ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

LITERATURA

- Holčík, J.: přednáškové prezentace
- Holčík, J.: Analýza a klasifikace signálů. [Učební texty VŠ], Brno, FE VUT 1992.

LITERATURA

- Bishop, C.M.: Pattern Recognition and Machine Learning, New York, Springer2006, 738s.
- Han, J., Kamber,M.: Data Mining . Concepts &Techniques. 2nd ed., Amsterdam, Elsevier2006, 770s.
- Tan P-N., Steibach M., Kumar V.: Introduction to Data Mining. Boston, Pearson-Addison Wesley 2006, 769s.

0. ČEM TO BUDE?

ANOTACE

Předmět poskytne informaci o základních metodách a algoritmech pro výběr popisu, hodnocení a klasifikaci biomedicínských dat. Zabývá se základním tříděním klasifikačních přístupů – příznakové a strukturální a uvádí principy obou přístupů. Dále se zabývá podrobně zejména metodami příznakovými. Klasifikace podle diskriminačních funkcí (princip a stanovení diskriminačních funkcí na základě statistických vlastností množiny obrazů) a minimální vzdálenosti. Sekvenční klasifikace. Volba a výběr příznaků. Selekce a extrakce příznaků – analýza hlavních a nezávislých komponent, faktorová analýza. Učení klasifikátorů. Shlukování – podobnost mezi obrazy, podobnost mezi shluky, metody shlukování. Klasifikace pomocí neuronových sítí. Základní přístupy jsou vysvětlovány ve spojitosti s praktickými úlohami.

V rámci cvičení studenti řeší samostatné úlohy (projekty), buď zadané učitelem nebo související s řešením jejich diplomové práce.

OSNOVA

1. Klasifikace dat – základní terminologie. Třídění klasifikačních algoritmů.
2. Příznakové metody. Klasifikace podle diskriminačních funkcí a minimální vzdálenosti.
3. Stanovení diskriminačních funkcí na základě statistických vlastností množiny obrazů.
4. Sekvenční klasifikace.
5. Volba a výběr příznaků.
6. Analýza hlavních komponent.
7. Analýza nezávislých komponent.
8. Faktorová analýza
9. Učení klasifikátorů. Metody odhadu hustot pravděpodobnosti a odhad apriorních pravděpodobností klasifikačních tříd.
10. Shlukování. Podobnost mezi obrazy a shluky.
11. Metody shlukování.
12. Klasifikace pomocí neuronových sítí.

OSNOVA

1. Klasifikace dat – základní terminologie. Třídění klasifikačních algoritmů.
2. Příznakové metody. Klasifikace podle diskriminačních funkcí a minimální vzdálenosti.
3. Stanovení diskriminačních funkcí na základě statistických vlastností množiny obrazů.
4. Sekvenční klasifikace.
5. Volba a výběr příznaků.
6. Analýza hlavních komponent.
7. Analýza nezávislých komponent.
8. Faktorová analýza
9. Učení klasifikátorů. Metody odhadu hustot pravděpodobnosti a odhad apriorních pravděpodobností klasifikačních tříd.
10. Shlukování. Podobnost mezi obrazy a shluky.
11. Metody shlukování.
12. Klasifikace pomocí neuronových sítí.

OSNOVA

1. Klasifikace dat – základní terminologie. Třídění klasifikačních algoritmů.
2. Příznakové metody. Klasifikace podle diskriminačních funkcí a minimální vzdálenosti.
3. Stanovení diskriminačních funkcí na základě statistických vlastností množiny obrazů.
4. Sekvenční klasifikace.
5. Volba a výběr příznaků.
6. Analýza hlavních komponent.
7. Analýza nezávislých komponent.
8. Faktorová analýza
9. Učení klasifikátorů. Metody odhadu hustot pravděpodobnosti a odhad apriorních pravděpodobností klasifikačních tříd.
10. Shlukování. Podobnost mezi obrazy a shluky.
11. Metody shlukování.
12. Klasifikace pomocí neuronových sítí.
13. Strukturální (syntaktická) klasifikace

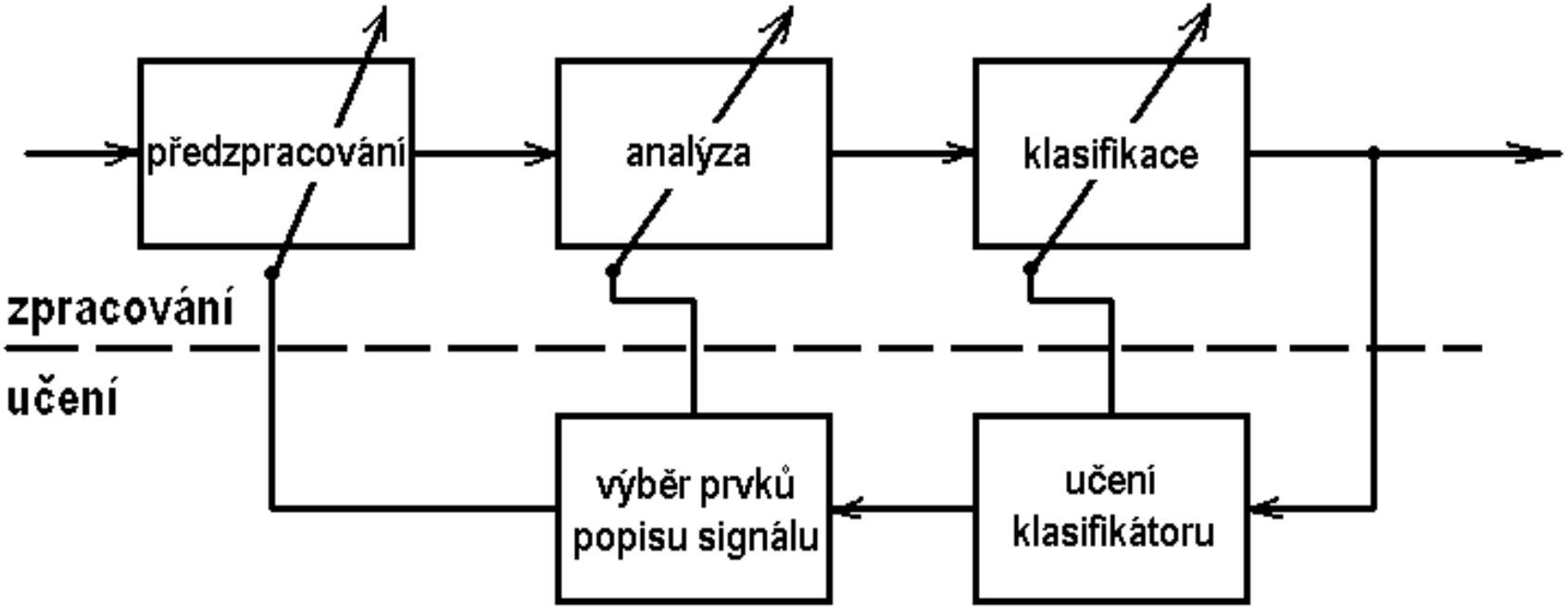
UKONČENÍ PŘEDMĚTU

Požadavky:

