Nr 113 Prace Naukowe Instytutu Górnictwa Politechniki Wrocławskiej Nr 113

Studia i Materiały Nr 31 2005

Data Mining, GIS, metody analizy danych

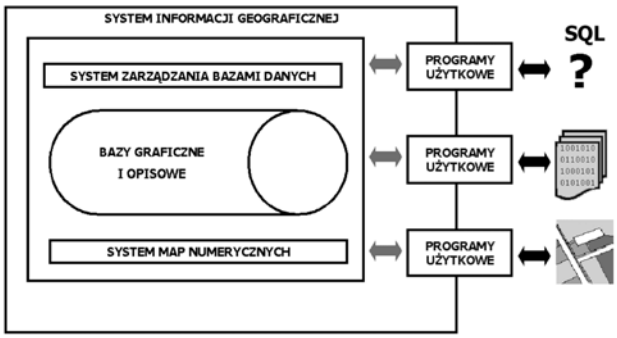
Justyna GÓRNIAK- ZIMROZ*, Józef WOŹNIAK*, Radosław ZIMROZ*

MOŻLIWOŚCI METOD DATA MINING W GEOGRAFICZNYCH SYSTEMACH INFORMACYJNYCH ZORIENTOWANYCH NA ZARZĄDZANIE ZASOBAMI ZIEMI

Metody Data Mining dostarczają nową jakość i zakres analiz danych w systemach geoinformacyjnych, poprzez odkrywanie dodatkowej wiedzy zawartej w zintegrowanych danych przestrzennych i nie przestrzennych. Stosowane są do tego celu coraz popularniejsze metody i techniki oparte na sieciach neuronowych, modelowaniu statystycznym, czy zbiorach rozmytych W artykule zaprezentowano wyniki prac prowadzonych w Instytucie Górnictwa Politechniki Wrocławskiej w zakresie metod przetwarzania danych w systemach GIS dotyczące analizy danych opisujących właściwości wyrobisk górniczych, w odniesieniu do podobnych prac realizowanych w wiodących ośrodkach na świecie

1. WPROWADZENIE

Dynamiczny wzrost liczby danych, znaczenie posiadania informacji oraz duża popularność komputerowych baz danych powodują rosnące zainteresowanie technikami automatycznego, inteligentnego i szybkiego przetwarzania danych w celu pozyskania wiedzy o obiektach opisanych przez dane znajdujące się w bazie danych. Szczególne miejsce w klasie systemów bazodanowych zajmują systemy informacji geograficznej przeznaczone do zbierania, weryfikowania, przechowywania, integrowania, analiz przestrzennych i wizualizacji



Rys. 1. Struktura systemów GIS
Fig. 1. Structure of GIS

* Instytut Górnictwa Politechniki Wrocławskiej, pl. Teatralny 2, 50-051 Wrocław, e-mail: justyna.gorniak-zimroz@pwr.wroc.pl

danych przestrzennie odniesionych do powierzchni Ziemi, w których każdy element jest opisany przez współrzędne X, Y lub opcjonalnie Z. Strukturę systemów GIS ukierunkowanych na analizy przestrzenne ilustruje rysunek 1.

Istotą systemów GIS w sensie funkcjonalnym są analizy przestrzenne realizowane albo poprzez predefiniowane w systemie raporty, zestawienia, wykresy wraz z wizualizacją przestrzenną lub przy użyciu strukturalnego języka zapytań (SQL) do zintegrowanej graficznej i opisowej bazy (Woźniak i Ferenc, 2004). Bazę graficzną tworzą numeryczne mapy tematyczne, ortofotomapy, numeryczne modele terenu, w zakresie tematycznym opracowania. Poszczególne obiekty bazy graficznej są połączone z bazą opisową, której zawartość wynika ze struktury fizycznej - wynikającej głównie z rodzaju i zakresu informacji zawartych w bazie danych. Obecnie w wielu krajach systemy GIS mają szerokie zastosowanie w rozwiązywaniu zagadnień związanych z planowaniem zagospodarowania przestrzennego, wspomaganiu zarządzania, a zwłaszcza z planowaniem działalności w strefie usług publicznych. Systemy te są stosowane w obszarach gdzie rozpatrywane zadania są szczególnie złożone i do ich rozwiązania potrzebne jest pozyskanie i przetwarzanie dużych zbiorów danych pochodzących z różnych źródeł (Gaździcki, 2001; Kraak i Ormeling, 1998; Urbański, 1997; Zapart, 1994).

W Instytucie Górnictwa Politechniki Wrocławskiej od wielu lat prowadzone są prace dotyczące projektowania i wdrażania systemów GIS w szeroko pojętym zarządzaniu zasobami ziemi. W ostatnich latach w prowadzonych badaniach szczególny nacisk położono na efektywne wykorzystanie systemów GIS, czyli na projektowanie i opracowanie modułów realizujących funkcje systemów wspomagania decyzji z wykorzystaniem oprogramowania światowych liderów - ESRI, Bentley i Intergraph, jak również inteligentnych metod opartych na sieciach neuronowych. Naturalnym rozwinięciem tego kierunku jest wykorzystywanie technik „Data Mining" do odkrywania wiedzy zawartej w danych tzn. do grupowania, klasyfikacji i asocjacji danych, czyli do odkrywania niejawnych zależności pomiędzy danymi czy wreszcie ich predykcji.

