Kompletny Gradientowy Algorytm Klasteryzacji: możliwości aplikacyjne w zagadnieniach analizy systemowej

Piotr Kulczycki [1,2], Małgorzata Charytanowicz [1,3], Piotr A. Kowalski [1,2], Szymon Łukasik [1,2]

[1] Polska Akademia Nauk, Instytut Badań Systemowych, Centrum Informatycznych Metod Analizy Danych
[2] Politechnika Krakowska, Katedra Automatyki i Technik Informacyjnych
[3] Katolicki Uniwersytet Lubelski, Katedra Analizy Numerycznej i Programowania

Streszczenie — Celem niniejszej publikacji jest zaprezentowanie aplikacyjnych aspektów i własności kompletnej postaci gradientowego algorytmu klasteryzacji, a także ich ilustracja dla konkretnych praktycznych zadań z zakresu analizy systemowej. Podstawową cechą powyższego algorytmu jest brak wymagań dotyczących arbitralnego ustalenia liczby klastrów, co umożliwia lepsze ich dopasowanie do rzeczywistej struktury danych. Możliwe jest wszakże wskazanie parametrów, których ewentualna zmiana wpływa na rząd wielkości liczby klastrów oraz proporcję między ich liczbą w obszarach zagęszczenia elementów zbioru danych oraz obszarach gdzie są one rzadkie. Ponadto prezentowany algorytm może być wykorzystany do wykrywania i ewentualnej eliminacji elementów nietypowych (odstających). Powyższe cechy okazały się bardzo cenne w prezentowanych zastosowaniach z zakresu bioinformatyki (kategoryzacja ziaren zbóż), zarządzania i marketingu (określenie strategii marketingowej operatora telefonii komórkowej) oraz inżynierii (synteza rozmytego regulatora PID), i mogą być równie użyteczne w wielu różnorodnych praktycznych zadaniach.

Słowa kluczowe: eksploracyjna analiza danych, klasteryzacja, metody nieparametryczne, estymatory jądrowe, kategoryzacja ziaren zbóż, strategia operatora telefonii komórkowej, rozmyty regulator PID.

1. WSTĘP

Klasteryzacja [Krzyśko *et al*; 2008] stanowi obecnie podstawową procedurę eksploracyjnej analizy danych. Brak jest tu jednak naturalnego aparatu matematycznego, jakim na przykład przy badaniu ekstremów jest rachunek różniczkowy. W tej sytuacji wieloznaczność zarówno interpretacji jak i poszczególnych ścisłych elementów definicyjnych (np. co oznacza „podobieństwo" elementów, czy też jak mierzyć jakość uzyskanego podziału) implikuje olbrzymią różnorodność koncepcji i w konsekwencji procedur klasteryzacji. Z jednej strony istotnie utrudnia to fazę badawczą, ale z drugiej pozwala lepiej dopasować stosowaną metodykę do specyfiki i wymagań konkretnego rozważanego zadania.

Celem niniejszego artykułu jest przedstawienie własności kompletnej postaci gradientowego algorytmu klasteryzacji w aspekcie jego aplikacyjnych możliwości, zilustrowanych na przykładzie praktycznych zadań z zakresu bioinformatyki, zarządzania i marketingu oraz inżynierii, dotyczących – odpowiednio – kategoryzacji ziaren zbóż dla potrzeb produkcji materiału siewnego, określenia strategii marketingowej operatora telefonii komórkowej, a także syntezy rozmytego regulatora PID w zakresie redukcji zbioru jego reguł.

Niech zatem dany będzie *m* elementowy zbiór danych składający się z *n*-wymiarowych wektorów

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

traktowany tu jako próba otrzymana z *n*-wymiarowej zmiennej losowej. W swym klasycznym już dziś artykule [Fukunaga, Hostetler; 1975], Fukunaga i Hostetler sformułowali naturalną i efektywną koncepcję klasteryzacji, wykorzystującą znaczne możliwości wchodzących wówczas do coraz szerszego użytku statystycznych estymatorów jądrowych [Kulczycki; 2005]. 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, a w konsekwencji „doliny" funkcji gęstości stanowią rozgraniczenie tak powstałych klastrów. Przedstawiona metoda została sformułowana jedynie w zakresie ogólnej idei, pozostawiając detale – zgodnie z powszechnie obowiązującym wówczas zwyczajem – do szczegółowej analizy użytkownika. Algorytm Fukunagi oraz Hostetlera został w ramach pracy [Kulczycki, Charytanowicz; 2010] uzupełniony i ostatecznie podany w wersji kompletnej, dogodnej do stosowania bez konieczności posiadania dogłębnej wiedzy statystycznej oraz prowadzenia żmudnych badań przedmiotowych. Charakteryzuje się on następującymi cechami:

