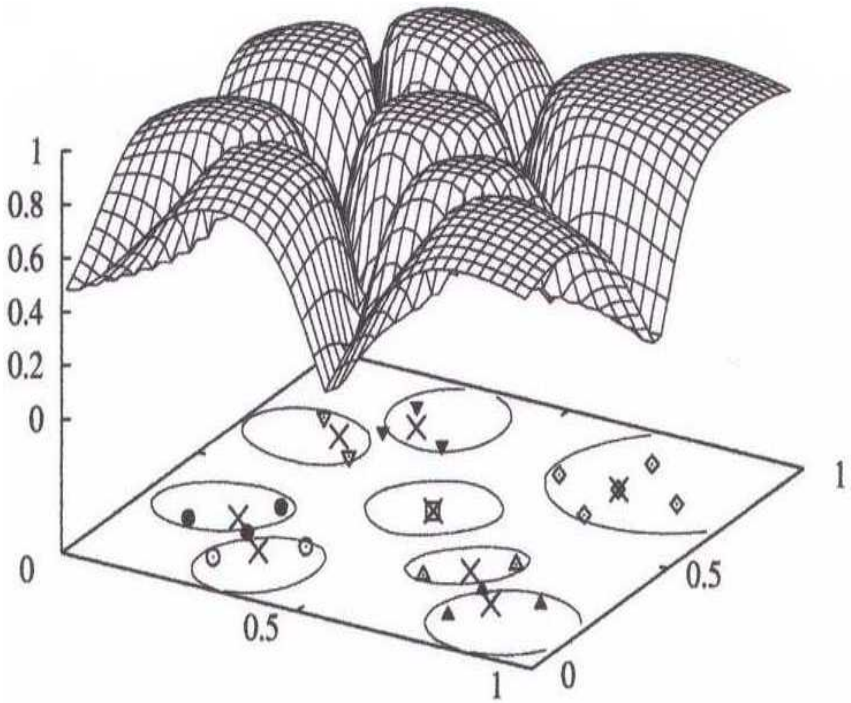


Fuzzy 6-rozčlenění vzorů



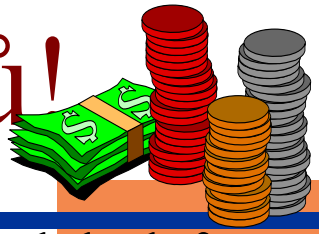
´×´ odpovídá centerům shluků, vzory ze stejných shluků jsou označeny stejně

Fuzzy 8-rozčlenění vzorů



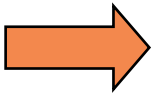
´×´ odpovídá centerům shluků, vzory ze stejných shluků jsou označeny stejně

Interpretace výsledků!



Characteristické vlastnosti nalezených shluků:

- centra shluků - *"fiktivní" vzory mimo předkládaná data*
- "kalibrace" shluků "nejreprezentativnějšími" vzory z trénovací množiny - *charakterizace podle jediného vzoru*
- **Určit význačné vlastnosti shluků:**
  - vzhledem k dalším vlastnostem příslušného shluku
  - vzhledem k charakteristickým vlastnostem ostatních shluků
  - výjimka: "oblasti u hranic"



fuzzy *c*-landmarks

Automatický výběr "landmarks"

Fuzzy c-landmark pro shluk i^*: $(j^*, v_{i^* j^*})$

- "fuzzy vzdálenost" od centra shluku by měla být malá
- "fuzzy vzdálenost" od všech ostatních center shluků by měla být velká

$$j^* = \arg \min_{1 \le j \le n} \frac{\sum_{p=1}^{P} u_{i^* p}^{s} \left| x_{pj} - v_{i^* j} \right|}{\sum_{p=1}^{P} u_{i^* p}^{s}} \Big/ \min_{\substack{1 \le i \le c \\ i \ne i^*}} \frac{\sum_{p=1}^{P} u_{i^* p}^{s} \left| x_{pj} - v_{ij} \right|}{\sum_{p=1}^{P} u_{i^* p}^{s}}$$

vstupní vzory

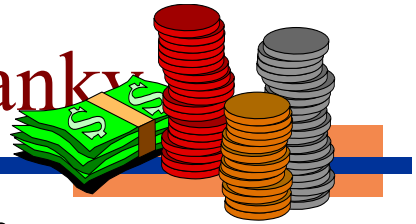
centra shluků

vstupy

shluky

stupeň členství

Experimenty: data ze Světové banky



- 99 států se 16 WDI-indikátory pro každý z nich
- economický a sociální potencial zemí a jejich obyvatel
- všechny indikátory jsou relativní vzhledem k velikosti populace
- po složkách transformace vzorů do intervalu (0,1) pomocí:

