UKD 622.333: 528.9

Perspektywy wykorzystania cyfrowej kartografii gleb w ocenie szkód w glebach terenów górniczych

Prospect of application of digital cartography of soils in the evaluation of soil pollution in post-mine areas

Prof. dr hab. inż. Stanisław Gruszczyński)*

Treść: W pracy przeanalizowano czynniki budowy odpowiednio wiarygodnego modelu środowisko-gleby na podstawie danych z rejonu GOP. Stwierdzono, że pojedyncze klasyfikatory w rzeczywistych warunkach zawodzą jako dostatecznie precyzyjne modele poszukiwanych zależności. Lepszym i skuteczniejszym rozwiązaniem jest wykorzystanie zespołu wyselekcjonowanych, słabych klasyfikatorów. Alternatywą dla tradycyjnego głosowania większościowego jako procesu decyzyjnego wskazania klasy przez zespół, jest procedura związana z interpretacją wektora wskazań jako wejścia do wyspecjalizowanego klasyfikatora decyzyjnego (*stacking*).

Abstract: This paper presents an analysis of factors in the development of a reliable environment-soil model on the basis of data from the Upper Silesian Industrial Region. It was stated that individual classifiers fail as a sufficiently precise models for seeking dependences. in realtime conditions. The use of selected weak classifier ensembles is a more effective solution. The procedure connected with interpretation of indication vector as the access to the specialized decisive classifier (stacking) is an alternative for the traditional majority voting as a decisive process of indicating the class.

Słowa kluczowe:

górnictwo, deformacja terenu, monitoring

Key words:

mining, area deformation, monitoring

1. Wprowadzenie

Cyfrowa kartografia gleb (*digital soil mapping*) jest realizacją koncepcji Hansa Jenny [10] z 1941 roku, opracowania modelu powiązań między czynnikami środowiskowymi a kształtowaniem się gleb i ich lokalnymi właściwościami. Ta koncepcja, reprezentowana w oryginale przez symboliczny zapis matematyczny z niezdefiniowaną postacią analityczną, mogła uzyskać walor praktyczny po upowszechnieniu się technologii informacyjnej i dojrzałej metodologii GIS. Prawdopodobnie najbardziej charakterystyczną konsekwencją takiego podejścia jest odwrót od „poligonalnej" prezentacji treści glebowej map tematycznych i zmierzanie w kierunku mniejszych powierzchniowo jednostek, z reguły o regularnym kształcie (kwadratów), które można uznać za jednorodne z punktu widzenia ich właściwości piksele (*picture element*), tworzącymi macierz obrazu zmienności cech. Produktem jest, bliski ciągłemu, obraz zróżnicowania gleb, jeżeli tylko wielkość pikseli jest dostatecznie mała w stosunku przestrzennego zróżnicowania ich właściwości.

Istnieją już różne, pod względem konstrukcji systemów wnioskowania, realizacje tej koncepcji. Na gruncie europejskim można do nich zaliczyć system mapy gleb Unii Europejskiej, który jednak z uwagi na rozmiar piksela (kwadrat o powierzchni 100 ha) nadaje się raczej do analiz w skalach regionalnych lub krajowych. Jest to system w pełni funkcjonujący, dobrze zdefiniowany, zaopatrzony we własny mechanizm wnioskowania (zdyskretyzowane klasy cech uzyskiwane za pomocą tzw. *pedotransfer rules* – np. klasy zasobów węgla organicznego w glebach, oraz ciągłe oceny właściwości hydrologicznych ujęte w formie tak zwanych *pedotransfer functions*). Ten mechanizm, bazujący na obserwowanych właściwościach gleb pozwala na produkcję map cech pochodnych, nierejestrowanych w bazach danych.

Zaawansowana jest też koncepcja nazywana modelem *SCORPAN* [12], najbliższa chyba pierwotnemu podejściu Hansa Jenny, forsowana przez silny zespół gleboznawców australijskich. Jako system wnioskowania przewiduje on różnego rodzaju modele (statystyczne, geoinformatyczne, ewolucyjne), odzwierciedlające powiązanie między czynnikami środowiskowymi i glebowymi, z uwzględnieniem różnych podejść wygładzających wyniki (autokorelacja, kriging, co-kriging).

W USA znany jest model o nazwie *SoLIM* (*Soil Land Inference Model*) [17], którego autorzy powołują się także na propozycje Jenny, lecz odmiennie, wobec modelu *SCORPAN*, podchodzą do wnioskowania o klasie ocenianej właściwości. *SoLIM* zakłada jako wynik wnioskowania kolekcję wartości funkcji przynależności wskazujących podobieństwo do określonych wzorców właściwości (np. typów lub klas). Taka konstrukcja wyniku wnioskowania oznacza odejście od ostrej

*) AGH w Krakowie

klasyfikacji gleb (typologicznej, użytkowej, bonitacyjnej) na rzecz podejścia rozmytego. Wymaga to oczywiście stosownych reguł interpretacji uzyskanego, zagregowanego zbioru rozmytego, zwłaszcza gdy jest on bardzo rozproszony.

Należy zwrócić uwagę na dwie kluczowe właściwości wymienionych podejść: uwzględnienie ciągłej w czasie i przestrzeni zmienności cech glebowych oraz wieloznaczność ocen klas (rozumianych jako jednostki klasyfikacyjne) w zależności od analizowanej cechy.

W okolicznościach, gdy główny mechanizm systemu wnioskowania polega na założeniu współzależności przestrzennych i czasowych między czynnikami środowiska a cechami gleb pojawia się możliwość wykorzystania go nie tylko do oceny aktualnego stanu gleb, lecz także prognozowania jego zmian pod wpływem przekształcenia elementów środowiska, w tym przekształceń antropogenicznych, na przykład pogórniczych. Konieczny jest jednak odpowiednio sprawny i wiarygodny sposób budowy samego modelu i interpretowania wyników jego wskazań. Ze względu na nieliniowość zależności między cechami środowiska a cechami gleb, a także niedostatek prototypów funkcji aproksymujących je naturalnym kierunkiem poszukiwania, analogicznie jak w modelach *SCORPAN* i *SoLIM*, są algorytmy ewolucyjne, rozwiązanie stosowane przy braku efektywnych algorytmów deterministycznych.