1. wartości wszystkich parametrów mogą być efektywnie wyznaczone z użyciem numerycznych procedur opartych na kryteriach optymalizacyjnych;
2. algorytm nie wymaga ścisłego ustalenia liczby klastrów, co pozwala lepiej dopasować ich liczbę do rzeczywistej struktury danych;
3. wyodrębniony został pojedynczy parametr odpowiadający za liczbę klastrów; pokazano jak ewentualne zmiany jego wartości – np. względem obliczonej z użyciem kryterium

optymalizacyjnego (por. punkt 1) – wpływają na zmniejszenie lub zwiększenie ilości klastrów, aczkolwiek nadal bez wskazania konkretnej ich liczby;

4. wskazany został kolejny parametr, którego wartość ma wpływ na proporcję pomiędzy liczbą klastrów w regionach zagęszczenia elementów zbioru danych oraz obszarach gdzie są one rzadkie; także tu określona została jego wartość oparta na kryterium optymalizacyjnym (por. punkt 1), ewentualnie poddawana modyfikacjom w celu zwiększenia ich ilości w regionach zagęszczenia danych kosztem redukcji w obszarach rzadszego ich występowania lub odwrotnie;
5. odpowiednia relacja pomiędzy powyższymi parametrami pozwala na redukcję lub nawet eliminację klastrów w rejonach rzadkiej lokalizacji danych, praktycznie bez wpływu na obszary gęste;
6. prezentowany algorytm kreuje również mało liczne lub nawet jednoelementowe klastry, których elementy mogą być traktowane jako elementy nietypowe (odosobnione) [Barnett, Lewis; 1994], co umożliwia ich usunięcie lub przypisanie do najbliższych liczniejszych klastrów, osiągane przykładowo przez opisane wcześniej zmiany wartości odpowiednich parametrów.

Własności zapisane w punkcie 4, i w konsekwencji 5, są szczególnie warte podkreślenia jako praktycznie niewystępujące w innych procedurach klasteryzacyjnych. W praktycznych zastosowaniach warta podkreślenia są implikacje połączenia punktów 1 i 2 oraz ewentualnie dodatkowo 3. Nietypowe możliwości otwiera także własność zapisana w punkcie 6.

Szerszy opis przedstawionego tu materiału dostępny będzie w ramach publikacji [Kulczycki *et al*; 2012].

2. KOMPLETNY GRADIENTOWY ALGORYTM KLASTERYZACJI

Pełny opis rozważanego tu Kompletnego Gradientowego Algorytmu Klasteryzacji, opartego na nieparametrycznej metodyce estymatorów jądrowych [Kulczycki; 2005], został zawarty w artykule [Kulczycki, Charytanowicz; 2010] oraz skrótowo w jego wcześniejszej konferencyjnej publikacji [Kulczycki, Charytanowicz; 2008]. Warto ponowić spostrzeżenie, że prezentowany algorytm klasteryzacji nie wymagał wstępnego, w praktyce często arbitralnego i nieuzasadnionego, 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.

I tak, wprowadzony w definicji estymatora jądrowego parametr wygładzania h bezpośrednio wpływa na liczbę klastrów. Jego zmniejszenie względem wartości podstawowej, uzyskiwanej zwykle w oparciu o kryterium średniokwadratowe, powoduje powiększenie rzędu ilości klastrów. Analogicznie, zwiększenie wartości parametru wygładzania skutkuje zmniejszeniem tego rzędu. Należy podkreślić, iż w obu przypadkach, pomimo wpływania na rząd ilości klastrów, ich ścisła liczba nadal zależeć będzie jedynie od wewnętrznej struktury danych (1).

Podobnie, wprowadzony w definicji procedury modyfikacji parametru wygładzania parametr c wpływa na proporcję ilości klastrów w obszarach zagęszczenia elementów zbioru danych (1) i rejonach gdzie są one rzadkie. Zwiększenie jego wartości względem otrzymanego na podstawie kryterium średniokwadratowego implikuje zmniejszenie ilość klastrów w obszarach rzadkiego występowania danych i jednocześnie zwiększa w obszarach ich zagęszczenia. Przeciwne efekty wystąpią w przypadku zmniejszenia wartości tego parametru.

W praktyce często pojawia się 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, powstałymi nierzadko przez różnego rodzaju błędy. Łącząc powyższe rozważania można zaproponować jednoczesne zwiększenie standardowej intensywności modyfikacji parametru wygładzania c oraz odpowiednie – określone konkretnym wzorem – powiększenie wartości parametru wygładzania h.

