Dr inż. Szymon Łukasik
Instytut Badań Systemowych
Polska Akademia Nauk
ul. Newelska 6, 01-447 Warszawa

Warszawa, 26 kwietnia 2019
Załącznik 3a

AUTOREFERAT

przedstawiający opis dorobku i osiągnięć naukowych w związku z ubieganiem się o nadanie stopnia doktora habilitowanego

1 Dane osobowe

Imię i nazwisko: Szymon Łukasik

Adres służbowy: Instytut Badań Systemowych
Polska Akademia Nauk
ul. Newelska 6
01-447 Warszawa
e-mail: slukasik@ibspan.waw.pl

oraz

Wydział Fizyki i Informatyki Stosowanej
Akademia Górniczo-Hutnicza
al. Mickiewicza 30
30-059 Kraków
e-mail: slukasik@agh.edu.pl

2 Posiadane dyplomy i stopnie naukowe

9.03.2012 — Instytut Badań Systemowych Polskiej Akademii Nauk
dr inż.; dziedzina: **Nauki techniczne**, dyscyplina: ***Informatyka***
Tytuł rozprawy: *Algorytm redukcji wymiaru i liczności próby dla celów procedur eksploracyjnej analizy danych* (z wyróżnieniem)

22.03.2006 — Wydział Inżynierii Elektrycznej i Komputerowej Politechniki Krakowskiej
mgr inż.; specjalność: ***Automatyka***
Tytuł pracy: *Identyfikacja rozkładu w systemach rzeczywistych za pomocą estymatorów jądrowych*

22.06.2005 — Wydział Inżynierii Elektrycznej i Komputerowej Politechniki Krakowskiej
mgr inż.; specjalność: ***Inżynieria teleinformatyczna***
Tytuł pracy: *Oprogramowanie testera wielokanałowych zasilaczy wysokiego napięcia dla detektorów krzemowych eksperymentu ATLAS w CERNie* (z wyróżnieniem)

3 Informacje o dotychczasowym zatrudnieniu w jednostkach naukowych

01.10.2014 – Akademia Górniczo-Hutnicza
Wydział Fizyki i Informatyki Stosowanej
adiunkt

01.07.2012 – Instytut Badań Systemowych Polskiej Akademii Nauk

	Pracownia Metod Statystycznych obecnie Centrum Informatycznych Metod Analizy Danych adiunkt
1.11.2012 – – 30.09.2014	Politechnika Krakowska Wydział Inżynierii Elektrycznej i Komputerowej adiunkt
1.11.2007 – – 31.06.2012	Instytut Badań Systemowych Polskiej Akademii Nauk Pracownia Metod Statystycznych asystent
1.11.2005 – – 31.10.2012	Politechnika Krakowska Wydział Inżynierii Elektrycznej i Komputerowej asystent
1.10.2004 – – 30.06.2005	Politechnika Krakowska Wydział Inżynierii Elektrycznej i Komputerowej asystent-stażysta

4 Podstawowe osiągnięcie naukowe

Jako osiągnięcie naukowe wynikające z art. 16 ust. 2 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz stopniach i tytule w zakresie sztuki (Dz. U. 2003 nr 65 poz. 595 z późn. zm.) wskazuję:

cykl publikacji, powiązanych tematycznie, pod wspólnym tytułem

Metody nienadzorowanego przetwarzania i wydobywania wiedzy ze zbiorów danych o znacznych rozmiarach

4.1 Publikacje wchodzące w skład osiągnięcia

Sumaryczny IF ośmiu publikacji przedstawionych w ramach cyklu według bazy JCR wynosi **9,846**. Sumaryczna liczba punktów według MNiSW, zgodnie z obowiązującym w roku wydania wykazem czasopism naukowych, jest równa **175**[1].

W skład wskazanego cyklu publikacji wchodzą następujące prace:

[H.1] S. Łukasik, A. Moitinho, P. A. Kowalski, A. Falcão, R. A. Ribeiro i P. Kulczycki. "Survey of Object-Based Data Reduction Techniques in Observational Astronomy". *Open Physics* vol. 14. nr 1 (2016), s. 579–586. DOI: 10.1515/phys-2016-0064.

JCR, Web of Science, IF: 0,745, MNiSW: 15 pkt.

Mój wkład w powstanie tej publikacji szacuję na **20%**, obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowywanie jej koncepcji i układu oraz sformułowanie celów badań; b) projekt i implementację algorytmów oraz wyprowadzenie wyników; c) analizę i interpretację wyników d) przygotowanie publikacji do druku.

[H.2] P. Kulczycki i S. Łukasik. "An algorithm for reducing the dimension and size of a sample for data exploration procedures". *International Journal of Applied Mathematics and Computer Science* vol. 24. nr 1 (2014), s. 133–149. DOI: 10.2478/amcs-2014-0011.

JCR, Web of Science, IF: 1,227, MNiSW: 25 pkt.

Mój wkład w powstanie tej publikacji szacuję na **50%**, obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowywanie jej koncepcji i układu oraz sformułowanie celów badań; b) projekt i implementację algorytmów oraz wyprowadzenie wyników; c) analizę i interpretację wyników d) przygotowanie publikacji do druku.

[1] W przypadku publikacji wydanych po roku 2017 przyjęto punktację z ostatniego ujednoliconego wykazu czasopism naukowych.

[H.3] S. Łukasik i P. Kulczycki. "Using Topology Preservation Measures for Multidimensional Intelligent Data Analysis in the Reduced Feature Space". *Artificial Intelligence and Soft Computing*. Red. L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh i J. M. Zurada. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, s. 184–193. DOI: 10.1007/978-3-642-38610-7_18.

Web of Science, MNiSW: 10 pkt.

Mój wkład w powstanie tej publikacji szacuję na **50%**, obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowywanie jej koncepcji i układu oraz sformułowanie celów badań; b) projekt i implementację algorytmów oraz wyprowadzenie wyników; c) analizę i interpretację wyników d) przygotowanie publikacji do druku.

[H.4] D. Domańska i S. Łukasik. "Handling high-dimensional data in air pollution forecasting tasks". *Ecological Informatics* vol. 34 (2016), s. 70–91. DOI: 10.1016/j.ecoinf.2016.04.007.

JCR, Web of Science, IF: 2,020, MNiSW: 25 pkt.

Mój wkład w powstanie tej publikacji szacuję na **50%**, obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowywanie jej koncepcji i układu oraz sformułowanie celów badań; b) projekt i implementację algorytmów oraz wyprowadzenie wyników; c) analizę i interpretację wyników d) przygotowanie publikacji do druku.

[H.5] P. Kulczycki, M. Charytanowicz, P. A. Kowalski i S. Łukasik. "Identification of Atypical (Rare) Elements – A Conditional, Distribution-Free Approach". *IMA Journal of Mathematical Control and Information* vol. 35 (2017), s. 923–937. DOI: 10.1093/imamci/dnx007.

JCR, Web of Science, IF: 1,358, MNiSW: 25 pkt.

Mój wkład w powstanie tej publikacji szacuję na **25%**, obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowywanie jej koncepcji i układu oraz sformułowanie celów badań; b) projekt i implementację algorytmów oraz wyprowadzenie wyników; c) analizę i interpretację wyników d) przygotowanie publikacji do druku.

[H.6] M. Charytanowicz, P. Kulczycki, P. A. Kowalski, S. Łukasik i R. Czabak-Garbacz. "An Evaluation of Utilizing Geometric Features for Wheat Grain Classification using X-ray Images". *Computers and Electronics in Agriculture* vol. 144 (2018), s. 266–268. DOI: 10.1016/j.compag.2017.12.004.

JCR, Web of Science, IF: 2,427, MNiSW: 40 pkt.

Mój wkład w powstanie tej publikacji szacuję na **20%**, obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowywanie jej koncepcji i układu oraz sformułowanie celów badań; b) projekt i implementację algorytmów oraz wyprowadzenie wyników; c) analizę i interpretację wyników d) przygotowanie publikacji do druku.

[H.7] P. Kulczycki, M. Charytanowicz, P. A. Kowalski i S. Łukasik. "The complete gradient clustering algorithm: properties in practical applications". *Journal of Applied Statistics* vol. 39. nr 6 (2012), s. 1211–1224. DOI: 10.1080/02664763.2011.644526.

JCR, Web of Science, IF: 0,449, MNiSW: 15 pkt.

Mój wkład w powstanie tej publikacji szacuję na **25%**, obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowywanie jej koncepcji i układu oraz sformułowanie celów badań; b) projekt i implementację algorytmów oraz wyprowadzenie wyników; c) analizę i interpretację wyników d) przygotowanie publikacji do druku.

[H.8] P. A. Kowalski i S. Łukasik. "Training neural networks with krill herd algorithm". *Neural Processing Letters* vol. 44. nr 1 (2016), s. 5–17. DOI: 10.1007/s11063-015-9463-0.

JCR, Web of Science, IF: 1,620, MNiSW: 20 pkt.

Mój wkład w powstanie tej publikacji szacuję na **50%**, obejmował on wszystkie etapy przygotowywania pracy, w szczególności: a) opracowywanie jej koncepcji i układu oraz sformułowanie celów badań; b) projekt i implementację algorytmów oraz wyprowadzenie wyników; c) analizę i interpretację wyników d) przygotowanie publikacji do druku.

4.2 Ogólne przedstawienie tematyki i głównego celu prac

Niniejsza część stanowi omówienie celu naukowego i wyników publikacji wchodzących w skład osiągnięcia wynikającego z art. 16 ust. 2 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz stopniach i tytule w zakresie sztuki (*Dz. U. 2003 nr 65 poz. 595 ze zm.*).

Przedmiotem prowadzonych badań, których rezultaty zostały przedstawione w postaci cyklu publikacji powiązanych tematycznie, jest zaprojektowanie klasy algorytmów mających na celu przetwarzanie i wydobywanie wiedzy ze zbiorów danych o znacznym rozmiarze. Szczególny nacisk w tym zakresie położono na procedury o charakterze nienadzorowanym, tj. zakładające brak dodatkowych informacji dotyczących przynależności poszczególnych elementów zbioru do wyróżnionych klas/kategorii i opierające się jedynie o dane, bez opisujących je etykiet [9].

Zagadnienia badawcze, rozważane przeze mnie po uzyskaniu stopnia naukowego doktora, dotyczyły w szczególności algorytmów kondensacji danych, redukcji ich wymiarowości, a także wykrywania elementów odosobnionych. Ponadto przedmiotem moich prac były również zagadnienia analizy skupień. W wymienionych obszarach wykorzystany został aparat statystyki nieparametrycznej, a także inspirowane naturą metody heurystycznej optymalizacji. W zakresie tych ostatnich prowadzono również szeroko zakrojone prace dotyczące własności wybranych algorytmów i ich zastosowania w optymalizacji ciągłej – wykraczające częściowo poza zakres tematyczny omawianego cyklu i będące przedmiotem zawartego w części 5 omówienia pozostałych osiągnięć naukowych. Przedstawiona tu tematyka badawcza wpisuje się we wszechobecne zapotrzebowanie na efektywne algorytmy, pozwalające na przetwarzanie i wydobywanie wiedzy ze zbiorów typu Big Data. Zasadniczymi cechami tych zbiorów, poza znacznymi rozmiarami, są również zmienność w czasie i różnorodność typów zgromadzonych danych [10].

