



**Politechnika
Śląska**

Wydział Automatyki, Elektroniki i Informatyki
Instytut Elektroniki

Sebastian Porębski

**WSPOMAGANIE DIAGNOZY MEDYCZNEJ NA PODSTAWIE
ROZMYTYCH REGUŁ WARUNKOWYCH**

Rozprawa doktorska napisana pod kierunkiem
dr hab. inż. Ewy Straszeckiej

Gliwice, 2019 r.

Składam podziękowania

*Pani dr hab. inż. **Ewie Straszeckiej**,
za wzorowe pełnienie roli opiekuna naukowego i promotora,
nieoszacowaną pomoc w pracy badawczej,
rzetelną weryfikację jej efektów
oraz trafne doradzanie w chwilach niepewności,*

*Panu prof. dr hab. inż. **Jackowi Łęskiemu**
za słuszne uwagi dotyczące mojej pracy
oraz za wsparcie w rozmaitych sprawach
w czasie moich studiów doktoranckich,*

***Pracownikom Instytutu Elektroniki**,
za okazaną pomoc i częste pytania o postępy w pracy,
które wpływały pozytywnie na moje zaangażowanie.*

***Moim Przyjaciółom**
za obecność, troskę i wsparcie.*

*rozprawę poświęcam **Rodzicom**
Józefie i Januszowi Porębskim*

Spis treści

Spis rysunków	v
Spis tabel	ix
Wykaz ważniejszych oznaczeń i skrótów	xi
1 Wprowadzenie	1
1.1 Tematyka i zakres rozprawy	2
1.2 Cel i teza rozprawy	9
1.3 Treść rozprawy	11
2 Reprezentacja niepewności i nieprecyzyjności wiedzy medycznej	13
2.1 Teoria zbiorów rozmytych	14
2.1.1 Rodzaje funkcji przynależności	16
2.2 Teoria możliwości	22
2.2.1 Rozkład możliwości	22
2.2.2 Pojęcie miary rozmytej	23
2.2.3 Miara możliwości	24
2.2.4 Miara konieczności	25
2.2.5 Związek między miarami możliwości i konieczności	25
2.3 Teoria Dempstera-Shafera	26
2.3.1 Miary przekonania i domniemania	27
2.3.2 Rozmyte elementy ogniskowe	29
3 Rozmyte reguły warunkowe oparte na teorii Dempstera-Shafera	31
3.1 Reprezentacja danych medycznych	31
3.2 Definiowanie reguły diagnostycznej	32
3.3 Wyznaczanie rozkładu bazowego prawdopodobieństwa	34
3.4 Wspomaganie diagnozy na podstawie miary przekonania	36
3.5 Modyfikacja metody obliczania miary przekonania	39
3.6 Analiza danych uczących i wyznaczanie funkcji przynależności w rozmytych elementach ogniskowych	41
3.6.1 Trójkątna funkcja przynależności	43
3.6.2 Trapezowa funkcja przynależności	44
3.6.3 Uogólniona dzwonowa funkcja przynależności	46
3.6.4 Gaussowska łączona funkcja przynależności	48
3.7 Sposoby oceny reguł diagnostycznych	49
3.7.1 Ocena reguł na podstawie wartości bazowego prawdopodobieństwa	49
3.7.2 Ocena na podstawie MCC	50
3.7.3 Ocena na podstawie współczynnika zgodności	51

3.8	Ocena reguł na przykładzie bazy <i>iris flowers</i>	52
3.9	Podsumowanie	56
4	Algorytmy wydobywania reguł diagnostycznych	57
4.1	Wyznaczanie początkowego zbioru rozmytych elementów ogniskowych	57
4.2	Algorytm eliminacji reguł	60
4.3	Algorytmy selekcji reguł	63
4.4	Metody oceny działania algorytmów wydobywania reguł diagnostycznych	66
4.4.1	Efektywność wspomaganie diagnozy	66
4.4.2	Złożoność zbioru reguł diagnostycznych	69
4.5	Przykład wydobywania reguł dla bazy <i>iris flowers</i>	70
4.6	Ocena zbioru reguł diagnostycznych	73
4.7	Podsumowanie	78
5	Wyniki przeprowadzonych badań	79
5.1	Wpływ kształtu funkcji przynależności na rezultat wydobywania reguł	80
5.2	Warunek zatrzymania algorytmu eliminacji reguł	86
5.3	Złożoność wydobywanych reguł diagnostycznych	90
5.4	Zmodyfikowana miara przekonania	92
5.5	Wpływ warunkowej selekcji na wyniki wydobywania reguł	96
5.6	Wybrane metody weryfikacji zdolności uogólniania	97
5.7	Wpływ metody oceny rozmytych elementów ogniskowych na jakość wspomaganie diagnozy oraz złożoność reguł diagnostycznych	100
5.8	Porównanie proponowanych algorytmów wydobywania reguł diagnostycznych	102
5.9	Wydobywanie reguł diagnostycznych z rzeczywistej bazy medycznej	112
5.9.1	Diagnoza zwółknienia wątroby	112
5.9.2	Funkcje przynależności i rozmyte elementy ogniskowe	114
5.9.3	Wyniki wydobywania reguł	114
5.10	Porównanie wyników wydobywania reguł diagnostycznych z metodami referencyjnymi dla baz benchmarkowych	121
6	Wnioski końcowe	125
A	Testowane bazy	129
A.1	Opis baz i przykładowe zbiory reguł	129
A.1.1	<i>iris flowers</i>	129
A.1.2	<i>thyroid disease</i>	129
A.1.3	<i>appendicitis</i>	130
A.1.4	<i>Pima Indians diabetes</i>	130
A.1.5	<i>Wisconsin breast cancer</i>	131
A.1.6	<i>Cleveland heart disease</i>	132
A.2	Funkcje przynależności	134
	Literatura	146

Spis rysunków

2.1	Kategorie diagnostyczne ESC-ESH [37] wskazujące na stopniowanie nadciśnienia tętniczego	14
2.2	Zbiór rozmyty „nadciśnienie tętnicze” odpowiadający kategoriom diagnostycznym z rysunku 2.1	15
2.3	Klasyczne zbiory utworzone jako : (a) nośnik, (b) rdzeń, (c) α -przekrój ($\alpha = 0.4$) zbioru A przedstawionego na rysunku 2.2	16
2.4	Przedziałowo-liniowe funkcje przynależności: (a) trójkątna, (b) trapezowa	17
2.5	Gaussowskie funkcje przynależności: (a) symetryczna, (b) łączona	18
2.6	Funkcje przynależności (a) uogólniona dzwonowa, (b) sigmoidalna, $\beta_1 > \beta_2 > \beta_3$	19
2.7	Funkcje przynależności: (a) harmoniczna, (b) wielomianowa	20
2.8	Funkcje przynależności: (a) typu S, (b) typu Z	21
2.9	Funkcja przynależności typu rozmyty singleton	21
2.10	Pojęcie „stan podgorączkowy” opisane gaussowską funkcją przynależności	23
2.11	Objawy i zbiór elementów ogniskowych wraz z wartościami przypisanych prawdopodobieństw. Elementy ogniskowe zostały otoczone pętlą ciągłą a zbiór wszystkich objawów (2.40) pętlą przerywaną.	28
3.1	Procedura wyznaczania rozkładu bazowego prawdopodobieństwa	35
3.2	Zbiory rozmyte w przesłankach diagnozy pierwszej (linia kropkowana) oraz drugiej (linia kreskowana). Liniami ciągłymi zaprezentowano wartości symptomów przypadku $\mathbf{p}^{(*)}$	38
3.3	Zmiany wartości miary przekonania dla konkurujących ze sobą diagnoz w zależności od zmiany wartości prognozy dopasowania η_b	39
3.4	Wyznaczenie parametrów koniecznych do zdefiniowania funkcji przynależności dla trzech diagnoz	43
3.5	Trójkątne funkcje przynależności dla symptomów opisanych funkcjami gęstości z rysunku 3.4	44
3.6	Trapezowe funkcje przynależności dla funkcji gęstości z rysunku 3.4	46
3.7	Trapezowe funkcje przynależności na podstawie danych opisanych przez $P^{(1)}$, $P^{(2)}$: (a) bez konieczności, (b) z koniecznością korekcji wartości jej parametrów	47
3.8	Uogólnione dzwonowe funkcje przynależności dla symptomów opisanych funkcjami gęstości z rysunku 3.4	47
3.9	Gaussowskie łączone funkcje przynależności określone dla symptomów opisanych funkcjami gęstości z rysunku 3.4	49
3.10	Przykład dopasowania danych do elementu ogniskowego. Okręgi reprezentują przypadki z l -tej diagnozy a romby przypadki z l' -tej diagnozy. Białe kształty wskazują na dopasowanie przypadku i elementu ogniskowego lepsze od przyjętego prognozy, a szare przeciwnie.	51

3.11	Znormalizowane wartości ocen opartych na m , MCC oraz CC dla elementów ogniskowych utworzonych dla drugiej klasy bazy <i>iris flowers</i>	55
3.12	Znormalizowane wartości ocen opartych na m , MCC oraz CC dla elementów ogniskowych utworzonych dla trzeciej klasy bazy <i>iris flowers</i>	55
4.1	Schemat blokowy procedury tworzenia początkowych zbiorów elementów ogniskowych	59
4.2	Schemat blokowy zmodyfikowanego algorytmu eliminacji reguł	62
4.3	Schemat blokowy algorytmu selekcji reguł	64
4.4	Schemat blokowy algorytmu warunkowej selekcji reguł	67
4.5	Zmiana liczby reguł i błędów podczas działania algorytmu eliminacji reguł (AER) dla bazy <i>iris flowers</i>	71
4.6	Zmiana liczby reguł i błędów podczas działania algorytmu selekcji reguł (ASR) dla bazy <i>iris flowers</i>	71
4.7	Zmiana liczby reguł i błędów podczas działania algorytmu warunkowej selekcji reguł (AWR) dla bazy <i>iris flowers</i>	71
4.8	Wpływ zmiany progów η_e i η_b na efektywność klasyfikacji zbiorów reguł otrzymanych w wyniku działania algorytmów: (a) AER, (b) ASR, (c) AWR	75
4.9	Zależność całkowitej liczby reguł od efektywności klasyfikacji uzyskanych dla wszystkich kombinacji progów η_e i η_b w algorytmie wydobywania reguł ASR	76
5.1	Zbiór elementów ogniskowych jako wynik algorytmu eliminacji reguł dla czterech kształtów funkcji przynależności – baza <i>thyroid disease</i>	82
5.2	Zbiór elementów ogniskowych jako wynik algorytmu eliminacji reguł dla czterech kształtów funkcji przynależności – baza <i>appendicitis</i>	83
5.3	Zbiór elementów ogniskowych jako wynik algorytmu eliminacji reguł dla czterech kształtów funkcji przynależności – baza <i>Pima Indians diabetes</i>	83
5.4	Zbiór elementów ogniskowych jako wynik algorytmu eliminacji reguł dla czterech kształtów funkcji przynależności – baza <i>Wisconsin breast cancer</i>	84
5.5	Zbiór elementów ogniskowych jako wynik algorytmu eliminacji reguł dla czterech kształtów funkcji przynależności – baza <i>Cleveland heart disease</i>	84
5.6	Zmiana liczby reguł oraz procentowej efektywności (4.2) w każdej iteracji algorytmu eliminacji reguł dla baz: (a) <i>thyroid disease</i> , (b) <i>appendicitis</i> , (c) <i>Pima Indians diabetes</i>	87
5.7	Zmiana liczby reguł oraz procentowej efektywności (4.2) w każdej iteracji algorytmu eliminacji reguł dla baz: (a) <i>Wisconsin breast cancer</i> , (b) <i>Cleveland heart disease</i>	88
5.8	Dynamika zmian liczby błędnych diagnoz w iteracjach algorytmu eliminacji reguł dla baz: (a) <i>thyroid disease</i> , (b) <i>appendicitis</i> , (c) <i>Pima Indians diabetes</i> , (d) <i>Wisconsin breast cancer</i> , (e) <i>Cleveland heart disease</i>	89
5.9	Porównanie wyników algorytmu selekcji reguł dla bazy <i>thyroid disease</i> otrzymanych za pomocą dwóch sposobów obliczania miary przekonania	93
5.10	Porównanie wyników algorytmu selekcji reguł dla bazy <i>appendicitis</i> otrzymanych za pomocą dwóch sposobów obliczania miary przekonania	94
5.11	Porównanie wyników algorytmu selekcji reguł dla bazy <i>Pima Indians diabetes</i> otrzymanych za pomocą dwóch sposobów obliczania miary przekonania	94
5.12	Porównanie wyników algorytmu selekcji reguł dla bazy <i>Wisconsin breast cancer</i> otrzymanych za pomocą dwóch sposobów obliczania miary przekonania	95

5.13	Porównanie wyników algorytmu selekcji reguł dla bazy <i>Cleveland heart disease</i> otrzymanych za pomocą dwóch sposobów obliczania miary przekonania	95
5.14	Schemat procedury rotacji ($\alpha = 0.5$)	98
5.15	Schemat procedury K -krotnej walidacji ($K = 5$)	99
5.16	Wynik procedury rotacji dla bazy <i>thyroid disease</i> w postaci wykresu „skrzynka z wąsami” dla badanych algorytmów przy zastosowaniu wartości CC jako wskaźnika oceny reguł: (a) rozkład efektywności, (b) rozkład liczebności reguł, (c) rozkład wartości QI	107
5.17	Wynik procedury rotacji dla bazy <i>appendicitis</i> w postaci wykresu „skrzynka z wąsami” dla badanych algorytmów przy zastosowaniu wartości CC jako wskaźnika oceny reguł: (a) rozkład efektywności, (b) rozkład liczebności reguł, (c) rozkład wartości QI	108
5.18	Wynik procedury rotacji dla bazy <i>Pima Indians diabetes</i> w postaci wykresu „skrzynka z wąsami” dla badanych algorytmów przy zastosowaniu wartości CC jako wskaźnika oceny reguł: (a) rozkład efektywności, (b) rozkład liczebności reguł, (c) rozkład wartości QI	109
5.19	Wynik procedury rotacji dla bazy <i>Wisconsin breast cancer</i> w postaci wykresu „skrzynka z wąsami” dla badanych algorytmów przy zastosowaniu wartości CC jako wskaźnika oceny reguł: (a) rozkład efektywności, (b) rozkład liczebności reguł, (c) rozkład wartości QI	110
5.20	Wynik procedury rotacji dla bazy <i>Cleveland heart disease</i> w postaci wykresu „skrzynka z wąsami” dla badanych algorytmów przy zastosowaniu wartości CC jako wskaźnika oceny reguł: (a) rozkład efektywności, (b) rozkład liczebności reguł, (c) rozkład wartości QI	111
5.21	Trójkątne funkcje przynależności zaprojektowane dla bazy dotyczącej diagnozy stopnia zwłóknienia wątroby – symptomy od 1 do 14	115
5.22	Trójkątne funkcje przynależności zaprojektowane dla bazy dotyczącej diagnozy stopnia zwłóknienia wątroby – symptomy od 15 do 26	116
5.23	Trójkątne funkcje przynależności zaprojektowane dla bazy dotyczącej diagnozy stopnia zwłóknienia wątroby – symptomy występujące w regułach przedstawionych w tabeli 5.18	119
A.1	Funkcje przynależności dla bazy <i>iris flowers</i> : (a) trójkątne, (b) trapezowe, (c) uogólnione dzwonowe, (d) gaussowskie łączone	134
A.2	Funkcje przynależności dla bazy <i>thyroid disease</i> : (a) trójkątne, (b) trapezowe, (c) uogólnione dzwonowe, (d) gaussowskie łączone	135
A.3	Funkcje przynależności dla bazy <i>appendicitis</i> : (a) trójkątne, (b) trapezowe	136
A.4	Funkcje przynależności dla bazy <i>appendicitis</i> : (a) uogólnione dzwonowe, (b) gaussowskie łączone	137
A.5	Funkcje przynależności dla bazy <i>Pima Indians diabetes</i> : (a) trójkątne, (b) trapezowe	138
A.6	Funkcje przynależności dla bazy <i>Pima Indians diabetes</i> : (a) uogólnione dzwonowe, (b) gaussowskie łączone	139
A.7	Funkcje przynależności dla bazy <i>Wisconsin breast cancer</i> : (a) trójkątne, (b) trapezowe	140
A.8	Funkcje przynależności dla bazy <i>Wisconsin breast cancer</i> : (a) uogólnione dzwonowe, (b) gaussowskie łączone	141
A.9	Trójkątne funkcje przynależności dla bazy <i>Cleveland heart disease</i>	142

- A.10 Trapezowe funkcje przynależności dla bazy *Cleveland heart disease* 143
- A.11 Uogólnione dzwonowe funkcje przynależności dla bazy *Cleveland heart disease* 144
- A.12 Gaussowskie łączone funkcje przynależności dla bazy *Cleveland heart disease* 145

Spis tabel

2.1	Kategorie diagnostyczne nadciśnienia tętniczego na podstawie wytycznych ESH-ESC [37]	13
3.1	Rozmyte elementy ogniskowe i odpowiadające im zbiory symptomów	33
3.2	Poziomy dopasowania siedmiu elementów ogniskowych i 10 danych uczących oraz obliczony na ich podstawie rozkład bazowego prawdopodobieństwa. Wartości $\eta_r^{(l)}(\mathbf{p}_i^{(l)})$ większe od $\eta_e = 0.4$ zostały pogrubione.	36
3.3	Poziomy dopasowania elementów ogniskowych do zadanego w przykładzie przypadku medycznego (3.17)	38
3.4	Liczba przypadków z bazy <i>iris flowers</i> dopasowanych do elementów ogniskowych należących do zbioru $S^{(2)}$ i odpowiadające im wartości bazowego prawdopodobieństwa	54
3.5	Bazowe prawdopodobieństwo, <i>Matthews Correlation Coefficient</i> i współczynnik zgodności obliczone dla zbioru elementów ogniskowych $S^{(2)}$	54
3.6	Bazowe prawdopodobieństwo, <i>Matthews Correlation Coefficient</i> i współczynnik zgodności obliczone dla zbioru elementów ogniskowych $S^{(3)}$	56
4.1	Macierz pomyłek klasyfikacji dwuklasowej	68
4.2	Macierz pomyłek dla l -tej diagnozy	68
4.3	Zbiory elementów ogniskowych $\mathcal{S}^{(l)}$, $l = 1, \dots, 3$ uzyskane dla bazy <i>iris flowers</i> dzięki algorytmom: AER, ASR, AWR. Poniżej oznaczeń rozmytych elementów ogniskowych zapisano ich wartość bazowego prawdopodobieństwa.	73
4.4	Zbiór reguł wydobyty na podstawie danych z bazy <i>iris flowers</i>	77
5.1	Opis użytych benchmarkowych baz medycznych	79
5.2	Wpływ kształtu funkcji przynależności na rezultat algorytmu eliminacji reguł dla testowanych baz	81
5.3	Porównanie najlepszych kształtów funkcji przynależności wybranych dla każdej z baz na podstawie wartości efektywności przedstawionych w tabeli 5.2	85
5.4	Porównanie najlepszych kształtów funkcji przynależności wybranych dla każdej z baz na podstawie liczebności zbiorów reguł diagnostycznych przedstawionych w tabeli 5.2	85
5.5	Złożoność uzyskanych zbiorów reguł diagnostycznych (4.14) dla każdej metody oceny w wyniku algorytmu selekcji reguł (ASR) dla bazy <i>thyroid disease</i>	91
5.6	Złożoność uzyskanych zbiorów reguł diagnostycznych (4.14) dla każdej metody oceny w wyniku algorytmu selekcji reguł (ASR) dla bazy <i>appendicitis</i>	91
5.7	Złożoność uzyskanych zbiorów reguł diagnostycznych (4.14) dla każdej metody oceny w wyniku algorytmu selekcji reguł (ASR) dla bazy <i>Pima Indians diabetes</i>	91

