Klasifikace antropologických dat

Martin Bruchanov

***Abstrakt*— Určování věku úmrtí z kosterních pozůstatků je jedna z důležitých úloh v oblasti antropologie. Nabízejí se otázky, jaké jsou možnosti její automatizace na základě parametrů kostry a využití výpočetních prostředků při výzkumu. Jako jedna z možných cest k nalezení odpovědi se nabízí využití neuronových sítí.**

I. Zadání

Prozkoumat možnosti neuronové sítě typu *RBF (Radial Basis Function)* jako klasifikátoru antropologických dat.

II. Úvod

A. Vstupní data

Jedná se o antropologická data s výčtem parametrů kosterních pozůstatků sloužící ke klasifikaci věku úmrtí. Jeden vstupní vektor obsahuje následující údaje:

- **Kontinent:** Europe, Africa, North America, Asia.
- **Národnost:** Portugal, Africaner, ZULU, USAW, Spain, Suisse, SOTO, Thailand, USAB.
- **Pohlaví:** Female, Male.
- **Číselné parametry:** PUSA, PUSB, PUSC, SSPIA, SSPIB, SSPIC, SSPID.
- **Třídy (věk)**, pro testování neuronovou sítí byly zvoleny tři hlavní soubory klasifikovaných tříd:

  T1: $29-$, $30-39$, $40-49$, $50-59$, $60-69$, $70+$;
  T2: $29-$, $30-49$, $50+$;
  T3: $29-$, $30-59$, $60+$;

  Vstupní soubory těchto tříd byli navíc dále rozděleny na soubory obsahující:
  - obě pohlaví,
  - pouze ženy,
  - pouze muže.

  Navíc byl zvlášť zkoumán soubor pro národnosti Evropanů.
  - Celkem 18 souborů pro T1, T2 a T3.
  - Vstupních vektorů: 953; ženy: 479, muži: 474.
  - Zastoupení tříd (T1):

Třída	$29-$	$30-39$	$40-49$	$50-59$	$60-69$	$70+$
Počet	115	168	189	176	149	156
Zastoupení	12,0%	18,6 %	19,8 %	18,4 %	15,6 %	16,4 %

B. Síť typu RBF

- Dvouvrstvá síť s dopředným šířením signálu.
- Učí se s učitelem (konkurence k perceptronovských sítím MLP a BP).
- Síť obsahuje rozvětvovací vrstvu a dvě vrstvy neuronů: skrytou a výstupní, vrstvy bývají úplně propojené.

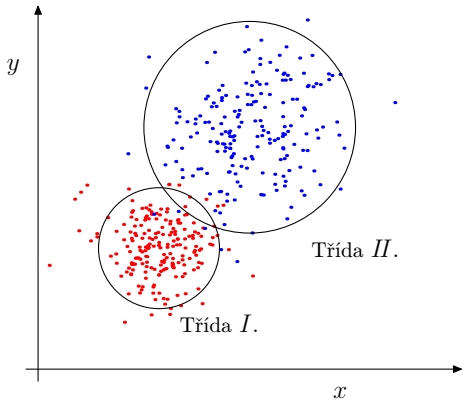
1) Učení různých sítí RBF:

C. Parametry RBF sítě

Použitý simulátor *Weka 3.4.11* má implementovánu normalizovanou Gaussovskou RBF síť a používá standardní algoritmus *K-means*. Tato neuronová síť se ve vstupních datech snaží nacházet shluky vektorů a příslušnost dané třídy určuje euklidovská vzdálenost od středu shluku.

Uživatel má možnost pro RBF síť nastavit následující parametry:

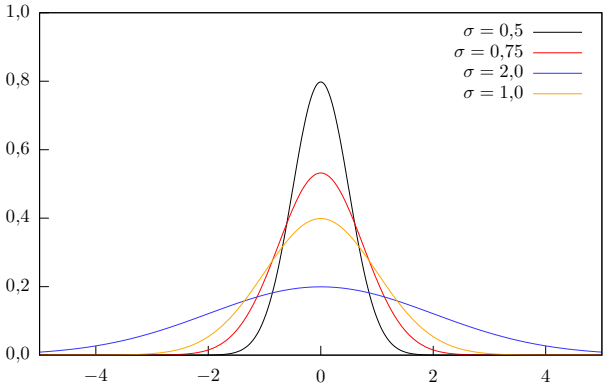
- `numClusters` n – Počet shluků, které by měl vygenerovat algoritmus K-means. Počet shluků je třeba odhadnout. Příklad klasifikace pro 2D prostor vstupních vektorů, který obsahuje dva shluky:



- `minStdDev` σ – minimální standardní odchylka f(x) pro shluky.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

Parametr σ ovlivňuje určování shluků. Velikost parametru má vliv na to jakým způsobem se síť bude učit. Příliš malá hodnota může způsobit to, že síť bude přeučená, příliš velká naopak to, že bude docházet k chybné klasifikaci.



Obr.: Vliv střední odchylky na normální rozdělení.

