MODELOVÁNÍ BONITY OBCÍ POMOCÍ KOHONENOVÝCH SAMOORGANIZUJÍCÍCH SE MAP A LVQ NEURONOVÝCH SÍTÍ

Vladimír Olej, Petr Hájek
Univerzita Pardubice, Fakulta ekonomicko-správní, Ústav systémového inženýrství a informatiky

Abstract: *The paper presents the design of municipal creditworthiness parameters. Further, a model is designed based on LVQ neural networks for municipal creditworthiness classification. The model is composed of Kohonen's self-organizing feature maps (unsupervised learning) whose outputs represent the input of the LVQ neural networks (supervised learning).*

Key words: *Municipal creditworthiness parameter, Kohonen's self-organizing feature maps, LVQ neural networks, classification.*

1 Úvod

Bonita obce [1,2,6,7] je schopnost obce řádně plnit krátkodobé a dlouhodobé závazky. Je určována na základě faktorů (parametrů), které se týkají ohodnocovaného objektu. Ohodnocování bonity obcí je v současnosti realizováno metodami kombinující matematicko-statistické metody a ohodnocování expertem [1]. Výstupem těchto metod je buď skóre (bodovací systémy) nebo přiřazení i-tého objektu $o_i \in O$, $O=\{o_1,o_2, \dots ,o_i, \dots ,o_n\}$ do j-té třídy $\omega_{i,j} \in \Omega$, $\Omega=\{\omega_{1,j},\omega_{2,j}, \dots ,\omega_{i,j}, \dots ,\omega_{n,j}\}$ (modely založené na ratingu). Rating obce je nezávislé ohodnocování expertem založené na komplexní analýze všech dostupných parametrů bonity obce. Jeho nevýhodou je subjektivita experta. Na základě uvedeného je možné pro ohodnocování bonity obcí doporučit takové metody, které jsou schopny zpracovat a naučit se znalosti experta, umožňují uživateli zevšeobecnění těchto znalostí a jsou zároveň vhodně interpretovatelné. Pro ohodnocování bonity obcí jsou proto vhodné např. následující metody: hierarchické struktury fuzzy inferenčních systémů [2,7], metody učení bez učitele [1,2,6] a neuro-fuzzy systémy [1]. Neuronové sítě [4] jsou vhodné pro schopnosti učit se, zevšeobecňovat a modelovat nelineární vztahy. Ohodnocování bonity obcí lze považovat za klasifikační problém. Ten je možné řešit pomocí metod učení bez učitele (pokud nejsou třídy $\omega_{i,j} \in \Omega$ předem známy) nebo metodami učení s učitelem (pokud jsou třídy $\omega_{i,j} \in \Omega$ předem známy). V článku je uveden návrh parametrů pro ohodnocování bonity obcí, přičemž jsou vybrány pouze parametry s nízkými korelačními závislostmi. Dále jsou v článku popsány Kohonenovy samoorganizující se mapy (KSOM) a LVQ (Learning Vector Quantization) neuronové sítě. Přínos článku spočívá v návrhu modelu ohodnocování bonity obcí, který realizuje výhody jak metod učení bez učitele (kombinací KSOM a algoritmu K-průměrů), tak metod učení s učitelem (LVQ neuronové sítě). Na závěr je uvedena analýza výsledků, porovnání s dalšími metodami klasifikace a prezentace klasifikace obcí $o_i \in O$ do tříd $\omega_{i,j} \in \Omega$.

2 Návrh parametrů bonity obcí

Pro ohodnocování bonity obcí jsou používány následující kategorie parametrů: ekonomické, dluhové, finanční a administrativní [1]. Ekonomické parametry mají vliv na dlouhodobou bonitu obcí. Obce s více diverzifikovanou ekonomikou a příznivými sociálně-ekonomickými podmínkami jsou lépe připraveny na ekonomickou recesi. Dluhové parametry zahrnují velikost a strukturu dluhu. Finanční parametry informují o kvalitě rozpočtového hospodaření. Návrh parametrů, založený na předchozí korelační analýze [1,6] a na doporučení

významných expertů v dané oblasti, je uveden v Tab. 1. Parametry x_3 a x_4 jsou definovány pro r-tý rok a parametry x_5 až x_{12} jako průměrné hodnoty za roky r a r-1.

