Uniwersytet im. Adama Mickiewicza w Poznaniu
Wydział Matematyki i Informatyki

Zastosowanie teorii zbiorów rozmytych w klasyfikacji dominującej dyscypliny naukowej autorów w naukometrii

Application of Fuzzy Set Theory in Classifying Author's Dominant Scientific Disciplines in Scientometrics

Łukasz Szymula

Rozprawa doktorska z dziedziny nauk ścisłych i przyrodniczych w dyscyplinie Informatyka

Promotorzy:
prof. UAM dr hab. Krzysztof Dyczkowski
prof. dr hab. Marek Kwiek

Poznań, 2024

Autor uzyskał środki finansowe w ramach finasowania z projektu badawczego z Narodowego Centrum Nauki PRELUDIUM BIS 1 nr UMO-2019/35/0/HS6/02591 pt. „Produktywność badawcza i wzorce współpracy w nauce globalnej. Ujęcie teoretyczne” kierowanego przez prof. dr. hab. Marka Kwieka.

Chciałbym złożyć serdeczne podziękowania
dla Prof. Krzysztofa Dyczkowskiego
oraz Prof. Marka Kwieka
za cenne uwagi i życzliwą pomoc
w trakcie pisania tej pracy.

Podziękowania kieruję również
moim Rodzicom i Dziadkom za wsparcie
w dążeniu do wyznaczonych celów.

Streszczenie

Jednym z głównych zagadnień w naukometrii jest klasyfikacja dyscyplin naukowych autorów, co ma znaczący wpływ na ocenę ich dorobku naukowego. Tradycyjne metody, oparte na prostych algorytmach i ograniczonej interpretacji danych, często prowadzą do niejednoznaczności w klasyfikacji. W pracy zaproponowano zastosowanie teorii zbiorów rozmytych, stworzonej przez Lotfiego Zadeha, jako rozwiązanie tego problemu. Teoria ta pozwala na modelowanie nieprecyzyjności informacji i oferuje nowe perspektywy dla uzyskania jednoznacznej klasyfikacji dyscyplin w dużej skali.

Głównym celem pracy jest zbadanie możliwości wykorzystania teorii zbiorów rozmytych do ulepszenia algorytmu klasyfikacji dominującej dyscypliny naukowej autorów posługując się wartością modalną. Przeprowadzone przeze mnie badania miały na celu nie tylko teoretyczne zrozumienie wpływu zastosowania tej teorii na klasyfikację, ale również praktyczne sprawdzenie skuteczności proponowanych modyfikacji na podstawie pełnej bibliometrycznej bazy danych Scopus udostępnionej w platformie ICSR Lab, Elsevier. Hipotezy badawcze skupiały się zarówno na możliwości zwiększenia jednoznaczności klasyfikacji poprzez identyfikację kluczowych pojęć z obszaru naukometrii, jak i na określeniu podobieństwa klasyfikacji uzyskanej przez zaproponowane modyfikacje z powszechnie stosowanym podejściem, które nie uwzględnia nieprecyzyjności informacji.

W badaniu wykorzystano najpopularniejsze metody z teorii zbiorów rozmytych, dzięki którym usprawniono proces klasyfikacji dyscyplin naukowych autorów. Pierwsze zaproponowane rozwiązanie obejmowało wykorzystanie zmiennych lingwistycznych reprezentowanych przez trzy poziomy intensywności: niskie, średnie i wysokie. Kolejnym skutecznym rozwiązaniem okazało się zbudowanie sterowników rozmytych do tworzenia bardziej elastycznych reguł klasyfikacji. Ostatnie podejście obejmowało agregacje wartości rozmytych operatorami OWA. Dzięki tym trzem propozycjom, uzyskano bardziej jednoznaczną klasyfikację autorów w naukometrii. Wykorzystanie tej metody pozwoliło na poprawę jednoznacznej klasyfikacji z około 69% do ponad 95%. Dzięki zastosowanym modyfikacjom jednoznaczna klasyfikacja autorów wzrosłą zatem o ponad 37% w porównaniu do podejścia, które nie uwzględnia nieprecyzyjności informacji. Wyniki moich badań wskazują na znaczącą rolę stosowania teorii zbiorów rozmytych w naukometrii, co otwiera drogę do dalszych badań w obu dziedzinach.

Abstract

One of the main issues in Scientometrics is the classification of authors' scientific disciplines, which has a significant impact on the evaluation of their scientific output. Traditional methods, based on simple algorithms and limited interpretation of data, often lead to ambiguities in classification. This thesis proposes the use of fuzzy set theory, developed by Lotfi Zadeh, as a solution to this problem. This theory allows to model the imprecision of information and offers new perspectives for achieving unambiguous classification of disciplines on a large scale.

The main purpose of the work is to explore the possibility of using fuzzy set theory to improve the classification algorithm of the authors' dominant scientific discipline using modal value. The study I conducted aimed not only to theoretically understand the impact of application of this theory on classification, but also to practically test the effectiveness of the proposed modifications on the basis of the full bibliometric database Scopus provided in the platform ICSR Lab, Elsevier. The research hypotheses focused both on the possibility of increasing the unambiguity of the classification by identifying key concepts from the field of Scientometrics, and on determining the similarity of the classification obtained by the proposed modifications with the commonly used approach that does not take into account imprecision of information.

The study used the most popular methods from fuzzy set theory to improve classification process of the authors' scientific disciplines. The first proposed solution involved the use of linguistic variables represented by three levels of intensity: low, medium and high. Another effective solution was the construction of fuzzy controllers to create more flexible classification rules. The last approach involved aggregating fuzzy values with OWA operators. With these three proposals, a more unambiguous classification of authors in Scientometrics has been achieved. The use of this method improved the unambiguous classification from about 69% to over 95%. Thus, with the modifications used, the unambiguous classification of authors increased by more than 37% compared to an approach that does not take into consideration the imprecision of information. The results of my research indicate a significant role for the application of fuzzy set theory in Scientometrics, which opens the possibilities for further research in both fields.

Spis treści

Rozdział 1.
Wprowadzenie

Naukometria, będąca interdyscyplinarną dziedziną analizującą procesy naukowe za pomocą metod matematycznych, statystycznych i informatycznych jest kluczowym narzędziem w ocenie dynamiki badań naukowych. Współczesny rozwój technologii i dostęp do bibliometrycznych baz danych pozwalają na wyznaczanie nowych kierunków badań, które mogą znacząco wpłynąć na zrozumienie i ocenę wartości naukowej publikacji, obecnych trendów oraz poznania wzorców współpracy między naukowcami.

Dla naukometrii jednym z kluczowych wyzwań jest klasyfikacja dyscyplin naukowych autorów publikacji, co ma bezpośredni wpływ na ocenę ich dorobku naukowego i analizy obszarów tematycznych. Przedstawiona w pracy metoda jej klasyfikacji na podstawie wartości modalnej z dyscyplin czasopism, nazywana w pracy podejściem bazowym, oparta jest na prostych algorytmach i ograniczonej interpretacji dla bogatych danych. Podejście to wiąże się z dużą niejednoznacznością w przypisywaniu dyscyplin do autorów. Propozycją, która może rozwiązać ten problem, jest zastosowanie teorii zbiorów rozmytych stworzonej przez Lotfiego Zadeha (Zadeh, 1965). Teoria ta, umożliwiając modelowanie nieprecyzyjności informacji, oferuje nowe perspektywy dla klasyfikacji dyscyplin naukowych, opierając się na różnych dostępnych rozwiązaniach. Taka modyfikacja podejścia, nazywana w pracy podejściem rozmytym, może znacząco poprawić jednoznaczność w dokonywanej przez algorytm klasyfikacji obszarów tematycznych.

W niniejszej pracy postawiono za cel zbadanie możliwości wykorzystania metod teorii zbiorów rozmytych do modyfikacji i ulepszenia algorytmów klasyfikacji dyscyplin naukowych autorów w oparciu o wartość modalną. Przeprowadzone badania mają na celu nie tylko teoretyczne zrozumienie wpływu zastosowania teorii zbiorów rozmytych na klasyfikację dyscyplin, ale również praktyczne sprawdzenie skuteczności proponowanych modyfikacji w kontekście prawdziwych, dostępnych danych bibliometrycznych. Główne hipotezy badawcze skupiają się na możliwości zwiększenia jednoznacznej klasyfikacji przez zastosowanie metod zbiorów rozmytych poprzez identyfikację kluczowych zmiennych lingwistycznych, sterowników rozmytych, metod agregacji i funkcji wagowych. Dodatkowo za sprawą metryk ewaluacyjnych

w badaniu porównano wyniki pochodzące z proponowanych podejść wykorzystujących wybrane metody modelowania nieprecyzyjności informacji z metodą bazową. Zabieg ten zastosowano w celu wyszczególnienia kluczowych rozwiązań, przypadków, których aplikacje dążyłyby do przypisywania obszarów tematycznych zgodnych z klasami podejścia bazowego, a co za tym idzie uzyskania poprawy jednoznacznej klasyfikacji bez wprowadzania ingerencji w etykiety autorów jednoznacznie klasyfikowanych dwoma podejściami.

Rozdział drugi przedstawia ogólne pojęcia z obszaru naukometrii wskazując występowanie w nim pojęć o charakterze nieprecyzyjnym. Opisano również specyfikę największych globalnych baz bibliometrycznych oraz wskaźników naukometrycznych. Kolejną część rozdziału stanowi opis algorytmu klasyfikacji dominującej dyscypliny przez wartość modalną (podejście bazowe). Przedstawiono jego najważniejsze wady wskazując skalę problemu niejednoznaczności klasyfikacji. Dokonano identyfikacji głównych powodów ich występowania wraz z przedstawieniem propozycji jego modyfikacji o aspekt nieprecyzyjności informacji. Zdefiniowane zostały główne hipotezy badawcze oraz wyszczególniono metody ewaluacji jakości proponowanego podejścia rozmytego.

Rozdział trzeci szczegółowo omawia podstawowe pojęcia teorii zbiorów rozmytych, które są niezbędne na etapie modelowania niejednoznaczności informacyjnych w procesie wprowadzania zaproponowanych modyfikacji. W tym rozdziale przedstawiono trzy kluczowe elementy, na których opiera się algorytm w podejściu rozmytym: zmienne lingwistyczne, sterowniki rozmyte oraz operatory agregacji OWA. Ostatnia część rozdziału stanowi opis głównej wykorzystanej charakterystyki teorii zbiorów rozmytych, którą jest obliczanie mocy skalarnej zbiorów rozmytych. Dodatkowo, zaprezentowano wybrane funkcje wagowe, które są kluczowym elementem analizowanej charakterystyki.

Kolejne trzy rozdziały przedstawiają propozycje algorytmu klasyfikacji dominującej dyscypliny przez wartość modalną w podejściu rozmytym. W każdym rozdziale stosuje się inne rozwiązanie znane z teorii zbiorów rozmytych. Każdy rozdział prezentuje metodologię zaproponowanego podejścia oraz przedstawia wyniki z przeprowadzonego badania jednoznaczności przypisywania dyscypliny oraz oceny jakości podejścia w porównaniu do podejścia bazowego. Wyniki badań obejmują analizę gęstości jądrowej dziedziny dla wybranych funkcji wagowych oraz opis uzyskanej klasyfikacji. Analizie poddawano pięć najlepszych wyników dla każdego prezentowanego podejścia.

W rozdziale czwartym zaprezentowano modyfikację algorytmu z wykorzystaniem zmiennych lingwistycznych. W badaniu zaproponowano 7 zmiennych lingwistycznych: Cytowanie, FWCI 4-letnie, FWCI 5-letnie, FWCI bez ram czasowych, Percentyl czasopisma, Rok oraz Zespół. Każda z prezentowanych zmiennych lingwistycznych wystąpiła w trzech termach reprezentując odpowiednio niską, ani niską ani wysoką oraz wysoką przynależność do zbioru rozmytego. W części metodologicznej przedstawiono proces konstrukcji zmiennych lingwistycznych, opisano szczegółowo algorytm w podejściu bazowym oraz wyszczególniono warianty funkcji wagowych użytych na etapie zliczania częstości występowania dyscyplin naukowych autora. Drugą część rozdziału stanowi opis wyników uzyskanych badań.

Rozdział piąty prezentuje zastosowanie sterowników rozmytych wysokiego prestiżu publikacji. W badaniu wykorzystano cztery warianty sterownika oparte na dwóch zmiennych wejściowych: Cytowanie i Percentyl czasopisma, FWCI 4-letnie i Percentyl czasopisma, FWCI 5-letnie i Percentyl czasopisma oraz dla kombinacji FWCI bez ram czasowych i Percentyla czasopisma. W części metodologicznej opisano proces konstruowania sterowników rozmytych, tworzenia zbioru reguł oraz zaprezentowano wersję algorytmu w podejściu bazowym. Analogicznie dla badania wykorzystania zmiennych lingwistycznych w algorytmie ostatnia część rozdziału stanowi opis wyników uzyskanych dla proponowanego podejścia.

W rozdziale szóstym przedstawiono modyfikację algorytmu bazowego o wykorzystanie operatorów agregacji OWA wartości dwóch zmiennych lingwistycznych: Cytowania oraz Percentyla czasopisma w każdym termie. Podobnie jak w przypadku poprzednich dwóch rozdziałów opisano metodologię dla proponowanego podejścia oraz przedstawiono wyniki uzyskanych badań.

Ostatni rozdział zawiera kompleksowe podsumowanie uzyskanych wyników badań, odnosząc się do hipotez postawionych na wstępie pracy. Ponadto, przedstawia perspektywy dalszego rozwoju omawianego podejścia w aspekcie eksploracji nowych rozwiązań w teorii zbiorów rozmytych.

Rozdział 2.
Problem klasyfikacji dyscypliny autora w naukometrii – propozycje modyfikacji stosowanego algorytmu

2.1 Wprowadzenie do naukometrii

W tym rozdziale przedstawiono główne pojęcia dotyczące obszaru badań oraz zdefiniowano główne hipotezy badawcze pracy.

2.1.1 Naukometria i nieprecyzyjność informacji

Naukometria jest interdyscyplinarną dziedziną nauki, która zajmuje się pomiarowymi analizami procesów naukowych. Wykorzystuje ona metody matematyczne, statystyczne i informatyczne do analizy danych naukowych (Leydesdorff i Milojević, 2015). Jej zakres tematyczny jest rozległy i oferuje niezliczone możliwości zrozumienia dynamiki nauki. Rozwój technologii i dostęp do coraz większych zbiorów danych sprawia, że naukometria staje się jeszcze bardziej dynamicznym obszarem dla nowych pytań i podejść badawczych. Wiele takich badań porusza kwestie najbardziej wpływowych, wartościowych publikacji, autorów oraz instytucji naukowych. Analizuje się różnorodne parametry, takie jak liczba cytowań, wskaźniki współpracy, prestiż instytucji, aby ocenić jakość i znaczenie prac naukowych. Wyszukiwanie "tego, co najlepsze" w kontekście badań bibliometrycznych obejmuje również określanie trendów naukowych, kluczowych tematów oraz ewolucji dyscyplin naukowych.

Jednym z najważniejszych tematów w naukometrii jest analiza cytowań. Heersmink i inni dokonali mapowania pojęć etyki komputerowej i informacyjnej ukazując najistotniejsze tematy i tendencje badawcze na podstawie najczęściej cytowanych prac (Heersmink i inni, 2011). Podobnie Faraji i inni określili strukturę intelektualną literatury dotyczącej kapitału intelektualnego analizując cytowania i współcytowania (Faraji i inni, 2022). W pracy Kuta i Pietruchy-Urbanik na podstawie analizy cytowań najpopularniejszych prac nad energią odnawialną wskazano perspektywy i dalsze kierunki branży energetycznej (Kuta i Pietrucha-Urbanik, 2024). Na podstawie cytowań publikacji z obszaru badań nad celami zrównoważonego rozwoju Sweileh

dokonał identyfikacji najbardziej wpływowych naukowców (Sweileh, 2020). Kolejnym szeroko poruszanym tematem jest badanie współpracy naukowej. Lee i Bozeman oceniali wpływ badań współautorskich na produktywność naukową, analizując czynniki, które mogą wpływać na skuteczność współpracy (Lee i Bozeman, 2005). Katz i Hicks określili związek pomiędzy typami współpracy geograficznej a cytowaniem (Katz i Hicks, 1997). Kwiek i Roszka wskazali różnice pomiędzy kobietami i mężczyznami w skłonności do publikowania bez zawierania współpracy (Kwiek i Roszka, 2022b). Inny aspekt badań dotyczy znaczenia prestiżowych instytucji w kształtowaniu globalnej nauki. Organizacje w oparciu o różne czynniki tworzą rankingi światowych instytucji naukowych, jak np. SCImago, ranking instytucji Times Higher Education, Ranking Szanghajski Uniwersytetów lub ranking Lejdejski, które są wykorzystywane w badaniach naukometrycznych (Jons i Hoyler, 2013; Visser i inni 2021).

Wspólną cechą przedstawionych powyżej badań jest operowanie pojęciami jak: najbardziej cytowane publikacje, największa produktywność, najbardziej prestiżowe czasopisma i instytucje lub mało znaczące publikacje. Są to pojęcia nieprecyzyjne (stopniowalne), których możliwości modelowania dostarcza dziedzina sztucznej inteligencji: teoria zbiorów rozmytych zaproponowana przez Lotfiego Zadeha (Zadeh, 1965). Wielokrotnie w określaniu przynależności obserwacji do grupy najbardziej wpływowych publikacji (w zależności od podejmowanego problemu badawczego) stosuje się metody odcinania lub tworzenia i rankingowania grup percentylowych. Rozwiązania zaproponowane przez teorie zbiorów rozmytych nie są podejmowane w badaniach naukometrycznych, a ze względu na ciągły charakter definiowanych pojęć stwarzana jest przestrzeń dla ich aplikacji.

Badania wykazały ogromny wpływ teorii zbiorów rozmytych na wiele dziedzin nauki, technologii i możliwości praktycznego zastosowania. Wykorzystując kontrolery rozmyte Dyczkowski zaproponował metody umożliwiające rozpoznawanie znaków drogowych w czasie rzeczywistym (Dyczkowski, 2011), a Eze i inni zaproponowali system kontroli sygnalizacji świetlnej na skrzyżowaniach w zależności od natężenia ruchu (Eze i inni, 2014). Doctor i inni zaproponowali kontrolery rozmyte do regulowania stanu pacjenta podczas śpiączki farmakologicznej (Doctor i inni, 2016). Podobnie również w pracach Dyczkowskiego oraz Żywicy podejmowano tematykę diagnozowania raka jajnika i proponowania metod leczenia z wykorzystaniem przedziałowych zbiorów rozmytych (Dyczkowski, 2018; Żywica i inni, 2016). Kacprzyk i Wilbik na przykładzie analiz notowań funduszy inwestycyjnych zaproponowali metodę

lingwistycznego podsumowywania szeregów czasowych z wykorzystaniem rozmytych kwantyfikatorów opartych na agregacji trendów (Kacprzyk i Wilbik, 2009). Azahar i inni oraz Szulczyński i inni wykorzystali klasyfikatory rozmyte do rozpoznawania mieszanin gazowych przez czujniki elektryczne w pomieszczeniach (Azahar i inni, 2015; Szulczyński i inni, 2018). Są to tylko nieliczne przykłady zastosowania teorii zbiorów rozmytych z wielu skutecznie stosowanych w praktyce.

2.1.2 Bibliometryczne bazy danych

Bibliometryczne bazy danych, nazywane inaczej jako bazy cytowań lub indeksy cytowań, to specjalistyczne zbiory informacji, które służą do analizy i oceny publikacji naukowych, autorów, instytucji i czasopism oraz ich wpływu w środowisku akademickim i badawczym. Głównym celem tych baz jest udostępnienie danych umożliwiających pomiar produktywności naukowej, ocenę jakości badań oraz analizę trendów naukowych i współpracy na wielu płaszczyznach.

Zbiory te, reprezentowane na poziomie publikacji naukowych, zawierają informacje bibliograficzne, takie jak tytuł, rok wydania, dane autorów, źródło, typ pracy, afiliacje, słowa kluczowe, streszczenia prac i język publikacji oraz ulokowanie publikacji w czasopiśmie. W przypadku największych baz danych dodatkowo mogą obejmować referencje, źródła finansowania, dane autorów korespondujących, identyfikatory wydawnictw i numery książek oraz nazwę kolekcji w przypadku, gdy zbiór stanowi agregację wielu indeksów cytowań (Birkle i inni, 2020; Noruzi, 2005).

Istnieje duża liczba baz bibliometrycznych. Obejmują one zasięgi globalne, narodowe (np. polska baza cytowań POL-index) lub wybrane obszary tematyczne (np. PubMed). Przykładowo do największych i najpopularniejszych globalnych multidyscyplinanych baz cytowań należą:

- Scopus – jest globalną bazą danych prowadzoną przez wydawnictwo Elsevier od 2004 roku. Obejmuje głównie literaturę z zakresu nauk ścisłych, przyrodniczych, inżynierii, medycyny (STEMM; science, technology, engieenering, mathematics, medicine) oraz w mniejszym stopniu z nauk społecznych, sztuki i humanistyki. Zawiera ponad 80 milionów rekordów obejmujących artykuły z ponad 25 tys. aktywnie działających recenzowanych czasopism, publikacje konferencyjne, książki, listy i patenty naukowe (Baas i inni, 2020; Burnham, 2006),
- Web of Science - jest najstarszym, globalnie uznawanym i autorytatywnym repozytorium publikacji naukowych i cytowań, założonym przez Eugene'a Garfielda w 1964 roku, aktualnie

prowadzona przez spółkę analityczną Clarivate. Zawiera ponad 155 milionów rekordów obejmujących artykuły z czasopism, książki, materiały konferencyjne, patenty oraz zestawy danych z 34 tys. tytułów źródłowych. Baza rozszerza zawartość Core Collection poprzez hosting baz danych innych dostawców, takich jak BIOSIS Citation Index, Chinese Science Citation Database, Russian Science Citation Index, oraz SciELO Citation Index (Birkle i inni, 2020; Schnell, 2017).

- OpenAlex – jest otwartą bazą danych, która ma na celu zapewnienie swobodnego dostępu do szerokiej gamy informacji naukowych obejmującej publikacje, autorów, instytucje, czasopisma oraz koncepty badawcze. Została stworzona jako alternatywa i kontynuacja projektu Microsoft Academic Graph. Oferuje darmowy i otwarty dostęp do swoich zasobów, co wspiera ideę otwartego dostępu do wiedzy naukowej. Wykorzystuje algorytmy przetwarzania języka naturalnego do indeksowania i organizacji danych. Jest największą bazą cytowań zawierającą ponad 240 milionów publikacji (Priem i inni, 2022; OpenAlex, 2024),
- Dimensions – baza wydawana przez Digital Science, łączy w sobie informacje na temat grantów, publikacji, cytatów, badań klinicznych, patentów oraz dokumentów politycznych. Rozwijana jest we współpracy z sześcioma firmami rodziny Digital Science. Zawiera około 153 milionów rekordów, w tym 100 milionów publikacji, 4,6 miliona grantów, 38 milionów patentów, 455 tysięcy badań klinicznych oraz 422 tysiące dokumentów politycznych. Charakteryzuje się szerokim zakresem dziedzin naukowych, dostarczając danych z ponad 340 instytucji finansujących oraz pokrywając metadane dla 100 milionów publikacji z ponad 50 tysięcy tytułów źródłowych (Hook i inni, 2018; Dimensions, 2019).

Bibliometryczne bazy danych udostępniane są poprzez platformy webowe, interfejsy programowania aplikacji lub w postaci plików tekstowych. Wybrane obiekty, takie jak np. publikacja, autor, źródło lub instytucja mogą być dodatkowo oznaczane za pomocą unikalnych identyfikatorów. Pozwalają one w takich bazach na wydajne zarządzanie, wyszukiwanie, aktualizowanie, powiązanie danych bibliometrycznych oraz zmniejszają skalę problemu wysokiej dezambiguacji. Przykładowo baza danych Scopus identyfikuje publikacje poprzez nadanie im identyfikatorów EID oraz autorów, przydzielając im Scopus Author Id. Podobnie w przypadku OpenAlex wykorzystuje się OpenAlex Id dla publikacji oraz OpenAlex Author Id dla autorów. Identyfikowanie nie występuje jednak w każdej bazie, co może powodować pewne ograniczania dla stosowania wybranych rozwiązań (On i inni, 2005; Milojević, 2013).

2.1.3 Wskaźniki w naukometrii na poziome publikacji

Rolą wskaźników naukowych jest zapewnienie zrównoważonego, wielowymiarowego ujęcia w ocenie publikowanych badań. Obecnie w naukometrii można spotkać szereg propozycji na ocenianie nauki. W zależności od indywidualnych przypadków użycia, wydawnictwa, organizacje, społeczność naukowa, bibliotekarze i inni przedstawiciele środowisk akademickich proponują wskaźniki, których przedmiot stanowi inna jednostka analizy: publikacja, czasopismo, autor lub instytucja naukowa.

Poszczególne metryki naukowe mogą być wykorzystywane niezależnie od badanego poziomu lub obejmować wybrane z nich. Przykładowo wyznaczenie sumarycznej liczby cytowań może dotyczyć wszystkich czterech poziomów, wartość indeksu Hirscha jest użyteczna wyłącznie na poziomie autora, instytucji lub czasopisma. Większość metryk naukowych może występować na innym poziomie poprzez użycie różnych metod agregacji. Przykładowo wyznaczanie liczby cytowań autora uzyskuje się na podstawie sumowania cytowań prac autora. Wyznaczanie wskaźnika znormalizowanego cytowania do dyscypliny autora polega na uśrednieniu tych wskaźników wyznaczonych na poziome publikacji, a określanie wskaźników geograficznej kolaboracji autora opiera się na ustalaniu odsetka prac z przypisanym danym typem geograficznej współpracy z pełnego portfolio autora. Możliwe jest również przypisywanie wartości wskaźnika wybranej jednostki na jednostkę innego poziomu. Przykładowo wskaźnik metryki percentyla wartości CiteScore czasopisma może reprezentować wartość atrybutu prestiżu w publikacji.

Poniżej zaprezentowano kilka wybranych metryk na poziomie publikacji, które zostały zaaplikowane w proponowanym algorytmie opisanym w rozdziale 2.3.3 z wykorzystaniem bazy Scopus (Elsevier, 2018):

- Cytowanie – metryka stanowiąca sumaryczną liczbę cytowań, odwołań publikacji w innych publikacjach z bazy, uzyskaną do momentu udostępnienia zbioru bibliometrycznego,
- Wskaźnik FWCI n-letni – indeks cytowań znormalizowany do dyscypliny naukowej. Wskaźnik został zaproponowany na potrzeby ewaluacji publikacji w bazie Scopus. Stanowi on stosunek liczby uzyskanych cytowań przez pracę do średniej liczby cytowań prac z tej samej dyscypliny oraz typu w tym samym roku i następujących n-1 lat. Metryka FWCI występuje w trzech wariantach ze względu na przedział czasowy: FWCI 4-letnie, FWCI 5-letnie oraz FWCI bez ram czasowych. FWCI publikacji równe jeden oznacza cytowanie nie większe niż cytowanie

prac podobnego typu z tego samego obszaru. Wartość wskaźnika FWCI ponad wartość jeden określa, o ile procent praca jest bardziej cytowana w porównaniu do globalnej średniej. Uzyskanie wartości z przedziału [0, 1) określa cytowanie mniejsze niż dla średniej globalnej.

- Liczba autorów – to wartość określająca liczbę autorów w pracy. Liczbę autorów określa się na podstawie wybranego atrybutu, na przykład poprzez zliczenie identyfikatorów autorów w pracy lub korzystając z pełnych danych autorów obejmujących ich imię i nazwisko. Wybór podejścia może wpływać na wynik ze względu na problem braku tworzenia identyfikatorów dla niektórych autorów pracy lub organizacji.
- Percentyl czasopisma publikacji – wartość centylu czasopisma metryki CiteScore, zaproponowana została jako autorska miara przez Elsevier'a. Metryka CiteScore wyliczana jest dla każdego roku. Zlicza cytowania otrzymane w danym roku dzieląc je przez liczbę dokumentów opublikowanych w trzech poprzedzających latach. Percentyl wartości metryki CiteScore wyznaczany jest dla każdej dyscypliny przypisanej do czasopisma na poziomie 334 dyscyplin ASJC. W celu nadania publikacji wartości metryki percentyla CiteScore wybiera się wartość tej metryki dla roku wydania publikacji. Jest to stosunkowo nowa metryka, wyliczana od 2010 roku. Aby przypisać tę wartość do wcześniejszych publikacji stosuje się metodę pierwszej niezerowej wartości, największej osiągniętej wartości lub percentyla wyznaczonego dla ostatniego dostępnego roku.

Liczba proponowanych metryk przez społeczność akademicką jest bardzo duża i z każdym rokiem rozszerza się o kolejne. Istniejące już metryki są stale udoskonalane poprze tworzenie ich kolejnych wariantów. Tematyka metod oceny badań naukowych jest podejmowana w takich pracach jak Harzing, 2010; Moed i inni, 2005, Moed, 2005; Garfield, 2006; Waltman i inni, 2017, Ding i inni, 2014; Ye, 2017; Sugimoto i Lariviere, 2017.

2.2 Podejścia w klasyfikacji dyscyplin w bazach bibliometrycznych na poziomie publikacji

Popularnym zagadnieniem obejmującym analizy w obszarze naukometrii jest badanie różnic na poziomie dyscyplin naukowych. Głównym podejściem przypisywania dyscypliny do publikacji jest jej przydział zgodnie z ustalonym schematem twórców danego zbioru. Poniżej zestawiono schematy zastosowane w wybranych największych bibliometrycznych bazach danych.

Baza Scopus wykorzystuje schemat ASJC (All Science Journal Classification). Obejmuje on 26 dyscyplin naukowych czterech głównych gałęzi wiedzy (nauki przyrodnicze, nauki fizyczne, nauki o zdrowiu oraz nauki społeczne i humanistyczne) oraz łącznie 334 poddziedziny (Elsevier, 2023). W przypadku darmowej wersji produktu Dimensions dostępny jest schemat Fields of Research. Płatny dostęp do bazy Dimensions rozszerza je o kolejne schematy: Research, Condition, and Disease Categorization; Health Research Classification System; Broad Research Areas; Health Research Areas; Common Scientific Outline oraz Units of Assessment (Dimensions, 2022). Web of Science (Clarivate) operuje dziewięcioma głównymi obszarami trzech nadrzędnych kolekcji: bazy publikacji nauk humanistycznych, nauk przyrodniczych oraz nauk społecznych z 250 poddziedzinami. Ponadto Clarivate pozwala na stosowanie metod mapowania dyscyplin naukowych pomiędzy innymi schematami a własną klasyfikacją, na przykład ze schematu Fields of Research, kategorii OECD, polskiego schematu klasyfikacji PL19 i innych (Clarivate, 2021). Największa bibliometryczna baza OpenAlex, pierwotnie klasyfikowała publikacje zgodnie z wielopoziomową hierarchią dyscyplin definiowaną przez Microsoft Academic Graph. Kategoryzacja dyscyplin na bieżąco ulega rozszerzaniu poprzez wyszukiwanie nowych klastrów obszarów tematycznych metodami sztucznej inteligencji (OpenAlex, 2024).

Kolejne podejście klasyfikacji dyscypliny publikacji opiera się na częstości termów wskazywanych bezpośrednio przez autorów w pracy. Metadane bibliometrycznych baz danych zawierają takie elementy jak tytuł, słowa kluczowe i streszczenie (abstrakt) publikacji. Przykładowo Meen i inni operowali na ponad dziesięciu tysiącach publikacji o ujednoliconym systemie języka medycznego (UMLS) z bazy Web of Science. Tworzyli oni terminy na podstawie słów kluczowych i abstraktów. Przypisując najczęściej występujący termin w publikacji dokonywali analizy grafowej współwystępowania słów kluczowych obszaru UMLS z wykorzystaniem oprogramowania VOSViewer (Meen i inni, 2020). Podobny zabieg (analiza częstości terminów) wykonali Purnomo i inni, Makabate i inni dla zbioru bazy Scopus, jednak odpowiednio dla pięciu tysięcy publikacji z zakresu analizy danych (Purnomo i inni, 2020) i tysiąca publikacji z obszaru modelowania budownictwa i projektowania konstrukcji (Makabate i inni, 2022). W przypadku większych zbiorów danych Bhatt i inni wykorzystali 85 tysięcy publikacji opisujących zastosowanie głębokiego uczenia w medycynie z bazy ScienceDirect do klasyfikacji modeli uczenia głębokiego (Bhatt i inni, 2021).

Poza metodami wykorzystującymi frekwencję wybranych terminów w słowach kluczowych i streszczeniach wykorzystuje się również zaawansowane metody uczenia maszynowego. Eykens i inni zastosowali dwa podejścia: naiwny klasyfikator bayesowski i wzmacnianie gradientowe do klasyfikacji publikacji z obszaru nauk społecznych dla pełnych danych tekstowych stu tysięcy prac z bazy ProQuest. Wyniki własnej klasyfikacji porównali z klasyfikacją VODS (Flemish Research Discipline Standard, FRIS) Dla metody naiwnej Bayesa uzyskali klasyfikację z wynikiem F1 do 0.42 i Accuracy do 0.24. W przypadku metody wyostrzenia gradientu osiągnęli do 0.55 wartości metryki F1 i 0.46 Accuracy (Eykens i inni, 2021). Kandimalla i inni wykorzystali zbiór 9 milionów publikacji z bazy Web of Science. Zaproponowali metodę klasyfikacji głębokimi uważnymi sieciami neuronowymi (DANN) opierając się wyłącznie na streszczeniach publikacji. Ewaluacji poddali 9 modeli wykorzystujących metody: lasów losowych, naiwnego klasyfikatora bayesowskiego, maszyny wektorów nośnych, regresji logistycznej i konwolucyjnych sieci neuronowych. Dla najlepszych modeli uzyskali klasyfikację z wartością do 0.76 metryki F1 w porównaniu do klasyfikacji ustalonej przez schemat bazy Web Of Science Core Collection ze 104 wybranymi obszarami badawczymi (Kandimalla i inni, 2021). Dunham i inni wykorzystali 4 modele BERT pretrenowane na pełnych tekstach z bazy Semantic Scholar uczące się na tekstach z korpusu publikacji arXiv. Klasyfikacji metodami BERT poddano 38 milionów publikacji pochodzących z baz Dimensions, Microsoft Academic Graph (MAG) oraz Web of Science z lat 2010 do 2019. Zaproponowane modele zostały porównane z metodą klasyfikacji kategorii dyscyplin w schemacie MAG. W wyniku ewaluacji otrzymali metrykę F1 w przedziale od 0.75 do 0.84 (Dunham i inni, 2020). Daradkeh i inni poddali ewaluacji konwolucyjne sieci neuronowe (CNN). Posłużyli się zbiorem siedmiu i pół tysiąca publikacji z platformy Scopus, ProQuest, i EBSCOhost z obszaru danologii, które zanotowali manualnie do 22 obszarów tematycznych. Zaproponowali oni 4 warianty konwolucyjnych sieci neuronowych. Najlepsze wyniki uzyskali dla modelu wykorzystującego terminy z tytułów, słów kluczowych, abstraktów i metadanych o afiliacjach autorów z F1 = 0.73. Podobnie wysoką klasyfikację uzyskali opierając się wyłącznie na tytułach i abstraktach z F1=0.70 oraz stosując listy terminów zatrzymania zbioru snowball F1=0.72. Własne modele CNN zestawili z trzema innymi podejściami: naiwnym klasyfikatorem bayesowskim, maszyną wektorów nośnych oraz metodą k-najbliższych sąsiadów otrzymując odpowiednio wartości F1=0.65, F1=0.58 oraz F1=0.50 (Daradkeh i inni, 2022).

Powyższe podejścia dokonują klasyfikacji dyscypliny na poziome publikacji. Wadą większości bibliometrycznych baz danych jest niskie ujednoznacznienie identyfikatora autora w metadanych publikacyjnych. Inny problem stanowi brak otwartego dostępu do dużych baz danych, takich jak Scopus czy Web of Science. Pozostałe otwarte bazy danych w większości zawierają jedynie imię i nazwisko autora, i nie obejmują ich własnymi identyfikatorami lub stosują metody probabilistyczne, co powoduje ich mniejsze ujednoznacznienie (On i inni, 2005; Milojević, 2013).

Jeżeli dysponuje się bibliometrycznymi bazami danych, w których podjęto problem ujednoznacznienia autorów, możliwe jest prowadzenie analiz na poziomie autora. Poniżej opisano najpopularniejszą metodę, która na podstawie publikacji autora klasyfikuje jego dominującą dyscyplinę w oparciu o wartość modalną.

2.3 Klasyfikacja dominującej dyscypliny autora w oparciu o wartość modalną

2.3.1 Algorytmy klasyfikacji dominującej dyscypliny autora

Jednym z najpopularniejszych sposobów przypisywania dyscypliny do autora, opierając się wyłącznie na danych pochodzących z bibliometrycznych baz danych jest zastosowanie wartości modalnej z portfolio publikacyjnego (nazywane dalej algorytmem lub podejściem bazowym). W takim podejściu dla każdej dyscypliny, która wystąpiła dla autora (identyfikowanego przy użyciu dowolnego wybranego atrybutu) zliczana jest częstość jej wystąpienia (każde wskazanie dyscypliny w publikacji liczone jest jako wartość 1). Metadane portfolio autora w większości bibliometrycznych baz danych zawierają atrybut określający dyscyplinę publikacji (zgodnie z przyjętą klasyfikacją opisaną w rozdziale 2.2) lub wybierane są przez badaczy w oparciu o inną preferowaną metodologię (na przykład klasyfikacja na podstawie słów kluczowych). Autorowi przypisywana jest dyscyplina, która najczęściej występowała w portfolio publikacyjnym (wartość modalna).

Dla takiego podejścia mogą wystąpić sytuacje, w których autor zostanie zaklasyfikowany do więcej niż jednej dyscypliny naukowej lub nie otrzyma ani jednej dyscypliny (wynikającej z braku przypisania dyscyplin na poziomie publikacji). W takiej sytuacji autor najczęściej nie jest

uwzględniany w badanej próbie ze względu na brak jednoznacznie zaklasyfikowanej etykiety. Alternatywnie stosuje się losowanie (Abramo, 2020; Boekhout i inni, 2022; Kwiek i Roszka, 2022a,b 2023a,b).

Algorytm 1. Algorytm klasyfikacji dominującej dyscypliny autora w podejściu bazowym

Wejście: zbiór publikacji w bazie Scopus
Wyjście: zbiór par autor-dyscyplina

1. Wykonaj normalizację tabeli do pierwszej postaci normalnej ze względu na identyfikatory autorów i listę dyscyplin
2. Dla każdego autora i jego dyscyplin oblicz częstość występowania dyscypliny
3. Dla każdego autora wyznacz maksymalną z wartości częstości dyscyplin (wartość modalna)
4. Dla każdego autora wybierz dyscypliny, dla których wartość częstości odpowiada wartości modalnej
5. Dokonaj filtracji tabeli do autorów posiadających jedną zaklasyfikowaną dyscyplinę

Dla klasyfikacji dyscypliny w podejściu bazowym dane wyjściowe stanowi portfolio publikacyjne autora. W przypadku badań naukometrycznych można spotkać się z różnymi podejściami określającymi przedział czasowy uwzględnianego portfolio publikacyjnego. Wybór podejścia zależy od indywidualnych preferencji, jednak definiuje typ powstałej klasyfikacji. Jedną z możliwości jest obejmowanie pełnego portfolio publikacyjnego (bez ograniczeń czasowych) interpretowane jako zaklasyfikowanie dyscypliny w oparciu o pełny dorobek publikacyjny (Abramo, 2020; Boekhout i inni, 2022; Kwiek i Roszka, 2022b, 2023a,b). Inna metoda obejmuje jedynie pierwszy rok działalności publikacyjnej autora. Z podejściem tym można spotkać się w przypadku raportów tworzonych przez wydawnictwo Elsevier (Elsevier, 2020a,b) i może być interpretowane jako pierwsza dyscyplina lub dyscyplina z początku kariery publikacyjnej. Inne podejścia obejmują ramy czasowe tożsame z zakresem lat, które obejmuje dane badanie (Kwiek i Roszka 2022a).

Inna metoda polega na klasyfikacji autora do dyscypliny na podstawie dyscyplin referencji z portfolio publikacyjnego. Metoda ta wprowadza jednak pewne ograniczenia na pojawiające się klasy-dyscypliny. Dla badanej bazy Scopus (jak również w innych bazach, na przykład Web of Science) główne referencje indeksowane są dla obszaru grupy STEMM, tj. nauka, technologia, inżynieria, matematyka i medycyna, co powoduje pozostawienie pozostałych referencji

bez zaklasyfikowanej dyscypliny, przeważnie z obszaru nauk humanistycznych i społecznych (Kwiek i Szymula 2023a,b, 2024).

2.3.2 Wady algorytmu

Efekt działania algorytmu sprawdzono w oparciu o bibliometryczną bazę danych Scopus, do której otrzymano dostęp w ramach umowy zawartej w sierpniu 2021 r. pomiędzy Centrum Studiów nad Polityką Publiczną Uniwersytetu im. Adama Mickiewicza w Poznaniu a International Center for the Study of Research (ICSR Lab) należącej do wydawnictwa Elsevier.

ICSR Lab to platforma, której celem jest wsparcie badań prowadzonych przez naukowców z obszaru naukometrii i pokrewnych nauk społecznych. Zapewnia ona dostęp do danych wydawnictwa Elsevier, takich jak Scopus, PlumX, SciVal oraz udostępnia środowisko do wykonywania obliczeń na dużych zbiorach danych w chmurze.

Dostęp do platformy odbywał się z wykorzystaniem środowiska Databricks z dostępem do pełnej bazy Scopus ulokowanej na usługach AWS (Amazon Web Services). Użytkownicy platformy ICSR Lab dysponowali klastrem w trybie standardowym z Databricks Runtime w wersji 12.2 LST ML, wykorzystującym technologię Apache Spark w wersji 3.3.2, Scala 2.12 oraz instancję i3.2xlarge z 61 GB pamięci, ośmioma rdzeniami, od jednego do sześciu procesów roboczych dla węzła roboczego oraz instancji c4.2xlarge z 15 GB pamięci, czterema rdzeniami dla węzła sterownika. Generowanie wyników i prowadzenie analiz możliwe było z poziomu notatnika poprzez tworzenie odpowiedniego kodu źródłowego w jednym z czterech dostępnych języków programowania (Python, SQL, Scala lub R).

Działanie algorytmu sprawdzono na podstawie pełnej bazy Scopus z datą udostępnienia zbioru 21 października 2022. Jako zbiór wyjściowy wybrano wszystkie publikacje z bazy Scopus do roku 2021 włącznie z przypisaną dyscypliną ASJC czasopisma (26 dyscyplin naukowych), N=60,987,987 (z wykluczeniem publikacji z 2022 roku ze względu na brak danych dla pełnego roku). Algorytm w podejściu bazowym zaklasyfikował dominujące dyscypliny do 36,010,088 autorów. Około 69.25% (N=24,938,113) z nich zostało zaklasyfikowanych do jednej dominującej dyscypliny, pozostałe 30.75% (N=11,071,975) uzyskało więcej niż jedną dyscyplinę (Tab. 2.1).