- `clusteringSeed` x – Náhodná inicializace pro K-means algoritmus, pro každé měření byla zvolena jiná inic. hodnota.
- `maxIts -1` – Maximální počet iterací pro splnění logistické regrese. Pouze pro problémy s diskrétními třídami.
- `ridge 1.0E-8` – hodnota vrcholu pro logistickou nebo lineární regresi.

D. Výsledky

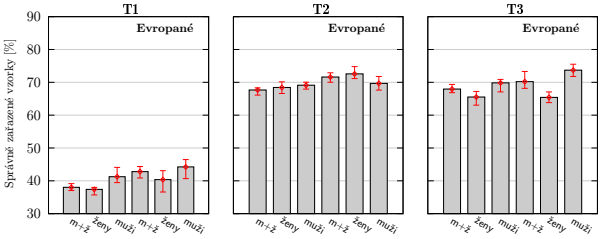
Byly testovány různé parametry sítě n (`numClusters`) a σ (`minStdDev`).

Pro vyčíslení chyby jsem využil metodu *cross-validation*, kdy je vstupní soubor rozdělen na 10 částí, kdy vždy jedna skupina se neúčastní učení sítě, ale slouží pro testování. Ve výsledích je ukázána průměrná hodnota výsledku, na sloupcových grafech je navíc ukázán rozptyl mezi nejhorším a nejlepším výsledkem pro dané měření.

Nejprve jsem provedl odhad parametru σ (`minStdDev`), nejlepších výsledů pro sítě s $n = 2$ a $n = 3$ bylo dosaženo pro $\sigma = 0{,}775$. S touto hodnotou jsem pak vyzkoušel měnit počet shluků n. Výpočetní složitost roste s počtem shluků přibližně $O(2^n)$ a čas výpočtu pro hodnoty $n > 4$ byl v řádově v hodinách až desítkách hodin (u souboru T1).

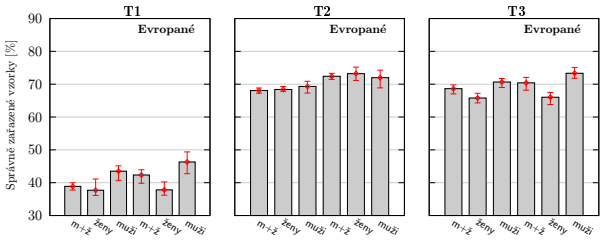
1) Výsledky měření $n = 2, \sigma = 0,775$:

				Evropané		
T1:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	38,0 %	37,4 %	41,2 %	42,8 %	40,4 %	44,2 %
T2:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	67,6 %	68,4 %	69,1 %	71,6 %	72,6 %	71,8 %
T3:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	67,9 %	65,5 %	69,8 %	70,2 %	65,4 %	73,7 %



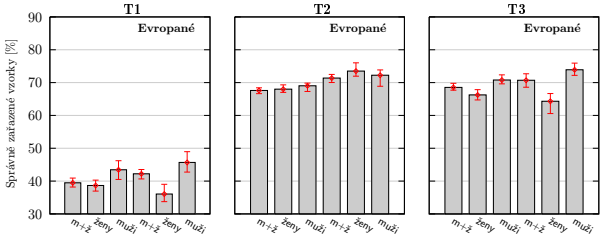
2) Výsledky měření $n = 3, \sigma = 0,775$:

				Evropané		
T1:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	38,9 %	37,7 %	43,5 %	42,3 %	37,8 %	46,3 %
T2:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	68,1 %	68,4 %	69,3 %	72,4 %	73,2 %	72,0 %
T3:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	68,6 %	65,8 %	70,7 %	70,4 %	66,0 %	73,3 %



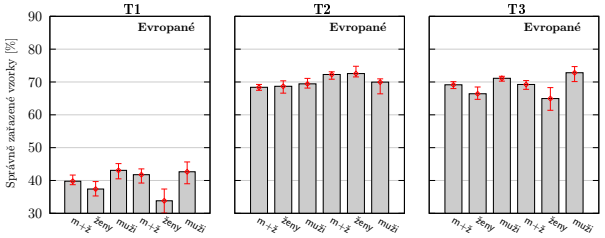
3) Výsledky měření $n = 4, \sigma = 0,775$:

				Evropané		
T1:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	39,5 %	38,7 %	43,4 %	42,2 %	36,1 %	45,7 %
T2:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	67,6 %	68,0 %	69,0 %	71,4 %	73,5 %	72,2 %
T3:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	68,5 %	66,2 %	70,8 %	70,7 %	64,3 %	73,9 %



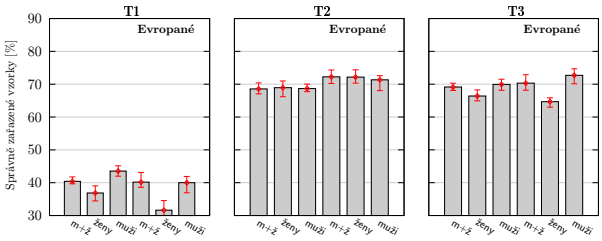
4) Výsledky měření $n = 5, \sigma = 0,775$:

				Evropané		
T1:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	39,8 %	37,4 %	43,1 %	41,8 %	33,8 %	42,7 %
T2:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	68,4 %	68,7 %	69,4 %	72,3 %	72,6 %	70,0 %
T3:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	69,1 %	66,4 %	71,1 %	69,2 %	65,0 %	72,8 %



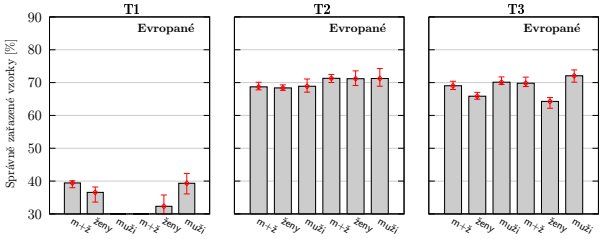
5) Výsledky měření $n = 6, \sigma = 0,775$:

				Evropané		
T1:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	40,4 %	36,8 %	43,5 %	40,2 %	31,6 %	40,0 %
T2:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	68,6 %	68,9 %	68,7 %	72,2 %	72,2 %	71,3 %
T3:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	69,1 %	66,4 %	69,9 %	70,3 %	64,7 %	72,7 %



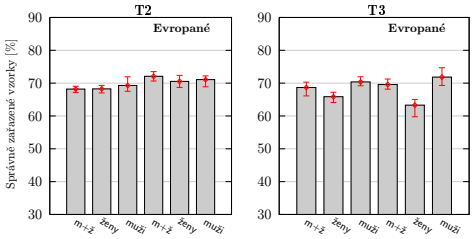
6) Výsledky měření $n = 7, \sigma = 0,775$:

	Evropané					
T1:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	39,4 %	36,6 %	—	—	32,3 %	39,3 %
T2:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	68,7 %	68,4 %	68,9 %	71,3 %	71,2 %	71,2 %
T3:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	69,0 %	65,8 %	70,1 %	69,8 %	64,3 %	72,1 %



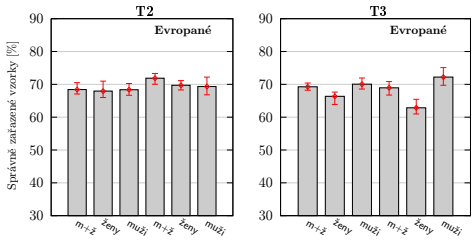
7) Výsledky měření $n = 8, \sigma = 0,775$:

	Evropané					
T1:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	—	—	—	—	—	—
T2:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	68,2 %	68,2 %	69,3 %	72,1 %	70,5 %	71,1 %
T3:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	68,7 %	65,9 %	70,4 %	69,6 %	63,3 %	71,8 %



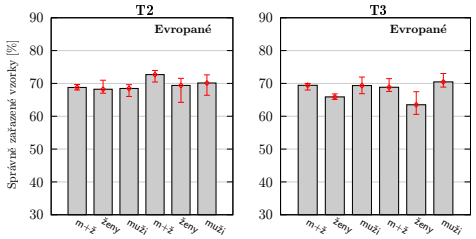
8) Výsledky měření $n = 9, \sigma = 0,775$:

	Evropané					
T1:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	—	—	—	—	—	—
T2:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	68,4 %	67,9 %	68,4 %	71,9 %	69,7 %	69,3 %
T3:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	69,3 %	66,3 %	70,0 %	68,9 %	62,8 %	72,2 %



9) Výsledky měření $n = 10, \sigma = 0,775$:

	Evropané					
T1:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	—	—	—	—	—	—
T2:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	68,8 %	68,2 %	68,5 %	72,7 %	69,4 %	70,1 %
T3:	obě pohl.	ženy	muži	obě pohl.	ženy	muži
	69,4 %	65,9 %	69,3 %	68,8 %	63,5 %	70,5 %



III. Závěr

- Změna parametrů sítí se projevovala spíše nevýrazně a dosažené výsledky se měnily v řádech okolo jednotek procent.
- Nejlepších výsledků bylo dosaženo pro $n = 4, \sigma = 0,775$. A to jak pro vstup s počtem tříd 4 (T2 a T3), tak i pro T1 s 6 třídami. S dále rostoucím počtem shluků, se dosahovalo stejných a nebo dokonce horších výsledků.
- Pro kostry žen Evropanek pro T1 byl naměřen nejlepší výsledek dokonce pouze pro $n = 2$. Předpoklad, že větší počet shluků zlepší výsledky klasifikace se nepotvrdil.
- Ve výsledcích je patrný výrazný rozdíl klasifikace koster žen v souborech T2 a T3. Ženy jsou v celkovém souboru rovnoměrně zastoupené, v souboru Evropanek je výraznější rozdíl v zastoupení tříd, viz následující histogram:

T3	ženy Evropanky	T2	ženy Evropanky
< 29	32	< 29	32
$30 - 59$	111	$30 - 49$	66
> 60	103	> 50	148

Literatura

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

[2] *Weka 3: Data Mining Software in Java*, `http://www.cs.waikato.ac.nz/ ml/weka/`