5.8	Złożoność uzyskanych zbiorów reguł diagnostycznych (4.14) dla każdej metody oceny w wyniku algorytmu selekcji reguł (ASR) dla bazy <i>Wisconsin breast cancer</i>	92
5.9	Złożoność uzyskanych zbiorów reguł diagnostycznych (4.14) dla każdej metody oceny w wyniku algorytmu selekcji reguł (ASR) dla bazy <i>Cleveland heart disease</i>	92
5.10	Procentowa zmiana efektywności uczenia (ΔE) i liczebności reguł diagnostycznych (ΔL) przy użyciu zmodyfikowanej miary przekonania (3.21) dla algorytmu selekcji reguł (ASR)	94
5.11	Procentowa efektywność diagnozy oraz liczebność reguł uzyskanych w wyniku algorytmów wydobywania reguł diagnostycznych ASR i AWR	96
5.12	Średnie efektywności uogólniania oraz złożoność reguł diagnostycznych uzyskiwanych w procedurach rotacji z walidacją ($K = 100, \alpha = 0.5$) dla wszystkich badanych baz benchmarkowych	101
5.13	Średnie wyniki efektywności, liczebności reguł oraz wskaźnika QI w procedurach rotacji z walidacją ($K = 100, \alpha = 0.5$) dla wszystkich badanych baz benchmarkowych	103
5.14	Średnie efektywności, liczebności złożonych elementów ogniskowych oraz wartości wskaźnika QI uzyskiwane w procedurze rotacji z walidacją przedstawione osobno dla każdej z baz benchmarkowych	105
5.15	Opis symptomów zebranych w bazie danych na temat pacjentów zarażonych wirusem HCV i procentowe liczby brakujących danych w całym zbiorze i w poszczególnych diagnozach	113
5.16	Zbiór rozmytych elementów ogniskowych dla diagnozy stopnia zwłóknienia wątroby	117
5.17	Macierz pomyłek i miary pochodne uzyskane dla wydobytych reguł diagnostycznych przedstawionych w tabeli 5.16	118
5.18	Zbiór rozmytych elementów ogniskowych dla diagnozy stopnia zwłóknienia wątroby – podejście z dwoma diagnozami	119
5.19	Macierz pomyłek i miary pochodne uzyskane dla wydobytych reguł diagnostycznych przedstawionych w tabeli 5.16	121
5.20	Średnie efektywności uczenia i testu oraz średnie liczby reguł otrzymanych w wyniku przeprowadzenia zmodyfikowanej procedury 10-F DOBSCV na testowych bazach danych za pomocą wybranych metod	122
5.21	Średnie stosunki liczb warunków logicznych do liczb reguł otrzymane dzięki analizie wyników procedury 10-F DOBSCV	122
A.1	Zbiór reguł diagnostycznych dotyczących stanu funkcjonowania tarczycy . . .	130
A.2	Zbiór reguł diagnostycznych dotyczących diagnozy stanu zapalnego wyrostka	130
A.3	Zbiór reguł diagnostycznych dotyczących diagnozy cukrzycy	131
A.4	Zbiór reguł diagnostycznych dotyczących diagnozy złośliwości nowotworu piersi	132
A.5	Zbiór reguł diagnostycznych dotyczących diagnozy choroby serca	133

Wykaz ważniejszych oznaczeń i skrótów

- x – element zbioru
- X, E, F – zbiory
- A, B – zbiory rozmyte
- $\mu_A(x)$ – funkcja przynależności zbioru rozmytego A
- $\text{Supp}(A)$ – nośnik zbioru rozmytego A
- $\text{Core}(A)$ – rdzeń zbioru rozmytego A
- A_α – α -przekrój zbioru rozmytego A
- \underline{A}_α – ostry α -przekrój zbioru rozmytego A
- $a, c, b, d, \sigma, \beta, x_0$ – parametry funkcji przynależności
- Π – rozkład możliwości
- \mathbb{X} – przestrzeń
- \mathcal{X} – zmienna
- $\pi_{\mathcal{X}}(x)$ – funkcja rozkładu możliwości zmiennej \mathcal{X}
- $g()$ – miara rozmyta
- \mathcal{C} – klasa podzbiorów
- \mathcal{M} – miara możliwości
- \mathcal{N} – miara konieczności
- m – wartość bazowego prawdopodobieństwa
- Bel – miara przekonania
- Pl – miara domniemania
- X – macierz z danymi pacjentów
- \mathbf{p} – wektor symptomów pacjenta
- C – liczba rozpatrywanych diagnoz
- \mathbf{x} – wektor wartości symptomu
- S – zbiór elementów ogniskowych
- ρ – liczba symptomów
- \mathbb{J} – wektor indeksów symptomów
- f – symptom niezwiązany z diagnozą
- η – poziom dopasowania
- η_e, η_b – progi dopasowania
- $I()$ – funkcja wskaźnikowa
- \underline{D} – etykieta diagnozy
- \widetilde{Bel} – zmodyfikowana miara przekonania
- P – funkcja gęstości prawdopodobieństwa
- tp, fn, fp, tn, ld – liczby dopasowań elementu ogniskowego
- N – liczba pacjentów
- w – maksymalna liczba warunków logicznych w przesłance
- V – wektor wartości ocen elementu ogniskowego
- v – wartość oceny elementu ogniskowego

\hat{S} – początkowy zbiór elementów ogniskowych
 \mathcal{S} – ostateczny zbiór elementów ogniskowych
 Θ – wektor liczby prostych i złożonych elementów ogniskowych
 θ – liczba elementów ogniskowych o jednakowej liczbie warunków logicznych
 QI – wskaźnik jakości reguł
 N_e – liczba błędnych diagnoz
 \mathbf{U} – zbiór uczący
 \mathbf{T} – zbiór testowy
 \bar{E} – średnia procentowa efektywność uczenia lub testu
 F – podzbiór w metodzie k -krotnej walidacji
 \bar{L} – średnia liczba reguł

AER – algorytm eliminacji reguł
ASR – algorytm selekcji reguł
AWR – algorytm warunkowej selekcji reguł
BPA – (ang. *Basic Probability Assignment*) – rozkład bazowego prawdopodobieństwa
CC – (ang. *Conformance Coefficient*) – współczynnik zgodności
SE – (ang. *Sensitivity*) – czułość
SP – (ang. *Specificity*) – swoistość
MAD – (ang. *Median Absolute Deviation*) – medianowe odchylenie bezwzględne
METAVIR – (ang. *Meta-analysis of Histological Data in Viral Hepatitis*) – metaanaliza danych histologicznych w wirusowym zapaleniu wątroby
MCC – (ang. *Matthews Correlation Coefficient*) – współczynnik korelacji Matthews'a
NPV – (ang. *Negative Prognostic Value*) – negatywna wartość prognostyczna
PPV – (ang. *Positive Prognostic Value*) – pozytywna wartość prognostyczna
QI – (ang. *Quality Index*) – wskaźnik jakości

TH – baza *thyroid disease*
AP – baza *appendicitis*
PI – baza *Pima Indians diabetes*
BC – baza *Wisconsin breast cancer*
HD – baza *Cleveland heart disease*

Rozdział 1

Wprowadzenie

Na przestrzeni wieków medycyna, obok innych aktywności człowieka, była traktowana jako jedna z gałęzi sztuki. Z drugiej strony, ważnym elementem rozwoju medycyny był postęp techniki umożliwiający między innymi obserwacje mikroskopowe, rejestrację aktywności elektrycznej ciała ludzkiego czy obrazowanie tkanek kostnych za pomocą promieniowania rentgenowskiego [74]. Te i inne osiągnięcia nierozdzielnie związały medycynę z nauką. Od tego momentu korzystanie z narzędzi techniki stanowi nieodłączny element w poznawaniu i weryfikacji wiedzy o człowieku.

Dzisiaj, tak jak w innych dziedzinach, tak i w medycynie mamy do czynienia z rosnącą ilością informacji. Cyfrowy zapis obrazów i wyników badań pacjenta umożliwia tworzenie coraz większych zbiorów danych. Przy rosnącej ilości informacji medycznej jej analiza przez lekarza staje się utrudniona [18]. Niemniej, taka analiza jest konieczna, jeśli procedury diagnostyczne wzbogacane są o kolejne wyniki testów lub następuje modyfikacja postępowania diagnostycznego. Inteligencja obliczeniowa systemów komputerowych może być pomocą w tej analizie [38].

Dalszy rozwój technik medycznych będzie nadal wzbogacał informację medyczną. Jednakże, wiedza medyczna pozostanie domeną lekarzy, bo tylko oni mogą brać odpowiedzialność za stawianą diagnozę, która ma bezpośredni wpływ na ludzkie zdrowie. Niemniej jednak, urządzenia elektroniczne oraz ich projektanci są po części odpowiedzialni za diagnozę pacjenta. Urządzenia dostarczają informacji niedoskonałej, tzn. wyników pomiarów zawsze obciążonych skończoną dokładnością i obrazów będących nieprecyzyjną próbą opisu rzeczywistego stanu pacjenta, które poddane są wstępnemu przetwarzaniu w systemach filtracyjnych. Lekarz na podstawie takich informacji podejmuje diagnozę, która ze względu na niedoskonałość urządzeń pomiarowych staje się podejmowaniem decyzji w obliczu niepewności. Dlatego niezmiernie ważnym wydaje się ujęcie aspektów nieprecyzyjności informacji i niepewności wiedzy medycznej w podejmowaniu diagnozy [105]. Tym, co może zaoferować inżynieria wiedzy, są systemy wspomaganie diagnozy [18]. Sama inżynieria wiedzy (ang. *knowledge engineering*) zajmuje się wydobywaniem wiedzy z danych, a ściślej budowaniem systemów opartych na wiedzy (ang. *knowledge-based systems*) [38]. W przypadku danych medycznych poprzez wydobywanie wiedzy należy rozumieć dążenie do opisu źródła

wiedzy, który w sposób zrozumiały pomaga określić związek objawów (symptomów) z diagnozą lub wśród wszystkich danych wybrać tę część, która najlepiej opisuje stan chorobowy i pozwala jasno odróżnić go od stanu zdrowia. Takie rozwiązania mogą sprawdzić się najbardziej w elektronicznym systemie monitorowania pacjenta, który oprócz akwizycji informacji biomedycznej mógłby w sposób obiektywny przetwarzać nierozzerwalną z pomiarami nieprecyzyjność i niepewność wyciąganych wniosków tj. sugerowanych decyzji diagnostycznych.

Główną przeszkodą sprawnego wydobywania wiedzy z danych, szczególnie w obszarze medycznym, jest niemożliwość uniknięcia ich niedoskonałości [58]. To słowo kryje kilka cech [105]. Pierwszą z nich jest niekompletność zbioru danych. Mamy z nią do czynienia zwłaszcza wtedy gdy jakaś baza medyczna była tworzona przez wiele lat a opis pacjentów w tym czasie zmieniał się [97]. Często pacjenci w gorszym stanie są poddawani większej liczbie testów [88]. Tym samym nie zostaje zachowana proporcja opisu pacjentów zaliczanych do tych samych a tym bardziej do różnych kategorii diagnostycznych (mówi się wtedy o danych niezbalansowanych) [97]. Obecność danych obarczonych dużym błędem pomiarowym lub omyłką ludzką to też nie rzadkość, podobnie jak dane odstające [58]. Dana odstająca znacznie odbiega od pozostałych wartości symptomu wszystkich pacjentów. Jej wykrycie z jednej strony jest zadaniem stosunkowo prostym i wykonywanym za pomocą szeregu metod [1]. z drugiej strony ocena czy wartość odstająca jest wynikiem błędu czy też osobniczej zmienności w przypadkach medycznych jest wyjątkowo trudna.

1.1 Tematyka i zakres rozprawy

Niniejsza praca porusza temat wydobywania wiedzy z danych medycznych. Szczególną uwagę poświęca się nieprecyzyjności objawów oraz niepewności wiedzy medycznej występującej podczas formułowania diagnozy. Prowadzone badania mają na celu zbadanie możliwości wykorzystania wydobytej wiedzy w zagadnieniach wspomagania diagnozy medycznej za pomocą reguł. Najefektywniejszym wykorzystaniem przeprowadzonych w rozprawie badań we wspomaganiu diagnozy medycznej jest dostarczenie elektronicznego systemu wspomagania diagnozy. W niniejszej rozprawie, korzystając z ogólnodostępnych danych medycznych w postaci tzw. danych benchmarkowych, weryfikuje się proponowane metody mające na celu udoskonalenie i ułatwienie procesu diagnozy na podstawie danych medycznych.

Na przestrzeni ostatnich kilkudziesięciu lat ukształtowało się kilka kierunków opracowywania danych niedoskonałych. Podejścia te można ująć pod wspólną nazwą reprezentacji wiedzy z niepewnością. Teoria prawdopodobieństwa, która powstała jako próba opisu zjawisk występujących w grach losowych [15], stanowiła przez długi czas najczęstsze podejście do przetwarzania danych niepewnych. Była to bowiem jedyna dostępna i akceptowana teoria [58]. Korzystając z klasycznego twierdzenia Bayesa możliwe jest określenie warunkowego prawdopodobieństwa pewnej hipotezy na podstawie dostępnych danych. Niestety taki sposób obliczania warunkowego prawdopodobieństwa wymaga znajomości rozkładów prawdopodobieństwa związanych z zagadnieniem zwanych prawdopodobieństwa-

mi *a priori* (łac. „z góry”, „z założenia”). Te rozkłady przeważnie nie mogą być wyznaczone w sposób analityczny. Metody estymacji prawdopodobieństwa takie jak filtry Kalmana [71], metoda Monte Carlo i łańcuchy Markowa [16] często związane są z założeniami uproszczającymi złożoność problemu, bądź przeciwnie, z dużym nakładem obliczeniowym potrzebnym do osiągnięcia założonej dokładności. Mimo tych ograniczeń, metody wykorzystujące miarę prawdopodobieństwa są przedmiotem badań nad przetwarzaniem informacji niepewnych [6, 62, 128].

Miara przekonania opisana przez G. Shafera [103] jako matematyczne sformalizowanie wniosków wysuniętych przez A. P. Dempstera [27, 28], jest stosowana do opisu danych niepewnych [49, 51, 58, 77, 78, 124]. Teoria Dempstera-Shafera definiuje tak zwany rozkład bazowego prawdopodobieństwa [21] (BPA – ang. *Basic Probability Assignment* [103]) w sposób odmienny do rozkładu prawdopodobieństwa w klasycznej teorii. Rozkład bazowego prawdopodobieństwa wyznacza się dla zbioru elementów ogniskowych [103], bez rozpatrywania ich wzajemnej zależności. Umożliwia to zastosowanie miary prawdopodobieństwa w sytuacji, w której zależność objawów nie jest możliwa do określenia, a tym bardziej nie jest możliwe obliczenie prawdopodobieństwa dla zależnych symptomów. W niniejszej rozprawie poprzez element ogniskowy rozumie się wystąpienie wybranego symptomu. Zgodnie z tą teorią istnieje możliwość przypisania wartości bazowego prawdopodobieństwa nie tylko pojedynczym symptomom, ale także każdej możliwej ich kombinacji (podzbiorem wszystkich symptomów). Tak więc określony podzbiór symptomów (jedno- lub wieloelementowy) dla wybranej diagnozy, któremu przypisano niezerową wartość bazowego prawdopodobieństwa, nazywa się elementem ogniskowym (ang. *focal element*) [103]. Jednakże, przypisanie bazowego prawdopodobieństwa wszystkim możliwym kombinacjom symptomów jest zadaniem, którego złożoność wzrasta wykładniczo wraz z liczbą rozpatrywanych symptomów. Oryginalna reguła Dempstera [27] łączenia rozkładów bazowego prawdopodobieństwa z różnych źródeł i jej modyfikacja stanowią aktywny obszar badań [13, 106, 133]. Proces wnioskowania z zastosowaniem teorii Dempstera-Shafera wiąże się z obliczeniem miary przekonania o danej hipotezie wyznaczonej na podstawie potwierdzenia elementów ogniskowych. Miara przekonania wskazuje jaka jest niepewność hipotezy i umożliwia porównanie różnych hipotez. Niemniej jednak, oryginalna teoria Dempstera-Shafera operowała na zdarzeniach opisywanych w sposób klasyczny, poprzez zbiory. Tym samym wyznaczanie miary przekonania, łączenie rozkładów bazowego prawdopodobieństwa były związane z analizą zdarzeń w klasycznej teorii zbiorów i w teorii nie było miejsca na ujęcie ich nieprecyzyjności [105]. Ten fakt spowodował, że oryginalna teoria doczekała się rozszerzeń i odmiennych interpretacji [30, 32, 101, 111].

Przełomem w przetwarzaniu informacji nieprecyzyjnej było wprowadzenie teorii zbiorów rozmytych przez L. A. Zadeha [130]. Teoria ta uznawana jest za rozszerzenie lub uogólnienie klasycznej teorii zbiorów, ponieważ zakłada ona, że przynależność elementu do zbioru może być częściowa. Opisanie funkcji przynależności odwzorowującej elementy zbioru na wartości z zakresu od 0 do 1 stworzyło podstawy matematyczne opisu pojęć nieprecyzyjnych. Odważny krok L. A. Zadeha spowodował, że inni badacze zaproponowali swoje teorie matematyczne mniej lub bardziej związane ze zbiorami rozmytymi lub nawet uznawane

za ich uogólnienia. Były to między innymi: zbiory rozmyte typu-2 i typu-n (wprowadzone przez samego twórcę zbiorów rozmytych [131, 134]), przedziałowe [82], przybliżone [90, 91], nieokreślone [45], intuicjonistyczne rozmyte [8], rozmyte multizbiory [123], zbiory cieniowane [92], miękkie [81] a także ich kombinacje [5, 25, 39, 76]. Choć dostarczenie wymienionych narzędzi pozwoliło w pewnych zagadnieniach osiągnąć lepsze rezultaty od zbiorów rozmytych, to jednak oryginalna idea L. A. Zadeha wyrażona przez niego m.in. jako „precyzyjna logika nieprecyzyjności” [52] znalazła do tej pory najwięcej zastosowań, wymienionych między innymi w [104].

Po raz pierwszy zagadnienie rozszerzenia teorii Dempstera-Shafera na zbiory rozmyte pojawia się już w pracach ich twórcy. W pracy [132] Zadeh opisał nową teorię możliwości. Najważniejszym pojęciem w tej teorii jest rozkład możliwości. W odróżnieniu od rozkładu prawdopodobieństwa, opisuje on stopień możliwości wystąpienia zdarzenia. Autor odnosi się również do teorii Dempstera-Shafera. Rozwija pierwotnie zaproponowaną przez A. P. Dempstera relację zgodności na tzw. warunkowy rozkład możliwości [132]. To rozwinięcie pozwala określić stopień zgodności dwóch zbiorów rozmytych, a nie jak dotąd, operować jedynie na określeniach ich pełnej zgodności lub niezgodności. Tym samym wprowadza rozmyte miary tzw. przewidywanej pewności oraz możliwości. Stanowią one uogólnienia miar przekonania i domniemania wprowadzonych przez G. Shafera [103, 129].

W pracy [122] R. R. Yager także opisuje miary przekonania i domniemania w oparciu o zbiory rozmyte. Opisując miarę przekonania, zauważa, że wyznaczanie inkluzji dwóch zbiorów rozmytych nie jest jednoznaczne, gdyż zależy od rodzaju zastosowanego operatora czyli t-normy. Swoje definicje inkluzji zbiorów rozmytych przedstawiali także inni, m.in. M. Ishizuka [54], H. Ogawa [87] i J. Yen [129]. J. Yen, podobnie jak inni, przypisuje wartość bazowego prawdopodobieństwa zbiorowi rozmytemu jako sumie bazowych prawdopodobieństw jego elementów. Jednakże opiera obliczanie inkluzji i wartości przekonania na dekompozycji rozmytego elementu ogniskowego na kilka α -przekrojów. Ta szczegółowa analiza związana jest jednak z dużym nakładem obliczeniowym.

Nieco inne podejście zaproponowali D. Dubois i H. Prade [32]. Wykorzystując opisane wcześniej przez L. A. Zadeha miary możliwości i konieczności wskazali metody porównywania zbiorów rozmytych. Dalszej analizy zagadnienia oraz odniesienia do teorii Dempstera-Shafera dokonano w pracy [33]. Autorzy opisują związek teorii możliwości wprowadzonej przez L.A. Zadeha i teorii zbiorów probabilistycznych z teorią Dempstera-Shafera. Porównują rozkład prawdopodobieństwa dla zbioru losowego z rozkładem bazowego prawdopodobieństwa, a także miarę możliwości z miarą domniemania. Obliczenia przy użyciu zbiorów probabilistycznych uznano za warte stosowania przy zagadnieniach przetwarzania danych niepewnych. Do podobnego wniosku doprowadziła szczegółowa analiza miar domniemania i możliwości w pracy [113]. W dalszych pracach związanych z teorią możliwości i teorią prawdopodobieństwa można zauważyć nierzadko odejście od teorii Dempstera-Shafera i niezależne badania [31].

