INŻYNIERIA ŚRODOWISKA • Tom 11 • Zeszyt 1 • 2006

Stanisław Gruszczyński*

Przydatność różnych typów sieci neuronowych w klasyfikacji gleb**

1. Wprowadzenie

Kartografia gleb w coraz większym stopniu bazuje na technice cyfrowej. Metody, algorytmy i zasady dokumentowania zmienności przestrzennej zdające egzamin w tradycyjnej technice analogowej mogą być niezadowalające w technologii informatycznej [12]. Nowego spojrzenia wymagają sposoby prezentowania cech glebowych zmieniających się w przestrzeni w sposób ciągły (uziarnienia, zawartości węgla organicznego, składników pokarmowych, miąższości poziomów diagnostycznych itp.), a także elementów z założenia dyskretnych: jednostek typologicznych, klas bonitacyjnych, kompleksów glebowych. Szczególnie istotna jest integracja elementów ciągłych z dyskretnymi, która powinna uwzględniać anizotropowość płatów terenu, jednorodnych pod względem klasyfikacyjnym. Naturalną konsekwencją takiego podejścia jest konieczność tworzenia kryteriów podobieństwa gleb do obiektów wzorcowych oraz miar stopnia przynależności obiektów do poszczególnych klas. Konsekwencją tego rozwiązania jest rozmyta klasyfikacja, cechująca się, na przykład częściową przynależnością obiektów do różnych jednostek systematycznych lub bonitacyjnych [1, 5, 13]. Wydaje się, że prosta digitalizacja materiałów analogowych, tworząca wyłącznie cyfrową kopię tradycyjnej mapy może stanowić dopiero początek transformacji, prowadzącej do pogłębienia znajomości przestrzennej zmienności gleb.

Algorytmy sztucznej inteligencji wykorzystano, między innymi, w problemach prognozowania skutków przekształceń gleb [6–9]. Istota tej propozycji sprowadza się do poszukiwania modelu zależności między możliwymi do ustalenia cechami morfologii terenu i gleb oraz zaliczenia gleb do odpowiednich jednostek klasyfikacji czy kwalifikacji. Można przypuszczać, że w zmienionych warunkach morfologii terenu, będących skutkiem deformacji poeksploatacyjnych, uzyskany

* Akademia Górniczo-Hutnicza, Wydział Geodezji Górniczej i Inżynierii Środowiska
** Praca wykonana w ramach badań statutowych AGH nr 11.11.150.171

model dostarczać będzie dobrego przybliżenie nowej klasyfikacji. Podjęte badania, obok zastosowania w dziedzinie prognozowania szkód górniczych, mają walor ogólniejszy, polegający na umożliwieniu oceny niepewności związanej z klasyfikacją gruntów na podstawie danych, dostępnych z przetworzenia istniejących materiałów kartograficznych.

Celem podjętych badań było określenie stopnia wiarygodności szacunku kwalifikacji gleb na dużym terenie, przy użyciu algorytmów eksploracji danych.

W podjętych badaniach wykorzystano bazę danych zbudowaną dla części terenu byłego Górnośląskiego Okręgu Przemysłowego. W celu budowy cyfrowego modelu powierzchni, zostały zdigitalizowane współrzędne x, y i z ponad 167 600 punktów, co w rzeczywistych granicach opracowania daje gęstość około jeden digitalizowany punkt na 1,3 ha. W granicach obszaru opracowania zdigitalizowano ponad 16 600 konturów glebowych, które obejmują około 200 wariantów kompleksów i typów glebowych oraz ponad 6000 wariantów budowy litologicznej profili glebowych oraz ich klasyfikacji. W ramach digitalizacji konturów glebowych wyznaczono współrzędne x i y ponad 520 000 punktów załamania granic konturów. Digitalizacją współrzędnych x, y i z objęto także ponad 90 000 punktów rzędnych zwierciadła wód powierzchniowych.