Tab. 1: Návrh parametrů bonity obcí

Parametry	
Ekonomické	$x_1 = PO_r$, PO_r je počet obyvatel v r-tém roce.
	$x_2 = PO_r/PO_{r-s}$, PO_{r-s} je počet obyvatel v roce r-s, a s je zvolený časový interval.
	$x_3 = U$, U je míra nezaměstnanosti v obci.
	$x_4 = \sum_{i=1}^{e} (PZO_i/PZ)^2$, PZO_i je počet obyvatel obce zaměstnaných v i-tém odvětví ekonomiky, i = 1,2, … ,e, PZ je celkový počet zaměstnaných obyvatel, e je počet ekonomických odvětví.
Dluhové	$x_5 = DS/OP$, $x_5 \in <0,1>$, DS je dluhová služba, OP jsou opakujících se
	$x_6 = CD/PO$, CD je celkový dluh.
	$x_7 = KD/CD$, $x_7 \in <0,1>$, KD je krátkodobý dluh.
Finanční	$x_8 = OP/BV$, $x_8 \in R^+$, BV jsou běžné výdaje.
	$x_9 = VP/CP$, $x_9 \in <0,1>$, VP jsou vlastní příjmy, CP jsou celkové příjmy.
	$x_{10} = KV/CV$, $x_{10} \in <0,1>$, KV jsou kapitálové výdaje, CV jsou celkové výdaje.
	$x_{11} = IP/CP$, $x_{11} \in <0,1>$, IP jsou investiční příjmy.
	$x_{12} = LM/PO$, [Kč], LM je velikost likvidního majetku obce.

Na základě uvedených skutečností je možno navrhnout následující datovou matici **P**

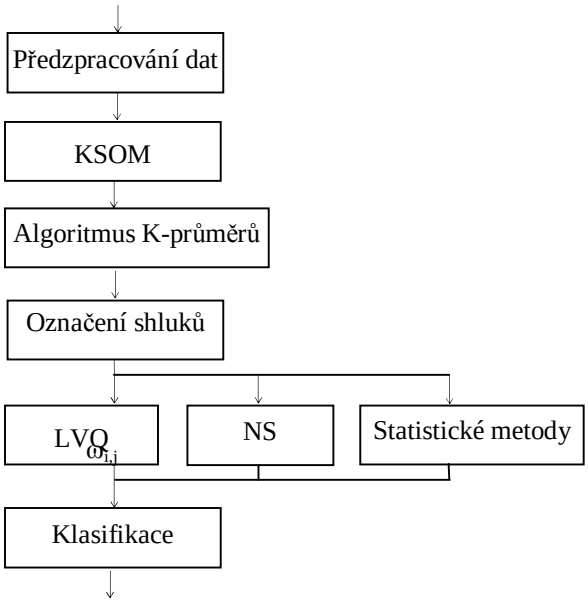
$$\mathbf{P} = \begin{array}{c|cccccc} & x_1 & \dots & x_k & \dots & x_m & \omega_{i,j} \\ o_1 & x_{1,1} & \dots & x_{1,k} & \dots & x_{1,m} & \omega_{1,j} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ o_i & x_{i,1} & \dots & x_{i,k} & \dots & x_{i,m} & \omega_{i,j} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ o_n & x_{n,1} & \dots & x_{n,k} & \dots & x_{n,m} & \omega_{n,j} \end{array},$$

kde $o_i \in O$, $O=\{o_1,o_2, \dots ,o_i, \dots ,o_n\}$ jsou objekty (obce), x_k je k-tý parametr, $x_{i,k}$ je hodnota parametru x_k pro i-tý objekt $o_i \in O$, $\omega_{i,j}$ je j-tá třída přiřazená i-tému objektu $o_i \in O$, $\mathbf{p}_i = (x_{i,1},x_{i,2}, \dots ,x_{i,k}, \dots ,x_{i,m})$ je i-tý vzor, $\mathbf{x}=(x_1,x_2, \dots ,x_k, \dots ,x_m)$ je vektor parametrů.