Warto jeszcze wspomnieć o możliwości, a w niektórych zagadnieniach praktycznych wręcz konieczności redukcji wymiaru n oraz ewentualnie liczności próby m. Dogodny algorytm desygnowany do zagadnień analizy danych ukaże się niebawem w ramach publikacji [Kulczycki, Łukasik; 2012].

3. ZASTOSOWANIA

Kompletny Gradientowy Algorytm Klasteryzacji został wszechstronnie przebadany zarówno na pseudolosowych danych statystycznych jak i ogólnie dostępnych benchmarkowych. Jego cechy predestynują go do szerokiego zakresu zastosowań. Poniżej, w kolejnych sekcjach, przedstawione zostaną wyniki uzyskane w trzech przedsięwzięciach badawczych z zakresu bioinformatyki, zarządzania i marketingu oraz inżynierii.

3.1. Kategoryzacja ziaren dla potrzeb produkcji materiału siewnego

Bioinformatyka – dyscyplina zajmująca się stosowaniem narzędzi matematycznych oraz informatycznych do rozwiązywania problemów nauk biologicznych – rozwija się obecnie wyjątkowo dynamicznie i w zróżnicowany sposób. Możliwości powstałe dzięki rozwojowi oraz powszechności techniki komputerowej spowodowały wyraźne zbliżenie i współdziałanie w ramach odmiennych dotychczas metod badawczych nauk ścisłych i przyrodniczych. Poniżej zostaną przedstawione wyniki badań prowadzonych w ramach szerszego projektu nad kategoryzacją ziaren, na podstawie geometrycznych cech nasion, uzyskanych ze zdjęcia rentgenowskiego ich ziarniaków.

W celu przedstawienia ilustracyjnej i porównawczej prezentacji aspektu badań, w którym stosowano Kompletny Gradientowy Algorytm Klasteryzacji, analizie poddana została próba ziaren pszenicy zebranych z eksperymentalnych poletek Instytutu Agrofizyki Polskiej Akademii Nauk w

Lublinie. Testowany zbiór składał się z ziaren trzech odmian pszenicy, *Kama*, *Rosa* oraz *Canadian*, po 70 losowo wybranych sztuk. Wysokiej jakości wizualizację ich wewnętrznej struktury uzyskano z wykorzystaniem miękkiej techniki rentgenowskiej, nie niszczącej testowanego materiału. Po zeskanowaniu uzyskanych zdjęć, otrzymano – przy zastosowaniu specjalnie do tego celu stworzonego oprogramowania ZIARNA – następujących siedem geometrycznych wielkości ziarniaków**:** jego powierzchnia A, obwód P, zwartość $C = 4\pi A / P^2$, długość, szerokość, współczynnik asymetrii, i wreszcie długość bruzdy ziarniaka. Każde testowane ziarno było zatem reprezentowane przez 7-wymiarowy wektor ($n = 7$), a ich zbiór tworzył 210-elementową próbę (1). We wstępnej fazie, wymiar został zredukowany do $n = 2$ z użyciem klasycznej analizy składowych głównych PCA.

W wyniku użycia Kompletnego Gradientowego Algorytmu Klasteryzacji ze standardowymi wartościami parametrów h oraz c, uzyskanymi za pomocą kryterium średniokwadratowego, otrzymano 7 klastrów o licznościach 76, 64, 57, 7, 3, 2, 1. Można wnioskować, że 3 pierwsze reprezentują trzy użyte do niniejszej analizy odmiany pszenicy, natomiast mało liczne pozostałe 4 klastry zawierają elementy nietypowe, niewykluczone iż uszkodzone mechanicznie. Jeśli pominąć 13 sztuk zawartych w owych 4 mało licznych klastrach (6% całej populacji), to liczba poprawnych klasyfikacji ziaren wynosiła kolejno 91%, 97%, 88% odpowiednio dla odmian pszenicy *Kama*, *Rosa* and *Canadian*. Warto zwrócić uwagę, iż uzyskanie powyższego wyniku nie wymagało założenia *a priori* ilości wymaganych klastrów, co w praktycznych problemach z zakresu biologii może być informacją trudno dostępną lub wręcz niemożliwą do uzyskania.