Mając do dyspozycji wspomniane interpretacje teorii Dempstera-Shafera rozszerzając ją na zbiory rozmyte można zauważyć pewne podejścia modyfikujące znane metody rozpo-

znawania i klasyfikacji. W swojej pracy [125], R. R. Yager podejmuje próbę ujęcia niepewności w systemie wnioskowania rozmytego. Używając modelu Mamdaniego-Assilana [72] opisuje reguły wnioskowania z użyciem elementów ogniskowych w ich konkluzjach. Oryginalny proces wnioskowania systemu jest zmodyfikowany tak, aby zarówno bazowe prawdopodobieństwa mogły zostać ujęte w procesie aktywacji reguł jak i rozmyte elementy ogniskowe w agregacji konkluzji. Aby wyjście systemu rozmytego miało wartość rzeczywistą, autor postuluje najpierw przekształcenie rozmytych elementów ogniskowych w wartości rzeczywiste za pomocą tak zwanej operacji wyostrzania (ang. *defuzzification*) [69], a później zsumowanie otrzymanych wartości ważonych przez przypisane im wartości bazowego prawdopodobieństwa. W systemie wspomaganie decyzji – w odróżnieniu do systemu rozmytego z wartością rzeczywistą jako wyjściem – odpowiedź powinna być raczej jednoznacznie określonym wynikiem, tj. decyzją, diagnozą. Mimo to trzeba zaznaczyć, że w opisanym powyżej podejściu autor w satysfakcjonujący sposób ujmuje niepewność wiedzy za pomocą określania wag reguł. Zbliżone podejścia do zagadnień klasyfikacji wspomagane przez teorię Dempstera-Shafera zaproponował też E. Binaghi w [19] oraz [20]. Warto wspomnieć o podejściu prezentuje także T. Denoeux używając miary przekonania do klasyfikacji za pomocą algorytmu k-najbliższych sąsiadów [29, 138] czy sieci neuronowych [30].

Zbiory rozmyte i teoria Dempstera-Shafera zostały wykorzystane także w systemie opartym na wnioskowaniu dowodowym RIMER [126] (ang. *Rule-base Inference Methodology using the Evidential Reasoning approach*). W systemie tym niepewna wiedza czy nieprecyzyjne dane są przetwarzane za pomocą bazy reguł, w której każdy z warunków logicznych zawartych w przesłankach ma przypisany stopień nieprecyzyjności. Następną charakterystyczną cechą reguły w podejściu RIMER jest to, że w konkluzji występuje pełny zbiór możliwych wniosków, a każdemu z nich przypisany jest stopień niepewności. To podejście wydaje się słuszne w zagadnieniach klasyfikacji, natomiast, ze względu na niemałą liczbę parametrów opisujących reguły, jest dyskusyjne w zastosowaniu do wspomaganie decyzji. Reguła z konkluzją odnoszącą się do wielu konkurujących decyzji może wydawać się nietrafna z punktu widzenia diagnozy medycznej. Na etapie projektu systemu można konstruować reguły odnoszące się do konkretnych diagnoz. Te reguły mogą dodatkowo w swoich przesłankach zawierać warunki logiczne odnoszące się do symptomów charakterystycznych dla jednej diagnozy. Symptomy, które jednakowo odpowiadają obu konkurującym diagnozom, nie niosą dodatkowej informacji w procesie wnioskowania. Wydaje się więc, że nie powinny występować w konkurujących ze sobą regułach.

W pracy [107] dokonano analizy symptomów branych pod uwagę przy wnioskowaniu diagnostycznym. Autorka uważa, że w procesie wspomaganie decyzji diagnostycznych, system powinien ujmować zarówno symptomy reprezentowane przez zbiory rozmyte [112], a jednocześnie te, które wykazują klasyczną wartość prawdy lub fałszu (wystąpienia bądź niewystąpienia danego objawu). W opisywanym podejściu do wnioskowania z użyciem teorii Dempstera-Shafera elementy ogniskowe są reprezentowane przez zbiory rozmyte. Wartość bazowego prawdopodobieństwa przypisuje się rozmytym elementom ogniskowym jako wartościom lingwistycznym opisującym symptomy. Nie definiuje się natomiast bazowego prawdopodobieństwa jako sumy prawdopodobieństw dla poszczególnych elementów zbioru

ru rozmytego, tak jak miało to miejsce w poprzednio opisywanych metodach [54, 87, 122, 129]. W odróżnieniu do wcześniej opisywanego podejścia w systemie RIMER, projektowane elementy ogniskowe związane są tylko z jedną z konkurujących diagnoz (są dla niej charakterystyczne). Proces wspomaganie diagnozy w zaproponowanym podejściu związany jest z wyznaczaniem wartości miary przekonania osobno dla konkurujących diagnoz i wybrania diagnozy o najwyższej wartości funkcji przekonania. W późniejszych pracach rozszerzono dyskusję na temat wspomaganie diagnozy przy użyciu teorii Dempstera-Shafera o sposobie tworzenia funkcji przynależności w przesłankach reguł [108], progowego doboru reguł nie tylko w procesie wnioskowania ale także w procesie obliczania bazowego prawdopodobieństwa [109], a także odmiennego podejścia do łączenia informacji w procesie wspomaganie diagnozy [110].

W [101, 127] zbadano przydatność we wspomaganie decyzji tzw. przedziałów przekonania, na które składa się wartość miary przekonania oraz domniemania. Należy zauważyć, że podejmowanie decyzji na ogół opiera się na bardziej krytycznej mierze przekonania, a miara domniemania jest najczęściej uwzględniana podczas analizy pewności podjętej decyzji. W [101] P. Sevastjanov postuluje aby wynikiem porównania dwóch przedziałów czy też liczb rozmytych nie była jedna wartość ale pewien przedział. Autor określa motodykę wyznaczania wspomnianych przedziałów. Podejście to wykorzystano m.in. we wspomaganie decyzji giełdowych [36] w systemie opartym na regułach. Dodatkowo w pracy [102] opisano ten model wnioskowania na przykładzie wspomaganie diagnozy oraz zaproponowano metodę łączenia informacji pochodzącej z różnych źródeł, m.in. wtedy kiedy jest ona reprezentowana przez nierozmyte, bądź rozmyte przedziały liczbowe. W tym modelu podobnie jak w podejściu RIMER, tworzone reguły mogły zawierać w konkluzji łączoną odpowiedź związaną z konkurującymi ze sobą diagnozami. Tym samym nie były to reguły dedykowane dla konkretnej diagnozy.

Powyższe podejścia dostarczają narzędzi, dzięki którym możliwe jest uwzględnienie niedoskonałości danych przy wspomaganie decyzji za pomocą reguł wnioskowania. Różnią się one zarówno opisem reguł jak i podejściem do przetwarzania danych nieprecyzyjnych i niepewnych. Aby użyć powyższych narzędzi do zagadnień wspomaganie diagnostyki medycznej koniecznym wydaje się wypracowanie sposobu tworzenia bazy reguł diagnostycznych. Mimo, że niektóre z podejść podejmowały również temat reprezentacji wiedzy w regułach warunkowych [125, 126], to w niewystarczającym stopniu odniosły się do zagadnienia tworzenia zbioru reguł. W pierwszych próbach tworzenia bazy reguł wykorzystywano bezpośrednio wiedzę eksperta ludzkiego czyli np. lekarza diagnostyka. Z uwagi na wspomniane wcześniej cechy wiedzy ludzkiej, rozwijano także podejścia do automatycznego określenia reguł z danych, jak i podejść łączących wiedzy eksperta z wiedzą ukrytą w danych [80].

Jednym z pomysłów wartych uwagi są próby przekształcenia parametrów klasyfikatorów w formę reguł. Dobrze sprawdzają się podejścia związane z najpopularniejszymi klasyfikatorami: opartymi na maszynie wektorów podtrzymujących, drzewach decyzyjnych lub sztucznych sieciach neuronowych [12, 35, 48, 56, 117]. Takie podejście zastosowano przy przekształceniu klasyfikatora neuronowego (wielowarstwowej sieci neuronowej) do postaci rozmytych oraz klasycznych reguł warunkowych [34]. Podejście to testowano na wielu

benchmarkowych bazach danych, również z dziedziny medycznej. Otrzymywano reguły warunkowe odnoszące się bezpośrednio do porównywania wybranych zmiennych w danej bazie. Wyniki, które okazały się być konkurencyjne w stosunku do metod referencyjnych, udowodniły, że możliwe jest wydobycie łatwych do interpretacji reguł wnioskowania zapewniające satysfakcjonującą niezawodność. Działanie wielowarstwowej sieci neuronowej nie jest trywialne z punktu widzenia użytkownika systemu wspomagania diagnozy, lecz mimo to, przedstawione rozwiązanie stanowi częsty punkt odniesienia.

Interesującym narzędziem, które pozwala tworzyć zbiory reguł są także drzewa decyzyjne [23, 68, 116, 128]. W [67] tworzono drzewo decyzyjne będące zbiorem reguł wykorzystujących tzw. aksjomatyczne zbiory rozmyte i nowy typ funkcji zgodnej przynależności. Tworzone drzewo reprezentuje na początku zestaw wszystkich reguł, które są dopasowane do danych na poziomie założonej wartości progowej. Pozwala to na wstępne wykluczenie reguł niezgodnych z danymi bądź nieistotnych, nie wpływających na wnioskowanie. Kluczowym etapem jest dalsza iteracyjna metoda eliminacji reguł w drzewie, polegająca na próbnym usuwaniu pojedynczej reguły i sprawdzeniu jakości klasyfikacji dla pozostałego zbioru. Taka metoda z pewnością stanowi wyzwanie z punktu widzenia złożoności obliczeniowej, niemniej jednak dostarcza zbiór możliwych do interpretacji reguł warunkowych i zwraca uwagę na aspekt kompromisu pomiędzy złożonością oraz efektywnością zbioru reguł. Dodatkowo, w ten sposób unika się konieczności żmudnego heurystycznego przeszukiwania.

Jeżeli inżynier wiedzy chce tworzyć zarówno reguły z prostymi i złożonymi przesłankami to ich liczba rośnie wykładniczo wraz z liczbą opisanych symptomów w danym zagadnieniu medycznym. Tworzenie reguł często sprowadza się najpierw do tzw. selekcji cech. Z punktu widzenia wspomagania decyzji diagnostycznych ta selekcja cech związana jest z ograniczeniem liczby symptomów w bazie tak aby móc zachować odpowiednią efektywność wspomagania diagnozy.

Poszukiwanie najlepszego zbioru symptomów a tym samym i reguł wynikających z danych wymaga osiągnięcia dwóch celów: efektywności systemu wspomagania decyzji i jego dostatecznie małej złożoności. Idealnym przypadkiem jest sytuacja gdy osiągnięto bardzo wysoką efektywność w odwzorowaniu wartości wybranych symptomów na poprawną diagnozę przy jak najmniejszej liczbie reguł. Osiąganie wysokiej efektywności i jednocześnie małej złożoności jest niestety kwestią kompromisu. Sama ocena efektywności jest najczęściej określana w prosty sposób. Oblicza się na przykład procent poprawnie rozpoznanych przypadków wśród wszystkich rozpoznawanych. Tę samą informację niesie podawana wprost liczba błędnie podjętych decyzji. Natomiast dla oceny złożoności systemu wspomagania diagnozy, a więc złożoności zbioru reguł bierze się pod uwagę takie wielkości jak: liczba reguł, liczba warunków występujących w przesłance reguły (wpływa ona na czytelność reguły), liczba wybranych symptomów, pokrycie dziedziny wartości symptomów przez reguły, liczba reguł pokrywających się i sprzecznych [42]. Pogodzenie wysokiej efektywności z małą złożonością zbioru reguł diagnostycznych wiąże się z obliczeniem dodatkowej wielkości, która może być współczynnikiem jednocześnie uwzględniającym ocenę efektywności i ocenę złożoności [53, 67]. Niemniej jednak, nie powstała dotąd metoda oceny czytelności systemu, która byłaby uznana za standard [42].

Wraz z rozwojem narzędzi inteligencji obliczeniowej, rośnie liczba algorytmów przeszukiwania proponowanych dla wydobywania reguł wnioskowania i wyboru symptomów. W [24] dokonano pewnego przeglądu rozwiązań w dziedzinie selekcji nie tylko symptomów ale i ogólnie cech w zagadnieniach rozpoznawania. Można wyróżnić dwa zasadnicze podejścia do problemu. Algorytmy z pierwszej grupy są określone jako metody filtracyjne. Są one stosowane niezależnie od procesu wnioskowania w systemie. Metody te oceniają parametry bazy danych według pewnych współczynników takich jak korelacja [46] czy wskaźnik wspólnej informacji [14]. Oba kryteria pozwalają wykryć pary symptomów, z których można usunąć jeden bez zmiany opisu bazy, albo można wskazać symptomy najbardziej istotne tj. najlepiej reprezentujące daną diagnozę. Zaletą takich metod jest z pewnością mała złożoność obliczeniowa. Istnieje jednak ryzyko przeoczenia istotnego wpływu jednego bądź kombinacji symptomów w przesłance reguły na ostateczny proces wnioskowania.

Do drugiej grupy metod zalicza się podejścia, które sprawdzają wpływ wyboru symptomów na wynik uczenia systemu, a także na wynik jego testowania. Nie jest trudno zauważyć, że metody te mogą w bardziej szczegółowy sposób ocenić wybraną grupę symptomów. Niektóre z nich poza wyborem symptomów przeprowadzają także strojenie dodatkowych parametrów wpływających na proces rozpoznawania. Wśród nich dwoma najpopularniejszymi podejściami są metody sekwencyjne [84] oraz metody heurystyczne [3]. Wiele z takich metod implementowanych jest w pakietach uczenia maszynowego KEEL [4], WEKA [40] czy w aplikacji *Classification Learner* z środowiska MATLAB [73].

Z powyższego przeglądu literatury, wynika, że mimo iż autorzy dostrzegają konieczność ograniczania liczby i złożoności reguł oraz potrzebę wskazania efektywnych sposobów radzenia sobie z danymi niepewnymi, to jednak często metody te traktują dane medyczne na równi z innymi danymi popularnymi przy testowaniu proponowanych metod klasyfikacji. Niestety zagadnienie wspomagania diagnostyki medycznej nie może być traktowane na równi ze zwykłą klasyfikacją. W szczególności, jak już nadmieniono, rzadko jest poruszany problem wspomagania diagnozy, w którym priorytetem jest współpraca projektanta systemu i lekarza. Takie ujęcie problemu warunkuje samodzielne i efektywne użytkowanie wdrażanego systemu przez lekarza. Tylko niektóre prace kierują szczególną uwagę na interpretowalność klasyfikatora opartego na regułach [42, 44, 70]. Tymczasem, powyższe zadanie wymaga odrębnego podejścia do reprezentacji danych, modelu wnioskowania oraz sposobu poprawy jakości diagnozowania przy ograniczaniu liczby reguł w bazie wiedzy, uwzględniają powyższy cel. Takie podejście reprezentuje niniejsza rozprawa.

1.2 Cel i teza rozprawy

Celem pracy jest zaproponowanie i przetestowanie metod konstrukcji reguł diagnostycznych na podstawie medycznych baz danych z uwzględnieniem miar niepewności i nieprecyzyjności. Spełnienie powyższego celu związane jest z następującymi warunkami:

- Konstrukcja zbioru reguł musi być poprzedzona staranną i zrozumiałą oceną jakości symptomów i ich wpływu na diagnozę.
- Stworzone reguły muszą być łatwe w interpretacji przez eksperta.
- Należy zapewnić czytelność proponowanego schematu wnioskowania.
- Opracowane metody muszą być wystarczająco skuteczne nawet dla niepełnych danych.

W zastosowaniu do wspomagania podejmowania decyzji, a w szczególności do wspomagania diagnozy, metody wyszczególnione w punkcie 1.1 często nie poddają się porównaniu. Każdą z nich wyróżnia odmienny sposób definiowania niepewności i/lub nieprecyzyjności. Na podstawie przeprowadzonych już badań, których wyniki są zawarte w opublikowanych przez autora pracach, teoria Dempstera-Shafera została wybrana jako metoda dostarczająca wygodnego mechanizmu wspomagania diagnozy tj. miary przekonania. Natomiast teoria zbiorów rozmytych oraz podstawy teorii możliwości zostaną wykorzystane jako sprawdzone narzędzia przetwarzania informacji nieprecyzyjnej. Rozszerzenie teorii Dempstera-Shafera na rozmyte elementy ogniskowe pozwala aby sprawdzony mechanizm wnioskowania w obliczu niepewności mógł zostać wzbogacony o aspekt przetwarzania czynnika nieprecyzyjności. Nieprecyzyjność będzie reprezentowana za pomocą funkcji przynależności i użyta do modelowania objawów. Dodatkowo, pojęcie miary możliwości zostanie wykorzystane do określenia dopasowania rozmytego elementu ogniskowego w przesłance reguły i informacji wejściowej jaką są dane na temat pacjenta. Dotychczasowe podejście do wyznaczania miary przekonania wykorzystujące miarę możliwości sprowadza się do sprawdzenia zgodności elementu ogniskowego i danych pacjenta lepszej od przyjętego progu. W niniejszej pracy zaproponowana i przetestowana zostanie modyfikacja obliczania miary przekonania. Wartość miary możliwości będzie użyta bezpośrednio tzn. do skalowania wartości prawdopodobieństwa elementów ogniskowych zgodnie z ich dopasowaniem do danych diagnozowanego pacjenta. Przedstawienie proponowanego modelu wspomagania diagnozy stanowi jeden element niniejszej rozprawy. Drugim, równie kluczowym aspektem, będzie analiza sposobu wyodrębniania rozmytych reguł warunkowych z danych medycznych. W punkcie 1.1 wspomniano o podejściach do selekcji reguł i cech w ogólnych zagadnieniach klasyfikacji. Wśród nich najbardziej obiecującą grupę stanowią metody sekwencyjne. Aby zapewnić odpowiednią zrozumiałość systemu wspomagania decyzji diagnostycznej i umożliwić przyszłą współpracę z ekspertem medycznym, koniecznym wydaje się aby sam mechanizm wydobywania reguł mógł zostać zinterpretowany i poddany krytyce. W pracy zostanie przedstawiony tok myślowy prowadzący do zaproponowania kolejnych algorytmów opartych na iteracyjnym

wydobywaniu reguł diagnostycznych. Część z metod [10, 17, 43, 44, 50, 118] implementowanych w pakiecie uczenia maszynowego [4] będzie potraktowana jako metody referencyjne do proponowanych w rozprawie.

Rozprawa zmierza do weryfikacji prawdziwości następującej tezy:

Wykorzystanie teorii Dempstera-Shafera i teorii zbiorów rozmytych poprzez zastosowanie iteracyjnych algorytmów wyodrębniania rozmytych reguł warunkowych na podstawie danych uczących poprawia jakość wspomaganie diagnozy medycznej. Jakość jest rozumiana jako efektywność uogólniania oraz możliwie mała złożoność uzyskanego zbioru rozmytych reguł warunkowych.

Aby wykazać słuszność powyższej tezy zrealizowano następujące zadania:

1. Opracowanie zmodyfikowanego algorytmu eliminacji reguł, w którym zbiór elementów ogniskowych powstaje z pewnego początkowego i nadmiarowego zbioru poprzez iteracyjne usuwanie elementów ogniskowych (przesłanek reguł diagnostycznych) z najniższą wartością bazowego prawdopodobieństwa.
2. Zaproponowanie metody wyznaczania kształtów funkcji przynależności dla elementów ogniskowych i zbadanie ich wpływu na efektywność algorytmu eliminacji reguł.
3. Zaproponowanie kryterium (wskaźnika) oceny reguły diagnostycznej, który pozwoli, w wyniku działania proponowanych algorytmów, otrzymać mniejszą liczbę reguł z zachowaniem zbliżonej efektywności wyznaczania diagnozy w porównaniu z dotychczasowym kryterium w postaci wartości bazowego prawdopodobieństwa.
4. Opracowanie algorytmu selekcji reguł, polegającego na iteracyjnym wyborze najlepszych elementów ogniskowych z ich założonego zbioru początkowego.
5. Modyfikacja algorytmu selekcji reguł poprzez warunkowe usuwanie reguł, które nie poprawiają efektywności diagnozy tj. uzyskanie algorytmu warunkowej selekcji reguł.
6. Wprowadzenie wskaźnika oceny wyniku działania proponowanych algorytmów wydobywania reguł diagnostycznych. Wskaźnik ten ma uwzględniać kompromis pomiędzy efektywnością wyznaczania poprawnej diagnozy a złożonością zbioru reguł diagnostycznych.
7. Wykorzystanie do wspomaganie diagnozy zmodyfikowanej miary przekonania, łączącej bezpośrednio wartość bazowego prawdopodobieństwa elementów ogniskowych z wartością ich dopasowania do danych diagnozowanego pacjenta.

Wymienione wyżej zadania zostały wykonane z wykorzystaniem kilku znanych porównawczych baz medycznych oraz jednej bazy danych zebranej przez lekarzy praktyków i wykorzystywanej w zagadnieniach klasyfikacji danych i selekcji cech przez pracowników Instytutu Informatyki Uniwersytetu Śląskiego w Katowicach [64, 97].

