

# Interpretace a evaluace

## 1. Deskriptivní úlohy

kritériem novost, zajímavost, užitečnost a srozumitelnost

Kvalitativní hodnocení:

- zřejmé znalosti, které jsou ve shodě se „zdravým selským rozumem“
- zřejmé znalosti, které jsou ve shodě se znalostmi experta z dané oblasti
- nové, zajímavé znalosti, které přinášejí nový pohled
- znalosti, které musí expert podrobit bližší analýze, neboť není zcela jasné co znamenají
- „znalosti“, které jsou v rozporu se znalostmi experta

Kvantitativní hodnocení:

- např. spolehlivost a podpora u asociačních pravidel  
pozor, ne vše co je statisticky významné je i zajímavé !

## 2. Klasifikační úlohy

kritériem úspěšnost klasifikace na datech

### **Testování modelů:**

- testování v celých trénovacích datech
- křížová validace (cross-validation)
- leave-one-out
- bootstrap
- testování na testovacích datech

Cílem je zjistit v kolika případech došlo ke shodě resp. neshodě modelu (systému) s informací od učitele

Matice záměn (confusion matrix)

Správné zařazení	Klasifikace systémem	
	+	-
+	TP	FN
-	FP	TN

- Celková správnost resp. celková chyba (overall accuracy a error)

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Err} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

celková správnost  $\in [\text{Acc}_{\text{def}}, \text{Acc}_{\text{max}}]$

kde

$\text{Acc}_{\text{def}}$  ... správnost při klasifikaci všech příkladů do majoritní třídy

$\text{Acc}_{\text{max}}$  ... maximální možná správnost pro daná data

Chyba bez ceny

$$\text{Err} = 1 - \text{Acc}$$

Chyba s cenami

$$\text{Err} = \text{FP} * c(\text{P},\text{n}) + \text{FN} * c(\text{N},\text{p})$$

- Správnost pro jednotlivé třídy

$$\text{Acc}_+ = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Acc}_- = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

vhodnější při „nevyvážených“ třídách

- Přesnost a úplnost (precision a recall)

$$\text{Přesnost} = \frac{TP}{TP + FP}$$

$$\text{Úplnost} = \frac{TP}{TP + FN}$$

Ize kombinovat do tzv. F-míry

$$F = \frac{2 * \text{přesnost} * \text{úplnost}}{\text{přesnost} + \text{úplnost}} = \frac{2TP}{2TP + FP + FN}$$

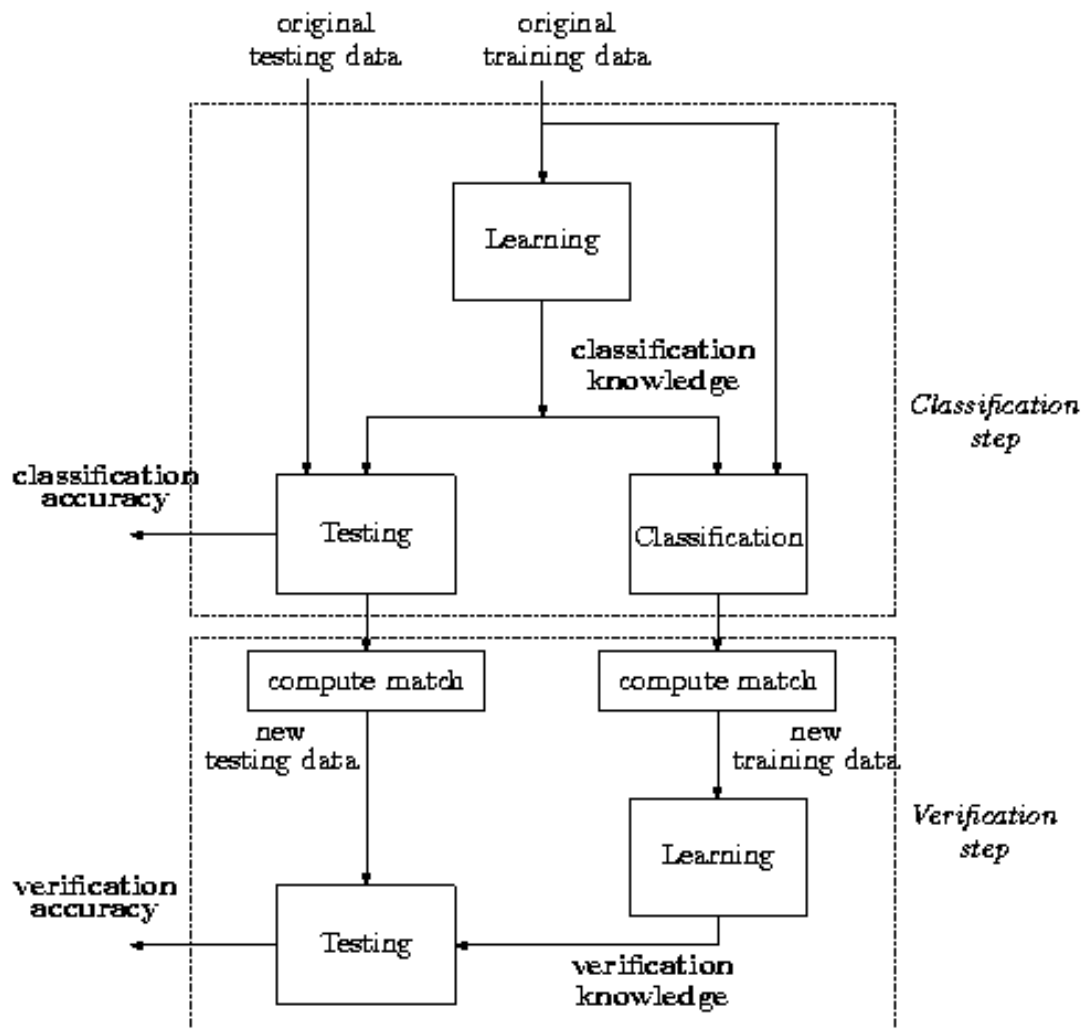
- Senzitivita a specificita (sensitivity a specificity)