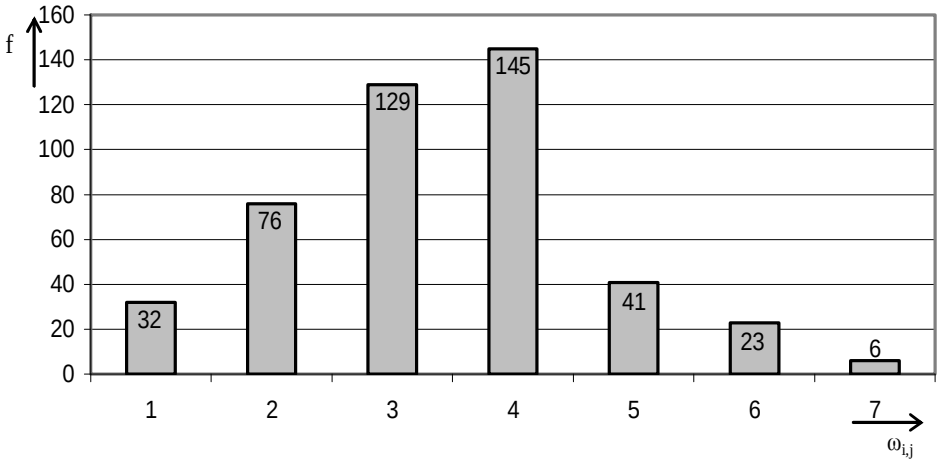
3 Návrh modelu pro klasifikaci obcí

Navržený model (Obr. 1) realizuje modelování bonity obcí. Předzpracování dat (standardizace) umožňuje vhodnou ekonomickou interpretaci výsledků. Dále jsou obce zařazeny do shluků pomocí KSOM. Shluky jsou označeny třídami $\omega_{i,j} \in \Omega$. Výstupy KSOM jsou použity jako vstupy LVQ neuronových sítí. Modelování bonity obcí představuje klasifikační problém, který možno definovat následujícím způsobem. Nechť $F(\mathbf{x})$ je funkce definovaná na množině A, která přiřazuje obraz $\hat{\mathbf{x}}$ (hodnota funkce z množiny B) každému prvku $\mathbf{x} \in A$, $\hat{\mathbf{x}} = F(\mathbf{x}) \in B$, $F: A \rightarrow B$. Takto definovaný problém lze řešit metodami učení s učitelem nebo metodami učení bez učitele. Pouze několik obcí v České republice má přiřazenu třídu $\omega_{i,j} \in \Omega$ specializovanými agenturami [1]. Proto je vhodné realizovat

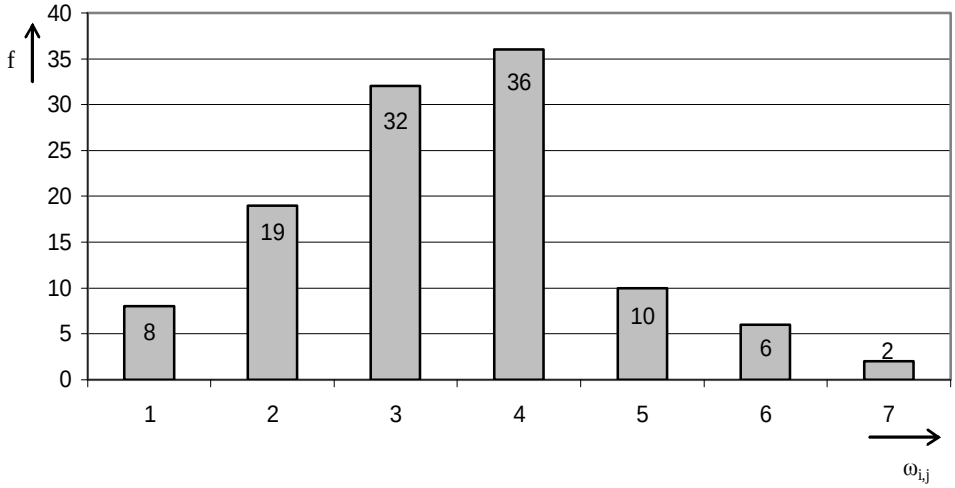
modelování bonity obcí metodami učení bez učitele. Na základě předchozí analýzy [2,6] je pro modelování bonity obcí vhodná kombinace KSOM a algoritmu K-průměrů. Jejich výstupy jsou použity jako vstupy LVQ neuronových sítí, resp. dalších struktur neuronových sítí (NS) a statistických metod, které realizují výhody metod učení s učitelem. Četnosti obcí v trénovací a testovací množině jsou uvedeny na Obr. 2 a Obr. 3.



