Kamila Migdał Najman*
Krzysztof Najman**

Ocena wpływu wartości stałej Minkowskiego na możliwość identyfikacji struktury grupowej danych o wysokim wymiarze przy występowaniu wartości nietypowych

Wstęp

W badaniach społeczno-ekonomicznych coraz częściej spotyka się przykłady zbiorów danych, które zawierają znaczną liczbę cech. Jednostki opisane są setkami, tysiącami, a nawet w szczególnych przypadkach milionami cech. Fakt ten, poza konsekwencjami czysto technicznymi związanymi z przechowywaniem i analizą takiego zbioru danych, skutkuje także określonymi problemami natury statystycznej. Wiele metod statystycznej analizy danych zakłada, że liczba cech powinna pozostawać w pewnej rozsądnej relacji do liczby analizowanych jednostek. I jakkolwiek nie ma formalnych zasad w tym względzie, to nie powinno dochodzić do sytuacji, w której liczba cech przekracza 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 inni [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 różne i nie ma prostej zasady rozstrzygającej problem, ile cech powinniśmy rozpatrywać przy danej liczbie jednostek.

Jeżeli jednak w badaniu empirycznym zdarzy się, że liczba cech jest bardzo duża (szczególnie w stosunku do liczby badanych jednostek), badacz może podjąć jedną z dwóch decyzji. Może dokonać redukcji liczby cech lub wręcz przeciwnie – pozostawić wszystkie cechy w badaniu. W sytuacji kiedy podejmie decyzję o redukcji liczby cech, można zastosować trzy podejścia, które pozwolą ustalić optymalny zbiór cech.

* Prof. UG dr hab., Katedra Statystyki, Wydział Zarządzania, Uniwersytet Gdański, ul. Armii Krajowej 101, 81–824, Sopot, kamila.migdal-najman@ug.edu.pl
** Prof. UG dr hab., Katedra Statystyki, Wydział Zarządzania, Uniwersytet Gdański, ul. Armii Krajowej 101, 81–824, Sopot, krzysztof.najman@ug.edu.pl

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, s. 106–118]. Decyzja druga zakłada wręcz przeciwną sytuację. Można pozostawić w badaniu wszystkie dostępne cechy. Decyzja ta może wynikać z ograniczeń czasowych analizy. Czasami nie ma czasu na analizowanie poszczególnych cech, gdy decyzja musi być podjęta (niemal) natychmiast. 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.

Dodatkowym problemem związanym z danymi o wysokim wymiarze jest znaczne prawdopodobieństwo pojawienia się w zbiorze danych jednostek nietypowych (*outliers*, jednostek skrajnych, ekstremalnych). Jednostką nietypową nazywa się taką jednostkę, która ze względu na przynajmniej jedną cechę ekstremalnie różni się od innych jednostek w zbiorze danych. W klasycznych, statystycznych metodach analizy danych jednostki takie są zwykle niepożądane. Nawet pojedyncze przypadki bardzo różniące się wartościami niektórych cech od innych mogą w znaczący sposób wpłynąć na uzyskane wyniki analizy, zniekształcając je. Problem taki jest dobrze widoczny w analizie skupień, w której zwykle dokonuje się pomiaru odległości między jednostkami w przestrzeni wielowymiarowej. Pomiar ten może być nieskuteczny lub mylący, gdy wymiar jest bardzo duży. Problemy te może potęgować występowanie wartości nietypowych. Celem prezentowanych badań jest ocena wpływu wartości stałej Minkowskiego na możliwość identyfikacji struktury grupowej danych o wysokim wymiarze przy występowaniu wartości nietypowych.

1. Przekleństwo wymiarowości

Decyzja badacza o postawieniu w zbiorze danych wszystkich cech ma określone konsekwencje analityczne, które zostały nazwane w literaturze: przekleństwem (problemem, klątwą) wymiarowości (*curse of dimensionality*). Konsekwencje te wynikają przede wszystkim z gwałtownego (wykładniczego) wzrostu objętości hiperprzestrzeni, w której znajdują się obserwowane jednostki, wraz ze wzrostem wymiaru. W 1961 roku Richard Ernest Bellman[1] (1920–1984, matematyk) 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*).