W opracowaniu przedstawiono wyniki dotychczasowych prac prowadzonych w Instytucie Górnictwa w zakresie metod przetwarzania danych w systemach GIS oraz omówiono kierunki aktualnie prowadzonych badań w odniesieniu do prac realizowanych w innych ośrodkach na świecie. Ze względu na częsty brak dobrych odpowiedników terminów w języku polskim autorzy artykułu zdecydowali podawać, oprócz terminów w języku polskim, oryginalne terminy w języku angielskim.

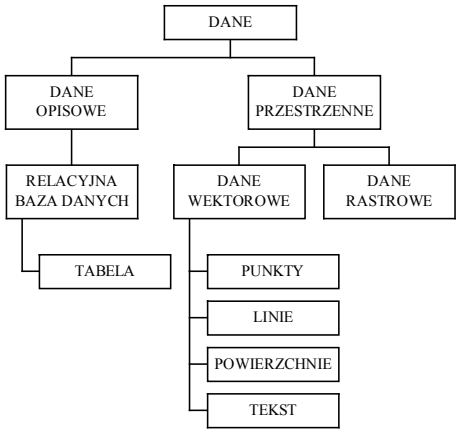
2. ISTNIEJĄCY STAN WIEDZY

Data Mining to proces analizy danych mający na celu odkrycie nieznanej dotąd wiedzy ukrytej w danych. Polskie odpowiedniki tego terminu to eksploracja danych

lub odkrywanie wiedzy w bazach danych. Źródłem dla poszukiwań wiedzy ukrytej w danych są zwykle duże zbiory danych (ang. *large databases*) lub hurtownie danych. System informacji geograficznej do zarządzania zasobami ziemi w ujęciu regionalnym opisany w artykule jest niewątpliwie przykładem dużej bazy danych, w której można zastosować Data Mining.

Generalnie proces Data Mining można podzielić na trzy etapy: pierwszy etap to przygotowanie danych, czyli oczyszczanie danych, ich integracja, selekcja, transformacja itp., drugi etap to zasadniczy proces Data Mining polegający na odkryciu wiedzy w danych, często z wykorzystaniem inteligentnych metod oraz trzeci etap to ocena i prezentacja wiedzy. Poprzez zasadniczy proces Data Mining rozumie się szereg metod takich jak: grupowanie, klasyfikacja, predykcja, asocjacja, wykrywanie odchyleń i inne. Szczególne miejsce w Data Mining zajmuje analiza szeregów czasowych polegająca na odnajdywaniu zależności i interesujących właściwości z poszukiwaniem sekwencji i wzorców.

Zadania Data Mining zależą od typów danych (rys. 2.). W zależności czy mówimy o danych opisowych, liczbowych czy szczególnym przypadku danych liczbowych – danych binarnych stosowane są specjalnie opracowane w tym celu algorytmy.



Rys. 2. Podział danych w systemach GIS
Fig. 2. Data division in GIS

Ze względu na specyficzny charakter systemów GIS specjalnie należy wyróżnić tutaj operacje na danych przestrzennych. Najbardziej zaawansowane prace w zakresie metod Data Mining w analizie danych przestrzennych prowadzone są w Stanach Zjednoczonych i w Kanadzie w ramach zleceń rządowych. W krajach tych w ostatnich latach przeprowadzono wiele badań z zakresu odkrywania wiedzy w relacyjnych bazach danych oraz analiz Data Mining dla danych przestrzennych (ang. *spatial data mining*), które podobnie jak dla danych opisowych odnoszą się do ekstrakcji ukrytej niejawnej wiedzy, zależności pomiędzy danymi przestrzennymi, czy innych wzorców nie wprost przechowywanych w bazach danych.

Według Hana i Kambera (Han i Kamber, 2000) dane przestrzenne mają wiele cech odróżniających je od danych przechowywanych w bazach relacyjnych. Dostarczają informacji topologicznych i/lub informacji o odległościach zwykle zorganizowanych jako struktury indeksowania. Podobne opinie wyrażono w pracy Estera i innych (Ester i in., 2000), w której autor twierdzi, że procedury Data Mining dla danych przestrzennych i danych opisowych znacznie się różnią. Jako główną różnicę

pomiędzy Data Mining w tradycyjnych relacyjnych bazach danych i przestrzennych bazach danych wskazuje fakt, że niektóre atrybuty znajdujące się w sąsiedztwie obiektu zainteresowań mogą wpływać na ten obiekt i dlatego muszą być uwzględniane w procesie wnioskowania. Wyrażone w sposób jawny położenie obiektów definiuje niejawne relacje przestrzennego sąsiedztwa - takie jak np. topologiczne odległościowe i kierunkowe współzależności między danymi.