Obr. 1: Model klasifikace obcí do tříd $\omega_{i,j}$



Obr. 2: Četnosti f obcí ve třídách ωi,j∈ Ω v trénovací množině



Obr. 3: Četnosti f obcí ve třídách ωi,j∈ Ω v testovací množině

Kohonenovy samoorganizující se mapy [1,5] jsou založeny na kompetiční strategii učení. Vstupní vrstva slouží k distribuci vstupních vzorů $\mathbf{p}_i$, i=1,2, … ,n. Neurony v kompetiční vrstvě jsou reprezentanty vstupních vzorů a jsou organizovány do topologické struktury. Ta určuje, které neurony spolu sousedí. Nejprve jsou vypočteny Euklidove vzdálenosti d_j mezi vzorem $\mathbf{p}_i$ a vahami synapsí $\mathbf{w}_{i,j}$ všech neuronů v kompetiční vrstvě. Je vybrán ten vítězný neuron s indexem j*, pro který je vzdálenost d_j od vzoru $\mathbf{p}_i$ minimální [5]. Výstup tohoto neuronu je aktivní, zatímco výstupy ostatních neuronů jsou neaktivní. Cílem učení KSOM je aproximovat hustotu pravděpodobnosti vstupních vektorů $\mathbf{p}_i \in R^n$ pomocí konečného počtu reprezentantů $\mathbf{w}_j \in R^n$, kde j=1,2, … ,s. Po nalezení reprezentantů $\mathbf{w}_j$ je každému vzoru $\mathbf{p}_i$ přiřazen reprezentant $\mathbf{w}_{j^*}$ vítězného neuronu. V procesu učení je definována funkce okolí h(j*,j), která určuje rozsah spolupráce mezi neurony, tj. kolik reprezentantů $\mathbf{w}_j$ v okolí vítězného neuronu bude adaptováno, a do jaké míry. Po nalezení vítězných neuronů je realizována adaptace vah synapsí $\mathbf{w}_{i,j}$. Principem sekvenčního trénovacího algoritmu [5] je ta skutečnost, že reprezentanti $\mathbf{w}_{j^*}$ vítězného neuronu a jeho topologického okolí se posouvají směrem k aktuálnímu vstupnímu vektoru $\mathbf{p}_i$ podle vztahu

$$\mathbf{w}_{i,j}(t+1) = \mathbf{w}_{i,j}(t) + \eta(t)h(j^*, j)[\mathbf{p}_i(t) - \mathbf{w}_{i,j}(t)], \quad (1)$$

kde $\eta(t) \in (0,1)$ je rychlost učení.