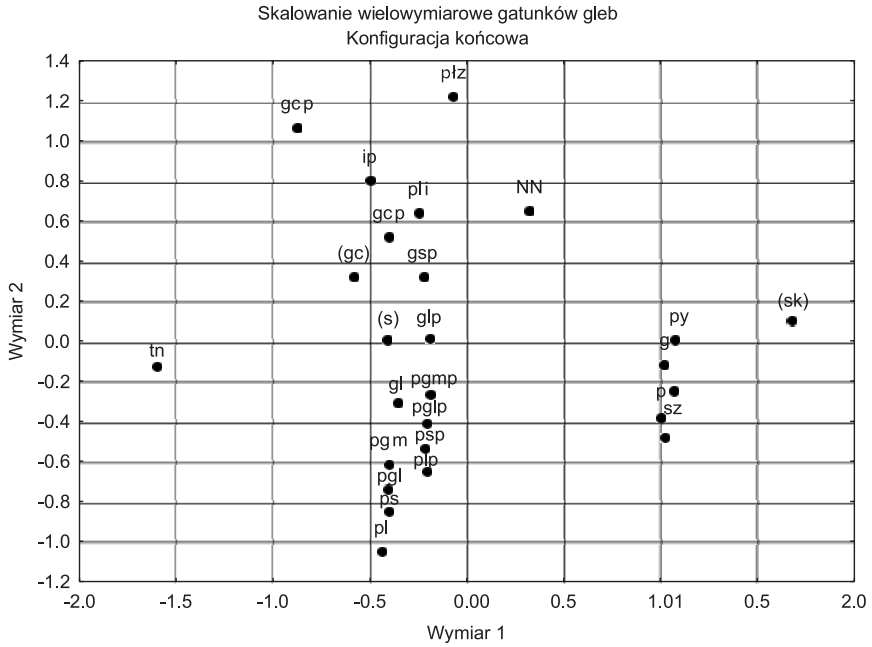
Analiza i prognoza przekształceń wymaga skwantowania przestrzeni na jednostki elementarne (JE), z założenia jednorodne pod względem wszystkich charakteryzujących je cech. Zdecydowano, że odpowiednie do zadania będą jednostki elementarne w postaci kwadratowych płatów terenu o powierzchni 4 arów (20 × 20 metrów). Założenie o jednorodności cech jednostki elementarnej oznacza, że pod względem zmiennych, mających charakter dyskretny (np. gatunek gleby, jej rodzaj lub klasa), zobrazowanych w dokumentacji jako kontury, całość jednostki (JE) ma atrybut zgodny z atrybutem konturu, wewnątrz którego leży środek JE. Podobnie w przypadku zmiennych mających charakter ciągły (np. rzędne, nachylenia) cała JE jest opisywana wartością zmiennej odpowiadającej środkowi jednostki.

Obszar testowy ma wymiary 59 × 44 km; w regularnym jego podziale na poszczególne jednostki elementarne występuje zatem 2950 kolumn oraz 2200 wierszy, co daje łącznie 6 490 000 JE. Należy zwrócić uwagę na fakt, że ze względu na nieregularności podziału arkuszy map glebowo-rolniczych i częściowe niezgodności przebiegu granic opracowania i granic dokumentacji glebowej, liczba wykorzystywanych w obliczeniach JE była znacznie mniejsza, niemniej opisany układ stanowił punkt odniesienia dla wszystkich obliczeń.

Projektowanie zawartości bazy danych przestrzennych musi uwzględniać charakter i wymagania dalszej ich obróbki. W przypadku wykorzystania sztucznych sieci neuronowych, to znaczy aplikacji poszukujących określonych regularności w obszernych bazach danych, konieczne jest ustalenie poprawnego i potencjalnie użytecznego kodowania składowych wektorów cech (wektorów wejściowych).

Większych wątpliwości nie ma w przypadku zmiennych wyrażanych w skalach porządkowych i ilorazowych (rzędne, nachylenia itp.). Natomiast rozważenia wymagają składowe wyrażane w skalach nominalnych: wówczas w rachubę, jako sposób kodowania, wchodzi kodowanie binarne typu 1 z *N* (co oznacza uaktywnienie, czyli „ustawienie" jednego z konkretnych *N* wejść, reprezentujących poszczególne składowe skali nominalnej). W bardzo rozbudowanych skalach nominalnych taki sposób kodowania może być bardzo uciążliwy przez nadmierne rozbudowanie wejściowej warstwy algorytmu. Innym sposobem jest przekształcenie skali nominalnej w porządkową lub ilorazową, co powinno zapewnić możliwość wykrycia regularności w odpowiedzi na zmiany jej wartości. Atrybuty jednostek elementarnych uzyskiwane w wyniku transformacji zmiennych nominalnych na ilorazowe będą nazywane w dalszym ciągu atrybutami pochodnymi.

Jak atrybut pochodny w stosunku do wartości nominalnych zostały potraktowane gatunki gleb. W celu doprowadzenia do usytuowania gatunków w skali porządkowej zastosowano procedurę skalowania wielowymiarowego (ang. *multidimensional scaling*), zawartą w pakiecie Statistica. Skalowanie przeprowadzono w dwu wymiarach, wykorzystując jako parametry porównania grup mechanicznych minimalne i maksymalne zawartości poszczególnych frakcji mechanicznych i substancji organicznej. Graficzny obraz skalowania prezentuje rysunek 1.