$$\text{Senzitivita} = \frac{TP}{TP + FN}$$

$$\text{Specificita} = \frac{TN}{TN + FP}$$

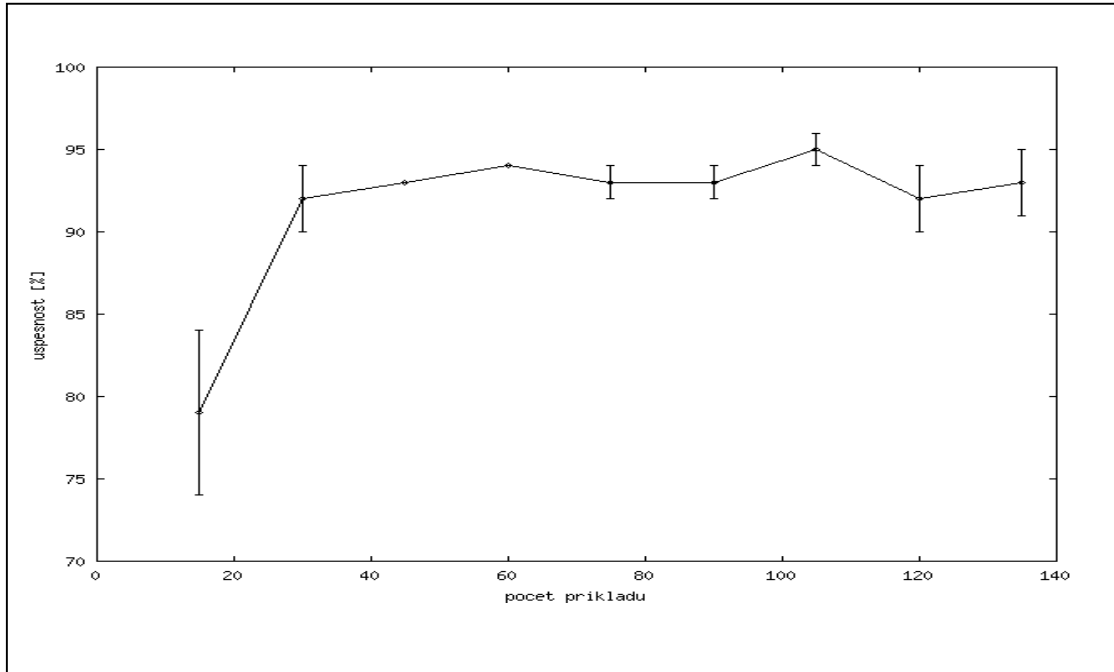
# Spolehlivost klasifikace na základě meta-učení

