KOMPLETNY ALGORYTM GRADIENTOWEJ KLASTERYZACJI

Piotr Kulczycki [1], Małgorzata Charytanowicz [2]

[1] Instytut Badań Systemowych
Polskiej Akademii Nauk
ul. Newelska 6, 01-447 Warszawa
kulczycki@ibspan.waw.pl

[2] Instytut Badań Systemowych
Polskiej Akademii Nauk
ul. Newelska 6, 01-447 Warszawa
malgorzata.charytanowicz@ibspan.waw.pl

Streszczenie: Celem niniejszej pracy jest podanie algorytmu gradientowej klasteryzacji w postaci kompletnej, dogodnej do bezpośredniego stosowania. Dzięki bliskiej intuicji interpretacji samej koncepcji powyższego algorytmu i jego teoretycznej bazy – statystycznych estymatorów jądrowych, zostanie podana ilustracyjna analiza znaczenia poszczególnych parametrów i wskazanie, jakie efekty uzyskuje się poprzez ich ewentualną zmianę w stosunku do wartości otrzymywanych poprzez zastosowanie kryteriów optymalizacyjnych. W proponowanym algorytmie nie jest wymagane ścisłe ustalenie żądanej ilości klastrów, lecz jedynie wskazanie jej rzędu wielkości, co pozwala lepiej dopasować uzyskiwaną ich liczbę do rzeczywistej struktury danych. Możliwy jest także wpływ na proporcję pomiędzy ilością klastrów w regionach zagęszczenia elementów zbioru danych oraz obszarach gdzie są one rzadkie.

1 Wstęp

Niech dany będzie *m*-elementowy zbiór *n*-wymiarowych wektorów:

$$x_1,\ x_2, \ldots, x_m \ \in \mathbb{R}^n \quad . \tag{1}$$

Zadanie klasteryzacji polega na dokonaniu podziału powyższego zbioru danych na podzbiory (klastry), z których każdy zawiera podobne do siebie elementy, ale istotnie różniące się między poszczególnymi podzbiorami [1, 3, 4]. Taka intuicyjnie oczywista, dostosowana do wymagań aplikacyjnych, definicja jest w równym stopniu niedogodna z teoretycznego jak i praktycznego punktu widzenia, gdyż zawiera ewidentne i ukryte nieścisłości. Przede wszystkim nie jest jednoznacznie określone co oznacza „podobieństwo" (a w konsekwencji „różnienie się") elementów, czy ilość klastrów ma być ustalona arbitralnie czy wynikać z samej struktury zbioru danych (1), jak mierzyć jakość dokonanych podziałów. Jeśli uwzględnić dodatkowo, że aparat

[1] Także: Politechnika Krakowska, Katedra Automatyki, ul. Warszawska 24, 31-155 Kraków.
[2] Także: Katolicki Uniwersytet Lubelski, Instytut Matematyki, al. Racławickie 14, 20-950 Lublin.

matematyczny nie dysponuje naturalną metodyką, jaką można by zastosować do rozwiązania zadania klasteryzacji, to oczywiste staje się istnienie wręcz nadmiernej ilości heurystycznych procedur iteracyjnych, z których każda charakteryzuje się odmiennymi zaletami i wadami oraz pewnymi własnościami, jakie w niektórych problemach mogą być korzystne, a w innych nie.

Odmienną trudność przy stosowaniu owych heurystycznych procedur iteracyjnych stanowi fakt, iż wiele z nich było opracowywanych 30-40 lat temu, gdy dostęp do komputera był możliwy jedynie dla wąskiej grupy specjalistów, dysponujących dogłębną wiedzą niezbędną do wszechstronnych analiz otrzymanych wyników. W licznych przypadkach szereg warunków, między innymi tak fundamentalnych jak założona liczba klastrów lub kryterium stopu, pozostawiano uznaniowości użytkownika. Podstawę analiz uzyskiwanych wyników stanowił często ogląd graficznej ich reprezentacji, niełatwy nawet w przypadku użycia specjalistycznych metod wizualizacji, zasadniczo sprowadzających się do wnioskowania na podstawie odpowiednio skonstruowanych 2- lub 3-wymiarowych przekrojów.

