SROVNÁNÍ POUŽITELNOSTI NEURONOVÝCH SÍTÍ A CLUSTEROVACÍCH METOD NA PŘÍKLADU KLASIFIKACE FINANČNÍ SITUACE PODNIKU

O. Trenz, V. Konečný

Došlo: 31. srpna 2010

Abstract

TRENZ, O., KONEČNÝ, V.: *Comparison of the applicability of neural networks and cluster classification methods on the example company's financial situation*. Acta univ. agric. et silvic. Mendel. Brun., 2010, LVIII, No. 6, pp. 579–586

The paper is focused on comparing the classification ability of the model with self-learning neural network and methods from cluster analysis. The emphasis is particularly on the comparison of different approaches to a specific application example of the commitment, the classification of then financial situation. The aim is to critically evaluate different approaches at the level of application and deployment options.

The verify the classification capability of the different approaches were used financial data from the database „Credit Info“, in particular data describing the financial situation of the two hundred eleven farms of homogeneous and uniform primary field.

Input data were from the methods used, modified and evaluated by appropriate methodology. Found the final solution showed that the used approaches do not show significant differences, and they can say that they are equivalent. Based on this finding can formulate the conclusion that the approach of artificial intelligence (self-learning neural network) is as effective as a partial methods in the field of cluster analysis. In both cases, these approaches can be an invaluable tool in decision making.

When the financial situation is evaluated by the expert, the calculation of liquidity, profitability and other financial indicators are making some simplification. In this respect, neural networks perform better, since these simplifications in them selves are not natively included. They can better assess and somewhat ambiguous cases, including businesses with undefined financial situation, the so-called data in the border region. In this respect, support and representation of the graphical layout of the resulting situation sorted out objects using software implemented neural network model.

neural network, self-learning, clustering, classification, financial situation of companies

Cílem příspěvku je ověřit schopnosti klasifikace rozsáhlejší množiny objektů s mnoha různorodými atributy využitím samoučících se neuronových sítí a porovnání výsledků s konvenčním přístupem z oblasti shlukové analýzy.

MATERIÁL A METODY

Pro porovnání klasifikačních schopností samoučících se neuronových sítí a konvenčního algoritmu byla zvolena náročná úloha klasifikace dvě stě jedenácti zemědělských podniků, podle patnácti ukaza-

telů (Konečný, Trenz; 2009a). Podstatou klasifikace byla snaha posoudit finanční situaci podniku na základě známých ukazatelů (Pokorná a kol., 2007). Pro značnou rozdílnost absolutních hodnot jednotlivých ukazatelů byla provedena jejich standardizace, tj. lineární transformace ukazatelů tak, aby standardní odchylka každého ukazatele byla rovna jedné a střední hodnota nule. Tím se vyrovnají váhy ukazatelů a eliminuje se nevhodný vliv velkých hodnot některých ukazatelů na výsledek klasifikace. Vzorek původních, nestandardizovaných, vstupních dat je uveden v tab. I.