V [5] je uvedeno několik verzí algoritmů učení pro struktury LVQ1, LVQ2, LVQ3 a OLVQ1 (Optimized Learning Vector Quantization) neuronových sítí. Liší se v procesu hledání optimálních hranic mezi třídami $\omega_{i,j}$. Neuronové sítě LVQ jsou variantou KSOM s tím rozdílem, že se jedná o metody učení s učitelem. Nechť existuje LVQ1 neuronová síť a známý počet tříd $\omega_{i,j} \in \Omega$. Třídy $\omega_{i,j} \in \Omega$ jsou v procesu inicializace LVQ neuronové sítě přiřazeny všem vzorům $\mathbf{p}_i$. Potom, cílem procesu učení je nalezení vítězného neuronu j*. Rozdíl oproti KSOM spočívá v té skutečnosti, že proces učení je ukončen, jestliže $\mathbf{p}_i$ a $\mathbf{w}_{j^*}$ náleží do stejné třídy $\omega_{i,j} \in \Omega$. Dále, nechť vstupní vektor $\mathbf{p}_i$ náleží do třídy ω_p a jeho reprezentant $\mathbf{w}_{j^*}$ je středem třídy ω_q. V procesu učení jsou adaptovány pouze váhy synapsí $\mathbf{w}_{j^*}(t)$ tímto způsobem

$$\mathbf{w}_{j^*}(t+1) = \mathbf{w}_{j^*}(t) + \eta(t)[\mathbf{p}_i(t) - \mathbf{w}_{j^*}(t)], \quad (2)$$

jestliže $\mathbf{p}_i(t)$ a $\mathbf{w}_{j^*}(t)$ náleží do stejné střídy, $\omega_q = \omega_p$,

$$\mathbf{w}_{j^*}(t+1) = \mathbf{w}_{j^*}(t) - \eta(t)[\mathbf{p}_i(t) - \mathbf{w}_{j^*}(t)], \quad (3)$$

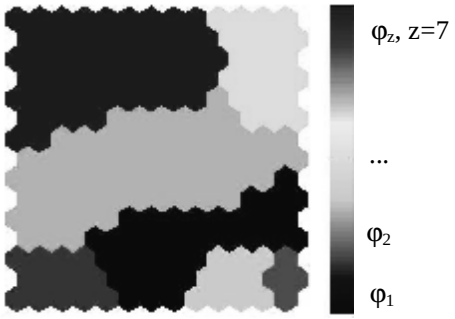
jestliže $\mathbf{p}_i(t)$ a $\mathbf{w}_{j^*}(t)$ náleží do různých tříd, $\omega_q \neq \omega_p$,

$\mathbf{w}_{i,j}(t+1) = \mathbf{w}_{i,j}(t)$ pro $j \neq j^*$, j=1,2, … ,s. (4)

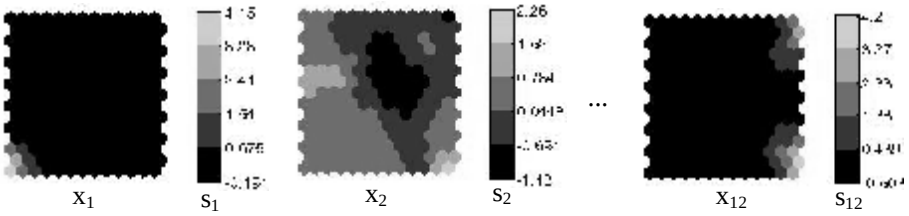
Neuronová síť OLVQ1 reprezentuje optimalizovanou verzi LVQ1 neuronové sítě, kde je ke každému reprezentantovi $\mathbf{w}_i$ přiřazena individuální rychlost učení $\eta_i(t)$. V procesu učení LVQ2 neuronové sítě jsou zároveň adaptovány ti dva reprezentanti $\mathbf{w}_i$ and $\mathbf{w}_j$, kteří jsou nejbližšími sousedy vstupního vektoru $\mathbf{p}_i$. Vektory $\mathbf{p}_i$ a $\mathbf{w}_j$ náleží do stejné třídy, zatímco $\mathbf{p}_i$ a $\mathbf{w}_i$ náleží do různých tříd. Učící algoritmus LVQ3 neuronové sítě zajišťuje tu skutečnost, že $\mathbf{w}_i$ pokračuje v aproximaci rozdělení tříd $\omega_{i,j} \in \Omega$.