„Spatial Data Mining" może być użyte do przeglądania baz danych przestrzennych, rozumienia danych przestrzennych, odkrywania relacji między danymi przestrzennymi oraz pomiędzy danymi przestrzennymi i opisowymi, reorganizacji bazy danych, projektowania bazy wiedzy, optymalizacji zapytań itd. Celowe zatem wydaje się rozpatrywanie procesu Data Mining w bazach GIS na trzech płaszczyznach:

- Data Mining tylko dla danych opisowych - niezależnych od położenia
- Data Mining dla danych przestrzennych
- Data Mining obejmujące współzależności pomiędzy danymi opisowymi i przestrzennymi

Według Andersa (Anders, 2001) w przetwarzaniu i analizie danych przestrzennych wyróżnić można następujące zadania: metody klasyfikacji, analiza skupień (2 metody – oparte na funkcji gęstości prawdopodobieństwa i na funkcji odległości), asocjacja, agregacja, aproksymacja, analiza szeregów czasowych, modelowanie współzależności, analiza odchyleń oraz predykcja.

Z kolei wspomniani wcześniej Han i Kamber (Han i Kamber, 2000) sugerują, że w procesie Data Mining metody odkrywania wiedzy z przestrzennych baz danych mogą przyjmować różne formy włączając reguły do ogólnego opisu danych przestrzennych, reguły klasyfikujące dane przestrzenne – opisujące przynależność obiektów do grupy obiektów o podobnych parametrach, reguły kojarzące daną/zbiór danych z inną daną/zbiorem danych, reguły opisujące odchylenia i trendy w procesach chwilowych zmian parametrów oraz reguły opisujące wyraźne struktury czy klasy. W pracach Koperskiego i innych (Koperski i in., 1998) oraz Hana i Kambera (Han i Kamber, 2000) można znaleźć szczegółowy opis algorytmów do przestrzennego Data Mining wraz z przykładami aplikacji metod do np. poszukiwania zależności pomiędzy danymi w procesie odkrywania wiedzy dotyczącego tropikalnych huraganów.

Systemy wykorzystujące inteligentne techniki analizy danych zwykle zbudowane są w sposób modułowy, a jednym z tych modułów jest moduł Data Mining. Moduł taki może wykorzystywać wspomniane metody statystyczne, metody uczenia maszynowego oraz metody Data Mining. Podejście statystyczne jest najczęściej używaną techniką w analizie danych przestrzennych. Zwykle podejście statystyczne opiera się na założeniu statystycznej niezależności przestrzennie rozmieszczonych danych, co może powodować sporo problemów ze względu na fakt współzależności tych danych w praktyce – tzn. obiekty przestrzenne są wzajemnie powiązane z sąsiadującymi obiektami. Zaawansowane metody statystyczne pozwalające na

uwzględnianie współzależności bądź eliminację lub minimalizację skutków pomijania faktu współzależności powodują radykalne zwiększenie złożoności obliczeniowej. Można zatem powiedzieć, że metody statystyczne nie powinny być stosowane bez przeprowadzenia wstępnych analiz wymagających sporej wiedzy z zakresu analizy danych, co praktycznie wyklucza ich powszechne stosowanie w postaci wbudowanej analizy.

Co więcej wielu autorów prac pokazuje, że modelowanie reguł współzależności o nieliniowym charakterze, czy przetwarzanie danych symbolicznych (np. „nazwa") przy wykorzystaniu metod statystycznych jest utrudnione. Prace prowadzone przez zespół Shekhara (Shekhar i in., 2003) z powodzeniem wykorzystują np. skomplikowane modele statystyczne (ang. *Spatial Autoregressive Models*). Więcej informacji na temat metod, algorytmów, oraz kierunków badań w analizie danych przestrzennych znaleźć można w pracy (Shekhar i in., 2003).

Przed rozpoczęciem procesu Data Mining, niejednokrotnie ze względu na niekompletność, niespójność, wielowymiarowość itp. danych należy odpowiednio je przygotować i proces ten według wielu autorów jest najważniejszym etapem przetwarzania danych (Pyle, 1999) - patrz następny rozdział.

3. ZASTOSOWANIE DATA MINING W GIS NA POTRZEBY ZARZĄDZANIA ZASOBAMI ZIEMI - CASE STUDY

3.1. PRZYGOTOWANIE DANYCH

W artykule skoncentrowano się na analizie danych opisowych przedstawionych zarówno w postaci numerycznej jak i tekstowej opisujących wyrobiska górnicze zlokalizowane na terenie Dolnego Śląska, ich otoczenie oraz wpływ wyrobiska na to otoczenie. Ze względu na dużą liczbę obiektów w tym regionie – ponad 700 obiektów – przeprowadzone badania empiryczne zawężono do 197 obiektów z byłego województwa wrocławskiego. Zastosowane ograniczenie liczby obiektów nie wpłynęło na zastosowaną w badaniach metodę analizy danych.