I: Vzorek vstupních dat
I: The sample input data

	Krát. záv.	Cel. zadlu.	Dob. obr. zás.	Dob. obr. pohl.	Spl. záv.	Běž. lik.	Poh. lik.	Oka. lik.	Ren. vl. kap.	Ren. celk. kap.	Ø M M	Prac. kap.	Stál. aktiv.	Fin. maj. [%]	Dlou. záv. [%]
x1	109859	40.54	47	569	988	0.15	0.14	0	40.34	23.99	22	-93776	95.12	0.05	7.23
x2	3628	28.15	0	1282	71	1.82	1.82	0.14	23.22	16.69	18	2975	75.74	1.72	13.59
x3	622	45	351	800	62	5.85	4.32	0.39	17.52	9.63	29	3017	39.24	4.02	25.63
x4	353	55.05	90	439	0	24.47	20.65	11.69	154.86	69.61	0	8285	81.05	7.05	54.45
x5	22071	75.35	0	50	118	0.79	0.79	0.29	49.52	12.21	21	-4718	45.47	19.8	3.54
x6	15688	51.66	6	53	27	1.92	1.82	0.5	27.41	13.25	14	14425	0.84	25.73	0
x7	6247	13.95	172	90	32	7.21	2.97	0.74	0.28	0.24	22	38809	58.59	4.26	8.21
x8	12965	30.39	115	32	30	3.02	1.16	0.58	12.24	8.52	21	26205	63.53	7.03	15.2
x9	5749	19.26	147	63	42	5.48	2.45	1.42	1.9	1.54	17	25778	63.59	9.16	12.81
x10	2008	49.35	234	144	0	1.41	0.93	0.63	14.71	7.45	0	825	45.58	24.28	0
x11	45322	21.18	79	42	45	2.05	0.74	0.06	1.22	0.96	18	47583	72.79	0.73	7.68
x12	17782	27.68	176	97	100	3.94	1.89	0.37	2.1	1.52	15	52259	60.33	3.73	15.73
x13	7412	21.05	209	58	61	5.29	1.56	0.65	3.13	2.47	18	31816	59.71	4.91	13.53
x14	34125	30.78	136	143	60	4.51	2.17	0.97	4.91	3.4	29	119615	43.65	10.34	8.65
x15	18142	36.97	83	153	37	2.63	1.83	0.54	14.88	9.38	32	29499	46.18	10.56	17.28

Neuronové sítě

Jednodušší klasifikační model bude realizován neuronovou sítí s jedním výstupem pro každou třídu, tj. počet výstupů je dán počtem požadovaných klasifikačních tříd – viz obr. 1a), druhý model neuronovou sítí s výstupními neurony uspořádanými do dvojrozměrné mapy s počtem neuronů rovným počtu klasifikovaných objektů – obr. 1b) a třetí model konvenční klasifikační metodou shlukové analýzy (K průměrů) v software STATISTIKA. Pro realizace neuronových sítí se využívá software vytvořený na Ústavu informatiky PEF MENDELU.

V obou neuronových modelech se realizuje samoučení s klasifikací vektorů vstupních objektů

do tří tříd. Každý neuron N_r je spojen s každým vstupem x_k váhou w_{kr} , které společně tvoří váhový vektor W_r neuronu N_r , a hodnoty x_k vstupní vektor neuronové sítě X , tj.

$$W_r = (w_{1r}, w_{2r}, \dots, w_{nr}), X_r = (x_1, x_2, \dots, x_n). \quad (1)$$

Základem algoritmu samoučení je pravidlo blízkosti, tj.

- pro libovolný vstupní vektor X_i bude aktivován ten výstupní neuron, jehož vektor W_r je nejbližší vektoru X_i ;
- blízké vstupní vektory X_i, X_j musejí mít blízké i obrazy reprezentované výstupními neurony.

a)

b)

1: Struktura neuronových modelů
1: The structure of neural models

V průběhu učení se provádí náhodný výběr vektorů X a neuron s vektorem W_r , pro který platí

$$|X_i - W_r| \leq |X_i - W_s| \quad \forall s \text{ a } s \neq r, \quad (2)$$

se nazývá vítězný. Jeho vektor se koriguje směrem k vektoru X_r , tj. nový vektor bude

$$\bar{W}_r = W_r + \alpha (X - W), \quad (3)$$

kde α je koeficient učení. Jeho počáteční hodnota je menší než jedna a s narůstajícím počtem epoch učení se zmenšuje. Podrobně je proces samoučení popsán v monografii (Konečný, Trenz; 2009b). V důsledku korekce vektorů W_r budou jejich hodnoty konvergovat:

- k hodnotám těch vektorů X_r , pro které jsou neurony s vektorem W_r vítězné,
- nebo k hodnotám těžiště bodů X_i (s jednotkovou hmotností), pro které mají společný vítězný neuron.

Dodržení pravidla vzdálenosti ve výstupním uspořádání neuronů je zabezpečováno korekcí neuronů sousedících s vítězným. S narůstající vzdáleností sousedních neuronů se koeficient učení zmenšuje, stejně jako s narůstajícím počtem epoch učení. Za jednu epochu učení lze považovat provedení korekce vítězných neuronů pro všechny vstupní vektory X_r .

