PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 384

Taksonomia 24
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2015

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronie internetowej Wydawnictwa
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 384 • 2015
Taksonomia 24 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Kamila Migdał-Najman
Uniwersytet Gdański
e-mail: kamila.migdal-najman@ug.edu.pl

OCENA WPŁYWU WARTOŚCI STAŁEJ MINKOWSKIEGO NA MOŻLIWOŚĆ IDENTYFIKACJI STRUKTURY GRUPOWEJ DANYCH O WYSOKIM WYMIARZE

Streszczenie: W analizie zróżnicowania jednostek w przestrzeni wielowymiarowej ważny jest wybór odpowiedniej miary odległości. Wybór ten nabiera znaczenia, gdy analizie poddany jest zbiór danych o dużej liczbie jednostek opisanych setkami cech. Najczęściej stosuje się miary odległości oparte na metryce potęgowej. W metryce tej istotny staje się wybór odpowiedniego poziomu stałej Minkowskiego. Celem prezentowanych badań jest ocena wpływu wartości stałej Minkowskiego i wymiaru przestrzeni na możliwą do uzyskania strukturę grupową. W artykule na podstawie przeprowadzonych badań symulacyjnych wykazano, że w przypadku wysokiego wymiaru przestrzeni zastosowanie ułamkowego poziomu wykładnika w normie potęgowej wpływa na możliwość identyfikacji istniejącej struktury grupowej badanych jednostek.

Słowa kluczowe: analiza skupień, przekleństwo wymiarowości, metryka potęgowa.

DOI: 10.15611/pn.2015.384.20

1. Wstęp

Analizowane zjawiska lub obiekty mogą być opisywane przez wiele różnych cech (j=1,...,p). Wiele metod statystycznych zakłada, że liczba cech powinna pozostawać w pewnej rozsądnej relacji do liczby analizowanych jednostek. I jakkolwiek nie ma określonych reguł w tym względzie, to raczej nie powinno dochodzić do sytuacji, aby liczba cech przekraczała liczbę jednostek. Taylor [1977] w kontekście analizy głównych składowych zwraca uwagę, że niektórzy użytkownicy metody niechętnie ją stosują, jeżeli nie ma ona 3-4 razy więcej jednostek niż cech. Hair i in. [1995] podaje regułę, że jednostek powinno być 5 razy więcej niż cech. Dodaje również, że bardziej akceptowalnym warunkiem byłby iloraz 1 do 10, a niektórzy proponują nawet 20 jednostek na każdą cechę. Sugestie w tej mierze są zresztą

różne i nie ma prostego przełożenia, ile cech powinniśmy rozpatrywać. Należy zgodzić się z zasadą, aby nie mnożyć liczby cech i operować oszczędnymi, przemyślanymi ich zbiorami.

Współcześnie jednak spotykamy przykłady zbiorów, które zawierają znaczną liczbę cech. Jednostki opisane są setkami, tysiącami, a nawet w szczególnych przypadkach milionami cech. W sytuacji tej możemy podjąć dwie decyzje. Możemy dokonać redukcji liczby cech lub wręcz przeciwnie pozostawić wszystkie cechy w badaniu. W sytuacji kiedy podejmujemy decyzję o redukcji liczby cech, możemy wyróżnić trzy podejścia, które pozwolą ustalić nam optymalny zbiór cech. Będzie to: 1) ważenie cech, gdzie każdej cesze nadaje się wagę mówiącą o jej relatywnej ważności w opisie badanego problemu; 2) selekcja cech, polegająca na tym, że ze zbioru cech eliminuje się te, których potencjał dyskryminacyjny wydaje się najmniejszy; podejście to może być uznane za szczególny przypadek podejścia pierwszego, gdzie wagi cech przyjmują jedynie wartości 0 dla cech odrzuconych i 1 dla wybranych; 3) zastąpienie cech oryginalnych przez cechy sztuczne, jest to klasyczne statystyczne podejście bazujące na analizie głównych składowych [Walesiak 2005]. Decyzja druga zakłada wręcz przeciwną sytuację. Chcemy pozostawić w badaniu wszystkie dostępne cechy. Decyzja ta może wynikać z faktu, że nie mamy czasu na analizowanie poszczególnych cech (zbiór składa się np. z kilku tysięcy cech). Analiza danych musi być bardzo szybka. Kolejnym ważnym aspektem, który wymaga pozostawienia w badaniu wszystkich cech jest ich dynamiczny charakter. Wynika on z bardzo dużej częstotliwości aktualizacji danych. W sieciach telekomunikacyjnych, w systemach rejestrujących transakcje bankowe lub zakupowe, zbiór danych może być aktualizowany kilkaset razy na sekundę. Cechy w takim zbiorze mogą więc także szybko zmieniać swoje znaczenie.

