PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 327

Taksonomia 22
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

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

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

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Andrzej Sokołowski, Magdalena Czaja
Uniwersytet Ekonomiczny w Krakowie

EFEKTYWNOŚĆ METODY *k*-ŚREDNICH W ZALEŻNOŚCI OD SEPAROWALNOŚCI GRUP

Streszczenie: W pracy przedstawiono wybrane wyniki badań symulacyjnych dotyczących efektywności metody *k*-średnich, mierzonej procentem poprawnie zaklasyfikowanych obserwacji w zależności od separowalności grup, błędnej specyfikacji liczby skupień oraz obecności obserwacji odstających. Znaleziono analityczną postać funkcji opisującej efektywność metody *k*-średnich przy poprawnym ustaleniu liczby grup.

Słowa kluczowe: analiza skupień, metoda *k*-średnich, efektywność.

1. Wstęp

Metoda *k*-średnich jest jedną z najpopularniejszych metod taksonomicznych. Niewiele osób, nawet w Polsce, wie, że jej zasady jako pierwszy zaproponował Hugo Steinhaus [1956]. W 1957 r. metodę opisał S. Lloyd w wewnętrznym opracowaniu dla Bell Laboratories. Opublikował ją dopiero w 1982 r. Lloyd [Lloyd 1982]. Najczęściej nazwa metody: *k*-średnich kojarzona jest z Jamesem MacQueenem [MacQueen 1967]. Istnieje wiele wersji i modyfikacji metody *k*-średnich. Dobry przegląd historyczny zawierają prace [Bock 2007 oraz Jain 2010].

Standardowy algorytm metody *k*-średnich obejmuje następujące kroki (przy zadanym *k*):

1) wybór wstępnych środków skupień,
2) przyporządkowanie każdego obiektu do najbliższego środka,
3) wyznaczenie nowych środków skupień,
4) powrót do punktu (2).

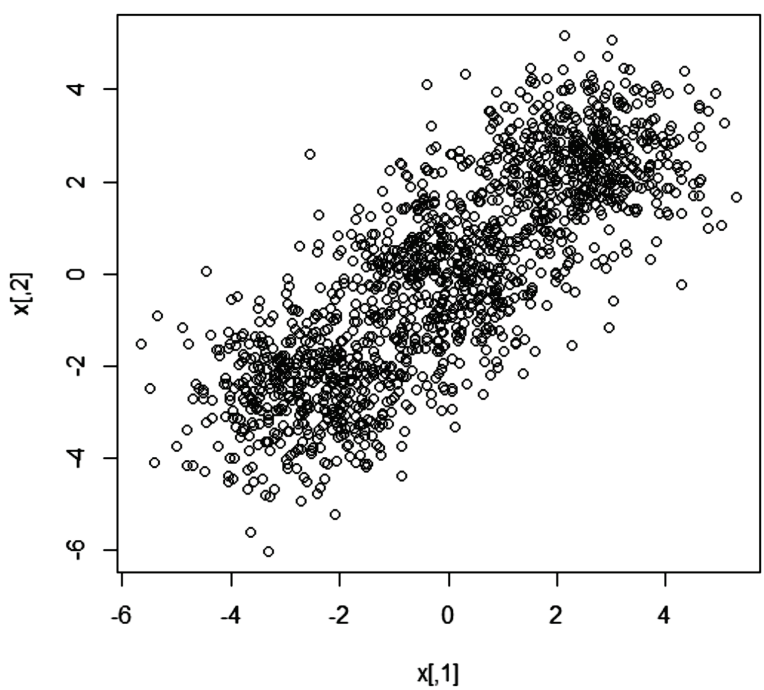
Przesuwanie punktów kończy się, gdy każdy punkt jest bliżej środka własnej grupy niż jakiejkolwiek innej.