Důsledkem pravidel samoučení je to, že po ukončení procesu samoučení budou vektory neuronů W_r shodné s vektory těžiště reprezentované množiny vstupních vektorů X_r . Jednoduše se lze přesvědčit, že právě těžiště představuje bod, ke kterému mají všechny objekty X_i nejbližší. Vektory výstupů modelu obr. 1a) obsahují souřadnice těžiště reprezentované množiny objektů X_i a tak vítězným neuronem pro vstup X_i je určena přímo klasifikační třída.

V případě modelu na obr. 1b) je nutno klasifikační třídu zjistit na základě vzdálenosti vektoru vítězného neuronu a vektoru těžiště třídy klasifikovaného objektu (Konečný, Trenz; 2008c). Výstupní neuron N_r náleží k té množině, jejíž vektor těžiště je k vektoru W_r nejbližší. Pokud reprezentant množiny (vektor těžiště) není znám, lze model na obr. 1a) použít k jeho stanovení i pro model na obr. 1b). Tento model, kromě klasifikace, poskytuje ještě informaci o poloze klasifikovaného objektu vzhledem k prvkům klasifikačních tříd a jejich reprezentantům.

Jedním z problémů využití neuronových sítí je otázka přesnosti rozhodování, která úzce souvisí s přesností naučení a obvykle se hodnotí součtem čtverců odchylek požadovaných hodnot od dosažených. V případě učení modelů na obr. 1 je úroveň naučení dána hodnotou chybové funkce

$$E_{rr} = \frac{1}{2} \sum_i |X_i - W_r|^2, \quad (4)$$

kde X_i je vstupní vektor a W_r vektor vítězného neuronu. Přeučení neuronové sítě nebo nedoučení může vést k její chybné funkci.

Shluková analýza

Pro potřeby srovnání přístupu z oblasti shlukové analýzy byl nasazen přístup K-means a metoda PAM (Partition around medoid), jež dosahuje velmi dobrých výsledků pro identifikaci reprezentantů jednotlivých identifikovaných shluků. Obě zde uvedené metody patří do nehierarchického shlukování (Řezánková a kol., 2007).

Podstatou metody K-means je rozklad vstupní matice X do předem známého počtu shluků minimalizací zvoleného kritéria. Nejčastěji se používá hodnota stopy matice vnitroskupinové variability $st(W)$, kde

$$W = \sum_{h=1}^k \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h) (x_{hi} - \bar{x}_h)^T \quad (5)$$

Symbol k představuje celkový počet shluků, n_h počet objektů v h -tém shluku, x_{hi} je i -tý objekt v h -tém shluku a $\bar{x}_h$ je vektor průměru h -tého shluku.

Naproti tomu přístup PAM jde cestou identifikace k reprezentativním objektům, které se nazývají medoidy. Tyto reprezentanti patří do vstupní matice X . Hodnota k je zvolena na základě popisu situace jako jedno z kritérií úlohy (Rousseeuw a kol., 1997). Metoda PAM je považována za metodu velmi stabilní a to zejména s ohledem na vliv náhodných skokových hodnot ve vstupním vzorku na výsledek.

Medoidy jsou určeny tak, aby celková vzdálenost všech objektů k reprezentantu skupiny medoidu byla minimální. Algoritmus lze popsat hledáním prvků m_i podmnožiny z objektů výchozí množiny x_i , který lze popsat následujícím vztahem

$$\sum_{i=1}^n \min_t D(x_i, m_t), \quad t = 1, 2, \dots, k. \quad (6)$$

Vlastní postup výpočtu v tomto modelu je velmi podobný přístupu samoučících se neuronových sítí, tj. v prvním kroku se provádí identifikace reprezentantů skupin (medoidů) a následně se provádí zařizování zbylých prvků do těchto podskupin vzniklých na základě nalezení jejich medoidů.