1.3 Treść rozprawy

Pierwszy rozdział pracy przybliży podjętą problematykę, przedstawia zakres tematyczny oraz tezę rozprawy. W rozdziale drugim przybliżone zostają klasyczne metody przetwarzania niepewności, szczególnie istotne z punktu widzenia wspomaganie diagnozy medycznej.

Rozdział trzeci rozpoczyna zasadniczą część rozprawy, w której najpierw przedstawiona zostaje wybrana metoda konstruowania reguł warunkowych oparta na teorii Dempstera-Shafera oraz analiza sposobów ich oceny z uwzględnieniem jakości wspomaganie diagnozy medycznej. Równocześnie zaproponowano nowy wskaźnik oceny reguł. W dalszej części pracy wyszczególniono proponowane podejścia do wydobywania reguł z danych medycznych. Zaproponowane algorytmy wydobywania reguł diagnostycznych stanowią oryginalny pomysł autora i są zilustrowane z użyciem przykładowej bazy. Te zagadnienia stanowią rozdział czwarty. Rozdział piąty poświęcono na przedstawienie wyników przeprowadzonych badań. Zilustrowano działanie proponowanych metod wspomaganie diagnozy oraz algorytmów a także porównano jakość wspomaganie diagnozy z podejściami referencyjnymi.

Zasadniczą część rozprawy podsumowuje rozdział szósty, w którym zawarte są najważniejsze wnioski płynące z próby potwierdzenia postawionej tezy. Jako dodatek umieszczono dokładny opis użytych baz benchmarkowych wraz z przedstawieniem przykładowych zbiorów reguł i funkcji przynależności, które razem stanowią bazę wydobytej wiedzy. Na końcu pracy wyszczególniono spis literatury.

Rozdział 2

Reprezentacja niepewności i nieprecyzyjności wiedzy medycznej

Normy diagnostyczne tradycyjnie modelowane są przy pomocy klasycznych zbiorów. Mamy z nimi do czynienia np. w wytycznych publikowanych jako wyniki wieloletnich obserwacji i badań [9,37]. Poniższy przykład zobrazuje, że normy medyczne mogą też wykazywać cechę stopniowania czy też różnicowania zagadnienia medycznego na więcej kategorii diagnostycznych niż samo odróżnienie stanu choroby od zdrowia. Po przedstawieniu definicji zbioru rozmytego okaże się, że dobrze nadaje się on do modelowania zagadnień diagnostycznych.

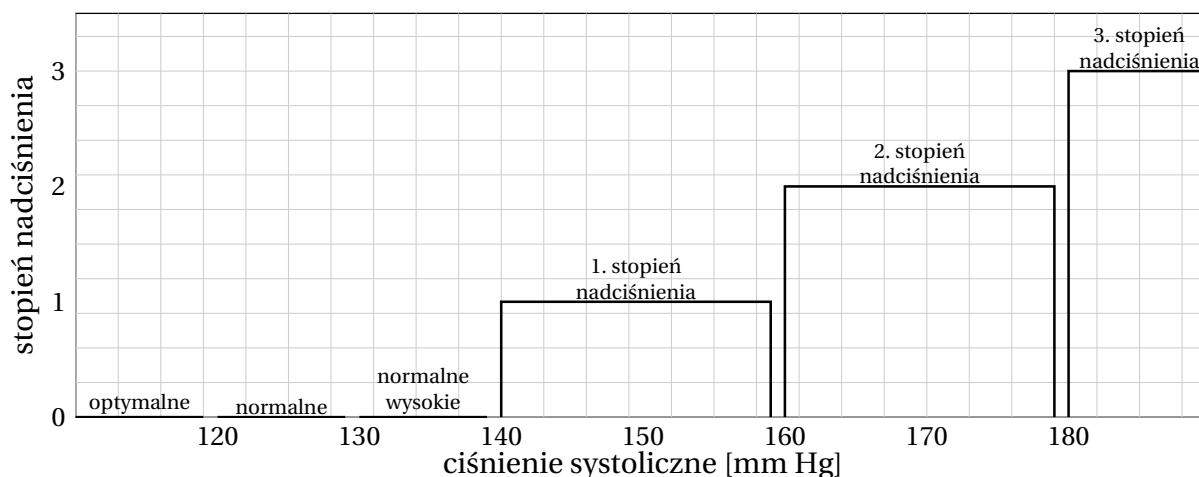
Wytyczne przedstawione przez ESH (ang. *European Society of Hypertension* – Europejskie Towarzystwo Nadciśnienia Tętniczego) i ESC (ang. *European Society of Cardiology* – Europejskie Towarzystwo Kardiologów) określają klasyfikację nadciśnienia tętniczego wyróżniającą siedem kategorii [37]. Dane te przedstawiono w tabeli 2.1.

Tabela 2.1: Kategorie diagnostyczne nadciśnienia tętniczego na podstawie wytycznych ESH-ESC [37]

kategoria	ciśnienie systoliczne [mm Hg]	ciśnienie diastoliczne [mm Hg]
ciśnienie optymalne	< 120	< 80
ciśnienie normalne	120–129	80–84
ciśnienie normalne wysokie	130–139	85–89
1. stopień nadciśnienia (łagodny)	140–159	90–99
2. stopień nadciśnienia (umiarkowany)	160–179	100–109
3. stopień nadciśnienia (ciężki)	≥ 180	≥ 110
nadciśnienie izolowane systoliczne	≥ 140	< 90

Tabela 2.1 wyróżnia trzy kategorie diagnostyczne pacjentów chorych na nadciśnienie i trzy kategorie pacjentów, których ciśnienie nie jest niepokojąco wysokie. W tabeli odróżniony został stan nietypowy tzn. taki, w którym u pacjenta występuje wysokie ciśnienie systoliczne z jednocześnie dobrym wynikiem ciśnienia diastolicznego. Poza tym stanem, wszystkie diagnozy mają ściśle określone i wzajemnie rozdzielone przedziały wartości ciśnień. Zaprezentowane wytyczne wskazują na stopniowanie stanów pacjenta (rysunek 2.1). Niemniej

jednak, takie jednoznaczne rozdzielanie kategorii diagnostycznych może być zgubne. Zaliczenie pacjenta z wynikiem pomiaru 139/89 do diagnozy „ciśnienie normalne wysokie” jest dyskusyjne, bo osiągnięcie wyniku większego o 1 mm Hg dla któregośkolwiek z ciśnień spowodowałoby (według wytycznych [37]) zakwalifikowanie pacjenta do diagnozy wskazującej na 1. stopień nadciśnienia. Przed taką nieściśłością chronić może modelowanie kategorii diagnostycznych za pomocą zbiorów rozmytych.



Rysunek 2.1: Kategorie diagnostyczne ESC-ESH [37] wskazujące na stopniowanie nadciśnienia tętniczego

2.1 Teoria zbiorów rozmytych

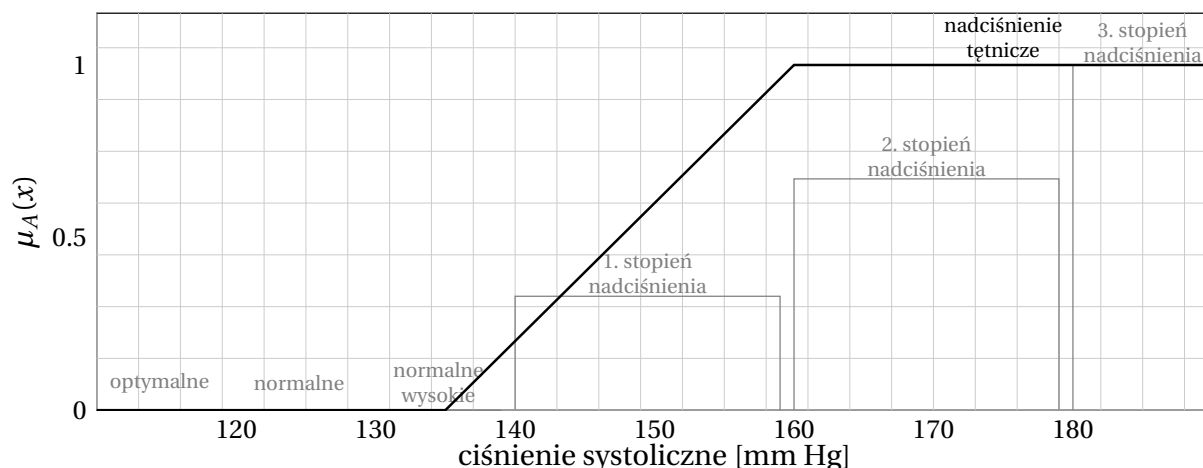
Zbiór rozmyty daje możliwość jednoczesnego określenia elementów należących do zbioru wraz ze stopniem przynależności każdego elementu do zbioru. Pojęcie zbioru rozmytego A zaproponowane przez Zadeha można przedstawić jako zbiór uporządkowanych par [130]

$$A = \{(\mu_A(x), x) \mid \forall x \in X\}, \quad (2.1)$$

gdzie x to element w zbiorze X , a $\mu_A(x)$ to funkcja przynależności odwzorowująca elementy zbioru X na stopień przynależności do zbioru A , w przedziale od 0 do 1, czyli

$$\mu_A(x) \in [0, 1]. \quad (2.2)$$

Na rysunku 2.2 przedstawiony został przykładowy zbiór rozmyty, który może subiektywnie być zinterpretowany jako wartość lingwistyczna „nadciśnienie tętnicze”. Zbiór rozmyty z rysunku 2.2 dla dyskretnych wartości pomiaru ciśnienia może zostać łatwo zapisany jako zbiór uporządkowanych par $\{(\mu_A(x), x)\}$. Taki zapis jest natomiast niemożliwy dla przestrzeni ciągłej. Jednak określenie funkcji przynależności $\mu_A(x)$ daje możliwość jednoznacznego zdefiniowania jej kształtu. Zbiór rozmyty z rysunku 2.2 można zdefiniować następującą funkcją



Rysunek 2.2: Zbiór rozmyty „nadciśnienie tętnicze” odpowiadający kategoriom diagnostycznym z rysunku 2.1

przynależności

$$\mu_A(x) = \begin{cases} 0 & \text{dla } x \leq 135, \\ \frac{x-135}{25} & \text{dla } 135 < x \leq 160, \\ 1 & \text{dla } x > 160. \end{cases} \quad (2.3)$$

Wracając do omawianego przykładu nietrudno zauważyć, że za pomocą jednego zbioru rozmytego możliwe jest uzyskanie modelu, który pozwala określać stopień, w jakim wartość pomiaru pasuje do diagnozy „nadciśnienie tętnicze”. Wartość 160 mm Hg ciśnienia systolicznego (początek przedziału wskazującego na kategorię „nadciśnienie umiarkowane” w tabeli 2.1) posłużyła tu jako punkt, od którego wyższe wyniki pomiarów w pełni przynależą do zbioru. Dodatkowo, użycie funkcji przynależności $\mu_A(x)$ pozwala odejść od ostrego przejścia pomiędzy kategoriami diagnostycznymi. Szczególnie istotne jest to np. dla wartości podobnych do granicznej, wspomnianej wcześniej, tzn. 139 mm Hg ciśnienia systolicznego. Ta wartość należy do zbioru rozmytego „nadciśnienie tętnicze” ze stopniem prawie 0.3. Interpretacja tego wyniku jest teraz jednoznaczna. tzn. wartość 139 mm Hg wskazuje już w pewnym stopniu na nadciśnienie tętnicze i wymaga uwagi. Powyższy przykład jest oczywiście subiektywny, natomiast pozwala dobrze zilustrować intuicyjną interpretację modelowania wiedzy medycznej za pomocą funkcji przynależności.

Poza opisaniem zbioru rozmytego funkcją przynależności (2.2) możliwe jest scharakteryzowanie go kilkoma następującymi wielkościami [69]:

- nośnik zbioru A

$$\text{Supp}(A) = \{x \in X \mid \mu_A(x) > 0\}, \quad (2.4)$$

- rdzeń zbioru A

$$\text{Core}(A) = \{x \in X \mid \mu_A(x) = 1\}, \quad (2.5)$$

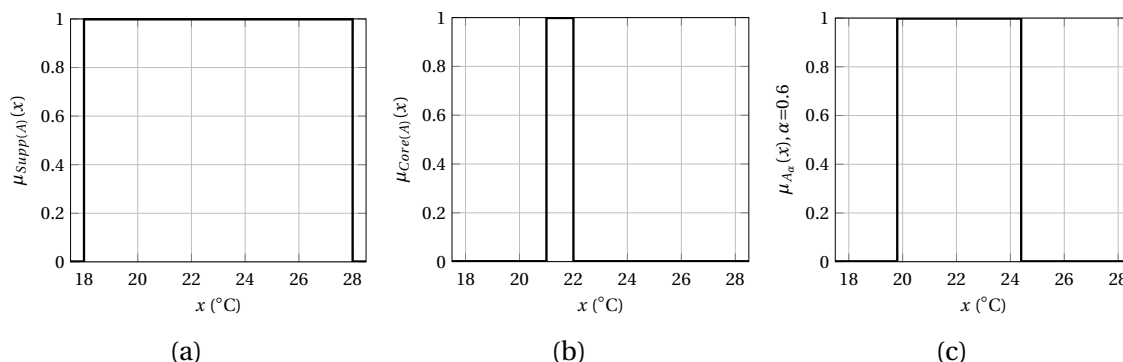
- α -przekrój zbioru A

$$A_\alpha = \{x \in X | \mu_A(x) \geq \alpha\}, \quad (2.6)$$

- ostry α -przekrój zbioru A

$$A_{\underline{\alpha}} = \{x \in X | \mu_A(x) > \alpha\}. \quad (2.7)$$

Przykładowe kształty funkcji charakterystycznych wyznaczających nośnik, rdzeń i α -przekrój zbioru A opisanego w (2.3) przedstawiono na rysunku 2.3.



Rysunek 2.3: Klasyczne zbiory utworzone jako : (a) nośnik, (b) rdzeń, (c) α -przekrój ($\alpha = 0.4$) zbioru A przedstawionego na rysunku 2.2

2.1.1 Rodzaje funkcji przynależności

W literaturze do opisu zbiorów rozmytych stosowanych jest kilka rodzajów funkcji przynależności. Każda z nich może dać odmienny wynik przy przetwarzaniu danych nieprecyzyjnych oraz inaczej być interpretowana przez użytkownika. Można także określić dla każdej funkcji czy jest ciągła, monotoniczna, zdefiniowana na skończonym przedziale czy nieskończonej dziedzinie. Poprzednio funkcję przynależności do zbioru A reprezentowano jako $\mu_A(x)$. Tym razem w indeksie dolnym oznaczenia zawarto symboliczne oznaczenie kształtu funkcji.

1. *Funkcje przynależności przedziałowo-liniowe* [93]. Ten rodzaj funkcji przynależności, nazywany także klasą funkcji wielokątnych, jest najprostszy ze względu na liniowy przebieg funkcji przynależności w danych przedziałach. Uznawane są one za funkcje najprostsze także ze względu na łatwość interpretacji i modyfikacji przez użytkownika. Poniżej opisano dwie najpopularniejsze.

a) *Trójkątna*

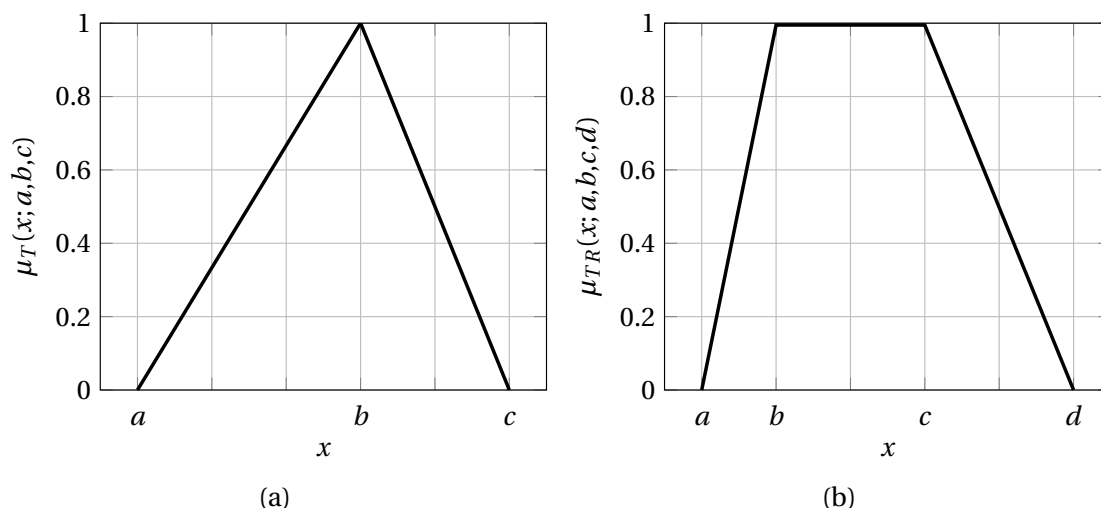
$$\mu_T(x; a, b, c) = \begin{cases} 0, & x \leq a, \\ \frac{x-a}{b-a}, & a < x \leq b, \\ \frac{c-x}{c-b}, & b < x \leq c, \\ 0, & x > c, \end{cases} \quad (2.8)$$

gdzie a , b i c definiują liniowe przedziały funkcji. Wykres funkcji dla przykładowych parametrów przedstawiono na rysunku 2.4a.

b) Trapezowa

$$\mu_{TR}(x; a, b, c, d) = \begin{cases} 0, & x \leq a, \\ \frac{x-a}{b-a}, & a < x \leq b, \\ 1, & b < x \leq c, \\ \frac{d-x}{d-c}, & c < x \leq d, \\ 0, & x > d, \end{cases} \quad (2.9)$$

gdzie podobnie jak w funkcji trójkątnej, a , b , c i d definiują liniowe przedziały funkcji. Wykres funkcji trapezowej dla przykładowych parametrów przedstawiono na rysunku 2.4b. Aby funkcja miała kształt trapezu musi zostać zachowany warunek, że $a \leq b < c \leq d$.



Rysunek 2.4: Przedziałowo-liniowe funkcje przynależności: (a) trójkątna, (b) trapezowa

Nietrudno zauważyć, że przyrównanie wartości b i c spowoduje, że uzyskany zostanie kształt trójkątny. Można więc uznać, że funkcja trapezowa stanowi uogólnienie dla trójkątnej funkcji przynależności. Dodatkowo, przyrównanie ze sobą parametrów określające nachylenie ramion trapezu ($a = b$ oraz $c = d$) przekształci trapezową funkcję przynależności w funkcję charakterystyczną określającą przynależność do klasycznego zbioru (przedziału). Do głównych zalet tego rodzaju funkcji przynależności zalicza się wspomnianą prostotę w interpretacji. Jednoznaczne odniesienie parametrów funkcji do jej kształtu ułatwia także modyfikację kształtu tych funkcji. Jedyną niedogodnością wspomnianą w literaturze [93] jest obecność punktów nieciągłości pochodnej funkcji w punktach określonych przez a , b , c oraz d . Takie punkty nieciągłości pochodnych mogą stwarzać wyzwanie obliczeniowe dla wykorzystujących je metod, które optymalizują kształty funkcji [69].