Pozyskiwanie danych - Data collection

Ze względu na złożoność rozpatrywanego zagadnienia do jego prawidłowego rozwiązania i funkcjonowania należało zebrać bardzo wiele danych z różnych źródeł. Ogólnie źródła pozyskiwania danych podzielono na siedem grup: administrację publiczną, instytucje państwowe, jednostki badawczo-rozwojowe, przedsiębiorstwa górnicze, archiwa i biblioteki, badania empiryczne oraz inne. Więcej na temat organizacji pozyskiwania danych można znaleźć w pracach Górniak-Zimroz (Górniak-Zimroz, 2004 a,b).

Przechowywanie danych - Data storage

Ze względu na olbrzymie ilości danych - docelowo dane będą dotyczyły około 700 obiektów górniczych - dane opisujące wyrobiska górnicze i ich otoczenie

zarchiwizowano w postaci relacyjnej bazy danych w środowisku Microsoft Access, a dane przestrzenne w ArcMap. Integrację tych dwóch baz wykonano w środowisku ArcGIS uwzględniającym przestrzenne parametry danych i umożliwiającym użytkownikowi lepsze zarządzanie danymi oraz nakładanie na siebie warstw z przefiltrowanymi przez użytkownika danymi w celu przeprowadzenia analiz przestrzennych pozwalających na wydobycie lub stworzenie na ich podstawie nowych informacji.

Odszumianie danych - Data cleaning

Większość istniejących baz danych zawiera dane zaszumione. Źródłem tych zakłóceń – w postaci szumu informacyjnego – mogą być nie zidentyfikowane w całości podczas kontroli jakości danych błędy instrumentów pomiarowych, rodzaju zastosowanej metody pomiaru, braku lub błędów kalibracji, błędy w odczytach w czasie pomiarów, błędy w transmisji danych, ograniczenia sprzętowe (efekt nasycenia – ang. *overload*), kodowanie danych i inne. Pod pojęciem „data cleaning" w literaturze rozumie się także usuwanie lub korygowanie danych niespójnych lub sprzecznych.

Postępowanie z brakującymi/niekompletnymi danymi - Missing value problem

Konsekwencją działań opisanych w poprzednim punkcie są niekompletne dane. Dodatkowo niekompletność danych może być wywołana innymi czynnikami takimi jak: brak dostępności danych, awarie sprzętu, tzw. „czynnik ludzki" i inne. W literaturze przedmiotu opisano metody postępowania w przypadku brakujących wartości (Pyle, 1999), czyli:

- ignorowanie kolumn z brakującymi danymi – najprostszy ale czasem zupełnie nieprzydatny sposób, zwłaszcza w przypadku dużych baz danych może się okazać, że każda kolumna zawiera brakujące dane
- ręczne wypełnianie pól – czasochłonne i wymagające danych do wprowadzenia
- metoda globalnej stałej – wypełnianie brakujących pól stałą np.: „nieznany" lub „∞". Podejście takie może błędnie wygenerować regułę, która w rzeczywistości nie istnieje
- metoda średniej globalnej – polega na wypełnianiu brakujących pól wartością średnią wyznaczoną z istniejących w bazie wartości tego samego atrybutu
- użycie średniej wartości atrybutów należących do tej samej klasy
- użycie najbardziej prawdopodobnej wartości.

W czasie prowadzonych badań autorzy również borykali się wielokrotnie z problemem niepełnych danych. Podczas wypełniania baz danymi wykorzystano tę ostatnią opcję, czyli ekspert na podstawie znajomości problemu szacował wartości danych wpisywanych do bazy danych opisowych.

Integracja danych - Data integration

Zwykle proces integracji danych sprowadza się do przetworzenia danych w jeden spójny format danych (ang. *data cube*) (Han i Kamber, 2000). Dane zbierane

i przechowywane przez rożne ośrodki - patrz pozyskiwanie danych - często zawierają te same informacje. Brak standaryzacji powoduje np. nadawanie różnych nazw tym samym atrybutom, np. ncustomer_id lub cust_id lub atrybuty o tych samych nazwach opisujące te same obiekty mogą przyjmować różne wartości np.: „Radek", „Radosław" lub po prostu „R". W procesie rozpoznawania dane opisowe po konwersji do formatu liczbowego przyjęłyby zupełnie różne wartości powodując np. rozbieżne wyniki klasyfikacji.