DOSAŽENÉ VÝSLEDKY

Vstupní data zemědělských podniků (vzorek viz tab. I) převzatá z databáze „Credit info“ nebyla vhodná pro přímý vstup modelů neuronové sítě z obr. 1 a jak již bylo uvedeno, příčinou jsou velké rozdíly v absolutních hodnotách ukazatelů. Tento nedostatek byl odstraněn provedením standardizace. Před vlastní standardizací byl soubor dat „očištěn“ o ty záznamy, které se evidentně jevily jako nesprávně zaznamenané. Jednalo se o záznamy neúplné a nestandardní, tj. takové, které byly svým rozsahem mimo meze definované pro jednotlivá kritéria.

Další úpravou byla vyloučena část záznamů, které s ohledem na použitý princip samoučení neuronové sítě si byly velmi blízké, a s ohledem na soft-

warová omezení byl vstupní soubor s tři sta šestnácti záznamy redukován na dvě stě jedenáct záznamů.

Vzhledem k tomu, že správná klasifikace není známa a nejsou známy ani vektory těžiště množin, resp. reprezentanti klasifikace, budou reprezentanti a klasifikační množiny stanoveny modelem obr. 1a. a souřadnice těžištěově ověřeny podle vztahu

$$x_i^T = \frac{1}{N} \sum_j x_j, \quad (7)$$

kde N je celkový počet vstupních vektorů. Požadovaná klasifikace vytvořená vzhledem k takto stanoveným reprezentantům klasifikačních množin byla získána na experimentálním modelu, jehož uživatelské rozhraní je uvedeno na obr. 2.

Barevné rozlišení výstupních neuronů – reprezentantů $r1$, $r2$ a $r3$ bude použito také pro rozlišení objektů odpovídajících množin, při rovinném uspořádání výstupních neuronů v modelu dle obr. 1b. V textovém okně VÝSTUPY je vidět (s ohledem na jejich velikost) pouze část výpisu klasifikačních množin reprezentantů $r1$, $r2$ a $r3$. Kompletní obsah množin je:

$r1 = [x7, x8, x9, x10, x11, x13, x14, x16, x19, x21, x22, x23, x24, x25, x26, x28, x29, x30, x31, x32, x34, x36, x41, x48, x53, x55, x59, x62, x65, x66, x68, x69, x71, x73, x75, x76, x78, x79, x80, x82, x83, x84, x85,$

$x87, x88, x89, x90, x91, x94, x95, x96, x97, x103, x105, x112, x118, x119, x120, x125, x126, x131, x147, x171, x183, x195, x196, x197, x199, x210, x218, x249, x250, x255]$

$r2 = [x81, x92, x98, x110, x111, x121, x123, x124, x127, x130, x132, x133, x134, x135, x136, x139, x140, x141, x143, x149, x150, x155, x156, x157, x161, x164, x165, x167, x168, x169, x172, x173, x181, x182, x189, x191, x203, x211, x212, x215, x219, x221, x224, x231, x232, x233, x235, x236, x237, x238]$

$r3 = [x5, x6, x38, x39, x40, x42, x43, x45, x47, x52, x57, x58, x60, x61, x63, x64, x70, x74, x77, x99, x100, x102, x104, x106, x108, x109, x113, x115, x116, x117, x128, x129, x137, x144, x151, x153, x154, x158, x159, x160, x163, x170, x174, x175, x176, x178, x179, x184, x187, x188, x190, x192, x193, x194, x198, x201, x202, x204, x205, x206, x207, x208, x209, x213, x214, x216, x217, x220, x222, x223, x225, x226, x228, x229, x230, x234, x239, x240, x241, x243, x244, x245, x246, x251, x252, x253, x257, x259]$

Klasifikace provedená softwarem Statistika, nehierarchickou metodou shlukování (K průměrů) je téměř shodná. Rozdíl je pouze v klasifikaci objektu $x88$, z celkového počtu dvě stě jedenácti záznamů. Obdobně se chovala i metoda PAM, byť v základu je tento přístup navržen v mírně pozměněné podobě.

