ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

VI.
SEKVENČNÍ KLASIFIKACE

ZAČÍNÁME

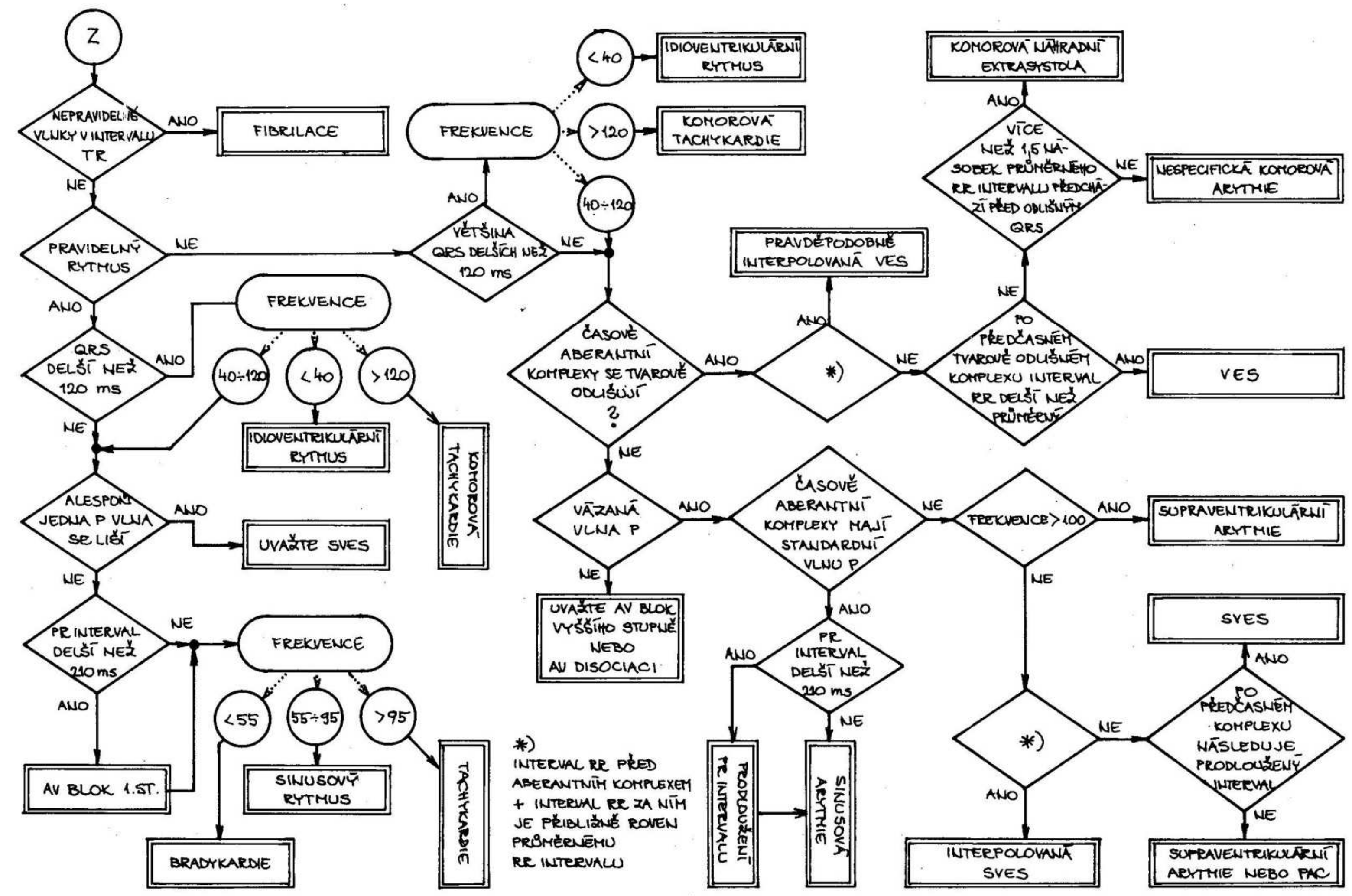
- až dosud (bayesovské klasifikátory, neuronové sítě, …) – pevný konstantní počet příznaků
- kolik a jaké příznaky ?
  - málo příznaků – možná chyba klasifikace;
  - moc příznaků – možná nepřiměřená pracnost, vysoké náklady;
  - použít příznaky, které nesou co nejvíce informace o klasifikační úloze;