Wśród rozważanych zagadnień pierwszą grupę tematów stanowiły te związane z nienadzorowaną redukcją liczności i wymiarowości danych. W tym zakresie prowadzono zarówno prace nad nowymi algorytmami jak i zastosowaniami istniejących technik w obszarach nauki i techniki, w szczególności w astronomii i inżynierii środowiska. W ramach badań podstawowych w tej tematyce opracowano i szczegółowo przebadano autorski algorytm redukcji wymiaru i liczności zbioru oparty o metodologię symulowanego wyżarzania. Ta oparta o liniową transformację technika – poza uzyskaniem zbioru o zredukowanej liczbie cech – umożliwia poprawę wyników uzyskiwanych przez procedury eksploracji danych w przestrzeni zredukowanej, poprzez eliminację lub przypisanie mniejszej wagi elementom które w wyniku transformacji zmieniają istotnie swe położenie. Koncepcja ta została twórczo rozwinięta i uogólniona do zastosowań dla dowolnej techniki redukcji wymiarowości. W toku dalszych badań realizowanych w tym nurcie zaproponowano użycie, dla celów kondensacji danych astronomicznych, algorytmu opartego o analizę lokalnej gęstości wyznaczonej dla każdej z obserwacji. Wynik ten jest o tyle wartościowy, że dla danych tego typu, reprezentujących przestrzennie rozmieszczone obiekty astronomiczne, stosowano dotąd – z dużo gorszym skutkiem – jedynie metody próbkowania losowego. Pracę tą poprzedził dogłębny przegląd literatury przedmiotu w zakresie metod redukcji liczności i wymiarowości danych w odniesieniu do danych reprezentujących obiekty obserwacji astronomicznych. Tutaj również poruszano się w obszarze dotąd nieeksplorowanym, jako że istniejące prace w tej tematyce dotyczą głównie redukcji danych na poziomie ich akwizycji, w toku przetwarzania sygnałów pochodzących z poszczególnych elementów instrumentarium astronomicznego. Obraz prac z zakresu redukcji wymiaru danych kompletuje przegląd nienadzorowanych metod redukcji wymiarowości zrealizowany w odniesieniu do zagadnienia prognozowania poziomów zanieczyszczeń powietrza. Prace badawcze w tej tematyce obejmowały – poza studiami literaturowymi – eksperymenty dla 16 technik redukcji wymiarowości, ze szczególnym uwzględnieniem metod pozwalających na uogólnienie wyników redukcji dla nowych, pojawiających się dynamicznie porcji danych, charakterystycznych dla zbiorów typu Big Data. Zaproponowano też, i pozytywnie zweryfikowano, koncepcję użycia dla celów prognozowania odległości ułamkowych – jako alternatywę do redukcji wymiaru zbioru. Takie porównawcze zestawienie tych dwóch strategii stanowiło również nowatorski wkład w dziedzinę eksploracyjnej analizy danych.

Drugi obszar tematyczny realizowanych prac dotyczył najczęściej spotykanych problemów nienadzorowanej eksploracji danych – wykrywania elementów odosobnionych i analizy skupień. W tym zakresie kontynuowano i zakończono pracę nad dwoma autorskimi technikami opartymi o nieparametryczną metodykę estymatorów jądrowych. Pierwsza z nich – dedykowana dla wykrywania elementów nietypowych – opiera się o jądrowy estymator kwantyla, a u jej fundamentów leży naturalna interpretacja wartości estymatora funkcji gęstości rozkładu prawdopodobieństwa wyznaczanego dla badanego elementu. Im mniejsza jest wartość tego estymatora, tym ów element można interpretować jako "mniej typowy", a ściślej: rzadziej występujący. Koncepcja ta w toku przeprowadzonych prac została poszerzona o przypadek warunkowy – w praktyce często występujący dla współczesnych zbiorów danych. Z kolei opracowany algorytm analizy skupień opiera się o wstępną relokację elementów zbioru w kierunku gradientu funkcji gęstości oraz realizowane w drugim kroku przypisanie elementów do skupień. W tej drugiej fazie pomocniczą rolę odgrywa – oszacowana na podstawie próby odległości między poszczególnymi elementami zbioru – połowa odległości między środkami dwóch skupień, leżących najbliżej siebie. Technika ta pozwala na uzyskanie podziału zbioru na skupienia bez arbitralnych założeń dotyczących ich liczby. Ostatnim – zorientowanym na praktyczne aspekty analizy danych – tematem badań w rozważanym

nurcie była analiza skupień (klasteryzacja) zbioru reprezentującego pomiary rentgenowskie ziaren pszenicy. W ramach prac badawczych – poza samym podziałem zbioru na skupienia – szczególną uwagę poświęcono reprezentatywności geometrycznych cech ziaren i ich dyskryminacyjnej istotności. O znacznej wartości uzyskanych wyników, a także interesującym charakterze opracowywanych danych, może świadczyć fakt dużej popularności publicznie udostępnionego w ramach tych studiów zbioru *seeds*[11]. Poza powszechnym użyciem w pracach naukowych dotyczących redukcji wymiaru i analizy skupień, stanowi on również ilustrację tych problemów w podręcznikach akademickich [12].

Trzecim, pomocniczym dla rozważanych wyżej zagadnień, obszarem tematycznym realizowanych prac badawczych była problematyka zastosowań technik inspirowanych naturą w analizie danych. W tym zakresie opracowano nowy wariant równoległego algorytmu szybkiego symulowanego wyżarzania wykorzystujący przetwarzanie wielowątkowe. Dzięki zastosowaniu automatycznego kryterium zatrzymania pracy algorytmu opartego o statystykę porządkową oraz adaptacyjny dobór temperatury wyżarzania algorytm zyskał pożądaną w praktycznych zastosowaniach wygodę użytkowania, związaną z niewielką liczba parametrów. Wzmiankowany algorytm został z sukcesem użyty w problemie nienadzorowanej redukcji wymiaru – opisywanym na początku tego podrozdziału. W tym nurcie ulokować można też prace nad rozwojem algorytmu kryla (ang.Krill Herd Algorithm). Prace w tym zakresie – choć zasadniczo poświęcone zastosowaniom algorytmu w nadzorowanym problemie klasyfikacji – prowadziły w zamierzeniu autorów, w kierunku ustalenia właściwych wartości parametrów czy też istotnych cech aplikacyjnych algorytmu, niezależnych od rozważanego problemu optymalizacji. O ich użyteczności może świadczyć fakt, że rzeczony algorytm znalazł również zastosowanie w problemie klasteryzacji, uzyskując dla zbiorów testowych bardzo dobre wyniki.

W publikacjach podsumowujących przeprowadzone prace badawcze starano się szczegółowo zarysować rozważany problem, jak i opisać zaproponowane rozwiązanie w taki sposób, by możliwe było jego jak najwierniejsze odtworzenie. Dla opracowanych procedur przeprowadzano badania dotyczące rekomendowanych wartości występujących w nich parametrów. Oprócz artykułów wymienionych w przedstawionym tu cyklu publikacji, część cząstkowych wyników realizowanych prac była przedmiotem wystąpień na renomowanych konferencjach naukowych wzbudzając duże zainteresowanie i merytoryczne dyskusje.

W ramach omawianych tu zagadnień badawczych przeprowadzono wnikliwe badania eksperymentalne z użyciem danych rzeczywistych – pozyskanych, bądź to z ogólnodostępnych repozytoriów, bądź od ekspertów dziedzinowych – a także zbiorów o charakterze losowym. Punktem wyjścia do porównań z istniejącym stanem wiedzy były metody o podobnych cechach aplikacyjnych, których użycie w przeważającej części przypadków dawało gorsze wyniki, zarówno pod kątem jakości uzyskiwanych rozwiązań jak i szybkości działania algorytmu.

W moim odczuciu, badania wymienione w ramach cyklu publikacji, stanowią twórczy wkład w dziedzinę wielowymiarowej analizy danych, poszerzając wachlarz dostępnych technik w najważniejszych obszarach nienadzorowanego przetwarzania i wydobywania wiedzy z danych. W ich efekcie zaprojektowano, zaimplementowano i przebadano kilka nowych algorytmów służących do rozwiązywania rzeczywistych problemów występujących między innymi w naukach technicznych, ekonomicznych czy biomedycznych. Istotnym wkładem omawianych tu prac jest również usystematyzowanie wiedzy nt. istniejących technik i ich wnikliwa eksperymentalna weryfikacja.

Do najważniejszych wyników uzyskanych w trakcie prowadzonych badań stanowiących oryginalny dorobek naukowy - a zarazem zasadniczą część osiągnięcia naukowego - zaliczam:

1. koncepcję kondensacji obserwacji astronomicznych z użyciem algorytmu opartego o gęstość danych;
2. ideę nienadzorowanej liniowej procedury redukcji wymiaru opartą o zachowanie odległości i optymalizację heurystyczną;
3. koncepcję wag – uwzględniających stopień względnego przesunięcia poszczególnych elementów zbioru w wyniku redukcji wymiaru – oraz ich użycia w procedurach analizy danych.
4. wyniki badań porównawczych nad nienadzorowanymi technikami redukcji wymiaru w problemach prognozowania zanieczyszczeń i klasyfikacji ziaren oraz ideę użycia w pierwszym z wymienionych zagadnień odległości ułamkowych;
5. opracowanie nowego wariantu algorytmu symulowanego wyżarzania o bardzo wysokiej wartości aplikacyjnej – także w problemach analizy danych;
6. sfinalizowanie prac nad kompletnymi algorytmami analizy skupień i wykrywania elementów odosobnionych opartych o nieparametryczną estymację funkcji gęstości rozkładu prawdopodobieństwa.

4.3 Opis rozważanych problemów badawczych i uzyskanych wyników

Moje osiągnięcie naukowe stanowi cykl publikacji powiązanych tematycznie pt. "Metody nienadzorowanego przetwarzania i wydobywania wiedzy ze zbiorów danych o znacznych rozmiarach".

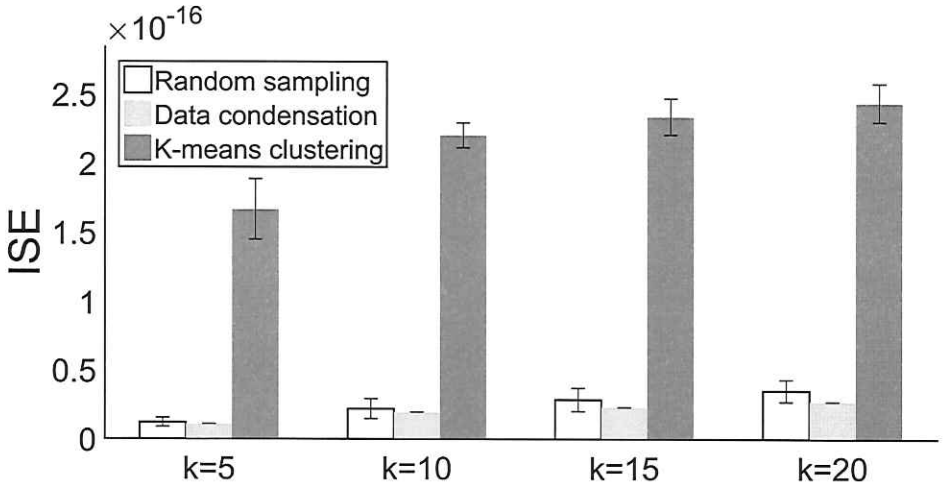
W niniejszym podsumowaniu prac badawczych w pierwszej kolejności omówiony zostanie artykuł [H.1] zatytułowany "Survey of Object-Based Data Reduction Techniques in Observational Astronomy". W pracy tej podjęto się analizy dostępnych metod redukcji rozmiaru danych, w odniesieniu do danych astronomicznych. Przy czym jako dane, rozumiane są nie sygnały niskiego poziomu (jak to zwykle przyjmuje się w literaturze astronomicznej) a konkretne obiekty o zadanych parametrach fizycznych i lokalizacji. Redukcja danych w tym przypadku ma dwa aspekty. Pierwszy jest związany ze zmniejszeniem liczności zbioru. Wprowadza się ją przede wszystkim by uzyskać kompaktową jego reprezentację dla celów wizualizacji. W tym celu wykorzystuje się przede wszystkim dane dotyczące rozlokowania przestrzennego badanych obiektów. Drugim zagadnieniem jest redukcja wymiarowości. Dotyczy ona atrybutów opisujących cechy fizyczne obiektów i może prowadzić do ich kategoryzacji.