2: Rozhraní experimentálního modelu
2: Experimental Model Interface

Klasifikace provedená modelem dle obr. 1b vykazuje oproti požadované klasifikaci rozdíl pouze u dvou objektů – x10 a x231. Oba tyto objekty se nachází na hranici oblastí – viz obr. 3, kde je graficky znázorněna klasifikace do tříd r1, r2, r3. Barevným rozlišením umožňuje jednoduše rozpoznat jak vlastní prvky třídy, tak i hraniční prvky.

Z výsledného zobrazení lze doplnit informaci k objektu x88 rozdílně klasifikovaného metodou K průměrů. Z obrázku je patrné, že i tento objekt se nachází na hranici oblastí, takže na jeho klasifikaci je nutno nahlížet obezřetně.

II: Porovnání použitých metod

II: Comparison of methods used

	Neurosít varianta a	Neurosít varianta b	Shlukování K – průměry	Shlukování PAM
Počet rozdílných klasifikací	-	2	1	1

Hraniční prvky třídy poskytují informaci o vzdálenosti objektu od reprezentantů tříd, o nejbližších hraničních prvcích, sousedících třídách atd. Na základě takového zobrazení lze soudit, které prvky s velkou a které s malou jistotou náleží do dané třídy. Detailní zkoumání vzdáleností objektů umožňuje experimentální model na základě výstupů dílčích procedur. Například zobrazení vzdáleností hraničního prvku x223 od reprezentantů všech tříd.

Vzdálenost x223 od všech r_i

x223	r3	2.5814
x223	r1	3.1005
x223	r2	3.1075

Z uvedených vzdáleností je očividné, že stačí malá změna v attributech objektu x223 a místo třídy r1 bude patřit do třídy r2. Nebezpečí nesprávné klasifikace a nutnost podrobnějšího zkoumání je automaticky signalizováno zobrazením na hranici třídy. Žádný z dalších dvou zkoumaných modelů toto neumožňuje.

Vypovídací schopnost modelu s grafickým výstupem je tedy větší než zbývajících modelů, přesnost klasifikace srovnatelná a bude tak pro praktické použití vhodnější.

Tabulka II obsahuje vyhodnocení jednotlivých metod a ve vztahu k výchozímu modelu neuronové sítě (obr. 1a). Protože není známá správná klasifikace, není používán pojem „chyba klasifikace“, ale „rozdíl klasifikace“.

I když samoučením lze rozdělit objekty do množin, nelze jim dát význam. Charakteristiku množin musí dodatečně vytvořit podle vlastností reprezentantů a případně prvků množin znalec klasifikovaných objektů. V daném případě objekty množiny r1 představují podkapitalizovaní r2 – překapitalizovaní r3 – nevyhraněnou situaci.

DISKUSE A ZÁVĚR

V textu použité metody klasifikace dat (modely samoučící se neuronové sítě, metoda k-průměrů, metoda PAM) umožňují provést třídění a zařizování dat ze vstupního souboru na základě splnění dílčích kritérií spjatých s danou konkrétní metodou. Jak bylo potvrzeno, použité metody mají obdobné klasifikační schopnosti, které se liší jen minimálně. Toto jen potvrzuje správnost přístupu a to zejména z oblasti samoučících se neuronových sítí.

Vzorek použitých dat popisující finanční situaci zemědělských podniků lze po úpravě brát za reprezentativní (úpravou nechť je chápáno odstranění neplatných a nevěrohodných záznamů). Cílem řešení bylo rozčlenit vstupní data do skupin reprezentujících podniky dle finančního hlediska podkapitalizované, překapitalizované a s nevyhraněnou finanční situací. Totoho cíle bylo také dosaženo všemi modely, nicméně pro podrobnější zkoumání klasifikace je vhodnější samoučící se neuronová síť s grafickým výstupem do dvojrozměrné mapy (resp. Kohonenova síť). Předpokladem řešení podobných úloh je existence uceleného vstupního souboru obsahujícího objekty všech požadovaných tříd a možnost konzultace výsledků s expertem pro stanovení sémantiky tříd.

