View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Portal Czasopism Naukowych (E-Journals)

AUTOMATYKA
CZASOPISMO TECHNICZNE
TECHNICAL TRANSACTIONS
AUTOMATIC CONTROL
WYDAWNICTWO POLITECHNIKI KRAKOWSKIEJ
1-AC/2012
ZESZYT 25
ROK 109
ISSUE 25
YEAR 109

SZYMON ŁUKASIK, PIOTR KULCZYCKI*

ZASTOSOWANIE MIAR ZACHOWANIA STRUKTURY TOPOLOGICZNEJ ZBIORU W WIELOWYMIAROWEJ ANALIZIE DANYCH W PRZESTRZENI ZREDUKOWANEJ

USING TOPOLOGY PRESERVATION MEASURES FOR HIGH-DIMENSIONAL DATA ANALYSIS IN A REDUCED FEATURE SPACE

Streszczenie

Przedmiotem niniejszego artykułu jest wielowymiarowa analiza danych, która realizowana jest poprzez uzupełnienie standardowych procedur ekstrakcji cech odpowiednimi miarami zachowania struktury topologicznej zbioru. Podejście to motywuje obserwacja, że nie wszystkie elementy zbioru pierwotnego w toku redukcji są właściwie zachowane w ramach reprezentacji w przestrzeni o zmniejszonej wymiarowości. W artykule przedstawiono najpierw istniejące miary zachowania topologii zbioru, a następnie omówiono możliwości ich włączenia w klasyczne procedury eksploracyjnej analizy danych. Załączono również ilustracyjne przykłady użycia omawianego podejścia w zadaniach analizy skupień i klasyfikacji.

Słowa kluczowe: zbiory wielowymiarowe, redukcja wymiaru, zachowanie topologii, analiza skupień, klasyfikacja

Abstract

This paper deals with high-dimensional data analysis accomplished through supplementing standard feature extraction procedures with topology preservation measures. This approach is based on an observation that not all elements of an initial dataset are equally preserved in its low-dimensional embedding space representation. The contribution first overviews existing topology preservation measures, then their inclusion in the classical methods of exploratory data analysis is discussed. Finally, some illustrative examples of presented approach in the tasks of cluster analysis and classification are given.

Keywords: multidimensional datasets, dimensionality reduction, topology preservation, cluster analysis, classification

* Dr inż. Szymon Łukasik, prof. dr hab. inż. Piotr Kulczycki, Katedra Automatyki i Technik Informacyjnych, Wydział Inżynierii Elektrycznej i Komputerowej, Politechnika Krakowska oraz Instytut Badań Systemowych, Polska Akademia Nauk.

Publikacja objęta jest prawem autorskim. Wszelkie prawa zastrzeżone. Kopiowanie i rozpowszechnianie zabronione.
Publikacja przeznaczona jedynie dla klientów indywidualnych. Zakaz rozpowszechniania i udostępniania w serwisach bibliotecznych.
http://www.ejournals.eu/Czasopismo-Techniczne/

1. Wstęp

Przedmiotem współczesnej analizy danych są przeważnie zbiory o dużej wymiarowości i znacznym rozmiarze próby. Jest to wynikiem dynamicznego wzrostu ilości informacji przechowywanych w hurtowniach danych oraz opracowania narzędzi pozwalających na wykorzystanie takich właśnie rozproszonych źródeł informacji [4]. Ekstrakcja wiedzy i wizualizacja danych w przypadku zbiorów wielowymiarowych stanowią wyzwanie, głównie ze względu na trudności metodologiczne występujące w przypadku danych o znacznej wymiarowości. Wynikają przede wszystkim z wielu zjawisk występujących w tego typu zbiorach, w literaturze znanych pod pojęciem „przekleństwa wielowymiarowości" [16]. Aby ograniczyć trudności z nich wynikające, opracowano liczne procedury redukcji wymiarowości zbioru. Niech zatem **X** oznacza macierz danych o wymiarze $n \times m$:

$$\mathbf{X} = [x_1 \quad x_2 \quad \ldots \quad x_m] \tag{1}$$

której kolumny reprezentują n-wymiarowe elementy próby zmiennej losowej o wartościach rzeczywistych. Każdy wymiar tej zmiennej będzie określany w niniejszym artykule mianem cechy – zgodnie z terminologią uczenia maszynowego. Celem redukcji wymiaru jest transformacja zbioru do nowej, N-wymiarowej reprezentacji, gdzie N jest znacznie mniejsze od n. Efekt ten można osiągnąć bądź to przez wybór N najistotniejszych cech (ang. *feature selection*), bądź przez ekstrakcję – konstrukcję zredukowanego, bazującego na pierwotnym – zestawu cech (ang. *feature extraction*). Drugą klasę metod można uznać za bardziej ogólną i będzie ona przedmiotem rozważań niniejszego artykułu. Spośród metod ekstrakcji cech wyróżnia się metody liniowe, dla których macierzowa postać zbioru wynikowego **Y** otrzymywana jest z użyciem liniowej transformacji:

$$\mathbf{Y} = \mathbf{AX} \tag{2}$$

gdzie **A** stanowi macierz transformacji ($N \times n$) oraz metody nieliniowe, dla których transformacja ta może być opisana nieliniową funkcją $g : R^n \rightarrow R^N$ (do tej grupy przypisuje się również metody, dla których taka zależność funkcyjna nie istnieje). Szczegółowe omówienie metod należących do obu klas, wraz z ich eksperymentalnym porównaniem można znaleźć w pracy [10].