2. Cel pracy

Jednym z celów niniejszej pracy była ocena efektywności metody *k*-średnich w zależności od odległości skupień. Jest oczywiste, że metoda tym trafniej będzie rozpoznawała obserwacje i dzieliła zbiór, im dalej od siebie położone są środki grup. Dlatego naszym zadaniem było poszukiwanie postaci analitycznej funkcji, która będzie dobrze opisywała zmianę efektywności metody w zależności od separowalności grup. Efektywność jest tu mierzona procentem prawidłowo zidentyfikowanych obiektów. W pracy przedstawiono też niektóre wyniki badań wpływu obserwacji odstających na zachowanie się metody *k*-średnich oraz zakłócenia efektywności wynikające z niewłaściwej identyfikacji liczby grup.

3. Badania symulacyjne

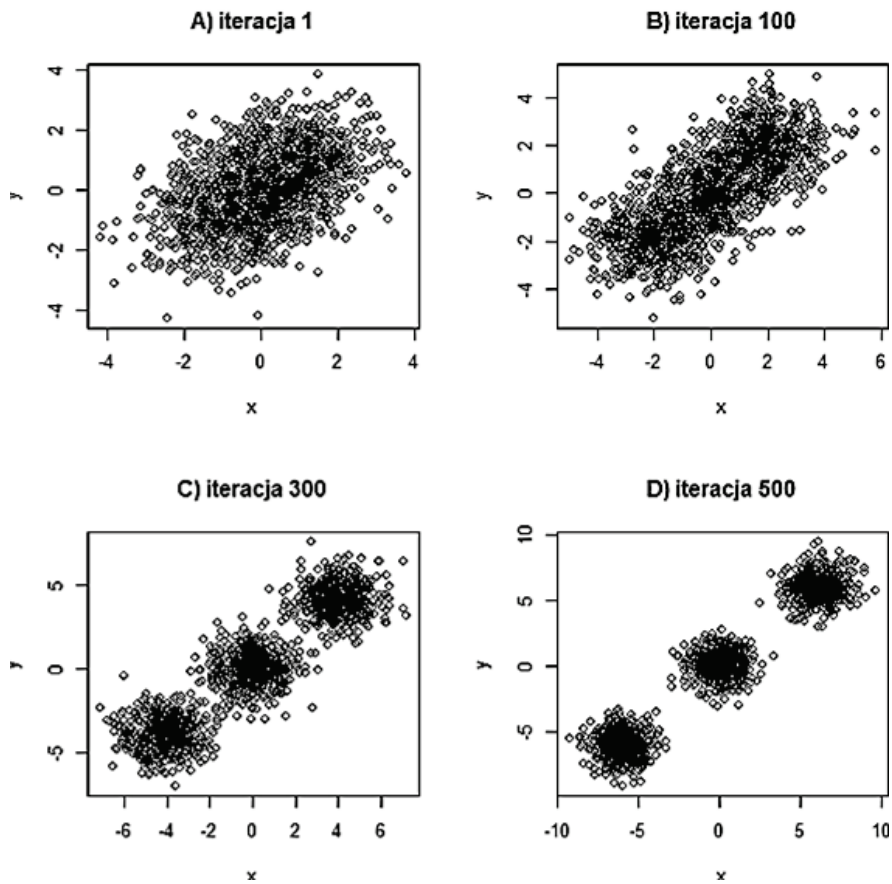
Model symulacyjny przewidywał analizy w przestrzeni dwuwymiarowej dla zadanej liczby trzech grup. Obserwacje z tych grup generowane są przez dwuwymiarowe rozkłady normalne o niezależnych składowych. Rozważano rozkłady o brzegowych odchyleniach standardowych 1 oraz 2. Grupy oddalają się od siebie wzdłuż prostej $y = x$, a odległość między środkami grup oznaczona jest przez α.



Rys. 1. Rozrzut punktów dla $k = 3$ oraz $\alpha = 2{,}5$

Źródło: opracowanie własne.

Generowano próby o liczebnościach 500 oraz 1000. W badaniach wykorzystano program napisany specjalnie w tym celu w środowisku R. Wstępne środki grup wybierane były w sposób losowy, a jako miarę odległości wzięto odległość euklidesową. Efektywność była mierzona frakcją obiektów poprawnie zakwalifikowanych. Przykładowy rozrzut punktów dla $k = 3$, $\alpha = 2{,}5$ przedstawiono na rys. 1. Efektywność wyniosła w tym przypadku 0,942.



