Algoritmy a struktury neuropočítačů
ASN – P7

- Syntéza neuronových sítí
- Optimalizace struktury
- Klestění neuronové sítě
- Výběr vstupních dat

Syntéza neuronových sítí

N je počet neuronů

- dělení soustav
  - kanonické $N=N_{krit}$
  - subkanonické $N < N_{krit}$
  - redundantní $N > N_{krit}$

N_{krit} většinou předem neznáme

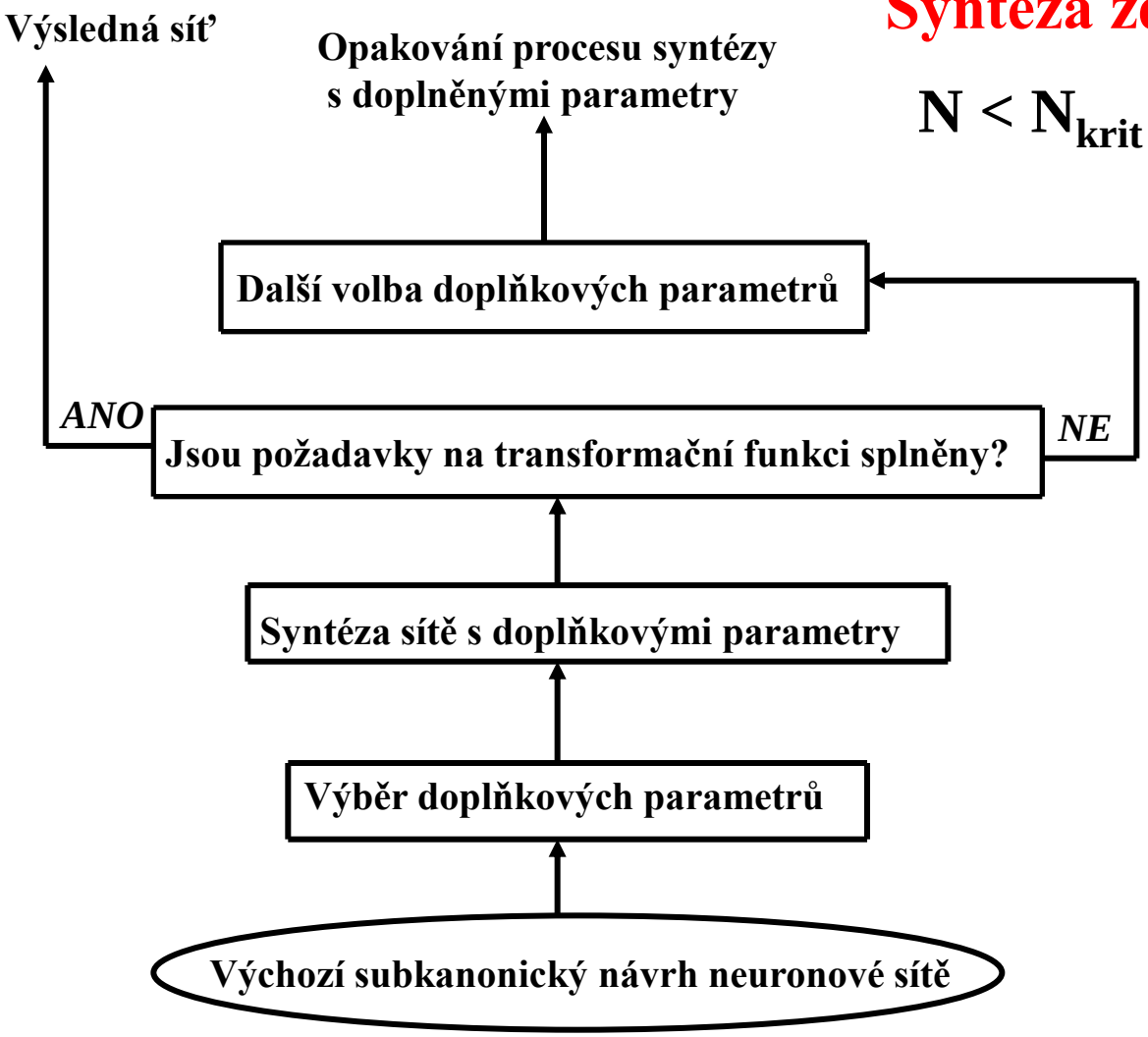
- Syntéza „zdola“ $N < N_{krit}$

nevýhoda : *při malé volbě N_{krit} je proces dlouhodobý*

otázky:
1. *které prvky přidat*
2. *na kterém místě je přidat*

Syntéza zdola

$N < N_{krit}$



Syntéza shora

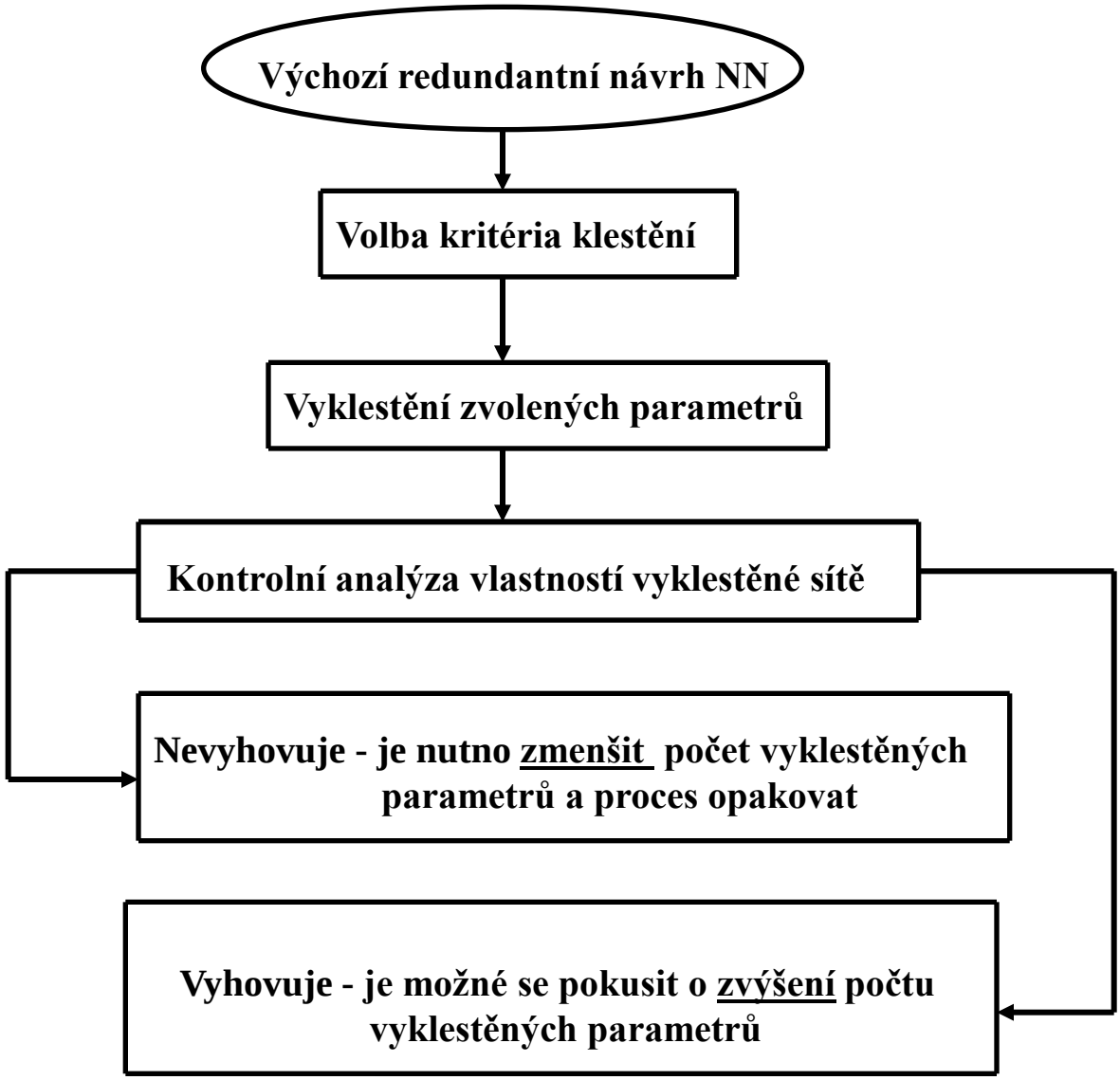