Charakterystyczną własnością wszystkich metod redukcji wymiaru jest naturalna kompresja informacji spowodowana zmniejszeniem liczby dostępnych cech. Stopień stratności tej kompresji może być zmierzony z użyciem odpowiednich miar zachowania struktury topologicznej zbioru określających ilościowo jej deformację. Niektóre z tych miar mogą być rozpatrywane w odniesieniu do każdego elementu rozważanego zbioru, co pozwala na określenie, w jakim stopniu dany element został zachowany – w sensie swego względnego położenia – w toku przeprowadzanej redukcji. Koncepcja ta jest przedmiotem rozważań niniejszego artykułu. Ponadto, proponuje się tu także użycie wspomnianych miar – określanych dalej wagami elementów analizowanego zbioru – w celu poprawy skuteczności procedur analizy danych przeprowadzanych w przestrzeni zredukowanej. Podejście to zostało po raz pierwszy zaproponowane w pracy [9] w kontekście nowatorskiej procedury redukcji wymiaru opartej na metaheurystyce symulowanego wyżarzania.

Publikacja objęta jest prawem autorskim. Wszelkie prawa zastrzeżone. Kopiowanie i rozpowszechnianie zabronione.
Publikacja przeznaczona jedynie dla klientów indywidualnych. Zakaz rozpowszechniania i udostępniania w serwisach bibliotecznych.
http://www.ejournals.eu/Czasopismo-Techniczne/

Dalsza część artykułu podzielona została na cztery części. Przedmiotem pierwszej z nich jest omówienie obecnych w istniejącej literaturze miar zachowania struktury topologicznej zbioru. Użycie części z wymienionych indeksów, w odniesieniu do poszczególnych elementów zbioru, do analizy danych w przestrzeni zredukowanej rozważono w Sekcji 3. Następnie – w Sekcji 4 – przedstawiono wyniki przeprowadzonych badań eksperymentalnych. W ostatniej część artykułu zawarto uwagi podsumowujące oraz propozycje dalszych prac w ramach rozważanej tematyki badawczej.

2. Miary zachowania struktury topologicznej zbioru

Niech **Y** zdefiniowane w sposób analogiczny do (1), czyli:

$$\mathbf{Y} = [y_1 \quad y_2 \quad ... \quad y_m] \tag{3}$$

oznacza macierzową reprezentację rozważanego zbioru w przestrzeni zredukowanej, o wymiarze $N \times m$. W dalszych rozważaniach niech dodatkowo d_{ij} oraz δ_{ij} oznaczają, dla $i,j \in \{1, 2, ..., m\}$, odległości euklidesowe między elementami analizowanego zbioru w przestrzeni pierwotnej i zredukowanej, określone następującymi wzorami:

$$d_{ij} = \left\| x_i - x_j \right\|_{R^n} \tag{4}$$

$$\delta_{ij} = \left\| y_i - y_j \right\|_{R^N} \tag{5}$$

Metody redukcji wymiaru często przyporządkowuje się – nie zawsze w sposób jednoznaczny – do jednej z dwóch ogólnych klas, związanych z ogólnym celem realizowanej procedury: technik lokalnych oraz technik globalnych [14]. Pierwsze z nich charakteryzuje dążenie w toku algorytmu redukcji wymiaru do zachowania relacji lokalnego sąsiedztwa elementów zbioru pierwotnego **X**. W przypadku drugich, nadrzędnym celem redukcji jest uzyskanie możliwie najlepszej zgodności odległości między wszystkimi elementami rozważanego zbioru w przestrzeni pierwotnej i zredukowanej.

