PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1076 2005
TAKSONOMIA 12
Klasyfikacja i analiza danych – teoria i zastosowania

Jerzy Korzeniewski
Uniwersytet Łódzki

PROPOZYCJA NOWEGO ALGORYTMU WYZNACZAJĄCEGO LICZBĘ SKUPIEŃ

1. Metoda średniego przesunięcia okna

W algorytmie prezentowanym w artykule jest wykorzystywana metoda średniego przesunięcia oszacowań maksimów lokalnych funkcji gęstości wektora losowego, zaproponowana przez Comaniciu i Meera [1999]. Idea tej metody jest następująca. Niech $\{x_i\}_{i=1,\ldots,n}$, będzie dowolnym zbiorem n punktów z d-wymiarowej przestrzeni euklidesowej. Estymator jądrowy wielowymiarowej gęstości, oparty na jądrze $K(x)$ i oknie o promieniu h, dany jest wzorem:

$$\hat{f}(x)=\frac{1}{nh^d}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right). \tag{1}$$

Jądrem optymalnym w sensie minimalizacji średniego odchylenia kwadratowego jest jądro Epanecznikowa dane wzorem

$$K_E(x)=\begin{cases}0.5c_d^{-1}(d+2)\left(1-x^Tx\right), & \text{gdy} \quad x^Tx<1\\ 0, & \text{w przeciwnym przypadku}\end{cases}, \tag{2}$$

gdzie c_d jest objętością sfery jednostkowej w d-wymiarowej przestrzeni euklidesowej. Ze wzorów tych łatwo można znaleźć estymator gradientu funkcji gęstości

$$\nabla\hat{f}(x)=\frac{1}{nh^d}\sum_{i=1}^{n}\nabla K\left(\frac{x-x_i}{h}\right). \tag{3}$$

Dla jądra Epanecznikowa otrzymamy wzór :

$$\nabla\hat{f}(x)=\frac{1}{nh^dc_d}\frac{d+2}{h^2}\sum_{x_i\in S_h(x)}[x-x_i]=\frac{n_x}{nh^dc_d}\frac{d+2}{h^2}\left(1/n_x\sum_{x_i\in S_h(x)}[x-x_i]\right). \tag{4}$$

Wielkość

$$M_h(x) = \left(1/n_x \sum_{x_i \in S_h(x)} [x - x_i]\right) = 1/n_x \sum_{x_i \in S_h(x)} x_i - x \qquad (5)$$

nazywana jest średnim przesunięciem okna/próby. Średnie przesunięcie próby zawsze przesuwa ją w kierunku największego wzrostu gęstości, w związku z czym, przesuwając próbę o wektor dany wzorem (5), uzyskamy zbieżność środka okna do lokalnego maksimum funkcji gęstości (por. [Comaniciu, Meer 1999]). Dla dowolnego punktu ze zbioru danych odpowiadającym mu punktem granicznym przy rozmiarze okna h nazywamy granicę ciągu środków okien (czyli środków ciężkości próby) występujących po prawej stronie wzoru (5).

Istotne jest to, że okno jest przesuwane w każdym kroku tej procedury w kierunku jakiegoś maksimum lokalnego funkcji gęstości, którego położenie zależy oczywiście od parametru h. Im mniejsza jest wartość tego parametru, tym maksimum bardziej lokalne, im jest ona większa, tym maksimum ma charakter bardziej globalny. Gdy szerokość okna przekroczy maksymalną odległość w zbiorze, każdemu punktowi zbioru danych będzie odpowiadał ten sam punkt graniczny.

2. Sformułowanie algorytmu

Formalne sformułowanie algorytmu jest następujące:

1. Dla $k = 2$ losujemy w sposób zależny 2 punkty ze zbioru danych i dla każdego punktu znajdujemy odpowiadający mu punkt graniczny w procedurze średniego przesunięcia dla ustalonej szerokości okna h.

2. Sprawdzamy, czy wśród wszystkich par utworzonych spośród k punktów granicznych (dla $k = 2$ będzie tylko jedna para) istnieje choć jedna, dla której odległość jest mniejsza od h.