Rys. 1. Wynik operacji skalowania wielowymiarowego w celu uzyskania charakterystycznych parametrów dla różnych rodzajów i gatunków gleb

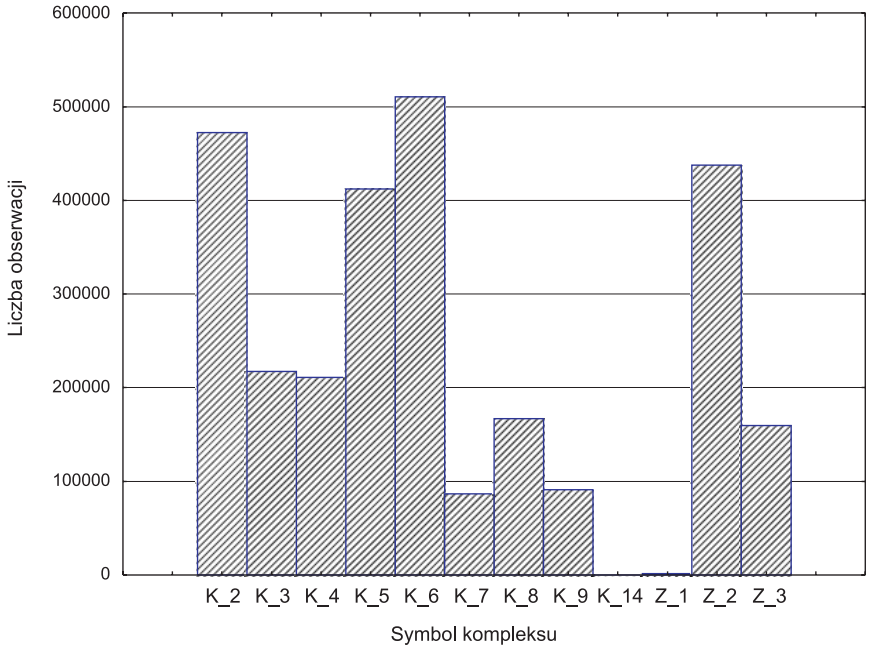
Procedura dostarcza informacji o współrzędnych położenia obiektów (którymi są gatunki wyróżniane na mapach glebowo-rolniczych) w umownej przestrzeni dwuwymiarowej. Po przeprowadzeniu skalowania każdy gatunek gleb jest reprezentowany przez dwie wielkości zmiennoprzecinkowe, jednoznacznie odzwierciedlające zespół wszystkich istotnych właściwości obiektu w stosunku do każdego innego.

O istotnych cechach gleb w poszczególnych punktach decydują czynniki lokalne, właściwe dla konkretnego punktu (lokalne nachylenie, uziarnienie, głębokość do wody itp.), lecz pewne znaczenie mają także czynniki, które można nazwać kontekstowymi, to znaczy kształtowanymi przez otoczenie jednostki elementarnej. Na przykład o bilansowaniu zasobów wodnych w profilu glebowym współdecyduje konfiguracja terenu, która w szczególności może przyjąć lub wykluczać napływ powierzchniowy. Inne warunki występują, kiedy wszystkie obszary otaczające są położone wyżej (przypuszczalnie zwiększony napływ powierzchniowy), a inne – na jednostkach położonych na wyniesieniach. W celu uwzględnienia tego faktu, obok nachylenia terenu, dla każdej jednostki elementarnej obliczono wskaźnik nazwany *konf*, będący średnim przewyższeniem ośmiu JE otaczających konkretną jednostkę. Do obliczenia tego wskaźnika wykorzystano moduł r.mapcalc zawarty w pakiecie GRASS, służący do realizacji operacji algebraicznych na zbiorach map cyfrowych, w tym do wykonywania operacji mających charakter kontekstowy.

Baza danych obejmuje następujące cechy (pola):

1) Numer JE – kolejny numer jednostki.
2) Współrzędna *X* środka JE.
3) Współrzędna *Y* środka JE.
4) Współrzędna *Z* środka JE.
5) Wartość nachylenia jednostki w procentach.
6) Wskaźnik *konf* określający konfigurację terenu w otoczeniu jednostki.
7) Identyfikator rodzaju użytkowania JE.
8) Identyfikator typu gleby w JE.
9) Identyfikator kompleksu rolniczej przydatności gleb JE.
10) Identyfikator uziarnienia w warstwie 0–25 JE.
11) Identyfikator uziarnienia w warstwie 25–75 JE.
12) Identyfikator uziarnienia w warstwie 75–125 JE.
13) Identyfikator uziarnienia w warstwie głębszej niż 125 JE.
14) Identyfikator zlewni.
15) Wielkość różnicy rzędnych JE i najniżej położonej jednostki w zlewni.
16) Odległość JE od najniżej położonej JE w zlewni.
17) Wskaźnik *aezb* JE.
18) Wskaźnik *aerz* JE.