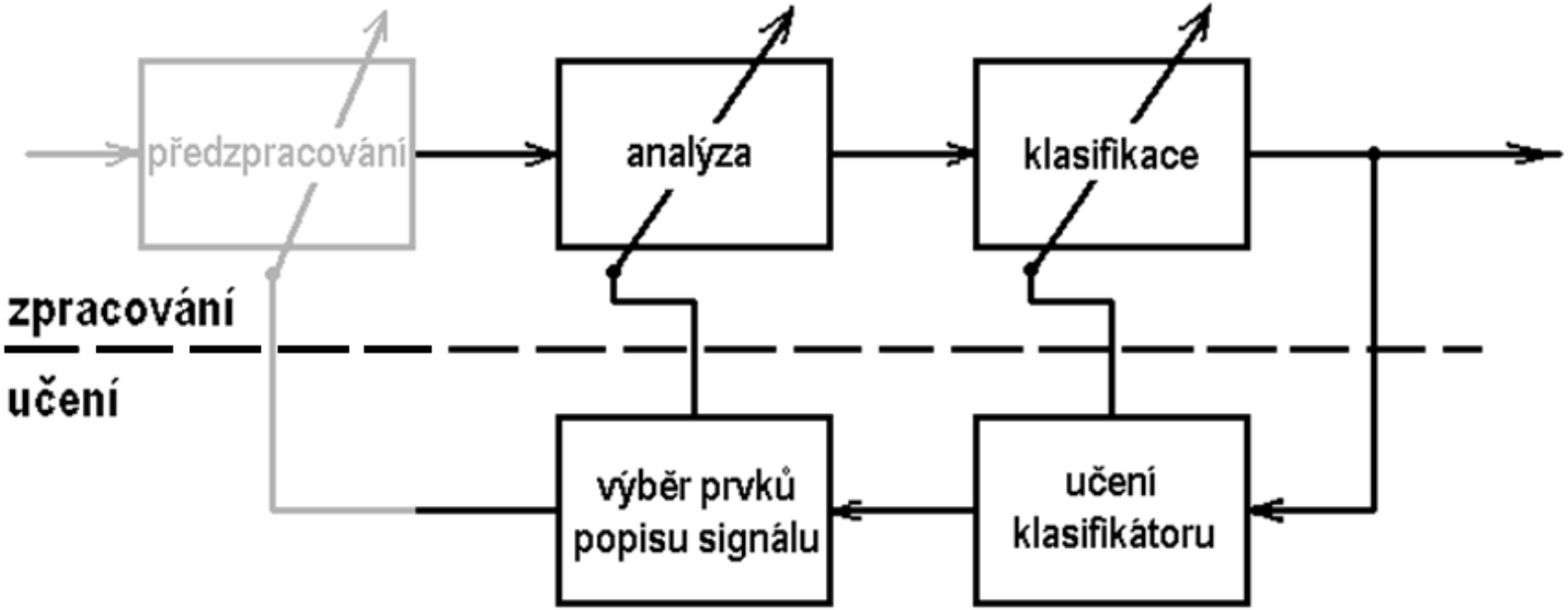
- ☑ ústní zkouška
  - ➔ dvě části:
    - učená rozprava o některém z témat, která budou náplní předmětu;
    - diskuze nad vyřešeným problémem týkajícím se problematiky klasifikace dat a používajícím některé z technik, které budou náplní předmětu;

I. ZAČÍNÁME

OBECNÉ SCHÉMA ZPRACOVÁNÍ SIGNÁLŮ (DAT)



OBECNÉ SCHÉMA ZPRACOVÁNÍ SIGNÁLŮ (DAT)



OBECNÉ SCHÉMA ZPRACOVÁNÍ DAT

ZPRACOVÁNÍ

- **předzpracování**
  - filtrace rušivých složek x zvýraznění užitečných složek signálu;
  - rekonstrukce a doplnění chybějících údajů;
  - konverze typu dat;
  - redukce dat;
  - (A/Č převod);
- **analýza dat**
  - určení hodnot příznaků (reprezentativních parametrů) – pro příznakové klasifikátory;
  - nalezení primitiv (charakteristických tvarových segmentů) – strukturální klasifikátory
- **klasifikátor** –
  - zatřídění do diagnostických kategorií

OBECNÉ SCHÉMA ZPRACOVÁNÍ DAT

ZPRACOVÁNÍ

- **předzpracování**
  - filtrace rušivých složek x zvýraznění užitečných složek signálu;
  - rekonstrukce a doplnění chybějících údajů;
  - konverze typu dat;
  - redukce dat;
  - (A/Č převod);
- **analýza dat**
  - určení hodnot příznaků (reprezentativních parametrů) – pro příznakové klasifikátory;
  - nalezení primitiv (charakteristických tvarových segmentů) – strukturální klasifikátory
- **klasifikátor –**
  - zatřídění do diagnostických kategorií

OBECNÉ SCHÉMA ZPRACOVÁNÍ DAT

- **Analýza** (z řečtiny – *rozbor, rozčlenění*) je vědecká metoda založená na dekompozici celku na elementární části. Cílem analýzy je identifikovat podstatné a nutné vlastnosti elementárních částí celku, poznat jejich podstatu a zákonitosti.
- **Syntéza** je obecné označení pro proces spojení dvou nebo více částí do jednoho celku. S tímto pojmem se lze setkat v různých spojeních: syntéza obrazu, syntéza řeči, syntéza zvuku, chemická syntéza, jaderná syntéza, termonukleární syntéza, syntéza látek, fotosyntéza, proteosyntéza, biosyntéza, evoluční syntéza.

ANALÝZA

Během analýzy se vytváří formální (abstraktní) popis zpracovávaných dat, který nese podstatnou informaci z hlediska kvality rozhodování při klasifikaci. Abstraktní popis se často nazývá obrazem $\Rightarrow$ rozpoznávání obrazů (pattern recognition). V datech je vybrána určitá množina elementárních vlastností, příp. jejich elementárních částí a jejich vazeb, jejichž způsob popisu je apriori znám.

KLASIFIKACE

- rozumí se rozdělení (konkrétní či teoretické) dané skupiny (množiny) předmětů či jevů na konečný počet dílčích skupin (podmnožin), v nichž všechny předměty či jevy mají dostatečně podobné společné vlastnosti. Vlastnosti podle nichž lze klasifikaci zadat či provádět, určují klasifikační kritéria. Předměty (jevy), které mají podobnou uvažovanou vlastnost tvoří třídu. Každá klasifikace musí být úplná, tzn., že každý předmět musí patřit do nějaké třídy a nemůže být současně ve dvou či více třídách.