Konwersja danych - Data conversion

W przeprowadzanych badaniach konwersja typów danych miała duże znaczenie, ponieważ przed wykonaniem dalszych analiz istniała potrzeba konwersji danych opisowych takich jak np.: duży, mały, średni lub korzystny, obojętny, niekorzystny na dane liczbowe typu: -1, 0, 1. W celu wykonania konwersji danych opracowano procedury automatycznej zamiany danych opisowych na zbiór wartości {-1 0 1}. Ponadto zautomatyzowano procedurę zamiany przecinka na kropkę – obliczenia wykonywane były w środowisku Matlab wymagającym używania kropki jako separatora w reprezentowaniu liczb rzeczywistych.

Wybór cech – Data selection – etap 1

Zaprojektowana w ramach badań baza danych ma charakter kompleksowy i zawiera dane o wyrobiskach oraz dane opisujące ich otoczenie. Do wykonania analizy dotyczącej rozpoznania i klasyfikacji wyrobisk górniczych pod kątem lokalizacji w nich składowisk odpadów komunalnych nie są potrzebne wszystkie dane znajdujące się w bazie danych. Powstaje zatem pytanie jakie parametry mogą wpływać na wynik procesu Data Mining? Na podstawie szczegółowej analizy danych literaturowych dotyczących określenia czynników decydujących o wyborze kierunku rekultywacji, czynników warunkujących wykorzystanie terenów poeksploatacyjnych do składowania odpadów, czynników uwzględnianych podczas wyboru terenu pod lokalizację składowiska oraz analizy prawa polskiego i unijnego dotyczącego ochrony środowiska naturalnego oraz gospodarki odpadami zebrano i wyselekcjonowano czynniki opisujące stan i przydatność wyrobiska górniczego w gospodarce odpadami, czyli określających tak zwaną „jakość wyrobiska". Czynniki te podzielono na cztery grupy: czynniki przestrzenne, czynniki, przyrodnicze, czynniki techniczne i czynniki społeczne. Więcej informacji na ten temat można odnaleźć w pracach Górniak-Zimroz (Górniak-Zimroz, 2004 a,b).

Wybór cech – Data selection – etap 2

W analizie danych – np. w przypadku klasyfikacji – rozpatrywanie zmiennej o stałej wartości – w szczególnym przypadku zerowej – która nie różnicuje w żaden sposób klas nie ma sensu. Przygotowane do analizy dane w postaci macierzy M{197 x 98} gdzie 197 oznacza liczbę rozpatrywanych wyrobisk, a 98 liczbę wyodrębnionych danych opisujących „jakość wyrobiska" poddano prostej analizie statystycznej w postaci wariancji. Analiza ta pozwoliła na wyodrębnienie zmiennych o zerowej wariancji, czyli danych o stałej wartości. Dane te podczas tworzenia danych

wejściowych można pominąć, ponieważ nie mają one żadnego wpływu na wynik klasyfikacji.

Transformacje danych - Data transformation

Najczęściej spotykaną formą transformacji danych jest normalizacja, pozwalająca na zwiększenie skuteczności i szybkości obliczeniowej algorytmów wnioskowania wykorzystujących pomiar odległości w przestrzeni N-wymiarowej. Inne znane transformacje danych - stosowane popularne zwłaszcza w przypadku analizy szeregów czasowych czy obrazów - to transformaty Fouriera i będąca jej uogólnieniem analiza falkowa. W opisywanym przypadku wykorzystano normalizację danych.

Redukcja wymiaru danych – Data reduction

Analiza danych wielowymiarowych jest zadaniem skomplikowanym obliczeniowo - tym bardziej im większy jest wymiar danych. Dodatkowo np. przy stosowaniu sieci neuronowych duży wymiar danych wymaga zwiększania liczby przypadków uczących. Niemożliwa jest także ich wizualizacja. Zatem uzasadnione jest redukowanie wymiaru danych poprzez agregacje, grupowanie, analizę współzależności, analizę zmienności itd.

Popularną metodą stosowaną do redukcji danych jest metoda składowych głównych (ang. *Principal Component Analysis* - PCA), która jest techniką przetwarzania danych wielowymiarowych w przestrzeni R^N (N liczba określająca ilość zmiennych). Metoda ta opiera się na zaawansowanym aparacie matematycznym polegającym na poszukiwaniu nowego układu współrzędnych opisującego przestrzeń R^D gdzie D<<N. Symulacje przeprowadzone w opisywanym przypadku wykonano w darmowym pakiecie SOM Toolbox pracującym w środowisku MatLab, przy pomocy którego wyznaczono udziały poszczególnych składowych głównych oraz udział skumulowany będący sumą kolejnych składowych głównych. Z przeprowadzonych obliczeń wynika, że już 38 składowych głównych reprezentuje w 95% zmienność zestawu danych, co jest wystarczające do przeprowadzenia klasyfikacji.