Niezależnie od ogólniejszych koncepcji zmierzających do systemów krajowych lub regionalnych, w Polsce także podejmowane są próby lokalnych sposobów wnioskowania bazujące na algorytmach mieszczących się w obrębie inteligencji obliczeniowej. Można tu przytoczyć próby podejmowane w tej dziedzinie od kilkunastu lat [5-7, 13].

2. Problem adekwatności modeli ewolucyjnych

Modele ewolucyjne służą do rozwiązywania zadań regresyjnych lub klasyfikacyjnych. Obydwa rodzaje zadań, należące do kategorii inteligencji obliczeniowej, mogą mieć zastosowanie w cyfrowej kartografii gleb.

Modele ewolucyjne reprezentują niektóre współzależności właściwe dla danych, tym samym wybór odpowiednich danych będących nośnikiem poszukiwanych informacji, jak też danych testowych i walidacyjnych jest kluczowym problemem inteligencji obliczeniowej grupujących wiele rodzajów algorytmów ewolucyjnych optymalizacji [2, 8, 15]. Istnieje wiele ogólnie sformułowanych zasad doboru danych, ich reprezentatywności i liczby w relacji do rozmiaru modelu. Złożone zadania klasyfikacyjne, do których należy także klasyfikacja gleb, nie mogą być rozwiązywane bezbłędnie, z powodu niejednoznaczności danych lub wad modeli. Rozróżnienie przyczyn błędów modeli może być przydatne w ich doskonaleniu.

Celem pracy jest przedstawienie niektórych uwarunkowań powstawania i unikania błędów modeli glebowych. Jest to zagadnienie potencjalnie istotne z powodu potrzeby uzyskiwania wiarygodnych podstaw prognozowania. Uwarunkowania te dotyczą struktury samych modeli wnioskowania (architektury klasyfikatorów, ich rozmiaru i sposobu konstruowania), jak też działania w przypadku ich nadmiernej niedoskonałości.

2.1. Dane

Dane wykorzystane w pracy są kompilacją digitalizacji materiałów kartograficznych: mapy glebowo-rolniczej, mapy topograficznej oraz danych hydrograficznych obszaru GOP. Z utworzonej w ten sposób bazy danych, której rekordy zawierały charakterystykę pikseli terenowych, przylegających do siebie kwadratów o boku 20 m [6].

Źródłem danych użytych do badań była mapa glebowo-rolnicza oraz topograficzna części Górnośląskiego Okręgu Przemysłowego o powierzchni 2596 kilometrów kwadratowych (prostokąt o rozmiarach 59 km w kierunku WE oraz 44 km w kierunku NS). Górnośląski Okręg Przemysłowy jest zróżnicowany pod względem morfologicznym, hydrologicznym i glebowym. Cechuje go rozwinięta sieć hydrograficzna. Miejscami gleby są przekształcone z powodu wieloletniego oddziaływania górnictwa. Mapa glebowo-rolnicza wyodrębnia w tym obszarze ponad 16 000 konturów jednostek glebowych.

Zróżnicowanie typologiczne obejmuje jednostki gleb bielicowych, brunatnych, rędzin, czarnych ziem, gleb hydrogenicznych i mad. Pod względem bonitacyjnym wyróżniono tu większość jednostek szeregu bonitacyjnego gruntów ornych (od klasy II do VI), oraz pełny szereg niżowych użytków zielonych. Poza kompleksem pszennym najlepszym wyodrębniono tu wszystkie kompleksy niżowe gruntów ornych oraz trwałych użytków zielonych.

Baza danych dotyczących zróżnicowania właściwości gleb na terenie GOP została opracowana na podstawie digitalizacji mapy glebowo-rolniczej w skali 1:25 000 oraz mapy topograficznej w tej samej skali. Dane dotyczące konturów glebowych zostały przekształcone w regularną siatkę przylegających do siebie kwadratowych płatów terenu o powierzchni 400 m^2. Z 2 767 890 płatów (pikseli) z kompletem informacji dotyczących gleb i morfologii terenu (typ gleby, kompleks glebowy, rozkład uziarnienia w profilu glebowym) wyselekcjonowano losowo próbę służącą do optymalizacji modelu liczącą ponad 36 000 przypadków. Próby podobnej liczebności wyselekcjonowano w celu testowania i walidacji modelu.

2.2. Metoda

Podstawową metodą dochodzenia do modelu klasyfikacyjnego (klasyfikatora) w postępowaniu ewolucyjnym jest realizacja wielu prób z różnymi architekturami i losowo wybranymi parametrami początkowymi. Możliwe są także różne strategie konstrukcji klasyfikatorów i metody optymalizacji. Ostateczny wybór modelu następuje po jego walidacji, pod warunkiem osiągnięcia dostatecznego poziomu generalizacji danych [2, 15].

Przy bardzo złożonych problemach modele mogą nie być dostatecznie wiarygodne; w każdym przypadku zależy to od rodzaju zadania. W przypadku prognozy przekształceń gleb na terenach górniczych wymagania wobec jakości modelu powinny być stosunkowo wysokie: następstwa przekształceń dotyczą często kilku procent powierzchni w zasięgu wpływów, porównywalny poziom błędu przekreśla jego przydatność do prognozowania.

Przyczynami niedoskonałości modeli mogą być wady danych wzorcowych (na przykład niekonsekwencje podstawowego algorytmu klasyfikacji) lub ograniczenia strukturalne (zbyt mały lub duży rozmiar, zła architektura).

W dalszej części pracy przeanalizowano ścieżkę dojścia do akceptowalnej jakościowo struktury modelu współzależności cech glebowych i środowiskowych, potencjalnie przydatnej w problemach prognozowania szkód górniczych w użytkach przyrodniczych na bazie domniemanej budowy cyfrowej kartografii gleb.

2.3. Wyniki

Tabela 1 przedstawia wyniki zbiorcze charakterystyki klasyfikatorów zbudowanych na podstawie danych z GOP. Podane charakterystyki dotyczą łącznie 26 klasyfikatorów neuronowych (10 typu MLP, 8 RBF i 8 PNN) o zróżnicowanych architekturach i algorytmach optymalizacji.