W przypadku konieczności przypisania każdego elementu do któregoś z odpowiednio licznych klastrów, efekt ten można uzyskać stosownie zmieniając wartości parametrów h and c względem standardowych wartości otrzymanych z kryteriów optymalizacyjnych. Jak wspomniano, zwiększając sukcesywnie wartość pierwszego z nich, potencjalnie zmniejsza się ilość klastrów, a z kolei zmniejszanie drugiego uniemożliwia rozbicie tworzonych w ten sposób dużych klastrów. Postępując w ten sposób, trzy liczne klastry uzyskano dla h zwiększonego o 75% oraz pięciokrotnie pomniejszonego parametru c. Liczba poprawnych klasyfikacji była w tym przypadku nieznacznie mniejsza od otrzymanej poprzednio i wynosiła dla poszczególnych odmian odpowiednio 91%, 96%, 88%.

Powyższe wyniki były porównywalne z uzyskanymi innymi metodami, między innymi klasycznej procedury k-średnich, aczkolwiek w tym przypadku wymagało to dodatkowego podania prawidłowej informacji o ilości klasyfikowanych odmian, co w praktyce nie zawsze jest informacją dostępną.

Podsumowując, zastosowanie Kompletnego Gradientowego Algorytmu Klasteryzacji, umożliwiło poprawną klasyfikację ziaren trzech odmian pszenicy bez apriorycznej informacji o ich liczbie. Co więcej, ze standardowymi wartościami parametrów powyższy algorytm umożliwiał również identyfikację nietypowych elementów testowanej próby, przykładowo uszkodzonych mechanicznie.

Powyższy ilustracyjny przykład dotyczący trzech odmian pszenicy może być uogólniony na inne zagadnienia kategoryzacji materiału siewnego o podobnych uwarunkowaniach.

Szczegółowe rozważania znaleźć można w pracy [Charytanowicz *et al*; 2010].

3.2. Strategia marketingowa operatora telefonii komórkowej

Notowany w ostatnich latach gwałtowny rozwój telefonii komórkowej w naturalny sposób wymusza na operatorach ciągłe dostosowywanie swojej strategii marketingowej do zróżnicowanych i zmieniających się wymagań klientów, przy jednoczesnej maksymalizacji spodziewanego zysku. Jednak brak spójnej koncepcji postępowania może skutkować niejednorodnością traktowania poszczególnych klientów, a w konsekwencji ich niezadowoleniem i przejściem do firm konkurencyjnych. W celu uniknięcia tych zagrożeń konieczne jest wypracowanie jednolitej strategii marketingowej, zwłaszcza wobec klientów kluczowych, z którymi prowadzone są indywidualne negocjacje ustalające warunki kontraktu. Poniżej zostaną przedstawione badania przeprowadzone dla jednego z polskich operatorów sieci komórkowej, dotyczących klientów biznesowych o dłuższym stażu, czyli takich którzy posiadają więcej niż 30 kart SIM przez co najmniej 2 lata.

W praktyce istnieje szerokie spektrum wielkości określających cechy poszczególnych abonentów. Po szczegółowej analizie ekonomicznych aspektów rozważanego zagadnienia przyjęto, że poszczególni klienci biznesowi mogą być skutecznie scharakteryzowani przez trzy z nich**:** średni miesięczny przychód z karty SIM, staż w sieci oraz liczbę aktywnych kart SIM. I tak, każdy z *m*-elementowej bazy danych x_1, $x_2, \ldots, x_m$ jest opisywany za pomocą 3-wymiarowego wektora:

$$x_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ x_{i,3} \end{bmatrix} \quad \text{dla} \quad i = 1, 2, \ldots, m \quad , \tag{2}$$

gdzie $x_{i,1}$ oznacza średni miesięczny przychód z karty SIM *i*-tego klienta, $x_{i,2}$ – jego staż w sieci, natomiast $x_{i,3}$ – liczbę aktywnych kart SIM.

We wstępnej fazie, ze zbioru x_1, $x_2, \ldots, x_m$ zostały wyeliminowane elementy nietypowe (odosobnione), zgodnie z procedurą zawartą w pracy [Kulczycki, Prochot; 2004], podobnie jak Kompletny Gradientowy Algorytm Klasteryzacji opartą na metodyce estymatorów jądrowych. Dzięki temu zwiększona została homogeniczność (jednorodność) zbioru danych, poprzez usunięcie mało wartościowej informacji, nie mającej pozytywnego wpływu na kolejne etapy badawcze.