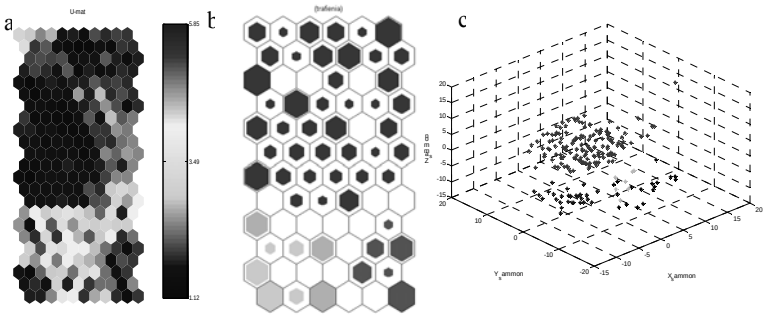
3.2 PROCES DATA MINING

Eksploracja danych – Data Mining

W analizowanym przykładzie nie wiemy czy w ogóle - a jeśli tak to ile występuje klas obiektów o zbliżonych/podobnych cechach. Dlatego też do analizy rozpoznania i klasyfikacji wyrobisk górniczych pod kątem lokalizacji w nich składowisk odpadów komunalnych wykorzystano jedną z najbardziej popularnych metod analizy danych wielowymiarowych, czyli wykorzystano projekcję przestrzeni N wymiarowej na przestrzeń 2 lub 3 wymiarową. W celu przeprowadzenia tej analizy w badaniach wykorzystano samoorganizującą się sieć Kohonena (przestrzeń 2-wymiarowa) oraz przekształcenie Sammona (przestrzeń 3-wymiarowa). Opis działania samoorganizującej się sieci Kohonena można znaleźć w pracach Korbicza i innych (Korbicz i in., 1994, 2002), Osowskiego (Osowski, 1994,1996) i Żurady i innych

(Żurada i in., 1996), a szczegółowy opis aparatu matematycznego odwzorowania Sammona w pracach Bartkowiak (Bartkowiak, 2002), Bartkowiak i innych (Bartkowiak i in., 2004) i Osowskiego (Osowski, 1996).

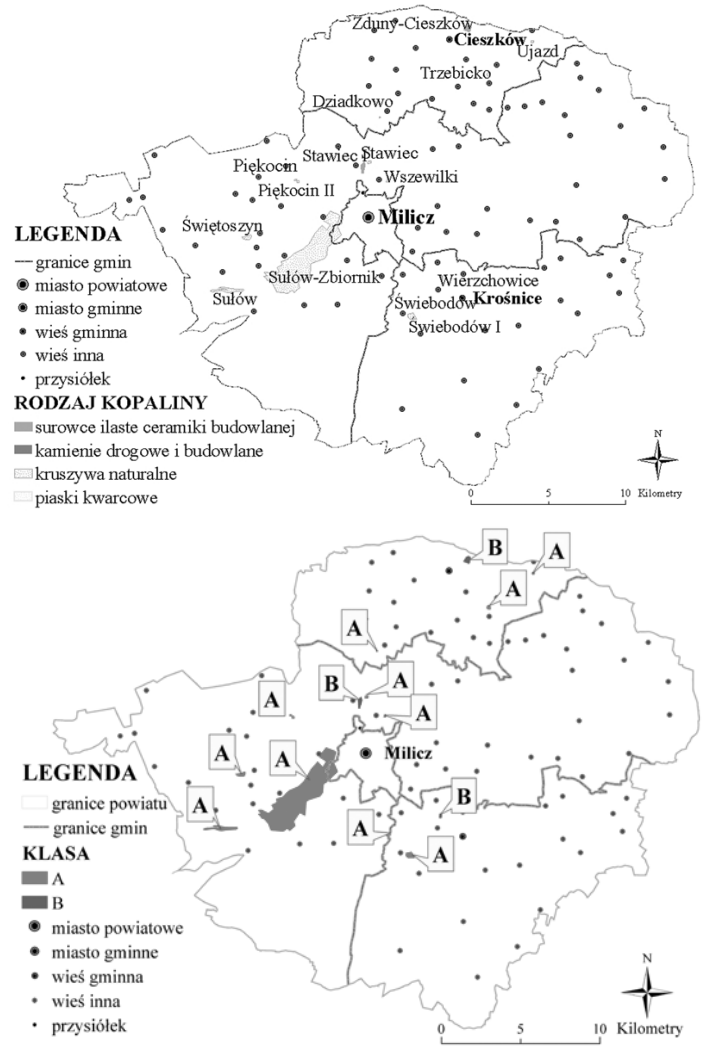
Na rysunku 3 przedstawiono otrzymane podczas analizy wyniki, czyli topologiczne mapy Kohonena oraz wynik odwzorowania Sammona. Na rysunku 3a i 3b przedstawiono mapy Kohonena, które są dwuwymiarowym obrazem 98 wymiarowej przestrzeni składającej się z heksagonalnych obszarów, w których środkach znajdują się wektory określające pozycję siatki neuronów odpowiadające wektorom wag danego neuronu umiejscowionym w przestrzeni 98 wymiarowej.