Rys. 2. Oddalanie się grup w modelu symulacyjnym

Źródło: obliczenia własne.

W analizach symulacyjnych badano zmiany efektywności związane ze stopniowym oddalaniem się grup od siebie, z niewielkim skokiem, równym 0,01 (pojedynczy skok to tzw. iteracja). Środkowa grupa ma obydwie wartości przeciętne równe zeru. Pierwsza grupa ma wartości przeciętne równe $\mu_1 = -1 - 0{,}01 \cdot j$, gdzie j jest numerem iteracji. Wartości przeciętne trzeciej grupy są równe $\mu_3 = +1 + 0{,}01 \cdot j$. Oddalanie się grup ilustruje rys. 2.

4. Wyniki analizy efektywności metody *k*-średnich

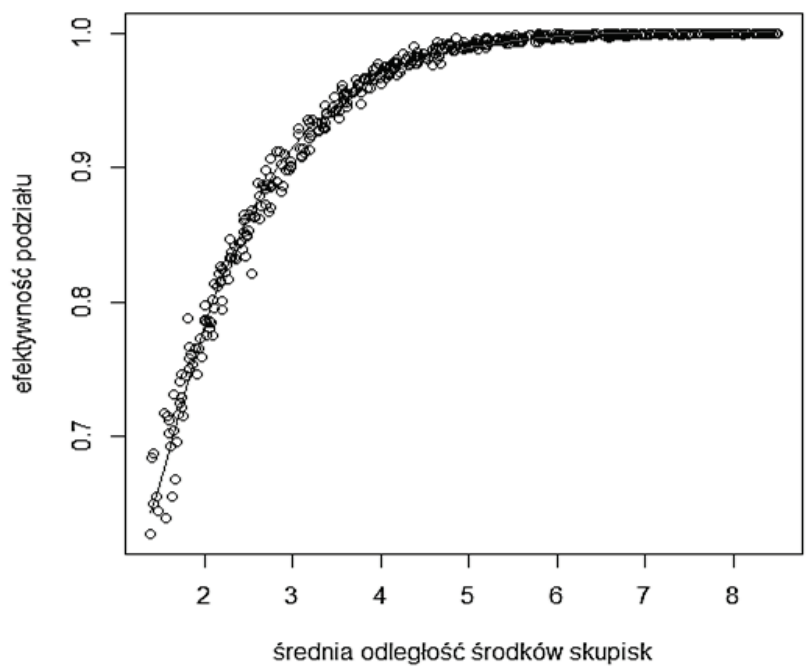
Na drodze analizy wyników symulacyjnych badań efektywności oraz prób dopasowywania różnych funkcji analitycznych stwierdzono, że efektywność metody *k*-średnich da się bardzo dobrze opisać za pomocą funkcji (przy założeniu prawidłowej specyfikacji liczby grup):

$$E = \beta_0 + \beta_1\alpha^{-1} + \beta_2\alpha^{-2} + \beta_3\alpha^{-3}.$$

Wykorzystując wyniki symulacji, oszacowano parametry tej funkcji metodą najmniejszych kwadratów, otrzymując:

$$\hat{E} = 0{,}953729 + 0{,}785544\alpha^{-1} - 3{,}568560\alpha^{-2} + 2{,}609153\alpha^{-3}.$$