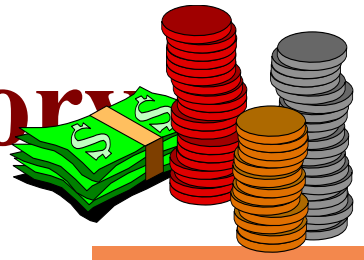
$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad \text{a} \quad x'' = \frac{1}{1 + e^{-k(x' - 1/2)}}$$

minimum přes všechny vzory

maximum přes všechny vzory

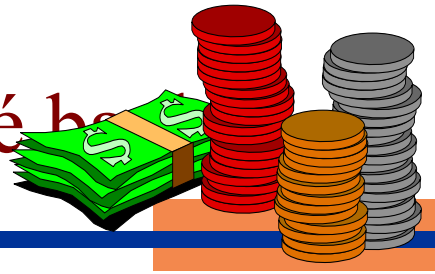
- volba dalších parametrů ($k=4;\ s=1.4;\ \varepsilon=0.05$)

Použité WDI-indikátory

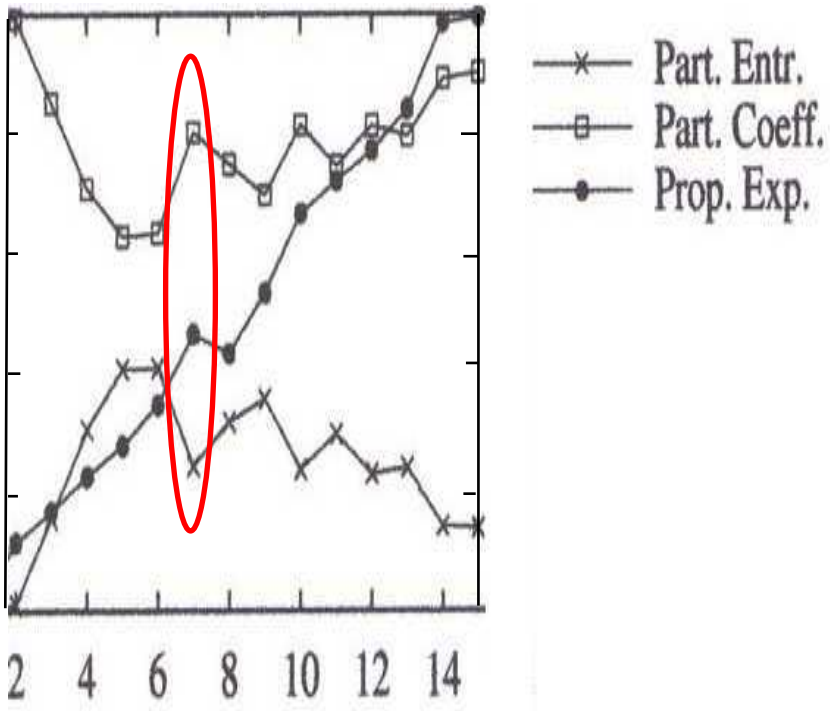


- HNP na obyvatele
- Parita kupní síly
- Růst HDP na obyvatele
- Implicitní deflace HDP
- Vnější zadluženost (% HNP)
- Celkvé náklady na zadlužení (% z exportu zboží a služeb)
- Export high-tech technologií (% z vyvážených výrobků)
- Výdaje na armádu a zbrojení (% HNP)
- Výdaje na výzk. a výv. (% HNP)
- Celk. výd. na zdrav. (% HDP)
- Veř. výd. na školství (% HNP)
- Očekávaná délka života u mužů
- Plodnost
- GINI-index (rozdělení příjmů a spotřeby)
- Uživ. internetu na 10000 obyvatel
- Počet mobilních telefonů na 1000 obyvatel

Experimenty: data ze Světové b

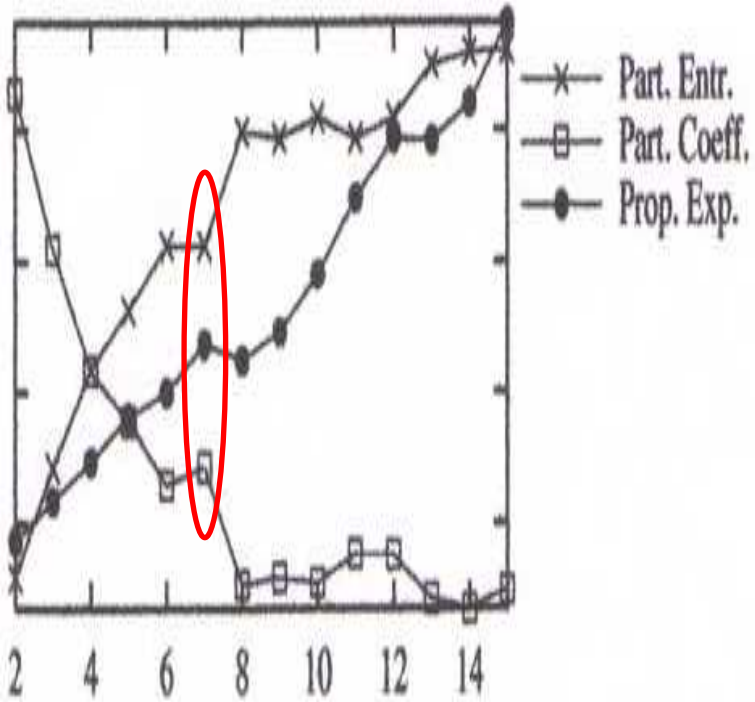


Indikátory validity shluků pro data ze SB



99 zemí se 16 indikátory
$s = 1.1$, $\varepsilon = 0.05$

Indikátory validity shluků pro data ze SB



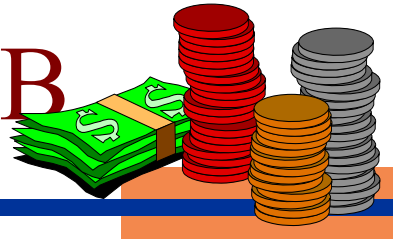
99 zemí se 16 indikátory
$s = 1.4$, $\varepsilon = 0.05$