W zakresie pierwszej strategii stosuje się głównie próbkowanie losowe, względnie algorytm klasteryzacyjny. W omawianej tu pracy zaproponowano użycie zmodyfikowanego algorytmu kondensacji danych opisanego w pracy [13]. Opiera się on o iteracyjną eliminację (tzw. pruning) tych elementów które znajdują się w otoczeniu punktu o największej gęstości – określanej przez odległość do k-tego najbliższego sąsiada. Parametr k określa przy tym pośrednio intensywność redukcji (zwiększenie k skutkuje uzyskaniem bardziej skondensowanej reprezentacji zbioru pierwotnego). Dzięki zastosowaniu drzew k-wymiarowych (ang. kd-tree) i przetwarzaniu równoległemu zastosowanie tej złożonej obliczeniowo strategii staje się możliwe, nawet dla zbiorów danych o znacznych rozmiarach. Oszacowano, że zrealizowanie kondensacji danych dla zbioru składającego się z 1 mln elementów na średniej klasy komputerze domowym (bez wykorzystania przetwarzania z użyciem GPU) trwa 61 godzin. Jest to czas akceptowalny biorąc pod uwagę konfigurację sprzętową i założenie dotyczące jednokrotnego przeprowadzenie tej procedury.

W pracy zweryfikowano także jakość rozwiązań generowanych przez zaproponowany algorytm. Do jej oceny użyto błędu wyznaczanego przez ISE (ang. Integrated Square Error) zdefiniowanego jako

$$ISE(\hat{f}(x)) = \sum_{i=1}^{m} (\hat{f}(x_i) - f(x_i))^2 \tag{1}$$

przy czym $f(x_i)$ reprezentuje wartość funkcji gęstości prawdopodobieństwa oszacowaną dla elementu x_i należącego pierwotnego zbioru o liczności m, z kolei $\hat{f}(x)$ określa tą samą gęstość obliczoną dla zbioru o zredukowanej liczności. Jako że po przeprowadzonej kondensacji część elementów zbioru reprezentuje sąsiadów z pierwotnej próby mu najbliższych, każdemu takiemu reprezentantowi przypisano wagę w_i równą liczbie elementów które on symbolizuje. Rozkłady zostały oszacowane z użyciem metodyki statystycznych estymatorów jądrowych. Porównanie uzyskanych wyników przeprowadzono w odniesieniu do próbkowania losowego i algorytmu k-średnich. Wyniki 30 powtórzeń procedury redukcji dla zbioru Hipparcos – jednego z powszechnie dostępnych katalogów obiektów astronomicznych [14] – ilustruje Rysunek 1. Można zauważyć, że zaproponowana metoda oferuje najlepsze odwzorowanie pierwotnej gęstości zbioru. Co więcej przewaga tego podejścia rośnie wraz ze wzrostem intensywności redukcji. W przypadku $k = 20$ otrzymane wyniki w 27 na 30 przypadków są lepsze niż dla próbkowania losowego. Godzien podkreślenia jest również deterministyczny charakter zaproponowanej tu procedury.



Rysunek 1: Wartości ISE obliczone dla oszacowań gęstości zredukowanego zbioru Hipparcos

W aspekcie redukcji wymiarowości zbadano możliwość użycia algorytmu t-Distributed Stochastic Neighbourhood Embedding (t-SNE) wprowadzonego przez Hintona i Roweisa [15]. Głównym problemem związanym z

zastosowaniem tej procedury dla celów przetwarzania danych astronomicznych jest spodziewany znaczny czas wymaganych obliczeń. W omawianej pracy zweryfikowano możliwość jej użycia – z użyciem zbioru Hipparcos. Zastosowano przy tym wariant Barnes-Hut tej procedury [16]. Stwierdzono, że złożoność obliczeniowa wynosi $O(m \log m)$ – co w odniesieniu do rzeczywistych danych astronomicznych jest zdecydowanie akceptowalne. Wspomniane tu prace były realizowane w ramach współpracy międzynarodowej z zespołem badawczym z centrum badawczego UNINOVA (Portugalia) biorącym udział w misji GAIA Europejskiej Agencji Kosmicznej [17]. Warto przy tym nadmienić że dalsze poszerzenie omawianej tu tematyki badawczej przedstawiono w przyjętym do druku artykule:

[18] S. Łukasik, K. Lalik, P. Sarna, P. A. Kowalski, M. Charytanowicz i P. Kulczycki. "Efficient Astronomical Data Condensation using Approximate Nearest Neighbors". *International Journal of Applied Mathematics and Computer Science* (2019). [Oczekujący na publikację].

JCR, Web of Science, IF: 1,694, MNiSW: 25 pkt.

Zawiera on wyniki działania procedury redukcji dla danych astronomicznych z repozytorium GAIA, z uwzględnieniem zrównoleglenia obliczeń i użyciem przybliżonego algorytmu k-najbliższych sąsiadów. Tekst tego artykułu włączono do zestawienia pozostałych osiągnięć naukowych.

Kolejnymi pracami które zostaną w tym miejscu omówiona są artykuły [H.2] pt. "An algorithm for reducing the dimension and size of a sample for data exploration procedures" oraz [H.3] pt. "Using Topology Preservation Measures for Multidimensional Intelligent Data Analysis in the Reduced Feature Space". Są one związane z zagadnieniem redukcji wymiaru i poprawą jakości wyników eksploracji danych w przestrzeni zredukowanej. W ramach pierwszej publikacji omówiono i eksperymentalnie zweryfikowano algorytm nienadzorowanej redukcji wymiaru oparty o nowatorski wariant algorytmu symulowanego wyżarzania. Aspekty związane z opracowaną strategią optymalizacji zostaną omówione w dalszej części niniejszego rozdziału. Przedmiotem przedstawianych tutaj rozważań jest jedynie jej użycie w problemie ekstrakcji cech.

Zaproponowana procedura opiera się o liniową transformację pierwotnego m-elementowego i n wymiarowego zbioru danych X:

$$X = [x_1|x_2|...|x_m]^T, \tag{2}$$

czyli operację:

$$Y = X \cdot A, \tag{3}$$

gdzie A jest macierzą transformacji o wymiarach $n \times N$. Elementy tej macierzy są określane z użyciem procedury heurystycznej optymalizacji. W tym celu definiuje się wskaźnik kosztu:

$$S_R(A) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (d_{ij} - \delta_{ij}(A))^2 . \tag{4}$$

w literaturze określanym surowym stresem (ang. raw stress)[19]. Przy czym d_{ij} oznacza odległość między elementami x_i i x_j zbioru pierwotnego, a $\delta_{ij}(A)$ reprezentuje tą samą odległość wyznaczoną w zbiorze zredukowanym. Dzięki zastosowaniu takiej prostej formy ekstrakcji cech zyskuje się możliwość jej uogólnienia na nowe elementy, nie znajdujące się w pierwotnym zbiorze danych (ang. out-of-sample extension). Ponadto możliwe jest określenie wkładu jaki transformacja każdego elementu i wnosi w ostateczną wartość wskaźnika (4). Jest on określony wzorem:

$$S_R(A)_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} (d_{ij} - \delta_{ij}(A))^2 , \tag{5}$$

Tak zdefiniowany wskaźnik leży u podstaw kolejnej koncepcji, uogólnionej w pracy [H.3] na inne metody redukcji wymiaru. Zakłada ona wykorzystanie odwróconej i unormowanej wartości $S_R(A)_i$ jako wagi – której użycie w procedurach analizy danych realizowanych w przestrzeni zredukowanej przynosi poprawę jakości uzyskiwanych wyników. Dzieje się tak ponieważ elementy, które w skutek transformacji uległy względnemu przesunięciu względem pozostałej części zbioru, zostają uznane za mniej istotne. Wagi są znormalizowane w taki sposób by ich średnia była równa 1. W toku realizowanych po redukcji wymiaru procedur eksploracji danych można całkowicie pominąć elementy których wartości wag są mniejsze od pewnej wartości progowej W. Taką modyfikację można bez przeszkód wprowadzić także do klasycznych algorytmów klasteryzacji (np. k-średnich), klasyfikacji (np. sieci neuronowych) czy wykrywania elementów odosobionych (bazującego chociażby o metody statystyki nieparametrycznej).

Omawiana wyżej metoda redukcji wymiaru została pozytywnie zweryfikowana w zagadnieniach ekstrakcji cech ze zbiorów zarówno losowych jak i pochodzących z UCI Machine Learning Repository [20]. Badano przy tym nie tylko minimalizację wskaźnika 4 ale wpływ redukcji i wprowadzenia wyżej zdefiniowanych wag na

jakość wyników uzyskiwanych w przestrzeni zredukowanej w toku procedur: wykrywania elementów odosobnionych, analizy skupień i klasyfikacji. Ilustrację skutecznosci zaproponowanego podejścia stanowi Tabela 1 w której zaprezentowano wyniki dla klasyfikacji z użyciem metody estymatorów jądrowych. Redukcja wymiaru była tu przeprowadzana dla zbioru uczącego a uogólniana na zbiór testujący. Wyniki dla algorytmów ewolucyjnych podano na podstawie pracy [21].

Tabela 1: Porównanie metod redukcji wymiaru dla problemu klasyfikacji (podano średni odsetek poprawnych klasyfikacji oraz jego odchylenie standardowe – dla 30 powtórzeń z procedurą walidacji krzyżowej z podziałem na 5 podzbiorów)

	Zbiór				
	glass	*wine*	*WBC*	*vehicle*	*seeds*
$\overline{I}_{kINIT}$	71.90	74.57	95.88	63.37	90.23
$\pm\sigma(I_{kINIT})$	±8.10	±5.29	±1.35	±3.34	±2.85
	Opracowany algorytm				
$\overline{I}_{kRED}$	69.29	75.14	95.66	63.85	89.29
$\pm\sigma(I_{kRED})$	±3.79	±6.46	±1.12	±2.98	±2.31
$time[s]$	4.5	27.1	10.0	400.2	4.7
	Redukcja z użyciem PCA				
$\overline{I}_{kRED}$	58.33	72.00	95.29	62.24	83.09
$\pm\sigma(I_{kRED})$	±6.37	±7.22	±2.06	±3.84	±7.31
	Redukcja z użyciem algorytmów ewolucyjnych				
$\overline{I}_{kRED}$	64.80	72.82	95.10	60.86	N/A
$\pm\sigma(I_{kRED})$	±4.43	±1.02	±0.80	±1.51	

Z kolei w Tabeli 2 zawarto wyniki dla procedur analizy skupień i klasyfikacji w przestrzeni zredukowanej stosujących wyżej wspomniane wagi. Poza uwzględnieniem klasycznej postaci wag podanej wzorem (5) uwzględniono również możliwość użycia do ich konstrukcji: stresu Sammona – dodatkowo unormowanej postaci stresu redukującej wpływ dużych odległości, ρ-Spearmana [22] – które pozwala na ocenę stopnia zachowania porządku odległości oraz współcznnika Mean Relative Rank Error (MRRE) – opisującego stopień zachowania lokalnych grafów sąsiedztwa [23]. Dla każdej z metod uwzględniającej użycie wag podano kombinację postaci wag i progu odcięcia W która zapewnia najwyższą jakość uzyskiwanych wyników.

Tabela 2: Analiza skupień i klasyfikacja w przestrzeni zredukowanej – porównanie standardowego podejścia z propozycją algorytmów z wagami ($I_c \cdot 100\%$ - indeks Randa dla klasteryzacji i I_k - odsetek poprawnych klasyfikacji). Podano też optymalne wartości progów odcięcia wag W oraz rodzaj wskaźnika używanego do ich konstrukcji.