informace o správnosti klasifikace jako nový atribut

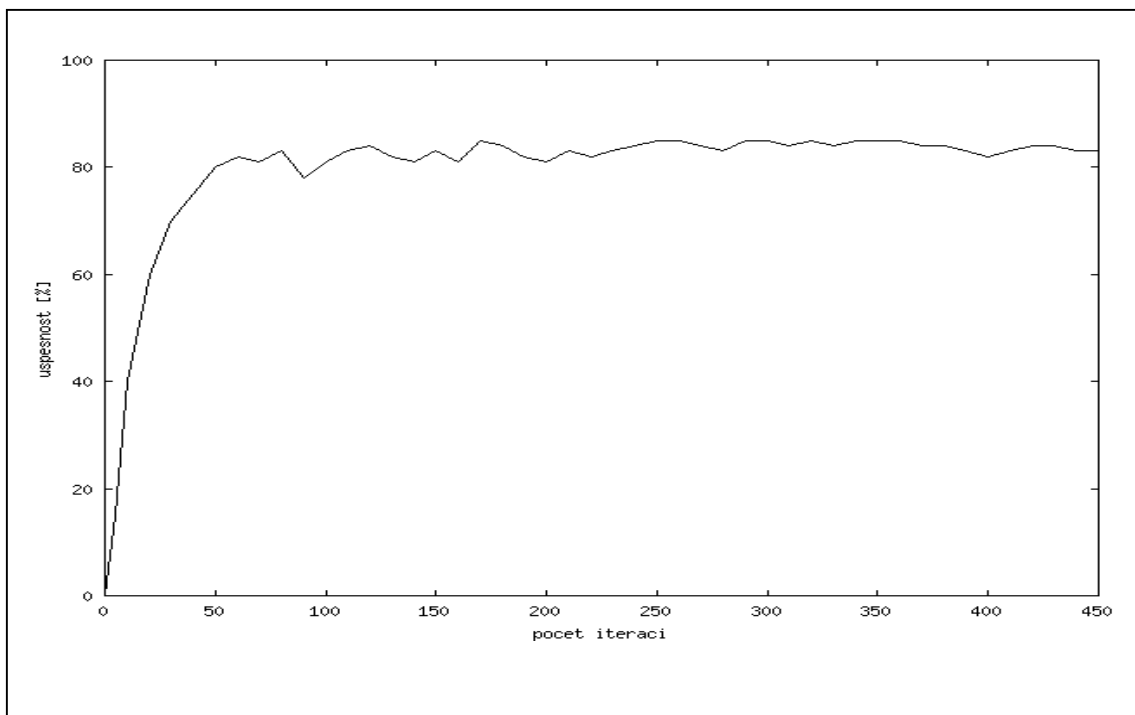


- Křivka učení (learning curve)