Wyjaśnienie tego zjawiska można pokazać na prostym przykładzie. Załóżmy, że w zbiorze danych znajduje się 10 jednostek opisanych przez jedną cechę, równomiernie rozłożonych w przestrzeni. Niech jednostki te reprezentują 10 (10^1) przedziałów (rysunek 1a). Powiemy wówczas, że każdy przedział jest reprezentowany przez jedną jednostkę. Opiszmy teraz te same jednostki dwiema cechami. Dla 10 jednostek w przestrzeni dwuwymiarowej uzyskamy 100 (10^2) kwadratów (przedziałów – rysunek 1b). Jednostki te zajmują teraz już tylko 10% przestrzeni dwuwymiarowej. Opiszmy te same jednostkami trzema cechami. Reprezentują one już 1000 (10^3) kostek (rysunek 1c), zajmując jedynie 1% przestrzeni trójwymiarowej. Wzrost objętość przestrzeni p-wymiarowej ($j = 1,...,p$) wraz ze wzrostem p powoduje, że ta sama liczba jednostek wypełnia coraz mniejszą część przestrzeni. Aby jednostki w kolejnych wymiarach reprezentowały całą przestrzeń, należałoby ich liczbę zwiększyć wykładniczo wraz ze wzrostem wymiaru.

Problem przekleństwa wymiarowości jest również wyzwaniem w grupowaniu i klasyfikacji danych. W badaniach społeczno-ekonomicznych skupienia często mają postać w przybliżeniu hiperkuli, w której gęstość jednostek zwiększa się w kierunku jej centrum. Wyobraźmy sobie skupienie jednostek w przestrzeni dwuwymiarowej, które ma postać koła. Jeżeli przestrzeń jest opisana dwiema cechami, to przy założeniu, że nie ma ograniczeń, otrzymamy płaszczyznę. Gdyby ją ograniczyć do średnicy koła równej jeden, to powierzchnia tego koła wypełniłaby 78,5%

[1] R. E. Bellman znany jest głównie jako twórca programowania dynamicznego. Jest to technika lub strategia projektowania algorytmów stosowana przeważnie do rozwiązywania zagadnień optymalizacyjnych.

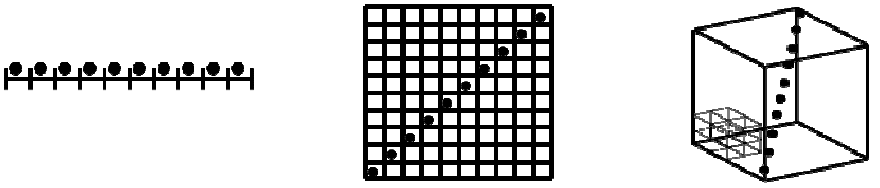
powierzchni kwadratu o boku równym jeden (rysunek 2a). Udział objętości kuli o średnicy równej jeden w objętości sześcianu o boku równym jeden stanowi 52,4% (rysunek 2b). Udział objętości hiperkuli o średnicy równej jeden w przestrzeni np. ośmiowymiarowej, w stosunku do objętości hipersześcianu o boku równym jeden wynosi jedynie 1,6%.

Wraz ze wzrostem wymiaru przestrzeni rośnie udział przestrzeni znajdującej się poza hiperkulą. Jeżeli wymiar 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:

$$\lim_{p\to\infty} \frac{d\max_p - d\min_p}{d\min_p} \to 0. \qquad (1)$$

gdzie: $d\ \min_P$ – jednostka położona najbliżej np. środka ciężkości skupienia, $d\ \max_P$ – jednostka położona najdalej w stosunku do tego samego punktu, np. środka ciężkości skupienia.