4 Analýza výsledků

Vstupní parametry KSOM jsou určeny s cílem minimalizace kvantizační a topografické chyby [1]. Pomocí KSOM lze analyzovat strukturu dat. Použitím algoritmu K-průměrů lze v naučené KSOM najít shluky tak, jak je to uvedeno na Obr. 4. Algoritmus K-průměrů patří mezi nehierarchické algoritmy shlukové analýzy, kdy vzory $\mathbf{p}_i$ jsou přiřazeny do shluků $\varphi_1,\varphi_2, \dots ,\varphi_z$. Interpretace shluků je realizována pomocí hodnot parametrů $x_1,x_2, \dots ,x_m$, m=12, pro jednotlivé reprezentanty $\mathbf{w}_j$ (Obr. 5). Na základě interpretace lze shluky $\varphi_1,\varphi_2, \dots ,\varphi_z$, z=7, označit třídami $\omega_{i,j}$, j=1,2, … ,7. Navržené struktury KSOM a výstup shlukování uvedený na Obr. 4 a Obr. 5 představují ukázku z množství struktur KSOM realizovaných v průběhu experimentů.



Obr. 4: Shlukování KSOM pomocí algoritmu K-průměrů



Obr. 5: Hodnoty vektoru parametrů x pro jednotlivé reprezentanty

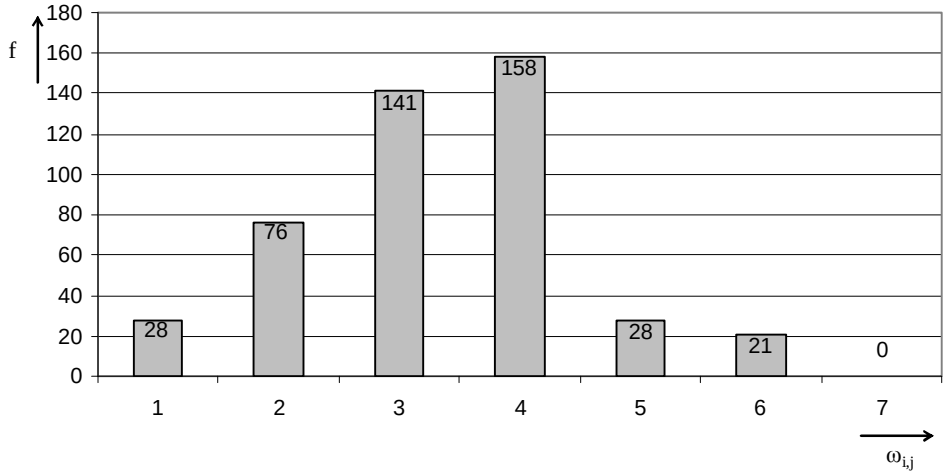
Legenda: $x_1,x_2, \dots ,x_k, \dots ,x_m$, m=12, jsou parametry, s_k je škála standardizovaných hodnot k-tého parametru

Vstupní parametry struktur LVQ neuronových sítí jsou uvedeny v Tab. 2, kde Eveninit je rovnoměrná inicializace, Propinit je proporcionální inicializace, s je počet reprezentantů, NN je počet sousedů v klasifikaci pomocí K-nejbližších sousedů (KNN), $W \in (0,1)$ je šířka okna a $S \in (0,1)$ je faktor stabilizace. Opět bylo navržena množina struktur LVQ1, LVQ2, LVQ3 a OLVQ1 neuronových sítí s různými hodnotami vstupních parametrů. Nejlepších výsledků bylo dosaženo s hodnotami parametrů uvedenými v Tab. 2.