Wielkości *aerz* i *aezb* są liczbowymi wskaźnikami określającymi położenie powierzchni terenu w odniesieniu do powierzchni interpolowanej z rzędnych brzegów zbiorników i cieków położonych w granicach terenu testowego. Zmienne nominalne wymienione w punktach 10)–13) zostały przetransformowane do postaci zmiennych ciągłych zgodnie z zasadą opisaną wcześniej. W wektorze wejściowym pominięto: numer JE, typ gleby (jako cechę obarczoną największym chyba błędem identyfikacji na mapach glebowo-rolniczych).



Rys. 2. Histogram rozkładu jednostek elementarnych w poszczególnych rodzajach kompleksów przydatności rolniczej

Rysunek 2 przedstawia histogram rozkładu JE w poszczególnych rodzajach kompleksów.

1.1. Zbiór uczący

Zbiór uczący (roboczy) składał się z 102 404 rekordów wylosowanych z danych liczących blisko 3 miliony rekordów. Technika poszukiwania rozmiaru sieci odpowiedniego do zadania została dostosowana do cech procedur treningowych. Pozostałe elementy zbioru tworzącego bazę danych były używane jako składniki testowe dla określania zdolności generalizujących sieci.

2. Klasyfikatory

Istotą metod adaptacyjnych jest wykrywanie regularności i związków deterministycznych w obszernych bazach danych. Obecnie nie ma możliwości określenia *a priori* stopnia złożoności problemu opisywanego wielu zmiennymi. Jedyną metodą znajdowania parametrów przetwarzania zapewniającego dostatecznie mały błąd realizacji zadania jest postępowanie iteracyjne.

Jako potencjalnie przydatne w zadaniu klasyfikacyjnym zostały przetestowane trzy typy klasyfikatorów neuronowych:

1) perceptron wielowarstwowy (MLP – ang. *Multilayer Perceptron*),
2) sieci z kołowymi funkcjami bazowymi (RBF – ang. *Radial Basis Function*),
3) sieci probabilistyczne (PNN – ang. *Probability Neural Network*).

Nie istnieje deterministyczna procedura umożliwiająca aprioryczny wybór odpowiedniej wielkości i architektury klasyfikatora, zaś ich poszukiwanie wymaga usystematyzowanego podejścia, czasu oraz przebadania wielu punktów startu algorytmu treningu.

2.1. Perceptrony wielowarstwowe: MLP

W badaniach przetestowano (w grupie sieci typu MLP) wiele architektur i metod treningu. Ich podstawowe parametry zostały zawarte w tabeli 1.

Skuteczność identyfikacji przez sieci z warstwami ukrytymi jest na ogół znacznie wyższa niż algorytmów pozbawionych jednostek ukrytych. Wyjątkiem jest sieć z jedną warstwą ukryta, w której funkcją celu była minimalizacja entropii. Pozostałe algorytmu treningu i architektury prowadziły do znacznie lepszych rezultatów. Podstawowym algorytmem treningu była klasyczna wsteczna propagacja błędu (wartość współczynnika uczenia $\eta = 0{,}15$) z momentem ($m = 0{,}7$). W niektórych przypadkach stosowano, jako uzupełnienie algorytmu wstecznej propagacji błędu, algorytm zmniejszania sprzężonego gradientu CGD (ang. *Conjugate Gradient Descent*). Ten algorytm wydatnie poprawiał skuteczność sieci, nie zdaje on jednak egzaminu w początkowych fazach treningu, ze względu liczne minima lokalne, z którymi radził sobie algorytm BP, natomiast nie radził algorytm CGD. Praktycznie zawiodła architektura sieci kwantowej QNN (ang. *Quantum Neural Network*).

Na podstawie zrealizowanych obliczeń można wyprowadzić następujące wnioski:

1. Architekturą klasyfikatora MLP o potencjalnie najwyższej przydatności w rozpoznawaniu kompleksów gleb na terenie testowym jest topologia z dwoma warstwami ukrytymi, złożona z 50 jednostek sigmoidalnych w pierwszej warstwie ukrytej i 35 jednostek sigmoidalnych w drugiej.