Tabela 1. Niektóre właściwości klasyfikatorów tworzących zespół
Table 1. Selected properties of classifiers for an ensemble

Typ klasyfikatora	Liczba wejść modelu	Liczba jednostek ukrytych	Zakres wartości TPerf	Zakres wartości VPerf	Zakres wartości TePerf
MLP	1-16	2-24	0,34-0,75	0,34-0,73	0,34-0,73
RBF	8-16	10-72	0,45-0,61	0,45-0,61	0,46-0,60
PNN	16	10000	0,53-0,93	0,51-0,72	0,53-0,72

Objaśnienia skrótów: MLP – MultiLayer Perceptron, RBF – Radial Basis Function, PNN – Probabilistic Neural Network, TPerf – poprawność identyfikacji części treningowej zbioru treningowego, VPerf – poprawność identyfikacji części walidacyjnej zbioru treningowego,TePerf – poprawność identyfikacji części testowej zbioru treningowego

Explanation of abbreviations: MLP – MultiLayer Perceptron, RBF – Radial Basis Function, PNN – Probabilistic Neural Network, TPerf – correctness of identification of the training part of the training ensemble, VPerf - correctness of identification of the validation part of the training ensemble, TePerf - correctness of identification of the test part of the training ensemble

10 klasyfikatorów MLP (*MultiLayer Perceptron*: Perceptron Wielowarstwowy) o różnej, testowanej liczbie zmiennych wejściowych oraz zróżnicowanej liczebności jednostek ukrytych, pozwalało na poprawną identyfikację maksymalnie do 75% przypadków, odpowiednio niższą dla danych testowych i walidacyjnych (do 73%).

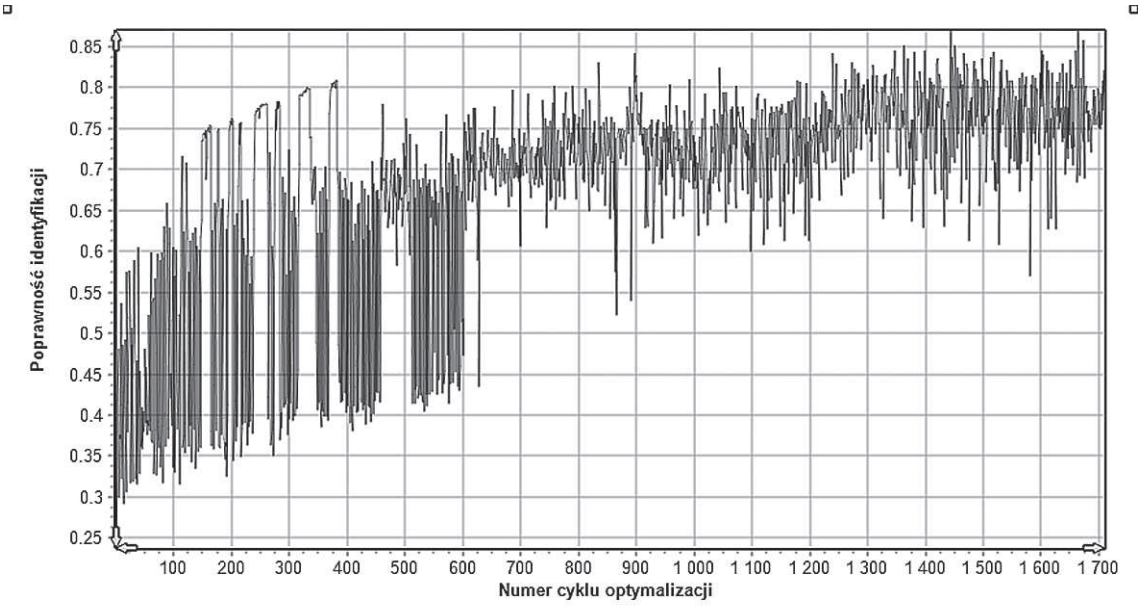
Konkurentami klasyfikatorów MLP (reprezentujących w zakresie funkcji przetwarzających wewnątrz jednostek tak zwany system z nielokalnymi funkcjami transferu) są klasyfikatory określane skrótem RBF (*Radial Basis Function*: Kołowe Funkcje Bazowe, lokalne funkcje transferu) generalnie wykazują gorsze właściwości od MLP, wskaźnik sukcesu nie przekracza tu 61%. Jako pojedyncze klasyfikatory modele typu RBF, przynajmniej w zakresie sprawdzonych rozmiarów, są zanadto zawodne.

Szczególnym rodzajem sieci są klasyfikatory typu PNN (*Probabilistic Neural Network*: Probabilistyczne Sieci Neuronowe), o budowie podobnej do RBF, lecz odmiennych funkcjach transferu w warstwie wyjściowej (softmax). Są to struktury na ogół bardzo duże, zaś ich nazwa pochodzi od prezentacji wyniku w postaci rozkładu prawdopodobieństwa szacowanych klas. Jako klasyfikatory, przy bardzo rozbudowanej architekturze, wykazują one cechy dość korzystne, choć ceną za jakość generalizacji jest czas przetwarzania danych. Jednak nawet ta wielka architektura sieci nie jest wystarczająca do uzyskania zadowalającego stanu generalizacji danych.

Stosunkowo nową alternatywą, wobec klasycznego sposobu budowy klasyfikatorów wymagającego określenia jego struktury *a priori*, są tak zwane algorytmy konstruktywistyczne [3, 9]. Architektura klasyfikatora ulega w nich modyfikacji w trakcie procesu optymalizacji. W szczególności ulega ona zmniejszeniu lub zwiększeniu w zależności od odległości od zadanego kryterium stopu. W skomplikowanych zadaniach algorytm prowadzi do wielkich rozmiarów sieci, wymagających starannej walidacji. Odpowiednio do zmian architektury i parametrów funkcji transferu wahają się chwilowe błędy klasyfikacji. Rysunek 1, obrazuje przebieg krzywej optymalizacji klasyfikatora FSM dla zadania klasyfikacji danych z rejonu GOP.