Tabela 2.1 Rozkład klasyfikacji autorów do liczby dominujących dyscyplin w podejściu bazowym, wyniki własne

Liczba dyscyplin	Liczba autorów	% kol.
1	24,938,113	69.25
2	7,655,307	21.26
3	2,538,076	7.05
4	615,292	1.71
5	198,175	0.55
6	48,216	0.13
7	11,773	0.03
8	3,528	0.01
9	1,382	0.00
10	178	0.00
11	36	0.00
12	11	0.00
13	1	0.00
Całkowicie	36,010,088	100

Liczba autorów, które zostały zaklasyfikowane do więcej niż jednej dyscypliny stanowi duży podzbiór badanej próby. Jednym z powodów, dla których autor zawsze zostanie zaklasyfikowany do kilku dominujących dyscyplin, jest posiadanie wyłącznie jednej publikacji w portfolio, która pochodzi z czasopisma zaklasyfikowanego do kilku dyscyplin naukowych. Takie obserwacje stanowią 68.59% (N=7,704,780) niejednoznacznie zaklasyfikowanych obserwacji. Drugim powodem jest posiadanie w portfolio publikacji, które zawsze klasyfikowane są do tych samych kilku dyscyplin naukowych. Taki przypadek obejmuje około 9.39% (N=1,039,587) autorów. Pozostałe 21.02% obejmuje sytuację, w której dla portfolio autora istnieją dwie lub więcej dominujące dyscypliny przy jednoczesnym występowaniu dyscyplin, które pojawiają się z mniejszą częstością niż dyscypliny dominujące (Tab. 2.2).

Tabela 2.2 Rozkład klasyfikacji autorów niejednoznacznie zaklasyfikowanych do dominującej dyscypliny naukowej (N=11,071,975), wyniki własne

Liczba dyscyplin	Liczba autorów z jedną publikacją w portfolio		Liczba autorów z więcej niż jedną publikacją w portfolio				Całkowicie
			wszystkie publikacje z tymi samymi dyscyplinami		co najmniej jedna dyscyplina niedominująca		
	N	% wiersz.	N	% wiersz.	N	% wiersz.	N
2	5,271,811	68.86	778,886	10.17	1,604,610	20.96	7,655,307
3	1,894,340	74.64	215,065	8.47	428,671	16.89	2,538,076
4	394,629	64.14	34,343	5.58	186,320	30.28	615,292
5	116,562	58.82	9,603	4.85	72,010	36.34	198,175
6	22,261	46.17	1,293	2.68	24,662	51.15	48,216
7	3,620	30.75	285	2.42	7,868	66.83	11,773
8	953	27.01	75	2.13	2,500	70.86	3,528
9	604	43.70	37	2.68	741	53.62	1,382
10	0	0.00	0	0.00	178	100.00	178
11	0	0.00	0	0.00	36	100.00	36
12	0	0.00	0	0.00	11	100.00	11
13	0	0.00	0	0.00	1	100.00	1
Całkowicie	7,704,780	69.59	1,039,587	9.39	2,327,608	21.02	11,071,975

Z powyższych analiz wynika, że algorytm wyznaczania dominującej dyscypliny przez wartość modalną nie pozwala na dokonanie jednoznacznej klasyfikacji dużej części autorów w bazie Scopus. Wykorzystanie podejścia bazowego nie umożliwia jednoznacznego zaklasyfikowania ponad 30% obserwacji. Główny problem (około 70% obserwacji) stanowią autorzy posiadający jedną publikację w portfolio publikacyjnym czyli tak zwani autorzy okazjonalni. W związku z tym podejście to wymaga ponownego zbadania, czy istnieją metody, których aplikacja w algorytmie bazowym pozwoliłyby na poprawę, czyli uzyskanie większej liczby jednoznacznie zaklasyfikowanych autorów do danej dyscypliny.

2.3.3 Modyfikacja algorytmu z uwzględnieniem nieprecyzyjności informacji i zdefiniowanie problemów badawczych

Ze względu na operowanie w naukometrii pojęciami o charakterze nieprecyzyjnym (rozdział 2.1.1) algorytm klasyfikacji dominującej dyscypliny w podejściu bazowym może zostać zaimplementowany w taki sposób, aby na etapie wyznaczania częstości danej dyscypliny wykorzystywał metody pochodzące z obszaru sztucznej inteligencji, a konkretniej z zakresu teorii zbiorów rozmytych, inteligentnych obliczeń w oparciu o wybrane atrybuty i wskaźniki naukowe opisane w rozdziale 2.1.2 oraz 2.1.3. Taka modyfikacja podejścia bazowego tworzy rozwiązanie nazywane algorytmem lub podejściem rozmytym. W związku z tym celem prowadzonych badań w pracy jest sprawdzenie, czy możliwe jest uzyskanie zwiększenia jednoznacznej klasyfikacji dyscyplin modyfikując algorytm w podejściu bazowym o metody pochodzące z obszaru teorii zbiorów rozmytych. W ramach dostępnych metod sprawdzeniu poddane zostaną trzy najpopularniejsze rozwiązania: (1) zmienne lingwistyczne, (2) sterowniki rozmyte, (3) agregacje wartości rozmytych z wykorzystaniem operatorów OWA. Główne hipotezy badawcze sformułowane są w następujący sposób:

- (H1) Czy zastosowanie metod z aparatu pojęciowego teorii zbiorów rozmytych w algorytmie klasyfikacji dominującej dyscypliny może jednoznacznie zaklasyfikować większą ilość obserwacji niż podejście bazowe?
- (H2) Które rozwiązania znane z teorii zbiorów rozmytych pozwalają na poprawę liczby jednoznacznie zaklasyfikowanych autorów w podejściu rozmytym?

- (H3) Które zmienne lingwistyczne, termy i funkcje wagowe są kluczowe w poprawie jednoznacznej klasyfikacji?
- (H4) Dla jakich parametrów wejściowych sterownika rozmytego i funkcji wagowych możliwe jest uzyskanie poprawy jednoznacznej klasyfikacji dyscyplin?
- (H5) Jak dobór wag w operatorze OWA wpływa na poprawę jednoznacznie zaklasyfikowanych autorów do dyscyplin?
- (H6) Czy klasyfikacja wykonana podejściem rozmytym przypisuje takie same klasy jak w przypadku podejścia bazowego?

2.3.4 Ograniczenia

Do przeprowadzania badania klasyfikacji dominującej dyscypliny w podejściu bazowym i rozmytym wykorzystano platformę ICSR Lab. Platforma oferowała dostęp do pełnej bibliometrycznej bazy danych Scopus poprzez środowisko Databricks. Jego zastosowanie oraz wymagania zawarte w umowie pomiędzy dostawcą danych i metod ich przetwarzania w chmurze, czyli Elsevier'em a Centrum Studiów nad Polityką Publiczną Uniwersytetu im. Adama Mickiewicza w Poznaniu wprowadziły pewne ograniczenia.

Dostęp do bazy Scopus możliwy był wyłącznie z poziomu Databricks. Nie dysponowano lokalną wersją bazy oraz nie było możliwe pobranie jej fragmentów. Wszelkie komendy odpowiadające za wyświetlanie danych w postaci tabelarycznej ograniczone były do maksymalnie pierwszych dwustu rekordów. Dane można było uzyskać w postaci wyników jako pliki csv lub poprzez ich wizualizację w środowisku Databricks. Wyniki stanowiące małą liczbę rekordów można było pozyskać przy użyciu komendy wyświetlającej daną tabelę. W pierwszej kolejności tworzenie skryptów bazowało na dostępie do zbioru 10% bazy Scopus. O możliwości dostępu do pełnej bazy i pobrania danych na poziome wyników decydowały osoby reprezentujące platformę ICSR Lab. Wykonywały one recenzję skryptu oraz sprawdzały, czy uzyskany wynik nie narusza warunków umowy lub nie wyprowadza zbyt dużej ilości danych poza platformę. Zespół Elsevier monitorował czas wykonania skryptów, ich złożoność oraz znaczenie wyników w badaniach naukometrycznych. Dodatkowo w umowie wskazany był maksymalny czas dostępu do platformy (użycia jednostek obliczeniowych Databricks) w danym miesiącu współpracy.

Zastosowanie środowiska Databricks oraz dostęp do bazy w modelu rozproszonym wymagał posługiwania się platformą programistyczną PySpark (lub rozwiązaniami z podobną

implementacją: Spark, RSpark lub SQL). Obecnie nie ma bibliotek wspierających implementację rozwiązań teorii zbiorów rozmytych, które operowałyby na strukturach wykorzystywanych przy obliczeniach rozproszonych. Zastosowanie jakiejkolwiek biblioteki języka Python (jako, że instrukcje tego języka można uruchamiać w platformie PySpark) wymagałoby przejścia na inną strukturę np. ramki danych Pandas lub Numpy, a co za tym idzie stratę rozproszoności obliczeń, wydłużenie czasu uruchomienia skryptów lub brak możliwości alokowania pamięci dla przypisanego klastra. W związku z tym skrypty pozyskujące wyniki wymagały własnej implementacji rozwiązań teorii zbiorów rozmytych.

Mając na uwadze czas i wymagania platformy podjęto pewne działania mające na celu optymalizację procesu implementacji skryptów. Do modelowania funkcji przynależności zbiorów rozmytych wykorzystano funkcje liniowe. Zaproponowane zmienne lingwistyczne występowały w trzech termach. Wyznaczano wartości rozmyte dla termu wysokiej przynależności do zbioru rozmytego. Wartość dla termu niskiej przynależności uzyskano poprzez zastosowanie negacji Łukasiewicza. Trzeci term zbudowano poprzez zastosowanie t-normy minimum. Tak przyjęta konwencja charakteryzowała się prostotą w implementacji oraz niską złożonością obliczeniową. Podobnie w sterownikach rozmytych wejście zostało organiczne do dwóch zmiennych oraz 9 reguł. Stosowano również proste rozwiązania operujące na t-normach minimum i t-konromach maksimum, które posiadały swoje przełożenie na instrukcje platformy PySpark.

Możliwości czasowe określały również ilość powstałych wariantów funkcji wagowych. W pierwszej kolejności zdecydowano się na wybór trzech wariantów wartości potęgowej: 1, 2, -1, jako że reprezentują jedne z głównych wykładników funkcji wielomianowych. Podobnie przy określaniu wartości progujących wybrano optymalną liczbę pięciu wartości z zakresu (0, 1), aby analiza obejmowała zarówno warianty z mniejszymi i większymi wartościami punktu odcięcia oraz wybrane wartości progowania nie tworzyły funkcji wagowych jednoznacznych z innymi typami. Ze względu na ograniczenia czasowe wybrano również pięć głównych funkcji wagowych, które różniły się znacząco swoimi wzorami. W pracy nie badano funkcji stosujących ostre progowanie ze względu na małą zmianę w implementacji skryptu oraz podobną budowę formuły dla wybranych wartości progowych.

Wszelkie powyżej przedstawione pojęcia związane z teorią zbiorów rozmytych opisano w następnym rozdziale. Dokładny proces tworzenia zmiennych lingwistycznych, sterowników rozmytych oraz agregacji operatorami OWA zawarto w rozdziałach opisujących metodologię

odpowiednich badań tj. odpowiednio rozdziały od 4. do 6. w zależności od stosowanego rozwiązania z teorii zbiorów rozmytych.

Ograniczenie się do korzystania z platformy programistycznej PySpark określało również dostępne biblioteki programistyczne. Do oceny jakości klasyfikacji wybrano główną bibliotekę do Uczenia Maszynowego MLlib, która oferowała klasę MulticlassClassificationEvaluator do ewaluacji klasyfikacji wieloklasowej. Dostarcza ona w prosty sposób w postaci atrybutów bogatą ofertę metryk ewaluacyjnych. Do wyznaczania wartości tych metryk biblioteka MLlib stosuje metodę ważonego-averging'u. Pełny opis metod ewaluacji klasyfikacji zaprezentowano w następnym podrozdziale.

Wykorzystanie platformy ICSR Lab zapewniło dostęp do dużych danych bazy Scopus. W badaniach operowano 60-cioma milionami rekordów w modelu rozproszonym, które wymagały zalokowania ponad 400 GB pamięci tymczasowej. Przełożenie takiej specyfikacji na urządzenia lokalne byłoby aktualnie niemożliwe ze względu na ograniczania w dostępie do sprzętu technicznego. Ponadto dostęp do platformy ICSR Lab był bezpłatny, co było jednym z czynników, który zdecydował o wyborze tego rozwiązania. Odwzorowanie takiej specyfikacji wymagałoby ponoszenia wysokich opłat oraz zapewnienia innej bazy bibliometrycznej, co wydłużyłoby również czas prowadzonych badań. Ponadto zespół Elsevier zapewniał wsparcie w zakresie dostępu do platformy oraz udostępniał dane w postaci pozwalającej na ich natychmiastowe wykorzystanie w skryptach.

2.3.5 Metody ewaluacji klasyfikacji

Zarówno w wyniku zastosowania algorytmu klasyfikacji w podejściu bazowym, jak i w podejściu rozmytym rezultatem jest przypisanie do autora jednej z 26-ciu dyscyplin. Taka klasyfikacja stanowi przypadek klasyfikacji wieloklasowej. Porównanie wyników otrzymanych podejściem rozmytym z wynikami otrzymanymi w podejściu bazowym możliwe jest za sprawą wyznaczenia wartości klasyfikacyjnych metryk ewaluacyjnych dla skonstruowanego zbioru testowego. Opisy metryk ewaluacyjnych stanowią nieodłączną część prac dedykowanych zagadnieniom uczenia maszynowego. Dokładne opisy metryk ewaluacyjnych znajdują się na przykład w następujących pracach Bishop, 2006; Witten 2011; Japkowicz i Shah 2011; Murphy, 2012.

Każdy wariant aplikujący inną metodę w podejściu rozmytym może zwrócić różną liczbę zaklasyfikowanych obserwacji pozostawiając pozostałą część próby bez jednoznacznie przypisanej dyscypliny. Takie obserwacje można oznaczyć umownie dodatkową klasą (na przykład „brak klasy") lub ignoruje się je w ewaluacji pozostawiając obserwacje, które otrzymały klasyfikację (jednoznaczną) zarówno podejściem bazowym jak i rozmytym. W zależności od indywidualnych preferencji możliwe jest zastosowanie dowolnego podejścia, jednak ze względu na skalę problemu w algorytmie bazowym w pracy metryki ewaluacyjne wyznaczane będą dla zbioru obserwacji, które otrzymały klasyfikację (jednoznaczną) zarówno podejściem bazowym jak i rozmytym, i stanowić będą zbiór testujący dla badanego podejścia rozmytego.

Optymistycznie ujmując liczba jednoznacznie zaklasyfikowanych obserwacji podejściem bazowym i rozmytym odpowiadać może liczbie obserwacji jednoznacznie zaklasyfikowanych w podejściu bazowym. Może wystąpić również sytuacja, w której w wyniku klasyfikacji podejściem rozmytym traci się jednoznaczne zaklasyfikowanie obserwacji zaklasyfikowanej podejściem bazowym. W związku z tym zaproponowano stworzenie metryki wskazującej procentową utratę zaklasyfikowanych obserwacji podejściem bazowym w klasyfikacji podejściem rozmytym wyznaczanej jako:

$$Percentage\ loss = \frac{C_B - C_{B,F}}{C_B},$$

gdzie $C_{B,F}$ oznacza liczbę jednoznacznie zaklasyfikowanych obserwacji w podejściu bazowym i rozmytym, C_B oznacza liczbę autorów jednoznacznie zaklasyfikowanych podejściem bazowym.

Pierwotnie metryki ewaluacji klasyfikacji zostały zaproponowane na potrzeby ewaluacji klasyfikacji binarnej. Aby stosować istniejące metryki dla przypadku klasyfikacji wieloklasowej, zaproponowano wyszczególnienie poszczególnych klas i wyznaczenie tych metryk dla klasyfikacji binarnej w każdej klasie (przypisanie lub nieprzypisanie danej klasy do obserwacji w zbiorze testującym). Efekt tego zabiegu stanowi uzyskanie k wartości wybranej metryki w każdej z występujących k klas. Wyznaczanie wartości dowolnej metryki ewaluacyjnej możliwe jest w oparciu o wartości poszczególnych typów błędów macierzy konfuzji (dla problemu klasyfikacji binarnej). W skład macierzy konfuzji zalicza się następujące cztery typy błędów obejmujące wszystkie elementy zbioru testującego:

1. TP (True Positive) – liczba obserwacji klasy pozytywnej zaklasyfikowanych jako pozytywne,
2. FP (False Positive) – liczba obserwacji klasy negatywnej zaklasyfikowanych jako pozytywne,
3. TN (True Negative) – liczba obserwacji klasy negatywnej zaklasyfikowanych jako negatywne,
4. FN (False Negative) – liczba obserwacji klasy pozytywnej zaklasyfikowanych jako negatywne.

Jako pojedynczą wartość, która reprezentowałaby ewaluację klasyfikacji wieloklasowej zaproponowano metodę uśredniania wyników w oparciu o mikro, makro oraz ważony-averging wszystkich *k* wyznaczonych wartości (Japkowicz i Shah 2011; Murphy, 2012). Ze względu na prowadzenie analiz za pośrednictwem platformy ICSR Lab w badaniu klasyfikacji podejścia rozmytego wykorzystano bibliotekę PySpark'a MulticlassClassificationEvaluator zwracającą metryki ewaluacyjne uśredniane metodą ważonego-averging'u. W pracy zostaną wykorzystane następujące metryki ewaluacji klasyfikacji wieloklasowej:

1. Accuracy (Acc) – określające średnią ważoną stosunku obserwacji zaklasyfikowanych poprawnie w każdej klasie do liczby wszystkich ewaluowanych elementów, w kontekście badania podejścia rozmytego oznaczające stosunek liczby autorów zaklasyfikowanych w ten sam sposób podejściem bazowym i rozmytym do liczby wszystkich ewaluowanych autorów:

$$Accuracy = \sum_{i=1}^{k} w_i \times \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i},$$

gdzie TP_i, TN_i, FP_i i FN_i odpowiada wartościom TP, TN, FP i FN *i*-tej klasy, *k* odpowiada liczbie występujących klas, w_i oznacza stosunek liczby poprawnie zaklasyfikowanych obserwacji w klasie do liczby wszystkich elementów.

2. Precision (Prec) – określana jako średnia ważona stosunku poprawnie zaklasyfikowanych pozytywnych obserwacji w każdej klasie do liczby obserwacji zaklasyfikowanych poprawnie w każdej klasie. Oznacza ona prawdopodobieństwo przewidywania właściwych dyscyplin do autorów, które podejście rozmyte zakłada jako właściwe w porównaniu do wszystkich przypadków, które przewiduje jako właściwe:

$$Precision = \sum_{i=1}^{k} w_i \times \frac{TP_i}{TP_i + FP_i},$$

3. Sensitivity (Sens) – oceniającą średnią ważoną stosunku poprawnie zaklasyfikowanych pozytywnych obserwacji w każdej klasie do liczby obiektów klasy pozytywnej. Wyznacza ona prawdopodobieństwo właściwie przypisanych autorów do dyscyplin spośród wszystkich autorów zaklasyfikowanych do tej dyscypliny:

$$Sensivity = \sum_{i=1}^{k} w_i \times \frac{TP_i}{TP_i + FN_i},$$

4. Specificity (Spec) – wyznacza liczbę poprawnie zaklasyfikowanych obserwacji klasy negatywnej w każdej klasie do liczby obserwacji klasy negatywnej. Może być interpretowana jako prawdopodobieństwo nieprzypisania dyscypliny w przypadku, gdy nie powinna ona być zaklasyfikowana do autora:

$$Specificity = \sum_{i=1}^{k} w_i \times \frac{TN_i}{FP_i + TN_i},$$

5. F1 – jest to metryka stanowiąca średnią harmoniczną wartości metryk Sensitivity i Precision. Stworzona została ze względu na brak dostatecznej możliwości opisu jakości modelu w oparciu o wyłącznie jedną z tych dwóch metryk. Metryka F1 stanowi wariant metryki F-Score z parametrem α=1 i oznacza, że wagi składowych metryk są takie same:

$$F1 = \sum_{i=1}^{k} w_i \times \frac{2TP_i}{2TP_i + FP_i + FN_i},$$

6. MCC – współczynnik korelacji Matthewsa, nazywany również współczynnikiem fi (φ). Określa liniową korelację pomiędzy dwoma zbiorami danych. Zwraca wartość z zakresu [-1, 1], gdzie wartość 1 oznacza idealną korelację, 0 nie lepszą predykcję niż podejście losowe, a -1 całkowity brak liniowości. Metryka ta jest wskazywana wraz z Accuracy i F1 jako najczęściej wybierane metody w ewaluacji klasyfikacji. Współczynnik MCC ze względu na generowanie wyniku w oparciu o wszystkie przypadki z macierzy konfuzji jest określany jako dokładniejsza metryka niż Accuracy i F1 w sytuacjach, gdy wszystkie klasy nie są równoliczne:

$$MCC = \sum_{i=1}^{k} w_i \times \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i + FP_i)(TP_i + FN_i)(TN_i + FP_i)(TN_i + FN_i)}},$$

7. Hamming Loss (Loss) – określany jako średnia ważona stosunku niepoprawnie zaklasyfikowanych obserwacji w każdej klasie do liczby wszystkich ewaluowanych elementów. Oznacza liczbę autorów, którzy zarówno dla klasyfikacji w podejściu bazowym i rozmytym uzyskali przypisanie innych dyscyplin:

$$Hamming\ Loss = \sum_{i=1}^{k} w_i \times \frac{FP_i + FN_i}{TP_i + FP_i + TN_i + FN_i}.$$

Dwie z zaprezentowanych powyżej siedmiu metryk (Accuracy i MCC) opisywane będą w części deskryptywnej na potrzeby oceny jakości klasyfikacji w podejściu rozmytym. Pozostałe pięć metryk zostanie wyznaczonych na potrzeby tabelarycznego przedstawienia wyników.

Metryki ewaluacji klasyfikacji pozwolą na określenie jakości klasyfikacji podejściem rozmytym, jednak nie wskażą odsetków jednoznacznie zaklasyfikowanych autorów z całej badanej próby. W tym celu w badaniu wykorzystane zostaną następujące metryki:

1. Percentage classified (Perc class) – metryka wyznaczająca odsetek jednoznacznie zaklasyfikowanych obserwacji dla pełnego zbioru wyjściowego autorów z bazy Scopus:

$$Percentage\ classified = \frac{C_F}{N_S},$$

gdzie C_F oznacza liczbę jednoznacznie zaklasyfikowanych obserwacji w podejściu rozmytym, N_S oznacza liczbę autorów w bazie Scopus zbioru wyjściowego, N_S = 36,010,088.

2. Percentage classified [over base] (Perc class [base]) – wyznaczająca odsetek jednoznacznie zaklasyfikowanych obserwacji w stosunku do liczby obserwacji zaklasyfikowanych podejściem bazowym:

$$Percentage\ classified\ [over\ base] = \frac{C_F}{C_B},$$

gdzie C_F oznacza liczbę jednoznacznie zaklasyfikowanych obserwacji w podejściu rozmytym, C_B oznacza liczbę autorów jednoznacznie zaklasyfikowanych podejściem bazowym, C_B = 24,938,113. Obejmuje ona zakres wartości od 0 (gdy C_F = 0) do około 1.444 (gdy dokonano jednoznacznej klasyfikacji dla wszystkich autorów ze zbioru wyjściowego, $C_F = N_S$). Wybór metryki ma na celu określenie, o ile procent metoda w podejściu rozmytym zaklasyfikowała więcej lub mniej obserwacji niż metoda bazowa.

Rozdział 3.

Elementy teorii zbiorów rozmytych

3.1 Zbiory rozmyte

3.1.1 Pojęcie zbioru rozmytego

Pojęcie zbioru rozmytego, będące rozwinięciem klasycznej definicji zbioru, zostało wprowadzone przez Lotfiego Zadeha w 1965 roku (Zadeh, 1965). Teoria zbiorów rozmytych dopuszcza sytuację, w której element x może należeć do zbioru rozmytego tylko w określonym stopniu (tzn. przyjmowanie wartości z przedziału [0,1] zamiast {0, 1}). Zgodnie z definicją, zbiór rozmyty A w wymiarze U nazywany jest funkcją:

$$A : U \rightarrow [0, 1] \tag{3.1}$$

Taka funkcja nazywana jest funkcją przynależności zbioru rozmytego A. Inna reprezentacja zbioru rozmytego obejmuje uporządkowaną parę:

$$\{\big(x, A(x)\big) : x \in U\} \tag{3.2}$$

gdzie $A(x)$ nazywane jest stopniem przynależności elementu x do zbioru rozmytego .

3.1.2 Operacje triangularne

Pojęcie normy triangularnej zostało zaproponowane na początku lat 40-tych przez Karla Mengera na potrzeby badań nad probabilistycznymi przestrzeniami metrycznymi (Menger, 1942). Stanowiła ona narzędzie do generalizacji nierówności trójkąta (stąd przyjmując tę nazwę). Prace nad tworzeniem zbioru aksjomatów kontynuowali w latach 60-tych B. Schweizer i A. Sklar (Schweizer i inni, 1960). Kolejne badania wykazały użyteczność tych norm w logikach wielowartościowych (opisywane w pracach Gottwald, 1999; Klement i inni, 2000).

Definicja 3.1 Operacja binarna $\boldsymbol{t} : [0, 1] \times [0, 1] \rightarrow [0, 1]$ nazywana jest normą triangularną (t-normą), jeżeli spełnia następujące własności dla każdego $a, b, c \in [0, 1]$:

1. $a \boldsymbol{t} b = b \boldsymbol{t} a$ (przemienność),

2. $(a \; \boldsymbol{t} \; b) \; \boldsymbol{t} \; c = a \; \boldsymbol{t} \; (b \; \boldsymbol{t} \; c)$ (łączność),
3. $a \leq b \Rightarrow a \; \boldsymbol{t} \; c \leq b \; \boldsymbol{t} \; c$ (monotoniczność),
4. $a \; \boldsymbol{t} \; 1 = 1$ (1 - element neutralny).

Definicja 3.2 Operacja binarna $\boldsymbol{s} : [0, 1] \times [0, 1] \rightarrow [0, 1]$ nazywana jest konormą triangularną (t-konormą), jeżeli spełnia następujące własności dla każdego $a, b, c \in [0, 1]$:

1. $a \; \boldsymbol{s} \; b = b \; \boldsymbol{s} \; a$ (przemienność),
2. $(a \; \boldsymbol{s} \; b) \; \boldsymbol{s} \; c = a \; \boldsymbol{s} \; (b \; \boldsymbol{s} \; c)$ (łączność),
3. $a \leq b \Rightarrow a \; \boldsymbol{s} \; c \leq b \; \boldsymbol{s} \; c$ (monotoniczność),
4. $a \; \boldsymbol{s} \; 0 = a$ (0 - element neutralny).

Normy i konormy triangularne łącznie nazywane sią operacjami triangularnymi (t-operacje). Do najpopularniejszych t-operacji należą:

1. t-norma minimum

$$a \wedge b = \min(a, b),$$

2. t-konorma maksimum

$$a \vee b = \max(a, b),$$

3. t-norma drastyczna $\boldsymbol{t}_d$

$$a \; \boldsymbol{t}_d \; b = \begin{cases} a \wedge b, & \text{jeżeli } a \vee b = 0, \\ 0, & w \; p. p \end{cases}$$

4. t-konorma drastyczna $\boldsymbol{s}_d$

$$a \; \boldsymbol{s}_d \; b = \begin{cases} a \vee b, & \text{jeżeli } a \wedge b = 0, \\ 0, & w \; p. p \end{cases}$$

5. t-norma algebraiczna $\boldsymbol{t}_a$

$$a \; \boldsymbol{t}_a \; b = ab,$$

6. t-konorma algebraiczna $\boldsymbol{s}_a$

$$a \; \boldsymbol{s}_a \; b = a + b - ab,$$

7. t-norma Łukasiewicza $\boldsymbol{t}_Ł$

$$a \; \boldsymbol{t}_Ł \; b = 0 \vee (a + b - 1),$$

8. t-konorma Łukasiewicza $\boldsymbol{s}_Ł$

$$a \; \boldsymbol{s}_Ł \; b = 1 \wedge (a + b).$$

Zarówno wszystkie t-normy $\boldsymbol{t}$ jak i t-konormy $\boldsymbol{s}$ dla $a, b, c \in [0, 1]$ posiadają następujące własności:

1. $a\ \boldsymbol{t}\ 0 = 0,\ a\ \boldsymbol{s}\ 1 = 1,$ (3.3)
2. $a\ \boldsymbol{t_d}\ b \leq a\ \boldsymbol{t}\ b \leq a \wedge b \leq a \vee b \leq a\ \boldsymbol{s}\ b \leq a\ \boldsymbol{s_d}\ b,$
3. $a\ \boldsymbol{t}\ a \leq a \leq a\ \boldsymbol{s}\ a,$
4. $a\ \boldsymbol{t}\ b = 1 \Leftrightarrow a = b = 1,\quad a\ \boldsymbol{s}\ b = 0 \Leftrightarrow a = b = 0,$
5. $a\ \boldsymbol{t}\ a = a \Leftrightarrow \boldsymbol{t} = \wedge,\quad a\ \boldsymbol{s}\ a = a \Leftrightarrow \boldsymbol{s} = \vee,$
6. $(a\ \boldsymbol{s}\ (b\ \boldsymbol{t}\ c) = (a\ \boldsymbol{s}\ b)\ \boldsymbol{t}\ (a\ \boldsymbol{s}\ c)) \Leftrightarrow \boldsymbol{t} = \wedge,$
7. $(a\ \boldsymbol{t}\ (b\ \boldsymbol{s}\ c) = (a\ \boldsymbol{t}\ b)\ \boldsymbol{s}\ (a\ \boldsymbol{t}\ c)) \Leftrightarrow \boldsymbol{s} = \vee,$

ponadto pomiędzy t-normą $\boldsymbol{t}$, a t-konormą $\boldsymbol{s}$ występuje zgodność bijektywna, w której dla t-normy $\boldsymbol{t}^*$ takiej, że:

$$a\ \boldsymbol{t}^*\ b = 1 - (1 - a)\ \boldsymbol{t}\ (1 - b) \tag{3.4}$$

$\boldsymbol{t^*}$ jest t-konormą $\boldsymbol{s}$. Dla t-konormy $\boldsymbol{s^*}$ takiej, że:

$$a\ \boldsymbol{s}^*\ b = 1 - (1 - a)\ \boldsymbol{s}\ (1 - b) \tag{3.5}$$

$\boldsymbol{s}^*$ jest t-normą $\boldsymbol{t}$.

3.1.3 Rodziny operacji triangularnych

Poza tym istnieje również szereg innych t-operacji, które wykorzystują parametr λ. Operacje te stanowią rodzinę operacji triangularnych.

1. rodzina t-operacji Schweizera z $\lambda > 0$

$$a\ \boldsymbol{t_{S,\lambda}}\ b = \left(0 \vee \left(a^\lambda + b^\lambda - 1\right)\right)^{\frac{1}{\lambda}},$$

$$a\ \boldsymbol{s_{S,\lambda}}\ b = 1 - \left(0 \vee \left((1-a)^\lambda + (1-b)^\lambda - 1\right)\right)^{\frac{1}{\lambda}},$$

2. rodzina t-operacji Yagera z $\lambda \geq 1$

$$a\ \boldsymbol{t_{Y,\lambda}}\ b = 1 - (1 \wedge ((1-a)^\lambda + (1-b)^\lambda)^{\frac{1}{\lambda}}),$$

$$a\ \boldsymbol{s_{Y,\lambda}}\ b = 1 \wedge \left(a^\lambda + b^\lambda\right)^{\frac{1}{\lambda}},$$

3. rodzina t-operacji Hamachera z $\lambda \geq 0$

$$a\ \boldsymbol{t_{H,\lambda}}\ b = \frac{ab}{\lambda + (1-\lambda)(a+b-ab)},$$

$$a\ \boldsymbol{s_{H,\lambda}}\ b = \frac{a+b-ab-(1-\lambda)ab}{1-(1-\lambda)ab},$$

4. rodzina t-operacji Franka z $\lambda > 0$, $\lambda \neq 1$

$$a\ \boldsymbol{t_{F,\lambda}}\ b = \log_\lambda\left(1 + \frac{(\lambda^a - 1)(\lambda^b - 1)}{\lambda - 1}\right),$$

$$a\ \boldsymbol{s_{F,\lambda}}\ b = 1 - \log_\lambda\left(1 + \frac{(\lambda^{1-a} - 1)(\lambda^{1-b} - 1)}{\lambda - 1}\right),$$

5. rodzina t-operacji Webera z $\lambda > -1$

$$a\ \boldsymbol{t_{W,\lambda}}\ b = 0 \wedge \left(\frac{a+b-1+\lambda ab}{1+\lambda}\right),$$

$$a\ \boldsymbol{s_{W,\lambda}}\ b = 1 \vee \left(\frac{(1+\lambda)(a+b)-\lambda ab}{1+\lambda}\right),$$

Dla rodziny t-operacji Hamachera ze wskaźnikiem $\lambda = 2$ przyjmuje się oznaczenie rodziny t-operacji Einsteina. Dodatkowo dla wybranych wartości λ niektóre rodziny t-operacji są tożsame z t-operacjami zaprezentowanymi w sekcji 3.1.2:

- dla t-operacji Hamachera z $\lambda = 1$ otrzymujemy t-operacje algebraiczne,
- t-operacje Schweizera z $\lambda = 1$ są tożsame z t-operacjami Yagera z $\lambda = 1$, Webera z $\lambda = 0$ oraz t-operacjami Łukasiewicza,
- t-operacje Franka dla λ dążącego do 0 są równoważne t-operacjom minimum/maksimum, dla λ dążącego do 1 t-operacjom algebraicznym oraz dla λ dążącego do ∞ odpowiadają t-operacjom Łukasiewicza.

3.1.4 Negacje

Definicja 3.3 Negacją nazywamy funkcję v: [0, 1] → [0, 1], kiedy $v(0) = 1$, $v(1) = 0$ oraz v jest funkcją nierosnącą. Negację nazywamy silną negacją, jeżeli dodatkowo jest funkcją ciągłą malejącą. Ścisłą negację nazywamy silną negacją, jeżeli jest dodatkowo inwolucyjna (tzn. $v(v(a)) = a$ dla $a \in [0, 1]$. Do najpopularniejszych negacji należą:

1. negacja progująca v_t dla $t \in [0, 1)$

$$v_t(a) = \begin{cases} 1, & jeżeli\ a \leq t, \\ 0, & w\ p.p., \end{cases},$$

2. negacja progująca v^t dla $t \in (0, 1]$

$$v^t(a) = \begin{cases} 1, & jeżeli\ a < t, \\ 0, & w\ p.p., \end{cases},$$

3. negacja Łukasiewicza $v_Ł$

$$v_Ł(a) = 1 - a,$$

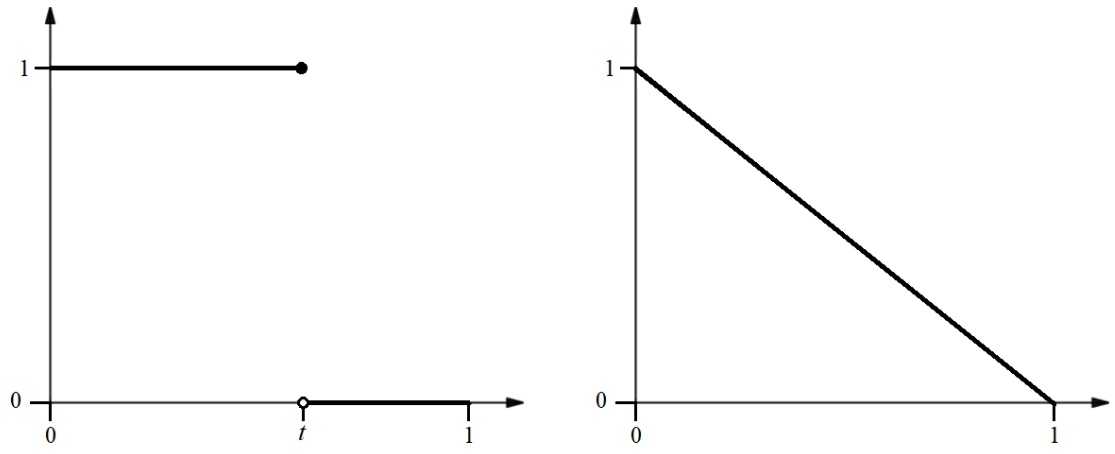
4. negacja Sugeno $v_{S,\lambda}$

$$v_{S,\lambda}(a) = \frac{1-a}{1+\lambda a}, \qquad \lambda > -1,$$

5. negacja Yagera

$$v_{Y,t}(a) = (1 - a^t)^{\frac{1}{t}}, \qquad t > 0.$$

Negację progującą v_t z t=0 oraz negację progującą v^t dla t=1 nazywa się negacjami ekstremalnymi (negacjami Godla). Negację Łukasiewicza nazywa się negacją standardową (Rys. 3.1).



Rysunek. 3.1 Przykład negacji progującej v_t w punkcie t (lewy wykres) oraz negacji Łukasiewicza (prawy wykres)

3.1.5 Rodzaje funkcji przynależności

Modelując wybrany zbiór rozmyty należy uwzględnić kształt jej funkcji przynależności. L. Zadeh definiując zbiory rozmyte przedstawił ich kilka podstawowych wariantów. Stanowiły one zarówno grupę funkcji liniowych jak i krzywych (Zadeh, 1965). Następni naukowcy w zależności od własnych potrzeb i subiektywizmu w konstrukcji zbiorów rozmytych opisali i wykorzystywali kolejne typy funkcji (zebrane między innymi w pracach Hong i Lee, 1996; Sivanandam 2007; Sadeghian i inni, 2013; Sambariya i Prasad, 2017). Obecnie najczęściej do modelowania funkcji przynależności *f* wykorzystuje się następujące typy funkcji:

1. funkcja trójkątna

$$triangle(x,a,b,c) = \begin{cases} \dfrac{x-a}{b-a}, & jeżeli\ a \le x \le b, \\ \dfrac{c-x}{c-b}, & jeżeli\ b \le x \le c, \\ 0, & w\ p.p., \end{cases} \tag{3.6}$$

2. funkcja trapezoidalna

$$trapezoid(x,a,b,c,d) = \begin{cases} 0, & jeżeli\ x < a, \\ \dfrac{x-a}{b-a}, & jeżeli\ a \le x \le b, \\ 1, & jeżeli\ b \le x \le c, \\ \dfrac{d-x}{d-c}, & jeżeli\ c \le x \le d, \\ 0, & jeżeli\ x > d, \end{cases} \tag{3.7}$$

3. funkcja liniowa *s*-kształtna

$$sShapeLineral(x,a,b) = \begin{cases} 0, & jeżeli\ x < a, \\ \dfrac{x-a}{b-a}, & jeżeli\ a \le x \le b, \\ 1, & jeżeli\ x > b, \end{cases} \tag{3.8}$$

4. funkcja liniowa *z*-kształtna

$$zShapeLineral(x,a,b) = \begin{cases} 1, & jeżeli\ x < a, \\ \dfrac{b-x}{b-a}, & jeżeli\ a \le x \le b, \\ 0, & jeżeli\ x > b, \end{cases} \tag{3.9}$$

5. funkcja Gaussowska

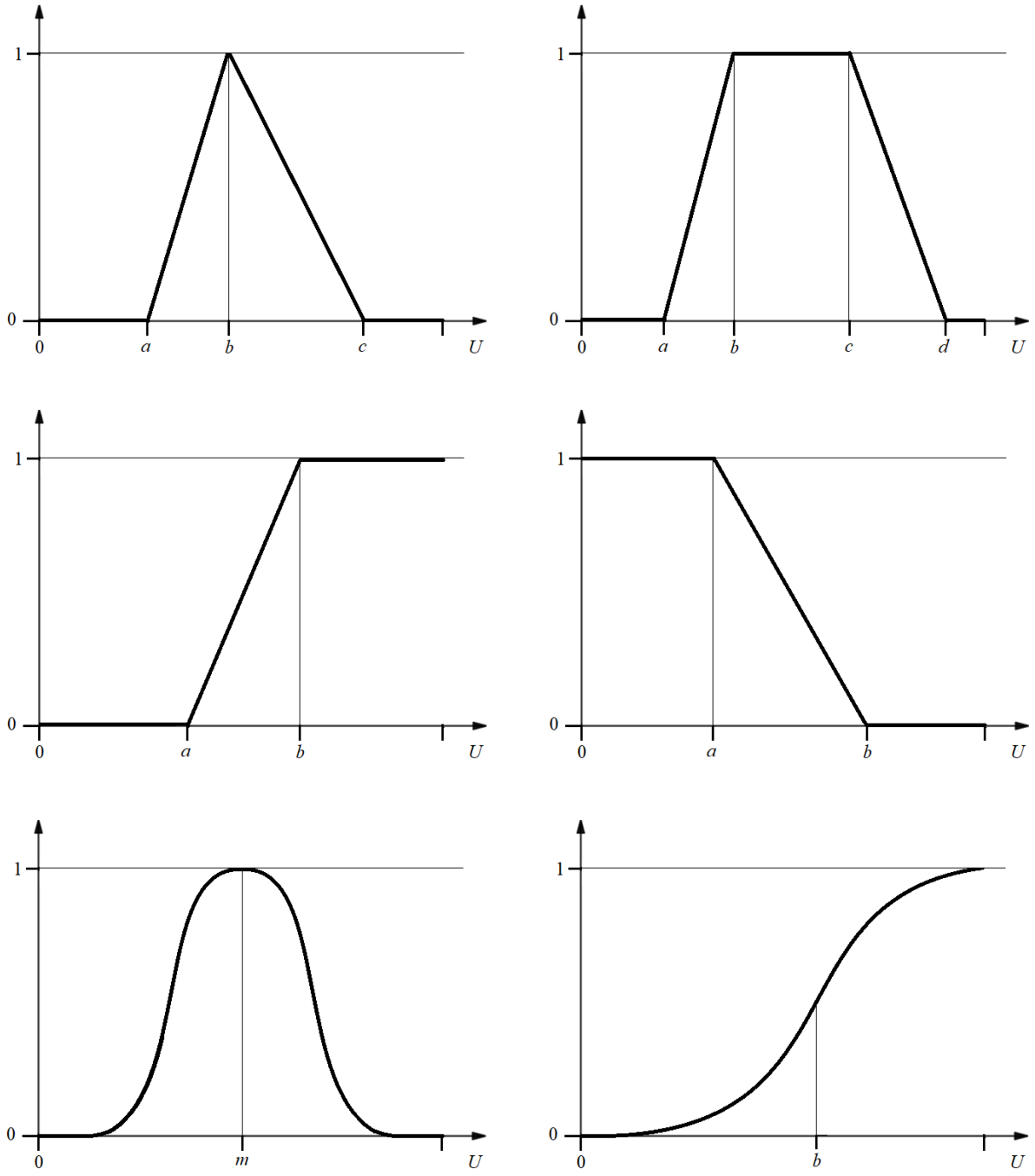
$$gaussian(x,m,\alpha) = e^{\frac{-(x-m)^2}{2\alpha^2}}, \tag{3.10}$$

gdzie *m* wyznacza centrum, a α odchylenie standardowe,

6. funkcja sigmoidalna

$$sig(x, m, b) = \frac{1}{1 + e^{-m(x-b)}}\ , \tag{3.11}$$

gdzie *m* oznacza nachylenie w skrzyżowaniu *x* = *b*. Graficzne reprezentacje przedstawia Rys. 3.2.



