

INSTYTUT BADAŃ SYSTEMOWYCH

POLSKIEJ AKADEMII NAUK

Nowe metody uwzględnienia niepewności
i nieprecyzyjności danych w klasyfikacji

Zastosowania w medycynie

New Methods for Handling Uncertainty and Imprecision in Classification

Applications in medicine

mgr inż. DAWID KOSIOR

ROZPRAWA DOKTORSKA

PROMOTOR:

dr hab. Barbara PEKALA, prof. UR

dziedzina: Nauk inżynierijsko-technicznych

dyscyplina: Informatyka techniczna i telekomunikacja

Warszawa, 2025

Streszczenie

W rozprawie zaproponowano nowe podejścia do uwzględniania niedoskonałości danych, rozumianej jako niepewność lub nieprecyzyjność, w klasyfikacji danych, analizowanej zarówno w ujęciu teoretycznym, jak i praktycznym. Badania koncentrowały się przede wszystkim na zastosowaniach praktycznych, ze szczególnym uwzględnieniem diagnostyki medycznej. Część teoretyczna obejmowała opracowanie nowych miar zawierania, podobieństwa oraz entropii dla przedziałowych zbiorów rozmytych, które zastosowano do reprezentacji danych obarczonych niepewnością/nieprecyzyjnością. W części aplikacyjnej zaproponowano metody wykorzystania tych miar na różnych etapach tworzenia klasyfikatora. Opracowane rozwiązania zostały dostosowane do specyfiki wspomagania podejmowania decyzji na podstawie danych medycznych. W ramach kompleksowego wspomagania klasyfikacji zaproponowano również nowe podejścia do uzupełniania brakujących danych oraz do redukcji atrybutów. Proponowane rozwiązania zostały przetestowane na rzeczywistych zbiorach danych klinicznych, dotyczących m.in. leczenia pacjentek z rakiem piersi, diagnostyki schorzeń sercowo-naczyniowych oraz prewencji depresji, czyli wczesnego wykrywania zagrożenia wystąpieniem epizodów depresyjnych.

Abstract

The dissertation proposes new approaches to handling data imperfections, understood as uncertainty or imprecision, in data classification, analyzed from both theoretical and practical perspectives. The research primarily focused on practical applications, with particular emphasis on medical diagnostics. The theoretical part involved the development of new measures of inclusion, similarity, and entropy for interval-valued fuzzy sets, which were applied to the representation of data affected by uncertainty and imprecision. In the applied part, methods for using these measures at various stages of classifier construction were proposed. The developed solutions were adapted to the specifics of decision support based on medical data. As part of comprehensive classification support, new approaches to missing data imputation and attribute reduction were also introduced. The proposed methods were tested on real clinical datasets, including breast cancer treatment, cardiovascular disease diagnostics, and depression prevention, i.e., early detection of the risk of depressive episodes.

Spis treści

1	Wstęp	7
1.1	Zakres tematyczny rozprawy	7
1.2	Motywacja i cel rozprawy	9
1.3	Struktura rozprawy	14
2	Podstawowe pojęcia teorii klasycznych i przedziałowych zbiorów rozmytych	17
2.1	Pojęcie zbioru rozmytego	17
2.2	Rozszerzenie zbiorów rozmytych do przedziałowych zbiorów rozmytych .	19
2.2.1	Relacje porządku na przedziałach	19
2.2.2	Przedziałowe agregacje rozmyte	22
2.2.3	Przedziałowe negacje rozmyte	31
2.2.4	Przedziałowe zbiory rozmyte	36
2.2.5	Przedziałowe relacje rozmyte - wybrane własności	37
3	Miary zawierania, podobieństwa i entropii uwzględniające nieprecyzyjność	39
3.1	Miary zawierania się przedziałów i przedziałowych zbiorów rozmytych . .	39
3.1.1	Wskaźnik pierwszeństwa klasyczny	40
3.1.2	Wskaźnik pierwszeństwa możliwości	42
3.1.3	Wskaźnik pierwszeństwa konieczności	43
3.2	Miara podobieństwa dla przedziałowych zbiorów rozmytych	44
3.2.1	Miara podobieństwa klasyczna	44
3.2.2	Miara podobieństwa możliwości	47
3.2.3	Miara podobieństwa konieczności	47
3.3	Entropia przedziałowych zbiorów rozmytych	48
3.3.1	Miara entropii klasyczna	49
3.3.2	Miara entropii możliwości	52
3.3.3	Miara entropii konieczności	54

4 Wybrane metody klasyfikacji	57
4.1 Przegląd wybranych metod klasyfikacji	57
4.2 Miary jakości klasyfikacji	64
5 Metody wspierające klasyfikację w diagnostyce medycznej oparte o technikę k najbliższych sąsiadów	69
5.1 Klasyfikacja z niepewnością	70
5.1.1 Propozycja modelu klasyfikacji IV-kNN	71
5.1.2 Opis danych - diagnostyka raka piersi	74
5.1.3 Analiza wyników badań wpływu uwzględnienia niepewności na jakość klasyfikacji	75
5.2 Uzupełnianie brakujących wartości w danych	79
5.2.1 Metoda uzupełniania brakujących danych z użyciem przedziałowej miary entropii	81
5.2.2 Analiza wyników badań wpływu brakujących danych na jakość klasyfikacji	82
5.3 Selekcja atrybutów	87
5.3.1 Nowy algorytm wyznaczania istotności atrybutów w zbiorze danych - IVRelief	87
5.3.2 Analiza wyników badań wpływu selekcji cech na jakość klasyfikacji	89
5.4 System prewencji depresji	92
5.4.1 Opis problemu i procesu budowy zbioru danych	92
5.4.2 Model wspomagania prewencyjnego wykrywania epizodów depresyjnych	95
5.4.3 Ocena Systemu Prewencji Depresji (SPD) i analiza wpływu algorytmu SENSDEPR na czułość diagnostyki depresji	97
6 Pozostałe kierunki badań	101
6.1 System wspierający diagnostykę choroby sercowo - naczyniowej	101
6.1.1 Opis danych i problemu	101
6.1.2 Stosowane metody i scenariusze badawcze	102
6.1.3 Ocena systemu i analiza wyników eksperymentów	108
6.2 Biblioteka dla systemów wnioskowania przedziałowego (Interval-Valued Fuzzy Inference System (IFIS))	109
7 Podsumowanie i plany na przyszłość	113
Bibliografia	119
Spis rysunków	133

Podstawowe oznaczenia

$$L^I = \{[\underline{x}, \bar{x}] : \underline{x}, \bar{x} \in [0, 1], \underline{x} \leq \bar{x}\};$$

$$w(p) = \bar{p} - \underline{p} \text{ dla } p = [\underline{p}, \bar{p}] \in L^I;$$

$$1_{L^I} = [1, 1];$$

$$0_{L^I} = [0, 0];$$

$$[\underline{a}, \bar{a}] \leq_{L^I} [\underline{b}, \bar{b}] \Leftrightarrow \underline{a} \leq \underline{b} \text{ i } \bar{a} \leq \bar{b};$$

$$[\underline{a}, \bar{a}] <_{L^I} [\underline{b}, \bar{b}] \Leftrightarrow [\underline{a}, \bar{a}] \leq_{L^I} [\underline{b}, \bar{b}] \text{ i } (\underline{a} < \underline{b} \text{ lub } \bar{a} < \bar{b});$$

$$\leq_{Adm} - \text{porządek dopuszczalny (liniowy) w } L^I;$$

$$[\underline{x}, \bar{x}] \leq_{XY} [\underline{y}, \bar{y}] \Leftrightarrow \underline{x} + \bar{x} < \underline{y} + \bar{y} \text{ lub } (\bar{x} + \underline{x} = \bar{y} + \underline{y} \text{ i } \bar{x} - \underline{x} \leq \bar{y} - \underline{y}) -$$

porządek Xu i Yagera;

$$[\underline{x}, \bar{x}] \leq_{Lex1} [\underline{y}, \bar{y}] \Leftrightarrow \underline{x} < \underline{y} \text{ lub } (\underline{x} = \underline{y} \text{ i } \bar{x} \leq \bar{y}) - \text{pierwszy porządek leksykograficzny};$$

$$[\underline{x}, \bar{x}] \leq_{Lex2} [\underline{y}, \bar{y}] \Leftrightarrow \bar{x} < \bar{y} \text{ lub } (\bar{x} = \bar{y} \text{ i } \underline{x} \leq \underline{y}) - \text{drugi porządek leksykograficzny};$$

$$[\underline{x}, \bar{x}] \leq_{nec} [\underline{y}, \bar{y}] \Leftrightarrow \bar{x} \leq \underline{y} - \text{relacja konieczności (pesymistyczna)};$$

$$[\underline{x}, \bar{x}] \leq_{pos} [\underline{y}, \bar{y}] \Leftrightarrow \underline{x} \leq \bar{y} - \text{relacja możliwości (optymistyczna)};$$

$\leq$ - porządek częściowy lub dopuszczalny (porządek klasyczny), a w dalszej kolejności relacje $\leq_{pos}$ (relacja możliwości) i $\leq_{nec}$ (relacja konieczności);

$IVFS$ - przedziałowy zbiór rozmyty;

$IVFS(X)$ - rodzina wszystkich przedziałowych zbiorów rozmytych na X ;

$A \preceq B \Leftrightarrow a_i \leq b_i$ dla $A, B \in IVFS(X)$, $i = 1, \dots, n$, gdzie $\leq$ jest tym samym rodzajem porządku (częściowym lub liniowym, lub relacją możliwości lub konieczności) dla każdego i oraz $a_i = A(x_i)$, $b_i = B(x_i)$.

ROZDZIAŁ 1

Wstęp

1.1 Zakres tematyczny rozprawy

Współczesne systemy informatyczne znajdują zastosowanie praktycznie w każdej dziedzinie życia, co wiąże się często z potrzebą gromadzenia i przetwarzania znacznych ilości danych. Ten początkowy etap eksploracji stanowi fundament dla wielu istotnych zastosowań, takich jak klasyfikacja, która opiera się na budowie modeli, zwanych klasyfikatorami. Po odpowiednim przetworzeniu dane mogą posłużyć do budowy systemów rozpoznawania obiektów lub systemów wspomagania podejmowania decyzji, które stanowią kluczowy element zarządzania w różnych dziedzinach, takich jak medycyna, finanse czy przemysł. Jednak gromadzone dane w systemach informatycznych często dotyczą złożonych procesów i zjawisk, które nie poddają się klasycznym metodom modelowania. Istniejące metody i podejścia mają swoje ograniczenia, zwłaszcza jeśli chodzi o efektywne modelowanie pojęć złożonych, w szczególności związanych z diagnostyką medyczną, które mogą być niepewne lub nieprecyzyjne, czyli niedoskonałe (ang. imperfect data). Teoria zbiorów rozmytych [1] oraz jej rozszerzenia stanowią skuteczne narzędzie do reprezentacji i przetwarzania takich danych. Klasyczna teoria zbiorów rozmytych pozwala opisywać pojęcia nieprecyzyjne (nieostre) poprzez wprowadzenie stopni przynależności, które wykraczają poza binarne ujęcie należenia elementu do zbioru i związane z tym binarne pojęcie prawdy i fałszu. Jednak w przypadku danych niepełnych lub niedoskonałych, gdzie dokładne określenie stopnia przynależności jest trudne lub niemożliwe, taka reprezentacja okazuje się niewystarczająca. Mamy wtedy do czynienia z nieprecyzyznością i związaną z nią niepewnością. W odpowiedzi na te ograniczenia, badacze opracowali szereg rozszerzeń klasycznej teorii zbiorów rozmytych, które umożliwiają lepsze modelowanie i przetwarzanie informacji obciążonej niepewnością i nieprecyzyznością.

Istnieje wiele uogólnień teorii zbiorów rozmytych, które wiążą się bezpośrednio z potrzebą bardziej elastycznego konstruowania funkcji przynależności. Pozwalają one bar-

dziej adekwatnie reprezentować obiekty opisane poprzez niedoskonałą informację, a także reprezentować brak wiedzy lub niepewność ekspertów. Wśród najbardziej znanych uogólnień teorii zbiorów rozmytych możemy wskazać: zbiory L-rozmyte [2] oparte o koncepcję kraty. W szczególności, możliwym rozszerzeniem zbioru rozmytego jest intuicjonistyczny zbiór rozmyty Atanassowa. Intuicjonistyczne zbiory rozmyte [3], wykazują zaletę polegającą na elastyczności w określeniu przynależności poprzez uwzględnienie także stopnia nieprzynależności. Przegląd różnych rozszerzeń teorii zbiorów rozmytych można znaleźć, na przykład, w pracy [4].

W niniejszej rozprawie koncentrujemy się na przedziałowych zbiorach rozmytych. W 1975 roku Sambuc i Zadeh [5, 6] przedstawili koncepcję zbioru rozmytego o przedziałowych wartościach funkcji przynależności, co stanowi szczególny przypadek zbiorów rozmytych typu 2, w wypadku których wartościami funkcji przynależności są zbiory rozmyte określone na przedziale $[0,1]$. **Przedziałowe zbiory rozmyte** (*ang. Interval-Valued Fuzzy Sets*) pozwalają na określenie nieprecyzyjności w sposób bardziej elastyczny i efektywny.

W zakresie rozwoju teorii przedziałowych zbiorów rozmytych wielu badaczy proponowało różne operatory i miary dla przedziałowych zbiorów rozmytych, w szczególności miary agregacji, podobieństwa, entropii czy miary zawierania (wskaźniki pierwszeństwa). Badano także różnego rodzaju zależności między nimi, por. np. [7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. Używano ich także w procesie wspomagania podejmowania decyzji [17, 18, 19, 20, 21, 22, 23, 24]. Reprezentacja niepewności jest niezwykle istotna w podejmowaniu decyzji w warunkach niepełnej informacji lub braku precyzyjnych danych. Wybór odpowiedniej metody reprezentacji zależy od kontekstu problemu i charakterystyki danych, jakimi dysponujemy.

Zastosowanie przedziałów do reprezentacji niedoskonałej informacji wykracza oczywiście poza tematykę przedziałowych zbiorów rozmytych. Często nasza wiedza co do wartości danego atrybutu/pomiaru jest nieprecyzyjna i możemy tylko powiedzieć, że znajduje się ona w pewnym przedziale. W literaturze pojawiły się propozycje modelowania nieprecyzyjności mającej umożliwić klasyfikację w warunkach niepełnych lub nieprecyzyjnych danych, co zostało omówione w wielu pracach [21, 22, 23, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36]. Niemniej jednak, nadal wymagane są poszukiwania efektywniejszych metod reprezentacji niepewności.

Badania teoretyczne i praktyczne w pracy skupiały się na aspekcie aplikacyjnym ze szczególnym uwzględnieniem diagnostyki medycznej. Dane medyczne są szczególnie obciążone niepewnością i nieprecyzyjnością wynikającą z błędów związanych z dokładnością mierzonych parametrów medycznych (np. rodzaju badanej tkanki, badań obrazowych, itp.), jak również z klasą sprzętu, jakim wykonywane jest badanie i jego dokładnością. Również opisy pewnych badań wykonywane przez ekspertów są obciążone niepewnością,

gdyż istnieje wiele czynników zewnętrznych, które wprowadzają niepewność co do wyników wykonanych pomiarów.

Przedział stanowiący wartość funkcji przynależności może być interpretowany w ujęciu dyzjunktywnym (epistemicznym), jeśli zakładamy że jeden z jego elementów jest faktycznym stopniem przynależności, ale nie wiemy który. Może też być interpretowany w ujęciu konjunktywnym - wszystkie elementy przedziału łącznie reprezentują stopień przynależności - na przykład jako efekt agregacji różnych punktów widzenia. Przyjmujemy w niniejszej pracy pierwszą z tych interpretacji, która jest bardziej naturalna, w szczególności w kontekście rozważanych zastosowań.

W niniejszej rozprawie zaproponowano nowe podejścia wspierające podejmowanie decyzji z niepewnością w dwóch obszarach: teoretycznym i praktycznym. Aspekt teoretyczny opiera się na zaproponowaniu nowych miar zawierania, podobieństwa czy entropii przedziałowych zbiorów rozmytych i operatorów agregacji przedziałowych uwzględniających nieprecyzyjność. Kluczowe w tych badaniach było podejście epistemiczne. Kolejnym istotnym kierunkiem prac był aspekt praktyczny koncentrujący się na zaproponowaniu sposobu zastosowania wprowadzonych operatorów i miar uwzględniających niepewność w trzech ujęciach:

- klasycznym (miary zgodne z porządkiem częściowym lub liniowym dla przedziałowych wartości rozmytych),
- optymistycznym (miary zgodne z relacją możliwości porównywania przedziałowych wartości rozmytych),
- pesymistycznym (miary zgodne z relacją konieczności porównywania przedziałowych wartości rozmytych)

do poprawy jakości klasyfikatorów na różnych etapach procesu budowy modelu. Opracowano nowe metody reprezentacji nieprecyzyjności w klasyfikacji z wykorzystaniem danych nieprecyzyjnych z punktu widzenia pomiarów, czy ocen eksperckich, a także metod uzupełniania wartości brakujących w danych oraz redukcji atrybutów w zbiorze danych. Następnie poddano je weryfikacji na rzeczywistych danych klinicznych dotyczących leczenia pacjentów chorych na raka piersi (baza UCI, [37]), danych dotyczących diagnostyki problemów zakrzepowo-sercowych (Hiszpania, Pampeluna [18]) oraz danych związanych z wykrywaniem epizodów depresyjnych [38]. Wyniki przeprowadzonych doświadczeń skuteczność opracowanych metod.

1.2 Motywacja i cel rozprawy

Analiza danych z nieprecyzyjnością napotyka trudności przy konstruowaniu efektywnie działających klasyfikatorów dla wielu rzeczywistych problemów. W związku z tym

pojawiły się propozycje uwzględnienia niepewności, która jest wszechobecna w procesach decyzyjnych. Zastosowanie wiedzy dziedzinowej z uwzględnieniem nieprecyzyjności danych czy ocen eksperckich ma istotny wpływ na wydajność stosowanych systemów decyzyjnych. Przykładowo w pracach [32, 36, 39, 40, 41, 42, 43, 44, 45] badano wpływ uwzględnienia niepewności na wyniki w takich metodach klasyfikacji jak: regresja logistyczna, sztuczne sieci neuronowe, metoda k -najbliższych sąsiadów k -NN (*ang. k -nearest neighbours*), naiwny klasyfikator Bayesa, drzewa decyzyjne, algorytmy ewolucyjne oraz metoda wektorów nośnych SVM (*ang. support vector machine*). Poprawa jakości klasyfikacji w porównaniu do modeli nieuwzględniających niepewności była znacząca dla metody k -NN. Wyjaśnialność tego typu klasyfikatorów w kontekście proponowanych nowych ww. miar uwzględniających niepewność przyczyniła się do skoncentrowania badań na tej metodzie. Ponadto taki stan rzeczy skłonił do postawienia pytania badawczego dotyczącego możliwości efektywnego zastosowania reprezentacji niepewności w zakresie wspomnianej klasyfikacji i opisu niepewności danych na innych niż dotychczas etapach procesu odkrywania wiedzy.

Głównym celem rozprawy jest opracowanie nowych metod i algorytmów uwzględniających niepewność i nieprecyzyjność danych w celu poprawy jakości klasyfikatorów tworzonych za pomocą metody k -najbliższych sąsiadów, bardzo szeroko stosowanej i dającej zwykle bardzo dobre wyniki. Główny cel rozprawy realizowano poprzez następujące cele szczegółowe, które obejmowały:

1. Opracowanie miar i metod uwzględniających niepewność i nieprecyzyjność, umożliwiających efektywną klasyfikację;
2. Opracowanie metod klasyfikacji uwzględniających dane niepewne, opartych o algorytm k -NN z wykorzystaniem nowych miar, a także wnioskowania rozmytego z nowymi agregacjami uwzględniającymi niepewność;
3. Wybór cech dystynktywnych dla poszczególnych klas obiektów - zaproponowanie efektywnej metody selekcji atrybutów przy wykorzystaniu miar uwzględniających niepewność, w szczególności nowej miary podobieństwa;
4. Oszacowanie wpływu brakujących wartości na jakość klasyfikacji - zaproponowanie efektywnej metody uzupełniania brakujących wartości z wykorzystaniem miar uwzględniających niepewność, w szczególności nowej miary entropii;
5. Weryfikację przydatności proponowanych metod reprezentacji i manipulacji danymi z niepewnością, która została przeprowadzona poprzez sprawdzenie ich wpływu na jakość klasyfikacji.

Główna teza rozprawy brzmi: **Uwzględnienie niepewności i nieprecyzyjności może znacząco polepszyć jakość działania klasyfikatorów opartych na idei k -najbliższych sąsiadów, zarówno w warunkach danych niekompletnych jak i w procesie optymalizacji zbioru cech.**

Tezę tę można uszczegółowić za pomocą następujących tez pomocniczych:

1. Proponowane w rozprawie metody porządkowania danych przedziałowych, a także agregacji i miar wiedzy/entropii oraz podobieństwa zgodne z epistemicznym (możliwości lub konieczności) punktem widzenia mogą znacząco wpłynąć na efektywność klasyfikacji.
2. Nowe metody reprezentacji niepewności oraz związane z nimi miary mogą być z powodzeniem stosowane do rozwiązywania rzeczywistych problemów, takich jak klasyfikacja raka piersi, prognozowanie zagrożenia depresją czy chorobami sercowo-naczyniowymi.

Główne wyniki

Główne wyniki rozprawy można podzielić na dwie kategorie: opracowanie metod reprezentacji nieprecyzyjności/niepewności danych, algorytmów klasyfikacji, uzupełniania brakujących wartości lub selekcji atrybutów uwzględniających dane z niepewnością oraz rozwiązanie konkretnych problemów związanych z predykcją w warunkach nieprecyzyjności. Metody służące realizacji celów rozprawy zostały opisane w rozdziałach 2, 3 i 5 i stanowią znaczące rozszerzenie wyników badań opublikowanych w artykułach: [38, 46, 47, 48, 49, 50, 51, 52].

Pierwszy zbadany problem skupia się na obszarze medycznym, a konkretnie dotyczy diagnostyki raka piersi. Warto zaznaczyć, że diagnostyka raka piersi stanowi bardzo ważny obszar medycyny, którego skuteczność wpływa na życie wielu pacjentów. Warto także zwrócić uwagę, że próby diagnostyki raka piersi mogą być oparte na danych, które nie zawsze są kompletne. Mogą pojawić się sytuacje, w których niektóre cechy są niedostępne, a odczytane parametry obarczone są pewnym stopniem niepewności. Taka sytuacja otwiera drzwi do stosowania metodologii opartej na teorii przedziałowych zbiorów rozmytych. W ramach tego zagadnienia, w pracy [46] został opracowany model klasyfikacji wykorzystujący ideę algorytmu k -NN, który został poddany uogólnieniu tak, aby uwzględniał aspekt niepewności. Badania w ww. artykule koncentrowały się na wprowadzeniu nowych miar podobieństwa opartych na relacji porządku dla wartości przedziałowych w sensie klasycznym, czyli dla porządku częściowego lub liniowego oraz zastosowaniu tych miar w algorytmie klasyfikacji. Analiza omawianego problemu reprezentacji niepewności/nieprecyzyjności danych przyczyniła się do rozszerzenia w ramach pracy doktorskiej prezentowanych w powyższym artykule badań na inne klasy relacji porządkujących i związanych z nimi miar agregacji, czy podobieństwa w sensie epistemicznym (optymistycznym/możliwości lub pesymistycznym/konieczności). W badaniach kluczowym elementem było zweryfikowanie wpływu proponowanych miar na klasyfikację w warunkach uwzględniających niepewność i nieprecyzyjność, czyli z wykorzystaniem teorii przedziałowych zbiorów rozmytych. Podejście uwzględniające niepewność pozwoliło na bardziej precyzyjną i rzetelną klasyfikację, co wykazały przeprowadzone eksperymenty. Dzięki temu, pacjenci mogą być objęci odpowiednią opieką medyczną, co przekłada się na zwiększenie szans na skuteczne leczenie i poprawę rokowań.

Ponadto w pracy doktorskiej badania teoretyczne nad wspomnianymi miarami podobieństwa (nowe metody konstrukcji, takie jak np. prezentowana w twierdzeniu 3.9) zostały rozszerzone na miary modelujące niepewność, czyli entropie przedziałowe (rozdziały 3.3.2 oraz 3.3.3) wraz z badaniem przedziałowych negacji rozmytych dla odpowiednich klas (rozdział 2.2.3). Wspomniane miary następnie zastosowano w rozwiązaniach dwóch kluczowych problemów w klasyfikacji, w szczególności w analizie danych, takich jak braki danych i selekcja atrybutów (rozdział 5).

Problem brakujących wartości jest często występującym szczególnie w danych medycznych. Badania dotyczące brakujących wartości, zapoczątkowane w pracach [47, 48], oraz [49], zostały rozszerzone o informacje teoretyczne i praktyczne dotyczące entropii optymistycznej (możliwości) i pesymistycznej (konieczności), zastosowanej w autorskim algorytmie do uzupełniania brakujących wartości (rozdział 5.2).

Selekcja atrybutów stanowi kluczowy etap w analizie danych, a jej istotność wynika z kilku powodów. Po pierwsze, w medycynie często mamy do czynienia z dużą liczbą cech lub parametrów, które mogą być zbierane od pacjenta. Nie wszystkie atrybuty mogą mieć równie istotny wpływ na proces diagnozowania czy podejmowania decyzji klinicznych. Selekcja atrybutów pozwala na identyfikację cech najbardziej istotnych pod względem niesionych przez nie informacji, które rzeczywiście wpływają na analizowany problem. Redukcja liczby atrybutów może znacząco wpłynąć na efektywność procesu analizy i klasyfikacji. Często mniejsza liczba atrybutów przekłada się na szybsze i bardziej efektywne przetwarzanie danych. W przypadku analizy medycznej, gdzie czas może być kluczowy dla pacjenta, taka optymalizacja jest szczególnie cenna. Dodatkowo pozwala na wyłonienie tych cech, które rzeczywiście mają diagnostyczne znaczenie. Warto zwrócić również uwagę na to, że mniejsza liczba badanych parametrów może oznaczać mniej inwazyjne badania lub mniejsze obciążenie dla pacjenta. Efektywne rozwiązania służące selekcji atrybutów z wykorzystaniem miar podobieństwa zaprezentowane zostały w pracach [50, 51].

Innym rozważanym zagadnieniem było opracowanie metody wczesnego wykrywania depresji opartej o autorski test diagnostyczny oraz metodę klasyfikacji (algorytmy 4 i 5), która dawała najefektywniejsze rezultaty dla stosowanej w niej miary podobieństwa z klasy optymistycznej (możliwości). Według szacunków WHO, na depresję cierpi 3,8% populacji, z czego 5% stanowią dorośli. W przypadku braku odpowiedniego leczenia, może ona prowadzić do samobójstwa, będącego częstą przyczyną śmierci wśród osób młodych. W celu skuteczniejszej prewencji depresji opracowano nowy test diagnostyczny. W porównaniu z testem CESD-R, test ten okazał się bardziej wrażliwy, umożliwiając wcześniejsze wykrycie epizodów depresyjnych u badanych osób. Potwierdzono to w procesie klasyfikacji [38].

Kolejne badania dotyczące przewidywania ryzyka zachorowania na chorobę sercowo-naczyniową (CVD) w ciągu dziesięciu lat były istotnym aspektem analiz badawczych

nad wpływem badanych klas agregacji na klasyfikację. Wartość ryzyka CVD została sklasyfikowana w trzech kategoriach: niska, średnia lub wysoka. W trakcie tych badań zapoczątkowanych w [52] oraz [53], a rozwijanych w pracy doktorskiej skoncentrowano się na rozwiązaniu tego problemu jako problemu klasyfikacji, wykorzystując system oparty na regułach rozmytych o wartościach przedziałowych w kontekście wpływu różnych klas agregacji: odpowiednio w sensie klasycznym (zgodne z porządkiem częściowym lub liniowym), pre-agregacyjnym (definicja 11), czy optymistycznym (możliwości) i pesymistycznym (konieczności) (definicje 7 i 8). Eksperymenty przeprowadzono na danych klinicznych z ośrodków opieki zdrowotnej w Pampelunie. Dane dostarczone przez lekarzy były niepełne, co skłoniło do zastosowania przedziałowych zbiorów rozmytych z wykorzystaniem funkcji ignorancji. Wyniki zastosowania uogólnionego złożenia przedziałowych relacji rozmytych okazały się bardzo skuteczne i zostały szczegółowo opisane w rozdziale 6.1 niniejszej rozprawy doktorskiej.

Ponadto w toku prac nad problemem wyznaczania ryzyka choroby sercowo-naczyniowej zaobserwowano brak publicznej biblioteki w języku Python obsługującej przedziałowe wnioskowanie rozmyte, dlatego we współpracy z zespołem stworzona została dedykowana biblioteka związana z przedziałowym wnioskowaniem rozmytym, która została z sukcesem zastosowana do badań rozważanych w rozdziale 6.1. Opis wspomnianej biblioteki przedstawiony został w artykułach [54, 55], a pierwsza wersja biblioteki dostępna jest na platformie GitHub: <https://github.com/PGrochowalski/ifis>.

Zatem można powiedzieć, że analizy skoncentrowały się na opracowywaniu metod, które wykorzystują wiedzę niepewną w celu usprawnienia klasyfikatorów oraz rozwiązywaniu problemów predykcji w warunkach nieprecyzyjności. W tym celu użyto trzech klas miar (klasycznej, optymistycznej i pesymistycznej) i poddano je testom oraz porównaniom.

Do głównych osiągnięć badawczych uzyskanych w ramach pracy doktorskiej można zaliczyć:

Opracowanie miar i metod uwzględniających niepewność i nieprecyzyjność umożliwiające efektywną diagnostykę medyczną oraz przeprowadzenie weryfikacji przydatności proponowanych metod poprzez sprawdzenie ich wpływu na jakość klasyfikacji. W szczególności:

- Opracowanie algorytmu/techniki progowania dla przypisania klas na podstawie ich oceny za pomocą wartości przedziałowych;
- Stworzenie modelu klasyfikacji uwzględniającego niepewność i nieprecyzyjność opartego o przedziałowe miary podobieństwa;
- Zaproponowanie nowej metody konstrukcji przedziałowej miary podobieństwa zgodnej z porządkiem częściowym lub liniowym, czyli w ujęciu klasycznym;
- Zaproponowanie nowych negacji przedziałowych z klasy epistemicznej (możliwości i konieczności);

- Opracowanie i eksperymentalna weryfikacja skuteczności autorskiej metody selekcji cech opartej o przedziałową miarę podobieństwa;
- Zaproponowanie metod konstrukcji nowych miar niepewności, czyli entropii przedziałowych w duchu epistemicznym;
- Opracowanie i eksperymentalna weryfikacja skuteczności autorskiej metody uzupełniania brakujących wartości opartej o przedziałową miarę entropii;
- Opracowanie algorytmu klasyfikacji dotyczącej prewencji depresji, czyli systemu wczesnego wykrywania zagrożenia depresją przy wykorzystaniu miar podobieństwa i entropii przy odpowiednim porządku. Badania uwzględniały porządek klasyczny, możliwości oraz konieczności;
- Zaproponowanie optymalnej agregacji przedziałowej dla przedziałowego klasyfikatora regulowego w problemie sercowo-naczyniowym.

1.3 Struktura rozprawy

Rozprawa składa się z sześciu rozdziałów, które można podzielić na trzy części tematyczne.

Część I - Wprowadzenie teoretyczne.

Rozdział 2 prezentuje podstawowe pojęcia i zagadnienia związane z teorią zbiorów rozmytych, szczególnie skupiając się na przedziałowych zbiorach rozmytych. Opisuje również różne operacje oraz własności ich dotyczące tj. relacje porządkujące, agregacje czy negacje przedziałowe, które są wykorzystywane w dalszej części rozprawy.

W trzecim rozdziale przedstawiono analizę teoretyczną dotyczącą wprowadzonych nowych miar uwzględniających niepewność. Wśród tych miar znajdują się wskaźnik pierwszeństwa, miara podobieństwa oraz miara entropii przedziałowych zbiorów rozmytych. Co istotne, każda z tych miar została poddana analizie w kontekście trzech kluczowych kategorii: klasycznej, optymistycznej (możliwości) oraz pesymistycznej (konieczności).

Część II - Metody klasyfikacji i metody pomiaru ich efektywności.

Rozdział 4 skupia się na przeglądzie różnych metod klasyfikacji, ilustrując podział tych metod oraz ich przegląd. Przedstawione są również konkretne przykłady metod, które zostały krótko opisane, aby zobrazować, jak funkcjonują w praktyce. W ramach tego rozdziału omawiane są również metody oceny efektywności klasyfikatorów, co pozwala na zrozumienie, w jaki sposób można mierzyć skuteczność różnych podejść.

Część III - Eksperymenty i wnioski.

Rozdział 5 przedstawia szczegółowy opis przeprowadzonych badań eksperymentalnych, charakterystykę danych użytych do testów oraz wyniki eksperymentów mających na celu sprawdzenie efektywności zaproponowanych metod (zob. 5.1.1, 5.2.1, 5.3.1, czy 5.4.2). W tym rozdziale dokonano uporządkowania wyników i zestawiono optymalne rezultaty, które posłużyły do wyciągnięcia wniosków końcowych i stanowi główny opis autorskiego

wkładu autora w proces badawczy.

Rozdział 6 zawiera opis podjętych innych obszarów zastosowań dla badanych uogólnionych operatorów agregacji i ich wpływu na efektywność systemu klasyfikacji w problemie sercowo naczyniowym.

Natomiast rozdział 7 stanowi podsumowanie rozprawy, prezentuje wnioski płynące z przeprowadzonych badań oraz wskazuje kierunki dalszych możliwych badań w dziedzinie modelowania systemów decyzyjnych z wykorzystaniem miar zawierania, podobieństwa czy entropii przedziałowych zbiorów rozmytych i operatorów agregacji przedziałowych uwzględniających niepewność/nieprecyzyjność.

ROZDZIAŁ 2

Podstawowe pojęcia teorii klasycznych i przedziałowych zbiorów rozmytych

Pojęcie zbioru rozmytego zostało zaproponowane przez Zadeha [1] w 1965 roku. W późniejszym czasie nastąpił dynamiczny rozwój teorii zbiorów rozmytych wraz z rozwojem kierunków ich zastosowań praktycznych. Jednym z pierwszych takich zastosowań było rozwiązanie problemu sterowania piecem do wytwarzania cementu [56]. Niewątpliwie wielkim sukcesem było opracowanie pierwszego na świecie systemu sterowania metrem wykorzystującego logikę rozmytą do kontroli jego prędkości. System ten zapewniał płynność startów i zatrzymań. Udało się w ten sposób uzyskać zużycie energii mniejsze o 10% w porównaniu z maszyną kontrolowaną przez człowieka. Logika rozmyta znalazła również zastosowanie w projektowaniu układów zasilających [57], a także w planowaniu rozwoju sieci elektroenergetycznych [58]. Późniejsze badania umożliwiły zastosowanie logiki rozmytej nie tylko w przemyśle ale też w urządzeniach codziennego użytku.

W tym rozdziale przedstawiono podstawowe pojęcia z teorii zbiorów rozmytych, dotyczące operacji na zbiorach, rozszerzeń tych zbiorów, jakimi są przedziałowe zbiory rozmyte, oraz relacji porządkujących.

2.1 Pojęcie zbioru rozmytego

Logika rozmyta [1] stanowi rozszerzenie klasycznego rozumowania na rozumowanie bliższe ludzkiemu. Zamiast dwóch wartości logicznych (prawda i fałsz) rozważa się nieskończenie wiele wartości prawdy zawartych w przedziale $[0,1]$ co pozwala modelować pojęcia nieprecyzyjne i wieloznaczne tj.: np. wysoki hałas, małe zarobki, niskie zużycie paliwa. Nieprecyzyjność tych pojęć wynika nie tylko z braku jasno określonej wartości

granicznej, ale też z zależności warunków, do jakich odnosimy się w konkretnej sytuacji np. "niskie zużycie paliwa" oznacza co innego w przypadku gdy odnosimy się do jazdy w ruchu miejskim, co innego podczas jazdy wyłącznie drogami szybkiego ruchu, a jeszcze inne podczas jazdy po torze.

Zbiór rozmyty A na przestrzeni $X \neq \emptyset$ można przedstawić jako zbiór uporządkowanych par $A = (x, \mu_A(x)) : x \in X$, gdzie $\mu_A : X \rightarrow [0, 1]$ jest funkcją przynależności, a $\mu_A(x)$ oznacza stopień przynależności elementu x do zbioru A . Dla $\mu_A(x)$ będziemy stosować oznaczenie $A(x)$.

Zatem, zbiór rozmyty jest utożsamiany często z funkcją przynależności $\mu_A : X \rightarrow [0, 1]$. Często do określania zbiorów rozmytych wykorzystuje się poniższy zapis.

$$A = a_1/x_1 + a_2/x_2 + \dots + a_k/x_k. \quad (2.1)$$

Oznacza on zbiór rozmyty w którym $\mu_A(x_i) = a_i$ dla każdego $i \in \{1, 2, \dots, k\}$ oraz $\mu_A(x) = 0$ dla pozostałych $x \in X$. Rodzinę wszystkich zbiorów rozmytych w danym uniwersum X oznaczamy przez $FS(X)$.

Podstawowe operacje na zbiorach rozmytych

Do podstawowych operacji na zbiorach rozmytych zalicza się: sumę zbiorów rozmytych, iloczyn i dopełnienie. Ważnym zagadnieniem wykorzystywanym w wielu kwestiach praktycznych, w szczególności w konstrukcjach wielu operacji rozmytych, jest negacja rozmyta.

Negacja rozmyta N jest uogólnieniem klasycznego dopełnienia lub negacji. Poniższe definicje można znaleźć w każdym wprowadzającym podręczniku do logiki rozmytej np. [59, 60, 61].

Definicja 1. Funkcja negacji rozmytej to funkcja nierosnąca $N : [0, 1] \rightarrow [0, 1]$ spełniająca warunki brzegowe $N(0) = 1$ i $N(1) = 0$. Silnie malejące i ciągłe negacje rozmyte nazywamy negacjami ścisłymi, natomiast inwolutywne negacje rozmyte (tzn. te dla wszystkich $x \in [0, 1]$ weryfikujące $N(N(x)) = x$) nazywamy negacjami silnymi i stanowią one podklasę negacji ścisłych.

Przykład 1 ([62]). Typowymi przykładami negacji rozmytych są:

- $N(x) = 1 - x$, jest silną negacją i jest nazywana klasyczną bądź standardową negacją,
- $N(x) = 1 - x^2$, jest ścisłą negacją, ale nie silną,
- $N_S^\lambda(x) = \frac{1-x}{1+\lambda x}$, rodzina Sugeno rozmytych (silnych) negacji, gdzie $\lambda \in (-1, \infty)$, a dla $\lambda = 0$ otrzymujemy klasyczną negację rozmytą.

Definicja 2 ([1]). Niech $A, B \in FS(X)$. Iloczyn dwóch zbiorów rozmytych A i B oznaczany „ $\cap$ ” jest zdefiniowany jako:

$$(A \cap B)(x) = \min(A(x), B(x)), x \in X. \quad (2.2)$$

Suma dwóch zbiorów A i B oznaczana „ $\cup$ ” jest określona jako:

$$(A \cup B)(x) = \max(A(x), B(x)), x \in X. \quad (2.3)$$

Dopełnienie A' do A względem N definiuje się następująco:

$$A'(x) = N(A(x)), x \in X. \quad (2.4)$$

2.2 Rozszerzenie zbiorów rozmytych do przedziałowych zbiorów rozmytych

2.2.1 Relacje porządku na przedziałach

Ogromną zaletą przedziałowych zbiorów rozmytych jest wprowadzenie możliwości wyrażania niepewności co do rzeczywistej wartości funkcji przynależności, która to może znajdować się gdzieś pomiędzy podanymi granicami przedziału. Aczkolwiek wtedy dużym wyzwaniem w aspektach praktycznych jest porównywalność przedziałów, czy przedziałowych zbiorów rozmytych.

Przyjmujemy w pracy następujące oznaczenia:

Przez L^I oznaczamy rodzinę wszystkich przedziałów zawartych w $[0,1]$,

$$L^I = \{[\underline{x}, \bar{x}] : \underline{x}, \bar{x} \in [0, 1], \underline{x} \leq \bar{x}\}. \quad (2.5)$$

Koncentrujemy się w tym rozdziale na tym jak porównać ze sobą przedziały.

Porządki w ujęciu klasycznym:

Porządek częściowy

Dobrze znanym i często używanym do porównywania przedziałów jest następujący porządek częściowy w L^I :

$$[\underline{a}, \bar{a}] \leq_{L^I} [\underline{b}, \bar{b}] \Leftrightarrow \underline{a} \leq \underline{b} \text{ i } \bar{a} \leq \bar{b}, \quad (2.6)$$

przy czym zwykle definiujemy dodatkowo relację $<_{L^I}$:

$$[\underline{a}, \bar{a}] <_{L^I} [\underline{b}, \bar{b}] \Leftrightarrow [\underline{a}, \bar{a}] \leq_{L^I} [\underline{b}, \bar{b}] \text{ i } (\underline{a} < \underline{b} \text{ lub } \bar{a} < \bar{b}).$$

Rodzina wszystkich przedziałów z porządkiem $\leq_{L^I}$ jest częściowo uporządkowana, a ponadto jest kratą. Operacje sumy (*ang. join*) i iloczynu (*ang. meet*) definiuje się następująco:

$$[\underline{a}, \bar{a}] \vee [\underline{b}, \bar{b}] = [\max(\underline{a}, \underline{b}), \max(\bar{a}, \bar{b})],$$

$$[\underline{a}, \bar{a}] \wedge [\underline{b}, \bar{b}] = [\min(\underline{a}, \underline{b}), \min(\bar{a}, \bar{b})].$$

Warto zauważyć, że struktura $(L^I; \vee; \wedge)$ jest ponadto kratą zupełną, z częściowym porządkiem $\leq_{L^I}$ i z elementem największym $1_{L^I} = [1, 1]$ oraz elementem najmniejszym $0_{L^I} = [0, 0]$.

Porządek liniowy

Ponieważ w wielu rzeczywistych problemach potrzebujemy porządku liniowego, aby móc porównywać dowolne dwa przedziały, zatem interesuje nas rozszerzenie porządku częściowego $\leq_{L^I}$ na porządek liniowy. Przydatne będzie pojęcie tzw. porządku dopuszczalnego (*ang. admissible*). Pojęcie porządku dopuszczalnego zostało wprowadzone w [63] i szerzej badane m.in. w pracach [64] i [65]. Liniowość porządku jest konieczna w wielu zastosowaniach w rzeczywistych problemach, aby móc porównywać dowolne dwie dane niepewne [66].

Definicja 3 ([63]). Porządek $\leq_{Adm}$ w L^I nazywany porządkiem dopuszczalnym, jeśli jest on liniowy i dla wszystkich $x, y \in L^I$, takich, że gdy $x \leq_{L^I} y$, wtedy zachodzi $x \leq_{Adm} y$.

W dalszej części pracy będą stosowane także inne oznaczenia dla porządku dopuszczalnego $\leq_{Adm}$ w zależności od przyjętej metody konstrukcji, np. $\leq_{B_{1,2}}$, czy w szczególności $\leq_{Lex1}$ lub $\leq_{Lex2}$.