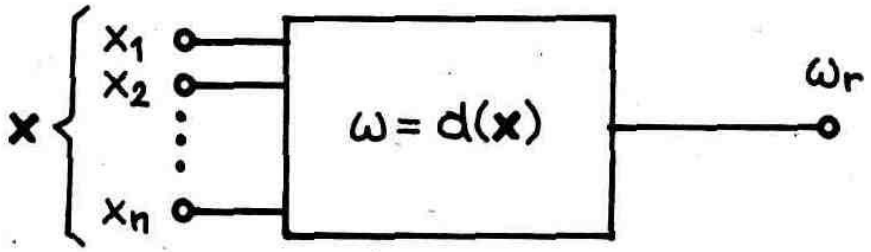
KLASIFIKÁTOR

- **Klasifikátor** je stroj (algoritmus,…) s jedním diskrétním výstupem, který udává třídu, do které klasifikátor zařadil vstupní reprezentaci dat

$$\omega_r = d(\mathbf{x})$$

$d(\mathbf{x})$ je funkce argumentu $\mathbf{x}$ představujícího reprezentaci vstupních dat, kterou nazýváme **rozhodovací pravidlo klasifikátoru**;

ω_r je **identifikátor klasifikační třídy;** $\omega_r|_{r=1,\dots,R} \in \Omega$



PRINCIPY KLASIFIKACE

- pomocí diskriminačních funkcí – funkcí, které určují míru příslušnosti k dané klasifikační třídě;
- pomocí definice hranic mezi jednotlivými třídami a logických pravidel;
- pomocí vzdálenosti od reprezentativních obrazů (etalonů) klasifikačních tříd;
- pomocí ztotožnění s etalony;

OBECNÉ SCHÉMA ZPRACOVÁNÍ DAT

UČENÍ

- **učení klasifikátoru**
  - nastavení klasifikačních kritérií;
    - s učitelem
      - dokonalým
      - nedokonalým
    - bez učitele – typicky shlukování
- **výběr prvků popisu dat**
  - stanovení reprezentativních charakteristických rysů zpracovávaného dat;

TYPY KLASIFIKÁTORŮ

Základní členění vychází z reprezentace vstupních dat

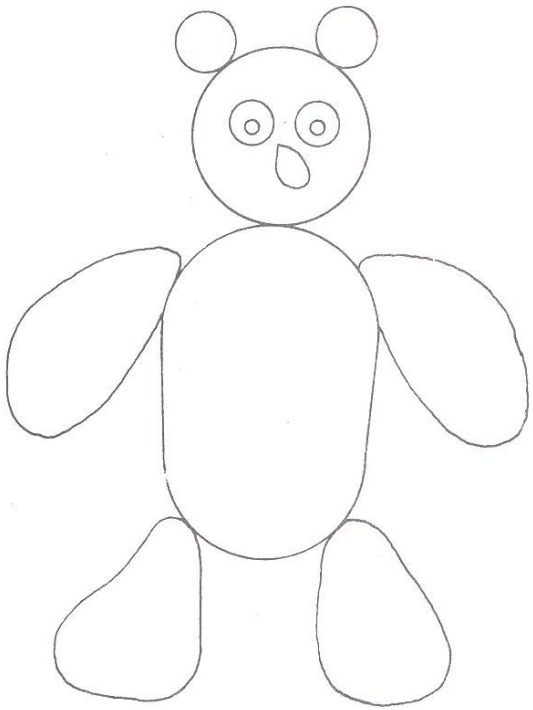
- příznakové – každý vstupní data jsou vyjádřena vektorem hodnot (příznaků);
- strukturální (syntaktické) – vstupní data jsou popsána relačními strukturami;
- kombinované – jednotlivá primitiva jsou doplněna příznakovým popisem

II. STRUKTURÁLNÍ KLASIFIKACE

STRUKTURÁLNÍ POPIS

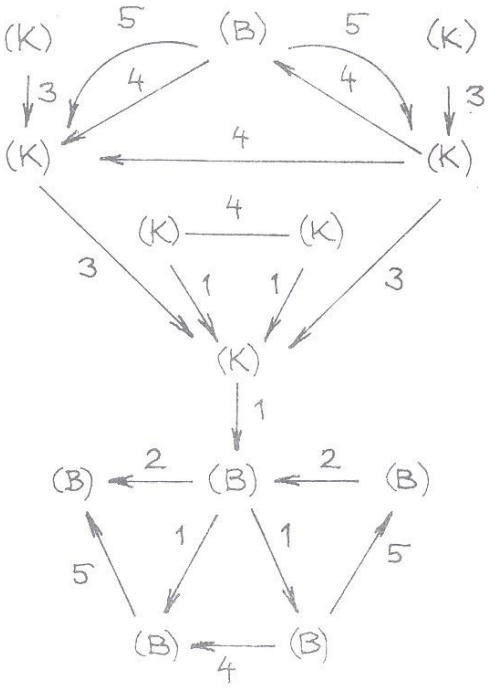
- relační struktura je vytvořena z určitých elementárních popisných částí dat, tzv. primitiv a vzájemných vztahů mezi nimi – relacemi;
- relační struktury zpravidla vyjadřujeme pomocí grafů;