SUMMARY

Results of classification of companies do not vary exceedingly due to all used methods in the paper. Thus, we can assume that they are equivalent. The advantage of plane projection of the objects in a multidimensional space in a plane enables transparent classification and so that a visual evaluation of a decision situation as well. We used this within a software enabling realization of a model of a neuron net and it seems very contributory even though a certain simplification has been used for the final display as the original space is 15-dimensional. Such a displayed classification can be very useful for experts for evaluating a company situation.

x187	x229	x43	x39		x42	x38	x226	x214	x241	x60	x47		x6	x259
x222	x252			x174		x159		x178	x220	x192	x108	x193	x57	x104
x158			x129	x77		x163	x137	x240	x253	x207	x206	x190		x5 x231
x216	x202		x117	x61		x115	x170	x245	x225	x194			x70	
x160		x113		x45	x102		r3				x228	x128		x182
x197	x144		x64	x100	x63	x58		x176	x209	x213	x205		x169	
x131	x198	x154				x99		x201	x217		x215	x124		x164
	x82	x66		x119	x75			x195	x223	x237	x167		x98	x127
x22		x55	x250	x16	x171	x30	x76	x105		x221	x173	x130	x181	x155
x36		x21				x103		x89	x147			x238		
	x97	x96	x24	x8	x78	x11	x95	x118	x126	x150		x161		x134
x62		x83	r1		x71	x26	x48	x218	x87		x135	r2	x156	x110
x32	x19		x23	x94	x53	x13	x84	x73	x85	x140	x157	x81	x149	
x79		x14	x199		x7	x9		x210			x136	x92	x203	x132
x80	x125	x28	x41	x34	x249	x88	x31	x25	x10 x10		x111	x211	x143	x172

3: Zobrazení klasifikace objektů

3: Display object classification

(Top-heavy, under capitalization, neutral situation) mostly with respect to marginal transitional objects which can be hardly expressed by means of numeral values. Generally, the objects situated at the marginal area are more difficult to be classified and their graphic representation to better assesses the situation descriptive.

The carried out experiments have confirmed the former experience with shorter sets of objects showing that vectors are useful if being standardized, and assortment of objects with respect to distance.

The experiment with subsequent normalization of vectors due to the edge length of a displayed quadratic map has not improved the quality of classification at all.

neuronová síť, samoučení, shlukování, klasifikace, finanční situace podniku

LITERATURA

- KONEČNÝ, V., TRENZ, O., SVOBODOVÁ, E., 2010: Classification of companies with assistance of self-learning neural networks. *Agricultural economics: Zemědělská ekonomika*. sv. 56, No. 2, p. 51–58. ISSN 0139-570X.
- KONEČNÝ, V., TRENZ, O., 2009: *Rozhodování s podporou umělé inteligence*. Brno: MZLU v Brně. 50 s. Folia, ročník II, č. 8. ISBN 978-80-7375-344-3.
- POKORNÝ, M., TRENZ, O., KONEČNÝ, V., 2007: Neuronový rozhodovací model o kapitalizaci podniku. *Acta of Mendel University of agriculture and forestry*. sv. LV, č. 3, s. 97–104. ISSN 1211-8516.
- KONEČNÝ, V., TRENZ, O., 2009: *Klasifikace objektů samoučením neuronové sítě*. In: Firma a konkurenční prostředí – 5. část. 1. vyd. Brno: MSD, s. r. o., s. 52–60. ISBN 978-80-7392-088-3.
- ŘEZÁNKOVÁ, H., HÚSEK, D., SNÁŠEL, V., 2007: *Shluková analýza dat*. Praha: Professional Publishing, 196 s. ISBN 978-80-86946-26-9.
- ROUSSEEUW, P., STRUYF, A., HUBERT, M., 1996: *Clustering in an Object-Oriented Environment*. Journal of Statistical Software, Volume 1, Issue 4.

Adresa

Ing. Oldřich Trenz, Ph.D., doc. Ing. Vladimír Konečný, CSc., Ústav informatiky, Mendelova univerzita v Brně, Zemědělská 1, 613 00 Brno, Česká republika, e-mail: oldrich.trenz@mendelu.cz, vladimir.konecny@mendelu.cz