Twierdzenie 2.1 ([63]). Niech $B_1, B_2 : [0, 1]^2 \rightarrow [0, 1]$ będą dwiema ciągłymi agregacjami rozmytymi [67] takimi, że dla wszystkich $x = [\underline{x}, \bar{x}], y = [\underline{y}, \bar{y}] \in L^I$, zachodzą równości $B_1(\underline{x}, \bar{x}) = B_1(\underline{y}, \bar{y})$ i $B_2(\underline{x}, \bar{x}) = B_2(\underline{y}, \bar{y})$ wtedy i tylko wtedy gdy $x = y$. Jeżeli $x \leq_{B_{1,2}} y$ wtedy i tylko wtedy gdy

$$B_1(\underline{x}, \bar{x}) < B_1(\underline{y}, \bar{y}) \text{ lub } (B_1(\underline{x}, \bar{x}) = B_1(\underline{y}, \bar{y}) \text{ i } B_2(\underline{x}, \bar{x}) \leq B_2(\underline{y}, \bar{y})),$$

wtedy $\leq_{B_{1,2}}$ jest porządkiem dopuszczalnym w L^I .

Przykład 2 ([63]). Przykładami dopuszczalnych porządków w L^I są:

- porządek Xu i Yagera

$$[\underline{x}, \bar{x}] \leqslant_{XY} [\underline{y}, \bar{y}] \Leftrightarrow \underline{x} + \bar{x} < \underline{y} + \bar{y} \text{ lub } (\bar{x} + \underline{x} = \bar{y} + \underline{y} \text{ i } \bar{x} - \underline{x} \leqslant \bar{y} - \underline{y}),$$

- pierwszy porządek leksykograficzny (w odniesieniu do dolnej granicy przedziału)

$$[\underline{x}, \bar{x}] \leqslant_{Lex1} [\underline{y}, \bar{y}] \Leftrightarrow \underline{x} < \underline{y} \text{ lub } (\underline{x} = \underline{y} \text{ i } \bar{x} \leqslant \bar{y}),$$

- drugi porządek leksykograficzny (w odniesieniu do górnej granicy przedziału)

$$[\underline{x}, \bar{x}] \leqslant_{Lex2} [\underline{y}, \bar{y}] \Leftrightarrow \bar{x} < \bar{y} \text{ lub } (\bar{x} = \bar{y} \text{ i } \underline{x} \leqslant \underline{y}),$$

- porządek $\alpha\beta$

$$\begin{aligned} [\underline{x}, \bar{x}] \leqslant_{\alpha\beta} [\underline{y}, \bar{y}] &\Leftrightarrow K_{\alpha}(\underline{x}, \bar{x}) < K_{\alpha}(\underline{y}, \bar{y}) \text{ lub} \\ &(K_{\alpha}(\underline{x}, \bar{x}) = K_{\alpha}(\underline{y}, \bar{y}) \text{ i } K_{\beta}(\underline{x}, \bar{x}) \leqslant K_{\beta}(\underline{y}, \bar{y})), \end{aligned}$$

dla pewnych $\alpha \neq \beta \in [0, 1]$ i $x, y \in L^I$, gdzie $K_{\alpha} : [0, 1]^2 \rightarrow [0, 1]$ jest zdefiniowane następująco $K_{\alpha}(x, y) = \alpha x + (1 - \alpha)y$.

Porządki $\leqslant_{XY}$, $\leqslant_{Lex1}$ i $\leqslant_{Lex2}$ są szczególnymi przypadkami porządku $\leqslant_{\alpha\beta}$ z odpowiednio $\leqslant_{0,5\beta}$ (dla $\beta > 0, 5$), $\leqslant_{1,0}$, $\leqslant_{0,1}$. Porządki $\leqslant_{XY}$, $\leqslant_{Lex1}$, $\leqslant_{Lex2}$, i $\leqslant_{\alpha\beta}$ są dopuszczalnymi porządkami liniowymi $\leqslant_{B_{1,2}}$ określonymi przez pary funkcji agregacji, czyli średnie ważone. W przypadku porządków $\leqslant_{Lex1}$ i $\leqslant_{Lex2}$ stosuje się agregacje, które są operacjami rzutowania.

Porządki możliwości lub konieczności (epistemiczne)

W 1983 roku F. Allen zaproponował zestaw 13 podstawowych relacji określających porównywalność przedziałów, które wcześniej – w 1979 roku – były również rozważane przez Moore'a.

Alternatywnym podejściem do problemu porównywania wartości przedziałowych jest interpretacja przedziałów jako zbiorów o charakterze epistemicznym (optymistycznym / możliwości lub pesymistycznym / konieczności), zgodnie z koncepcjami przedstawionymi przez Dubois [68, 69].

Zbiór epistemiczny (dysjunktywny) Z można interpretować jako reprezentację niekompletnej informacji o wartości pewnej zmiennej x , która jest nieznana, ale zawiera się w tym zbiorze, czyli $x \in Z$. Natomiast zbiory opisujące elementy składające się na dany obiekt nazywane są ontycznymi (koniunktywnymi).

Wspomniane relacje porządkujące możliwości i konieczności można wykorzystać na przykład w modelu decyzyjnym do reprezentowania niepewności. Wiele procesów decyzyjnych odbywa się w środowisku, w którym informacje nie są dokładnie znane. W związku z tym wartości przedziałowe można uznać za odpowiedni format reprezentacji, który pozwala lepiej uchwycić niepewne informacje. Definiuje się następujące relacje porządku (porównywalności) na zbiorze L^I :

- **Relacja konieczności - pesymistyczna** (*ang. necessary*).

Definiujemy następującą relację porządku na przedziałach:

$$[\underline{x}, \bar{x}] \leq_{nec} [\underline{y}, \bar{y}] \Leftrightarrow \bar{x} \leq \underline{y}.$$

W L^I relacja $\leq_{nec}$ jest antysymetryczna, przechodnia i ma własność Ferrersa.

- **Relacja możliwości - optymistyczna** (*ang. possible*).

Zdefiniowana jest następująco:

$$[\underline{x}, \bar{x}] \leq_{pos} [\underline{y}, \bar{y}] \Leftrightarrow \underline{x} \leq \bar{y}.$$

W strukturze $(L^I, \leq_{pos})$ relacja $\leq_{pos}$ jest zupełna i ma własność Ferrersa (porządek przedziałowy). Jeśli $[\underline{x}, \bar{x}]$ jest nieprecyzyjnym opisem wartości zmiennej x , a $[\underline{y}, \bar{y}]$ jest nieprecyzyjnym opisem zmiennej y , to $[\underline{x}, \bar{x}] \leq_{pos} [\underline{y}, \bar{y}]$ oznacza, że jest możliwe, iż wartość zmiennej x jest mniejsza lub równa od wartości zmiennej y [70].

Możemy zauważyć, następującą zależność:

$$x \leq_{nec} y \Rightarrow x \leq_{L^I} y \Rightarrow x \leq_{pos} y.$$

2.2.2 Przedziałowe agregacje rozmyte

Pierwszą rozważaną tu grupą operacji na przedziałach, zdefiniowaną względem relacji porządku, są przedziałowe funkcje agregacji w L^I . Początkowo przedstawione zostaną przedziałowe funkcje agregacji zarówno względem $\leq_{L^I}$, jak i $\leq_{Adm}$.

Uwaga 2.2. W dalszej części pracy wykorzystano notację $\leq$ zarówno dla częściowego, jak i dopuszczalnego porządku (porządek klasyczny), a w dalszej kolejności dla relacji $\leq_{pos}$ (relacja możliwości) i $\leq_{nec}$ (relacja konieczności), przy czym 0_{L^I} i 1_{L^I} oznaczają odpowiednio, minimalny i maksymalny element w L^I , czyli odpowiednio, przedziały $[0,0]$ i $[1,1]$.

Przyjmujemy następujące nazewnictwo: operacje zdefiniowane względem relacji $\leq_{L^I}$ (porządek częściowy) lub pewnego porządku liniowego tworzą rodzinę operatorów klasycznych, a zdefiniowane względem relacji możliwości i konieczności, odpowiednio klasę

operatorów odpowiednio, możliwości i konieczności.

Agregacje klasyczne

Definicja 4 ([65, 71]). Operacja $\mathcal{A} : (L^I)^n \rightarrow L^I$ jest nazywana przedziałową funkcją agregacji, jeśli jest rosnąca względem każdego argumentu w sensie relacji porządku klasycznego na przedziałach $\leq$:

$$\forall_{x_i, y_i \in L^I} x_i \leq y_i \Rightarrow \mathcal{A}(x_i) \leq \mathcal{A}(y_i) \quad (2.7)$$

i spełnia warunki brzegowe:

$$\mathcal{A}(\underbrace{[0, 0], \dots, [0, 0]}_{n \times}) = [0, 0] = 0_{L^I}, \quad \mathcal{A}(\underbrace{[1, 1], \dots, [1, 1]}_{n \times}) = [1, 1] = 1_{L^I}.$$

Specjalną klasą przedziałowych funkcji agregujących są reprezentowalne przedziałowe funkcje agregacji.

Definicja 5 ([7, 72]). Przedziałową funkcją agregacji $\mathcal{A} : (L^I)^n \rightarrow L^I$ nazywamy reprezentowalną, jeśli istnieją rozmyte funkcje agregacji $A_1, A_2 : [0, 1]^n \rightarrow [0, 1]$ takie, że

$$\mathcal{A}(x_1, \dots, x_n) = [A_1(\underline{x}_1, \dots, \underline{x}_n), A_2(\bar{x}_1, \dots, \bar{x}_n)]$$

dla wszystkich $x_1, \dots, x_n \in L^I$.

Przykład 3. Operacje kratowe $\vee$ i $\wedge$ na L^I :

$$\begin{aligned} [\underline{x}, \bar{x}] \vee [\underline{y}, \bar{y}] &= [\max(\underline{x}, \underline{y}), \max(\bar{x}, \bar{y})], \\ [\underline{x}, \bar{x}] \wedge [\underline{y}, \bar{y}] &= [\min(\underline{x}, \underline{y}), \min(\bar{x}, \bar{y})] \end{aligned}$$

są przykładami reprezentowalnych funkcji agregacji na L^I względem częściowego porządku $\leq_{L^I}$, z $A_1 = A_2 = \max$ w pierwszym przypadku i $A_1 = A_2 = \min$ w drugim. Jednak $\wedge$ i $\vee$ nie są przedziałowymi funkcjami agregacji względem $\leq_{Lex1}$, $\leq_{Lex2}$ czy $\leq_{XY}$. Poniżej podane zostały inne przykłady reprezentowalnych przedziałowych funkcji agregacji względem $\leq_{L^I}$.

- Rzutowanie:

$$\mathcal{A}_L([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\underline{x}, \bar{x}], \quad \mathcal{A}_R([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\underline{y}, \bar{y}].$$

- Produkt reprezentowalny:

$$\mathcal{A}_{prod}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\underline{xy}, \bar{xy}].$$

- Reprezentowalna średnia arytmetyczna:

$$\mathcal{A}_{mean}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \left[\frac{\underline{x} + \underline{y}}{2}, \frac{\bar{x} + \bar{y}}{2} \right].$$

- Reprezentowalna średnia geometryczna:

$$\mathcal{A}_{gmean}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\sqrt{\underline{x}\underline{y}}, \sqrt{\bar{x}\bar{y}}].$$

- Reprezentowalna średnia harmoniczna:

$$\mathcal{A}_{hmean}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \begin{cases} [0, 0], & \text{jeśli } x = y = [0, 0], \\ \left[\frac{2\underline{x}\underline{y}}{\underline{x} + \underline{y}}, \frac{2\bar{x}\bar{y}}{\bar{x} + \bar{y}} \right], & \text{w przeciwnym przypadku.} \end{cases}$$

- Reprezentowalna średnia kwadratowa:

$$\mathcal{A}_{powermean}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \left[\sqrt{\frac{\underline{x}^2 + \underline{y}^2}{2}}, \sqrt{\frac{\bar{x}^2 + \bar{y}^2}{2}} \right].$$

- Reprezentowalna średnia potęgowa:

$$\mathcal{A}_{meanpow}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \left[\frac{\underline{x} + \underline{y}}{2}, \sqrt{\frac{\bar{x}^2 + \bar{y}^2}{2}} \right].$$

- Reprezentowalny średni produkt:

$$\mathcal{A}_{prodmean}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \left[\underline{x}\underline{y}, \frac{\bar{x} + \bar{y}}{2} \right].$$

- Reprezentowalna średnia maksymalna:

$$\mathcal{A}_{meanmax}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \left[\frac{\underline{x} + \underline{y}}{2}, \max(\bar{x}, \bar{y}) \right].$$

- Agregacja Łukasiewicza:

$$\mathcal{A}_{Łukasiewicz}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\max(0, \underline{x} + \underline{y} - 1), \max(0, \bar{x} + \bar{y} - 1)].$$

Reprezentowalność nie jest jedynym możliwym sposobem konstruowania przedziałowych funkcji agregujących względem porządku $\leq_{L^I}$ lub $\leq_{Adm}$.

Przykład 4. Niech $A : [0, 1]^2 \rightarrow [0, 1]$ będzie rozmytą funkcją agregacji.

- Funkcja $\mathcal{A}_1 : (L^I)^2 \rightarrow L^I$, gdzie

$$\mathcal{A}_1(x, y) = \begin{cases} [1, 1], & \text{gdy } (x, y) = ([1, 1], [1, 1]), \\ [0, A(\underline{x}, \underline{y})], & \text{w innym przypadku,} \end{cases}$$

jest niereprezentowalną przedziałową funkcją agregacji względem $\leq_{L^I}$.

- Funkcja $\mathcal{A}_2, \mathcal{A}_3 : (L^I)^2 \rightarrow L^I$ ([73]), gdzie

$$\mathcal{A}_2(x, y) = \begin{cases} [1, 1], & \text{gdy } (x, y) = ([1, 1], [1, 1]) \\ [0, A(\underline{x}, \underline{y})], & \text{w innym przypadku,} \end{cases}$$

$$\mathcal{A}_3(x, y) = \begin{cases} [0, 0], & \text{gdy } (x, y) = ([0, 0], [0, 0]) \\ [A(\underline{x}, \underline{y}), 1], & \text{w innym przypadku,} \end{cases}$$

są niereprezentowalnymi przedziałowymi funkcjami agregacji względem $\leq_{Lex1}$.

- Funkcja $\mathcal{A}_4, \mathcal{A}_5 : (L^I)^2 \rightarrow L^I$ ([73]), gdzie

$$\mathcal{A}_4(x, y) = \begin{cases} [1, 1], & \text{gdy } (x, y) = ([1, 1], [1, 1]) \\ [0, A(\overline{x}, \overline{y})], & \text{w innym przypadku,} \end{cases}$$

$$\mathcal{A}_5(x, y) = \begin{cases} [0, 0], & \text{gdy } (x, y) = ([0, 0], [0, 0]) \\ [A(\overline{x}, \overline{y}), 1], & \text{w innym przypadku,} \end{cases}$$

są niereprezentowalnymi przedziałowymi funkcjami agregacji względem $\leq_{Lex2}$.

Często dyskutowanym w literaturze i stosowanym w praktyce operatorem agregacji jest operator OWA (uporządkowana średnia ważona) wprowadzony przez Yagera [19]. Koncepcja operatora OWA została rozszerzona do agregacji przedziałów przez Zhou i innych [74], gdzie wagi są reprezentowane za pomocą zbiorów rozmytych.

Definicja 6 ([66]). Niech $\leq$ będzie dopuszczalnym porządkiem na L^I , a $w = (w_1, \dots, w_n) \in [0, 1]^n$, przy czym $w_1 + \dots + w_n = 1$. Przedziałowy operator uporządkowanej średniej ważonej IVOWA (*ang. Interval-Valued Ordered Weighted Averaging Operator*) związany z $\leq$ i w jest odwzorowaniem $IVOWA_{\leq, w} : (L^I)^n \rightarrow L^I$, przez

$$IVOWA_{\leq, w}([x_1, \overline{x}_1], \dots, [x_n, \overline{x}_n]) = \sum_{i=1}^n w_i \cdot [x_i, \overline{x}_i],$$

gdzie $[x_i, \bar{x}_i], i = 1, \dots, n$, oznacza i -ty największy argument w sensie porządku $\leq$, i $w \cdot [\underline{x}, \bar{x}] = [w\underline{x}, w\bar{x}], [\underline{x}_1, \bar{x}_1] + [\underline{x}_2, \bar{x}_2] = [\underline{x}_1 + \underline{x}_2, \bar{x}_1 + \bar{x}_2]$.

Motywacją wprowadzenia w [66] operatora $IVOWA_{\leq, w}$ względem porządku dopuszczalnego był fakt, że $IVOWA_{\leq, w}$ nie jest przedziałową funkcją agregacji względem $\leq_{L^I}$. Ponadto badania wykazały, że także dobór porządku liniowego nie był bez znaczenia, co obrazuje następująca uwaga.

Uwaga 2.3 ([66]). Operator $IVOWA_{\leq_{\alpha\beta}, w}$ jest przedziałową funkcją agregacji w L^I względem porządku $\leq_{\alpha\beta}$.

W szczególności można rozpatrywać funkcję

$$\mathcal{A}_\alpha(x, y) = [\alpha\underline{x} + (1 - \alpha)\underline{y}, \alpha\bar{x} + (1 - \alpha)\bar{y}],$$

która jest przedziałową funkcją agregacji w L^I względem $\leq_{Lex1}$, $\leq_{Lex2}$ i $\leq_{XY}$ dla $x, y \in L^I$ i $\alpha \in [0, 1]$ (por. [65]). Ponadto można zauważyć, że dla $\alpha = \frac{1}{2}$ $\mathcal{A}_\alpha$ sprowadza się do $\mathcal{A}_{mean}$, która jest także funkcją agregacji w odniesieniu do $\leq_{\alpha\beta}$ (por. [64]).

Agregacje możliwości (optymistyczne) i konieczności (pesymistyczne)

Kolejnymi interesującymi i ważnymi klasami agregacji są agregacje zgodne z optymistyczną (*ang. possible*) lub pesymistyczną (*ang. necessary*) relacją porządku.

Definicja 7 ([17]). Funkcja $\mathcal{A} : (L^I)^n \rightarrow L^I$ jest pos-agregacją/agregacją możliwości (inaczej optymistyczną) (*ang. possible aggregation*), jeśli

$$\forall_{x_i, y_i \in L^I} x_i \leq_{pos} y_i \Rightarrow \mathcal{A}(x_1, \dots, x_n) \leq_{pos} \mathcal{A}(y_1, \dots, y_n)$$

i $\mathcal{A}$ posiada warunki brzegowe jak w definicji 4.

Poniżej przedstawiono kilka przykładów agregacji możliwości, w których $A : [0, 1]^2 \rightarrow [0, 1]$ będzie rozmytą funkcją agregacji:

$$\mathcal{A}_{pos1}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \begin{cases} [1, 1], & \text{jeśli } x = y = [1, 1], \\ \left[\frac{\underline{y} + \bar{x}}{2}, \frac{\bar{x} + \bar{y}}{2} \right], & \text{w przeciwnym przypadku,} \end{cases}$$

$$\mathcal{A}_{pos2}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \begin{cases} [1, 1], & \text{jeśli } x = y = [1, 1], \\ \left[\frac{\underline{x} + \bar{y}}{2}, \frac{\bar{x} + \bar{y}}{2} \right], & \text{w przeciwnym przypadku,} \end{cases}$$

$$\mathcal{A}_{pos3}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \begin{cases} [1, 1], & \text{jeśli } x = y = [1, 1], \\ [0, A(\bar{x}, \bar{y})], & \text{w przeciwnym przypadku,} \end{cases}$$

$$\mathcal{A}_{pos4}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \begin{cases} [0, 0], & \text{jeśli } x = y = [0, 0], \\ [A(\underline{x}, \underline{y}), 1], & \text{w przeciwnym przypadku,} \end{cases}$$

$$\mathcal{A}_{pos5}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \begin{cases} [1, 1], & \text{jeśli } x = y = [1, 1], \\ [0, A(\underline{x}, \bar{y})], & \text{w przeciwnym przypadku,} \end{cases}$$

$$\mathcal{A}_{pos6}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \begin{cases} [1, 1], & \text{jeśli } x = y = [1, 1], \\ [0, A(\bar{x}, \underline{y})], & \text{w przeciwnym przypadku,} \end{cases}$$

$$\mathcal{A}_{pos7}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \begin{cases} [0, 0], & \text{jeśli } x = y = [0, 0], \\ [A(\underline{x}, \bar{y}), 1], & \text{w przeciwnym przypadku,} \end{cases}$$

$$\mathcal{A}_{pos8}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \begin{cases} [0, 0], & \text{jeśli } x = y = [0, 0], \\ [A(\bar{x}, \underline{y}), 1], & \text{w przeciwnym przypadku,} \end{cases}$$

$$\mathcal{A}_{pos9}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \begin{cases} [1, 1], & \text{jeśli } x = y = [1, 1], \\ \left[0, \frac{\bar{x}^2 + \bar{y}^2}{\bar{x} + \bar{y}}\right], & \text{w przeciwnym przypadku,} \end{cases}$$

$$\mathcal{A}_{pos10}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \begin{cases} [0, 0], & \text{jeśli } x = y = [0, 0], \\ \left[\frac{\underline{x}^2 + \underline{y}^2}{\underline{x} + \underline{y}}, 1\right], & \text{w przeciwnym przypadku,} \end{cases}$$

$$\mathcal{A}_{pos11}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\underline{x} \cdot |2\underline{y} - 1|, \bar{x}],$$

$$\mathcal{A}_{pos12}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\underline{x} \cdot |2\underline{x} - 1|, \bar{x}],$$

$$\mathcal{A}_{pos13}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\underline{x} \cdot |2\bar{y} - 1|, \bar{x}],$$

$$\mathcal{A}_{pos14}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\underline{x} \cdot |2\bar{x} - 1|, \bar{x}],$$

$$\mathcal{A}_{pos15}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\underline{x} - (\max(0, \underline{x} - \underline{y}))^2, \bar{x}],$$

$$\mathcal{A}_{pos16}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\underline{x} - (\max(0, \underline{x} - \bar{y}))^2, \bar{x}],$$

$$\mathcal{A}_{pos17}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \left[\sqrt{\underline{x}\underline{y}}, \frac{\bar{x} + \bar{y}}{2} \right],$$

$$\mathcal{A}_{pos18}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \left[\min(\underline{x}\bar{y}, \bar{x}\underline{y}), \frac{\bar{x} + \bar{y}}{2} \right],$$

gdzie A w agregacjach $\mathcal{A}_{pos3}$ i $\mathcal{A}_{pos4}$ to średnia geometryczna lub potęgowa. Ponadto agregacjami możliwości są również agregacje $\wedge, \vee, \mathcal{A}_{mean}, \mathcal{A}_{meanmax}$.

Definicja 8 ([17]). Funkcja $\mathcal{A} : (L^I)^n \rightarrow L^I$ jest nec-agregacją/agregacją konieczności (inaczej pesymistyczną) (*ang. necessary aggregation*), jeśli

$$\forall_{x_i, y_i \in L^I} x_i \leq_{nec} y_i \Rightarrow \mathcal{A}(x_1, \dots, x_n) \leq_{nec} \mathcal{A}(y_1, \dots, y_n)$$

i $\mathcal{A}$ posiada warunki brzegowe jak w definicji 4.

Przykład 5. Poniżej przedstawiono kilka przykładów agregacji konieczności:

$$\mathcal{A}_{nec1}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \left[\frac{\underline{x} + \underline{y}}{2}, \max\left(\frac{\underline{x} + \bar{y}}{2}, \frac{\bar{x} + \underline{y}}{2}\right) \right],$$

$$\mathcal{A}_{nec2}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \left[\min\left(\frac{\bar{x} + \underline{y}}{2}, \frac{\underline{x} + \bar{y}}{2}\right), \frac{\bar{x} + \bar{y}}{2} \right],$$

$$\mathcal{A}_{nec3}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\underline{x}\underline{y}, \max(\underline{x}\bar{y}, \bar{x}\underline{y})],$$

$$\mathcal{A}_{nec4}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\min(\underline{x}\bar{y}, \bar{x}\underline{y}), \bar{x}\bar{y}],$$

$$\mathcal{A}_{nec5}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\underline{x}\bar{x}, \underline{x}\bar{x}],$$

$$\mathcal{A}_{nec6}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\underline{x}^2, \underline{x}\bar{x}],$$

$$\mathcal{A}_{nec7}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = [\underline{x}\bar{x}, \bar{x}^2],$$

$$\mathcal{A}_{nec8}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \left[\frac{\underline{y} + \frac{\underline{x} + \bar{x}}{2}}{2}, \frac{\bar{x} + \bar{y}}{2} \right],$$

$$\mathcal{A}_{nec9}([\underline{x}, \bar{x}], [\underline{y}, \bar{y}]) = \left[\frac{\underline{x} + \bar{x}}{2}, \frac{\underline{x}^2 + \bar{x}^2}{\underline{x} + \bar{x}} \right].$$

Również agregacje $\wedge, \vee, \mathcal{A}_{mean}$ są nec-agregacjami.

W dalszej części pracy przez $\mathcal{A}_{pos}$ i $\mathcal{A}_{nec}$ będziemy oznaczać odpowiednio rodziny agregacji możliwości i konieczności.

Pre-agregacje

Ważnym do dalszych rozważań będzie pojęcie kierunkowej monotoniczności operatorów przedziałowych, która posłużyła do zdefiniowania pre-agregacji i zbadania potencjału aplikacyjnego tej klasy agregacji. W szczególności w rozdziale 5 przedstawiono potencjał aplikacyjny wskazanej pre-agregacji w uogólnionym wnioskowaniu rozmytym.

Definicja 9 ([75]). Niech $v = ((a_1, b_1), \dots, (a_n, b_n)) \in (R^2)^n$ taki, że $(a_i, b_i) \neq \vec{0}$ dla niektórych $i \in \{1, \dots, n\}$. Funkcja $F : (L^I)^n \rightarrow L^I$ uważa się, że jest v -rosnąca (v -malejąca), jeżeli dla wszystkich $x \in (L^I)^n$ i $c > 0$ takich, że $x + cv \in (L^I)^n$, spełniony jest warunek

$$F(x) \leq_{L^I} F(x + cv) \quad (F(x) \geq_{L^I} F(x + cv)).$$

W przypadku, gdy F jest jednocześnie v -rosnąca i v -malejąca, oznacza to, że jest v -stała.

Twierdzenie 2.4 ([75]). Niech $a, b > 0$ i $u, v \in (R^2)^n \setminus \{0\}$ takie, że dla wszystkich $x \in (L^I)^n$ i $c > 0$ co spełnia $x + c(au + bv) \in (L^I)^n$, to też zawiera $x + cau \in (L^I)^n$ lub $x + cbv \in (L^I)^n$. Wówczas, jeśli funkcja $F : (L^I)^n \rightarrow L^I$ jest zarówno u -rosnąca i v -rosnąca, wtedy F jest $(au + bv)$ -rosnąca.

Podobnie, można rozszerzyć pojęcie słabej monotoniczności na zbiory przedziałowe.

Definicja 10 ([75]). Niech $\vec{0} \neq (a, b) \in R^2$. Funkcja $F : (L^I)^n \rightarrow L^I$ jest (a, b) -słabo rosnąca ((a, b) -słabo malejąca), jeśli dla wszystkich $x \in (L^I)^n$ i $c > 0$ takich, że $x + c(a, b, \dots, a, b) \in (L^I)^n$, spełniony jest warunek

$$F(x) \leq_{L^I} F(x + c(a, b, \dots, a, b)) \quad (F(x) \geq_{L^I} F(x + c(a, b, \dots, a, b))).$$

Przykład 6. Niech $F : (L^I)^2 \rightarrow L^I$ będzie przedziałową funkcją określoną dla $[\underline{x}_1, \overline{x}_1], [\underline{x}_2, \overline{x}_2] \in L^I$ następująco

$$F([\underline{x}_1, \overline{x}_1], [\underline{x}_2, \overline{x}_2]) = \frac{1}{2} [\underline{x}_1 + \underline{x}_2, \min(\underline{x}_1 + \overline{x}_2, \overline{x}_1 + \underline{x}_2)].$$

Dla $a, b \in R$, F jest (a, b) -słabo rosnąca wtedy i tylko wtedy gdy $a > 0$ i $a + b \geq 0$, lub $a = 0$ i $b > 0$. Jest to konsekwencją faktu, że

$$F([\underline{x}_1, \overline{x}_1], [\underline{x}_2, \overline{x}_2]) + c(a, b, a, b) = F([\underline{x}_1, \overline{x}_1], [\underline{x}_2, \overline{x}_2]) + c \left[a, \frac{a+b}{2} \right]$$

dla $c > 0$.

W artykule [52] zbadano wpływ na wnioskowanie rozmyte nowego przedziałowego operatora, którym jest przedziałowa funkcja pre-agregacji (*ang. IV pre-aggregation function*).

Został on wykorzystany w uogólnionym złożeniu używanym w algorytmie wnioskowania. Główną przyczyną zastosowania wprowadzonego operatora była jego zdolność do zawężania wartości przedziału. Wprowadzony operator nie jest monotoniczny pod względem rozpatrywanych porządków ($\leq_{L^I}$, $\leq_{Adm}$ lub $\leq_{nec}$, $\leq_{pos}$). Natomiast związany jest z monotonicznością kierunkową, co prowadzi do koncepcji funkcji pre-agregacji.

Definicja 11 ([75]). Funkcja $F : (L^I)^n \rightarrow L^I$ jest przedziałową funkcją pre-agregacji, jeżeli spełnia następujące warunki:

- I $F([0, 0], \dots, [0, 0]) = [0, 0]$;
- II $F([1, 1], \dots, [1, 1]) = [1, 1]$;
- III Istnieje wektor $v = ((a_1, b_1), \dots, (a_n, b_n)) \in ((R^+)^2)^n$, taki, że F jest v -rosnąca.

Wprowadzono nowy operator przedziałowy nazwany słabym operatorem agregacji:

Definicja 12 ([52]). Niech $n \in \mathbb{N}$, $i \in \{1, \dots, n\}$ i $[x_i, y_i] \subset [0, 1]$ będzie operatorem przedziałowym. Operator zdefiniowany następującym wzorem

$$W(\underline{x}, \bar{x}, \underline{y}, \bar{y}) = \left[\frac{\underline{x} + \underline{y} + \bar{x}(\bar{x} - \underline{x}) + \bar{y}(\bar{y} - \underline{y})}{2 + (\bar{x} - \underline{x}) + (\bar{y} - \underline{y})}, \frac{\bar{x} + \bar{y} + \underline{x}(\bar{x} - \underline{x}) + \underline{y}(\bar{y} - \underline{y})}{2 + (\bar{x} - \underline{x}) + (\bar{y} - \underline{y})} \right] \quad (2.8)$$

będziemy nazywać słabym operatorem agregacji.

Na początku rozważono kierunkową monotoniczność słabego operatora z wektorami bazowymi. Doprowadziło to do braku monotoniczności. Dlatego w dalszej części pracy rozważono inne wektory dla monotoniczności kierunkowej.

Przykład 7. Niech $x = [2/8, 3/8]$, $y = [0, 0]$ i $c = 0.08$.

Wtedy $W(x, y) = [0.139, 0.191]$;

$W(x + c[1, 0], y + c[0, 0]) = [0.169, 0.190]$,

tj. W nie jest $((1, 0), (0, 0))$ - rosnąca.

Na podstawie powyższego przykładu otrzymano

Uwaga 2.5. [52] Niech $W : L^2 \rightarrow L$ będzie słabym operatorem określonym przez (2.8).

- W nie jest $((1, 0), (0, 0))$ - rosnąca;
- W nie jest $((0, 1), (0, 0))$ - rosnąca;
- W nie jest $((0, 0), (1, 0))$ - rosnąca;
- W nie jest $((0, 0), (0, 1))$ - rosnąca.

Natomiast można zauważyć następującą własność.

Twierdzenie 2.6 ([52]). Niech $W : L^2 \rightarrow L$ jest słabym operatorem określonym przez (2.8).

W jest $(a, b, 0, 0)$ - rosnąca dla wszystkich $a > 0$, $b > 0$ takich, że $\frac{1}{3}a \leq b \leq 3a$.

W podobny sposób można otrzymać następujące twierdzenie.

Twierdzenie 2.7 ([52]). *Niech $W : L^2 \rightarrow L$ jest słabym operatorem określonym przez (2.8).*

W jest (a, b) - słabo rosnący dla wszystkich $a \geq 0, b \geq 0$.

Wniosek 2.8. Słaby operator $W : L^2 \rightarrow L$ dany przez (2.8) jest przedziałową funkcją pre-agregacji.

Ponadto, porównując słaby operator z min i max, możemy zauważyć:

Twierdzenie 2.9 ([52]). *Niech $W : L^2 \rightarrow L$ jest słabym operatorem określonym przez (2.8). Wtedy*

$$\underline{\min}(x, y) \leq \underline{W}(x, y), \quad \overline{W}(x, y) \leq \overline{\max}(x, y).$$

Przykład 8. Niech $x = [0, 3/8]$, $y = [0, 1/8]$, $u = [0, 1/2]$, $v = [1/2, 1/2]$. Wtedy

$$\min(x, y) = [0, 1/8], \quad W(x, y) = [1/16, 1/5],$$

$$\min(u, v) = [0, 1/2], \quad W(x, y) = [1/4, 1/3].$$

Oznacza to, że W i min są nieporównywalne.

Analogicznie możemy otrzymać, że W i max są nieporównywalne.

2.2.3 Przedziałowe negacje rozmyte

Przedstawiona zostanie definicja przedziałowej negacji rozmytej zgodnej z porządkiem klasycznym (częściowym lub liniowym) oraz możliwości (optymistycznej) i konieczności (pesymistycznej).

Negacje klasyczne

Poniżej przedstawiona została definicja przedziałowej negacji rozmytej, która jest zgodna z porządkiem klasycznym $\leq$, czyli częściowym ($\leq_{L^I}$) lub liniowym ($\leq_{Adm}$).

Definicja 13 ([9, 76]). Przedziałowa negacja rozmyta jest funkcją $N_{IV} : L^I \rightarrow L^I$ malejącą względem porządku $\leq$, czyli taką, że

$$\forall_{x, y \in L^I} x \leq y \Rightarrow N_{IV}(y) \leq N_{IV}(x)$$

i spełniającą własności $N_{IV}(0_{L^I}) = 1_{L^I}$ i z $N_{IV}(1_{L^I}) = 0_{L^I}$. Przedziałową negację rozmytą nazywa się za inwolutywną (silną), jeżeli spełnia ona warunek $N_{IV}(N_{IV}(x)) = x$ dla każdego $x \in L^I$.

Przykład 9 ([65]). Funkcja $N_{IV}(x) : L^I \rightarrow L^I$ określona przez

$$N_{IV}([x, \bar{x}]) = [1 - \bar{x}^2, 1 - x^2]$$

jest przedziałową negacją rozmytą względem porządku $\leq_{L^I}$.

$$N_{XY}([x, \bar{x}]) = \begin{cases} [1, 1], & \text{jeśli } x = [0, 0], \\ [0, 0], & \text{jeśli } x = [1, 1], \\ \left[\frac{1-x}{2}, 1 - \frac{\bar{x}}{2}\right], & \text{w przeciwnym przypadku} \end{cases}$$

jest przedziałową negacją rozmytą względem porządków $\leq_{XY}$ i $\leq_{\alpha\beta}$.

$$N_{lex1}([x, \bar{x}]) = \begin{cases} [0, 0], & \text{jeśli } x = [1, 1], \\ [1 - x, 1], & \text{w przeciwnym przypadku} \end{cases}$$

jest przedziałową negacją rozmytą względem porządku $\leq_{lex1}$.

$$N_{lex2}([x, \bar{x}]) = \begin{cases} [1, 1], & \text{jeśli } x = [0, 0], \\ [0, 1 - \bar{x}], & \text{w przeciwnym przypadku} \end{cases}$$

jest przedziałową negacją rozmytą względem porządku $\leq_{lex2}$.

Poniżej przedstawiono pojęcie reprezentowalnej przedziałowej negacji rozmytej. Inne reprezentacje przedziałowych negacji można znaleźć w [77].

Definicja 14 ([78]). N_{IV} nazywamy reprezentowalną przedziałową negacją rozmytą, jeśli istnieją negacje rozmyte N_1, N_2 takie, że $N_{IV}(x) = [N_1(\bar{x}), N_2(x)]$.

Mówi się, że N_1, N_2 są skojarzonymi negacjami rozmytymi do danej przedziałowej negacji rozmytej.

Twierdzenie 2.10 ([7]). *Niech N_1, N_2 będą negacjami rozmytymi takimi, że $N_1 \leq N_2$. Operacja $N_{IV} : L^I \rightarrow L^I$ taka, że $N_{IV}(x) = [N_1(\bar{x}), N_2(x)]$ jest reprezentowalną, przedziałową negacją zgodną z $\leq_{L^I}$.*

Poniższy wynik łączy silne, przedziałowe negacje rozmyte (w sensie reprezentowalności) z silnymi negacjami rozmytymi.

Twierdzenie 2.11 ([77]). *Niech $N_{IV}(x) = [N_1(\bar{x}), N_2(x)]$ będzie silną, inwolutywną, przedziałową negacją rozmytą względem porządku $\leq_{L^I}$. Wtedy N_1 i N_2 są silnymi negacjami rozmytymi i $N_1 \leq N_2$.*

Ponadto,

Twierdzenie 2.12 ([65]). *Niech $N : [0, 1] \rightarrow [0, 1]$ będzie silną negacją. Niech $N_{IV} : L^I \rightarrow L^I$ będzie funkcją zdefiniowaną przez*

$$N_{IV}(x) = [N(\bar{x}), N(\underline{x})].$$

Jeśli N_{IV} jest przedziałową negacją rozmytą zgodną z $\leq$ w L^I , wtedy $\leq$ nie jest porządkiem liniowym.

Dogłębną analizę negacji przedziałowych w odniesieniu do dopuszczalnych porządków można znaleźć w [76].

Negacje możliwości i konieczności

Kolejnymi klasami negacji są negacje oparte o relacje konieczności (*ang. necessary*) i możliwości (*ang. possible*). Negacje te nazywamy odpowiednio negacjami możliwości (negacjami optymistycznymi) i negacjami konieczności (negacjami pesymistycznymi). Poniżej przedstawione zostały ich definicje.

Definicja 15. Negacja możliwości jest funkcją $N_{pos} : L^I \rightarrow L^I$ malejącą względem porządku $\leq_{pos}$, czyli taka, że

$$\forall_{x,y \in L^I} x \leq_{pos} y \Rightarrow N_{pos}(y) \leq_{pos} N_{pos}(x)$$

oraz spełnia następujące warunki brzegowe $N_{pos}(1_{L^I}) = 0_{L^I}$ i $N_{pos}(0_{L^I}) = 1_{L^I}$.

Definicja 16. Negacja konieczności jest funkcją $N_{nec} : L^I \rightarrow L^I$ malejącą względem porządku $\leq_{nec}$, czyli taka, że

$$\forall_{x,y \in L^I} x \leq_{nec} y \Rightarrow N_{nec}(y) \leq_{nec} N_{nec}(x)$$

oraz spełnia następujące warunki brzegowe $N_{nec}(1_{L^I}) = 0_{L^I}$ i $N_{nec}(0_{L^I}) = 1_{L^I}$.

Przez N_{pos} i N_{nec} w pracy oznaczono odpowiednio rodzinę negacji możliwości (optymistycznych) i negacji konieczności (pesymistycznych).

Twierdzenie 2.13. *Niech $N_1, N_2, N_3, N_4, N_5, N_6 : L^I \rightarrow L^I$, wtedy*

$$N_1(x) = [1 - \bar{x}, 1 - \underline{x}]$$

jest negacją względem $\leq_{L^I}$, $\leq_{pos}$, $\leq_{nec}$ i jest nazywaną negacją klasyczną.

$$N_2(x) = \begin{cases} [1, 1], & \text{jeśli } x = [0, 0], \\ [0, 1 - \bar{x}], & \text{w przeciwnym przypadku} \end{cases}$$

jest negacją względem $\leq_{pos}$.

$$N_3(x) = \begin{cases} [0, 0], & \text{jeśli } x = [1, 1], \\ [1 - \underline{x}, 1], & \text{w przeciwnym przypadku} \end{cases}$$

jest negacją względem $\leq_{pos}$.

$$N_4(x) = \begin{cases} [1, 1], & \text{jeśli } x = [0, 0], \\ [0, 0], & \text{jeśli } x = [1, 1], \\ \left[\frac{1-\underline{x}}{2}, 1 - \frac{\bar{x}}{2}\right], & \text{w przeciwnym przypadku} \end{cases}$$

jest negacją względem: $\leq_{L^I}, \leq_{XY}, \leq_{pos}$.

$$N_5(x) = [1 - \bar{x}, 1 - \bar{x}]$$

jest zgodna z porządkiem $\leq_{nec}$.

$$N_6(x) = \begin{cases} [1, 1], & \text{jeśli } x = [0, 0], \\ [0, 0], & \text{jeśli } x = [1, 1], \\ [0, \underline{x}], & \text{w przeciwnym przypadku} \end{cases}$$

jest negacją względem $\leq_{pos}$.

Dowód.

I. Początkowo wykażemy, że N_1 jest negacją zgodną z porządkiem częściowym: Jeżeli

$$\forall_{x,y \in L^I} x \leq_{L^I} y \Leftrightarrow \underline{x} \leq \underline{y}, \bar{x} \leq \bar{y},$$

wtedy otrzymujemy:

$$1 - \underline{y} \leq 1 - \underline{x} \quad \text{i} \quad 1 - \bar{y} \leq 1 - \bar{x},$$

zatem $N_1(y) \leq_{L^I} N_1(x)$. Ponadto, $N_1([1, 1]) = [0, 0]$; $N_1([0, 0]) = [1, 1]$.

II. Następnie wykażemy, że negacja N_1 jest zgodna z porządkiem $\leq_{pos}$:

1° Jeśli $x \leq_{pos} y$, np. $\underline{x} \leq \bar{y} \Leftrightarrow 1 - \bar{y} \leq 1 - \underline{x}$ wtedy

$$N_1(y) = [1 - \bar{y}, 1 - \underline{y}] \leq_{pos} N_1(x) = [1 - \bar{x}, 1 - \underline{x}];$$

2° $N_1([1, 1]) = [0, 0]$;

3° $N_1([0, 0]) = [1, 1]$.

III. Negacja N_1 jest zgodna z porządkiem $\leq_{nec}$:

- 1° Jeśli $x \leq_{nec} y$, czyli $\bar{x} \leq \underline{y} \Leftrightarrow 1 - y \leq 1 - \bar{x}$ zatem $N_1(y) \leq_{nec} N_1(x)$;
- 2° $N_1([1, 1]) = [0, 0]$;
- 3° $N_1([0, 0]) = [1, 1]$.

Kolejną rozważaną negacją będzie N_4 :

I. Najpierw wykazano, że jest ona zgodna z porządkiem częściowym: Jeżeli

$$\forall_{x,y \in L^I} x \leq_{L^I} y \Leftrightarrow \underline{x} \leq \underline{y}, \bar{x} \leq \bar{y},$$

wtedy otrzymano:

$$1 - \frac{\bar{y}}{2} \leq 1 - \frac{\bar{x}}{2} \quad \text{i} \quad \frac{1 - y}{2} \leq \frac{1 - x}{2},$$

zatem $N_4(y) \leq_{L^I} N_4(x)$.