2. Gaussowskie funkcje przynależności

a) Symetryczna gaussowska funkcja przynależności [69]

$$\mu_G(x; \sigma, c) = e^{-\frac{(x-c)^2}{2\sigma^2}}, \quad (2.10)$$

gdzie c jest środkiem funkcji (wartością średnią), a σ to odchylenie standardowe. Przykładowa gaussowska funkcja przynależności przedstawiona jest na rysunku 2.5a. Łącząc łatwość definicji trójkątnej i ciągłość gaussowskiej funkcji przynależności, w [69] zaproponowano przekształcenie jednego typu funkcji na drugi. Taki zabieg ułatwia automatyczną modyfikację funkcji przynależności jeżeli istnieje potrzeba wyznaczania pochodnych funkcji.

b) Niesymetryczna gaussowska funkcja przynależności [93]

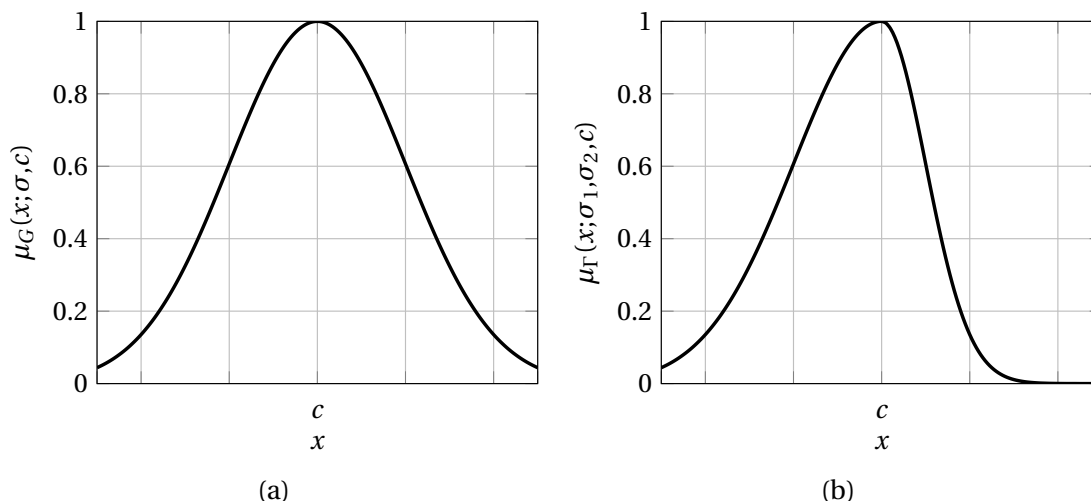
Kolejna funkcja gaussowska, nazywana także gaussowską łączoną funkcją przynależności, pozwala na osobne określenie lewej i prawej części jej kształtu. Tym samym zależność opisująca gaussowską łączoną funkcję przynależności można przedstawić jako

$$\mu_\Gamma(x; \sigma_1, \sigma_2, c) = t \cdot e^{-\frac{(x-c)^2}{2\sigma_1^2}} + (1-t) \cdot e^{-\frac{(x-c)^2}{2\sigma_2^2}}, \quad (2.11)$$

gdzie

$$t = \begin{cases} 1 & \text{dla } -\infty < x \leq c, \\ 0 & \text{dla } c < x < +\infty. \end{cases} \quad (2.12)$$

Wszystkie funkcje gaussowskie określają pewną niezerową wartość przynależności dla wszystkich wartości rzeczywistych (rysunek 2.5). Dzięki temu, z jednej strony są różniczkowalne w całej dziedzinie wartości, ale z drugiej strony nie można określić zamkniętego przedziału stosowalności użytej funkcji gaussowskiej.



Rysunek 2.5: Gaussowskie funkcje przynależności: (a) symetryczna, (b) łączona

3. Uogólniona dzwonowa funkcja przynależności [69]

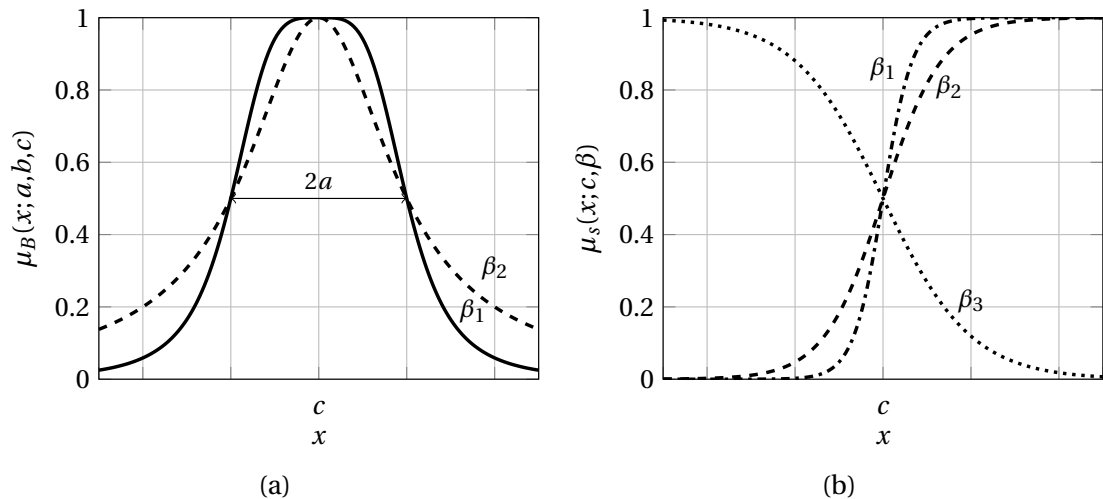
$$\mu_B(x; a, b, c) = \frac{1}{1 + \left| \frac{x-c}{a} \right|^{2b}}, \quad (2.13)$$

gdzie c określa środkową wartość funkcji (wartość średnią), a wpływa na szerokość funkcji a b na nachylenie zboczy funkcji. Przykładowa funkcja przynależności przedstawiona jest na rysunku 2.6a. Z jednej strony funkcja ta pozwala na uzyskanie szerokiego przedziału pełnej przynależności do zbioru natomiast może to skutkować uzyskaniem bardzo stromych przejść pomiędzy dużymi a małymi wartościami przynależności.

4. Sigmoidalna funkcja przynależności [69]

$$\mu_s(x; c, \beta) = \frac{1}{1 + e^{(-\beta(x-c))}}, \quad (2.14)$$

gdzie c określa punkt przynależący do zbioru rozmytego z wartością 0.5, a β to nachylenie funkcji. Funkcja ta jest monotoniczna: malejąca dla $\beta < 0$ i rosnąca dla $\beta > 0$. Ta funkcja stanowi alternatywę dla funkcji gaussowskiej, szczególnie jeśli chce się opisać skrajnie położone zbiory rozmyte, czyli te które opisują ekstremalne przedziały opisywanych wartości. Dodatkowo, parametr β umożliwia intuicyjnie łatwą reprezentację pojęć heurystycznych. Funkcja sigmoidalna dopiero w granicy przyjmuje wartości pełnej przynależności. Przykład funkcji sigmoidalnej przedstawiono na rysunku 2.6b.



Rysunek 2.6: Funkcje przynależności (a) uogólniona dzwonowa, (b) sigmoidalna, $\beta_1 > \beta_2 > \beta_3$

5. Harmoniczna funkcja przynależności [93]

$$\mu_H(x; a, b) = \begin{cases} 0 & \text{dla } x < (b-a), \\ 0.5 \cdot (1 + \cos(\pi \frac{x-b}{a})) & \text{dla } (b-a) \leq x \leq (b+a), \\ 0 & \text{dla } x > (b+a). \end{cases} \quad (2.15)$$

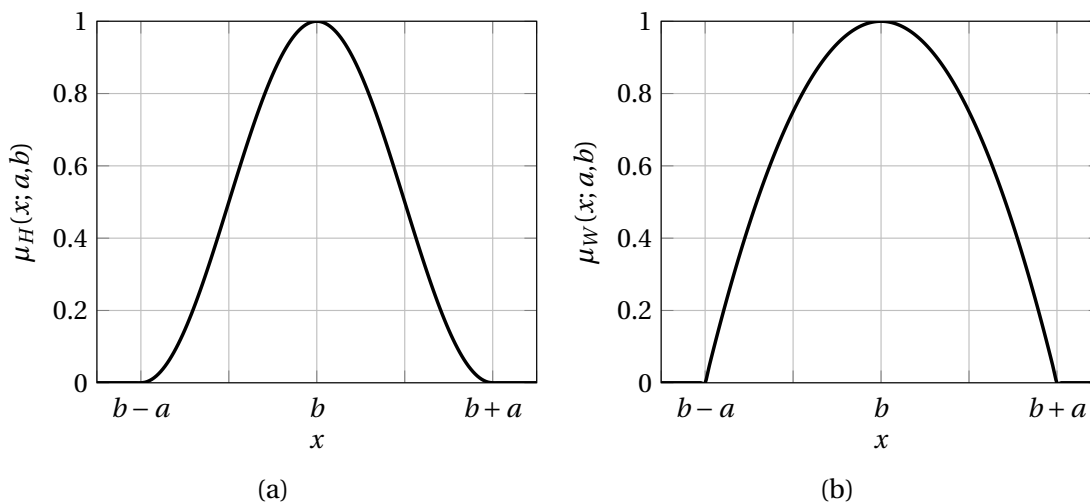
Jak można zauważyć na rysunku 2.7a funkcja ta jest odwzorowaniem pełnego okresu funkcji sinusoidalnej. W odróżnieniu do poprzednich funkcji gaussowskich oraz sigmoidalnej,

jej dziedziną (nośnikiem) jest przedział domknięty. Naturalnie, poprzez modyfikację (2.15) możliwe jest otwarcie lewo- bądź prawostronne funkcji harmoniczej.

6. Wielomianowa funkcja przynależności [93]

$$\mu_W(x; a, b) = \begin{cases} 0 & \text{dla } x < (b - a), \\ 1 - \left(\frac{x-b}{a}\right)^2 & \text{dla } (b - a) \leq x \leq (b + a), \\ 0 & \text{dla } x > (b + a). \end{cases} \quad (2.16)$$

Przykładowa wielomianowa funkcja przynależności jest przedstawiona na rysunku 2.7b. Ma ona podobne cechy jak poprzednia funkcja harmoniczna. Określenie przedziału funkcji też stanowi trywialne zagadnienie z punktu widzenia użytkownika. Identycznie jak poprzednio, możliwe jest otwarcie funkcji dla skrajnie małych i dużych wartości x .



Rysunek 2.7: Funkcje przynależności: (a) harmoniczna, (b) wielomianowa

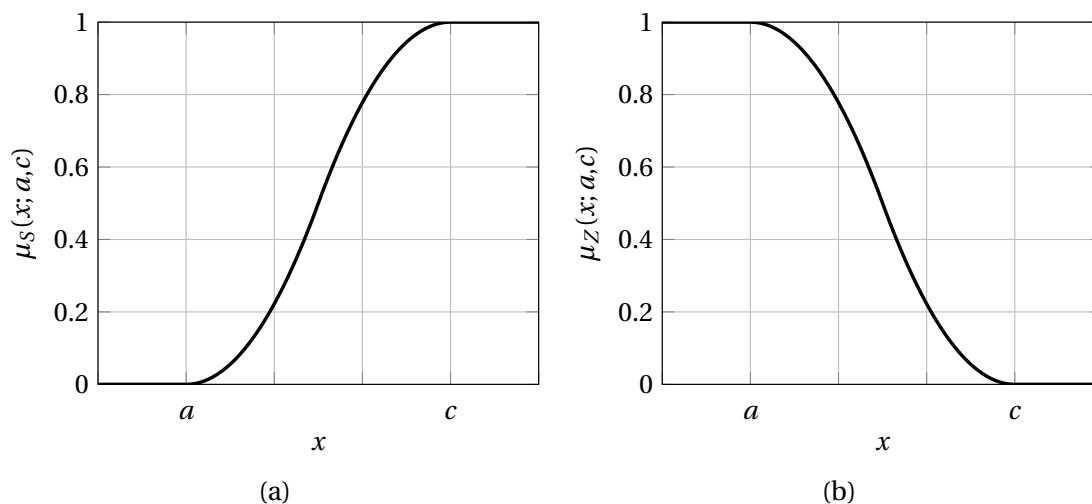
7. Funkcje przynależności typu S i Z [69]

$$\mu_S(x; a, c) = \begin{cases} 0 & \text{dla } x \leq a, \\ 2\left(\frac{x-a}{c-a}\right)^2 & \text{dla } a < x \leq b, \\ 1 - 2\left(\frac{x-c}{c-a}\right)^2 & \text{dla } b < x \leq c, \\ 1, & \text{dla } x > c, \end{cases} \quad (2.17)$$

gdzie a, c są parametrami a wartość $b = (a+c)/2$ określa punkt dla którego funkcja przyjmuje wartość 0.5 (rysunek 2.8a). Funkcja typu Z (rysunek 2.8b) stanowi dopełnienie funkcji S, tym samym jej wzór można określić jako

$$\mu_Z(x; a, c) = \begin{cases} 1 & \text{dla } x \leq a, \\ 1 - 2\left(\frac{x-a}{c-a}\right)^2 & \text{dla } a < x \leq b, \\ 2\left(\frac{x-c}{c-a}\right)^2 & \text{dla } b < x \leq c, \\ 0, & \text{dla } x > c. \end{cases} \quad (2.18)$$

Opisane wyżej funkcje, mimo że mają podobny przebieg do funkcji sigmoidalnej, charakteryzują się jedną istotną różnicą. Ich parametry a i c jednoznacznie wyrażają przedział zmienności od wartości 0 do 1. Dla wskazanych przez nie punktów funkcja S przyjmuje wartości odpowiednio 0 i 1, a funkcja Z, przeciwnie.

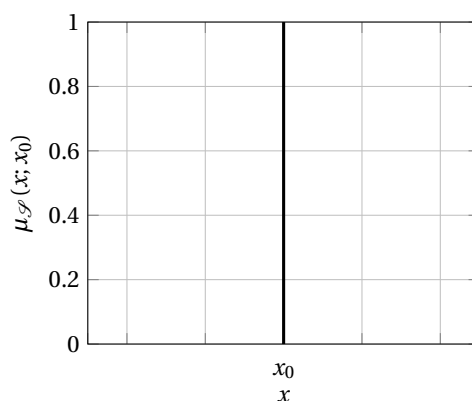


Rysunek 2.8: Funkcje przynależności: (a) typu S, (b) typu Z

8. Rozmyty singleton [69]

$$\mu_{\mathcal{F}}(x; x_0) = \begin{cases} 1 & \text{dla } x = x_0, \\ 0 & \text{dla } x \neq x_0, \end{cases} \quad (2.19)$$

gdzie x_0 określa jedyną wartość x o pełnej przynależności do zbioru. Ten typ funkcji przynależności interpretuje też pomiary lub zmienne nominalne. Funkcja ta daje możliwość przetwarzania informacji o wystąpieniu bądź niewystąpieniu pewnego zjawiska istotnego dla rozpatrywanego zagadnienia (rysunek 2.9).



Rysunek 2.9: Funkcja przynależności typu rozmyty singleton

Pomimo, że wymieniono niemałą liczbę klas funkcji przynależności, scharakteryzowano z matematycznego punktu widzenia ich zalety i wady to zauważa się, że najczęściej stosowanymi funkcjami przynależności są liniowo-przedziałowe i gaussowskie [2, 22, 65, 137].

Powodem stosowania tych pierwszych wydaje się być najprostszy sposób opisanie ich przez eksperta ludzkiego. Funkcje gaussowskie używane są najczęściej przy zagadnieniach ich automatycznego strojenia [70].

2.2 Teoria możliwości

Wprowadzenie teorii możliwości przez L. A. Zadeha [135] było następstwem jego słusznej obserwacji, że w świecie rzeczywistym podejmowanie decyzji związane jest z charakterystycznymi dla człowieka nieprecyzyjnymi stwierdzeniami i niepewną wiedzą. Jednocześnie, bardzo rzadko jest wykorzystywany dostępny aparat matematyczny oparty na klasycznej teorii prawdopodobieństwa. Wynika to m.in. z istotnych trudności w wyznaczaniu wartości prawdopodobieństwa niezbędnych do zastosowania w teorii. Natomiast możliwe jest użycie zbiorów rozmytych do modelowania zagadnień związanych z przetwarzaniem niepewnej i nieprecyzyjnej informacji. Na przykład, aby skorzystać z teorii prawdopodobieństwa przy wspomaganiu diagnozy, należy znać precyzyjne rozkłady prawdopodobieństwa dla pojedynczych symptomów oraz wszystkich ich kombinacji. Zwykle ten warunek nie może być spełniony. Tymczasem, ekspert, bazując na doświadczeniu, może zaproponować zbiory rozmyte dla reprezentacji nieprecyzyjności symptomów. Można również wyznaczyć funkcje przynależności z danych.

2.2.1 Rozkład możliwości

Mając do czynienia ze zmienną \mathcal{X} przyjmującą wartości x z przestrzeni \mathbb{X} , przykładowy rozkład możliwości można określić jako [132]

$$\Pi_{\mathcal{X}} = A, \tag{2.20}$$

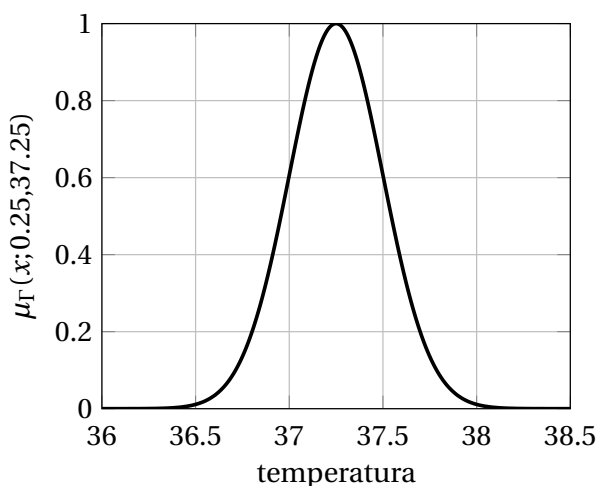
gdzie A to zbiór rozmyty opisany funkcją przynależności $\mu_A(x)$ przyporządkowującą elementom przestrzeni \mathbb{X} wartości z zakresu $[0,1]$. Rozkład możliwości $\Pi_{\mathcal{X}}$ opisuje możliwość, że zmienna x przyjmie wartość z przestrzeni \mathbb{X} , czyli określa stopień podobieństwa wartości x z przykładowym stwierdzeniem „ \mathcal{X} jest A ”. Funkcja przynależności $\mu_A(x)$ użyta do opisu rozkładu możliwości nazywana jest w literaturze funkcją rozkładu możliwości i oznaczana dla odróżnienia symbolem $\pi_{\mathcal{X}}(x)$ choć jest numerycznie równa $\mu_A(x)$ [132]. Interpretacja intuicyjna wartości funkcji rozkładu możliwości jest następująca:

- $\pi_{\mathcal{X}}(x) = 1$ oznacza całkowitą możliwość przyjęcia przez zmienną \mathcal{X} wartości $x \in \mathbb{X}$,
- $\pi_{\mathcal{X}}(x) = 0$ oznacza całkowitą niemożliwość przyjęcia przez zmienną \mathcal{X} wartości $x \in \mathbb{X}$,
- $0 < \pi_{\mathcal{X}}(x) < 1$ oznacza pewien stopień możliwości przyjęcia przez zmienną \mathcal{X} wartości $x \in \mathbb{X}$.

Aby objaśnić wprowadzone powyżej zależności i stwierdzenia, posłużono się przykładem. Oznaczmy $x \in \mathbb{X}$ jako wartość temperatury oraz zbiór rozmyty „stan podgorączkowy” jako A . Można go opisać następującą gaussowską funkcją przynależności

$$\mu_A(x) = \exp\left(-\frac{x - 37.25}{0.125}\right), \quad (2.21)$$

Rozkład możliwości dla zmiennej \mathcal{X} w tym wypadku to $\pi_{\mathcal{X}}(x) = \mu_A(x)$ (2.21). Jest on przedstawiony na rysunku 2.10. Stwierdzenie „pacjent jest w stanie podgorączkowym” jest rozmytym ograniczeniem nałożonym na wartość temperatury ciała. Dla przykładowej temperatu-



Rysunek 2.10: Pojęcie „stan podgorączkowy” opisane gaussowską funkcją przynależności

ry ciała pacjenta równej 37.6°C otrzymana zostaje wartość przynależności równa $\mu_A(37.6) = 0.38$. Uzyskując dla tej wartości temperatury przynależność równą 0.38 można stwierdzić, że stwierdzenia „temperatura ciała pacjenta jest równa 37.6°C ” i „pacjent jest w stanie podgorączkowym” są ze sobą zgodne w stopniu 0.38. Można także, rozumując w odwrotnym kierunku, określić, że dla stwierdzenia „pacjent jest w stanie podgorączkowym”, wartość 0.38 oznacza możliwość, że temperatura pacjenta wynosi 37.6 (lub 36.8°C).

2.2.2 Pojęcie miary rozmytej

Kluczową rolę w teorii możliwości i jej zastosowaniu pełnią tzw. miary możliwości i miary konieczności. Są to przykłady miar rozmytych. Aby jednoznacznie opisać wspomniane miary, koniecznym wydaje się zdefiniowanie w tym miejscu pojęcia miary rozmytej. Zachowując przyjęte oznaczenia dla zmiennej i przestrzeni, oznaczono dowolną klasę podzbiorów przestrzeni \mathbb{X} jako \mathcal{C} . Dodatkowo niech dwa zbiory $E, F \in \mathcal{C}$. Funkcja g odwzorowująca \mathcal{C} na wartości z przedziału $[0, +\infty]$ jest miarą rozmytą na $(\mathbb{X}, \mathcal{C})$ [114, 119] jeżeli spełnia następujące warunki:

W1 $g(\emptyset) = 0$, (zeruje się dla zbioru pustego),

W2 $E \in \mathcal{C}, F \in \mathcal{C}$ i $E \subseteq F$ to $g(E) \leq g(F)$, (monotoniczna),

W3 $\{E_n\} \subset \mathcal{C}, E_1 \subset E_2 \subset \dots$, i $\bigcup_{n=1}^{\infty} E_n \in \mathcal{C}$ to $\lim_n g(E_n) = g\left(\bigcup_{n=1}^{\infty} E_n\right)$ (ciągła od dołu),

W4 $\{E_n\} \subset \mathcal{C}, E_1 \supset E_2 \supset \dots$, $g(E_1) < \infty$ i $\bigcap_{n=1}^{\infty} E_n \in \mathcal{C}$ to $\lim_n g(E_n) = g\left(\bigcap_{n=1}^{\infty} E_n\right)$ (ciągła od góry).