$N > N_{krit}$

nevýhoda: neekonomické řešení

redukce → klestění (*pruning*)

kritérium pro volbu parametrů

biologické sítě
- vysoce redundantní (*nekanonické, nadbytečné)*
- nízká citlivost


Výchozí redundantní návrh NN
Volba kritéria klestění
Vyklestění zvolených parametrů
Kontrolní analýza vlastností vyklestěné sítě
Nevyhovuje - je nutno zmenšit počet vyklestěných parametrů a proces opakovat
Vyhovuje - je možné se pokusit o zvýšení počtu vyklestěných parametrů

Klestění a optimalizace NN

Většina NN je vysoce redundantní (*obsahuje nadbytečný počet neuronů)*

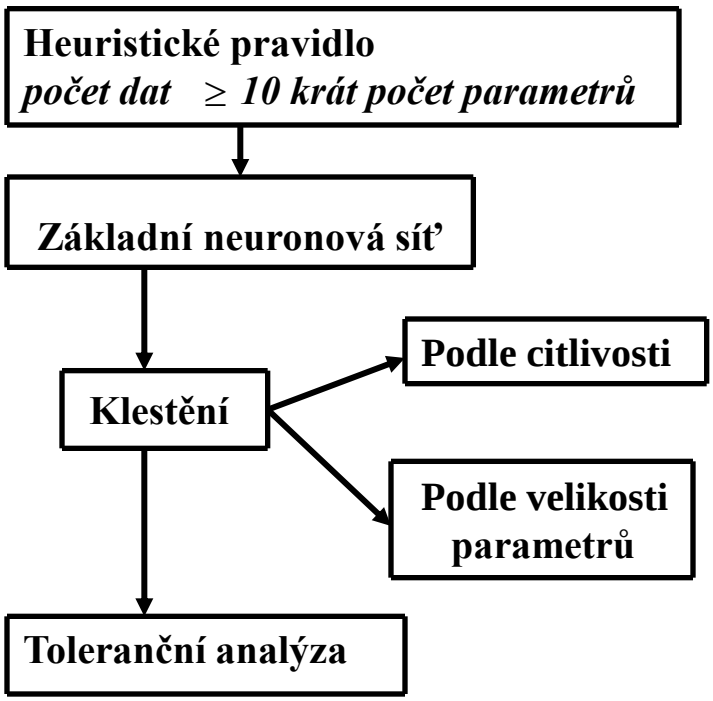
Co je klestění ?