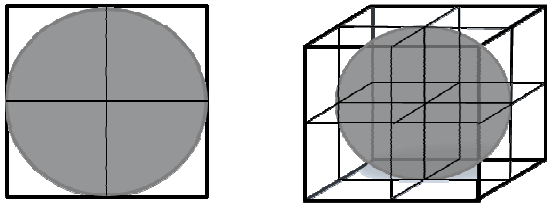
Rysunek 1. 10 jednostek w przestrzeni jedno-, dwu- i trójwymiarowej



a) 10 jedn. w przestrzeni jednowymiarowej | b) 10 jedn. w przestrzeni dwuwymiarowej | c) 10 jedn. w przestrzeni trójwymiarowej

Źródło: Opracowanie własne.

Rysunek 2. Udział hiperkuli w hiperprzestrzeni w przestrzeni dwuwymiarowej (p=2) i trójwymiarowej (p=3)



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

Źródło: Opracowanie własne.

Dlatego też miary odległości najczęściej stosowane w badaniach empirycznych, takie jak odległość euklidesowa, zaczynają tracić swoją skuteczność jako miary oceniające zróżnicowanie jednostek w przestrzeniach o wysokim wymiarze. 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.

2. 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, którą definiuje się następująco:

$$L_k = d_{rs}^{(k)} = \sqrt[k]{\sum_{j=1}^{p} \left| x_{rj} - x_{sj} \right|^k} \tag{2}$$

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, Kriegel, 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 liczbą wymiarów szybciej pogarsza się kontrast między zdefiniowanymi jednostkami.

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*). Szczegółowe wyniki analizy wpływu wartości stałej Minkowskiego na jakość grupowania jednostek o wysokim wymiarze prezentowała K. Migdał-Najman w 2015 roku. Z badań eksperymentalnych wynika, że dla zbiorów liczących setki cech faktycznie korzystne jest stosowanie ułamkowych wartości *k*. W cytowanym badaniu nie brano jednak pod uwagę występowania w zbiorze danych wartości nietypowych. Jest to jednak sytuacja, z którą badacz ma stosunkowo często do czynienia. Jest to tym bardziej prawdopodobne, im więcej cech danej jednostki bierze się pod uwagę.

3. Wyniki badań symulacyjnych

W celu oceny wpływu wielkości wymiaru przestrzeni i poziomu wykładnika normy potęgowej na ocenę jakości uzyskanej struktury grupowej w sytuacji występowania wartości nietypowych przeprowadzono badania symulacyjne, analogiczne jak [Migdał-Najman, 2015, s. 191–199]. Niech zbiór danych składa się z czterech 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. Dodatkowo każdorazowo do zbioru danych dołączono jedną jednostkę nietypową, różniącą się wartością przynajmniej jednej cechy pięciokrotnie od dotychczas ekstremalnej jednostki. Ocena zgodności wyników grupowania ze znaną przynależnością jednostek do skupień zostanie przeprowadzona w oparciu o skorygowany wskaźnik Randa. Dla uśrednienia uzyskanych wyników grupowanie i jego ocena dla każdego wymiaru i każdego poziomu wykładnika *k* zostaną powtórzone 10-krotnie.

Jeżeli w zbiorze danych nie występowały wartości nietypowe (tablica 1) można zauważyć, że w przypadku wysokiego wymiaru przestrzeni grupowanych jednostek, zastosowanie ułamkowego poziomu *k* normy potęgowej znacząco wpływa na poprawę jakości uzyskanej struktury grupowej. Im wyższy wymiar przestrzeni i mniejszy (ułamkowy) poziom

wykładnika *k*, tym uzyskano wyższą jakość grupowania. Obserwacja ta jest zgodna z oczekiwaniami wobec ułamkowego poziomu *k* normy potęgowej i przypuszczeniami innych badaczy.