Z wyróżnionych powyżej ogólnych kryteriów oceny metod redukcji wymiaru wynika konieczność zdefiniowania odpowiednich miar zachowania struktury topologicznej zbioru. Jedną z ważniejszych miar biorących pod uwagę globalny kontekst redukcji jest tzw. surowy stres (ang. *raw stress*), powszechnie używany w ramach wielu wariantów skalowania wielowymiarowego [1], który dany jest ściślej następującą zależnością:

$$S_R = \sum_{i=1}^{m} \sum_{j=i+1}^{m-1} (d_{ij} - \delta_{ij})^2 \tag{6}$$

Często stosowany jest również zaproponowany przez Sammona [12] wskaźnik stresu, w ramach którego mniejszy nacisk kładzie się na duże odległości, zdefiniowany według wzoru:

Publikacja objęta jest prawem autorskim. Wszelkie prawa zastrzeżone. Kopiowanie i rozpowszechnianie zabronione.
Publikacja przeznaczona jedynie dla klientów indywidualnych. Zakaz rozpowszechniania i udostępniania w serwisach bibliotecznych.
http://www.ejournals.eu/Czasopismo-Techniczne/

$$S_S = \frac{1}{\sum_{i=1}^{m}\sum_{j=i+1}^{m-1} d_{ij}} \sum_{i=1}^{m}\sum_{j=i+1}^{m-1} \frac{(d_{ij} - \delta_{ij})^2}{d_{ij}} \tag{7}$$

Przedmiotem badań eksperymentalnych zawartych w niniejszym artykule będzie użycie przedstawionych powyżej wskaźników stresu, jednak zaproponowane tu podejście można zastosować również w przypadku innych miar zachowania struktury topologicznej zbioru. W globalnym ujęciu możliwe jest tu użycie m.in. współczynnika korelacji rang Spearmana (inaczej: rho Spearmana). Pozwala on na ilościowe określenie zachowania porządku odległości w przestrzeni zredukowanej, w odniesieniu do porządku tych samych odległości wyznaczonych w przestrzeni pierwotnej. Rho Spearmana stanowi estymator współczynnika korelacji rang [13]. W kontekście redukcji wymiaru wskazuje on zatem, w jakim stopniu przeprowadzana transformacja zachowuje porządek odległości wzajemnych między poszczególnymi elementami analizowanej próby. Współczynnik ten oblicza się z użyciem następującego wzoru:

$$\rho_{SP} = 1 - \frac{6\sum_{p=1}^{M}(r_{pd} - r_{p\delta})^2}{M^3 - M} \tag{8}$$

gdzie $M = m(m-1)/2$ oznacza łączną liczbę odległości podlegających porównaniu, natomiast r_{p_d} i r_{p_δ} stanowią rangi uporządkowanych rosnąco odległości (gdy $i = 1, 2, ..., M$) w przestrzeni pierwotnej oraz zredukowanej. Wartość współczynnika ρ_{SP} równa 1 odpowiada perfekcyjnemu zachowaniu porządku odległości, w ogólnym zaś przypadku $\rho_{SP} \in [-1, 1]$.

Ocenę realizacji redukcji wymiaru o charakterze lokalnym przeprowadza się zwykle poprzez weryfikację zgodności grafów lokalnego sąsiedztwa. Istnieje wiele miar wykorzystujących tego typu podejście – przykładem może być tu miara Koniga [7]. W ramach niniejszego artykułu proponowane jest użycie prostej, wymagającej podania tylko jednego parametru, miary średniego względnego błędu rang MRRE (ang. *Mean Relative Rank Error*) [8]. Niech zatem $N_k(x_i)$ oznacza zbiór k – najbliższych sąsiadów elementu x_i, a R_{ij_d} i R_{ij_δ} stanowią rangi odległości d_{ij} oraz δ_{ij} określone dla elementu x_i względem reszty analizowanego zbioru. Współczynnik MRRE jest wtedy zdefiniowany w sposób następujący:

$$\mathrm{MRRE} = \frac{1}{C}\sum_{i=1}^{m}\sum_{x_j \in N_k(x_i)}^{m} \frac{\left|R_{ij_d} - R_{ij_\delta}\right|}{R_{ij_d}} \tag{9}$$

przy czym występująca w powyższej zależności stała normalizująca C, zapewniająca by MRRE $\in [0, 1]$, jest określana według wzoru:

$$C = m\sum_{p=1}^{k}\frac{\left|2p - m - 1\right|}{p} \tag{10}$$

Publikacja objęta jest prawem autorskim. Wszelkie prawa zastrzeżone. Kopiowanie i rozpowszechnianie zabronione.
Publikacja przeznaczona jedynie dla klientów indywidualnych. Zakaz rozpowszechniania i udostępniania w serwisach bibliotecznych.
http://www.ejournals.eu/Czasopismo-Techniczne/

Tak zdefiniowana miara jest podobna do współczynnika ciągłości i równa się zeru, gdy w zbiorach najbliższych sąsiadów wyznaczonych dla każdego z elementów próby występuje taka sama kolejność w przestrzeni pierwotnej i zredukowanej [8].

Bardziej szczegółowe omówienie i porównanie wymienionych wyżej miar zachowania struktury topologicznej zbioru można znaleźć w pracy [5]. Następna część artykułu poświęcona będzie ich zastosowaniu w analizie danych realizowanej w zredukowanej przestrzeni cech.