Procedura/zbiór	*glass*	*wine*	*WBC*	*vehicle*	*seeds*
PCA, K-średnich	69.41	71.37	92.51	64.21	87.28
	±2.43	±1.07	±0.13	±1.91	±0.15
PCA, ważone K-średnich	71.46	71.40	93.21	64.21	87.93
	±1.88	±0.98	±0.00	±1.80	±0.00
Najlepsze W	W=0.6	W=0.2	W=0	W=0.1	W=0.6
Typ wag	Surowy stres	ρ-Spearmana	Stres Sammona	ρ-Spearmana	Surowy stres
PCA, sieć neuronowa	59.99	74.51	96.78	55.78	88.34
	±2.79	±3.08	±0.62	±1.21	±1.80
PCA, ważona sieć neuronowa	63.80	76.58	96.82	58,13	90.24
	±2,69	±2.57	±0.63	±1.40	±1.88
Najlepsze W	W=0.5	W=0	W=0.1	W=0	W=0.7
Typ wag	MRRE	Surowy stres	MRRE	Stres Sammona	Surowy stres

W kolejnej pracy [H.4], pt. "Handling high-dimensional data in air pollution forecasting tasks" rozważano praktyczny problem wielowymiarowej analizy danych – predykcję poziomu zanieczyszczeń powietrza na obszarze Górnego Śląska. W tym celu wykorzystywane były rzeczywiste dane pochodzące ze stacji meteorologicznych oraz historyczne pomiary ze stacji mierzących koncentracje CO, NO_2, O_3, PM_{10} i SO_2. Predykcja realizowana była z użyciem autorskiego algorytmu wykorzystującego wnioskowanie rozmyte. Łącznie korzystano z 27 cech.

Zasadniczym celem przeprowadzonych badań było zaproponowanie optymalnej strategii rozwiązywania problemu wielowymiarowości dostępnych danych. Założono przy tym, że zastosowana metoda ma mieć charakter nienadzorowany. Preferowane były też techniki które pozwalały na uogólnienie transformacji na obserwacje nie należące do pierwotnego (dostępnego na etapie redukcji) zbioru danych (wspomniane już tzw. "out-of-sample extension"). O pionierskim charakterze przeprowadzonych badań może świadczyć porównanie aż 16 metod redukcji. Ponadto w porównaniu tym uwzględniono również możliwość przeprowadzenia wnioskowania na podstawie całego niezredukowanego zbioru – z użyciem odległości ułamkowych. Bardzo wartościowym wynikiem omawianej pracy było wskazanie, że tego rodzaju podejście zapewnia najwyższą jakość realizowanej predykcji.

Tabela 3 zawiera podsumowanie uzyskanych wyników dla stacji Katowice (KTW), Wodzisław Śląski (WOD) i Cieszyn (CIE). Jakość prognozy oceniano na podstawie błędu RMSE liczonego bądź to godzinowo (prognoza na 72 godziny, błąd liczony co godzinę) bądź względem średniej dziennej. Użyto również 3 wariantów konstrukcji szeregów czasowych dla celów prognozowania (określonych w tabeli jako $avg.$, α i $\alpha\beta$). Więcej szczegółów na temat samego algorytmu predykcyjnego można znaleźć w pracy [26]. Samo porównanie przeprowadzono poprzez ranking metod dla każdej z prognoz. Ostatnie kolumny tabeli zawierają uśredniony ranking oraz jego odchylenie standardowe Ponadto podano również względny błąd – liczony w odniesieniu do standardowej metody opartej o pełne 27-wymiarowe wektory cech i odległości euklidesowe. Jak już wcześniej zasygnalizowano najlepsze wyniki prognozowania uzyskuje się dla algorytmu pracującego w oparciu o odległości ułamkowe (z $p = 0,5$). Skuteczność wyższą niż wspomniane standardowe podejście zapewniały także – w każdym z badanych przypadków – metody Isomap, Landmark Isomap i Analiza Czynnikowa (ang. Factor Analysis). Na koniec warto wspomnieć o jeszcze jednym ciekawym wyniku zawartym w opisywanej pracy. Jest nim demonstracja, że postać transformacji redukująca wymiar otrzymana dla danych z jednej stacji może być z powodzeniem użyta w prognozowaniu dla innej stacji.

Tabela 3: Podsumowanie wyników predykcji poziomu koncentracji zanieczyszczeń powietrza dla 16 metod redukcji wymiaru, metryki euklidesowej i ułamkowej.

		KTW	KTW	KTW	WOD	WOD	WOD	CIE	CIE	CIE	WOD	WOD	WOD	Ranking	
		godz. avg.	godz. α	godz. $\alpha\beta$	godz. avg.	godz. α	godz. $\alpha\beta$	śr.dz. avg.	śr.dz. α	śr.dz. $\alpha\beta$	śr.dz. avg.	śr.dz. α	śr.dz. $\alpha\beta$	średnia	σ
Ułamkowa	Ranking	1	1	2	1	1	1	1	9	2	1	1	1	1.8	2.3
	$\overline{RMSE_{vsEucl.}}$	-9.4%	-16.0%	-5.4%	-8.9%	-13.2%	-12.5%	-14.3%	-8.4%	-11.0%	-16.3%	-20.5%	-20.3%	-13.0%	4.7%
IM	Ranking	2	4	3	3	2	2	3	15	3	2	3	2	3.7	3.6
	$\overline{RMSE_{vsEucl.}}$	-8.2%	-14.6%	-5.0%	-7.7%	-11.0%	-9.6%	-10.1%	-2.6%	-7.9%	-16.2%	-17.0%	-14.7%	-10.4%	4.5%
LIM	Ranking	3	5	4	2	3	3	2	14	4	3	2	3	4.0	3.3
	$\overline{RMSE_{vsEucl.}}$	-8.1%	-14.4%	-4.9%	-7.8%	-10.8%	-9.2%	-10.2%	-3.7%	-7.3%	-15.5%	-17.0%	-13.5%	-10.2%	4.2%
FA	Ranking	5	3	6	4	4	4	5	13	5	4	4	7	5.3	2.6
	$\overline{RMSE_{vsEucl.}}$	-7.4%	-14.9%	-4.7%	-6.1%	-9.9%	-8.7%	-6.0%	-3.8%	-7.3%	-14.0%	-14.6%	-12.8%	-9.2%	4.0%
PCA	Ranking	4	2	5	5	5	6	4	17	16	5	5	15	7.4	5.3
	$\overline{RMSE_{vsEucl.}}$	-8.0%	-15.0%	-4.8%	-5.6%	-9.1%	-7.8%	-6.7%	7.6%	0.7%	-12.7%	-13.5%	-9.8%	-7.1%	6.3%
ProbPCA	Ranking	6	6	1	6	6	5	6	18	17	6	6	14	8.1	5.2
	$\overline{RMSE_{vsEucl.}}$	-6.3%	-14.2%	-5.5%	-2.7%	-9.0%	-8.3%	-5.1%	13.4%	0.5%	-9.3%	-12.7%	-9.9%	-5.8%	7.3%
Euklidesowa	Ranking	7	18	8	7	15	18	7	16	15	7	18	18	12.8	5.1
	$\overline{RMSE_{vsEucl.}}$	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Laplacian	Ranking	10	7	7	8	7	8	8	8	1	8	7	8	7.3	2.1
	$\overline{RMSE_{vsEucl.}}$	14.7%	-8.6%	-2.5%	27.3%	-2.7%	-4.8%	19.0%	-9.0%	-12.1%	16.3%	-11.0%	-12.5%	1.2%	14.1%
LPP	Ranking	8	9	9	11	8	7	10	4	8	9	8	6	8.1	1.8
	$\overline{RMSE_{vsEucl.}}$	12.8%	-6.2%	0.3%	28.5%	-2.7%	-5.0%	20.6%	-11.7%	-6.7%	17.3%	-10.6%	-13.0%	2.0%	14.1%
LTSA	Ranking	11	8	11	15	9	11	11	7	9	14	9	4	9.9	3.0
	$\overline{RMSE_{vsEucl.}}$	16.2%	-6.4%	0.9%	31.3%	-1.6%	-4.4%	21.3%	-9.6%	-6.4%	19.5%	-9.0%	-13.3%	3.2%	14.8%
LLTSA	Ranking	13	11	13	14	12	12	9	1	7	13	11	10	10.5	3.6
	$\overline{RMSE_{vsEucl.}}$	16.7%	-5.3%	1.9%	31.0%	-0.6%	-3.9%	20.4%	-13.6%	-7.0%	19.1%	-7.2%	-11.8%	3.3%	14.6%
NPE	Ranking	16	12	16	12	11	10	13	2	6	11	10	5	10.3	4.2
	$\overline{RMSE_{vsEucl.}}$	18.7%	-4.7%	4.7%	29.8%	-1.1%	-4.4%	22.7%	-11.8%	-7.2%	17.9%	-8.5%	-13.1%	3.6%	14.8%
tSNE	Ranking	9	10	10	10	10	16	14	10	13	10	13	17	11.8	2.7
	$\overline{RMSE_{vsEucl.}}$	14.3%	-5.6%	0.7%	27.6%	-1.2%	-2.5%	24.0%	-7.6%	-3.8%	17.8%	-5.9%	-7.3%	4.2%	13.0%
LMDS	Ranking	12	13	12	17	14	13	12	3	11	16	12	9	12.0	3.5
	$\overline{RMSE_{vsEucl.}}$	16.4%	-4.6%	1.4%	33.1%	-0.1%	-3.6%	22.2%	-11.8%	-4.1%	19.8%	-6.9%	-12.1%	4.2%	14.8%
RP	Ranking	14	15	14	9	13	9	17	12	18	15	16	13	13.8	2.8
	$\overline{RMSE_{vsEucl.}}$	16.8%	-3.6%	2.1%	27.3%	-0.1%	-4.6%	26.9%	-5.3%	1.3%	19.6%	-2.8%	-10.4%	5.6%	13.3%
Kernel PCA	Ranking	15	14	15	13	16	14	16	11	14	17	15	16	14.7	1.6
	$\overline{RMSE_{vsEucl.}}$	18.1%	-3.7%	4.1%	30.8%	0.4%	-3.3%	26.8%	-7.5%	-2.1%	20.2%	-4.9%	-9.6%	5.8%	14.2%
LLE	Ranking	18	17	18	16	18	15	15	6	12	12	17	12	14.7	3.6
	$\overline{RMSE_{vsEucl.}}$	24.4%	-1.0%	7.9%	32.3%	1.8%	-2.7%	24.1%	-10.6%	-4.0%	18.1%	-2.8%	-10.4%	6.4%	14.7%
SPE	Ranking	17	16	17	18	17	17	18	5	10	18	14	11	14.8	4.1
	$\overline{RMSE_{vsEucl.}}$	23.5%	-2.2%	4.7%	35.0%	1.5%	-2.0%	27.8%	-11.3%	-4.7%	21.2%	-5.6%	-10.7%	6.4%	16.0%

W dwóch kolejnych pracach ([H.7] i [H.5]) pt. "The complete gradient clustering algorithm: properties in practical applications" oraz "Identification of Atypical (Rare) Elements – A Conditional, Distribution-Free Approach" przedstawiono kompletne procedury uczenia nienadzorowanego – analizy skupień i wykrywania elementów odosobnionych – zdefiniowane z użyciem estymatorów jądrowych.

Pierwszy z algorytmów opiera się o koncepcję Gradientowego Algorytmu Klasteryzacji wprowadzonego przez Fukunagę i Hostetlera [24]. Na wstępie elementy próby są przesuwane zgodnie z kierunkiem gradientu.