Postupné nebo jednorázové vyloučení nadbytečných prvků a synaptických vazeb.

Co klestit ?

synaptické spoje
neurony
parametry aktivačních funkcí
prahy

Optimalizace



Optimalizace NN

- **výběr parametrů (*feature selection*)**
  určování důležitých parametrů, které mají vliv na nějakou významnou vlastnost systému (prozodii, zdravotní stav pacientů)

Jak? ***pomocí statistických metod – nejčastěji (PCA, K-means,…)***
pomocí UNS - nejlépe
pomocí metod matematické logiky

- **získávání znalostí z dat (*data mining*) –**
  z experimentálních dat

Data Mining

- **proces umožňující nalezení nových informací nebo ověření vzájemných vztahů ve velkých databázích, např. v obchodě bankovnictví, pojišťovnictví…**

- **pomocí neuronových sítí (často Kohonenových)**
- **pomocí fuzzy – logických obvodů, kombinovaných fuzzy neuronových sítí**
- **pomocí metody založené na matematické logice a statistice GUHA**

Optimalizace struktur

výběr množiny vzorů
správné přiřazení hodnot parametrů

protiklad : potřeba velkého počtu vzorů - velký počet vstupních neuronů
co nejkratší doba trenování

Počet neuronů ve vstupní vrstvě:

vyhledávání markerů - hledání parametrů nejvíce ovlivňujících výstupní hodnoty; souvisí s velikostí vstupní vrstvy

Počet neuronů ve výstupní vrstvě:

je dán požadavkem na výstupní veličiny

Počet neuronů ve skrytých vrstvách:

není přesně určen

obvykle větší počet na začátku učení, po naučení se některé extrahují; během tohoto procesu je třeba sledovat vývoj chybové funkce (nesmí dojít k jejímu vzrůstu)

Hledáme kanonické sítě.

PCA - principal component analysis

- **analýza hlavních komponent - umožňuje redukovat dimenze dat (minimální ztráta informací)**

Vytváří nové parametry dat - ortogonální lineární souřadnicový systém

- **první parametr (PCA 1 – principal component 1)**

obsahuje největší množství informací původních dat

- **poslední parametr - obsahuje nejmenší množství informací**
- **Metoda hledá takový parametr, který má v původních parametry největší rozptyl.**
- **Nové parametry jsou vypočteny ze všech původních parametrů - z vlastních vektorů kovariační matice původních dat.**

korelační tabulka porovnává chyby detekce

35% segmentů „o“ detekováno jako „u“

	a	e	i	o	u
a	99,8	0,2	0	0	0
e	4,6	95	0	0	0
i	0	0	99,8	0	0,2
o	0	0	0,7	65	35
u	0	0	9,1	0	91

- **PCA garantuje nezávislost dat pouze pro normální rozdělení**

PCA … Karhunen-Loève transformace (KLT)

Klestění synaptických vah

- **váhy s minimálními hodnotami**
- **podle citlivostí (*velmi obtížné*) – *viz [SAN05] - SpeechLab***

Postup klestění :

- **nalezení vah alespoň o řád menší než je jejich střední hodnota**
- **pokud takové váhy existují, vyklestit je a síť přetrénovat**
- **vlastnosti se nezhorší (podstatně), pak rozbor citlivostí na parametry s nejmenšími hodnotami**
- **parametry s malými citlivostmi vyklestíme**
- **přeučení sítě** *z hodnot vah a prahů z předchozího kroku*
- **opakování procesu**

Pozor! ***I malé změny některého parametru (malé citlivosti) mohou mít sice malý vliv na výslednou transformační funkci, ale úplné vyklestění může zásadně ovlivnit konečnou funkci.***

Vliv pracovního bodu na klestění vah

- **Důvod: nelinearita aktivační funkce – problém oblast saturace**

Poloha pracovního bodu:

$$\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y[n]$$

Odchylka pracovního bodu:

$$\sigma_y = \left(\frac{1}{N-1}\sum_{n=1}^{N} [\bar{y} - y[n]]^2 \right)^{1/2}$$