Powyższe pozostaje obecnie w całkowitej sprzeczności z wymaganiami większości współczesnych użytkowników tych metod. Powszechność techniki komputerowej sprawia, iż często nie dysponują oni specjalistyczną wiedzą do wszechstronnej analizy. Najkorzystniejszym dla nich rozwiązaniem jest podanie kompletnego algorytmu, z uwzględnieniem „automatycznych" procedur wyznaczania wszystkich występujących tam wielkości, zarówno funkcji jak i parametrów, a także obrazowych informacji dotyczących ich wpływu na otrzymywane wyniki i – w konsekwencji – korzyści i strat wynikłych z ich ewentualnych zmian. Dodatkową trudność klasycznej analizy zagadnienia na podstawie oglądu graficznej reprezentacji wyników stanowi istotny w ostatnich latach wzrost wartości parametru n, stanowiącego o wymiarowości używanych danych (1).

W swym klasycznym już dziś artykule [2], Fukunaga oraz Hostetler sformułowali naturalną i efektywną koncepcję klasteryzacji, wykorzystującą znaczne możliwości wchodzących wówczas do coraz szerszego użytku statystycznych estymatorów jądrowych, dziś wiodącej metody estymacji nieparametrycznej. Podstawą powyższej koncepcji jest uznanie zbioru danych (1) za próbę losową pozyskaną z pewnej n-wymiarowej zmiennej losowej, wyznaczenie estymatora jądrowego gęstości jej rozkładu i przyjęcie naturalnego założenia, iż poszczególne klastry odpowiadają modom (lokalnym maksimom) uzyskanego estymatora. Przedstawiona metoda została sformułowana jedynie w zakresie ogólnej idei, pozostawiając detale – zgodnie z powszechnie obowiązującym wówczas zwyczajem – do szczegółowej analizy użytkownika. Jej naturalność i przejrzystość interpretacji spowodowała, iż doczekała się ona wielu różnorodnych specjalistycznych zastosowań (por. [15]), a nawet bezwiednego powtórzenia samej idei [14].

Celem niniejszej pracy jest podanie gradientowego algorytmu klasteryzacji, opartego na koncepcji Fukunagi oraz Hostetlera, w wersji kompletnej, dogodnej do stosowania bez konieczności posiadania dogłębnej wiedzy statystycznej i prowadzenia żmudnych badań przedmiotowych ze strony użytkownika. Wykorzystując bliską intuicji interpretację samej koncepcji algorytmu gradientowego oraz jego teoretycznej bazy – estymatorów jądrowych, zostanie podana ilustracyjna analiza znaczenia poszczególnych parametrów i wskazanie jakie efekty uzyskuje się poprzez ich ewentualną zmianę w stosunku do wartości otrzymywanych z użyciem kryteriów optymalizacyjnych. Podstawową cechą opracowanego tu algorytmu będzie brak

wymagania ścisłego ustalenia ilości klastrów, a jedynie wskazanie jej rzędu wielkości, co pozwala dopasować ich liczbę do rzeczywistej struktury danych. Powyższe jest realizowane poprzez wyodrębnienie parametru odpowiadającego za ilość klastrów, po czym wskazanie jego wartości poprzez zastosowanie kryteriów optymalizacyjnych, a następnie skutków jej ewentualnej modyfikacji implikującej zmniejszenie lub zwiększenie rzędu ilości klastrów (aczkolwiek nadal bez wskazania na ich konkretną liczbę). Co więcej, wyodrębniony zostanie kolejny parametr, którego wartość będzie mieć wpływ na proporcję pomiędzy ilością klastrów w regionach zagęszczenia elementów zbioru danych oraz obszarach gdzie są one rzadkie (z probabilistycznego punktu widzenia – na „ogonach" rozkładów). W szczególności odpowiednia relacja między oboma powyższymi parametrami pozwoli na zmniejszenie ilości lub nawet wyeliminowanie klastrów w obszarach rzadszego występowania elementów zbioru danych – co w konsekwencji umożliwia usunięcie wpływu elementów nietypowych (odosobnionych) – praktycznie bez ingerencji w ilość klastrów w obszarach ich zagęszczenia.