W związku z tym położenie elementu x_i w iteracji k jest zdefiniowane następująco:

$$x_i^{k+1} = x_i^k + b\frac{\nabla \hat{f}(x_i^k)}{\hat{f}(x_i^k)}, \tag{6}$$

gdzie b jest dodatnim parametrem zależnym od wymiarowości zbioru i parametru wygładzania użytego do konstrukcji estymatora $\hat{f}$. Proces relokacji elementów zbioru jest realizowany do momentu gdy zmiana odległości pomiędzy elementami zbioru w dwóch następujących po sobie iteracjach nie będzie przekraczała zadanej wartości progowej. W kolejnym kroku konstruuje się estymator wzajemnych odległości między poszczególnymi przypadkami. Pierwsze, większe od 0 lokalne minimum tego estymatora traktuje się jako odległość między środkami dwóch skupień, leżących najbliżej siebie, oznaczaną jako d_{min}. Podział na skupienia jest dokonywany na zbiorze po przeprowadzonej relokacji według następującej procedury:

1. wybierz dowolny element zbioru x_p i stwórz z niego jednoelementowe skupienie;
2. znajdź element inny od x_p – znajdujący się w odległości d_{min} od x_p – i dodaj go do skupienia, jeśli takiego elementu nie ma idź do kroku 4;
3. znajdź element zbioru inny niż te już przypisane do skupienia, a znajdujący się w odległosci d_{min} od nich i przypisz je do skupienia które reprezentują, powtarzaj krok 3 do momentu gdy żaden element nie zostanie odnaleziony;
4. usuń skupienie i elementy w nim zawarte ze zbioru, gdy zbiór nie jest pusty – idź do kroku 1; gdy zbiór jest pusty – zakończ algorytm i zwróć otrzymane skupienia;

Otrzymany algorytm posiada kilka bardzo atrakcyjnych cech z punktu widzenia aplikacyjnego. Po pierwsze niewielką liczbę parametrów, dla których podano rekomendowane wartości uzasadnione teoretycznie, bądź wynikające z praktyki obliczeniowej. Po drugie: brak potrzeby zdefiniowania arbitralnych założeń dotyczących oczekiwanej liczby skupień. Nie oznacza to przy tym braku możliwości dostosowywania uzyskiwanego rozwiązania do własnych potrzeb. Przykładowo zmieniając wartość parametru wygładzania względem tej otrzymywanej na podstawie procedury walidacji krzyżowej można uzyskać zwiększoną/zmniejszoną liczbę skupień. Fakt ten ilustruje Rysunek 2. Ponadto poprzez zmianę wartości parametru c obecnego w procedurze modyfikacji parametru wygładzania można także wpływać na liczbę skupień w obszarach peryferyjnych. I wreszcie dzięki zastosowaniu przybliżonego wyznaczania wartości estymatora w połączeniu z przetwarzaniem równoległym (np. w oparciu o GPU) możliwe jest jego zastosowanie także dla zbiorów danych o znacznych rozmiarach [25].

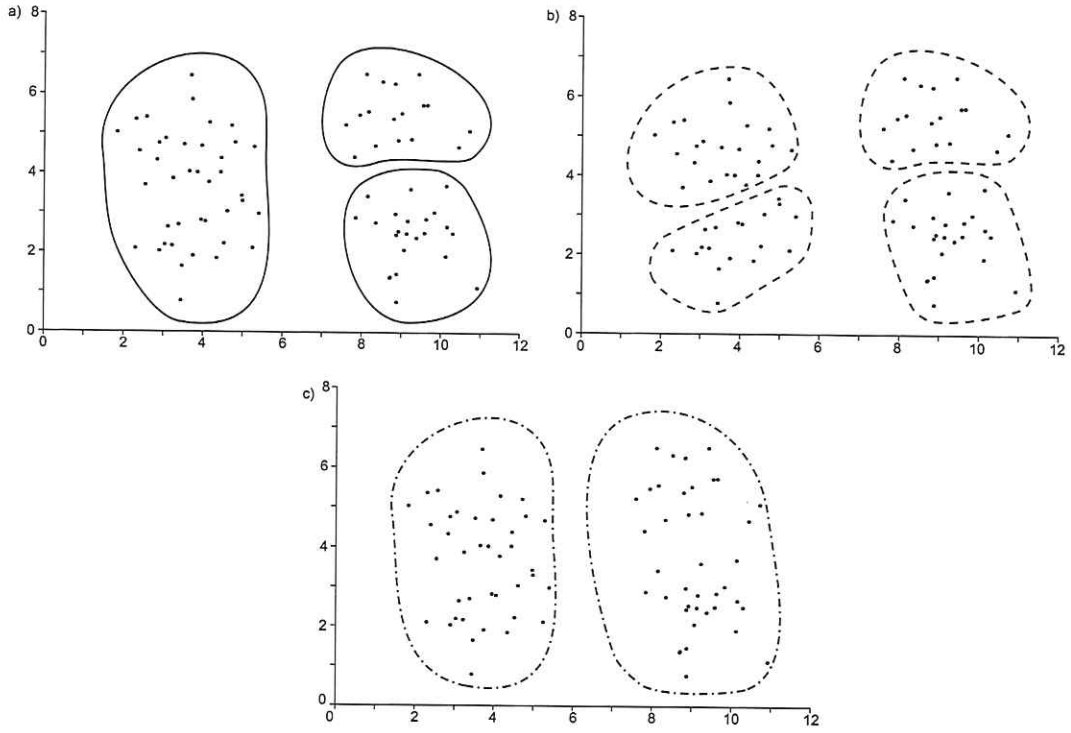
Z kolei zawarty w pracy [H.5] algorytm wykrywania elementów odosobnionych pozwala w próbie $y_1, y_2, ..., y_m$ (będącej realizacją zmiennej losowej Y) otrzymanej dla próby czynników warunkujących $w_1, w_2, ..., w_m$ (będących realizacją zmiennej warunkującej W) zidentyfikować elementy które będą pojawiać się rzadko, dla przyjętej wartości zmiennej warunkującej w^*. W tym celu wykorzystywana jest jednowymiarowa próba składająca się z wartości estymatora $\{f_{Y|W=w^*}(y_i)\}_{i=1...m}$. Wartość tego estymatora określa prawdopodobieństwo wystąpienia elementu y_i, przy założeniu wartości zmiennej warunkującej w^*. Dla tak skonstruowanej próby używany jest estymator pozycyjny kwantyla zadanego rzędu (np. 0,01; 0,05 lub 0,1) – pozwalający na określenie wartości progowej funkcji gęstości rozkładu prawdopodobieństwa, poniżej której element będzie klasyfikowany jako odstający. Podejście to pozwala – po uzyskaniu wspomnianego progu – na dokonywanie samego wykrywania w trybie online, co jest bardzo istotne w przypadku zbiorów z którymi musi się mierzyć współczesna eksploracja danych.

Kolejna praca z omawianego cyklu ([H.6]), pt. "An Evaluation of Utilizing Geometric Features for Wheat Grain Classification using X-ray Images" poświęcona jest redukcji wymiaru zbioru, reprezentującego cechy geometryczne ziaren pszenicy, dla celów jego klasteryzacji. W ramach opisywanych badań pozyskano wyniki 288 pomiarów dla 3 gatunków ziaren (Canadian, Kama i Rosa). Każdy pomiar zawierał 12 cech – zidentyfikowanych na podstawie przetworzenia obrazów rentgenowskich ziaren. Jako że poszczególne zmienne zbioru były ze sobą silnie skorelowane (co demonstruje Tabela 4) istotne było znalezienie minimalnego zestawu cech pozwalającego na uzyskanie skupień odpowiadającym gatunkom ziaren.

W tym celu przeprowadzono analizę głównych składowych (ang. Principal Components Analysis) w której wyniku stwierdzono, że pierwsze 3 składowe zawierają prawie 90% łącznej wariancji. Fakt ten ilustruje również wykres osypiskowy (Rys. 3).

W wyniku analizy czynnikowej stwierdzono, że pierwsza ze składowych odpowiada podstawowym cechom geometrycznym ziaren opisujących ich kształt. Druga składowa opisuje proporcje między zalążkiem a ziarnem. Z kolei trzecia stanowi uzupełnienie dla pierwszej w zakresie zasadniczych cech geometrii ziaren.

Po przeprowadzonej analizie zweryfikowano adekwatność zredukowanej reprezentacji zbioru w odniesieniu do problemu analizy skupień. Uzyskiwane wartości indeksu Randa – liczonego względem znanych etykiet klas –



Rysunek 2: Wynik procedury klasteryzacji a) przy wartości parametru wygładzania uzyskanej metodą walidacji krzyżowej, b) przy wartości parametru wygładzania zmniejszonej o 25%, c) przy wartości parametru wygładzania zwiększonej o 50%.

są bliskie 0.9. Można wnioskować zatem o wysokiej użyteczności trójwymiarowej postaci analizowanego zbioru w problemie dyskryminacji ziaren.

Kolejny artykuł [H.8] zatytułowany "Training neural networks with krill herd algorithm" poświęcony został metaheurystyce kryla (szczętki) czyli Krill Herd Algorithm (KHA). Jego podstawowym celem – poza studiami nad aplikacyjnymi aspektami użycia tej techniki – było jej zastosowanie do modyfikacji połączeń pomiędzy neuronami w wielowarstwowej sieci neuronowej. Choć zasadniczo eksperymentalna część tej pracy poświęcona jest technice uczenia nadzorowanego – klasyfikacji, to wyniki w niej zawarte stanowiły punkt wyjścia do dalszych badań, poświęconych m.in. zastosowaniom KHA w analizie skupień (omówione w Podrozdziale 5.1).

Opracowany w 2012 przez A.H. Gandomiego i A. H. Alaviego algorytm [27] jest jedną z nowszych procedur optymalizacyjnych o heurystycznym charakterze. U jej podstaw leży inspiracja naturą – a śliślej zachowaniem ławic kryla antarktycznego (*Euphausia superba*), występującego w Oceanie Południowym.

Metaheurystyka KHA (rozumiana jako ogólny algorytm do rozwiązywania zadań obliczeniowych) została opracowana przede wszystkim dla celów rozwiązywania problemów optymalizacji ciągłej. Zadanie takie polega na znalezieniu wartości x^* w A, gdzie $A \subset \mathbb{R}^n$, tak aby dla każdego x należącego do A prawdziwa była relacja:

$$f(x) \geq f(x^*) \quad \text{/dla zadania minimalizacji/} \tag{7}$$

gdy f odpowiada tzw. funkcji kosztu lub:

$$f(x) \leq f(x^*) \quad \text{/dla zadania maksymalizacji/} \tag{8}$$

jeśli f jest rozumiane jako funkcja przystosowania/jakości. W obu przypadkach funkcja f jest zdefiniowana jako:

$$f : A \longmapsto \mathbb{R}, \tag{9}$$

Procedura KHA należy do kategorii metaheurystyk populacyjnych, które w celu rozwiązania problemów (7) lub (8) wykorzystują populację rozwiązań. Odpowiada ona w przypadku tego algorytmu całej ławicy składającej się z M kryli. Każdy osobnik ($i = 1, ..., M$) reprezentuje rozwiązanie problemu optymalizacji – definiowane jako jego położenie X_i w przestrzeni rozwiązań dopuszczalnych. Przestrzeń ta jest przez niego eksplorowana na podstawie następującego równania opisującego dynamikę ruchu poszczególnych osobników ławicy:

$$X_i(t + \delta t) = X_i(t) + \Delta t \frac{dX_i}{dt} \tag{10}$$

Tabela 4: Macierz korelacji poszczególnych cech reprezentujących własności geometryczne ziaren. Gwiazdką oznaczono wyniki statystycznie istotne, na poziomie istotności 0,01.