3. Opis proponowanej procedury

Ubocznym efektem redukcji wymiaru może być znaczna deformacja położenia niektórych elementów analizowanego zbioru, co zasygnalizowano wstępnie w pierwszej części niniejszego opracowania. Wpływ tej deformacji na skuteczność realizacji dalszych procedur analizy danych może niwelować niezaprzeczalny zysk wynikający z uzyskania zredukowanej reprezentacji rozważanych danych. Celowe wydaje się zatem ilościowe określenie stopnia tej deformacji dla każdego elementu analizowanego zbioru. Wskaźnik taki, nazywany wagą i oznaczany w_i, może być następnie użyty dla celów poprawienia skuteczności procedur analizy danych realizowanych w przestrzeni zredukowanej.

Aby wyznaczyć wartości wag dla poszczególnych elementów należy na wstępie obliczyć odpowiadający im wkład w ostateczną wartość indeksu deformacji struktury topologicznej. Wkład ten oznaczony będzie jako w_i^*, a metoda jego obliczenia wynika bezpośrednio ze wzorów (6)–(9). I tak odpowiednio, gdy wagi mają być wyznaczone na podstawie surowego stresu, wkład ów wyznacza się w następujący sposób:

$$w_i^* = S_{R_i} = \sum_{j=1}^{m} (d_{ij} - \delta_{ij})^2 \tag{11}$$

Natomiast w przypadku rozważania stresu Sammona należy skorzystać ze wzoru:

$$w_i^* = S_{S_i} = \frac{1}{\sum_{i=1}^{m} \sum_{j=i+1}^{m-1} d_{ij}} \sum_{j=1}^{m} \frac{(d_{ij} - \delta_{ij})^2}{d_{ij}} \tag{12}$$

a dla współczynnika Rho Spearmana:

$$w_i^* = 1 - \rho_{SP_i} = \frac{6 \sum_{p=1}^{m-1} (r_{Pd} - r_{P\delta})^2}{M^3 - M} \tag{13}$$

W końcu, gdy wagi mają być otrzymywane na podstawie średniego względnego błędu rang, to:

Publikacja objęta jest prawem autorskim. Wszelkie prawa zastrzeżone. Kopiowanie i rozpowszechnianie zabronione.
Publikacja przeznaczona jedynie dla klientów indywidualnych. Zakaz rozpowszechniania i udostępniania w serwisach bibliotecznych.
http://www.ejournals.eu/Czasopismo-Techniczne/

$$w_i^* = \mathrm{MRRE}_i = \frac{1}{C} \sum_{x_j \in \mathrm{N}_k(x_i)}^{m} \frac{\left| R_{ij_d} - R_{ij_\delta} \right|}{R_{ij_d}} \tag{14}$$

W rzadko spotykanej w praktycznych zagadnieniach analizy danych sytuacji, gdy dla dowolnego elementu $w_i^* = 0$, należy dokonać dodatkowej modyfikacji wartości w_i^*, dodając do niej stałą $w_{\min}{}^* = \min_i w_i^*$, z zachowaniem dodatkowego założenia $w^*_{\min} \neq 0$.

W każdym z przedstawionych wyżej przypadków nie jest wymagane, by suma wkładów dla poszczególnych elementów zbioru stanowiła ostateczną wartość rozpatrywanego indeksu. Wagi otrzymywane są bowiem na podstawie wzorów (11)–(14) przez przeprowadzenie dodatkowej normalizacji:

$$w_i = \frac{m\left(w_i^*\right)^{-1}}{\sum_{i=1}^{m}\left(w_i^*\right)^{-1}} \tag{15}$$

która, dla $i = 1, 2, ..., m$ zapewnia, że:

$$\sum_{i=1}^{m} w_i = m \tag{16}$$

Wagi zdefiniowane według wzoru (15) pozwalają na uwzględnienie deformacji struktury topologicznej zbioru, która występuje w toku redukcji wymiaru. Elementy o dużej wadze powinny być traktowane jako bardziej adekwatne w ramach dalszej analizy danych przeprowadzanej w przestrzeni zredukowanej. Co więcej, z użyciem zaproponowanego tu schematu można istotnie zmniejszyć wpływ znacząco zdeformowanych elementów zredukowanego zbioru poprzez ustalenie wartości $w_i = 0$ dla wszystkich elementów, dla których zachodzi $w_i < W$, gdzie $W \in R^+$ jest wartością progową, nazywaną również współczynnikiem kompresji. Pozostałe wagi należy wtedy dodatkowo znormalizować lub ustalić $w_i = 1$. W ramach przedstawionych tu badań przyjęto drugi z wariantów proponowanego algorytmu.

Kolejne dwie części niniejszego rozdziału poświęcone będą modyfikacjom dwóch standardowych procedur eksploracyjnej analizy danych – klasteryzacji z użyciem algorytmu *K*-średnich i klasyfikacji z zastosowaniem techniki *k*-najbliższych sąsiadów – które uwzględniają użycie omawianego schematu wag.