2. Powiększanie rozmiaru klasyfikatora ponad topologię opisaną w pierwszym punkcie nie prowadzi już do znaczącej poprawy skuteczności identyfikacyjnej, stwarzając równocześnie zagrożenie pojawienia się wadliwych rozpoznań.
3. W przeciwieństwie do innych topologii, sieć MLP jest perspektywiczna w sensie potencjalnej poprawy jej właściwości poprzez wzrost liczby danych uczących oraz uzupełnień zmiennych wejściowych.

Tabela 1. Zestawienie wyników treningów sieci typu MLP. Skuteczność identyfikacji zbioru treningowego (*T*) oraz walidacyjnego (*W*) wyrażona jako ułamek elementów poprawnie zidentyfikowanych. Symbol BP oznacza algorytm wstecznej propagacji błędu, QProp oznacza algorytm szybkiej propagacji, CGD oznacza algorytm zmniejszania sprzężonego gradientu

Typ	Ukryta (1)	Ukryta (2)	Skuteczność (*T*)	Skuteczność (*W*)	Trening
MLP	0	0	0,523	0,511	BP
MLP	45	0	0,265	0,220	Entr
MLP	21	0	0,747	0,732	BP
MLP	14	0	0,757	0,743	BP
MLP	14	14	0,776	0,765	BP
MLP	21	18	0,796	0,000	BP
MLP	21	19	0,814	0,795	BP
MLP	100	0	08,29	0,809	BP
MLP	31	21	0,829	0,801	BP
MLP	35	21	0,842	0,812	BP
MLP	70	0	0,844	0,820	BP
MLP	84	36	0,851	0,841	BP
MLP	42	36	0,854	0,843	QProp
MLP	145	75	0859	0,845	BP
MLP	50	35	0,888	0,855	BP + CGD
MLP	100	50	0,889	0,859	BP + CGD
MLP	100	75	0,895	0,867	BP + CGD
QNN	75	0	0,851	0,832	QNN

2.2. Sieci z kołowymi funkcjami bazowymi: RBF

Sieci RBF umożliwiają zastosowanie wielu algorytmów treningu i określanie rozmiaru ich struktury. Ze względu na brak ogólnej zasady dotyczącej ustalania dostatecznej wielkości sieci zbadano warianty zawierające od 2000 do 20 000 jednostek w warstwie ukrytej. Wypróbowywany konstruktywistyczny algorytm mapowania przestrzeni cech (FSM) został przerwany przy stanie 2100 jednostek

w warstwie ukrytej. Charakteryzując czasowy aspekt obliczeń, należy podkreślić, że pomimo stosunkowo prostych algorytmów treningu, duża liczba danych znacznie wydłuża czas uzyskiwania odpowiedniego stanu sieci (optymalizacja warstwy wyjściowej). Zgodnie z oczekiwaniami, generalnie nie zdały egzaminu warianty treningu polegające na kwantyzacji wektorowej zbioru uczącego; w literaturze [11] podkreśla się zazwyczaj słabą przydatność tej metody treningu dla danych ze zmiennymi wielomodalnymi, jak prawdopodobnie jest w tym wypadku. Tabela 2 zawiera podstawowe informacje dotyczące oceny przydatności przebadanych sieci RBF do rozwiązywania postawionego zadania. Sieć oznaczona symbolem FSM jest realizacją jednego z nowszych konstruktywistycznych algorytmów sieciowych [4]. Jego wadą jest długi czas treningu, zwłaszcza przy dużej liczbie danych uczących. Sieć CANFIS (ang. *Co-Active Neuro-Fuzzy Inference System*), algorytm udostępniany przez pakiet NeuroSolution, jest siecią hybrydową rozmyto-neuronową. Pakiet stawia ograniczenia wobec rozmiaru wektora wejściowego, w związku z tym został zmniejszony do pięciu elementów poprzez redukcję za pomocą procedury analizy składowych głównych (PCA), co jednak prowadzi do zmniejszenia pojemności informacyjnej wektora (wektor po przekształceniu obejmował około 75% zmienności wektora oryginalnego).