II. W pracach [76, 63, 65] wykazano, że N_4 jest zgodna z porządkiem $\leq_{XY}$.

III. Następnie wykazemy, że omawiana negacja N_4 jest zgodna z porządkiem $\leq_{pos}$, czyli dla

$$x \leq_{pos} y \Leftrightarrow \underline{x} \leq \bar{y} \tag{2.9}$$

zachodzi

$$N_4(y) \leq_{pos} N_4(x) \Leftrightarrow \frac{1 - y}{2} \leq 1 - \frac{\bar{x}}{2}. \tag{2.10}$$

Rozważono przypadki:

- 1° gdy $x = y = [0, 0]$ warunek (2.10) jest spełniony;
- 2° gdy $x = [0, 0]$, $y = [1, 1]$ warunek (2.10) jest spełniony;
- 3° gdy $x = [0, 0]$, $y \neq [1, 1]$, $y \neq [0, 0]$ warunek (2.10) jest spełniony, bo $\frac{1-y}{2} \leq 1$;
- 4° gdy $x \neq [0, 0]$, $y = [1, 1]$ warunek (2.10) jest spełniony;
- 5° gdy $x = y = [1, 1]$ warunek (2.10) jest spełniony;
- 6° gdy $x \neq [0, 0]$, $y \neq [0, 0]$, $y \neq [1, 1]$ warunek (2.10) jest spełniony, bo $\underline{y} \geq \bar{x} - 1$.

Dowód dla pozostałych negacji przebiega podobnie. □

Kolejne twierdzenie wskazuje na zależność negacji konieczności (*ang. neccessary*) z metodą konstrukcji przedziałowych negacji jaką jest reprezentowalność.

Twierdzenie 2.14. *Niech N_{IV} będzie operacją $N_{IV} = [N, N]$, także N jest negacją rozmytą, wtedy:*

- N_{IV} jest reprezentowalną negacją konieczności względem $\leq_{nec}$.

Dowód. Jeżeli $x \leq_{nec} y \Leftrightarrow \bar{x} \leq \underline{y}$, to $N_{IV}(y) \leq_{nec} N_{IV}(x)$ zachodzi, gdyż $N(\bar{y}) \leq N(\underline{y}) \leq N(\bar{x}) \leq N(\underline{x})$, czyli $N(\bar{y}) \leq N(\underline{x})$. Co kończy dowód. $\square$

2.2.4 Przedziałowe zbiory rozmyte

Koncepcja przedziałowego zbioru rozmytego pojawiła się w 1975. Sambuc w swojej pracy [6] wprowadził pojęcie zbioru rozmytego (*ang. H-fuzzy set*). Natomiast Zadech w tym samym czasie zaproponował pojęcie przedziałowego zbioru rozmytego w pracy [5] jako szczególny przypadek zbiorów rozmytych typu 2, stanowiącego uogólnienie zbioru rozmytego poprzez wprowadzenie dodatkowego określenia nieprecyzyjności informacji, przez co okazał się bardzo przydatny w reprezentacji niepewności. W kolejnych latach dzięki pracom Gorzałczanego i Türksena [26, 79, 80, 81, 82, 83] przedziałowe zbiory rozmyte zyskały swoją ostateczną nazwę oraz znaczenie.

Definicja 17 ([5, 6]). Przedziałowym zbiorem rozmytym (*ang. Interval-valued fuzzy set (IVFS)*) A w X jest odwzorowanie $A : X \rightarrow L^I$ takie, że $A(x) = [\underline{A}(x), \bar{A}(x)] \in L^I$, gdzie $x \in X$.

Podstawowymi operacjami jakie można wykonać na dwóch przedziałowych zbiorach rozmytych są:

- suma

$$A \cup B(x) = [\max\{\underline{A}(x), \underline{B}(x)\}, \max\{\bar{A}(x), \bar{B}(x)\}],$$

- iloczyn

$$A \cap B(x) = [\min\{\underline{A}(x), \underline{B}(x)\}, \min\{\bar{A}(x), \bar{B}(x)\}],$$

- dopełnienie

$$A^N(x) = [N(\bar{A}(x)), N(\underline{A}(x))],$$

gdzie N jest negacją rozmytą; w przypadku negacji klasycznej otrzymujemy

$$A^c(x) = [1 - \bar{A}(x), 1 - \underline{A}(x)].$$

Rodzina wszystkich przedziałowych zbiorów rozmytych na X jest oznaczana przez $IVFS(X)$.

Niech $X \neq \emptyset$ i $\text{card}(X) = n$, $n \in \mathbb{N}$. Dla $A, B \in IVFS(X)$ będę posługiwał się następującym pojęciem częściowego porządku

$$A \preceq B \Leftrightarrow a_i \leq b_i \quad (2.11)$$

dla $i = 1, \dots, n$, gdzie $\leq$ jest tym samym rodzajem porządku (częściowym lub liniowym, lub relacją możliwości lub konieczności) dla każdego i oraz $a_i = A(x_i)$, $b_i = B(x_i)$.

Przedziałowe zbiory rozmyte są często wykorzystywane w praktycznych zastosowaniach. Przykładami takich zastosowań są: diagnostyka medyczna zobrazowana w pracach Sambuc'a [6], Kohout'a [28], Dyczkowskiego [24], przetwarzanie obrazów u Bigand'a i Colot'a [84] oraz Bustince i innych [85]. W ostatnich latach nastąpił znaczący przyrost prac naukowych w tym zakresie. Ponadto, przedziałowe zbiory rozmyte [4] w wielu przypadkach poprawiają wyniki zbiorów rozmytych w różnych zastosowaniach [18, 73, 86, 76], ponieważ pozwalają na uwzględnienie nieprecyzyjności także związanej z konstrukcją funkcji przynależności.

2.2.5 Przedziałowe relacje rozmyte - wybrane własności

Niech X i Y będą zbiorami niepustymi.

Przedziałową relacją rozmytą [5, 6] R na iloczynie kartezjańskim $X \times Y$ nazywamy funkcję: $R : X \times Y \rightarrow L^I$. Rodzinę przedziałowych relacji rozmytych na $X \times Y$ oznaczamy $IVFR(X \times Y)$ ($IVFR(X)$ dla $X = Y$). Zwrócimy uwagę na wybrane własności rozmytych relacji przedziałowych, a następnie złożenie, które miały wpływ na zastosowanie wybranych operatorów agregacji w rozwiązaniu problemu przedstawionego w rozdziale 6.1. Szczególną uwagę zwracamy na słaby operator agregacji W .

Definicja 18. Relacja $R \in IVFR(X)$ nazywana jest

- I zwrotną, jeśli $R(x, x) = [1, 1]$;
- II przeciw zwrotną, jeśli $R(x, x) = [0, 0]$,
- III W -przechodnią, jeżeli $W(R(x, z), R(z, y)) \leq R(x, y)$;
- IV symetryczną, jeżeli $R(x, y) = R(y, x)$; dla każdego $x, y, z \in X$.

Można zauważyć, w oparciu o wyniki przedstawione w pracach [52, 73], że agregacje przedziałowe oraz słaby operator W zachowują wspomniane własności.

Twierdzenie 2.15. Niech $R_1, \dots, R_n \in IVFR(X)$, $\mathcal{A} : (L^I)^n \rightarrow L^I$ będzie przedziałową funkcją agregacji i $W : (L^I)^n \rightarrow L^I$ słabym operatorem dla $n \in N$.

- Jeśli dla wszystkich $R_1, \dots, R_n$ są zwrotne/przeciw zwrotne, to $\mathcal{A}(R_1, \dots, R_n)$ ($W(R_1, \dots, R_n)$) jest również zwrotna/przeciw zwrotna,
- Jeśli relacje $R_1, \dots, R_n \in IVFR(X)$ są symetryczne, to $\mathcal{A}(R_1, \dots, R_n)$ ($W(R_1, \dots, R_n)$) jest również symetryczna,
- $R \in IVFR(X)$ jest W -przechodnią wtedy i tylko wtedy, gdy R^{-1} jest W -przechodnią i $R^{-1}(x, y) = R(y, x)$.

Rozważono kilka podstawowych własności związanych ze słabym operatorem W , interesujących ze względu na potencjalny ich wpływ na zastosowanie w diagnostyce medycznej

(rozdział 6.1), tj. np. symetrię, zwrotność czy przechodniość.

Kolejną, obok agregacji, ważną operacją w naszych rozważaniach praktycznych zawartych w dalszej części pracy, jest złożenie przedziałowych relacji rozmytych. Badania nad uogólnieniami złożenia zbiorów rozmytych zapoczątkował w swojej pracy Goguen [2] i były kontynuowane przez innych badaczy (np. [53, 87, 88]).

Przypomnimy definicję uogólnionego złożenia przedziałowych zbiorów rozmytych dla zbiorów skończonych $X, Y, Z \neq \emptyset$ przedstawioną w [73]. Gdy przedziałowe funkcje agregacji $\mathcal{A} : (L^I)^n \rightarrow L^I$ oraz $\mathcal{B} : (L^I)^2 \rightarrow L^I$ są zgodne z porządkiem $\leq$, wtedy otrzymujemy $\mathcal{A} - \mathcal{B}$ złożenie relacji P i R , oznaczane przez $P \circ_{\mathcal{AB}} R \in IVFR(X \times Z)$, następująco

$$(P \circ_{\mathcal{AB}} R)_{ik} = \mathcal{A}_{j=1}^n(\mathcal{B}(P_{ij}, R_{jk})),$$

gdzie $P \in IVFR(X \times Y)$ and $R \in IVFR(Y \times Z)$.

ROZDZIAŁ 3

Miary zawierania, podobieństwa i entropii uwzględniające nieprecyzyjność

W tym rozdziale przedstawiono różne miary, które uwzględniają niepewność. Omówione zostaną: wskaźnik pierwszeństwa, miara podobieństwa oraz entropia. Każda z wspomnianych miar będzie rozważana w trzech ujęciach: klasycznym, możliwości (optymistycznym) i konieczności (pesymistycznym). Zastosowanie miar uwzględniających niepewność jest konieczne, aby operować na danych przedziałowych i wykorzystać je efektywnie w procesie klasyfikacji poprzez modyfikację znanych algorytmów eksploracji danych. Przedstawione w niniejszym rozdziale miary podobieństwa i entropii rozważane są, ze względu na ich potencjał aplikacyjny, w rodzinie $IVFS(X)$, gdzie $card(X) = n$, $n \in N$.

3.1 Miary zawierania się przedziałów i przedziałowych zbiorów rozmytych

Pierwszą miarą jakiej poświęcimy uwagę będzie miara inkluzji, czyli zawierania zbioru rozmytego A w B . Zaproponowano kilka wskaźników wyrażających stopień, w jakim przedziałowy zbiór rozmyty jest podzbiorem innego przedziałowego zbioru rozmytego [86, 65, 89, 90].

Zanim zostaną przedstawione miary zawierania przedziałowych zbiorów rozmytych, a także miary podobieństwa i entropii w środowisku przedziałowych zbiorów rozmytych, należy wprowadzić wskaźnik pierwszeństwa, czyli miarę zawierania się przedziałów.

3.1.1 Wskaźnik pierwszeństwa klasyczny

W pierwszej kolejności zostanie przedstawiona koncepcja miary zawierania się przedziałów, nazywana wskaźnikiem pierwszeństwa, wykorzystująca porządek liniowy lub częściowy, a także pojęcie szerokości przedziału:

$$w(p) = \bar{p} - \underline{p} \text{ dla } p = [\underline{p}, \bar{p}] \in L^I.$$

Definicja 19 ([91], por. [92]). Funkcja $\text{Prec} : (L^I)^2 \rightarrow L^I$ jest wskaźnikiem pierwszeństwa (*ang. precedence indicator*) względem porządku $\leq$ (częściowego lub liniowego) określonego na L^I , jeżeli spełnia następujące warunki dla każdego $p, q, u \in L^I$:

P1 Jeśli $p = 1_{L^I}$ i $q = 0_{L^I}$ wtedy $\text{Prec}(p, q) = 0_{L^I}$;

P2 Jeśli $p < q$, wtedy $\text{Prec}(p, q) = 1_{L^I}$;

P3 $\text{Prec}(p, p) = [1 - w(p), 1]$;

P4 Jeśli $p \leq q \leq u$, wtedy $\text{Prec}(u, p) \leq \text{Prec}(q, p)$ i $\text{Prec}(u, p) \leq \text{Prec}(u, q)$.

Uwaga 3.1. Jeśli $p = q$ i $w(p) = 0$, wtedy $\text{Prec}(p, q) = 1_{L^I}$.

W pracy [46] rozważano szczególny przypadek powyższej definicji, w której aksjomat **P4** miał postać:

P4' Jeśli $p \leq q \leq u$ i $w(p) = w(q) = w(u)$, wtedy $\text{Prec}(u, p) \leq \text{Prec}(q, p)$ i $\text{Prec}(u, p) \leq \text{Prec}(u, q)$, dla każdego $p, q, u \in L^I$.

Jedną z metod konstrukcji wskaźnika pierwszeństwa zgodnego z definicją 19 jest wtedy następująca metoda oparta na szerokości przedziałów i ich wzajemnej odległości.

Twierdzenie 3.2. *Funkcja*

$$\text{Prec}_w(a, b) = \begin{cases} 1_{L^I}, & a <_{L^I} b, \\ [1 - \max(w(a), r(a, b)), 1 - r(a, b)], & \text{w przeciwnym przypadku} \end{cases}$$

jest wskaźnikiem pierwszeństwa względem $\leq_{L^I}$, gdzie $r(a, b) = \max\{|\underline{a} - \underline{b}|, |\bar{a} - \bar{b}|\}$ dla $a, b \in L^I$ spełniającym aksjomat **P4'**.

Natomiast poniżej przedstawiono metodę konstrukcji wskaźnika pierwszeństwa spełniającą definicję 19.

Twierdzenie 3.3 ([92]). Dla $x, y \in L^I$ funkcja $\text{Prec}_A : (L^I)^2 \rightarrow L^I$,

$$\text{Prec}_A(x, y) = \begin{cases} [1 - w(x), 1], & x = y, \\ 1_{L^I}, & x <_{L^I} y, \\ \mathcal{A}(N_{IV}(x), y), & \text{w przeciwnym przypadku} \end{cases}$$

jest wskaźnikiem pierwszeństwa dla przedziałowej negacji rozmytej N_{IV} , takiej że

$$N_{IV}(x) = [n(\bar{x}), n(\underline{x})] \leq [1 - \bar{x}, 1 - \underline{x}],$$

gdzie n jest negacją rozmytą i $\mathcal{A}$ jest reprezentowalną przedziałową agregacją taką, że $\mathcal{A} \leq \vee$.

Przykład 10. Poniższe funkcje są wskaźnikami pierwszeństwa względem $\leq_{L^I}$:

$$\begin{aligned} \text{Prec}_{\mathcal{A}_{meanL^I}}(x, y) &= \begin{cases} [1 - w(x), 1], & x = y, \\ 1_{L^I}, & x <_{L^I} y, \\ \left[\frac{1 - \bar{x} + y}{2}, \frac{1 - \underline{x} + \bar{y}}{2} \right], & \text{w przeciwnym przypadku,} \end{cases} \\ \text{Prec}_{\mathcal{A}_{meanpow}}(x, y) &= \begin{cases} [1 - w(x), 1], & x = y, \\ 1_{L^I}, & x <_{L^I} y, \\ \left[\frac{1 - \bar{x} + y}{2}, \sqrt{\frac{(1 - \underline{x})^2 + \bar{y}^2}{2}} \right], & \text{w przeciwnym przypadku,} \end{cases} \\ \text{Prec}_{\mathcal{A}_{meanmax}}(x, y) &= \begin{cases} [1 - w(x), 1], & x = y, \\ 1_{L^I}, & x <_{L^I} y, \\ \left[\frac{1 - \bar{x} + y}{2}, \max(1 - \underline{x}, \bar{y}) \right], & \text{w przeciwnym przypadku,} \end{cases} \end{aligned}$$

gdzie $N_{IV}(x) = [1 - \bar{x}, 1 - \underline{x}]$.

Ponadto, poniższa funkcja jest wskaźnikiem pierwszeństwa względem $\leq_{Lex2}$

$$\text{Prec}_{\mathcal{A}_{meanLex2}}(x, y) = \begin{cases} [1 - w(x), 1], & x = y, \\ 1_{L^I}, & x <_{Lex2} y, \\ \left[\frac{y}{2}, \frac{1 - \bar{x} + \bar{y}}{2} \right], & \text{w przeciwnym przypadku,} \end{cases}$$

i wykorzystując przedziałową funkcję agregacji $\mathcal{A}_\alpha$ dla $\alpha \in [0, 1]$, otrzymujemy wskaźnik pierwszeństwa

$$\text{Prec}_{\mathcal{A}_\alpha Lex2}(x, y) = \begin{cases} [1 - w(x), 1], & x = y, \\ 1_{L^I}, & x <_{Lex2} y, \\ [(1 - \alpha)y, \alpha(1 - \bar{x}) + (1 - \alpha)\bar{y}], & \text{w przeciwnym przypadku,} \end{cases}$$

gdzie

$$N_{IV}(x) = \begin{cases} 1_{L^I}, & x = [0, 0], \\ [0, 1 - \bar{x}], & \text{w przeciwnym przypadku} \end{cases}$$

jest przedziałową negacją rozmytą względem $\leq_{Lex2}$.

Miara stopnia zawierania dla przedziałowych zbiorów rozmytych przyjmuje następującą postać.

Definicja 20 ([91], por. [92]). Niech $A_g : [0, 1]^n \rightarrow [0, 1]$ będzie funkcją agregacji i $A, B, C \in \text{IVFS}(X)$. Miara stopnia zawierania oparta na agregacji (*ang. aggregation-based indicator of the grade of inclusion*) przedziałowych zbiorów rozmytych jest funkcją $\text{Inc} : (\text{IVFS}(X))^2 \rightarrow L^I$ która spełnia następujące warunki:

$$\text{IMV1 } \text{Inc}(\mathbf{1}, \mathbf{0}) = 0_{L^I};$$

$$\text{IMV2 } \text{Jeśli } A \prec B, \text{ wtedy } \text{Inc}(A, B) = 1_{L^I};$$

$$\text{IMV3 } \text{Inc}(A, A) = [1 - A_g(w(a_1), \dots, w(a_n)), 1];$$

$$\text{IMV4 } \text{Jeśli } A \preceq B \preceq C, \text{ wtedy } \text{Inc}(C, A) \leq \text{Inc}(C, B) \text{ i } \text{Inc}(C, A) \leq \text{Inc}(B, A),$$

gdzie $A \prec B \Leftrightarrow A \preceq B$ i $\exists_{1 \leq k \leq n} a_k < b_k$ oraz $\mathbf{1} = \{1_{L^I}\}$, $\mathbf{0} = \{0_{L^I}\}$.

Twierdzenie 3.4 ([91]). Niech $\mathcal{A} = [\Psi, \Upsilon] : (L^I)^n \rightarrow L^I$ będzie agregacją przedziałową, w której Ψ jest auto-odwracalna i niech $\text{Prec} : (L^I)^2 \rightarrow L^I$ będzie wskaźnikiem pierwszeństwa względem tego samego porządku. Wówczas odwzorowanie $\text{Inc}_{\mathcal{A}, \text{Prec}} : (\text{IVFS}(X))^2 \rightarrow L^I$ zdefiniowane dla dowolnego $A, B \in \text{IVFS}(X)$ przez

$$\text{Inc}_{\mathcal{A}, \text{Prec}}(A, B) = \mathcal{A}\left(\text{Prec}(A(x_1), B(x_1)), \dots, \text{Prec}(A(x_n), B(x_n))\right) \quad (3.1)$$

jest opartą na agregacji miarą stopnia zawierania dla przedziałowych zbiorów rozmytych.

3.1.2 Wskaźnik pierwszeństwa możliwości

Przedstawiony teraz zostanie wskaźnik pierwszeństwa możliwości zwany także optymistycznym wskaźnikiem pierwszeństwa [51] ze względu na zależność od relacji możliwości porównywania przedziałów, która jest ogólniejszą klasą miar zawierania od klasycznej.

Definicja 21. Funkcja $\text{Prec}_{\text{pos}} : (L^I)^2 \rightarrow L^I$ jest nazywana wskaźnikiem pierwszeństwa możliwości (*ang. possible precedence indicator*), jeśli spełnia następujące warunki dla dowolnego $a, b, c \in L^I$:

$$\text{PP1 } \text{Jeśli } a = 1_{L^I} \text{ i } b = 0_{L^I}, \text{ wtedy } \text{Prec}_{\text{pos}}(a, b) = 0_{L^I};$$

$$\text{PP2 } \text{Jeśli } a <_{\text{pos}} b \text{ i } a \neq b, \text{ wtedy } \text{Prec}_{\text{pos}}(a, b) = 1_{L^I};$$

$$\text{PP3 } \text{Prec}_{\text{pos}}(a, a) = [1 - w(a), 1], \text{ gdzie } w(a) = \bar{a} - \underline{a} \text{ oznacza szerokość przedziału } a \in L^I;$$

$$\text{PP4 } \text{Jeśli } a \leq_{\text{pos}} b \leq_{\text{pos}} c, \text{ wtedy } \text{Prec}_{\text{pos}}(c, a) \leq_{\text{pos}} \text{Prec}_{\text{pos}}(b, a) \text{ i } \text{Prec}_{\text{pos}}(c, a) \leq_{\text{pos}} \text{Prec}_{\text{pos}}(c, b).$$

Twierdzenie 3.5. *Niech N_{pos} oznacza przedziałową negację rozmytą możliwości (tzn. operację malejącą, która spełnia warunki $N_{pos}(0_{L^I}) = 1_{L^I}$ i $N_{pos}(1_{L^I}) = 0_{L^I}$), takie, że $N_{pos}(a) \leq_{pos} [1 - \bar{a}, 1 - \underline{a}]$, gdzie $\mathcal{A}_{pos}$ jest reprezentowalną agregacją przedziałową możliwości, taką że $\mathcal{A}_{pos} \leq_{pos} \vee$. Wtedy*

$$\text{Prec}_{\mathcal{A}_{pos}}(a, b) = \begin{cases} [1 - w(a), 1], & \text{jeśli } a = b, \\ 1_{L^I}, & \text{jeśli } a <_{pos} b, \\ \mathcal{A}_{pos}(N_{pos}(a), b), & \text{w przeciwnym wypadku} \end{cases}$$

jest wskaźnikiem pierwszeństwa możliwości.

3.1.3 Wskaźnik pierwszeństwa konieczności

W [51] został zaproponowany wskaźnik pierwszeństwa konieczności zwany także pesymistycznym wskaźnikiem pierwszeństwa, który jest zgodny z pesymistyczną relacją porównywalności i tworzy klasę miar bardziej restrykcyjną od klas możliwości i w ujęciu klasycznym.

Definicja 22. Funkcja $\text{Prec}_{nec} : (L^I)^2 \rightarrow L^I$ jest nazywana wskaźnikiem pierwszeństwa konieczności (*ang. necessary precedence indicator*), jeśli spełnia następujące warunki dla dowolnych $a, b, c \in L^I$:

PN1 Jeśli $a = 1_{L^I}$ i $b = 0_{L^I}$ wtedy $\text{Prec}_{nec}(a, b) = 0_{L^I}$;

PN2 Jeśli $a <_{nec} b$, wtedy $\text{Prec}_{nec}(a, b) = 1_{L^I}$;

PN3 $\text{Prec}_{nec}(a, a) = [1 - w(a), 1]$ dla każdego $a \in L^I$, gdzie $w(a) = \bar{a} - \underline{a}$ oznacza szerokość przedziału $a \in L^I$;

PN4 Jeśli $a \leq_{nec} b \leq_{nec} c$, wtedy $\text{Prec}_{nec}(c, a) \leq_{nec} \text{Prec}_{nec}(b, a)$ i $\text{Prec}_{nec}(c, a) \leq_{nec} \text{Prec}_{nec}(c, b)$.

Następujące twierdzenie przedstawia metodę konstrukcji wskaźnika pierwszeństwa konieczności.

Twierdzenie 3.6. *Funkcja*

$$\text{Prec}_{nec}(a, b) = \begin{cases} [1 - w(a), 1], & \text{if } a = b, \\ 1_{L^I}, & \text{if } a <_{nec} b, \\ [1 - \max(w(a), r(a, b)), 1 - r(a, b)], & \text{w przeciwnym przypadku} \end{cases}$$

jest wskaźnikiem pierwszeństwa konieczności, gdzie $r(a, b) = \max(|\underline{a} - \underline{b}|, |\bar{a} - \bar{b}|)$.

3.2 Miara podobieństwa dla przedziałowych zbiorów rozmytych

W tym rozdziale przedstawiona zostanie miara podobieństwa pomiędzy przedziałowymi zbiorami rozmytymi. Motywacją do badań było uwzględnienie ważnych dla modelowania niepewności miar porównywalności przedziałów oraz wskaźników pierwszeństwa w klasach odpowiednio klasycznej, możliwości i konieczności. Aby skonstruować podobieństwo w rodzinie przedziałowych zbiorów rozmytych, zostaną wykorzystane przedziałowe funkcje agregacji obok wskaźników pierwszeństwa (*ang. precedence indicator*) zaprezentowanych we wcześniejszym rozdziale, które uwzględniają szerokość przedziałów. Ze względu na aspekt aplikacyjny przyjęto w dalszych rozważaniach $\text{card}(X) = n$, $n \in N$.

3.2.1 Miara podobieństwa klasyczna

Klasyczna miara podobieństwa dla przedziałowych zbiorów rozmytych oparta jest na porządku częściowym $\leq_{L^I}$ lub liniowym $\leq_{Adm}$ (zob. rozdz. 2.2.1).

Definicja 23 ([91]). Niech $A_1 : [0, 1]^n \rightarrow [0, 1]$ będzie funkcją agregacji. Wówczas operacja $S : IVFS(X) \times IVFS(X) \rightarrow L^I$, spełniająca następujące warunki:

SIMv1 $S(A, B) = S(B, A)$ dla $A, B \in IVFS(X)$;

SIMv2 $S(A, A) = [1 - A_1(w(A(x_1)), \dots, w(A(x_n))), 1]$;

SIMv3 $S(A, B) = 0_{L^I}$, jeśli $\{A(x_i), B(x_i)\} = \{0_{L^I}, 1_{L^I}\}$, dla $i=1, \dots, n$;

SIMv4 Jeśli $A \preceq B \preceq C$, wtedy $S(A, C) \leq S(A, B)$ i $S(A, C) \leq S(B, C)$

jest nazywana przedziałową miarą podobieństwa pomiędzy przedziałowymi zbiorami rozmytymi, gdzie $\leq \in \{\leq_{L^I}, \leq_{Adm}\}$.

Zaprezentowana w powyższej definicji własność **SIMv4** jest bardziej ogólna niż odpowiednia własność zaproponowana przez autorów artykułu [93], gdzie dodatkowo zakłada się równość szerokości przedziałów $w(A) = w(B) = w(C)$.

Następująca metoda konstrukcji miary podobieństwa wykorzystuje wskaźnik pierwszeństwa w sensie klasycznym wprowadzonym w poprzednim rozdziale.

Twierdzenie 3.7 ([91]). Niech Prec będzie wskaźnikiem pierwszeństwa. Jeśli $\mathcal{A} = [A_1, A_2]$, $\mathcal{B} = [B_1, B_2]$ są reprezentowalnymi przedziałowymi funkcjami agregacji, gdzie A_1 jest auto-odwracalna ($1 - (1 - A_1(x_1, \dots, x_n)) = A_1(x_1, \dots, x_n)$), $\mathcal{B}$ jest symetryczna z elementem neutralnym 1_{L^I} i B_1 jest idempotentną funkcją agregacji, to operacja

$S : IVFS(X) \times IVFS(X) \rightarrow L^I$:

$$S(A, B) = \mathcal{A}_{i=1}^n(\mathcal{B}(\text{Prec}(A(x_i), B(x_i)), \text{Prec}(B(x_i), A(x_i))))$$

jest przedziałową miarą podobieństwa.

Wniosek 3.8. Niech Prec będzie wskaźnikiem pierwszeństwa. Jeśli $\mathcal{A}$ jest symetryczną reprezentowalną przedziałową funkcją agregacji z elementem neutralnym 1_{L^I} i A_1 jest auto-odwracalna i idempotentna, wówczas operacja $S : IVFS(X) \times IVFS(X) \rightarrow L^I$:

$$S(A, B) = \mathcal{A}_{i=1}^n(\mathcal{A}(\text{Prec}(A(x_i), B(x_i)), \text{Prec}(B(x_i), A(x_i))))$$

jest przedziałową miarą podobieństwa.

Przykład 11. Operacja $S : IVFS(X) \times IVFS(X) \rightarrow L^I$:

$$S(A, B) = \mathcal{A}_{i=1}^n(\text{Prec}_{\mathcal{A}}(A(x_i), B(x_i)) \wedge \text{Prec}_{\mathcal{A}}(B(x_i), A(x_i)))$$

jest miarą podobieństwa, gdzie $\mathcal{A} \in \{\mathcal{A}_{mean}, \mathcal{A}_{meanpow}, \mathcal{A}_{meanmax}\}$.

Kolejne twierdzenie przedstawia nową metodę konstrukcji przedziałowej miary podobieństwa opartą o miarę stopnia zawierania przedziałowych zbiorów rozmytych Inc zgodną z $\leq_{L^I}$ użytą w $\preceq$ (zob. 2.11).

Metoda konstrukcji podobieństwa S uwzględniająca Inc .

Twierdzenie 3.9. Dla symetrycznej reprezentowalnej agregacji przedziałowej $\mathcal{A} = [\varrho, \delta]$ z elementem zerowym 0_{L^I} , gdzie ϱ jest agregacją idempotentną i reprezentowalnej agregacji $\mathcal{B} = [\Psi, \Gamma]$ z elementem zerowym 0_{L^I} i z auto-odwracalnością Ψ użytej w konstrukcji Inc (miary stopnia zawierania przedziałowych zbiorów rozmytych) operacja $S : (IVFS)^2 \rightarrow L^I$,

$$S(A, B) = \mathcal{A}(Inc(A, B), Inc(B, A))$$

jest przedziałową miarą podobieństwa.

Dowód. Sprawdzono kolejno spełnienie aksjomatów dla miary podobieństwa przedstawionych w definicji 23:

SIMv1. $S(A, B) = S(B, A)$ zachodzi wobec symetryczności agregacji $\mathcal{A}$.

SIMv2. Wykażemy $S(A, A) = [1 - \Psi(w_A(x_1), \dots, w_A(x_n)), 1]$. Dla

$$S(A, A) = \mathcal{A}(Inc(A, A), Inc(A, A) =$$

$$\mathcal{A}([1 - \Psi(w_A(x_1), \dots, w_A(x_n)), 1], [1 - \Psi(w_A(x_1), \dots, w_A(x_n)), 1])$$

z idempotentności ϱ otrzymano

$$S(A, A) = \mathcal{A}(Inc(A, A), Inc(A, A)) = [1 - \Psi(w_A(x_1), \dots, w_A(x_n)), 1],$$

co dowodzi spełnienia **SIMv2**.

SIMv3. $S(A, B) = 0_{L^I}$, jeżeli $\{A(x_i), B(x_i)\} = \{0_{L^I}, 1_{L^I}\}$.

Rozważono przypadki:

1. Jeżeli $(A(x_i) = 0_{L^I} \ \forall_{i=1, \dots, n} \text{ i } B(x_i) = 1_{L^I} \ \forall_{i=1, \dots, n})$
lub $(A(x_i) = 1_{L^I} \ \forall_{i=1, \dots, n} \text{ i } B(x_i) = 0_{L^I} \ \forall_{i=1, \dots, n})$, to
 $Inc(A, B) = 1_{L^I}$ i $Inc(B, A) = 0_{L^I}$ lub $Inc(A, B) = 0_{L^I}$
i $Inc(B, A) = 1_{L^I}$ wtedy dla $\mathcal{A}$ z elementem zerowym 0_{L^I}
SIMv3 jest spełniony.
2. Jeżeli $\forall_{i=1, \dots, n} [(A(x_i) = 0_{L^I} \text{ i } B(x_i) = 1_{L^I}) \text{ lub } (A(x_i) = 1_{L^I} \text{ i } B(x_i) = 0_{L^I})]$, to
 $Prec(A(x_i), B(x_i)) = 1_{L^I}$ i $Prec(B(x_i), A(x_i)) = 0_{L^I}$ lub
 $Prec(A(x_i), B(x_i)) = 0_{L^I}$ i $Prec(B(x_i), A(x_i)) = 1_{L^I}$ a dla $\mathcal{B}$
z elementem zerowym 0_{L^I} $S(A, B) = 0_{L^I}$.

SIMv4. Jeżeli $A \preceq B \preceq C$, to $S(A, C) \leq S(A, B)$, $S(A, C) \leq S(B, C)$.

Niech

1. $A \prec B \prec C$, wtedy $\exists$ przynajmniej jedno $k \in \{1, \dots, n\}$,
że $A(x_k) < B(x_k)$ i $\exists$ przynajmniej jedno $z \in \{1, \dots, n\}$,
że $B(x_z) < C(x_z)$, wtedy $A(x_k) < C(x_k)$ i $A(x_z) < C(x_z)$.

Zakładając, że dla l elementów zachodzi silna nierówność.

$S(A, C) = \mathcal{A}(Inc(A, C), Inc(C, A)) \leq S(A, B)$, bo $Inc(A, C) = 1_{L^I}$
i $Inc(C, A) \leq Inc(B, A)$, $Inc(C, A) \leq Inc(C, B)$ dla $A \prec B \prec C$
oraz $S(A, C) \leq S(B, C) = S(C, B)$.

2. Niech $A = B \prec C$. Wtedy

$$S(A, C) = \mathcal{A}(Inc(A, C), Inc(C, A)) = \mathcal{A}[1_{L^I}, Inc(C, B)] = S(B, C) = S(C, B),$$

$$S(A, C) \leq S(A, B) = S(A, A)$$

na podstawie twierdzenia 3.4, z auto-odwracalności Ψ oraz według $Inc(C, A) \leq Inc(A, A)$.

Pozostałe przypadki, czyli $A \prec B = C$ i $A = B = C$ są dowodzone w analogiczny sposób.

□

Przez zastosowanie wskaźników pierwszeństwa w sensie możliwości i konieczności (optymistycznym i pesymistycznym), można uzyskać również podobieństwa w adekwatnych przypadkach tj. możliwości i konieczności (zob. [51]). Ogólniejsza miara możliwości odegra kluczową rolę w części aplikacyjnej, czyli w rozdziale 5.

3.2.2 Miara podobieństwa możliwości

Zaproponowana nowa miara podobieństwa oparta jest na relacji możliwości. Nazywamy ją miarą podobieństwa możliwości (optymistyczną miarą podobieństwa).

Definicja 24. Niech $A_1 : [0, 1]^n \rightarrow [0, 1]$ będzie funkcją agregacji. Funkcja $S_{pos} : (IVFS(X))^2 \rightarrow L^I$, która spełnia następujące warunki:

$$\text{SP1 } S_{pos}(A, B) = S_{pos}(B, A);$$

$$\text{SP2 } S_{pos}(A, A) = [1 - A_1(w(A(x_1)), \dots, w(A(x_n))), 1];$$

$$\text{SP3 } S_{pos}(A, B) = 0_{L^I}, \text{ jeśli } \{A(x_i), B(x_i)\} = \{0_{L^I}, 1_{L^I}\}, \text{ dla } i = 1, \dots, n;$$

$$\text{SP4 } \text{Jeśli } A \preceq_{pos} B \preceq_{pos} C, \text{ wtedy } S_{pos}(A, C) \leq_{pos} S_{pos}(A, B), S_{pos}(A, C) \leq_{pos} S_{pos}(B, C)$$

jest nazywana **miarą podobieństwa możliwości** (ang. *possible similarity measure*) pomiędzy przedziałowymi zbiorami rozmytymi, gdzie

$$A \preceq_{pos} B \Leftrightarrow a_i \leq_{pos} b_i \text{ dla } i = 1, \dots, n.$$

Twierdzenie 3.10. Niech Prec_{pos} będzie wskaźnikiem pierwszeństwa możliwości. Jeśli $\mathcal{A}_{pos} = [A_1, A_2]$, $\mathcal{B}_{pos} = [B_1, B_2]$ są przedziałowymi agregacjami możliwości dla których A_1 jest auto-odwracalna, $\mathcal{B}_{pos}$ jest symetryczna z elementem neutralnym 1_{L^I} i B_1 jest idempotentną funkcją agregacji, wtedy $S_{pos} : IVFS(X) \times IVFS(X) \rightarrow L^I$:

$$S_{pos}(A, B) = \mathcal{A}_{pos_{i=1}}^n(\mathcal{B}_{pos}(\text{Prec}_{pos}(A(x_i), B(x_i)), \text{Prec}_{pos}(B(x_i), A(x_i))))$$

jest miarą podobieństwa możliwości.

Przykład 12. Operacja $S : IVFS(X) \times IVFS(X) \rightarrow L^I$:

$$S_{pos}(A, B) = \mathcal{A}_{i=1}^n \mathcal{B}(\text{Prec}_{pos}(A(x_i), B(x_i)), \text{Prec}_{pos}(B(x_i), A(x_i)))$$

jest miarą podobieństwa możliwości, gdzie $\mathcal{A} = \mathcal{A}_{mean}$, $\mathcal{B} = \wedge$ dla Prec_{pos} spełniającego twierdzenie 3.5.

3.2.3 Miara podobieństwa konieczności

Kolejną przedstawioną nową miarą podobieństwa, opartą na relacji porównywalności konieczności jest miara podobieństwa konieczności (pesymistyczna miara podobieństwa).

Definicja 25. Niech $A_1 : [0, 1]^n \rightarrow [0, 1]$ będzie funkcją agregacji. Funkcja $S_{nec} : (IVFS(X))^2 \rightarrow L^I$ która spełnia następujące warunki

$$\text{SN1 } S_{nec}(A, B) = S_{nec}(B, A);$$

$$\text{SN2 } S_{nec}(A, A) = [1 - A_1(w(A(x_1)), \dots, w(A(x_n))), 1];$$

$$\text{SN3 } S_{nec}(A, B) = 0_{L^I}, \text{ jeśli } \{A(x_i), B(x_i)\} = \{0_{L^I}, 1_{L^I}\}, \text{ dla } i = 1, \dots, n;$$

$$\text{SN4 } \text{Jeśli } A \preceq_{nec} B \preceq_{nec} C, \text{ wtedy } S(A, C) \leq_{nec} S(A, B) \text{ i } S(A, C) \leq_{nec} S(B, C),$$

jest nazywana **miarą podobieństwa konieczności** (ang. necessary similarity measure) pomiędzy przedziałowymi zbiorami rozmytymi, gdzie

$$A \preceq_{nec} B \Leftrightarrow a_i \leq_{nec} b_i \text{ dla } i = 1, \dots, n.$$

Przykład 13. Dla $\mathcal{A} = \mathcal{A}_{mean}$, $\mathcal{B} = \wedge$ operacja $S : IVFS(X) \times IVFS(X) \rightarrow L^I$:

$$S_{nec}(A, B) = \mathcal{A}_{i=1}^n \mathcal{B}(\text{Prec}_{nec}(A(x_i), B(x_i)), \text{Prec}_{nec}(B(x_i), A(x_i)))$$

jest miarą podobieństwa konieczności.

3.3 Entropia przedziałowych zbiorów rozmytych

Entropia to pojęcie, które funkcjonuje w naukach przyrodniczych, socjologii, ekonomii, czy teorii informacji i odpowiednio może określać nieuporządkowanie, niejednorodność, przypadkowość, czy ilość informacji. Krótko mówiąc, im większa entropia, tym większa niepewność. Pierwszy raz entropia została zdefiniowana w XIX wieku przez fizyków w termodynamice jako funkcja stanu określająca kierunek zachodzących zmian i przepływu energii [94]. Bardzo szybko zaczęto koncepcję entropii termodynamicznej przekładać na inne dziedziny naukowe oraz łączyć z tak abstrakcyjnym tworem jak informacja.

Entropia w teorii informacji została wprowadzona w 1948 roku przez Claude'a Shannona [95], który zdefiniował ją dla zmiennej losowej jako średni poziom informacji lub niepewności związany z możliwymi wynikami zmiennej. Entropia informacji (zwana też entropią Shannona) zdefiniowana jest za pomocą wzoru:

$$H = - \sum_{i=1}^n p_i \cdot \log_b p_i,$$

gdzie:

p_i - prawdopodobieństwo zajścia i-tego zdarzenia;

$\log_b$ - logarytm o podstawie b , gdzie $b = 2$ dla informacji binarnej.

Miara entropii pełni kluczową rolę w teorii zbiorów rozmytych ze względu na swoją wysoką użyteczność jako miara poziomu niepewności lub stopnia rozmycia danych [6, 96]. Stąd wynika także zainteresowanie badaniami nad rozszerzeniem pojęcia miary entropii w celu zastosowania go do przedziałowych zbiorów rozmytych [89, 97, 98]. Badania podjęte w ramach pracy doktorskiej koncentrowały się na porównaniu i analizie różnych podejść do definiowania entropii. Do budowy entropii wykorzystujemy wskaźnik pierwszeństwa oraz miarę podobieństwa pomiędzy przedziałowymi zbiorami rozmytymi, które odzwierciedlają niepewność.

Warty podkreślenia jest fakt, że rozważania prowadzone w ramach rozprawy dotyczą miary entropii przedziałowej, czyli nie tylko wyznaczamy entropię dla przedziałowych zbiorów rozmytych, ale także wynik jest przedziałem, co uogólnia najczęściej stosowane podejścia z literatury.

Rozważymy miarę entropii dla przedziałowych zbiorów rozmytych, analogicznie jak miarę podobieństwa w trzech klasach: standardowej (klasycznej), możliwości oraz konieczności. Aby skonstruować nową entropię przedziałową, potrzebujemy przedziałowej funkcji agregacji oraz wskaźnika pierwszeństwa zgodnego z odpowiednim porządkiem (relacją porównywalności), a także musimy uwzględnić szerokość przedziałów.

W kolejnej części pracy przyjęto oznaczenie $\mathbf{P} \in IVFS(X)$, gdy $\mathbf{P}(x) = p$ dla $p \in L^I$ i $x \in X$.

3.3.1 Miara entropii klasyczna

Wzorując się na [96] w pracy [48] przedstawiono następującą definicję entropii dla porządku częściowego lub liniowego.

Definicja 26. Niech N będzie silną, przedziałową negacją z punktem równowagi $e \in L^I$ (gdzie $N(e) = e$). Funkcja $E : IVFS(X) \rightarrow L^I$ jest przedziałową miarą entropii w $IVFS(X)$ w odniesieniu do negacji N , jeśli dla $A, B \in IVFS(X)$:

E1 $E(A) = 0_{L^I}$, jeśli A jest zbiorem nierozmytym (ostrym);

E2 $E(\mathbf{E}) = [1 - w(e), 1]$, gdzie $\mathbf{E}(x) = e$ dla każdego $x \in X$;

E3 $E(A) \leq E(B)$, jeśli $A(x) \leq B(x) \leq e$ lub $A(x) \geq B(x) \geq e$ dla każdego $x \in X$.