Próba budowy klasyfikatora metodą FSM (*Feature Space Mapping*: Mapowanie Przestrzeni Cech), została przerwana po czterech dniach optymalizacji przy stanie 1750 jednostek przetwarzających (architekturą sieci optymalizowaną metodą FSM jest generalnie RBF), oraz współczynniku sukcesu identyfikacji wynoszącym około 82%.

Przegląd wyników optymalizacji klasyfikatorów wskazuje, że pojedynczy algorytm ewolucyjny raczej nie doprowadzi do zadowalającej struktury generalizującej problem.



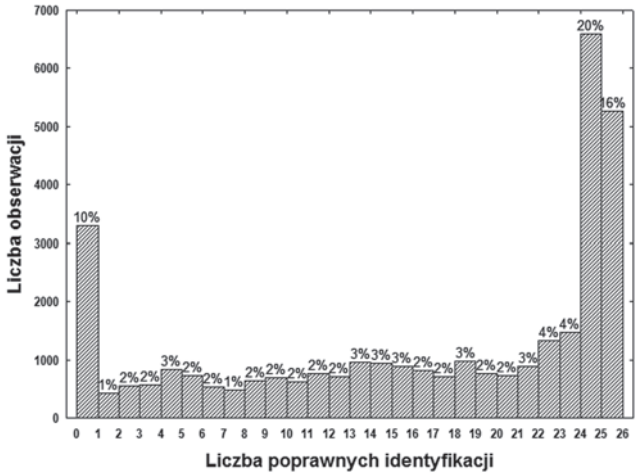
Rys. 1. Krzywa uczenia klasyfikatora FSM dla zadania identyfikacji kompleksów glebowo-rolniczych na terenie GOP

Fig. 1. Learning curve of the FSM classifier for the task of identifying the agricultural-soil complexes in the Upper Silesian Industrial Region

W tej sytuacji pojawia się perspektywa wprowadzenia bardziej złożonego podejścia, zespołu klasyfikatorów.

Uzasadnieniem teoretycznym dla stosowania zespołu klasyfikatorów (określanych mianem słabych klasyfikatorów) jest ich potencjalna komplementarność: słabość klasyfikacji w pewnym obszarze cech jednego klasyfikatora może być niwelowana przez inny klasyfikator, dobrze sprawdzający się w tym obszarze [11,14]. Sensownym założeniem jest budowa zespołu złożonego z klasyfikatorów o możliwie dobrym stopniu generalizacji lecz słabo ze sobą skorelowanych pod względem wskazań.

Uzasadniona wydaje się próba wykorzystania, jako podstawy takiego zespołu, stosunkowo słabych klasyfikatorów wypróbowanych uprzednio indywidualnie. Wśród metod wyboru klasyfikacji wskazywanej przez zespół klasyfikatorów zazwyczaj przyjmie się tę, którą wskazuje większość klasyfikatorów w wyniku głosowania większościowego. W tym przypadku, przyjmując, że poprawnie wskazywane są przypadki identyfikowane przez przynajmniej 13 z 26 klasyfikatorów, wskaźnik sukcesu wynosi zaledwie 68,5%. Jest to wynik znacząco gorszy od najlepszego klasyfikatora indywidualnego. Wykres na rysunku 2, ilustruje rozkład liczebności poprawnych wskazań przez zespół 26 klasyfikatorów.

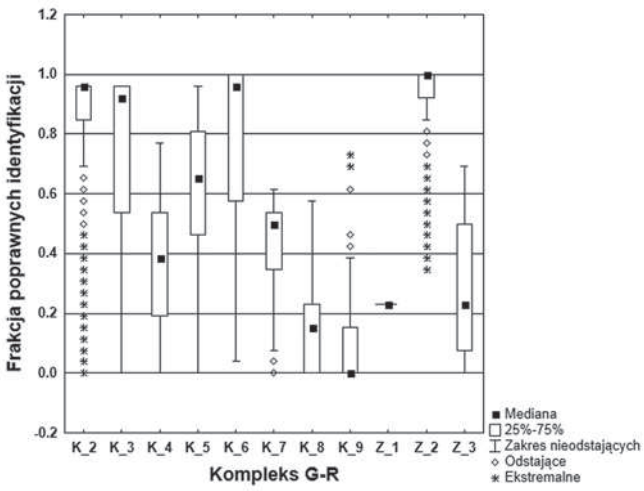


Rys. 2. Histogram rozkładu liczby poprawnych wskazań przez składowe zespołu klasyfikatorów

Fig. 2. Histogram of partition of a number of correct indications by the components of classifier ensembles

Wynika z tego, że zaledwie 16% elementów zbioru treningowego jest poprawnie identyfikowana przez wszystkie klasyfikatory.

Pouczający jest wykres ilustrujący rozkład frakcji poprawnych wskazań w poszczególnych kompleksach glebowo-rolniczych (rys. 3). Obok kompleksów, gdzie dominują poprawne wskazania (K_2, K_3, K_6, Z_2), występuje większa liczba kompleksów słabo rozróżnialnych, po części z powodu słabszej reprezentacji w zbiorze danych (mniejsza powierzchnia kompleksu) oraz podobieństwa fizjograficznego do liczniej reprezentowanego kompleksu. Ogólnie można zauważyć, że słabo rozróżniane są kompleksy o zbliżonych właściwościach oraz mało liczne przypadki (Z_1). Wypada podkreślić, że szczególnie niekorzystne są błędy popełniane w obrębie kompleksów skrajnych: podmokłych i suchych (K_7, K_8, K_9, Z_3). Ta właściwość przekreśla przydatność modelu bazującego na zespole wszystkich klasyfikatorów i głosowaniu większościowym.



Rys. 3. Rozkład frakcji poprawnych wskazań przez zespół klasyfikatorów w podziale na kompleksy przydatności rolniczej w GOP

Fig. 3. Fractional decomposition of correct indications by classifier ensembles and the agricultural usability complexes division in Upper Silesian Industrial Region

Inną możliwością interpretacji wskazań wszystkich klasyfikatorów jest wprowadzenie rozwiązania określonego mianem *stacking*. Polega ono na użyciu wskazań indywidualnych klasyfikatorów jako wartości wejściowych do wyspecjalizowanego klasyfikatora decyzyjnego, który z konfiguracji wskazań (w odniesieniu do poprawnej klasyfikacji) podejmuje decyzję o klasie obiektu [16]. Należy zwrócić uwagę, że oznacza to zazwyczaj odejście od głosowania większościowego, na rzecz badania konfiguracji wskazań. Można przypuszczać, że nawet błędne wskazania niektórych (lub wszystkich) klasyfikatorów, mogą być przesłanką poprawnej klasyfikacji końcowej (przez klasyfikator decyzyjny), o ile są one konsekwentne, a klasyfikatory tworzące zespół dostatecznie wrażliwe na zmiany oryginalnych wejść.