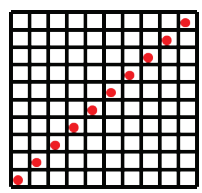
2. Przekleństwo wymiarowości

Decyzja badacza o pozostawieniu w badaniu wszystkich cech może spowodować różnego rodzaju problemy. Trudności wynikające z analizy danych opisanych dużą liczbą cech (wymiarów) określono w literaturze pojęciem: przekleństwo (problem, klątwa) wymiarowości (*curse of dimensionality*). Trudności te wynikają przede wszystkim z gwałtownego (wykładniczego) wzrostu objętości hiperprzestrzeni wraz ze wzrostem jej wymiaru. W 1961 roku Richard Ernest Bellman[1] w opracowaniu *Adaptive control processes* po raz pierwszy użył pojęcia „przekleństwo wymiarowości". Pojęcie to pojawiło się następnie w pracach: White'a [1989], Bishopa [1995] a także w pracach Scotta i Thompsona [1983], Silvermana [1986] pod pojęciem „zjawisko pustej przestrzeni" (*empty space phenomenon*).

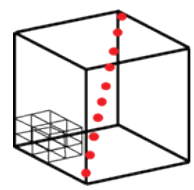
[1] R.E. Bellman (1920-1984), matematyk, znany głównie jako twórca programowania dynamicznego. Jest to technika lub strategia projektowania algorytmów stosowana przeważnie do rozwiązywania zagadnień optymalizacyjnych.

Rozważmy prosty przykład. Załóżmy, że posiadamy 10 jednostek opisanych przez jedną cechę. Niech jednostki te reprezentują 10 (10^1) przedziałów (por. rys. 1a). Powiemy wówczas, że każda część przestrzeni jest reprezentowana przez jedną jednostkę. Opiszmy teraz te same jednostki dwiema cechami. Dla 10 jednostek w przestrzeni dwuwymiarowej uzyskamy 100 (10^2) kwadratów (por. rys. 1b). Teraz 10 jednostek reprezentuje 100 kwadratów. Jednostki te zajmują teraz już tylko 10% przestrzeni dwuwymiarowej. Opiszmy te same jednostki trzema cechami. Jednostki te reprezentują 1000 (10^3) kostek (por. rys. 1c). Jednostki zajmują 1% przestrzeni trójwymiarowej. Objętość przestrzeni p-wymiarowej (j=1,...,p) wraz ze wzrostem p powoduje, że te same jednostki wypełniają coraz mniejszą część przestrzeni. Aby jednostki w kolejnych wymiarach reprezentowały całą przestrzeń, należałoby ich liczbę zwiększyć wykładniczo wraz ze wzrostem wymiaru.