3.1. Zastosowanie w analizie skupień z użyciem algorytmu *k*-średnich

Zadanie analizy skupień (klasteryzacji) polega na podziale rozważanego zbioru danych na podgrupy zawierające elementy do siebie podobne, ale istotnie różniące się między poszczególnymi podgrupami. *K*-średnich jest iteracyjnym algorytmem klasteryzacji, który realizuje minimalizację błędu kwadratowego, w kontekście techniki tej równoważnego sumie odległości elementów zbioru od najbliższego im środka klastra $C_i = [c_{i1}, c_{i2}, ..., c_{iN}]$, dla $i = 1, 2, ..., K$. Każda iteracja procedury *K*-średnich, w jej standardowym wariancie, składa się z dwóch kroków: przypisania elementów zbioru do odpowiednich skupień i aktualizacji położenia środków klastrów [3]. Włączenie przedstawionego powyżej schematu wag

Publikacja objęta jest prawem autorskim. Wszelkie prawa zastrzeżone. Kopiowanie i rozpowszechnianie zabronione.
Publikacja przeznaczona jedynie dla klientów indywidualnych. Zakaz rozpowszechniania i udostępniania w serwisach bibliotecznych.
http://www.ejournals.eu/Czasopismo-Techniczne/

deformacji struktury topologicznej można osiągnąć w drugim z tych etapów. Każdy środek klastra jest wtedy wyznaczony jako ważony środek ciężkości, według następującego wzoru:

$$c_{ij} = \frac{1}{\sum_{y_l \in C_i} w_l} \sum_{y_l \in C_i} w_l y_{lj} \tag{17}$$

gdzie: $i = 1, 2, ..., K$, $j = 1, 2, ..., N$ i $\sum_{y_l \in C_i} w_l \neq 0$ (w przeciwnym wypadku nie następuje aktualizacja położenia danego środka). Procedurę tego typu określa się ogólnie mianem ważonego algorytmu *K*-średnich [6].

3.2. Zastosowanie w klasyfikacji z użyciem algorytmu *k*-najbliższych sąsiadów

Zadanie klasyfikacji polega na przypisaniu badanego elementu $\tilde{x}$ do jednej z klas z danymi próbami wzorcowymi (zbiór uczący) w postaci podobnej do (3). Metoda *k*-najbliższych sąsiadów (ang. *k-Nearest Neighbor*, w skrócie: *k*-NN) jest nieskomplikowaną techniką opracowaną dla tego zadania. Algorytm, w swym najprostszym i rozpatrywanym tu wariancie, czyli dla $k = 1$, przypisuje element $\tilde{x}$ do klasy, do której należy najbliższy mu element ze zbioru uczącego. Zmodyfikowana procedura, uwzględniająca przedstawiony powyżej schemat wag, podobnego przypisania dokonuje na podstawie ważonych odległości, czyli podzielonych dodatkowo przez wartość w_i. Ten sposób postępowania można uogólnić na przypadek $k > 1$, co prowadzi do syntezy znanego z literatury przedmiotu ważonego klasyfikatora *k*-najbliższych sąsiadów [11].

4. Wyniki badań eksperymentalnych

Skuteczność zaproponowanej tu techniki została wstępnie zweryfikowana w toku procedur eksploracyjnej analizy danych przeprowadzonych dla pięciu wielowymiarowych zbiorów pozyskanych z UCI Machine Learning Repository [15] oraz pracy [2]. Ich charakterystykę przedstawiono w tabeli 1.

W ramach przeprowadzonych eksperymentów redukcja wymiaru była realizowana z użyciem metody składowych głównych PCA (ang. *Principal Components Analysis*). Wartość parametru wymiaru ukrytego *N* została ustalona w toku wcześniejszych eksperymentów. Skuteczność klasteryzacji, z użyciem techniki *K*-średnich, była określana za pomocą indeksu Randa I_C obliczanego względem dostępnych etykiet klas. Dla klasyfikacji według reguły najbliższego sąsiada obliczono natomiast średnią dokładność klasyfikacji I_K w trakcie krzyżowego uwiarygodniania z podziałem na 5 zbiorów (ang. *5-fold cross-validation*). Wszystkie eksperymenty powtórzono 30 razy, odnotowując wartości średnie i odchylenia standardowe uzyskiwanych wyników (które przedstawiono w niniejszym artykule w notacji średnia±odchylenie standardowe). W toku przeprowadzonych eksperymentów użyto schematu wag opartego o surowy stres dany wzorem (6).