Jeżeli g dla nieskończonej przestrzeni spełnia tylko warunki (W1, W2 i W3) to nazywa się ją półciągłą od dołu, a jeśli spełnia (W1, W2 i W4) to nazywa się ją półciągłą od góry. Dodatkowo, jeżeli $g(\mathbb{X}) = 1$ to miarę można określić mianem znormalizowanej miary rozmytej [69]. W odróżnieniu do miar klasycznych, miara rozmyta zawiera mniej restrykcyjny warunek monotoniczności zamiast warunku addytywności.

Inaczej niż w teorii prawdopodobieństwa, gdzie tylko jedna miara, tj. miara prawdopodobieństwa opisuje jak prawdopodobne jest dane zdarzenie, w teorii możliwości analogiczną rolę pełni para miar: możliwości i konieczności.

2.2.3 Miara możliwości

Jeżeli zdarzenie jest opisane klasycznym nierozmytym zbiorem A_k i jest znany rozkład możliwości $\Pi_{\mathcal{X}}$ związany ze zmienną \mathcal{X} to miarę możliwości \mathcal{M} wystąpienia zdarzenia A_k można obliczyć jako

$$\mathcal{M}\{x \in A_k\} = \mathcal{M}_{\mathcal{X}}(A_k) = \sup_{x \in A_k} [\pi_{\mathcal{X}}(x)], \quad (2.22)$$

gdzie $\pi_{\mathcal{X}}(x)$ to funkcja rozkładu możliwości związanego ze zmienną \mathcal{X} . Równanie (2.22) może być łatwo zinterpretowane jako możliwość wystąpienia zdarzenia A_k (przyjęcia przez zmienną \mathcal{X} wartości w zbiorze A_k) znając możliwość przyjęcia przez zmienną \mathcal{X} wartości z przestrzeni \mathbb{X} . Jeżeli funkcja rozkładu możliwości będzie funkcją charakterystyczną, tj. $\pi_{\mathcal{X}}(x) \in \{0, 1\}$ a zmienna \mathcal{X} przyjmuje wartości ze zbioru X to powyższy wzór sprowadza się do

$$\mathcal{M}_{\mathcal{X}}\{A_k\} = 1 \text{ jeżeli } X \cap A_k \neq \emptyset. \quad (2.23)$$

Jeżeli natomiast wyrazi się funkcję rozkładu możliwości za pomocą funkcji przynależności zbioru rozmytego A to ogólna postać miary możliwości przedstawia się jako

$$\mathcal{M}_{\mathcal{X}}(A) = \sup_{x \in \mathbb{X}} [\min(\mu_A(x), \pi_{\mathcal{X}}(x))]. \quad (2.24)$$

Odnosząc się do przykładu z punktu 2.2.1 ($\pi_{\mathcal{X}}(x) = \mu_A(x)$) zdefiniujmy klasyczny (nierozmyty) zbiór B odnoszący się do stwierdzenia „temperatura niższa niż 37°C” czyli $B = (-\infty, 37)$. Miara możliwości tego, że „pacjent w stanie podgorączkowym ma temperaturę ciała mniejszą niż 37°C” obliczana jest w następujący sposób

$$\mathcal{M}_{\mathcal{X}}\{A \in B\} = \mathcal{M}_A\{B\} = \sup_{x \in \mathbb{X}} [\min(\mu_B(x), \mu_A(x))] = \sup_{x \in B} [\mu_A(x)] = 0.61, \quad (2.25)$$

gdzie $\mu_A(x)$ to funkcja przynależności zaprezentowana na rysunku 2.10. Wartość (2.25) mówi nam o stopniu możliwości wystąpienia u pacjenta w stanie podgorączkowym temperatury ciała mniejszej niż 37°C .

2.2.4 Miara konieczności

Drugą z podstawowych miar rozmytych jest miara konieczności [60, 69, 132], zdefiniowana jako

$$\mathcal{N}_{\mathcal{X}}\{x \in B\} = \mathcal{N}_{\mathcal{X}}(B) = 1 - \mathcal{M}_{\mathcal{X}}(\overline{B}) = \inf_{x \notin B} [1 - \pi_{\mathcal{X}}(x)], \quad (2.26)$$

Ponownie, jeśli $\pi_{\mathcal{X}}(x) \in \{0, 1\}$ i zdarzenie zostanie opisane zbiorem A_k to

$$\mathcal{N}_{\mathcal{X}}(A_k) = 1 \quad \overline{A_k} \cap X = \emptyset. \quad (2.27)$$

Inaczej o konieczności zdarzenia B można mówić jako o stopniu niemożliwości wystąpienia odwrotnego zdarzenia. Dla przedstawionego w tym punkcie przykładu konieczność (pewność) tego, że „pacjent w stanie podgorączkowym ma temperaturę ciała mniejszą niż 37°C ”

$$\mathcal{N}_{\mathcal{X}}\{A \in B\} = \mathcal{N}_A(B) = \inf_{x \notin B} [1 - \mu_A(x)] = 0, \quad (2.28)$$

co dla powyższego przykładu oznacza, że na podstawie zgromadzonej wiedzy nie ma żadnej pewności co do tego zdarzenia, choć na podstawie (2.25) zdarzenie było w pewnym stopniu możliwe.

2.2.5 Związek między miarami możliwości i konieczności

Posiłkując się przedstawionym przykładem, można dostrzec, że zdarzenie, które nie jest w pełni możliwe ($\mathcal{M}_A(B) < 1$), jest jednocześnie całkowicie niepewnym ($\mathcal{N}_A(B) = 0$). Powyższa zależność jest jedną z zależności wiążących miarę możliwości i konieczności dla zdarzeń opisanych zbiorami klasycznymi (B) [60]. Można zauważyć, że skoro $\mathcal{M}_{\mathcal{X}}(\mathbb{X}) = 1$ to opisując zdarzenie klasycznym zbiorem $A_k \in \mathbb{X}$, zachodzi zależność

$$\max(\mathcal{M}_{\mathcal{X}}(A_k), \mathcal{M}_{\mathcal{X}}(\overline{A_k})) = 1. \quad (2.29)$$

Zależność ta mówi o tym, że jedno z dwóch przeciwstawnych zdarzeń musi być możliwe. Taki stan rzeczy wpływa na następujące zależności:

$$\mathcal{N}_{\mathcal{X}}(A_k) \leq \mathcal{M}_{\mathcal{X}}(A_k) \quad \forall A_k \in \mathbb{X}, \quad (2.30)$$

$$\mathcal{N}_{\mathcal{X}}(A_k) > 0 \Rightarrow \mathcal{N}_{\mathcal{X}}(\overline{A_k}) = 0 \quad \Leftrightarrow \quad \mathcal{M}_{\mathcal{X}}(A_k) = 1, \quad (2.31)$$

$$\mathcal{M}_{\mathcal{X}}(A_k) < 1 \Rightarrow \mathcal{M}_{\mathcal{X}}(\overline{A_k}) = 1 \quad \Leftrightarrow \quad \mathcal{N}_{\mathcal{X}}(A_k) = 0. \quad (2.32)$$

Pierwsza własność wyraża jasno, że zdarzenie konieczne jest zawsze możliwe, ale nie odwrotnie [60]. Dwie pozostałe wiążą fakt, że zdarzenie w jakimś stopniu pewne jest w pełni możliwe, a zdarzenie nie w pełni możliwe jest całkowicie niepewne. Ten ostatni przypadek wystąpił w przykładzie opisanym w punktach 2.2.3 i 2.2.4.

2.3 Teoria Dempstera-Shafera

G. Shafer w swojej pracy „*Mathematical theory of evidence*” wprowadził pojęcie funkcji przekonania [103]. Jej definicja jest bezpośrednią konsekwencją i matematycznym sformalizowaniem wniosków wysuniętych przez A. P. Dempstera w pracach [27, 28], gdzie uogólnia on definicję bayesowskiego prawdopodobieństwa warunkowego [15].

W teorii Dempstera-Shafera kluczowym zagadnieniem jest sposób przypisywania wartości prawdopodobieństwa różny od podejścia klasycznego określony przez G. Shafera jako *basic probability assignment* (BPA) [103]. W polskojęzycznej literaturze spotykane są jednak odmienne sposoby tłumaczenia tego przypisania: funkcja alokacji, m-funkcja [121], podstawowe przypisanie prawdopodobieństwa [69] a także rozkład bazowego prawdopodobieństwa czy bazowy rozkład prawdopodobieństwa [21, 61]. W niniejszej rozprawie stosowane będą określenia zgodne z tym ostatnim czyli „bazowe prawdopodobieństwo” zamiennie z „wartość bazowego prawdopodobieństwa” (używanego też przez autora teorii jako *basic probability number* [103]).

Założmy ogólnie, że mamy do czynienia z elementami we wcześniej zdefiniowanej przestrzeni \mathbb{X} (w [103] zwanej *frame of discernment* czyli tzw. ramą rozróżniania). BPA przypisuje wszystkim możliwym podzbiорom \mathbb{X} wartość m zgodnie z [103]

$$m(\emptyset) = 0, \quad (2.33)$$

gdzie \emptyset jest zbiorem pustym oraz

$$\sum_{E \subset \mathbb{X}} m(E) = 1, \quad (2.34)$$

gdzie E jest podzbiorem \mathbb{X} jak w 2.2.2. Zbiór wszystkich możliwych podzbiорów $E \subset \mathbb{X}$ określa zbiór potęgowy $2^{\mathbb{X}}$. Jeżeli przykładowo

$$\mathbb{X} = \{a, b, c\}, \quad (2.35)$$

to

$$2^{\mathbb{X}} = \{\emptyset, a, b, c, \{a, b\}, \{a, c\}, \{b, c\}, \mathbb{X}\}. \quad (2.36)$$

Jeżeli wartość $m(E) > 0$, E nazywa się elementem ogniskowym (ang. *focal element*). Zbiór wszystkich $E \subset \mathbb{X}$, dla których $m(E) > 0$ nazywany jest zbiorem elementów ogniskowych S .

W ogólnym przypadku zbiór S może być mniej liczny niż zbiór $2^{\mathbb{X}}$, czyli $S \subsetneq 2^{\mathbb{X}}$ lub $S \subseteq 2^{\mathbb{X}}$. Bazowy rozkład prawdopodobieństwa m można traktować jako funkcję odwzorowującą elementy zbioru potęgowego $2^{\mathbb{X}}$ na wartości z przedziału $[0,1]$ z zachowaniem (2.33) i (2.34). Tym samym (2.34) wskazuje, że suma prawdopodobieństwa przypisana wszystkim elementom ogniskowym wynosi jeden. Samo $m(E)$ interpretuje się jako tzw. miarę przekonania związaną dokładnie z E .

2.3.1 Miary przekonania i domniemania

Z uwagi na to, że mogą istnieć zbiory F , takie że $F \subset E$ i mają one przypisaną wartość prawdopodobieństwa $m(F)$ to miara przekonania związana z E jest sumą

$$Bel(E) = \sum_{F \subset E} m(F). \quad (2.37)$$

Tym samym miara przekonania Bel podobnie jak m jest funkcją odwzorowującą elementy zbioru E na wartości z przedziału $[0,1]$. W oryginalnej pracy G. Shafera podana jest także definicja miary Bel bez odnoszenia się do definicji rozkładu bazowego prawdopodobieństwa m [103]. Dodatkową miarą opartą na rozkładzie bazowego prawdopodobieństwa jest miara Pl (ang. *plausibility*) określana miarą domniemania [111, 121]. Ta miara jest ściśle powiązana z miarą przekonania dlatego, że

$$Pl(E) = 1 - Bel(\bar{E}) = \sum_{F \subset \mathbb{X}} m(F) - \sum_{F \subset \bar{E}} m(F) = \sum_{F \cap E \neq \emptyset} m(F). \quad (2.38)$$

Co więcej, na podstawie (2.37) i (2.38) nietrudno zauważyć, że

$$Bel(E) \leq Pl(E). \quad (2.39)$$

Ma to swoje odzwierciedlenie w przyjętej nomenklaturze. Miara domniemania Pl wskazuje na możliwość niewykluczenia hipotezy wyrażonej przez E , ze względu na przynajmniej częściową „zgodność” F i E wyrażoną jako $F \cap E \neq \emptyset$. Miara przekonania nie jest od niej nigdy większa, dlatego, że bazuje tylko na w pełni „zgodnych” dowodach i wiedzy (tj. gdy zachodzi inkluzja zbiorów oznaczona jako $F \subset E$). Aby zilustrować obliczenie wartości miar Bel i Pl warto posłużyć się przykładem.

Przykład 2.1. Przyjmijmy, że nasza przestrzeń \mathbb{X} jest zbiorem objawów występujących w pewnym stanie chorobowym. Przedstawmy je symbolicznie jako

$$\mathbb{X} = \{o_1, o_2, o_3, o_4\}. \quad (2.40)$$

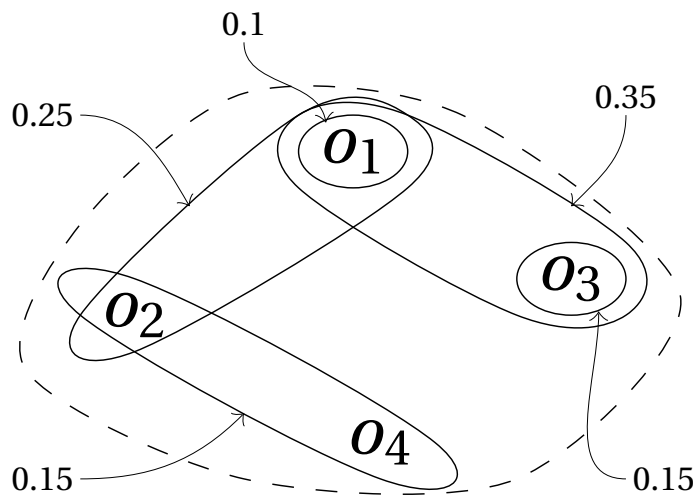
Niech o_r , $r = 1, \dots, 4$ są stwierdzeniami np. „gorączka”, „poziom cukru w normie”, itd. U podstaw teorii przyjęto, że te stwierdzenia związane są ze zbiorami klasycznymi. Lekarz diagnostyk może przypisać każdemu pojęciu jako podzbiorowi \mathbb{X} wartość bazowego prawdopodobieństwa m , która odpowiada jego stopniowi przekonania o związku symptomu lub ich

kombinacji z daną diagnozą. Przyjmujemy zatem dla wybranego przez niego zbioru elementów ogniskowych

$$S = \{o_1, \{o_1, o_2\}, \{o_1, o_3\}, o_3, \{o_2, o_4\}\} \quad (2.41)$$

następujący rozkład bazowego prawdopodobieństwa

$$\begin{aligned} m(\{o_1\}) &= 0.10, m(\{o_1, o_2\}) = 0.25, m(\{o_1, o_3\}) = 0.35, \\ m(\{o_3\}) &= 0.15, m(\{o_2, o_4\}) = 0.15. \end{aligned} \quad (2.42)$$



Rysunek 2.11: Objawy i zbiór elementów ogniskowych wraz z wartościami przypisanych prawdopodobieństw. Elementy ogniskowe zostały otoczone pętlą ciągłą a zbiór wszystkich objawów (2.40) pętlą przerywaną.

Opisany rozkład bazowego prawdopodobieństwa został także przedstawiony na rysunku 2.11. Warto zauważyć, że np. pojedynczy objaw o_2 nie jest uznawany przez eksperta za istotny w podjęciu decyzji diagnostycznej. Ten pojedynczy objaw nie jest więc według niego elementem ogniskowym (nie został zawarty w S). Niemniej jednak, złożone elementy ogniskowe ($\{o_1, o_2\}$ i $\{o_2, o_4\}$) wyrażają wspólne wystąpienie obu objawów a jednym z nich jest o_2 . Oznacza to tyle, że wystąpienie objawu o_2 ma znaczenie dla lekarza, ale tylko wtedy gdy występuje razem z innymi objawami co w rzeczywistości jest często spotykane [110]. Podobne znaczenie ma objaw o_4 .

Teraz przyjmijmy, że chcemy skonfrontować pacjenta, u którego stwierdzono objawy o_1 i o_3 , z omawianą diagnozą. Miary przekonania i domniemania będą wynosiły odpowiednio:

$$Bel(\{o_1, o_3\}) = m(\{o_1\}) + m(\{o_1, o_3\}) + m(\{o_3\}) = 0.60, \quad (2.43)$$

$$Pl(\{o_1, o_3\}) = m(\{o_1\}) + m(\{o_1, o_2\}) + m(\{o_1, o_3\}) + m(\{o_3\}) = 0.85. \quad (2.44)$$

2.3.2 Rozmyte elementy ogniskowe

Modelowanie nieprecyzyjnych objawów jest wygodne za pomocą funkcji przynależności oraz z wykorzystaniem teorii możliwości tak jak opisano w poprzednich punktach. W niniejszej rozprawie chcąc wykorzystać teorię Dempstera-Shafera do modelowania niepewnej wiedzy medycznej i jednocześnie przetwarzać nieprecyzyjne dane pacjenta, konieczne jest użycie rozmytego elementu ogniskowego. Rozmyty element ogniskowy może być opisany funkcją przynależności $\mu_A(x)$ a w przypadku złożonego rozmytego elementu ogniskowego (dotyczącego kombinacji nieprecyzyjnych objawów) kilkoma funkcjami przynależności modelującej każdy nieprecyzyjny objaw, którego rozmyty element ogniskowy dotyczy.

Teraz jednak, posługując się rozmytym elementem ogniskowym, nie operuje się wyłącznie z jednoznacznym potwierdzeniem albo odrzuceniem objawu u danego pacjenta. Przedstawiona w poprzednich punktach teoria możliwości, proponując odpowiednie miary, pozwala określić poziom dopasowania pomiędzy informacją wejściową na temat pacjenta oraz symptomem modelowanym rozmytym elementem ogniskowym. Przywołując definicję miary możliwości (2.24), mając na uwadze, że funkcja rozkładu możliwości dla danego pacjenta jest funkcją typu rozmyty singleton $\mu_{\mathcal{F}}(x; x^{(*)})$ wskazującą jednoznacznie np. na wartość otrzymanego pomiaru czy testu diagnostycznego $x^{(*)}$, to miara możliwości przyjmuje formę

$$\mathcal{M}_{\mathcal{F}}(A) = \sup_{x \in \mathbb{X}} [\min(\mu_A(x), \mu_{\mathcal{F}}(x; x^{(*)}))] = \mu_A(x^{(*)}). \quad (2.45)$$

Naturalnie rozmyty element ogniskowy może dotyczyć kombinacji objawów i może być związany z kilkoma funkcjami przynależności opisującymi zbiory A_i , $i = 1, \dots, n$. Dlatego poziom dopasowania w ogólnym przypadku (dla $x_i^{(*)}$, $i = 1, \dots, n$ wartości opisujących pacjenta) agreguje poszczególne wartości (2.45) dla każdego A_i . Przedstawia się on jako [111]

$$\eta = \min_{1 \leq i \leq n} \mathcal{M}_{\mathcal{F}}(A_i) = \min_{1 \leq i \leq n} \mu_{A_i}(x_i^{(*)}). \quad (2.46)$$

Powyższa definicja poziomu dopasowania pozwala ustalić w jakim stopniu przesłanki reguł dotyczących wybranej diagnozy są zgodne ze stanem diagnozowanego pacjenta. W niniejszej rozprawie zostaną przetestowane dwa sposoby użycia poziomu dopasowania rozmytych elementów ogniskowych: jedno związane z progowym wyborem elementów ogniskowych, drugie, jako modyfikacja obliczania miary przekonania. Należy zaznaczyć, że w tym rozdziale przedstawiono pokrótce podstawowe metody, które zostaną wykorzystane do modelowania danych medycznych i w proponowanych sposobach wspomagania diagnozy. Opis zastosowanych podejść zostanie szczegółowo i wyczerpująco przedstawiony w dalszej części rozprawy.