3. Pierwsze dwa kroki powtarzamy 2000 razy w celu ustalenia prawdopodobieństwa spełnienia warunku z kroku drugiego.

4. Pierwsze trzy kroki powtarzamy dla wszystkich możliwych szerokości okna h (ustalając pewien mały przyrost Δh) z zakresu od 0 do największej odległości w badanym zbiorze danych. Dla $k = 2$ otrzymujemy w ten sposób pseudodystrybuantę pewnej zmiennej losowej.

5. Pierwsze cztery kroki powtarzamy kolejno dla $k = 3, 4, 5, ..., kk$, gdzie kk jest krzywą mającą fazę poziomą „na wysokości" prawdopodobieństwa równego 1.

6. Za prawidłową ilość skupień uznajemy tę, która odpowiada ostatniej (względem k) pseudodystrybuancie, której wykres ma pewne cechy charakterystyczne (szczególnie „fazę poziomą") takie same, jakie mają wykresy pseudodystrybuant dla mniejszych k.

Przez fazę poziomą na wysokości prawdopodobieństwa równego p rozumiemy każdą część, od $h = a$ do $h = b$, wykresu pseudodystrybuanty spełniającą następujące warunki:

a) pseudodystrybuanta w punkcie $h = a$ ma wartość równą $p < 1$,

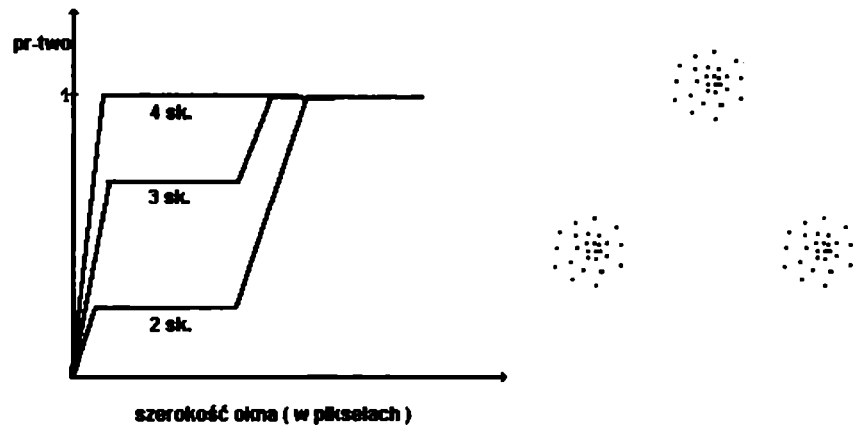
b) w każdym punkcie h z odcinka $(a,b]$ dystrybuanta ma wartość mniejszą lub równą $p + 0{,}02$,

c) długość fazy poziomej tj. $b - a$, jest większa od 1/20 mediany odległości pomiędzy dwoma punktami zbioru.

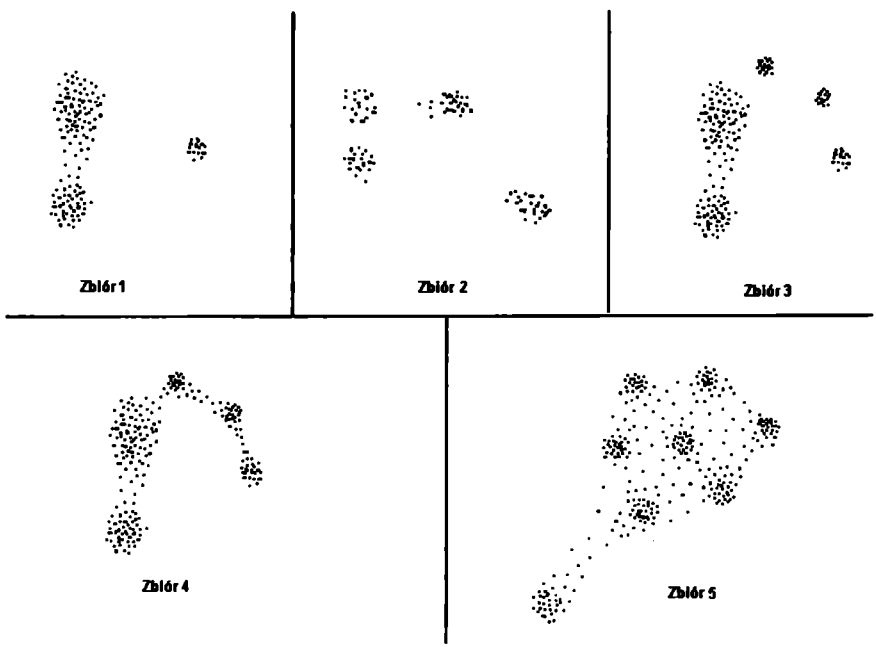
Jak wynika z powyższej definicji fazy poziomej, potrzebna jest znajomość mediany odległości pomiędzy dwoma punktami zbioru. Wyznaczania mediany odległości nie wyróżniono jako osobnego punktu algorytmu, gdyż procedura ta jest bardzo prosta i ma charakter pomocniczy. Medianę wyznaczamy w następujący sposób: gdy zbiór składa się z mniej niż 200 punktów, obliczamy wszystkie odległości pomiędzy wszystkimi parami punktów i wybieramy ich medianę; jeśli zbiór jest większy, losujemy w sposób zależny 300 par punktów i znajdujemy medianę odległości dla tych 300 par.