Proponowana tu metoda została pozytywnie zweryfikowana między innymi w zastosowaniu do zadania syntezy regulatorów rozmytych przy wyodrębnianiu zbioru ich reguł [10], a także w ramach złożonej i zróżnicowanej metodologicznie procedury wspomagania strategii marketingowej operatora telefonii komórkowej [9], jak również do polepszenia jakości samego estymatora jądrowego funkcji gęstości poprzez klasteryzację próby losowej (1) i zastosowanie odrębnych procedur wobec każdej z wyróżnionych niniejszym grup [8].

2 Statystyczne estymatory jądrowe

Niech dana będzie *n*-wymiarowa zmienna losowa X, o rozkładzie posiadającym gęstość *f*. Jej estymator jądrowy $\hat{f}:\mathbb{R}^n\to[0,\infty)$, wyznacza się na podstawie eksperymantalnie uzyskanych wartości *m*-elementowej próby losowej x_1, $x_2,\ldots,x_m$, i w podstawowej postaci definiowany jest on jako

$$\hat{f}(x)=\frac{1}{mh^n}\sum_{i=1}^{m}K\left(\frac{x-x_i}{h}\right)\ , \tag{2}$$

gdzie $m\in\mathbb{N}\setminus\{0\}$, współczynnik $h>0$ zwany jest parametrem wygładzania, natomiast mierzalną funkcję $K:\mathbb{R}^n\to[0,\infty)$ spełniającą warunek $\int_{\mathbb{R}^n}K(x)\mathrm{d}x=1$, symetryczną względem zera i mającą w tym punkcie słabe maksimum globalne, określa się mianem jądra. Wyboru postaci jądra K oraz wyznaczenia wartości parametru wygładzania h dokonuje się najczęściej na podstawie kryterium minimalizacji scałkowanego błędu średniokwadratowego.

I tak, w praktyce wybór postaci jądra nie ma istotnego znaczenia z punktu widzenia statystyki i dzięki temu możliwe jest uwzględnienie przede wszystkim własności otrzymanego estymatora (np. klasa regularności, przyjmowanie dodatnich wartości) lub aspektów obliczeniowych, korzystnych z punktu widzenia konkretnego problemu aplikacyjnego [5, 13]. Najpopularniejsze wśród praktyków jest jądro normalne:

$$K(x) = (2\pi)^{-n/2} e^{-x^{\mathrm{T}}x/2} \quad . \tag{3}$$

Jest ono klasy $\mathcal{C}^{\infty}$ i przyjmuje dodatnie wartości w całej swej dziedzinie.

Duże znaczenie dla jakości estymacji ma natomiast wyznaczenie parametru wygładzania. Zbyt mała jego wartość powoduje pojawienie się znacznej ilości ekstremów lokalnych estymatora $\hat{f}$, co jest sprzeczne z faktycznymi własnościami realnych populacji. Z drugiej strony, za duże wartości parametru h skutkują nadmiernym wygładzeniem tego estymatora, maskującym specyficzne cechy badanego rozkładu. Zgodnie z formułą najbardziej uniwersalnej metody krzyżowego uwiarygodnienia [5, 12], może być ona obliczona jako wartość realizująca minimum funkcji $g : (0, \infty) \to \mathbb{R}$ postaci

$$g(h) = \frac{1}{m^2 h^n} \sum_{i=1}^{m} \sum_{j=1}^{m} \tilde{K}\left(\frac{x_j - x_i}{h} \right) + \frac{2}{mh^n} K(0) \quad , \tag{4}$$

gdzie $\tilde{K}(x) = K^{*2}(x) - 2K(x)$, natomiast K^{*2} oznacza kwadrat splotowy funkcji K. Dla jądra normalnego (3):

$$K^{*2}(x) = (4\pi)^{-n/2} e^{-x^{\mathrm{T}}x/4} \quad . \tag{5}$$

Przy specyficznych uwarunkowaniach opracowano szereg innych metod doboru parametru wygładzania. W szczególności, w przypadku jednowymiarowym polecić można prostą i skuteczną metodę podstawień [5, 13].