ZAČÍNÁME

sekvenční klasifikace - kompromis mezi velikostí klasifikační chyby a cenou určení příznaků

- klasifikace na základě klasifikačního stromu;

ZAČÍNÁME



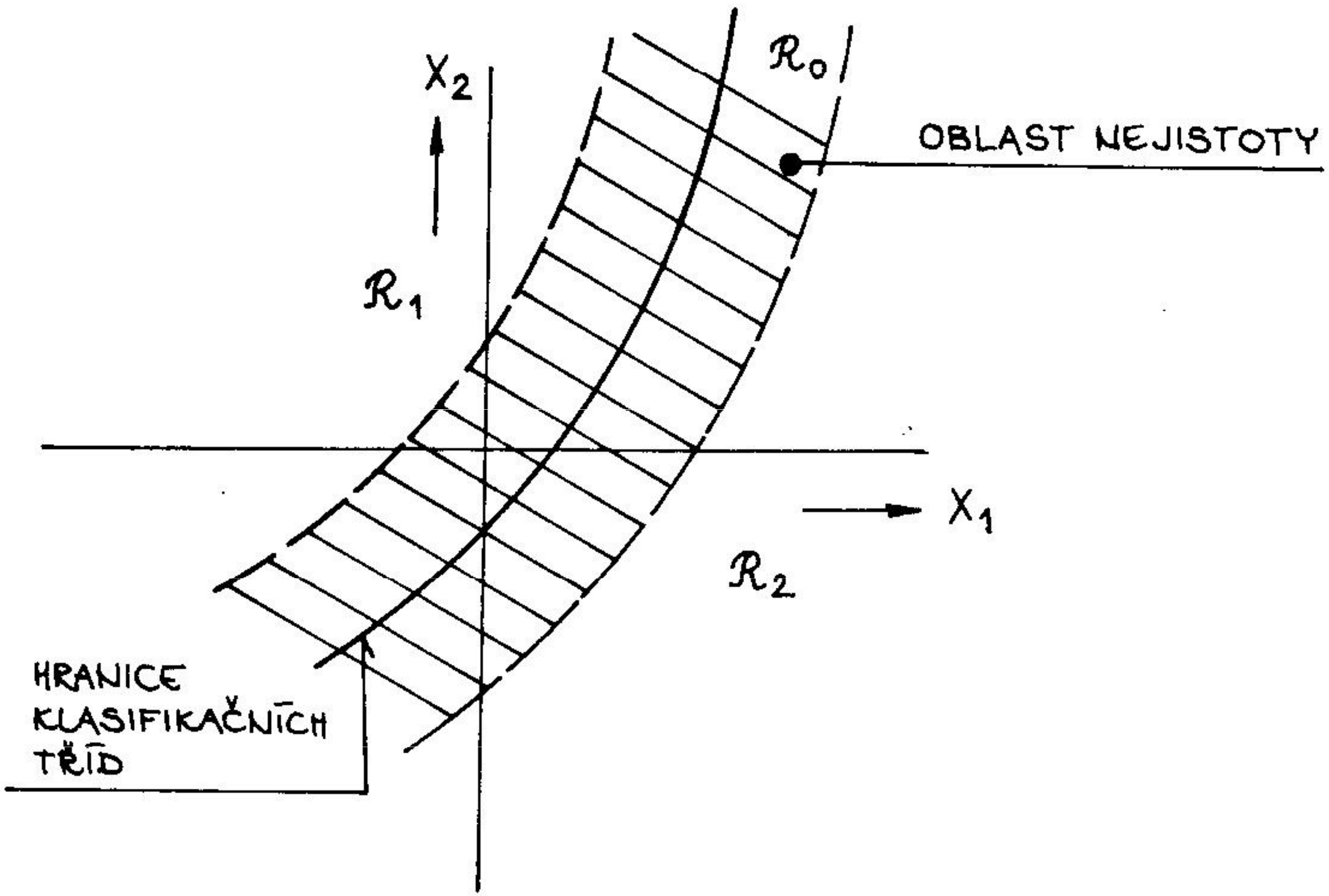
Rozhodovací strom pro klasifikaci rytmu signálu EKG