Publikacja objęta jest prawem autorskim. Wszelkie prawa zastrzeżone. Kopiowanie i rozpowszechnianie zabronione.
Publikacja przeznaczona jedynie dla klientów indywidualnych. Zakaz rozpowszechniania i udostępniania w serwisach bibliotecznych.
http://www.ejournals.eu/Czasopismo-Techniczne/

Tabela 1

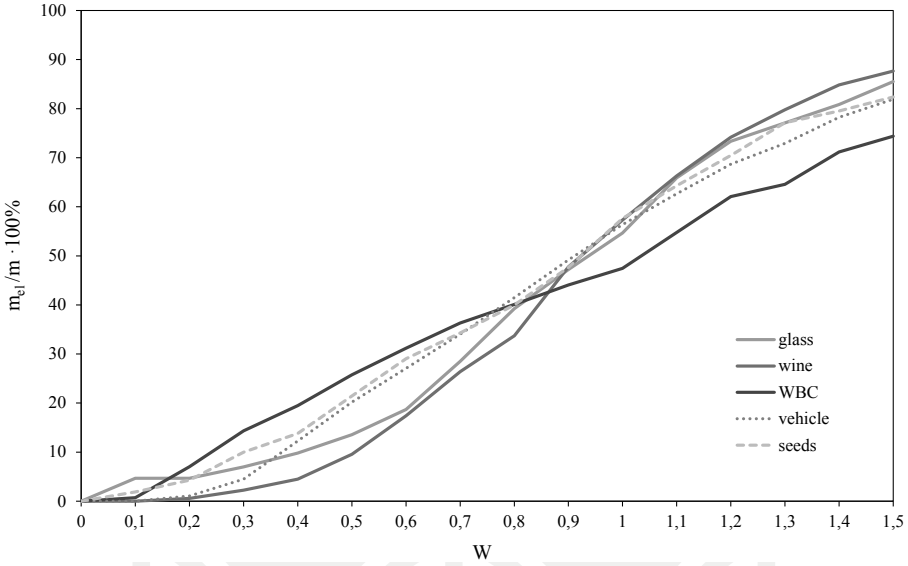
Charakterystyka użytych zbiorów

Zbiór	m	n	N	Klasy	Opis klas	
					Nazwa klasy	Liczność
glass	214	9	4	6	*building_windows_float_processed*	70
					building_windows_non_float_processed	76
					vehicle_windows_float_processed	17
					containers	13
					tableware	9
					headlamps	29
wine	178	13	5	3	*producer_1*	59
					producer_2	72
					producer_3	47
WBC	683	9	4	2	*benign*	444
					malign	239
vehicle	846	18	10	4	*Opel*	212
					Saab	217
					bus	218
					van	199
seeds	210	7	2	3	*Kama*	70
					Rosa	70
					Canadian	70

W pierwszej fazie badań celem przeprowadzonych testów było określenie rozkładu wartości wag obliczonych na podstawie wzorów (11) i (15). Aby to osiągnąć, zmieniając wartość współczynnika kompresji W w zakresie {0,1; 0,2; ...; 1,5} odnotowywano odpowiadającą mu procentową liczbę elementów podlegających redukcji m_{el} (ze względu na spełnienie warunku $w_i < W$). Wyniki tych eksperymentów przedstawiono na rys. 1. Można zauważyć, że rozkład wartości wag nie ma charakteru jednostajnego. Dla wszystkich rozważanych zbiorów około 50% elementów próby odpowiada jednak wartość $w_i < 1$ (czyli mniejsza od średniej).

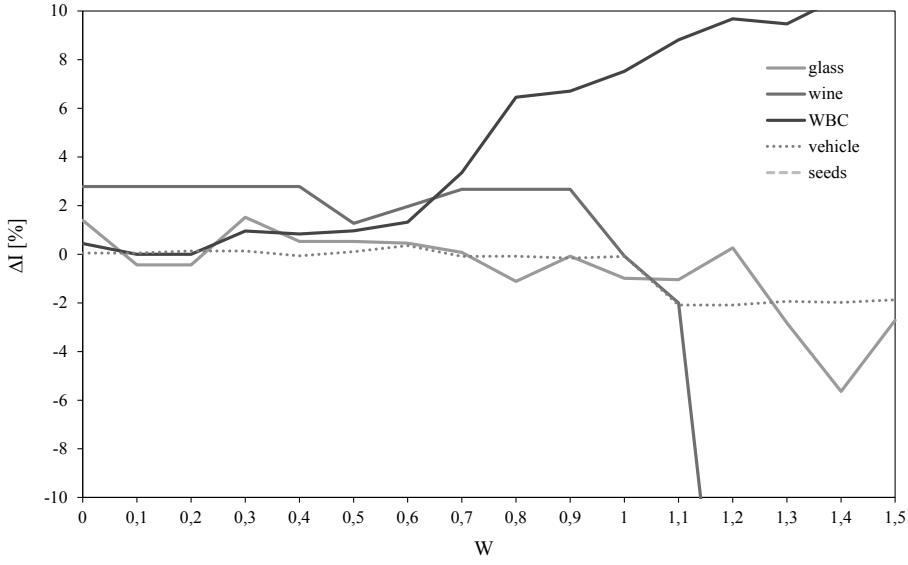
Następna seria eksperymentów miała na celu określenie zmian skuteczności klasteryzacji w przypadku użycia ważonego algorytmu *K*-średnich dla różnych wartości progowych W. Najpierw w celu porównania zbadano efektywność standardowego algorytmu *K*-średnich, odnotowując procentową wartość indeksu Randa I_C uzyskaną na tym etapie badań. Następnie przeprowadzono kolejne testy – tym razem stosując zmniejszenie znaczenia zdeformowanych elementów zredukowanego zbioru, ze zmienną wartością W. Rysunek 2 obrazuje uzyskane różnice $\Delta I_C = I_C - I_{CW}$ między skutecznością standardowego algorytmu, a wariantami z redukcją o różnej intensywności. Uzyskane wyniki można uznać za obiecujące. Dla wszystkich zbiorów samo wprowadzenie schematu wag (z $W = 0$) skutkuje poprawą skuteczności klasteryzacji. W większości przypadków wskazane jest także pominięcie w toku analizy