STRUKTURÁLNÍ POPIS



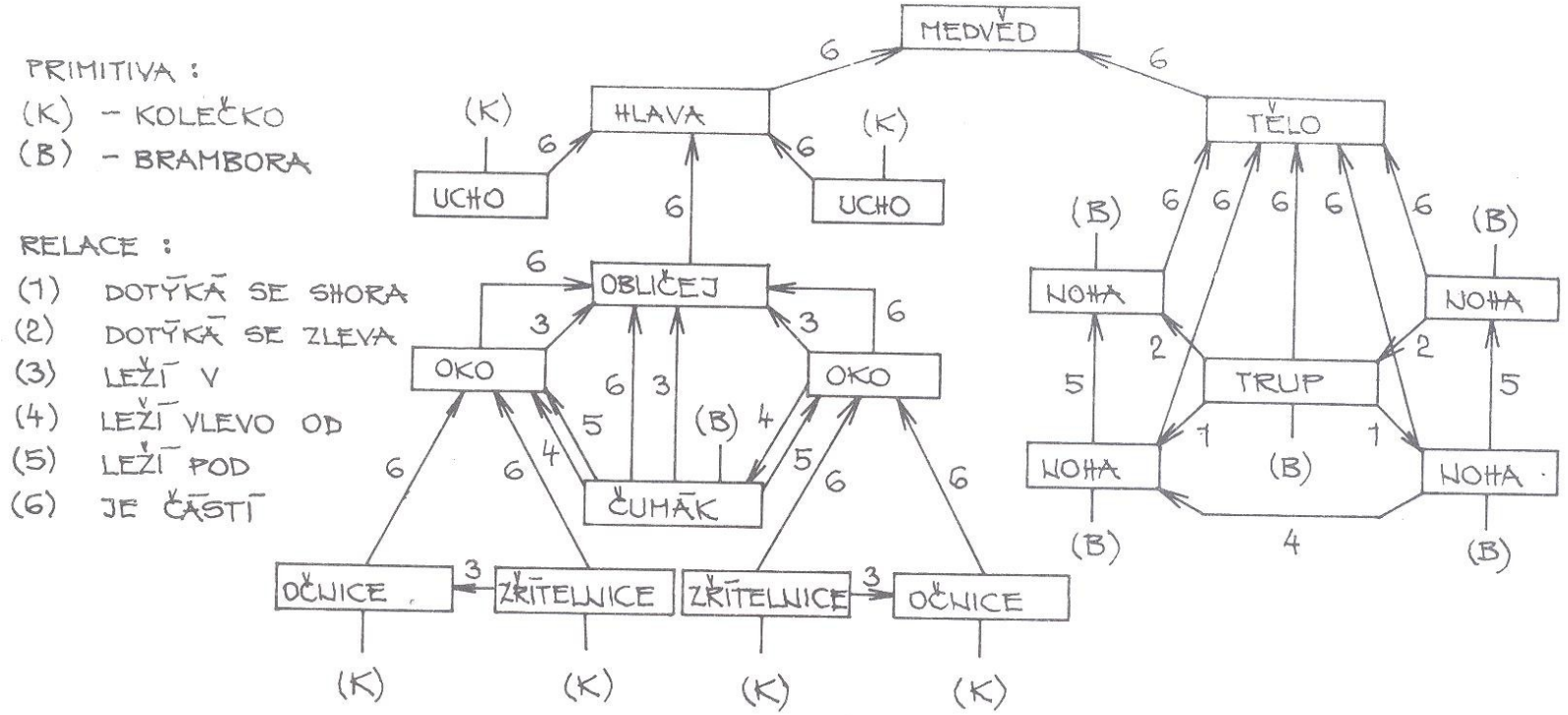
PRIMITIVA:
(K) – KOLEČKO
(B) – BRAMBORA

RELACE:
(1) – DOTÝKÁ SE SHORA
(2) – DOTÝKÁ SE ZLEVA
(3) – LEŽÍ UVNITŘ
(4) – LEŽÍ VLEVO OD
(5) – LEŽÍ POD



Obr. 3.1 Primitiva, relace a relační struktura čarové kresby

STRUKTURÁLNÍ POPIS

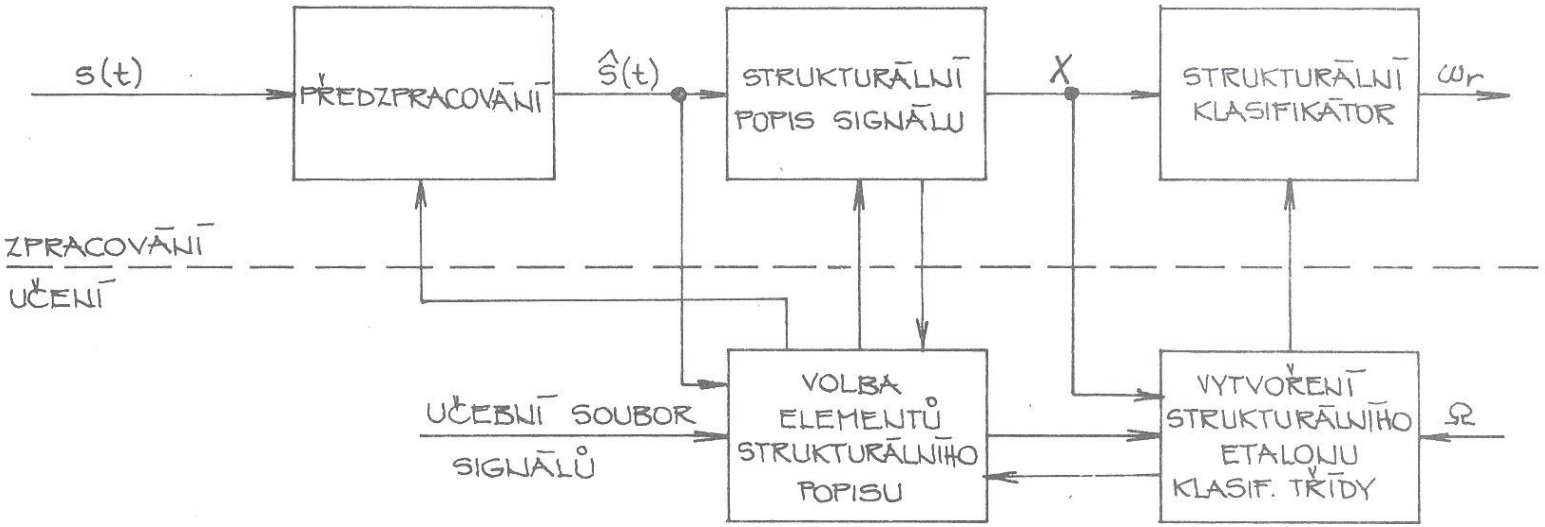


Obr. 3. 4 Hierarchická relační struktura kresby z obr. 3. 1

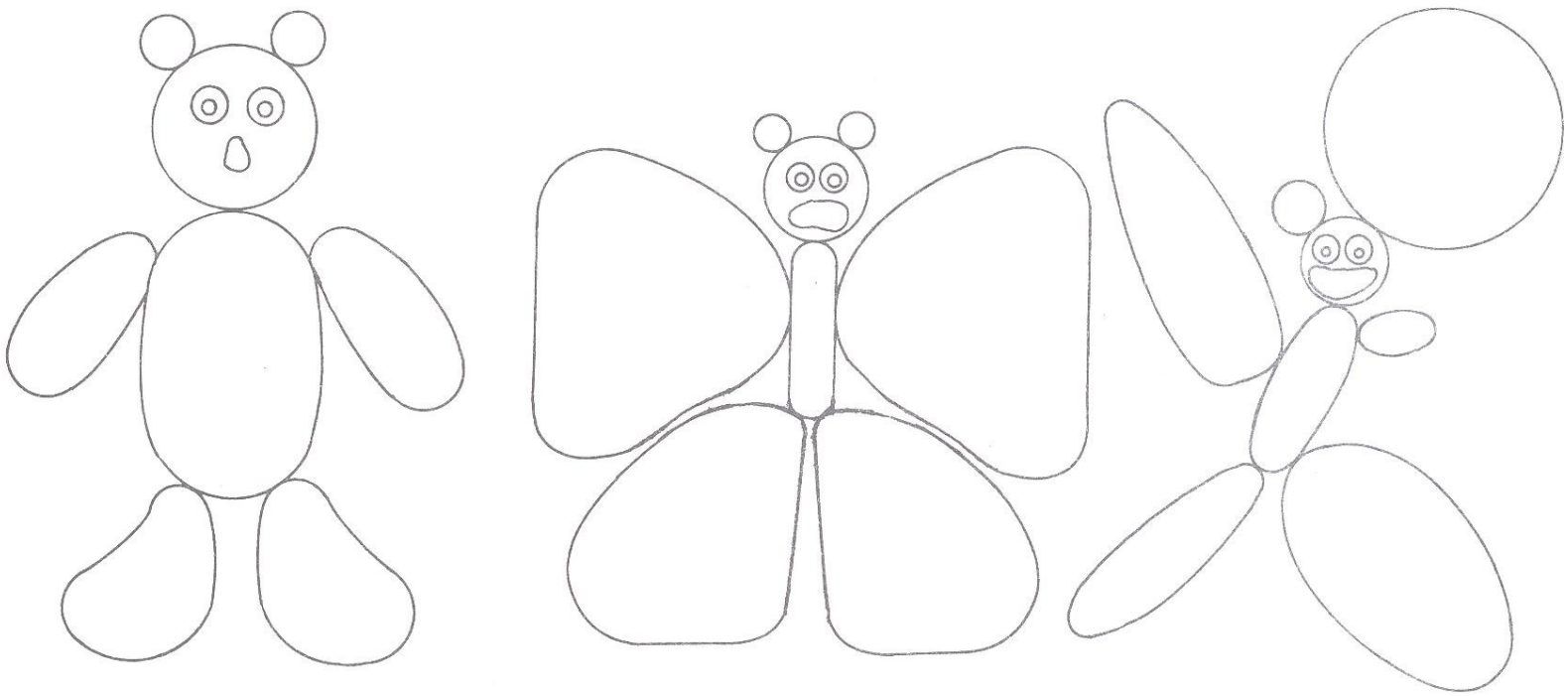
TYPY RELAČNÍCH STRUKTUR

- **řetězce** (uvuuyzuvw)
- **pole** (především pro reprezentaci 2D obrazů)
- **stromy** (relační struktura neobsahující cykly a paralelní cesty)
- **obecné grafy**

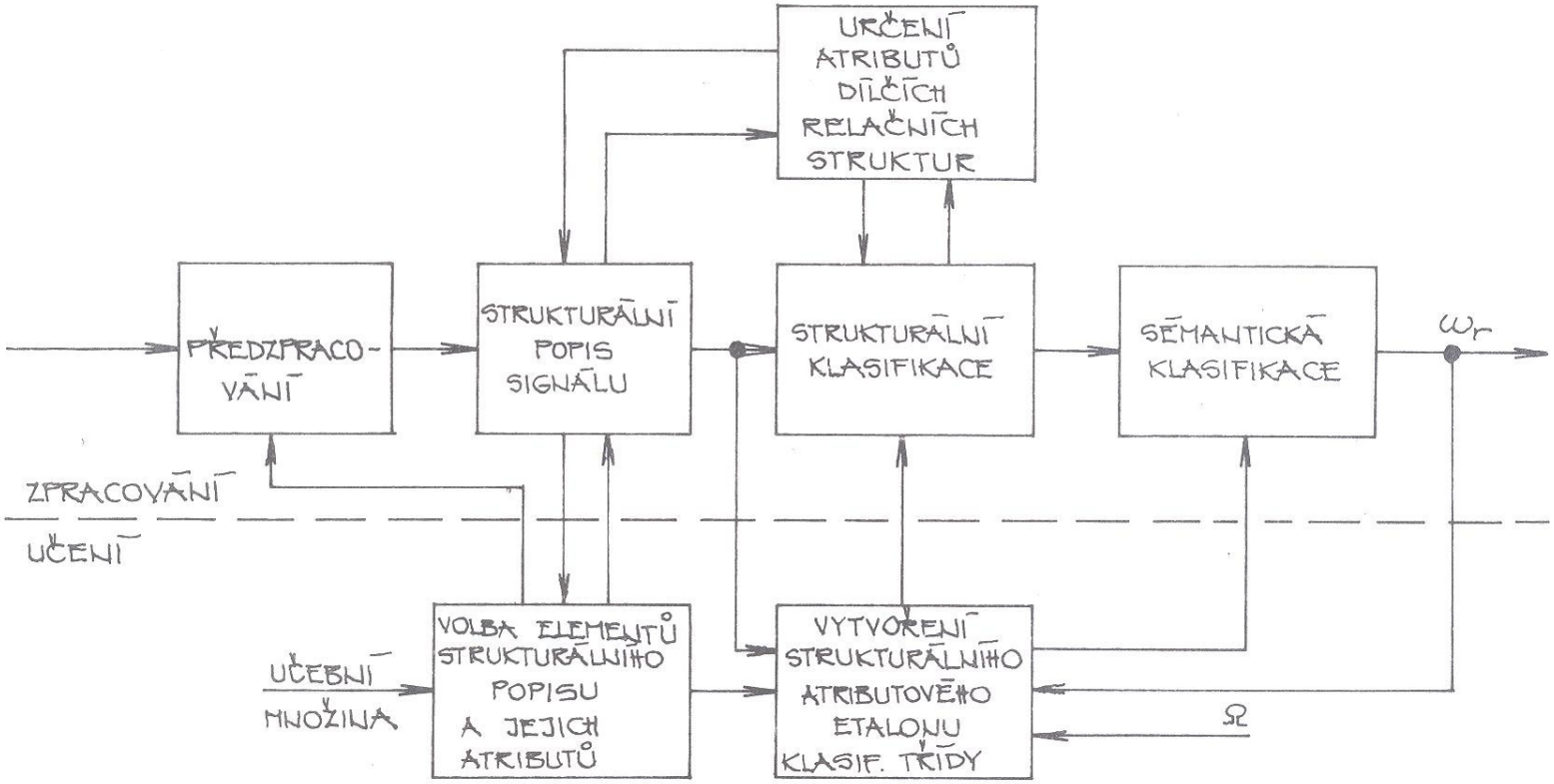
STRUKTURÁLNÍ POPIS



STRUKTURÁLNÍ POPIS



KOMBINOVANÝ STRUKTURÁLNÍ POPIS



Obr. 3.5 Blokové schéma atributového strukturálního klasifikátoru

REPREZENTACE KLASIFIKAČNÍ TŘÍDY

- výčtem relačních struktur

  (může být bohužel velký až nekonečný)

REPREZENTACE KLASIFIKAČNÍ TŘÍDY

- ☑ výčtem relačních struktur

  (může být bohužel velký až nekonečný)
- ☑ generátorem relačních struktur – gramatika
  - ➔ Gramatika je čtveřice $G = (V_n, V_t, P, S)$, kde V_n a V_t jsou konečné disjunktní množiny (abecedy), přičemž prvky množiny V_n nazývají neteminální pomocné symboly a prvky V_t terminální symboly, $S \in V_n$ je tzv. axiom gramatiky nebo také počáteční symbol a P je množina substitučních pravidel tvaru $\alpha \rightarrow \beta$, které definují způsob náhrady dílčí relační struktury α novou strukturou β.

REPREZENTACE KLASIFIKAČNÍ TŘÍDY

- výčtem relačních struktur

  (může být bohužel velký až nekonečný)
- generátorem relačních struktur – gramatika
  - Příklad gramatiky:

    G = ((A,B), (0,1), P, A), P = (A → 0B1, 0B→00B, B → „e“).