Velká odchylka – neuron přebuzen – malá síť

Velmi malá odchylka – význam neuronu je potlačen

redundantnost, neuron lze vyklesttit

Redukce vstupních parametrů

GUHA metoda

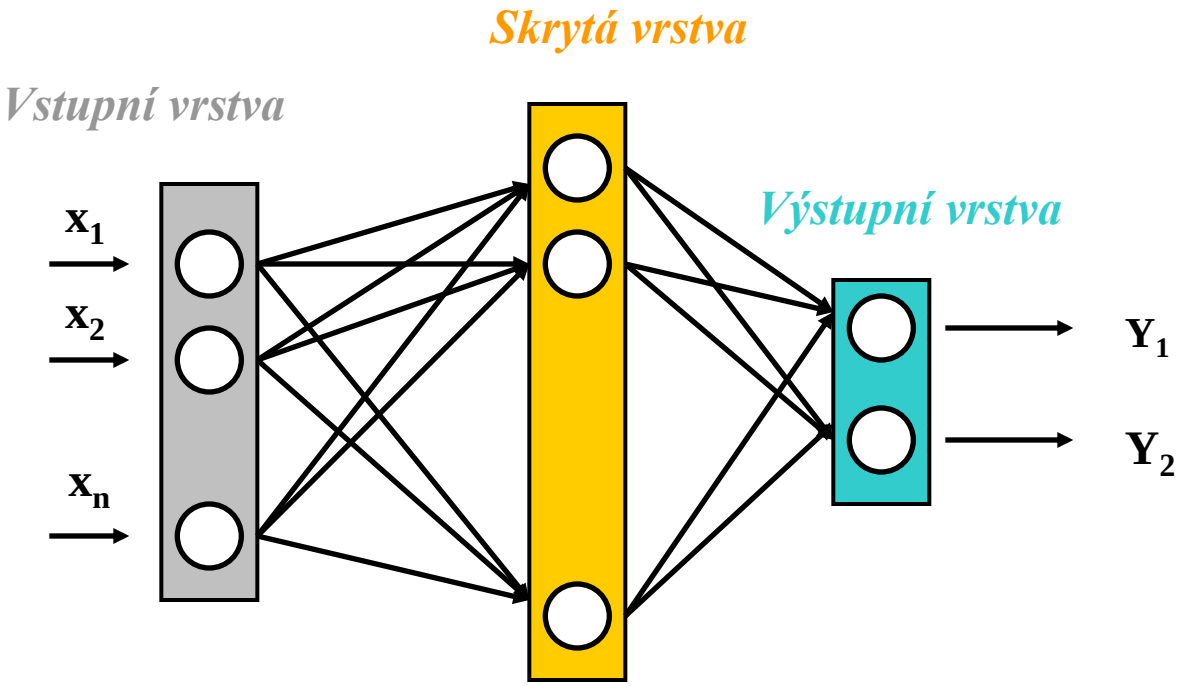
General Unary Hypotheses Automaton

Zpracovává data formou rectangulární matice:

různé objekty, parametry - řádky (např. hlásky)

zkoumané vlastnosti - sloupce (např. základní kmitočet, trvání hlásky)

Optimalizace struktury



Data Mining (GUHA)

Klestění

GUHA

Principy byly formulovány v r.1966 v publikaci [HAJ66] .
Autory byli P. Hájek, I.Havel a Chytil.
Autory teoretického a praktického rozvoje základních principů metody jsou P. Hájek a T.Havránek (1978).

GUHA je založena na počítačovém generování všech možných hypotéz, které berou v úvahu souvislosti mezi zpracovávanými daty.

Hypotézy popisují vztahy mezi vlastnostmi objektů.

Počet hypotéz vypovídá o důležitosti jednotlivých parametrů.

Během let došlo k mnoha změnám:

- **vylepšení počítačů**
- **praktické aplikace v nejrůznějších oblastech lidského počínání.**

Není masově rozšířena pro svoji náročnost.

GUHA zpracovává data ve formě rectangulární matice, řádky odpovídají různým objektům (např. hláskám), sloupce odpovídající zkoumaným vlastnostem (např. základní kmitočet, trvání hlásky).

Rozdělení vlastností:

- **antecedenty** ***(např. charakteristické vlastnosti češtiny)***

  → **vstupní parametry**

- **succedenty** ***(např. F0 a trvání)***

  → **výstupní parametry**

Program generuje a porovnává hypotézy o vztazích

$$A \longrightarrow S$$

A … souvislost mezi antecedenty
S … souvislost mezi succedenty
→ … implikace (pro odhad podmíněné pravděpodobnosti P(S|A)

- **Je třeba specifikovat počet elementů v antecedentech a succedentech.**
- **V souhlasu s počtem věrohodných hypotéz můžeme určit důležitost vstupních parametrů.**
- **Definice jednotlivých kvantifikátorů je založena na statistických testech, ale jejich interpretace není statistická.**

Příklad - Guha

Parametry charakterizující češtinu

fokus

P1	P11	P21	Identifikace pausy.
P2	P12	P22	Identifikace přízvuku.
P3	P13	P23	Identifikace jádra slabiky. [1]
P4	P14	P24	Identifikace znamínka.
P5	P15	P25	Identifikace fonému.
P6	P16	P26	Výška samohlásky.
P7	P17	P27	Délka samohlásky.
P8	P18	P28	Znělost souhlásky .
P9	P19	P29	Způsob tvoření souhlásky.
P10	P20	P30	Počet fonémů ve slově obsahujícím focus-foném

červené – zůstávají, černé – vynechávají se

[1] V češtině není jednoznačně definována slabika, jádro identifikujeme podle samohlásky, diftongu nebo sonoru

Počet hypotéz v závislosti na korelacích mezi vstupními parametry – příklad .