Publikacja objęta jest prawem autorskim. Wszelkie prawa zastrzeżone. Kopiowanie i rozpowszechnianie zabronione.
Publikacja przeznaczona jedynie dla klientów indywidualnych. Zakaz rozpowszechniania i udostępniania w serwisach bibliotecznych.
http://www.ejournals.eu/Czasopismo-Techniczne/



Rys. 1. Rozkład wartości wag

Fig. 1. Weights' values distribution



Rys. 2. Porównanie skuteczności ważonego algorytmu *K*-średnich i standardowego algorytmu *K*-średnich w przestrzeni zredukowanej

Fig. 2. Performance comparison of weighted *K*-means algorithm and standard *K*-means in the reduced feature space

Publikacja objęta jest prawem autorskim. Wszelkie prawa zastrzeżone. Kopiowanie i rozpowszechnianie zabronione.
Publikacja przeznaczona jedynie dla klientów indywidualnych. Zakaz rozpowszechniania i udostępniania w serwisach bibliotecznych.
http://www.ejournals.eu/Czasopismo-Techniczne/

skupień części zdeformowanych elementów. Zwiększenie wartości *W* powyżej 1 prowadzi jednak do nieprzewidywalnych efektów (np. dla zbioru WBC zaobserwowano spektakularne zwiększenie skuteczności klasteryzacji, co nie miało miejsca dla pozostałych z badanych zbiorów).

Podsumowanie wyników uzyskanych dla analizy skupień oraz klasyfikacji zawarto w tab. 2. Po raz kolejny warto zauważyć, że zaproponowana technika przynosi pozytywne efekty w odniesieniu do procedur analizy danych przeprowadzanych w przestrzeni zredukowanej. Szczególnie wskazane jest użycie eliminacji zdeformowanych elementów zbioru dla klasyfikatora najbliższego sąsiada. Jest to wynikiem słabej odporności tego klasyfikatora na obecność zdeformowanych elementów w zbiorze uczącym, które – dzięki zastosowaniu wprowadzonej tu procedury – mogą zostać pominięte w algorytmie klasyfikacyjnym.

Tabela 2

Skuteczność klasteryzacji i klasyfikacji w przestrzeni zredukowanej – porównanie standardowych algorytmów i zaproponowanej w artykule metodyki

	glass	**wine**	**WBC**	**vehicle**	**seeds**
PCA+*K*-średnich	67,71	92,65	66,16	64,16	88,96
PCA+ważone *K*-średnich (W = 0)	69,11	95,43	66,61	64,21	89,88
PCA+ważone *K*-średnich (najlepsze W)	69,23 (W = 0,3)	95,43 (W = 0,4)	76,74 (W = 1,4)	64,52 (W = 0,6)	91,64 (W = 0,5)
PCA+klasyfikator *k*-NN	61,42 ±8,98	69,43 ±7,73	55,59 ±4,50	54,71 ±3,57	84,84 ±5,98
PCA+ważony klasyfikator *k*-NN (W = 0)	53,73 ±11,98	69,52 ±7,84	44,34 ±5,19	54,55 ±3,75	83,49 ±6,02
PCA+ważony klasyfikator *k*-NN (najlepsze W)	61,82 ±8,76 (W = 0,3)	69,52 ±7,84 (W = 0)	65,71 ±4,62 (W = 0,5)	54,79 ±3,41 (W = 0,4)	86,58 ±5,94 (W = 0,8)