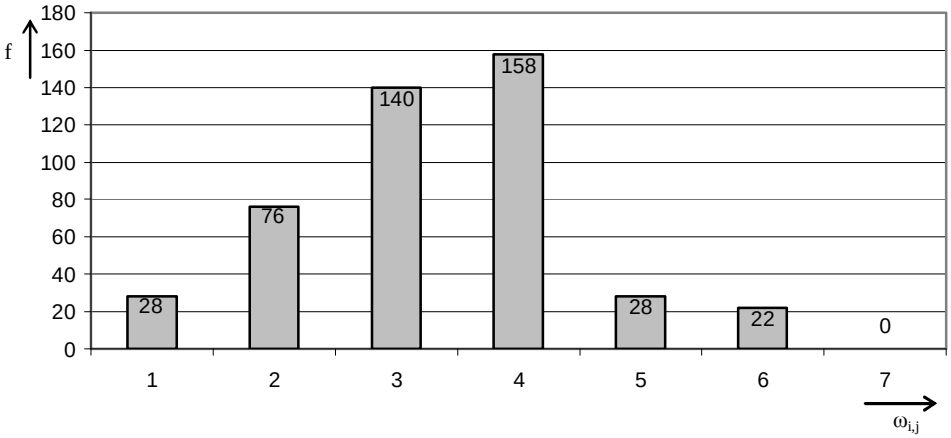
Tab. 2: Vstupní parametry LVQ neuronových sítí

Struktura	Inicializace	s	NN	η(t')	W	S	Epochy
LVQ1	Propinit	200	5	0.05	-	-	10000
LVQ2	Eveninit	200	5	0.05	0.3	-	100
LVQ3	Eveninit	200	5	0.05	0.3	0.1	1000
OLVQ1	Propinit	200	5	-	-	-	1000

Četnosti výskytu f obcí ve třídách $\omega_{i,j} \in \Omega$ (klasifikace) pro neuronovou síť LVQ1 a OLVQ1 jsou znázorněny na Obr. 6 a Obr. 7.



Obr. 6: Četnosti f obcí ve třídách ωi,j∈ Ω pro neuronovou síť LVQ1



Obr. 7: Četnosti f obcí ve třídách ωi,j∈ Ω pro neuronovou síť OLVQ1

Neuronové sítě LVQ1 a OLVQ1 mají na testovací množině dat nejlepší výsledky ze všech LVQ neuronových sítí (Tab. 3). Neuronová síť LVQ1 má maximální správnost klasifikace ξ_{max}=92.92[%], průměrnou správnost klasifikace ξ_a=91.33[%] a směrodatnou odchylku SD=0.97[%]. Lepších výsledků nebylo dosaženo ani aplikací LVQ2 a LVQ3 neuronových sítí na výsledky LVQ1 neuronové sítě.

Tab. 3: Správnost klasifikace ξ[%] LVQ neuronových sítí na testovacích datech

	LVQ1	LVQ2	LVQ3	OLVQ1
ξ_{max}[%]	92.92	92.04	92.04	92.92
ξ_a[%]	91.33	89.91	90.09	90.44
SD[%]	0.97	1.61	1.45	1.45

V Tab. 4 je uvedeno srovnání správnosti klasifikace ξ[%] na testovací množině dat s dalšími navrženými a analyzovanými strukturami neuronových sítí a zástupci statistických modelů. Konkrétně byly použity neuronové sítě ARTMAP (Adaptive Resonance Theory and Mapfield), standardní dopředné neuronové sítě (FFNN), neuronové sítě RBF (Radial Basis Function), lineární neuronové sítě (LNN) a pravděpodobnostní neuronové sítě (PNN). Dále bylo srovnání realizováno se statistickými modely SVM (Support Vector Machines), K-nejbližších sousedů (KNN) a logistickým regresním modelem MLRM (Multinomial Logistic Regression Model). Nejhorších výsledků bylo dosaženo pomocí MLRM, zatímco nejlepších pomocí RBF neuronových sítí s maximální správností klasifikace ξ_{max}=94.69[%], průměrnou správností klasifikace ξ_a= 89.93[%] a směrodatnou odchylkou SD=2.88[%].