a) 10 jedn. w przestrzeni jednowymiarowej



b) 10 jedn. w przestrzeni dwuwymiarowej



c) 10 jedn. w przestrzeni trójwymiarowej

Rys. 1. 10 jednostek w przestrzeni jedno-, dwu- i trójwymiarowej

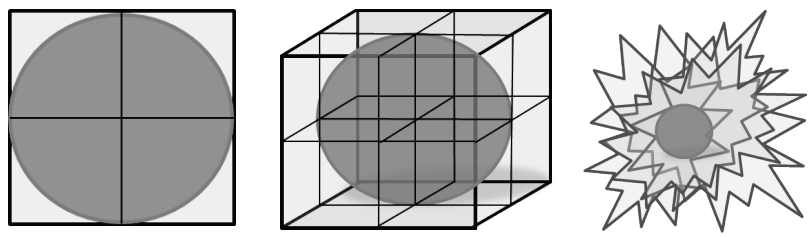
Źródło: opracowanie własne.

Problem przekleństwa wymiarowości jest również wyzwaniem w grupowaniu i klasyfikacji danych. W badaniach społeczno-ekonomicznych skupienia najczęściej mają postać w przybliżeniu hiperkuli, w której gęstość jednostek zwiększa się w kierunku jej centrum. Podobnie jak wyżej, wyobraźmy sobie skupienie jednostek w przestrzeni dwuwymiarowej, które ma postać koła. Jeżeli przestrzeń jest opisana dwiema cechami, to przy założeniu, że nie ma ona ograniczeń, otrzymamy płaszczyznę. Gdyby ją ograniczyć do średnicy koła równej jeden, to powierzchnia tego koła wypełniłaby 78,5% powierzchni kwadratu o boku równym jeden (por. rys. 2a). Udział objętości kuli o średnicy równej jeden w objętości sześcianu o boku równym jeden stanowi 52,4% (por. rys. 2b). Udział objętości hiperkuli o średnicy równej jeden w przestrzeni ośmiowymiarowej w stosunku do objętości hipersześcianu o boku równym jeden wynosi jedynie 1,6% (por. rys. 2c).

Wraz ze wzrostem wymiarowości rośnie odsetek jednostek znajdujących się poza hiperkulą. Jeżeli wymiarowość dąży do nieskończoności, stosunek różnicy odległości euklidesowej między jednostką położoną najdalej i najbliżej środka ciężkości hiperkuli do odległości jednostki położonej najbliżej środka ciężkości hiperkuli dąży do zera:

$$\lim_{p\to\infty} \frac{d\max_p - d\min_p}{d\min_p} \to 0.$$

Dlatego też miary odległości zaczynają tracić swoją skuteczność jako miary oceniające zróżnicowanie jednostek w przestrzeniach wysoce wymiarowych. Stają się mniej dyskryminacyjne wraz ze wzrostem wymiaru. Efekt ten w aspekcie „przekleństwa wymiarowości" w klasyfikacji (i nie tylko) nazywamy: efektem koncentracji L_k normy.



a) $p = 2$ b) $p = 3$ c) $p = 8$

Rys. 2. Udział hiperkuli w hiperprzestrzeni w przestrzeni dwuwymiarowej ($p = 2$), trójwymiarowej ($p = 3$) i ośmiowymiarowej ($p = 8$)

Źródło: opracowanie własne.

3. Efekt koncentracji L_k normy

Efekt ten ma istotne znaczenie w analizie skupień, szczególnie gdy liczba wymiarów jest rzędu setek i więcej. Wiele metod grupowania w swojej konstrukcji zawiera pomiar odległości. Do grupy tej należą między innymi metody hierarchiczne, metody optymalizacyjno-podziałowe czy wybrane topologie sztucznych sieci neuronowych. Najczęściej stosuje się miary odległości oparte na ogólnej metryce potęgowej. Zdefiniujmy metrykę potęgową następująco:

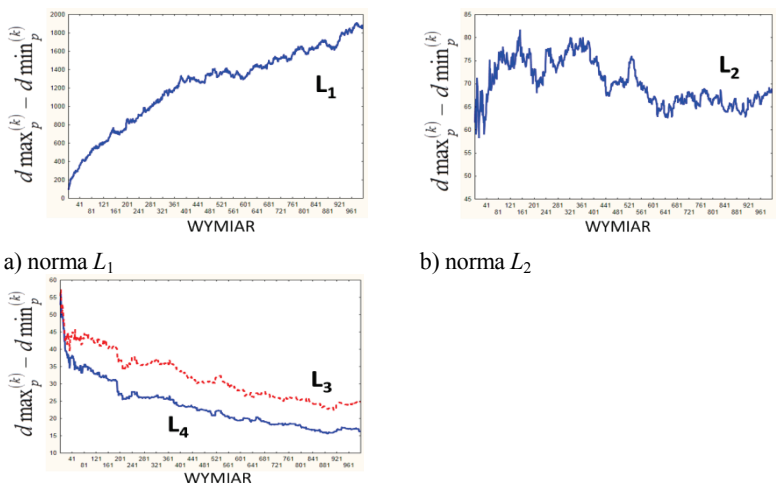
$$L_k = d_{rs}^{(k)} = \sqrt[k]{\sum_{j=1}^{p} \left| x_{rj} - x_{sj} \right|^k},$$

gdzie k jest dowolną liczbą naturalną (zwaną stałą Minkowskiego), specyfikującą wykładnik potęgi. Metryka miejska L_1 definiowana jest przez wykładnik k =1. Metryka euklidesowa L_2 definiowana jest przez wykładnik k =2.

Wybór odpowiedniego poziomu wykładnika k okazuje się szczególnie ważny w przypadku jednostek opisywanych bardzo dużą liczbą cech (*High-Dimensional Data, HDD*). Badania takie prowadzili: Beyer, Goldstein, Ramakrishnan, Shaft [1999], Hinneburg, Aggarwal, Keim [2000; 2001], Verleysen, François [2005], Houle, Krie-

gel, Kröger, Schubert, Zimek [2010], Schnitzer, Flexer [2014]. Ich autorzy sugerują, że norma L_k zastosowana w przypadku danych wysoce wymiarowych przyjmująca poziom $k = 1$ lub $k = 2$ pozwala uzyskać wyższą ocenę jakości struktury grupowej niż norma przyjmująca poziom $k > 3$. Zaobserwowano, że wraz ze wzrostem wymiaru następuje zaburzenie odległości między: jednostką położoną najbliżej np. środka ciężkości skupienia $-d\min_p^{(k)}$ a jednostką położoną najdalej w stosunku do tego samego punktu – $d\max_p^{(k)}$. Wraz ze wzrostem poziomu wykładnika k i liczby wymiarów szybciej pogarsza się kontrast między zdefiniowanymi jednostkami. Istotny wpływ na ten efekt ma zastosowana norma L_k.

Rozważmy następujący przykład. Niech zbiór danych składa się z 500 jednostek. Załóżmy, że jednostki te w każdym wymiarze mają rozkład normalny. Dla kolejnych wymiarów, od 2 do 1000, wyznaczmy wektor średnich[2]. Dla każdego wymiaru poszukajmy jednostki, która jest najbliżej i najdalej w stosunku do tego punktu. Obliczmy dla każdego wymiaru odległości tych punktów do wyznaczonego wektora średnich, stosując odpowiednią normę L_k (k =1,2,3 i 4). Dla każdego wymiaru obliczmy różnicę odległości między jednostką położoną najdalej i najbliżej wektora średnich (por. rys. 3).