Następnie, poprzez zastosowanie wobec tak zredukowanego zbioru Kompletnego Gradientowego Algorytm Klasteryzacji, uzyskany został podział danych reprezentujących zróżnicowanych klientów na grupy o podobnej charakterystyce. Wyniki uzyskane dla standardowej wartości parametru c, wskazywały na zbytnią ilość klastrów o niedużej liczności ulokowanych w obszarach małego zagęszczenia elementów próby, zawierających najczęściej mało istotnych acz specyficznych klientów, a także zbyt liczne główne skupienie zawierające ponad połowę elementów. Zgodnie z przedstawionymi

własnościami stosowanego algorytmu, wartość tego parametru została powiększona dwukrotnie. Uzyskano dzięki temu żądany efekt**:** liczba „peryferyjnych" klastrów istotnie zmniejszyła się, a główne skupienie uległo rozbiciu. Otrzymana liczba klastrów była zadowalająca, co spowodowało, iż ewentualna zmiana wartości parametru h okazała się zbędna. Ostatecznie uzyskano podział rozważanej w tej fazie 1639-elementowej próby na 26 klastrów o następujących licznościach**:** 488, 413, 247, 128, 54, 41, 34, 34, 33, 28, 26, 21, 20, 14, 13, 12, 10, dwa klastry 4-elementowe, trzy 3-elementowe, dwa 2-elementowe i dwa klastry 1-elementowe. Warto zwrócić uwagę na wyraźnie zarysowane cztery grupy**:** pierwsza z nich to dwa duże klastry o licznościach 488 i 413, następnie dwa średnie 247- i 128-elementowe, po czym małe – dziewięć liczących od 20 do 54 oraz wreszcie 13 klastrów zawierających mniej niż 20 elementów. Następnie przystąpiono do eliminacji klastrów o licznościach mniejszych od 20, ale z wyłączeniem tych, które zawierały klientów kluczowych (klastry 14-, 13- i 10-elementowe) i składających się co najmniej w połowie z klientów prestiżowych (klaster 12-elementowe). Ostatecznie do dalszej analizy pozostało 17 klastrów.

Następnie, dla każdego z tak określonych klastrów, wyznaczane były optymalne – z punktu widzenia oczekiwanych zysków operatora – reguły postępowania wobec należących do niego abonentów. Ze względu na fakt, iż wykorzystywana do tego celu ocena ekspertów jest ze swej natury nieprecyzyjna, użyte zostały elementy logiki rozmytej [Kacprzyk, 1986] oraz teorii preferencji [Fodor, Rubens; 1994] – szczegóły tej fazy wykraczają jednak poza zakres badawczy prezentowany w niniejszej publikacji i z tego powodu zostaną pominięte. Warto zaznaczyć, iż powyższe operacje nie są przeprowadzane w czasie rzeczywistym podczas negocjacji z klientem, lecz powinny być jedynie uaktualniane co pewien okres czasu, rzędu 1-6 miesięcy.

Klient, z którym prowadzone są negocjacje scharakteryzowany jest – w nawiązaniu do wzoru (2) – podobnie jak pozostali, przez 3-wymiarowy wektor. Dane te nawiązują do historii abonenta w sieci, w przypadku prowadzenia renegocjacji warunków kontraktu, lub mogą pochodzić z innej sieci, jeśli podejmuje się próbę przejęcia. Rzeczony klient zostaje przyporządkowany do odpowiedniej grupy podobnych mu abonentów (w ramach podziału otrzymanego uprzednio w wyniku klasteryzacji) z zastosowaniem algorytmu klasyfikacji bayesowskiej także przy użyciu metodyki estymatorów jądrowych (literaturę przedmiotową można znaleźć w artykule [Kulczycki, Kowalski; 2010]). Biorąc pod uwagę, że reguły postępowania dla poszczególnych klastrów zostały już wcześniej określone, ostatecznie kompletuje to prezentowaną tu procedurę wspomagania strategii marketingowej operatora telefonii komórkowej wobec klienta biznesowego.

Szczegółowe rozważania znaleźć można w artykule [Kulczycki, Daniel; 2009].

3.3. Synteza rozmytego regulatora PID

Rozmyte regulatory typu PID stanowią cenne – z aplikacyjnego punktu widzenia – uogólnienie powszechnie stosowanych, dokładnie przebadanych i poznanych przez praktyków klasycznych regulatorów PID. Wersja rozmyta jest szczególnie użyteczna dla układów nietrywialnych, przykładowo

silnie nieliniowych i/lub nieokreślonych, gdyż dzięki większej liczbie stopni swobody, regulatory takie potrafią lepiej dopasować się do specyfiki takiego obiektu. Z drugiej jednak strony, zbyt duża liczba owych stopni swobody może spowodować trudności w prawidłowym ustaleniu funkcji i parametrów regulatora, implikując nieprawidłową pracę systemu, a w skrajnym przypadku nadmierne rozbudowanie jego struktury i w konsekwencji wręcz – jego praktyczną nierealizowalność. Zagadnienie możliwie dużego, lecz nie pogarszającego jakości, uproszczenia struktury rozmytych regulatorów PID ma zatem podstawowe aplikacyjne znaczenie.