Idea przedstawionego algorytmu jest następująca. Wyobraźmy sobie zbiór danych w przestrzeni dwuwymiarowej (taki jak na rys. 1), tzn. zbiór składający się z 3 identycznych, jednomodalnych skupień, których mody są jednakowo oddalone od siebie, np. o 80 pikseli dla odległości euklidesowej. Każdy wylosowany punkt zostanie sprowadzony przez procedurę średniego przesunięcia (już dla małego rozmiaru okna np. $h = 5$ pikseli) do odpowiadającego mu punktu granicznego. Będzie nim moda skupienia, do którego należy, ponieważ, jak widać na rysunku, skupienia są „idealnie" skonstruowane, tzn. lokalna gęstość maleje wraz z oddalaniem się od mody skupienia. Zatem, jeśli wylosujemy 2 punkty, warunek z punktu 2 algorytmu będzie spełniony z określonym prawdopodobieństwem (które łatwo można policzyć, gdy znamy liczbę punktów w każdym skupieniu) stałym, niezależnie od tego, czy szerokość okna będzie równa 5, 30, 40, czy 50 pikseli (aż do 80 pikseli), gdyż w tym przedziale nie ma szans na to, że dwie różne mody skupień będą oddalone od siebie o mniej niż 80 pikseli. Po przekroczeniu przez rozmiar okna 80 pikseli prawdopodobieństwo spełnienia warunku z punktu 2 algorytmu wyniesie 1, gdyż okno obejmie wówczas cały zbiór danych i dowolny wylosowany punkt tego zbioru będzie miał ten sam punkt graniczny. Analogicznie w przypadku wylosowania trzech punktów prawdopodobieństwo spełnienia warunku z punktu 2 pozostanie stałe w przedziale do 80 pikseli (będąc oczywiście wyższe niż w przypadku losowania dwóch punktów), po czym wzrośnie szybko do 1. Gdy wylosujemy cztery punkty ze zbioru, co najmniej dwa muszą wpaść do tego samego skupienia (co za tym idzie – będą miały ten sam punkt graniczny) i warunek z punktu 2 będzie spełniony dla dowolnej szerokości okna z prawdopodobieństwem równym 1. Na wykresie przedstawionym na rys. 1 widać, którą liczbę skupień należy uznać za prawidłową. Będzie to liczba odpowiadająca ostatniej krzywej mającej fazę poziomą na wysokości mniejszej od 1. Istnienie fazy poziomej na wysokości mniejszej od 1 dla ustalonego k świadczy bowiem o tym, że są szanse (równe jeden minus wysokość odpowiadająca fazie poziomej) na to, że wylosowane punkty mają punkty graniczne należące do różnych k skupień. Długość fazy poziomej jest zwią-