Poniżej przedstawiono sposób konstrukcji przedziałowej entropii zgodnej z porządkiem częściowym wykorzystującym wskaźnik pierwszeństwa i przedziałowe miary agregacji w ujęciu klasycznym.

Twierdzenie 3.11 ([48]). *Niech $Prec$ będzie wskaźnikiem pierwszeństwa, N będzie silną, przedziałową negacją z punktem równowagi e i $\mathcal{A}$ będzie przedziałową, idempotentną*

funkcją agregacji. Wówczas funkcja $E : IVFS(X) \rightarrow L^I$:

$$E(A) = \mathcal{A}_{i=1}^n \text{Prec}_{\mathcal{A}}(A(x_i) \vee A_N(x_i), N(A(x_i) \vee A_N(x_i))) \quad (3.2)$$

jest przedziałową miarą entropii względem N .

Poniżej przedstawiono uzupełniony dowód w stosunku do dowodu prezentowanego w [48].

Dowód. Jeśli A jest ostry, wtedy $(A \vee A_N)(x_i) = 1_{L^I}$ dla każdego $x_i \in X$, zatem $\text{Prec}_{\mathcal{A}}(A(x_i) \vee A_N(x_i), N(A(x_i) \vee A_N(x_i))) = 0_{L^I}$ i z własności funkcji agregacji otrzymujemy $E(A) = 0_{L^I}$ oraz aksjomat **E1** jest dowiedziony.

Natomiast **E2** można otrzymać na podstawie **P3** (z definicji 19) i idempotentności funkcji agregacji $\mathcal{A}$.

Jeśli $A \preceq B \preceq \mathbf{E}$, wówczas $A \preceq B \preceq \mathbf{E} \preceq B_N \preceq A_N$ a w konsekwencji aksjomatu **P4** otrzymujemy:

$$\begin{aligned} E(A) &= \mathcal{A}_{i=1}^n \text{Prec}_{\mathcal{A}}(A(x_i) \vee A_N(x_i), N(A(x_i) \vee A_N(x_i))) = \\ &\mathcal{A}_{i=1}^n \text{Prec}_{\mathcal{A}}(A_N(x_i), A(x_i)) \leq \mathcal{A}_{i=1}^n \text{Prec}_{\mathcal{A}}(B_N(x_i), A(x_i)) \\ &\leq \mathcal{A}_{i=1}^n \text{Prec}_{\mathcal{A}}(B_N(x_i), B(x_i)) = E(B) \end{aligned}$$

dla $i = 1, \dots, n$.

W podobny sposób przebiega dowód przypadku $A \succeq B \succeq \mathbf{E}$. Jeśli $A \succeq B \succeq \mathbf{E}$, wówczas $A \succeq B \succeq \mathbf{E} \succeq B_N \succeq A_N$ a w konsekwencji **P4** otrzymujemy:

$$\begin{aligned} E(A) &= \mathcal{A}_{i=1}^n \text{Prec}_{\mathcal{A}}(A(x_i) \vee A_N(x_i), N(A(x_i) \vee A_N(x_i))) = \\ &\mathcal{A}_{i=1}^n \text{Prec}_{\mathcal{A}}(A(x_i), A_N(x_i)) \leq \mathcal{A}_{i=1}^n \text{Prec}_{\mathcal{A}}(B(x_i), A_N(x_i)) \\ &\leq \mathcal{A}_{i=1}^n \text{Prec}_{\mathcal{A}}(B(x_i), B_N(x_i)) = E(B) \end{aligned}$$

dla $i = 1, \dots, n$. Co kończy dowód. □

Przykład 14. Dla przedziałowej funkcji agregacji $\mathcal{A}_{mean}$, wskaźnika pierwszeństwa

$$\text{Prec}_{\mathcal{A}} \in \{\text{Prec}_{\mathcal{A}_{meanLI}}, \text{Prec}_{\mathcal{A}_{meanpow}}, \text{Prec}_{\mathcal{A}_{meanmax}}\}$$

spełniającego twierdzenie 3.3 i przedziałowej negacji $N(x) = [1 - \bar{x}, 1 - \underline{x}]$ otrzymujemy entropię spełniającą równanie (3.2).

Poniżej została przedstawiona druga metoda konstrukcji przedziałowej entropii. Jest

ona zgodna z porządkiem częściowym i liniowym oraz oparta na przedziałowej mierze podobieństwa.

Twierdzenie 3.12 ([48]). *Niech S będzie miarą podobieństwa spełniającą warunki twierdzenia 3.7, gdzie A_1 jest idempotentną funkcją agregacji. Wówczas funkcja $E : IVFS(X) \rightarrow L^I$:*

$$E(A) = S(A, A_N) \quad (3.3)$$

jest przedziałową entropią w odniesieniu do N (inwolutywnej, przedziałowej negacji z punktem równowagi e).

Poniżej przedstawiono uzupełniony dowód w stosunku do [48].

Dowód. Jeśli A jest ostry, wówczas

$$\{\text{Prec}(A(x_i), B(x_i)), \text{Prec}(B(x_i), A(x_i))\} \in \{0_{L^I}, 1_{L^I}\}$$

i ponieważ $\mathcal{B}$ ma element neutralny 1_{L^I} (zobacz twierdzenie 3.7) otrzymujemy $E(A) = 0_{L^I}$, co dowodzi **E1** dla wszystkich $x_i \in X$.

E2 możemy otrzymać na podstawie **SIMv2** i idempotentności funkcji agregacji A_1 .

Jeśli $A \preceq B \preceq \mathbf{E}$, wtedy $A \preceq B \preceq \mathbf{E} \preceq B_N \preceq A_N$ a w konsekwencji **SIMv4** otrzymujemy

$$E(A) = S(A, A_N) \leq S(A, B_N) \leq S(B, B_N) = E(B).$$

W podobny sposób korzystając z aksjomatu **SIMv1** można udowodnić przypadek $A \succeq B \succeq \mathbf{E}$.

Jeśli $A \succeq B \succeq \mathbf{E}$, wówczas $A \succeq B \succeq \mathbf{E} \succeq B_N \succeq A_N$ a w konsekwencji **SIMv4** otrzymujemy

$$E(A) = S(A, A_N) \geq S(A, B_N) \geq S(B, B_N) = E(B).$$

Co kończy dowód **E3** i twierdzenie 3.12.

□

Przykład 15. Wykorzystując miary podobieństwa z przykładu 11 z $\text{Prec}_{\mathcal{A}_{meanLI}}$, $\text{Prec}_{\mathcal{A}_{meanpow}}$, $\text{Prec}_{\mathcal{A}_{meanmax}}$, otrzymamy trzy miary entropii spełniające równanie (3.3) w odniesieniu do $N(x) = [1 - \bar{x}, 1 - \underline{x}]$.

Analogicznie, stosując miary podobieństwa optymistyczną (możliwości) i pesymistyczną (konieczności), można uzyskać dwie kolejne klasy entropii: optymistyczną (możliwości) i pesymistyczną (konieczności).

3.3.2 Miara entropii możliwości

W tym podrozdziale przedstawiono nową miarę entropii możliwości nazywaną również optymistyczną miarą entropii uwzględniającą relację możliwości porównywania przedziałów i będącą ogólniejszą miarą entropii niż miary w ujęciu klasycznym (zob. zależności w rozdziale 2.2.1).

Definicja 27. Niech N_{pos} będzie silną (inwolutywną) negacją możliwości z punktem równowagi (*ang. equilibrium point*) $e \in L^I$ ($N_{pos}(e) = e$). Funkcja $E_{pos} : IVFS(X) \rightarrow L^I$ jest miarą entropii możliwości w IVFS(X) w stosunku do negacji N , jeśli dla $A, B \in IVFS(X)$:

EP1 $E_{pos}(A) = 0_{L^I}$, jeśli A jest ostry;

EP2 $E_{pos}(\mathbf{E}) = [1 - w(e), 1]$;

EP3 $E_{pos}(A) \leq_{pos} E_{pos}(B)$, jeśli $A(x) \preceq_{pos} B(x) \preceq_{pos} e$ lub $A(x) \succeq_{pos} B(x) \succeq_{pos} e$ dla wszystkich $x \in X$, gdzie $\mathbf{E}(x) = e$.

W kolejnym twierdzeniu przedstawiona zostanie metoda konstrukcji przedziałowej miary entropii możliwości wykorzystująca wskaźnik pierwszeństwa.

Twierdzenie 3.13. Niech Prec_{pos} będzie wskaźnikiem pierwszeństwa możliwości, N_{pos} będzie silną (inwolutywną) negacją możliwości z punktem równowagi e i $\mathcal{A}_{pos}$ będzie idempotentną funkcją agregacji możliwości. Wtedy funkcja $E_{pos} : IVFS(X) \rightarrow L^I$:

$$E_{pos}(A) = \mathcal{A}_{pos, i=1}^n \text{Prec}_{\mathcal{A}_{pos}}(A(x_i) \vee A_{N_{pos}}(x_i), N_{pos}(A(x_i) \vee A_{N_{pos}}(x_i))) \quad (3.4)$$

jest miarą entropii możliwości w odniesieniu do N_{pos} .

Dowód. Jeśli A jest ostry, wtedy $(A \vee A_{N_{pos}})(x_i) = 1_{L^I}$ dla każdego i , więc $\text{Prec}_{\mathcal{A}_{pos}}(A(x_i) \vee A_{N_{pos}}(x_i), N_{pos}(A(x_i) \vee A_{N_{pos}}(x_i))) = 0_{L^I}$ dla każdego i przez co w konsekwencji otrzymujemy z własności granicznych funkcji agregacji $\mathcal{A}_{pos}$: $E_{pos}(A) = 0_{L^I}$, co dowodzi **EP1**. **EP2** można otrzymać przez **PP3** i idempotentność funkcji agregacji $\mathcal{A}_{pos}$. Jeśli $A \preceq_{pos} B \preceq_{pos} \mathbf{E}$, wtedy $A \preceq_{pos} B \preceq_{pos} \mathbf{E} \preceq_{pos} B_{N_{pos}} \preceq_{pos} A_{N_{pos}}$ i w konsekwencji **PP4** przez $A_{N_{pos}} \preceq_{pos} A \vee A_{N_{pos}}$, więc $A \succeq_{pos} N_{pos}(A \vee A_{N_{pos}})$, a także uwzględniając specyfikę $\preceq_{pos}$ mamy $B \vee B_{N_{pos}} \preceq_{pos} B_{N_{pos}}$ i $N_{pos}(B \vee B_{N_{pos}}) \succeq_{pos} B$ otrzymujemy kolejno:

$$E_{pos}(A) = \mathcal{A}_{pos, i=1}^n \text{Prec}_{\mathcal{A}_{pos}}(A(x_i) \vee A_{N_{pos}}(x_i), N_{pos}(A(x_i) \vee A_{N_{pos}}(x_i))) \leq_{pos}$$

$$\mathcal{A}_{pos, i=1}^n \text{Prec}_{\mathcal{A}_{pos}}(A_{N_{pos}}(x_i), A(x_i)) \leq_{pos} \mathcal{A}_{pos, i=1}^n \text{Prec}_{\mathcal{A}_{pos}}(B_{N_{pos}}(x_i), A(x_i))$$

$$\leq_{pos} \mathcal{A}_{pos, i=1}^n \text{Prec}_{\mathcal{A}_{pos}}(B_{N_{pos}}(x_i), B(x_i)) \leq_{pos}$$

$$\mathcal{A}_{pos_{i=1}}^n \text{Prec}_{\mathcal{A}_{pos}}(B(x_i) \vee B_{N_{pos}}(x_i), N_{pos}(B(x_i) \vee B_{N_{pos}}(x_i))) = E_{pos}(B)$$

dla $i = 1, \dots, n$.

W podobny sposób można udowodnić przypadek $A \succeq_{pos} B \succeq_{pos} \mathbf{E}$. Co kończy dowód. $\square$

Przykład 16. Wykorzystując $\mathcal{A} = \mathcal{A}_{mean}$ funkcję agregacji możliwości i wskaźnik pierwszeństwa

$$\text{Prec}_{\mathcal{A}_{pos}} \in \{\text{Prec}_{\mathcal{A}_{pos_1}}, \text{Prec}_{\mathcal{A}_{pos_2}}\}$$

otrzymujemy entropię możliwości spełniającą równanie (3.4) przy zastosowaniu standardowej negacji przedziałowej N_{pos} (kończności).

Poniżej przedstawiono drugą metodę konstrukcji przedziałowej entropii możliwości, opartą na miarze podobieństwa możliwości.

Twierdzenie 3.14. *Niech S_{pos} będzie miarą podobieństwa możliwości spełniającą twierdzenie 3.10, gdzie A_1 będzie idempotentną funkcją agregacji. Wówczas funkcja $E : IVFS(X) \rightarrow L^I$:*

$$E_{pos}(A) = S_{pos}(A, A_{N_{pos}}) \quad (3.5)$$

jest przedziałową optymistyczną entropią w odniesieniu do N_{pos} (inwolutywnej, przedziałowej optymistycznej negacji z punktem równowagi e).

Dowód. Jeśli A jest ostry, wtedy

$$\{\text{Prec}_{pos}(A(x_i), A_{N_{pos}}(x_i)), \text{Prec}_{pos}(A_{N_{pos}}(x_i), A(x_i))\} \in \{0_{L^I}, 1_{L^I}\}$$

i ponieważ $\mathcal{B}_{pos}$ (zob. twierdzenie 3.10) posiada element neutralny 1_{L^I} otrzymujemy $E_{pos}(A) = 0_{L^I}$, co dowodzi **EP1** dla każdego $x_i \in X$.

EP2 można wykazać przy wykorzystaniu **SP2** i idempotentności A_1 .

Jeśli $A \preceq_{pos} B \preceq_{pos} \mathbf{E}$, wtedy $A \preceq_{pos} B \preceq_{pos} \mathbf{E} \preceq_{pos} B_{N_{pos}} \preceq_{pos} A_{N_{pos}}$ a w konsekwencji **SP4** otrzymujemy

$$E_{pos}(A) = S_{pos}(A, A_{N_{pos}}) \leq_{pos} S_{pos}(A, B_{N_{pos}}) \leq_{pos} S_{pos}(B, B_{N_{pos}}) = E_{pos}(B).$$

W podobny sposób i przez **SP1** można udowodnić przypadek $A \succeq_{pos} B \succeq_{pos} \mathbf{E}$. Co kończy dowód **EP3** i twierdzenie 3.14. $\square$

Przykład 17. Jeśli wykorzystana zostanie miara podobieństwa możliwości z przykładu 12 wówczas otrzymamy miarę entropii, która spełnia równanie (3.5) w odniesieniu do N_{pos} .

3.3.3 Miara entropii konieczności

Kolejną klasą entropii jest miara entropii konieczności nazywana również pesymistyczną miarą entropii, która wykorzystuje relację konieczności do porównywania przedziałów i jest zatem bardziej restrykcyjną od entropii możliwości i klasycznej.

Definicja 28. Niech N_{nec} będzie silną (inwolutywną), negacją konieczności z punktem równowagi $e \in L^I$. Funkcja $E_{nec} : IVFS(X) \rightarrow L^I$ jest miarą entropii konieczności w IVFS(X) w odniesieniu do negacji N_{nec} , jeśli dla $A, B \in IVFS(X)$:

EN1 $E_{nec}(A) = 0_{L^I}$, jeśli A jest ostry;

EN2 $E_{nec}(\mathbf{E}) = [1 - w(e), 1]$;

EN3 $E_{nec}(A) \leq E_{nec}(B)$, jeśli $A(x) \leq_{nec} B(x) \leq_{nec} e$ lub $A(x) \geq_{nec} B(x) \geq_{nec} e$ dla każdego $x \in X$ i $\mathbf{E}(x) = e$.

W kolejnym twierdzeniu przedstawiono metodę konstrukcji miary entropii konieczności wykorzystującą wskaźnik pierwszeństwa.

Twierdzenie 3.15. Niech $Prec_{nec}$ będzie wskaźnikiem pierwszeństwa konieczności, N_{nec} będzie silną (inwolutywną), negacją konieczności z punktem równowagi e i $\mathcal{A}_{nec}$ będzie idempotentną funkcją agregacji konieczności. Wtedy, funkcja $E_{nec} : IVFS(X) \rightarrow L^I$:

$$E_{nec}(A) = \mathcal{A}_{nec_{i=1}^n} Prec_{\mathcal{A}_{nec}}(A(x_i) \vee A_{N_{nec}}(x_i), N_{nec}(A(x_i) \vee A_{N_{nec}}(x_i))) \quad (3.6)$$

jest miarą entropii konieczności w odniesieniu do N_{nec} .

Dowód. Jeśli A jest ostry, wtedy $(A \vee A_{N_{nec}})(x_i) = 1_{L^I}$ dla każdego i , więc $Prec_{\mathcal{A}_{nec}}(A(x_i) \vee A_{N_{nec}}(x_i), N_{nec}(A(x_i) \vee A_{N_{nec}}(x_i))) = 0_{L^I}$ dla każdego i przez co jako konsekwencja z własności funkcji agregacji otrzymujemy $E_{nec}(A) = 0_{L^I}$ i **EN1** jest udowodniony. **EN2** można uzyskać przy spełnieniu **PN3** i idempotentności funkcji agregacji $\mathcal{A}_{nec}$.

Jeśli $A \preceq_{nec} B \preceq_{nec} \mathbf{E}$, wtedy $A \preceq_{nec} B \preceq_{nec} \mathbf{E} \preceq_{nec} B_N \preceq_{nec} A_N$ a w konsekwencji **PN4** otrzymujemy

$$\begin{aligned} E_{nec}(A) &= \mathcal{A}_{nec_{i=1}^n} Prec_{\mathcal{A}_{nec}}(A(x_i) \vee A_{N_{nec}}(x_i), N_{nec}(A(x_i) \vee A_{N_{nec}}(x_i))) = \\ &\mathcal{A}_{nec_{i=1}^n} Prec_{\mathcal{A}_{nec}}(A_{N_{nec}}(x_i), A(x_i)) \leq_{nec} \mathcal{A}_{nec_{i=1}^n} Prec_{\mathcal{A}_{nec}}(B_{N_{nec}}(x_i), A(x_i)) \\ &\leq_{nec} \mathcal{A}_{nec_{i=1}^n} Prec_{\mathcal{A}_{nec}}(B_{N_{nec}}(x_i), B(x_i)) = E_{nec}(B) \end{aligned}$$

dla $i = 1, \dots, n$.

W podobny sposób można udowodnić przypadek $A \succeq_{nec} B \succeq_{nec} \mathbf{E}$. Co kończy dowód. $\square$

Przykład 18. Wykorzystując $\mathcal{A}_{nec}$ przedziałową funkcję agregacji konieczności z przykładu 5, wskaźnik pierwszeństwa $\text{Prec}_{\mathcal{A}_{nec}}$ i negację przedziałową $N(x) = [1 - \bar{x}, 1 - \underline{x}]$ otrzymujemy entropię spełniającą równanie (3.6).

Poniżej przedstawiono kolejną metodę konstrukcji przedziałowej entropii konieczności wykorzystującą miarę podobieństwa konieczności.

Twierdzenie 3.16. *Niech S_{nec} jest miarą podobieństwa konieczności z przykładu 13, gdzie $\mathcal{A}_1$ będzie idempotentną funkcją agregacji. Wówczas funkcja $E_{nec} : IVFS(X) \rightarrow L^I$:*

$$E_{nec}(A) = S_{nec}(A, A_{N_{nec}}) \quad (3.7)$$

jest przedziałową entropią konieczności w odniesieniu do N_{nec} (inwolutywnej, przedziałowej negacji konieczności z punktem równowagi e).

Dowód. Jeśli A jest ostry, wtedy dla każdego $x_i \in X$

$$\{\text{Prec}_{nec}(A(x_i), B(x_i)), \text{Prec}_{nec}(B(x_i), A(x_i))\} \in \{0_{L^I}, 1_{L^I}\}$$

i ponieważ $\mathcal{B}_{nec}$ posiada element neutralny 1_{L^I} otrzymujemy $E_{nec}(A) = 0_{L^I}$.

EN2 możemy otrzymać bezpośrednio przez **SN2** i idempotentność funkcji agregacji $\mathcal{A}_{nec}$. Jeśli $A \preceq_{nec} B \preceq_{nec} \mathbf{E}$, wtedy $A \preceq_{nec} B \preceq_{nec} \mathbf{E} \preceq_{nec} B_{N_{nec}} \preceq_{nec} A_{N_{nec}}$ i w konsekwencji **SN4** otrzymujemy

$$E_{nec}(A) = S_{nec}(A, A_{N_{nec}}) \leq_{nec} S_{nec}(A, B_{N_{nec}}) \leq_{nec} S_{nec}(B, B_{N_{nec}}) = E_{nec}(B).$$

W podobny sposób i przez **SN1** można udowodnić przypadek $A \succeq_{nec} B \succeq_{nec} \mathbf{E}$. Co kończy dowód **EN3** i twierdzenie 3.16. □

Przykład 19. Jeśli użyje się miar podobieństwa z przykładu 13 z Prec_{nec} z twierdzenia 3.6, wówczas otrzymujemy miarę entropii spełniającą równanie (3.7) w stosunku do N_{nec} .

ROZDZIAŁ 4

Wybrane metody klasyfikacji

Klasyfikacja jest jedną z najczęściej stosowanych metod w analizie i eksploracji danych. Celem klasyfikacji jest konstrukcja uogólnionego modelu decyzyjnego, umożliwiającego przypisanie nieznanym obiektów do jednej z ustalonych klas na podstawie zbioru danych uczących. Opracowany model wykorzystywany jest następnie do predykcji przynależności klasowej nowych obiektów, dla których etykieta klasowa nie została uprzednio określona [99]. Sytuacja gdy klasyfikator może nie zaproponować żadnej wartości decyzji powinna zasugerować, że wybór decyzji jest trudny i należy rozważyć szerszy kontekst, jeśli to możliwe, aby pozyskać więcej informacji, które mogą pomóc w podjęciu decyzji. Przykładem takiej sytuacji może być zbiór danych medycznych, w którym obiektami są pacjenci opisani wartościami wyników badań, jakie reprezentowane są przez poszczególne atrybuty warunkowe. W przypadku, gdy klasyfikator nie proponuje jednoznacznej decyzji dla pacjenta, sugeruje to konieczność przeprowadzenia dodatkowych badań diagnostycznych i rozszerzenia zbioru danych o nowe atrybuty.

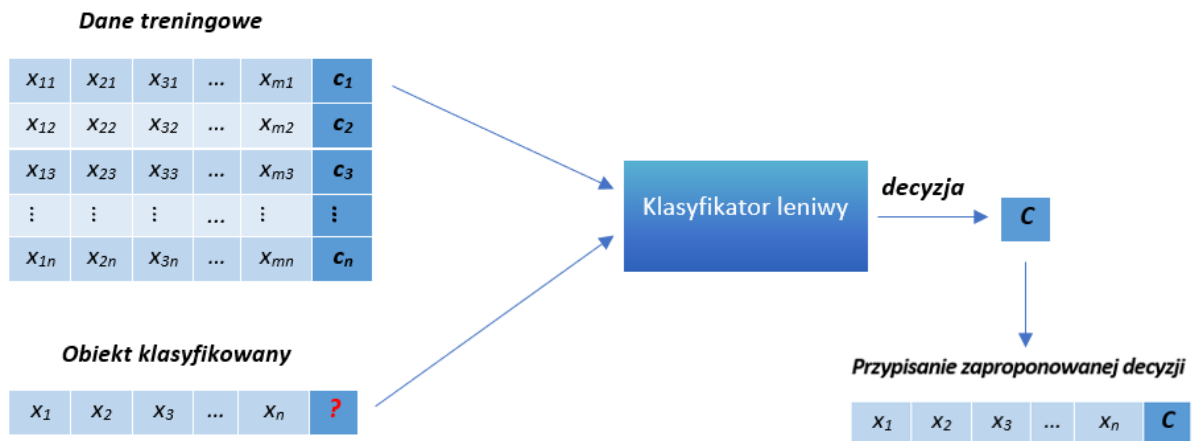
4.1 Przegląd wybranych metod klasyfikacji

Klasyfikacja jest jedną z najstarszych i jednocześnie najczęściej stosowanych metod analizy oraz eksploracji danych. Celem całego procesu klasyfikacji jest zdefiniowanie modelu klasyfikacyjnego dla określonego zestawu klas obiektów na podstawie dostępnego zbioru danych "uczących" o znanych klasach, a następnie wykorzystanie tego modelu do predykcji klasy nowych obiektów, dla których klasa nie jest jeszcze znana.

Metody klasyfikacji znajdują obecnie zastosowanie w niemal wszystkich dziedzinach, takich jak bankowość, medycyna, marketing i wiele innych.

Pierwszym rodzajem klasyfikatorów, który zostanie omówiony, jest grupa klasyfikatorów tzw. leniwych [100]. Są to klasyfikatory, które nie tworzą wewnętrznej reprezentacji wiedzy, lecz poszukają rozwiązania po prezentacji każdego wzorca testowego przeszuku-

jąc dane uczące. Główną zaletą klasyfikatorów leniwych jest natychmiastowa gotowość do klasyfikacji nowych obiektów oraz możliwość uzupełniania zbioru uczącego w trakcie działania algorytmu. Poniższy rysunek przedstawia schemat przebiegu procesu klasyfikacji z wykorzystaniem klasyfikatora leniwego.



Rysunek 4.1: Schemat procesu klasyfikacji przeprowadzonej z użyciem klasyfikatorów leniwych

Na prezentowanym schemacie zbiór obiektów treningowych zawiera n obiektów opisanych przez m atrybutów warunkowych. Każdy z obiektów należy do jednej z klas C .

Algorytm k -najbliższych sąsiadów (k -NN)

Przykładem klasyfikatora leniwego jest algorytm k -NN [101]. Metoda k najbliższych sąsiadów (k -NN) wymaga przechowywania pełnego zbioru danych uczących, względem którego obliczana jest odległość do nowego (testowego) obiektu. Działanie algorytmu opiera się na dwóch podstawowych założeniach:

- obiekty znajdujące się blisko siebie w przestrzeni cech mają wysokie prawdopodobieństwo należenia do tej samej klasy decyzyjnej,
- obiektowi nieoznaczonemu (bez przypisanej decyzji) należy przypisać etykietę klasy na podstawie etykiet jego k najbliższych sąsiadów, czyli najbardziej podobnych obiektów.

W algorytmie występują dwa parametry, które można modyfikować w zależności od potrzeb, a są nimi:

- metoda oceny podobieństwa obiektów względem siebie,
- liczba obiektów, które mają być wyznaczane jako podobne do obiektu rozważanego.

Algorytm k -NN klasyfikuje nowy obiekt przez przypisanie decyzji, jaką posiada większa liczba spośród wyznaczonych k najbliższych sąsiadów. W przypadku klasyfikatorów

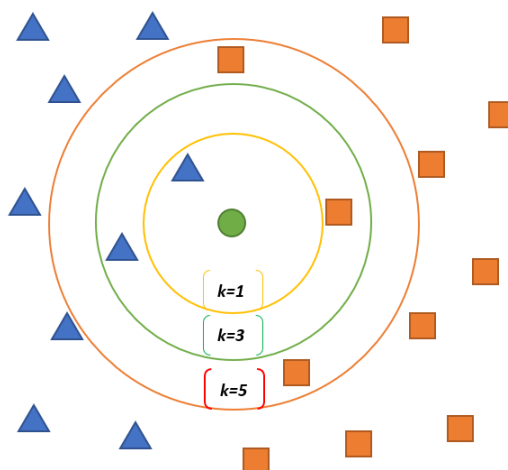
binarnych, aby otrzymać większość obiektów przynależących do jednej z klas decyzyjnych, stosuje się nieparzystą liczbę najbliższych obiektów. Jeśli chodzi o wybór metody oceny podobieństwa, stosuje się często, w przypadku klasycznym, poniżej zamieszczone przykłady metryk dla danych numerycznych:

- metryka euklidesa

$$D_{Euk}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}, \quad (4.1)$$

- metryka Manhattan (miejska)

$$D_{Man}(x, y) = \sum_{i=1}^n |x_i - y_i|.$$



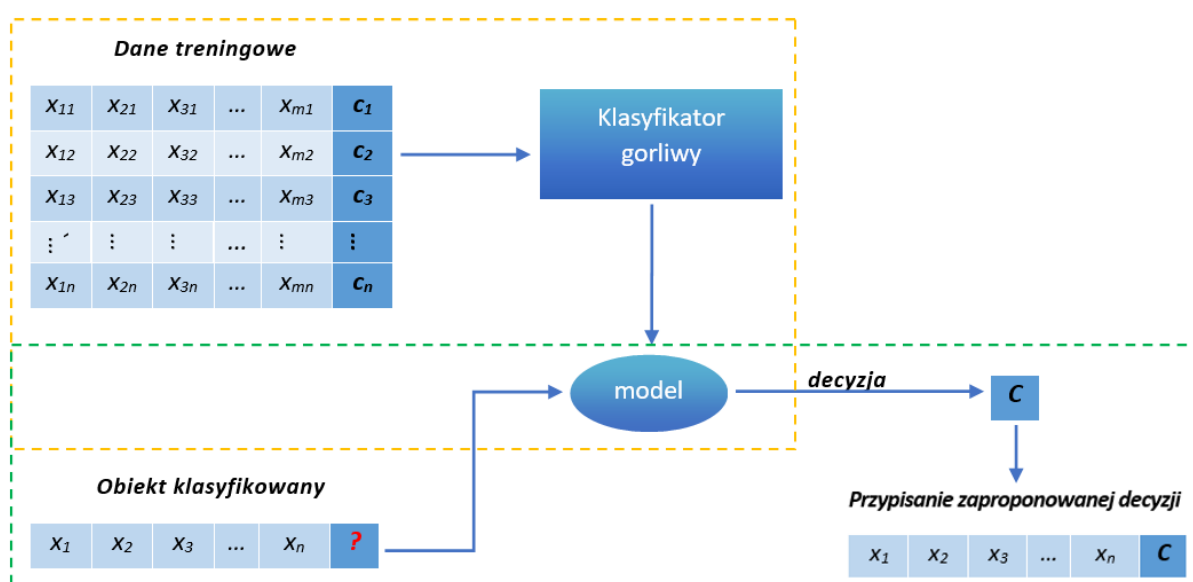
Rysunek 4.2: Schemat znajdowania najbliższych sąsiadów w otoczeniu obiektu klasyfikowanego z użyciem metryki euklidesowej

Rysunek 4.2 przedstawia schemat działania algorytmu k -NN. Znajdują się na nim obiekty z dwóch klas: trójkąt i kwadrat. Zielone koło to nowy obiekt, który ma zostać sklasyfikowany, czyli należy przypisać mu jedną z istniejących klas. W tym celu sprawdzane jest jego otoczenie przez wyznaczenie obiektów do niego najbardziej podobnych. Służą do tego wcześniej przytoczone metryki lub badane w pracy miary podobieństwa. Jak można zauważyć, przynależność do klasy rozważanego obiektu może być zależna od wcześniej przyjętej liczby poszukiwanych najbliższych obiektów k , dla których chcemy zaproponować decyzję.

Do głównych zalet algorytmu k -NN należy zaliczyć prosty proces użycia i implementacji, łatwość w wyjaśnieniu, jak doszło do ustalenia (predykcji) klasy, oraz odporność na występowanie obiektów izolowanych dzięki ocenie najbliższych sąsiadów. Ze względu na

wspomniane zalety algorytmu k -NN w dalszej części rozprawy zostanie zaproponowany algorytm oparty na technice k -NN, ale z uwzględnieniem danych i operatorów czy miar z niepewnością. Wadą jest konieczność przechowywania całego zbioru uczącego i jego przeglądania każdorazowo w chwili klasyfikowania nowego obiektu.

Algorytmy uczenia się, zwane gorliwymi, to takie, które najpierw rozwijają model, a następnie wykorzystują go do określonego zadania obliczeniowego, np. klasyfikacji. Klasyfikatory gorliwe są grupą klasyfikatorów, które tworzą modele klasyfikacji opisujące zależności odkryte w zestawie danych treningowych. Mowa tutaj o zależnościach pomiędzy konkretną klasą decyzyjną a atrybutami obiektów, a dokładniej - wartościami tych atrybutów. Gdy model zostanie utworzony, nie ma konieczności przechowywania danych treningowych, ponieważ już nie będą one używane. Warto jednak w przyszłości ponowić proces uczenia na poszerzonym zbiorze danych, wzbogaconym o nowe zestawy obiektów uczących. Zaletą tego typu klasyfikacji jest stosunkowo krótki czas klasyfikacji. Rysunek 4.3 przedstawia schemat procesu klasyfikacji z wykorzystaniem klasyfikatorów gorliwych.



Rysunek 4.3: Schemat procesu klasyfikacji przeprowadzonej z użyciem klasyfikatorów gorliwych

Schemat podzielony jest na dwa obszary. Częścią wspólną tych obszarów jest model klasyfikacji, dzięki czemu klasyfikacja staje się dwuetapowa. Pierwszym koniecznym etapem (oznaczonym pomarańczową linią) jest proces definiowania klasyfikatora oraz proces jego uczenia na danych treningowych, a następnie utworzenie modelu klasyfikacji. Drugi etap (oznaczony zieloną linią) dotyczy wykorzystania modelu do klasyfikacji nowego obiektu i przypisania mu decyzji zaproponowanej przez model. Od tej pory, aby sklasyfikować nowy obiekt, wystarczy odwołać się do utworzonego wcześniej modelu, bez konieczności tworzenia go na nowo. Poniżej przedstawiono opis kilku przykładowych rodzin klasyfikatorów gorliwych.

Klasyfikatory regułowe

Popularną techniką budowy modeli klasyfikacji są klasyfikatory regułowe. Model klasyfikacji jest w tym wypadku oparty o zestaw reguł decyzyjnych, które przyjmują postać:

$$rule_x : \text{if } \ll \text{warunek} \gg \text{ then } \ll \text{konkluzja} \gg,$$

można je również zapisać w formie alternatywnej:

$$rule_x : \ll \text{warunek} \gg \rightarrow \ll \text{konkluzja} \gg .$$

Jeśli obiekt spełnia poprzednik reguły, mówi się, że jest on pokryty przez regułę $rule_x$. Poprzednik reguły, będący warunkiem, jest koniunkcją testów zdefiniowanych na atrybutach opisujących obiekty, a następnik reguły określa klasę, do której przynależy obiekt spełniający warunki poprzednika. W fazie treningu generowany jest zbiór reguł D na podstawie tablicy decyzyjnej. Istnieje wiele rodzajów reguł, m.in.: logiczne, decyzyjne (inaczej klasyfikacyjne), regresyjne, asocjacyjne i inne [102]. Zasadę działania zilustrowano na poniższym rysunku.

atrybuty warunkowe				decyzja
id	status	dochód	dzieci	ryzyko
1	kawaler	niski	0	wysokie
2	żonaty	niski	1	wysokie
3	kawaler	wysoki	0	niskie
4	rozwiedziony	niski	2	wysokie
5	żonaty	średni	1	niskie
6	rozwiedziony	wysoki	2	niskie
7	rozwiedziony	niskie	2	wysokie
8	żonaty	średni	2	niskie
9	żonaty	wysoki	2	wysokie
10	rozwiedziony	wysoki	3	niskie

Przykładowy zbiór reguł	
r1:	if dochód = wysoki and dzieci = 0 and status = kawaler then ryzyko = niskie
r2:	if dochód = średni and dzieci = 1 ... 2 then ryzyko = niskie
r3:	if dochód = niski then ryzyko = wysokie

Rysunek 4.4: Przykład zbioru danych i zbioru reguł klasyfikacyjnych.

Przykład dotyczy klasyfikacji oceny ryzyka udzielenia kredytu potencjalnym kredytobiorcom. Warunkiem reguły $r1$ jest koniunkcja trzech kryteriów zdefiniowanych na podstawie atrybutów warunkowych: *dochód*, *status* i *dzieci*. Następnik reguły $r1$ określa decyzję w tym przypadku jako *niskie* ryzyko przyznania kredytu dla osób scharakteryzowanych przez konkretne wartości na poszczególnych cechach zdefiniowanych w warunku reguły. Z kolei w regule $r3$ warunek składa się z jednego kryterium opartym na atrybucie *dochód*, a następnik określa *ryzyko* jako *wysokie*. W powyższym przykładzie można zauważyć, że:

1. reguła $r1$ pokrywa tylko obiekt 3;
2. reguła $r2$ pokrywa obiekty 5 i 8;

3. reguła r_3 pokrywa obiekty 1, 2, 4 i 7.

Ocenę jakości reguł klasyfikacyjnych najczęściej przeprowadza się za pomocą dwóch podstawowych miar: trafności i pokrycia. Pokrycie reguły klasyfikacyjnej r jest definiowane jako stosunek liczby obiektów, które są pokrywane przez tę regułę, do łącznej liczby obiektów w zbiorze danych.

$$\text{pokrycie}(r) = \frac{D_r}{D},$$

gdzie D_r to liczba obiektów ze zbioru D , które są pokrywane przez regułę. Natomiast trafność reguły to stosunek liczby obiektów, które spełniają zarówno poprzednik, jak i następnik reguły ($D_{r \cap y}$) do liczby obiektów, które spełniają poprzednik reguły D_r .

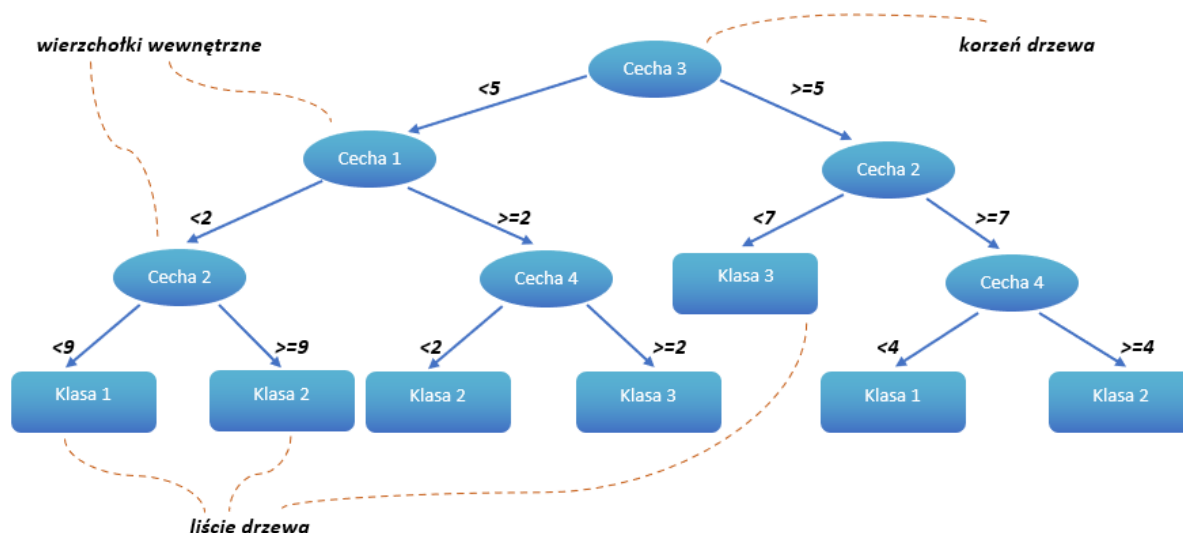
$$\text{trafność}(r) = \frac{D_{r \cap y}}{D_r}.$$

Wyznaczanie klasy decyzyjnej dla nowego obiektu odbywa się poprzez znalezienie wszystkich reguł, które pokrywają klasyfikowany obiekt. Możliwe są następujące scenariusze:

- jeśli nie udało się znaleźć żadnej reguły pokrywającej klasyfikowany obiekt, nie ma możliwości jego klasyfikacji;
- jeśli znaleziono jedną regułę, obiektowi przypisywana jest klasa decyzyjna wskazana przez tę regułę;
- jeśli znaleziono więcej niż jedną regułę, decyzja wyznaczana jest na podstawie pewnego ustalonego schematu głosowania.

Drzewa decyzyjne

Algorytmy indukcji drzew decyzyjnych są kolejnym z przykładów klasyfikatorów gorliwych. Drzewo decyzyjne to model klasyfikacyjny, który można zobrazować jako schemat decyzji w formie grafu, w którym nie ma cykli i każda gałąź prowadzi do wyniku. Konstrukcja drzewa doskonale prezentuje metodę *dziel i zwyciężaj*, która zakłada podział problemu na mniejsze podproblemy, a następnie scalanie ich rozwiązań otrzymując w rezultacie jako wynik rozwiązania pierwotnego, większego problemu. Drzewo decyzyjne zbudowane jest z dwóch typów wierzchołków (tzw. wierzchołków wewnętrznych i liści) oraz krawędzi, co zaprezentowane zostało na rysunku 4.5.



Rysunek 4.5: Schemat ogólny drzewa decyzyjnego.

Wyróżnić można także korzeń drzewa, będący wierzchołkiem wewnętrznym, który nie posiada krawędzi wchodzących, natomiast musi posiadać co najmniej dwie krawędzie wychodzące. Każdy węzeł opisuje jedną cechę (atrybut), a krawędzie reprezentację wartości atrybutu. Liście drzewa to węzły, które nie posiadają krawędzi wychodzących i zawierają etykiety klas decyzyjnych. Zatem do wybranego liścia trafiają wyłącznie obiekty z jednej klasy decyzyjnej. Zaprezentowany przykład widoczny na rysunku 4.5 jest szczególnym rodzajem drzewa decyzyjnego, w którym z każdego z atrybutów wychodzą najwyżej dwie krawędzie; takie drzewo nazywa się drzewem binarnym. Najpopularniejszymi algorytmami reprezentacji drzewa decyzyjnego są algorytmy *CART* [103], *ID3* [104] oraz jego późniejsza modyfikacja *C4.5* [105]. Zalety, jakie płyną z wykorzystywania drzew decyzyjnych to przejrzysta forma i opis procesu klasyfikacji, natomiast główną wadą są trudności w dostosowaniu modelu klasyfikacji do nowych obiektów wykorzystywanych w procesie uczenia. Z zapisu ścieżek przejść po węzłach drzewa i poszczególnych jego krawędziach można zdefiniować reguły, które mogą tworzyć klasyfikator regałow.

