ANN - Klasifikace

Menu:	QCExpert	Prediktivní metody	ANN-Klasifikace

Tento modul využívá neuronovou síť ke klasifikaci, tedy modelování nečíselné (kategorické, faktorové) odezvy *Z* na základě jednoho nebo více číselných prediktorů *X*. Nečíselná odezva nemá numerickou hodnotu, ale takzvanou úroveň. Odezvy mají obvykle formu textu s nejméně dvěmi úrovněmi, například (A, B, C, D, E); (YES, NO); (zelená, modrá, žlutá); („Elytrigia repens“, „Lolium perenne“, „Phleum pratense“) a podobně. Přitom není definováno pořadí těchto úrovní. Odezva může obsahovat i čísla, ta jsou však interpretována jako text bez numerické hodnoty. Úkolem tohoto modulu je nalézt vztah mezi hodnotami prediktoru a úrovní odezvy a případně využít tohoto vztahu pro predikci neznámé úrovně odezvy na základě známých hodnot prediktoru. Například podle numerických hodnot analýzy krve rozpoznat diagnózu pacienta (za krev lze případně dosadit mechanické, fyzikální, chemické, optické parametry, vlastnosti půdy, míra stresorů atd., za diagnózu pacienta vadu výrobku nebo materiálu, typ poruchy ve stroji či procesu, klasifikaci ekologické jednotky, či biologický druh, atd, atd.

Typická data budou tedy vypadat například takto:

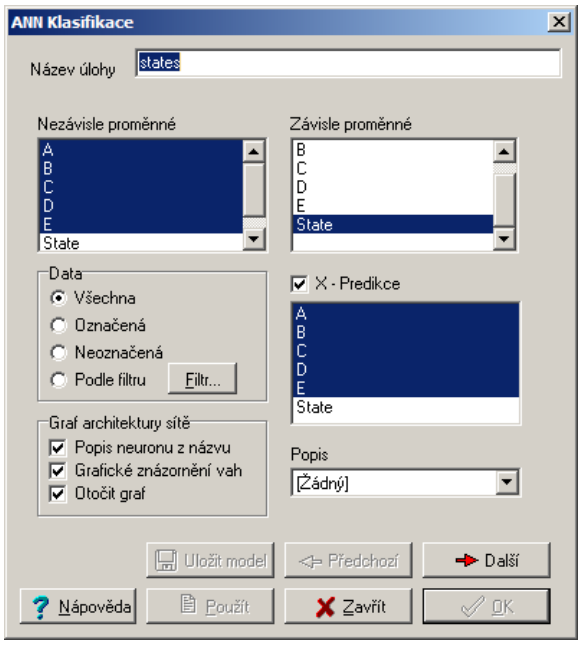
Prediktor					**Odezva**
X1	**X2**	**X3**	**X4**	**X5**	
2.4	3.2	3.9	3.8	2.2	NY
-1.2	2.6	3.6	3.7	2.1	CA
3.5	1.4	6.9	4.3	0.1	NE
1.8	1.3	3.3	0.6	1.1	WA
-0.4	2.6	1.8	2.2	2.8	FL
1.2	0.5	4.6	3.1	3.7	NY
0.1	2	4.2	2.6	3.9	CA
4.7	2.3	7.5	3.6	1.9	NE
2.8	2.1	4.8	1.4	-1	WA
0.8	1	1.3	2.6	4.2	FL
3.3	3.2	4.8	5.2	2.6	NY
…	…	…	…	…	…

Počet úrovní *R* odezvy (zde je počet úrovní odezvy *R*=5) přitom nesouvisí s počtem prediktorů *M* (zde náhodou rovněž *M*=5). Pokud odezva závisí na prediktorech, klasifikační model se pokusí tuto závislost popsat. Tento model je pak možno použít ještě během výpočtu, nebo později pro odhad odezvy při zadaných hodnotách prediktoru. Odhad má přitom formu pravděpodobností přiřazených každé z úrovní odezvy. Predikce tedy říká, s jakou pravděpodobností nastane za daných hodnot prediktoru každá z definovaných úrovní faktoru. Ta úroveň, která má největší pravděpodobnost se pak prohlásí za predikovanou úroveň a to i v případě, že jiná úroveň má jen nepatrně nižší pravděpodobnost. Princip této metody spočívá v nahrazení odezvy *P* umělými binárními “dummy” indikátorovými proměnnými a model je pak ANN-analogií vícefaktorové logistické regrese.