a) norma L_1

b) norma L_2

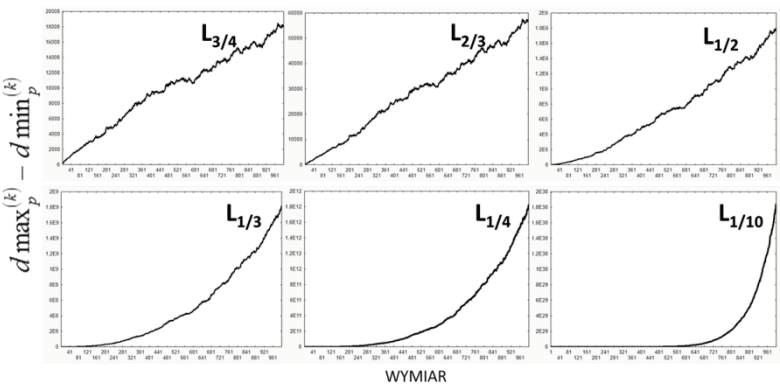
c) norma L_3 i L_4

Rys. 3. Wpływ wymiaru i normy L_k na różnicę odległości między jednostką położoną najdalej i najbliżej wektora średnich

Źródło: opracowanie własne.

[2] Wszystkie prezentowane w artykule symulacje komputerowe przeprowadzone zostały w systemie Matlab na podstawie autorskiego oprogramowania.

Dla kolejnych wymiarów różnica odległości między jednostką położną najbliżej i najdalej w stosunku do wyznaczonego wektora średnich w przypadku zastosowanej normy L_1 rośnie liniowo (por. rys. 3a). Zastosowana norma L_2 wykazuje zmienną tendencję (por. rys. 3b), a norma L_3 i L_4 tendencję malejącą (por. rys. 3c). Oznacza to, że wraz ze wzrostem poziomu wykładnika k i liczby wymiarów pogarsza się kontrast między zdefiniowanymi jednostkami.



Rys. 4. Wpływ wymiaru i normy ułamkowej L_k na różnicę odległości między jednostką położoną najdalej i najbliżej wektora średnich

Źródło: opracowanie własne.

Jeżeli wraz ze wzrostem wymiaru przestrzeni kontrast między jednostkami maleje najwolniej przy niskich wartościach k, to może należałoby rozważyć poziom k mniejszy od 1. Normę, gdzie k jest wartością z przedziału (0;1), nazywamy **normą ułamkową** (*fractional distance metrics*). Na rysunku 4 zaprezentowano różnice odległości między jednostką położoną najbliżej i najdalej wektora średnich dla kolejnych wymiarów. Dla każdego wymiaru obliczono normę L_k (k = 3/4, 2/3, 1/2, 1/3, 1/4 i 1/10)[3]. Najwolniejszy spadek kontrastu następował dla normy o najniższym poziomie k (por. rys. 4). Wraz ze wzrostem wymiaru jednostki w przestrzeni zbliżają się do siebie. Norma ułamkowa o niskim poziomie wykładnika k wraz ze wzrostem wymiaru pozwala uzyskać większe zróżnicowanie, kontrast między jednostkami.

4. Ocena jakości struktury grupowej

Jeżeli wartość stałej Minkowskiego ma tak duże znaczenie w ocenie zróżnicowania jednostek w przestrzeni o wysokim wymiarze, to jej wybór powinien mieć duże znaczenie dla każdej metody grupowania, która korzysta z metryki potęgowej. Przykła-

[3] Wybór wykładnika k jest subiektywny i służy ilustracji omawianych prawidłowości.

dem są najczęściej stosowane metody grupowania, takie jak aglomeracyjne metody hierarchiczne czy metoda *k*-średnich.

Aby zweryfikować to przypuszczenie, wykonano odpowiednie badania symulacyjne. Niech zbiór danych składa się z 4 sferycznych skupień, o gęstości jednostek wzrastającej w kierunku centrum skupienia. W każdym skupieniu znajduje się 1000 jednostek. Skupienia są całkowicie separowalne w niektórych wymiarach, a w innych nie. Jednostki zostaną umieszczone w przestrzeni: 50-, 100-, 150-, 200-, 250-, 300-, 350-, 400-, 450- i 500-wymiarowej. Grupowanie jednostek zostanie przeprowadzone metodą *k*-średnich. W metodzie *k*-średnich wykorzystana zostanie norma potęgowa z zadanym poziom *k* równym: 2, 1, 3/4, 1/2, 1/4, 1/10 i 1/20[4]. Ocena zgodności wyników grupowania z założoną *a priori* przynależnością jednostek do skupień zostanie przeprowadzona na podstawie skorygowanego wskaźnika Randa. Dla uśrednienia uzyskanych wyników grupowanie i jego ocena dla każdego wymiaru i każdego poziomu wykładnika *k* zostanie powtórzona 10-krotnie[5].