W przypadku podstawowej definicji estymatora jądrowego (2), wpływ parametru wygładzania na poszczególne jądra jest jednakowy. Wyjątkowo korzystne rezultaty uzyskuje się dzięki zindywidualizowaniu tego wpływu, co realizuje się poprzez procedurę tzw. modyfikacji parametru wygładzania. Polega ona na określeniu dodatnich parametrów modyfikujących $s_1, s_2, \ldots, s_m$ przypisanych poszczególnym jądrom, definiowanych formułą

$$s_i = \left(\frac{\hat{f}_X(x_i)}{\tilde{s}} \right)^{-c} \quad , \tag{6}$$

gdzie $c \in [0, \infty)$, natomiast $\tilde{s}$ jest średnią geometryczną liczb $\hat{f}_X(x_1)$, $\hat{f}_X(x_2), \ldots, \hat{f}_X(x_m)$, i ostatecznie zdefiniowaniu estymatora jądrowego z modyfikacją parametru wygładzania w następującej postaci:

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} \frac{1}{s_i^n} K\left(\frac{x - x_i}{hs_i} \right) \quad . \tag{7}$$

Dzięki powyższej procedurze obszary w których estymator jądrowy przyjmuje małe wartości (np. w zakresie „ogonów") są dodatkowo wygładzane, a obszary związane z dużymi wartościami – wyostrzone, co pozwala lepiej ukazać specyficzne cechy

rozkładu. Parametr c stanowi o intensywności procedury modyfikacji. Bazując na przesłankach wynikłych z kryterium minimum scałkowanego błędu średnio-kwadratowego, można wstępnie przyjąć

$$c = 0{,}5 \quad . \tag{8}$$

Zgodnie z podstawową postacią definicji estymatora jądrowego (2), parametr wygładzania ma taki sam wpływ na poszczególne współrzędne badanej zmiennej. Biorąc pod uwagę możliwość znacznych różnic w skalach powyższych współrzędnych, dla części z nich wartość tego parametru może okazać się za mała, natomiast dla innych zbyt duża. W związku z tym zalecana jest liniowa transformacja

$$X \equiv RY \quad , \tag{9}$$

przy czym macierz R jest dodatnio określona. W niniejszej pracy używana będzie jej postać diagonalna

$$R = \begin{bmatrix} \sqrt{Var(X_1)} & 0 & \cdots & 0 \\ 0 & \sqrt{Var(X_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{Var(X_n)} \end{bmatrix} , \tag{10}$$

gdzie $Var(X_i)$ oznacza wariancję i-tej współrzędnej zmiennej X. Po zastosowaniu transformacji (9), estymator jądrowy przyjmuje postać

$$\hat{f}(x) = \frac{1}{mh^n \det(R)} \sum_{i=1}^{m} K\left(R^{-1} \frac{x - x_i}{h} \right) \quad . \tag{11}$$

Naturalność i przejrzystość formuły estymatorów jądrowych pozwala na łatwe dostosowywanie ich cech do uwarunkowań badanego problemu, np. poprzez ograniczenie nośnika funkcji $\hat{f}$. Szczegółowo zostanie teraz omówiony stosowany w niniejszej pracy przypadek lewostronnego ograniczenia jednowymiarowej zmiennej losowej, a zatem gdy spełniony ma być warunek $\hat{f}(x) = 0$ dla każdego $x < x_*$, przy dowolnie ustalonym $x_* \in \mathbb{R}$. Koncepcja proponowanej procedury sprowadza się do dokonania symetrycznego „odbicia" względem ograniczenia x_* fragmentu dowolnego i-tego jądra, który znajduje się poza przedziałem $[x_*, \infty)$ i traktowanie go jako fragmentu jądra „zaczepionego" w symetrycznym „odbiciu" elementu x_i względem ograniczenia x_*, czyli w punkcie $2x_* - x_i$. Podstawowa definicja estymatora jądrowego (2) przyjmuje wtedy postać

$$\hat{f}(x) = \frac{1}{mh} \sum_{i=1}^{m} \chi_{[x_*, \infty)}(x) \left[K\left(\frac{x - x_i}{h} \right) + K\left(\frac{x + x_i - 2x_*}{h} \right) \right] , \tag{12}$$

gdzie $\chi_{[x_*, \infty)}$ oznacza funkcję charakterystyczną przedziału $[x_*, \infty)$. Fragmenty

poszczególnych jąder „ucinane” poza założonym nośnikiem są zatem „uzupełniane” wewnątrz nośnika w bezpośrednim sąsiedztwie ograniczenia, a więc w zakresie akceptowalnego najczęściej w praktyce błędu.