Vztah mezi úspěšností klasifikace a počtem příkladů

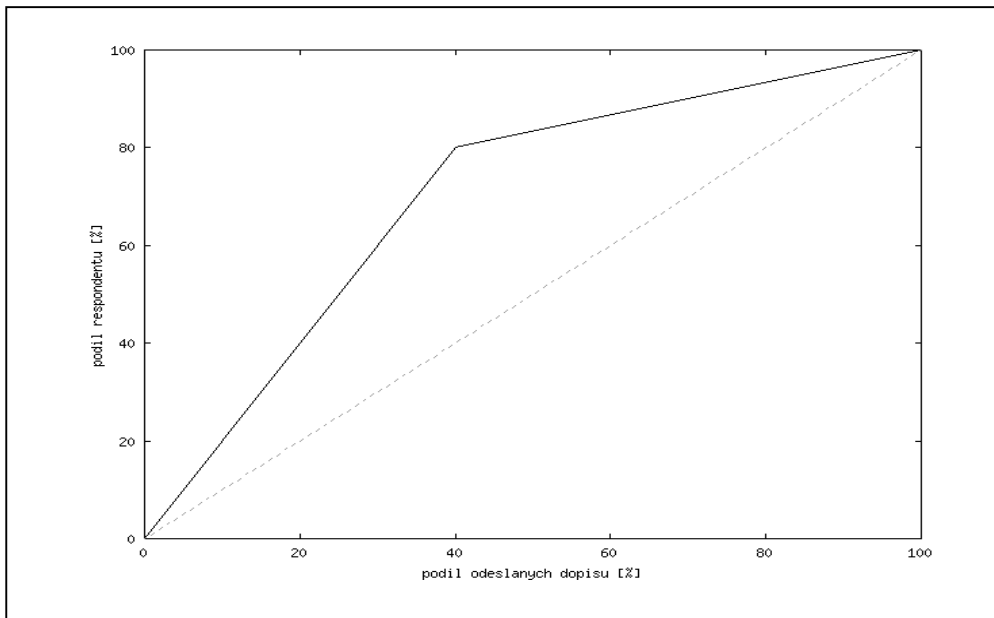


Vztah mezi úspěšností klasifikace a počtem iterací



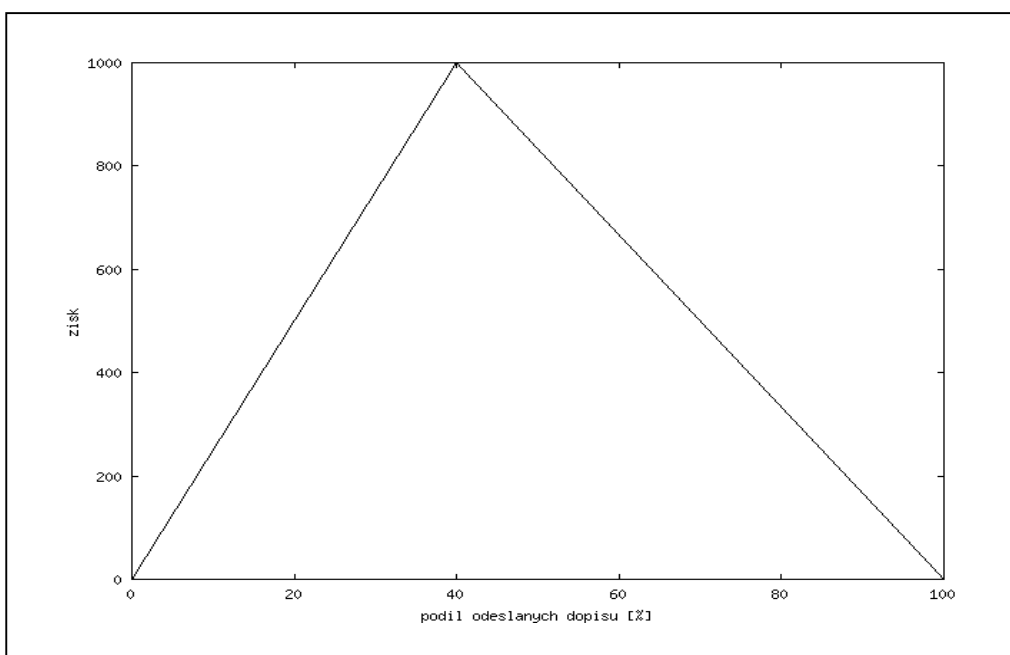
- Křivka navýšení (lift curve)

Vztah mezi počtem úspěšných klasifikací a váhou klasifikace



- Křivka návratnosti investic (ROI curve)