Klasyfikatory bayesowskie

Ostatnim przykładem z grupy klasyfikatorów gorliwych są klasyfikatory Bayesowskie, które modelują relacje prawdopodobieństwa pomiędzy zbiorem atrybutów warunkowych a atrybutem decyzyjnym. Typowym przykładem jest tzw. naiwny klasyfikator Bayesa, model oparty na twierdzeniu Bayesa oraz założeniu warunkowej niezależności cech. To założenie można opisać w następujący sposób:

$$P(X|C_i) = \prod_{i=1}^n P(A_i = x_i|C_i), \quad (4.2)$$

gdzie $X = (A_1 = x_1, A_2 = x_2, \dots, A_n = x_n)$ oznacza obiekt, $A = (A_1, A_2, \dots, A_n)$ to zbiór atrybutów warunkowych, a $C = (C_1, C_2, \dots, C_n)$ jest zbiorem klas decyzyjnych. Przyjmuje się również, że liczba atrybutów jest stała dla wszystkich obiektów, jednak w niektórych obiektach, zarówno treningowych jak i klasyfikowanych, może brakować wartości pewnych atrybutów. Na podstawie danych treningowych można oszacować prawdopodobieństwo *a priori* dla każdej klasy C_i oznaczane jako $P(C_i)$, przez stosunek liczby obiektów zbioru treningowego należących do klasy C_i do liczby wszystkich obiektów ze zbioru treningowego. W celu utworzenia modelu klasyfikatora do predykcji decyzji dla obiektu X , rozważając zbiór atrybutów warunkowych A oraz atrybut decyzyjny C jako zmienne losowe, zależność między nimi można opisać jako prawdopodobieństwo warunkowe $P(C|A)$. Naiwny klasyfikator Bayesowski maksymalizuje prawdopodobieństwo warunkowe *a posteriori* $P(C_i|X)$, które można wyznaczyć na podstawie wzoru Bayesa:

$$P(C_i|X) = \frac{P(X|C_i) \cdot P(C_i)}{P(X)},$$

gdzie $P(C_i|X)$ to prawdopodobieństwo *a posteriori*, że $C = C_i$ przy znajomości wartości zbioru atrybutów warunkowych A dla obiektu X natomiast $P(C_i)$ to prawdopodobieństwo *a priori*, że $C = C_i$, bez uwzględniania jakiegokolwiek wiedzy o wartościach na atrybutach dla obiektu X . Natomiast $P(X)$ jest prawdopodobieństwem wystąpienia wartości zbioru atrybutów A obiektu X . Zgodnie z zasadą maksymalizacji prawdopodobieństwa *a posteriori*, klasyfikator przypisuje obiekt X do klasy C_i , która maksymalizuje wartość iloczynu $P(X|C_i) \cdot P(C_i)$.

Zatem obiekt X , dla którego klasa decyzyjna nie jest jeszcze znana, przypisuje się do klasy C_i , która maksymalizuje wartość wyrażenia $P(C_i) \prod_{j=1}^n P(A_j = x_j|C_i)$, co zapisujemy w postaci:

$$C = \arg \max_{C_i} P(C_i) \prod_{j=1}^n P(A_j = x_j|C_i).$$

Zaletami naiwnego klasyfikatora Bayesowskiego są prostota implementacji, szybkość działania oraz możliwość określenia wartości funkcji wsparcia dla każdej z klas decyzyjnych. Jego wady to brak możliwości wizualizacji procesu klasyfikacji, oraz trudność w objaśnieniu procesu podejmowania decyzji.

4.2 Miary jakości klasyfikacji

Ocena jakości modelu klasyfikującego jest kluczowym krokiem w procesie uczenia maszynowego. Przeprowadza się ją, porównując wyniki klasyfikacji obiektów testowych z ich rzeczywistymi etykietami klas decyzyjnych. Pozwala to na określenie, jak skutecznie model radzi sobie z przypisywaniem obiektów do odpowiednich klas. Jednym z podstawowych

narzędzi używanych do oceny jakości klasyfikacji jest macierz pomyłek (*ang. confusion matrix*). Jest to kwadratowa macierz o wymiarach $m \times m$, gdzie m oznacza liczbę różnych etykiet klas decyzyjnych. Wiersze macierzy odpowiadają rzeczywistym etykietom klas obiektów testowych, natomiast kolumny - klasom przypisanym tym obiektom przez klasyfikator. Przykładowo dla dwóch klas element macierzy f_{ij} reprezentuje liczbę obiektów z klasy C_i , które zostały przypisane do klasy C_j . Wtedy wartość f_{12} oznacza liczbę obiektów z klasy C_1 , które klasyfikator błędnie przypisał do klasy C_2 . Macierz pomyłek pozwala na dokładną analizę jakości klasyfikacji, dzięki czemu można łatwo zidentyfikować, które klasy są najczęściej mylone, oraz obszary wymagające poprawy w modelu klasyfikacyjnym.

Klasa rzeczywista	Klasa przewidywana	
	C_1	C_2
C_1	TP (f_{11})	FN (f_{12})
C_2	FP (f_{21})	TN (f_{22})

Poszczególne elementy macierzy pomyłek można opisać za pomocą następujących kategorii, które pomagają zrozumieć, jak model klasyfikacyjny działa w odniesieniu do danych testowych:

- *TP* (*ang. true positive*) to liczba przypadków, w których obiekt należący do klasy C_1 ("pozytywnej") został poprawnie sklasyfikowany jako C_1 . Innymi słowy, są to poprawnie przypisane obiekty do danej klasy.
- *FP* (*ang. false positive*) to liczba przypadków, w których obiekt należący do klasy innej niż C_1 został błędnie przypisany do klasy C_1 . Są to obiekty, które klasyfikator przypisał do klasy C_1 , choć w rzeczywistości należą do innych klas.
- *TN* (*ang. true negative*) to liczba przypadków, w których obiekt należący do innej klasy niż C_1 został poprawnie sklasyfikowany jako nie należący do C_1 . Innymi słowy, klasyfikator poprawnie odrzucił obiekt, który nie należy do klasy C_1 .
- *FN* (*ang. false negative*) to liczba przypadków, w których obiekt należący do klasy C_1 został błędnie przypisany do innej klasy. Są to obiekty, które klasyfikator nieprawidłowo zaklasyfikował jako nie należące do C_1 , mimo że powinny do niej należeć.

Analizując macierz pomyłek, można obliczyć łączną liczbę poprawnych zaklasyfikowań obiektów, sumując elementy na głównej przekątnej macierzy, gdzie wiersze odpowiadają rzeczywistym etykietom, a kolumny – etykietom przypisanym przez klasyfikator. Jednocześnie elementy poza główną przekątną wskazują na liczbę błędnych zaklasyfikowań, co pozwala na przeprowadzenie analizy, jakie typy błędów najczęściej popełnia klasyfikator, np. czy myli obiekty między konkretnymi klasami. Tego typu analiza jest kluczowa do

zrozumienia, w jakich sytuacjach klasyfikator ma tendencję do popełniania błędów i jakie są główne przyczyny nieprawidłowych decyzji. Niech f_{test} oznacza liczbę przypadków testowych. Łączna liczba poprawnie zaklasyfikowanych obiektów testowych $f_{correct}$ wynosi $f_{correct} = \sum_{i=1}^m f_{ii}$, a łączna liczba wszystkich błędnych klasyfikacji to $f_{err} = f_{test} - \sum_{i=1}^m f_{ii}$. Wykorzystując macierz pomyłek można przeprowadzić szczegółową analizę błędnych zaklasyfikowań obiektów dla różnych typów błędów. Ma to istotne znaczenie dla szeregu dziedzin zastosowań. Przykładowo w diagnostyce medycznej błędne zaklasyfikowanie pacjenta chorego na chorobę X (klasa C_i - "pozytywna") do klasy pacjentów chorych na chorobę Y (klasa C_j - "negatywna") może skutkować zastosowaniem nieprawidłowego leczenia, które może mieć tragiczne skutki dla pacjenta. Z drugiej strony, błędne zaklasyfikowanie pacjenta chorego na chorobę Y do klasy pacjentów chorych na chorobę X może nie mieć poważniejszych skutków. W takim przypadku bardziej istotna jest liczba błędnych zaklasyfikowań f_{ij} niż f_{ji} . Podobnie, różne mogą być koszty błędnych decyzji w analizie finansowej, na przykład w ocenie ryzyka kredytowego. Z punktu widzenia banku, poważniejsze skutki będą miały błędne zaklasyfikowania osób ubiegających się o kredyt, które mają wysokie ryzyko kredytowe do grupy kredytobiorców z niskim ryzykiem kredytowym niż odwrotnie. Ma to także znaczenie dla analizy zbiorów danych o nierównomiernym rozkładzie obiektów pomiędzy klasami, gdy użytkownikowi zależy na poprawnym klasyfikowaniu obiektów z klasy posiadającej mniejszą liczbę obiektów. W takiej sytuacji należy posłużyć się analizą zawartości macierzy pomyłek rozszerzonej o dodatkową informację na temat kosztów błędnej klasyfikacji. Niech c_{ij} oznacza koszt błędnego zaklasyfikowania do klasy C_j obiektu należącego do klasy C_i . Koszt poprawnej klasyfikacji c_{ii} przyjmuje się równy 0. Koszt błędnej klasyfikacji $C(f_{err})$ jest łączną sumą wszystkich błędnych klasyfikacji:

$$C(f_{err}) = \sum_{i=1}^m \sum_{j=1}^m f_{ij} \cdot c_{ij}. \quad (4.3)$$

W przypadku gdy błędne zaklasyfikowania obiektów mają identyczne znaczenie dla każdej pary klas, do oceny jakości klasyfikatora można wykorzystać zagregowane miary oceny jakości klasyfikacji, takie jak trafność/dokładność klasyfikacji (*ang. accuracy*) i błąd klasyfikacji (*ang. error rate*). Trafność klasyfikacji na zbiorze danych testowych definiuje się jako stosunek liczby poprawnie zaklasyfikowanych obiektów do całkowitej liczby obiektów testowych:

$$accuracy = \frac{\text{liczba poprawnie zaklasyfikowanych obiektów testowych}}{\text{łączna liczba obiektów testowych}};$$

$$accuracy = \frac{TN + TP}{TP + TN + FP + FN}. \quad (4.4)$$

Alternatywnie, jakość klasyfikacji można wyrazić za pomocą miary błędu klasyfikacji,

która opisuje proporcję błędnie zaklasyfikowanych obiektów w całkowitym zbiorze danych testowych. Wzór na miarę błędu klasyfikacji (*ang. error rate*) jest następujący:

$$Class_{err} = \frac{\text{liczba błędnie zaklasyfikowanych obiektów testowych}}{\text{łączna liczba obiektów testowych}};$$

$$Class_{err} = \frac{f_{err}}{t_{test}} = 1 - \frac{f_{correct}}{f_{test}} = 1 - \frac{TP + TN}{TP + TN + FP + FN} = \frac{FP + FN}{TP + TN + FP + FN}. \quad (4.5)$$

W literaturze zaproponowano szereg innych miar oceny jakości klasyfikacji, które można bezpośrednio wywieść z macierzy pomyłek dla przypadku binarnego (przyjmując każdą z klas z osobna jako pozytywną w przypadku wieloklasowym). Najważniejsze z nich to:

- Czulość (*ang. true positive rate, recall*), która wyraża zdolność klasyfikatora do poprawnego wykrywania pozytywnych przypadków. Oblicza się ją jako stosunek TP do sumy TP i FN :

$$sensitivity = \frac{TP}{TP + FN}; \quad (4.6)$$

Czulość odpowiada na pytanie, jaki odsetek liczby obiektów należących do klasy pozytywnej został poprawnie zaklasyfikowany.

- Specyficzność (*ang. true negative rate, specificity*) wyrażająca zdolność klasyfikatora do poprawnego wykrywania negatywnych przypadków. Oblicza się ją jako stosunek TN do sumy TN i FP :

$$specificity = \frac{TN}{TN + FP}; \quad (4.7)$$

Specyficzność wskazuje, jak dobrze klasyfikator radzi sobie z wykrywaniem przypadków negatywnych.

- Precyzja (*ang. precision, positive predictive value*), która jest miarą tego, jak wiele z zaklasyfikowanych pozytywnych rzeczywiście nimi jest. Oblicza się ją jako stosunek TP do liczby obiektów łącznie zaklasyfikowanych przez klasyfikator jako należące do klasy pozytywnej:

$$precision = \frac{TP}{TP + FP}; \quad (4.8)$$

Precyzja odpowiada na pytanie, jaki odsetek przypadków zaklasyfikowanych jako pozytywne rzeczywiście jest pozytywny.

- Swoistość (*ang. false positive rate, fall-out*), która wskazuje, jaki odsetek przypadków negatywnych został błędnie zaklasyfikowany jako pozytywny. Jest to stosunek FP do sumy FP i TN :

$$FPR = \frac{FP}{TN + FP};$$

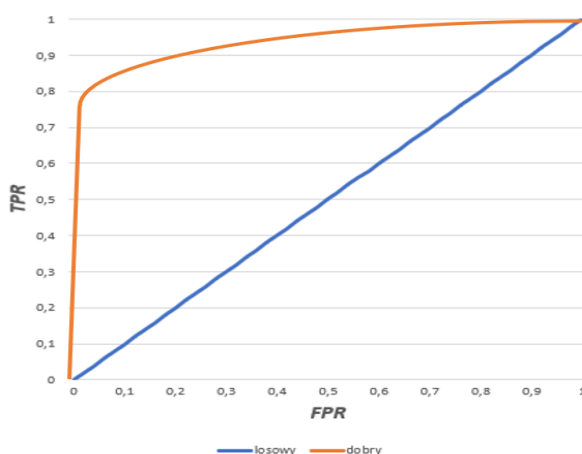
- F-miara (*ang.* *F1-score*) będąca miarą harmoniczną precyzji i czułości, co pozwala na zbalansowaną ocenę obu wskaźników. Wzór na F-miarę to:

$$F1 - score = \frac{2 \cdot precision \cdot sensitivity}{precision + sensitivity}.$$

F-miara jest szczególnie użyteczna, gdy zależy nam na równoczesnym optymalizowaniu precyzji i czułości.

Te miary pozwalają na bardziej szczegółową ocenę jakości klasyfikatora niż sama trafność, zwłaszcza w przypadku, gdy klasy są niezbalansowane (np. w przypadku rzadkich zdarzeń).

Inną popularną metodą oceny klasyfikatorów jest krzywa ROC [106]. Pozwala ona na wyciągnięcie ogólnych wniosków dotyczących jakości klasyfikatora. Krzywa ROC ilustruje zależność między dwoma kluczowymi wskaźnikami: czułością (*ang.* *sensitivity*, *TPR*) a swoistością (*ang.* *false positive rate*, *FPR*).



Rysunek 4.6: Przykładowe krzywe ROC

Krzywa ROC pokazuje zależność pomiędzy czułością a swoistością.

Punkt $(0,0)$ odpowiada sytuacji, w której wszystkie obiekty sklasyfikowane są do klasy negatywnej, natomiast punkt $(1,1)$ sytuację, gdzie wszystkie obiekty sklasyfikowane są do klasy pozytywnej. Klasyfikator jest idealny, gdy krzywa przechodzi przez punkt $(0,1)$, natomiast krzywa będąca przekątną przestrzeni ROC odpowiada klasyfikatorowi, który losowo klasyfikuje obiekty do klasy pozytywnej lub negatywnej. Stąd można wyciągnąć wniosek, że im większy obszar znajduje się pod krzywą ROC, tym klasyfikator jest lepszy i cechuje się wyższą trafnością klasyfikacji.

ROZDZIAŁ 5

Metody wspierające klasyfikację w diagnostyce medycznej oparte o modyfikacje techniki k najbliższych sąsiadów

W niniejszym rozdziale przedstawione zostaną metody wykorzystania zaprezentowanych wcześniej miar zawierania, podobieństwa oraz entropii, uwzględniających niepewność, do tworzenia algorytmów wspomagających proces podejmowania decyzji w różnych aspektach życia. Niewątpliwie istotnym zagadnieniem jest zdrowie, dlatego też skoncentrowano się na podnoszeniu komfortu i jakości życia pacjentów poprzez opracowanie algorytmów, które wspierają proces diagnozowania chorób. Niepewność i nieprecyzyjność w danych medycznych może być natury obiektywnej (spowodowana złożonością lub naturą zjawiska), subiektywnej (spowodowana osobistą opinią, interpretacją, czy brakiem przekonania decydenta) lub spowodowana niską jakością informacji (niedoskonały układ pomiarowy mierzonej wielkości lub przybliżenie i uproszczenie założeń związanych z metodą i procedurą pomiaru). Dodatkowo z samej natury opisy medyczne są często nieprecyzyjne i niejednoznaczne, np. zależą od stosowanego sprzętu medycznego lub interpretacji lekarza. Taki stan rzeczy wymaga zastosowania nieklasycznych metod modelowania danych i wnioskowania, czyli metod mających na względzie nieprecyzyjność. Pomimo wielu prac z tego zakresu (np. [86, 66, 65]) wciąż brak efektywnych metod obsługujących ten rodzaj nieprecyzyjności w medycynie, czy przemyśle. W ramach badań wzięto pod uwagę teorię przedziałowych zbiorów rozmytych umożliwiającą efektywną reprezentację niepewności i modelowanie decyzji w warunkach niepewności. Kluczowym aspektem badań było poszukiwanie optymalnych operatorów agregacji przedziałowych, miar wiedzy lub ilości informacji czy zależności (miary zawierania, czy podobieństwa oraz entropii).

Przedstawię zaproponowane w ramach pracy doktorskiej efektywne modele podejmowania decyzji wykorzystujące wspomniane operacje w ujęciu klasycznym (zgodne z porządkiem częściowym lub liniowym), możliwości i konieczności:

1. Przedziałowy klasyfikator uogólniający algorytm k-najbliższych sąsiadów (patrz rozdz. 5.1.1).
2. Metoda uzupełniania braków w danych wykorzystująca miarę entropii przedziałowej (patrz rozdz. 5.2.1).
3. Przedziałowy algorytm wyznaczania istotności atrybutów (patrz rozdz. 5.3.1).
4. Metoda prewencji (wczesnego wykrywania) depresji (patrz rozdz. 5.4.2).

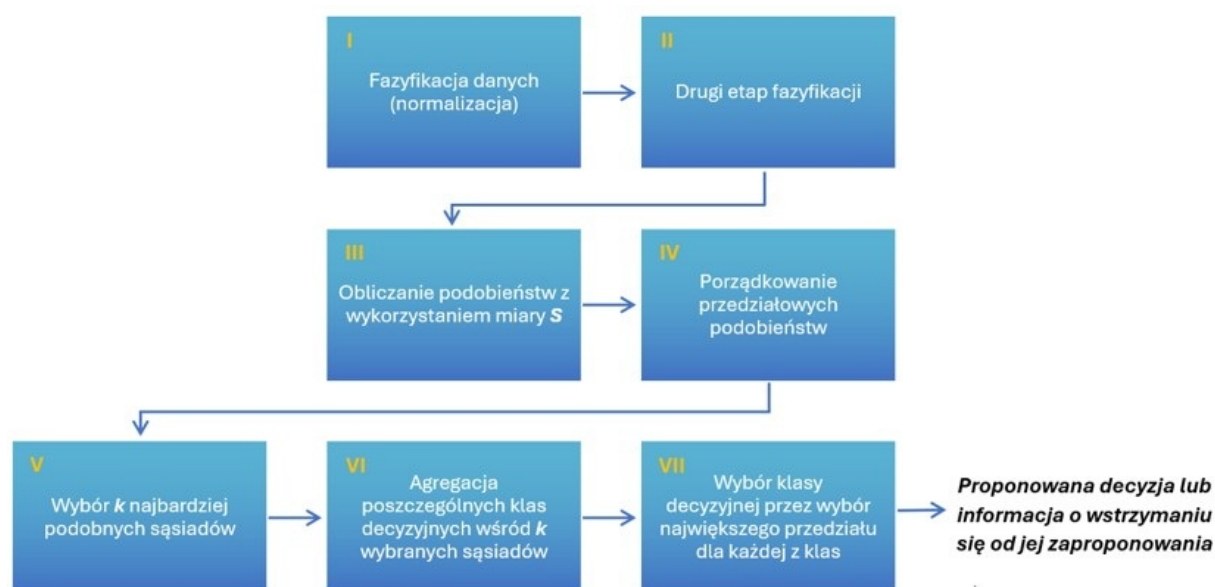
Cechą, która wyróżnia zaproponowane metody jest pełne wsparcie dla niepewności i nieprecyzyjności danych. Dzięki wykorzystaniu miar uwzględniających niepewność, zaproponowane metody klasyfikacji i/lub selekcji atrybutów czy uzupełniania brakujących wartości bez problemu radzą sobie z niepewnością typu epistemicznego zarówno w zbiorze uczącym, jak i podczas klasyfikacji. Opracowywane metody tworzone były z myślą o wykorzystaniu w systemach diagnostyki medycznej. Dlatego jednym z podstawowych założeń było zachowanie prostoty koncepcyjnej. Było to również motywacją wyboru rozwiązań w dużej mierze bazujących na metodzie najbliższych sąsiadów, która bardzo przypomina stosowaną w medycynie metodę opartą na studium przypadku.

5.1 Klasyfikacja z niepewnością

Problem klasyfikacji staje się znacznie bardziej skomplikowany, gdy dopuszczamy niepewność danych. Niepewność w sensie epistemicznym dopuszcza reprezentację danych przez wartości przedziałowe, gdy nie jest możliwe określenie dokładnej wartości, jaką przyjmuje obiekt na danym atrybucie. W takich warunkach zaprojektowanie efektywnego klasyfikatora za pomocą klasycznych metod może być bardzo trudne lub wręcz niemożliwe. W ramach pracy zaproponowano model klasyfikacji uwzględniający niepewność i nieprecyzyjność oparty na jednej z najczęściej stosowanych technik klasyfikacji, czyli klasyfikatorze k -NN (*ang. k-nearest neighbours*). Jest to jedna z najpopularniejszych metod uczenia nadzorowanego, będąca jednocześnie jednym z dziesięciu algorytmów uznawanych za najlepsze w Data Mining [107], stanowiąca integralną część wielu zastosowań uczenia maszynowego w różnych dziedzinach [108]. W tej strategii obiekt poddawany procesowi klasyfikacji należy do klasy, do której należy większość z jego k najbliższych sąsiadów. Przez najbliższych sąsiadów klasyfikowanego obiektu rozumie się te obiekty ze zbioru referencyjnego, które są do niego najbardziej podobne pod względem przyjętej miary podobieństwa.

5.1.1 Propozycja modelu klasyfikacji IV-kNN

Pierwszym problemem badawczym i aplikacyjnym było stworzenie uogólnionego rozmytego klasyfikatora wspomagającego procesy decyzyjne, opartego na nieprecyzyjnych (niepewnych) danych, tj. operującego na danych przedziałowych. Głównym celem było opracowanie kompleksowego i efektywnego podejścia, które umożliwi modelowanie i przetwarzanie danych wejściowych, a następnie prezentację wyników w sposób pozwalający na zachowanie istotnych informacji dotyczących poziomu niepewności na każdym etapie procesu. W odróżnieniu od rozmytego k -NN [109], proponowana koncepcja klasyfikacji opiera się na nowej definicji miary podobieństwa, która uwzględnia szerokość przedziałów (zob. rozdział 3.2: przedziałowe miary podobieństwa klasyczne, możliwości i konieczności). Poniżej przedstawiono schemat opisujący poszczególne etapy proponowanego algorytmu IV-kNN.



Rysunek 5.1: Ogólny schemat algorytmu IV-kNN

Poniżej przedstawiony został bardziej szczegółowy opis prezentowanego algorytmu. Składa się on z siedmiu kluczowych etapów.

Etap I Fazyfikacja danych. Fazyfikacja jest procesem umożliwiającym reprezentację danych rzeczywistych za pomocą wartości nieostrych z przedziału $[0,1]$ i pozwala na lepsze modelowanie niepewności i zmienności danych, w szczególności danych medycznych, rozwijane w etapie II.

Dane w naszym podejściu są konwertowane do przedziału $[0, 1]$ zgodnie ze wzorem wykorzystującym minimalne m_i i maksymalne M_i wartości dla i -tego atrybutu o rzeczywistej wartości b_i

$$a_i = \frac{b_i - m_i}{M_i - m_i}. \quad (5.1)$$

Etap II Reprezentacja przedziałowa. Drugi etap fazyfikacji, w którym wartości rozmyte atrybutów są zamieniane na wartości przedziałowe. Jedną z zastosowanych technik użytą do budowy przedziałowych zbiorów rozmytych dla każdej instancji oparto o funkcję ignorancji. Jej koncepcja, zdefiniowana w pracy [110], służy do pomiaru stopnia ignorancji lub braku wiedzy eksperta, gdy przypisuje on różne wartości liczbowe jako stopnie przynależności obiektu do różnych klas i została wybrana spośród innych metod konstrukcji [85, 111, 112] ze względu na jej poniżej prezentowane własności szczególnie przydatne w danych medycznych. Adaptacja tej koncepcji z wykorzystaniem zmiennej i jej negacji została opisana w poniższej definicji.

Definicja 29. Słaba funkcja ignorancji to ciągle odwzorowanie $g : [0, 1] \rightarrow [0, 1]$ o następujących własnościach:

1. $g(x) = g(1 - x)$ dla wszystkich $x \in [0, 1]$;
2. $g(x) = 0$ wtedy i tylko wtedy, gdy $x = 0$ lub $x = 1$;
3. $g(0.5) = 1$.

Przykład 20. Poniższa funkcja jest słabą funkcją ignorancji

$$g(x) = 2 \cdot \min(x, 1 - x) \text{ dla wszystkich } x \in [0, 1]. \quad (5.2)$$

Powyższą funkcję ignorancji można wykorzystać do konstrukcji przedziałowych zbiorów rozmytych (przedziałów) następująco: dla zbioru rozmytego $A \in FS(X)$ i słabej funkcji ignorancji g , dla każdego $x_i \in X$ przyjmując $g(\mu_A(x_i))$, $\delta(x_i)$, $\gamma(x_i) \in [0, 1]$, zbiór

$$A_{IV} = \{(x_i, A_{IV}(x_i)) | x_i \in X\},$$

gdzie $A_{IV}(x_i) = G(\mu_A(x_i), g(\mu_A(x_i)), \delta(x_i), \gamma(x_i))$ jest przedziałowym zbiorem rozmytym na X , gdzie $\delta(x_i), \gamma(x_i) \in [0, 1]$ i $G(x, y, \delta, \gamma) = [x \cdot (1 - \delta \cdot y), x \cdot (1 - \delta \cdot y) + \gamma \cdot y]$.

Uwzględniając badania z pracy [9] uwzględniamy w eksperymentach optymalne wartości parametrów δ i γ równe 0,25. Zatem dla $x \in [0, 1]$ i funkcji g użyto:

$$[x(1 - 0.25g(x)), x(1 - 0.25g(x)) + 0.25g(x)] \in L^I.$$

Etap III Wyznaczanie podobieństw. W tym kroku algorytmu do pomiaru podobieństwa badanych obiektów względem siebie wykorzystano jedną z miar podobień-

stwa, przedstawionych w rozdziale 3.2. Uzyskane przedziały mogą mieć różną szerokość, co pozwala uwzględnić niepewność w danych.

Etap IV **Porządkowanie przedziałów.** Polega na uporządkowaniu rozmytych podobieństw przedziałowych za pomocą jednego z porządków z przyjętych klas: klasycznej, możliwości i konieczności.

Etap V **Wybór k najbliższych (najbardziej podobnych) sąsiadów.** Dla wartości k (podanej jako parametr wejściowy do algorytmu), gdzie $k \in \{1, \dots, n\}$, $n \in N$, wskazanie k najbardziej podobnych sąsiadów w odniesieniu do obiektu podlegającego klasyfikacji.

Etap VI **Agregacja w klasach decyzyjnych.** Wartości miary podobieństwa agregowane są dla każdej z klas decyzyjnych osobno. Metody agregacji uwzględniają ten sam porządek, co w etapie IV.

Etap VII **Wybór decyzji.** Dla uzyskanych dwóch przedziałów odpowiadających każdej z klas decyzyjnych zastosowano następującą metodę prowadzącą do wyboru decyzji. Dla wspomnianych klas zdefiniowano przedziały oznaczone odpowiednio przez k_0 i k_1 , wyrażające stopień przynależności klasyfikowanego obiektu do każdej z klas w przypadku binarnym. Dodatkowo istnieje możliwość wstrzymania się od zaproponowania decyzji wówczas algorytm zwróci *Brak decyzji*. Metoda wyboru decyzji opiera się na porównaniu szerokości przedziałów (im mniejsza szerokość, tym lepiej) oraz ich położenia względem wartości $[0.5, 0.5]$. Założono, że większa wartość końców przedziałów niż 0,5 sugeruje lepszą jednoznaczność decyzji. Algorytm 1 przedstawia metodę stosowaną do wyboru proponowanej decyzji. Proponowana metoda może być uogólniona na wieloklasową.

Algorithm 1: Algorytm progowania - wybór decyzji

Variables: k_0 - zagregowana wartość klasy decyzyjnej 0 spośród k najbliższych sąsiadów
 k_1 - zagregowana wartość klasy decyzyjnej 1 spośród k najbliższych sąsiadów

Output : Klasa decyzyjna dla rozważanego obiektu

```

if  $w(k_1) < w(k_0)$  then
    if  $\underline{k_1} \geq 0.5$  then return 1;
    else if  $\underline{k_0} \geq 0.5$  then return 0;
    else if  $\overline{k_1} \geq 0.5$  then return 1;
    else if  $\overline{k_0} \geq 0.5$  then return 0;
    else return "brak decyzji";
end
else if  $w(k_1) > w(k_0)$  then
    if  $\underline{k_0} \geq 0.5$  then return 0;
    else if  $\underline{k_1} \geq 0.5$  then return 1;
    else if  $\overline{k_0} \geq 0.5$  then return 0;
    else if  $\overline{k_1} \geq 0.5$  then return 1;
    else return "brak decyzji";
end
else if  $w(k_1) == w(k_0)$  then
    if  $k_1 \leq k_0$  then
        if  $\underline{k_0} \geq 0.5$  then return 0;
        else if  $\underline{k_1} \geq 0.5$  then return 1;
        else if  $\overline{k_0} \geq 0.5$  then return 0;
        else if  $\overline{k_1} \geq 0.5$  then return 1;
        else return "brak decyzji";
    end
    if  $k_0 < k_1$  then
        if  $\underline{k_1} \geq 0.5$  then return 1;
        else if  $\underline{k_0} \geq 0.5$  then return 0;
        else if  $\overline{k_1} \geq 0.5$  then return 1;
        else if  $\overline{k_0} \geq 0.5$  then return 0;
        else return "brak decyzji";
    end
end

```

5.1.2 Opis danych - diagnostyka raka piersi

Zaproponowana metoda klasyfikacji została wykorzystana w problemie diagnostyki medycznej, w szczególności w diagnostyce raka piersi. Rak piersi (rak gruczołu piersiowego, rak sutka) to najczęstszy nowotwór złośliwy u kobiet wywodzący się głównie z tkanki nabłonkowej gruczołu piersiowego. Mimo, że choroba ta dotyczy przede wszystkim kobiet,

około 1% zachorowań występuje także u mężczyzn. Jest to groźna choroba nowotworowa, ponieważ zazwyczaj przez długi czas nie daje żadnych objawów, a w momencie jej wykrycia często występują już przerzuty komórek nowotworowych do innych narządów, takich jak płuca, wątroba, mózg czy kości. Dlatego najważniejszym czynnikiem warunkującym wyleczenie jest wczesne postawienie diagnozy, co możliwe jest dzięki odpowiedniej profilaktyce i regularnym badaniom, w tym samokontroli prowadzonej przez pacjenta. Zbór danych dotyczący diagnozowania raka piersi (*Wisconsin (diagnostic) breast cancer*), który posłużył do przeprowadzenia testów tworzonych algorytmów, pochodzi z ogólnodostępnego repozytorium UCI [37]. Repozytorium UCI zawiera zbiory danych z różnych dziedzin. Wybrany zbiór danych jest jednym z bardziej popularnych znajdujących się w tym repozytorium. Został opublikowany w listopadzie 1995 roku i składa się z 569 obiektów z przypisanymi decyzjami. Cechy zostały obliczone na podstawie zdigitalizowanego obrazu aspiratu cienkoigłowego (FNA) z masy piersi, opisując charakterystykę jąder komórkowych obecnych na obrazie. Dla każdego jądra komórkowego obliczono dziesięć cech o wartościach rzeczywistych:

- promień (średnia odległości od środka do punktów na obwodzie),
- tekstura (odchylenie standardowe wartości w skali szarości),
- obwód,
- obszar,
- gładkość (lokalna zmienność długości promienia),
- zwartość (powierzchnia obwodu - 1:02),
- wklęsłość (stopień wklęsłości części konturu),
- punkty wklęsłe (liczba wklęsłych części konturu),
- symetria,
- wymiar fraktalny (aproksymacja linii brzegowej - 1).

Atrybuty warunkowe nie posiadają wartości brakujących, a ich wartości są liczbami rzeczywistymi. Atrybut decyzyjny przechowuje informacje o diagnozie: czy wykryty guz jest guzem złośliwym, czy łagodnym. Guz złośliwy reprezentowany jest przez wartość "0", a guz łagodny przez "1". Zbiór danych składa się z 212 obiektów z diagnozą złośliwą oraz 357 obiektów z diagnozą łagodną.

5.1.3 Analiza wyników badań wpływu uwzględnienia niepewności na jakość klasyfikacji

W procesie ewaluacji algorytmu IV-kNN zbiór danych został podzielony na część treningową i testową, przy czym 70% obiektów było losowo (z zachowaniem zbalansowanego rozkładu klas) przypisanych do części treningowej, a pozostałe 30% do części testowej.

Algorytm IV-kNN był testowany przy użyciu 10-krotnej walidacji krzyżowej dla określonych wartości parametru k (rozważano $k \in \{1, 3, 5\}$ oraz dla każdej kombinacji porządków, wskaźników pierwszeństwa i agregacji, a na ich podstawie tworzonych miar podobieństwa z poszczególnych klas. Każda iteracja algorytmu korzystała z innej pary zbiorów treningowych i testowych. Na podstawie macierzy pomyłek obliczano dokładność (4.4), czułość (4.6), specyficzność (4.7) i precyzję (4.8) dla każdej iteracji algorytmu uruchomionej na danym zestawie danych wejściowych i ustalonych parametrach. Wyniki miar jakości klasyfikacji były następnie uśredniane. Warto zaznaczyć, że przy obliczaniu miar jakości klasyfikacji nie uwzględniano sytuacji, w których klasyfikator wstrzymał się od podjęcia decyzji, ponieważ były one rzadkie i nie miały istotnego wpływu na ostateczne wyniki.

Analiza algorytmu została przeprowadzona w trzech głównych scenariuszach opartych o przedziałowe miary podobieństwa klasyczne, możliwości i konieczności:

- **Model 1.** Model oparty o miary klasyczne (zgodne z porządkiem częściowym lub liniowym) oraz porównanie modelu przedziałowego do przypadku bazowego, czyli rozmytego;
- **Model 2.** Model oparty o miary możliwości;
- **Model 3.** Model oparty o miary konieczności.

Ponadto ewaluacja modelu przebiegała dla różnych wariantów, oceniono skuteczność klasyfikacji wariantów metody wykorzystano następujące komponenty i wartości parametrów:

- 1) Dla wartości współczynnika k , $k \in \{1, 3, 5\}$ badano wpływ liczby najbliższych sąsiadów na efektywność klasyfikacji w różnych wariantach algorytmu.
- 2) Konstrukcja miar podobieństwa dla poszczególnych klas (etap III algorytmu IV-kNN):
 - dla klasycznej miary podobieństwa testowane były
agregacje: $\vee, \wedge, \mathcal{A}_{prod}, \mathcal{A}_{mean}, \mathcal{A}_{meanpow}, \mathcal{A}_{prodmean}, \mathcal{A}_{meanmax}, \mathcal{A}_{\alpha}$;
wskaźniki pierwszeństwa: $Prec_{\mathcal{A}}, Prec_w$;
 - dla miary podobieństwa konieczności testowane były
agregacje: $\vee, \wedge, \mathcal{A}_{mean}, \mathcal{A}_{nec1}, \mathcal{A}_{nec2}$;
wskaźniki pierwszeństwa: $Prec_{nec}$;
 - dla miary podobieństwa możliwości testowane były
agregacje: $\vee, \wedge, \mathcal{A}_{mean}, \mathcal{A}_{pos1}, \mathcal{A}_{pos2}, \mathcal{A}_{pos3}, \mathcal{A}_{pos4}$;
wskaźniki pierwszeństwa: $Prec_{\mathcal{A}_{pos}}$;
agregacje użyte do budowy $Prec_{\mathcal{A}_{pos}}$: $\mathcal{A}_{pos1}, \mathcal{A}_{pos2}, \mathcal{A}_{pos3}, \mathcal{A}_{pos4}$;
negacje: N_1, N_2, N_3, N_4 .

- 3) Wybór relacji porządkującej (etap IV i jeden z przypadków w VII etapie algorytmu IV-kNN) spośród jednej z klas (klasyczna, możliwości, konieczności), czyli ze zbioru:

$$\mathcal{O} = \{\leq_{LI}, \leq_{XY}, \leq_{Lex1}, \leq_{Lex2}, \leq_{pos}, \leq_{nec}\}.$$

W VI etapie wykorzystano tą samą funkcję agregacji, co w etapie III.

Wyniki testów przeprowadzonych z wykorzystaniem miar porządkowych, agregacji i podobieństwa klasycznych, możliwości i konieczności

Ze względu na liczbę przeprowadzonych testów wynikającą z wielu możliwych kombinacji parametrów wejściowych algorytmu, w tabelach wynikowych przedstawiono jedynie najlepsze rezultaty. Nowa propozycja algorytmu IV-kNN umożliwiającą operowanie na wartościach przedziałowych oraz wykorzystującą miary podobieństwa w trzech różnych klasach, została również porównana z wynikami zwróconymi przez podstawowy algorytm k-NN. Wyniki dla standardowego (rozmytego) klasyfikatora k-NN, uzyskane przy 10-krotnej krosvalidacji, prezentuje tabela 5.1. Prezentowane w niej wartości są wartościami referencyjnymi i zostały uzyskane przy zastosowaniu dobrze znanej metryki euklidesowej (4.1). Dla poprawy czytelności najwyższe wartości uzyskane dla poszczególnych miar oznaczone zostały pogrubioną czcionką.

Model rozmyty - rozmyty k-NN

accuracy	sensitivity	specificity	precision	k
0,91	0,97	0,82	0,89	1
0,93	0,96	0,88	0,94	3
0,93	0,94	0,94	0,95	5

Tabela 5.1: Wyniki uzyskane przy użyciu standardowego (rozmytego) algorytmu k-NN

W tabeli 5.2 przedstawiono zestawienie kluczowych wyników uzyskanych w wyniku przeprowadzonych testów, które uwzględniają trzy różne klasy miar podobieństwa, agregacji i relacji porządkowania w uogólnionym algorytmie IV-kNN, uwzględniającym niepewność poprzez reprezentację przedziałową danych.

Model przedziałowo-rozmyty - IV-kNN

klasa podobieństwa	accuracy	sensitivity	specificity	precision	k
klasyczna [46]	0,94	0,99	0,94	0,96	1
	0,93	0,99	0,90	0,95	3
	0,90	0,99	0,85	0,92	5
konieczności	0,95	0,96	0,94	0,96	1
	0,95	0,98	0,89	0,93	3
	0,92	0,98	0,81	0,89	5
możliwości	0,95	0,98	0,94	0,97	1
	0,94	0,94	0,94	0,96	3
	0,93	0,93	0,92	0,95	5

Tabela 5.2: Wyniki IV-kNN dla różnych klas miar podobieństwa

Wyniki uzyskane z wykorzystaniem klasycznej miary podobieństwa S (zob. rozdz. 3.2.1) otrzymano przy zastosowaniu następujących parametrów wejściowych: wskaźnik pierwszeństwa $Prec_w$, porządek $\leq_{XY}$, natomiast do agregacji podobieństw w kroku VI algorytmu IV-kNN zastosowano $\mathcal{A}_{mean}$ dla różnych k . Należy zauważyć, że w przypadku $k=1$ nie ma potrzeby agregacji klas decyzyjnych, ponieważ w takim przypadku w procesie agregacji bierze udział tylko jeden przedział.

W przypadku miary podobieństwa konieczności (pesymistycznej) S_{nec} (zob. rozdz. 3.2.3) najlepsze wyniki otrzymano przy zastosowaniu wskaźnika pierwszeństwa $Prec_{nec}$ oraz związanego z nim $\leq_{nec}$ wspieranego przy nieporównywalności przez porządek $\leq_{XY}$. Do agregacji podobieństw zastosowano agregację $\mathcal{A}_{mean}$, zgodną z ww. porządkami.

Natomiast najlepsze wyniki dla algorytmu wykorzystującego miary podobieństwa możliwości (optymistyczne) S_{pos} , przedstawione w tabeli 5.2, uzyskano przy następującym zestawie parametrów: wskaźnik pierwszeństwa $Prec_{Apos}$, negacja N_1 wykorzystana do budowy wskaźnika pierwszeństwa i miary podobieństwa oraz agregacja $\mathcal{A}_{mean}$ do agregacji podobieństw.

Analiza wyników

Wyniki przeprowadzonych testów ukazują, że zaproponowane uogólnienie algorytmu IV-kNN przyniosło poprawę dokładności klasyfikacji, szczególnie w przypadku użycia miary **podobieństwa możliwości (optymistycznej)**. Poprawa obejmuje zwiększenie kluczowych wskaźników oceny klasyfikacji w porównaniu zarówno do rozmytego klasyfikatora (zob. tab. 5.1), jak i do przedziałowego (IV-kNN) dla operatorów oraz miar podobieństwa z klasy klasycznej i konieczności (zob. tab. 5.2). Możemy zaobserwować wzrost zarówno dokładności (accuracy), jak i precyzji (precision), przy jednoczesnym zachowaniu dobrego poziomu czułości (sensitivity) oraz swoistości (specificity), co świadczy o stabilności poprawy wyników klasyfikacji. Zauważalny wzrost dokładności i precyzji o 1% względem klasycznego podejścia do budowy miar uwzględniających niepewność, bądź 2% w porównaniu z klasyfikatorem rozmytym, ma istotne znaczenie w kontekście

diagnostyki medycznej. Tego rodzaju poprawa może przekładać się na bardziej trafne i niezawodne decyzje medyczne, co w praktyce klinicznej może prowadzić do szybszego i skuteczniejszego leczenia, zmniejszając ryzyko błędnych diagnoz oraz podnosząc ogólną jakość opieki zdrowotnej.

Obserwuje się również, że uzyskanie lepszych wyników wiąże się z mniejszą liczbą poszukiwanych obiektów podobnych, co jest szczególnie istotne w przypadku danych medycznych, które często charakteryzują się ograniczoną liczbą obiektów. Ta cecha algorytmu stanowi jego dużą zaletę, gdyż pozwala na bardziej efektywne wykorzystanie dostępnych danych bez potrzeby zwiększania zbioru treningowego. **Miara możliwości (optymistyczna), Model 2**, charakteryzuje się **największą stabilnością** wyników, niezależnie od wartości k . Dodatkowo, daje najwyższą precyzję dla $k=1$, co czyni ją dobrą opcją w sytuacjach, w których kluczowe jest minimalizowanie błędnych decyzji pozytywnych.

Wszystkie te obserwacje wskazują na potrzebę głębszego zrozumienia i odpowiedniego dostosowania różnych miar podobieństwa w kontekście modelowania danych niepewnych, zwłaszcza w medycynie. W tej dziedzinie dane są często obarczone niepewnością wynikającą z różnorodnych źródeł, takich jak ograniczona liczba próbek, różnice w pomiarach czy subiektywna interpretacja wyników. Wyniki uzyskane dla różnych klas miar podobieństwa sugerują, że odpowiedni dobór miary oraz liczby sąsiadów (k) może mieć istotny wpływ na efektywność klasyfikacji.