Koncepcje wyrażone formułami (7), (11) i (12) mogą być łączone w naturalny sposób.

Szczegółowe informacje o estymatorach jądrowych można znaleźć w monografiach [5, 12, 13]. Przykładowe zastosowania przedstawiono w publikacjach [6, 7].

3 Kompletny algorytm gradientowej klasteryzacji

Niech – zgodnie z poczynionym na wstępie założeniem – dany będzie *m*-elementowy zbiór *n*-wymiarowych wektorów (1). Traktowany on będzie jako próba losowa otrzymana z *n*-wymiarowej zmiennej losowej X, której rozkład posiada gęstość f. Korzystając z metodyki przedstawionej w punkcie 2, można utworzyć jej estymator jądrowy $\hat{f}$. Przyjmijmy naturalne założenie, iż poszczególne klastry odpowiadają modom, czyli lokalnym maksimom funkcji $\hat{f}$, a przyporządkowywanie im poszczególnych elementów zbioru (1) następuje poprzez przesuwanie tych elementów w kierunku gradientu $\nabla\hat{f}$, z odpowiednio dobranym krokiem.

Powyższe realizowane jest iteracyjnie poprzez algorytm gradientowej klasteryzacji [2], oparty na klasycznej koncepcji algorytmu Newtona, określony wzorami

$$x_j^0 = x_j \quad \text{dla} \quad j = 1, 2, \ldots, m \tag{13}$$

$$x_j^{k+1} = x_j^k + b\frac{\nabla\hat{f}(x_j^k)}{\hat{f}(x_j^k)} \quad \text{dla} \quad j = 1, 2, \ldots, m \quad \text{oraz} \quad k = 0, 1, \ldots, k^* \ , \tag{14}$$

gdzie $b > 0$ i $k^* \in \mathbb{N} \setminus \{0\}$. W praktyce rekomenduje się $b = h^2/(n+2)$.

Aby doprecyzować powyższą ideę do postaci kompletnego algorytmu, należy sformułować i szczegółowo zanalizować następujące aspekty:

1. formułę konstrukcji estymatora jądrowego $\hat{f}$;
2. ustalenie warunku stopu (i w konsekwencji liczby kroków k^*);
3. określenie procedury tworzenia klastrów i zaliczania do nich poszczególnych elementów zbioru (1), po wykonaniu ostatniego, k^*-tego kroku;
4. analizę wpływu wartości poszczególnych parametrów na uzyskiwane wyniki.

Powyższe zagadnienia będą teraz przedmiotem rozważań kolejnych podpunktów.

3.1 Formuła konstrukcji estymatora jądrowego

Dla potrzeb dalszej części prezentowanej tu koncepcji został przyjęty estymator jądrowy $\hat{f}$, skonstruowany przy zastosowaniu modyfikacji parametru wygładzania z jej standardową intensywnością (8), a także transformacją liniową z diagonalną macierzą (10). Jądro K rekomendowane jest w postaci normalnej (3) z uwagi na jego

dużą efektywność, a jednocześnie różniczkowalność w całej dziedzinie, dogodność obliczeń analitycznych przy wyznaczaniu gradientu, jak również przyjmowanie dodatnich wartości, co w każdym przypadku zabezpiecza przed dzieleniem przez zero we wzorze (14).