    Příklad generování řetezce:

    A → 0B1 → 00B1 → 000B1 → 0001

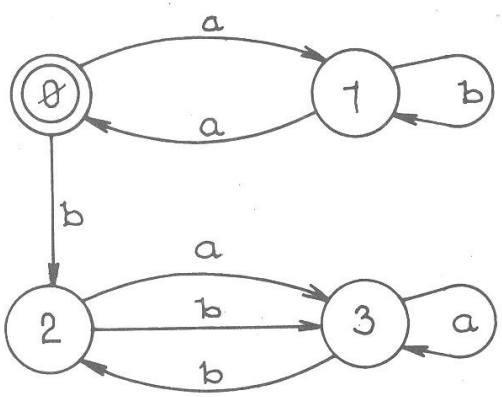
REPREZENTACE KLASIFIKAČNÍ TŘÍDY

- ☑ výčtem relačních struktur
  (může být bohužel velký až nekonečný)
- ☑ generátorem relačních struktur – gramatika
- ☑ příjemcem relačních struktur – automat
  - ➔ různé typy automatů podle charakteru relační struktury a substitučních pravidel – nejjednodušší konečný automat
  - ➔ Konečný stavový automat A je pětice A = (X, S, s_0, S_c, □), kde X je konečná vstupní abeceda, S je množina vnitřních stavů, s_0 je počáteční stav automatu, S_c je množina cílových stavů automatu a □: X x S → S je přechodová funkce.

REPREZENTACE KLASIFIKAČNÍ TŘÍDY

- ☑ výčtem relačních struktur
  (může být bohužel velký až nekonečný)
- ☑ generátorem relačních struktur – gramatika
- ☑ příjemcem relačních struktur – automat
  - ➔ Příklad konečného stavového automatu

δ	0	1	2	3
a	1	0	3	3
b	2	1	3	2



REPREZENTACE KLASIFIKAČNÍ TŘÍDY

- výčtem relačních struktur
  (může být bohužel velký až nekonečný)
- generátorem relačních struktur – gramatika
- příjemcem relačních struktur – automat
- ekvivalence gramatiky a automatu –
  gramatika a automat jsou ekvivalentní, pokud množina relačních struktur generovaná gramatikou a množina akceptovaná automatem jsou stejné

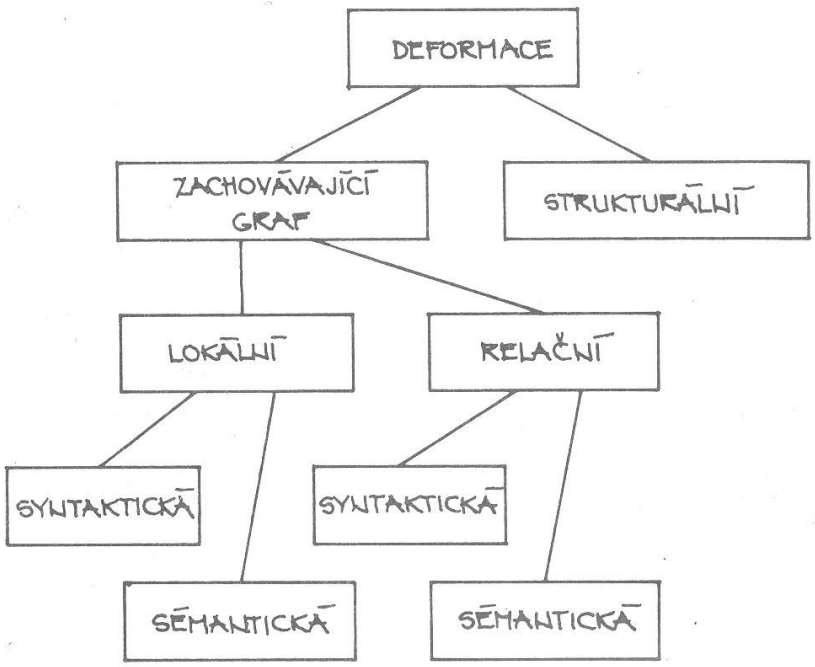
STRUKTURÁLNÍ KLASIFIKACE

nedeformované relační struktury

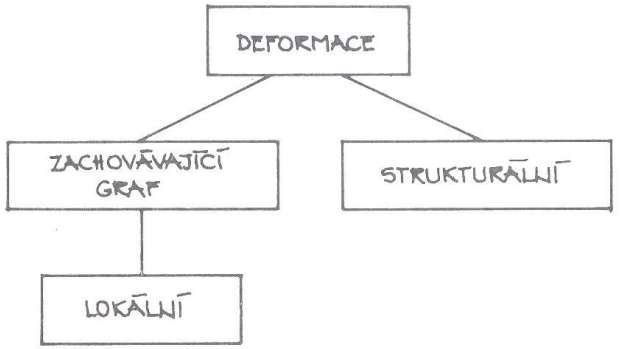
- ztotožnění s reprezentativními relačními strukturami;
- přijetí automatem

STRUKTURÁLNÍ KLASIFIKACE

deformované relační struktury

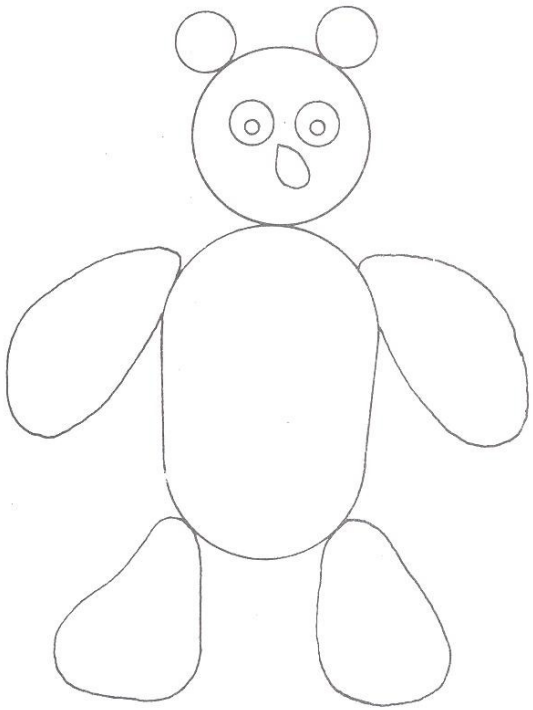


Obr. 3. 21 Obecné strukturální deformační schéma



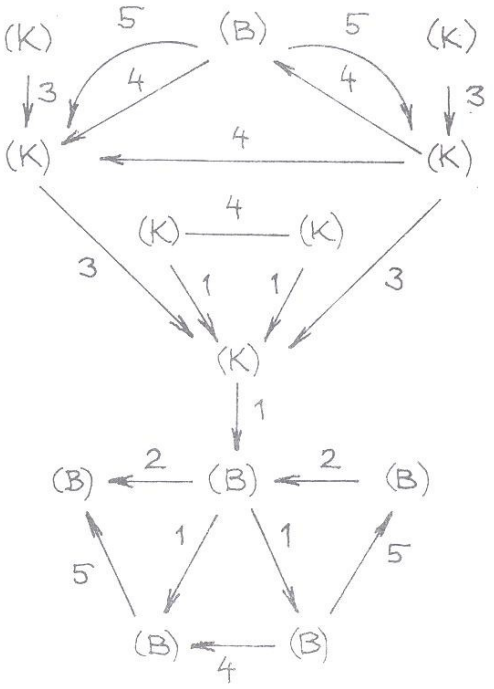
Obr. 3. 22 Deformační schéma pro řetězce bez sémantické informace