Tablica 1. Ocena jakości struktury grupowej dla przyjętego wymiaru i poziomu wykładnika normy potęgowej uzyskana w oparciu o skorygowany wskaźnik 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	**1,000**	0,876	0,876	**1,000**	0,809	0,753	0,658	**1,000**	0,781
1/10	**0,934**	**0,926**	**0,926**	**0,971**	0,897	0,832	**1,000**	**1,000**	**1,000**	**1,000**
1/20	**1,000**	**1,000**	0,832	**1,000**	**1,000**	0,714	**1,000**	**1,000**	**1,000**	**1,000**

Źródło: Opracowanie własne na podstawie [Migdał-Najman, 2015, s. 190–199].

Jeżeli jednak w zbiorze danych pojawia się choćby pojedyncza wartość nietypowa, wyniki analiz ulegają zasadniczej zmianie (tablica 2). We wszystkich analizowanych przypadkach zaobserwowano spadek zgodności uzyskanej klasyfikacji ze wzorcem. Spadek ten jest tym większy im mniejszy poziom *k* normy potęgowej.

Tablica 2. Ocena jakości struktury grupowej dla przyjętego wymiaru i poziomu wykładnika normy potęgowej uzyskana w oparciu o skorygowany wskaźnik Randa. Jedna wartość nietypowa

k	Wymiar									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
2	**0,995**	0,445	0,463	0,565	**0,927**	**0,760**	**0,975**	0,560	**0,995**	0,524
1	0,995	**0,975**	0,445	**0,771**	0,670	0,543	0,658	**0,680**	0,562	**0,541**
3/4	0,445	0,437	0,526	0,614	0,516	0,568	0,518	0,572	0,574	0,439
1/2	0,504	0,437	0,438	0,552	0,543	0,532	0,495	0,504	0,439	0,440
1/4	0,438	0,504	**0,572**	0,572	0,504	0,502	0,499	0,533	0,498	0,508
1/10	0,497	0,444	0,497	0,496	0,441	0,440	0,500	0,501	0,572	0,496
1/20	0,439	0,437	0,437	0,568	0,496	0,438	0,443	0,568	0,572	0,438

Źródło: Opracowanie własne.

Wynik ten jest na tyle zaskakujący, że poddano szczegółowej analizie wszystkie badane jednostki w tych częściach przestrzeni, w których powstały największe błędy klasyfikacji. Zaobserwowano tu interesujące zjawisko. Zastosowanie ułamkowej normy potęgowej faktycznie powoduje wzrost kontrastu między jednostkami. Jednak istnienie wartości nietypowych powoduje, że zmienia się on w nieliniowy sposób. Jednostki znajdujące się bliżej centrum skupienia zwiększają swój kontrast w stosunku do jednostek odległych znacznie szybciej. W konsekwencji centra skupień stają się jakby bardziej gęste. Jednostki leżące najdalej od centrum skupienia zachowują słaby kontrast. Nasuwa się tu analogia do fizycznego zjawiska grawitacji. Jednostki o tej samej, niewielkiej masie przyciągane przez inną, dużą masę spadają z prędkością zależącą od odległości, w jakiej się od niej znajdują. Jednostki położone blisko centrum masy spadają szybciej niż te położone daleko. Powoduje to, że w pewnym momencie jednostki, które wyjściowo były blisko centrum, już spadły, a te położone daleko dopiero zaczynają spadać. W efekcie końcowym te odległe jednostki stają się na tyle niepodobne do tych blisko centrum skupienia, że są niewłaściwie identyfikowane (stają się dodatkowymi jednostkami nietypowymi).