3.2 Ustalenie warunku stopu

Przyjęto, iż algorytm (13)-(14) ulega zakończeniu, gdy po wykonaniu kolejnego, *k*-tego kroku spełniony jest warunek

$$\left|D_k - D_{k-1}\right| \leq aD_0 \quad , \tag{15}$$

gdzie $a > 0$ oraz

$$D_0 = \sum_{i=1}^{m} \sum_{j=i+1}^{m} d(x_i, x_j) \tag{16}$$

$$D_{k-1} = \sum_{i=1}^{m} \sum_{j=i+1}^{m} d(x_i^{k-1}, x_j^{k-1}) \quad , \qquad D_k = \sum_{i=1}^{m} \sum_{j=i+1}^{m} d(x_i^{k}, x_j^{k}) \quad , \tag{17}$$

przy czym *d* oznacza metrykę euklidesową w $\mathbb{R}^n$. Tak więc D_0 oraz D_{k-1} i D_k reprezentują sumy odległości między poszczególnymi elementami zbioru (1) – odpowiednio – przed rozpoczęciem algorytmu oraz po wykonaniu $k-1$ i *k*-tego kroku. Rekomenduje się wartość $a = 0{,}001$. Jej ewentualne zmniejszenie nie wpływa istotnie na otrzymywane wyniki, natomiast zwiększenie wymaga indywidualnej weryfikacji ich poprawności.

Ostatecznie: jeśli po wykonaniu *k*-tego kroku spełniony jest warunek (15), to przyjmuje się $k^* = k$ i w konsekwencji krok ten traktuje jako ostatni.

3.3 Procedura tworzenia klastrów i zaliczanie do nich poszczególnych elementów

We wstępnym etapie rozważany jest zbiór

$$x_1^{k^*} \, , \; x_2^{k^*} \, , \ldots , x_m^{k^*} \quad , \tag{18}$$

złożony z elementów zbioru (1) przemieszczonych po wykonaniu k^*-tego kroku algorytmu (13)-(14). Następnie należy utworzyć zbiór wzajemnych odległości tych elementów:

$$\left\{ d(x_i^{k^*}, x_j^{k^*}) \right\}_{\substack{i=1,2,\ldots,m \\ j=i+1,i+2,\ldots,m}} \quad . \tag{19}$$

Jego liczność wynosi $m_d = m(m-1)/2$.

Traktując zbiór (19) jako próbę losową 1-wymiarowej zmiennej losowej, należy następnie utworzyć pomocniczy estymator jądrowy $\hat{f}_d$ wzajemnych odległości elementów zbioru (18). W nawiązaniu do metodyki estymatorów jądrowych opisanej

w punkcie 2, ponownie proponowane jest jądro normalne (3) oraz stosowanie procedury modyfikacji parametru wygładzania dla standardowej wartości parametru (8), a także dodatkowo lewostronne ograniczenie nośnika do przedziału $[0,\infty)$.

Kolejną czynność stanowi wyznaczenie lokalnego minimum funkcji $\hat{f}_d$ dla najmniejszej wartości argumentu należącej do przedziału $[0, \max_{\substack{i=1,2,...,m \\ j=i+1,i+2,...,m}} d(x_i, x_j)]$, ale z pominięciem ewentualnego lokalnego minimum w zerze. W tym celu należy potraktować zbiór (19) jako próbę losową i obliczyć jej odchylenie standardowe σ_d, a następnie obliczać kolejno wartości

$$\hat{f}_d(0),\ \hat{f}_d(0.01\sigma_d),\ \hat{f}_d(0.02\sigma_d),... \tag{20}$$

aż do wyznaczenia najmniejszej liczby rzeczywistej $x_d \in (0,\infty)$, takiej że

$$\hat{f}_d(x_d - 0.01\sigma_d) > \hat{f}_d(x_d) \quad \text{oraz} \quad \hat{f}_d(x_d) \le \hat{f}_d(x_d + 0.01\sigma_d)\ . \tag{21}$$

Wyznaczoną w ten sposób wartość można interpretować jako połowę odległości miedzy „środkami" pary najbliżej siebie położonych potencjalnych klastrów. (Jeśli taka wartość nie istnieje, to należy uznać istnienie jednego klastra i zakończyć algorytm.)

I wreszcie, zostaną teraz wyznaczone klastry. W tym celu należy:

1. wziąć dowolny element zbioru (18) i wstępnie utworzyć jednoelementowy klaster złożony z tego elementu;
2. poszukać elementu zbioru (18) różnego od elementu zawartego w klastrze, odległego od niego nie więcej niż x_d; jeśli takiego elementu nie ma, to przejść do punktu 4, a jeśli jest, to dodać go do klastra;
3. poszukać elementu zbioru (18) różnego od elementów zawartych w klastrze, odległego od co najmniej jednego z nich nie więcej niż x_d; jeśli taki jest, to dodać go do klastra i powtórzyć punkt 3;
4. dodać uzyskany klaster do „listy klastrów" i usunąć ze zbioru (18) elementy tego klastra; jeśli tak zredukowany zbiór (18) nie jest pusty, to należy wrócić do punktu 1; jeśli jest – zakończyć algorytm.