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}
V_1	.	.	.	.	.	.	.	.	.	.	.	.
V_2	0.99*	.	.	.	.	.	.	.	.	.	.	.
V_3	0.62*	0.56*	.	.	.	.	.	.	.	.	.	.
V_4	0.96*	0.98*	0.42*	.	.	.	.	.	.	.	.	.
V_5	0.97*	0.96*	0.76*	0.89*	.	.	.	.	.	.	.	.
V_6	−0.40*	−0.39*	−0.43*	−0.36*	−0.41*	.	.	.	.	.	.	.
V_7	0.90*	0.92*	0.31*	0.96*	0.81*	−0.25*	.	.	.	.	.	.
V_8	0.83*	0.84*	0.31*	0.84*	0.76*	−0.22*	0.84*	.	.	.	.	.
V_9	0.65*	0.66*	0.14*	0.72*	0.56*	−0.23*	0.69*	0.86*	.	.	.	.
V_{10}	−0.08	−0.05	−0.45*	0.03	−0.16*	0.26*	0.12	0.49*	0.53*	.	.	.
V_{11}	−0.08	−0.07	−0.25*	−0.01	−0.12*	0.05	0.00	0.34*	0.69*	0.73*	.	.
V_{12}	0.57*	0.51*	0.95*	0.34*	0.73*	−0.32*	0.24*	0.29*	0.08	−0.40*	−0.25*	.

gdzie Δt reprezentuje krok w dziedzinie czasu (domyślnie równy 1 i odpowiadający jednej iteracji algorytmu). Wektor prędkości $\frac{dX_i}{dt}$ zwany przez autorów Lagrangianem jest zdefiniowany następująco:

$$\frac{dX_i}{dt} = N_i + F_i + D_i, \tag{11}$$

i obejmuje trzy komponenty stanowiące o istocie ruchu danego osobnika: N_i symbolizuje ruch wywołany przez obecność innych osobników w ławicy, F_i jest związany z poszukiwaniem pokarmu, natomiast D_i określa fizyczną dyfuzję i-tego elementu ławicy.

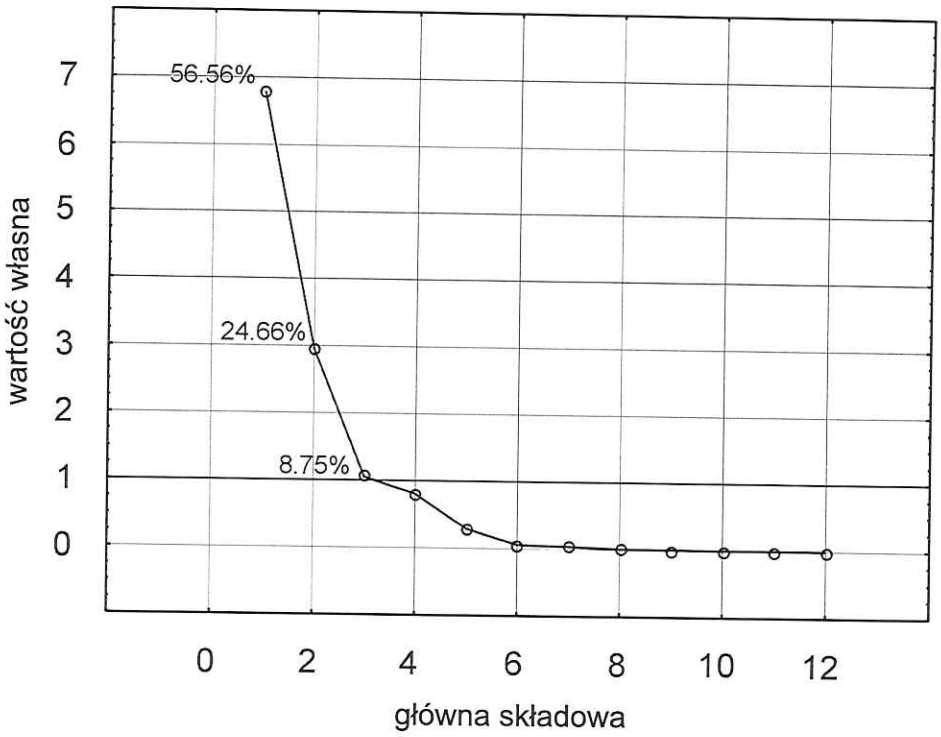
Pełny schemat blokowy określający zasadę działania algorytmu KHA został przedstawiony na Rysunku 4). Na schemacie poza obecnymi we wzorze (11) wspomnianymi elementami związanych z zachowaniem stadnym (zgrupowanymi na rysunku jako "zjawiska ruchu"), umieszczono znane z klasycznych operatorów ewolucyjnych genetyczne operatory krzyżowania i mutacji. Ich użycie jest opcjonalnym elementem procedury KHA.

Technika optymalizacji ciągłej bazująca o Algorytm Kryla została użyta w problemie uczenia wielowarstwowej sieci neuronowej stosowanej w zagadnieniu klasyfikacji. Rozwiązaniem stanowił wektor określający wszystkie wagi i wartości progów dla całej sieci, o przyjętej z góry strukturze. Początkowe położenia osobników były generowane z użyciem generatora liczb losowym o rozkładzie jednostajnym z zadanymi ograniczeniami. W toku realizowanych prac przeprowadzono badania temat rekomendowanych wartości parametrów algorytmu – w tym rozmiaru populacji. Uzyskane wyniki posłużyły do przeprowadzenia ostatecznej serii eksperymentów – dla celów badań porównawczych.

Jako funkcję kosztu f, przyjęto kombinację dwóch miar (każdej z wagą 1/2): dokładności rozumianej jako odsetek błędnych klasyfikacji elementów próby uczącej oraz błędu – określanego przez sumę kwadratów błędów obliczanych dla poszczególnych elementów wzorca. W ramach przeprowadzonych badań przedmiotem bardziej szczegółowych rozważań była również szybkość zbieżności algorytmu. Opracowana technika była przedmiotem weryfikacji numerycznej przeprowadzonej dla wybranych zbiorów pochodzących z UCI Machine Learning Repository. Jako punkt odniesienia do badań porównawczych użyto klasycznego algorytmu Propagacji Wstecznej (ang. Back Propagation, BP), Algorytmu Genetycznego (ang. Genetic Algorithm, GA) oraz Wyszukiwania z użyciem Harmonii (ang. Harmony Search, HS) [28]. Rozważany problem dzięki swej wielomodalnej charakterystyce i znacznej wymiarowości (przykładowo dla zbioru *Ionosphere* optymalizacja dotyczyła wektora 378 zmiennych) stanowił bardzo cenny przyczynek do badań aplikacyjnych nad algorytmem KHA. Opracowana technika optymalizacji parametrów sieci neuronowej zapewnia bardzo dobre rezultaty w odniesieniu do użytej funkcji kosztu, a także przyspiesza znacznie proces uczenia. Dla zbioru Iris czas jego wykonania był równy połowie czasu trenowania z użyciem metody GA. Porównanie z algorytmem BP daje jeszcze bardziej jednoznaczne rezultaty – uczenie sieci tą techniką trwa 42 razy dłużej. Podobne wyniki uzyskano dla pozostałych, uwzględnionych w badaniu, zbiorach danych.

W omawianym tu nurcie badań nad algorytmami optymalizacyjnymi nie można pominąć wkładu pracy [H.2] omawianej już w kontekście zawartej w niej koncepcji algorytmu redukcji wymiaru. Procedura ta opiera się bowiem o nowatorski wariant klasycznego algorytmu Szybkiego Symulowanego Wyżarzania (ang. Fast Simulated Annealing, FSA) [29]. Algorytm FSA w celu zmian generacji rozwiązania sąsiedniego korzysta z generatora opartego o rozkład Cauchy'ego, z równoczesną szybką zmianą temperatury – i związanej z nim prawdopodobieństwem akceptacji rozwiązań gorszych od bieżącego w danej iteracji.

Opracowany nowy wariant tej techniki rozwiązuje kilka istotnych problemów praktycznych. Po pierwsze zaproponowano technikę pozwalającą na ustalenie wartości temperatury początkowej odpowiadającej zadane-



Rysunek 3: Wykres osypiskowy demonstrujący główne składowe zbioru pomiarów ziaren i ich procentowy wkład w łączną wariancję.

mu prawdopodobieństwu akceptacji rozwiązań gorszych (z równoczesnym uniezależnieniem się od rozważanej instancji problemu optymalizacji). Po drugie, zamiast powszechnie stosowanego w praktyce generowania każdego z wymiarów wektora kroku z jednowymiarowego generatora o rozkładzie Cauchy'ego wprowadzono alternatywną strategię – użycie generatora wielowymiarowego opartego o transformację kartezjańskiego układu współrzędnych do układu współrzędnych sferycznych. I wreszcie wprowadzono warunek zatrzymania pracy algorytmu opierający się o analizę otrzymywanych wartości funkcji kosztu, pozwalającą na wnioskowanie o stopniu zaawansowania procedury przeszukiwania. W tym celu konstruowany jest estymator oczekiwanej wartości globalnego minimum oparty o statystykę porządkową. Wspomniana koncepcja, wzbogacona dodatkowo o równoległą generacją rozwiązań sąsiednich, stanowi atrakcyjny efektywny algorytm optymalizacyjny z minimalną liczbą parametrów i ich intuicyjną interpretacją.

5 Pozostałe osiągnięcia naukowe

5.1 Algorytmy inspirowane naturą w problemach optymalizacji ciągłej

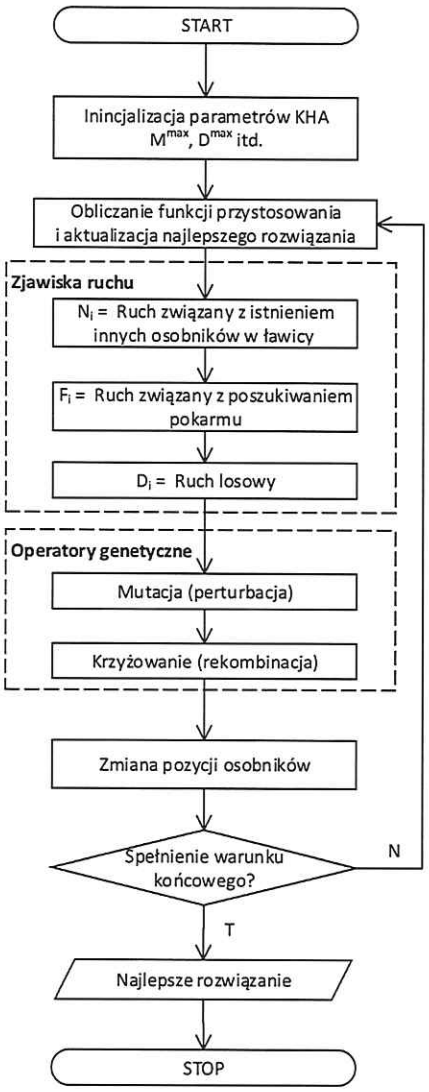
Zarysowany w poprzednim rozdziale problem optymalizacji ciągłej był w ramach prowadzonych prac badawczych rozważany w odniesieniu do procedur optymalizacji inspirowanych naturą. W tym zakresie podjęto prace badawcze w zakresie aplikacyjnych własności algorytmów: Kryla (Krill Herd Algorithm, KHA) oraz zapylania kwiatów (Flower Pollination Algorithm, FPA). W ten nurt wpisują się następujące prace:

[O.35] P. Kopciewicz i S. Łukasik. "Exploiting Flower Constancy in Flower Pollination Algorithm: Improved Biotic Flower Pollination Algorithm and its Experimental Evaluation". *Neural Computing and Applications* (2019). DOI: 10.1007/s00521-019-04179-9.

JCR, Web of Science, IF: 4,213, MNiSW: 25 pkt.