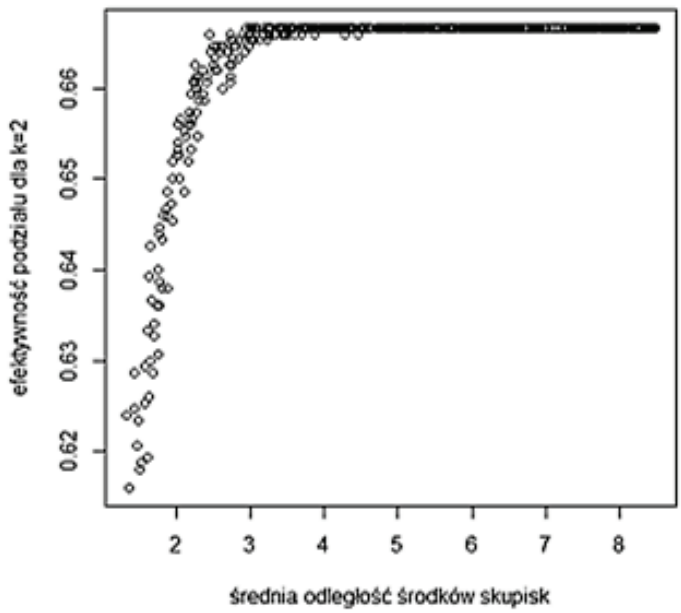
Parametry tej funkcji są wysoce istotne statystycznie (co stwierdzono testem istotności parametrów strukturalnych modelu regresji wielorakiej, wykorzystującym statystykę *t*-Studenta), a jej przebieg ilustruje rys. 3.



Rys. 3. Efektywność metody *k*-średnich w zależności od odległości skupisk

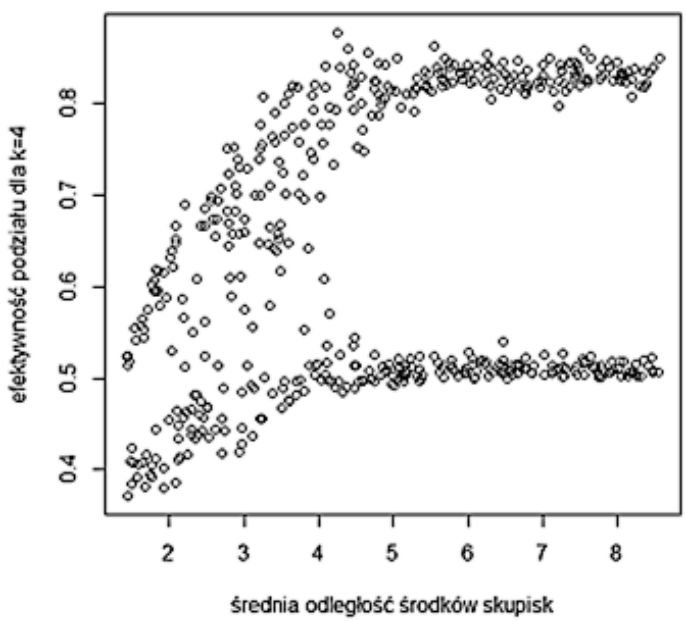
Źródło: obliczenia własne.

Wykonano szereg analiz symulacyjnych, pozwalających na ocenę efektywności metody *k*-średnich przy niewłaściwym ustaleniu liczby skupisk oraz w przypadku obecności obserwacji odstających. Tutaj prezentujemy tylko wybrane, typowe wy-



Rys. 4. Efektywność metody *k*-średnich w przypadku przyjęcia $k = 2$ zamiast $k = 3$

Źródło: obliczenia własne.

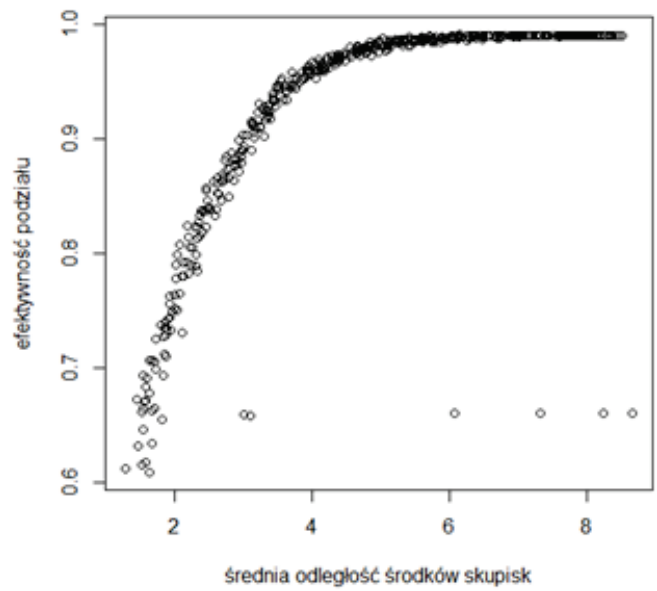


Rys. 5. Efektywność metody *k*-średnich przy błędnym podaniu $k = 4$ zamiast $k = 3$