Wykorzystanie jako klasyfikatora decyzyjnego struktury MLP (20-30 jednostek w warstwie ukrytej) pozwoliło na poprawną klasyfikację do 78% wzorców walidacyjnych. Oznacza to znaczącą poprawę możliwości generalizujących systemu klasyfikacyjnego (z około 68% przy głosowaniu większościowym do ponad 78% w systemie z klasyfikatorem decyzyjnym).

Należy zauważyć, że znaczna liczba klasyfikatorów w zespole, zwłaszcza klasyfikatorów bardzo słabych, o niestabilnych wskazaniach może obniżać sprawność systemu poprzez zakłócanie wektora wskazań znaczną liczbą identyfikacji losowych. W takich przypadkach można przeprowadzić selekcję klasyfikatorów, pozostawiając najsilniejsze i w miarę możliwości słabo skorelowane. Przy okazji uzyskuje się mniejszy wektor wskazań, co oznacza także przyspieszenie i uproszczenia działania całego systemu, Jednym z kryteriów wyboru dobrego zbioru klasyfikatorów jest wskaźnik Q [1]

$$Q_{a.b} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \qquad (2)$$

gdzie:

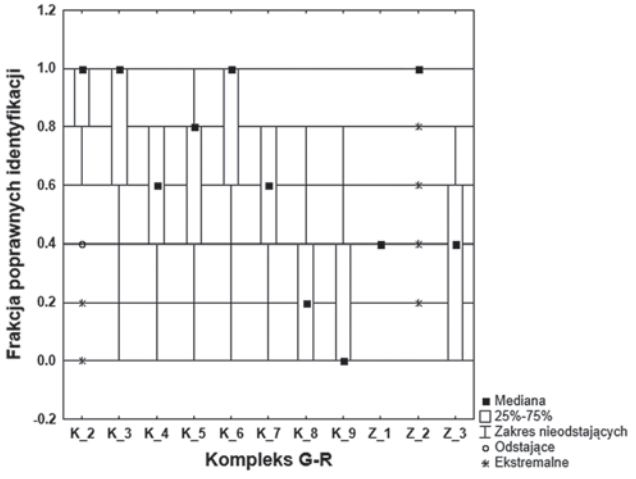
- N^{11} jest liczbą przypadków poprawnie sklasyfikowanych przez oba klasyfikatory,
- N^{00} oznacza liczbę przypadków sklasyfikowanych błędnie przez oba klasyfikatory,
- N^{10} oraz N^{01} oznaczają, odpowiednio, liczbę przypadków sklasyfikowanych poprawnie przez klasyfikator a i

błędnie przez *b* oraz liczbę przypadków o przeciwnych właściwościach,

Statystyka *Q* przyjmuje wartość 1,0 gdy, porównywane klasyfikatory dają błędne wskazania tych samych przypadków. Wartość ujemną osiąga, gdy klasyfikatory dają błędne wskazania dla rożnych przypadków. Zgodnie z tym właściwościami najlepszym zespołem będzie zbiór klasyfikatorów o najmniejszej wartości statystyki *Q*. Algorytm wyboru składników zespołu klasyfikatorów z puli 26 klasyfikatorów-kandydatów obejmował włączenie do niego najsilniejszego klasyfikatora, a następnie włączanie kolejnych charakteryzujących się najniższą, średnią wartością wskaźników *Q*, wobec klasyfikatorów obecnych już w zespole.

Rysunek 4, przedstawia rozkład frakcji poprawnych wskazań w poszczególnych kompleksach glebowo-rolniczych dla zredukowanego zestawu klasyfikatorów. W stosunku do pełnego zbioru zespołu widoczne jest ostrzejsze zróżnicowanie wskazań. Trzeba też zaznaczyć, że usunięcie najsłabszych klasyfikatorów spowodowało nieznaczny wzrost poprawnych wskazań według kryteriów głosowania większościowego do 71,6%, W dalszym ciągu zespół, którego wskazania bazują na ocenie większościowej, nie wykazuje zadowalających właściwości generalizujących.

Usunięcie najsłabszych składników zespołu wymaga ponownego podjęcia procedury *stacking* dla zmienionego zbioru danych. Zastosowanie jako klasyfikatora decyzyjnego MLP nie daje podnosi zdolności identyfikacyjnej systemu wnioskowania, MLP (8-30 jednostek w warstwie ukrytej) decyzyjny rozpoznaje poprawnie i stabilnie około 78% zbioru danych. Znaczący postęp umożliwia zastosowanie jako klasyfikatora decyzyjnego algorytmu konstruktywistycznego FSM. Po osiągnięciu 31 jednostek w warstwie ukrytej dokonuje on poprawnej klasyfikacji wszystkich elementów zbioru testowego, a co ważniejsze, prawie wszystkich elementów zbioru walidacyjnego (niewykorzystywanych wcześniej w optymalizacji i testowaniu). Oznacza to, że zredukowany do stabilnych wskazań wektor wejściowy systemu RBF zoptymalizowany algorytmem FSM wykorzystuje jako wzorcowe zaledwie 31 konfiguracji identyfikacji wzorców do praktycznie bezbłędnego rozpoznania elementów walidacyjnych (8 błędnych rozpoznań na ponad 33 tysiące elementów zbioru).