STRUKTURÁLNÍ POPIS



PRIMITIVA:
(K) – KOLEČKO
(B) – BRAMBORA

RELACE:
(1) – DOTÝKÁ SE SHORA
(2) – DOTÝKÁ SE ZLEVA
(3) – LEŽÍ UVNITŘ
(4) – LEŽÍ VLEVO OD
(5) – LEŽÍ POD



Obr. 3.1 Primitiva, relace a relační struktura čarové kresby

STRUKTURÁLNÍ KLASIFIKACE

deformované relační struktury

- ☑ podle vzdálenosti od etalonu

jak vzdálenost určíme?

STRUKTURÁLNÍ KLASIFIKACE

deformované relační struktury

- ☑ podle vzdálenosti od etalonu

jak vzdálenost určíme?

vzdálenost vs. metrika

?

METRIKA - VZDÁLENOST

Metrický prostor je neprázdná množina X spolu s funkcí $\rho: X \quad X \to R$ splňující:

- 1. totožnost: $\rho(x, y) = 0 \Leftrightarrow x = y, \forall x,y \in X;$
- 2. symetrie: $\rho(x, y) = \rho(y, x), \forall x,y \in X;$
- 3. trojúhelníková nerovnost:

$$\rho(x, z) \leq \rho(x, y) + \rho(y, z), \forall x,y,z \in X.$$

ρ je nezáporná funkce.

Funkci ρ nazýváme metrika na X.

Vzdálenost je hodnota určená podle metriky.

STRUKTURÁLNÍ VZDÁLENOST

V případě řetězců lze deformační vlivy vyjádřit (na úrovni primitiv) trojicí tzv. elementárních deformačních transformací – eliminací, substitucí a inzercí

a) eliminační deformační transformace

$$T_E: \quad \omega_1 a \omega_2 \xrightarrow{W_E(a)} \omega_1 \omega_2;$$

b) substituční deformační transformace

$$T_S: \quad \omega_1 a \omega_2 \xrightarrow{W_S(a,b)} \omega_1 b \omega_2;$$

c) inzerční deformační transformace

$$T_I: \quad \omega_1 \omega_2 \xrightarrow{W_I(b)} \omega_1 b \omega_2,$$

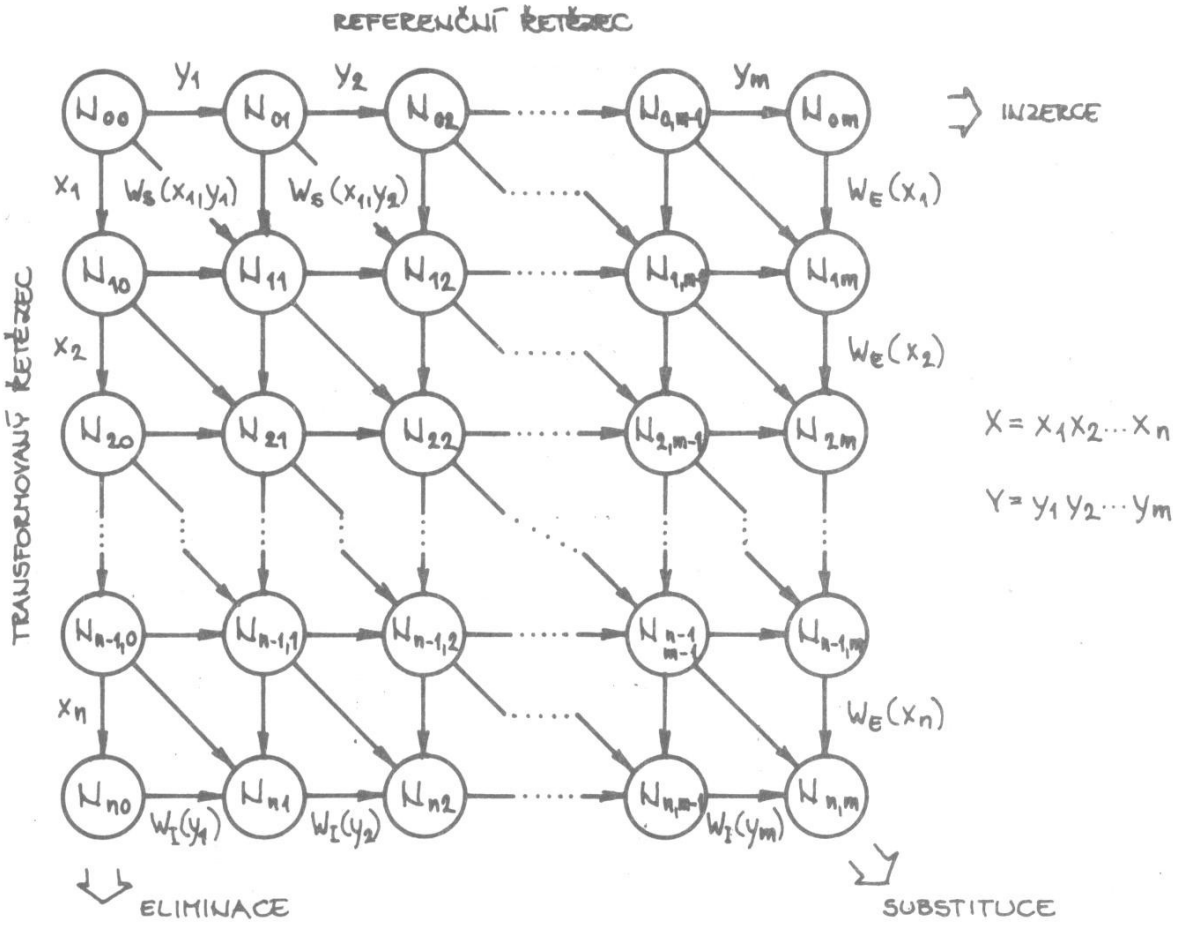
VÁHOVANÁ LEVENŠTEJNOVA METRIKA

Je-li $\mathcal{S} = \{T_1, T_2, \ldots, T_n\}$, $n \geq 0$, $T_i \in \{T_E, T_S, T_I\}$ posloupnost elementárních deformačních transformací taková, že pro libovolná konečná slova X,Y nad abecedou $\mathcal{V}_{te} = \mathcal{V}_t \cup \{"e"\}$ je $Y = \mathcal{S}(X)$, pak váhovaná Levenštejnova metrika je definována vztahem

$$d_{WL}(X,Y) = \min_{\mathcal{S}} \left\{ \sum_{\forall a\ [T_E(a) \in \mathcal{S}]} W_E(a) + \sum_{\forall a,b\ [T_S(a,b) \in \mathcal{S}]} W_S(a,b) + \sum_{\forall b\ [T_I(b) \in \mathcal{S}]} W_I(b) \right\} \quad (3\text{-}28)$$

Aby byly splněny všechny tři základní axiomy metrik (axiom totožnosti, symetričnosti a trojúhelníková nerovnost) je třeba, aby platilo $W_I(a) = W_E(a)$ a $W_S(a,b) = W_S(b,a)$ pro všechny terminální symboly a,b. Metrika splňující tyto požadavky je pravá metrika.

