SZYMON ŁUKASIK[1,2] , PIOTR KULCZYCKI[1,2]

MIARY ZACHOWANIA STRUKTURY TOPOLOGICZNEJ ZBIORU I ICH UŻYCIE W PROBLEMACH WIELOWYMIAROWEJ ANALIZY DANYCH

TOPOLOGY PRESERVATION MEASURES AND THEIR APPLICATION IN PROBLEMS OF MULTIDIMENSIONAL DATA ANALYSIS

Streszczenie

Przedmiotem artykułu jest użycie miar zachowania struktury topologicznej zbioru w problemach wielowymiarowej analizy danych. Zaproponowana metodologia jest inspirowana obserwacją, że nie wszystkie elementy zbioru pierwotnego w toku redukcji są właściwie zachowane w ramach reprezentacji w przestrzeni o zmniejszonej wymiarowości. W pracy omówiono na wstępie istniejące miary zachowania topologii zbioru, a następnie przedstawiono możliwości ich włączenia w klasyczne procedury eksploracyjnej analizy danych.

Słowa kluczowe: wielowymiarowa analiza danych, redukcja wymiaru, zachowanie topologii

Abstract

The article focuses on the use of topology preservation measures in multidimensional data analysis. Proposed methodology 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 being discussed.

Keywords: multidimensional datasets, dimensionality reduction, topology preservation

[1] Katedra Automatyki i Technik Informacyjnych, Politechnika Krakowska
[2] Instytut Badań Systemowych, Polska Akademia Nauk
Adres do korespondencji: szymonl@pk.edu.pl

1. Wstęp

Współczesna analiza danych musi podejmować się przetwarzania zbiorów o dużej wymiarowości i znacznym rozmiarze próby. Wynika to z szybkiego wzrostu ilości informacji przechowywanych w hurtowniach danych oraz opracowania narzędzi pozwalających na wykorzystanie takich właśnie rozproszonych źródeł informacji [3]. Ekstrakcja wiedzy i wizualizacja danych w przypadku zbiorów wielowymiarowych stanowią wyzwanie, głównie ze względu na trudności metodologiczne mające miejsce w przypadku danych o znacznej wymiarowości. Wynikają przede wszystkim z szeregu zjawisk występujących w tego typu zbiorach, w literaturze znanych pod pojęciem „przekleństwa wielowymiarowości" [14]. Aby ograniczyć trudności z nich wynikające opracowano liczne procedury redukcji wymiarowości zbioru. Celem redukcji wymiaru jest transformacja zbioru do nowej, N - wymiarowej reprezentacji, gdzie N jest znacznie mniejsze od n (czyli pierwotnej wymiarowości rozważanych danych). Efekt ten można osiągnąć między innymi przez ekstrakcję – konstrukcję zredukowanego, bazującego na pierwotnym, zestawu cech (ang. feature extraction). Szczegółowe omówienie metod redukcji wymiaru należących do tej klasy, 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 niniejszej pracy. Ponadto, proponuje się tu także użycie wspomnianych miar – określanych wagami elementów analizowanego zbioru – dla celów poprawy skuteczności procedur analizy danych przeprowadzanych w przestrzeni zredukowanej. Podejście to zostało po raz pierwszy zaproponowane w pracy [7] w kontekście nowatorskiej procedury redukcji wymiaru opartej o metaheurystykę symulowanego wyżarzania.

2. Miary zachowania struktury topologicznej zbioru

Niech Y oznacza macierzową reprezentację rozważanego zbioru w przestrzeni zredukowanej, o wymiarze $N \times m$:

$$Y = \left[y_1 \; y_2 \; \ldots \; y_m \right], \tag{1}$$

a X podobną reprezentację zbioru pierwotnego (m oznacza liczność próby). Dla celów dalszych rozważań niech dodatkowo d_{ij} oraz δ_{ij} oznaczają, dla $i, j \in \{1, 2, \ldots, m\}$, odległości euklidesowe między elementami analizowanego zbioru w przestrzeni pierwotnej i zredukowanej.

Jedną z ważniejszych miar zachowania struktury topologicznej, biorących pod uwagę globalny kontekst redukcji – tj. możliwie najlepszą zgodność odległości między wszystkimi elementami rozważanego zbioru w przestrzeni pierwotnej i zredukowanej – jest tzw.

surowy stres (ang. raw stress), powszechnie używany w ramach wielu wariantów skalowania wielowymiarowego [1]. Jest on dany ściślej następującą zależnością:

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

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

$$S_S = \frac{1}{\sum_{i=1}^{m} \sum_{j=i+1}^{m-1} d_{ij}} \sum_{i=1}^{m} \sum_{j=i+1}^{m-1} \frac{(d_{ij} - \delta_{ij})^2}{d_{ij}} \ . \tag{3}$$