5.2 Uzupełnianie brakujących wartości w danych

Bez wątpienia proces klasyfikacji jest podatny na wszelkie wady związane zarówno z danymi testowymi, jak i treningowymi. Wady danych mogą znacząco wpływać na wiarygodność i użyteczność uzyskanych wyników. Mogą one występować na różnych etapach zbierania, przechowywania lub analizowania danych, a ich przyczyny mogą być różnorodne. Dlatego ważne jest, aby zidentyfikować i zrozumieć te wady, aby skutecznie je wyeliminować i uniknąć błędnych wniosków. Jedną z najczęściej występujących wad jest brak informacji o wartościach atrybutów konkretnego obiektu, który bierze udział w procesie klasyfikacji. Dane z brakującymi wartościami mogą powodować, że większość algorytmów nie działa poprawnie, gdyż wymagają one kompletnych danych. Istnieje kilka możliwości radzenia sobie z brakami w danych. Można je usunąć, jednak przy tym usuwamy również inne informacje. Często usunięcie obiektu z brakującą wartością może uniemożliwić algorytmowi odkrycie wzorców, które były dostępne tylko w danym obiekcie lub atrybucie. Dlatego lepszym, choć bardziej skomplikowanym rozwiązaniem, jest uzupełnienie brakujących wartości. Istnieje wiele metod uzupełniania braków w danych, które zależą od rodzaju danych, rozmiaru brakujących wartości i celu analizy. Oto kilka przykładów

metod uzupełniania wartości brakujących:

- uzupełnianie średnią - polega na wstawieniu do brakujących komórek średniej wartości dla danego atrybutu, może być stosowane w przypadku danych numerycznych, ale nie jest dobrym rozwiązaniem, gdy w danych znajdują się wartości odstające;
- uzupełnianie medianą - podobne do uzupełniania średnią, ale w tym przypadku wstawiana jest mediana zamiast średniej, może być lepszą opcją w sytuacjach, gdy dane mają duży rozrzut lub zawierają wartości odstające;
- uzupełnianie najczęstszą wartością - wstawiana jest najczęściej występująca wartość dla danej zmiennej, ta metoda może być stosowana w przypadku danych kategorycznych;
- czasami wprowadza się 0 lub inną stałą wartość, aby ograniczyć wpływ wstawianej wartości na proces klasyfikacji.

W literaturze wykazano [40, 113], że w praktyce lekarskiej pobranie kompletnego zbioru danych przez lekarza podczas badań może być bardzo problematyczne ze względu na ograniczenia techniczne instytucji opieki zdrowotnej, wysokie koszty lub duże ryzyko pogorszenia się stanu zdrowia pacjenta po potencjalnym badaniu. Brak danych natomiast utrudnia korzystanie z tradycyjnych rozwiązań wspomagających diagnostykę i dlatego istnieje pilna potrzeba rozwiązania tego problemu. Jednym z możliwych podejść do zarządzania niekompletnością danych jest wykorzystanie ugruntowanych metod imputacji danych [34, 114] lub [115, 116]. Niewątpliwie w wielu obszarach badawczych takie podejście jest wystarczające. Jednak w zastosowaniach medycznych, gdzie w grę wchodzi ludzkie życie, nie jest to takie oczywiste czy można wprowadzić nowe dane, które mogą być obciążone niewielkim, ale istotnym błędem. Dlatego w ramach prac badawczych zdecydowano się poszukać metody, która umożliwi adaptację i integrację istniejących, ugruntowanych metod diagnostycznych, aby umożliwić ich zastosowanie w przypadku niekompletnych danych oraz danych z niepewnością.

Bezpośrednią motywacją była potrzeba wsparcia lekarzy onkologów. W dziedzinie danych medycznych niepełne dane są zjawiskiem powszechnym i często występującym. Może to wynikać z ograniczeń w możliwości przeprowadzania badań lub pomiarów dla każdego pacjenta, szczególnie na sprzęcie, który nie jest dostępny we wszystkich placówkach medycznych. W takich przypadkach niektóre wyniki badań mogą być niedostępne dla konkretnych pacjentów. Oprócz braku dostępu do odpowiedniej aparatury, niekompletność danych może być także spowodowana innymi ograniczeniami, niekoniecznie związanymi ze sprzętem. Przykładem mogą być indywidualne cechy pacjenta, takie jak ekstremalna otyłość, która może utrudnić lub uniemożliwić przeprowadzenie niektórych badań, z obawy o uszkodzenie sprzętu. Innym przypadkiem są wcześniejsze zabiegi ortopedyczne, które wykluczają pewne rodzaje badań, na przykład ze względu na obecność metalowych

elementów w ciele pacjenta. Wiele sytuacji uniemożliwia przeprowadzenie określonych badań ze względów związanych zarówno z dobrem pacjenta, jak i integralnością aparatury medycznej. W takich okolicznościach, stajemy przed trudnością reprezentowania pełnej informacji. W kontekście tych różnorodnych ograniczeń niekompletność danych staje się powszechnie występującym problemem.

W ramach prac badawczych przedstawionych w rozprawie sprawdzono skuteczność różnych podstawowych metod uzupełniania brakujących danych w warunkach nieprecyzyjności, czyli dla przedziałowych wartości rozmytych (z wykorzystaniem teorii przedziałowych zbiorów rozmytych) i zestawionych w [47]:

Metoda I W przypadku brakującej wartości wstawiany był przedział $[0, 1]$, tak jak opisano w pracy [117].

Metoda II Znalezienie k najbardziej podobnych obiektów i wybór wartości dla brakującego atrybutu spośród nich. Następnie brakująca wartość była zastąpiona przedziałem skonstruowanym w jeden z poniższych sposobów:

$$\text{a) } [\min(\underline{a}_1, \dots, \underline{a}_k), \max(\bar{a}_1, \dots, \bar{a}_k)],$$

$$\text{b) } [\text{avg}(\underline{a}_1, \dots, \underline{a}_k), \text{avg}(\bar{a}_1, \dots, \bar{a}_k)].$$

Metoda III Znalezienie najbardziej podobnych obiektów l , analogicznie do metody II, ale tylko z tej samej klasy decyzyjnej, do której należy obiekt z brakującą wartością.

Następnie powyższe metody porównano z zaproponowaną metodą w [48], opartą na mierze entropii w sensie klasycznym (przypadek klasyczny, czyli konstrukcja oparta na porządku częściowym lub liniowym), a następnie rozszerzoną w ramach prac badawczych doktoratu o przypadek entropii możliwości i konieczności (zob. rozdz. 5.2.1).

5.2.1 Metoda uzupełniania brakujących danych z użyciem przedziałowej miary entropii

W przypadku brakujących wartości kluczowym aspektem badań było zastosowanie miary entropii w procesie uzupełniania wartości oraz przeprowadzenie testów mających na celu zbadanie skuteczności tej metody w kontekście klasyfikacji. Poniżej przedstawiono proponowaną metodę oznaczoną jako **Metoda IV**:

Metoda IV Wykorzystanie miary entropii do wyboru wartości użytej do imputacji z obiektu o najmniejszej niepewności danych spośród najbardziej podob-

nych obiektów do obiektu obarczonego brakiem - zgodnie z algorytmem 2 przedstawionym poniżej.

Metoda IV opiera się na wyznaczaniu w pierwszej kolejności k najpodobniejszych obiektów do obiektu z brakującą wartością, aby w kolejnym etapie wskazać obiekt spośród nich, który ma najmniejszą wartość entropii, czyli niepewności.

Algorithm 2: Algorytm wyznaczający wartość dla atrybutu z brakującą wartością wykorzystujący do tego celu miarę entropii

Input : Obiekt O_z z brakującą wartością dla atrybutu t ;
 l - liczba obiektów najbardziej podobnych do O_z z tej samej klasy co O_z

Output: wyznaczona wartość $O_z(t)$

```

    minEntropy = MaxInteger;
    // Inicjalizacja zmiennej minEntropy jakąś dużą wartością
    noObject := 1;
    // przechowywanie numeru obiektu o najmniejszej entropii
    minEntropy :=  $E(O_1)$ ;
    for  $p := 2, \dots, l$  do
        if  $E(O_p) < minEntropy$  then
            minEntropy :=  $E(O_p)$ ;
            noObject :=  $p$ ;
        end
    end
     $O_z(t) := O_{noObject}(t)$ ;
    /* Na koniec wartość brakującego atrybutu jest uzupełniana przez
    pobranie wartości na tym atrybucie ze zwycięskiego obiektu      */

```

5.2.2 Analiza wyników badań wpływu brakujących danych na jakość klasyfikacji

W przypadku brakujących wartości w danych obserwuje się spadek wydajności klasyfikatorów. Jednak dzięki zastosowaniu entropii do uzupełniania brakujących wartości nastąpił wzrost wydajności klasyfikacji w stosunku do innych metod imputacji. By ocenić skuteczność proponowanego algorytmu imputacji opartego na mierze entropii, wykorzystano wcześniej opisaną metodą klasyfikacji IV-kNN (omówioną w rozdziale 5.1). Badania przeprowadzono na zbiorze danych dotyczących diagnozowania raka piersi, który został opisany w rozdziale 5.1.2.

Proces ewaluacji przebiegał w następujących aspektach:

- Przypadek 1** Wpływ na efektywność klasyfikacji różnych metod imputacji danych z niepewnością: Metoda I-IV;
- Przypadek 2** Wpływ na efektywność klasyfikacji różnych metod wyznaczania miary entropii stosowanej w Metodzie IV;
- Przypadek 3** Wpływ na efektywność klasyfikacji różnych klas entropii przedziałowej.

Dla każdego przypadku przebadano określony odsetek brakujących wartości. Polegało to na usunięciu wartości jednego losowego atrybutu dla określonego procentu obiektów (tj. 5%, 25% lub 50%). Wybrany poziom braków w danych znajduje się w tabelach w kolumnie oznaczonej jako MS , podczas gdy kolumna k zawiera informacje o liczbie najbardziej podobnych obiektów, dla których uzyskano przedstawione najlepsze wyniki. Ponadto w procesie oceny efektywności wykorzystano 10-krotną walidację krzyżową.

Wyniki testów przeprowadzonych z wykorzystaniem klasycznej entropii

Przypadek 1. Ważnym elementem ewaluacji było zbadanie wpływu na klasyfikację zaproponowanej metody imputacji w porównaniu z często stosowanymi podejściami. Poniżej zaprezentowane zostaną wyniki weryfikacji opisanych wcześniej czterech metod imputacji dla brakującej wartości atrybutu.

Ocenę efektywności klasyfikacji uzyskaną po wcześniejszym uzupełnieniu brakujących wartości za pomocą metod I-IV opisano częściowo w pracach [47, 48]. Tabela 5.3 prezentuje porównanie wyników przeprowadzonych testów różnych metod uzupełniania wartości brakujących omówionych wcześniej. Z analizy wynika, że metoda IV oparta na mierze entropii (algorytm 2) osiągnęła najlepszą skuteczność, zarówno pod względem dokładności klasyfikacji, jak i stabilności uzyskanych wyników. W szczególności charakteryzowała się mniejszym spadkiem dokładności klasyfikacji wraz ze wzrostem odsetka brakujących wartości.

W przypadku metody IV (w klasycznym ujęciu) najlepsze i najbardziej stabilne wyniki (charakteryzujące się najwolniejszym spadkiem dokładności klasyfikacji wraz ze wzrostem braków w danych) osiągnięto przy zastosowaniu porządku $\leq_{XY}$, agregacji podobieństw za pomocą operatora $\mathcal{A}_{mean}$, oraz agregacji klas przy użyciu $\mathcal{A}_{meanpow}$. Ponadto weryfikację metody IV przeprowadzono z wykorzystaniem następujących parametrów: jako wskaźnik pierwszeństwa zastosowano $Prec_A$, który posłużył do budowy miary entropii oraz miary podobieństwa opartej na ww. agregacjach i porządku.

Przypadek 2. Ważnym rozszerzeniem procesu ewaluacji było zbadanie wpływu na efektywność klasyfikacji różnych metod wyznaczania miary entropii stosowanych w Metodzie IV, która okazała się najefektywniejsza w przypadku 1. Porównano zaproponowaną

Metoda	accuracy	sensitivity	specificity	precision	k	MS
I	0,8592185	0,9549781	0,7949657	0,8893751	5	5%
	0,8196330	0,9296112	0,7294092	0,8595327	1	25%
	0,8172602	0,9301064	0,7363129	0,8578783	1	50%
II a)	0,8562544	0,9509575	0,7952890	0,8894207	5	5%
	0,8418038	0,9415083	0,7681922	0,8783080	1	25%
	0,8412828	0,9522094	0,7606850	0,8713155	5	50%
II b)	0,8540611	0,9471567	0,7927968	0,8881783	1	5%
	0,8417425	0,9416843	0,7677751	0,8781215	1	25%
	0,8403609	0,9439721	0,7696108	0,8750467	1	50%
III	0,8555894	0,9531871	0,7877616	0,8871398	5	5%
	0,8450197	0,9529377	0,7587463	0,8750599	5	25%
	0,8399786	0,9511847	0,7513039	0,8683283	5	50%
IV	0,8875097	0,9370335	0,8024434	0,8914741	5	5%
	0,8800044	0,9423466	0,7698350	0,8792120	5	25%
	0,8778067	0,9372094	0,7770936	0,8775432	5	50%

Tabela 5.3: Tabela z porównaniem poszczególnych metod uzupełniania brakujących wartości

w rozdziale 3.3.1 metodę konstrukcji entropii przedziałowej (w ujęciu klasycznym) z wybranymi metodami znanymi i popularnymi w literaturze, a reprezentującymi przypadek gdy wyznaczona entropia jest wartością rzeczywistą, pomimo danych przedziałowych, w odróżnieniu od zaproponowanej w ramach pracy doktorskiej entropii, gdzie entropia ma wartość przedziałową:

- 1) Zastosowanie nowych miar entropii zaproponowanych w rozdziale 3.3.1 (oznaczanych dalej jako E_I).
- 2) Zastosowanie entropii opartej na szerokości przedziałów i przedstawionej w [97]:

$$E_{II}(F) = \sum_{i=1, \dots, n} w_{F(x_i)}$$

dla $F \in IVFS(X)$, $card(X) = n$.

- 3) Zastosowanie entropii opartej na liczebności (cf. [98]):

Przy zastosowaniu odległości Hamminga, entropia przedziałowych zbiorów rozmytych jest stosunkiem największych liczebności ($maxCount$) uwzględniający F i F^c (dopełnienie F):

$$E_{III}(F) = \frac{1}{n} \sum_{i=1}^n \frac{maxCount(F(x_i) \wedge F^c(x_i))}{maxCount(F(x_i) \vee F^c(x_i))},$$

gdzie $F \in IVFS(X)$. Tak więc w odniesieniu do [98] w praktyce obliczano:

Jeśli $d(F, 1_{LI}) > d(F, 0_{LI})$, wtedy

$$E_{III}(F) = \frac{1}{n} \sum_{i=1}^n \frac{\overline{F}(x_i)}{1 - \underline{F}(x_i)},$$

w przeciwnym przypadku

$$E_{III}(F) = \frac{1}{n} \sum_{i=1}^n \frac{1 - \underline{F}(x_i)}{\overline{F}(x_i)},$$

gdzie dla dowolnych $A, B \in IVFS(X)$ otrzymujemy m.in. odległość Hamminga

$$d(A, B) = \frac{1}{n} \sum_{i=1}^n (|\underline{A}(x_i) - \underline{B}(x_i)| + |\overline{A}(x_i) - \overline{B}(x_i)| + |w_A(x_i) - w_B(x_i)|).$$

Entropia	accuracy	sensitivity	specificity	precision	MS
E_I	0,8875097	0,9370335	0,8024434	0,8914741	5%
	0,8800044	0,9370166	0,7698350	0,879212	25%
	0,8778067	0,9370094	0,7770936	0,8775432	50%
E_{II}	0,8870261	0,9369867	0,8011517	0,8907949	5%
	0,8742604	0,9352522	0,7523640	0,8717101	25%
	0,8683938	0,9342521	0,7482776	0,864492	50%
E_{III}	0,8864288	0,9346241	0,8038857	0,8913553	5%
	0,8808355	0,9314817	0,777911	0,872423	25%
	0,8768056	0,9165073	0,8091392	0,8910009	50%

Tabela 5.4: Tabela dla różnych entropii używanych w algorytmie 2

Tabela 5.4 przedstawia rezultaty uzyskane po zastosowaniu ww. entropii. Jak można zauważyć, nowo zaproponowana entropia (E_I) osiągnęła najlepsze wyniki, a także wykazała największą stabilność pod względem dokładności klasyfikacji w miarę wzrostu liczby brakujących wartości w porównaniu do pozostałych testowanych entropii. Uzyskane wyniki wskazujące na większą efektywność klasyfikacji po zastosowaniu entropii przedziałowej E_I zaproponowanej w ramach prac badawczych doktorskich potwierdzają postawioną hipotezę, że uwzględnienie niepewności poprzez reprezentację entropii za pomocą przedziału daje lepsze rezultaty.

Wyniki testów przeprowadzonych z wykorzystaniem entropii optymistycznych i pesymistycznych

Przypadek 3. Uzupełnieniem procesu ewaluacji przedstawionej w Przypadku 1 i 2 było zbadanie wpływu na efektywność klasyfikacji różnych klas entropii przedziałowej, czyli dla Metody IV rozszerzono badania o analizę klas możliwości i konieczności zastosowanych miar entropii.

Wskazaniem dla tych badań był fakt, że metoda oparta na klasycznej entropii E_I przyniosła najlepsze wyniki zawarte w tabeli 5.4. Wyniki obejmujące analizę działania algorytmu z miarami entropii w ujęciu optymistycznym i pesymistycznym przedstawiono w tabelach 5.5 i 5.6.

accuracy	sensitivity	specificity	precision	k	MS
0,894465563	0,938157946	0,820501827	0,899514567	3	5%
0,886472679	0,942453568	0,789790122	0,88755312	3	25%
0,891047909	0,938782115	0,811451422	0,893704812	3	50%

Tabela 5.5: Tabela przedstawiająca wyniki uzyskane po zastosowaniu entropii możliwości do uzupełniania brakujących wartości

accuracy	sensitivity	specificity	precision	k	MS
0,881693138	0,965405237	0,739066044	0,863301194	5	5%
0,872484279	0,966860351	0,706181046	0,853122574	5	25%
0,872049124	0,967259661	0,711917726	0,85021783	5	50%

Tabela 5.6: Tabela przedstawiająca wyniki uzyskane po zastosowaniu entropii pesymistycznej do uzupełniania brakujących wartości

Tabele prezentują wybrane, najlepsze wyniki uzyskane dla konkretnych wartości parametrów k i MS , przy użyciu odpowiednich agregacji w kolejnych krokach algorytmu. W przypadku entropii pesymistycznej rezultaty były nieznacznie gorsze od tych uzyskanych przy zastosowaniu klasycznej miary entropii. Dla każdego poziomu braków w danych najlepsze wyniki osiągnięto, stosując agregację A_{mean} zarówno do agregacji podobieństw, jak i decyzji dotyczących badanego obiektu. Dodatkowo użyto relacji $\leq_{nec}$ przy budowie wskaźnika pierwszeństwa.

Proces weryfikacji uzupełniania brakujących wartości z wykorzystaniem **entropii optymistycznej** dał najlepsze rezultaty przy zastosowaniu agregacji A_{mean} zarówno do agregacji podobieństw, jak i potencjalnych decyzji dla analizowanego obiektu. Prezentowane wyniki uzyskano wykorzystując do metody konstrukcji entropii możliwości (zob. twierdzenie 3.13) konstrukcję wskaźnika pierwszeństwa z twierdzenia 3.5 wykorzystując negację N_1 , agregację $\mathcal{A}_1$ oraz relację porządkującą $\leq_{pos}$.

Jak można zauważyć, zastosowanie różnych klas entropii znacząco wpływa na wyniki klasyfikacji. Zastosowanie **entropii optymistycznej (możliwości)** przyniosło najlepsze rezultaty pod względem dokładności klasyfikacji, w szczególności widzimy progres dla miar efektywności dokładności, specyficzności i precyzji. Warto także podkreślić, że te wyniki uzyskano przy wykorzystaniu mniejszej liczby najbliższych sąsiadów, co nie tylko poprawia efektywność obliczeniową, ale również jest korzystne w kontekście danych medycznych, gdzie charakterystyka niektórych schorzeń często uniemożliwia zebranie dużej liczby pomiarów dla licznej populacji.

5.3 Selekcja atrybutów

W obróbce danych częstym problemem jest ich rozmiar. Ważnym zagadnieniem praktycznym w klasyfikacji jest wybór odpowiednich atrybutów, na podstawie których algorytmy uczą się rozpoznawać wzorce. W literaturze [118, 119, 120, 121, 122] wyróżnia się różne kryteria podziału metod selekcji cech, jednak najczęściej stosowany jest następujący podział:

- metody rankingowe (znane także jako filtry lub metody filtracji) - opierają się na niezależnej od modelu klasyfikacji ocenie atrybutów za pomocą ogólnych miar charakteryzujących dane;
- metody opakowujące (*ang. wrapper methods*) - oceniają poszczególne podzbiory atrybutów za pomocą algorytmu uczenia maszynowego, a następnie wyznaczają podzbiór, który zostanie wykorzystany do klasyfikacji danych;
- metody wbudowane - są integralną częścią algorytmu klasyfikacyjnego, który na etapie tworzenia modelu decyzyjnego przypisuje wagi poszczególnym cechom lub je eliminuje.

W tym rozdziale omówiono zastosowanie klasycznych, optymistycznych i pesymistycznych miar podobieństwa w algorytmie, który umożliwia wybór atrybutów reprezentowanych za pomocą wartości przedziałowych i daje najlepsze rezultaty w klasyfikacji, w szczególności diagnostyce medycznej (wsparcie diagnostyki raka piersi, zob. opis zbioru danych w rozdziale 5.1.2).

5.3.1 Nowy algorytm wyznaczania istotności atrybutów w zbiorze danych - IVRelief

Proponowany algorytm IVRelief jest rozszerzoną wersję algorytmu Relief, dostosowaną do danych przedziałowych na podstawie wskaźnika pierwszeństwa oraz podobieństwa przedziałowych zbiorów rozmytych. Algorytm ten w oryginalnej wersji (w przypadku klasycznym) został przez nas opracowany i opisany w pracy [50]. Następnie zbadano jego efektywność w przypadkach możliwości (optymistycznym) i konieczności (pesymistycznym) w [51]. Efektywność doboru atrybutów oceniono za pomocą algorytmu IV-kNN przedstawionego w rozdziale 5.1. Wybór algorytmu Relief [50, 119] jako bazy nowo tworzonego algorytmu służącego do selekcji cech reprezentujących dane z niepewnością był podyktowany jego wyjaśnialnością i zastosowania miar uwzględniających niepewność. Nowością proponowanego podejścia jest wykorzystanie miary podobieństwa pomiędzy przedziałowymi zbiorami rozmytymi, która uwzględnia niepewność informacji w aspekcie epistemicznym.

Wielu badaczy zaproponowało różne operatory i miary dla przedziałowych zbiorów rozmytych, badając różnorodne zależności między nimi oraz praktyczne zastosowania w rozwiązywaniu rzeczywistych problemów, w tym diagnostyki medycznej, np. [17, 24, 73, 86, 76, 65, 71, 89, 91, 92, 93]. Jednak zagadnienie modelowania niepewności wymaga wciąż poszukiwania nowych efektywnych rozwiązań. Zaproponowane przedziałowe miary podobieństwa w klasach możliwości i konieczności w problemie szukania minimalnego rozmiaru podzbioru cech, który jest niezbędny i wystarczający do reprezentowania obiektów, aby zmaksymalizować dokładność predykcji lub zmniejszyć złożoność modelu bez znacznego obniżenia dokładności klasyfikacji wypełnia tę lukę w badaniach.

Zastosowanie przedziałów do reprezentowania niepewności w danych oraz miary porównywalności związanych z rachunkiem przedziałowym uwzględnia różne podejścia epistemiczne, mniej lub bardziej restrykcyjne (optymistyczne lub pesymistyczne).

Głównym celem badań było wsparcie diagnostyki raka piersi poprzez:

- 1) Wykorzystanie przedziałowych miar podobieństwa w aspektach: klasycznym, możliwości i konieczności, w procedurze selekcji cech (algorytm IVRelief).
- 2) Weryfikacja skuteczności IVRelief za pomocą algorytmu IV-kNN (za pomocą algorytmu IV-kNN).

Relief, jako metoda selekcji atrybutów, oblicza statystykę dla każdego atrybutu, która może posłużyć do oszacowania jego istotności w później przeprowadzonym procesie przewidywania decyzji. Te statystyki atrybutów są określane jako wagi atrybutów (gdzie W_i oznacza wagę atrybutu A_i), a ich wartości mogą wahać się od -1 (najgorszy) do +1 (najlepszy). Warto jednak zauważyć, że oryginalny algorytm Relief był ograniczony do problemów klasyfikacji binarnej i nie uwzględniał mechanizmów obsługi danych niepewnych. W odpowiedzi na te ograniczenia zaproponowano modyfikację algorytmu Relief, nazwaną IVRelief, która działa na przedziałowych zbiorach rozmytych, uwzględniając niepewność w danych i zobrażowana została za pomocą pseudo kodu w algorytmie 3.

Jako dane wejściowe algorytm przyjmuje n -elementowy zbiór danych P , w którym każdy obiekt składa się z a cech (atrybutów) z wartościami z L^I oraz przypisaną decyzją pozytywną lub negatywną. Na początku zbiór P jest dzielony na dwa rozłączne zbiory: P^+ dla obiektów z pozytywną decyzją oraz P^- dla obiektów z negatywną decyzją.

Algorytm rozpoczyna działanie od utworzenia wektora W , który zawiera wagi poszczególnych atrybutów, początkowo zainicjowanych zerami. Dla każdego obiektu w zbiorze P algorytm wyszukuje najbardziej podobny obiekt z P^+ , oznaczany jako "*nearHit*" oraz najbardziej podobny obiekt z P^- , oznaczany jako "*nearMiss*".

Następnie wektor wag jest aktualizowany za pomocą funkcji *diff*:

$$diff(a, b) = \begin{cases} 0, & D_{IV}(a, b) \leq \gamma, \\ 1, & \text{w przeciwnym wypadku,} \end{cases}$$

Algorithm 3: Algorytm IVRelief wyznaczający wagi dla poszczególnych atrybutów w zbiorze danych P

input : n -elementowy zbiór danych P opisanych a atrybutami o wartościach przedziałowych
 przedziałowa miara odległości $D_{IV} = N_{IV}(S)$
output: wektor wag W o rozmiarze a

$W \leftarrow (0, 0, \dots, 0);$
for *each* $x \in P$ **do**
 znajdź najbliższy wektor "nearHit" w P^+ używając D_{IV} ;
 znajdź najbliższy wektor "nearMiss" w P^- używając D_{IV} ;
 for $j \leftarrow 1$ **to** a **do**
 $W_j \leftarrow W_j - \text{diff}(x_j, \text{nearHit}_j)/n + \text{diff}(x_j, \text{nearMiss}_j)/n$
 end
end

gdzie $a, b \in L^I$, a parametr γ jest małą stałą wartością progową, która określa, przy jakim poziomie przedziały są uważane za nierozróżnialne. Dodatkowo, przedziałowa odległość D_{IV} jest definiowana na podstawie przedziałowej miary podobieństwa S , co można zapisać jako $D_{IV} = N_{IV}(S)$, gdzie N_{IV} jest przedziałową negacją rozmytą. W wyniku działania algorytmu, otrzymujemy wektor W zawierający wagi, które określają istotność poszczególnych atrybutów. W kontekście przyszłej klasyfikacji, atrybuty są wybierane, jeśli ich wartość istotności przekracza ustalony próg istotności θ , przy czym $0 \leq \theta \leq 1$.

5.3.2 Analiza wyników badań wpływu selekcji cech na jakość klasyfikacji

Do testów algorytmu wykorzystano zbiór danych dotyczący diagnozowania raka piersi, opisany szerzej w rozdziale 5.1.2. Rozważono tutaj niepewność w sensie epistemicznym, co oznacza, że dane są reprezentowane przez wartości przedziałowe. Dla każdej wartości atrybutów podano odchylenie standardowe oraz średnią wartość pomiarów atrybutu dla danego pacjenta.

Proces ewaluacji i analiza algorytmu została przeprowadzona w trzech głównych scenariuszach:

- Model 1 oparty o miary klasyczne (zgodne z porządkiem częściowym lub liniowym) oraz porównanie modelu przedziałowego do przypadku bazowego, czyli bez redukcji cech;
- Model 2 oparty o miary możliwość;
- Model 3 oparty o miary konieczności.

W związku z wnioskami płynącymi z analiz w rozdziale 5.1 koncentrujemy na przedziałowych miarach podobieństwa będących efektywniejszym od rozmytych, szczególnie w klasyfikacji dla danych medycznych dzięki uwzględnieniu niepewności i nieprecyzyjności.

Problem klasyfikacyjny posłużył do oceny skuteczności metody doboru atrybutów. Zatem weryfikacja działania algorytmu IVRelief została przeprowadzona za pomocą algorytmu IV-kNN, opisanego w rozdziale 5.1, w którym kilka etapów zostało zmodyfikowanych. Mianowicie:

Etap II z IV-kNN W etapie II w algorytmie IV-kNN (zob. rozdział 5.1) zastosowano inną metodę konstrukcji przedziałowych wartości dobrze modelującą niepewność i wykorzystującą naturalny potencjał danych poprzez istnienie: średniej wartości pomiaru (oznaczonej jako m_{vm}) i odchylenia standardowego pomiaru (oznaczanego jako σ_{vm}) dla każdego parametru dla badanych pacjentów. Zatem konstruowana jest wartość przedziału następująco:

$$[m_{vm} - \sigma_{vm}, m_{vm} + \sigma_{vm}].$$

Oba końce przedziałów są następnie normalizowane za pomocą wzoru (5.1), przy jednoczesnej kontroli warunku: dolny kraniec jest mniejszy lub równy od górnego krańca przedziału.

Rozważono trzy scenariusze analizy dotyczących usunięcia atrybutu o najmniejszej wadze spośród wskazanych przez parametr θ , na podstawie trzech rodzajów podobieństw:

- podobieństwo klasyczne S , Model 1,
- podobieństwo możliwości S_{pos} , Model 2,
- podobieństwo konieczności S_{nec} , Model 3.

Ocena skuteczności klasyfikacji algorytmu IVRelief we wszystkich scenariuszach opierała się na następujących założeniach:

- dane zostały podzielone na część treningową i testową, gdzie 90% obiektów stanowiło zbiór treningowy, a 10% zbiór testowy;
- algorytm przetestowano, stosując 10-krotną walidację krzyżową;
- z dziesięciu uzyskanych wyników dla dokładności (*ang. accuracy*), czułości (*ang. sensitivity*), specyficzności (*ang. specificity*) i precyzji (*ang. precision*) obliczono średnią arytmetyczną, którą przedstawiono jako ostateczny wynik algorytmu dla zadanych parametrów;
- algorytm sprawdzono dla różnych wartości $k \in \{1, 3, 5\}$.

Badając algorytm IVRelief przetestowano różne wartości dla parametru γ oraz k . Poniższa tabela przedstawia optymalne wyniki dla $\gamma = 0.01$, $0 \leq \theta \leq 0.01$ oraz $k=5$.

W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}
0.012	0.042	0.007	0.075	-0.05	0.04	0.054	0.117	0.128	-0.035

Tabela 5.7: Wyniki IVRelief - wektor wag uzyskany z wykorzystaniem miary podobieństwa konieczności

Parametry, które zapewniły najlepsze wyniki przy wykorzystaniu klasycznej miary podobieństwa S , to: agregacja $\mathcal{A}_{mean}$, wskaźnik pierwszeństwa oparty na agregacji $\mathcal{A}_{0.4}$, oraz porządek $\leq_{XY}$ stosowany do porównywania wartości przedziałowych w krokach IV i VII algorytmu IV-kNN.

Podobieństwo	Badany przypadek	accuracy	sensitivity	specificity	precision
S	Bez redukcji atrybutów	0.961	0.927	0.993	0.992
	Z redukcją atrybutów (IVRelief)	0.964	0.934	0.993	0.993
S_{pos}	Z redukcją atrybutów (IVRelief)	0.928	0.967	0.750	0.937
S_{nec}	Z redukcją atrybutów (IVRelief)	0.982	0.968	0.972	0.994

Tabela 5.8: Porównanie metod redukcji atrybutów opartych na różnych klasach miar podobieństwa

W przypadku **miary podobieństwa konieczności** S_{nec} z agregacją $\mathcal{A}_{mean}$ i wskaźnikiem $Prec_{nec}$ z przykładu 13, zaobserwowano wzrost skuteczności metody w porównaniu z metodami opartymi na klasycznej oraz optymistycznej budowie miary podobieństwa. Uzyskano najlepszy poziom dokładności klasyfikacji na poziomie 98,2% (zwiększając dokładność o około 2%) poprzez odfiltrowanie atrybutu o najmniejszej wadze (atrybut nr 5) wyznaczonej przez algorytm 3.

5.4 System prewencji depresji

5.4.1 Opis problemu i procesu budowy zbioru danych

Depresja jest schorzeniem coraz bardziej powszechnym we współczesnych społeczeństwach i dotyczy zarówno osoby młode, jak i starsze. Szacuje się, że do 2030 roku stanie się ona drugą najczęstszą przyczyną niepełnosprawności w krajach wysoko rozwiniętych, a w krajach nisko rozwiniętych zajmie trzecie miejsce [123]. Według danych WHO na depresję cierpi szacunkowo 5% dorosłej populacji, co stanowi około 390 milionów ludzi na świecie. W przypadku umiarkowanych lub ciężkich epizodów depresji, schorzenie to może stanowić zagrożenie życia pacjenta. W najgorszym przypadku może prowadzić do samobójstwa, z którego powodu co roku ponad 800 000 osób na świecie traci życie. Samobójstwo jest czwartą najczęstszą przyczyną zgonów w grupie wiekowej 15-29 lat [124].

Zgodnie z Międzynarodową Statystyczną Klasyfikacją Chorób i Problemów Zdrowotnych (ICD-10), w kategorii zaburzeń psychicznych i zachowania wyróżnia się szeroką grupę zaburzeń nastroju (F30-F39). Różnorodność kliniczna tych zaburzeń uzasadnia podział na konkretne diagnozy: epizod maniakałny (F30), zaburzenia afektywne dwubiegunowe (F31), epizod depresyjny (F32), nawracające zaburzenia depresyjne (F33), uporczywe zaburzenia nastroju (F34), inne zaburzenia nastroju (F38) oraz zaburzenia nastroju nieokreślone (F39). W grupie zaburzeń nastroju wyróżnia się depresję jednobiegunową i depresję w przebiegu choroby afektywnej dwubiegunowej. Depresję jednobiegunową klasyfikuje się jako epizod dużej depresji (F32, F33), dystymię (F34.1) oraz inne, łagodniejsze formy depresji. Depresja dwubiegunowa natomiast występuje w przebiegu choroby afektywnej dwubiegunowej, w której pacjent naprzemiennie doświadcza epizodów depresji i manii (typ I) lub hipomanii (typ II) [125]. Obecnie stosowane metody diagnostyki depresji opierają się m.in. na testach diagnostycznych, które mogą stanowić pomocne narzędzia dla psychiatrów. Do najczęściej wykorzystywanych należą m.in. test Becka, skala oceny depresji Hamiltona, PHQ-9, QIDS-SR, HDRS, GDS, CES-D i CESD-R. Wiele z tych narzędzi spotyka się jednak z krytyką, np. ze względu na niską trafność, jak w przypadku skali sceny depresji Hamiltona [126]. Jedną z najnowszych skal diagnostycznych jest skala CESD-R, która została opracowana w 2004 roku jako uaktualnienie skali CES-D przez Williama Eatona i współpracowników [127]. Jest to jedno z najczęściej stosowanych narzędzi w epidemiologii psychiatrycznej. Skala ma charakter samoopisowy – respondenci wybierają jedną z pięciu możliwych odpowiedzi na każde z 20 stwierdzeń opisujących ich samopoczucie lub zachowanie. Odpowiedzi są punktowane od 0 (brak objawów) do 4 (objawy codzienne przez co najmniej dwa tygodnie), a łączna liczba punktów może wynosić od 0 do 80. Wynik 16 punktów lub więcej może sugerować konieczność konsultacji psychiatrycznej lub psychologicznej [126]. Odpowiedzi są klasyfikowane na podstawie dziewięciu grup objawów depresji [128], zgodnie z definicją zawartą w [129].

1. Smutek (dysforia).
2. Utrata zainteresowania (anhedonia).
3. Zmniejszony apetyt lub chudnięcie bez diety.
4. Bezsenność lub hipersomnia.
5. Niepokój psychosomatyczny lub spowolnienie.
6. Zmęczenie lub utrata energii.
7. Problemy z koncentracją.
8. Poczucie bezwartościowości lub winy.
9. Myśli samobójcze.

Wspomniana wcześniej wartość graniczna 16 punktów w diagnostyce, która wywołała obawy w środowisku medycznym (zwłaszcza w przypadku pacjentów z wynikami bliskimi tej wartości), była motywacją do podjęcia badań nad opracowaniem nowego testu i systemu diagnostycznego. System ten miałby wspierać wczesną diagnostykę depresji oraz prewencję u pacjentów znajdujących się w tej newralgicznej grupie. Celem tych badań było stworzenie bardziej precyzyjnych narzędzi diagnostycznych, które pozwoliłyby skuteczniej identyfikować osoby wymagające interwencji, zanim ich stan ulegnie pogorszeniu.

Nie zaobserwowano w literaturze funkcji prewencji u żadnego z dostępnych testów diagnostyki depresji. Dlatego opierając się na analizie dostępnych miar depresji (w szczególności na skali CESD-R), podjęto się w ramach pracy doktorskiej stworzenia nowego, bardziej czułego testu, dostosowanego do aktualnej sytuacji społeczno-gospodarczej [38], czyli wskazującego pacjentów zagrożonych w najbliższym czasie objawami depresji. W wyniku tych badań opracowano wraz z ekspertem dziedzinowym (psychiatrą) autorski test SenDD, składający się z 22 pytań, z których dwa pełnią funkcję weryfikacyjną, oceniając zagrożenie depresją w sytuacji bliskości granicy decyzji. Wtedy pytania 21 i 22 o dualnym wydźwięku emocjonalnym w stosunku do pierwszych 20 pytań ewaluują ostateczną decyzję (algorytm 5) tworząc tym samym bardziej czuły na zagrożenia depresją system diagnostyczny.

Pytania zawarte w nowym teście SenDD to:

1. Miałem(am) nastrój depresyjny (pesymizm w odniesieniu do przyszłości, poczucie smutku, skłonność do płaczu).
2. Miałem(am) poczucie winy.
3. Miałem(am) kłopoty ze snem.
4. Czułem(am) się beznadziejny(a) i bezwartościowy(a).
5. Nie miałem(am) siły wyjść z domu i podejmować się nowych zajęć.
6. Czułem(am), że moje życie jest puste.
7. Byłem(am) zrezygnowany(a) i nie widziałem(am) sensu życia.
8. Czułem(am) się gorszy(a) od innych.

9. Byłem(am) nerwowo(a) i rozdrażniony(a).
10. Czułem(am) nieuzasadniony strach.
11. Miałem(am) kłopoty z koncentracją na prostych czynnościach (np. oglądanie telewizji).
12. Miałem(am) zmienione łaknienie.
13. Nic mnie nie interesowało, większość rzeczy była dla mnie obojętna.
14. Bardziej niż zwykle martwiłem(am) się swoim stanem zdrowia.
15. Czułem(am), że zawodzę innych ludzi.
16. Męczyła mnie obecność innych ludzi.
17. Zaniedbywałem(am) codzienne obowiązki.
18. Miałem(am) zmniejszone libido.
19. Straciłem(am) na wadze.
20. Znałomi/rodzina zauważyli obniżenie mojego nastroju.
21. Wykonywanie przyjemnych dla mnie czynności sprawiało, że czułem(am) radość (np. wyjście na spacer, sport).
22. Potrzebowałem(am) spotkania ze znajomym, aby poprawić swój nastrój.

Respondenci zostali poproszeni o zaznaczenie odpowiedzi, która najlepiej odzwierciedla ich stan psycho-fizyczny w ciągu ostatnich dwóch tygodni. W zależności od odpowiedzi, mogli uzyskać od 0 do 4 punktów za każde pytanie. Przy czym zauważmy, że punktacja odzwierciedla nasilenie stanu emocjonalnego, rośnie wraz ze zwiększaniem negatywnych odczuć. Każdemu pytaniu przyporządkowano następujące możliwości odpowiedzi wraz z odpowiednią punktacją:

- zdecydowanie tak: 4 punkty;
- raczej tak: 3 punkty;
- trudno powiedzieć: 2 punkty;
- prawdopodobnie nie: 1 punkt;
- zdecydowanie nie: 0 punktów.

Powyższy test posłużył do stworzenia bazy danych. Zatem zbiór danych na którym testowano nowy system diagnostyczny powstał w ramach badania przeprowadzonego techniką CAWI dwukrotnie (w odstępach 4 tygodniowych) na grupie 750 studentów. W obu badaniach uczestniczyli studenci Uniwersytetu Rzeszowskiego. Wśród 750 respondentów 59,6% stanowiły kobiety, a 40,4% mężczyźni. Ponad połowa z nich mieszkała na wsi (57,7%).

Przebadana została grupa 750 studentów Uniwersytetu Rzeszowskiego i przeprowadzono to na dwa sposoby:

- przy użyciu standaryzowanego CESD-R;
- przy użyciu nowego testu.

Dane zebrano przy użyciu obu testów z przypisanymi diagnozami według 20 atrybutów plus atrybut wyjściowy, diagnoza dla testu CESD-R, natomiast diagnostyka w nowym teście o nazwie SenDD oparta jest na metodzie sztucznej inteligencji, klasyfikacji, po wykorzystaniu wiedzy eksperckiej (22 atrybuty plus diagnoza, <https://tiny.pl/92s9g>). Zatem zbiór danych zawiera 750 obiektów, w tym 501 z decyzją 0 i 249 z decyzją 1 (klasy) „0” (brak zagrożenia depresją) i „1” (zagrożenie depresją). Pełny opis zebranych danych można zobaczyć w [38].

5.4.2 Model wspomagania prewencyjnego wykrywania epizodów depresyjnych

Zaproponowany **System Prewencji Depresji (SPD)** składa się z dwóch etapów opartych o test SenDD:

- (i) Wstępnego przypisania diagnozy (algorytm 4);
- (ii) Weryfikacja diagnozy względem pytań 21 i 22 celem zwiększenia czułości klasyfikacji i wczesnej - prewencji objawów depresji (algorytm 5).

Poniżej przedstawiono algorytm szukania wstępnej decyzji w diagnostyce depresji (etap 1 systemu), gdzie w procesie budowy przedziałów (IVFS(a)) - reprezentacja niepewności dała wartości atrybutu danego obiektu) stosowano regułę:

$$a = [0, 1] \text{ jeśli brak wartości } a, \text{ w przeciwnym przypadku } a = [a, a].$$

Następnie w procesie normalizacji zgodnie ze wzorem (5.1) i zasadą $\underline{a} \leq \bar{a}$ otrzymujemy reprezentację za pomocą przedziałowych zbiorów rozmytych każdego obiektu.

Aby w kolejnym kroku, w algorytmie 4, przy wykorzystaniu przedziałowych miar podobieństwa, wyznaczyć wstępną diagnozę/klasę decyzyjną, którą w przypadku decyzji bliskiej granicznej zostanie poddana weryfikacji za pomocą algorytmu 5 (SENSDEPR).