VÁHOVANÁ LEVENŠTEJNOVA METRIKA



Obr. 3. 23 Výpočet vdhované Levenštejnovy vzdálenosti

DALŠÍ STRUKTURÁLNÍ METRIKY

- řetězce
  - prostá (neváhovaná) Levenštejnova metrika
  - Hammingova metrika
- stromy

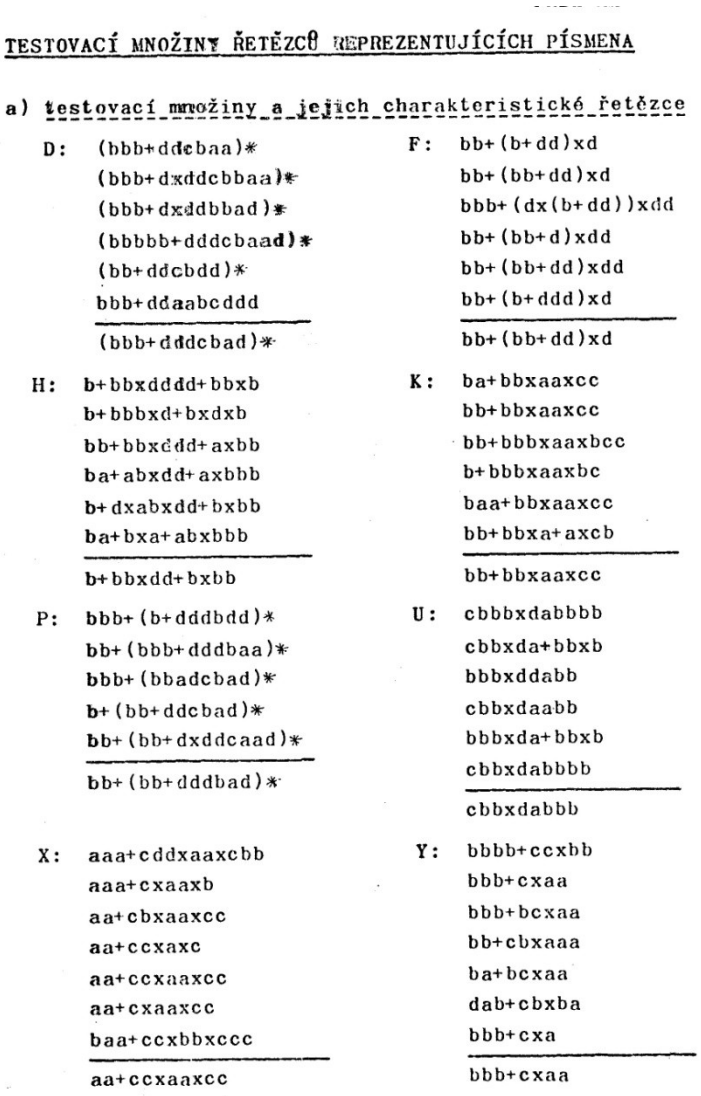
:

:

KLASIFIKACE DEFORMOVANÝCH STRUKTUR

- výpočet vzdálenosti mezi reprezentativními strukturami (etalony) klasifikační třídy a klasifikovanou strukturou;
- začlenění deformačních pravidel do substitučních pravidel gramatiky (resp. automatu);

VYTVOŘENÍ STRUKTURÁLNÍHO ETALONU MNOŽINY PÍSMEN



TESTOVACÍ MNOŽINY ŘETĚZCŮ REPREZENTUJÍCÍCH PÍSMENA

a) testovací množiny a jejich charakteristické řetězce

D:
(bbb+ddcbaa)*
(bbb+dxddcbbaa)*
(bbb+dxddbbad)*
(bbbbb+dddcbaad)*
(bb+ddcbdd)*
bbb+ddaabcddd

(bbb+dddcbad)*

F:
bb+(b+dd)xd
bb+(bb+dd)xd
bbb+(dx(b+dd))xdd
bb+(bb+d)xdd
bb+(bb+dd)xdd
bb+(b+ddd)xd
---
bb+(bb+dd)xd

H:
b+bbxdddd+bbxb
b+bbbxd+bxdxb
bb+bbxddd+axbb
ba+abxdd+axbbb
b+dxabxdd+bxbb
ba+bxa+abxbbb
---
b+bbxdd+bxbb

K:
ba+bbxaaxcc
bb+bbxaaxcc
bb+bbbxaaxbcc
b+bbbxaaxbc
baa+bbxaaxcc
bb+bbxa+axcb
---
bb+bbxaaxcc

P:
bbb+(b+dddbdd)*
bb+(bbb+dddbaa)*
bbb+(bbadcbad)*
b+(bb+ddcbad)*
bb+(bb+dxddcaad)*
---
bb+(bb+dddbad)*

U:
cbbbxdabbbb
cbbxda+bbxb
bbbxddabb
cbbxdaabb
bbbxda+bbxb
cbbxdabbbb
---
cbbxdabbb

X:
aaa+cddxaaxcbb
aaa+cxaaxb
aa+cbxaaxcc
aa+ccxaxc
aa+ccxaaxcc
aa+cxaaxcc
baa+ccxbbxccc
---
aa+ccxaaxcc

Y:
bbbb+ccxbb
bbb+cxaa
bbb+bcxaa
bb+cbxaaa
ba+bcxaa
dab+cbxba
bbb+cxa
---
bbb+cxaa

VYTVOŘENÍ STRUKTURÁLNÍHO ETALONU MNOŽINY PÍSMEN

SHRNUTÍ – POKUD MOŽNO CO NEJOBECNĚJI

- základní klasifikační úloha je zatřídit (z hlediska prostoru i času) (matematický, abstraktní) popis daného klasifikovaného objektu do odpovídající třídy/kategorie;
- děje se to na základě klasifikačního pravidla, pomocí kterého je definována klasifikační třída;
- klasifikační třída může být definována:
  - výčtem prvků do ní patřících;
  - vzdáleností/podobností (od) vzorů té které třídy;
  - hranicemi, vymezujícími prostor dané třídy.