Poniżej rozważane będą regulatory rozmyte PID typu Takagi-Sugeno [Yager, Filev; 1994]. Ich koncepcja opiera się na zbiorze (bazie) k rozmytych reguł postaci

$$\text{IF } (x \text{ is } A_j) \text{ THEN } (y = f_j(x)) \quad \text{dla} \quad j = 1,2,...,k \ . \tag{3}$$

Jeżeli – zgodnie z charakterem ujęcia rozmytego – element x należy do wielu zbiorów w stopniu określanym wartościami ich funkcji przynależności, czyli poprzez $\mu_{A_j}(x)$, to ostatecznie y przyjmuje postać unormowanej średniej

$$y = \frac{\sum_{j=1}^{k} \mu_{A_j}(x)\, f_j(x)}{\sum_{j=1}^{k} \mu_{A_j}(x)} \ . \tag{4}$$

W przypadku rozmytych regulatorów typu PID, współrzędne wektora x związane są z uchybem oraz jego całką i pochodną, a zmienna y stanowi wygenerowane sterowanie. Nawet jeżeli zakłada się proste trójkątne lub trapezowe funkcje przynależności μ_{A_j}, a funkcje f_j w postaci liniowej, to duża liczba występujących w takim zagadnieniu parametrów może grozić brakiem możliwości efektywnego ustalenia ich wartości. Odpowiednia redukcja liczności zbioru rozmytych reguł (3) staje się zatem podstawowym problemem, zwłaszcza w przypadku złożonych aplikacyjnych przypadków. Do rozwiązania zadania redukcji rozmytych reguł używa sie wielu metod współczesnych technik informacyjnych [Kulczycki *et al*; 2007], przede wszystkim algorytmów ewolucyjnych [Arabas; 2001], systemów neuronowo-rozmytych [Rutkowski; 2005], czy też ujęcia statystycznego [Hryniewicz; 2004]. Przedstawiony w niniejszej publikacji Kompletny Gradientowy Algorytm Klasteryzacji, został z powodzeniem zastosowany do tego celu.

Niech zatem dany będzie wektor $\begin{bmatrix} x \\ y \end{bmatrix}$ oraz m pomiarów jego wartości uzyskanych podczas działania systemu sterowania z użyciem rozmytego regulatora PID w pierwotnej postaci, czyli bez redukcji zbioru reguł:

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}, \ldots, \begin{bmatrix} x_m \\ y_m \end{bmatrix} \ . \tag{5}$$

Traktując powyższy zbiór jako próbę losową (1) dokonać można klasteryzacji z użyciem Kompletnego Gradientowego Algorytmu Klasteryzacji. Niech

$$\begin{bmatrix} \tilde{x}_1 \\ \tilde{y}_1 \end{bmatrix}, \begin{bmatrix} \tilde{x}_2 \\ \tilde{y}_2 \end{bmatrix}, \ldots, \begin{bmatrix} \tilde{x}_{\tilde{c}} \\ \tilde{y}_{\tilde{c}} \end{bmatrix} \tag{6}$$

reprezentują centra otrzymanych w wyniku jego działania $\tilde{c}$ klastrów. Każdy element $\tilde{x}_i$ dla $i = 1,2,\ldots,\tilde{c}$ może stanowić podstawę i-tej rozmytej reguły z odpowiadającą mu funkcją przynależności

$$\mu_i(x) = \exp\left(-\left\| \frac{x - \tilde{x}_i}{d} \right\|^2 \right) , \tag{7}$$

gdzie parametr $d > 0$ charakteryzuje zdolność uogólniania wynikłą z wnioskowania rozmytego w ramach projektowanego systemu sterowania. Przeprowadzone badania eksperymentalne wskazują, iż jako standardową można użyć wartość $d = \tilde{c} / 2$. W konsekwencji wzór (4) przyjmuje postać

$$y = \frac{\sum_{i=1}^{\tilde{c}} \mu_i(x)\, f_i(x)}{\sum_{i=1}^{\tilde{c}} \mu_i(x)} , \tag{8}$$

przy czym f_i są funkcjami liniowymi, których parametry mogą być wyznaczone na podstawie klasycznego średniokwadratowego zadania estymacji.