ZAČÍNÁME

sekvenční klasifikace - kompromis mezi velikostí klasifikační chyby a cenou určení příznaků

- klasifikace na základě klasifikačního stromu;
- klasifikace s rostoucím počtem příznaků, přičemž okamžik ukončení klasifikační procedury stanoví klasifikátor sám podle předem daného kritéria pro kvalitu rozhodnutí (tj. na základě vlastností klasifikačních tříd, resp. obrazů v nich);

PRINCIP



X_2
$\mathcal{R}_0$
OBLAST NEJISTOTY
$\mathcal{R}_1$
X_1
$\mathcal{R}_2$
HRANICE KLASIFIKAČNÍCH TŘÍD

WALDOVO KRITÉRIUM

- předpokládejme dichotomický klasifikátor obrazů popsaných příznakovými vektory $(x_1, x_2, \ldots)$;
- nechť $p(x_1, x_2, \ldots, x_i|\omega_1)$ a $p(x_1, x_2, \ldots, x_i|\omega_2)$ jsou i-rozměrné hustoty pravděpodobnosti výskytu obrazu $\mathbf{x} = (x_1, x_2, \ldots, x_i)$ v i-tém klasifikačním kroku v třídách ω_1 a ω_2;
- nechť A a B jsou konstanty $(0<B<1<A<\infty)$;

WALDOVO KRITÉRIUM

- pokud pro věrohodnostní poměr

$$\Lambda_i = \frac{p(x_1, x_2, \ldots, x_i \mid \omega_1)}{p(x_1, x_2, \ldots, x_i \mid \omega_2)}$$

je $\Lambda_i \leq B$, pak $d_w(\mathbf{x}) = \omega_2$, je-li $\Lambda_i \geq A$, pak $d_w(\mathbf{x}) = \omega_1$; když $\Lambda_i \in (B,A)$, pokračujeme s dalším příznakem

WALDOVO KRITÉRIUM

- jsou-li dány pravděpodobnosti chybné klasifikace

$$\alpha = \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1)\, d\mathbf{x} \quad \text{a} \quad \beta = \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2)\, d\mathbf{x}$$

můžeme určit meze A a B podle vztahů

$$A \cong \frac{1-\alpha}{\beta} \quad \text{a} \quad B \cong \frac{\alpha}{1-\beta}$$

WALDOVO KRITÉRIUM

OPTIMALITA

- ☑ pro libovolné kritérium s pevným počtem n příznaků a s pravděpodobnostmi α a β chybných rozhodnutí platí pro n, že je větší nebo rovno střední hodnotě počtu kroků podle W.k.;
- ☑ pro libovolné sekvenční kritérium je k rozhodnutí potřeba průměrný počet kroků větší než je průměrný počet kroků podle W.k.

MODIFIKOVANÉ WALDOVO KRITÉRIUM

přes optimální vlastnosti Waldova kritéria může nastat:

- ☑ počet kroků potřebných k přijetí rozhodnutí může být pro některé obrazy příliš velký, i když střední hodnota počtu kroků pro všechny obrazy je nízká;
- ☑ střední hodnota počtu kroků potřebných k rozhodnutí může být příliš velká, požadujeme-li malé pravděpodobnosti chybných rozhodnutí;

MODIFIKOVANÉ WALDOVO KRITÉRIUM

PRAXE

Po určitém počtu kroků se sekvenční výpočet přeruší a dokončí se na základě nějakého rozhodnutí vycházejícího z nějakého kritéria založeného na pevném počtu příznaků.