5. Podsumowanie

W niniejszym artykule zaproponowano nowatorski algorytm dedykowany dla zagadnień wielowymiarowej analizy danych. Bazuje on na obserwacji, że redukcja wymiaru powoduje znaczną modyfikację struktury topologicznej zbioru. Jego istotą jest wprowadzenie miar zachowania struktury topologicznej w celu poprawy skuteczności metod eksploracyjnej analizy danych realizowanych w zredukowanej przestrzeni cech. Wstępne eksperymenty obliczeniowe przeprowadzone dla surowego stresu i wybranych zbiorów testowych dowodzą, że zastosowanie zaproponowanego tu podejścia daje obiecujące rezultaty. Dalsze prace w zakresie tematycznym artykułu będą dotyczyć wykorzystania pozostałych z zasugerowanych miar zachowania struktury topologicznej zbioru. Ponadto przedmiotem rozważań będzie wykorzystanie

Publikacja objęta jest prawem autorskim. Wszelkie prawa zastrzeżone. Kopiowanie i rozpowszechnianie zabronione.
Publikacja przeznaczona jedynie dla klientów indywidualnych. Zakaz rozpowszechniania i udostępniania w serwisach bibliotecznych.
http://www.ejournals.eu/Czasopismo-Techniczne/

opisywanej tu procedury w odniesieniu do różnorakich technik redukcji wymiaru, a także przeanalizowanie zagadnienia doboru właściwej wartości współczynnika kompresji *W*.

Badanie zrealizowano dzięki dofinansowaniu w ramach stypendium naukowego z projektu pn. „Technologie informacyjne: badania i ich interdyscyplinarne zastosowania" współfinansowanego ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego, Program Operacyjny Kapitał Ludzki (Umowa nr UDA-POKL.04.01.01-00-051/10-00).

Literatura

[1] Borg I., Groenen P.J.F., *Modern Multidimensional Scaling: Theory and Applications*, Springer, Heidelberg 2010.
[2] Charytanowicz M., Niewczas J., Kulczycki P., Kowalski P.A., Łukasik S., Żak S., *Complete Gradient Clustering Algorithm for Features Analysis of X-Ray Images*, Advances in Intelligent and Soft Computing, vol. 69, 2010, 15-24.
[3] Everitt B.S., Landau S., Leese M., Stahl D., *Cluster Analysis*, Wiley, New York 2011.
[4] Furht B., Escalante A. (red.), *Handbook of Data Intensive Computing*, Springer, Heidelberg 2011.
[5] Karbauskaite R., Dzemyda G., *Topology Preservation Measures in the Visualization of Manifold-Type Multidimensional Data*, Informatica, vol. 20, 2009, 235-254.
[6] Kerdprasop K., Kerdprasop N., Sattayatham P., *Weighted K-Means for Density-Biased Clustering*, Lecture Notes in Computer Science, vol. 3589, 2005, 488-497.
[7] Konig A., *Interactive visualization and analysis of hierarchical neural projections for data mining*, IEEE Transactions on Neural Networks, vol. 11/3, 2000, 615-624.
[8] Lee J.A., Verleysen M., *Nonlinear Dimensionality Reduction*, Springer, New York 2007.
[9] Łukasik S., Kulczycki P., *An Algorithm for Sample and Data Dimensionality Reduction Using Fast Simulated Annealing*, Lecture Notes in Artificial Intelligence, vol. 7120, 2011, 152-161.
[10] Maaten L.J.P.v., Postma E.O., Herik H.J., *Dimensionality Reduction: A Comparative Review*, Tilburg University Technical Report, TiCC-TR 2009-005, 2009.
[11] Parvin H., Alizadeh H., Minati B., *A Modification on K-Nearest Neighbor Classifier*, Global Journal of Computer Science and Technology, vol. 10, 2010, 37-41.
[12] Sammon J.W., *A Nonlinear Mapping for Data Structure Analysis*, IEEE Transactions on Computers, vol. 18, 1969, 401-409.
[13] Sammut C., Webb G.I. (red.), *Encyclopedia of Machine Learning*, Springer, New York 2011.
[14] Silva V.D., Tenenbaum J.B., *Global versus local methods in nonlinear dimensionality reduction*, [w:] Becker S., Thrun S., Obermayer K. (red.), *Advances in Neural Information Processing Systems*, vol. 15, MIT Press, Cambridge 2003, 705-712.
[15] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/
[16] Verleysen M., François D., *The Curse of Dimensionality in Data Mining and Time Series Prediction*, [w:] Cabestany J., Prieto A., Sandoval F. (red.), *Computational Intelligence and Bioinspired Systems. Lecture Notes in Computer Science*, vol. 3512, Springer, Heidelberg 2005, 758-770.

Publikacja objęta jest prawem autorskim. Wszelkie prawa zastrzeżone. Kopiowanie i rozpowszechnianie zabronione.
Publikacja przeznaczona jedynie dla klientów indywidualnych. Zakaz rozpowszechniania i udostępniania w serwisach bibliotecznych.
http://www.ejournals.eu/Czasopismo-Techniczne/