Rys. 3. Wizualizacja wyników przedstawionych w postaci mapy U-matrix (a, b) oraz w postaci przestrzeni trójwymiarowej (c)
Fig. 3. Visualisation of the results presented as the U-matrix map (a) and in a three-dimensional space (b)

Rysunek 3a przedstawia macierz U-matrix reprezentującą neurony oraz odległość między sąsiednimi neuronami na mapie. Widoczne są na niej trzy obszary – klastry – o ciemniejszym zabarwieniu, które zostały utworzone z neuronów reagujących na podobne dane wejściowych. Neurony te położone są blisko siebie i tworzą pewne skupienie danych o podobnych właściwościach. Na mapie średnie odległości między neuronami są obrazowane kolorami: obszary bliskie są oznaczone ciemnym kolorem, a kolory jaśniejsze oznaczają duże odległości i oznaczają granice obszarów – klasterów. W macierzy U-matrix widoczne są wyraźne granice między trzema obszarami na mapie oraz niewielkie różnice wewnątrz tych trzech jednolitych obszarów. Na rysunku 3b w polach odpowiadających neuronom warstwy wyjściowej sieci podano przypadki, dla których dany neuron był zwycięzcą - zwycięstwo neuronu wyrażone za pomocą wielkości i koloru heksagonu. Można tutaj zobaczyć, że niektóre neurony są puste, a inne zdołały przyciągnąć po kilka wektorów. Z otrzymanych map Kohonena odczytano w jaki sposób zostały pogrupowane analizowane wyrobiska górnicze, czyli ile i jakie obiekty tworzą daną klasę. Widoczny jest tutaj wyraźny podział potencjalnych wyrobisk górniczych na trzy klasy, które umownie nazwano klasą: A, B i C. Klasę A tworzą wyrobiska surowców takich jak kruszywa naturalne i piaski kwarcowe, klasę B tworzą surowce ilaste ceramiki budowlanej, surowce kaolinowe oraz gliny ogniotrwałe, a klasę C kamienie drogowe i budowlane, surowce skaleniowe oraz kwarcyty. Wokół klasy C widoczne są również przypadki odstające od tej klasy, ale jednocześnie znajdujących się w niezbyt dużej odległości – skupisko w porównaniu do poprzednich jest nieco rozproszone. Przypadki te to niektóre złoża

kamieni drogowych i budowlanych, kwarcytów i kwarców żyłowych. Zostały one nazwane klasą N.



Rys. 4. Wizualizacja przestrzenna wyrobisk górniczych zlokalizowanych w powiecie milickim wyrobiska górnicze (a), wyrobiska górnicze sklasyfikowane przez sieci neuronowe (b)
Fig. 4. Spatial visualization of the mining voids in the Milicz County mining voids (a), mining voids classified by neuron networks (b)

Na rysunku 3c pokazano również wynik zastosowania odwzorowania Sammona, na którym możemy zauważyć wyraźne trzy skupiska danych Największe skupisko jest odpowiednikiem najliczniejszej klasy A, mniejsze skupisko odpowiada klasie B,

a trzecie skupisko można łatwo odseparować od klasy B natomiast, jeśli chodzi o klasę A można zauważyć tu raczej płynne przejście pomiędzy klasami A i C.

Na rysunku 4 przedstawiono wyrobiska górnicze zlokalizowane w analizowanym powiecie milickim oraz wynik rozpoznania i klasyfikacji tych wyrobisk otrzymany podczas przeprowadzanych badań.

4. WNIOSKI

Zastosowanie technik Data Mining w analizach jakości danych gromadzonych w systemach GIS daje nowe możliwości pełniejszej oceny dokładności i klasyfikacji tych danych oraz generowania nowej jakości informacji na podstawie wykrywania niejawnych zależności między nimi. Techniki te stanowić mogą, wraz z oprogramowaniem funkcjonującym na rynku IT/GIS, nowy kierunek budowy bardziej funkcjonalnych systemów geoinformacyjnych, dostarczających pełniejszą i bardziej wiarygodną wiedzę o otaczającej rzeczywistości, niezbędną w procesach decyzyjnych.

Metody pozyskiwania wiedzy z danych w postaci procesu Data Mining można podzielić na trzy etapy. W artykule szczególny nacisk położono na proces przygotowania danych, który zdaniem wielu autorów jest kluczowy, zwłaszcza w przypadku baz danych w GIS ze względu na różnorodność typów danych.

W ramach przeprowadzonych badań, wykorzystując samoorganizujące się sieci Kohonena, odkryto grupy obiektów - wyrobisk - o podobnych właściwościach, przeanalizowano te właściwości pod kątem zastosowania wyrobisk w gospodarce odpadami. Prace dotyczące odkrywania wiedzy w systemach GIS przez autorów są kontynuowane. W artykule przeanalizowano literaturę przedmiotu z wiodących ośrodków na świecie i zdefiniowano kierunki dalszych badań.