Rysunek 3.2 Przykłady typów funkcji przynależności, od lewej od góry: funkcja trójkątna, funkcja trapezoidalna, funkcja liniowa s-kształtna, funkcja liniowa z-kształtna, funkcja Gaussowska i funkcja sigmoidalna

3.1.6 Operacje na zbiorach rozmytych

Wraz z powstaniem teorii zbiorów rozmytych L. Zadeh zaproponował standardowe operacje dla zbiorów rozmytych: sumę, przekrój i dopełnienie (Zadeh, 1965; kolejne operacje opisane w pracach o tematyce zbiorów rozmytych np. Wygralak, 2013; Dyczkowski, 2018).

Niech A i B będą zbiorami rozmytymi $U \rightarrow [0,1]$. Zbiór rozmyty $A \cap_t B$ stanowi przekrój zbiorów rozmytych A, B indukowany przez t-normę $\boldsymbol{t}$ jeżeli:

$$(A \cap_t B)(x) = A(x)\ \boldsymbol{t}\ B(x) \quad dla\ każdego\ x \in U. \tag{3.12}$$

Sumą zbiorów rozmytych A i B indukowaną przez t-konormę $\boldsymbol{s}$ jest zbiór $A \cup_s B$ taki, że:

$$(A \cup_s B)(x) = A(x)\ \boldsymbol{s}\ B(x) \quad dla\ każdego\ x \in U. \tag{3.13}$$

Produktem kartezjańskim zbiorów A i B takich, że A: $U \rightarrow [0,1]$ i B: $V \rightarrow [0,1]$ indukowanym przez t-normę $\boldsymbol{t}$ jest zbiór rozmyty $A \times_t B$: $U \times V \rightarrow [0, 1]$ uporządkowanej pary $U \times V$, kiedy:

$$(A \times_t B)(x, y) = A(x)\ \boldsymbol{t}\ B(y) \quad dla\ każdej\ pary\ (x, y) \in U \times V. \tag{3.14}$$

Dopełnienie zbioru rozmytego A indukowane przez negację v stanowi A^v jeżeli:

$$A^v(x) = v(A(x)) \quad dla\ każdego\ x \in U. \tag{3.15}$$

Różnicę pomiędzy zbiorem A i B, zapisywaną jako A \ B lub $A \cap B'$ uzyskuje się jeżeli

$$(A \backslash B)(x) = A(x) \wedge \big(1 - B(x)\big) \quad dla\ każdego\ x \in U. \tag{3.16}$$

Zbiór rozmyty A zawiera się w zbiorze rozmytym B (A $\subset$ B) jeżeli:

$$A(x) \leq B(x) \quad dla\ każdego\ x \in U. \tag{3.17}$$

Dla powyższych operacji stosując t-normę minimum oraz t-konormę maksimum w uproszczeniu wprowadza się następujące terminologie:

$$A \cap B = A \cap_{\wedge} B \quad (przekrój) \tag{3.18}$$
$$A \cup B = A \cup_{\vee} B \quad (suma)$$
$$A \times B = A \times_{\wedge} B \quad (produkt\ kartezjański)$$
$$A' = A^{v_{\text{Ł}}} \quad (dopełnienie\ zbioru\ A)$$

Na podstawie Definicji 3.1 i 3.2 i własności (3.4) stosując operacje przekroju, sumy indukowanej przez dowolną normę triangularną t, s spełnione są następujące własności:

1. $A \cap_t B = B \cap_t A$, $A \cup_t B = B \cup_t A$, (przemienność),
2. $A \cap_t (B \cap_t C) = (A \cap_t B) \cap_t C$, (łączność),
   $A \cup_t (B \cup_t C) = (A \cup_t B) \cup_t C$,
3. $A \cap_t A = A$, $A \cap_t A = A$, (idempotentność),
4. $A \cup_t (A \cap_t B) = A$, (absorbcja),
   $A \cap_t (A \cup_t B) = A$
5. $A \cap_t (B \cup_t C) = (A \cap_t B) \cup_t (A \cap_t C)$, (rozdzielność),
   $A \cup_t (B \cap_t C) = (A \cup_t B) \cap_t (A \cup_t C)$,
6. $A \cap_t 1_U = A$, $A \cup_t 1_\emptyset = A$ (elementy neutralne)

3.1.7 Charakterystyki zbiorów rozmytych

Na potrzeby opisywania i analizy zbiorów rozmytych zaproponowano kilka pomocniczych pojęć stanowiących charakterystykę zbiorów rozmytych.

T-warstwą zbioru rozmytego *A* nazywamy zbiór rozmyty A_t taki, że:

$$A_t(x) = \{x \in U: A(x) \geq t\}, \quad t \in (0,1]. \tag{3.19}$$

T-warstwą ostrą zbioru rozmytego *A* nazywamy zbiór rozmyty A^t taki, że:

$$A^t(x) = \{x \in U: A(x) > t\}, \quad t \in (0,1] \tag{3.20}$$

Nośnik zbioru rozmytego *A* stanowi *supp*(*A*) taki, że:

$$supp(A) = \{x \in U: A(x) > 0\} \tag{3.21}$$

Jądro zbioru rozmytego *A* stanowi *core*(*A*) taki, że:

$$core(A) = \{x \in U: A(x) = 1\} \tag{3.22}$$

Jeżeli nośnik zbioru rozmytego *A* jest zbiorem skończonym, nazywany jest skończonym zbiorem rozmytym (ang. Finite Fuzzy Set, *FFS*).

W przypadku gdy nośniki zbioru rozmytego A są jednoelementowe, o zbiorze mówi się, że jest singletonem (zapisywanym w postaci a/x, gdzie a > 0). Korzystając z postaci zapisu dla singletonu dowolny zbiór rozmyty można przedstawić w postaci sumy singletonów:

$$A = \bigcup_{x \in supp(A)} \frac{A(x)}{x} \tag{3.23}$$

Szczególnie w przypadku skończonych zbiorów rozmytych, dla których *supp*(*A*) = {x_1,x_2,…,x_n}, $n \geq 1$ można wykorzystać zapis:

$$A = \frac{A(x_1)}{x_1} \cup \frac{A(x_2)}{x_2} \cup \ldots \cup \frac{A(x_n)}{x_n} \tag{3.24}$$

Po zamianie operatora ∪ na operator arytmetyczny + uzyskuje się zapis:

$$A = \frac{A(x_1)}{x_1} + \frac{A(x_2)}{x_2} + \cdots + \frac{A(x_n)}{x_n} \tag{3.25}$$

nazywany notacją singletonową zbioru rozmytego.

3.1.8 Operator implikacji

Definicja 3.4 Operator implikacji (oznaczany jako →) stanowi relację logiczną [0, 1] x [0, 1] → [0, 1], jeżeli spełnia następujące warunki dla *a*, *b*, *c*, *d* ∈ [0, 1]:

1. $a \rightarrow b \geq c \rightarrow b$, gdy $a \leq c$, (nierosnącość w pierwszym argumencie)
2. $a \rightarrow b \leq a \rightarrow d$, gdy $b \leq d$, (niemalejącość w drugim argumencie)
3. $0 \rightarrow b = 1$; $a \rightarrow 1 = 1$; $1 \rightarrow 0 = 0$ (warunki brzegowe).

Stosuje je się, miedzy innymi w systemach wykorzystujących reprezentacje wiedzy (sterowniki rozmyte, systemy podejmowania decyzji wielokryterialnej itd.). Szczegółową charakterystykę i klasyfikację zestawiono w pracach Dubois i Prade, 1993; Trillas i Valverde, 1993; Rojas i inni, 1998; Cordon i inni, 2000; Gottwald, 2001; Ross, 2010, Wygralak 2013. Do najpopularniejszych operatorów implikacji należą:

1. operator implikacji Łukasiewicza

$$a \rightarrow b = 1 \wedge (1 - a + b)$$

2. operator implikacji Godla

$$a \rightarrow b = \begin{cases} 1, & jeżeli\ a \leq b \\ b, & w\ p.p. \end{cases}$$

3. operator implikacji Goguena

$$a \rightarrow b = \begin{cases} 1 \wedge \frac{b}{a}, & jeżeli\ a \neq 0 \\ 1, & w\ p.p. \end{cases}$$

4. operator implikacji Kleenea-Dienesa

$$a \rightarrow b = (1 - a) \vee b$$

5. operator implikacji Zadeha

$$a \rightarrow b = (1 - a) \vee (a \wedge b)$$

6. operator implikacji Reichenbacha

$$a \rightarrow b = 1 - a + ab$$

3.2 Liczby rozmyte i zmienne lingwistyczne

Jeden z fundamentalnych elementów w teorii zbiorów rozmytych oraz w systemach wnioskowania rozmytego stanowi zmienna lingwistyczna. W tej sekcji opisano pojęcie zmiennej lingwistycznej i liczb rozmytych, które wykorzystywane są przy tworzeniu interpretacji wartości zmiennych lingwistycznych.

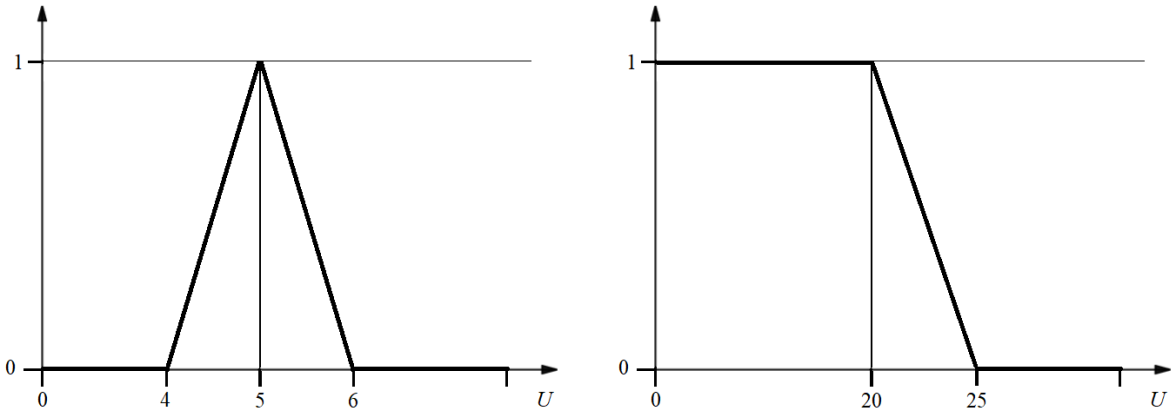
3.2.1 Liczby rozmyte

Modelowanie pojęć poprzez liczby rozmyte wykorzystuje się w przypadku opisywania zjawisk przy użyciu nieprecyzyjnych danych numerycznych. Liczbą rozmytą nazywamy zbiór rozmyty w podzbiorze liczb rzeczywistych $\mathbb{R}$, czyli funkcję działającą ze zbioru $\mathbb{R} \rightarrow [0, 1]$.

Do modelowania funkcji przynależności liczby rozmytej wykorzystuje się kształty znane z najpopularniejszych typów funkcji przynależności (sekcja 3.1.5). Wybór kształtu funkcji zależy od pojęcia nieprecyzyjnego, które wymaga reprezentacji w postaci zbioru rozmytego. Dla liczby rzeczywistej *x* w przypadku formułowania pojęcia „około *x*" wykorzystuje się funkcje trójkątne lub Gaussowskie osiągające centrum w punkcie *x* (na przykład około 5 stopni, średnio 2 osoby).

Do modelowania pojęcia „od około *x* do około *y*" lub „w przybliżeniu pomiędzy *x* i *y*" i innych podobnych wskazujących przedziały liczbowe stosuje się funkcje trapezoidalne, w których przedział [*x, y*] stanowi jądro zbioru rozmytego (np. od około 10 do 15 kilometrów, pomiędzy około 35 a 50 lat, średnio pomiędzy 175 a 180 cm). Zarówno funkcja trójkątna, jak i trapezoidalna wykorzystywana jest do reprezentacji wartości „średnio". Wybór funkcji zależy od przedziału, który użytkownik akceptuje jako jądro zbioru rozmytego.

Przy reprezentacji nieprecyzyjnej wartości numerycznej „co najmniej około *x*” ma zastosowanie funkcja s-kształtna liniowa lub sigmoidalna z jądrem zbioru rozmytego dla wartości $\geq x$ (np. co najmniej około 20 lat). Dla pojęcia „co najwyżej około *y*” wykorzystuje się funkcję z-kształtną liniową z jądrem zbioru rozmytego dla wartości $\leq y$ (np. co najwyżej około 30 osób). Jeżeli użytkownik posiada ściśle zdefiniowane wartości, które uważa za spełniające dane pojęcie nieprecyzyjne (stanowiące jądro zbioru rozmytego) można stosować funkcje s-kształtne/sigmoidalne dla wartości „wysoko”, „dużo” oraz z-kształtne dla wartości „nisko”, „mało” (Rys. 3.3).



Rysunek 3.3. Interpretacja liczby rozmytej około 5 (lewa), co najwyżej około 20 (prawa)

Poza definiowaniem typów liczb rozmytych ze względu na kształt ich funkcji przynależności można przypisać im również cechy arytmetyczne. Jeżeli nośnik liczby rozmytej zawiera się w przedziale $(0, +\infty)$ nazywana jest ona liczbą rozmytą pozytywną. Dla nośnika występującego w przedziale $(-\infty, 0)$ stanowi ona liczbę rozmytą negatywną. Gdy wartość 0 zawiera się w nośniku, liczba rozmyta jest zerową liczbą rozmytą.

3.2.2 Zmienne lingwistyczne

Zmienna lingwistyczna to zmienna, która przyjmuje wartości lingwistyczne, terminologie języka naturalnego. Pojęcie to wprowadził L. Zadeh w 1975 roku (Zadeh 1975a, b, c). Ze względu na jej zdolności do wyrażania nieprecyzyjnych pojęć, wykorzystuje się je do konstrukcji zbiorów rozmytych. Zmienną lingwistyczną można wyrazić w postaci czwórki:

$$(\mu, U, T, I),$$

gdzie μ jest nazwą zmiennej lingwistycznej, T stanowi listę wartości (termów), I obejmuje zbiór interpretacji termów. Dla każdej interpretacji termu definiowany jest odpowiedni zbiór rozmyty będący funkcją działającą z $U \rightarrow [0, 1]$. Pojęciami nieprecyzyjnymi może być na przykład: wiek, wzrost, prędkość, odległość, powierzchnia, populacja, opady deszczu, a termami odpowiednio: młody, szybko, daleko, duży, wysoki, niski. Przy tworzeniu interpretacji często bazuje się na typach funkcji przynależności (sekcja 3.1.4) lub modeluje się je w postaci liczb rozmytych (sekcja 3.2.1) (Zadeh, 1965).

Przy interpretowaniu termów często stosuje się zasadę tworzenia kolejnych bazując na stworzonych funkcjach przynależności innych wartości dla tej samej zmiennej. Niech μ będzie zmienną posiadającą interpretację wartości α w postaci zbioru rozmytego A, wartości β w postaci zbioru rozmytego B, wartości γ dla zbioru rozmytego C, t i s stanowią normy triangularne, a v negację. Posiadając zbiór rozmyty A dla wartości α możemy wygenerować interpretację:

1. dla β jako A^v (B stanowi negację A).
2. dla γ jako $A \cap_t B$ (C stanowi przekrój A i B indukowany przez t i oznacza α i β),
3. dla γ jako $A \cup_t B$ (C stanowi sumę A i B indukowany przez s i oznacza α lub β).

Najpopularniejszy schemat tworzenia zmiennej lingwistycznej obejmuje uwzględnienie co najmniej trzech termów (na przykład *niskie*, *wysokie*, *średnie*). Dla zamodelowanego termu *wysokie* modeluje się term *niskie* poprzez dopełnienie zbioru rozmytego *wysokie* (najczęściej dopełnienie będące negacją Łukasiewicza). Term *średnie* stanowi przekrój zbioru *niskie* i *wysokie* indukowany przez t-normę minimum (przyjmuje się również nazwę *ani niskie ani wysokie*).

3.3 Operatory agregacji

W tej części opisano pojęcie operatorów agregacji oraz wskazano najpopularniejsze przykłady operatorów agregacji.

Definicja 3.5 Operator agregacji jest funkcją:

$$Aggr: \bigcup_{n \geq 1} [0,1]^n \rightarrow [0,1],$$

kiedy spełnia następujące warunki:

1. $Aggr(a_1, \ldots, a_n) \leq Aggr(b_1, \ldots, b_n)$, gdy $a_i \leq b_i$ dla $i \in \{1, \ldots, n\}$, (monotoniczność)
2. $Aggr(a) = a$, dla każdego $a \in [0, 1]$, (tożsamościowość)
3. $Aggr(0, \ldots, 0) = 0$, $Aggr(1, \ldots, 1) = 1$ (warunki brzegowe)

Operatory agregacji można przyporządkować do następujących klas:

1. Operatory uśredniające, jeżeli:

$$(a_1 \wedge \ldots \wedge a_n) \leq Aggr(a_1, \ldots, a_n) \leq (a_1 \vee \ldots \vee a_n), \tag{3.26}$$

2. Operatory koniunktywne, jeżeli

$$Aggr(a_1, \ldots, a_n) \leq (a_1 \wedge \ldots \wedge a_n), \tag{3.27}$$

3. Operatory dysjunktywne, jeżeli

$$Aggr(a_1, \ldots, a_n) \geq (a_1 \vee \ldots \vee a_n), \tag{3.28}$$

4. pozostałe /mieszane, jeżeli nie spełniają powyższych warunków.

Operatory agregacji mają na celu uzyskanie jednej wartości zdolnej do reprezentacji większej ilości danych. Dokładną charakterystykę operatorów agregujących opisano w pracy Beliakov i inni, 2007. Poniżej zestawiono najpopularniejsze operatory agregacji.

3.3.1 Operatory minimum/maksimum

Dla formuł (3.26) - (3.28) można wyszczególnić dwie operacje triangularne, dzięki którym operator stanowi zarówno klasę operatorów uśredniających, jak i koniunktywnych, i dysjunktywnych.

Niech dany będzie wektor *x* i t-norma minimum, wówczas operator agregacji minimum stanowi funkcja:

$$Aggr_{min}(x) = (x_1 \wedge \ldots \wedge x_n), \tag{3.29}$$

Dla t-konormy maksimum uzyskujemy operator agregacji maksimum jako funkcję:

$$Aggr_{max}(x) = (x_1 \vee \ldots \vee x_n). \tag{3.30}$$

Kolejne operatory (3.3.2 oraz 3.3.3) stanowią przykład operatorów uśredniających.

3.3.2 Średnie

1. Średnia arytmetyczna

$$Aggr_A(x) = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \tag{3.31}$$

2. Średnia geometryczna

$$Aggr_G(x) = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}, \tag{3.32}$$

3. Średnia harmoniczna

$$Aggr_H(x) = n\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}, \tag{3.33}$$

4. R-średnia

$$Aggr_R(x) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^r\right)^{\frac{1}{r}}, \tag{3.34}$$

dla dowolnego r ∈ ℝ. Średnia arytmetyczna jest równoważna wariantowi R-średniej z $r = 1$.

3.3.3 Średnie ważone

Niech w będzie wektorem $w = (w_1, \dots, w_n)$ takim, że dla każdego $i \in \{1, \dots, n\}$ $w_i > 0$ i $\sum_{i=1}^{n} w_i = 1$

1. Ważona średnia arytmetyczna

$$Aggr_{WA}(x, w) = \sum_{i=1}^{n} w_i x_i, \tag{3.35}$$

2. Ważona średnia geometryczna

$$Aggr_{WG}(x, w) = \prod_{i=1}^{n} {x_i}^{w_i}, \tag{3.36}$$

3. Ważona średnia harmoniczna

$$Aggr_{WH}(x, w) = \left(\sum_{i=1}^{n} \frac{w_i}{x_i}\right)^{-1}, \tag{3.37}$$

4. Ważona R-średnia

$$Aggr_{WR}(x, w) = \left(\frac{1}{n}\sum_{i=1}^{n} w_i x_i^r\right)^{\frac{1}{r}}, \tag{3.38}$$

dla dowolnego r ∈ ℝ. Ważona średnia arytmetyczna jest równoważna ważonej R-średniej z $r = 1$.

3.3.4 Operatory uporządkowanej średniej ważonej (OWA)

Poprzednie operatory (3.3.3) przypisywały odpowiednie wagi do odpowiadających im atrybutom. Przykładem operatorów, które przypisują wagi do uporządkowanych wartości sa operatory uporządkowanej średniej ważonej (ang. Ordered weighted averging. OWA). Zostały one zaproponowane przez Yagera oraz prezentowane w badaniach kolaboracyjnych z Kacprzykiem (Yager, 1988; Yager i Kacprzyk, 1997). Operatory OWA dzięki swojej elastyczności w dostosowywaniu wag, a tym samym różnych poziomów pewności lub istotności oraz możliwości efektywnej agregacji danych dla różnych kryteriów, znajdują zastosowanie w systemach wielokryterialnego podejmowania decyzji.

Niech d_1, …, d_n będzie sekwencją uzyskaną przez nierosnące uporządkowanie wektora $x = (x_1, \ldots, x_n)$ tzn. $d_1 \geq \ldots \geq d_n$. i niech w stanowi wektor $(w_1, \ldots, w_n)$ taki, że dla każdego $i \in \{1, \ldots, n\}$ $w_i \geq 0$ i $\sum_{i=1}^{n} w_i = 1$. Operator OWA jest odwzorowaniem:

$$Aggr_{OWA}(x, w) = \sum_{i=1}^{n} w_i d_i \tag{3.39}$$

W przypadku gdy wszystkie wagi są takie same operator OWA jest równoważny średniej arytmetycznej. Jeżeli wagi operatora OWA odpowiadają wektorowi (1, 0, …, 0) jest on równoważny operatorowi maximum, ponieważ wybierany jest tylko największy element sekwencji, a dla wektora wag (0, …, 0, 1) odpowiada on operatorowi minimum, ponieważ wybierany jest tylko najmniejszy element sekwencji.

3.3.5 Miękkie operatory triangularne

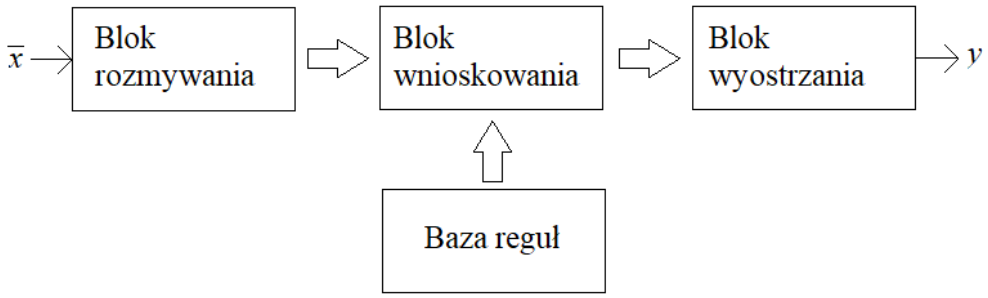
Niech dana będzie dowolna t-norma *t*, t-konorma *s* oraz współczynnik $\lambda \in [0, 1]$. Miękką t-normę oraz t-konormę stanowią odpowiednio:

$$Aggr_t(x, \lambda) = (1 - \lambda) \sum_{i=1}^{n} x_i + \lambda(x_1 \; t \; \ldots t \; x_n) \tag{3.40}$$

$$Aggr_s(x, \lambda) = (1 - \lambda) \sum_{i=1}^{n} x_i + \lambda(x_1 \; s \; \ldots s \; x_n) \tag{3.41}$$

3.4 Sterowniki rozmyte

Kolejny element teorii zbiorów rozmytych oraz systemów wnioskowania rozmytego stanowią sterowniki rozmyte. Podejście to zostało zaproponowane przez L. Zadeha oraz kontynuowane przez Mamdaniego w latach 70. (Zadeh L, 1972; Mamdani 1976). Sterowniki rozmyte polegają na mechanizmach decyzyjnych, które uwzględniają niepewność i subiektywność analizowanych danych wejściowych i wyjściowych. Za sprawą zmiennych lingwistycznych i regułowości tworzą alternatywę dla kontroli systemów lub procesów, w sytuacji gdy nie jest możliwe skonstruowanie dokładnego modelu matematycznego. Sterowniki rozmyte posiadają następującą strukturę:



Rysunek 3.4. Struktura sterowników rozmytych

Wejście sterownika stanowi wektor x n-elementowy zawierający wartości rzeczywiste dla n zmiennych lingwistycznych.

Blok rozmywania: przeprowadza proces kojarzenia rzeczywistych wartości wejściowych z wartościami (termami) odpowiadających im zmiennych lingwistycznych. Każda wartość rzeczywista może być skojarzona z jednym lub kilkoma termami.

Baza reguł: stanowi listę reguł wnioskowania, opisujące działania, które należy podjąć w zależności od kombinacji wartości wejściowych. Reguły przyjmują formę:

$$IF \ \propto \ IS \ A \ THEN \ \beta \ IS \ B \tag{3.42}$$

Gdzie $\alpha = A$ stanowi przesłankę, a $\beta = B$ konkluzję, człon WTEDY pełni rolę operatora implikacji we wnioskowaniu dedukcyjnym *modus ponens*. Przesłanka może stanowić formułę złożoną z dysjunkcji lub koniunkcji kilku wyrażeń logicznych. Zakładając, że przesłanka stanowi koniunkcję dwóch wyrażeń, forma (3.26) przyjmie postać:

$$IF \ \propto \ IS \ A \ \text{AND} \ \beta \ IS \ B \ THEN \ \gamma \ IS \ C \tag{3.43}$$

Reguły sterownika rozmytego składają się z kombinacji wartości zmiennych lingwistycznych dla zmiennych wejściowych oraz jednej wartości zmiennej lingwistycznej wyjściowej. Zapis reguły przyjęto zgodnie z normą IEC 1131 Międzynarodowej Komisji Elektrotechnicznej. Określono w niej sposoby implementacji aplikacji sterowania rozmytego w językach sterowników programowalnych (International Electrotechnical Commission, 2000).

Blok wnioskowania: obejmuje skonstruowanie wyjściowych zbiorów rozmytych $B^{(1)},\ldots, B^{(k)}$, gdzie $k \in \{1, 2, \ldots, k\}$ dla każdej z k reguł znajdujących się w bazie reguł. Aplikując poszczególne reguły, wartość dla przesłanki p w przypadku formuły złożonej z kilku wyrażeń logicznych może stanowić wynik dowolnej operacji triangularnej. Przy wyznaczaniu wyjściowego zbioru rozmytego $B^{(k)}$ najczęściej stosuje się jeden wybrany spośród dostępnych operator implikacji (3.1.7). Najpopularniejszy wariant stanowi operator Mamdaniego interpretowany jako t-norma minimum, operator implikacji Larssena jako t-norma algebraiczna. Następnie zbiory wyjściowe są agregowane do jednego zbioru rozmytego B^* z wykorzystaniem operacji sumy zbiorów rozmytych implementowanej przez t-konormę maksimum.

Blok wyostrzania: stanowi ostatni etap działania sterownika rozmytego. W tym kroku następuje proces defuzyfikacji, czyli wyznaczenia wyjściowej wartości rzeczywistej x^* na podstawie zagregowanego zbioru rozmytego z bloku wnioskowania. Do wyboru wartości wyjściowej stosuje się najczęściej następujące metody:

1. metoda wysokości zbioru rozmytego

$$A(x^*) \geq A(x)\ dla\ każdego\ x \in X,$$

   wyznacza wartość x^*, dla której zbiór rozmyty osiąga najwyższy stopień przynależności. Metoda stosowana w przypadku, gdy A osiąga jedno ekstremum globalne.

2. metoda pierwszego maksimum

$$x^* = min\{x_i \mid \exists_{x_i \in X} : A(x_i) = \max_{x_i \in X} A(x_i)\},$$

   zwraca pierwszą wartość x, dla której zbiór rozmyty osiąga najwyższy stopień przynależności.

3. metoda ostatniego maksimum

$$x^* = max\{x_i \mid \exists_{x_i \in X} : A(x_i) = \max_{x_i \in X} A(x_i)\},$$

zwraca ostatnią wartość x, dla której zbiór rozmyty osiąga najwyższy stopień przynależności.

4. metoda średniej z maksimum

$$x^* = \frac{\sum_{x_i \in M} x_i}{|M|},$$

$$gdzie\ M = \{x_i \mid \exists_{x_i \in X} : A(x_i) = \max_{x_i \in X} A(x_i)\},$$

5. metoda środka ciężkości

1. $$x^* = \frac{\sum_{i=1}^{n} x_i A(x_i)}{\sum_{i=1}^{n} A(x_i)}$$ lub 2. $$x^* = \frac{\int x_i A(x_i)}{\int A(x_i)}$$

gdzie x_i stanowi element ze zbioru rozmytego, a $A(x_i)$ stopień przynależności x_i do zbioru rozmytego. Dla zbioru rozmytego w postaci dyskretnej występuje formuła 1., dla ciągłej 2.

6. metoda średniej ważonej

Jeżeli zbiór rozmyty A przedstawiony jest w postaci $A = A_1 \cup \ldots \cup A_k$

$$x^* = \frac{\sum_{i=1}^{k} x_i A_i(x_i)}{\sum_{i=1}^{k} A_i(x_i)},$$

gdzie x_i wyznacza wartość środkową A_i. Metoda ta wykorzystywana jest głównie w przypadku symetrycznych funkcji przynależności.

Metody defuzyfikacji stanowią rozdziały prac m. in.: Leekwijck i Keree, 1999; Kacprzyk 2001; Timothy, 2004; Nowicki 2009; Sivanandam 2007).

3.5 Moc zbioru rozmytego

Wraz z narastającą złożonością zbiorów danych, ich zliczanie staje się kluczowym wyzwaniem w analizie danych. W kontekście nieprecyzyjnego zliczania (wyznaczania mocy zbioru rozmytego) stosuje się uwzględnienie wartości stopni przynależności ich elementów, co pozwala na bardziej elastyczne podejście w interpretacji danych (niż w przypadku zbiorów klasycznych). Takie metody zliczania pozwalają na wydobycie rozmytych zależności między danymi, co może prowadzić

do lepszego zrozumienia analizowanych danych i bardziej precyzyjnych wniosków. Tematyka liczności zbiorów rozmytych podejmowana była od początków powstania teorii zbiorów rozmytych przez L. Zadeha (Zadeh, 1965, Zadeh 1977) oraz De Luca i Termini (De Luca i Termini, 1972), Dubois i Prade (Dubois i Prade, 1993). Rozszerzeniem tematyki zajmowali się również Wygralak (Wygralak 2012; Wygralak 2013) oraz Dyczkowski (Dyczkowski 2018). W tej części zaprezentowano pojęcia mocy zbiorów rozmytych w podejściu skalarnym oraz przedstawiono najpopularniejsze warianty funkcji wagowych, które uwzględniane są na etapie wyznaczania liczności zbiorów rozmytych.

3.5.1 Moc skalarna

Definicja 3. 7 Funkcja σ: FFS → [0, ∞) nazywana jest mocą skalarną, jeżeli spełnia następujące warunki dla każdego $a, b \in [0, 1]$, $A, B \in$ FFS oraz $x, y \in U$:

1. $\sigma(1/x) = 1$, (zgodność)
2. $\sigma(a/x) \leq \sigma(b/y)$, gdy $a \leq b$, (monotoniczność)
3. $\sigma(A \cup B) = \sigma(A) + \sigma(B)$, gdy $A \cap B = 1_D$. (addytywność)

Kiedy σ spełnienia powyższe warunki σ(A), nazywa się mocą skalarną zbioru rozmytego A.

Definicja 3. 8 Funkcja $f : [0, 1] \rightarrow [0, 1]$ nazywana jest funkcją wagową (lub funkcją wzorcową), jeżeli spełnia następujące warunki dla każdego $a, b \in [0, 1]$:

1. $f(a) \leq f(b)$, gdy $a \leq b$, (monotoniczność)
2. $f(0) = 0$; $f(1) = 1$. (warunki brzegowe)

Stosując funkcje wagowe (z definicji 3.7), przy wyznaczaniu mocy skalarnej (3.8), otrzymujemy moc skalarną indukowaną przez funkcję wagową *f*.

Definicja 3.9 Funkcja $\sigma_f : FFS \rightarrow [0, 1]$ nazywana jest mocą skalarną indukowaną przez funkcję wagową *f* zdefiniowaną jako:

$$\sigma_f = \sum_{x \in supp(A)} f(A(x)), \tag{3.44}$$

gdzie *A* jest skończonym zbiorem rozmytym, a funkcja *f* jest funkcją wagową spełniającą warunki wskazane w definicji 3.8.

3.5.2 Funkcje wagowe

Rolą funkcji wagowych jest określenie, jak istotny jest dany element zbioru rozmytego przy wyznaczaniu jego liczności. Poniżej zestawiono najpopularniejsze funkcje wagowe stanowiące kluczowy element przy określaniu mocy skalarnych zbiorów rozmytych (Wygralak 2012; Wygralak 2013; Dyczkowski 2018).

1. Zliczanie przez progowanie, gdzie $t \in (0,1]$ (Rys. 3.5).

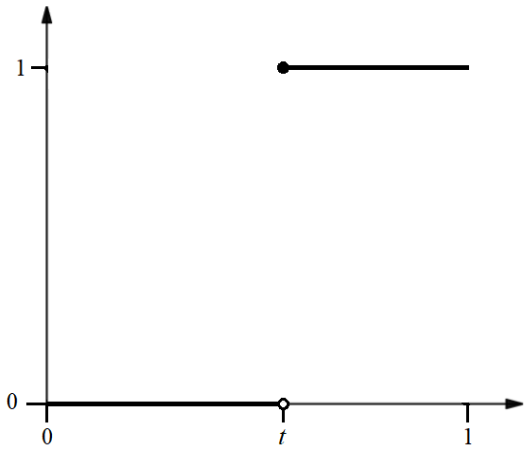
$$f_{1,t}(x) = \begin{cases} 1, & gdy\ x \geq t \\ 0, & w\ p.p. \end{cases} \tag{3.45}$$

2. Zliczanie przez łączenie, gdzie $p \geq 0$ (Rys 3.6).

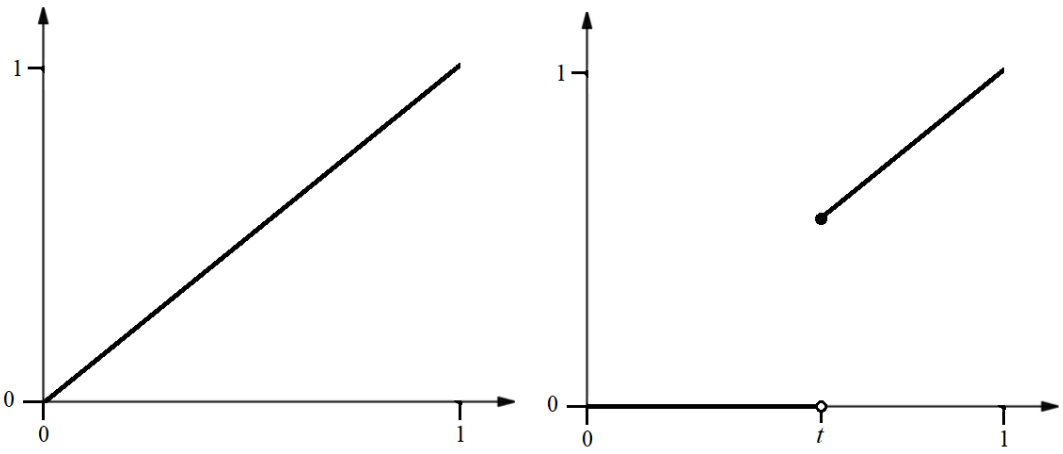
$$f_{2,p}(x) = x^p \tag{3.46}$$

3. Zliczanie przez progowanie i łączenie, gdzie $t \in (0,1]$, $p \geq 0$ (Rys 3.6).

$$f_{3,p,t}(x) = \begin{cases} x^p, & gdy\ x \geq t \\ 0, & w\ p.p. \end{cases} \tag{3.47}$$



Rysunek 3.5. Funkcja wagowa zliczania przez progowanie $f_{1,t}$ w punkcie *t*



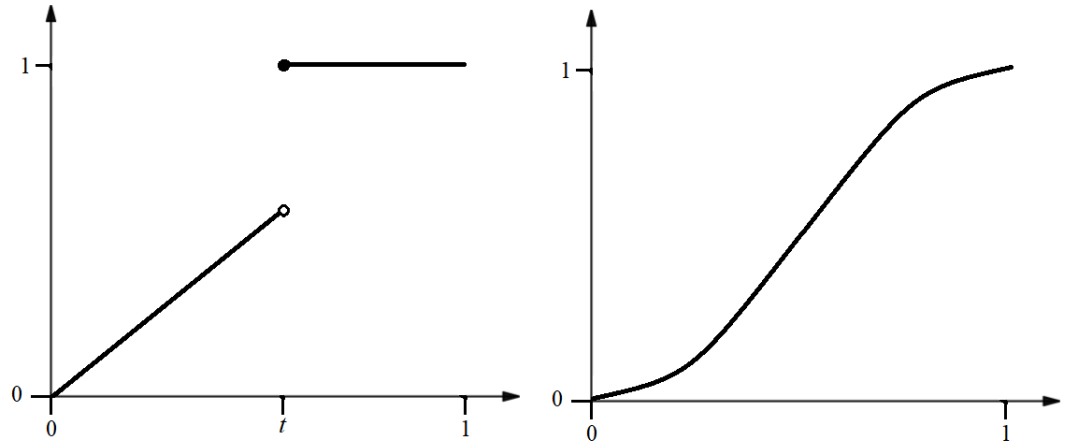
Rysunek 3.6. Funkcja wagowa zliczania przez łączenie $f_{2,p}$ z p=1 (lewa), funkcja wagowa zliczania przez progowanie i łączenie $f_{2,p,t}$ z p=1 w punkcie t (prawa)

4. Zliczanie przez wyostrzenie, gdzie $t \in (0,1], p \geq 0$.

$$f_{4,p,t}(x) = \begin{cases} 1, & gdy\ x \geq t \\ x^p, & w\ p.p. \end{cases} \tag{3.48}$$

5. Wyostrzenie kontrastu (Rys 3.7).

$$f_5(x) = \begin{cases} 2x^2, & gdy\ x \leq 0{,}5 \\ 1 - 2(1-x)^2, & w\ p.p. \end{cases} \tag{3.49}$$



Rysunek 3.7. Funkcja wagowa zliczania przez wyostrzenie $f_{4,p,t}$ z p=1 w punkcie t (lewa), funkcja wagowa wyostrzenia kontrastu f_5 (prawa)

Dla funkcji wagowych, które uwzględniają progowanie alternatywnie można zastosować progowanie ostre (nie będą jednak one wykorzystywane w badaniu klasyfikacji dominującej dyscypliny podejściem rozmytym). Dla funkcji 3.45, 3.47 oraz 3.48 uzyskujemy odpowiednio:

1. Zliczanie przez ostre progowanie, gdzie $t \in [0,1)$.

$$f_{6,t}(x) = \begin{cases} 1, & gdy\ x > t \\ 0, & w\ p.p. \end{cases} \tag{3.50}$$

2. Zliczanie przez ostre progowanie i łączenie, gdzie $t \in [0,1)$, $p \geq 0$.

$$f_{7,p,t}(x) = \begin{cases} x^p, & gdy\ x > t \\ 0, & w\ p.p. \end{cases} \tag{3.51}$$

3. Zliczanie przez ostre wyostrzenie, gdzie $t \in [0,1)$, $p \geq 0$.

$$f_{8,p,t}(x) = \begin{cases} 1, & gdy\ x > t \\ x^p, & w\ p.p. \end{cases} \tag{3.52}$$

Niektóre funkcje wagowe wykorzystujące progowanie ostre lub nieostre z wartościami skrajnymi t=0 lub t=1 są tożsame z innymi wariantami funkcji wagowych. Zliczanie przez progowanie i łączenie z t=1 oraz przez wyostrzenie z t=1 odpowiada funkcji zliczania przez progowanie z t=1. Zliczanie przez ostre progowanie i łącznie z parametrem t=0 oraz przez ostre wyostrzenie z t=0 odpowiada funkcji zliczania przez łączenie. W związku z tym opisane warianty funkcji wzorcowych nie stanowią przedmiotu badań w pracy.

Wyznaczając moc zbioru rozmytego dobór funkcji wagowej określa nam, jak istotny jest dany element w procesie zliczania zgodnie z następującymi zasadami:

- zliczanie przez progowanie oraz przez ostre progowanie uwzględnia jedynie elementy (z taką samą wyższością; wartość 1 dla każdego elementu), które stanowią odcięcie zbioru rozmytego i ignoruje wartości poniżej ustalonego progu przypisując wartość 0,
- zliczanie przez łączenie uwzględnia wszystkie elementy sumując stopnie przynależności, dodatkowo stosując wyższy parametr p uzyskuje się mniejszą wartość funkcji wagowej, w szczególności dla p=1 jest to funkcją tożsamościowa,
- zliczanie przez progowanie i łączenie oraz ostre progowanie i łączenie dla elementów stanowiących odcięcie zbioru rozmytego uwzględnia je w ten sam sposób, jak w przypadku zliczania przez łączenie, poniżej odcięcia elementy otrzymują wartość 0,

- zliczanie przez wyostrzenie oraz ostre wyostrzenie promuje wartości powyżej ustalonego progu przypisując im wartość 1, pozostałe wartości uwzględnia zgodnie z wartościami ich funkcji przynależności,
- wyostrzenie kontrastu uwydatnia wartości większe od 0.5 oraz zmniejsza wartości pozostałych elementów.