Algorithm 4: Algorytm wyznaczania wstępnej diagnozy

Variables: *sim* - tablica z podobieństwami pomiędzy parami obiektów

P - *n* - elementowy zbiór danych *P* zbudowany z *a* wektorów i decyzji

k - liczba branych pod uwagę najbliższych sąsiadów

Output : Klasa dla obiektu testowanego *t*

```

for x ∈ P do
    for ax do
        ax ← FS(ax) ;           // normalizacja wartości atrybutu
        ax ← IVFS(ax) ;       // zamiana wartości atrybutu na przedział
    end
end
for x ∈ P do
    | sim(x) ← S(x, t);
end
sort(sim);
sim = sim[1, ..., k];
for i := 1, 2, ..., k do
    | Agregacja każdej klasy decyzyjnej wśród k najbardziej podobnych sąsiadów;
end
return run(algorithm 1);

```

Algorytm 5 został zaprojektowany w celu weryfikacji decyzji podejmowanej przez algorytm 4. Zaproponowane decyzje są weryfikowane poprzez sprawdzenie, czy dany obiekt nie znajduje się na granicy przyjętego poziomu diagnostycznego depresji, tj. 16 punktów (w oparciu o miarę odległości podobieństw najbliższych obiektów z decyzją "depresja") poprzez wykorzystanie pytań weryfikacyjnych 21 i 22 o odwróconej intencji emocjonalnej. Czynnikiem weryfikującym były odpowiedzi pacjentów na dwa pytania weryfikujące (pytania 21 i 22) zawarte w teście. Jeśli wartość udzielonych odpowiedzi była większa niż 2, co oznacza, że pacjent wybrał jedną z opcji *zdecydowanie tak* lub *raczej tak*, algorytm kwalifikował go jako osobę zagrożoną ryzykiem wystąpienia depresji, aby wcześniej wykryć to ryzyko.

Wykorzystując nowe klasy miar podobieństwa oraz wyniki zwracane przez algorytm 4, algorytm 5 identyfikuje osoby z potencjalnym ryzykiem wystąpienia depresji. Oznacza to, że u danej osoby w danym momencie nie występują epizody depresyjne, ale jej odpowiedzi sugerują, że w niedalekiej przyszłości mogą się one pojawić. Celem algorytmu 5 jest więc wyłonienie takich osób, aby możliwe było podjęcie działań zapobiegawczych odpowiednio wcześniej. Zatem przez większą czułość klasyfikacji proponowanego systemu będziemy rozumieć wcześniejsze, w stosunku do innych metod diagnostyki, wskazanie zagrożenia depresją, czyli prewencja depresji przed wystąpieniem poważnych objawów.

Algorithm 5: Algorytm SENSDEPR - zwiększanie czułości diagnostyki

Variables: D jest przedziałową miarą odległości
 o_x - obiekt, który jest najbardziej podobny do obiektu o_y , który klasyfikujemy
 o_z - obiekt, dla którego szukamy decyzji
 o_z - obiekt z decyzją "1", który jest najbardziej podobny do obiektu o_y
 ϵ - parametr czułości/bliskości do decyzji "1" (zagrożenia depresją), $\epsilon \in [0, 1]$

Output : decyzja dla obiektu o_y

```

for  $o_x = 0$  to  $N$  do
    if  $o_y[21] < 2$  and  $o_y[22] < 2$  and  $o_y[dec] == 0$  then
        for  $o_z = 0$  to  $N$  do
            | znajdowanie najbardziej podobnego sąsiada którego  $o_z[dec] = 1$ ;
        end
    end
    if  $D(S(o_y, o_x), S(o_y, o_z)) = \max(|\underline{S}(o_y, o_x) - \underline{S}(o_y, o_z)|, |\overline{S}(o_y, o_x) - \overline{S}(o_y, o_z)|) < \epsilon$ 
        then
            | return 1
        else
            | return 0
        end
    end
end

```

5.4.3 Ocena Systemu Prewencji Depresji (SPD) i analiza wpływu algorytmu SENSDEPR na czułość diagnostyki depresji

Na początku zauważmy, że wzorowanie się w systemie, w algorytmie 4 na technice związanej z algorytmem k najbliższych sąsiadów (kNN) jest podyktowane faktem, że chcemy porównać nasz System Prewencji Depresji i oparty na nim test SenDD do testu CESD-R, który jest jednym z najefektywniejszych testów diagnozujących depresję w środowisku medycznym [126, 130]. A najwyższą wartość Precyzji test CESD-R uzyskuje dla właśnie algorytmu kNN. Porównanie wpływu różnych metod uczenia maszynowego na diagnostykę ww. testem możemy zaobserwować w [130] i w tabeli 5.9.

Algorytmy uczenia maszynowego	PREC
Random Forest	65.21
kNN	95
Rule generation	80.2

Tabela 5.9: Różne metody uczenia maszynowego dla CESD-R

Ewaluacja systemu prewencji depresji prowadzona była w następujących scenariuszach:

- Scenariusz 1** Sprawdzenie skuteczności systemu w zależności od wykorzystywanych klas miar podobieństwa (klasycznej, optymistycznej, pesymistycznej) we wczesnym wykryciu pacjentów, u których w niedalekiej przyszłości istnieje ryzyko wystąpienia depresji;
- Scenariusz 2** Analiza porównawcza wytypowanych pacjentów z zagrożeniem depresją za pomocą Systemu Prewencji Depresji (SPD) i testu SenDD z rezultatami uzyskanymi po przeprowadzeniu testu CESD-R.

Testy zrealizowano na danych, w których część treningowa zawierała 70% obiektów, a część testowa 30%. Algorytm testowany był z wykorzystaniem 10-krotnej walidacji krzyżowej. W wyniku każdej z 10 prób uzyskano wartości współczynników: dokładności (4.4), czułości (4.6), specyficzności (4.7) oraz precyzji (4.8). Następnie obliczono ich średnią arytmetyczną, która przyjęta została jako rezultat procesu klasyfikacji dla konkretnego zestawu zadanych parametrów wejściowych. W procesie ewaluacji analizowano także różne wartości k oraz ϵ .

Scenariusz 1.

W tabeli 5.10 przedstawiono zestawienie najlepszych wyników uzyskanych w procesie klasyfikacji w zależności od używanej klasy miary podobieństwa i związanymi z nimi parametrami takimi jak porządki, agregacje, czy wskaźniki pierwszeństwa dla klas: klasycznej, możliwości i konieczności (zob. rozdział 2 i 3).

	accuracy	sensitivity	specificity	precision	k
Podobieństwo klasyczne	0,913978	0,817692	0,978723	0,975	1
	0,905914	0,815217	0,974681	0,893377	3
	0,897849	0,804348	0,989362	0,986667	5
Podobieństwo pesymistyczne	0,600395	0,4	0,6	0,408195	1
	0,636638	0,501405	0,613394	0,505135	3
	0,611682	0,390069	0,664689	0,421291	5
Podobieństwo optymistyczne	0,875584	0,44444	0,875	0,94444	1
	0,961244	0,556689	0,976716	0,96032	3
	0,970486	0,817692	0,981067	0,98	5

Tabela 5.10: Wyniki efektywności Systemu Prewencji Depresji dla różnych klas miar podobieństwa

Dla klasycznej miary podobieństwa uzyskano najlepsze rezultaty dla następujących wartości parametrów wejściowych algorytmu: porządek $\leq_{LI}$ oraz agregację $\mathcal{A}_{meanmax}$ do agregacji klas decyzyjnych, a do konstrukcji miary podobieństwa wykorzystano ponadto wskaźnik pierwszeństwa $Prec_{\mathcal{A}}$.

Natomiast, najlepsze rezultaty użycia pesymistycznej miary podobieństwa zgodnej z $\leq_{nec}$ otrzymano dla agregacji $\mathcal{A}_{meanpow}$ i wskaźnika pierwszeństwa $Prec_{nec}$.

Aczkolwiek najwyższe rezultaty dla klasyfikacji w stosunku do klas klasycznej i konieczności uzyskano, wykorzystując **optymistyczną miarę podobieństwa**, S_{pos} . Miarę podobieństwa oparto o wskaźnik pierwszeństwa $Prec_{\mathcal{A}_{pos}}$, gdzie wykorzystano agregację $\mathcal{A}_1$ i negację N_1 dla $k=5$.

We wszystkich przypadkach jako agregację podobieństw wykorzystano agregację $\mathcal{A}_{mean}$ zgodną z porządkiem we wszystkich trzech klasach. Możemy zauważyć że uzyskano najwyższą wartość dokładności swoistości i precyzji w klasie miar możliwości.

Uzyskane rezultaty pokazały, że stosując **operatory i miary możliwości** poprawiają efektywność diagnostyki depresji w dwóch wymiarach: względem innych testów diagnozowania (zob. tabele 5.9) oraz różnych modeli decyzyjnych/klasyfikacji (różnych klas użytych operatorów i miar). Powyższe rezultaty potwierdzają celowość zaproponowania i stosowania do innych rzeczywistych problemów nowych miar podobieństwa, czyli miar podobieństwa możliwości.

Scenariusz 2.

Prowadzone badania dotyczyły porównania obiektów, które system SPD 5 wyznaczył jako przypadki zagrożone ryzykiem wystąpienia depresji, z klasyfikacją tych obiektów za pomocą testu CESD-R. Aby to zweryfikować, pacjenci odpowiadali na pytania zarówno z testu CESD-R, jak i z zaproponowanego nowego testu w tym samym czasie. Zestawiono wyniki uzyskane z testu CESD-R i SenDD, szczególnie zwracając uwagę na obiekty wskazane jako pacjenci, u których należy spodziewać się problemów z depresją. Weryfikacja tych pacjentów miała miejsce po czterech tygodniach. We współpracy ze specjalistą psychiatrii okazało się, że test SenDD we wcześniejszym etapie trafnie wykrył o **20%** pacjentów więcej, u których występuje ryzyko depresji. To potwierdza, że zaproponowany nowy test jest wrażliwszy na wykrywanie problemów związanych z depresją i umożliwia wcześniejszą diagnozę oraz wdrożenie leczenia jeszcze przed wystąpieniem zauważalnych objawów związanych z zachowaniem pacjenta. Nowy system przyczyni się to do poprawy jakości życia osób narażonych na epizody depresyjne poprzez prewencję tych epizodów, umożliwiając wykrywanie zagrożenia depresją przed wystąpieniem objawów klinicznych.

ROZDZIAŁ 6

Pozostałe kierunki badań

Rozdział ten, w odróżnieniu do poprzedniego, nie jest związany z wykorzystaniem zmodyfikowanej metody k najbliższych sąsiadów, a dotyczy ewaluacji oraz implementacji systemu regułowego, w szczególności uogólnionego (przedziałowego) systemu wnioskowania rozmytego.

6.1 System wspierający diagnostykę choroby sercowo - naczyniowej

6.1.1 Opis danych i problemu

W tym rozdziale przedstawione zostaną osiągnięcia dotyczące rzeczywistego problemu, jakim jest przewidywanie ryzyka zachorowania na chorobę sercowo-naczyniową (ang. Cardiovascular Disease (CVD)) w ciągu 10 lat. Choroby sercowo-naczyniowe zazwyczaj wynikają z problemów utrudniających swobodny przepływ krwi, co może prowadzić do zaburzeń w pracy serca. Te schorzenia wiążą się z wysokim ryzykiem wystąpienia ciężkich chorób, takich jak zawał serca czy zakrzepica, co skutkuje wysoką śmiertelnością w wielu krajach rozwiniętych [131]. Dlatego ważne jest, aby przewidzieć ryzyko zachorowania na choroby sercowo-naczyniowe tak, by lekarze mogli odpowiednio leczyć swoich pacjentów. W celu oszacowania ryzyka hiszpańscy lekarze korzystają z tabel REGICOR [132], które zawierają liczby reprezentujące prawdopodobieństwo zachorowania na choroby sercowo-naczyniowe na podstawie pewnych zmiennych wejściowych, takich jak płeć, wiek, obecność lub brak cukrzycy, ciśnienie tętnicze skurczowe i rozkurczowe, całkowite ciśnienie krwi oraz cholesterol całkowity lub HDL. Na podstawie tych liczb ryzyko zachorowania na CVD może być skategoryzowane jako niskie, średnie lub wysokie.

System ma na celu wsparcie młodych, niedoświadczonych lekarzy w diagnostyce CVD w ramach rutynowych badań, dlatego rozważono tylko dwie zmienne wejściowe: wiek

i skurczowe ciśnienie krwi pacjenta. Na wyjściu uzyskano oszacowanie ryzyka zachorowania na choroby sercowo-naczyniowe w ciągu kolejnych dziesięciu lat. Wykorzystany zbiór danych składa się z 898 przypadków klinicznych uzyskanych z rejestrów siedmiu ośrodków podstawowej opieki zdrowotnej w Pampelunie (Navarra, Hiszpania) w 2008 roku. Wiele danych dostarczonych przez lekarzy charakteryzuje się niedoskonałością informacji, dlatego zastosowano przedziałowe zbiory rozmyte, podobnie jak w artykule [133]. Po wyznaczeniu ryzyka pacjenci są dzieleni na trzy kategorie ryzyka, w szczególności:

- pacjenci o niskim ryzyku zachorowania (LR), u których zakres ryzyka wynosi ≤ 5 ;
- pacjenci średniego ryzyka (MR), których ryzyko mieści się w przedziale (5, 10);
- pacjenci wysokiego ryzyka (HR), których zakres ryzyka wynosi > 10 .

Celem prac podjętych w ramach rozprawy związanych z problemem diagnostyki CVD była optymalizacja systemu wnioskowania przybliżonego przedstawionego w pracach [53] lub [73] poprzez wykorzystanie klas agregacji możliwości, konieczności i pre-agregacji użytych w jego konstrukcji.

6.1.2 Stosowane metody i scenariusze badawcze

Podstawą klasyfikacji i działania Systemu wspomagającego diagnostykę chorób sercowo-naczyniowych jest model przedziałowego wnioskowania rozmytego. Poniżej zostanie omówione zastosowanie uogólnionego systemu opartego na przedziałowych regułach rozmytych do automatycznego przewidywania wartości ryzyka zachorowania, uwzględniającego niepewność informacji i decyzji dla klas stosowanych operatorów możliwości oraz konieczności. Potwierdzona została skuteczność użycia tych operatorów w porównaniu z operatorami klasycznymi rozważonymi w pracy [53]. Zatem w ramach pracy doktorskiej rozszerzono badania przedstawione w [53], gdzie zaprezentowano uogólniony model wnioskowania przybliżonego (rozmytego), a jego efektywność była analizowana dla różnych klasycznych operatorów agregacji (dla porządków częściowego lub liniowych) stosowanych w procesie agregacji przesłanek oraz dedykowanym złożeniu, będącym elementem wnioskowania poprzez zastosowanie klas operatorów możliwości i konieczności.

Poniższy rysunek przedstawia uogólniony schemat wnioskowania rozmytego wykorzystany do analizowanego problemu związanego z przewidywaniem ryzyka zachorowania na chorobę sercowo-naczyniową.



Rysunek 6.1: Ogólny schemat wnioskowania

Przedziałowe wnioskowanie rozmyte

Jednym z najbardziej znanych obszarów zastosowań logiki rozmytej i jej rozszerzeń jest wnioskowanie przybliżone (rozmyte), w którym z nieprecyzyjnych danych wejściowych oraz rozmytych reguł otrzymujemy często nieprecyzyjne wnioski [88]. Wnioskowanie przybliżone za pomocą zbiorów rozmytych obejmuje szeroki zakres schematów wnioskowania i znalazło szerokie zastosowanie w wielu dziedzinach, w szczególności w podejmowaniu decyzji i systemach eksperckich. Schematy wnioskowania w rozumowaniu przybliżonym opierają się na własności modus ponens, zwanej również A-warunkowością. W związku z tym wymaga się, aby przedziałowe funkcje agregujące, wykorzystywane w procesie wnioskowania w systemach opartych na przedziałowych regułach rozmytych, spełniały tę własność, co staje się kluczowe w ogólnym wnioskowaniu przybliżonym. Rozważane są różne metody stosowane do wnioskowania w rozumowaniu przybliżonym opartym na przedziałowych zbiorach rozmytych, ale wciąż wymagają efektywniejszych podejść uwzględniających niepewność i nieprecyzyjność.

Klasyczny schemat modus ponens został zastąpiony w modelu przedziałowego wnioskowania przez uogólniony modus ponens (GMP):

$$R_1 : \text{If } x_1 \text{ is } \mathcal{D}_1^{(1)} \text{ and/or } \dots \text{ and/or } x_n \text{ is } \mathcal{D}_1^{(n)} \text{ then } y \text{ is } \mathcal{E}_1$$

.....

$$R_n : \text{If } x_1 \text{ is } \mathcal{D}_n^{(1)} \text{ and/or } \dots \text{ and/or } x_n \text{ is } \mathcal{D}_n^{(n)} \text{ then } y \text{ is } \mathcal{E}_n.$$

$$\text{Fakt} : x \text{ is } \mathcal{D}'$$

$$y \text{ is } \mathcal{E}',$$

gdzie przesłanki ($\mathcal{D}$) i wnioski ($\mathcal{E}$) są przedziałowymi zbiorami rozmytymi w przestrzeni Y .

Przedziałowa relacja rozmyta R wyraża związek pomiędzy zmiennymi x i y , które są użyte w przesłankach. Istnieją różne metody [61, 73] pozwalające na zbudowanie R .

Dla procesu przedziałowego wnioskowania rozmytego zaproponowano następującą metodę, wykorzystującą kompozycję relacji rozmytych "o":

$$\mathcal{E}' = \mathcal{D}' \circ_{\mathcal{A}\mathcal{B}} R, \quad (6.1)$$

gdzie $\mathcal{E}', \mathcal{D}' \in IVFS(X)$ i $\mathcal{A}, \mathcal{B}$ są przedziałowymi agregacjami rozmytymi. W warunkach rozmytych zdarzają się sytuacje, w których eksperci mają problemy z określeniem stopni przynależności elementów do rozważanych zbiorów rozmytych [4]. Gdy tak się dzieje, wskazane jest stosowanie rozszerzeń teorii zbiorów rozmytych. Jednym z najczęściej stosowanych rozszerzeń jest teoria przedziałowych zbiorów rozmytych [5, 6] (lub teoria intuicjonistycznych zbiorów rozmytych Atanassov'a [134]). Dlatego analizujemy GMP, gdy zbiory $\mathcal{D}$ i $\mathcal{D}'$ są przedziałowymi zbiorami rozmytymi na zbiorze X , a zbiory $\mathcal{E}$ i $\mathcal{E}'$ są przedziałowymi zbiorami rozmytymi na zbiorze Y . Oczywiście, w tym przypadku relacja R użyta do reprezentacji reguły jest przedziałową relacją rozmytą na $X \times Y$ (zapoczątkowane np. w pracach [18, 135]).

Wykorzystując algorytm 6:

Algorithm 6: Algorytm Uogólnionego wnioskowania rozmytego

Variables: $\mathcal{R}$ - zbiór n reguł

$$R_i \in IVFR(X \times Y)$$

Inputs : Przesłanki $\mathcal{D}_1, \dots, \mathcal{D}_n, \mathcal{D}' \in IVFS(X)$;

Wnioski $\mathcal{E}_1, \dots, \mathcal{E}_n, \mathcal{E}' \in IVFS(Y)$;

Przedziałowe funkcje agregacji $\mathcal{A}, \mathcal{A}_k, k \in \{1, 2, 3\}$;

N_{IV} – przedziałowa negacja

Output : $\mathcal{E}'$

for all $r \in \mathcal{R}$ **do**

$$R_i(x, y) = \mathcal{A}_2(N_{IV}(\mathcal{D}_i(x)), \mathcal{E}_i(y))$$

$$\mathcal{E}'_i(y) = \mathcal{A}_{x \in X}(\mathcal{A}_1(\mathcal{D}'(x), R_i(x, y))),$$

end

return $\mathcal{E}' = \mathcal{A}_3(\mathcal{E}'_1, \dots, \mathcal{E}'_n)$

zaproponowano następującą metodę wyznaczania $\mathcal{E}'$, gdzie $\mathcal{A} = \mathcal{A}_3 = \vee$, natomiast $\mathcal{A}_1$ oraz $\mathcal{A}_2$ będą odpowiednio agregacjami z klas: klasycznej, możliwości (optymistycznej) lub konieczności (pesymistycznej):

1. Dla każdej reguły budowana jest związana z nią przedziałowa relacja rozmyta $R_i \in IVFR(X \times Y)$ gdzie $i = 1, \dots, n$;

2. Dla każdej reguły obliczany jest (GMP):

$$\mathcal{E}'_i(y) = \bigvee_{x \in X} (\mathcal{A}_1(\mathcal{D}'(x), R_i(x, y))), \text{ gdzie } i = 1, \dots, n.$$

3. Obliczamy

$$\mathcal{E}' = \bigvee_{i=1}^n \mathcal{E}'_i.$$

Rozważono ogólniejszą metodę do [53] i badano złożenie przedziałowych relacji rozmytych w tym samym duchu co [88], ale używając przedziałowych funkcji agregacji z różnych ww. klas zamiast przedziałowych t-norm często stosowanych w literaturze.

Powodami do rozważenia tych celów były:

- Stwierdzenie, że istnieją zastosowania, w których użycie w GMP przedziałowych zbiorów rozmytych zamiast zbiorów rozmytych oraz zastąpienie t-normy T w (6.1) funkcją agregującą $\mathcal{A}_1$ większą od minimum, a więc nie będącą t-normą, prowadzi do zmniejszenia błędu wyników.
- Ponadto jeśli przyjmiemy $\mathcal{A}_1 \geq T = \wedge$, wtedy

$$\mathcal{E}'_{\mathcal{A}}(y) = \bigvee_{x \in X} \mathcal{A}_1(\mathcal{D}'(x), R_i(x, y)) \geq \mathcal{E}'_T(y) = \bigvee_{x \in X} (\mathcal{D}'(x) \wedge R_i(x, y)) \text{ dla każdego } y \in Y.$$

W ten sposób, choć nie można zagwarantować, że pierwszy aksjomat Baldwina [136] jest spełniony, $\mathcal{E} \geq \mathcal{E}'$, można stwierdzić, że przy stosowanych konstrukcjach uzyskany wniosek jest zawsze większy lub równy wnioskowi uzyskanemu za pomocą reguły kompozycyjnej Zadeha. Oznacza to, że proces wnioskowania nigdy nie jest bardziej restrykcyjny niż zastosowana funkcja prawdy. Biorąc pod uwagę, że problem obejmuje dwie zmienne wejściowe, a dla każdej z nich rozważane są trzy etykiety lingwistyczne, otrzymujemy 9 reguł, które zostały przedstawione w tabeli 6.1.

Skurczowe ciśnienie krwi			
Wiek	Niskie(Low)	Średnie(Medium)	Wysokie(High)
Niska(Low)	LR	LR	LR
Średnia(Medium)	LR	MR	MR
Wysoka(High)	LR	MR	HR

Tabela 6.1: Reguły rozmyte wykorzystane w niniejszej pracy

Interpretacja reguł z powyższej tabeli jest następująca:

Rule1: If Age is Low and Syst.blood pres. is Low then Risk is Low

Rule2: If Age is Medium and Syst.blood pres. is Low then Risk is Low

...

Rule9: If Age is High and Syst.blood pres. is High then Risk is High

W celu skonstruowania relacji związanej z każdą regułą, fazyfikowano wartości wszystkich elementów do odpowiadających im przedziałowych zbiorów rozmytych dla tej reguły. Następnie łączono przedziałowe zbiory rozmyte poprzedników i wykonano iloczyn kartezjański z przedziałowym zbiorem rozmytym występującym w następniku reguły. W efekcie relacja R powiązana z rozważaną regułą jest reprezentowana przez przedziałową macierz o 1796 wierszach i 898 kolumnach. Aby pracować z przedziałami, zastosowano te same operacje zarówno do dolnych, jak i górnych granic podziałów, co oznacza użycie reprezentowalnych operacji przedziałowych. Ostatecznie uzyskano 9 relacji, które są reprezentowane przez 9 macierzy przedziałowych.

Aby przewidzieć ryzyko zachorowania dla nowego pacjenta, postępowano zgodnie z następującymi krokami:

- W pierwszej kolejności dla każdego atrybutu tworzone przedziałowy zbiór rozmyty w dwóch krokach K1 i K2. Przeprowadzono podwójną fazyfikację, polegającą na konstrukcji przedziałów z wykorzystaniem funkcji ignorancji. W przypadku systemu opartego na regułach rozmytych o wartościach przedziałowych rozważono użycie trzech lingwistycznych etykiet dla każdej zmiennej rozmytej (*niska*, *średnia* i *wysoka*), które są modelowane za pomocą przedziałowych zbiorów rozmytych. Szerokość przedziału przynależności wykorzystano do modelowania niepewności związanej z przypisaniem pojedynczej liczby jako stopnia przynależności elementów do zbiorów rozmytych.

K1 Generowane są zbiory rozmyte, wykonując jednorodny podział przestrzeni wejściowej dla każdej zmiennej. W tym celu stosowane są:

- funkcja przynależności typu L, która ma następującą postać

$$A(x; a, b) = \begin{cases} 1, & \text{dla } x \leq a, \\ \frac{b-x}{b-a}, & \text{dla } a < x \leq b, \\ 0, & \text{dla } x > b, \end{cases}$$

dla której przyjęto parametry (25, 35) dla wieku oraz (120, 130) dla skurczowego ciśnienia krwi,

- funkcja przynależności typu R, która ma następującą postać

$$A(x; a, b) = \begin{cases} 0, & \text{dla } x \leq a, \\ \frac{x-a}{b-a}, & \text{dla } a < x \leq b, \\ 1, & \text{dla } x > b, \end{cases}$$

dla której przyjęto parametry (45, 65) dla wieku i (140, 160) dla skurczowego ciśnienia krwi

- trapezoidalna funkcja przynależności, przyjmująca następującą postać

$$A(x; a, b, c, d) = \begin{cases} 0, & \text{dla } x \leq a, \\ \frac{x-a}{b-a}, & \text{dla } a < x \leq b, \\ 1, & \text{dla } b < x \leq c, \\ \frac{d-x}{d-c}, & \text{dla } c < x \leq d, \\ 0, & \text{dla } x > d, \end{cases}$$

przyjmująca parametry (25, 35, 45, 65) dla wieku i (120, 130, 140, 160) dla skurczowego ciśnienia krwi.

K2 Dla każdego zbioru rozmytego tworzony jest odpowiadający mu przedziałowy zbiór rozmyty, stosując metodę konstrukcji opisaną w rozdziale 5.1.1, która opiera się na funkcji ignorancji (definicja 29).

- Następnie obliczono przedziałowe relacje rozmyte odpowiadające każdej regule.
- Przedziałowe zbiory rozmyte - następniki wszystkich reguł agregowano przy użyciu operacji maksimum dla poszczególnych klas decyzji.

Na koniec obliczono wartość będącą przewidywanym ryzykiem zachorowania na CVD dla badanego pacjenta (por. [137]):

$$\underline{y} = \frac{\sum_{k=1}^N \underline{R}_k(\bar{y}_k) \bar{y}_k}{\sum_{k=1}^N \underline{R}_k(\bar{y}_k)},$$

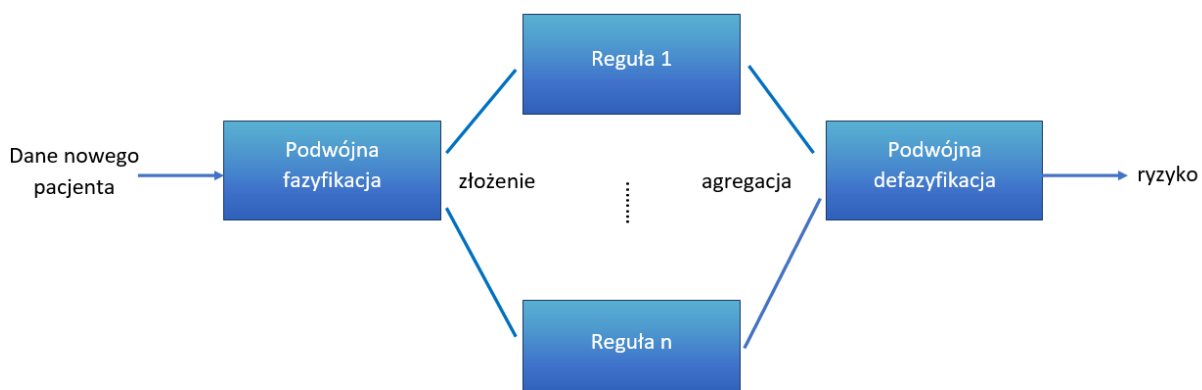
$$\bar{y} = \frac{\sum_{k=1}^N \bar{R}_k(\bar{y}_k) \bar{y}_k}{\sum_{k=1}^N \bar{R}_k(\bar{y}_k)},$$

gdzie $\bar{y}_k$ jest wartością centralną zbioru z maksymalnymi wartościami ryzyka w k -tej regule, $R_k(\bar{y}_k)$ jest maksymalną wartością w k -tej regule przy danym ryzyku.

Ostatecznie otrzymano

$$y' = \frac{\underline{y} + \bar{y}}{2}.$$

W procesie optymalizacji systemu diagnostycznego w problemie oceny ryzyka zachorowania na CVD przyjęto następujące scenariusze badawcze, w których podstawą było



Rysunek 6.2: Schemat oceny ryzyka zachorowania na CVD dla nowego pacjenta

rozważenie miar agregacji $\mathcal{A}_1$ i $\mathcal{A}_2$ z klas: klasycznej, możliwości, konieczności i pre-agregacji, odpowiednio użytych jak przedstawiono w algorytmie 6:

- Model 1 - oparty o miary agregacji klasyczne (zgodne z porządkiem częściowym lub liniowym);
- Model 2 - oparty o miary agregacji możliwości;
- Model 3 - oparty o miary agregacji konieczności;
- Model 4 - oparty o miary pre-agregacji.

6.1.3 Ocena systemu i analiza wyników eksperymentów

W ramach eksperymentów sprawdzono wpływ różnych klas agregacji w procesie wnioskowania (przy budowie relacji R -agregacji przesłanek oraz szukaniu końcowej wartości $\mathcal{E}'$) na efektywność klasyfikacji ryzyka zachorowania na chorobę sercowo-naczyniową.

Analiza algorytmu została przeprowadzona w ww. scenariuszach (Model 1 -4), w których podstawą było uwzględnienie różnych klas dla miar agregacji $\mathcal{A}_1$ i $\mathcal{A}_2$ w algorytmie 6.

Rozważono następujące agregacje klasyczne, optymistyczne (możliwości) i pesymistyczne (konieczności):

$\mathcal{A}_{prod}$, $\mathcal{A}_{hmean}$, $\mathcal{A}_{gmean}$, $\mathcal{A}_{Lukasiewicz}$, $\wedge$, $IVOWA$, $\mathcal{A}_{mean}$, $\mathcal{A}_{pos1}$, $\mathcal{A}_{pos2}$, $\mathcal{A}_{pos18}$, $\mathcal{A}_{nec2}$, $\mathcal{A}_{nec4}$.
A także zbadano wpływ na klasyfikację pre-agregacji z wykorzystaniem operatora W .

Na podstawie uzyskanych wyników można zauważyć, że najmniejszy poziom błędów różni się w zależności od poziomu ryzyka zachorowania oraz zastosowanej metody agregacji. Wyniki przedstawione w tabeli 6.2 są zestawieniem najlepszych rezultatów uzyskanych dla poszczególnych klas agregacji. W przypadku niskiego ryzyka zachorowania, najmniejszy błąd osiągnięto, stosując t-normę Łukasiewicza (klasyczna agregacja). Dla

Agregacja	Poziom błędu w prognozowaniu ryzyka CVD			
	Niski	Średni	Wysoki	Całkowity
Model 1 - $\mathcal{A}_{classic}$	13.20	1.33	22.55	10.19
Model 2 - $\mathcal{A}_{pos}$	15.28	1.64	20.18	11.62
Model 3 - $\mathcal{A}_{nec}$	13.25	1.33	22.58	10.23
Model 4 - W	13.28	1.31	23.13	10.89

Tabela 6.2: Wyniki uzyskane w procesie wnioskowania w zależności od zastosowanych agregacji: $Class_{err}$

średniego ryzyka najlepsze wyniki dał operator W (patrz. (2.8), pre-agregacja), natomiast dla wysokiego ryzyka najniższy poziom błędu uzyskano przy zastosowaniu $\mathcal{A}_{mean}$ lub $\mathcal{A}_{pos1}$ (**agregacja możliwości - optymistyczna**, zob. rozdz. 2.2.2). Biorąc pod uwagę, że błędna klasyfikacja w przypadku wysokiego ryzyka CVD niesie ze sobą poważne konsekwencje dla pacjentów, można wyciągnąć wniosek, że agregacje możliwości (**Model 2**) ponownie okazały się najskuteczniejsze w procesie klasyfikacji medycznej.

6.2 Biblioteka dla systemów wnioskowania przedziałowego (Interval–Valued Fuzzy Inference System (IFIS))

W ramach prowadzonych prac badawczych nad problemem wyznaczania ryzyka zachorowania na choroby sercowo-naczyniowe opracowano bibliotekę w języku Python dedykowaną obsłudze modeli przedziałowego wnioskowania rozmytego. Motywacją był brak takiej ogólnodostępnej/publicznej biblioteki. Pierwsze wyniki tych prac zostały zaprezentowane w artykułach [54, 55], a pierwszą wersję biblioteki można znaleźć na platformie GitHub pod adresem: <https://github.com/PGrochowalski/ifis>. Zaproponowana biblioteka umożliwia obsługę Systemów wnioskowania przedziałowego (Interval–Valued Fuzzy Inference System (IFIS)) została zaimplementowana w języku programowania Python 3. Biblioteka IFIS została z powodzeniem zastosowana w badaniach opisanych w tym rozdziale. Jest ona rozszerzeniem biblioteki Simpyful [138] i wykorzystuje popularne biblioteki Numpy oraz Scipy do swojego działania. System rozszerza klasyczne wnioskowanie rozmyte o obsługę wartości przedziałowych, co pozwala na lepsze odwzorowanie niepewności w modelach decyzyjnych. Dzięki temu w systemie dane przedziałowe mogą być używane jako przesłanki w regułach, a proces agregacji odbywa się z wykorzystaniem przedziałowych operatorów agregacji. Biblioteka umożliwia również wczytywanie danych wejściowych z plików CSV lub pickle, co ułatwia jej integrację z różnymi źródłami danych.

Przygotowanie danych obejmuje dwie ścieżki:

- rozmycie danych rzeczywistych poprzez zdefiniowanie przedziałowych zbiorów roz-

mytych za pomocą standardowych funkcji przynależności (np. trójkątnej, trapezowej, gaussowskiej) lub funkcji zdefiniowanych przez użytkownika;

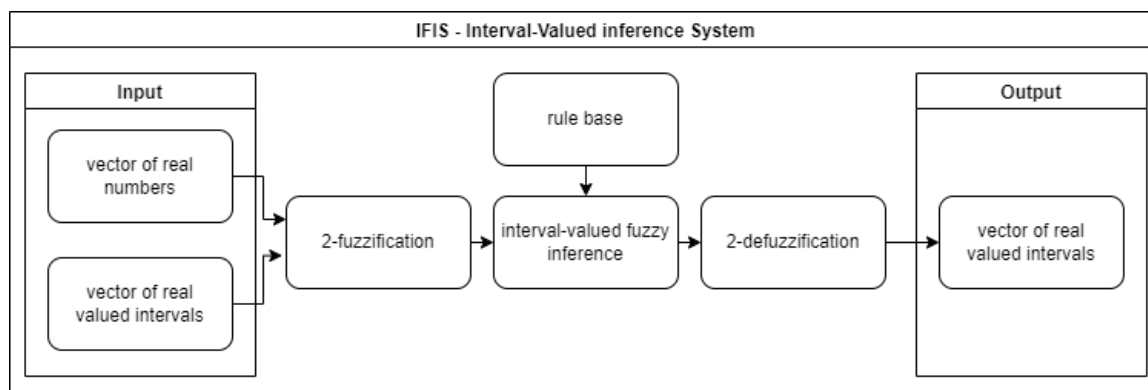
- bezpośrednią obsługę danych przedziałowych, w przypadku gdy dane wejściowe są podane jako przedziały, możliwe jest przedstawienie zbioru rozmytego za pomocą pojedynczej funkcji przynależności.

Ogólny schemat przedziałowego wnioskowania rozmytego przedstawiono w 6.1.2.

Konstrukcja systemu wnioskowania opiera się na rozszerzeniu przedziałowym biblioteki *Simpful* i obejmuje następujące kroki:

- definiowanie przedziałowych zbiorów rozmytych jako obiektów klasy rozszerzającej *FuzzySet*, z granicami dziedziny określonymi dla każdej zmiennej;
- tworzenie zmiennych lingwistycznych o wartościach przedziałowych na podstawie zdefiniowanych zbiorów rozmytych;
- definiowanie reguł wnioskowania, które mogą korzystać z operatorów logicznych AND, OR, NOT oraz przedziałowych funkcji agregacji;
- wykonanie wnioskowania (Mamdani lub Takagi–Sugeno) z podanymi wartościami zmiennych oraz, w przypadku Takagi–Sugeno, zdefiniowanie funkcji wyjściowych;
- defuzyfikacja wyników przy użyciu centroidu (Mamdani) lub średniej ważonej (Takagi–Sugeno).

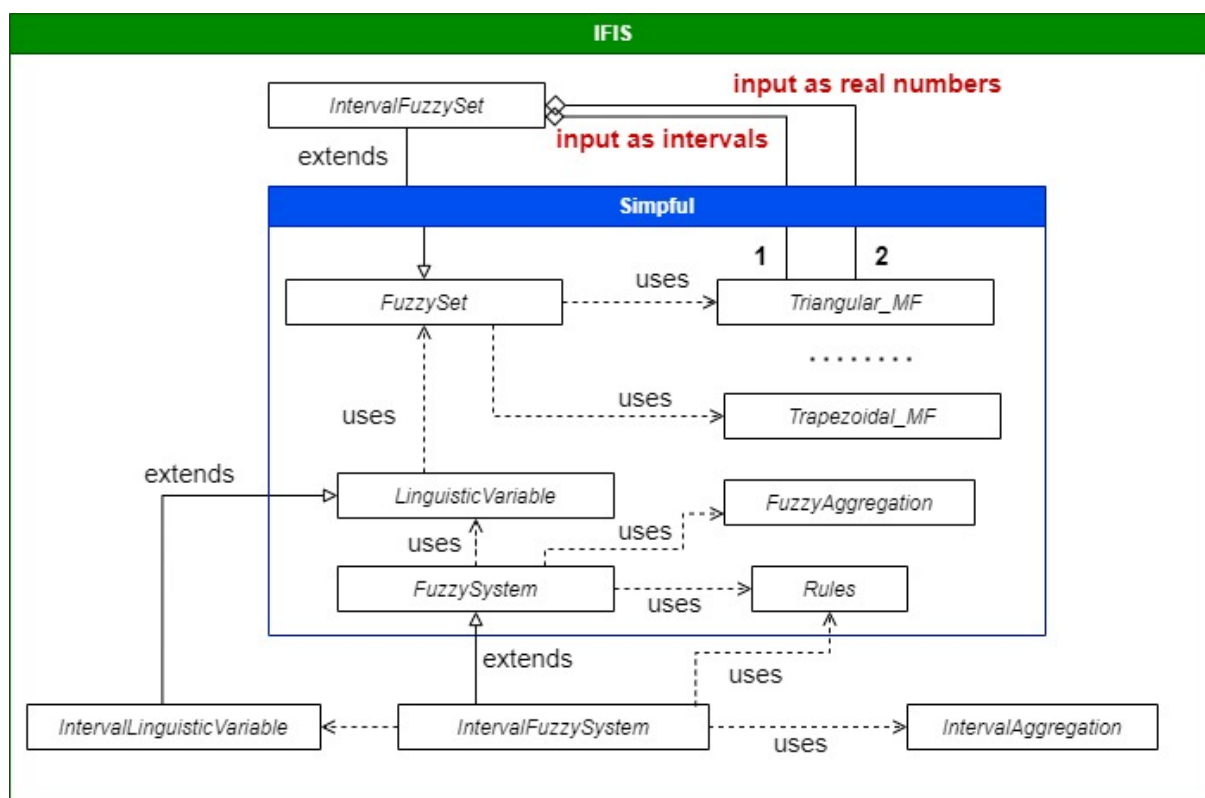
Struktura systemu wnioskowania została przedstawiona na rysunku 6.3.



Rysunek 6.3: IFIS (Interval-Valued Fuzzy Inference System)

Szczegółowy opis zaimplementowanych klas, metod oraz przykłady ich zastosowania, wraz z pełnym kodem źródłowym znajduje się w repozytorium GitHub dostępnym pod adresem: <https://github.com/PGrochowalski/ifis>.

Rysunek 6.4 przedstawia przegląd struktury klas w bibliotece IFIS, podkreślając rozszerzenia wprowadzone względem *Simpful*. Struktura klas w IFIS opiera się na bazowej strukturze *Simpful*, z dodatkowymi klasami, które umożliwiają obsługę wnioskowania rozmytego o wartościach przedziałowych.



Rysunek 6.4: Schemat oceny ryzyka zachorowania na CVD dla nowego pacjenta

Rozszerzenia w bibliotece wprowadzają nowe komponenty, takie jak funkcje przynależności o wartościach przedziałowych, reguły rozmyte z wartościami przedziałowymi oraz możliwość obsługi dowolnych agregacji przedziałowych. Te dodatki pozwalają na implementację wnioskowania rozmytego o wartościach przedziałowych zarówno za pomocą metod Mamdaniego, jak i Takagi-Sugeno.

Co istotne, w przypadku braku niepewności, wyniki uzyskiwane za pomocą IFIS są zgodne z wynikami wersji rozmytej biblioteki Simpful. IFIS jest narzędziem, które pozwala użytkownikom na definiowanie przedziałowych zbiorów rozmytych, reguł oraz przeprowadzanie uogólnionego wnioskowania rozmytego. Zamieszczone na repozytorium przykłady zastosowań demonstrują praktyczną użyteczność biblioteki, co czyni ją wartościowym narzędziem w oprogramowaniu typu open source.

ROZDZIAŁ 7

Podsumowanie i plany na przyszłość

W rozprawie doktorskiej wykazano, że uwzględnienie niepewności i nieprecyzyjności może znacząco polepszyć jakość działania klasyfikatorów. Wprowadzono modyfikację algorytmu k -NN, która bierze pod uwagę niepewność w danych poprzez zastosowanie teorii przedziałowych zbiorów rozmytych. Tego rodzaju reprezentacja danych wymagała rozważenia i opracowania różnego rodzaju operacji przedziałowych oraz miar, które uwzględniają niepewność. W części rozprawy poświęconej zagadnieniom teoretycznym (rozdziały 2 i 3) szczegółowo opisano wspomniane operacje oraz miary. Każda z miar była rozważana w trzech kategoriach: klasycznej, możliwości i konieczności, co spełniło założenia pierwszego z celów niniejszej pracy.

