Klasifikace diabetes u indiánek kmene Pima

Martin Bruchanov

***Abstrakt*— Zkoumání výskytu diabetes na základě kritérií stanovených World Health Organization. Parametry jako např. množství glukózy v krevní plazmě, atd. byly zkoumány u indiánských žen kmene Pima, starších 21 let, u populace žijící blízko Phoenixu v Arizoně, USA.**

I. Zadání

Prozkoumat vlastnosti učení neuronové sítě typu *Backpropagation* s různým počtem neuronů a vnitřních vrstev.

II. Úvod

Určování výskytu diabetes je rozhodovací problém mezi dvěma třídami (ano/ne), který využívá nasledujících 8 číselných parametrů:

1) počet těhotenství,
2) koncentrace plazmové glukózy 2 hodiny po provedení testu tolerance na glukózu,
3) diastolický krevní tlak (mm·Hg),
4) tloušťka tukové řasy pod tricepsem (mm),
5) 2 hodinový sérový inzulín (mu·U·ml^{-1}),
6) body mass index BMI (hmotnost v kg / (výška v m)^2),
7) Diabetes pedigree function,
8) věk.

Vstupní soubor obsahuje 768 položek bez chybějících parametrů. Třídy jsou 0 a 1 (hodnota 1 interpretována jako pozitivní test pro výskyt diabetes).

TABULKA I
ZATOUPENÍ TŘÍD

Třída	Počet výskytů
0	500
1	268

III. Experimenty

Vstupní data bylo nutné před zpracování sítí normalizovat. Normalizace byla provedena vůči středním hodnotám jednotlivých parametrů vstupního souboru.

TABULKA II
PRŮMĚRNÉ PARAMETRY SOUBORU

Parametr	1.	2.	3.	4.	5.	6.	7.	8.
Střed. hod.	3,8	120,9	69,1	20,5	79,8	32,0	0,5	33,2

A. Učení různých sítí

Vstupní soubor byl rozdělen na dvě části:

- Soubor vzorků pro učení sítě, celkem 512 (2/3 celkového souboru):
  - počet vzorků 0: 327 (63,8 %)
  - počet vzorků 1: 185 (36,1 %)
- Soubor vzorků pro testování sítě, celkem 256 (1/3 celkového souboru):
  - počet vzorků 0: 173 (67,5 %)
  - počet vzorků 1: 83 (32,4 %)

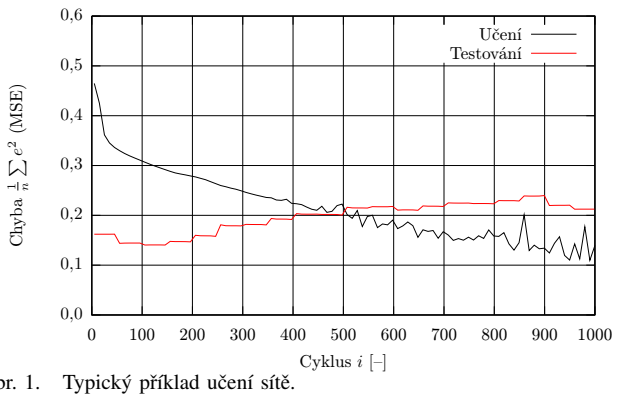
Parametry BP-sítě:

- $\eta = 0,2, d_{max} = 0.1$
- Chyba: $\frac{1}{n}\sum e^2$ (MSE)

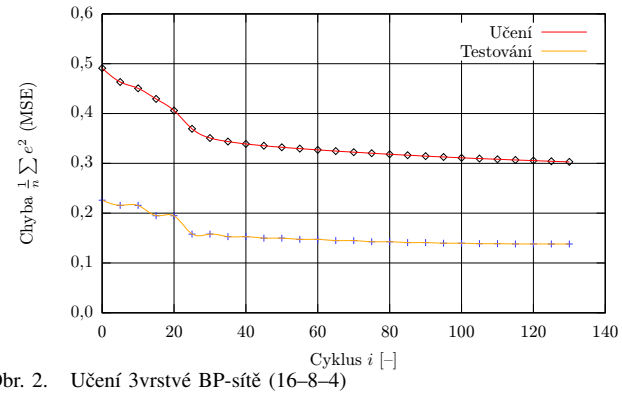
Každá síť obsahovala 8 vstupních neuronů, 1 výstupní a žádnou nebo více vnitřních vrstev různých konfigurací. Síť byla nejprve inicializována náhodnými váhami a byla zjištěna *počáteční chyba*.

Další sledované parametry sítě jsou počet učících cyklů (dokud není síť evidentně přeučená a dalšími cykly navíc nedochází ke zlepšení chyby validačních vzorků) a chyba v momentě nejlepšího naučení.

1) Příklad průběhu učení sítě: Na obr. 1 je vidět typický průběh učení sítě typu backpropagation. Ačkoliv je ukázáno učení pro 1000 cyklů, bylo pro zkoumání rozhodujících prvních asi 150 cyklů. V tomto průběhu rychle klesá chyba učení a klesá i chyba testování na nezávislém vzorku. Přibližně od 150 cyklu začíná chyba u testovacího vzorku vzrůstat – v tomto místě je dosaženo nejlepšího naučení. Ačkoliv s dalšími cykly učení chyba učení nadále klesá, chyba u testovacích dat se buď nemění a nebo dokonce vzrůstá – síť je evidentně *přeučená* a příliš ovlivněná učícími vzorky.



Obr. 1. Typický příklad učení sítě.



Obr. 2. Učení 3vrstvé BP-sítě (16–8–4)

2) Výsledky:

TABULKA III

VÝSLEDKY RŮZNÝCH TYPŮ SÍTÍ

Počet vnitř. vrstev	Konfigurace neuronů	Počáteční chyba		Učících cyklů	Chyba	
		učení	validace		učení	validace
0	1 výstup	0,673	0,346	10	0,345	0,151
1	8	0,479	0,228	100	0,313	0,139
1	16	0,650	0,293	155	0,291	0,137
1	32	0,570	0,295	55	0,320	0,141
1	48	0,666	0,298	55	0,316	0,136
2	8–8	0,842	0,443	160	0,307	0,136
2	16–8	0,619	0,321	115	0,305	0,147
2	32–16	1,041	0,549	105	0,301	0142
2	48–24	0,608	0,310	55	0,317	0,152
3	16–8–4	0,491	0,226	130	0,303	0,138

IV. ZÁVĚR

U BP-sítě s rostoucím počtem neuronů ve vrstvách klesal počet učících cyklů pro sítě různých vrstev. Největší vliv pro naučení sítě mělo obvykle několik počátečních cyklů sítě (max. 30), poté se síť ještě dále učila, ale v kvalitě naučení již nedocházelo k markantnímu zlepšení, viz. obr. 2.

LITERATURA

[1] Šnorek M., Jiřina M.: *Neuronové sítě a neuropočítače*. Vydání první. Praha ČVUT, 1996

[2] Vincent Sigillito: *Pima Indians Diabetes Database*, vstupní soubor, 1990