Tabela 1. Ocena jakości struktury grupowej dla przyjętego wymiaru i poziomu wykładnika normy potęgowej uzyskana na podstawie skorygowanego wskaźnika Randa

k	WYMIAR									
	50	100	150	200	250	300	350	400	450	500
2	0,815	0,877	0,852	0,861	**0,935**	0,753	**0,908**	0,877	0,869	0,855
1	0,738	0,823	0,758	0,803	0,934	0,803	0,824	0,906	0,869	0,846
3/4	0,889	0,823	**0,963**	0,943	0,783	**0,849**	0,860	0,906	0,860	**0,918**
1/2	0,775	0,918	0,861	0,882	0,918	**0,918**	0,815	**0,929**	**0,929**	0,882
1/4	0,753	**1**	0,876	0,876	**1**	0,809	0,753	0,658	**1**	0,781
1/10	**0,934**	**0,926**	**0,926**	**0,971**	0,897	0,832	**1**	**1**	**1**	**1**
1/20	**1**	**1**	0,832	**1**	**1**	0,714	**1**	**1**	**1**	**1**

Źródło: opracowanie własne.

W tabeli 1 zaprezentowano uzyskaną ocenę zgodności struktury grupowej ze znanym wzorcem. Można zauważyć, że w przypadku normy $L_{1/20}$ i $L_{1/10}$ uzyskano najwyższą zgodność klasyfikacji. W przypadku wysokiego wymiaru przestrzeni grupowanych jednostek zastosowanie ułamkowego poziomu *k* normy potęgowej znacząco wpływa na możliwość identyfikacji struktury grupowej. Im wyższy wymiar przestrzeni i mniejszy (ułamkowy) poziom wykładnika *k,* tym uzyskano wyższą jakość grupowania. Nieznaczne zaburzenia tej tendencji, np. dla wymiaru 150, 250 i

[4] Prezentowane symulacje wymagają kilkudziesięciu godzin pracy bardzo wydajnego komputera. Z tego powodu badany zbiór danych ma niewielką liczbę skupień i prostą strukturę przestrzenną. Z tego samego powodu zaprezentowano wyniki jedynie dla wybranych wartości *k* i liczby wymiarów przestrzeni.

[5] W metodzie *k*-średnich procedura grupowania rozpoczyna się od losowo wybranych, hipotetycznych centrów skupień.

300, wynikają z niedoskonałości samej metody *k*-średnich, która, mimo że skupienia były sferyczne, przy niekorzystnej inicjalizacji nie potrafiła poprawnie zidentyfikować istniejącej struktury grupowej badanych jednostek.

5. Zakończenie

W badaniach empirycznych coraz częściej pojawiają się zbiory danych o wysokim wymiarze, rzędu setek i tysięcy cech. Tak duża liczba cech w istotny sposób zmienia skalę problemów stojących przed analizą skupień. Między innymi zaobserwowano, że wraz ze wzrostem wymiaru następuje zaburzenie różnicy odległości między jednostką położoną najbliżej i najdalej od dowolnej innej jednostki, np. środka ciężkości skupienia. Wraz ze wzrostem poziomu wykładnika *k* w normie potęgowej i wzrostem liczby wymiarów pogarsza się kontrast między obserwowanymi jednostkami w przestrzeni. Wykazano, że w takiej sytuacji należałoby rozważyć zastosowanie ułamkowego poziomu wykładnika *k*, dzięki czemu możliwe stanie się zachowanie znacznie lepszego kontrastu między jednostkami. Zastosowanie ułamkowego poziomu wykładnika w normie potęgowej wpływa na możliwość identyfikacji istniejącej struktury grupowej badanych jednostek. Obserwacja powyższa może mieć istotne znaczenie w grupowaniu jednostek każdą metodą wykorzystującą w swojej konstrukcji pomiar odległości między jednostkami.