Rozdział 4.
Klasyfikacja dominującej dyscypliny z wykorzystaniem zmiennych lingwistycznych

4.1 Wprowadzenie

W tym rozdziale opisano proces badania skuteczności klasyfikacji dominującej dyscypliny z wykorzystaniem zmiennych lingwistycznych. W części metodologicznej przedstawiono proces konstruowania zmiennych z odpowiednimi termami oraz funkcjami przynależności. Dodatkowo opisano metody pozyskiwania danych z platformy ICSR Lab, które umożliwiły konstrukcję zaproponowanych zmiennych. Kolejna część metodologii stanowi opis modyfikacji algorytmu wyznaczania dominującej dyscypliny z uwzględnieniem zmiennych lingwistycznych (podejście rozmyte) oraz prezentuje listę wybranych do badania wariantów funkcji wagowych na etapie wyznaczania mocy zbiorów rozmytych w zmodyfikowanej wersji algorytmu. Kolejna część rozdziału prezentuje wyniki z przeprowadzonych badań. Podzielono je na dwie części. W pierwszej części opisano rozkłady gęstości obserwacji stanowiących dziedzinę badanych funkcji wagowych. Druga część wyników stanowi opis rezultatów uzyskanych dla algorytmu klasyfikacji dominującej dyscypliny w podejściu rozmytym. Dla algorytmu w podejściu rozmytym, wykorzystującego każdy wariant zmiennej lingwistycznej i termu opisano 5 wariantów funkcji wagowych, dla których dokonano najwyższej klasyfikacji oraz wykonano ewaluację wyników uzyskanych w podejściu rozmytym z wynikami z podejścia bazowego. Ostatnia część rozdziału stanowi dyskusję nad wynikami badania. *(W tym rozdziale nazwy zmiennych lingwistycznych zapisano pogrubioną kursywą w celu ich wyszczególnienia.)*

4.2 Metodologia

4.2.1 Konstrukcja zmiennych lingwistycznych

Na potrzeby badania skonstruowano 7 zmiennych lingwistycznych (***Cytowanie***, ***FWCI 4-letnie***, ***FWCI 5-letnie***, ***FWCI bez ram czasowych***, ***Percentyl***, ***Zespół*** i ***Rok***). Każda zmienna lingwistyczna występowała w wariancie dyscypliny naukowej (26 dyscyplin na każdą zmienną). Uniwersum wartości zmiennych lingwistycznych w dyscyplinach obejmował przedział wartości

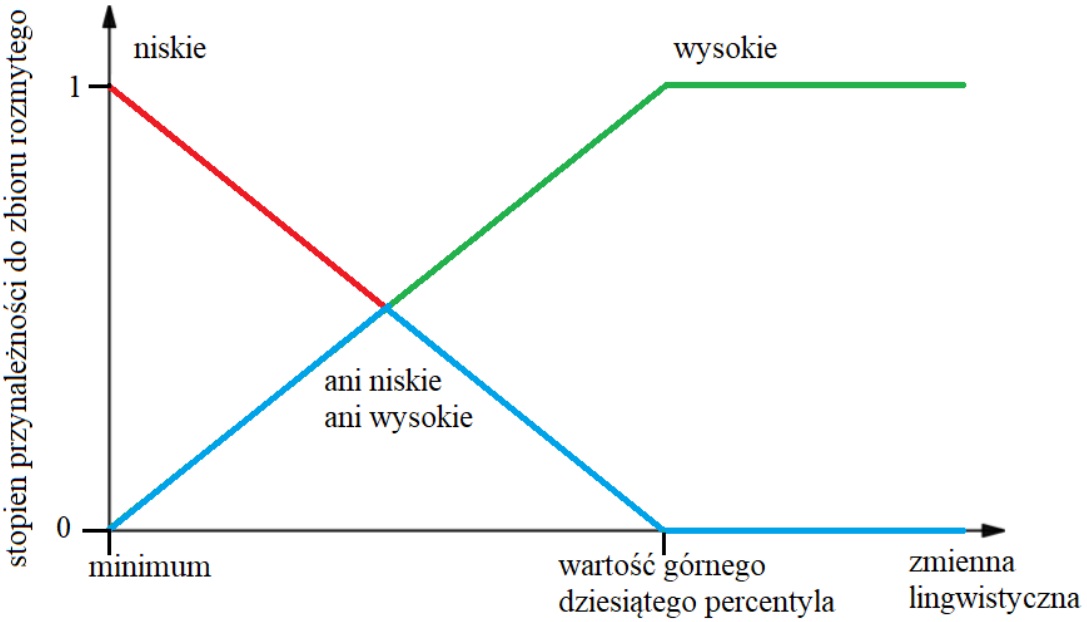
od minimum do maksimum wyznaczonego na podstawie tabeli publikacji z bazy Scopus. Każda zmienna lingwistyczna występowała w trzech termach, które modelowały odpowiednio niską, ani niską ani wysoką oraz wysoką przynależność do zbioru rozmytego (Tab. 4.1).

Tabela 4.1. Lista wybranych zmiennych lingwistycznych wraz z termami

Nazwa zmiennej	Termy
Cytowanie	{niskie, ani niskie ani wysokie, wysokie}
FWCI 4-letnie	{niskie, ani niskie ani wysokie, wysokie}
FWCI 5-letnie	{niskie, ani niskie ani wysokie, wysokie}
FWCI bez ram czasowych	{niskie, ani niskie ani wysokie, wysokie}
Percentyl	{niski, ani niski ani wysoki, wysoki}
Rok	{najstarszy, ani najstarszy ani najnowszy, najnowszy}
Zespół	{najmniejszy, ani najmniejszy ani największy, największy}

Modelowanie funkcji przynależności przebiegało w ten sam sposób w przypadku każdej zmiennej lingwistycznej i dyscypliny naukowej. Dla termów oznaczających wysoką przynależność do zbioru rozmytego funkcja przynależności osiągała wartość 1 od wartości górnego 10-tego percentyla wyznaczonego na podstawie publikacji w bazie Scopus dla wyznaczanej dyscypliny. Podejście uznania górnego 10-tego percentyla z wartością 1 przyjęto jako wiedzę ekspercką na podstawie prac Kwiek i Roszka, 2024; Narayan i Chogtu, 2021; Qua i Pelfrey, 2020 oraz Marzuqi i inni, 2019. Dla pozostałego dolnego 90-tego percentyla była to rosnąca funkcja liniowa w przedziale od wartości minimalnej dla danej zmiennej w danej dyscyplinie do wartości górnego 10-tego percentyla. Dla termów oznaczających niską przynależność do zbioru rozmytego funkcję przynależności utworzono poprzez negację funkcji przynależności dla termu wysokiej przynależności do zbioru rozmytego. Funkcja przyjmowała wartość 0 od wartości górnego 10-tego percentyla. Dla pozostałego dolnego 90-tego percentyla była to malejąca funkcja liniowa w przedziale od wartości minimalnej dla danej zmiennej w danej dyscyplinie do wartości górnego 10-tego percentyla.

Termy oznaczające ani niską ani wysoką przynależność do zbioru rozmytego stanowiły wartość t-normy minimum z wartości funkcji dla termów niskiej przynależności do zbioru rozmytego i wysokiej przynależności do zbioru rozmytego. W związku z tym term ani niskiej ani wysokiej przynależności przyjmował zbiór wartości stopni przynależności z przedziału od 0 do 0.5 włącznie. Schemat tworzenia funkcji przynależności wykorzystany dla każdej zmiennej lingwistycznej i ustalonych trzech termów przedstawiono na rysunku 4.1. Pełna lista wartości dla górnego 10-tego percentyla w ujęciu dyscyplin i zaproponowanych zmiennych lingwistycznych przedstawia tabela 4.2.



Rysunek 4.1. Interpretacja dowolnej zmiennej lingwistycznej w badaniu z termami niskie, ani niskie ani wysokie, wysokie

Tabela 4.2. Wartości górnego 10-tego percentyla dla zmiennych lingwistycznych na podstawie publikacji z bazy Scopus oraz liczba publikacji po dyscyplinach

Dyscyplina	Cytowanie	FWCI 4-letnie	FWCI 5-letnie	FWCI bez ram czasowych	Percentyl	Zespół	Rok
AGRI	53	2.36	2.34	2.33	92	7	2020
ARTS	26	2.59	2.53	2.49	91	3	2020
BIOC	74	2.55	2.53	2.46	91	9	2019
BUSI	54	2.98	2.96	2.91	96	4	2020
CENG	66	2.94	2.92	2.88	97	7	2020
CHEM	60	2.67	2.64	2.57	94	7	2019
COMP	45	2.84	2.81	2.68	95	5	2020
DECI	56	2.95	2.95	2.95	95	4	2020
DENT	45	2.42	2.40	2.38	83	7	2019
EART	55	2.49	2.47	2.35	91	7	2019
ECON	46	2.75	2.70	2.60	92	4	2020
ENER	53	3.09	3.08	3.04	97	7	2020
ENGI	40	2.83	2.79	2.62	94	6	2020
ENVI	55	2.79	2.76	2.69	97	7	2020
HEAL	43	2.62	2.59	2.52	93	7	2020
IMMU	73	2.57	2.54	2.48	90	10	2019
MATE	50	2.80	2.78	2.66	94	7	2020
MATH	35	2.48	2.44	2.32	90	4	2020
MEDI	44	2.19	2.19	2.21	82	8	2018
NEUR	80	2.57	2.55	2.47	93	9	2019
NURS	40	2.59	2.56	2.47	90	7	2020
PHAR	47	2.33	2.31	2.24	85	8	2019
PHYS	51	2.70	2.68	2.50	91	7	2019
PSYC	64	2.64	2.63	2.58	92	6	2020
SOCI	34	2.6	2.56	2.46	92	4	2020
VETE	33	2.46	2.4	2.31	86	8	2019

4.2.2 Dołączenie dziedziny dla zmiennych lingwistycznych do tabeli publikacji

Przy konstrukcji wybranych zmiennych lingwistycznych wymagane było pozyskanie wartości, które stanowiłyby dziedzinę zbiorów rozmytych. W związku z tym przy konstrukcji tabeli publikacji uwzględniono kolumny zawierające wartości cytowania publikacji, współczynników FWCI, percentyla czasopisma, liczności zespołu oraz roku wydania publikacji.

W przypadku wartości cytowania i roku wydania publikacji skorzystano z kolumn znajdujących się w tabeli publikacji. Wartość wskaźnika FWCI 4-letniego, FWCI 5-letniego oraz FWCI bez ograniczeń czasowych została dołączona do tabeli publikacji poprzez złączenie lewostronne z tabelą wskaźników FWCI ze względu na unikalny identyfikator publikacji (lewa strona: tabela publikacji, prawa strona: tabela wskaźników FWCI). W celu pozyskania wartości percentyla czasopisma dla publikacji skorzystano z tabeli czasopism. Z tabeli czasopism wykorzystano kolumnę z identyfikatorem czasopisma oraz kolumnę zwierającą metrykę CiteScore. Następnie dla każdego czasopisma wybrano percentyl metryki CiteScore z roku 2021. Wartość percentyla czasopisma została dołączona do tabeli publikacji poprzez złączenie lewostronne ze względu na unikalny identyfikator czasopisma (lewa strona: tabela publikacji, prawa strona: tabela czasopism). Aby określić liczność zespołu wykorzystano kolumnę z identyfikatorami autorów z tabeli publikacji. Na podstawie liczby identyfikatorów autorów wyznaczono ich liczność. W związku z wykonywanymi złączeniami lewostronnymi na tabeli publikacji liczba rekordów w tabeli publikacji nie zmieniła się w porównaniu do tabeli, na podstawie której dokonano klasyfikacji dominującej dyscypliny autora w podejściu bazowym.

Ze względu na występujące braki w wartościach komórek niektórych kolumn zastosowano ich wypełnienie wartościami domyślnymi. W przypadku rekordów, które nie posiadały wartości z liczbą cytowań lub wartości odpowiedniego wskaźnika FWCI, przypisano im wartość zero (oznaczające brak uzyskanych cytowań i zerową wartość wskaźnika FWCI). Rekordom, które nie posiadały wartości percentyla czasopisma, przypisano wartość jeden (najniższy percentyl). Ze względu na charakter zbierania metadanych do bazy Scopus przez firmę Elsevier w tabeli publikacji nie występowały rekordy, które nie posiadały wartości w kolumnie rok. Kolumna nie była uzupełniana o brakujące wartości. Ze względu na wymóg do operowania na danych na poziome autora nie było wymagane rozpatrywanie braku wartości liczności zespołu w niektórych publikacjach.

4.2.3 Algorytm wyznaczania dominującej dyscypliny z wykorzystaniem zmiennej lingwistycznej

Algorytm wyznaczania dominującej dyscypliny w podejściu bazowym zmodyfikowano tak, aby wykorzystywał wybraną zmienną lingwistyczną, term i wyznaczenia mocy zbiorów rozmytych z zastosowaniem badanych funkcji wagowych (wariant algorytmu nazywany podejściem rozmytym). Na początku do kolumn zawierających unikalny identyfikator publikacji, listę identyfikatorów autorów oraz listę dyscyplin przypisanych do czasopisma, z którego pochodziła publikacja dołączono kolumnę reprezentującą wartości ostre zmiennej lingwistycznej. Następnie, podobnie jak w podejściu bazowym, dokonano normalizacji tabeli do pierwszej postaci normalnej ze względu na listę identyfikatorów autorów i listę dyscyplin. Kolejny krok obejmował wyznaczenie i dołączenie dla każdego rekordu wartości stopnia przynależności do zbioru rozmytego dla wybranej zmiennej lingwistycznej, termu i przypisanej rekordowi dyscypliny naukowej. Po wyznaczeniu stopnia przynależności przystąpiono do wyznaczania wartości dla badanej funkcji wagowej oraz usunięto rekordy, dla których wartość tej funkcji wynosiła zero (pozostawiając rekordy, w których stopień przynależności do zbioru rozmytego stanowił jego nośnik). Następnie dokonano wyznaczenia mocy zbiorów rozmytych dla każdej dyscypliny autora.

Kolejne części algorytmu przebiegały podobnie do wariantu w podejściu bazowym, z tym że zamiast wyznaczać maksymalną wartość z liczby publikacji w dyscyplinie autora wyznaczono maksymalną wartość ze wszystkich wartości mocy zbiorów rozmytych dyscyplin autora. Z tabeli wybrano tylko te rekordy, dla których moc zbioru rozmytego dla dyscypliny autora była równa wartości maksymalnej mocy zbioru dla dyscypliny autora. Następnie zbiór autorów ograniczono jedynie do tych, którzy posiadali jedną dyscyplinę, dla której osiągnięto maksymalną wartość mocy zbioru rozmytego zmiennej lingwistycznej w przypisanym termie.

Algorytm 2. Algorytm klasyfikacji dominującej dyscypliny z wykorzystaniem zmiennej lingwistycznej

Wejście: zbiór publikacji w bazie Scopus,
Parametry: zmienna lingwistyczna, term, funkcja wagowa
Wyjście: zbiór par autor-dyscyplina

1. Dołącz do tabeli kolumnę z wartościami rzeczywistymi dla zmiennej lingwistycznej
2. Wykonaj normalizację tabeli do pierwszej postaci normalnej ze względu na identyfikatory autorów i listę dyscyplin

3. Wyznacz wartość stopnia przynależności do zbioru rozmytego dla zmiennej lingwistycznej i termu w dyscyplinie
4. Oblicz wartość wybranej funkcji wagowej dla wartości wyznaczonej w kroku 3.
5. Wyznacz nośnik wybranego zbioru rozmytego
6. Dla każdego autora i jego dyscyplin oblicz moc zbioru rozmytego
7. Dla każdego autora wyznacz maksymalną z wartości mocy zbiorów rozmytych
8. Dla każdego autora wybierz dyscypliny, dla których moc zbioru rozmytego odpowiada maksymalnej z wartości mocy zbiorów rozmytych
9. Dokonaj filtracji tabeli do autorów posiadających jedną zaklasyfikowaną dyscyplinę

4.2.4 Wybór badanych funkcji wagowych

Do badania klasyfikacji algorytmu w podejściu rozmytym wybrano 5 funkcji wagowych: zliczanie przez progowanie, zliczanie przez łączenie, zliczanie przez progowanie i łączenie, zliczanie przez wzmocnienie oraz wzmocnienie kontrastu. W przypadku zliczania przez łączenie, progowanie i łączenie oraz wzmocnienie wybrano 3 warianty wartości p (0.5, 1 i 2). Dla zliczania przez progowanie, progowanie i łączenie i przez wzmocnienie wybrano 4 warianty wartości t (0.2, 0.4, 0.6, 0.8). Dla termu niskiej i wysokiej przynależności do zbioru rozmytego możliwe było badanie wszystkich dostępnych kombinacji funkcji wagowych (34 warianty funkcji wagowych na term). W przypadku termu ani niskiej ani wysokiej przynależności do zbioru rozmytego możliwe było badanie jedynie kombinacji funkcji wagowych z t=0.2 lub t=0.4 ze względu na ograniczenie termu w zakresie przyjmowanych wartości stopnia przynależności do zbioru rozmytego do przedziału [0, 0.5] (22 warianty funkcji wagowych). Łącznie dla algorytmu wyznaczania dominującej dyscypliny w podejściu rozmytym uzyskano 88 badanych sposobów w każdej zmiennej lingwistycznej na wykonanie klasyfikacji (Tab. 4.3).

Tabela 4.3. Lista wybranych funkcji wagowych i wartości progowania w ujęciu termów

Funkcja wagowa	t	Term		
		niska	ani niska ani wysoka	wysoka
Zliczanie przez progowanie	0.2	✓	✓	✓
Zliczanie przez progowanie	0.4	✓	✓	✓
Zliczanie przez progowanie	0.6	✓		✓
Zliczanie przez progowanie	0.8	✓		✓
Zliczanie przez łączenie p=0.5	-	✓	✓	✓
Zliczanie przez łączenie p=1	-	✓	✓	✓
Zliczanie przez łączenie p=2	-	✓	✓	✓
Zliczanie przez progowanie i łączenie p=0.5	0.2	✓	✓	✓
Zliczanie przez progowanie i łączenie p=0.5	0.4	✓	✓	✓
Zliczanie przez progowanie i łączenie p=0.5	0.6	✓		✓

Zliczanie przez progowanie i łączenie p=0.5	0.8	✓		✓
Zliczanie przez progowanie i łączenie p=1	0.2	✓	✓	✓
Zliczanie przez progowanie i łączenie p=1	0.4	✓	✓	✓
Zliczanie przez progowanie i łączenie p=1	0.6	✓		✓
Zliczanie przez progowanie i łączenie p=1	0.8	✓		✓
Zliczanie przez progowanie i łączenie p=2	0.2	✓	✓	✓
Zliczanie przez progowanie i łączenie p=2	0.4	✓	✓	✓
Zliczanie przez progowanie i łączenie p=2	0.6	✓		✓
Zliczanie przez progowanie i łączenie p=2	0.8	✓		✓
Zliczanie przez wzmocnienie p=0.5	0.2	✓	✓	✓
Zliczanie przez wzmocnienie p=0.5	0.4	✓	✓	✓
Zliczanie przez wzmocnienie p=0.5	0.6	✓		✓
Zliczanie przez wzmocnienie p=0.5	0.8	✓		✓
Zliczanie przez wzmocnienie p=1	0.2	✓	✓	✓
Zliczanie przez wzmocnienie p=1	0.4	✓	✓	✓
Zliczanie przez wzmocnienie p=1	0.6	✓		✓
Zliczanie przez wzmocnienie p=1	0.8	✓		✓
Zliczanie przez wzmocnienie p=2	0.2	✓	✓	✓
Zliczanie przez wzmocnienie p=2	0.4	✓	✓	✓
Zliczanie przez wzmocnienie p=2	0.6	✓		✓
Zliczanie przez wzmocnienie p=2	0.8	✓		✓
Wzmocnienie kontrastu p=0.5	-	✓	✓	✓
Wzmocnienie kontrastu p=1	-	✓	✓	✓
Wzmocnienie kontrastu p=2	-	✓	✓	✓

4.3 Wyniki

4.3.1 Badanie rozkładu gęstości dziedziny funkcji wagowych

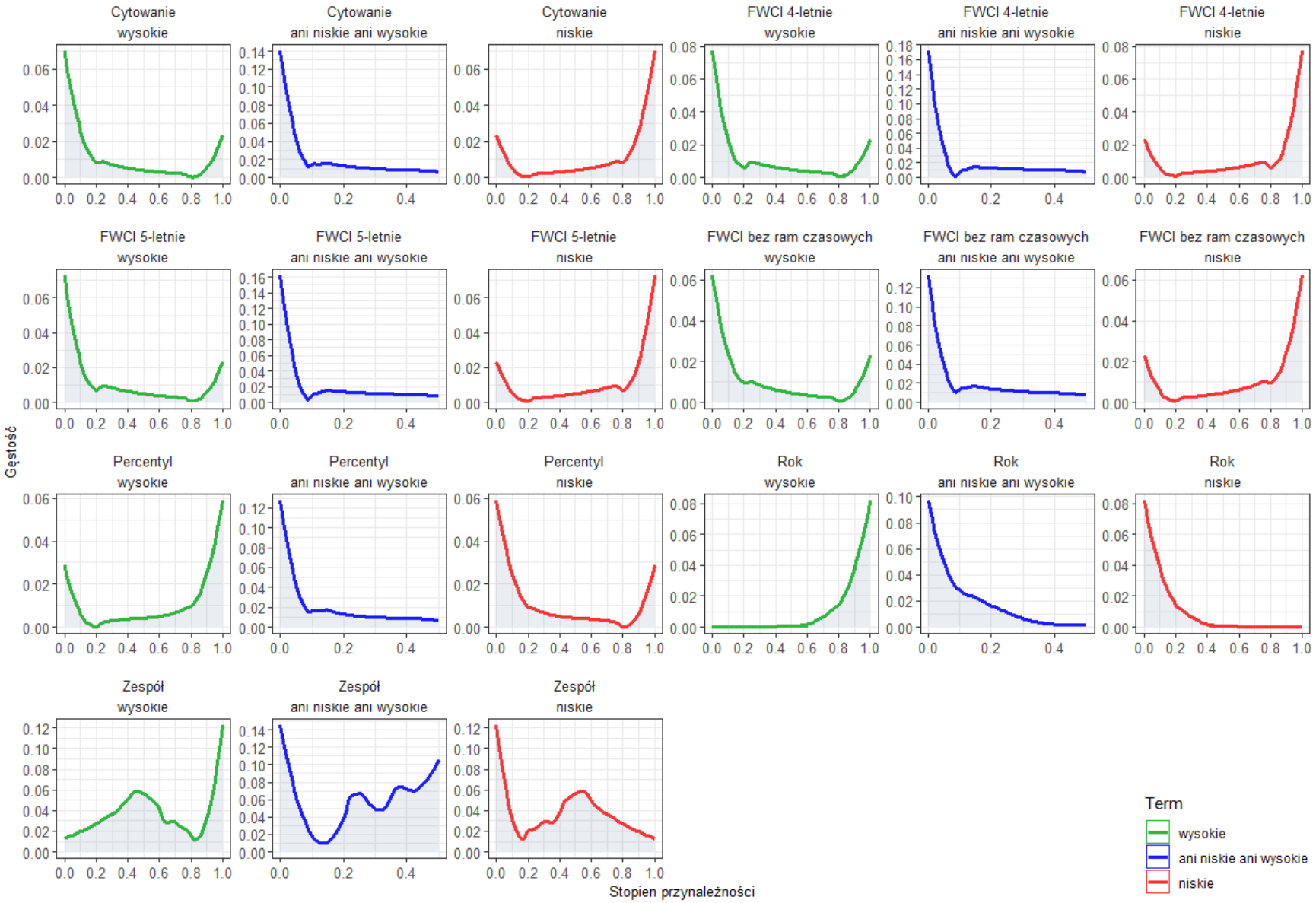
W pierwszym etapie przeanalizowano rozkłady wartości dziedziny funkcji wagowych dla każdej zmiennej lingwistycznej w każdym termie. Głównej analizie poddano termy wysokiej przynależności do zbioru rozmytego oraz ani niskiej ani wysokiej przynależności do zbioru rozmytego. Rozkłady gęstości jądrowych dla termu niskiej przynależności do zbioru rozmytego stanowiły symetrię rozkładu wysokiej przynależności do zbioru rozmytego względem punktu osi X w punkcie 0.5.

Dla czterech badanych zmiennych lingwistycznych (***Cytowanie***, ***FWCI 4-letenie***, ***FWCI 5-letnie oraz FWCI bez ram czasowych***) w termie wysokiej przynależności występuje najwyższa koncentracja obserwacji w przedziale od 0 do 0.2 (odpowiednio najwyższa koncentracja w przedziale od 0.8 do 1 dla termu niskiej przynależności). Pomiędzy trzema wariantami zmiennej reprezentującej wskaźnik ***FWCI*** można zauważyć spadek w różnicy gęstości jądrowej w dominującym przedziale (dla ***FWCI 4-letniego*** gęstości w przedziale od około 0.6% do 8%,

dla ***FWCI 5-letniego*** w przedziale od około 0.6% do 7% oraz dla ***FWCI bez ram czasowych*** w przedziale od 1% do 6%). Te same zmienne w termie ani niskiej ani wysokiej przynależności osiągnęły najwyższą gęstość jądrową w przedziale od około 0 do 0.08% zmniejszając swój zakres odpowiednio dla ***FWCI 4-letniego*** od 0% do 18%, ***FWCI 5-letniego*** od około 0% do 16% i ***FWCI bez ram czasowych*** od 1% do 13%.

Dla zmiennej lingwistycznej ***Rok*** w termie wysokiej przynależności uzyskano największą koncentrację obserwacji w przedziale od 0.9 do 1 (odpowiednio najwyższej dla przedziału od 0 do 0.1 dla termu niskiej przynależności). Rozkład gęstości zmiennej w termie stanowi funkcję rosnącą z największym wzrostem w przedziale od 0.7 do 1. Dla przedziału od 0 do 0.6 uzyskano znikomą gęstość rozkładu obserwacji. W przypadku termu ani niskiej ani wysokiej przynależności rozkład jest funkcją malejącą dla przedziału od 0 do 0.5. Zmienna lingwistyczna ***Percentyl*** w termie wysokiej przynależności uzyskała najwyższą koncentrację obserwacji w przedziale od 0.8 do 1 (odpowiednio od 0 do 0.2 dla termu niskiej przynależności). Rozkład obserwacji jest funkcją rosnącą w przedziale od 0.2 do 0.8 oraz malejącą w przedziale od 0 do 0.2. Dla obserwacji w termie ani niskiej ani wysokiej przynależności uzyskano największy rozkład obserwacji w przedziale od 0 do 0.1. W przypadku zmiennej lingwistycznej ***Zespół*** w termie wysokiej przynależności największy rozkład obserwacji występuje w przedziale od 0.4 do 0.6 oraz w punkcie 1 (odpowiednio w przedziale od 0.4 do 0.6 oraz wartości 0 dla termu niskiej przynależności). Dla termu ani niskiej ani wysokiej przynależności uzyskano największą koncentrację obserwacji w przedziale od 0.2 do 0.5 oraz w punkcie 0.

Dla czterech zmiennych (***Cytowanie*** i 3 warianty wskaźnika ***FWCI***) występuje większa redukcja obserwacji w nośniku zbioru rozmytego dla termu wysokiej przynależności niż niskiej przynależności. Dla zmiennej ***Cytowanie*** i ***FWCI 5-letniego*** jest to różnica w wartości estymowanej gęstości jądrowej pomiędzy 7% a 2%, dla ***FWCI 4-letniego*** pomiędzy 8% a 2.5%, a dla ***FWCI bez ram czasowych*** pomiędzy 6% a 2%. Dla trzech pozostałych zmiennych (***Percentyl***, ***Rok***, ***Zespół***) występuje mniejsza redukcja obserwacji dla termu wysokiej przynależności niż niskiej przynależności. Dla zmiennej ***Percentyl*** jest to różnica pomiędzy 3% (największa wartość dla opisywanych zmiennych) a 6%, dla ***Roku*** pomiędzy około 0% a 8%, dla zmiennej ***Zespół*** pomiędzy 1% a 12%. W przypadku każdej zmiennej dla termu ani niskiej ani wysokiej przynależności w wyniku wyznaczenia nośnika zbioru rozmytego występuje większa redukcja obserwacji niż w termie niskiej przynależności i wysokiej przynależności (Rys 4.2).



Rys 4.2. Wykres gęstości jądrowej dla dziedziny funkcji wagowych w ujęciu zmiennych lingwistycznych i termów

4.3.2 Ewaluacja algorytmu klasyfikacji dominującej dyscypliny w podejściu rozmytym

Poniższe wyniki przedstawiają ewaluacje klasyfikacji uzyskanej przez algorytm klasyfikacji dominującej dyscypliny z wykorzystaniem zmiennych lingwistycznych. Analizie poddano 5 wariantów funkcji wagowych, których wykorzystanie w algorytmie dokonało najwyższej klasyfikacji, spośród badanych grup funkcji wagowych (zliczania przez progowanie, zliczania przez łączenie, zliczania przez progowanie i łączenie, zliczania przez wyostrzenie oraz wyostrzenia kontrastu). Pełne wyniki zawierające listę wszystkich wariantów, które dokonały wyższej klasyfikacji niż w podejściu bazowym umieszczono w załączniku w sekcji 1. Dla każdego wariantu funkcji wagowej wyznaczono z algorytmu procent zaklasyfikowanych autorów oraz procentowy wzrost liczby zaklasyfikowanych obserwacji w stosunku do liczby zaklasyfikowanych obserwacji w podejściu bazowym. W celu zbadania podobieństwa pomiędzy klasyfikacją dokonaną metodą bazową, a metodą rozmytą wyznaczono metryki klasyfikacyjne dla zbioru testującego. Obejmował on obserwacje zaklasyfikowane zarówno podejściem bazowym i rozmytym. W celu zbadania redukcji klasyfikacji uzyskanej podejściem rozmytym, obliczono procent obserwacji, których nie zaklasyfikowano metodą rozmytą a zaklasyfikowano metodą bazową. Łącznie badano 88 warianty funkcji wagowych dla każdej zmiennej lingwistycznej.

Dla zmiennej lingwistycznej ***Cytowanie*** uzyskano zwiększenie liczby zaklasyfikowanych obserwacji stosując wszystkie 3 termy. Najwięcej obserwacji pozyskano w przypadku niskiego cytowania dla funkcji wagowych zliczania przez łączenie oraz wyostrzenia kontrastu (klasyfikację uzyskano dla 90.66% do 90.67% obserwacji, zwiększając ich liczbę o około 30.91% do 30.93% w stosunku do metody bazowej). Zwiększenie liczby pozyskanych obserwacji uzyskano przy jednoczesnej ewaluacji na dużym zbiorze testującym (od 2.76% do 2.78% redukcji). W przypadku tych funkcji uzyskano wysoką wartość wskaźnika Accuracy (od 88.9% do 92.0%) i MCC (od 89.1% do 92.1%). Dużą liczbę zaklasyfikowanych obserwacji (około 89.6%) otrzymano również w przypadku zliczania przez progowanie i łącznie dla p=0.5 i t=0.2 przy Accuracy i MCC 92.5%. W przypadku termu „ani niskie ani wysokie” i „wysokie” największej klasyfikacji dokonano dla zliczania przez łączenie, wyostrzenia kontrastu oraz zliczania przez wyostrzenie p=0.5 dla t=0.4 ani wysokiego ani niskiego cytowania oraz t=0.8 dla wysokiego cytowania (uzyskując od około 9.62% do 11.29% więcej obserwacji dla termu „ani niskie ani wysokie”

i od około 13.02% do 13.80% dla termu „wysokie"). Przy zastosowaniu tych dwóch termów można zauważyć spadek wartości metryk klasyfikacyjnych w porównaniu do zastosowania wariantu niskiego cytowania. Dla około 35% obserwacji w przypadku zliczania przez łącznie *p*=2 oraz wyostrzania kontrastu w termie „ani niskie ani wysokie" (35.8% Loss) i 50% obserwacji w termie „wysokie" (50.2% Loss) uzyskano klasyfikację innej dyscypliny naukowej. Ze wszystkich badanych funkcji wagowych większą liczbę zaklasyfikowanych obserwacji, niż w przypadku podejścia bazowego, uzyskano dla 51 z 88 badanych wariantów funkcji wagowych (Tab. 4.4).

Tabela 4.4. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej ***Cytowanie*** (5 najlepszych wyników w termie)

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niskie											
Zliczanie przez łączenie *p*=0.5	-	92.0	92.4	92.0	99.8	92.2	92.1	8.0	2.76	90.67	130.93
Zliczanie przez łączenie *p*=1	-	91.6	92.1	91.6	99.8	91.8	91.6	8.4	2.77	90.66	130.92
Zliczanie przez łączenie *p*=2	-	88.9	89.7	88.9	99.6	89.2	89.1	11.1	2.78	90.66	130.91
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	92.5	93.0	92.5	99.8	92.7	92.5	7.5	3.66	89.60	129.38
Wyostrzenie kontrastu	-	90.6	91.2	90.6	99.7	90.9	90.7	9.4	2.77	90.67	130.92
term = ani niskie ani wysokie											
Zliczanie przez łączenie *p*=0.5	-	89.5	89.9	89.5	99.3	89.7	89.3	10.5	22.90	77.07	111.29
Zliczanie przez łączenie *p*=1	-	82.3	82.8	82.3	98.8	82.5	82.2	17.7	22.94	77.04	111.24
Zliczanie przez łączenie *p*=2	-	64.2	65.4	64.2	97.1	64.8	65.0	35.8	22.91	77.06	111.28
Zliczanie przez wyostrzenie *p*=0.5	0.4	83.5	84.0	83.5	99.1	83.7	83.6	16.5	23.11	75.91	109.62
Wyostrzenie kontrastu	-	64.2	65.4	64.2	97.1	64.8	65.0	35.8	22.91	77.06	111.28
term = wysokie											
Zliczanie przez łączenie *p*=0.5	-	86.9	87.1	86.9	99.2	86.9	86.7	13.1	20.12	78.81	113.80
Zliczanie przez łączenie *p*=1	-	67.1	66.7	67.1	97.3	66.9	67.5	32.9	20.18	78.76	113.73
Zliczanie przez łączenie *p*=2	-	49.8	48.5	49.8	96.2	49.1	51.9	50.2	20.19	78.75	113.72
Zliczanie przez wyostrzenie *p*=0.5	0.8	86.9	87.1	86.9	99.2	86.9	86.7	13.1	20.15	78.27	113.02
Wyostrzenie kontrastu	-	50.0	48.9	50.0	96.3	49.4	52.3	50.0	20.17	78.77	113.75

Dla zmiennej lingwistycznej ***FWCI 4-letnie*** uzyskano zwiększenie liczby zaklasyfikowanych obserwacji jedynie w przypadku zastosowania termu „niskie". Najwięcej obserwacji pozyskano w przypadku zliczania przez łączenie oraz wyostrzenia kontrastu (około od 86.70% do 86.73%, czyli o około od 25.19% do 25.24% więcej obserwacji niż w przypadku podejścia bazowego) z niskim ubytkiem obserwacji w warstwie testującej (od 3.48% do 3.52%). W przypadku najlepszych wariantów funkcji wagowych, których użycie zwiększyło liczbę pozyskanych obserwacji, uzyskano wysokie wartości metryki Accuracy (od 89.9% do 92.9%) i MCC (od 90.0% do 92.9%). Ze wszystkich badanych funkcji wagowych większą liczbę zaklasyfikowanych obserwacji niż w przypadku podejścia bazowego uzyskano dla 22 z 88 badanych wariantów funkcji wagowych (Tab. 4.5).

Tabela 4.5. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej ***FWCI 4-letnie*** (5 najlepszych wyników w termie)

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niskie											
Zliczanie przez łączenie *p*=0.5	-	92.9	93.3	92.9	99.8	93.1	92.9	7.1	3.48	86.73	125.24
Zliczanie przez łączenie *p*=1	-	91.7	92.2	91.7	99.7	91.9	91.7	8.3	3.49	86.72	125.22
Zliczanie przez łączenie *p*=2	-	89.9	90.7	89.9	99.6	90.3	90.0	10.1	3.52	86.70	125.19
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	90.8	91.3	90.8	99.7	91.0	90.9	9.2	4.78	85.14	122.93
Wyostrzenie kontrastu	-	90.5	91.1	90.5	99.7	90.8	90.6	9.5	3.49	86.72	125.22

Wykorzystanie zmiennej lingwistycznej ***FWCI 5-letnie*** w algorytmie w podejściu rozmytym spowodowało wystąpienie czterech wariantów funkcji wagowych dla termu „wysokie" nieznacznie poprawiając klasyfikację (od 1.08% do 1.15% więcej niż w metodzie bazowej) z wysoką redukcją obserwacji w zbiorze testującym (od 28.90% do 28.96%). Stosując w algorytmie term „niskie" klasyfikacja zwiększyła się do około 26.45% w porównaniu do metody bazowej z niską redukcją obserwacji w zbiorze testującym (od około 3.47% do 4.77%). W termie „niskie", podobnie jak w zmiennej lingwistycznej ***FWCI 4-letnie***, dla wszystkich najlepszych wariantów funkcji wagowych, których użycie przyczyniło się do zwiększenia liczby pozyskanych obserwacji, uzyskano wysokie wartości metryki Accuracy (od 89.2% do 92.9%) i MCC (od 89.3% do 92.9%). Ze wszystkich badanych funkcji wagowych większą liczbę zaklasyfikowanych obserwacji, niż w przypadku podejścia bazowego, uzyskano dla 26 z 88 badanych wariantów funkcji wagowych (Tab. 4.6).

Tabela 4.6. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej ***FWCI 5-letnie*** (5 najlepszych wyników w termie)

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niskie											
Zliczanie przez łączenie *p*=0.5	-	92.9	93.3	92.9	99.8	93.1	92.9	7.1	3.47	87.57	126.45
Zliczanie przez łączenie *p*=1	-	91.6	92.2	91.6	99.7	91.9	91.7	8.4	3.48	87.56	126.44
Zliczanie przez łączenie *p*=2	-	89.2	90.0	89.2	99.6	89.6	89.3	10.8	3.50	87.54	126.41
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	90.7	91.3	90.7	99.7	91.0	90.9	9.3	4.77	85.98	124.15
Wyostrzenie kontrastu	-	90.5	91.1	90.5	99.7	90.8	90.6	9.5	3.48	87.56	126.44
term = wysokie											
Zliczanie przez łączenie *p*=0.5	-	94.0	94.0	94.0	99.4	94.0	93.5	6.0	28.90	70.05	101.15
Zliczanie przez łączenie *p*=1	-	87.7	87.8	87.7	98.9	87.7	87.2	12.3	28.93	70.03	101.12
Zliczanie przez łączenie *p*=2	-	64.7	64.5	64.7	97.2	64.5	65.5	35.3	28.96	70.00	101.08
Wyostrzenie kontrastu	-	72.8	72.6	72.8	97.6	72.7	72.7	27.2	28.93	70.02	101.11

Stosując zmienną lingwistyczną ***FWCI bez ram czasowych*** uzyskano klasyfikację większej liczby obserwacji niż w podejściu bazowym dla wszystkich trzech termów. W przypadku zastosowania termu „niskie" dokonano klasyfikacji największej liczby obserwacji dla zliczania przez łączenie, wyostrzenia kontrastu (od około 90.24% do około 90.26%; od około 30.30% do 30.33% więcej obserwacji niż wykorzystując podejście bazowe) oraz zliczania przez progowanie i łączenie z p=0.5 i t=0,2 (około 88.73%,; czyli około 28.12% więcej obserwacji niż wykorzystując podejście bazowe). W przypadku termu „ani niskie ani wysokie" uzyskano od 7.27% do 10.42% więcej obserwacji niż w podejściu wykorzystującym podejście bazowe.

Największa zmiana w porównaniu do zmiennej lingwistycznej ***FWCI 5-letnie*** nastąpiła dla termu „wysokie", gdzie dla użycia funkcji wagowej zliczania przez łączenie oraz wyostrzenia kontrastu klasyfikacja wzrosła z około 70% (70.00%-70.05%) do ponad 79% (78.77%-78.83%, o 13.74% do 13.82% więcej niż w podjęciu bazowym). Stosując term „wysokie" dla zliczania przez wyostrzenie z p=0.5 i t=0.8 uzyskano klasyfikację około 78.05% obserwacji (około 12.70% więcej). Ze wszystkich badanych funkcji wagowych większą liczbę zaklasyfikowanych obserwacji niż w przypadku podejścia bazowego uzyskano dla 42 z 88 badanych wariantów funkcji wagowych (Tab. 4.7).

Tabela 4.7. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej ***FWCI bez ram czasowych*** (5 najlepszych wyników w termie)

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niskie											
Zliczanie przez łączenie *p*=0.5	-	95.8	96.0	95.8	99.8	95.9	95.7	4.2	3.51	90.26	130.33
Zliczanie przez łączenie *p*=1	-	90.8	91.4	90.8	99.7	91.1	90.9	9.2	3.52	90.25	130.32
Zliczanie przez łączenie *p*=2	-	86.9	87.5	86.9	99.6	87.2	87.2	13.1	3.53	90.24	130.30
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	87.5	88.0	87.5	99.7	87.7	87.8	12.5	4.76	88.73	128.12
Wyostrzenie kontrastu	-	87.2	87.8	87.2	99.6	87.5	87.5	12.8	3.52	90.25	130.32
term = ani niskie ani wysokie											
Zliczanie przez łączenie *p*=0.5	-	91.7	91.9	91.7	99.4	91.8	91.4	8.3	23.71	76.47	110.42
Zliczanie przez łączenie *p*=1	-	88.3	88.6	88.3	99.2	88.4	88.0	11.7	23.71	76.47	110.42
Zliczanie przez łączenie *p*=2	-	65.1	65.6	65.1	98.0	65.3	66.8	34.9	23.71	76.47	110.42
Zliczanie przez wyostrzenie *p*=0.5	0.4	91.4	91.5	91.4	99.3	91.4	91.0	8.6	24.06	74.29	107.27
Wyostrzenie kontrastu	-	65.1	65.6	65.1	98.0	65.3	66.8	34.9	23.71	76.47	110.42
term = wysokie											
Zliczanie przez łączenie *p*=0.5	-	92.1	92.1	92.1	99.4	92.1	91.7	7.9	20.06	78.83	113.82
Zliczanie przez łączenie *p*=1	-	84.9	85.0	84.9	98.9	84.9	84.6	15.1	20.09	78.80	113.78
Zliczanie przez łączenie *p*=2	-	69.7	69.6	69.7	97.7	69.6	70.2	30.3	20.12	78.77	113.74
Zliczanie przez wyostrzenie *p*=0.5	0.8	92.1	92.1	92.1	99.4	92.1	91.7	7.9	20.11	78.05	112.70
Wyostrzenie kontrastu	-	70.2	70.1	70.2	97.9	70.1	70.9	29.8	20.10	78.79	113.77

Zastosowanie zmiennej lingwistycznej ***Percentyl*** pozwoliło na uzyskanie największej liczby obserwacji w przypadku wartości niskiego percentyla. Największy wzrost zaklasyfikowanych obserwacji osiągnięto stosując funkcję wagową zliczania przez łącznie, zliczania przez wyostrzenie z p=0.5 i t=0.8 oraz wyostrzenia kontrastu (od 81.13% do 81.70%, czyli od 17.15% do 17.98% więcej niż w podejściu bazowym). Wysokie wartości metryki Accuracy i MCC otrzymano w przypadku użycia 3 z 5 najlepszych wariantów funkcji wagowych (od około 90.6% do około 92.9% Accuracy i od 90.2% do 92.7% MCC). Drugi term, którego użycie poprawiło klasyfikację, stanowił term „wysokie". Największy wzrost zaklasyfikowanych obserwacji osiągnięto w przypadku zastosowania zliczania przez łącznie oraz wyostrzenia kontrastu (od 75.11% do 75.78%, czyli od 9.38% do 9.43% więcej niż w podejściu bazowym).