Fuzzy 7-rozčlenění dat ze SB

26	Egypt A.R.	0.02	0.90	0.01	0.05	0.01	0.00	0.01
27	El Salvador	0.01	0.08	0.01	0.11	0.01	0.00	0.78
28	Estonia	0.04	0.13	0.01	0.06	0.67	0.01	0.09
29	Ethiopia	0.00	0.00	0.97	0.02	0.00	0.00	0.00
30	Finland	0.00	0.00	0.00	0.00	0.02	0.96	0.00
31	France	0.00	0.00	0.00	0.00	0.01	0.99	0.00
32	Georgia	0.96	0.02	0.00	0.01	0.01	0.00	0.01
33	Germany	0.00	0.00	0.00	0.00	0.03	0.97	0.00
34	Ghana	0.00	0.07	0.08	0.82	0.00	0.00	0.02
35	Greece	0.01	0.05	0.00	0.02	0.85	0.04	0.03
36	Guatemala	0.01	0.09	0.18	0.37	0.01	0.00	0.34
37	Guinea	0.00	0.00	0.99	0.01	0.00	0.00	0.00
38	Honduras	0.01	0.03	0.02	0.09	0.01	0.00	0.86
39	Hungary	0.03	0.24	0.01	0.04	0.65	0.01	0.02
40	India	0.01	0.85	0.01	0.11	0.01	0.00	0.02
41	Indonesia	0.06	0.43	0.10	0.20	0.05	0.01	0.16
42	Ireland	0.01	0.02	0.01	0.01	0.13	0.79	0.02
43	Italy	0.00	0.00	0.00	0.00	0.01	0.99	0.00
44	Jamaica	0.07	0.47	0.01	0.14	0.10	0.00	0.22
45	Japan	0.00	0.00	0.00	0.00	0.01	0.98	0.00
46	Jordan	0.09	0.24	0.06	0.26	0.14	0.02	0.20
47	Kazakhstan	0.84	0.10	0.00	0.03	0.01	0.00	0.01
48	Kenya	0.01	0.04	0.19	0.67	0.01	0.00	0.07
49	Korea	0.04	0.09	0.02	0.05	0.38	0.38	0.05

Ukázka fuzzy 7-rozčlenění dat ze Světové banky:
99 zemí se 16 indikátory; $s = 1.4$, $\varepsilon = 0.05$

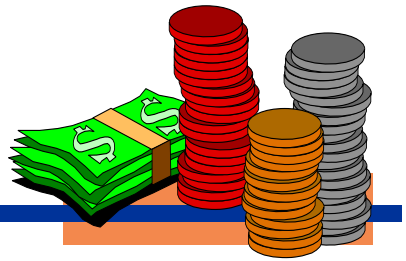
Landmarks pro data ze SB



No.	Representant	1. char. feature	2. char. feature	3. char. feature
1	Uzbekistan	GDP impl. defl. 330 % ann. growth	Hi-Tech exports 4 % of annual exp.	Gini-index 33.90
2	Vietnam	Fertility rate 2.57	Gini-index 36.73	Total exp. on health 4.94 % of GDP
3	Guinea	Internet hosts 0 per 10000 people	PPP per capita 1276 USD	GNP per capita 441.43 USD
4	Ghana	Fertility rate 3.94	Life exp. (males) 57.62 years	Gini-index 42.61
5	Slovenia	PPP per capita 13485 USD	Mobile phones 270 per 1000 people	Expend. on R&D 0.98 % of GNP
6	Netherlands	GDP impl. defl. 2.3 % of ann. growth	Ext. debt 1.1 % of GNP	Tot. debt serv. 0.47 % of export
7	Peru	Gini-index 48.98	GDP growth rate -1.92 % per capita	Life exp. (males) 66.95 years

"Reprezentativní vzory" a fuzzy 7-landmarks pro data ze Světové banky:
99 zemí se 16 indikátory; $s = 1.4$, $\varepsilon = 0.05$

FCM-klastrování: závěr



- FCM-klastrování
  - efektivita a kritéria pro validitu shluků
  - volba fuzzifikačního parametru
  - kategorizace a rozdělení ekonomik států (World Bank, Ip, Kohonen, Deboeck)
- Vizualizace
  - hodnoty členské funkce
  - topologické vztahy
- Landmarks a interpretace výsledků
  - formulace kritérií pro “vymezení tříd”