Vztah mezi ziskem z úspěšných klasifikací a váhou klasifikace



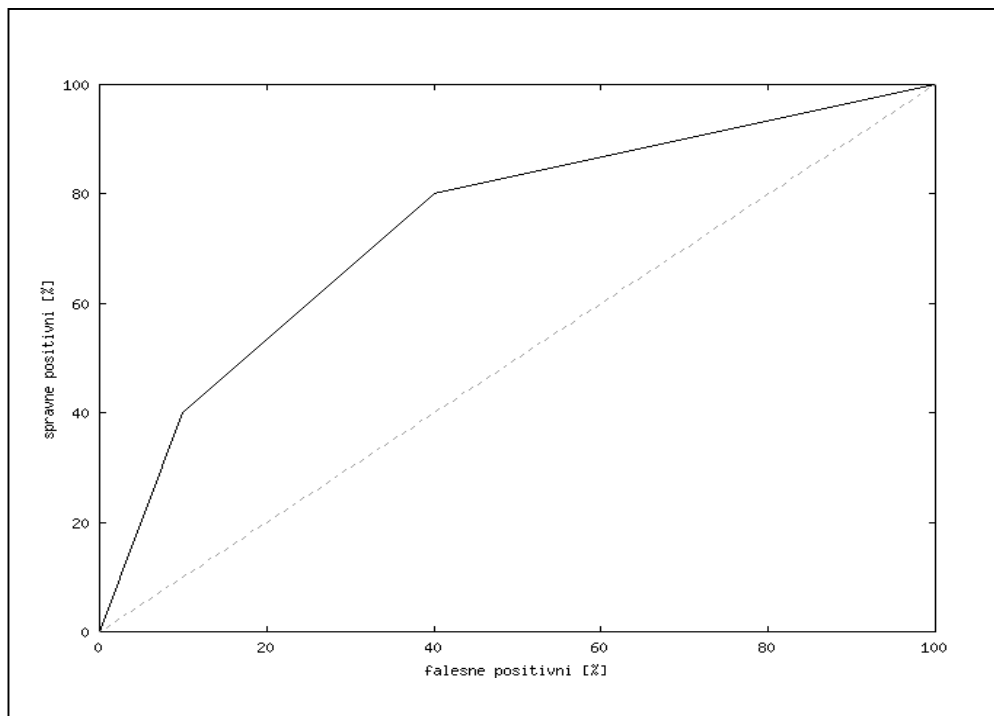
- Křivka ROC

Vztah mezi TP a FP pro různá nastavení klasifikátoru

$$TP\% = \frac{TP}{TP + FN}$$

$$FP\% = \frac{FP}{FP + TN}$$

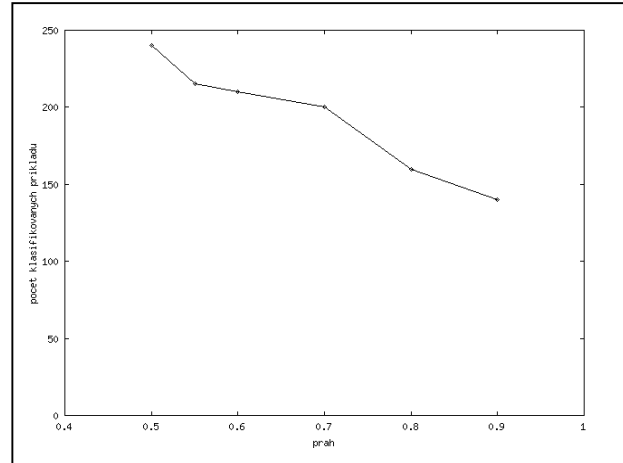
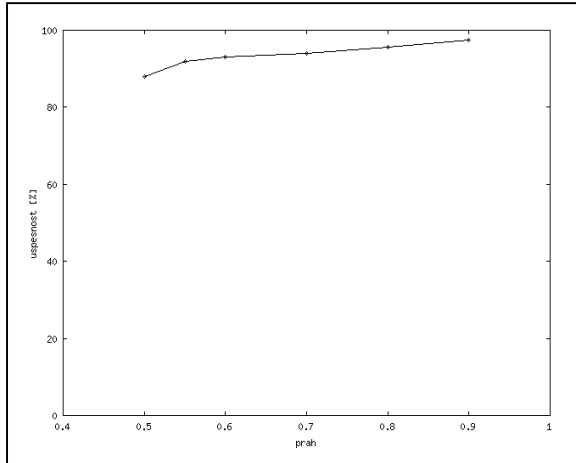
$TP\% = \text{Senzitivita}$ ,  $1 - FP\% = \text{Specifická}$





## Varianta (KEX)

Závislost správnosti a počtu rozhodnutí na prahu  $\alpha$ ,  
rozhodnutí, jen když  $w \geq \alpha$



Numerické třídy ( $p_i$  predikovaná hodnota a  $s_i$  skutečná hodnota)

- $$\text{MSE} = \frac{(p_1 - s_1)^2 + \dots + (p_n - s_n)^2}{n}$$

- $$\text{RMSE} = \sqrt{\frac{(p_1 - s_1)^2 + \dots + (p_n - s_n)^2}{n}}$$

- $$\text{MAE} = \frac{|p_1 - s_1| + \dots + |p_n - s_n|}{n}$$

- $$\text{RSE} = \frac{(p_1 - s_1)^2 + \dots + (p_n - s_n)^2}{(s_1 - s')^2 + \dots + (s_n - s')^2},$$

$$\text{kde } s' = \frac{\sum_i s_i}{n}$$

- $$\rho = \frac{S_{ps}}{\sqrt{S_p^2 S_s^2}},$$

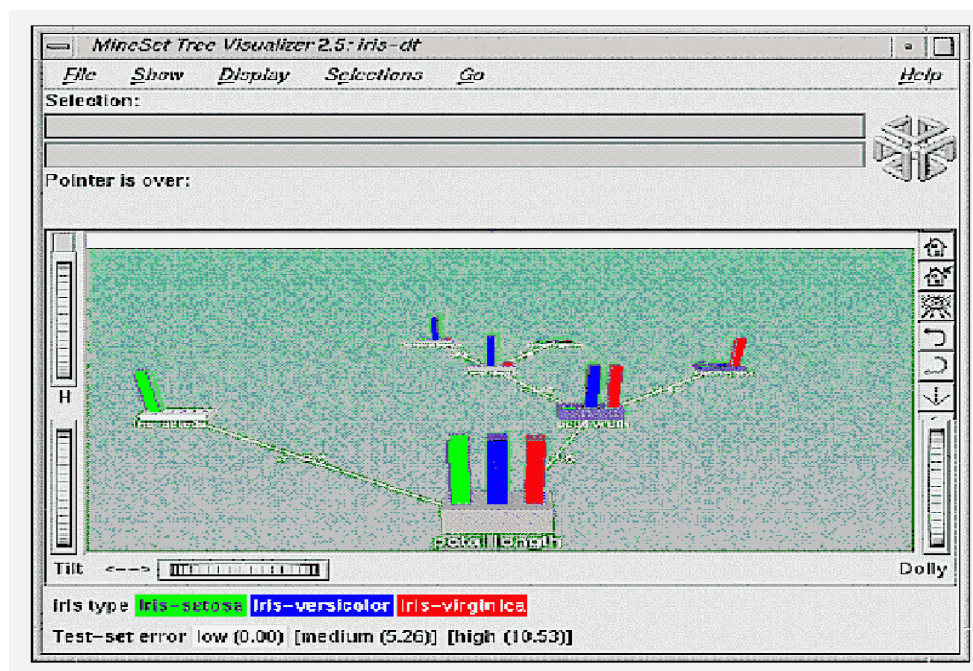
$$\text{kde } S_{ps} = \frac{\sum_i (p_i - p')(s_i - s')}{n-1}, \quad S_p^2 = \frac{\sum_i (p_i - p')^2}{n-1},$$