Přerušení:

- ☑ předepsaný počet kroků;
- ☑ zavedení proměnných hranic A(i) a B(i).

MODIFIKOVANÉ WALDOVO KRITÉRIUM

Nechť A(i) a B(i) jsou nezáporné nerostoucí, resp. nekladné neklesající funkce počtu klasifikačních kroků.

Je-li $\Lambda_i \leq e^{B(i)}$, pak $d_w(\mathbf{x}) = \omega_2$, je-li $\Lambda_i \geq e^{A(i)}$, pak $d_w(\mathbf{x}) = \omega_1$; když $\Lambda_i \in (e^{B(i)}, e^{A(i)})$, pokračujeme s dalším příznakem

MODIFIKOVANÉ WALDOVO KRITÉRIUM

URČENÍ FUNKCÍ A(i) A B(i)

- obvykle experimentálně;
- jestliže pro mezní funkce platí $A(i_{max}) = B(i_{max})$, pak nejpozději pro $i = i_{max}$ je klasifikace ukončena, přičemž střední počet potřebných kroků je menší než u W.k. za cenu snížení kvality rozhodnutí

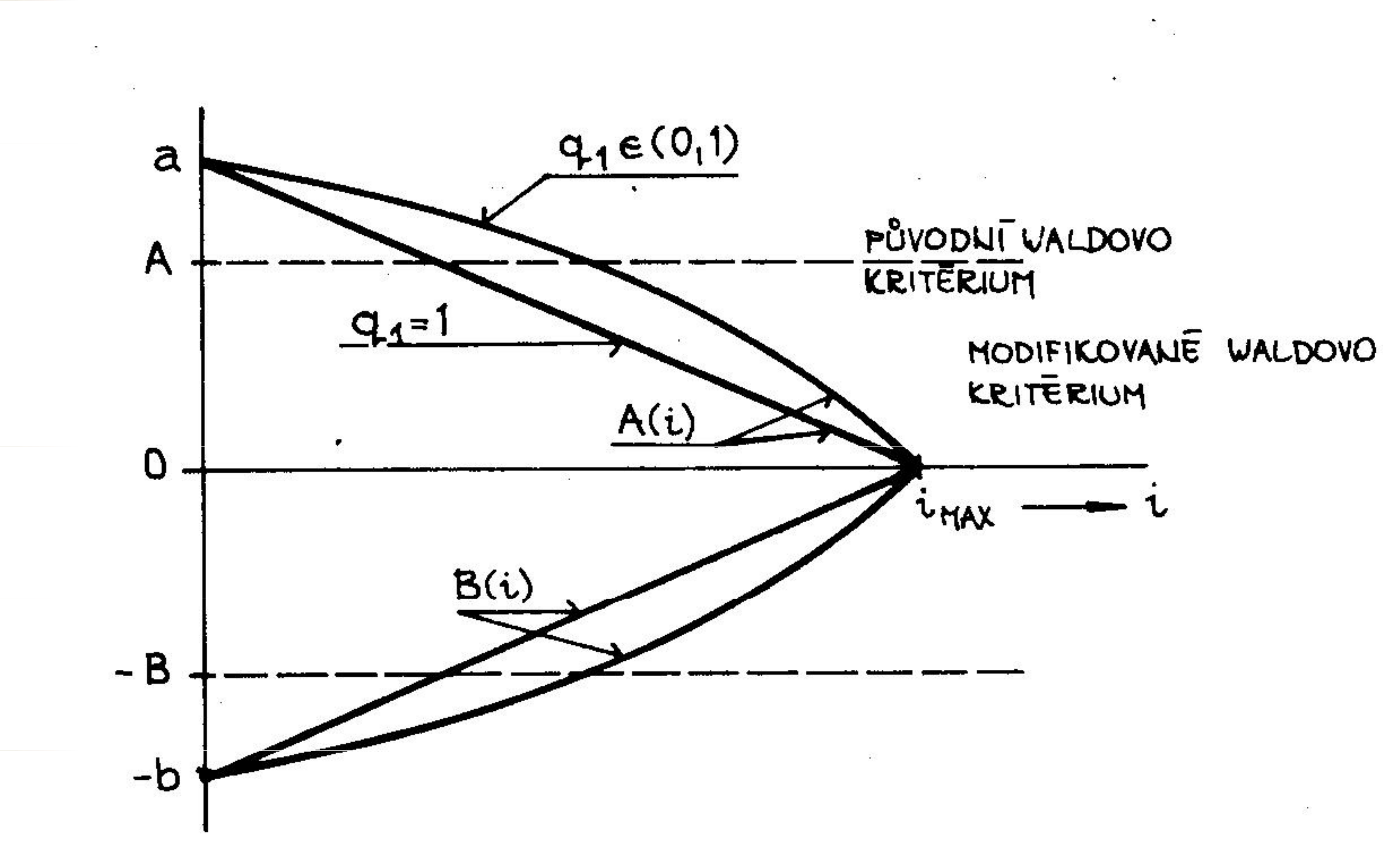
MODIFIKOVANÉ WALDOVO KRITÉRIUM

URČENÍ FUNKCÍ A(i) A B(i)

$$A(i) = a.\left(1 - \frac{i}{i_{max}}\right)^{q_1} \text{ a } B(i) = -b.\left(1 - \frac{i}{i_{max}}\right)^{q_2}$$

kde $a, b > 0$ a $q_1, q_2 \in (0,1\rangle$

MODIFIKOVANÉ WALDOVO KRITÉRIUM



REEDOVO KRITÉRIUM

- pro obecný počet tříd zobecněný věrohodnostní poměr

$$\Lambda_i(\mathbf{x} \mid \omega_r) = \frac{p(x_1, x_2, \ldots, x_i \mid \omega_r)}{\left[\prod_{s=1}^{R} p(x_1, x_2, \ldots, x_i \mid \omega_s)\right]^{1/R}}, \; r = 1,2,\ldots,R$$