Tabela 2. Zestawienie wyników badań sieci z kołowymi funkcjami bazowymi. Metoda treningu „Próba'" oznacza procedurę losowego kopiowania danych treningowych do jednostek przetwarzających

Typ	Ukryta (1)	Trening	Skuteczność (*T*)	Skuteczność (*W*)
RBF	2000	Próba	0,5678	0,5213
RBF	10 000	Próba	0,6775	0,6243
RBF	20 000	Próba	0,7893	0,7578
RBF	2000	*K*-średnich	0,5756	0,5345
RBF	10 000	*K*-średnich	0,6876	0,6323
RBF	20 000	*K*-średnich	0,8145	0,7981
RBF	2000	LVQ	0,4413	0,4098
RBF	10 000	LVQ	0,5357	0,5678
RBF	20 000	LVQ	0,6560	0,6234
FSM	2100	FSM	0,5436	0,5276
CANFIS	5	CANFIS	0,4867	0,4677

Wyniki uzyskane z badań sieci z kołowymi funkcjami bazowymi upoważniają do sformułowanie następujących wniosków:

1. Sieci z kołowymi funkcjami bazowymi mają relatywnie gorsze właściwości identyfikujące zbiór treningowy niż sieci typu MLP.

2. Podstawowym problemem technicznym jest dobór centrów jednostek kołowych, relatywnie najlepsze wynik daje tutaj metoda *K*-średnich.
3. Algorytmy treningu sieci RBF przy posiadanej liczbie danych przestają być atrakcyjne wobec algorytmów możliwych do zastosowania w przypadku sieci MLP.

2.3. Sieci probabilistyczne: PNN

Badania w zakresie sieci probabilistycznych ze względu na ich właściwości wymagają: ustalenia dostatecznej do zadania wielkości sieci oraz wielkości parametru σ decydującego o zdolności generalizacji sieci. Zbiorem treningowym był plik ,,roboczy'', z tym wyjątkiem, że jeżeli rozmiar warstwy ukrytej obejmował wszystkie elementy zbioru uczącego, to walidację sieci prowadzono na dodatkowo wyselekcjonowanym z głównej bazy danych zbiorze o rozmiarze 100 000 rekordów. Wzrost rozmiaru sieci prowadzi do znacznego wydłużenia czasu klasyfikacji co jest istotnym mankamentem algorytmu PNN. Orientacyjny czas potrzebny do ustalenia parametrów sieci probabilistycznej liczącej około 100 000 jednostek ukrytych (trening sieci) przez system taktowany zegarem 2,4 GHz, wynosi około 4 godzin. Klasyfikacja podobnego co do rozmiarów zbioru także wymaga od 4 do 8 godzin (zależnie od celu obliczeń – klasyfikacja czy też wartości aktywacji warstwy wyjściowej). Jednorazowa klasyfikacja pełnego zbioru danych (2,76 mln danych) trwała ponad 4 doby nieprzerwanej pracy systemu. Łączny czas pracy systemów komputerowych w związku z realizowanymi w tym zakresie badaniami wynosił około 300 godzin. Podstawowe parametry prowadzące do oceny poszczególnych architektur sieci probabilistycznych zawiera tabela 3.

Tabela 3. Zestawienie wybików treningu sieci typu probabilistycznego

Typ	σ	Ukryta (1)	Skuteczność (T)	Skuteczność (W)
PNN	0,3	102 404	0,6542	0,6326
PNN	0,3	2000	0,5478	0,5275
PNN	0,3	3300	0,5405	0,5345
PNN	0,3	20 000	0,5580	0,5539
PNN	0,15	20 000	0,8126	0,7589
PNN	0,1	20 000	0,9383	0,8340
PNN	0,05	20 000	0,9973	0,8162
PNN	0,1	2000	0,5741	0,5586
PNN	0,1	3300	0,5741	0,5586
PNN	0,1	102 404	0,9834	0,8078

Z analizy danych zawartych w tabeli 3 można wyprowadzić następujące wnioski:

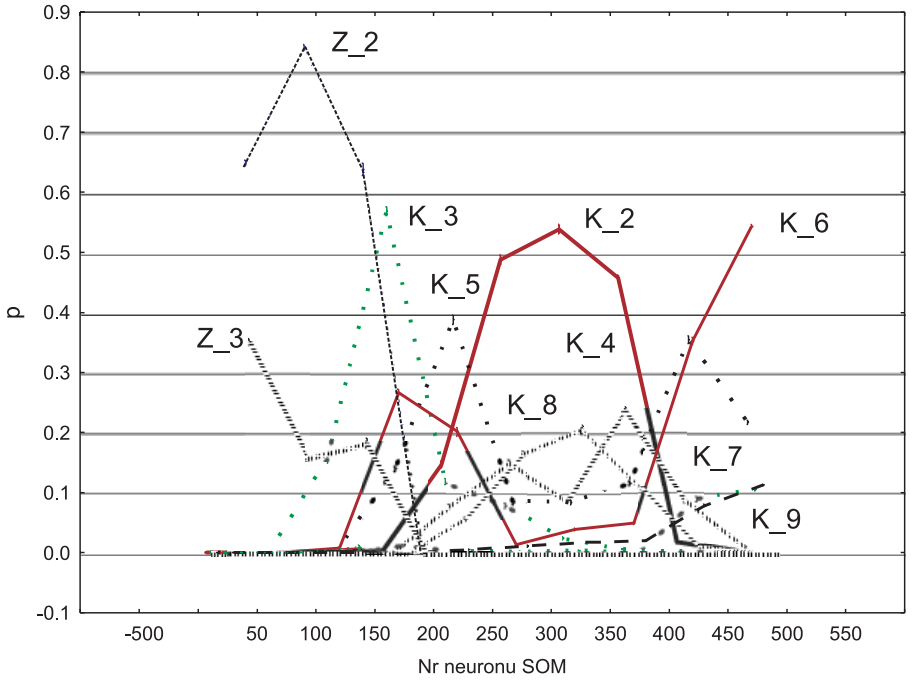
1. Architekturą dającą relatywnie najlepsze efekty zastosowania jest sieć złożona z 20 000 jednostek ukrytych oraz z parametrem $\sigma = 0{,}1$.
2. Dalsze zwiększanie rozmiaru sieci nie prowadzi do znaczącej poprawy jej skuteczności identyfikacyjnej. Sieć zawierająca ponad 100 000 jednostek w warstwie ukrytej dawała wyniki porównywalne z dużo mniejszymi jej wersjami.
3. Zbliżoną do optymalnej wartością parametru przetwarzania σ jest $\sigma = 0{,}1$. Podwyższanie tej wartości pogarsza skuteczność rozpoznawania elementów zbioru treningowego i testowego, natomiast jej zmniejszanie skutkuje znaczącym wzrostem skuteczności rozpoznawania zbioru treningowego przy bardzo znaczącym spadku skuteczności rozpoznawania zbioru testowego.
4. Argumentem przemawiającym przeciw zastosowaniu sieci probabilistycznej jako klasyfikatora jest wyjątkowo długi czas klasyfikacji, nawet przy bardzo wydajnych systemach przetwarzających.
5. Korzystną stroną stosowania sieci probabilistycznych jest uzyskiwanie informacji dotyczących stopnia wiarygodności indywidualnego wskazania klasyfikacyjnego. W niektórych przypadkach daje to możliwość utworzenia obrazu klasyfikacji powiązanego z oszacowaniem niepewności klasyfikacji.

3. Podsumowanie i wnioski

Współczesna technika umożliwia pełniejsze wykorzystanie czasochłonnych, iteracyjnych algorytmów klasyfikacji, pozwalających na bardziej wszechstronną analizę dużych zbiorów danych. Jedną z użytecznych technik eksploracji dużych zbiorów danych są algorytmy sztucznych sieci neuronowych. Dostarczają one heurystycznych modeli powiązań pomiędzy poszczególnymi cechami obiektów. Szczególnie zastosowanie algorytmów probabilistycznych daje możliwość pogłębienia obrazu zróżnicowania gleb.

Coraz większe zastosowanie w dziedzinie klasyfikacji obiektów przyrodniczych znajduje logika rozmyta, jako metoda alternatywna wobec logiki ostrej [2, 15, 16]. Szczególnie interesująca może być analiza jednoznaczności klasyfikacji gleb na określonych obszarach na podstawie skończonej listy kryteriów. Przykład takiej analizy prezentuje rysunek 3. Zawiera on wieloboki funkcji przynależności do zbiorów rozmytych reprezentujących kompleksy przydatności rolniczej w obszarze testowym. Zostały one oszacowane za pomocą funkcji aktywacji neuronów wyjściowych sieci PNN. Jak wiadomo, aktywacja neuronów wyjściowych sieci PNN generuje oszacowanie rozkładu prawdopodobieństwa przynależności wek-

tora wejściowego (reprezentującego obraz obiektu) do określonej kategorii. W wielowymiarowych przypadkach klasyfikacji, jak ma to miejsce w tym przypadku, trudność sprawia ustalenie „obszaru rozważań" (ang. *universe of discourse*) zbiorów rozmytych, czyli uporządkowanie wzdłuż jakiejś osi wektorów cech, co jest zadaniem banalnym w przypadku zbiorów rozmytych bazujących na pojedynczej cesze. Z pomocą przychodzi tutaj algorytm samoorganizującej się mapy SOM (ang. *Self Organizing Map*) [3, 10, 14], pozwalający na uporządkowanie wektorów cech w przestrzeni dwu- lub jednowymiarowej. W tym przypadku uporządkowanie dotyczy tych samych wektorów cech, które stanowią wejścia do sieci PNN. Skutkiem procedury jest sytuowanie się w sieci SOM wektorów podobnych obok siebie, zaś wektory znacznie się różniące położone są w dużych odległościach. Krzywe na rysunku 3 ilustrują podobieństwa i różnice w występowaniu poszczególnych kompleksów w obszarze testowym. Charakterystyczne jest sygnalizowane podobieństwo kompleksów K-2, K-4 i K-8 oraz klasteryzacja kompleksów K-5 i K-6. Wykres uwidacznia ograniczenia we wnioskowaniu na podstawie dostępnych z dokumentacji kartograficzno-glebowej informacji, polegające na trudności w jednoznacznym rozdzieleniu poszczególnych kompleksów.