Rozdział 3

Rozmyte reguły warunkowe oparte na teorii Dempstera-Shafera

3.1 Reprezentacja danych medycznych

Baza medyczna związana z pewnym zagadnieniem medycznym jest zbiorem danych pacjentów przypisanych do różnych, zwykle wykluczających się diagnoz. Na początku, warto przedstawić zbiór danych dotyczący jednej diagnozy. Można go opisać jako macierz

$$\mathbf{X}^{(l)} = \begin{bmatrix} x_{11}^{(l)} & x_{12}^{(l)} & \cdots & x_{1j}^{(l)} & \cdots & x_{1\rho}^{(l)} \\ x_{21}^{(l)} & x_{22}^{(l)} & \cdots & x_{2j}^{(l)} & \cdots & x_{2\rho}^{(l)} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1}^{(l)} & x_{i2}^{(l)} & \cdots & x_{ij}^{(l)} & \cdots & x_{i\rho}^{(l)} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n_l1}^{(l)} & x_{n_l2}^{(l)} & \cdots & x_{n_lj}^{(l)} & \cdots & x_{n_l\rho}^{(l)} \end{bmatrix}. \quad (3.1)$$

Macierz $\mathbf{X}^{(l)}$ posiada n_l wierszy, a każdy z nich zawiera opis jednego pacjenta. Jednocześnie, każda kolumna macierzy $\mathbf{X}^{(l)}$ odnosi się do zbioru wartości jednego z symptomów dla wszystkich n_l pacjentów. I tak, pierwsza kolumna może zawierać wartości temperatur zmierzonych u wszystkich pacjentów a inna – wartości ciśnień skurczowych. Dlatego indeks j będzie oznaczać numer symptomu spośród ρ symptomów występujących w bazie (stąd ρ kolumn w macierzy $\mathbf{X}^{(l)}$). Macierz $\mathbf{X}^{(l)}$ dotyczy pacjentów przypisanych jednej diagnozie oznaczonej indeksem l . Liczba możliwych diagnoz dla danego problemu medycznego będzie oznaczona jako C , tzn. $l = 1, 2, \dots, C$. Wektor symptomów i -tego pacjenta przypisanego l -tej diagnozie, na podstawie (3.1) jest oznaczony jako

$$\mathbf{p}_i^{(l)} = \left[x_{i1}^{(l)} \quad x_{i2}^{(l)} \quad \cdots \quad x_{ij}^{(l)} \quad \cdots \quad x_{i\rho}^{(l)} \right], \quad (3.2)$$

gdzie $x_{ij}^{(l)}$ jest wartością j -tego symptomu dla i -tego pacjenta. Tak więc $\mathbf{p}_i^{(l)}$ zawiera wartości symptomów i -tego pacjenta (3.2) i jest i -tym wierszem macierzy $\mathbf{X}^{(l)}$. Co więcej, j -ta kolumna macierzy $\mathbf{X}^{(l)}$ zawierająca wartości j -tego symptomu dla wszystkich pacjentów przypisanych l -tej diagnozie będzie oznaczona jako

$$\mathbf{x}_j^{(l)} = \left[x_{1j}^{(l)} \quad x_{2j}^{(l)} \quad \cdots \quad x_{ij}^{(l)} \quad \cdots \quad x_{n_{lj}}^{(l)} \right]^T. \quad (3.3)$$

3.2 Definiowanie reguły diagnostycznej

Reguła diagnostyczna rozumiana jako rozmyta reguła warunkowa składa się z przesłanki i konkluzji. W przesłance występuje stwierdzenie określające objaw np. „temperatura jest wysoka”. Jeżeli stwierdzenie to dotyczy jednego objawu, to jest to tak zwana przesłanka prosta. Konkluzją w przypadku reguły diagnostycznej jest określenie diagnozy np. „pacjent zdrowy”. Oznaczmy l -tą diagnozę jako $D^{(l)}$. Przesłanka może być również złożona (ze złożonym stwierdzeniem będącym kombinacją pewnych objawów) np. „temperatura jest wysoka i poziom CRP jest normalny”. Tym samym reguła może dotyczyć jednego bądź kombinacji symptomów, które mogą być charakterystyczne dla danej diagnozy. Niech prosta przesłanka (z jednym stwierdzeniem) opisana będzie w sposób następujący

$$X_j \text{ jest } A_j^{(l)}, \quad (3.4)$$

gdzie X_j to zmienna lingwistyczna opisująca badaną wielkość, np. „temperatura”, a $A_j^{(l)}$ to zbiór rozmyty opisany funkcją przynależności $\mu_j^{(l)}$ (2.2), która określa wartość lingwistyczną np. „wysoka”. Z kolei najbardziej złożona przesłanka, dotycząca wszystkich ρ symptomów w bazie $\mathbf{X}^{(l)}$ jest postaci

$$X_1 \text{ jest } A_1^{(l)} \text{ i } X_2 \text{ jest } A_2^{(l)} \text{ i } \cdots \text{ i } X_j \text{ jest } A_j^{(l)} \text{ i } \cdots \text{ i } X_\rho \text{ jest } A_\rho^{(l)}. \quad (3.5)$$

Możliwe jest również zdefiniowanie przesłanki dotyczącej podzbioru symptomów, np.

$$\begin{aligned} X_1 \text{ jest } A_1^{(l)} \text{ i } X_3 \text{ jest } A_3^{(l)} \text{ i } \cdots \text{ i } X_{j-1} \text{ jest } A_{j-1}^{(l)} \text{ i} \\ \text{i } X_{j+1} \text{ jest } A_{j+1}^{(l)} \text{ i } \cdots \text{ i } X_k \text{ jest } A_k^{(l)}; k < \rho. \end{aligned} \quad (3.6)$$

Przedstawione powyżej przesłanki w teorii Dempstera-Shafera określane są jako rozmyte elementy ogniskowe. Każdy element ogniskowy dotyczy pewnego podzbioru ρ symptomów. Liczba rozmytych elementów ogniskowych jest związana z liczbą opisanych symptomów w bazie. Jeżeli oznaczymy zbiór rozmytych elementów ogniskowych l -tej diagnozy jako $S^{(l)}$ a rozmyte elementy ogniskowe l -tej diagnozy jako $s_r^{(l)}$ (gdzie r to numer elementu ogniskowego w $S^{(l)}$) to maksymalna ich liczba w $S^{(l)}$ wynosi $\sum_{k=1}^{\rho} \binom{\rho}{k} = 2^\rho - 1$. Używając oznaczenia $s_r^{(l)}$ dla r -tego elementu ogniskowego w zbiorze elementów ogniskowych l -tej diagnozy, rozmytą

regułę warunkową można zdefiniować następująco

$$\text{jeżeli } s_r^{(l)} \text{ to } D^{(l)}, \quad (3.7)$$

Liczba elementów ogniskowych w $S^{(l)}$ będzie oznaczona jako $n^{(l)}$. W ogólnym przypadku liczba ta dla zbioru $S^{(l)}$ może być mniejsza od maksymalnej, czyli

$$n^{(l)} \leq 2^\rho - 1. \quad (3.8)$$

Oznaczmy dodatkowo jako $\mathbb{J}_r^{(l)}$ zbiór indeksów symptomów, dla których występują stwierdzenia w r -tym elemencie ogniskowym. Użycie tego zbioru będzie przydatne między innymi przy określaniu jakości reguł. Przykładowo, pełny zbiór elementów ogniskowych $S^{(l)}$ dla $\rho = 3$ symptomów zawiera siedem elementów ogniskowych

$$S^{(l)} = \{s_1^{(l)}, s_2^{(l)}, s_3^{(l)}, s_4^{(l)}, s_5^{(l)}, s_6^{(l)}, s_7^{(l)}\}, \quad (3.9)$$

Reprezentacja rozmytych elementów ogniskowych $s_r^{(l)}$ jak i odpowiadających im zbiorom indeksów symptomów $\mathbb{J}_r^{(l)}$ przedstawiona jest w tabeli 3.1.

Tabela 3.1: Rozmyte elementy ogniskowe i odpowiadające im zbiory symptomów

r	$s_r^{(l)}$	$\mathbb{J}_r^{(l)}$
1	$X_1 \text{ jest } A_1^{(l)}$	{1}
2	$X_2 \text{ jest } A_2^{(l)}$	{2}
3	$X_3 \text{ jest } A_3^{(l)}$	{3}
4	$X_1 \text{ jest } A_1^{(l)} \text{ i } X_2 \text{ jest } A_2^{(l)}$	{1, 2}
5	$X_1 \text{ jest } A_1^{(l)} \text{ i } X_3 \text{ jest } A_3^{(l)}$	{1, 3}
6	$X_2 \text{ jest } A_2^{(l)} \text{ i } X_3 \text{ jest } A_3^{(l)}$	{2, 3}
7	$X_1 \text{ jest } A_1^{(l)} \text{ i } X_2 \text{ jest } A_2^{(l)} \text{ i } X_3 \text{ jest } A_3^{(l)}$	{1, 2, 3}

W przypadku baz medycznych liczba opisywanych symptomów jest różna. Najczęściej jednak baza medyczna tworzona i zarządzana przez lekarza zawiera od kilku do kilkudziesięciu symptomów. Jest oczywiste, że liczba możliwych rozmytych elementów ogniskowych rośnie wykładniczo wraz z liczbą symptomów. Naturalnie, nie zawsze wiązanie ze sobą kombinacji wielu symptomów ma sens. Uważa się, że siedem warunków logicznych z psychologicznego punktu widzenia jest maksymalną ich liczbą, która może być jeszcze zrozumiała i efektywnie zinterpretowana w złożonej przesłance [42]. Ujmowanie kombinacji symptomów (objawów) ma bardzo ważne znaczenie w postępowaniu diagnostycznym. Niemniej jednak, inżynier wiedzy, mając na uwadze łatwą interpretację reguł, powinien starać się ograniczyć występowanie przesłanek o znacznej złożoności. Ten temat zostanie poddany szczególnej uwadze w rozdziale opisującym przeprowadzone eksperymenty na danych medycznych.

3.3 Wyznaczanie rozkładu bazowego prawdopodobieństwa

Opisane w punkcie 2.3 podstawy teorii Dempstera-Shafera stosowane będą przy wyznaczeniu rozkładu bazowego prawdopodobieństwa dla tworzonych przesłanek reguł diagnostycznych jako elementów ogniskowych. Podstawy teorii przybliżano stosując tylko pojęcia zbiorów. Dla wprowadzonych w punkcie 3.2 oznaczeń, zależności wiążące rozkład bazowego prawdopodobieństwa (2.33) i (2.34) można zapisać analogicznie, jako

$$m(f) = 0, \quad \sum_{s_r^{(l)} \in S^{(l)}} m(s_r^{(l)}) = 1. \quad (3.10)$$

Występujące oznaczenie f interpretowane jako fałszywa przesłanka. Można za jej pomocą modelować „brak jakichkolwiek objawów choroby” lub objawy niezwiązane z l -tą diagnozą. Analogicznie, drugi warunek (3.10) określa, że maksymalna wartość bazowego prawdopodobieństwa dla l -tej diagnozy (czyli suma bazowego prawdopodobieństwa dla wszystkich elementów ogniskowych) wynosi jeden.

Wartość bazowego prawdopodobieństwa może stanowić o niepewności reguły. W aspekcie wspomaganiania diagnozy słuszne jest aby ekspert ludzki określił niepewność reguły jako subiektywną ocenę związku pewnego objawu lub ich kombinacji (reprezentowanego przez rozmyty element ogniskowy) z daną diagnozą. Jeśli natomiast dostępne są dane uczące, rozkład bazowego prawdopodobieństwa może być określony dla tworzonych reguł w oparciu o częstość występowania objawów. Problem obliczania bazowego prawdopodobieństwa poruszony został m.in. w [109, 136].

W punkcie 2.3.2 przedstawiono sposób wyznaczania poziomu dopasowania informacji wejściowej i elementu ogniskowego dzięki wykorzystaniu miary możliwości. W niniejszej rozprawie zależność (2.46) zgodnie z wprowadzonymi oznaczeniami sprowadza się do [109]

$$\eta_r^{(l)}(\mathbf{p}_i^{(l)}) = \min_{j \in \mathcal{J}_r^{(l)}} \mu_j^{(l)}(x_{ij}^{(l)}), \quad (3.11)$$

gdzie $\mu_j^{(l)}(x_{ij}^{(l)})$ to przynależność wartości $x_{ij}^{(l)}$ do zbioru rozmytego $A_j^{(l)}$. Aby uzyskać rozkład bazowego prawdopodobieństwa można znaleźć na przykład liczbę dopasowanych przypadków z l -tej diagnozy. Aby uznać i -ty przypadek za dopasowany należy określić próg wartości $\eta_r^{(l)}$, od którego poziom dopasowania jest wystarczająco dobry. Wartość progu dopasowania przy obliczaniu BPA będzie oznaczona jako η_e . Podsumowując, dla l -tej diagnozy liczba przypadków dopasowanych do elementu ogniskowego $s_r^{(l)}$ może być wyrażona jako $\sum_{i=1}^{n_l} I(\eta_r^{(l)}(\mathbf{p}_i^{(l)}) > \eta_e)$ gdzie $I(\eta_r^{(l)}(\mathbf{p}_i^{(l)}) > \eta_e)$ jest funkcją wskaźnikową przyjmującą wartość 0 lub 1 w zależności od spełnienia warunku dopasowania. Procedura 1 przedstawia kroki postępowania przy obliczaniu rozkładu bazowego prawdopodobieństwa dla elementu ogniskowego l -tej diagnozy.

PROCEDURA 1

1. Oblicz poziom dopasowania r -tego elementu ogniskowego (3.11) z każdym przypadkiem $\mathbf{p}_i^{(l)}$ z macierzy danych $\mathbf{X}^{(l)}$.
2. Określ liczbę przypadków $\mathbf{p}_i^{(l)}$, dla których poziom dopasowania (3.11) przekroczył zadany próg niepewności η_e tj.

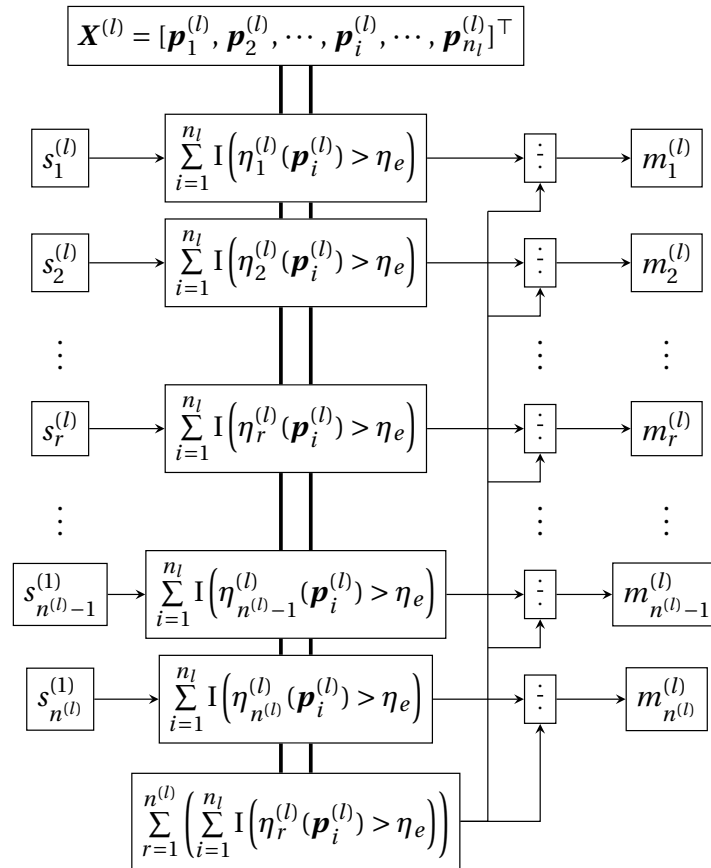
$$\eta_r^{(l)}(\mathbf{p}_i^{(l)}) > \eta_e. \quad (3.12)$$

3. Znormalizuj otrzymaną wartość poprzez podzielenie jej przez całkowitą liczbę dopasowań wszystkich elementów ogniskowych $s_r^{(l)}$ występujących w zbiorze $S^{(l)}$.

Stosując procedurę 1 otrzymana zostaje wartość bazowego prawdopodobieństwa

$$m(s_r^{(l)}) = \frac{\sum_{i=1}^{n_l} \mathbf{I}(\eta_r^{(l)}(\mathbf{p}_i^{(l)}) > \eta_e)}{\sum_{r=1}^{n^{(l)}} \left(\sum_{i=1}^{n_l} \mathbf{I}(\eta_r^{(l)}(\mathbf{p}_i^{(l)}) > \eta_e) \right)}. \quad (3.13)$$

Na rysunku 3.1 przedstawiono wyznaczanie rozkładu bazowego prawdopodobieństwa dla zbioru elementów ogniskowych zachowując wprowadzone oznaczenia. W rozprawie stosowane będzie skrócone oznaczenie wartości bazowego prawdopodobieństwa $m(s_r^{(l)}) \hat{=} m_r^{(l)}$.



Rysunek 3.1: Procedura wyznaczania rozkładu bazowego prawdopodobieństwa

Przykład 3.1. Tabela 3.2 przedstawia procedurę obliczenia rozkładu bazowego prawdopodobieństwa zakładając, że w zbiorze $S^{(l)}$ jest siedem elementów ogniskowych, tak jak zdefiniowano w (3.9). Z uwagi na to, że $\eta_r^{(l)}(\mathbf{p}_i^{(l)}) \in [0, 1]$, przykładowe wartości w głównej części tabeli zostały wygenerowane jako wartości zmiennej losowej o rozkładzie jednostajnym w powyższym przedziale. Do obliczenia BPA ustalono wymagany poziom dopasowania $\eta_e = 0.4$. Dlatego wartości $\eta_r^{(l)}(\mathbf{p}_i^{(l)})$ większe niż η_e zostały pogrubione. Na podstawie tabeli 3.2 można zauważyć, że elementy ogniskowe z największą liczbą dopasowań do przypadków z $X^{(l)}$ otrzymują najwyższą wartość $m_r^{(l)}$. Trzeci punkt procedury wyznaczania BPA, tj. podzielenie liczby dopasowań przez liczbę wszystkich dopasowań elementów ogniskowych w zbiorze $S^{(l)}$, podyktowany jest drugim warunkiem w (3.10).

Tabela 3.2: Poziomy dopasowania siedmiu elementów ogniskowych i 10 danych uczących oraz obliczony na ich podstawie rozkład bazowego prawdopodobieństwa. Wartości $\eta_r^{(l)}(\mathbf{p}_i^{(l)})$ większe od $\eta_e = 0.4$ zostały pogrubione.

$i \backslash r$	1	2	3	4	5	6	7
1	0.55	0.76	0.06	0.22	0.85	0.14	0.14
2	0.19	0.6	0.93	0.66	0.63	0.69	0.1
3	0.53	0.81	0.12	0.52	0.78	0.3	0.39
4	0.92	1	0.24	0.22	0.56	0.83	0.45
5	0.4	0.61	0.89	0.74	0.05	0.54	0.65
6	0.58	0.9	0.87	0.84	0.8	0.03	0.81
7	0.62	0.49	0.29	0.61	0.29	0.9	0.48
8	0.7	0.74	0.99	0.64	0.46	0.01	0
9	0.95	0.57	0.28	0.27	0.98	0.13	0.24
10	0.29	0.46	0.55	0.65	0.08	0.6	0.98
$\sum_{i=1}^{n_l} \mathbb{I}(\eta_r^{(l)}(\mathbf{p}_i^{(l)}) > \eta_e)$	7	8	5	7	6	5	3
$m_r^{(l)}$	0.171	0.195	0.122	0.171	0.146	0.122	0.073

3.4 Wspomaganie diagnozy na podstawie miary przekonania

Wykorzystana w niniejszej rozprawie miara przekonania, oznaczana jako $Bel^{(l)}(\mathbf{p}^{(*)})$ będzie wskazywać, jak wysoce jest prawdopodobna l -ta diagnoza dla pacjenta, którego dane opisane są przez

$$\mathbf{p}^{(*)} = \left[p_1^{(*)} \quad p_2^{(*)} \quad \dots \quad p_j^{(*)} \quad \dots \quad p_\rho^{(*)} \right]. \quad (3.14)$$

Oznaczenie wektora pacjenta poddanego diagnozie jest analogiczne do opisu przypadków z bazy $X^{(l)}$ (3.1) i same dane pacjenta mogą być z niej zaczerpnięte, lecz zakłada się, że diagnoza dla tego pacjenta jest nieznaną. Wartość $Bel^{(l)}$ jest sumą bazowego prawdopodobieństwa rozmytych elementów ogniskowych, które zostały potwierdzone u tego pacjenta.