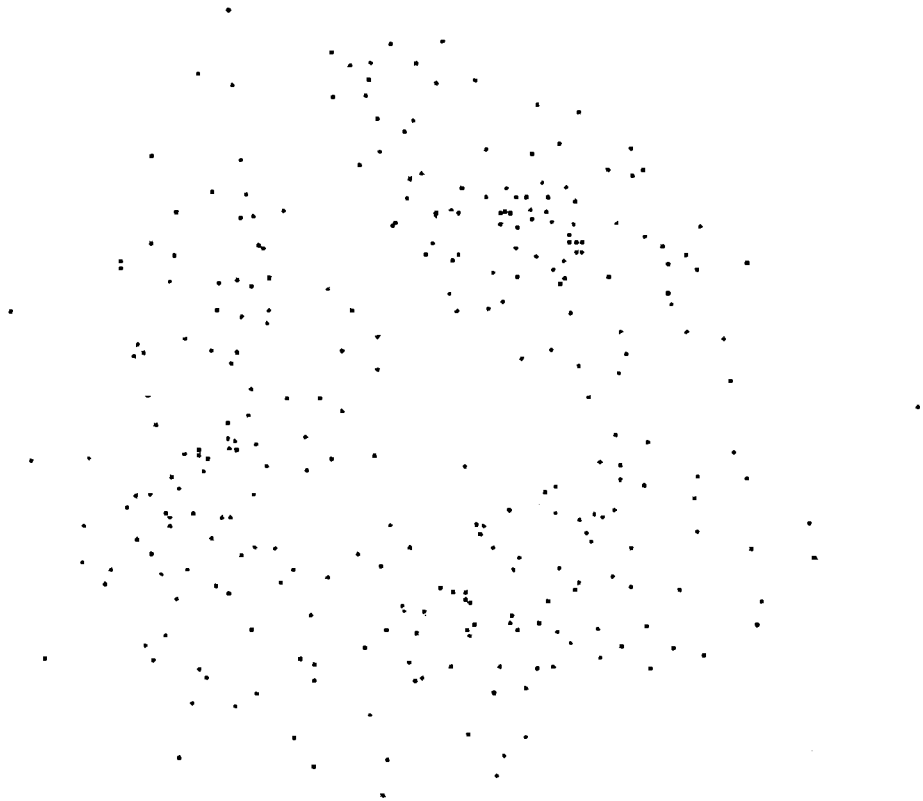
zana z odległościami pomiędzy modami skupień, a prawdopodobieństwo, na wysokości którego występuje faza pozioma, zależy od liczebności skupień w stosunku do liczebności całego zbioru danych.



Rys. 1. Przykładowy zbiór trzech równoodległych skupień z przestrzeni dwuwymiarowej (po prawej) oraz przybliżony wykres pseudodytrybuant dla tego zbioru (po lewej)
Źródło: opracowanie własne.



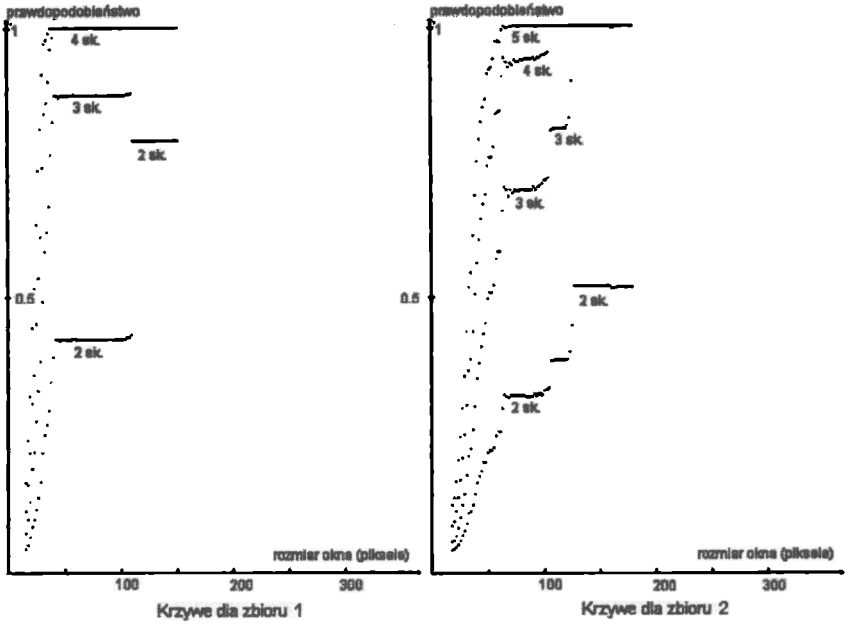
Rys. 2. Pięć pierwszych zbiorów danych z przestrzeni dwuwymiarowej. Największą średnicę ma zbiór piąty (ok. 300 pikseli), najmniejszą zbiór pierwszy (ok. 150 pikseli)
Źródło: opracowanie własne.