W przypadku termu „ani niski ani wysoki" użycie zliczania przez łącznie oraz przez wyostrzenie kontrastu pozwoliło na klasyfikację około 4% więcej obserwacji niż stosując podejście bazowe (od 72.40% do 72.51%, czyli od 4,55% do 4.71% więcej). Ze wszystkich badanych funkcji wagowych większą liczbę zaklasyfikowanych obserwacji, niż w przypadku podejścia bazowego, uzyskano dla 30 z 88 badanych wariantów funkcji wagowych (Tab. 4.8).

Tabela 4.8. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej ***Percentyl*** (5 najlepszych wyników w termie)

Funkcja wagowa	t	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niski											
Zliczanie przez łączenie p=0.5	-	92.9	93.1	92.9	99.5	93.0	92.7	7.1	4.07	81.70	117.98
Zliczanie przez łączenie p=1	-	90.6	90.9	90.6	99.3	90.8	90.2	9.4	4.20	81.59	117.81
Zliczanie przez łączenie p=2	-	78.4	78.7	78.4	98.7	78.5	78.8	21.6	4.29	81.57	117.79
Zliczanie przez wyostrzenie p=0.5	0.8	92.9	93.1	92.9	99.5	93.0	92.7	7.1	4.08	81.13	117.15
Wyostrzenie kontrastu	-	78.7	78.9	78.7	98.8	78.8	79.1	21.3	4.27	81.59	117.82
term = ani niski ani wysoki											
Zliczanie przez łączenie p=0.5	-	84.4	84.9	84.4	99.4	84.6	84.7	15.6	17.68	72.51	104.71
Zliczanie przez łączenie p=1	-	77.2	77.9	77.2	99.1	77.5	78.1	22.8	17.80	72.40	104.55
Zliczanie przez łączenie p=2	-	72.1	73.1	72.1	98.7	72.6	73.5	27.9	17.86	72.42	104.58
Zliczanie przez wyostrzenie p=0.5	0.4	84.2	84.7	84.2	99.3	84.4	84.5	15.8	17.87	70.87	102.33
Wyostrzenie kontrastu	-	72.1	73.1	72.1	98.7	72.6	73.5	27.9	17.86	72.42	104.57
term = wysoki											
Zliczanie przez łączenie p=0.5	-	84.8	85.1	84.8	99.3	84.9	85.0	15.2	13.43	75.78	109.43
Zliczanie przez łączenie p=1	-	84.5	84.8	84.5	99.3	84.6	84.7	15.5	13.46	75.76	109.40
Zliczanie przez łączenie p=2	-	82.8	83.2	82.8	99.2	83.0	83.1	17.2	13.50	75.75	109.38
Zliczanie przez progowanie i łączenie p=2	0.2	86.2	86.6	86.2	99.2	86.4	86.2	13.8	17.80	72.11	104.13
Wyostrzenie kontrastu	-	84.3	84.6	84.3	99.2	84.4	84.5	15.7	13.47	75.77	109.41

W przypadku zmiennej lingwistycznej ***Rok*** algorytm w podejściu rozmytym zaklasyfikował najwięcej obserwacji dla termu „najnowszy". Największy procent zaklasyfikowanych obserwacji osiągnięto w przypadku zliczania przez łącznie, zliczania przez łączenie i progowanie z p=2 i t=0.2 oraz wyostrzenia kontrastu (od 91.77% do 91.83%, czyli od 32.52% do 32.60% więcej niż w podejściu bazowym). Klasyfikacji dokonano wraz z uzyskanymi wysokimi wartościami metryki Accuracy (od około 98.4% do około 100%) i MCC (od około 98.3% do około 100%) i wysoką liczbą obserwacji w zbiorze testowym (od około 0.00% do około 0.07% redukcji). Podobne wyniki w zakresie liczby pozyskanych obserwacji i metryk ewaluacyjnych dla zbiorów testujących uzyskano stosując term „najstarszy" i „ani najstarszy ani najnowszy". W przypadku użycia tych termów pozyskano pomiędzy 83.93% a 84.06% obserwacji (od 21.19% do 21.38% więcej niż w podejściu bazowym). Klasyfikacji dokonano wraz z Accuracy od 46.8% do 78.9% i MCC od 50.4% do 80.1% na zbiorach testowych stanowiących około 88% obserwacji zaklasyfikowanych metodą wykorzystującą podejście bazowe (od 12.05 do 12.08% redukcji). Ze wszystkich badanych funkcji wagowych większą liczbę zaklasyfikowanych obserwacji, niż w przypadku podejścia bazowego, uzyskano dla 54 z 88 badanych wariantów funkcji wagowych (Tab. 4.9).

Tabela 4.9. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej ***Rok*** (5 najlepszych wyników w termie)

Funkcja wagowa	t	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = najstarszy											
Zliczanie przez łączenie p=0.5	-	78.8	81.6	78.8	99.5	80.1	80.1	21.2	12.05	84.05	121.37
Zliczanie przez łączenie p=2	-	46.8	52.8	46.8	97.4	49.5	51.2	53.2	12.07	84.06	121.38
Zliczanie przez wyostrzenie p=0.5	0.8	78.8	81.6	78.8	99.5	80.1	80.1	21.2	12.05	84.04	121.35
Zliczanie przez wyostrzenie p=2	0.8	46.8	52.8	46.8	97.4	49.5	51.2	53.2	12.07	84.05	121.37
Wyostrzenie kontrastu	-	46.8	52.8	46.8	97.4	49.5	51.2	53.2	12.07	84.05	121.36
term = ani najstarszy ani najnowszy											
Zliczanie przez łączenie p=0.5	-	78.9	81.6	78.9	99.5	80.2	80.1	21.1	12.06	84.04	121.35
Zliczanie przez łączenie p=1	-	67.8	71.5	67.8	99.1	69.5	70.4	32.2	12.08	84.00	121.29
Zliczanie przez łączenie p=2	-	46.8	52.8	46.8	97.4	49.6	51.2	53.2	12.07	84.06	121.38
Zliczanie przez wyostrzenie p=0.5	0.4	45.8	52.2	45.8	97.4	48.8	50.4	54.2	12.08	83.93	121.19
Wyostrzenie kontrastu	-	46.8	52.8	46.8	97.4	49.6	51.2	53.2	12.07	84.07	121.39
term = najnowszy											
Zliczanie przez łączenie p=0.5	-	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.00	91.81	132.58
Zliczanie przez łączenie p=1	-	99.9	99.9	99.9	100.0	99.9	99.9	0.1	0.00	91.80	132.55
Zliczanie przez łączenie p=2	-	98.4	98.4	98.4	99.9	98.4	98.3	1.6	0.00	91.83	132.60
Zliczanie przez progowanie i łączenie p=2	0.2	98.4	98.4	98.4	99.9	98.4	98.3	1.6	0.07	91.77	132.52
Wyostrzenie kontrastu	-	99.9	99.9	99.9	100.0	99.9	99.9	0.1	0.00	91.82	132.58

Dla zmiennej lingwistycznej ***Zespół*** najlepsze wyniki uzyskano w przypadku wykorzystania w algorytmie termu „największy". Najwięcej zaklasyfikowanych obserwacji osiągnięto dla zliczania przez progowanie i łączenie z p=0.5 i t=0.2 (87.29%, o 26.04% więcej) oraz zliczanie przez łączenie z p=0.5 (87.28%, o 26.03% więcej). Klasyfikacji dokonano wraz z uzyskaniem wysokiej wartości metryki Accuracy i MCC (około 99.7%) i niezauważalną redukcją obserwacji w zbiorze testującym (do około 0.01%). Wysoką klasyfikację uzyskano również w przypadku pozostałych wariantów zliczania przez łączenie oraz wyostrzenie kontrastu (od 87.07% do 87.12%, o 25.73% do 25.81% więcej) z równie wysokimi wartościami metryki Accuracy (93.1% do 96.9%) i MCC (92.9% do 96.8%). W przypadku aplikacji termu „najmniejszy" najlepsze wyniki uzyskano dla zliczania przez łączenie z p=0.5 i p=2, zliczania przez wyostrzenie z p=0.5 i p=2 dla t=0.8 oraz wyostrzenia kontrastu (od 77.05% do 77.26%, czyli od 11.26% do 11.56% więcej).

W przypadku termu „ani najmniejszy ani największy" najlepsze wyniki uzyskano dla zliczania przez łączenie, zliczania przez wyostrzenie z p=0.5 dla t=0.4 oraz wyostrzenia kontrastu (od 73.47% do 77.23%, czyli od 6.10% do 11.53% więcej). Stosując term „najmniejszy" oraz „ani najmniejszy ani największy" wraz z opisywanymi wariantami funkcji wagowych pojawiały się niższe wartości metryki Accuracy i MCC (od 51.1% do 84.8% Accuracy i 54.0% do 85.0% MCC dla termu „najmniejszy"; od 58.7% do 84.9% Accuracy i 60.6% do 85.2% MCC dla termu „ani najmniejszy ani największy"). Ze wszystkich badanych funkcji wagowych większą liczbę zaklasyfikowanych obserwacji niż w przypadku podejścia bazowego uzyskano dla 44 z 88 badanych wariantów funkcji wagowych (Tab. 4.10).

Tabela 4.10. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej ***Zespół*** (5 najlepszych wyników w termie)

Funkcja wagowa	t	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = najmniejszy											
Zliczanie przez łączenie p=0.5	-	84.8	85.8	84.8	99.3	85.2	85.0	15.2	14.32	77.26	111.56
Zliczanie przez łączenie p=2	-	51.7	54.1	51.7	96.8	52.8	54.4	48.3	14.53	77.06	111.28
Zliczanie przez wyostrzenie p=0.5	0.8	84.8	85.8	84.8	99.3	85.2	85.0	15.2	14.32	77.25	111.55
Zliczanie przez wyostrzenie p=2	0.8	51.6	54.1	51.6	96.8	52.8	54.3	48.4	14.53	77.05	111.26
Wyostrzenie kontrastu	-	51.1	53.5	51.1	96.9	52.3	54.0	48.9	14.52	77.06	111.28
term = ani najmniejszy ani największy											
Zliczanie przez łączenie p=0.5	-	84.9	86.0	84.9	99.4	85.4	85.2	15.1	14.35	77.23	111.53
Zliczanie przez łączenie p=1	-	84.3	85.5	84.3	99.3	84.8	84.5	15.7	14.58	76.96	111.13
Zliczanie przez łączenie p=2	-	58.7	61.2	58.7	97.2	59.9	60.6	41.3	14.54	77.05	111.26
Zliczanie przez wyostrzenie p=0.5	0.4	84.4	85.4	84.4	99.3	84.8	84.6	15.6	14.70	73.47	106.10
Wyostrzenie kontrastu	-	58.7	61.2	58.7	97.2	59.9	60.6	41.3	14.54	77.05	111.26

term = największy											
Zliczanie przez łączenie p=0.5	-	99.7	99.7	99.7	100.0	99.7	99.7	0.3	0.00	87.28	126.03
Zliczanie przez łączenie p=1	-	96.9	96.9	96.9	99.9	96.9	96.8	3.1	0.14	87.07	125.73
Zliczanie przez łączenie p=2	-	93.1	93.1	93.1	99.6	93.1	92.9	6.9	0.21	87.07	125.73
Zliczanie przez progowanie i łączenie p=0.5	0.2	99.7	99.7	99.7	100.0	99.7	99.7	0.3	0.01	87.29	126.04
Wyostrzenie kontrastu	-	96.4	96.5	96.4	99.8	96.5	96.3	3.6	0.10	87.12	125.81

4.3.3 Dyskusja

Zastosowanie zmiennych lingwistycznych w algorytmie klasyfikacji dominującej dyscypliny pozwoliło zwiększyć procent zaklasyfikowanych obserwacji w stosunku do podejścia bazowego we wszystkich siedmiu zaproponowanych zmiennych. W przypadku wszystkich skonstruowanych zmiennych algorytm zwiększył klasyfikację w stosunku do podejścia bazowego posługując się termem niskiej przynależności do zbioru rozmytego. Dla propozycji wykorzystania termu ani niskiej ani wysokiej przynależności do zbioru rozmytego uzyskano klasyfikację dla sześciu proponowanych zmiennych (wszystkie zmienne z wyjątkiem ***FWCI 4-letniego***). W przypadku termu wysokiej przynależności do zbioru rozmytego klasyfikację dokonano wykorzystując 5 zmiennych lingwistycznych (wszystkie proponowane z wyjątkiem ***FWCI 4-letniego*** i ***FWCI 5-letniego***).

Najwyższej klasyfikacji dokonano posługując się termem najnowszego roku publikacji (zmienna lingwistyczna ***Rok***). W przypadku tego podejścia zaklasyfikowano powyżej 91% obserwacji. Stosując funkcje wagowe zliczania przez łączenie, łączenie i progowanie z p=2 i t=0.8 oraz wyostrzenie kontrastu zaklasyfikowano od 91.77% do 91.83% obserwacji (od 32.52% do 32.60% więcej niż w podejściu bazowym). Wykorzystując najlepszy wariant funkcji wagowej osiągnięto wysokie wartości metryki Accuracy i MCC (98.4% Accuracy i 98.3% MCC). Była to jedyna zmienna w termie wysokiej przynależności, której najlepsze warianty funkcji wagowych dokonały klasyfikacji powyżej 90% obserwacji.

W przypadku wykorzystania zmiennych w termie ani niskiej ani wysokiej przynależności również najlepszy wynik uzyskano w przypadku zmiennej lingwistycznej ***Rok***. Klasyfikacji dokonano dla około 84% obserwacji. Najlepsze warianty funkcji wagowych w tym podejściu zwracały jednak różne wartości metryki Accuracy i MCC. W przypadku zliczania przez łączenie z p=0.5 osiągnięto 78.9% Accuracy i 80.1% MCC. Dla zliczania przez łącznie z p=2, zliczanie przez wyostrzenie z p=0.5 i t=0.4 oraz wyostrzenie kontrastu osiągnięto około 46% Accuracy

i 50% MCC.). Była to jedyna zmienna w termie ani niskiej ani wysokiej przynależności, której najlepsze warianty funkcji wagowych dokonały klasyfikacji powyżej 80% obserwacji.

Posługując się termem niskiej przynależności do zbioru rozmytego największej klasyfikacji dokonano z wykorzystaniem zmiennej lingwistycznej ***Cytowanie*** oraz ***FWCI bez ram czasowych***. Dla zmiennej lingwistycznej ***Cytowanie*** zaklasyfikowano do 90.67% obserwacji (stosując zliczanie przez łącznie z p=0.5 i wyostrzenie kontrastu) z wysoką wartością metryki Accuracy (odpowiednio 92.0% i 90.6%) i MCC (92.1% i 90.7%). Dla zmiennej lingwistycznej ***FWCI bez ram czasowych*** zaklasyfikowano do 90.26% obserwacji (stosując zliczanie przez łącznie z p=0.5) z wyższą wartością metryki Accuracy niż w przypadku zmiennej lingwistycznej ***Cytowanie*** (95.8%) i MCC (95.7%).

W przypadku wskaźników ***FWCI*** w termie wysokiej przynależności do zbioru rozmytego można zauważyć, że wraz ze wzrostem ramy czasowej zwiększa się procent klasyfikacji w algorytmie stosującym podejście rozmyte. W przypadku ***FWCI 4-letniego*** nie występują funkcje wagowe, które dokonały klasyfikacji większej liczby obserwacji niż stosując podejście bazowe. Biorąc ramę czasową o rok dłuższą możliwe było uzyskanie nieznacznego wzrostu odsetka zaklasyfikowanych obserwacji (do 1.15% więcej niż w podejściu bazowym). Stosując brak ram czasowych możliwe było wykonanie największej klasyfikacji (do 13.82% więcej niż w podjęciu bazowym).

Różnice w wartościach metryki Accuracy i MCC występowały dla algorytmu w podejściu rozmytym zarówno pomiędzy zmiennymi lingwistycznymi jak i pomiędzy termami w ramach jednej zmiennej lingwistycznej. W przypadku zmiennej ***Rok*** oraz ***Zespół*** uzyskano wyższe wartości metryki Accuracy i MCC dla termu wysokiej przynależności do zbioru rozmytego niż termu niskiej przynależności do zbioru rozmytego. Dla zmiennej ***Cytowanie***, trzech wariantów zmiennej ***FWCI*** i zmiennej ***Percentyl*** wyższe wartości metryki Accuracy i MCC uzyskano dla termu niskiej przynależności do zbioru rozmytego niż dla termu wysokiej przynależności. Największe różnice pomiędzy termami niskiej przynależności do zbioru i wysokiej przynależności do zbioru w ramach jednej zmiennej występowały w przypadku zliczania przez łącznie p=2 i wyostrzenia kontrastu dla ***Cytowania*** (różnica pomiędzy około 50% a 90% Accuracy i MCC), ***Roku*** (pomiędzy 47% a 98% Accuracy i 51% a 95% MCC) oraz ***Zespołu*** (pomiędzy 51% a 93% lub więcej dla Accuracy i 54% a 93% lub więcej dla MCC).

Grupa zmiennych, do której wyznaczenia wykorzystuje się wartość liczności cytowania (***Cytowanie***, wskaźniki ***FWCI***, ***Percentyl***) charakteryzowała się najwyższą liczbą zaklasyfikowanych obserwacji dla termu niskiej przynależności do zbioru rozmytego. Z wykresu gęstości jądrowej można zauważyć, że w przypadku tej grupy zmiennych największy rozkład obserwacji występuje w przedziale 0.8 do 1 dla termu niskiej przynależności (odpowiednio od 0 do 0.2 dla termu wysokiej przynależności). Wyjątek stanowi zmienna ***Percentyl***, która dodatkowo przyjmuje wysoki rozkład dla wartości 0 (odpowiednio dodatkowo dla wartości 1 dla termu wysokiej przynależności). Odwrotna sytuacja występuje w przypadku zmiennej ***Rok*** i ***Zespół***. Dla zmiennej najnowszego roku największy rozkład obserwacji występuje w przedziale od 0.9 do 1 (odpowiednio od 0 do 0.1 dla termu niskiej przynależności do zbioru). Dla zmiennej ***Zespół*** uzyskano największy rozkład w przedziale od 0.4 do 0.6.

Rozkład obserwacji w termach wyznacza również użyteczność funkcji wagowych w celu uzyskania najwyżej klasyfikacji. W przypadku zmiennej ***Cytowanie*** i trzech zmiennych ze wskaźnikami ***FWCI*** w termie niskiej przynależności do zbioru rozmytego najlepsze wyniki algorytm w podejściu rozmytym uzyskał stosując zliczanie przez łączenie, wyostrzenie kontrastu oraz progowanie i łączenie. Dla zmiennej ***Percentyl***, ***Rok*** i ***Zespół*** te funkcje wagowe dominowały w termie wysokiej przynależności do zbioru rozmytego. W przypadku zmiennej ***Cytowanie*** i trzech zmiennych ze wskaźnikami ***FWCI*** w termie wysokiej przynależności do zbioru rozmytego najlepsze wyniki uzyskano stosując zliczanie przez łączenie, wyostrzenie kontrastu oraz przez wyostrzenie. Dla zmiennej ***Percentyl***, ***Rok*** i ***Zespół*** te funkcje wagowe dominowały w termie niskiej przynależności do zbioru rozmytego. W sytuacji, gdy największa gęstość obserwacji występowała w przedziale od 0 do 0.2, stosowanie zliczania przez progowanie i łączenie zwracało wyższą klasyfikację niż w przypadku stosowania zliczania przez wyostrzenie. W przeciwnej sytuacji stosowanie zliczania przez wyostrzenie zwracało wyższą klasyfikację niż stosując zliczanie przez progowanie i łącznie.

Podsumowując klasyfikacja dominującej dyscypliny w podejściu rozmytym przebiegła pomyślnie dla wielu wariantów zmiennych lingwistycznych. Stosując najlepszą zmienną - ***Rok*** możliwe było zaklasyfikowanie od 91.77% do 91.83% obserwacji (od 32.52% do 32.60% więcej niż w podejściu bazowym). Dla tak zmodyfikowanego algorytmu klasyfikację osiągano z wysokim podobieństwem do klasyfikacji algorytmu w podejściu bazowym (uzyskując nawet do 98.4% wartości metryki Accuracy i 98.3% MCC).

Rozdział 5.
Klasyfikacja dominującej dyscypliny z wykorzystaniem sterowników rozmytych prestiżu publikacji

5.1 Wprowadzenie

W tym rozdziale opisano proces badania skuteczności klasyfikacji dominującej dyscypliny z wykorzystaniem sterowników rozmytych prestiżu publikacji. W części metodologicznej przedstawiono proces budowy sterowników z wybranymi zmiennymi wejściowymi. Ostatnia część metodologii stanowi opis modyfikacji algorytmu wyznaczania dominującej dyscypliny z uwzględnieniem sterowników rozmytych prestiżu publikacji (podejście rozmyte) oraz prezentuje listę wybranych do badania wariantów funkcji wagowych wykorzystanych na etapie wyznaczania mocy zbiorów rozmytych w zmodyfikowanej wersji algorytmu. Kolejny podrozdział prezentuje wyniki z przeprowadzonych badań. Część tę opisano w ten sam sposób jak w przypadku rozdziału 4. (klasyfikacja dominującej dyscypliny z wykorzystaniem zmiennych lingwistycznych). Wyniki badań zostały podzielone na dwie części. W pierwszej części opisano rozkłady gęstości obserwacji stanowiących dziedzinę badanych funkcji wagowych. Druga część wyników stanowi opis rezultatów uzyskanych dla algorytmu klasyfikacji dominującej dyscypliny w podejściu rozmytym. Dla algorytmu w podejściu rozmytym, wykorzystującego każdy wariant zaproponowanego sterownika rozmytego, opisano 5 wariantów funkcji wagowych, dla których dokonano najwyższej klasyfikacji oraz wykonano ewaluację wyników uzyskanych w podejściu rozmytym z wynikami z podejścia bazowego. Ostatnia część rozdziału stanowi dyskusję nad wynikami badania. *(W tym rozdziale nazwy zmiennych lingwistycznych oraz sterowników rozmytych zapisano pogrubioną kursywą w celu ich wyszczególnienia.)*

5.2 Metodologia

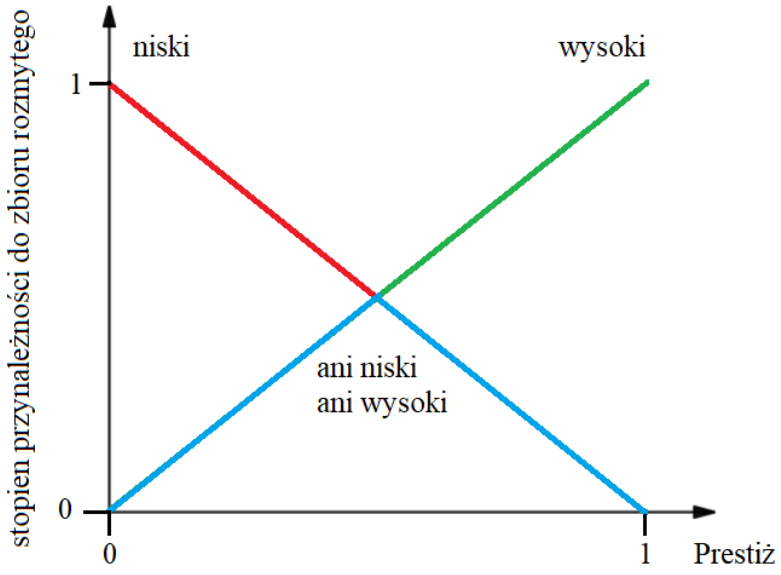
5.2.1 Budowa kontrolerów rozmytych

Na potrzeby badania skonstruowano 4 sterowniki rozmyte prestiżu publikacji. Każdy sterownik rozmyty przyjmował 2 zmienne wejściowe. Pierwszą zmienną wejściową był jeden wariant metryki cytowania (***Cytowanie***, ***FWCI 4-letnie***, ***FWCI 5-letnie***, ***FWCI bez ram czasowych***).

Drugą zmienną wejściową stanowił ***Percentyl*** czasopisma publikacji. Sterowniki zwracały wartość zbioru rozmytego wysokiego prestiżu publikacji. Skonstruowano 4 sterowniki rozmyte (sterownik rozmyty prestiżu w oparciu o ***Cytowanie i Percentyl***, sterownik rozmyty prestiżu w oparciu o ***FWCI 4-letnie i Percentyl***, sterownik rozmyty prestiżu w oparciu o ***FWCI 5-letnie i Percentyl*** oraz sterownik rozmyty prestiżu w oparciu o ***FWCI bez ram czasowych i Percentyl***).

Sterowniki rozmyte na wejściu przyjmowały dwie wartości (wartość jednej wybranej metryki cytowania oraz percentyl publikacji). Blok rozmywania wyznaczał stopnie przynależności wejść kontrolera dla odpowiednich zbiorów rozmytych. Na etapie fuzyfikacji zbiory rozmyte tworzono wykorzystując wcześniej skonstruowane odpowiednie zmienne lingwistyczne (zaprojektowane na potrzeby algorytmu klasyfikacji dominującej dyscypliny z wykorzystaniem zmiennych lingwistycznych; rozdział 4, podrozdział 2.1 oraz 2.2).

Zmienna wyjściowa prestiżu publikacji obejmowała uniwersum od 0 do 1 z trzema termami: niskiego prestiżu, ani niskiego ani wysokiego prestiżu i wysokiego prestiżu. Dla każdego termu zamodelowano odpowiednią funkcję przynależności. W przypadku termu wysokiego prestiżu była to rosnąca funkcja liniowa w przedziale od 0 do 1. Term niskiego prestiżu stanowiła negacja zbioru wartości dla termu wysokiego prestiżu (malejąca funkcja liniowa w przedziale od 0 do 1). Term ani niskiego ani wysokiego prestiżu stanowił wartość t-normy minimum z wartości funkcji przynależności dla termu niskiego prestiżu i wysokiego prestiżu. Term ani niskiego ani wysokiego prestiżu przyjmował zbiór wartości z przedziału od 0 do 0.5 włącznie (Rys 5.1).



Rysunek 5.1. Interpretacja zmiennej lingwistycznej prestiżu publikacji z termami niskiego, wysokiego oraz ani niskiego ani wysokiego prestiżu

Do przeprowadzenia operacji w bloku wnioskowania posłużono się zbiorem 9 reguł w postaci rozmytych instrukcji warunkowych (z operatorem Mamdaniego dla wyjściowego zbioru rozmytego). W przypadku gdy jedna zmienna przyjmowała niską przynależność, a druga zmienna osiągała wartość niskiej lub ani niskiej ani wysokiej przynależności do zbioru rozmytego, zmienna wyjściowa osiągała wartość niskiego prestiżu. Dla sytuacji, w której jedna zmienna przyjmowała niską przynależność, a druga zmienna osiągała wysoką przynależność lub obie zmienne osiągały ani niską ani wysoką przynależność do zbioru rozmytego, zmienna wyjściowa przyjmowała ani niski ani wysoki prestiż. W przypadku gdy jedna zmienna osiągała wysoką przynależność, a druga zmienna osiągała wysoką lub ani niską ani wysoką przynależność do zbioru rozmytego zmienna wyjściowa przyjmowała wysoki prestiż. Dla sterownika rozmytego, w którym pierwszą zmienną wejściową była metryka cytowania (***Cytowanie***, ***FWCI 4-letnie***, ***FWCI 5-letnie*** lub ***FWCI bez ram czasowych***), a drugą zmienną wejściową był ***Percentyl***, lista reguł wyglądała następująco:

IF *Metryka*$_{Cytowania}$ IS *niska* AND *Percentyl* IS *niski* THEN *Prestiż* IS *niski*
IF *Metryka*$_{Cytowania}$ IS *niska* AND *Percentyl* IS *ani niski ani wysoki* THEN *Prestiż* IS *niski*
IF *Metryka*$_{Cytowania}$ IS *ani niska ani wysoka* AND *Percentyl* IS *niski* THEN *Prestiż* IS *niski*
IF *Metryka*$_{Cytowania}$ IS *niska* AND *Percentyl* IS *wysoki* THEN *Prestiż* IS *ani niski ani wysoki*
IF *Metryka*$_{Cytowania}$ IS *ani niska ani wysoka* AND *Percentyl* IS *ani niski ani wysoki*
THEN *Prestiż* IS *ani niski ani wysoki*
IF *Metryka*$_{Cytowania}$ IS *wysoka* AND *Percentyl* IS *niski* THEN *Prestiż* IS *ani niski ani wysoki*
IF Metryka$_{Cytowania}$ IS *ani niska ani wysoka* AND *Percentyl* IS *wysoki* THEN *Prestiż* IS *wysoki*
IF *Metryka*$_{Cytowania}$ IS *wysoka* AND *Percentyl* IS *ani niski ani wysoki* THEN *Prestiż* IS *wysoki*
IF *Metryka*$_{Cytowania}$ IS *wysoka* AND *Percentyl* IS *wysoki* THEN *Prestiż* IS *wysoki*

Zbiór tych reguł przedstawiono również w postaci tabelarycznej (Tab. 5.1). Następnie zbiory rozmyte pochodzące z 9 reguł zostały zagregowane do jednego zbioru rozmytego. Do wykonania agregacji wykorzystano operację sumy zbiorów rozmytych $\cup_s$ z t-konormą maksimum.

Tabela 5.1. Tabela reguł dla zmiennej wyjściowej prestiżu publikacji

Metryka$_{Cytowania}$ → Percentyl ↓	niska	ani niska ani wysoka	wysoka
niski	niski	niski	ani niski ani wysoki
ani niski ani wysoki	niski	ani niski ani wysoki	wysoki
wysoki	ani niski ani wysoki	wysoki	wysoki

Ostatnim etapem w budowie sterowników rozmytych prestiżu publikacji było wybranie metody defuzyfikacji w bloku wyostrzania. Ze względu na prowadzenie badań w środowisku Databricks oraz operowanie na dużej ilości danych w podejściu rozproszonym wybrano metodę, której implementacja była najmniej kosztowna pod względem obliczeniowym. W celu wyznaczenia wartości wyjściowej prestiżu publikacji użyto metody średniej z maksimów.

5.2.2 Algorytm wyznaczania dominującej dyscypliny z wykorzystaniem sterownika rozmytego prestiżu publikacji

Algorytm wyznaczania dominującej dyscypliny w podejściu bazowym zmodyfikowano tak, aby wykorzystywał wybrany sterownik rozmyty prestiżu publikacji i wyznaczenia mocy zbiorów rozmytych z zastosowaniem badanych funkcji wagowych (wariant algorytmu nazywany podejściem rozmytym). Na początku do kolumn zawierających unikalny identyfikator publikacji, listę identyfikatorów autorów oraz listę dyscyplin przypisanych do czasopisma, z którego pochodziła publikacja, dołączono dwie kolumny reprezentujące zmienne wejściowe. Następnie, podobnie jak w podejściu bazowym, dokonano normalizacji tabeli do pierwszej postaci normalnej ze względu na listę identyfikatorów autorów i listę dyscyplin. Kolejny krok obejmował wyznaczenie i dołączenie dla każdego rekordu wartości wyjściowej sterownika rozmytego prestiżu publikacji. Po wyznaczeniu wartości wyjściowej sterownika przystąpiono do wyznaczania wartości dla badanej funkcji wagowej oraz usunięto rekordy, dla których wartość tej funkcji wynosiła zero (pozostawiając rekordy, w których stopień przynależności do zbioru rozmytego stanowił jego nośnik). Następnie dokonano wyznaczenia mocy zbiorów rozmytych dla każdej dyscypliny autora.

Kolejne części algorytmu przebiegały podobnie do wariantu w podejściu bazowym, z tym że zamiast wyznaczać maksymalną wartość z liczby publikacji w dyscyplinie autora, wyznaczono maksymalną wartość ze wszystkich wartości mocy zbiorów rozmytych dyscyplin autora. Z tabeli wybrano tylko te rekordy, dla których moc zbioru rozmytego dla dyscypliny autora była równa wartości maksymalnej mocy zbioru dla dyscypliny autora. Następnie zbiór autorów ograniczono jedynie do tych, którzy posiadali jedną dyscyplinę, dla której osiągnięto maksymalną wartość mocy zbioru rozmytego wysokiego prestiżu publikacji.

Algorytm 3. Algorytm klasyfikacji dominującej dyscypliny z wykorzystaniem sterownika rozmytego prestiżu publikacji

Wejście: zbiór publikacji w bazie Scopus,
Parametry: pierwsza zmienna wejściowa, druga zmienna wejściowa, funkcja wagowa
Wyjście: zbiór par autor-dyscyplina

1. Dołącz do tabeli kolumnę z wartościami dla pierwszej i drugiej zmiennej wejściowej
2. Wykonaj normalizację tabeli do pierwszej postaci normalnej ze względu na identyfikatory autorów i listę dyscyplin
3. Wyznacz wartość sterownika rozmytego prestiżu publikacji
4. Oblicz wartość wybranej funkcji wagowej dla wartości wyznaczonej w kroku 3.
5. Wyznacz nośnik zbioru rozmytego wysokiego prestiżu publikacji
6. Dla każdego autora i jego dyscyplin oblicz moc zbioru rozmytego
7. Dla każdego autora wyznacz maksymalną z wartości mocy zbiorów rozmytych
8. Dla każdego autora wybierz dyscypliny, dla których moc zbioru rozmytego odpowiada maksymalnej z wartości mocy zbiorów rozmytych
9. Dokonaj filtracji tabeli do autorów posiadających jedną zaklasyfikowaną dyscyplinę

5.2.3 Wybór badanych funkcji wagowych

Do badania klasyfikacji algorytmu w podejściu rozmytym wybrano 5 funkcji wagowych: zliczanie przez progowanie, zliczanie przez łączenie, zliczanie przez progowanie i łączenie, zliczanie przez wzmocnienie oraz wzmocnienie kontrastu. W przypadku zliczania przez łączenie, progowanie i łączenie oraz wzmocnienie wybrano 3 warianty wartości p (0.5, 1 i 2). Dla zliczania przez progowanie, progowanie i łączenie i przez wzmocnienie wybrano 4 warianty wartości t (0.2, 0.4, 0.6, 0.8). Łącznie dla algorytmu wyznaczania dominującej dyscypliny w podejściu rozmytym uzyskano 34 badane sposoby w każdym sterowniku rozmytym na wykonanie klasyfikacji.

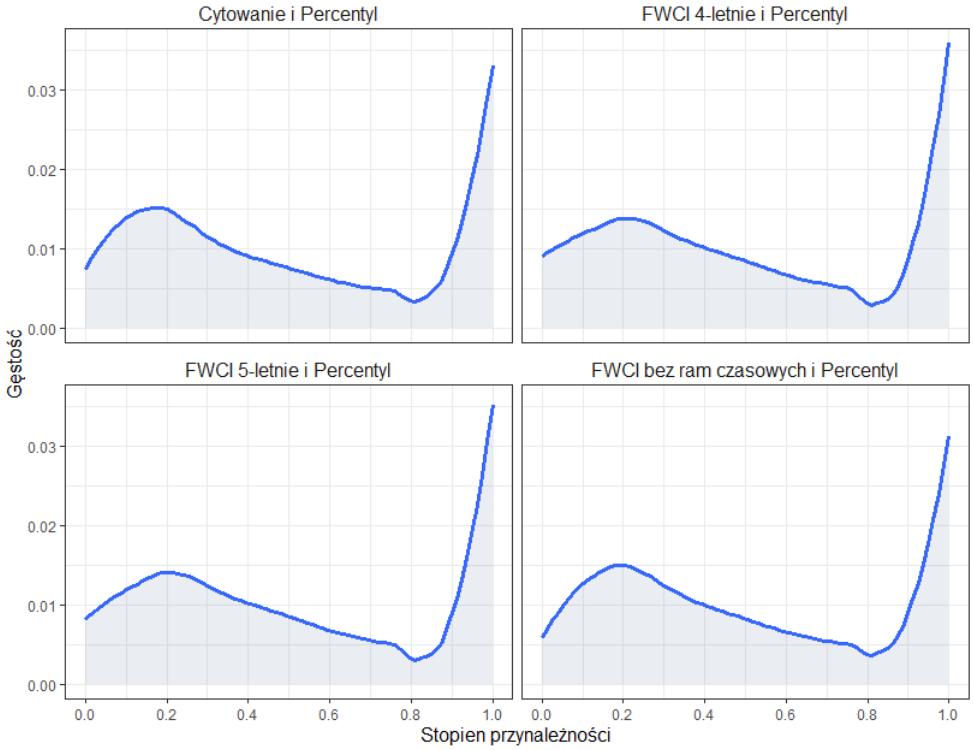
5.3 Wyniki

5.3.1 Badanie rozkładu gęstości dziedziny funkcji wagowych

W pierwszym etapie przeanalizowano rozkłady wartości dziedziny funkcji wagowych pozyskane jako wartości wyjściowe skonstruowanych sterowników rozmytych. W przypadku wszystkich czterech wariantów kontrolerów rozmytych uzyskano podobny kształt rozkładu gęstości jądrowej wartości dziedziny funkcji wagowych. Największy rozkład obserwacji występuje w przedziale od 0 do 0.4 oraz w punkcie 1. We wszystkich wariantach w przedziale od 0.2 do 0.8. wraz

ze wzrostem wartości stopnia przynależności do zbioru prestiżowych publikacji maleje gęstość występowania obserwacji.

Dla dwóch wariantów sterownika rozmytego prestiżu publikacji (przyjmującego jako zmienne wejściowe ***FWCI 4-letnie i Percentyl*** oraz przyjmującego jako zmienne wejściowe ***FWCI 5-letnie i Percentyl***) w kontekście wyznaczania nośnika zbioru rozmytego występuje większa redukcja obserwacji niż w przypadku pozostałych dwóch wariantów sterowników rozmytych (przyjmującego jako zmienne wejściowe ***Cytowanie i Percentyl*** oraz przyjmującego jako zmienne wejściowe ***FWCI bez ram czasowych i Percentyl***) (Rys 4.2).



Rysunek 4.2. Wykres gęstości jądrowej dla dziedziny funkcji wagowych w ujęciu sterowników rozmytych

5.3.2 Ewaluacja algorytmu klasyfikacji dominującej dyscypliny w podejściu rozmytym

Poniższe wyniki przedstawiają ewaluacje klasyfikacji uzyskanej przez algorytm klasyfikacji dominującej dyscypliny z wykorzystaniem sterowników rozmytych. Analizie poddano 5 wariantów funkcji wagowych, których wykorzystanie w algorytmie dokonało najwyższej klasyfikacji spośród badanych grup funkcji wagowych (zliczania przez progowanie, zliczania przez łączenie, zliczania przez progowanie i łączenie, zliczania przez wyostrzenie oraz wyostrzenia

kontrastu). Pełne wyniki zawierające listę wszystkich wariantów, które dokonały wyższej klasyfikacji niż w podejściu bazowym umieszczono w załączniku w sekcji 2. Dla każdego wariantu funkcji wagowej wyznaczono z algorytmu procent zaklasyfikowanych autorów oraz procentowy wzrost liczby zaklasyfikowanych obserwacji w stosunku do liczby zaklasyfikowanych obserwacji w podejściu bazowym. W celu zbadania podobieństwa pomiędzy klasyfikacją dokonaną metodą bazową, a metodą rozmytą stosującą sterowniki rozmyte prestiżu publikacji, wyznaczono metryki klasyfikacyjne dla zbioru testującego. Obejmował on obserwacje zaklasyfikowane zarówno podejściem bazowym i rozmytym. W celu zbadania redukcji klasyfikacji uzyskanej podejściem rozmytym, obliczono procent obserwacji, których nie zaklasyfikowano metodą rozmytą a zaklasyfikowano metodą bazową. Łącznie badano 34 warianty funkcji wagowych dla każdego sterownika rozmytego prestiżu publikacji.

W przypadku każdego opisywanego sterownika rozmytego największą klasyfikację wykonały te same warianty funkcji wagowych z tą samą kolejnością ich występowania (odpowiednio zliczania przez łączenie z p=0.5, zliczania przez łączenie z p=1, wyostrzenia kontrastu, zliczania przez łączenie z p=2 oraz zliczania przez wyostrzenie z p=0.5 i t=0.8).

Stosując sterownik rozmyty prestiżu publikacji ze zmienną wejściową ***Cytowanie i Percentyl*** w przypadku pięciu najlepszych wariantów funkcji wagowych dokonano klasyfikacji od 92.63% do 95% obserwacji (od 33.75% do 37.17% więcej niż stosując podejście bazowe). Wykorzystując 4 z pięciu badanych wariantów funkcji wagowych uzyskano klasyfikację z wartościami metryki Accuracy od 91.2% do 98.9% i MCC od 90.4% do 98.6%. Dla jednego badanego wariantu z Accuracy 84.3% i MCC 83.8%. Klasyfikacji dokonano wraz z niską redukcją obserwacji w zbiorze testującym (od 0.56% do 0.63%). Ze wszystkich badanych funkcji wagowych więcej zaklasyfikowanych obserwacji, niż w przypadku podejścia bazowego, uzyskano w przypadku zastosowania 19 z 34 wybranych wariantów funkcji wagowych (Tab. 4.2).

Tabela 4.2. Wyniki metryk ewaluacyjnych dla sterownika rozmytego prestiżu publikacji ze zmienną wejściową ***Cytowanie i Percentyl*** (5 najlepszych wyników)

Funkcja wagowa	t	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
Zliczanie przez łączenie p=0.5	-	98.9	98.9	98.9	99.7	98.9	98.6	1.1	0.56	95.00	137.18
Zliczanie przez łączenie p=1	-	93.0	92.9	93.0	99.3	92.9	92.5	7.0	0.59	94.98	137.15
Zliczanie przez łączenie p=2	-	84.3	84.1	84.3	98.6	84.2	83.8	15.7	0.63	94.94	137.10
Zliczanie przez wyostrzenie p=0.5	0.8	98.9	98.9	98.9	99.7	98.8	98.6	1.1	0.59	92.63	133.75
Wyostrzenie kontrastu	-	91.2	91.1	91.2	98.9	91.1	90.4	8.8	0.60	94.97	137.13

Dla zastosowania sterownika rozmytego prestiżu publikacji ze zmienną wejściową ***FWCI 4-letnie i Percentyl*** w przypadku najlepszych wariantów funkcji wagowych dokonano klasyfikacji od 91% do 92.84% obserwacji (od 31.40% do 34.06% więcej). Wykorzystując najlepsze badane warianty funkcji wagowych uzyskano klasyfikację z wysoką wartością metryki Accuracy i MCC (od 92.7% do 99.2% dla Accuracy i od 91.9% do 99% MCC). Klasyfikację wykonano wraz z niską redukcją obserwacji w zbiorze testującym (od 1.12% do 1.20%). Ze wszystkich badanych funkcji wagowych więcej zaklasyfikowanych obserwacji, niż w przypadku podejścia bazowego, uzyskano w przypadku zastosowania 19 z 34 wybranych wariantów funkcji wagowych (Tab. 4.3).