[O.36] P. A. Kowalski i S. Łukasik. "Experimental Study of Selected Parameters of the Krill Herd Algorithm". *Intelligent Systems'2014*. Red. P. Angelov, K. Atanassov, L. Doukovska, M. Hadjiski, V. Jotsov, J. Kacprzyk, N. Kasabov, S. Sotirov, E. Szmidt i S. Zadrożny. Cham: Springer International Publishing, 2015, s. 473–485. DOI: 10.1007/978-3-319-11313-5_42.

Web of Science, MNiSW: 15 pkt.



Rysunek 4: Schemat blokowy procedury KHA

[O.37] S. Łukasik i P. A. Kowalski. "Study of Flower Pollination Algorithm for Continuous Optimization". *Intelligent Systems'2014*. Red. P. Angelov, K. Atanassov, L. Doukovska, M. Hadjiski, V. Jotsov, J. Kacprzyk, N. Kasabov, S. Sotirov, E. Szmidt i S. Zadrożny. Cham: Springer International Publishing, 2015, s. 451–459. DOI: 10.1007/978-3-319-11313-5_40.

Web of Science, MNiSW: 15 pkt.

Oba algorytmy wnikliwie przetestowano korzystając w tym zakresie z dobrze znanego i uznanego zestawu funkcji testowych CEC17 [30]. Opracowano również rekomendacje w zakresie sugerowanych wartości dostępnych parametrów. W pracy [O.35] zaproponowano również modyfikację algorytmu FPA, która poprzez alternatywny schemat generacji rozwiązań zapewnia zarówno większą skuteczność optymalizacji jak i zmniejszoną złożoność obliczeniową.

Wyżej wymienione techniki inspirowane naturą zastosowano również z sukcesem w dwóch problemach praktycznych – binaryzacji obrazów pozyskanych na podstawie mikrotomografii komputerowej oraz projektowania strategii zachowań gracza komputerowego w grze wyścigowej. Podsumowanie wyników tych badań zawierają prace:

[O.38] P. A. Kowalski, J. Kamiński, S. Łukasik, J. Świebocka-Wiek, D. Gołuńska, J. Tarasiuk i P. Kulczycki. *Application of the Flower Pollination Algorithm in the Analysis of Micro-CT Scans*. Red. L. T. Kóczy, J. Medina-Moreno i E. Ramírez-Poussa. Cham, sty. 2019. DOI: 10.1007/978-3-030-00485-9_1.

Web of Science, MNiSW: 15 pkt.

[O.39] P. A. Kowalski, S. Łukasik, M. Charytanowicz i P. Kulczycki. "On the Use of Nature Inspired Metaheuristic in Computer Game". *Federated Conference on Computer Science and Information Systems 2017 (FedCSIS 2017)*. Red. M. Ganzha, L. Maciaszek i M. Paprzycki. T. 11. Annals of Computer Science and Information Systems. Praga (Republika Czeska): IEEE, wrz. 2017, s. 29–37. DOI: 10.15439/2017F385.

Web of Science, MNiSW: 15 pkt.

Ponadto, jako szczególny przypadek optymalizacji ciągłej rozpatrzono też zadanie klasteryzacji, gdzie rozwiązanie jest reprezentowane przez środki skupień – o założonej liczności (choć uogólnienie do przypadku gdy a priori przyjmuje się maksymalną liczbę skupień jest również możliwe). Jako funkcję oceny przyjęto indeks klasteryzacyjny Calinskiego-Harabasza – rozważając przy tym inne tego typu wskaźniki. W nurt prac badawczych związanych z tą tematyką wpisują się publikacje z użyciem algorytmów Krill Herd Algorithm, Glowworm Optimization Algorithm [31] oraz Flower Pollination Algorithm [32] wymienione poniżej:

[C.40] P. A. Kowalski, S. Łukasik, M. Charytanowicz i P. Kulczycki. "Nature Inspired Clustering – Use Cases of Krill Herd Algorithm and Flower Pollination Algorithm". *Interactions Between Computational Intelligence and Mathematics*. Red. L. T. Kóczy, J. Medina-Moreno i E. Ramírez-Poussa. Cham: Springer International Publishing, 2019, s. 83–98. DOI: 10.1007/978-3-030-01632-6_6.

MNiSW: 5 pkt.

[C.41] S. Łukasik, P. A. Kowalski, M. Charytanowicz i P. Kulczycki. "Data Clustering with Grasshopper Optimization Algorithm". *Federated Conference on Computer Science and Information Systems 2017 (FedCSIS 2017)*. Red. M. Ganzha, L. Maciaszek i M. Paprzycki. T. 11. Annals of Computer Science and Information Systems. Prague (Republika Czeska): IEEE, wrz. 2017, s. 71–74. DOI: 10.15439/2017F340.

Web of Science, MNiSW: 15 pkt.

W każdej z w/w prac dokonano porównań opracowanych algorytmów z klasyczną techniką k-średnich, opierając się w tym zakresie na zbiorach z repozytorium UCI, oraz na indeksie Randa R – jako metodzie oceniającej zgodność uzyskanego wyniku ze znanymi etykietami klas [33]. W Tabeli 5 pokazano wyniki takiego porównania dla algorytmów KHA i FPA. Tabela zawiera również wyniki testów t (statystycznie lepsze wyniki pod względem średniej wartości indeksu wyróżniono wytłuszczeniem) – zrealizowanych dla badanych technik i metody k-średnich. Jak można zauważyć w szczególności procedura FPA oferuje wysoką jakość uzyskiwanych rozwiązań analizy skupień oraz ich znaczną stabilność.

5.2 Metody wnioskowania rozmytego w modelowaniu i generacji podsumowań lingwistycznych

Ostatnim z rozważanych tu obszarów badawczych – zaliczanych do pozostałych osiągnięć naukowych – są studia nad zastosowaniami logiki rozmytej w modelowaniu, a także generowaniu podsumowań lingwistycznych.

Tabela 5: Wyniki analizy skupień wybranymi metodami inspirowanymi naturą – w zestawieniu z klasteryzacją techniką k-średnich.

	Klasteryzacja k-średnich		Klasteryzacja KHA		Klasteryzacja FPA	
Zbiór	$\overline{R}$	σ_R	$\overline{R_{KHA}}$	$\sigma_{R_{KHA}}$	$\overline{R_{FPA}}$	$\sigma_{R_{FPA}}$
S1	0.9748	0.0093	0.9782	0.0078	**0.9950**	0.0018
S2	0.9760	0.0072	**0.9839**	0.0053	0.9837	0.0037
S3	0.9522	0.0072	0.9548	0.0053	**0.9583**	0.0026
S4	0.9454	0.0056	0.9484	0.0048	**0.9487**	0.0023
Iris	0.8458	0.0614	0.8872	0.0145	**0.8931**	0.0000
Ionosphere	0.5945	0.0004	0.5573	0.0124	**0.5946**	0.0000
Seeds	0.8573	0.0572	0.8709	0.0156	**0.8839**	0.0000
Sonar	0.5116	0.0016	**0.5145**	0.0078	0.5128	0.0000
Vehicle	0.5843	0.0359	0.6076	0.0194	**0.6101**	0.0006
WBC	0.5448	0.0040	**0.5456**	0.0000	**0.5456**	0.0000
Wine	0.7167	0.0135	0.7257	0.0073	**0.7299**	0.0000
Thyroid	**0.5844**	0.0982	0.4535	0.0339	0.5128	0.0000

W ramach pierwszego zagadnienia rozważano przede wszystkim możliwość automatycznego generowania reguł rozmytego układu wnioskowania na podstawie dostępnych danych, w taki sposób by stworzony model możliwie wiernie odtwarzał funkcjonowanie procesu który owe dane reprezentują. W tym celu zostały użyte zarówno metody statystyki nieparametrycznej – a ściślej wstępny wariant algorytmu klasteryzacyjnego omówionego w poprzednim rozdziale – jak i metody klasteryzacji inspirowane naturą, konkretnie oparte o algorytm roju cząstek (z funkcją kosztu w postaci indeksu Davisa-Bouldina [34]). W obu przypadkach przedmiotem grupowania były łącznie: dane wejściowe modelowanego systemu jak i jego oczekiwane wyjście. W ten sposób każde skupienie mogło wprost odpowiadać jednej regule układu rozmytego. Koncepcja ta została z powodzeniem zweryfikowana w toku eksperymentów z zakresu modelowania i identyfikacji procesów (także nieliniowych). W ten nurt badawczy wpisują się prace:

[F.42] D. Falkiewicz i S. Łukasik. "Modelowanie rozmyte z zastosowaniem algorytmu optymalizacji rojem cząstek". *Czasopismo Techniczne. Automatyka* vol. 1-AC (2012), s. 41–54.

MNiSW: 5 pkt.

[F.43] P. A. Kowalski, S. Łukasik, M. Charytanowicz i P. Kulczycki. "Data-driven fuzzy modeling and control with kernel density based clustering technique". *Polish Journal of Environmental Studies* vol. 17. nr 4C (2008), s. 83–87.

JCR, IF:0.963 MNiSW: 10 pkt.

Ponadto podjęto nowatorskie badania na temat generowania rozmytych podsumowań lingwistycznych dla wyników rekomendacji. Tego typu procedura może być pomocna w "wyjaśnianiu" listy opcji wskazanych przez system rekomendacyjny, co warte podkreślenia, niezależnie od jego konstrukcji. Podsumowania lingwistyczne są generowane na podzbiorze opcji, który może być uzyskany zarówno wybraną metodą filtrowania kolaboracyjnego jak i rekomendacji bazującej na treści. Wspomniane rozwiązanie cechuje znaczna wartość komercyjna wynikająca ze zorientowania tego podejścia na potrzeby użytkownika końcowego. Dla celów weryfikacji tak sformułowanej idei stworzono własny zbiór zawierający oceny filmów pozyskane na próbie uczestników testu. Został on udostępniony publicznie by umożliwić szerszą replikację uzyskanych wyników. Przykładowy wynik działania algorytmu podsumowującego zademonstrowano w tabeli 6. Choć zawarte w niej podsumowania wydają się elementarne to ich automatyczne wygenerowanie, dla zbioru rekomendacji, może stanowić cenne uzupełnienie dla przedstawionych użytkownikowi opcji.
Szersze wyniki opisywanych tu prac zawarto w publikacji:

[F.44] S. Łukasik, M. Smęt i J. Królewski. "Generating Textual Descriptions for Recommendation Results using Fuzzy Linguistic Summaries". *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Lip. 2018, s. 1–6. DOI: 10.1109/FUZZ-IEEE.2018.8491601.

Web of Science, MNiSW: 15 pkt.

Będzie stanowić ona punkt wyjścia do dalszych rozważań w tej tematyce – w szczególności w zakresie badań komercyjnych prowadzonych we współpracy ze spółką Synerise.

Tabela 6: Wybrane podsumowania lingwistyczne pierwszego rzędu uzyskane dla wyników rekomendacji względem zmiennej "kompozytor muzyki"

Nr	Opis
1	Żadna z rekomendowanych opcji nie posiada muzyki skomponowanej przez Daft Punk
2	Wiele z rekomendowanych opcji zawiera muzykę skomponowaną przez Hansa Zimmer
3	Wiele z rekomendowanych opcji zawiera muzykę skomponowaną przez Johna Williamsa

6 Podsumowanie dorobku publikacyjnego

Liczba publikacji

W nawiasach dla pierwszych dwóch kolumn podano liczbę publikacji w czasopismach z listy Journal Citation Reports (JCR) oraz indeksowanych w bazie Web of Science (WoS). W tym, jak i dalszych zestawieniach nie uwzględniono prac przyjętych i oczekujących na publikację.