Powyższa metoda została pozytywnie zweryfikowana w kilku praktycznych zagadnieniach. Poniżej zostaną podane porównawcze wyniki otrzymane dla układu sterowania serwomotorem dysku twardego, przedstawionego w artykule [Tan *et al*; 2007]. Użyto tam następującego jego modelu:

$$\begin{bmatrix} \dot{s}(t) \\ \dot{v}(t) \end{bmatrix} = \begin{bmatrix} 1 & 1.664 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} s(t) \\ v(t) \end{bmatrix} + \begin{bmatrix} 1.384 \\ 1.664 \end{bmatrix} u(t) , \tag{9}$$

gdzie u stanowi wejście elementu wykonawczego (w woltach), natomiast s oraz v oznaczają pozycję (w ścieżkach) i prędkość głowicy twardego dysku. Analizowany był problem precyzyjnego pozycjonowania, przy czym wyjście stanowiła wielkość $s(t)$. Jako typowy w takich zastosowaniach, rozważany był regulator typu PD [Mudi, Pal; 1999].

Najpierw określony został standardowy regulator rozmyty PD z 49 regułami uwarunkowanymi na szybką odpowiedź na skok jednostkowy. Otrzymany w ten sposób 121-elementowy zbiór (5):

$$\begin{bmatrix} e_1 \\ \dot{e}_1 \\ u_1 \end{bmatrix}, \begin{bmatrix} e_2 \\ \dot{e}_2 \\ u_2 \end{bmatrix}, \ldots, \begin{bmatrix} e_{121} \\ \dot{e}_{121} \\ u_{121} \end{bmatrix} , \tag{10}$$

gdzie e reprezentuje uchyb, został potraktowany jako próba losowa (1) i poddany działaniu

Kompletnego Gradientowego Algorytmu Klasteryzacji. W wyniku otrzymano rozmyty regulator PD z bazą zredukowaną do 38 reguł.

W celu porównania przebiegów generowanych przez klasyczny regulator PD, rozmyty regulator PD z niezredukowaną 49-elementową bazą reguł [Mudi, Pal; 1999] oraz otrzymany z użyciem Kompletnego Gradientowego Algorytmu Klasteryzacji regulator rozmyty z bazą zredukowaną do 38 reguł, otrzymano – w odpowiedzi na skok jednostkowy – wartość pierwiastka z kwadratowego wskaźnika jakości, odpowiednio, 0,291, 0,198, 0,111 oraz procentową wartość przeregulowania 78%, 92%, 15%. W obu przypadkach zdecydowanie najlepsze wyniki otrzymano z zastosowaniem rozmytego regulatora PD ze zredukowaną bazą reguł. Podobne rezultaty uzyskiwano dla innych uwarunkowań i wskaźników jakości.

Dodatkowe badania wykonano dla systemu poddanego działaniu regulatora rozmytego z bazą reguł redukowaną przy zastosowaniu Kompletnego Gradientowego Algorytmu Klasteryzacji, dla różnych – odmiennych od otrzymanych z użyciem kryteriów optymalizacyjnych – wartości parametrów h oraz c. Otóż, najkorzystniejsze wyniki uzyskiwano przy wartości drugiego z nich dwukrotnie zmniejszonej względem optymalnej. Efekt taki można zinterpretować powiększeniem liczby klastrów peryferyjnych charakteryzujących stany nietypowe, „niebezpieczne” z punktu widzenia poprawności działania systemu. Co więcej, główny klaster zawierał przeważnie nawet 80% elementów zbioru (5), reprezentujących „bezpieczne” stany typowe i jego ewentualne rozdzielenie nie skutkowało pozytywnymi zmianami. Podobnie jak w zagadnieniu przedstawionym w poprzedniej sekcji nie było zatem potrzeby zmieniać – względem optymalnej – wartości parametru wygładzania h. Wskazuje to ponownie na dobre samoistne dopasowywanie się Kompletnego Gradientowego Algorytmu Klasteryzacji do rzeczywistej struktury danych.

Szczegółowe rozważania w tym zakresie znaleźć można w publikacjach [Kowalski *et al*; 2008; Łukasik *et al*; 2008].

4. PODSUMOWANIE

Przedstawione w niniejszej publikacji wyniki uzyskane w trakcie stosowania Kompletnego Gradientowego Algorytmu Klasteryzacji [Kulczycki, Charytanowicz; 2010] potwierdziły jego praktyczną użyteczność, w szczególności wymienione we wstępie 6 podstawowych cech. Zwraca uwagę brak zasadności istotnych zmian wartości parametru h, stanowiącego bezpośrednio o liczbie uzyskiwanych klastrów, co wskazuje na prawidłowe dostosowywane się procedury do rzeczywistej struktury danych. Bardzo cenna w praktyce okazała się możliwość zmiany wartości parametru c, stanowiącego o relacji ilości klastrów w obszarach zagęszczenia elementów próby wobec rejonów rzadkiego ich występowania. We wszystkich trzech przeprowadzonych badaniach zmiana taka umożliwiła uzyskanie istotnie korzystniejszych – z aplikacyjnego punktu widzenia – wyników. Jest to szczególnie warte podkreślenia, gdyż możliwość kształtowania powyższej relacji nie występuje w innych znanych algorytmach

klasteryzacji.