LITERATURA

ANDERS K.-H., *Data Mining for automated GIS data collection*, Photogrammetric Week, 2001, 263-272

BARTKOWIAK A., *Sieci neuronowe*, wykłady umieszczona na stronie internetowej Instytutu Informatyki Uniwersytetu Wrocławskiego, www.ii.uni.wroc.pl/~aba/, 2002, 1-129

BARTKOWIAK A., CEBRAT S., MACKIEWICZ P., *Probabilistic PCA and neutral networks in search of representative features for some yeast genome data*, AI-METH, Gliwice, 2004

ESTER M., FROMMELT A., KRIEGEL H.-P., SANDER J., *Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support*, specjalne wydanie *Integration of Data Mining with Database Technology, Data Mining and Knowledge Discovery*, Kluwer Academic Publishers, Nr 4, 2000, 193-216

GAŹDZICKI J., *Leksykon geomatyczny*, Polskie Towarzystwo Informacji Przestrzennej, Wydawnictwo „Wieś Jutra", Warszawa, 2001, 1-136

GÓRNIAK-ZIMROZ J., *Koncepcja budowy systemu wspomagania decyzji podejmowanych podczas zarządzania terenami pogórniczymi w kontekście gospodarki odpadami komunalnymi*, praca dyplomowa wykonana w ramach podyplomowego studium Systemy Informacji Geograficznej na Politechnice Wrocławskiej, praca niepublikowana, Wrocław, 2004a

GÓRNIAK-ZIMROZ J., *Zintegrowana gospodarka odpadami komunalnymi i wyrobiskami pogórniczymi*, rozprawa doktorska, praca niepublikowana, Politechnika Wrocławska, Wrocław, 2004b, 1-189

HAN J., KAMBER M., *Data Mining: Concepts and Techniques*, Morgan Kaufman Publishers, 2000, 1-550
KOPERSKI K., HAN J., ADHIKARY J., *Mining Knowledge in Geographical Data*, Communications of ACM, 1998
KORBICZ J., KOSCIELNY J., KOWALCZUK Z., CHOLEWA W., *Diagnostyka procesów. Modele. Metody sztucznej inteligencji Zastosowania*, Wydawnictwa Naukowo-Techniczne, Warszawa, 2002, 1-828
KORBICZ J., OBUCHOWICZ A., UCIŃSKI D., *Sztuczne sieci neuronowe – podstawy i zastosowanie*, Akademicka Oficyna Wydawnicza PLJ, Warszawa, 1994, 1-251
KRAAK M.-J., ORMELING F., *Kartografia – wizualizacja danych przestrzennych*, tłumaczenie Żyszkowska W., Wydawnictwo Naukowe PWN, Warszawa, 1998, 1-274
OSOWSKI S., *Sieci neuronowe*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa, 1994, 1-241
OSOWSKI S., *Sieci neuronowe w ujęciu algorytmicznym*, Wydawnictwa Naukowo-Techniczne, Warszawa, 1996, 1-349
Podstawy GIS, materiały umieszczone na stronie internetowej Zakładu Systemów Informacji Przestrzennej i Geodezji Leśnej w Katedrze Urządzania Lasu, Geomatyki i Ekonomiki Leśnictwa Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, www.wl.sggw.waw.pl, 2003, 1-39
PYLE D., *Data Preparation for Data Mining*, Morgan Kaufman Publishers, 1999
SHEKHAR S., ZHANG P., HUANG Y., VATSAVAI R. R., *Trends in Spatial Data Mining*, w: *Data Mining: Next Generation Challenges and Future Directions*, praca pod redakcją Hillol Kargupta i Anupam Joshi, AAAI/MIT Press, 2003
URBAŃSKI J., *Zrozumieć GIS – Analiza informacji przestrzennej*, Wydawnictwo Naukowe PWN, Warszawa, 1997, 1-144
WOŹNIAK J., FERENC J., *Budowa systemów geoinformacyjnych w zakładach górniczych,* Prace Naukowe Instytutu Górnictwa Politechniki Wrocławskiej Nr 106, 2004, 225-232
ZAPART P., *GIS - komputerowe systemy informacji przestrzennej*, Intersoftland, Warszawa, 1994, 1-94
ŻURADA J., BARSKI M., JĘDRUCH W., *Sztuczne sieci neuronowe: podstawy teorii i zastosowania*, Wydawnictwo Naukowe PWN, Warszawa, 1996, 1-374

Data Mining, GIS, data analysis techniques

APPLICATION OF DATA MINING TECHNIQUES FOR NATURAL RESOURCES MANAGEMENT SUPPORTING IN GIS

Data Mining approach can provide new set of data analysis techniques based on statistical analysis, artificial neural network or fuzzy logic that can be used both for spatial and non-spatial data in GIS. The main purpose is to get unknown knowledge that is hidden in data. Discovered knowledge can be used in GIS for decision making process supporting. Some examples of work related to data mining that were done in Institute of Mining Engineering have been presented in this paper. The emphasis has been put on data preparation and unsupervised classification of data set used SOM. Research tasks based on literature analysis has been defined