Rys. 4. Rozkład frakcji wskazań kompleksów glebowo-rolniczych w GOP przez zespół 5 klasyfikatorów wyselekcjonowanych według kryterium Q

Fig. 4. Fractional decomposition of indications of agricultural-soil complexes in Upper Silesian Industrial Region by the ensemble of 5 classifiers selected according to the Q criterion

3. Podsumowanie

Konwersja istniejącej dokumentacji kartograficznej, w tym map tematycznych, w kierunku systemów informacji przestrzennej, jest w najbliższej perspektywie nieunikniona. Konwersja bazująca na digitalizacji istniejących materiałów może mieć sens, przynajmniej jeżeli analizuje się użyteczność nowo powstającej dokumentacji, tylko w przypadku zamiaru stopniowego uzupełniania baz danych o elementy uwzględniające ciągłość zjawisk i obiektów przyrodniczych. Modele zmienności cech glebowych są nieodłączną częścią projektów cyfrowej dokumentacji gleb. Dokonane spostrzeżenia wskazują, że pojedyncze klasyfikatory, zasadniczo przypuszczalnie wystarczające do opisu zmienności gleb jako obiektów przyrodniczych, mogą być niewystarczające do wymagających większej precyzji zadań. Poszukiwanie pojedynczych klasyfikatorów osiągających około 70% poprawnych wskazań może być pośrednim etapem prowadzącym do bardziej złożonych, zespołów klasyfikatorów, pod warunkiem wykorzystania komplementarności ich wskazań. Analiza właściwości i selekcja klasyfikatorów może prowadzić do zadowalających modeli generalizujących. W prezentowanym przykładzie wykorzystano procedurę *stacking* do podejmowania decyzji o prognozowanej klasie (w rozumieniu kategorii obiektu, w tym przypadku kompleksu przydatności) gleby. Ostateczny model, złożony z zespołu pięciu klasyfikatorów neuronowych, z dodatkowym elementem decyzyjnym, bazującym na klasyfikatorze zbudowanym z użyciem algorytmu konstruktywistycznego FSM. Klasyfikator decyzyjny, złożony z 31 jednostek ukrytych, ustala prognozowaną klasę (kategorię) na podstawie konfiguracji wskazań pięciu klasyfikatorów zespołu. Z obserwacji wynika, że wskazania ostateczne są poprawne także przy braku wskazań niektórych klasyfikatorów zespołu.

Według analizy danych użytych w badaniach w zbiorze walidacyjnym obejmującym około 33 tys. rekordów danych, po przetworzeniu ich przez klasyfikatory zespołu powstaje odpowiednio zbiór pięcioelementowych wektorów wskazań. Analiza wykazała, że obejmują one 886 wariantów konfiguracji wskazań, generalnie wystarczających do poprawnej klasyfikacji praktycznie wszystkich rekordów danych, przy użyciu 31 jednostek ukrytych w klasyfikatorze decyzyjnym. Rozkład liczby wariantów konfiguracji wskazań dla poszczególnych kompleksów zawiera tabela 2.

Tablica 2. Relacja liczby wektorów wzorcowych wytworzonych przez zespól klasyfikatorów do liczb elementów walidacyjnych kompleksu

Table 2. Relation of a number of standard vectors created by classifier ensembles to the number of validation components of the complex

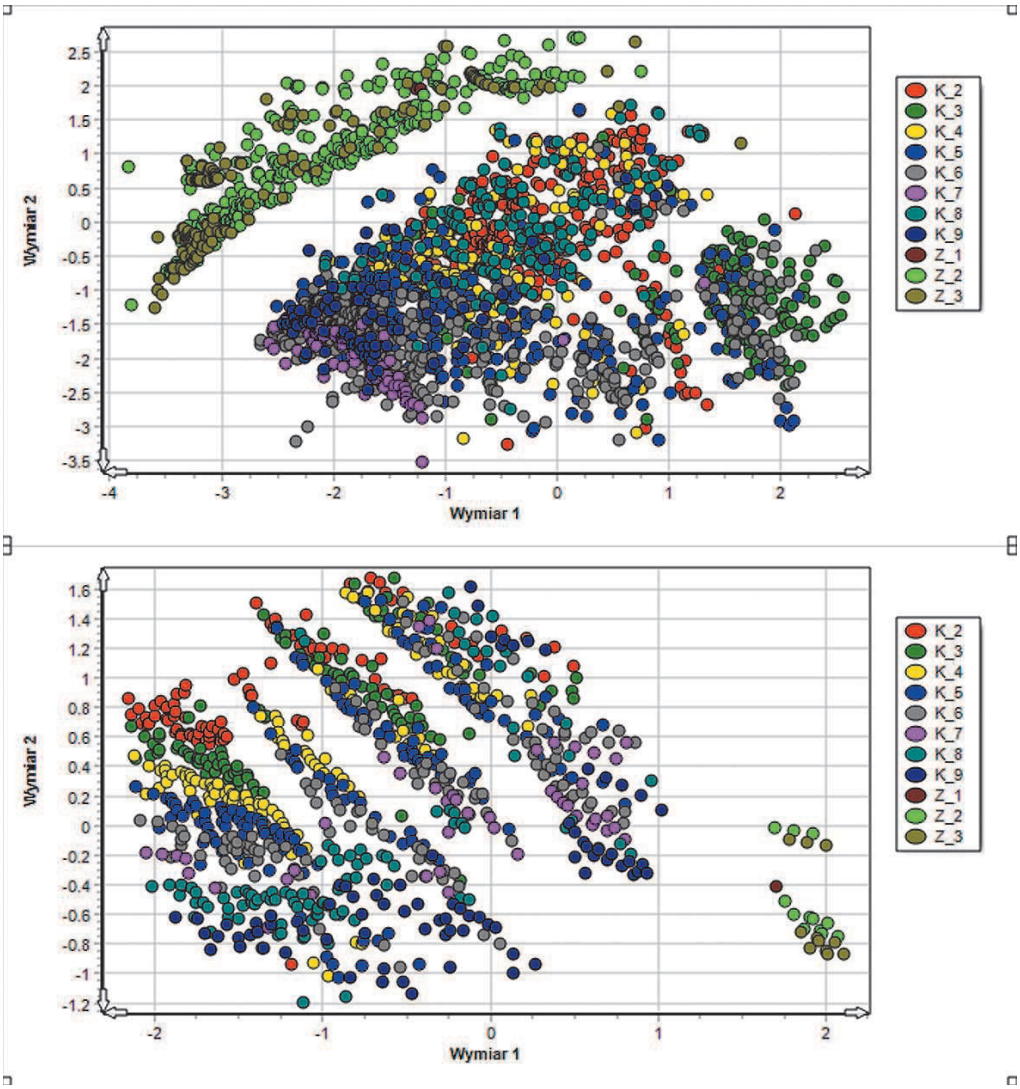
Oznaczenie kompleksu	Liczba wariantów konfiguracji wektorów	% liczby wariantów unikatowych	% liczby elementów walidacyjnych
K_2	101	11,4	17,4
K_3	109	12,3	7,8
K_4	122	13,8	7,8
K_5	169	19.1	14,7
K_6	132	14,9	18,5
K_7	49	5.5	3,1
K_8	98	11,0	6,1
K_9	83	9,4	3,3
Z_1	1	0,1	0,6
Z_2	12	1,3	15,8
Z_3	10	1,1	5,4