Tabela 4.3. Wyniki metryk ewaluacyjnych dla sterownika rozmytego prestiżu publikacji ze zmienną wejściową ***FWCI 4-letnie i Percentyl*** (5 najlepszych wyników)

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
Zliczanie przez łączenie *p*=0.5	-	99.2	99.2	99.2	99.8	99.2	99.0	0.8	1.12	92.84	134.06
Zliczanie przez łączenie *p*=1	-	97.9	97.9	97.9	99.6	97.9	97.4	2.1	1.16	92.80	134.01
Zliczanie przez łączenie *p*=2	-	92.7	92.5	92.7	99.1	92.6	91.9	7.3	1.20	92.78	133.97
Zliczanie przez wyostrzenie *p*=0.5	0.8	99.1	99.1	99.1	99.8	99.1	99.0	0.9	1.15	91.00	131.40
Wyostrzenie kontrastu	0	94.8	94.8	94.8	99.3	94.8	94.2	5.2	1.17	92.79	133.99

Zastosowanie sterownika rozmytego prestiżu publikacji ze zmienną wejściową ***FWCI 5-letnie i Percentyl*** w przypadku najlepszych wariantów funkcji wagowych pozwoliło na zaklasyfikowanie od 91.37% do 93.31% obserwacji (od 31.93% do 34.74% więcej). Poprzez użycie najlepszych wariantów funkcji wagowych klasyfikacji dokonano z wysokim Accuracy i MCC (od 92.7% do 99.2% dla Accuracy i od 92% do 99% MCC). Klasyfikację uzyskano wraz z niskim ubytkiem obserwacji w zbiorze testującym (od 1.06% do 1.14%). Ze wszystkich badanych funkcji wagowych więcej zaklasyfikowanych obserwacji, niż w przypadku podejścia bazowego, uzyskano w przypadku zastosowania 19 z 34 wybranych wariantów funkcji wagowych (Tab. 4.4).

Tabela 4.4. Wyniki metryk ewaluacyjnych dla sterownika rozmytego prestiżu publikacji ze zmienną wejściową ***FWCI 5-letnie i Percentyl*** (5 najlepszych wyników w termie)

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
Zliczanie przez łączenie *p*=0.5	-	99.2	99.2	99.2	99.8	99.2	99.0	0.8	1.06	93.31	134.74
Zliczanie przez łączenie *p*=1	-	97.9	97.9	97.9	99.6	97.9	97.5	2.1	1.09	93.28	134.69
Zliczanie przez łączenie *p*=2	-	92.7	92.6	92.7	99.1	92.6	92.0	7.3	1.14	93.25	134.65
Zliczanie przez wyostrzenie *p*=0.5	0.8	99.2	99.2	99.2	99.8	99.2	99.0	0.8	1.09	91.37	131.93
Wyostrzenie kontrastu	-	93.5	93.4	93.5	99.3	93.5	93.0	6.5	1.11	93.27	134.68

Stosując sterownik rozmyty prestiżu publikacji ze zmienną wejściową ***FWCI bez ram czasowych i Percentyl*** w przypadku najlepszych wariantów funkcji wagowych dokonano klasyfikacji od 92.52% do 95.05% obserwacji (od 33.6% do 37.25% więcej niż stosując podejście bazowe). Wykorzystując najlepsze badane warianty funkcji wagowych uzyskano klasyfikację z wysoką wartością metryki Accuracy i MCC (od 92.7% do 99.2% dla Accuracy i od 92.6% do 99.1% MCC). Klasyfikację uzyskano wraz z niskim ubytkiem obserwacji w zbiorze testującym (od 0.76% do 0.83%). Ze wszystkich badanych funkcji wagowych więcej zaklasyfikowanych obserwacji, niż w przypadku podejścia bazowego, uzyskano w przypadku zastosowania 19 z 34 wybranych wariantów funkcji wagowych (Tab. 4.5).

Tabela 4.5. Wyniki metryk ewaluacyjnych dla sterownika rozmytego prestiżu publikacji ze zmienną wejściową ***FWCI bez ram czasowych i Percentyl*** (5 najlepszych wyników w termie)

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
Zliczanie przez łączenie *p*=0.5	-	99.2	99.2	99.2	99.8	99.2	99.1	0.8	0.76	95.05	137.25
Zliczanie przez łączenie *p*=1	-	96.6	96.5	96.6	99.6	96.6	96.2	3.4	0.79	95.02	137.21
Zliczanie przez łączenie *p*=2	-	92.7	92.6	92.7	99.1	92.6	92.0	7.3	0.83	95.00	137.17
Zliczanie przez wyostrzenie *p*=0.5	0.8	99.2	99.2	99.2	99.8	99.2	99.0	0.8	0.80	92.52	133.60
Wyostrzenie kontrastu	-	93.5	93.4	93.5	99.3	93.5	93.0	6.5	0.80	95.01	137.20

5.3.3 Dyskusja

Zastosowanie sterowników rozmytych w algorytmie klasyfikacji dominującej dyscypliny wykazało bardzo wysoką skuteczność. Najlepsze wynik uzyskano w przypadku użycia sterownika prestiżu publikacji ze zmienną wejściową ***FWCI bez ram czasowych i Percentyl*** oraz funkcji wagowej zliczania przez łączenie *p*=0.5. Użycie tego rozwiązania pozwoliło na przejście z 69% zaklasyfikowanych obserwacji do około 95.05% obserwacji (czyli o 37.25% więcej niż stosując podejście bazowe). Klasyfikacji dokonano z uzyskaniem najwyższej wartości metryki Accuracy (99.2%) i MCC (99.1%). Opisywany sterownik rozmyty dokonał równie wysokiej klasyfikacji (ponad 95%) dla zliczania przez łączenie z *p*=1 i *p*=2 oraz wyostrzenia kontrastu, jednak z niższymi wartościami metryk ewaluacyjnych (do 92.7% Accuracy i 92% MCC).

Równie wysokie klasyfikacje uzyskano w przypadku użycia sterownika prestiżu publikacji ze zmienną wejściową ***Cytowanie i Percentyl***. W przypadku użycia funkcji wagowej zliczania przez łączenie *p*=0.5 zwiększono klasyfikację z 69% do 95% (o około 37.18%) z Accuracy 98.9%

i MCC 98.6%. Zadowalające (klasyfikację niższą jedynie o około 0.02% do 0.06%) wyniki uzyskano również w przypadku zliczania przez łączenie z p=1 i p=2 oraz wyostrzenia kontrastu.

Pomimo, że sterowniki rozmyte zbudowane na zmiennej wejściowej ***FWCI 4-letnie i Percentyl*** oraz ***FWCI 5-letnie i Percentyl*** nie zaklasyfikowały największej liczby obserwacji, nie oznacza, że nie uzyskano również wysokich wyników. Dla sterownika rozmytego stosującego ***FWCI 4-letnie i Percentyl*** i funkcję wagową zliczania przez łączenie p=0.5 zaklasyfikowano aż 92.84% obserwacji z równie wysoką wartością Accuracy jak w przypadku najlepszego z proponowanych sterowników rozmytych. Dla sterownika rozmytego stosującego ***FWCI 5-letnie i Percentyl*** w tej samej funkcji wagowej zaklasyfikowano aż 93.31% obserwacji z tą samą wartością metryki Accuracy.

Spośród czterech dostępnych sterowników rozmytych opisywane dwa najlepsze warianty posiadały najmniejszy procent obserwacji, które zaklasyfikowano w podejściu bazowym i których ni udało się zaklasyfikować w podejściu rozmytym. Dla sterownika na zmiennych wejściowych ***Cytowanie i Percentyl*** od 0.56% do 0.63%, na zmiennej ***FWCI bez ram czasowych i Percentyl*** od 0.76% do 0.83% dla pięciu najlepszych wariantów funkcji wagowych. Dla sterowników operujących na ***FWCI 4-letnim i Percentyl*** oraz ***FWCI 5-letnim i Percentyl*** było to odpowiednio od 1.15% do 1.2% oraz 1.06% do 1.14%, czyli nawet o około 214% większej różnicy pomiędzy najmniej stratnym i najbardziej stratnym podejściu w opisywanych wariantach sterowników rozmytych i funkcji wagowych. Ujmują to również różnice w gęstościach jądrowych w punkcie 0 (najniższa wartość dla sterownika stosującego ***Cytowanie i Percentyl*** oraz ***FWCI bez ram czasowych i Percentyl***, najwyższa dla ***FWCI 4-letniego i Percentyl*** oraz ***FWCI 5-letniego i Percentyl***).

Najwyższa skuteczność dla tych samych funkcji wagowych (zliczania przez łączenie, zliczania przez wyostrzenie i wyostrzenie kontrastu) wynika z występowania podobnego rozkładu gęstości obserwacji. Najwyższa gęstość występująca w przedziale 0 do 0.4 nie pozwala na stosowanie jakichkolwiek wartości progowania w funkcjach, które wykonując ten proces przypisują obserwacjom wartość zero. Stąd wśród najlepszych metod pojawia się brak występowania funkcji wagowej zliczania przez łączenie i progowanie oraz zliczania przez progowanie już od najmniejszych wartości progowania t (od wartości t=0.2). Możliwe jest jednak stosowanie funkcji wagowych, które progując nie obniżają wartości tych funkcji (na przykład zliczanie przez wyostrzenie z t=0.8).

Podsumowując klasyfikacja dominującej dyscypliny w podejściu rozmytym przebiegła pomyślnie dla każdego wariantu sterownika rozmytego. Zastosowanie tego podejścia pozwoliło na uzyskanie bardzo wysokiej klasyfikacji (do 95.05% obserwacji) z wysokim podobieństwem do klasyfikacji stosującej podejście bazowe (Accuracy do 99.2% i MCC do 99.1%). Algorytmy nie zaklasyfikowały do dominującej dyscypliny bardzo niskiej liczby obserwacji (około 4.95% wszystkich naukowców w bazie Scopus, w tym nie klasyfikując do 1.14% obserwacji klasyfikowanych podejściem bazowym).

Rozdział 6.
Klasyfikacja dominującej dyscypliny z wykorzystaniem agregacji zmiennej lingwistycznej Cytowanie i Percentyl

6.1 Wprowadzenie

W tym rozdziale opisano proces badania skuteczności klasyfikacji dominującej dyscypliny z wykorzystaniem agregacji zmiennej lingwistycznej ***Cytowanie i Percentyl***. W części metodologicznej przedstawiono proces agregowania wartości zmiennych lingwistycznych w termie wysokiej przynależności do zbioru rozmytego oraz tworzenia wyjściowej zmiennej lingwistycznej prestiżu publikacji. Kolejna część metodologii stanowi opis modyfikacji algorytmu wyznaczania dominującej dyscypliny z uwzględnieniem wyjściowej zmiennej lingwistycznej (podejście rozmyte) oraz prezentuje listę wybranych do badania wariantów funkcji wagowych. Kolejna część rozdziału prezentuje wyniki z przeprowadzonych badań. Podobnie jak w przypadku poprzednich dwóch rozdziałów wyniki badań zostały podzielone na dwie części. W pierwszej części opisano rozkłady gęstości obserwacji stanowiących dziedzinę badanych funkcji wagowych. Druga część wyników stanowi opis rezultatów uzyskanych dla algorytmu klasyfikacji dominującej dyscypliny w podejściu rozmytym. Analogicznie do poprzednich badań analizie poddano 5 wariantów funkcji wagowych, dla których dokonano najwyższej klasyfikacji oraz wykonano ewaluację wyników uzyskanych w podejściu rozmytym z wynikami z podejścia bazowego. Ostatnia część rozdziału stanowi dyskusję nad wynikami badania. *(W tym rozdziale nazwy zmiennych lingwistycznych oraz wektory wag zapisano pogrubioną kursywą w celu ich wyszczególnienia.)*

6.2 Metodologia

6.2.1 Agregacja wartości zmiennych lingwistycznych

W celu wykonania badania skonstruowano 3 zmienne lingwistyczne ***Prestiżu publikacji*** będące wynikiem trzech różnych agregacji stopni przynależności dwóch zmiennych lingwistycznych (***Cytowanie i Percentyl***) w termie wysokiej przynależności do zbioru rozmytego. W procesie

agregacji wykorzystano wartości nieostre wybranych zmiennych lingwistycznych, które zamodelowano w ten sam sposób jak w rozdziale 4.2.1. Następnie wyniki zagregowano operatorem uporządkowanej średniej ważonej (OWA). Do badania wybrano 3 warianty wag. Pierwsza waga przyporządkowywała największej wartości wartość 0, a najmniejszej wartość 1 tj. wektor ***(0, 1)***. Ten dobór wag spowodował powstanie agregacji, która tożsama jest z operacją minimum (rozdział 3.3.1). Drugi wektor wag przyporządkowywał największej wartości wartość 0 i najmniejszej wartość 1. Uzyskano w ten sposób wektor ***(1, 0)*** utożsamiany z operacją maksimum. Eksperymentalnie dla ostatniego wektora wag wybrano wektor ***(0.8, 0.2)***.

Wyznaczenie mocy skalarnej indukowanej przez funkcję wagową *f* (zgodnie z definicją 3.9) wymaga dysponowania skończonym zbiorem rozmytym, a w wyniku agregacji (zgodnie z definicją 3.5) uzyskuje się wartości ostre z przedziału [0, 1]. W związku z tym wynik agregacji dla wartości w termie wysokiej przynależności do zbioru rozmytego stanowił element wyjściowej zmiennej lingwistycznej Prestiżu publikacji (skonstruowanej analogicznie do zmiennej wyjściowej sterownika rozmytego prestiżu publikacji opisanej w sekcji 5.2.1).

6.2.2 Algorytm wyznaczania dominującej dyscypliny z wykorzystaniem agregacji zmiennej lingwistycznej Cytowanie i Percentyl

Algorytm wyznaczania dominującej dyscypliny w podejściu bazowym zmodyfikowano tak, aby wykorzystywał operator agregacji OWA wartości dwóch zmiennych lingwistycznych (***Cytowanie***, ***Percentyl***) i wyznaczenia mocy zbiorów rozmytych z zastosowaniem badanych funkcji wagowych (wariant algorytmu nazywany podejściem rozmytym). Na początku do kolumn zawierających unikalny identyfikator publikacji, listę identyfikatorów autorów oraz listę dyscyplin przypisanych do czasopisma, z którego pochodziła publikacja, dołączono dwie kolumny reprezentujące wartości ostre dwóch zmiennych lingwistycznych. Następnie, dokonano normalizacji tabeli do pierwszej postaci normalnej ze względu na listę identyfikatorów autorów i listę dyscyplin. Kolejny krok obejmował wyznaczenie i dołączenie dla każdego rekordu wartości stopnia przynależności do zbioru rozmytego dwóch zmiennych lingwistycznych, termów i przypisanej rekordowi dyscypliny naukowej. Dwie wartości nieostre zostały zagregowane operatorem OWA zawierającym dwuelementowy wektor wag. Następnie przystąpiono do wyznaczania wartości dla badanej funkcji wagowej w oparciu o wartość uzyskaną na etapie agregacji oraz usunięto rekordy, dla których wartość tej funkcji wynosiła zero (uzyskując nośnik

zbioru rozmytego). Następnie dokonano wyznaczenia mocy zbiorów rozmytych dla każdej dyscypliny autora.

Kolejne kroki algorytmu przebiegały podobnie do wariantu w podejściu bazowym, z tym że zamiast wyznaczać maksymalną wartość z liczby publikacji w dyscyplinie autora, wyznaczono maksymalną wartość ze wszystkich wartości mocy zbiorów rozmytych dyscyplin autora. Z tabeli wybrano tylko te rekordy, dla których moc zbioru rozmytego dla dyscypliny autora była równa wartości maksymalnej mocy zbioru dla dyscypliny autora. Następnie zbiór autorów ograniczono jedynie do tych, którzy posiadali jedną dyscyplinę, dla której osiągnięto maksymalną wartość mocy zbioru rozmytego wysokiego prestiżu publikacji.

Algorytm 4. Algorytm klasyfikacji dominującej dyscypliny z wykorzystaniem agregacji zmiennej lingwistycznej Cytowanie i Percentyl

Wejście: zbiór publikacji w bazie Scopus,
Parametry: pierwsza zmienna wejściowa (Cytowanie), druga zmienna wejściowa (Percentyl), wektor wag, funkcja wagowa
Wyjście: zbiór par autor-dyscyplina

1. Dołącz do tabeli kolumnę z wartościami dla pierwszej i drugiej zmiennej wejściowej
2. Wykonaj normalizację tabeli do pierwszej postaci normalnej ze względu na identyfikatory autorów i listę dyscyplin
3. Wyznacz wartość stopnia przynależności do zbioru rozmytego dla zmiennych lingwistycznych i termu w dyscyplinie
4. Dokonaj agregacji dwóch wartości rozmytych operatorem OWA z wektorem wag
5. Oblicz wartość wybranej funkcji wagowej dla wartości wyznaczonej w kroku 4.
6. Wyznacz nośnik powstałego zbioru rozmytego
7. Dla każdego autora i jego dyscyplin oblicz moc zbioru rozmytego
8. Dla każdego autora wyznacz maksymalną z wartości mocy zbiorów rozmytych
9. Dla każdego autora wybierz dyscypliny, dla których moc zbioru rozmytego odpowiada maksymalnej z wartości mocy zbiorów rozmytych
10. Dokonaj filtracji tabeli do autorów posiadających jedną zaklasyfikowaną dyscyplinę

6.2.3 Wybór badanych funkcji wagowych

Wybór funkcji wagowych w tej części badania pokrywał się z wyborem funkcji dla badania algorytmu klasyfikacji dominującej dyscypliny z wykorzystaniem zmiennych lingwistycznych (Rozdział 4). W badaniu wykorzystano 5 wariantów funkcji wagowych: zliczanie przez progowanie, zliczanie przez łączenie, zliczanie przez progowanie i łączenie, zliczanie przez

wzmocnienie oraz wzmocnienie kontrastu. W celu ustalenia wartości *p* oraz progowania *t* polegano na tej samej zasadzie jak w przypadku badania z poprzedniego rozdziału. Pełna lista badanych wariantów funkcji wagowych została opisana oraz wylistowana w rozdziale 4.2.4. Łącznie dla algorytmu wyznaczania dominującej dyscypliny w podejściu rozmytym uzyskano 88 badanych sposobów w każdej z wybranej kombinacji wag operatora OWA na wykonanie klasyfikacji.

6.3 Wyniki

6.3.1 Badanie rozkładu gęstości dziedziny funkcji wagowych

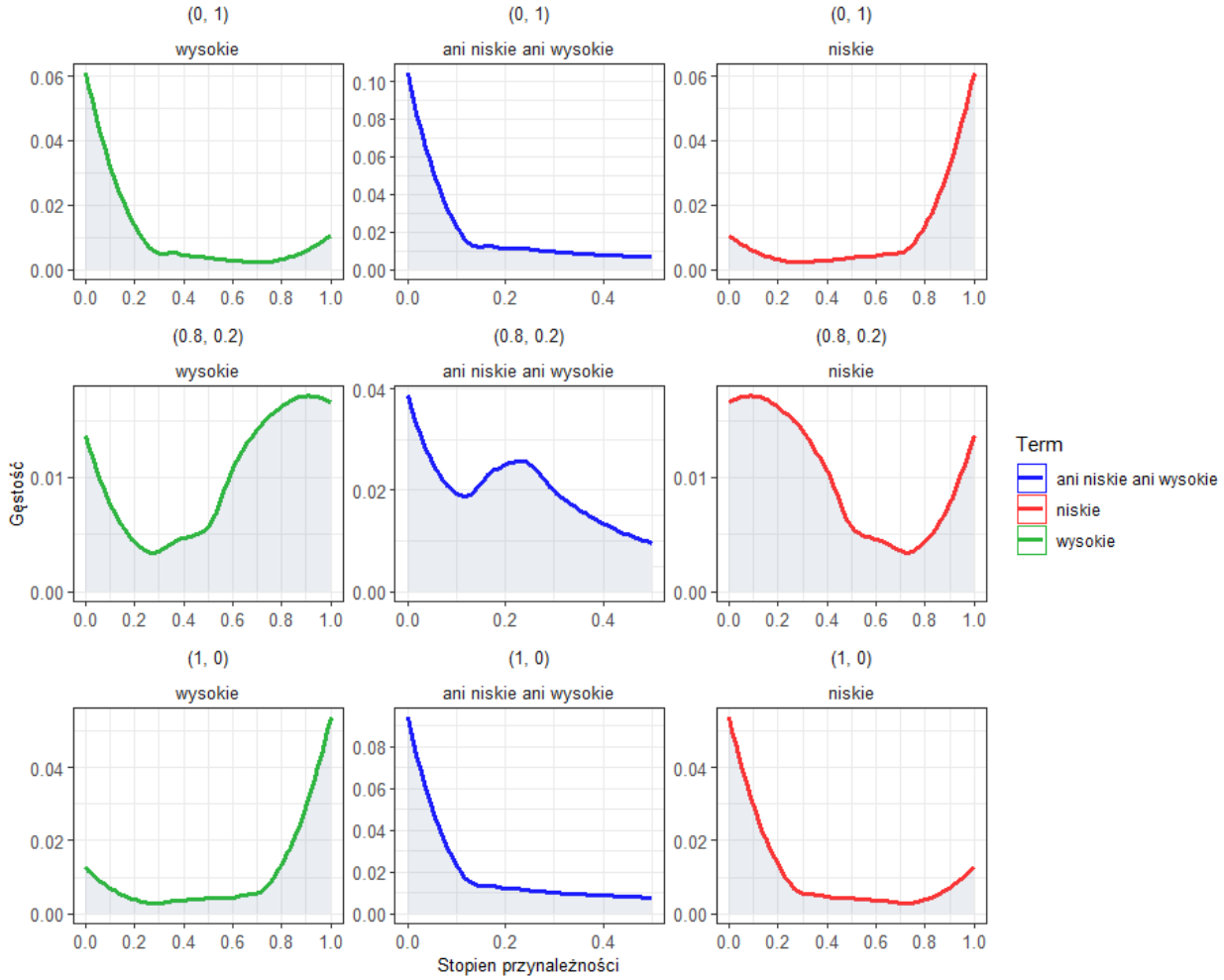
Podobnie jak w przypadku algorytmu wyznaczania dominującej dyscypliny w podejściu rozmytym poprzez zmienne lingwistyczne i sterowniki rozmyte zbadano rozkłady wartości dziedziny funkcji wagowych. Ze względu na charakter zastosowanej zmiany (aplikację zmiennej lingwistycznej występującej z trzema wartościami) analizie poddano termy wysokiej przynależności do zbioru rozmytego oraz ani niskiej ani wysokiej przynależności do zbioru rozmytego. Term niskiej przynależności do zbioru stanowił symetrię termu wysokiej przynależności do zbioru (podobnie jak w przypadku badania w rozdziale czwartym).

Rozkłady gęstości jądrowej dziedziny dla funkcji wagowych różniły się w zależności od wybranego wektora wag. W przypadku wektora ***(1, 0)*** w termie wysokiej przynależności do zbioru rozmytego najwięcej obserwacji występowało w przedziale od 0 do 0.2 (odpowiednio 0.8 do 1 dla termu niskiej przynależności). Odwrotna sytuacja wystąpiła dla wektora ***(1, 0)***, gdzie najwyższą gęstość uzyskano dla przedziału [0.8, 1] dla termu wysokiej przynależności (odpowiednio [0, 0.2] dla termu niskiej przynależności). W termie ani niskiej ani wysokiej przynależności zarówno dla wektora ***(1, 0)*** jak i ***(0, 1)*** obserwacje dominowały w przedziale od 0 do 0.1.

W przypadku wektora wag ***(0.8, 0,2)*** w termie wysokiej przynależności najwięcej obserwacji występowało w przedziale 0 do 0.2 oraz 0.6 do 1 (odpowiednio 0 do 04 oraz 0.8 do 1 w termie niskiej przynależności). Dla termu ani niskiej ani wysokiej przynależności największą gęstość uzyskano w przedziale od 0 do 0.1 oraz dla wartości 0.2.

Najwięcej obserwacji, które podlegały wykluczeniu przy wyznaczaniu mocy zbiorów rozmytych dla termu wysokiej przynależności i ani niskiej ani wysokiej przynależności zaobserwowano w wektorze wag ***(0, 1)*** tj. 6% i 10%; dodatkowo w termie ani niskiej ani wysokiej

przynależności dla wektora ***(1,0)***. W termie niskiej przynależności najwięcej wykluczonych obserwacji występowało dla wektora wag ***(1, 0)*** (Rys. 6.2).



Rys 6.2. Wykres gęstości jądrowej dla dziedziny funkcji wagowych w ujęciu wektorów wag operatora OWA i termów

6.3.2 Ewaluacja algorytmu klasyfikacji dominującej dyscypliny w podejściu rozmytym

Poniższe wyniki przedstawiają ewaluacje klasyfikacji uzyskanej przez algorytm klasyfikacji dominującej dyscypliny z wykorzystaniem operatorów agregacji wartości zmiennej lingwistycznej ***Cytowanie*** i ***Percentyl***. Podobnie jak w przypadku podejścia z rozdziału trzeciego oraz czwartego analizie poddano 5 wariantów funkcji wagowych, których wykorzystanie w algorytmie dokonało najwyższej klasyfikacji spośród badanych grup funkcji wagowych. Pełne wyniki zawierające listę wszystkich wariantów, które dokonały wyższej klasyfikacji niż w podejściu bazowym umieszczono w załączniku w sekcji 2.

Dla kombinacji wag ***(0, 1)*** wyższą klasyfikację niż w przypadku podejścia bazowego uzyskano dla dwóch z trzech badanych termów. Zarówno dla termu niskiej przynależności jak i wysokiej przynależności najlepsze klasyfikacje dokonano dla tych samych wariantów funkcji wagowych (3 wersje zliczania przez łączenie, wyostrzenie kontrastu oraz zliczanie przez wyostrzenie z p=0.5 i t=0.8). Wykorzystując term niskiej przynależności uzyskano wysokie wartości metryki Accuracy i MCC dla 3 wariantów funkcji wagowych (od 87.9% do 89.4% Accuracy i 87.7% do 89.3% MCC). Dla pozostałych dwóch wariantów uzyskano niższe wyniki metryk ewaluacyjnych o około 10% (77.0%-77.05% Accuracy i 77.6%-78% MCC). Średnio klasyfikacja w termie niskiej przynależności wynosiła od około 80.13% do 81.72% (o około 15.7% do 18.01% więcej niż w podejściu bazowym). Dla termu wysokiej przynależności uzyskano nieznacznie wyższą klasyfikację niż w przypadku podejścia bazowego tj. około 4.84% do 6.21% więcej zaklasyfikowanych obserwacji. Dla dwóch wariantów funkcji wagowych klasyfikacji dokonano z około 86% wynikiem metryki Accuracy i MCC (86.6% i 86.7% Accuracy, 86.4% i 86.5% MCC). Dla zliczania przez łączenie z p=2 oraz wyostrzenia kontrastu około 50% obserwacji uzyskało klasyfikację innej dyscypliny naukowej (50.1% i 50.3% Accuracy, 52.3% i 52.6% MCC). Klasyfikacji dokonano z wysoką redukcją obserwacji w zbiorze testującym (około 26%). Ze wszystkich badanych funkcji wagowych więcej zaklasyfikowanych obserwacji, niż w przypadku podejścia bazowego, uzyskano w przypadku zastosowania 29 z 88 wybranych wariantów funkcji wagowych (Tab. 6.1).

Tabela 6.1. Wyniki metryk ewaluacyjnych dla agregacji operatorem OWA zmiennej lingwistycznej Cytowanie i Percentyl z wektorem wag ***(0, 1)*** (5 najlepszych wyników w termie)

Funkcja wagowa	t	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niskie											
Zliczanie przez łączenie p=0.5	0	89.3	89.9	89.3	99.5	89.6	89.3	10.7	6.21	81.72	118.01
Zliczanie przez łączenie p=1	0	87.9	88.6	87.9	99.2	88.2	87.7	12.1	6.30	81.64	117.89
Zliczanie przez łączenie p=2	0	77.5	78.2	77.5	98.7	77.8	78.0	22.5	6.38	81.62	117.86
Zliczanie przez wyostrzenie p=0.5	0.8	89.4	89.9	89.4	99.5	89.6	89.3	10.6	6.29	80.13	115.70
Wyostrzenie kontrastu	0	77.0	77.7	77.0	98.8	77.3	77.6	23.0	6.36	81.64	117.88
term = wysokie											
Zliczanie przez łączenie p=0.5	0	86.6	86.9	86.6	99.1	86.8	86.4	13.4	26.08	73.55	106.21
Zliczanie przez łączenie p=1	0	79.9	80.1	79.9	98.5	79.9	79.8	20.1	26.11	73.52	106.16
Zliczanie przez łączenie p=2	0	50.1	48.8	50.1	96.3	49.4	52.3	49.9	26.13	73.50	106.14
Zliczanie przez wyostrzenie p=0.5	0.8	86.7	86.9	86.7	99.1	86.8	86.5	13.3	26.17	72.60	104.84
Wyostrzenie kontrastu	0	50.3	49.1	50.3	96.4	49.7	52.6	49.7	26.11	73.52	106.16

Uwzględniając wektor wag ***(1, 0)*** uzyskano poprawę klasyfikacji we wszystkich badanych termach. Najwyższą klasyfikacje uzyskano agregując wartości termu niskiej przynależności zmiennych lingwistycznych. W przypadku tego podejścia zaklasyfikowano około 90% obserwacji (88.6% do 90.87%, czyli o około 27.94% do 31.21% więcej niż w podejściu bazowym). Dla termu ani niskiej ani wysokiej przynależności uzyskano średnio niższą klasyfikację o 5 punktów procentowych. W tym przypadku zaklasyfikowano około 83% do 85% obserwacji (83.37% do 85.91%, czyli o około 20.38% do 24.06% więcej niż w podejściu bazowym). Wysoką klasyfikację uzyskano również w przypadku termu wysokiej przynależności. Stosując tę wartość w czterech najlepszych wariantach funkcji wagowej zaklasyfikowano około 81% obserwacji (około 17% więcej niż w podejściu bazowym). Dla każdego badanego termu w pięciu najlepszych wariantach funkcji wagowych uzyskano klasyfikację obserwacji z wysokimi wartościami metryk ewaluacyjnych (od 86.8% do 99.1% Accuracy i 86.7% do 99% MCC). Ze wszystkich badanych funkcji wagowych więcej zaklasyfikowanych obserwacji, niż w przypadku podejścia bazowego, uzyskano w przypadku zastosowania 51 z 88 wybranych wariantów funkcji wagowych (Tab. 6.2).

Tabela 6.2. Wyniki metryk ewaluacyjnych dla agregacji operatorem OWA zmiennej lingwistycznej Cytowanie i Percentyl z wektorem wag ***(1, 0)*** (5 najlepszych wyników w termie)

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niskie											
Zliczanie przez łączenie *p*=0.5	0	93.7	94.0	93.7	99.9	93.8	93.7	6.3	0.92	90.87	131.21
Zliczanie przez łączenie *p*=1	0	92.0	92.4	92.0	99.8	92.2	92.0	8.0	0.93	90.86	131.20
Zliczanie przez łączenie *p*=2	0	88.5	89.2	88.5	99.7	88.8	88.7	11.5	0.95	90.85	131.18
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	91.9	92.4	91.9	99.8	92.1	92.0	8.1	3.13	88.60	127.94
Wyostrzenie kontrastu	0	89.1	89.7	89.1	99.7	89.4	89.4	10.9	0.93	90.86	131.19
term = ani niskie ani wysokie											
Zliczanie przez łączenie *p*=0.5	0	94.7	94.8	94.7	99.7	94.7	94.5	5.3	9.00	85.91	124.06
Zliczanie przez łączenie *p*=1	0	92.4	92.7	92.4	99.5	92.6	92.1	7.6	9.03	85.89	124.02
Zliczanie przez łączenie *p*=2	0	87.1	87.6	87.1	99.1	87.4	86.8	12.9	9.06	85.87	124.00
Zliczanie przez wyostrzenie *p*=0.5	0.4	93.2	93.3	93.2	99.6	93.2	92.9	6.8	9.17	83.37	120.38
Wyostrzenie kontrastu	0	87.1	87.6	87.1	99.1	87.4	86.8	12.9	9.07	85.86	123.98
term = wysokie											
Zliczanie przez łączenie *p*=0.5	0	99.1	99.1	99.1	99.9	99.1	99.0	0.9	7.58	81.17	117.21
Zliczanie przez łączenie *p*=1	0	97.2	97.2	97.2	99.8	97.2	97.1	2.8	7.62	81.14	117.17
Zliczanie przez łączenie *p*=2	0	86.8	87.0	86.8	99.2	86.9	86.7	13.2	7.67	81.12	117.14
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	88.2	88.5	88.2	99.3	88.3	88.0	11.8	16.06	73.75	106.50
Wyostrzenie kontrastu	0	87.3	87.5	87.3	99.3	87.4	87.2	12.7	7.63	81.10	117.11

Dla kombinacji ***(0.8, 0.2)*** wektora wag operatora OWA uzyskano wysoką klasyfikację we wszystkich badanych termach. Dla termu niskiej przynależności zaklasyfikowano nawet 94%

obserwacji (od 92.35% do 94.89%), czyli nawet o około 37% (33.35% do 37.02%) więcej niż w podejściu bazowym. Dla termu wysokiej przynależności 4 badane funkcje wagowe dokonały klasyfikacji ponad 90% obserwacji, czyli o około 30% więcej niż w podejściu bazowym. Podobne wyniki do wartości wysokiej przynależności uzyskano dla termu ani niskiej ani wysokiej przynależności. W tym przypadku zaklasyfikowano ponad 89% obserwacji, czyli o około 29% więcej niż w podejściu bazowym. Podobnie jak w przypadku wektora wag ***(1, 0)*** uzyskano wysokie wartości metryk ewaluacyjnych w każdym termie (od około 86.9% do 97.7% Accuracy i 86.2% do 97.6% MCC). Ze wszystkich badanych funkcji wagowych więcej zaklasyfikowanych obserwacji, niż w przypadku podejścia bazowego, uzyskano w przypadku zastosowania 54 z 88 wybranych wariantów funkcji wagowych (Tab. 6.3).

Tabela 6.3. Wyniki metryk ewaluacyjnych dla agregacji operatorem OWA zmiennej lingwistycznej Cytowanie i Percentyl z wektorem wag ***(0.8, 0.2)*** (5 najlepszych wyników w termie)

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niskie											
Zliczanie przez łączenie *p*=0.5	0	93.7	93.9	93.7	99.9	93.8	93.7	6.3	0.90	94.89	137.02
Zliczanie przez łączenie *p*=1	0	91.9	92.3	91.9	99.8	92.1	91.9	8.1	0.90	94.89	137.02
Zliczanie przez łączenie *p*=2	0	89.6	90.3	89.6	99.6	90.0	89.7	10.4	0.91	94.88	137.01
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	94.0	94.5	94.0	99.8	94.2	93.9	6.0	3.30	92.35	133.35
Wyostrzenie kontrastu	0	90.2	90.8	90.2	99.7	90.5	90.3	9.8	0.90	94.89	137.02
term = ani niskie ani wysokie											
Zliczanie przez łączenie *p*=0.5	0	94.6	94.8	94.6	99.7	94.7	94.5	5.4	8.93	89.55	129.31
Zliczanie przez łączenie *p*=1	0	92.4	92.7	92.4	99.5	92.5	92.1	7.6	8.93	89.55	129.31
Zliczanie przez łączenie *p*=2	0	87.1	87.6	87.1	99.0	87.3	86.8	12.9	8.94	89.54	129.30
Zliczanie przez wyostrzenie *p*=0.5	0.4	93.6	93.8	93.6	99.7	93.7	93.4	6.4	8.95	89.06	128.60
Wyostrzenie kontrastu	0	87.1	87.6	87.1	99.0	87.3	86.8	12.9	8.94	89.54	129.30
term = wysokie											
Zliczanie przez łączenie *p*=0.5	0	97.7	97.7	97.7	99.9	97.7	97.6	2.3	7.46	90.49	130.66
Zliczanie przez łączenie *p*=1	0	97.0	97.0	97.0	99.8	97.0	96.8	3.0	7.47	90.49	130.66
Zliczanie przez łączenie *p*=2	0	86.4	86.5	86.4	99.1	86.4	86.2	13.6	7.49	90.48	130.65
Zliczanie przez wyostrzenie *p*=0.5	0.8	97.7	97.7	97.7	99.9	97.7	97.6	2.3	7.54	84.38	121.84
Wyostrzenie kontrastu	0	86.9	87.1	86.9	99.2	87.0	86.8	13.1	7.48	90.48	130.66

6.3.3 Dyskusja

Klasyfikacja dominującej dyscypliny autora z wykorzystaniem operatorów OWA do agregacji wartości ***Cytowania*** oraz ***Percentyla*** czasopisma również wykazała dużą skuteczność. Liczba zaklasyfikowanych obserwacji różniła się znacząco w zależności od doboru wektora wag oraz termu zmiennej wyjściowej. Najlepsze wyniki dla każdego z trzech termów uzyskano dla wektora

wag ***(0.8, 0.2)***, gdzie zaklasyfikowano do 95% obserwacji (94.89% dla termu niskie, 90.49% dla termu wysokie oraz 89.55% dla termu ani niskie ani wysokie). W każdym przypadku najlepsze wyniki otrzymano dla zliczania przez łączenie w trzech kombinacjach wartości *p* oraz wyostrzenia kontrastu.

Dla wektorów ***(1, 0)*** oraz ***(0.8, 0.2)***, czyli wektorów uwzględniających z większą wagą większe wartości funkcji wzorcowych uzyskano wysokie wartości metryk ewaluacyjnych. W tym przypadku wartości metryk Accuracy i MCC wynosiły co najmniej około 86% (od 86.8% do 99.1% Accuracy i 86.7% do 99% MCC). Odwrotna sytuacja wystąpiła dla wektora wag ***(0, 1)***, czyli agregacji utożsamianej z operacją minimum, gdzie w termie wysokiej przynależności tylko 50% obserwacji uzyskało taką samą klasyfikację jak w przypadku podejścia bazowego (50.1% i 50.3% Accuracy, 52.3% i 52.6% MCC).

Wyższą klasyfikację dla operatorów promujących wysokie wartości (dwa powyżej opisywane warianty) uzasadnia również rozkład obserwacji estymacji gęstości jądrowej. Wybierając te wektory wag uzyskano małą redukcję obserwacji w dziedzinie dla funkcji wagowych (około 1% estymowanej gęstości w termie wysokiej przynależności oraz 1% dla kombinacji 0.8 , 0.2 i 5% dla kombinacji 1 i 0). Podobnie niską redukcję uzyskano dla zbiorów testujących, gdzie ubytek ten wynosił mniej niż 1% dla termu niskiej przynależności i do około 9% dla pozostałych termów. Dla operatora promującego niskie wartości procent obserwacji niestanowiący nośnik zbioru był dużo wyższy. Dla wektora ***(0, 1)*** uzyskano najwyżej estymowaną gęstość jądrową w punkcie 0 dla termu niskiej przynależności (około 6%) i ani niskiej ani wysokiej przynależności (ponad 10%). Ten wynik przełożył się również na procent obserwacji, których nie zaklasyfikował algorytm w podejściu rozmytym prowadząc do 26-procentowej redukcji obserwacji wykorzystując term wysokiej przynależności. Ponadto spowodował całkowity brak występowania wariantów funkcji wagowych podnoszących klasyfikację w stosunku dla podejścia bazowego w termie ani niskiej ani wysokiej przynależności.

Podsumowując, stosowanie operatorów OWA pozwoliło na uzyskanie wyższej klasyfikacji niż w przypadku zastosowania podejścia bazowego. Za sprawą tego podejścia klasyfikacja obserwacji wzrosła z 69% do ponad 95%. Szczególny wpływ na poprawę wyników uzyskano dobierając wektory wag w taki sposób, aby promowały największe wartości.

Rozdział 7.
Podsumowanie

7.1 Podsumowanie wyników badań

Praca ta podejmuje kluczowe wyzwanie w obszarze naukometrii, jakim jest klasyfikacja dyscyplin naukowych autorów, proponując istotne modyfikacje w algorytmie przez uwzględnienie metod opartych na teorii zbiorów rozmytych. Niejednoznaczność klasyfikacji autorów do dominującej dyscypliny stworzyła zapotrzebowanie na wykorzystanie bardziej zaawansowanych metod uwzględniających nieprecyzyjność i wielowymiarowość danych naukowych.

Badania wykazały, że stosując podejście rozmyte możliwe jest jednoznaczne zaklasyfikowanie większej ilości obserwacji, niż w przypadku podejścia bazowego. W wyniku zastosowania tej modyfikacji uzyskano klasyfikację aż do ponad 95% wszystkich autorów z badanej próby w porównaniu do około 69% obserwacji zaklasyfikowanych podejściem bazowym (H1). Dla każdego z trzech zastosowanych podejść: zmiennych lingwistycznych, sterowników rozmytych uzyskano warianty, które wykazywały wysoką jednoznaczną klasyfikację (H2). W pracy zaprezentowano zmienne, termy i funkcje wagowe, które pozwoliły na poprawę jednoznacznej klasyfikacji dyscyplin. Podobnie wskazano rolę sterowników rozmytych i ich kluczowe parametry wejściowe. Omówiono również zastosowanie operatorów OWA z uwzględnieniem metod promowania wartości przez dobór odpowiednich wektorów wag (H3-H5). Wyniki badań wykazały, że w zależności od wybranego podejścia klasyfikacja metodą rozmytą niekoniecznie pokrywa się z klasyfikacją w podejściu bazowym (H6).

7.2 Kontynuacja badań

Zaproponowana w pracy modyfikacja algorytmu klasyfikacji dominującej dyscypliny miała charakter eksperymentalny. W pracy zbadano wybrane i najszerzej znane rozwiązania teorii zbiorów rozmytych, aby sprawdzić możliwość ich zastosowania w opisywanym algorytmie. Wysoka skuteczność zaproponowanych rozwiązań, jednak z nadal występującą

niejednoznacznością w mniejszej skali, sprawia, że należy szukać kolejnych metod, które dokonywałyby większą jednoznaczną klasyfikację.

Wyznaczanie dominującej dyscypliny obejmowało zliczanie częstości ich występowania na podstawie pełnego portfolio publikacyjnego autora. Kolejne warianty mogą obejmować inne ramy czasowe działalności publikacyjnej. Poza zastosowaniem pełnego portfolio możliwy jest na przykład wybór pełnych pierwszych lub ostatnich lat. Następnie w zależności od skali problemu możliwe jest zaproponowanie modyfikacji podejścia, aby podobnie jak w przypadku tej pracy uwzględniała nieprecyzyjność informacji.