	P_2	P_3	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}	P_{17}	P_{18}	P_{19}	P_{20}	P_{21}	P_{24}	Σ
P_2	52	81	67	69	68	75	77	56	34	44	30	36	22	28	29	32	32	20	65	72	166
P_3	82	99	145	159	139	167	169	86	51	32	28	54	21	20	22	30	30	10	99	102	327
P_5	51	111	62	133	96	112	116	55	75	29	34	80	25	15	17	37	37	15	64	64	286
P_6	77	151	156	125	113	113	118	82	134	31	69	139	44	33	40	86	86	15	97	97	397
P_7	71	137	118	132	77	132	135	74	46	30	25	49	16	19	19	30	30	11	91	91	258
P_8	79	155	125	114	124	131	152	81	36	62	52	39	46	68	56	38	38	38	93	99	331
P_9	79	157	137	114	128	152	131	81	44	62	58	47	51	68	57	43	43	38	93	98	358
P_{10}	50	66	56	57	60	58	58	45	55	22	1	55	0	0	0	1	1	38	63	65	111
P_{11}	29	58	82	86	47	53	56	45	209	37	122	253	120	107	87	165	163	37	30	39	507
P_{12}	18	38	50	41	27	33	35	23	105	60	118	111	100	97	97	123	125	83	12	18	297
P_{13}	21	26	44	44	20	33	34	21	138	100	132	137	155	180	162	200	199	103	21	26	371
P_{14}	34	86	104	112	67	87	90	51	242	42	133	257	130	118	94	172	171	44	40	47	612
P_{15}	21	24	34	36	23	26	26	19	106	71	128	113	78	140	117	137	135	73	21	28	279
P_{16}	28	56	44	35	44	80	80	25	95	91	164	98	145	130	138	130	131	92	30	35	348
P_{17}	10	36	29	26	23	53	53	7	83	88	160	84	127	149	107	157	155	92	9	14	319
P_{18}	3	4	26	39	6	0	4	1	166	112	207	164	159	161	168	177	195	116	0	3	357
P_{19}	3	4	26	36	6	0	4	1	177	112	209	175	169	163	170	197	177	116	0	3	370
P_{20}	3	0	0	0	0	0	0	21	30	67	87	28	74	77	80	79	79	58	0	3	114
P_{21}	88	98	87	100	99	95	95	86	48	41	27	48	25	27	23	24	25	17	93	108	183
P_{24}	85	112	100	115	107	111	111	89	76	46	44	78	45	53	43	49	50	27	94	105	284
Σ	107	229	248	263	192	259	268	123	345	145	272	242	265	227	284	287	152	152	112	131	1216

P11 Identifikace pausy.

P14 Identifikace znamínka (větší počet hypotéz)

Počet hypotéz o vztahu vstupních a výstupních parametrů

P	#hyp. F_0	# hyp. D	# hyp. celkem	P	#hyp. F_0	# hyp. D	# hyp. celkem
P_1	124	106	178	P_{16}	442	430	650
P_2	223	220	332	P_{17}	429	414	631
P_3	253	231	361	P_{18}	451	407	640
P_4	158	133	235	P_{19}	428	400	613
P_5	217	198	311	P_{20}	322	351	504
P_6	300	253	412	P_{21}	408	368	569
P_7	258	237	367	P_{22}	114	96	156
P_8	293	272	419	P_{23}	160	126	205
P_9	301	272	427	P_{24}	450	412	633
P_{10}	209	205	309	P_{25}	194	165	263
P_{11}	222	198	333	P_{26}	201	175	276
P_{12}	339	348	517	P_{27}	140	121	193
P_{13}	467	418	664	P_{28}	72	57	94
P_{14}	265	242	400	P_{29}	73	47	87
P_{15}	375	354	549	P_{30}	68	54	88

P … parametr

Limit je 300 hypotéz, lze eliminovat parametry P1, P4, P22, P23 a všechny od P25 do P30

Vyloučení parametru podle počtu hypotéz

~~P_{11}~~	P_{14}	19 vstupních parametrů
P_{13}	~~P_1~~	18 vstupních parametrů - optimální
~~P_{19}~~	P_{13}	17 vstupních parametrů

Nutné vždy ověřit!!!

Důsledky optimalizace

UNS	*# neuronů ve vrstvách*	*# neuronů*	*# vah*
Počáteční stav	30-25-2	57	800
Optimální stav	18-22-2	42	440

Literatura:

[HAJ95] Hajek,P.,Sochorova,A.,Zvarová,J.:GUHA for personal computers., *Computational Statistics and Data Analysis*, Vol.19, 1995, North Holland, pp.149-153.