LITERATURA

Arabas J. (2001) Wykłady z algorytmów ewolucyjnych, WNT, Warszawa.

Barnett V., Lewis T. (1994). *Outliers in Statistical Data*, Wiley, Chichester.

Charytanowicz M., Niewczas J., Kulczycki P., Kowalski P.A., Łukasik S., Żak S.. (2010). Gradient Clustering Algorithm for Features Analysis of X-Ray Images; w: E. Pietka, J. Kawa (red.) *Information Technologies in Biomedicine*, vol. 2, ss. 15-24, Springer-Verlag, Berlin.

Fodor J., Roubens M. (1994). *Fuzzy Preference Modelling and Multicriteria Decision Support*, Kluwer, Dordrecht.

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*, vol. 21, ss. 32-40.

Hryniewicz O. (2004). *Wykłady ze statystyki dla studentów informatycznych technik zarządzania*, Wydawnictwo WSISiZ, Warszawa.

Kacprzyk J. (1986). *Zbiory rozmyte w analizie systemowej*, PWN, Warszawa.

Kowalski P.A., Łukasik S., Charytanowicz M., Kulczycki P. (2008). Data-Driven Fuzzy Modeling and Control with Kernel Density Based Clustering Technique, *Polish Journal of Environmental Studies*, vol. 17, ss. 83-87.

Krzyśko M., Wołyński W., Górecki T., Skorzybut M. (2008). *Systemy uczące się – rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*, WNT, Warszawa.

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

Kulczycki P., Hryniewicz O., Kacprzyk J. (red.) (2007). *Techniki informacyjne w badaniach systemowych*, WNT, Warszawa.

Kulczycki P., Charytanowicz M. (2008). Kompletny algorytm gradientowej klasteryzacji, XVI Krajowa Konferencja Automatyki, Szczyrk, 11-15 maja 2008. Post-proceedings: „Sterowanie i automatyzacja: aktualne problemy i ich rozwiązania", Malinowski K., Rutkowski L. (red.), ss. 312-321, EXIT, Warszawa, 2008.

Kulczycki P., Charytanowicz M. (2010). A Complete Gradient Clustering Algorithm Formed with Kernel Estimators, *International Journal of Applied Mathematics and Computer Science*, vol. 20, ss. 123-134.

Kulczycki P., Charytanowicz M., Kowalski P.A., Łukasik S. (2012). The Complete Gradient Clustering Algorithm: Properties in Practical Applications, *Journal of Applied Statistics*, w trakcie publikowania.

Kulczycki P., Daniel K. (2009). Metoda wspomagania strategii marketingowej operatora telefonii komorkowej, *Przegląd Statystyczny*, vol. 56, ss. 116-134.

Kulczycki P., Kowalski P.A. (2011). Bayes Classification of Imprecise Information of Interval Type,

Control and Cybernetics, vol. 40, ss. 101-123.

Kulczycki P., Łukasik S. (2012) An Algorithm for Reducing Dimension and Size of Sample for Data Exploration Procedures, w trakcie publikowania.

Kulczycki P., Prochot C. (2004). Wykrywanie elementów odosobnionych za pomocą metod estymacji nieparametrycznej; w: Kulikowski R., Kacprzyk J., Słowinski R. (red.) *Badania operacyjne i systemowe: podejmowanie decyzji – podstawy teoretyczne i zastosowania*, EXIT, Warszawa, ss. 313-328.

Łukasik S., Kowalski P.A., Charytanowicz M., Kulczycki P. (2008). Fuzzy Models Synthesis with Kernel-Density Based Clustering Algorithm, *Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, Ma J., Yin Y., Yu J., Zhou S. (red.), Jinan (Chiny), 18-20 października 2008, vol. 3, ss. 449-453.

Mudi R., Pal N. R. (1999). A Robust Self-Tuning Scheme for PI and PD Type Fuzzy Controllers, *IEEE Transactions on Fuzzy Systems*, vol. 7, ss. 2-16.

Rutkowski L. (2005) Metody i techniki sztucznej inteligencji, PWN, Warszawa.

Tan K.C., Sathikannan R., Tan W.W., Loh A.P. (2007). Evolutionary Design and Implementation of a Hard Disk Drive Servo Control System, *Soft Computing*, vol. 11, ss. 131-139.

Yager R.R., Filev D.P. (1995). *Podstawy modelowania i sterowania rozmytego*, WNT, Warszawa.