Inny aspekt badań może obejmować dobór kolejnych metryk naukowych i parametrów na potrzeby modelowania nieprecyzyjności informacji. Przy konstrukcji zmiennych lingwistycznych i sterowników rozmytych można uwzględnić inne parametry wejściowe. Mogą one obejmować na przykład liczbę stron, umiędzynarodowienie zespołu badawczego, liczbę referencji lub doświadczenie akademickie zespołu definiując średni wiek akademicki autorów pracy (Kwiek i Roszka, 2022a,b). Podobnie w przypadku operatorów agregacji OWA można uwzględnić większą ilość agregowanych wartości i kombinacji wektorów wag.

W badaniu wykorzystano pięć funkcji wagowych: zliczanie przez progowanie, zliczanie przez łączenie, zliczanie przez progowanie i łączenie, zliczanie przez wyostrzenie oraz wyostrzenie kontrastu. Poza tymi funkcjami istnieją inne warianty, których zastosowanie w algorytmie mogłoby wykazać równie wysoką jednoznaczność klasyfikacji dyscypliny. Jako kolejne funkcje wagowe można zastosować generatory Archimedejskiej t-normy (Wygralak, 2013) z uwzględnieniem bardziej zaawansowanych rodzin norm triangularnych prezentowanych w rozdziale 3.1.3.

Dodatkowo przedstawione w pracy funkcje wagowe można rozszerzyć o inne wartości parametrów progowania. Trzy z wykorzystanych funkcji wagowych: zliczanie przez progowanie, zliczanie przez progowanie i łączenie oraz zliczanie przez wyostrzenie przyjmowały jako parametr wartość progową *t*. Alternatywnym rozwiązaniem mogłoby być wykorzystanie tych funkcji z zastosowaniem progowania ostrego. Przykłady takich funkcji wskazano w rozdziale 3.5.2. Inny przypadek może stanowić wykorzystywanie przedziałowości progowania w funkcjach wagowych (Dyczkowski, 2018). Poza rozbudową funkcji wagowych o inne metody progowania analizie można poddać również inne wartości parametru potęgującego.

Bibliografia

1. Abramo, G., Aksnes, D. W. i D'Angelo, C. A. (2020). Comparison of research performance of Italian and Norwegian professors and universities. Journal of Informetrics, 14(2), Elsevier. doi.org/10.1016/j.joi.2020.101023,
2. Azahar, T., Norlaila, A. i Nuwairah, A. (2015). Development of Fuzzy Logic Based Odor Detection. Journal of Science & Engineering Technology JSET Vol: 02 No: 02. ir.unikl.edu.my/xmlui/handle/123456789/11993,
3. Baas, J., Schotten, M., Plume, A., Côté, G. i Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative Science Studies, 1(1), 377–386, MIT Press. doi.org/10.1162/qss_a_00019,
4. Beliakov, G., Pradera, A. i Calvo, T. (2007). Aggregation Functions: A Guide for Practitioners, Studies in Fuzziness and Soft Computing, 221, Springer. doi.org/10.1007/978-3-540-73721-6,
5. Bhatt, C., Kumar, I. i Vijayakumar, V. (2021). The state of the art of deep learning models in medical science and their challenges. Multimedia Systems 27, 599–613, Springer. doi.org/10.1007/s00530-020-00694-1,
6. Birkle, C., Pendlebury, D., Schnell, J. D. i Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. Quantitative Science Studies, 1(1), 363–376, MIT Press. doi.org/10.1162/qss_a_00018,
7. Bishop, C. (2006). Pattern Recognition and Machine Learning, Information Science and Statistic, Springer. doi.org/10.1007/978-0-387-45528-0,
8. Boekhout, H., van der Weijden, I. i Waltman L. (2022). Gender differ-ences in scientific careers: A large-scale bibliometric analysis. arxiv.org/abs/2106.12624,
9. Burnham, J. F. (2006). Scopus database: a review. Biomedical Digital Libraries, 3(1), Springer. doi.org/10.1186/1742-5581-3-1,
10. Clarivate (2021). Web of Science Research Areas. incites.help.clarivate.com/Content/Research-Areas/research-areas.htm
11. Cordón, O., Herrera, F. i Peregrín, A. (2000). Searching for basic properties obtaining robust implication operators in fuzzy control. Fuzzy Sets and Systems, 111(2), 237–251, Elsevier. doi.org/10.1016/s0165-0114(97)00402-8,

12. Daradkeh M, Abualigah L, Atalla S, Mansoor W. (2022). Scientometric Analysis and Classification of Research Using Convolutional Neural Networks: A Case Study in Data Science and Analytics, Electronics, 11(13), MDPI. doi.org/10.3390/electronics11132066,
13. De Luca, A. i Termini, S. (1972). A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. Information and Control, 20(4), 301–312. doi.org/10.1016/s0019-9958(72)90199-4,
14. Dimensions (2019). Dimensions Raport. A Guide to the Dimensions Data Approach. dimensions.ai/resources/a-guide-to-the-dimensions-data-approach,
15. Dimensions (2022). Which research categories and classification schemes are available in Dimensions? dimensions.freshdesk.com/support/solutions/articles/23000018820-which-research-categories-and-classification-schemes-are-available-in-dimensions-
16. Ding, Y., Rousseau, R. i Wolfram, D. (2014). Measuring scholarly impact. Springer. doi.org/10.1007/978-3-319-10377-8,
17. Doctor, F., Syue, C., Liu, Y., Shieh, J. i Iqbal, R. (2016). Type-2 fuzzy sets applied to multivariable self-organizing fuzzy logic controllers for regulating anesthesia. Applied Soft Computing, 38, 872–889, Elsevier. doi.org/10.1016/j.asoc.2015.10.014,
18. Dubois, D. i Prade, H. (1993). A THEOREM ON IMPLICATION FUNCTIONS DEFINED FROM TRIANGULAR NORMS. Readings in Fuzzy Sets for Intelligent Systems, 105–111, Morgan Kaufmann Publishers. doi.org/10.1016/b978-1-4832-1450-4.50014-6,
19. Dunham, J., Melot, J. i Murdick, D. (2020). Identifying the Development and Application of Artificial Intelligence in Scientific Text, arXiv. doi.org/10.48550/arXiv.2002.07143,
20. Dyczkowski K., Gadecki P. i Kułakowski A. (2011). Traffic Signs Recognition System. Proceedings of the World Confernce on Soft Computing, San Francisco State University, USA. min.wmi.amu.edu.pl/wp-content/uploads/2010/10/TSRS.pdf,
21. Dyczkowski, K. (2018). Intelligent Medical Decision Support System Based on Imperfect Information. The Case of Ovarian Tumor Diagnosis, Studies in Computational Intelligence, 735, Springer. doi.org/10.1007/978-3-319-67005-8,
22. Elsevier (2018). Research Metrics Guidebook. elsevier.com/products/scopus/Metrics
23. Elsevier (2020a), The Researcher Journey Through a Gender Lens. An examination of research participation, career progression and perceptions across the globe. Retrieved from elsevier.com/gender-report,

24. Elsevier (2020b), The Researcher Journey Through a Gender Lens: Focus on Japan. An examination of research participation, career progression and perceptions across the globe. Retrieved from elsevier.com/gender-report,
25. Elsevier (2023). What are Scopus subject area categories and ASJC codes? service.elsevier.com/app/answers/detail/a_id/12007/supporthub/scopus/related/1/
26. Eykens, J., Guns, R. i Engels, T. (2021). Fine-grained classification of social science journal articles using textual data: A comparison of supervised machine learning approaches. Quantitative Science Studies, 2(1), 89–110, MIT Press. doi.org/10.1162/qss_a_00106,
27. Eze, U., Emmanuel, I. i Stephen, E. (2014). Fuzzy logic model for traffic congestion. IOSR Journal of Mobile Computing & Application, 1(1), 15–20. doi.org/10.9790/0050-0111520,
28. Faraji, O., Asuaeu, K., Razaee, Z., Bontis, N. i Dolatzarei, E. (2022). Mapping the conceptual structure of intellectual capital research: A co-word analysis,' Journal of Innovation and Knowledge, 7(3), p. 100202, Elsevier. doi.org/10.1016/j.jik.2022.100202,
29. Garfield, E. (2006). Citation indexes for science. A new dimension in documentation through association of ideas†. International Journal of Epidemiology, 35(5), 1123–1127, Oxford Academic. doi.org/10.1093/ije/dyl189,
30. Gottwald, S. (1999). Many-Valued Logic And Fuzzy Set Theory. The Handbooks of Fuzzy Sets Series, 3, Springer. doi.org/10.1007/978-1-4615-5079-2_2,
31. Harzing, A. (2010). The Publish or Perish Book : Your guide to effective and responsible citation analysis. ISBN 978-0-9808485-0-2, Tarma Software Research Pty Ltd,
32. Heersmink, R., van den Hoven, J. i van Eck, N.J. (2011). Bibliometric mapping of computer and information ethics. Ethics and Information Technology, 13(3), pp. 241–249, Springer. doi.org/10.1007/s10676-011-9273-7,
33. Hong, T. P. i Lee, C. Y. (1996). Induction of fuzzy rules and membership functions from training examples. Fuzzy Sets and Systems, 84(1), 33–47, Elsevier. doi.org/10.1016/0165-0114(95)00305-3,
34. Hook, D.W., Porter, S. i Herzog, C. (2018) Dimensions: building context for search and evaluation. Frontiers in Research Metrics and Analytics, 3, Frontiers Media. doi.org/10.3389/frma.2018.00023.
35. International Electrotechnical Commission (2000). IEC 1131 - Programmable controllers. Part 7: Fuzzy control programming,

36. Japkowicz, N. i Shah, M. (2011). Evaluating learning algorithms, Cambridge University Press. doi.org/10.1017/cbo9780511921803,
37. Jöns, H. i Hoyler, M. (2013). Global geographies of higher education: The perspective of world university rankings. Geoforum, 46, pp.45-59, Elsevier. doi.org/10.1016/j.geoforum.2012.12.014,
38. Kacprzyk, J. (2001). Wieloetapowe sterowanie rozmyte. Wydawnictwo Naukowo-Techniczne,
39. Kacprzyk, J. i Wilbik, A. (2009). Using fuzzy linguistic summaries for the comparison of time series: an application to the analysis of investment fund quotations. IFSA-EUSFLAT Conference, 1321–1326. eusflat.org/materials/proceeding-eusflat-2009.pdf,
40. Kandimalla, B., Rohatgi, S., Wu J. i Giles, C. (2021). Large Scale Subject Category Classification of Scholarly Papers With Deep Attentive Neural Networks. Frontiers in Research Metrics and Analytics. 5:600382, Frontiers Media. doi.org/10.3389/frma.2020.600382,
41. Klement, E. P., Mesiar, R. i Pap, E. (2000). Triangular Norms. Trends in Logic, Springer. doi.org/10.1007/978-94-015-9540-7,
42. Kut, P. i Pietrucha-Urbanik, K. (2024). Bibliometric Analysis of Renewable Energy Research on the Example of the Two European Countries: Insights, Challenges, and Future Prospects. Energies, 17, 176, MDPI. doi.org/10.3390/en17010176,
43. Kwiek, M. i Roszka, W. (2022a). Academic vs. biological age in research on academic careers: a large-scale study with implications for scientifically developing systems. Scientometrics, 127(6), 3543–3575, Springer. doi.org/10.1007/s11192-022-04363-0,
44. Kwiek, M. i Roszka, W. (2022b). Are female scientists less inclined to publish alone? The gender solo research gap. Scientometrics, 127(4), 1697–1735, Springer. doi.org/10.1007/s11192-022-04308-7,
45. Kwiek, M. i Roszka, W. (2023a). The young and the old, the fast and the slow: a large-scale study of productivity classes and rank advancement. Studies in Higher Education, 1–16, Routledge. doi.org/10.1080/03075079.20,
46. Kwiek, M. i Roszka, W. (2023b). Once highly productive, forever highly productive? Full professors' research productivity from a longitudinal perspective. Higher Education, 87(3), 519–549, Springer. doi.org/10.1007/s10734-023-01022-y23.2288172,

47. Kwiek, M. i Roszka, W. (2024). Top Research Performance Over Time: A Multidimensional Micro-Data Approach. ArXiv. doi.org/10.31235/osf.io/uyzrx,
48. Kwiek, M. i Szymula, L. (2023a). Young male and female scientists: A quantitative exploratory study of the changing demographics of the global scientific workforce. Quantitative Science Studies, 4(4), 902–937, MIT Press. doi.org/10.1162 /qss_a_00276,
49. Kwiek, M. i Szymula, L. (2023b). Profesja akademicka w ujęciu globalnym: co Big Data mówią nam o udziale kobiet w nauce? Nr 2, Nauka, Polska Akademia Nauk. doi.org/10.24425/nauka.2023.146057,
50. Kwiek, M. i Szymula, L. (2024). Znikający naukowcy. Co ustrukturyzowane Big Data mówią nam o rezygnacji z nauki w 38 krajach OECD? Nr 1, Nauka, Polska Akademia Nauk. doi.org/ 10.24425/nauka.2024.149837,
51. Lee, S. i Bozeman, B. (2005). The Impact of Research Collaboration on Scientific Productivity. Social Studies of Science, 35(5), 673-702, SAGE Publications. doi.org/10.1177/0306312705052359,
52. Leekwijck, W. V. i Kerre, E. E. (1999). Defuzzification: criteria and classification. Fuzzy Sets and Systems, 108(2), 159–178, Elsevier. doi.org/10.1016/s0165-0114(97)00337-0,
53. Leydesdorff, L. i Milojević, S. (2015). Scientometrics. International Encyclopedia of the Social & Behavioral Sciences, 322–327, Elsevier. doi.org/10.1016/b978-0-08-097086 -8.85030-8,
54. Makabate, C., Musonda, I., Okoro, C.S. i Chileshe, N. (2022). Scientometric analysis of BIM adoption by SMEs in the architecture, construction and engineering sector, [w:] Engineering, Construction and Architectural Management, 29(1), 179-203, Emerald Publishing. doi.org/10.1108/ECAM-02-2020-0139,
55. Mamdani, E. (1976). Advances in the linguistic synthesis of fuzzy controllers. International Journal of Man-Machine Studies, 8(6), 669–678, Elsevier. doi.org/10.1016/s0020-7373(76)80028-4,
56. Marzouqi, A., Alameddine, M., Sharif, A. i Alsheikh-Ali, A. (2019). Research productivity in the United Arab Emirates: A 20-year bibliometric analysis. Heliyon, 5(12), Elsevier. doi.org/10.1016/j.heliyon.2019.e02819,

57. Meen, K., Seojin, N., Fei, W. i Yongjun, Z. (2020). Mapping scientific landscapes in UMLS research: a scientometric review, Journal of the American Medical Informatics Association, 27(10), 1612–1624, Oxford Academic. doi.org/10.1093/jamia/ocaa107,
58. Menger, K. (1942). Statistical Metrics. Proceedings of the National Academy of Sciences, 28(12), 535–537, PNAS. doi.org/10.1073/pnas.28.12.535,
59. Milojević, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. Journal of Informetrics,, 7(4), 767–773, Elsevier. doi.org/10.1016/j.joi.2013.06.006,
60. Moed, H. (2005). Citation analysis in research evaluation. Springer, doi.org/10.1007/1-4020-3714-7,
61. Moed, H., Glänzel, W. i Schmoch, U. (2005). Handbook of Quantitative Science and Technology Research. Springer. doi.org/10.1007/1-4020-2755-9,
62. Murphy, K. (2012). Machine Learning: a Probabilistic Perspective, MIT Press.
63. Narayan, A. I., Chogtu, B., Janodia, M. D. i Venkata, S. K. (2021). A bibliometric study on the research outcome of Brazil, Russia, India, China, and South Africa. F1000Research, 10, 213. doi.org/10.12688/f1000research.51337.1,
64. Nowicki R. (2009). Rozmyte systemy decyzyjne w zadaniach z ograniczoną wiedzą. Problemy Współczesnej Nauki i Techniki : teoria i zastosowania. Informatyka. Akademicka Oficyna Wydawnicza Exit,
65. On, B.-W., Lee, D., Kang, J. i Mitra, P. (2005). Comparative study of name disambiguation problem using a scalable blocking-based framework. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL ’05. doi.org/10.1145/1065385.1065463,
66. OpenAlex (2024). Topics. dostęp 2024, help.openalex.org/how-it-works/topics
67. OpenAlex (2024), OpenAlex technical documentation. dostęp 2024, docs.openalex.org,
68. Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. ArXiv. arxiv.org/abs/2205.01833,
69. Purnomo, A., Rosyidah, E., Firdaus, M., Asitah, N. i Septiano, A. (2020) Data Science Publication: Thirty-Six Years Lesson of Scientometric Review. Proceedings of the 2020 International Conference on Information Management and Technology (ICIMTech), Bandung, Indonesia,

70. Qua, K. i Pelfrey, C. M. (2020). Using bibliometrics to evaluate translational science training: evidence for early career success of KL2 scholars. Journal of Clinical and Translational Science, 5(1), Cambride University Press. doi.org/10.1017/cts.2020.516,
71. Rojas, I., Valenzuela, O., Anguita, M. i Prieto, A. (1998). Analysis of the operators involved in the definition of the implication functions and in the fuzzy inference process. International Journal of Approximate Reasoning, 19(3–4), 367–389, Elsevier. doi.org/10.1016/s0888-613x(98)10016-6,
72. Ross, T. J. (2010). Fuzzy Logic with Engineering Applications, John Wiley & Sons, Ltd. doi.org/10.1002/9781119994374,
73. Sadeghian, A., Mendel, J. M. i Tahayori, H. (Eds.). (2013). Advances in Type-2 Fuzzy Sets and Systems. Theory and Applications. Studies in Fuzziness and Soft Computing, Springer. doi.org/10.1007/978-1-4614-6666-6,
74. Sambariya, D. K. i Prasad, R. (2016). Selection of Membership Functions Based on Fuzzy Rules to Design an Efficient Power System Stabilizer. International Journal of Fuzzy Systems, 19(3), 813–828, Springer. doi.org/10.1007/s40815-016-0197-6,
75. Schnell, J. D. (2017). Web of Science: The First Citation Index for Data Analytics and Scientometrics. Dans F. J. Cantú-Ortiz (dir.), Research Analytics: Boosting University Productivity and Competitiveness Through Scientometrics (1re éd., p. 15-30). Auerbach Publications. doi.org/10.1201/9781315155890-2,
76. Schweizer, B. i Sklar, A. (1960). Statistical metric spaces. Pacific Journal of Mathematics, 10(1), 313–334. doi.org/10.2140/pjm.1960.10.313,
77. Sivanandam, S. N., Sumathi, S. i Deepa, S. N. (2007). Introduction to Fuzzy Logic using MATLAB. Springer. doi.org/10.1007/978-3-540-35781-0,
78. Sugimoto, C. R., & Larivière, V. (2017). Measuring Research: What Everyone Needs to know. Oxford Univeristy Press,
79. Sweileh, W.M. (2020). Bibliometric analysis of scientific publications on “sustainable development goals” with emphasis on “good health and well-being” goal (2015–2019). Global Health 16, 68, Springer. doi.org/10.1186/s12992-020-00602-2,
80. Szulczyński, B., Gębicki, J. i Namieśnik, J. (2018). Application of fuzzy logic to determine the odour intensity of model gas mixtures using electronic nose. E3S Web of Conferences, 28, 01036. doi.org/10.1051/e3sconf/20182801036,

81. Trillas, E. i Valverde, L. (1993). ON IMPLICATION AND INDISTINGUISHABILITY IN THE SETTING OF FUZZY LOGIC**la fond remembrance of Xarier Domingo. Readings in Fuzzy Sets for Intelligent Systems, 97–104, Morgan Kaufmann Publishers, Inc. doi.org/10.1016/b978-1-4832-1450-4.50013-4,
82. Visser, M., van Eck, N. i Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Sci-ence, Dimensions, Crossref, and Microsoft Academic. Quantitative Science Studies, 2(1), pp.20-41, MIT Press. doi.org/10.1162/qss_a_00112,
83. Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., Visser, M. S. i Van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. Journal of Informetrics, 5(1), 37–47, Elsevier. doi.org/10.1016/j.joi.2010.08.001,
84. Witten, I. H., Frank, E. i Hall, M. (2011). Data Mining: practical machine learning tools and techniques, Elsevier. doi.org/10.1016/c2009-0-19715-5
85. Wygralak, M. (2012). Cardinalities of Fuzzy Sets. Studies in Fuzziness and Soft Computing, 118, Springer. doi.org/10.1007/978-3-540-36382-8,
86. Wygralak, M. (2013). Intelligent Counting Under Information Imprecision. Applications to Intelligent Systems and Decision Support. Studies in Fuzziness and Soft Computing, 292, Springer. doi.org/10.1007/978-3-642-34685-9,
87. Yager, R. (1988). On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Transactions on Systems, Man, and Cybernetics, 18(1), 183-190, IEE. doi.org/ 10.1109/21.87068,
88. Yager, R. R., i Kacprzyk, J. (2012). The Ordered Weighted Averaging Operators. Theory and Applications. Springer. doi.org/10.1007/978-1-4615-6123-1,
89. Ye, F. (2017). Scientific metrics: towards analytical and quantitative sciences. Understanding complex systems, Springer. doi.org/10.1007/978-981-10-5936-0,
90. Zadeh, L. (1965). Fuzzy sets. Information and Control, 8(3), 338–353, Elsevier. doi.org/10.1016/s0019-9958(65)90241-x,
91. Zadeh, L. (1972). A Rationale for Fuzzy Control. Journal of Dynamic Systems, Measurement, and Control, 94(1), 3–4, ASME. doi.org/10.1115/1.3426540,
92. Zadeh, L. (1975a). The concept of a linguistic variable and its application to approximate reasoning-I. Information Sciences, 8(3), 199–249, Elsevier. doi.org/10.1016/0020-0255(75)90036-5,

93. Zadeh, L. (1975b). The concept of a linguistic variable and its application to approximate reasoning-II. Information Sciences, 8(4), 301–357, Elsevier. doi.org/10.1016/0020-0255(75)90046-8,
94. Zadeh, L. (1975c). The concept of a linguistic variable and its application to approximate reasoning-III. Information Sciences, 9(1), 43–80, Elsevier. doi.org/10.1016/0020-0255(75)90017-1,
95. Zadeh, L. (1977). Theory of Fuzzy Sets, Memo UCB/ERL M77/1. University of California, Berkeley,
96. Żywica, P., Dyczkowski, K., Wójtowicz, A., Stachowiak, A., Szubert, S. i Moszyński, R. (2016). Development of a fuzzy-driven system for ovarian tumor diagnosis. Biocybernetics and Biomedical Engineering, 36(4), 632–643, Elsevier. doi.org/10.1016/j.bbe.2016.08.003.

Załączniki

1. Załączniki do rozdziału czwartego

Tabela 1. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej Cytowanie

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niskie											
Zliczanie przez łączenie *p*=0.5	-	92.0	92.4	92.0	99.8	92.2	92.1	8.0	2.76	90.67	130.93
Zliczanie przez łączenie *p*=1	-	91.6	92.1	91.6	99.8	91.8	91.6	8.4	2.77	90.66	130.92
Zliczanie przez łączenie *p*=2	-	88.9	89.7	88.9	99.6	89.2	89.1	11.1	2.78	90.66	130.91
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	92.5	93.0	92.5	99.8	92.7	92.5	7.5	3.66	89.60	129.38
	0.4	90.3	91.0	90.3	99.7	90.6	90.4	9.7	5.20	87.78	126.75
	0.6	85.3	86.3	85.3	99.5	85.7	85.7	14.7	8.20	84.22	121.61
	0.8	74.2	75.8	74.2	98.9	74.9	75.4	25.8	15.71	75.50	109.02
Zliczanie przez progowanie i łączenie *p*=1	0.2	92.3	92.8	92.3	99.7	92.5	92.3	7.7	3.66	89.60	129.38
	0.4	90.2	90.9	90.2	99.7	90.5	90.3	9.8	5.20	87.77	126.74
	0.6	85.3	86.3	85.3	99.5	85.7	85.7	14.7	8.20	84.22	121.61
	0.8	74.2	75.8	74.2	98.9	74.9	75.4	25.8	15.71	75.50	109.02
Zliczanie przez progowanie i łączenie *p*=2	0.2	88.9	89.6	88.9	99.6	89.2	89.1	11.1	3.67	89.59	129.36
	0.4	88.7	89.5	88.7	99.6	89.0	88.8	11.3	5.21	87.77	126.73
	0.6	85.2	86.2	85.2	99.5	85.6	85.5	14.8	8.21	84.21	121.60
	0.8	74.2	75.8	74.2	98.9	74.9	75.4	25.8	15.71	75.50	109.02
Zliczanie przez wyostrzenie *p*=0.5	0.4	93.6	93.9	93.6	99.8	93.7	93.6	6.4	3.66	70.55	101.87
	0.6	93.4	93.7	93.4	99.8	93.5	93.4	6.6	3.36	73.13	105.60
	0.8	92.2	92.6	92.2	99.8	92.3	92.2	7.8	3.06	77.84	112.40
Zliczanie przez wyostrzenie *p*=1	0.4	93.4	93.7	93.4	99.8	93.6	93.4	6.6	3.68	70.53	101.84
	0.6	91.9	92.4	91.9	99.8	92.1	92.0	8.1	3.42	73.08	105.52
	0.8	91.7	92.2	91.7	99.8	91.9	91.7	8.3	3.10	77.80	112.35
Zliczanie przez wyostrzenie *p*=2	0.4	93.3	93.7	93.3	99.8	93.5	93.3	6.7	3.68	70.53	101.84
	0.6	91.5	92.1	91.5	99.7	91.8	91.6	8.5	3.43	73.06	105.50
	0.8	89.7	90.4	89.7	99.7	90.0	89.8	10.3	3.15	77.75	112.27
Wyostrzenie kontrastu	-	90.6	91.2	90.6	99.7	90.9	90.7	9.4	2.77	90.67	130.92
term = ani niskie ani wysokie											
Zliczanie przez łączenie *p*=0.5	-	89.5	89.9	89.5	99.3	89.7	89.3	10.5	22.90	77.07	111.29
Zliczanie przez łączenie *p*=1	-	82.3	82.8	82.3	98.8	82.5	82.2	17.7	22.94	77.04	111.24
Zliczanie przez łączenie *p*=2	-	64.2	65.4	64.2	97.1	64.8	65.0	35.8	22.91	77.06	111.28
Zliczanie przez wyostrzenie *p*=0.5	0.2	84.7	85.0	84.7	99.1	84.8	84.6	15.3	23.93	70.70	102.08
	0.4	83.5	84.0	83.5	99.1	83.7	83.6	16.5	23.11	75.91	109.62
Zliczanie przez wyostrzenie *p*=1	0.2	82.6	83.1	82.6	98.9	82.8	82.6	17.4	23.99	70.65	102.02
	0.4	80.5	81.2	80.5	98.6	80.8	80.4	19.5	23.25	75.80	109.46
Zliczanie przez wyostrzenie *p*=2	0.2	81.0	81.6	81.0	98.7	81.3	81.1	19.0	23.98	70.66	102.03
	0.4	60.7	62.2	60.7	96.9	61.4	61.8	39.3	23.26	75.80	109.45
Wyostrzenie kontrastu	-	64.2	65.4	64.2	97.1	64.8	65.0	35.8	22.91	77.06	111.28
term = wysokie											
Zliczanie przez łączenie *p*=0.5	-	86.9	87.1	86.9	99.2	86.9	86.7	13.1	20.12	78.81	113.80
Zliczanie przez łączenie *p*=1	-	67.1	66.7	67.1	97.3	66.9	67.5	32.9	20.18	78.76	113.73
Zliczanie przez łączenie *p*=2	-	49.8	48.5	49.8	96.2	49.1	51.9	50.2	20.19	78.75	113.72
Zliczanie przez wyostrzenie *p*=0.5	0.2	85.9	86.2	85.9	99.2	86.0	85.9	14.1	20.66	71.40	103.11
	0.4	86.6	86.9	86.6	99.2	86.7	86.5	13.4	20.35	75.54	109.08

	0.6	86.8	87.0	86.8	99.2	86.9	86.6	13.2	20.22	77.31	111.64
	0.8	86.9	87.1	86.9	99.2	86.9	86.7	13.1	20.15	78.27	113.02
Zliczanie przez wyostrzenie *p*=1	0.2	83.9	84.1	83.9	99.0	84.0	83.9	16.1	20.71	71.36	103.05
	0.4	66.4	66.2	66.4	97.4	66.3	67.1	33.6	20.46	75.45	108.94
	0.6	67.1	66.7	67.1	97.3	66.9	67.6	32.9	20.33	77.23	111.51
	0.8	67.1	66.7	67.1	97.3	66.9	67.6	32.9	20.24	78.19	112.91
Zliczanie przez wyostrzenie *p*=2	0.2	68.3	68.5	68.3	97.5	68.4	68.9	31.7	20.70	71.37	103.06
	0.4	63.3	63.1	63.3	97.0	63.1	64.1	36.7	20.46	75.45	108.95
	0.6	50.0	48.8	50.0	96.3	49.4	52.2	50.0	20.34	77.22	111.50
	0.8	49.9	48.6	49.9	96.2	49.2	52.0	50.1	20.26	78.18	112.89
Wyostrzenie kontrastu	-	50.0	48.9	50.0	96.3	49.4	52.3	50.0	20.17	78.77	113.75

Tabela 2. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej FWCI 4-letnie

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niskie											
Zliczanie przez łączenie *p*=0.5	-	92.9	93.3	92.9	99.8	93.1	92.9	7.1	3.48	86.73	125.24
Zliczanie przez łączenie *p*=1	-	91.7	92.2	91.7	99.7	91.9	91.7	8.3	3.49	86.72	125.22
Zliczanie przez łączenie *p*=2	-	89.9	90.7	89.9	99.6	90.3	90.0	10.1	3.52	86.70	125.19
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	90.8	91.3	90.8	99.7	91.0	90.9	9.2	4.78	85.14	122.93
	0.4	84.6	85.4	84.6	99.6	85.0	85.1	15.4	7.06	82.34	118.90
	0.6	84.0	85.3	84.0	99.4	84.6	84.5	16.0	11.34	77.06	111.27
Zliczanie przez progowanie i łączenie *p*=1	0.2	90.6	91.2	90.6	99.7	90.8	90.6	9.4	4.79	85.13	122.92
	0.4	84.5	85.4	84.5	99.5	84.9	85.0	15.5	7.06	82.34	118.89
	0.6	84.0	85.3	84.0	99.4	84.6	84.4	16.0	11.34	77.06	111.27
Zliczanie przez progowanie i łączenie *p*=2	0.2	89.9	90.6	89.9	99.6	90.2	90.0	10.1	4.81	85.10	122.89
	0.4	84.2	85.1	84.2	99.5	84.6	84.7	15.8	7.08	82.32	118.86
	0.6	84.7	85.9	84.7	99.4	85.3	85.1	15.3	11.35	77.05	111.26
Zliczanie przez wyostrzenie *p*=0.5	0.4	96.5	96.6	96.5	99.9	96.5	96.4	3.5	4.48	70.56	101.89
	0.6	94.3	94.6	94.3	99.8	94.4	94.3	5.7	4.10	73.98	106.82
	0.8	93.1	93.4	93.1	99.8	93.2	93.1	6.9	3.74	79.52	114.82
Zliczanie przez wyostrzenie *p*=1	0.4	94.5	94.8	94.5	99.8	94.6	94.5	5.5	4.51	70.54	101.85
	0.6	92.9	93.2	92.9	99.8	93.1	92.9	7.1	4.15	73.93	106.76
	0.8	92.6	93.0	92.6	99.7	92.8	92.6	7.4	3.77	79.48	114.77
Zliczanie przez wyostrzenie *p*=2	0.4	92.6	92.9	92.6	99.8	92.7	92.6	7.4	4.51	70.53	101.85
	0.6	91.6	92.1	91.6	99.7	91.8	91.6	8.4	4.19	73.90	106.71
	0.8	90.1	90.8	90.1	99.6	90.4	90.1	9.9	3.84	79.42	114.68
Wyostrzenie kontrastu	-	90.5	91.1	90.5	99.7	90.8	90.6	9.5	3.49	86.72	125.22

Tabela 3. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej FWCI 5-letnie

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niskie											
Zliczanie przez łączenie *p*=0.5	-	92.9	93.3	92.9	99.8	93.1	92.9	7.1	3.47	87.57	126.45
Zliczanie przez łączenie *p*=1	-	91.6	92.2	91.6	99.7	91.9	91.7	8.4	3.48	87.56	126.44
Zliczanie przez łączenie *p*=2	-	89.2	90.0	89.2	99.6	89.6	89.3	10.8	3.50	87.54	126.41
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	90.7	91.3	90.7	99.7	91.0	90.9	9.3	4.77	85.98	124.15
	0.4	84.6	85.4	84.6	99.6	85.0	85.1	15.4	6.99	83.23	120.18
	0.6	84.7	86.0	84.7	99.4	85.3	85.1	15.3	11.33	77.90	112.49

Zliczanie przez progowanie i łączenie p=1	0.2	90.5	91.1	90.5	99.7	90.8	90.6	9.5	4.77	85.97	124.14
	0.4	84.5	85.4	84.5	99.5	84.9	85.0	15.5	6.99	83.22	120.17
	0.6	84.7	86.0	84.7	99.4	85.3	85.1	15.3	11.33	77.90	112.49
Zliczanie przez progowanie i łączenie p=2	0.2	85.7	86.4	85.7	99.5	86.0	86.0	14.3	4.80	85.95	124.11
	0.4	84.2	85.1	84.2	99.5	84.6	84.7	15.8	7.01	83.21	120.15
	0.6	84.6	85.9	84.6	99.4	85.2	85.0	15.4	11.34	77.89	112.48
Zliczanie przez wyostrzenie p=0.5	0.4	96.5	96.6	96.5	99.9	96.5	96.4	3.5	4.49	70.55	101.87
	0.6	96.2	96.3	96.2	99.8	96.3	96.1	3.8	4.11	74.00	106.85
	0.8	93.1	93.4	93.1	99.8	93.2	93.1	6.9	3.74	79.64	115.01
Zliczanie przez wyostrzenie p=1	0.4	96.3	96.4	96.3	99.8	96.3	96.2	3.7	4.52	70.53	101.84
	0.6	92.9	93.2	92.9	99.8	93.1	92.9	7.1	4.16	73.95	106.79
	0.8	93.6	93.9	93.6	99.8	93.8	93.5	6.4	3.77	79.61	114.96
Zliczanie przez wyostrzenie p=2	0.4	96.2	96.3	96.2	99.8	96.2	96.1	3.8	4.52	70.52	101.83
	0.6	92.4	92.9	92.4	99.7	92.6	92.4	7.6	4.20	73.92	106.74
	0.8	90.1	90.8	90.1	99.6	90.4	90.1	9.9	3.85	79.55	114.86
Wyostrzenie kontrastu	-	90.5	91.1	90.5	99.7	90.8	90.6	9.5	3.48	87.56	126.44
term = wysokie											
Zliczanie przez łączenie p=0.5	-	94.0	94.0	94.0	99.4	94.0	93.5	6.0	28.90	70.05	101.15
Zliczanie przez łączenie p=1	-	87.7	87.8	87.7	98.9	87.7	87.2	12.3	28.93	70.03	101.11
Zliczanie przez łączenie p=2	-	64.7	64.5	64.7	97.2	64.5	65.5	35.3	28.96	70.00	101.08
Wyostrzenie kontrastu	-	72.8	72.6	72.8	97.6	72.7	72.7	27.2	28.93	70.02	101.11

Tabela 4. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej FWCI bez ram czasowych

Funkcja wagowa	t	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niskie											
Zliczanie przez łączenie p=0.5	-	95.8	96.0	95.8	99.8	95.9	95.7	4.2	3.51	90.26	130.33
Zliczanie przez łączenie p=1	-	90.8	91.4	90.8	99.7	91.1	90.9	9.2	3.52	90.25	130.32
Zliczanie przez łączenie p=2	-	86.9	87.5	86.9	99.6	87.2	87.2	13.1	3.53	90.24	130.30
Zliczanie przez progowanie i łączenie p=0.5	0.2	87.5	88.0	87.5	99.7	87.7	87.8	12.5	4.76	88.73	128.12
	0.4	85.9	86.8	85.9	99.6	86.3	86.3	14.1	6.88	86.11	124.34
	0.6	84.0	85.5	84.0	99.5	84.7	84.5	16.0	11.04	81.03	117.01
Zliczanie przez progowanie i łączenie p=1	0.2	87.3	87.8	87.3	99.6	87.5	87.6	12.7	4.76	88.72	128.11
	0.4	85.8	86.7	85.8	99.6	86.2	86.2	14.2	6.89	86.10	124.33
	0.6	84.0	85.4	84.0	99.5	84.7	84.4	16.0	11.04	81.03	117.01
Zliczanie przez progowanie i łączenie p=2	0.2	87.6	88.1	87.6	99.6	87.8	87.8	12.4	4.78	88.71	128.09
	0.4	85.5	86.4	85.5	99.5	85.9	85.9	14.5	6.90	86.09	124.31
	0.6	83.9	85.3	83.9	99.4	84.6	84.3	16.1	11.05	81.02	117.00
Zliczanie przez wyostrzenie p=0.5	0.4	96.6	96.7	96.6	99.9	96.6	96.5	3.4	4.61	70.32	101.54
	0.6	96.3	96.4	96.3	99.8	96.3	96.2	3.7	4.24	73.69	106.40
	0.8	96.0	96.2	96.0	99.8	96.1	95.9	4.0	3.85	79.43	114.70
Zliczanie przez wyostrzenie p=1	0.4	96.4	96.5	96.4	99.8	96.4	96.3	3.6	4.64	70.30	101.51
	0.6	95.9	96.1	95.9	99.8	96.0	95.7	4.1	4.28	73.64	106.34
	0.8	93.6	93.9	93.6	99.8	93.8	93.6	6.4	3.88	79.40	114.65
Zliczanie przez wyostrzenie p=2	0.4	96.3	96.4	96.3	99.8	96.3	96.2	3.7	4.64	70.30	101.51
	0.6	92.6	92.9	92.6	99.7	92.8	92.6	7.4	4.32	73.61	106.30
	0.8	86.9	87.6	86.9	99.6	87.2	87.2	13.1	3.95	79.34	114.56
Wyostrzenie kontrastu	-	87.2	87.8	87.2	99.6	87.5	87.5	12.8	3.52	90.25	130.32
term = ani niskie ani wysokie											
Zliczanie przez łączenie p=0.5	-	91.7	91.9	91.7	99.4	91.8	91.4	8.3	23.71	76.47	110.42
Zliczanie przez łączenie p=1	-	88.3	88.6	88.3	99.2	88.4	88.0	11.7	23.71	76.47	110.42
Zliczanie przez łączenie p=2	-	65.1	65.6	65.1	98.0	65.3	66.8	34.9	23.71	76.47	110.42

Zliczanie przez wyostrzenie *p*=0.5	0.4	91.4	91.5	91.4	99.3	91.4	91.0	8.6	24.06	74.29	107.27
Zliczanie przez wyostrzenie *p*=1	0.4	81.2	81.7	81.2	98.9	81.4	81.3	18.8	24.14	74.22	107.17
Zliczanie przez wyostrzenie *p*=2	0.4	64.5	65.1	64.5	97.8	64.7	66.1	35.5	24.18	74.19	107.13
Wyostrzenie kontrastu	-	65.1	65.6	65.1	98.0	65.3	66.8	34.9	23.71	76.47	110.42
term = wysokie											
Zliczanie przez łączenie *p*=0.5	-	92.1	92.1	92.1	99.4	92.1	91.7	7.9	20.06	78.83	113.82
Zliczanie przez łączenie *p*=1	-	84.9	85.0	84.9	98.9	84.9	84.6	15.1	20.09	78.80	113.78
Zliczanie przez łączenie *p*=2	-	69.7	69.6	69.7	97.7	69.6	70.2	30.3	20.12	78.77	113.74
Zliczanie przez wyostrzenie *p*=0.5	0.4	92.0	92.1	92.0	99.4	92.0	91.6	8.0	20.41	74.23	107.18
	0.6	92.1	92.1	92.1	99.4	92.1	91.7	7.9	20.21	76.72	110.78
	0.8	92.1	92.1	92.1	99.4	92.1	91.7	7.9	20.11	78.05	112.70
Zliczanie przez wyostrzenie *p*=1	0.4	85.2	85.2	85.2	98.9	85.2	84.9	14.8	20.48	74.16	107.09
	0.6	85.0	85.0	85.0	98.9	85.0	84.6	15.0	20.28	76.66	110.69
	0.8	85.0	85.0	85.0	98.9	85.0	84.6	15.0	20.16	78.00	112.64
Zliczanie przez wyostrzenie *p*=2	0.4	72.6	72.7	72.6	98.1	72.6	73.1	27.4	20.50	74.15	107.07
	0.6	70.1	70.0	70.1	97.9	70.0	70.7	29.9	20.32	76.63	110.65
	0.8	69.8	69.7	69.8	97.8	69.7	70.4	30.2	20.20	77.97	112.59
Wyostrzenie kontrastu	-	70.2	70.1	70.2	97.9	70.1	70.9	29.8	20.10	78.79	113.77

Tabela 5. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej Percentyl

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niski											
Zliczanie przez łączenie *p*=0.5	0	92.9	93.1	92.9	99.5	93.0	92.7	7.1	4.07	81.70	117.98
Zliczanie przez łączenie *p*=1	0	90.6	90.9	90.6	99.3	90.8	90.2	9.4	4.20	81.59	117.81
Zliczanie przez łączenie *p*=2	0	78.4	78.7	78.4	98.7	78.5	78.8	21.6	4.29	81.57	117.79
Zliczanie przez wyostrzenie *p*=0.5	0.2	92.8	93.0	92.8	99.5	92.9	92.5	7.2	4.36	75.49	109.01
	0.4	92.8	93.0	92.8	99.5	92.9	92.5	7.2	4.19	77.98	112.60
	0.6	92.9	93.1	92.9	99.5	93.0	92.6	7.1	4.12	79.76	115.17
	0.8	92.9	93.1	92.9	99.5	93.0	92.7	7.1	4.08	81.13	117.15
Zliczanie przez wyostrzenie *p*=1	0.2	89.0	89.4	89.0	99.3	89.2	88.8	11.0	4.47	75.41	108.89
	0.4	90.6	90.9	90.6	99.2	90.7	90.2	9.4	4.34	77.85	112.42
	0.6	90.6	90.9	90.6	99.2	90.7	90.2	9.4	4.27	79.62	114.97
	0.8	90.6	90.9	90.6	99.3	90.7	90.2	9.4	4.21	81.01	116.98
Zliczanie przez wyostrzenie *p*=2	0.2	80.0	80.2	80.0	99.0	80.1	80.5	20.0	4.49	75.43	108.92
	0.4	78.9	79.1	78.9	98.8	79.0	79.3	21.1	4.40	77.85	112.42
	0.6	78.5	78.7	78.5	98.8	78.6	78.9	21.5	4.37	79.61	114.95
	0.8	78.4	78.7	78.4	98.7	78.5	78.8	21.6	4.31	81.00	116.96
Wyostrzenie kontrastu	0	78.7	78.9	78.7	98.8	78.8	79.1	21.3	4.27	81.59	117.82
term = ani niski ani wysoki											
Zliczanie przez łączenie *p*=0.5	0	84.4	84.9	84.4	99.4	84.6	84.7	15.6	17.68	72.51	104.71
Zliczanie przez łączenie *p*=1	0	77.2	77.9	77.2	99.1	77.5	78.1	22.8	17.80	72.40	104.55
Zliczanie przez łączenie *p*=2	0	72.1	73.1	72.1	98.7	72.6	73.5	27.9	17.86	72.42	104.58
Zliczanie przez wyostrzenie *p*=0.5	0.4	84.2	84.7	84.2	99.3	84.4	84.5	15.8	17.87	70.87	102.33
Zliczanie przez wyostrzenie *p*=1	0.4	77.8	78.4	77.8	99.0	78.1	78.5	22.2	18.02	70.74	102.14
Zliczanie przez wyostrzenie *p*=2	0.4	71.8	72.7	71.8	98.6	72.2	73.1	28.2	18.09	70.74	102.14
Wyostrzenie kontrastu	0	72.1	73.1	72.1	98.7	72.6	73.5	27.9	17.86	72.42	104.57
term = wysoki											
Zliczanie przez łączenie *p*=0.5	0	84.8	85.1	84.8	99.3	84.9	85.0	15.2	13.43	75.78	109.43
Zliczanie przez łączenie *p*=1	0	84.5	84.8	84.5	99.3	84.6	84.7	15.5	13.46	75.76	109.40
Zliczanie przez łączenie *p*=2	0	82.8	83.2	82.8	99.2	83.0	83.1	17.2	13.50	75.75	109.38