[SEB01] Sebesta,V.,Tuckova,J.:Application of Feature Extraction in Text-to-Speech Processing. Proc. of ICANGA-Int. Conf. on Artificial Neural Nets and Genetic Algorithms, Springer-Verlag/Wien, ISBN 3-211-83651-9, Prague, April 2001, pp.145-148.

[SAN05] Santarius, J.: Systémy pro modelování prozódie syntetické řeči. Disertační práce, FEL ČVUT v Praze, 2005.

Data Mining

- **dobývání znalostí z dat**
- **dolování dat**
- **vytěžování dat**

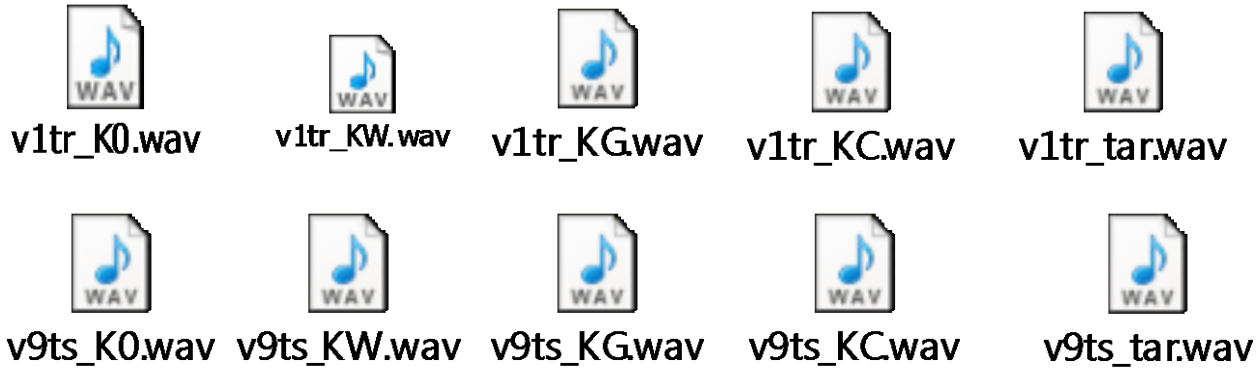
Získávají se skryté, předem neznámé informace z dat, cílem je praktická využitelnost výsledků.

Úlohy Data Miningu (nejčastější):

- **klasifikace**
- **predikce**
- **shluková analýza**
- **analýza závislostí**

Metody:

- **expertní systémy a rozhodovací stromy**
- **grafické a statistické techniky**
- **neuronové sítě (např. variantyKSOM)**
  pro mnoho typů úloh (i složitých)
  dlouhá doba učení
  potřeba velkého množství vzorů
- **genetické algoritmy**