Rys. 3. Wieloboki hipotetycznych funkcji przynależności do poszczególnych kompleksów w obszarze testowym oszacowane na podstawie analizy wskazań sieci PNN w uporządkowaniu sieci SOM Kohonena

Analizując przydatność poszczególnych algorytmów klasyfikacji w zadaniach bazujących na informacjach uzyskanych z przetworzenia tradycyjnych dokumentacji kartograficzno-glebowych, można sformułować następujące wnioski:

1. Poszczególne algorytmy dostarczają modeli o zbliżonej wiarygodności. Dla dużych zbiorów danych podstawowym problem jest selekcja reprezentatywnego zbioru uczącego.
2. Relatywnie dobre wyniki dają sieci MLP, generujące względnie najmniejsze błędy klasyfikacji.
3. Dobre wyniki klasyfikacji daje sieć typu probabilistycznego; jej mankamentem jest stosunkowo długi czas przetwarzania.
4. Połączenie wyników działania sieci PNN i SOM Kohonena pozwala na analizę podobieństw i źródeł błędów w klasyfikacji gleb.

Literatura

[1] Ameskamp M.: *Three-Dimensional Rule-Based Continuous Soil Modelling*. Christian-Albrechts-Universität Kiel, Institut für Informatik und Praktische Mathematik, February 1997 (praca doktorska)

[2] Baja S., Chapman D.M., Dragovich D.: *A conceptual model for defining and assessing land management units using a fuzzy modeling approach in gis environment*. Environmental Management, 29(5), April 2002, 647–661

[3] Budinich M., Taylor J.G.: *On the ordering conditions for self-organizing maps*. Neural Computation, 7(2), 1995, 284–289

[4] Duch W., Diercksen G.H.F.: F*eature space mapping as a universal adaptive system*. Computer Physics Communications, 341–371, 1994, 87(16)

[5] Enea M., Salemi G.: *Fuzzy approach to the environmental impact evaluation*. Ecological Modelling, (135), 2001, 131–147

[6] Gruszczyński S.: *Ocena zagrożenia gleb w rejonach górniczych za pomocą sztucznych sieci neuronowych*. Geoinformatica Polonica, 1(1), 1999, 45–62

[7] Gruszczyński S.: *Symulacja skutków przekształceń gleb na terenach górniczych za pomocą klasyfikatorów neuronowych*. Kraków, Uczelniane Wydawnictwa Naukowo-Dydaktyczne AGH 2000

[8] Gruszczyński S.: *Szkody górnicze na terenach użytkowanych przyrodniczo – niektóre problemy dokumentowania, prognozy i oceny*. Prace Naukowe GIG, 2002, 167–172

[9] Gruszczyński S.: *Szkody górnicze w użytkach rolnych – prognozowanie*. Inżynieria Rolnicza, 8 (41), 2002, 231–241

[10] Luttrell S.P.: *A Bayesian analysis of self-organizing maps*. Neural Computation, 6(5), 1994, 767–794

[11] Masters T.: *Sieci neuronowe w praktyce*. *Programowanie w języku C++*. Warszawa, Wydawnictwa Naukowo-Techniczne 1996

[12] McBratney A., Mendonça Santos M., Minasny B.: *On digital soil mapping*. Geoderma, 117 (1–2), November 2003, 3–52

[13] Silvert W.: *Ecological impact classification with fuzzy sets*. Ecological Modelling, (96), 1997, 1–10

[14] Yin H., Allinson N.M.: *On the distribution and the convergence of feature space in self-organizing maps*. Neural Computation, 7(6), 1995, 1178–1187

[15] Zadeh L.A.: *Fuzzy sets*. Information and Control, 8(3), June 1965, 338–353

[16] Zhu A.-X., Hudson B., Burt J., Lubich K., Simonson D.: S*oil mapping using GIS, expert knowledge, and fuzzy logic*. Soil Science Society of America Journal, 65(1), September – October 2001, 1463–1472