$$S_s^2 = \frac{\sum_i (s_i - s')^2}{n-1}$$

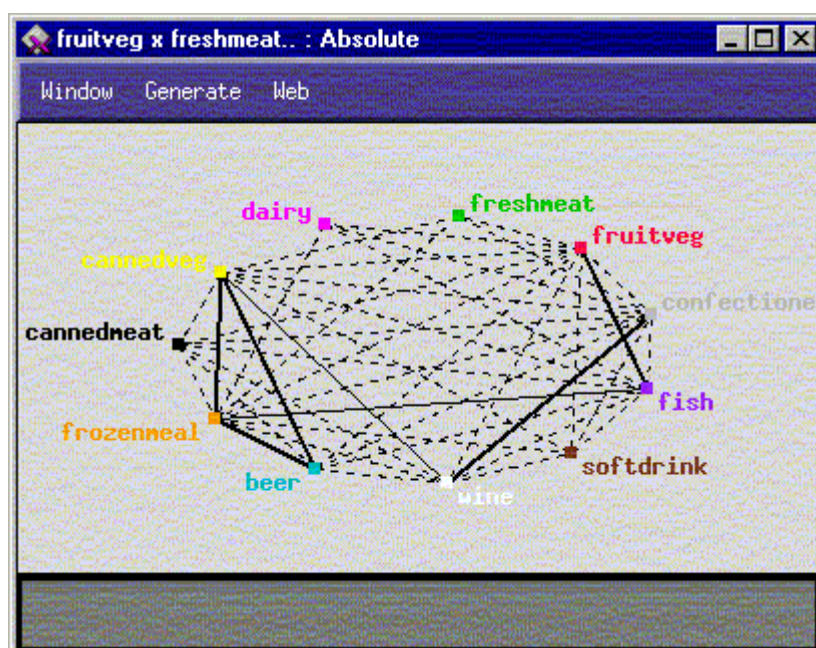
# Vizualizace

## 1. Vizualizace modelů

- Rozhodovací stromy (MineSet)



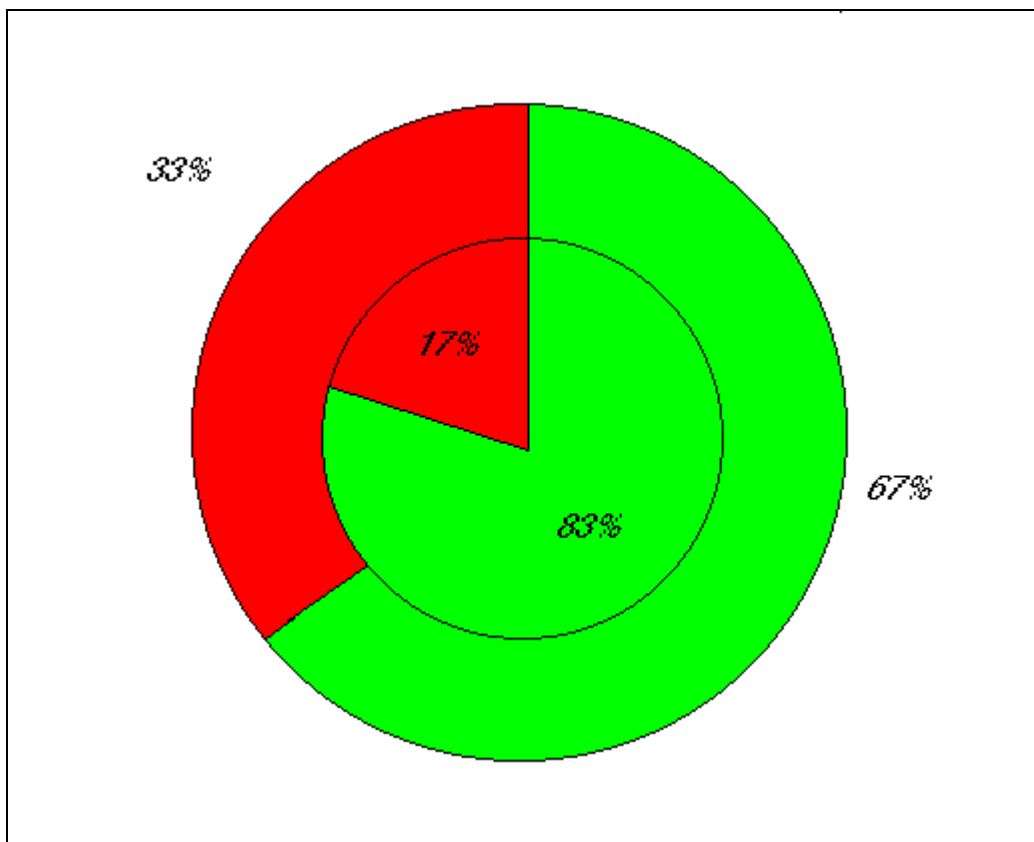
- Asociace (Clementine)