Zakończenie

W badaniach empirycznych coraz częściej pojawiają się zbiory danych o wysokim wymiarze. Obserwuje się jednostki opisane setkami a nawet tysiącami cech. Tak duży wymiar 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 do 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. Analizę skupień dodatkowo komplikuje pojawianie się w zbiorze danych jednostek nietypowych. Przeprowadzone badania symulacyjne pokazały, że o ile przy braku jednostek nietypowych zastosowanie ułamkowego wykładnika k dla normy potęgowej znacząco poprawia zdolność do identyfikowania skupień, to pojawienie się przynajmniej jednej jednostki nietypowej znacząco zaburza ten proces. W takim przypadku kontrast między jednostkami zamienia się nieliniowo, pozwalając dobrze zidentyfikować centra skupień, ograniczając jednocześnie poprawną identyfikację jednostek leżących dalej od centrum. Wydaje się, że dokładniejsza identyfikacja tego

procesu wymagała będzie dalszych, pogłębionych badań. Obserwacja powyższa może mieć istotne znaczenie w grupowaniu jednostek każdą metodą wykorzystującą w swojej konstrukcji pomiar odległości między jednostkami.

Literatura

1. Bellman R. E. (1961), *Adaptive control processes*, A Guided Tour, Princeton University Press, Princeton, New Jersey.
2. Beyer K., Goldstein J., Ramakrishnan R., Shaft U. (1999), *When is „nearest neighbor" meaningful*, International Conference on Database Theory, Jerusalem, Israel.
3. Bishop C. M. (1995), *Neural networks for pattern recognition*, Clarendon Press, Oxford.
4. 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.).
5. Hinneburg A., Aggarwal C. C., Keim D. A. (2000), *What is the nearest in high dimensional spaces*, The VLDB Journal, Bibliothek der Universität Konstanz.
6. 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, Springer, Heidelberg.
7. 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.
8. Migdał-Najman K. (2015), *Ocena wpływu wartości stałej Minkowskiego na możliwość identyfikacji struktury grupowej danych o wysokim wymiarze*, w: Jajuga K., Walesiak M. (red.), „Taksonomia" nr 24, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe UE we Wrocławiu.
9. 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.
10. 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.

11. Silverman B. (1986), *Density estimation for statistics and data analysis*, Chapman and Hall, London.
12. 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.
13. Walesiak M. (2005), *Problemy selekcji i ważenia zmiennych w zagadnieniach klasyfikacji*, „Taksonomia" nr 12, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 1076.
14. White H. (1989), *Learning in artificial neural networks: a statistical perspective*, „Neural Computation", Vol. 1.
15. Verleysen M., François D. (2005), *The curse of dimensionality in data mining and time series prediction*, 8th International Workshop on Artificial Neural Networks, IWANN.

Streszczenie

Ważną decyzją w analizie zróżnicowania jednostek w przestrzeni wielowymiarowej jest wybór odpowiedniej dla danego problemu miary odległości. W badaniach empirycznych najczęściej stosuje się miary odległości oparte na metryce potęgowej. W metryce tej, gdy jednostki opisane są bardzo dużą liczbą cech, istotny staje się wybór odpowiedniego poziomu stałej Minkowskiego. Wybór ten jest bardzo ważny, ponieważ istotnie wpływa na własności metryki we właściwym różnicowaniu jednostek. Własności te zmieniają wraz ze wzrostem wymiaru przestrzeni. W artykule poddano analizie wpływ wartości stałej Minkowskiego na poprawność identyfikacji skupień dla danych o wysokim wymiarze, przy występowaniu jednostek nietypowych.

Słowa kluczowe

analiza skupień, przekleństwo wymiarowości, metryka potęgowa, jednostki nietypowe, ocena jakości struktury grupowej

An evaluation of the impact of a Minkovsky constant on the possibility of identification of the group structure in high dimensional in the presence of outliers (Summary)

An important decision, in the analysis of the variability of units in a multidimensional space, is the choice the measurement of distance which is accurate for a given problem. In the empirical studies, the most used measure of distance is the exponential metric. When the units are described by very large number of features, the relevant in appropriate the choose the Minkovsky constant in the exponential metric. The choice is very important, because has to effect on the properties of the exponential metric. With the increase of dimensionality, the properties of the metric may change. The aim of this paper is to estimation of

the influence of the Minkovsky constant and high dimensional space on obtained group structure, in the presence of outliers.

Keywords
cluster analysis, curse of dimensionality, exponential metric, outliers, cluster validity