W globalnym ujęciu możliwe jest tu użycie między innymi 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:

$$\rho_{SP} = 1 - \frac{6 \sum_{p=1}^{M} (r_{p_d} - r_{p_\delta})^2}{M^3 - M}, \tag{4}$$

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 [5]. W ramach niniejszej pracy proponowane jest użycie prostej, wymagającej podania tylko jednego parametru, miary średniego względnego błędu rang MRRE (ang. Mean Relative Rank Error) [6]. Niech zatem $\mathrm{N}_k(x_i)$ 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:

$$MRRE = \frac{1}{C}\sum_{i=1}^{m}\sum_{x_j \in \mathrm{N}_k(x_i)}^{m} \frac{\left|R_{ij_d} - R_{ij_\delta}\right|}{R_{ij_d}}, \tag{5}$$

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

$$C = m\sum_{p=1}^{k}\frac{|2p-m-1|}{p}. \tag{6}$$

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 [6]

Bardziej szczegółowe omówienie i porównanie wymienionych wyżej miar zachowania struktury topologicznej zbioru można znaleźć w pracy [4]. 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 (2-5). W każdym z przedstawionych przypadków nie jest wymagane by suma wkładów dla poszczególnych elementów zbioru stanowiła ostateczną wartość rozpatrywanego indeksu. Wagi w_i otrzymywane są bowiem z przeprowadzeniem dodatkowej normalizacji wartości w_i^*, tak by $\sum_{i=1}^{m} w_i = m$. Pozwalają one 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 zredukować 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 $w_i = 1$.

Wagi w zaproponowanej postaci można użyć między innymi w zadaniach analizy skupień (klasteryzacji) oraz klasyfikacji. W pierwszym przypadku użycie wag w standardowym wariancie popularnego algorytmu procedury K-średnich jest możliwe na etapie aktualizacji położenia środków klastrów [2]. Są one wtedy wyznaczane jako ważone środki ciężkości. W zadaniu klasyfikacji wagi mogą być użyte m.in. w stworzeniu alternatywnego wariantu ważonego klasyfikatora k-najbliższych sąsiadów [11]. Zmodyfikowana procedura, uwzględniająca przedstawiony powyżej schemat wag, dokonuje przypisania elementów do klas na podstawie ważonych odległości, czyli podzielonych dodatkowo przez wartość w_i. Ten sposób postępowania można uogólnić na przypadek $k > 1$.

4. Podsumowanie

W niniejszym artykule metodologię dedykowaną dla zagadnień wielowymiarowej analizy danych. Bazuje ona na obserwacji, że redukcja wymiaru powoduje znaczną modyfikację struktury topologicznej zbioru. Jej istotą jest wprowadzenie miar zachowania struktury topologicznej w celu poprawy skuteczności metod eksploracyjnej analizy danych realizowanych w zredukowanej przestrzeni cech. Przeprowadzone eksperymenty obliczeniowe dowodzą, że zastosowanie zaproponowanego tu podejścia daje obiecujące rezultaty.

Dalsze informacje na temat przedstawionej tu koncepcji można znaleźć w pracach [8,9].

* * *

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] Everitt, B. S., Landau, S., Leese, M., Stahl, D., *Cluster Analysis*, Wiley, New York, 2011.

[3] Furht, B., Escalante, A. (red.), *Handbook of Data Intensive Computing*, Springer, Heidelberg, 2011.

[4] Karbauskaite, R., Dzemyda, G., *Topology Preservation Measures in the Visualization of Manifold-Type Multidimensional Data*, Informatica, vol. 20, 235-254, 2009.
[5] Konig, A., *Interactive visualization and analysis of hierarchical neural projections for data mining*, IEEE Transactions on Neural Networks, vol. 11/3, 615-624, 2000.
[6] Lee, J.A., Verleysen, M., *Nonlinear Dimensionality Reduction*, Springer, New York, 2007.
[7] Łukasik, S., Kulczycki, P., *An Algorithm for Sample and Data Dimensionality Reduction Using Fast Simulated Annealing*, Lecture Notes in Artificial Intelligence, vol. 7120, 152-161, 2011.
[8] Łukasik, S., Kulczycki, P., *Zastosowanie miar zachowania struktury topologicznej zbioru w wielowymiarowej analizie danych w przestrzeni zredukowanej*, Czasopismo Techniczne, seria: Automatyka, vol. 1-AC, ss. 5-16, 2012.
[9] Łukasik, S., Kulczycki, P., *Using Topology Preservation Measures for Multidimensional Intelligent Data Analysis in the Reduced Feature Space*, Lecture Notes in Artificial Intelligence, vol. 7120, 152-161, 2011.
[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, 37-41, 2010.
[12] Sammon, J. W., *A Nonlinear Mapping for Data Structure Analysis*, IEEE Transactions on Computers, vol. 18, 401-409, 1969.
[13] Sammut, C., Webb, G.I. (red.), *Encyclopedia of Machine Learning*, Springer, New York, 2011.
[14] 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, 758-770, Springer, Heidelberg, 2005.