- takto vypočítaný poměr se srovná s mezní hodnotou r-té třídy $A(\omega_r)$, určenou jako

$$A(\omega_r) = \frac{1 - P_{rr}}{\left[\prod_{s=1}^{R} (1 - P_{rs})\right]^{1/R}}, \; r = 1,2,\ldots,R$$

kde P_{rs} je pravděpodobnost, že obraz ze třídy ω_s zatřídíme do ω_r.

REEDOVO KRITÉRIUM

- pokud pro třídu ω_P platí

$$\Lambda_i(\mathbf{x}|\omega_P) \leq A(\omega_P),\ p=1,2,\ldots,R,$$

pak předpokládáme, že obraz **x** nepatří do třídy ω_P, kterou lze z dalších úvah vyloučit;
- po vyloučení všech možných tříd se spočítají nové hodnoty věrohodnostních poměrů pro zbylé třídy a proces se opakuje;
- není-li možné vyloučit další třídu, zvýší se počet příznaků a klasifikace pokračuje, dokud nezbude jediná klasifikační třída;

REEDOVO KRITÉRIUM

- pro R=2 je Reedovo kritérium ekvivalentní kritériu Waldovu a má tytéž optimální vlastnosti;
- pro R>2 nebyla optimalita prokázána;

MODIFIKOVANÉ REEDOVO KRITÉRIUM

- ☑ stejně jako Waldova kritéria lze použít proměnných mezí
- ☑ proměnných práh je zpravidla definován vztahem

$$G_r(i) = g_r \cdot \left(1 - \frac{i}{i_{max}}\right)^{q_i}, r = 1, \ldots, R$$