Uzyskana w ten sposób „lista klastrów" zawiera wszystkie wyróżnione w powyższej procedurze klastry. Rozważana procedura przyjmuje zatem formę kompletnego gradientowego algorytmu klasteryzacji, w jego podstawowej postaci – ewentualne modyfikacje i ich wpływ na otrzymywane wyniki zostaną przedstawione w kolejnym podpunkcie.

3.4 Analiza wpływu wartości poszczególnych parametrów na uzyskiwane wyniki

Warto ponowić spostrzeżenie, że prezentowany algorytm klasteryzacji nie wymagał wstępnego, w praktyce często arbitralnego, ustalenia liczby klastrów – ich ilość zależy jedynie od wewnętrznej struktury danych, określonej w postaci zbioru (1). Poprzez odpowiedni dobór wartości parametrów estymatora jądrowego możliwy jest

jednak wpływ na sam rząd wielkości liczby klastrów, a także na proporcje ich występowania w regionach zagęszczenia elementów tego zbioru względem obszarów gdzie są one rzadkie.

Otóż, jak wspomniano w punkcie 2, zbyt mała wartość parametru wygładzania h powoduje pojawienie się znacznej ilości ekstremów lokalnych estymatora jądrowego, natomiast za duże jego wartości skutkują nadmiernym wygładzeniem tego estymatora. W tej sytuacji, zmniejszenie wartości parametru h względem uzyskanej za pomocą procedur opartych na kryterium scałkowanego błędu średniokwadratowego spowoduje w konsekwencji powiększenie rzędu ilości klastrów. Analogicznie, zwiększenie wartości parametru wygładzania skutkuje zmniejszeniem rzędu ilości klastrów. Należy podkreślić, iż w obu przypadkach, pomimo wpływania na rząd ilości klastrów, ich ścisła liczba nadal zależeć będzie jedynie od wewnętrznej struktury danych. Na podstawie przeprowadzonych badań można rekomendować zmianę wartości parametru wygładzania w zakresie od –25% do +50%. Poza tym zakresem otrzymywane wyniki wymagają indywidualnej weryfikacji.

Jak zostało wspomniane w punkcie 2, intensywność procedury modyfikacji parametru wygładzania implikowana jest wartością parametru c - standardowo danej wzorem (8). Jej zwiększenie wygładza estymator jądrowy w tych obszarach w których elementy zbioru (1) są rzadkie, a także dodatkowo wyostrza go w rejonach ich zagęszczenia – w konsekwencji, jeżeli wartość parametru c zostanie zwiększona, to zmniejszy się ilość klastrów w obszarach rzadkiego występowania danych i jednocześnie zwiększy w obszarach ich zagęszczenia. Przeciwne efekty wystąpią w przypadku zmniejszenia wartości tego parametru. Na podstawie przeprowadzonych badań można rekomendować zmianę wartości parametru c w zakresie od 0 (co oznacza brak modyfikacji) do 1,5. Zwiększenie ponad 1,5 wymaga indywidualnej weryfikacji poprawności otrzymywanych wyników. Szczególnie rekomendowane jest $c=1$.

W praktyce bardzo często występuje jednak żądanie pozostawienia bez zmian klastrów w obszarach zagęszczenia danych – najważniejszych z praktycznego punktu widzenia – i jednoczesnego zredukowania lub wręcz zlikwidowania klastrów w obszarach gdzie elementy występują rzadko, gdyż są one głównie związane z elementami nietypowymi (odosobnionymi), powstałymi nierzadko przez różnego rodzaju błędy. Łącząc powyższe rozważania można zaproponować jednoczesne zwiększenie standardowej intensywności modyfikacji parametru wygładzania (8) oraz powiększenie wartości parametru wygładzania h, wyznaczonej w oparciu o kryterium scałkowanego błędu średniokwadratowego, do wartości h^* danej wzorem