Opracowano modyfikację metody klasyfikacji uwzględniającą dane niepewne i nieprecyzyjne, opartą na algorytmie k -NN, czyli IV- k NN, w którym wykorzystano nowe miary podobieństwa. Efekty tych działań zostały zweryfikowane na przykładzie diagnostyki raka piersi w procesie klasyfikacji. W porównaniu ze standardowym, rozmytym algorytmem k -NN uzyskano lepsze wyniki, szczególnie przy zastosowaniu miary podobieństwa możliwości (optymistycznej). Można zaobserwować wzrost zarówno dokładności (accuracy), jak i precyzji (precision), przy jednoczesnym utrzymaniu dobrego poziomu czułości (sensitivity) oraz swoistości (specificity), co świadczy o stabilności poprawy wyników klasyfikacji. Zaobserwowano wzrost dokładności (95%) i precyzji (97%) o 1% w porównaniu do klasycznego podejścia opartego na miarach uwzględniających niepewność, a także o 2% względem klasyfikatora rozmytego. Co istotne, poprawa wyników została osiągnięta przy mniejszej liczbie podobnych obiektów, co ma duże znaczenie w analizie danych medycznych, gdzie liczba dostępnych danych bywa ograniczona. Szczegółowy opis tego zadania przedstawiono w rozdziale 5.1.

Opracowano nową bibliotekę open-source do wnioskowania przybliżonego opartego na wartościach przedziałowych, która została zastosowana do przewidywania ryzyka chorób sercowo-naczyniowych. Rozszerzając w rozprawie badania o analizy wpływu różnych klas

agregacji na proces wnioskowania przedstawiono wpływ klas możliwości i konieczności na proces diagnostyki ryzyka chorób sercowo-naczyniowych. Wyniki analiz wykazały, że agregacja możliwości (optymistyczna) charakteryzowała się najniższym poziomem błędu dla wysokiego ryzyka zachorowania na choroby sercowo-naczyniowe. Szczegóły tego zagadnienia przedstawiono w rozdziale 6.1. W związku z uzyskanymi wynikami, można stwierdzić, że z powodzeniem zrealizowano drugi z założonych celów badawczych.

Zaproponowano efektywną metodę selekcji atrybutów, wykorzystując miary uwzględniające niepewność i nieprecyzyjność, zweryfikowaną w ujęciu klasycznym, możliwości (optymistycznym) oraz konieczności (pesymistycznym). Proces ten obejmował analizę danych z diagnostyki raka piersi z zastosowaniem algorytmu IVRelief, szczegółowo omówionego w rozdziale 5.3. Testy potwierdziły wysoką skuteczność algorytmu IVRelief w określaniu wag atrybutów. Po weryfikacji trafności redukcji atrybutów stwierdzono, że wykorzystanie miary podobieństwa konieczności przyniosło najlepszy poziom dokładności klasyfikacji (98,2%), co oznacza wzrost o około 2% w porównaniu do klasyfikacji opartej na pełnym zestawie atrybutów. W związku z tym zrealizowano trzeci cel pracy, potwierdzając skuteczność zaproponowanej metody selekcji atrybutów z uwzględnieniem niepewności, co przyczyniło się do zwiększenia dokładności klasyfikacji w analizowanych danych.

Zaproponowano również efektywną metodę uzupełniania brakujących wartości z wykorzystaniem miary entropii w ujęciu klasycznym, możliwości (optymistycznym) oraz konieczności (pesymistycznym). Przeprowadzono testy algorytmu dla trzech poziomów brakujących danych, w których 5%, 25% i 50% obiektów ze zbioru danych posiadało brakujące wartości na jednym z atrybutów. Badania obejmowały dwa scenariusze: pierwszy porównywał klasyczną miarę entropii z literatury do nowo zaproponowanej entropii, a drugi oceniał wyniki klasyfikacji przy zastosowaniu entropii we wszystkich trzech różnych ujęciach (klasycznym, możliwości i konieczności). W pierwszym scenariuszu najlepsze wyniki uzyskano dla zaproponowanej nowej entropii w ujęciu klasycznym, osiągając dokładność klasyfikacji 88,7%, 88,0% oraz 87,7% dla odpowiednio 5%, 25% i 50% brakujących danych. W drugim scenariuszu entropia możliwości okazała się najskuteczniejsza, dając wynik 89,4% dla próbki z 5% wartości brakujących (przy zachowaniu dobrych wartości dokładności dla 25 i 50%). Szczegóły tego badania znajdują się w rozdziale 5.2. W związku z osiągniętymi rezultatami z powodzeniem zrealizowano cel czwarty rozprawy.

Zaproponowany i zweryfikowany algorytm IV-kNN (rozdział 5.1) stanowił fundament dla weryfikacji poprawności działania innych algorytmów, które były stosowane w kontekście uzupełniania brakujących wartości, selekcji atrybutów oraz prewencyjnego wykrywania epizodów depresyjnych. W kontekście wczesnego wykrywania epizodów depresyjnych przeprowadzono badania mające na celu ocenę dokładności klasyfikacji w zależności od zastosowanej klasy miar podobieństwa (klasycznej, możliwości lub konieczności) oraz we-

ryfikację zaproponowanej diagnozy, mającej na celu identyfikację pacjentów z ryzykiem wystąpienia depresji w niedalekiej przyszłości. Wspólnie z lekarzem psychiatrą opracowano czulszy test diagnostyczny (SenDD), który okazał się bardziej skuteczny w porównaniu do powszechnie znanych testów stosowanych w praktyce lekarskiej. Najlepsze wyniki uzyskano dla miar podobieństwa możliwości (optymistycznej), a dodatkowo wykazano, że test SenDD wraz z autorskim systemem klasyfikacji prewencji depresji (alg. 4-5), stosowany z czterotygodniowym wyprzedzeniem, skutecznie zidentyfikował około 20% więcej pacjentów z ryzykiem depresji. Więcej szczegółów na temat wykrywania epizodów depresyjnych można znaleźć w rozdziale 5.4.

W związku z powyższymi osiągnięciami można stwierdzić, że wszystkie cele postawione w rozprawie zostały osiągnięte. Realizacja ww. celów stanowi potwierdzenie tezy postawionej w niniejszej rozprawie, która brzmi: uwzględnienie niepewności i nieprecyzyjności może znacząco polepszyć jakość działania klasyfikatorów opartych na idei k-najbliższych sąsiadów, nawet w warunkach danych niekompletnych lub nadmiarowych. Każdy etap badawczy, poczynając od opracowania metod reprezentacji danych z niepewnością, aż po implementację algorytmów, wykazał, że integracja niepewności w analizach ma kluczowe znaczenie dla poprawy jakości klasyfikacji i wnioskowania. Wyniki eksperymentów wskazują, że proponowane podejścia skutecznie radzą sobie z problemami związanymi z niepewnością i nieprecyzyjnością danych.

W przyszłości planowane jest rozwinięcie omawianych zagadnień oraz rozszerzenie ich zastosowań o ideę uczenia federacyjnego [139]. Uczenie federacyjne to innowacyjne podejście w dziedzinie uczenia maszynowego, umożliwiające trenowanie globalnych modeli na zdecentralizowanych, rozproszonych danych, bez konieczności i możliwości ich udostępniania kooperującym jednostkom. Jest to szczególnie istotne w kontekście ochrony prywatności i bezpieczeństwa danych, zwłaszcza w przypadku informacji medycznych.

Trwają prace nad stworzeniem platformy online, która rozwija proponowany w pracy doktorskiej model prewencji depresji, uwzględniając problem dostępności danych oraz dostosowując ją do potrzeb instytucji takich jak szkoły. Platforma ta ma na celu umożliwienie wykorzystania narzędzi diagnostycznych w rzeczywistym środowisku, co pozwoli na skuteczną identyfikację i wsparcie osób potrzebujących pomocy. Dzięki intuicyjnemu interfejsowi i łatwej dostępności użytkownicy będą mogli korzystać z funkcji platformy w codziennej pracy, przyczyniając się do poprawy zdrowia psychicznego w ich społecznościach.

Jak już wspomniano, we współpracy z zespołem, powstała biblioteka dedykowana wnioskowaniu przybliżonemu obsługującemu niepewność, czyli rachunek przedziałowych zbiorów rozmytych. Pierwsze efekty tych prac zostały przedstawione w artykułach [54, 55], a pierwszą wersję biblioteki można znaleźć w repozytorium GitHub pod adresem: <https://github.com/PGrochowalski/ifis>. Obecnie pracujemy nad rozwojem biblioteki w kie-

runku uwzględnienia uczenia federacyjnego. Ten krok umożliwi jeszcze szersze zastosowanie opisanych metod w różnych dziedzinach, zachowując jednocześnie wysoki poziom ochrony danych.

Paradygmat uczenia federacyjnego ma kluczowe znaczenie nie tylko w aspekcie ochrony danych, ale również w pozyskiwaniu informacji dotyczących rzadkich, specjalistycznych danych związanych z rzadkimi schorzeniami. W sytuacjach, w których częstość występowania tych chorób jest niska, zbieranie dostatecznej ilości danych może stanowić wyzwanie. Uczenie federacyjne stanowi rozwiązanie, które umożliwia współpracę między posiadaczami tych danych bez konieczności ich jawnej wymiany. Dzięki temu można skutecznie analizować i diagnozować rzadkie schorzenia, co ma kluczowe znaczenie dla poprawy opieki zdrowotnej w tych obszarach.

Zatem rozszerzenie omawianych zagadnień związanych z reprezentacją i modelowaniem niepewności o uczenie federacyjne będzie głównym przedmiotem dalszych badań.

Skorowidz

- agregacja (*ang. aggregation*), 23
- agregacja: klasyczna (*ang. classic aggregation*), 24
- agregacja: konieczności (*ang. necessary aggregation*), 28
- agregacja: możliwości (*ang. possible aggregation*), 26
- czułość klasyfikacji (*ang. sensitivity*), 67
- entropia: klasyczna (*ang. classic entropy measure*), 49
- entropia: konieczności (*ang. necessary entropy measure*), 54
- entropia: możliwości (*ang. possible entropy measure*), 52
- funkcja ignorancji (*ang. ignorance function*), 72
- inkluzja (*ang. inclusion*), 42
- Metryka Euklidesa, 59
- Metryka Manhattan, 59
- miara podobieństwa: klasyczna (*ang. classic similarity measure*), 44
- miara podobieństwa: konieczności (*ang. necessary similarity measure*), 47
- miara podobieństwa: możliwości (*ang. possible similarity measure*), 47
- negacja: klasyczna (*ang. classic negation*), 31
- negacja: konieczności (*ang. necessary negation*), 33
- negacja: możliwości (*ang. possible negation*), 33
- operacja: iloczyn (*ang. meet*), 20
- operacja: maksimum (*ang. maximum*), 23
- operacja: minimum (*ang. minimum*), 23
- operacja: suma (*ang. join*), 20
- operator: IVOWA (*ang. Interval-Valued Ordered Weighted Averaging Operator*), 25
- operator: słaby W (*ang. weak W operator*), 30
- pokrycie reguły *ang. rule coverage*, 62
- porządek: dopuszczalny (*ang. admissible order*), 20
- porządek: liniowy (*ang. linear order*), 20
- pre-agregacja (*ang. pre-aggregation*), 29
- precyzja klasyfikacji (*ang. precision*), 67
- przeciwwrotność (*ang. irreflexivity*), 37
- przedziałowy zbiór rozmyty (*ang. Interval-valued fuzzy set*), 36
- relacja: konieczności (*ang. necessary relation*), 22

-
- relacja: możliwości (*ang. possible relation*), 22
- specyficzność klasyfikacji (*ang. specificity*), 67
- symetria (*ang. symmetry*), 37
- trafność klasyfikacji (*ang. accuracy*), 66
- trafność reguły *ang. rule accuracy*, 62
- wskaźnik pierwszeństwa: klasyczny (*ang. classic precedence indicator*), 40
- wskaźnik pierwszeństwa: konieczności (*ang. neccessary precedence indicator*), 43
- wskaźnik pierwszeństwa: możliwości (*ang. possible precedence indicator*), 42
- zwrotność (*ang. refrefxivity*), 37

Bibliografia

- [1] L. A. Zadeh, “Fuzzy Sets,” *Information and Control*, vol. 8, pp. 338–353, 1965.
- [2] J. A. Goguen, “L-fuzzy sets,” *Journal of Mathematical Analysis and Applications*, vol. 18, no. 1, pp. 145–174, 1967.
- [3] K. T. Atanassov, “Intuitionistic fuzzy sets,” *Fuzzy Sets and Systems*, vol. 20, no. 1, pp. 87–96, 1986.
- [4] H. Bustince, E. Barrenechea, M. Pagola, J. Fernandez, Z. Xu, B. Bedregal, J. Montero, H. Hagraas, F. Herrera, and B. D. Baets, “A Historical Account of Types of Fuzzy Sets and Their Relationships,” *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 1, pp. 179–194, 2016.
- [5] L. Zadeh, “The concept of a linguistic variable and its application to approximate reasoning—I,” *Information Sciences*, vol. 8, no. 3, pp. 199–249, 1975.
- [6] R. Sambuc, “Fonctions ϕ -floues: Application à l’aide au diagnostic en pathologie thyroïdienne,” *Université de Marseille*, 1975.
- [7] G. Deschrijver and E. E. Kerre, “Implicators based on binary aggregation operators in interval-valued fuzzy set theory,” *Fuzzy Sets and Systems*, vol. 153, no. 2, pp. 229–248, 2005.
- [8] B. Baets, S. Janssens, and H. Meyer, “On the transitivity of a parametric family of cardinality-based similarity measures,” *Int. J. Approx. Reason.*, vol. 50, pp. 104–116, 2009.
- [9] E. Barrenechea, J. Fernandez, M. Pagola, F. Chiclana, and H. Bustince, “Construction of interval-valued fuzzy preference relations from ignorance functions and fuzzy preference relations. Application to decision making,” *Knowledge-Based Systems*, vol. 58, pp. 33–44, 2014.
- [10] K. Bosteels and E. E. Kerre, “On a reflexivity-preserving family of cardinality-based fuzzy comparison measures,” *Information Sciences*, vol. 179, no. 14, pp. 2342–2352, 2009.

- [11] B. Jayaram and R. Mesiar, "I-Fuzzy equivalence relations and I-fuzzy partitions," *Information Sciences*, vol. 179, no. 9, pp. 1278–1297, 2009.
- [12] E. S. Palmeira, B. Bedregal, H. Bustince, D. Paternain, and L. D. Miguel, "Application of Two Different Methods for Extending Lattice-Valued Restricted Equivalence Functions Used for Constructing Similarity Measures on L-Fuzzy Sets," *Information Sciences*, vol. 441, p. 95–112, 2018.
- [13] B. D. Baets, H. D. Meyer, and H. Naessens, "A class of rational cardinality-based similarity measures," *Journal of Computational and Applied Mathematics*, vol. 132, no. 1, pp. 51–69, 2001.
- [14] Z. Xu and R. Yager, "Intuitionistic and interval-valued intuitionistic fuzzy preference relations and their measures of similarity for the evaluation of agreement within a group," *Fuzzy Optim. Decis. Mak.*, vol. 8, pp. 123–139, 06 2009.
- [15] P. Żywica, "Modelling medical uncertainties with use of fuzzy sets and their extensions," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications*, pp. 369–380, Springer International Publishing, 2018.
- [16] P. Żywica, K. Dyczkowski, A. Wójtowicz, A. Stachowiak, S. Szuber, and R. Mośzyński, "Development of a fuzzy-driven system for ovarian tumor diagnosis," *Bio-cybernetics and Biomedical Engineering*, vol. 36, no. 4, pp. 632–643, 2016.
- [17] U. Bentkowska, "New types of aggregation functions for interval-valued fuzzy setting and preservation of pos-B and nec-B-transitivity in decision making problems," *Information Sciences*, vol. 424, pp. 385–399, 2018.
- [18] U. Bentkowska, H. Bustince, A. Jurio, M. Pagola, and B. Pękala, "Decision making with an interval-valued fuzzy preference relation and admissible orders," *Applied Soft Computing*, vol. 35, pp. 792–801, 2015.
- [19] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decisionmaking," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 18, no. 1, pp. 183–190, 1988.
- [20] J. A. Sanz, H. Bustince, A. Fernández, and F. Herrera, "IIVFDT: Ignorance functions based interval-valued fuzzy decision tree with genetic tuning," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 20, pp. 1–30, 10 2012.
- [21] G. li Xu and F. Liu, "An approach to group decision making based on interval multiplicative and fuzzy preference relations by using projection," *Applied Mathematical Modelling*, vol. 37, no. 6, pp. 3929–3943, 2013.

- [22] F. Liu, W.-G. Zhang, and L.-H. Zhang, “A group decision making model based on a generalized ordered weighted geometric average operator with interval preference matrices,” *Fuzzy Sets and Systems*, vol. 246, pp. 1–18, 2014.
- [23] H. Chen and L. Zhou, “An approach to group decision making with interval fuzzy preference relations based on induced generalized continuous ordered weighted averaging operator,” *Expert Systems with Applications*, vol. 38, no. 10, pp. 13432–13440, 2011.
- [24] K. Dyczkowski, *Intelligent Medical Decision Support System Based on Imperfect Information. The Case of Ovarian Tumor Diagnosis*, vol. 735. Springer, 2018.
- [25] D. Paternain, A. Jurio, E. Barrenechea, H. Bustince, B. Bedregal, and E. Szmidt, “An alternative to fuzzy methods in decision-making problems,” *Expert Systems with Applications*, vol. 39, no. 9, pp. 7729–7735, 2012.
- [26] A. Dziech and M. B. Gorzałczany, “Decision Making in Signal Transmission Problems with Interval-Valued Fuzzy Sets,” *Fuzzy Sets and Systems*, vol. 23, p. 191–203, Aug. 1987.
- [27] Z. Xu, “Intuitionistic preference relations and their application in group decision making,” *Information sciences*, vol. 177, no. 11, pp. 2363–2379, 2007.
- [28] L. J. Kohout and W. Bandler, *Fuzzy Interval Inference Utilizing the Checklist Paradigm and BK-Relational Products*, pp. 291–335. Springer, 1996.
- [29] M. Maadia, U. Aickelin, and H. A. Khorshidi, “A new interval-based aggregation approach based on bagging and Interval Agreement Approach (IAA) in ensemble learning,” *arXiv preprint arXiv:2101.10267*, 2020.
- [30] P. Singh and Y.-P. Huang, “A four-way decision-making approach using interval-valued fuzzy sets, rough set and granular computing: a new approach in data classification and decision-making,” *Granular Computing*, vol. 5, pp. 397–409, 2020.
- [31] J. Fumanal-Idocin, Z. Takáč, L. Horanská, H. Bustince, and O. Cordon, “Fuzzy Clustering to Encode Contextual Information in Artistic Image Classification,” in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 355–366, Springer, 2022.
- [32] J. A. Sanz and H. Bustince, “A wrapper methodology to learn interval-valued fuzzy rule-based classification systems,” *Applied Soft Computing*, vol. 104, p. 107249, 2021.

- [33] S. Zeraatkar and F. Afsari, “Interval-valued fuzzy and intuitionistic fuzzy-KNN for imbalanced data classification,” *Expert Systems with Applications*, vol. 184, p. 115510, 2021.
- [34] W. Shu, W. Qian, Y. Xie, and Z. Tang, “An efficient uncertainty measure-based attribute reduction approach for interval-valued data with missing values,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 27, no. 06, pp. 931–947, 2019.
- [35] T. da Cruz Asmus, J. A. Sanz, G. P. Dimuro, B. Bedregal, J. Fernández, and H. Bustince, “N-Dimensional admissibly ordered interval-valued overlap functions and its influence in interval-valued fuzzy-rule-based classification systems,” *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 4, pp. 1060–1072, 2021.
- [36] J. Dai, Z. Wang, and W. Huang, “Interval-valued fuzzy discernibility pair approach for attribute reduction in incomplete interval-valued information systems,” *Information Sciences*, vol. 642, p. 119215, 2023.
- [37] “UCI Machine Learning Repository.” <https://archive.ics.uci.edu/datasets>.
- [38] B. Pękala, K. Garwol, J. Czuma, L. Zaręba, M. Chyła, and D. Kosior, “Early detection of the risk of depressive episodes using a proprietary diagnostic test by new epistemic similarity measures,” *Applied Soft Computing*, vol. 148, p. 110910, 2023.
- [39] G. P. Chander and S. Das, “Multi-attribute decision making using interval-valued pythagorean fuzzy set and differential evolutionary algorithm,” in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6, 2021.
- [40] X. Ma, Y. Han, H. Qin, and P. Wang, “KNN Data Filling Algorithm for Incomplete Interval-Valued Fuzzy Soft Sets,” *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, p. 30, 2023.
- [41] X. Zhang and J. Li, “Incremental feature selection approach to interval-valued fuzzy decision information systems based on λ -fuzzy similarity self-information,” *Information Sciences*, vol. 625, pp. 593–619, 2023.
- [42] D. Zhao, X. Hu, S. Xiong, J. Tian, J. Xiang, J. Zhou, and H. Li, “k-means clustering and kNN classification based on negative databases,” *Applied Soft Computing*, vol. 110, p. 107732, 2021.
- [43] İ. Atacak, O. Çıtlak, and İ. A. Doğru, “Application of interval type-2 fuzzy logic and type-1 fuzzy logic-based approaches to social networks for spam detection with combined feature capabilities,” *PeerJ Computer Science*, vol. 9, 2023.

- [44] Y. Pan, L. Zhang, Z. Li, and L. Ding, "Improved Fuzzy Bayesian Network-Based Risk Analysis With Interval-Valued Fuzzy Sets and D-S Evidence Theory," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 9, pp. 2063–2077, 2020.
- [45] J. Chen, Z. Li, X. Wang, and J. Zhai, "A hybrid monotone decision tree model for interval-valued attributes," *Advances in Computational Intelligence*, vol. 2, no. 1, p. 12, 2022.
- [46] B. Pękala, E. Rak, D. Kosior, M. Mrukowicz, and J. G. Bazan, "Application of similarity measures with uncertainty in classification methods," *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8, 2020.
- [47] D. Kosior and B. Pękala, "Influence of Interval-Valued Measures on Classification Methods with Missing Values," in *Uncertainty and Imprecision in Decision Making and Decision Support: New Advances, Challenges, and Perspectives* (K. T. Atanassov, V. Atanassova, J. Kacprzyk, A. Kałuszko, M. Krawczak, J. W. Owsinski, S. S. Sotirov, E. Sotirova, E. Szmidt, and S. Zadrozny, eds.), (Cham), pp. 15–27, Springer International Publishing, 2022.
- [48] B. Pękala, D. Kosior, K. Dyczkowski, and J. Szkoła, "Application of entropy measures with uncertainty in classification methods with missing data problem," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8, IEEE, 2021.
- [49] P. Grzegorzewski, B. Pękala, K. Dyczkowski, and D. Kosior, "A New Look at the Entropy of Interval-Valued Fuzzy Sets - Theory and Applications," in *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, pp. 1–7, 2023.
- [50] B. Pękala, K. Dyczkowski, J. Szkoła, and D. Kosior, "Classification of uncertain data with a selection of relevant features based on similarities measures of Interval-Valued Fuzzy Sets," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8, 2021.
- [51] B. Pękala, K. Dyczkowski, J. Szkoła, and D. Kosior, "Selection of Relevant Features Based on Optimistic and Pessimistic Similarities Measures of Interval-Valued Fuzzy Sets," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems* (D. Ciucci, I. Couso, J. Medina, D. Ślęzak, D. Petturiti, B. Bouchon-Meunier, and R. R. Yager, eds.), (Cham), pp. 307–319, Springer International Publishing, 2022.
- [52] P. Drygaś, B. Pękala, K. Balicki, and D. Kosior, "Influence of new interval-valued pre-aggregation function on medical decision making," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8, 2020.

- [53] M. Elkano, J. A. Sanz, M. Galar, B. Pękala, U. Bentkowska, and H. Bustince, “Composition of interval-valued fuzzy relations using aggregation functions,” *Information Sciences*, vol. 369, pp. 690–703, 2016.
- [54] K. Dyczkowski, P. Grochowalski, D. Kosior, D. Gil, W. Koziół, B. Pękala, U. Kaymak, C. Fuchs, and M. S. Nobile, “Python library for interval-valued fuzzy inference,” *SoftwareX*, vol. 26, p. 101730, 2024.
- [55] B. Pękala, P. Grochowalski, D. Kosior, D. Gil, W. Koziół, K. Dyczkowski, U. Kaymak, C. Fuchs, and M. S. Nobile, “Applications of IFIS python library in interval-valued fuzzy inference problems,” in *2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–9, IEEE, 2024.
- [56] E. H. Mamdani and S. Assilian, “An experiment in linguistic synthesis with a fuzzy logic controller,” *International Journal of Man-Machine Studies*, vol. 7, p. 1–13, 1975.
- [57] J. M. Mendel, “Fuzzy logic systems for engineering: a tutorial,” *Proceedings of the IEEE*, vol. 83, pp. 345 – 377, 1995.
- [58] J. A. Momoh, X. W. Ma, and K. Tomsovic, “Overview and Literature Survey of Fuzzy Set Theory in Power Systems,” *IEEE Transaction on Power Systems*, vol. 10, pp. 1676 – 1690, 1995.
- [59] E. P. Klement, R. Mesiar, and E. Pap, *Triangular Norms*. Dordrecht: Kluwer Academic Publishing, 2000.
- [60] J. Fodor and M. Roubens, *Fuzzy Preference Modelling and Multicriteria Decision Support*. Dordrecht: Kluwer Academic Publishing, 1994.
- [61] G. J. Klir and B. Yuan, “Fuzzy sets and fuzzy logic: theory and applications,” *Possibility Theory versus Probab. Theory*, vol. 32, no. 2, pp. 207–208, 1996.
- [62] A. Pradera, G. Beliakov, H. Bustince, and B. D. Baets, “A review of the relationships between implication, negation and aggregation functions from the point of view of material implication,” *Information Sciences*, p. 357–380, 2016.
- [63] H. Bustince, J. Fernandez, A. Kolesárová, and R. Mesiar, “Generation of linear orders for intervals by means of aggregation functions,” *Fuzzy Sets and Systems*, vol. 220, pp. 69–77, 2013.
- [64] M. Asiain, H. Bustince, B. Bedregal, Z. Takáč, M. Baczyński, D. Paternain, and G. P. Dimuro, “About the Use of Admissible Order for Defining Implication Operators,” in *Modeling Decisions for Artificial Intelligence*, (Cham), pp. 126–134, Springer International Publishing, 2016.

- [65] H. Zapata, H. Bustince, S. Montes, B. Bedregal, G. P. Dimuro, Z. Takáč, M. Baczynski, and J. Fernandez, “Interval-valued implications and interval-valued strong equality index with admissible orders,” *International Journal of Approximate Reasoning*, vol. 88, pp. 91–109, 2017.
- [66] H. Bustince, M. Galar, B. Bedregal, A. Kolesárová, and R. Mesiar, “A New Approach to Interval-Valued Choquet Integrals and the Problem of Ordering in Interval-Valued Fuzzy Set Applications,” *IEEE Transactions on Fuzzy Systems*, vol. 21, no. 6, pp. 1150–1162, 2013.
- [67] M. Grabisch, J. Marichal, R. Mesiar, and E. Pap, *Aggregation Functions*. Encyclopedia of Mathematics and its Applications, Cambridge: Cambridge University Press, 2009.
- [68] D. Dubois and H. Prade, “Gradualness, uncertainty and bipolarity: Making sense of fuzzy sets,” *Fuzzy Sets and Systems*, vol. 192, pp. 3–24, 2012.
- [69] D. Dubois, L. Godo, and H. Prade, “Weighted logics for artificial intelligence – an introductory discussion,” *International Journal of Approximate Reasoning*, vol. 55, no. 9, pp. 1819–1829, 2014.
- [70] B. Pękala, U. Bentkowska, and B. D. Baets, “On Comparability Relations in the Class of Interval-Valued Fuzzy Relations,” *Tatra Mountains Mathematical Publications*, vol. 66, no. 1, pp. 91–101, 2017.
- [71] M. Komorníková and R. Mesiar, “Aggregation functions on bounded partially ordered sets and their classification,” *Fuzzy Sets and Systems*, vol. 175, no. 1, pp. 48–56, 2011.
- [72] G. Deschrijver, “Quasi-arithmetic means and OWA functions in interval-valued and Atanassov’s intuitionistic fuzzy set theory,” in *Advances in Intelligent Systems Research* (Galichet, S and Montero, J and Mauris, G, ed.), pp. 506–513, Atlantis Press, 2011.
- [73] B. Pękala, *Uncertainty data in interval-valued fuzzy set theory. Properties, Algorithms and Applications*. Springer International Publishing, 2019.
- [74] S. M. Zhou, F. Chiclana, R. I. John, and J. M. Garibaldi, “Type-1 OWA operators for aggregating uncertain information with uncertain weights induced by type-2 linguistic quantifiers,” *Fuzzy Sets and Systems*, vol. 159, no. 24, pp. 3281–3296, 2008.

- [75] M. Sesma-Sara, L. D. Miguel, R. Mesiar, J. Fernandez, and H. Bustince, "Interval-valued pre-aggregation functions: a study of directional monotonicity of interval-valued functions," in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6, 2019.
- [76] M. J. Asiain, H. Bustince, R. Mesiar, A. Kolesárová, and Z. Takáč, "Negations With Respect to Admissible Orders in the Interval-Valued Fuzzy Set Theory," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 2, pp. 556–568, 2018.
- [77] B. C. Bedregal, G. P. Dimuro, R. H. N. Santiago, and R. H. S. Reiser, "On Interval Fuzzy S-Implications," *Inf. Sci.*, vol. 180, p. 1373–1389, apr 2010.
- [78] G. Deschrijver and C. Cornelis, "Representability in interval-valued fuzzy set theory," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 15, no. 03, pp. 345–361, 2007.
- [79] M. B. Gorzalczany, "A method of inference in approximate reasoning based on interval-valued fuzzy sets," *Fuzzy Sets and Systems*, vol. 21, no. 1, pp. 1–17, 1987.
- [80] M. B. Gorzalczany, "Interval-valued fuzzy inference involving uncertain (inconsistent) conditional propositions," *Fuzzy Sets and Systems*, vol. 29, no. 2, pp. 235–240, 1989.
- [81] M. B. Gorzalczany, "An interval-valued fuzzy inference method — Some basic properties," *Fuzzy Sets and Systems*, vol. 31, no. 2, pp. 243–251, 1989.
- [82] I. B. Turksen, "Interval valued fuzzy sets based on normal forms," *Fuzzy Sets and Systems*, vol. 20, no. 2, pp. 191–210, 1986.
- [83] I. B. Turksen and Z. Zhong, "An Approximate Analogical Reasoning Schema Based on Similarity Measures and Interval-Valued Fuzzy Sets," *Fuzzy Sets and Systems*, vol. 34, p. 323–346, Feb. 1990.
- [84] A. Bigand and O. Colot, "Fuzzy filter based on interval-valued fuzzy sets for image filtering," *Fuzzy Sets and Systems*, vol. 161, pp. 96–117, 2010.
- [85] H. Bustince, E. Barrenechea, M. Pagola, and J. Fernandez, "Interval-valued fuzzy sets constructed from matrices: Application to edge detection," *Fuzzy Sets and Systems*, vol. 160, no. 13, pp. 1819–1840, 2009.
- [86] H. Bustince, "Indicator of inclusion grade for interval-valued fuzzy sets. Application to approximate reasoning based on interval-valued fuzzy sets," *International Journal of Approximate Reasoning*, vol. 23, no. 3, pp. 137–209, 2000.

- [87] B. Pękala, “Preservation of properties of interval-valued fuzzy relations,” *Carvalho, J. P. et al. (eds.) IFSA/EUSFLAT*, p. 1206–1210, 2009.
- [88] H. Bustince and P. Burillo, “Mathematical analysis of interval-valued fuzzy relations: Application to approximate reasoning,” *Fuzzy Sets and Systems*, vol. 113, no. 2, pp. 205–219, 2000.
- [89] Z. Takáč, M. Minárová, J. Montero, E. Barrenechea, J. Fernandez, and H. Bustince, “Interval-valued fuzzy strong S-subsethood measures, interval-entropy and P-interval-entropy,” *Information Sciences*, vol. 432, pp. 97–115, 2018.
- [90] W. Y. Zeng and P. Guo, “Normalized distance, similarity measure, inclusion measure and entropy of interval-valued fuzzy sets and their relationship,” *Information Sciences*, vol. 178, no. 5, pp. 1334–1342, 2008.
- [91] B. Pękala, K. Dyczkowski, P. Grzegorzewski, and U. Bentkowska, “Inclusion and similarity measures for interval-valued fuzzy sets based on aggregation and uncertainty assessment,” *Information Sciences*, vol. 547, pp. 1182–1200, 2021.
- [92] B. Pękala, U. Bentkowska, M. Sesma-Sara, J. Fernandez, J. Lafuente, A. Altalhi, M. Knap, H. Bustince, and J. M. Pintor, “Interval Subsethood Measures with Respect to Uncertainty for the Interval-Valued Fuzzy Setting,” *International Journal of Computational Intelligence Systems*, vol. 3, pp. 167–177, 2020.
- [93] H. Bustince, C. Marco-Detchart, J. Fernandez, C. Wagner, J. Garibaldi, and Z. Takáč, “Similarity between interval-valued fuzzy sets taking into account the width of the intervals and admissible orders,” *Fuzzy Sets and Systems*, vol. 390, pp. 23–47, 2020.
- [94] R. Clausius, *The mechanical theory of heat*. Macmillan, 1879.
- [95] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [96] Z. Takáč, H. Bustince, J. M. Pintor, C. Marco-Detchart, and I. Couso, “Width-Based Interval-Valued Distances and Fuzzy Entropies,” *IEEE Access*, vol. 7, pp. 14044–14057, 2019.
- [97] P. Burillo and H. Bustince, “Entropy on intuitionistic fuzzy sets and on interval-valued fuzzy sets,” *Fuzzy Sets and Systems*, vol. 78, no. 3, pp. 305–316, 1996.
- [98] E. Szmidt and J. Kacprzyk, “Entropy for intuitionistic fuzzy sets,” *Fuzzy Sets and Systems*, vol. 118, no. 3, pp. 467–477, 2001.

-
- [99] T. Morzy, *Eksploracja danych. Metody i algorytmy*. Wydawnictwo Naukowe PWN, 2013.
- [100] L. Bottou and V. Vapnik, “Local learning algorithms,” *Neural computation*, vol. 4, no. 6, pp. 888–900, 1992.
- [101] B. V. Dasarathy, “Nearest neighbor (NN) norms: NN pattern classification techniques,” *IEEE Computer Society Tutorial*, 1991.
- [102] S. M. Weiss and N. Indurkha, “Rule-based machine learning methods for functional prediction,” *Journal of Artificial Intelligence Research*, vol. 3, pp. 383–403, 1995.
- [103] L. Breiman, *Classification and Regression Trees (1st ed.)*. New York: Routledge, 1984.
- [104] J. R. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, pp. 81–106, Mar 1986.
- [105] S. L. Salzberg, “C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993,” *Machine Learning*, vol. 16, pp. 235–240, Sep 1994.
- [106] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [107] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, “Top 10 algorithms in data mining,” *Knowledge and Information Systems*, vol. 14, pp. 1–37, 2008.
- [108] G. Shakhnarovich, T. Darrell, and P. Indyk, “Nearest-Neighbor Methods in Learning and Vision,” *IEEE Transactions on Neural Networks*, vol. 19, p. 377, 01 2008.
- [109] G. Bhattacharya, K. Ghosh, and A. S. Chowdhury, “An affinity-based new local distance function and similarity measure for kNN algorithm,” *Pattern Recognition Letters*, vol. 33, no. 3, pp. 356–363, 2012.
- [110] H. Bustince, M. Pagola, E. Barrenechea, J. Fernandez, P. Melo-Pinto, P. Couto, H. Tizhoosh, and J. Montero, “Ignorance functions. An application to the calculation of the threshold in prostate ultrasound images,” *Fuzzy Sets and Systems*, vol. 161, no. 1, pp. 20–36, 2010.

- [111] E. Barrenechea, H. Bustince, B. De Baets, and C. Lopez-Molina, "Construction of interval-valued fuzzy relations with application to the generation of fuzzy edge images," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 5, pp. 819–830, 2011.
- [112] J. A. Sanz, A. Fernández, H. Bustince, and F. Herrera, "Improving the performance of fuzzy rule-based classification systems with interval-valued fuzzy sets and genetic amplitude tuning," *Information Sciences*, vol. 180, no. 19, pp. 3674–3685, 2010.
- [113] H. Qin and X. Ma, "Data analysis approaches of interval-valued fuzzy soft sets under incomplete information," *IEEE Access*, vol. 7, pp. 3561–3571, 2018.
- [114] T. D. Waal, J. Pannekoek, and S. Scholtus, "Handbook of statistical data editing and imputation," *John Wiley Sons*, vol. 563, 2011.
- [115] C. M. Salgado, C. Azevedo, H. Proença, and S. M. Vieira, "Missing data," *Secondary analysis of electronic health records*, pp. 143–162, 2016.
- [116] R. J. Little, J. R. Carpenter, and K. J. Lee, "A Comparison of Three Popular Methods for Handling Missing Data: Complete-Case Analysis," *Sociological Methods and Research*, no. 53(3), pp. 1105–1135, 2024.
- [117] A. Wójtowicz, P. Żywica, A. Stachowiak, and K. Dyczkowski, "Solving the problem of incomplete data in medical diagnosis via interval modeling," *Applied Soft Computing*, vol. 47, pp. 424–437, 2016.
- [118] K. Kira and L. A. Rendell, "A Practical Approach to Feature Selection," in *Machine Learning Proceedings 1992* (D. Sleeman and P. Edwards, eds.), pp. 249–256, San Francisco (CA): Morgan Kaufmann, 1992.
- [119] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of Biomedical Informatics*, vol. 85, pp. 189–203, 2018.
- [120] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1200–1205, 2015.
- [121] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data classification: Algorithms and applications*, pp. 37–70, 2014.
- [122] S. Srivastava, "A review paper on feature selection methodologies and their applications," *International Journal of Engineering Research and Development*, vol. 7, pp. 57–61, 2013.

-
- [123] C. D. Mathers and D. Loncar, “Projections of Global Mortality and Burden of Disease from 2002 to 2030,” *PLOS Medicine*, vol. 3, pp. 1–20, 11 2006.
- [124] “WHO - Depressive disorder (depression).” <https://www.who.int/news-room/fact-sheets/detail/depression>.
- [125] C. L. Hammen, *Depresja*. Gdańskie Wydawnictwo Psychologiczne, 2006.
- [126] K. Koziara, “Assessment of depressiveness in population. psychometric evaluation of the polish version of the cesd-r,” *Psychiatria Polska*, vol. 50, no. 6, p. 1109–1117, 2016.
- [127] W. Eaton, C. Smith, M. Ybarra, M. Carles, and A. Tien, “Center for Epidemiologic Studies Depression Scale: Review and revision (CESD and CESD-R),” *Instruments for Adults*, vol. 3, pp. 363–377, 2004.
- [128] A. Schmitt, “German version of the Center for Epidemiologic Studies Depression Scale Revised (CESD-R),” *Deutsche Fassung*, 2016.
- [129] P. Gałęcki, “Diagnostic and statistical manual of mental disorders,” American psychiatric association Washington, 2013.
- [130] J. Hong, J. Kim, S. Kim, J. Oh, D. Lee, S. Lee, J. Uh, J. Yoon, and Y. Choi, “Depressive symptoms feature-based machine learning approach to predicting depression using smartphone,” *Healthcare*, vol. 10, no. 7, 2022.
- [131] M. Grau and J. Marrugat, “Risk Functions and the Primary Prevention of Cardiovascular Disease,” *Revista Española de Cardiología (English Edition)*, vol. 61, no. 4, pp. 404–416, 2008.
- [132] J. Sala, J. Marrugat, R. Masiá, and M. Porta, “Improvement in survival after myocardial infarction between 1978–85 and 1986–88 in The REGICOR Study,” *European Heart Journal*, vol. 16, no. 6, pp. 779–784, 1995.
- [133] J. A. Sanz, M. Galar, A. Jurio, A. Brugos, M. Pagola, and H. Bustince, “Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system,” *Applied Soft Computing*, vol. 20, pp. 103–111, 2014.
- [134] P. Burillo and H. Bustince, “Construction theorems for intuitionistic fuzzy sets,” *Fuzzy Sets and Systems*, vol. 84, no. 3, pp. 271–281, 1996.
- [135] H. Sola and P. Burillo, “Interval-valued fuzzy relations in a set structures,” *The Journal of Fuzzy Mathematics*, vol. 4, 01 1996.

-
- [136] J. F. Baldwin and B. Pilsworth, “Axiomatic approach to implication for approximate reasoning with fuzzy logic,” *Fuzzy Sets and Systems*, vol. 3, no. 2, pp. 193–219, 1980.
 - [137] N. N. Karnik, J. M. Mendel, and Q. Liang, “Type-2 fuzzy logic systems,” *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 6, pp. 643–658, 1999.
 - [138] S. Spolaor, C. Fuchs, P. Cazzaniga, U. Kaymak, D. Besozzi, and M. Nobile, “Simpful: A user-friendly Python library for fuzzy logic,” in *International Journal of Computational Intelligence Systems*, vol. 13, p. 1687–1698, 2020.
 - [139] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

Spis rysunków

4.1	Schemat procesu klasyfikacji przeprowadzonej z użyciem klasyfikatorów leniwych	58
4.2	Schemat znajdowania najbliższych sąsiadów w otoczeniu obiektu klasyfikowanego z użyciem metryki euklidesowej	59
4.3	Schemat procesu klasyfikacji przeprowadzonej z użyciem klasyfikatorów gorliwych	60
4.4	Przykład zbioru danych i zbioru reguł klasyfikacyjnych.	61
4.5	Schemat ogólny drzewa decyzyjnego.	63
4.6	Przykładowe krzywe ROC	68
5.1	Ogólny schemat algorytmu IV-kNN	71
6.1	Ogólny schemat wnioskowania	103
6.2	Schemat oceny ryzyka zachorowania na CVD dla nowego pacjenta	108
6.3	IFIS (Interval–Valued Fuzzy Inference System)	110
6.4	Schemat oceny ryzyka zachorowania na CVD dla nowego pacjenta	111

Spis tabel

5.1	Wyniki uzyskane przy użyciu standardowego (rozmytego) algorytmu k -NN	77
5.2	Wyniki IV-kNN dla różnych klas miar podobieństwa	78
5.3	Tabela z porównaniem poszczególnych metod uzupełniania brakujących wartości	84
5.4	Tabela dla różnych entropii używanych w algorytmie 2	85
5.5	Tabela przedstawiająca wyniki uzyskane po zastosowaniu entropii możliwości do uzupełniania brakujących wartości	86
5.6	Tabela przedstawiająca wyniki uzyskane po zastosowaniu entropii pesymistycznej do uzupełniania brakujących wartości	86
5.7	Wyniki IVRelief - wektor wag uzyskany z wykorzystaniem miary podobieństwa konieczności	91
5.8	Porównanie metod redukcji atrybutów opartych na różnych klasach miar podobieństwa	91
5.9	Różne metody uczenia maszynowego dla CESD-R	97
5.10	Wyniki efektywności Systemu Prewencji Depresji dla różnych klas miar podobieństwa	98
6.1	Reguły rozmyte wykorzystane w niniejszej pracy	105
6.2	Wyniki uzyskane w procesie wnioskowania w zależności od zastosowanych agregacji: $Class_{err}$	109