Rodzaj publikacji	Liczba publikacji sumaryczna	Liczba publikacji po doktoracie	przed doktoratem
Artykuły w czasopismach	17 (JCR=9, WoS=10)	12 (JCR=8, WoS=10)	5
Monografie	1	1	0
Rozdziały w pracach zbiorowych	16 (WoS=10)	9 (WoS=5)	7
Publikacje konferencyjne	20 (WoS=8)	16 (WoS=7)	4
Redakcja książek oraz czasopism[2]	7 (WoS=2)	7 (WoS=2)	0

Impact Factor oraz Liczba punktów MNiSW

Wskaźnik IF podany został według bazy JCR zgodnie z dostępem z dnia 20 kwietnia 2019 r. – z uwzględnieniem wskaźników obowiązujących w roku publikacji (oraz IF za 2017 rok dla prac opublikowanych w kolejnych latach). Liczbę punktów przyznawaną za publikację wyszczególniono zgodnie z obowiązującym w roku wydania wykazem czasopism naukowych, ogłoszonym przez Ministra Nauki i Szkolnictwa Wyższego oraz poniższymi przepisami dotyczącymi kategoryzacji jednostek naukowych:

- Rozporządzenie MNiSW z dnia 12 grudnia 2016 r.,
- Rozporządzenie MNiSW z dnia 27 października 2015 r.,
- Rozporządzenie MNiSW z dnia 13 lipca 2012 r.,
- Rozporządzenie MNiSW z dnia 17 października 2007 r.,
- Rozporządzenie MNiSW z dnia 4 sierpnia 2005 r.

Sumaryczny Impact Factor według bazy Journal Citation Reports (JCR), obejmujący publikacje po uzyskaniu stopnia doktora: **14,059**. Sumaryczny Impact Factor, obejmujący także publikacje przed uzyskaniem stopnia doktora: **15,022**.

[2]Wliczono pozycje dotyczące redakcji monografii wieloautorskich i zeszytów specjalnych czasopism.

Rodzaj publikacji	sumaryczna	Liczba punktów MNiSW po doktoracie	przed doktoratem
Artykuły w czasopismach	250	221	29
Monografie	25	25	0
Rozdziały w książkach	137	90	47
Publikacje konferencyjne	112	105	7
Redakcja książek oraz czasopism[2]	35	35	0
Suma	559	476	83

Liczba cytowań oraz index Hirscha

Wyszukiwanie w oparciu o dane autora publikacji w bazie **Web of Science**:

- liczba publikacji: 29,
- liczba cytowań: 354,
- indeks Hirscha: 8.

Odnośnik do profilu autora:
`http://www.researcherid.com/rid/A-3799-2013`

Wyszukiwanie w oparciu o dane autora publikacji w bazie **Scopus**:

- liczba publikacji: 31,
- liczba cytowań: 468 (410 bez autocytowań),
- indeks Hirscha: 9.

Odnośnik do profilu autora:
`https://www.scopus.com/authid/detail.uri?authorId=24385431300`

Wyszukiwanie w oparciu o dane autora publikacji w **Google Scholar**:

- liczba publikacji: 78,
- liczba cytowań: 845,
- indeks Hirscha: 13,
- i10-indeks: 14.

Odnośnik do profilu autora:
`https://scholar.google.pl/citations?user=-O9fFQwAAAAJ&hl=pl&oi=ao`

7 Podsumowanie dorobku naukowo-badawczego, dydaktycznego oraz organizacyjnego

Pełny wykaz opublikowanych prac naukowych, jak również szczegółowe informacje o osiągnięciach naukowo-badawczych, dydaktycznych, współpracy naukowej oraz popularyzacji nauki (z uwzględnieniem wymagań określonych w Rozporządzeniu Ministra Nauki i Szkolnictwa Wyższego z dania 1 września 2011 roku w sprawie kryteriów oceny osiągnięć osoby ubiegającej się o nadanie stopnia doktora habilitowanego tj. Dziennik ustaw 2011 nr 196 poz. 1165) został przedstawiony w załączniku 5.

Do najważniejszych osiągnięć w omawianej kategorii zaliczam:

1) kierowanie jednym projektem badawczo-rozwojowym współfinansowanym ze Funduszy Europejskich (budżet 4,3 mln PLN) oraz trzema projektami własnymi,

[2]Wliczono pozycje dotyczące redakcji monografii wieloautorskich i zeszytów specjalnych czasopism.

2) otrzymanie nagród Rektora Politechniki Krakowskiej (za osiągnięcia naukowe) oraz Rektora Akademii Górniczo-Hutniczej w Krakowie (za osiągnięcia dydaktyczne i osiągnięcia organizacyjne),

3) odbycie dziewięciu zagranicznych staży naukowych m.in. w IRIDIA (Bruksela), UNINOVA (Portugalia), University of California, Berkeley (USA) i National Laboratory of Pattern Recognition (Chiny),

4) uczestnictwo w rządowym programie "Top 500 Innovators: Science – Management – Commercialization",

5) odbycie stażu podoktorskiego w Instytucie Podstaw Informatyki PAN,

6) sprawowanie opieki nad jednym doktorantem,

7) opieka nad dwudziestoma pięcioma dyplomantami (prace magisterskie oraz inżynierskie),

8) recenzowanie artykułów dla 26 międzynarodowych czasopism naukowych - głównie indeksowanych w Journal Citation Reports (w sumie 162 recenzje; m.in. *IEEE Transactions on Evolutionary Computation*, *Information Sciences*, *Neural Computing and Applications*, *Fuzzy Sets and Systems*, *Applied Soft Computing*), oraz dla wielu renomowanych konferencji naukowych,

9) wygłoszenie 23 referatów na konferencjach naukowych,

10) członkostwo w komitetach naukowych konferencji (24 razy, w tym osiem razy jako *event chair* lub *co-chair* konferencji),

11) udział w dwóch komitetach redakcyjnych czasopism naukowych,

12) pełnienie funkcji eksperta dla Komisji Europejskiej i Narodowego Centrum Badań i Rozwoju,

13) prowadzenie zajęć dydaktycznych dla studentów kierunku Informatyka, Informatyka Stosowana oraz Informatyka Społeczna z zakresu analizy danych oraz szeroko pojętej sztucznej inteligencji,

14) pełnienie roli prelegenta w ramach cyklu zaproszonych wykładów popularyzacyjnych prace badawcze z zakresu sztucznej inteligencji i analizy danych (wystąpienia na konferencji „Akademia Marketingu" czy „Digital Economy Forum" organizowanym przez dziennik „Rzeczpospolita").

Pozostałe cytowane prace

[9] B. Clarke, E. Fokoue i H. Zhang. *Principles and Theory for Data Mining and Machine Learning*. Springer Series in Statistics. Springer New York, 2009.

[10] L. Grandinetti, G. Joubert, M. Kunze i V. Pascucci. *Big Data and High Performance Computing*. Advances in Parallel Computing. IOS Press, 2015.

[11] M. Charytanowicz, J. Niewczas, P. Kulczycki, P. Kowalski, S. Łukasik i S. Żak. *UCI Machine Learning Repository: seeds Dataset*. 2010. URL: https://archive.ics.uci.edu/ml/datasets/seeds.

[12] G. Ciaburro. *Regression Analysis with R: Design and develop statistical nodes to identify unique relationships within data at scale*. Packt Publishing, 2018.

[13] P. Mitra, C. Murthy i S. K. Pal. "Density-Based Multiscale Data Condensation". *IEEE Transactions on Pattern Analysis & Machine Intelligence* vol. 24 (2002), s. 734–747. DOI: 10.1109/TPAMI.2002.1008381.

[14] M. Hernández-Pajares i J. Floris. "Classification of the Hipparcos input catalogue using the Kohonen network". *Monthly Notices of the Royal Astronomical Society* vol. 268. nr 2 (1994), s. 444–450. DOI: 10.1093/mnras/268.2.444. eprint: http://mnras.oxfordjournals.org/content/268/2/444.full.pdf+html.

[15] G. Hinton i S. Roweis. "Stochastic Neighbor Embedding". *Advances in Neural Information Processing Systems*. T. 15. Cambridge: The MIT Press, 2002, s. 833–840.

[16] L. Maaten van der. "Accelerating t-SNE using Tree-Based Algorithms". *Journal of Machine Learning Research* vol. 15 (2014), s. 3221–3245.

[17] *GAIA mission*. https://www.cosmos.esa.int/gaia. dostęp 20.04.2019. 2018.

[19] I. Borg, P. J. F. Groenen i P. Mair. *Applied Multidimensional Scaling and Unfolding*. 2 wyd. New York: Springer, 2018.

[20] K. Bache i M. Lichman. *UCI Machine Learning Repository*. Spraw. tech. School of Information, Computer Science, University of California, Irvine, CA, USA, School of Information i Computer Sciences, 2015.

[21] A. Saxena, N. Pal i M. Vora. "Evolutionary methods for unsupervised feature selection using Sammon's stress function". *Fuzzy Information and Engineering* vol. 2 (2010), s. 229–247.

[22] C. Sammut i G. I. Webb, red. *Encyclopedia of Machine Learning and Data Mining*. 2 wyd. Springer Reference. New York: Springer, 2017. DOI: 10.1007/978-1-4899-7687-1.

[23] J. A. Lee i M. Verleysen. *Nonlinear Dimensionality Reduction*. 1st. Springer, 2007.

[24] K. Fukunaga i L. Hostetler. "The estimation of the gradient of a density function, with applications in pattern recognition". *IEEE Transactions on Information Theory* vol. 21. nr 1 (1975), s. 32–40. DOI: 10.1109/TIT.1975.1055330.

[25] A. Gramacki. *Nonparametric Kernel Density Estimation and Its Computational Aspects*. Studies in Big Data. Springer International Publishing, 2017.

[26] D. Domanska i M. Wojtylak. "Application of fuzzy time series models for forecasting pollution concentrations". *Expert Systems with Applications* (2012). DOI: 10.1016/j.eswa.2012.01.023.

[27] A. H. Gandomi i A. H. Alavi. "Krill herd: A new bio-inspired optimization algorithm". *Communications in Nonlinear Science and Numerical Simulation* vol. 17. nr 12 (2012), s. 4831–4845. DOI: 10.1016/j.cnsns.2012.05.010.

[28] S. Kulluk, L. Ozbakir i A. Baykasoglu. "Training neural networks with harmony search algorithms for classification problems". *Engineering Applications of Artificial Intelligence* vol. 25. nr 1 (2012), s. 11–19. DOI: https://doi.org/10.1016/j.engappai.2011.07.006.

[29] H. Szu i R. Hartley. "Fast simulated annealing". *Physics Letters A* vol. 122. nr 3 (1987), s. 157–162. DOI: https://doi.org/10.1016/0375-9601(87)90796-1.

[30] H. Awad, M. Z. Ali, J. J. Liang, B. Y. Qu i P. N. Suganthan. *Problem Definitions and Evaluation Criteria for the CEC 2017 Special Session and Competition on Single Objective Bound Constrained Real-Parameter Numerical Optimization*. Spraw. tech. Nanyang Technological University, 2016.

[31] K. Kaipa i D. Ghose. *Glowworm Swarm Optimization: Theory, Algorithms, and Applications*. Studies in Computational Intelligence. Springer International Publishing, 2017.

[32] X.-S. Yang. "Flower Pollination Algorithm for Global Optimization". *Unconventional Computation and Natural Computation*. Red. J. Durand-Lose i N. Jonoska. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, s. 240–249. DOI: 10.1007/978-3-642-32894-7_27.

[33] W. M. Rand. "Objective Criteria for the Evaluation of Clustering Methods". *Journal of the American Statistical Association* vol. 66. nr 336 (1971), s. 846–850. DOI: 10.1080/01621459.1971.10482356.

[34] D. L. Davies i D. W. Bouldin. "A Cluster Separation Measure". *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. PAMI-1. nr 2 (1979), s. 224–227. DOI: 10.1109/TPAMI.1979.4766909.