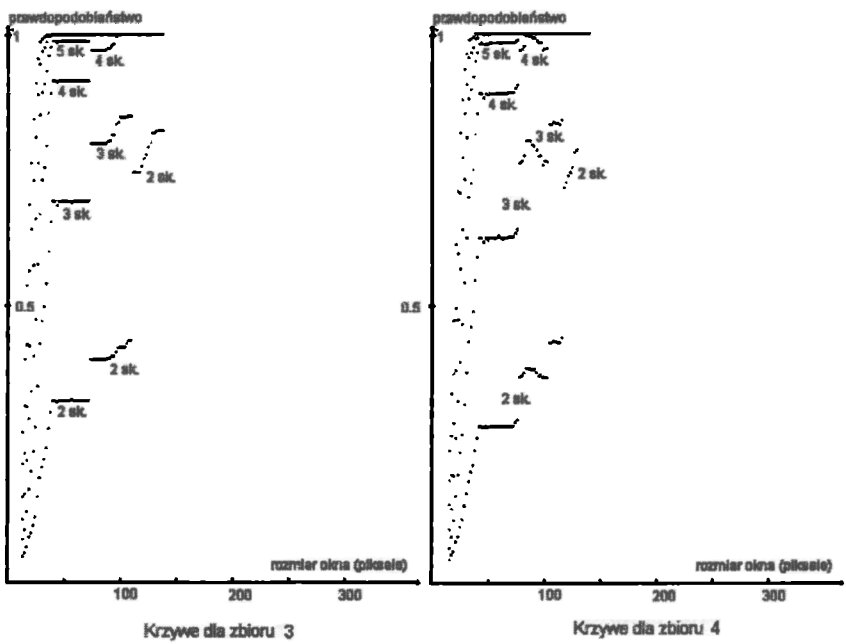
Rys. 3 Zbiór sześciu danych dwuwymiarowych Gordona.
Zbiór ma średnicę ok. 700 pikseli. Wygenerowano go z trzech różnych rozkładów normalnych (100 punktów z każdego rozkładu, zob. [Gordon 1999]).
Źródło: opracowanie własne.

3. Przykłady zastosowania algorytmu

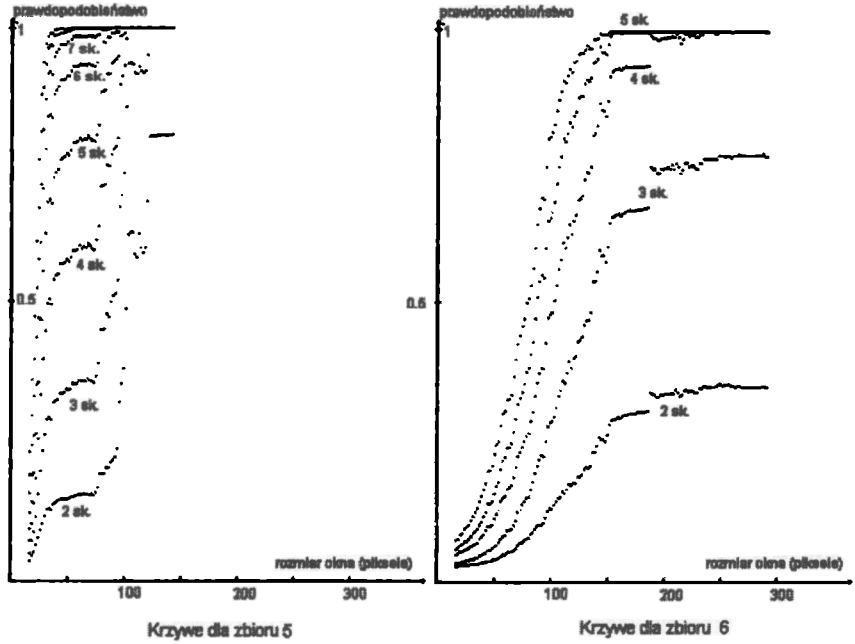
W punkcie tym zbadano zachowanie algorytmu dla kilku sztucznie wygenerowanych zbiorów danych z dwuwymiarowej przestrzeni euklidesowej. Zbiory zostały tak skonstruowane bądź wybrane, by objąć możliwie jak największe zróżnicowanie charakteru skupień, tzn. małe oraz duże ilości skupień, skupienia o zbliżonej liczbie punktów oraz skupienia zróżnicowane pod względem rozmiarów, skupienia dobrze zdefiniowane (oddzielone) i skupienia rozmyte. Przy generowaniu krzywych przyjęto Δh równe 1 piksel.