$$h^* = \left(\frac{3}{2}\right)^{c-0,5} h \quad . \tag{22}$$

Łączne działanie obu tych czynników skutkuje podwójnym wygładzeniem funkcji $\hat{f}$ w obszarach gdzie elementy zbioru (1) są rzadkie, co implikuje znaczną redukcję klastrów powstałych w tych obszarach. Równocześnie czynniki te znoszą się w obszarach zagęszczenia danych, praktycznie nie wpływając na identyfikację tamtejszych klastrów. Na podstawie przeprowadzonych badań można rekomendować zmianę wartości parametru c w zakresie od 0,5 do 1,25. Jej zwiększenie ponad 1,25 wymaga indywidualnej weryfikacji poprawności otrzymywanych wyników.

Szczególnie rekomendowane jest $c = 0{,}75$.

Warto jeszcze skomentować możliwość redukcji zbioru (19). Jego liczność m_d jest w praktyce zbyt duża, nie tylko ze względu na zależność kwadratową względem liczności zbioru (1), ale także fakt, iż estymator $\hat{f}_d$ jest związany z 1-wymiarową zmienną losową, a $\hat{f}$ najczęściej z wielowymiarową, ze swej natury wymagającą istotnie liczniejszej próby. Przy bardzo dużej liczności próby (19) warto użyć znanych z literatury procedur kompresji danych (np. [11]).

4 Bibliografia

[1] Everitt B.S., Landau S., Leese M. (2001) *Cluster Analysis,* Arnold, London.

[2] Fukunaga K., Hostetler L.D. (1975) The estimation of the gradient of a density function, with applications in Pattern Recognition, *IEEE Transactions on Information Theory*, **21**, 32-40.

[3] Jain A.K., Dubes R.C. (1988) *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs.

[4] Koronacki J., Mielniczuk J. (2001) *Statystyka*, WNT, Warszawa.

[5] Kulczycki P. (2005) *Estymatory jądrowe w analizie systemowej*, WNT, Warszawa.

[6] Kulczycki P. (2007) Estymatory jądrowe w badaniach systemowych, w: Kulczycki P., Hryniewicz O., Kacprzyk J. (red.) *Techniki informacyjne w badaniach systemowych*, WNT, Warszawa, 79-105.

[7] Kulczycki P. (2008) Kernel Estimators in Industrial Applications, w: Prasad B. (ed.) *Soft Computing Applications in Industry*, Springer-Verlag, Berlin, 69-91.

[8] Kulczycki P., Charytanowicz M. (2008) A Complete Gradient Clustering Algorithm Formed with Kernel Estimators, w druku.

[9] Kulczycki P., Daniel K. (2006) Algorytm wspomagania strategii marketingowej operatora telefonii komórkowej, w: Kacprzyk J., Budziński R. (red.) *Badania operacyjne i systemowe; metody i techniki*, EXIT, Warszawa, 245-256.

[10] Łukasik Sz., Kowalski P., Charytanowicz M., Kulczycki P. (2007) Fuzzy Model Identification Using Kernel-Based-Density Clustering, *Proc. Sixth International Workshop on Intuitionic Fuzzy Sets and Generalized Net*, Warszawa, 5 października 2007, EXIT, Warszawa, w druku.

[11] Pal S.K., Mitra P. (2004) *Pattern Recognition Algorithms for Data Mining*, Chapman and Hall, London.

[12] Silverman B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[13] Wand M.P., Jones M.C. (1994) *Kernel Smoothing*, Chapman and Hall, London.

[14] Wang W.-J., Tan Y.-X., Jiang J.-H., Lu J.-Z., Shen G.-L., Yu R.-Q. (2004) Clustering based on kernel density estimation: nearest local maximum searching algorithm, *Chemometrics and intelligent laboratory systems*, **72**, 1-8.

[15] Zhang K., Tang M., Kwok J.T. (2005) Applying Neighborhood Consistency for Fast Clustering and Kernel Density Estimation, *Proc. IEEE International Conference on Vision and Pattern Recognition*, San Diego (USA), 20-25 czerwca 2005, **2**, 1001-1007.