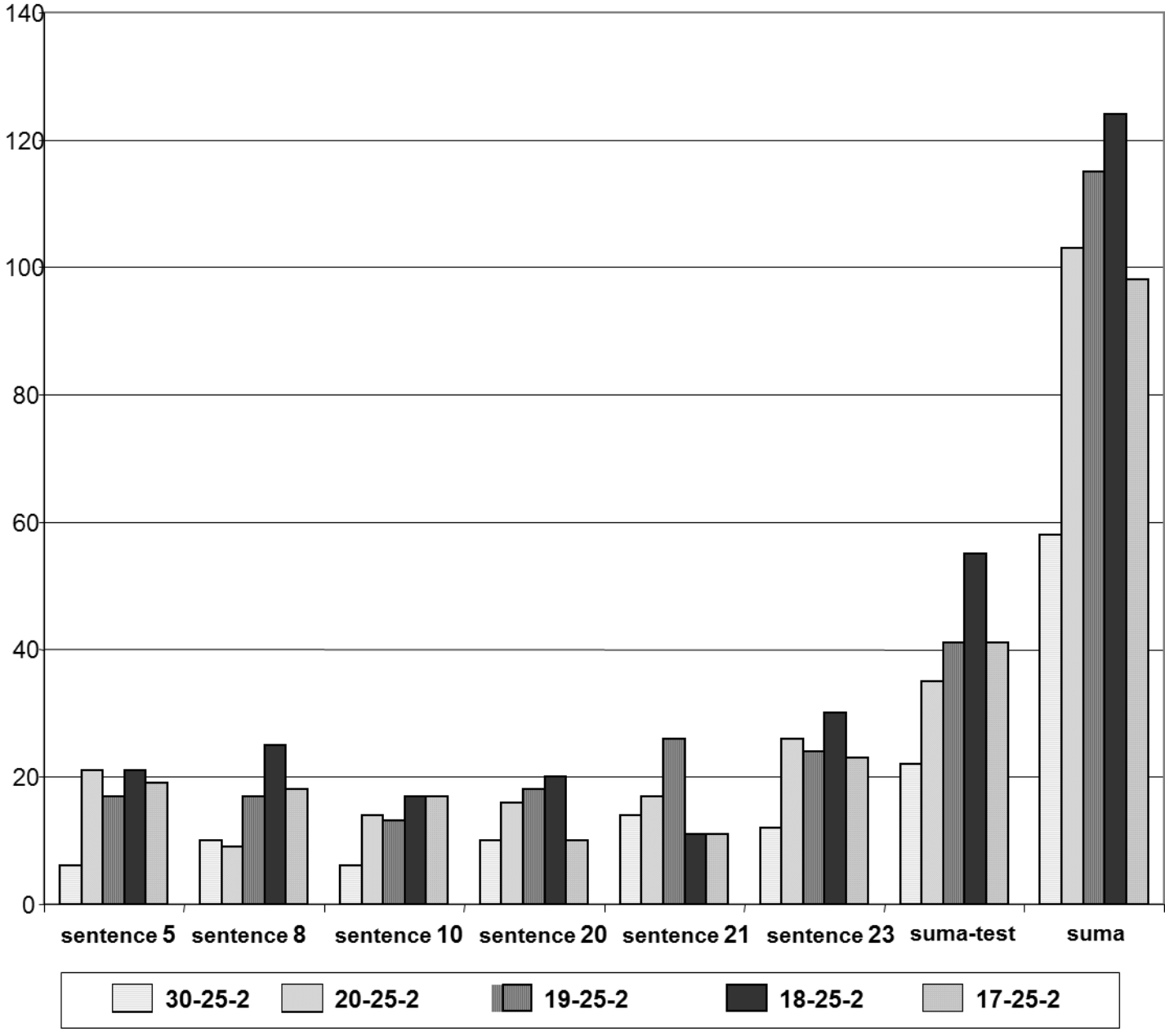


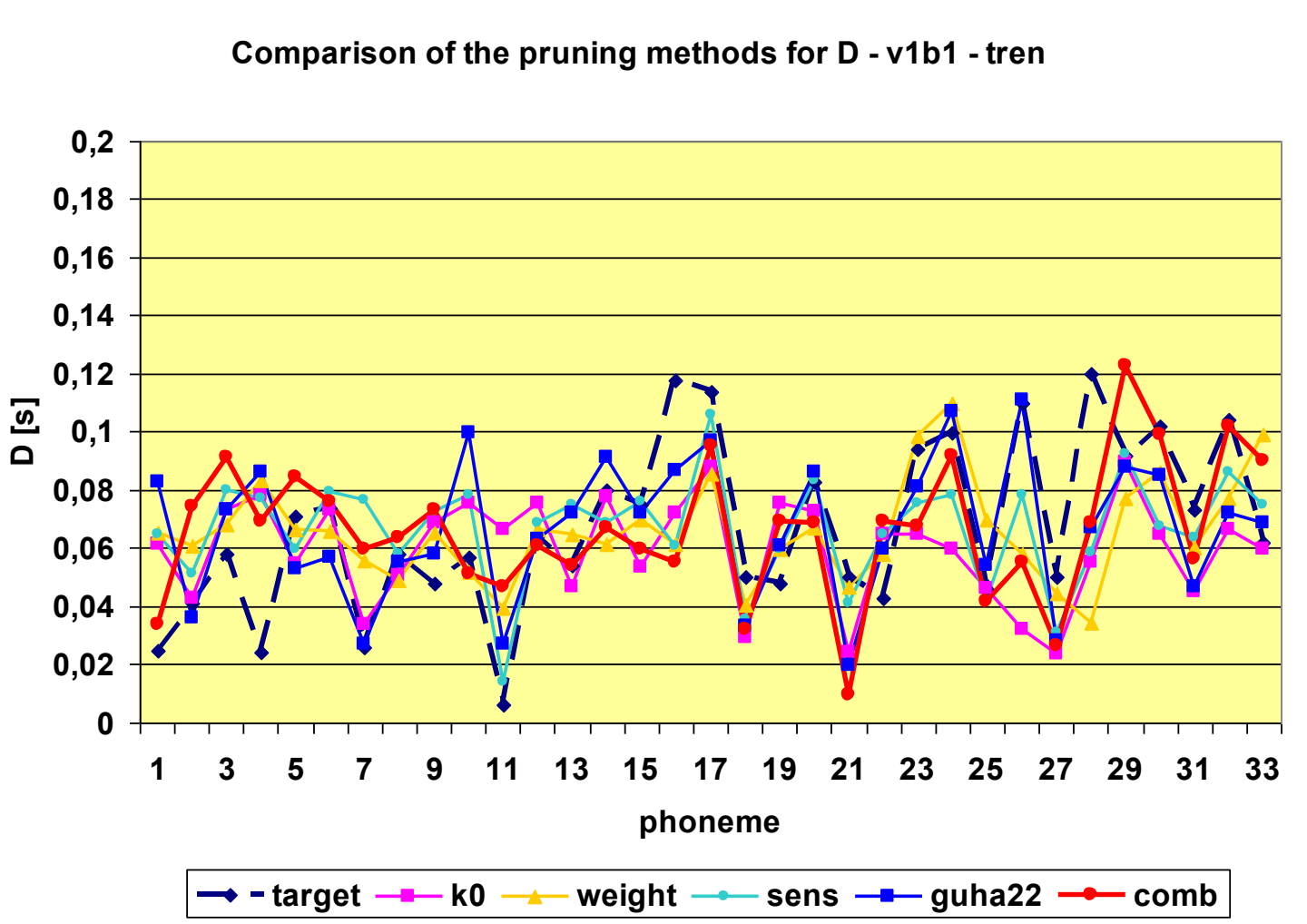
v1 ***Předpověď počasí na noc a zítřek.***