- Pravidlo

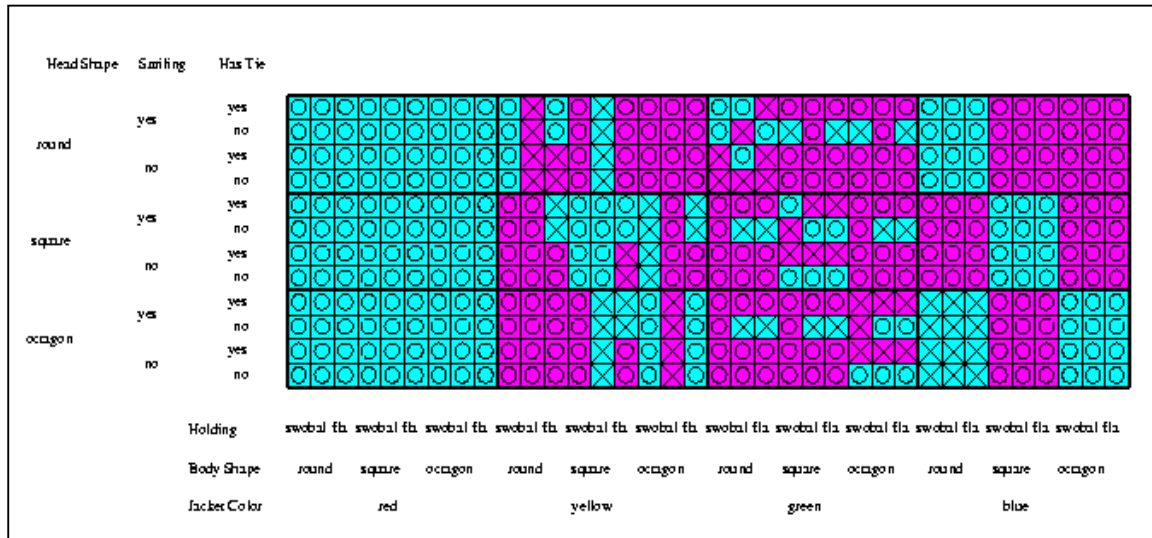
IF nezamestnany(ne) THEN uver(ano)