Data a parametry

Struktura dat je popsána výše. V dialogovém okně se vyberou nezávisle proměnné (prediktory) a jedna nezávisle proměnná, která obsahuje nečíselné úrovně odezvy. Po zaškrtnutí políčka X-Predikce lze vybrat sloupce, z nichž se mají odhadnout predikované

úrovně odezvy a jejich pravděpodobnosti. Další postup je shodný s ostatními moduly ANN. Tlačítkem *Uložit model* se aktuálně vypočítaný model uloží do souboru pro pozdější použití v modulu *Predikce*, případně k automatickým predikcím v inteligentní databázi QCE-DataCenter®.



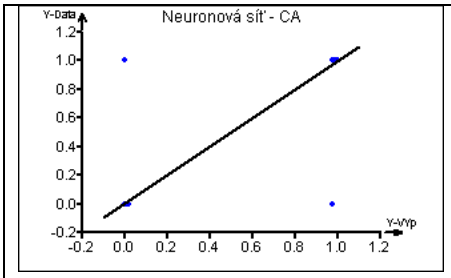
Obrázek 1 Nastavení a výber proměnných pro ANN klasifikaci

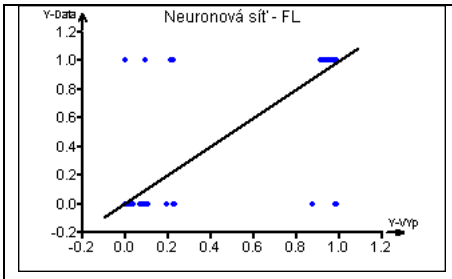
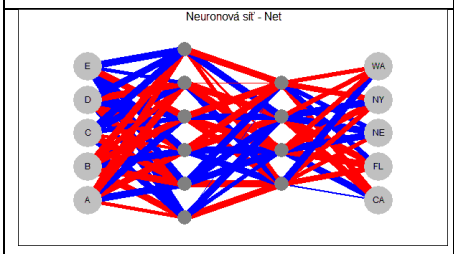
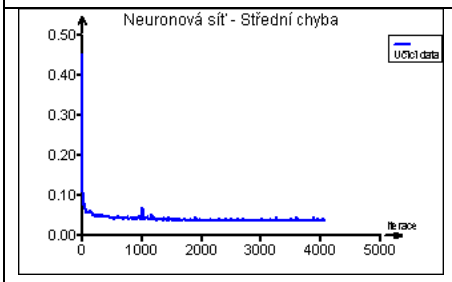
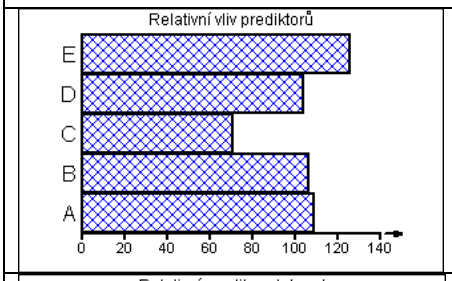
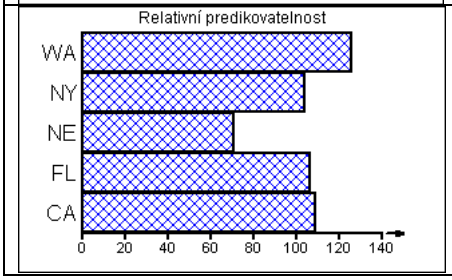
Protokol