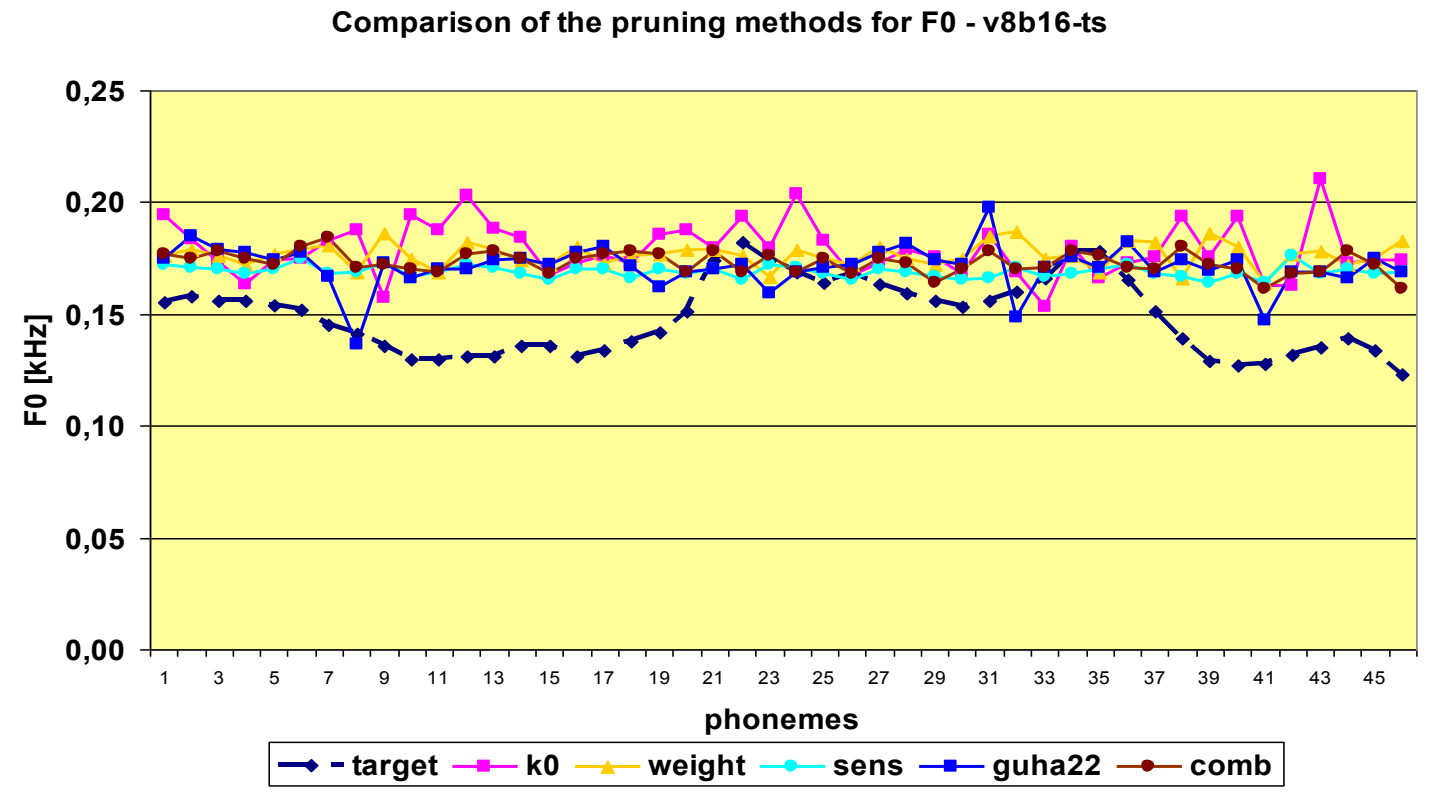
v9 ***To byly zprávy Českého rozhlasu 1 – Radiožurnálu.***

1. **Prozodie trénovaná neklestěnou NN**
2. **Prozodie trénovaná NN klestěnou pomocí synaptických vah**
3. **Prozodie trénovaná NN klestěnou pomocí GUHA**
4. **Prozodie trénovaná NN klestěnou současnou kombinací dvou metod**
5. **Řeč syntetizovaná pomocí target hodnot F0 a trvání fonému**

Histogram závislosti počtu nejlepších promluv na počtu parametrů




Comparison of the pruning methods for D - v1b1 - tren
D [s]
0,2
0,18
0,16
0,14
0,12
0,1
0,08
0,06
0,04
0,02
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33
phoneme
target
k0
weight
sens
guha22
comb


Comparison of the pruning methods for F0 - v8b16-ts
F0 [kHz]
0,25
0,20
0,15
0,10
0,05
0,00
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45
phonemes
target
k0
weight
sens
guha22
comb