Zliczanie przez progowanie i łączenie *p*=0.5	0.2	86.2	86.6	86.2	99.2	86.4	86.2	13.8	17.80	72.11	104.13
Zliczanie przez progowanie i łączenie *p*=1	0.2	81.3	81.8	81.3	99.1	81.5	81.7	18.7	17.82	72.10	104.11
Zliczanie przez progowanie i łączenie *p*=2	0.2	80.8	81.4	80.8	99.0	81.1	81.2	19.2	17.86	72.08	104.09
Wyostrzenie kontrastu	0	84.3	84.6	84.3	99.2	84.4	84.5	15.7	13.47	75.77	109.41

Tabela 6. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej Rok

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = najstarszy											
Zliczanie przez łączenie *p*=0.5	-	78.8	81.6	78.8	99.5	80.1	80.1	21.2	12.05	84.05	121.37
Zliczanie przez łączenie *p*=1	-	67.7	71.5	67.7	99.1	69.5	70.4	32.3	12.08	84.00	121.29
Zliczanie przez łączenie *p*=2	-	46.8	52.8	46.8	97.4	49.5	51.2	53.2	12.07	84.06	121.38
Zliczanie przez wyostrzenie *p*=0.5	0.2	74.6	77.9	74.6	99.4	76.2	76.3	25.4	12.11	81.63	117.87
	0.4	78.7	81.6	78.7	99.5	80.1	80.0	21.3	12.06	83.82	121.03
	0.6	78.8	81.6	78.8	99.5	80.1	80.0	21.2	12.05	84.00	121.29
	0.8	78.8	81.6	78.8	99.5	80.1	80.1	21.2	12.05	84.04	121.35
Zliczanie przez wyostrzenie *p*=1	0.2	61.8	65.7	61.8	98.8	63.5	65.3	38.2	12.18	81.56	117.77
	0.4	66.8	70.9	66.8	99.1	68.7	69.7	33.2	12.09	83.77	120.96
	0.6	67.7	71.5	67.7	99.1	69.4	70.4	32.3	12.08	83.95	121.22
	0.8	67.7	71.5	67.7	99.1	69.5	70.4	32.3	12.08	83.99	121.27
Zliczanie przez wyostrzenie *p*=2	0.2	45.3	52.1	45.3	97.3	48.5	49.9	54.7	12.20	81.59	117.82
	0.4	45.8	52.2	45.8	97.4	48.7	50.3	54.2	12.09	83.82	121.03
	0.6	46.7	52.8	46.7	97.4	49.5	51.1	53.3	12.07	84.00	121.30
	0.8	46.8	52.8	46.8	97.4	49.5	51.2	53.2	12.07	84.05	121.37
Wyostrzenie kontrastu	-	46.8	52.8	46.8	97.4	49.5	51.2	53.2	12.07	84.05	121.36
term = ani najstarszy ani najnowszy											
Zliczanie przez łączenie *p*=0.5	-	78.9	81.6	78.9	99.5	80.2	80.1	21.1	12.06	84.04	121.35
Zliczanie przez łączenie *p*=1	-	67.8	71.5	67.8	99.1	69.5	70.4	32.2	12.08	84.00	121.29
Zliczanie przez łączenie *p*=2	-	46.8	52.8	46.8	97.4	49.6	51.2	53.2	12.07	84.06	121.38
Zliczanie przez wyostrzenie *p*=0.5	0.2	74.6	77.9	74.6	99.4	76.2	76.3	25.4	12.11	81.66	117.92
	0.4	78.7	81.6	78.7	99.5	80.1	80.0	21.3	12.06	83.92	121.17
Zliczanie przez wyostrzenie *p*=1	0.2	61.8	65.7	61.8	98.8	63.5	65.3	38.2	12.18	81.59	117.82
	0.4	66.9	70.9	66.9	99.1	68.7	69.7	33.1	12.09	83.87	121.10
Zliczanie przez wyostrzenie *p*=2	0.2	45.3	52.1	45.3	97.3	48.5	49.9	54.7	12.20	81.62	117.86
	0.4	45.8	52.2	45.8	97.4	48.8	50.4	54.2	12.08	83.93	121.19
Wyostrzenie kontrastu	-	46.8	52.8	46.8	97.4	49.6	51.2	53.2	12.07	84.07	121.39
term = najnowszy											
Zliczanie przez łączenie *p*=0.5	-	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.00	91.81	132.58
Zliczanie przez łączenie *p*=1	-	99.9	99.9	99.9	100.0	99.9	99.9	0.1	0.00	91.80	132.55
Zliczanie przez łączenie *p*=2	-	98.4	98.4	98.4	99.9	98.4	98.3	1.6	0.00	91.83	132.60
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.07	91.77	132.51
	0.4	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.34	91.55	132.19
	0.6	98.5	98.5	98.5	99.9	98.5	98.4	1.5	1.35	90.68	130.94
	0.8	90.1	90.0	90.1	99.5	90.1	90.0	9.9	13.10	80.68	116.50
Zliczanie przez progowanie i łączenie *p*=1	0.2	99.9	99.9	99.9	100.0	99.9	99.9	0.1	0.07	91.75	132.48
	0.4	98.5	98.5	98.5	100.0	98.5	98.5	1.5	0.34	91.53	132.17
	0.6	98.4	98.4	98.4	99.9	98.4	98.4	1.6	1.35	90.66	130.91
	0.8	89.4	89.2	89.4	99.4	89.3	89.3	10.6	13.10	80.67	116.48
	0.2	98.4	98.4	98.4	99.9	98.4	98.3	1.6	0.07	91.77	132.52

Zliczanie przez progowanie i łączenie *p*=2	0.4	98.4	98.4	98.4	99.9	98.4	98.3	1.6	0.34	91.54	132.19
	0.6	98.3	98.3	98.3	99.9	98.3	98.2	1.7	1.35	90.68	130.94
	0.8	89.4	89.1	89.4	99.4	89.2	89.2	10.6	13.10	80.69	116.51
Zliczanie przez wyostrzenie *p*=0.5	0.2	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.00	69.29	100.05
	0.4	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.00	69.39	100.19
	0.6	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.00	69.81	100.80
	0.8	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.00	73.19	105.68
Zliczanie przez wyostrzenie *p*=1	0.2	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.00	69.29	100.05
	0.4	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.00	69.39	100.19
	0.6	100.0	100.0	100.0	100.0	100.0	99.9	0.0	0.01	69.80	100.80
	0.8	99.9	99.9	99.9	100.0	99.9	99.9	0.1	0.00	73.18	105.68
Zliczanie przez wyostrzenie *p*=2	0.2	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.00	69.29	100.05
	0.4	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.00	69.39	100.19
	0.6	99.9	99.9	99.9	100.0	99.9	99.9	0.1	0.01	69.80	100.79
	0.8	98.3	98.3	98.3	99.9	98.3	98.2	1.7	0.02	73.18	105.67
Wyostrzenie kontrastu	-	99.9	99.9	99.9	100.0	99.9	99.9	0.1	0.00	91.82	132.58

Tabela 7. Wyniki metryk ewaluacyjnych dla zmiennej lingwistycznej Zespół

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = najmniejszy											
Zliczanie przez łączenie *p*=0.5	-	84.8	85.8	84.8	99.3	85.2	85.0	15.2	14.32	77.26	111.56
Zliczanie przez łączenie *p*=1	-	59.5	61.6	59.5	97.3	60.5	61.3	40.5	14.57	76.98	111.15
Zliczanie przez łączenie *p*=2	-	51.7	54.1	51.7	96.8	52.8	54.4	48.3	14.53	77.06	111.28
Zliczanie przez progowanie i łączenie p=0.5	0.2	70.1	70.7	70.1	98.5	70.3	71.6	29.9	22.22	69.28	100.04
Zliczanie przez wyostrzenie *p*=0.5	0.4	84.5	85.6	84.5	99.3	84.9	84.8	15.5	14.82	70.48	101.78
	0.6	84.6	85.6	84.6	99.3	85.0	84.8	15.4	14.41	75.36	108.82
	0.8	84.8	85.8	84.8	99.3	85.2	85.0	15.2	14.32	77.25	111.55
Zliczanie przez wyostrzenie *p*=1	0.4	59.0	61.6	59.0	97.2	60.2	60.7	41.0	15.04	70.30	101.51
	0.6	59.2	61.4	59.2	97.2	60.3	61.0	40.8	14.91	74.92	108.18
	0.8	59.5	61.6	59.5	97.3	60.5	61.3	40.5	14.58	76.96	111.14
Zliczanie przez wyostrzenie *p*=2	0.4	58.2	61.3	58.2	97.0	59.7	60.0	41.8	15.01	70.34	101.57
	0.6	53.6	56.5	53.6	96.9	55.0	56.1	46.4	14.80	75.05	108.37
	0.8	51.6	54.1	51.6	96.8	52.8	54.3	48.4	14.53	77.05	111.26
Wyostrzenie kontrastu	-	51.1	53.5	51.1	96.9	52.3	54.0	48.9	14.52	77.06	111.28
term = ani najmniejszy ani największy											
Zliczanie przez łączenie *p*=0.5	-	84.9	86.0	84.9	99.4	85.4	85.2	15.1	14.35	77.23	111.52
Zliczanie przez łączenie *p*=1	-	84.3	85.5	84.3	99.3	84.8	84.5	15.7	14.58	76.96	111.13
Zliczanie przez łączenie *p*=2	-	58.7	61.2	58.7	97.2	59.9	60.6	41.3	14.54	77.05	111.26
Zliczanie przez wyostrzenie *p*=0.5	0.4	84.4	85.4	84.4	99.3	84.8	84.6	15.6	14.70	73.47	106.10
Zliczanie przez wyostrzenie *p*=1	0.4	58.6	61.4	58.6	97.1	59.9	60.3	41.4	15.00	73.25	105.77
Zliczanie przez wyostrzenie *p*=2	0.4	57.5	60.9	57.5	96.9	59.1	59.2	42.5	14.96	73.30	105.85
Wyostrzenie kontrastu	-	58.7	61.2	58.7	97.2	59.9	60.6	41.3	14.54	77.05	111.26
term = największy											
Zliczanie przez progowanie	0.2	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.01	69.27	100.03
Zliczanie przez łączenie *p*=0.5	-	99.7	99.7	99.7	100.0	99.7	99.7	0.3	0.00	87.28	126.03
Zliczanie przez łączenie *p*=1	-	96.9	96.9	96.9	99.9	96.9	96.8	3.1	0.14	87.07	125.73
Zliczanie przez łączenie *p*=2	-	93.1	93.1	93.1	99.6	93.1	92.9	6.9	0.21	87.07	125.73
Zliczanie przez progowanie	0.2	99.7	99.7	99.7	100.0	99.7	99.7	0.3	0.01	87.29	126.04
i łączenie *p*=0.5	0.4	95.2	95.2	95.2	99.8	95.1	95.0	4.8	10.81	78.70	113.64

Zliczanie przez progowanie i łączenie p=1	0.2	96.9	96.9	96.9	99.9	96.9	96.8	3.1	0.14	87.07	125.73
	0.4	95.4	95.4	95.4	99.7	95.4	95.2	4.6	10.93	78.52	113.38
Zliczanie przez progowanie i łączenie p=2	0.2	93.2	93.1	93.2	99.6	93.1	92.9	6.8	0.21	87.07	125.72
	0.4	87.6	87.7	87.6	99.4	87.6	87.6	12.4	10.98	78.52	113.38
Zliczanie przez wyostrzenie p=0.5	0.2	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.00	69.28	100.04
	0.4	99.8	99.8	99.8	100.0	99.8	99.7	0.2	0.03	72.69	104.96
	0.6	98.3	98.3	98.3	100.0	98.3	98.2	1.7	0.02	78.83	113.83
	0.8	98.3	98.3	98.3	100.0	98.3	98.3	1.7	0.01	84.57	122.12
Zliczanie przez wyostrzenie p=1	0.2	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.00	69.28	100.04
	0.4	97.3	97.3	97.3	99.9	97.3	97.2	2.7	0.09	72.64	104.89
	0.6	96.8	96.9	96.8	99.9	96.8	96.7	3.2	0.36	78.52	113.38
	0.8	96.8	96.8	96.8	99.9	96.8	96.7	3.2	0.23	84.30	121.73
Zliczanie przez wyostrzenie p=2	0.2	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.00	69.28	100.04
	0.4	97.0	97.1	97.0	99.9	97.0	97.0	3.0	0.09	72.64	104.90
	0.6	93.1	93.1	93.1	99.8	93.0	93.0	6.9	0.31	78.58	113.47
	0.8	94.7	94.7	94.7	99.7	94.7	94.4	5.3	0.29	84.31	121.75
Wyostrzenie kontrastu	-	96.4	96.5	96.4	99.8	96.5	96.3	3.6	0.10	87.12	125.81

2. Załączniki do rozdziału piątego

Tabela 8. Wyniki metryk ewaluacyjnych dla kontrolera rozmytego wysokiego prestiżu publikacji ze zmiennymi wejściowymi Cytowanie i Percentyl

Funkcja wagowa	t	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
Zliczanie przez łączenie p=0.5	-	98.9	98.9	98.9	99.7	98.9	98.6	1.1	0.56	95.00	137.18
Zliczanie przez łączenie p=1	-	93.0	92.9	93.0	99.3	92.9	92.5	7.0	0.59	94.98	137.15
Zliczanie przez łączenie p=2	-	84.3	84.1	84.3	98.6	84.2	83.8	15.7	0.63	94.94	137.10
Zliczanie przez progowanie i łączenie p=0.5	0.2	84.3	84.4	84.3	99.0	84.3	84.2	15.7	11.24	81.33	117.43
Zliczanie przez progowanie i łączenie p=1	0.2	83.1	83.1	83.1	98.8	83.0	82.8	16.9	11.26	81.31	117.41
Zliczanie przez progowanie i łączenie p=2	0.2	81.5	81.5	81.5	98.4	81.5	81.1	18.5	11.30	81.28	117.36
Zliczanie przez wyostrzenie p=0.5	0.2	98.8	98.8	98.8	99.8	98.8	98.6	1.2	0.73	79.42	114.69
	0.4	96.8	96.8	96.8	99.6	96.8	96.5	3.2	0.67	86.14	124.38
	0.6	98.7	98.7	98.7	99.7	98.7	98.4	1.3	0.62	90.06	130.05
	0.8	98.9	98.9	98.9	99.7	98.8	98.6	1.1	0.59	92.63	133.75
Zliczanie przez wyostrzenie p=1	0.2	98.4	98.4	98.4	99.7	98.4	98.1	1.6	0.76	79.40	114.65
	0.4	94.4	94.3	94.4	99.3	94.3	93.8	5.6	0.79	86.04	124.24
	0.6	92.9	92.7	92.9	99.3	92.8	92.3	7.1	0.73	89.97	129.91
	0.8	93.0	92.8	93.0	99.3	92.9	92.4	7.0	0.66	92.56	133.66
Zliczanie przez wyostrzenie p=2	0.2	98.1	98.1	98.1	99.6	98.1	97.8	1.9	0.76	79.40	114.65
	0.4	93.3	93.2	93.3	99.1	93.2	92.6	6.7	0.81	86.02	124.21
	0.6	84.8	84.6	84.8	98.7	84.6	84.3	15.2	0.82	89.90	129.82
	0.8	84.4	84.2	84.4	98.6	84.2	83.9	15.6	0.75	92.50	133.56
Wyostrzenie kontrastu	-	91.2	91.1	91.2	98.9	91.1	90.4	8.8	0.60	94.97	137.13

Tabela 9. Wyniki metryk ewaluacyjnych dla kontrolera rozmytego wysokiego prestiżu publikacji ze zmiennymi wejściowymi FWCI 4-letnie i Percentyl.

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
Zliczanie przez łączenie *p*=0.5	-	99.2	99.2	99.2	99.8	99.2	99.0	0.8	1.12	92.84	134.06
Zliczanie przez łączenie *p*=1	-	97.9	97.9	97.9	99.6	97.9	97.4	2.1	1.16	92.80	134.01
Zliczanie przez łączenie *p*=2	-	92.7	92.5	92.7	99.1	92.6	91.9	7.3	1.20	92.78	133.97
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	94.7	94.8	94.7	99.4	94.7	94.2	5.3	9.19	82.71	119.44
Zliczanie przez progowanie i łączenie *p*=1	0.2	94.1	94.1	94.1	99.2	94.1	93.4	5.9	9.21	82.69	119.40
Zliczanie przez progowanie i łączenie *p*=2	0.2	85.7	85.8	85.7	98.8	85.7	85.2	14.3	9.26	82.65	119.35
Zliczanie przez wyostrzenie *p*=0.5	0.2	99.3	99.3	99.3	99.9	99.3	99.2	0.7	1.30	76.21	110.04
	0.4	99.0	99.0	99.0	99.8	99.0	98.8	1.0	1.25	83.68	120.83
	0.6	99.1	99.1	99.1	99.8	99.1	98.9	0.9	1.19	88.22	127.39
	0.8	99.1	99.1	99.1	99.8	99.1	99.0	0.9	1.15	91.00	131.40
Zliczanie przez wyostrzenie *p*=1	0.2	99.1	99.1	99.1	99.8	99.1	98.9	0.9	1.32	76.20	110.02
	0.4	97.9	97.9	97.9	99.6	97.9	97.5	2.1	1.34	83.60	120.72
	0.6	97.7	97.7	97.7	99.5	97.7	97.3	2.3	1.29	88.13	127.26
	0.8	97.8	97.8	97.8	99.5	97.8	97.4	2.2	1.22	90.94	131.32
Zliczanie przez wyostrzenie *p*=2	0.2	99.0	99.0	99.0	99.8	99.0	98.8	1.0	1.32	76.20	110.04
	0.4	95.5	95.5	95.5	99.4	95.5	95.0	4.5	1.36	83.59	120.71
	0.6	94.3	94.2	94.3	99.2	94.3	93.6	5.7	1.36	88.08	127.19
	0.8	94.0	93.9	94.0	99.1	93.9	93.2	6.0	1.29	90.89	131.24
Wyostrzenie kontrastu	-	94.8	94.8	94.8	99.3	94.8	94.2	5.2	1.17	92.79	133.99

Tabela 10. Wyniki metryk ewaluacyjnych dla kontrolera rozmytego wysokiego prestiżu publikacji ze zmiennymi wejściowymi FWCI 5-letnie i Percentyl

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
Zliczanie przez łączenie *p*=0.5	-	99.2	99.2	99.2	99.8	99.2	99.0	0.8	1.06	93.31	134.74
Zliczanie przez łączenie *p*=1	-	97.9	97.9	97.9	99.6	97.9	97.5	2.1	1.09	93.28	134.69
Zliczanie przez łączenie *p*=2	-	92.7	92.6	92.7	99.1	92.6	92.0	7.3	1.14	93.25	134.65
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	94.7	94.8	94.7	99.4	94.7	94.2	5.3	9.04	83.16	120.08
Zliczanie przez progowanie i łączenie *p*=1	0.2	92.5	92.5	92.5	99.2	92.5	91.9	7.5	9.06	83.13	120.04
Zliczanie przez progowanie i łączenie *p*=2	0.2	85.7	85.8	85.7	98.8	85.7	85.2	14.3	9.11	83.10	119.99
Zliczanie przez wyostrzenie *p*=0.5	0.2	99.3	99.3	99.3	99.9	99.3	99.2	0.7	1.23	76.34	110.23
	0.4	99.0	99.0	99.0	99.8	99.0	98.8	1.0	1.18	83.94	121.20
	0.6	99.1	99.1	99.1	99.8	99.1	98.9	0.9	1.12	88.55	127.86
	0.8	99.2	99.2	99.2	99.8	99.2	99.0	0.8	1.09	91.37	131.93
Zliczanie przez wyostrzenie *p*=1	0.2	99.1	99.1	99.1	99.8	99.1	99.0	0.9	1.25	76.32	110.21
	0.4	97.9	97.9	97.9	99.6	97.9	97.5	2.1	1.27	83.86	121.09
	0.6	97.7	97.7	97.7	99.5	97.7	97.3	2.3	1.23	88.46	127.73
	0.8	97.9	97.8	97.9	99.6	97.8	97.4	2.1	1.15	91.31	131.85
Zliczanie przez wyostrzenie *p*=2	0.2	99.0	99.0	99.0	99.8	99.0	98.9	1.0	1.25	76.33	110.22
	0.4	95.5	95.5	95.5	99.4	95.5	95.0	4.5	1.29	83.85	121.07
	0.6	94.3	94.3	94.3	99.2	94.3	93.6	5.7	1.30	88.41	127.65

	0.8	94.0	93.9	94.0	99.1	94.0	93.2	6.0	1.23	91.25	131.77
Wyostrzenie kontrastu	-	93.5	93.4	93.5	99.3	93.5	93.0	6.5	1.11	93.27	134.68

Tabela 11. Wyniki metryk ewaluacyjnych dla kontrolera rozmytego wysokiego prestiżu publikacji ze zmiennymi wejściowymi FWCI bez ram czasowych i Percentyl

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
Zliczanie przez łączenie *p*=0.5	-	99.2	99.2	99.2	99.8	99.2	99.1	0.8	0.76	95.05	137.25
Zliczanie przez łączenie *p*=1	-	96.6	96.5	96.6	99.6	96.6	96.2	3.4	0.79	95.02	137.21
Zliczanie przez łączenie *p*=2	-	92.7	92.6	92.7	99.1	92.6	92.0	7.3	0.83	95.00	137.17
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	94.9	94.9	94.9	99.4	94.9	94.4	5.1	8.83	84.22	121.62
Zliczanie przez progowanie i łączenie *p*=1	0.2	93.4	93.4	93.4	99.3	93.4	92.8	6.6	8.85	84.20	121.59
Zliczanie przez progowanie i łączenie *p*=2	0.2	86.6	86.7	86.6	98.8	86.6	86.1	13.4	8.89	84.17	121.54
Zliczanie przez wyostrzenie *p*=0.5	0.2	99.3	99.3	99.3	99.9	99.3	99.2	0.7	0.91	77.24	111.54
	0.4	99.0	99.0	99.0	99.8	99.0	98.8	1.0	0.88	84.97	122.70
	0.6	99.1	99.1	99.1	99.8	99.1	98.9	0.9	0.83	89.63	129.42
	0.8	99.2	99.2	99.2	99.8	99.2	99.0	0.8	0.80	92.52	133.60
Zliczanie przez wyostrzenie *p*=1	0.2	99.1	99.1	99.1	99.8	99.1	98.9	0.9	0.93	77.23	111.52
	0.4	98.0	97.9	98.0	99.6	97.9	97.6	2.0	0.97	84.89	122.59
	0.6	97.7	97.7	97.7	99.6	97.7	97.3	2.3	0.93	89.54	129.29
	0.8	96.5	96.5	96.5	99.6	96.5	96.1	3.5	0.86	92.46	133.52
Zliczanie przez wyostrzenie *p*=2	0.2	99.0	99.0	99.0	99.8	99.0	98.8	1.0	0.93	77.23	111.52
	0.4	95.5	95.5	95.5	99.4	95.5	95.0	4.5	0.99	84.88	122.56
	0.6	94.3	94.3	94.3	99.2	94.3	93.7	5.7	1.00	89.48	129.21
	0.8	92.8	92.7	92.8	99.2	92.7	92.1	7.2	0.94	92.41	133.43
Wyostrzenie kontrastu	-	93.5	93.4	93.5	99.3	93.5	93.0	6.5	0.80	95.01	137.20

3. Załączniki do rozdziału szóstego

Tabela 12. Wyniki metryk ewaluacyjnych dla agregacji operatorem OWA zmiennej lingwistycznej Cytowani i Percentyl z wektorem wag (0. 1).

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niski											
Zliczanie przez łączenie *p*=0.5	-	89.3	89.9	89.3	99.5	89.6	89.3	10.7	6.21	81.72	118.01
Zliczanie przez łączenie *p*=1	-	87.9	88.6	87.9	99.2	88.2	87.7	12.1	6.30	81.64	117.89
Zliczanie przez łączenie *p*=2	-	77.5	78.2	77.5	98.7	77.8	78.0	22.5	6.38	81.62	117.86
Zliczanie przez wyostrzenie *p*=0.5	0.2	90.1	90.6	90.1	99.5	90.3	90.0	9.9	6.75	74.00	106.86
	0.4	89.3	89.8	89.3	99.4	89.5	89.2	10.7	6.49	76.59	110.60
	0.6	89.3	89.9	89.3	99.5	89.6	89.3	10.7	6.37	78.49	113.34
	0.8	89.4	89.9	89.4	99.5	89.6	89.3	10.6	6.29	80.13	115.70
Zliczanie przez wyostrzenie *p*=1	0.2	88.4	89.0	88.4	99.3	88.7	88.2	11.6	6.85	73.93	106.75
	0.4	88.0	88.6	88.0	99.2	88.2	87.7	12.0	6.62	76.48	110.43
	0.6	87.9	88.6	87.9	99.2	88.2	87.7	12.1	6.51	78.38	113.17
	0.8	87.9	88.6	87.9	99.2	88.2	87.7	12.1	6.41	80.03	115.56
Zliczanie przez wyostrzenie *p*=2	0.2	84.0	84.7	84.0	99.1	84.3	84.1	16.0	6.87	73.94	106.76

	0.4	82.9	83.7	82.9	98.9	83.2	82.9	17.1	6.69	76.46	110.40
	0.6	76.9	77.5	76.9	98.7	77.2	77.5	23.1	6.61	78.34	113.13
	0.8	76.7	77.4	76.7	98.7	77.0	77.3	23.3	6.50	80.00	115.52
Wyostrzenie kontrastu	-	77.0	77.7	77.0	98.8	77.3	77.6	23.0	6.36	81.64	117.88
term = wysoki											
Zliczanie przez łączenie *p*=0.5	-	86.6	86.9	86.6	99.1	86.8	86.4	13.4	26.08	73.55	106.21
Zliczanie przez łączenie *p*=1	-	79.9	80.1	79.9	98.5	79.9	79.8	20.1	26.11	73.52	106.16
Zliczanie przez łączenie *p*=2	-	50.1	48.8	50.1	96.3	49.4	52.3	49.9	26.13	73.50	106.14
Zliczanie przez wyostrzenie *p*=0.5	0.4	86.5	86.8	86.5	99.1	86.6	86.3	13.5	26.48	69.89	100.92
	0.6	86.6	86.9	86.6	99.1	86.7	86.4	13.4	26.28	71.61	103.41
	0.8	86.7	86.9	86.7	99.1	86.8	86.5	13.3	26.17	72.60	104.84
Zliczanie przez wyostrzenie *p*=1	0.4	80.7	81.0	80.7	98.6	80.8	80.6	19.3	26.58	69.81	100.81
	0.6	79.9	80.2	79.9	98.5	79.9	79.8	20.1	26.38	71.53	103.29
	0.8	79.9	80.2	79.9	98.5	79.9	79.9	20.1	26.24	72.55	104.76
Zliczanie przez wyostrzenie *p*=2	0.4	50.8	49.8	50.8	96.6	50.3	53.3	49.2	26.59	69.81	100.80
	0.6	50.3	49.1	50.3	96.4	49.7	52.6	49.7	26.41	71.51	103.26
	0.8	50.2	48.9	50.2	96.3	49.5	52.4	49.8	26.28	72.52	104.71
Wyostrzenie kontrastu	-	50.3	49.1	50.3	96.4	49.7	52.6	49.7	26.11	73.52	106.16

Tabela 13. Wyniki metryk ewaluacyjnych dla agregacji operatorem OWA zmiennej lingwistycznej Cytowanie i Percentyl z wektorem wag (1. 0).

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niski											
Zliczanie przez łączenie *p*=0.5	-	93.7	94.0	93.7	99.9	93.8	93.7	6.3	0.92	90.87	131.21
Zliczanie przez łączenie *p*=1	-	92.0	92.4	92.0	99.8	92.2	92.0	8.0	0.93	90.86	131.20
Zliczanie przez łączenie *p*=2	-	88.5	89.2	88.5	99.7	88.8	88.7	11.5	0.95	90.85	131.18
Zliczanie przez progowanie i łączenie p=0.5	0.2	91.9	92.4	91.9	99.8	92.1	92.0	8.1	3.13	88.60	127.94
	0.4	89.3	90.0	89.3	99.7	89.7	89.5	10.7	4.87	86.61	125.06
	0.6	88.3	89.3	88.3	99.6	88.8	88.5	11.7	7.69	83.23	120.18
	0.8	77.8	79.2	77.8	99.0	78.5	78.6	22.2	14.65	75.00	108.30
Zliczanie przez progowanie i łączenie p=1	0.2	91.7	92.2	91.7	99.8	91.9	91.7	8.3	3.14	88.60	127.93
	0.4	89.3	90.0	89.3	99.7	89.6	89.5	10.7	4.88	86.61	125.06
	0.6	88.3	89.3	88.3	99.6	88.8	88.5	11.7	7.69	83.23	120.18
	0.8	77.8	79.2	77.8	99.0	78.5	78.6	22.2	14.65	75.00	108.30
Zliczanie przez progowanie i łączenie p=2	0.2	89.3	90.0	89.3	99.7	89.6	89.5	10.7	3.15	88.58	127.91
	0.4	88.2	89.0	88.2	99.6	88.6	88.4	11.8	4.89	86.60	125.05
	0.6	88.2	89.2	88.2	99.6	88.6	88.4	11.8	7.70	83.22	120.17
	0.8	77.8	79.2	77.8	99.0	78.5	78.6	22.2	14.65	75.00	108.30
Zliczanie przez wyostrzenie *p*=0.5	0.2	97.5	97.5	97.5	99.9	97.5	97.4	2.5	1.31	70.93	102.43
	0.4	95.8	95.9	95.8	99.9	95.8	95.7	4.2	1.21	72.58	104.81
	0.6	93.8	94.0	93.8	99.9	93.9	93.8	6.2	1.12	74.86	108.09
	0.8	93.7	94.0	93.7	99.9	93.9	93.8	6.3	1.02	79.32	114.54
Zliczanie przez wyostrzenie *p*=1	0.2	95.9	96.0	95.9	99.9	96.0	95.9	4.1	1.32	70.93	102.42
	0.4	93.6	93.9	93.6	99.8	93.7	93.6	6.4	1.23	72.56	104.78
	0.6	92.2	92.6	92.2	99.8	92.4	92.2	7.8	1.17	74.82	108.03
	0.8	92.0	92.4	92.0	99.8	92.2	92.1	8.0	1.06	79.29	114.49
Zliczanie przez wyostrzenie *p*=2	0.2	95.9	96.0	95.9	99.9	95.9	95.9	4.1	1.32	70.94	102.43
	0.4	92.4	92.8	92.4	99.8	92.6	92.5	7.6	1.23	72.57	104.79
	0.6	91.8	92.3	91.8	99.8	92.0	91.8	8.2	1.19	74.81	108.02
	0.8	90.0	90.6	90.0	99.7	90.3	90.1	10.0	1.11	79.25	114.43
Wyostrzenie kontrastu	-	89.1	89.7	89.1	99.7	89.4	89.4	10.9	0.93	90.86	131.19

term = ani niski ani wysoki											
Zliczanie przez łączenie *p*=0.5	-	94.7	94.8	94.7	99.7	94.7	94.5	5.3	9.00	85.91	124.06
Zliczanie przez łączenie *p*=1	-	92.4	92.7	92.4	99.5	92.6	92.1	7.6	9.03	85.89	124.02
Zliczanie przez łączenie *p*=2	-	87.1	87.6	87.1	99.1	87.4	86.8	12.9	9.06	85.87	124.00
Zliczanie przez wyostrzenie p=0.5	0.2	94.1	94.2	94.1	99.6	94.2	93.9	5.9	9.60	75.03	108.34
	0.4	93.2	93.3	93.2	99.6	93.2	92.9	6.8	9.17	83.37	120.38
Zliczanie przez wyostrzenie *p*=1	0.2	92.7	92.9	92.7	99.5	92.8	92.4	7.3	9.65	75.00	108.29
	0.4	88.7	89.0	88.7	99.2	88.9	88.3	11.3	9.32	83.25	120.21
Zliczanie przez wyostrzenie *p*=2	0.2	92.4	92.6	92.4	99.5	92.5	92.1	7.6	9.65	75.00	108.30
	0.4	85.5	86.0	85.5	98.8	85.7	85.1	14.5	9.39	83.21	120.15
Wyostrzenie kontrastu	0	87.1	87.6	87.1	99.1	87.4	86.8	12.9	9.07	85.86	123.98
term = wysoki											
Zliczanie przez łączenie *p*=0.5	-	99.1	99.1	99.1	99.9	99.1	99.0	0.9	7.58	81.17	117.21
Zliczanie przez łączenie *p*=1	-	97.2	97.2	97.2	99.8	97.2	97.1	2.8	7.62	81.14	117.17
Zliczanie przez łączenie *p*=2	-	86.8	87.0	86.8	99.2	86.9	86.7	13.2	7.67	81.12	117.14
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	88.2	88.5	88.2	99.3	88.3	88.0	11.8	16.06	73.75	106.50
Zliczanie przez progowanie i łączenie *p*=1	0.2	87.9	88.2	87.9	99.2	88.1	87.7	12.1	16.08	73.74	106.47
Zliczanie przez progowanie i łączenie *p*=2	0.2	85.7	85.9	85.7	99.1	85.8	85.6	14.3	16.13	73.72	106.44
Zliczanie przez wyostrzenie *p*=0.5	0.6	99.2	99.2	99.2	99.9	99.2	99.1	0.8	7.78	70.45	101.73
	0.8	99.2	99.2	99.2	99.9	99.2	99.1	0.8	7.71	73.65	106.35
Zliczanie przez wyostrzenie *p*=1	0.6	97.4	97.4	97.4	99.8	97.4	97.3	2.6	7.86	70.39	101.65
	0.8	97.3	97.3	97.3	99.8	97.3	97.2	2.7	7.78	73.60	106.28
Zliczanie przez wyostrzenie *p*=2	0.6	97.0	97.0	97.0	99.8	97.0	96.8	3.0	7.92	70.36	101.59
	0.8	87.0	87.1	87.0	99.2	87.1	86.9	13.0	7.88	73.55	106.20
Wyostrzenie kontrastu	-	87.3	87.5	87.3	99.3	87.4	87.2	12.7	7.63	81.10	117.11

Tabela 14. Wyniki metryk ewaluacyjnych dla agregacji operatorem OWA zmiennej lingwistycznej Cytowanie i Percentyl z wektorem wag (0.8. 0.2).

Funkcja wagowa	*t*	Acc	Prec	Sens	Spec	F1	MCC	Loss	Perc lost	Perc class	Perc class [over]
term = niski											
Zliczanie przez łączenie *p*=0.5	-	93.7	93.9	93.7	99.9	93.8	93.7	6.3	0.90	94.89	137.02
Zliczanie przez łączenie *p*=1	-	91.9	92.3	91.9	99.8	92.1	91.9	8.1	0.90	94.89	137.02
Zliczanie przez łączenie *p*=2	-	89.6	90.3	89.6	99.6	90.0	89.7	10.4	0.91	94.88	137.01
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	94.0	94.5	94.0	99.8	94.2	93.9	6.0	3.30	92.35	133.35
	0.4	88.9	89.7	88.9	99.7	89.3	89.1	11.1	5.51	89.69	129.51
	0.6	87.1	88.3	87.1	99.5	87.6	87.2	12.9	10.36	83.73	120.90
Zliczanie przez progowanie i łączenie *p*=1	0.2	93.7	94.3	93.7	99.8	94.0	93.7	6.3	3.30	92.35	133.35
	0.4	88.8	89.6	88.8	99.6	89.2	89.0	11.2	5.51	89.69	129.51
	0.6	87.0	88.2	87.0	99.5	87.6	87.2	13.0	10.36	83.73	120.90
Zliczanie przez progowanie i łączenie *p*=2	0.2	90.3	91.0	90.3	99.6	90.6	90.3	9.7	3.31	92.34	133.34
	0.4	88.4	89.3	88.4	99.6	88.8	88.5	11.6	5.52	89.69	129.51
	0.6	86.9	88.1	86.9	99.4	87.5	87.0	13.1	10.37	83.73	120.90
Zliczanie przez wyostrzenie *p*=0.5	0.2	95.9	96.0	95.9	99.9	96.0	95.9	4.1	1.29	71.27	102.91
	0.4	93.8	94.0	93.8	99.9	93.9	93.8	6.2	1.17	73.40	105.99
	0.6	93.6	93.9	93.6	99.9	93.8	93.7	6.4	1.06	77.22	111.50
	0.8	93.7	93.9	93.7	99.8	93.8	93.7	6.3	0.94	86.56	124.99
Zliczanie przez wyostrzenie *p*=1	0.2	95.8	96.0	95.8	99.9	95.9	95.8	4.2	1.29	71.27	102.91
	0.4	92.3	92.7	92.3	99.8	92.5	92.4	7.7	1.20	73.38	105.96

	0.6	91.8	92.3	91.8	99.8	92.1	91.9	8.2	1.11	77.18	111.44
	0.8	91.8	92.3	91.8	99.8	92.1	91.9	8.2	0.97	86.54	124.96
Zliczanie przez wyostrzenie *p*=2	0.2	95.8	95.9	95.8	99.9	95.8	95.8	4.2	1.29	71.27	102.92
	0.4	92.2	92.6	92.2	99.8	92.4	92.2	7.8	1.20	73.38	105.96
	0.6	90.0	90.6	90.0	99.7	90.3	90.1	10.0	1.16	77.14	111.39
	0.8	89.5	90.2	89.5	99.6	89.8	89.5	10.5	1.03	86.49	124.88
Wyostrzenie kontrastu	-	90.2	90.8	90.2	99.7	90.5	90.3	9.8	0.90	94.89	137.02
term = ani niski ani wysoki											
Zliczanie przez łączenie *p*=0.5	-	94.6	94.8	94.6	99.7	94.7	94.5	5.4	8.93	89.55	129.31
Zliczanie przez łączenie *p*=1	-	92.4	92.7	92.4	99.5	92.5	92.1	7.6	8.93	89.55	129.31
Zliczanie przez łączenie *p*=2	-	87.1	87.6	87.1	99.0	87.3	86.8	12.9	8.94	89.54	129.30
Zliczanie przez wyostrzenie *p*=0.5	0.2	93.7	93.9	93.7	99.6	93.8	93.5	6.3	9.50	78.38	113.18
	0.4	93.6	93.8	93.6	99.7	93.7	93.4	6.4	8.95	89.06	128.60
Zliczanie przez wyostrzenie *p*=1	0.2	90.3	90.6	90.3	99.4	90.4	90.0	9.7	9.55	78.35	113.13
	0.4	88.5	88.9	88.5	99.3	88.7	88.3	11.5	8.99	89.03	128.56
Zliczanie przez wyostrzenie *p*=2	0.2	89.8	90.1	89.8	99.3	90.0	89.5	10.2	9.55	78.34	113.12
	0.4	85.0	85.6	85.0	98.8	85.3	84.6	15.0	9.02	89.01	128.53
Wyostrzenie kontrastu	-	87.1	87.6	87.1	99.0	87.3	86.8	12.9	8.94	89.54	129.30
term = wysoki											
Zliczanie przez łączenie *p*=0.5	-	97.7	97.7	97.7	99.9	97.7	97.6	2.3	7.46	90.49	130.66
Zliczanie przez łączenie *p*=1	-	97.0	97.0	97.0	99.8	97.0	96.8	3.0	7.47	90.49	130.66
Zliczanie przez łączenie *p*=2	-	86.4	86.5	86.4	99.1	86.4	86.2	13.6	7.49	90.48	130.65
Zliczanie przez progowanie i łączenie *p*=0.5	0.2	83.0	83.4	83.0	99.1	83.2	83.2	17.0	17.11	81.79	118.10
	0.4	75.5	75.6	75.5	98.6	75.6	76.2	24.5	26.61	72.70	104.98
Zliczanie przez progowanie i łączenie *p*=1	0.2	82.8	83.1	82.8	99.0	82.9	82.9	17.2	17.11	81.79	118.10
	0.4	75.4	75.5	75.4	98.6	75.5	76.1	24.6	26.61	72.70	104.98
Zliczanie przez progowanie i łączenie *p*=2	0.2	80.4	80.7	80.4	98.9	80.6	80.7	19.6	17.13	81.78	118.08
	0.4	73.1	73.2	73.1	98.5	73.1	74.0	26.9	26.61	72.70	104.97
Zliczanie przez wyostrzenie *p*=0.5	0.4	99.2	99.2	99.2	99.9	99.2	99.1	0.8	7.82	69.97	101.04
	0.6	97.7	97.7	97.7	99.9	97.7	97.6	2.3	7.73	74.18	107.11
	0.8	97.7	97.7	97.7	99.9	97.7	97.6	2.3	7.54	84.38	121.84
Zliczanie przez wyostrzenie *p*=1	0.4	97.5	97.5	97.5	99.8	97.5	97.3	2.5	7.88	69.92	100.97
	0.6	97.2	97.2	97.2	99.8	97.2	97.0	2.8	7.81	74.10	107.00
	0.8	97.0	97.0	97.0	99.8	97.0	96.8	3.0	7.58	84.35	121.80
Zliczanie przez wyostrzenie *p*=2	0.4	97.2	97.2	97.2	99.8	97.2	97.1	2.8	7.90	69.91	100.95
	0.6	96.5	96.6	96.5	99.7	96.5	96.3	3.5	7.89	74.05	106.93
	0.8	84.4	84.5	84.4	99.1	84.4	84.4	15.6	7.67	84.28	121.70
Wyostrzenie kontrastu	-	86.9	87.1	86.9	99.2	87.0	86.8	13.1	7.48	90.48	130.66