Ostatnia kolumna tabeli 2. zawiera procent liczby elementów danego kompleksu znajdującą się w zbiorze walidacyjnym (i treningowym). Należałoby oczekiwać, że przy większym udziale powierzchniowym kompleksu także liczba potencjalnych wariantów właściwych dla ustalenia kompleksu powinna być większa. Ta zależność jednak się nie sprawdza. Przy znaczącym udziale kompleksu K_2 w powierzchni terenu, liczba unikatowych wariantów wskazań jest znacznie mniejsza, odmiennie niż jest to w przypadku kompleksu K_3, co może być pewną niespodzianką. Widoczna jest także nadreprezentacja wariantów kompleksu K_8, który zapewne jest bardziej zróżnicowany niż K_6. Charakterystyczna jest mało liczna reprezentacja kompleksów trwałych użytków zielonych (w tym obrębie występują jednak pojedyncze błędy rozpoznania).

Skutek wykorzystania, w miejsce danych surowych (nieprzetworzonych) wektorów wskazań zespołu klasyfikatorów można zaobserwować na wykresach MDS (*Multi Dimensional Scalling*) przedstawionych na rys. 5. Efektem przetworzenia danych przez zespół jest wyraźniejsze rozdzielenie danych (obraz MDS jest przekształceniem obrazu wielowymiarowego wzajemnego położenia obiektów do obrazu jedno- lub dwuwymiarowego.

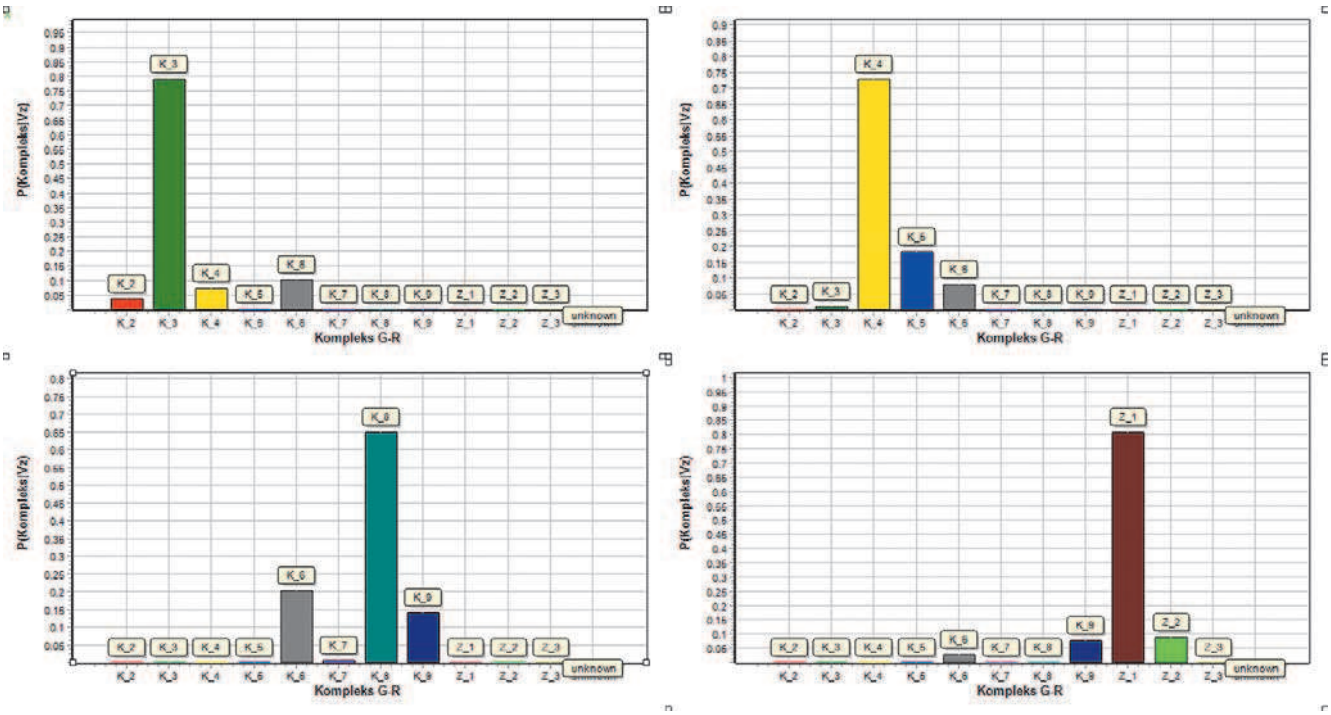
Dalszą konsekwencją odmiennego podejścia do problemu klasyfikacji gleb jest rozmycie klasyfikacji. Zgodnie z tym jednoznaczne przypisanie fragmentowi terenu jednej klasy (kategorii, kompleksu przydatności rolniczej lub typu siedliskowego) może być rzadkością. Praktycznie każdy fragment charakteryzuje się jedynie mniejszym lub większym podobieństwem do określonej kategorii. Oczywiście potrzeba jednolitej klasyfikacji nakazuje przyjęcie jako reprezentacji fragmentu kategorii o największym prawdopodobieństwie podobieństwa do wzorca klasy, jednak w przypadku prognozowania skutków przekształceń gleb dostęp do wektora prawdopodobieństw klas (w miejsce ostro wyznaczonej klasy) daje większe możliwości oceny skutków przekształceń. Przykładowe rozkłady prawdopodobieństw kompleksów dla różnych wektorów wskazań zespołu obrazuje rysunek 6.