Název úlohy	Zadaný název úlohy
Data	Vybrané řádky
Nezávisle proměnné	Použité sloupce prediktoru
Typ transformace	Zvolená transformace prediktoru
Závisle proměnné	Sloupec odezvy
Predikce	Sloupce pro které se počítá predikce
Vrstva, Neuronů	Číslo vrstvy a počet neuronů ve vrstvě
Strmost sigmoidy	Zadaná strmost sigmoidy (přechodové funkce)
Moment	Zadaný moment
Rychlost učení	Zadaná rychlost učení
Ukončit při chybě	Kritérium chyby pro ukončení
Procent dat pro učení (%)	Procent dat pro učení (%), je-li vybráno
Podmínky ukončení optimalizace	Zadané terminační podmínky pro ukončení výpočtu
Výpočet	Informace o průběhu výpočtu
Počet iterací	Skutečný počet iterací
Maximální chyba pro učící data	Dosažená maximální chyba pro učící data
Střední chyba pro učící data	Dosažená střední chyba pro učící data
Maximální chyba pro testovací data	Dosažená maximální chyba pro testovací data
Střední chyba pro testovací data	Dosažená střední chyba pro testovací data

Počet dat	Počet řádků x počet úrovní odezvy
Počet vah	Počet parametrů modelu (počet vah neuronové sítě).
Počet řádků	Počet vstupních řádků.
Počet úrovní	Počet úrovní odezvy.
Celkový součet čtverců	Součet čtverců bez modelu.
Reziduální součet čtverců	Součet čtverců odchylek pro výsledný model.
Vysvětlený součet čtverců	(Součet čtverců bez modelu) – (Součet čtverců odchylek pro výsledný model).
F-statistika	Vypočítaná F-statistika modelu.
F-krit	Kritický kvantil F-rozdělení.
P-hodnota	(1 – pravděpodobnost) F-rozdělení pro F-statistiku (je-li menší než 0.05, lze model považovat za významný.
Klasifikační pravděpodobnosti	Odhady pravděpodobnosti jednotlivých predikovaných úrovní odezvy vypočítané z modelu.
Predikce, Data	Predikovaná (nejpravděpodobnější) úroveň podle modelu a skutečná úroveň z dat.
Tabulka misklasifikací	Tabulka celkového počtu správných a chybných klasifikací a tabulka chybných a správných klasifikací pro jednotlivé úrovně odezvy. V ideálním případě diagonální matice (žádná chybná predikce).
Váhy	Tabulka parametrů neuronové sítě.
Relativní vliv	Relativní vliv jednotlivých prediktorů na odezvu, součet čtverců vah vycházejících ze vytupních neuronů.
Predikce	Je-li vybrána predikce, obsahuje tato tabulka modelem predikované odezvy pro zadané hodnoty prediktorů.

Grafy

	Graf predikovaných a skutečných úrovní pro jednotlivé úrovně prediktorů. V ideálním případě leží všechny body na přímce v bodech (0,0) a (1,1).

	
	Grafické vyjádření architektury sítě. Byla-li při výpočtu vybrána možnost *Grafické znázornění vah*, pak tloušťka synapsí znázorňuje absolutní velikost odpovídající váhy a tím také množství informace, která teče mezi dvěma neurony. Podle tloušťky synapse se tak dá usuzovat na významnost nezávisle proměnné (vychází z ní tlusté synapse) a také na kvalitu predikce jednotlivých závisle proměnných (vchází do ní tlusté synapse). Barva synapsí určuje pouze znaménko váhy (červená = záporná váha, modrá = kladná váha). Byla-li vybrána možnost *Popis neuronu z názvu*, použije se název prediktoru a název úrovně jako popis vstupních a výstupních neuronů.
	Graf učicího procesu, pokles součtu čtverců rozdílů predikce a skutečných hodnot závisle proměnné v závislosti na počtu iterací. Na obrázku je typický úspěšný učicí proces, který postupně zlepšoval model pro zadaná data.
	Relativní vliv prediktorů na predikci vyjádřený jako součet absolutních vah jednotlivých prediktorových proměnných. Tato statistika je pouze mírou vlivu proměnných na odezvu, není testem významnosti. Tento graf má reálnou vypovídací schopnost pouze při vhodné volbě modelu.
	Relativní predikovatelnost je součet absolutních vah jednotlivých úrovní odezvy a vyjadřuje míru jejich ovlivnění zvolenými prediktory. Tato statistika není potvrzením statistické významnosti vztahu mezi predikcí a odezvou. Tento graf má reálnou vypovídací schopnost pouze při vhodné volbě modelu.