Tab. 4: Správnost klasifikace ξ[%] na testovací množině dosažená ostatními modely neuronových sítí a statistickými metodami

	ARTMAP	FFNN	RBF	LNN	PNN	SVM	KNN	MLRM
ξ_{max}[%]	93.36	92.04	94.69	85.84	85.84	91.15	90.27	86.73
ξ_a[%]	90.34	90.56	89.93	84.60	83.34	89.76	87.46	81.42
SD[%]	3.81	1.33	2.88	0.79	1.89	1.83	3.38	5.31

5 Závěr

V článku je navržen vektor parametrů bonity obcí. Dále, navržený model realizuje ohodnocování bonity obcí. Předchozí analýza metod učení bez učitele (KSOM, neuronové sítě typu ART, shluková a fuzzy shluková analýza) [3] ukázala, že KSOM je pro modelování bonity obcí z uvedených metod nejvhodnější. Konkrétně, vizualizace shluků pomocí KSOM umožňuje vhodnou ekonomickou interpretaci výsledků. Hodnoty indexů kvality shlukování [8] jsou příznivé a navíc, odlehlé objekty nemají vliv na výsledky KSOM. Výstupy KSOM jsou použity jako vstupy LVQ neuronových sítí. Struktury LVQ neuronových sítí byly navrženy a analyzovány s cílem klasifikace obcí do tříd $\omega_{i,j} \in \Omega$ vzhledem k vysoké maximální správnosti klasifikace ξ_{max}[%] a průměrné správnosti klasifikace ξ_a[%] s nízkou hodnotou směrodatné odchylky SD[%]. Klasifikace pomocí LVQ neuronových sítí byla uskutočnena v prostředí LVQ_PAK, ostatní struktury neuronových sítí v prostředí MATLAB 7.1 a statistické modely v prostředí Weka 3.4 pod operačním systémem MS Microsoft Windows.

Použitá literatura:

[1]HÁJEK, P. Municipal Creditworthiness Modelling by Computational Intelligence Methods. Ph.D. Thesis, University of Pardubice, 2006.

[2]HÁJEK, P., OLEJ, V. Municipal Creditworthiness Modelling by means of Fuzzy Inference Systems and Neural Networks. In Proc. of 4th Int. Conference on Information Systems and Technology Management, TECSI-FEA USP, Sao Paulo, Brazil, May 30-June 01, 2007, pp. 586-608.

[3]Hájek, P., Olej, V. Municipal Creditworthiness Modelling by Clustering Methods. In Proc. of the 10th International Conference on Engineering Applications of Neural Networks,

EANN 2007, Margaritis, Illiadis, Eds., Thessaloniky, Greece, August 29-31, 2007, pp. 168-177.
[4]HAYKIN, S. S. Neural Networks: A Comprehensive Foundation. Prentice-Hall, Upper Saddle River, 1999.
[5]KOHONEN, T. Self-Organizing Maps. Springer-Verlag, New York, 2001.
[6]OLEJ, V., HÁJEK, P. Modelling of Municipal Rating by Unsupervised Methods. WSEAS Transactions on Systems, WSEAS Press, Issue 7, Vol.6, 2006, pp. 1679-1686.
[7]OLEJ, V., HÁJEK, P. Hierarchical Structure of Fuzzy Inference Systems Design for Municipal Creditworthiness Modelling. WSEAS Transactions on Systems and Control, WSEAS Press, Issue 2, Vol.2, 2007, pp. 162-169.
[8]STEIN, B., MEYER ZU EISSEN, S., WISSBROCK, F. On Cluster Validity and the Information Need of Users. In Proc. of the International Conference on Artificial Intelligence and Applications (AIA 03), Benalmádena, Spain, 2003, pp. 216-221.

Kontaktní adresa:

prof. Ing. Vladimír Olej, CSc., Ing. Petr Hájek, Ph.D.
Ústav systémového inženýrství a informatiky
Fakulta ekonomicko-správní
Univerzita Pardubice
Studentská 84, 532 10 Pardubice
Email: vladimir.olej@upce.cz, petr.hajek@upce.cz
Tel.: 466 036 004