Całość procedury daje się opisać jako kolejne konwersje prowadzące do oceny prognozowanego stanu gleby: poczynając od wektora $\boldsymbol{L}_{xy}$ obejmującego fizyczną charakterystykę gleb w konkretnym punkcie przestrzeni (uziarnienie, nachylenie, uwilgotnienie, konfiguracja otoczenia itp.), który w wyniku wykorzystania zoptymalizowanych i wyselekcjonowanych klasyfikatorów, jest przekształcany w wektor $\boldsymbol{V}_{xy}$ wskazań składowych zespołu, a ten z kolei w wektor $\boldsymbol{P(K}_{xy}$



Rys. 5. Analiza podobieństwa danych metodą MDS. Wykres górny z uwzględnieniem danych surowych, wykres dolny - obraz po przetworzeniu przez zespół pięciu klasyfikatorów

Fig. 5. Assessing the similarities of data by MDS metod. The upper chart with raw data, the lower chart – image after processing by the ensemble of 5 classifier



Rys. 6. Rozkłady prawdopodobieństw przynależności wektorów wskazań zespołu do określonego kompleksu
Fig. 6. Distribution function for the ensemble indication vectors fitting into a particular complex

V_{xy}) prawdopodobieństw określonego stanu gleb (przynależności do określonej klasy lub kompleksu). Oznacza to pośrednio nieco inne podejście do problemu prognozowania szkód przemysłowych (w tym górniczych), w którym granice między stanami: aktualnym i prognozowanym, są nieostre, wyrażane za pomocą różnic prawdopodobieństw określonych stanów. Prawdopdobną niedogodnością jest wykraczanie prognozowanego (nowego) stanu poza zakres charakterystyk całości terenu, co może prowadzić do nieznanego przyszłego stanu.

Należy podkreślić, że prezentowane podejście jest substytutem rozwiązania suboptymalnego. Digitalizacja istniejących materiałów kartograficznych oznacza przeniesienie dyskretnego obrazu zróżnicowania do bazy danych. Należy oczekiwać, że w dążeniu do uelastycznienia obrazu zmienności gleb powinny się pojawiać coraz obszerniejsze fragmenty terenu objęte celową inwentaryzacją stanu gleb z odpowiednią rozdzielczością, albo co najmiej, dostarczające obszerniejszych zbiorów danych treningowych, użytecznych w tworzeniu koniecznych, ciągłych modeli glebowych. Nawet jednak ten substytut jest pewnym postępem, na co wskazuje koncentrowanie się prognozowanych przekształceń zazwyczaj w otoczeniu granic konturów glebowych, czyli w peryferyjnych częściach zamienność cech glebowych określonych klas.

Literatura

1. *Aksela M.*: Comparison of Classifier Selection Methods for Improving Committee Performance, in *Proceedings of MCS2003*, 2003, pp, 84-93.
2. *Bishop C.M.*: Neural Networks for Pattern Recognition, (Oxford University Press, Oxford), 1995.
3. *Duch W., Grabczewski W.*: Heterogeneous adaptive systems, Neural Networks, IJCNN '02, Proceedings of the 2002 International Joint Conference.
4. *Duch W., Setiono R., Żurada J.M.*: Computational Intelligence Methods for Rule-Based Data Understanding, *Proceedings of the IEEE, vol, 92, no, 5*, 2004, pp, 771-805.
5. *Gruszczyński S.*: Symulacja skutków przekształceń gleb na terenach górniczych za pomocą klasyfikatorów neuronowych, Uczelniane Wydawnictwa Naukowo-Dydaktyczne, Rozprawy Monografie, AGH Kraków, 2000.
6. *Gruszczyński S.*: Application of a set of heterogeneous neural networks to modelling soil classification in mining regions. W: MPES 2009; SWEMP 2009 [Dokument elektroniczny]: Mine Planning and Equipment Selection and Environmental issues and waste management in energy and mineral production: Banff, Alberta, Canada, November 16–19, 2009 : proceedings of the eighteenth international symposium and the eleventh international symposium, 244–254.
7. *Gruszczyński S.*: An ensemble of neural classifiers and constructivist algorithms in the identification of agricultural suitability complexes of soils on the basis of physiographic information ISRN Soil Science ISSN 2090-875X, 2012 art,\,ID 610567 s, 1–9, Tryb dostępu: {http://www,i-srn,com/journals/ss/2012/610567/} [2012-05-18], Bibliogr, 2012, s, 9.
8. *Hecht-Nielsen R.*: Neurocomputing, Addison-Wesley, Reading 1991
9. *Jankowski N.*: Ontogeniczne sieci neuronowe, O sieciach zmieniających swoją strukturę, Warszawa 2004.
10. *Jenny H.*: Factors of Soil Formation, A System of Quantitative Pedology, New York: Dover Press, (Reprint, with Foreword by R, Amundson, of the 1941 McGraw-Hill publication) 1994.
11. *Kuncheva L., Bezdek J., Duin R.*: Decision templates for multiple classifier fusion: an experimental comparison, Pattern Recognition, vol, 34 (2), 2001, pp, 299-314.
12. *McBratney A.B., Mendonca Santos M.L., Minasny B.*: On digital soil mapping, Geoderma, v. 117, 2003, pp. 3-52.
13. *Olszewski*: Kartograficzne modelowanie rzeźby terenu metodami inteligencji obliczeniowej na Wyd, Politechniki Warszawskiej, 2009.
14. *Pal S.K., Mitra P.*: Pattern recognition algorithms for data mining, Scalability, knowledge discovery and soft granular computing, Chapman and Hall/CRC Press Company, Boca Raton-London-New York-Washington D,C 2004.
15. *Tadeusiewicz R.*: Sieci neuronowe. Akademicka Oficyna Wydawnicza, Warszawa 1993.
16. *Wolpert D.*: Stacked generalization, Neural Networks, vol. 5, 1992 pp, 241-259.
17. *Zhu A-X.*: Mapping soil landscape as spatial continua: the neural network approach, Water Resources Research, vol. 36 (3), 2000 pp, 633-677.