Literatura

Bellman R.E., 1961, *Adaptive Control Processes*, A Guided Tour, Princeton University Press, Princeton, New Jersey.

Beyer K., Goldstein J., Ramakrishnan R., Shaft U., 1999, *When Is "Nearest Neighbor" Meaningful*, International Conference on Database Theory, Jerusalem, Israel, s. 217-235.

Bishop C.M., 1995, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.

Hair J.F., Anderson R.E., Tatham R.L., Black W.C., 1995, *Multivariate Data Analysis with Readings*, Prentice Hall International, Ltd., London (4th ed.).

Hinneburg A., Aggarwal C.C., Keim D.A., 2001, *On the Surprising Behavior of Distance Metrics in High Dimensional Space*, [w:] Van den Bussche, Vianu V. (eds.), International Conference on Database Theory, LNCS, vol. 1973, Springer, Heidelberg, s. 420-434.

Hinneburg A., Aggarwal C.C., Keim D.A., 2000, *What is the Nearest in High Dimensional Spaces*, The VLDB Journal, Bibliothek der Universität Konstanz, s. 506-515.

Houle M.E., Kriegel H.P., Kröger P., Schubert E., Zimek A., 2010, *Can Shared-Neighbor Distances Defeat the Curse of Dimensionality*, [w:] Proceedings of the 22nd International Conference on Scientific and Statistical Database Management, Heidelberg, s. 482-500.

Schnitzer D., Flexer A., Tomasev N., 2014, *Choosing the metric in high-dimensional spaces based on hub analysis*, European Symposium on Artificial Neural Networks ESANN.

Scott D., Thompson J., (1983*), Probability Density Estimation in Higher Dimensions*, [w:] Gentle J. (ed.), *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the* Interface, s. 173-179.

Silverman B., 1986, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Taylor C.C., 1977, *Principal Component and Factor Analysis*, [w:] O'Muircheartaigh C.A., Payne C. (eds.), *The Analysis of Survey Data*, vol. I: Exploring data structures, Wiley&Sons, New York, s. 89-123.

Walesiak M., 2005, *Problemy selekcji i ważenia zmiennych w zagadnieniach klasyfikacji*, Taksonomia 12, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 1076, s. 106-118.

White H., 1989, *Learning in artificial neural networks: a statistical perspective*, Neural Computation, vol. 1, s. 425-464.

Verleysen M., François D., 2005, *The curse of dimensionality in data mining and time series prediction*, 8th International Workshop on Artificial Neural Networks, IWANN, s. 758-770.

THE ASSESSMENT OF IMPACT VALUE OF MINKOWSKI'S CONSTANT FOR THE POSSIBILITY OF GROUP STRUCTURE IDENTIFICATION IN HIGH DIMENSIONAL DATA

Summary: An important decision in the analysis of the variability of units in themultidimensional space, is the choice of the measurement of distance which is accurate for a given problem. This choice is of particular importance, when we have data sets which are described by hundreds of features. In the empirical studies, the most used measure of distance is the exponential metric measure. When the units are described by a very large number of features, an appropriate choice of the Minkowski's constant level is important because has to affect the properties of the exponential metrics. With the increase of dimensionality, the properties of the metrics may change. The aim of this paper is to estimate the influence of the Minkowski's constant and high dimensional space on the group structure which may be obtained. Based on simulation studies the author of this paper shows that the high dimension of the space application of fractional exponential metrics affects the ability to identify the group structure.

Keywords: cluster analysis, curse of dimensionality, exponential metrics.