Źródło: obliczenia własne.

niki. Jeżeli błędnie przyjęto $k = 2$ (zamiast $k = 3$), to efektywność nieznacznie tylko przekracza 2/3, gdyż jedna grupa (środkowa) jest sztucznie rozdzielana na dwie części (rys. 4).

Przy podaniu zbyt dużej liczby skupisk jedno z nich jest sztucznie rozdzielane, a efektywność (przy przyjęciu prawidłowej identyfikacji jako tej, która dobrze przyporządkowuje kod grupy) niejako rozdziela się i nie poprawia nawet przy dużej separowalności (rys. 5).



Rys. 6. Efektywność metody *k*-średnich w przypadku obecności 1% punktów odstających

Źródło: obliczenia własne.

Ciekawe, lecz raczej spodziewane wyniki przyniosły analizy efektywności przy obecności obserwacji odstających. W eksperymencie symulacyjnym zamieniano 1% losowo wybranych punktów na punkty o współrzędnych (15, 15). Jeżeli w przebiegu metody *k*-średnich taki punkt odstający zostaje wylosowany jako wstępny środek skupienia, wówczas metoda nie jest w stanie go już opuścić. Dla pozostałych punktów brakuje więc jednego środka i efektywność nie przekracza 2/3. Gdy punkt odstający nie jest wybierany jako wstępny środek, to jest przyporządkowywany do trzeciej grupy, a punkty odstające obniżają tylko graniczną, maksymalną efektywność.

5. Podsumowanie

Badania symulacyjne wykazały oczywiste i spodziewane zjawisko poprawiania się efektywności metody *k*-średnich w miarę wzrostu separowalności grup. Osiągnięciem pracy jest zidentyfikowanie analitycznej postaci funkcji (i oszacowanie jej parametrów), która bardzo dobrze opisuje zmiany tej efektywności. Prawdopodobieństwo zakłócenia poprawności działania metody *k*-średnich przez obserwacje odstające wzrasta wraz ze wzrostem udziału tych obserwacji w ogólnej liczbie klasyfikowanych obiektów, gdyż wówczas wzrasta prawdopodobieństwo tego, że taki punkt izolowany zostanie wylosowany jako wstępny środek grupy. Eksperymenty symulacyjne potwierdziły też kluczowe znaczenie właściwego wyboru liczby grup.

Literatura

Bock H.-H. (2007), *Clustering methods: A history of k-means algorithms*, [w:] *Selected Contributions in Data Analysis and Classification*, Springer, Berlin – Heidelberg, s. 161-172.

Jain A.K. (2010), *Data clustering: 50 years beyond k-means*, Pattern Recognition Letters, 31 sierpnia, s. 651-666.

Lloyd S. (1982), *Least squares quantization in PCM*, IEEE Trans. Inform. Theory, 28, s. 129-137.

MacQueen J. (1967), *Some methods for classification and analysis of multivariate observations*, Fifth Berkeley Symposium on Mathematics, Statistics and Probability. University California Press, s. 281-297.

Steinhaus H. (1956), *Sur la division des corps materiel en parties*, Bull. Acad. Polon. Sci. Cl. III. 4, s. 801-804.

CLUSTER SEPARABILITY AND THE EFFECTIVENESS OF *K*-MEANS METHOD

Summary: Selected results of simulation analysis on *k*-means method effectiveness is presented in the paper. The effectiveness is measured by the percentage of correctly identified observations. The effectiveness has been studied depending on group separability, wrongly identified number of clusters and the presence of outliers. The analytical function describing the effectiveness has been found and estimated.

Keywords: cluster analysis, *k*-means, effectiveness.