	uver(ano)	uver(ne)	
nezamestnany(ne)	5	1	6
nezamestnany (ano)	3	3	6
	8	4	12

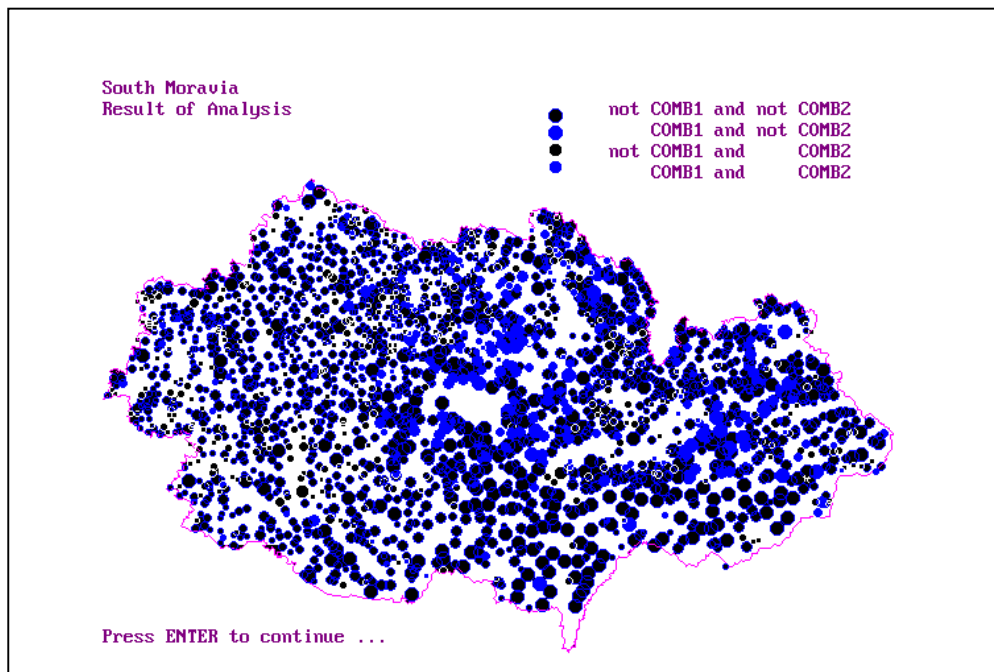


# Vizualizace klasifikací

## Obecné logické diagramy (Michalski)



## spojení s GIS (KEX)



## Porovnávání modelů

- T-test

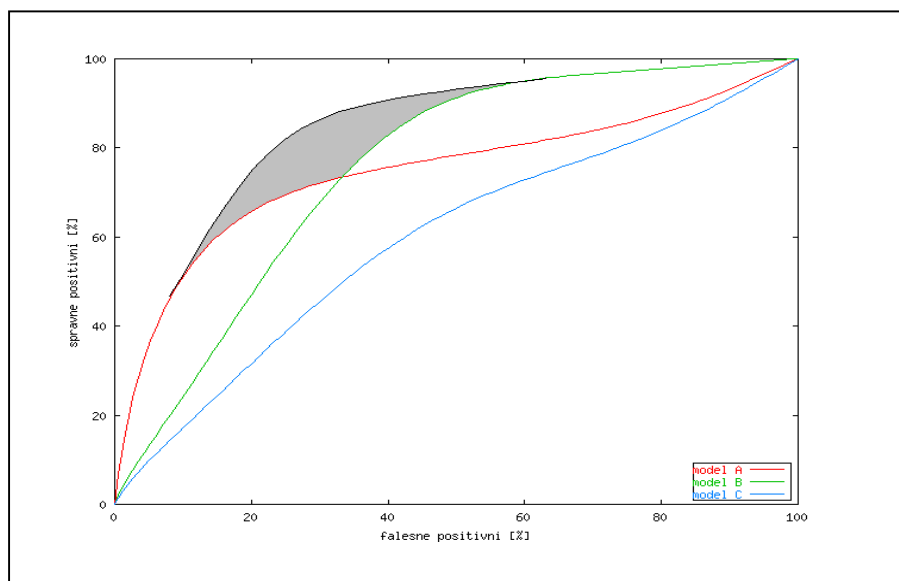
$$t(\mathbf{x}, \mathbf{y}) = \frac{x' - y'}{s(\mathbf{x}, \mathbf{y}) \sqrt{1/m + 1/n}},$$

$$\text{kde } x' = \frac{\sum_i x_i}{m}, \quad y' = \frac{\sum_i y_i}{n} \quad \text{a} \quad s^2(\mathbf{x}, \mathbf{y}) = \frac{(m-1) s_x^2 + (n-1) s_y^2}{m+n-2}$$

Model A bude lepší než model B, pokud

$$t(\mathbf{Acc}_A, \mathbf{Acc}_B) \geq t(1 - \alpha/2, m + n - 2)$$

- ROC křivky



- Occamova břitva

minimum description length, MDL

# Volba nejvhodnějšího algoritmu

- charakteristiky algoritmů vs. charakteristiky dat
  - způsob reprezentace příkladů,
  - vyjadřovací síla,
  - schopnost práce s numerickými atributy,
  - schopnost práce se zašuměnými a chybějícími daty,
  - schopnost práce s maticí cen,
  - předpoklad nezávislosti mezi atributy,
  - ostrá vs. neostrá klasifikace
- empirické studie  
Metaučení nad výsledky dosaženými jednotlivými systémy
  - STATLOG (1991-1994)
    - pro rozsáhlá data se hodí diskriminační analýza (lineární, kvadratická),
    - není velký rozdíl mezi „obyčejnou“ a logistickou diskriminační analýzou,
    - na rozsáhlých datech je nejpomalejší metoda  $k$  nejbližších sousedů,
    - použité algoritmy na tvorbu rozhodovacích stromů se chovaly zhruba stejně; nezdá se tedy, že by nějak zvlášť záleželo na kritériu pro volbu větvení,
    - neuronové sítě dávaly výborné výsledky u dat, kde se napoužívala matice cen.
  - METAL (2000 - )  
  
důraz i na předzpracování

# Kombinování modelů

různé varianty hlasování

- Bagging (bootstrap aggregating)
  - několik stejně velkých trénovacích množin pomocí náhodného výběru s opakováním (bootstrap)
  - všechny modely rovný hlas
- Boosting
  - následující model na data chybně klasifikovaná předcházejícím modelem
  - postupně se vytvářejí modely se stále větší vahou hlasu

## AdaBoost algoritmus

### učení

1. Přiřaď stejnou váhu všem trénovacím příkladům,
2. Pro každou iteraci (vytvářený model)
  - 2.1. Vytvoř model
  - 2.2. Spočítej chybu  $err$  na vážených datech
  - 2.3. If  $err=0$  nebo  $err \geq 0.5$  konec
  - 2.4. Pro každý příklad  
If klasifikace je správně then váha  $w=w*err/(1-err)$
  - 2.5. Normalizuj váhy příkladů (součet nových vah stejný jako součet původních)

### klasifikace jednoho příkladu

1. Přiřaď váhy 0 všm třídám
2. Pro každý model  
Přiřaď třídě určené modelem váhu  $w=w-\log(err/1-err)$
3. Vydej třídu s nejvyšší vahou

- Stacking

rozpoznat spolehlivost jednotlivých modelů na základě meta učení