Rys. 4. Wykresy pseudodystrybuant dla zbiorów pierwszego oraz drugiego
Źródło: opracowanie własne.



Rys. 5. Wykresy pseudodystrybuant dla zbiorów trzeciego oraz czwartego
Źródło: opracowanie własne.



Rys. 6. Wykresy pseudodystrybuant dla zbiorów piątego oraz szóstego
Źródło: opracowanie własne.

Dla zbioru pierwszego (rys. 4) krzywe w oczywisty sposób wskazują na istnienie trzech skupień, co, jak się wydaje, jest dobrą liczbą skupień. Dla zbioru drugiego sytuacja jest również oczywista: krzywe wskazują wyraźnie na cztery skupienia, co jest liczbą poprawną. Dla zbioru trzeciego oraz czwartego (rys. 5) krzywe wskazują wyraźnie na istnienie pięciu skupień, co jest jak najbardziej poprawne. Dla zbioru piątego (rys. 6) krzywe wskazują na istnienie siedmiu skupień, być może nawet ośmiu, co byłoby wskazaniem prawidłowym. Jednak faza pozioma dla krzywej reprezentującej osiem skupień znajduje się na wysokości tak bliskiej jedności, że bez konkretnej miary liczbowej trudno jest wnioskować o istnieniu ośmiu skupień. Dla zbioru szóstego (rys. 6) krzywe wskazują na istnienie czterech skupień. Trudno to wskazanie uznać za błędne, gdyż dane Gordona są tak rozmyte, że wśród nich w ogóle nie ma skupień wyraźnych.

4. Ocena algorytmu

Powyższe rozważania prowadzą do następujących wniosków.

1. Algorytm jest nieparametryczny, tzn. może być stosowany do dowolnego zbioru danych z przestrzeni euklidesowej.

2. Parametry liczbowe występujące w algorytmie, tj. liczba powtórzeń równa 2000, długość fazy poziomej równa co najmniej 1/20 mediany oraz parametr odchylenia w definicji fazy poziomej (równy 0,02) mają oczywiście duży wpływ na działanie algorytmu. Jak wynika z przedstawionych przykładów, w zbiorach z dość wyraźnie zdefiniowanymi skupieniami spisują się one dobrze.

3. Algorytm prawidłowo reaguje na jakość występujących skupień tzn. im lepiej (wyraźniej) zaznaczone skupienia tym bardziej zdecydowane wskazania algorytmu.

4. Czas generowania pięciu krzywych dla zbioru 300 punktów z przestrzeni dwuwymiarowej dla h z przedziału od 0 do 150 pikseli, przy kroku $\Delta h = 1$ piksel jest równy ok. 15 sek. (na komputerze 1 Mhz).

Literatura

Gordon A.D., *Classification*, Chapman & Hall 1999.

Comaniciu D., Meer P., *Mean Shift Analysis and Applications*, IEEE Int. Conf. Computer Vision (ICCV'99), Kerkyra, Greece, 1999, s. 1197-1203.

PROPOSAL OF NEW ALGORITHM FOR DETERMINING THE NUMBER OF CLUSTERS

Summary

The new algorithm is based on the comparison of pseudo cumulative distribution functions of a certain random variable. This variable is defined as follows. For a fixed window size we draw k different points and for every point we find the corresponding limiting point in the mean shift procedure. Then we check if the distance (e.g. Euclidean) between every pair of the limiting points is smaller than the window size. The probability of meeting this condition is the value of the pseudo cumulative distribution function at the point equal to the window size. Analogously we determine the pseudo cumulative distribution functions for different numbers k of clusters. The proper number of clusters is the one that corresponds to the last (with respect to k) curve to possess a horizontal phase at the altitude smaller than 1.