

Rozpoznávání ručně psaných číslic

Pavel Kotrč

12. 1. 2006

Data

- pocházejí z *UCI Repository*:
www.ics.uci.edu/~mlearn/MLRepository.html
- databáze *pendigits*
- 250 číslic od 44 osob
- trénovací sada: 7494 vzorů (30 osob)
- testovací sada: 3498 vzorů (14 osob)

Získání dat

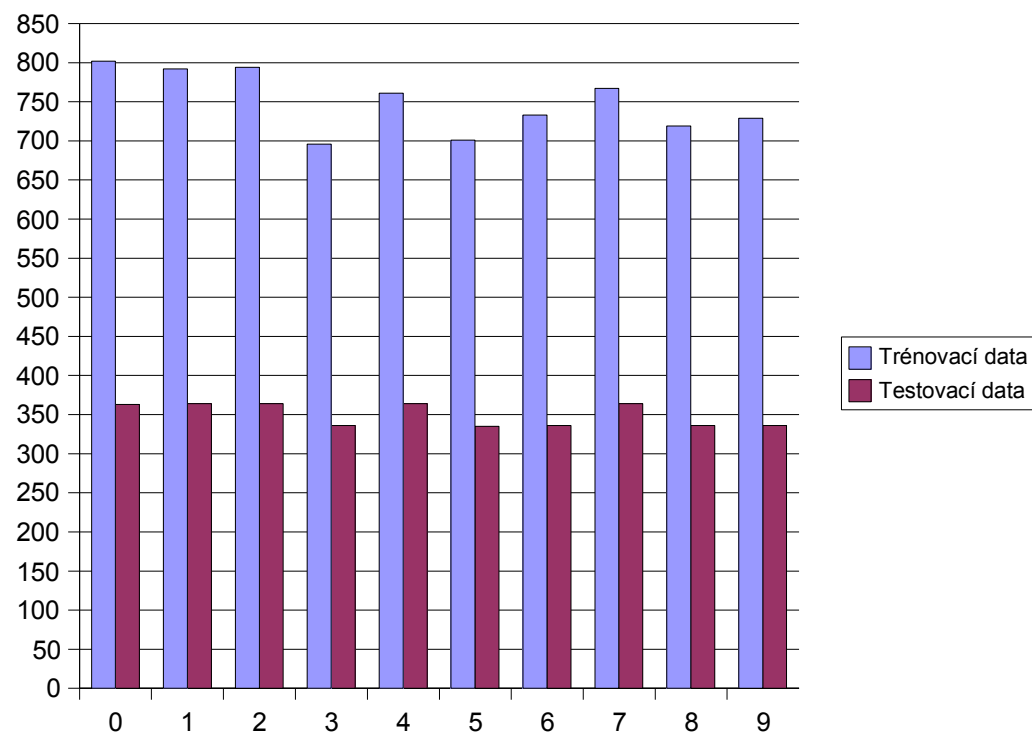
- hardware: tablet Wacom PL-100v
- 250 číslic v náhodném pořadí
- každá číslice zapsaná v oblasti 500 x 500 pixelů
- každých 100 ms se ukládá poloha pera
- tlak pera se zanedbává

Zpracování dat

- normalizace kvůli posunu a změně měřítka
- → hodnoty jsou mezi 0 a 100
- některé číslice jsou „kratší“ než jiné (např. 1 a 8) → nutno **převzorkovat**
- zde použita lineární interpolace mezi dvojicemi bodů
- výsledkem jsou vektory konstantní délky: 16 „datových“ složek
- označení třídy (0 – 9) se při učení vynechává

Rozložení dat

- rozložení dat podle jednotlivých číslic:



- metoda k -NN dává pro $k = 1, \dots, 11$ přesnost mezi 97,34 a 97,80 %

Učení

- jde o identifikaci shluků, tedy kompetiční učení
- 10 neuronů, kohonenovské učení (**learnk**) s konstantou učení 0,1 (1. případ) a 0,8 (2. případ)
- trénovací funkce **learnr**, tzn. v každé epoše je náhodně vybrán vektor, o který se „soutěží“
- 500 epoch učení

Jak počítat chybu?

- naučená kompetiční síť vrací vektor tříd (shluků), do kterých zařadil jednotlivé vstupní vektory
- čísla těchto tříd ovšem neodpovídají původním číslicím
- úprava: pro každou třídu se spočítá, kolik je ve výstupním vektoru reprezentantů dané číslice (0 – 9)
- nejčastější číslice se vybere jako ta, kterou daná třída reprezentuje (viz následující tabulky)

Kompetiční učení (1)

- parametr učení = 0,1
- úspěšnost
 - trénovací data: 77,4 %
 - testovací data: 67,2 %
- rozložení tříd (vlevo trénovací, vpravo testovací sada), červeně označeny vybrané třídy

	0	1	2	3	4	5	6	7	8	9
0	0	50	1	687	0	74	0	52	62	7
1	0	142	749	1	1	0	1	15	40	10
2	269	0	0	0	0	0	1	0	325	0
3	3	9	0	0	14	8	714	0	15	0
4	499	0	0	0	1	0	0	0	18	12
5	0	35	17	0	0	0	0	609	132	0
6	6	0	0	1	726	0	3	0	0	69
7	3	524	13	29	27	1	0	95	19	116
8	0	0	0	0	0	459	1	7	105	0
9	0	19	0	1	11	178	0	0	3	505

	0	1	2	3	4	5	6	7	8	9
0	0	6	0	313	0	34	0	20	0	2
1	10	141	338	0	0	0	0	10	18	0
2	133	0	0	0	0	0	0	3	156	0
3	13	2	0	0	22	1	335	0	0	0
4	206	0	0	0	0	0	0	0	2	0
5	0	15	19	0	0	0	0	270	58	1
6	0	1	0	0	317	0	1	0	0	20
7	1	141	7	21	14	0	0	58	1	31
8	0	0	0	0	0	168	0	3	100	0
9	0	58	0	2	11	132	0	0	1	282

Závěr

- kompetiční učení se zdá být logickou volbou pro tento typ úloh (identifikace shluků)
- **ale:**
 - je třeba vyladit parametry (viz nepříznivá hodnota chyby ve druhém případě)
 - je nutné naučit mnoho vzorů z každé třídy (zde zhruba 50 – to je zřejmě málo)
 - učení trvá dlouho (500 epoch trvalo cca 60 minut)
- → jednodušší je použít klasickou klastrovací metodu *k-NN*, která dává i lepší výsledky