Do potwierdzenia zgodności objawów pacjenta i elementów ogniskowych użyty zostanie, podobnie jak przy wyznaczaniu rozkładu bazowego prawdopodobieństwa, poziom dopasowania. Tym samym wyznaczenie tej miary przekonania sprowadza się do obliczenia

$$Bel^{(l)}(\mathbf{p}^{(*)}) = \sum_{r=1}^{n^{(l)}} I(\eta_r^{(l)}(\mathbf{p}^{(*)}) > \eta_b) \cdot m_r^{(l)}, \quad (3.15)$$

gdzie $I(\eta_r^{(l)}(\mathbf{p}^{(*)}) > \eta_b)$ to funkcja wskaźnikowa przyjmująca wartość 0 bądź 1 w zależności od tego czy poziom dopasowania przekracza zadany próg η_b a $m_r^{(l)}$ to wartość bazowego prawdopodobieństwa r -tego elementu ogniskowego dla l -tej diagnozy. Jeżeli wszystkie elementy ogniskowe są dopasowane do testowanego przypadku $\mathbf{p}^{(*)}$ to $Bel^{(l)}(\mathbf{p}^{(*)}) = 1$. Jeżeli żaden, to $Bel^{(l)}(\mathbf{p}^{(*)}) = 0$. Jeśli rozpatruje się C konkurujących diagnoz, o regułach zawartych w zbiorach $S^{(l)}$, $l = 1, \dots, C$, określenie ostatecznej diagnozy wiąże się z obliczeniem $Bel^{(l)}(\mathbf{p}^{(*)})$ dla wszystkich tych diagnoz i wybranie tej dla której uzyskano największą wartość

$$D(\mathbf{p}^{(*)}) = \begin{cases} D^{(l)} & \text{jeżeli } \exists! l = \arg \max_{1 \leq l \leq C} (Bel^{(l)}(\mathbf{p}^{(*)})), \\ \text{nieokreślona} & \text{w przeciwnym wypadku.} \end{cases} \quad (3.16)$$

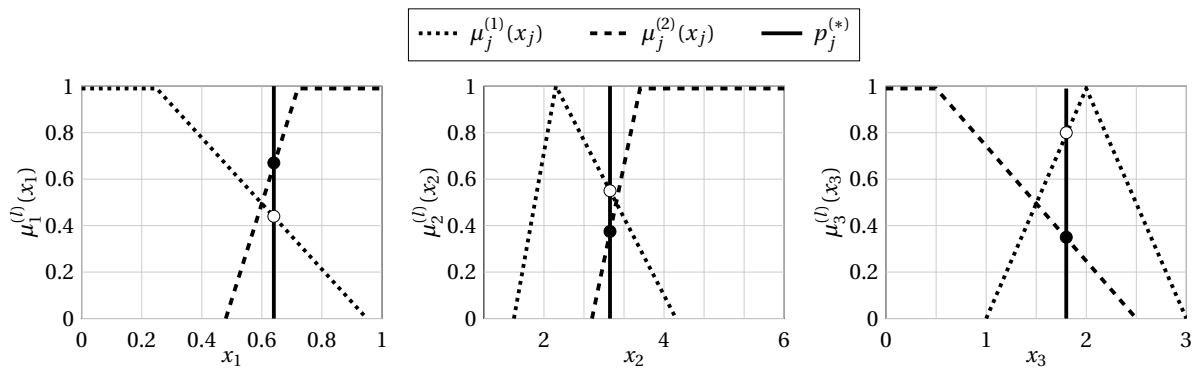
Można zauważyć, że zależność (3.16) pozwala wybrać diagnozę jeżeli tylko jedna wartość miary przekonania (3.21) ma wartość maksymalną. Niemniej jednak, może zająć okoliczność, w której nietypowy przypadek poddany diagnozie, otrzyma dwie identyczne maksymalne wartości miary przekonania. Najczęściej, w zagadnieniach klasyfikacji podejmuje się w takiej sytuacji decyzję arbitralną. W proponowanej metodzie, z uwagi na to, że przetwarzane są dane medyczne, przypadek dla którego porównanie wartości $Bel^{(l)}(\mathbf{p}^{(*)})$ nie daje jednoznacznej odpowiedzi, zostaje uznany jako nierozstrzygnięty i wymagający dalszej uwagi eksperta (lekarza). Aby zilustrować metodę wnioskowania wykorzystującą teorię Dempstera-Shafera rozszerzoną na rozmyte elementy ogniskowe, poniżej podano przykład wyznaczania diagnozy dla syntetycznego przypadku medycznego opisanego trzema symptomami z bazą reguł w postaci kilku elementów ogniskowych.

Przykład 3.2. Zadanie polega na wskazaniu, która z dwu diagnoz jest bardziej wiarygodna dla wybranego pacjenta. Załóżmy, że do zdiagnozowania stanu chorobowego wykonuje się trzy testy laboratoryjne, których wyniki są dane liczbowo. Dla każdej z diagnoz utworzono bazę wiedzy zawierającą trzy reguły rozmyte (rozmyte elementy ogniskowe). Przesłanki tych reguł przedstawiają się następująco:

$S^{(1)}$: przesłanki w regułach dla 1. diagnozy („zdrowy”):

- $s_1^{(1)}$: wynik testu 1. jest niski,
- $s_2^{(1)}$: wynik testu 2. jest w normie,
- $s_3^{(1)}$: wynik testu 3. jest w normie.

Powyższym elementom ogniskowym przypisano następujące wartości bazowego prawdopodobieństwa: $m_1^{(1)} = 0.5$, $m_2^{(1)} = 0.3$, $m_3^{(1)} = 0.2$.



Rysunek 3.2: Zbiory rozmyte w przesłankach diagnozy pierwszej (linia kropkowana) oraz drugiej (linia kreskowana). Liniami ciągłymi zaprezentowano wartości symptomów przypadku $\mathbf{p}^{(*)}$.

$S^{(2)}$: przesłanki w regułach dla 2. diagnozy („chory”):

$s_1^{(2)}$: wynik testu 1. jest wysoki,

$s_2^{(2)}$: wynik testu 2. jest wysoki i wynik testu 3. jest niski,

$s_3^{(2)}$: wynik testu 3. jest niski.

Powyższym elementom ogniskowym przypisano następujące wartości bazowego prawdopodobieństwa: $m_1^{(2)} = 0.3$, $m_2^{(2)} = 0.3$, $m_3^{(2)} = 0.4$. Na rysunku 3.2 zaprezentowano funkcje przynależności zbiorów rozmytych opisanych w elementach ogniskowych. Załóżmy następujący przypadek medyczny zawierający informację o numerycznych wynikach trzech testów

$$\mathbf{p}^{(*)} = \begin{bmatrix} p_1^{(*)} & p_2^{(*)} & p_3^{(*)} \end{bmatrix} = \begin{bmatrix} 0.61 & 3.1 & 1.8 \end{bmatrix}. \quad (3.17)$$

Wartości dla przypadku $\mathbf{p}^{(*)}$ zostały oznaczone na rysunku 3.2 jako trzy singletony o położeniach określonych przez wyniki testów laboratoryjnych. W tabeli 3.3 przedstawiono poziomy dopasowania (3.11) przypadku $\mathbf{p}^{(*)}$ dla każdego elementu ogniskowego. Wybierając próg dopasowania $\eta_b = 0.5$, miary przekonania dla konkurujących ze sobą diagnoz wynoszą, odpowiednio:

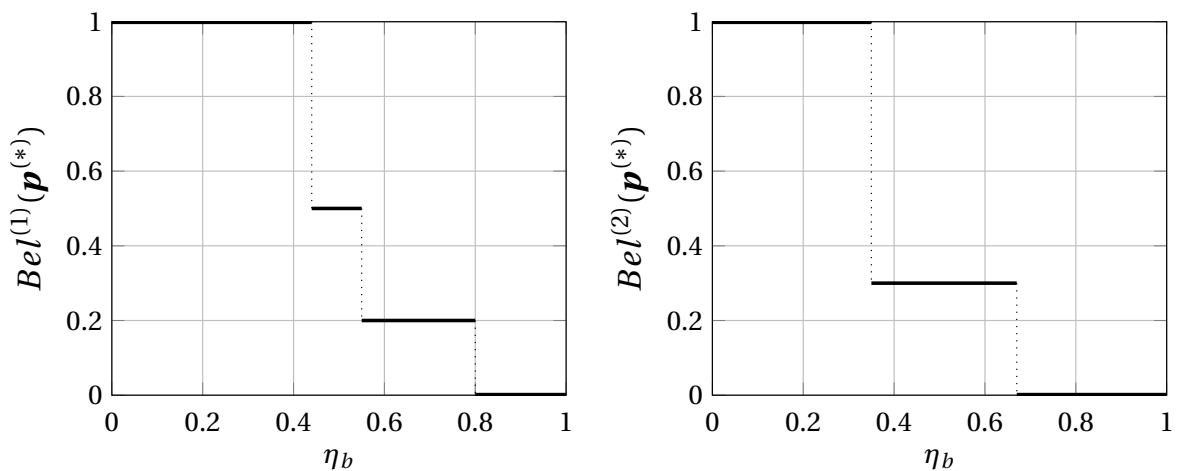
$$\begin{aligned} Bel^{(1)}(\mathbf{p}^{(*)}) &= m_2^{(1)} + m_3^{(1)} = 0.5, \\ Bel^{(2)}(\mathbf{p}^{(*)}) &= m_1^{(2)} = 0.3. \end{aligned} \quad (3.18)$$

Tabela 3.3: Poziomy dopasowania elementów ogniskowych do zadanego w przykładzie przypadku medycznego (3.17)

1. diagnoza	2. diagnoza
$\eta_1^{(1)}(\mathbf{p}^{(*)}) = \mu_1^{(1)}(p_1^{(*)}) = 0.44$	$\eta_1^{(2)}(\mathbf{p}^{(*)}) = \mu_1^{(2)}(p_1^{(*)}) = 0.67$
$\eta_2^{(1)}(\mathbf{p}^{(*)}) = \mu_2^{(1)}(p_2^{(*)}) = 0.55$	$\eta_2^{(2)}(\mathbf{p}^{(*)}) = \min[\mu_2^{(2)}(p_2^{(*)}), \mu_3^{(2)}(p_3^{(*)})] = 0.35$
$\eta_3^{(1)}(\mathbf{p}^{(*)}) = \mu_3^{(1)}(p_3^{(*)}) = 0.8$	$\eta_3^{(2)}(\mathbf{p}^{(*)}) = \mu_3^{(2)}(p_3^{(*)}) = 0.35$

Na podstawie podanych reguł diagnostycznych, dla przypadku medycznego opisanego przez (3.17), wyznaczona diagnoza to diagnoza pierwsza (pacjent „zdrowy”). Wybór progu η_b jest kluczowy i warunkuje nie tylko wymagany poziom dopasowania objawów do elementu ogniskowego dla także wpływa na ostateczną wartość miary $Bel^{(l)}(\mathbf{p}^{(*)})$. Dla powyższego przykładu można prześledzić zmianę wartości obu miar przekonania dla progu η_b mogącego zmieniać się w przedziale $[0,1]$. Na podstawie wykresów zależności miary $Bel^{(l)}(\mathbf{p}^{(*)})$ przedstawionych na rysunku 3.3 można zauważyć, że trzy różne decyzje (tzn. diagnoza „zdrowie”, „choroba” oraz diagnoza „nierozstrzygnięta”) można uzyskać dla następujących przedziałów wartości parametru η_b :

- diagnoza „zdrowie” dla $\eta_b \in (0.35, 0.55] \cup (0.67, 0.8]$,
- diagnoza „choroba” dla $\eta_b \in (0.55, 0.67]$,
- diagnoza „nierozstrzygnięta” dla $\eta_b \in [0, 0.35] \cup (0.8, 1]$.



Rysunek 3.3: Zmiany wartości miary przekonania dla konkurujących ze sobą diagnoz w zależności od zmiany wartości progu dopasowania η_b

Zmiana wymaganego progu dopasowania może prowadzić do zmiany ostatecznej diagnozy. Wyszukanie najlepszego progu η_b jest konieczne na etapie uczenia aby uzyskać pożądaną jakość wspomaganie diagnozy.

3.5 Modyfikacja metody obliczania miary przekonania

Wybór progów dopasowania (η_e i η_b) rozmytych elementów ogniskowych daje możliwość określenia wymaganego stopnia pewności podczas wyznaczania bazowego prawdopodobieństwa i miary przekonania. Jednakże, z punktu widzenia automatycznego wydobywania reguł diagnostycznych, potrzeba poszukiwania najlepszej pary progów η_e oraz η_b stanowi pewne wyzwanie optymalizacyjne. Wartość η_e występuje tylko przy wstępnej ocenie

wszystkich rozmytych elementów ogniskowych. Większej uwagi wymaga próg η_b , gdyż bierze on bezpośredni udział w etapie wspomagania diagnozy. Weźmy pod uwagę prosty rozmyty element ogniskowy „ X_j jest $A_j^{(l)}$ ”. Skoro funkcja wskaźnikowa w (3.15) pozwala określić spełnione dopasowanie elementu ogniskowego z danym przypadkiem $\mathbf{p}^{(*)}$ zgodnie z (3.11), to

$$\eta_r^{(l)}(\mathbf{p}^{(*)}) = \mu_j^{(l)}(p_j^{(*)}), j \in \mathbb{J}_r^{(l)}, \quad (3.19)$$

a tym samym

$$\mu_j^{(l)}(p_j^{(*)}) > \eta_b \Leftrightarrow p_j^{(*)} \in [A_j^{(l)}]_{\underline{\eta_b}}, \quad (3.20)$$

gdzie $[A_j^{(l)}]_{\underline{\eta_b}}$ jest ostrym η_b -przekrojem (2.7) zbioru $A_j^{(l)}$. Na podstawie (3.20) można zauważyć, że użycie progu η_b sprowadza warunek logiczny ze zbiorem rozmytym $A_j^{(l)}$ do warunku logicznego z ostrym α -przekrojem (2.7) zbioru $A_j^{(l)}$ ($\alpha = \eta_b$). W ten sposób otrzymuje się warunek logiczny w postaci $(p_j^{(*)} > \min[A_j^{(l)}]_{\underline{\eta_b}}) \wedge (p_j^{(*)} < \max[A_j^{(l)}]_{\underline{\eta_b}})$, który prowadzi do klasycznej reguły warunkowej. Nie należy jednak zapominać, że do jej powstania konieczna była definicja rozmytych elementów ogniskowych.

Z uwagi na powyższe obserwacje, przedstawiona zostanie idea obliczania miary przekonania, w której nie jest konieczne określanie czy dopasowanie jest dostatecznie dobre (większe od progu η_b). W takim podejściu definicja miary przekonania $Bel^{(l)}$ (3.15) powinna być zmodyfikowana. Niech będzie ona oznaczona jako $\widetilde{Bel}^{(l)}$. Można ją obliczyć jako

$$\widetilde{Bel}^{(l)}(\mathbf{p}^{(*)}) = \sum_{r=1}^{n^{(l)}} \eta_r^{(l)}(\mathbf{p}^{(*)}) \cdot m_r^{(l)}, \quad (3.21)$$

gdzie $\eta_r^{(l)}(\mathbf{p}^{(*)})$ jest poziomem dopasowania przypadku $\mathbf{p}^{(*)}$ do r -tego elementu ogniskowego zdefiniowanym w (3.11). Dla rozmytych elementów ogniskowych z jednym warunkiem logicznym (prostych przesłanek) obliczenie $\eta_r^{(l)}(\mathbf{p}^{(*)})$ sprowadza się do wyznaczenia przynależności wartości symptomu $p_j^{(*)}$ do rozmytego zbioru $A_j^{(l)}$. Miara przekonania (3.21) jest obliczana jako suma bazowych prawdopodobieństw wszystkich elementów ogniskowych pomnożonych przez wartość przynależności (w ogólnym przypadku poziom dopasowania). Takie podejście zaproponował Yen w swojej pracy dotyczącej uogólnienia teorii Dempstera-Shafera na zbiory rozmyte [129]. Zaproponowana definicja (3.21) jest tożsama ze sposobem obliczenia inkluzji dwóch zbiorów rozmytych, którą zaproponował H. Ogawa [87]. W obecnym podejściu poziom dopasowania wartości symptomu do rozmytego elementu ogniskowego jest także wynikiem obliczenia inkluzji, ponieważ wartość symptomu reprezentowana jest jednoelementowym zbiorem rozmytym (singletonem) a rozmyty element ogniskowy opisany jest funkcją przynależności. Z punktu widzenia wyznaczania diagnozy, \widetilde{Bel} pozwala ograniczyć potrzebę poszukiwania najlepszego progu dopasowania do wyznaczania jedynie wartości progu wymaganej w początkowej ocenie reguł i wyznaczeniu rozkładu bazowego prawdopodobieństwa. W eksperymentach przedstawionych w dalszej części rozprawy zo-

stanie porównany wpływ stosowania miar Bel i \widetilde{Bel} na wynik wspomaganie diagnozy jak i wydobywania reguł diagnostycznych. Obecnie zaprezentowany zostanie przykład wyznaczania diagnozy analogiczny do poprzedniego, tym razem jednak opierający wniosek na obliczonych wartościach $\widetilde{Bel}(\mathbf{p}^{(*)})$.

Przykład 3.3. Posiłkując się takim samym zbiorem reguł jak z przykładu 3.2, przypadek opisany w (3.17) zostanie poddany diagnozie wyznaczonej na podstawie porównania wartości przekonania $\widetilde{Bel}^{(l)}(\mathbf{p}^{(*)})$. Biorąc pod uwagę poziomy dopasowania z tabeli 3.3 obliczenie miar $\widetilde{Bel}^{(l)}(\mathbf{p}^{(*)})$ dla konkurujących diagnoz sprowadza się do:

$$\begin{aligned}\widetilde{Bel}^{(1)}(\mathbf{p}^{(*)}) &= \eta_1^{(1)}(\mathbf{p}^{(*)}) \cdot m_1^{(1)} + \eta_2^{(1)}(\mathbf{p}^{(*)}) \cdot m_2^{(1)} + \eta_3^{(1)}(\mathbf{p}^{(*)}) \cdot m_3^{(1)} = \\ &= 0.44 \cdot 0.5 + 0.55 \cdot 0.3 + 0.8 \cdot 0.2 = 0.220 + 0.165 + 0.160 = 0.545,\end{aligned}$$

$$\begin{aligned}\widetilde{Bel}^{(2)}(\mathbf{p}^{(*)}) &= \eta_1^{(2)}(\mathbf{p}^{(*)}) \cdot m_1^{(2)} + \eta_2^{(2)}(\mathbf{p}^{(*)}) \cdot m_2^{(2)} + \eta_3^{(2)}(\mathbf{p}^{(*)}) \cdot m_3^{(2)} = \\ &= 0.67 \cdot 0.3 + 0.35 \cdot 0.3 + 0.35 \cdot 0.4 = 0.201 + 0.105 + 0.140 = 0.446.\end{aligned}$$

W wyniku zastosowania zmodyfikowanej miary przekonania postawiona zostaje ta sama (pierwsza) diagnoza jak przy użyciu Bel z progiem dopasowania η_b równym 0.5. Warto zauważyć, że w przypadku tradycyjnego obliczania miary przekonania Bel , pierwsza diagnoza jest preferowana dla szerszego przedziału wartości η_b , niż diagnoza druga. Uzyskanie tego wyniku także dla \widetilde{Bel} jest zgodne z tą preferencją, mimo że strojnie wartości progu nie było konieczne.

3.6 Analiza danych uczących i wyznaczanie funkcji przynależności w rozmytych elementach ogniskowych

W punkcie 2.1.1 przedstawiono proponowane w teorii zbiorów rozmytych definicje funkcji przynależności. Kształt funkcji przynależności ma znaczenie w interpretacji symptomów i wpływ na wyniki wnioskowania. Do wyznaczenia funkcji przynależności dla symptomów charakterystycznych dla poszczególnych diagnoz konieczna jest ich analiza statystyczna. Dlatego punkt ten poświęcono na przedstawienie metody wyznaczania funkcji przynależności w elementach ogniskowych, aby później przeanalizować wpływ tych funkcji na działanie proponowanej metody ekstrakcji reguł i wspomaganie diagnozy.

Do konstrukcji funkcji przynależności dla j -tego symptomu w l -tej diagnozie służy zdefiniowany w (3.3) wektor wszystkich wartości j -tego symptomu występujących u pacjentów przypisanych l -tej diagnozie. Liczba elementów $\mathbf{x}_j^{(l)}$ jest tożsama z liczbą przypadków przypisanych l -tej diagnozie, czyli n_l .

W naukach medycznych na ogół zakłada się, że wartości pomiarów w różnych kategoriach diagnostycznych podlegają rozkładowi normalnemu. Jest to zgodne z centralnym twierdzeniem granicznym. Jeśli to założenie jest spełnione, analiza sprowadza się do wyznaczenia kilku istotnych parametrów użytych później do określenia funkcji przynależności:

1. średnia wartość j -tego symptomu w l -tej diagnozie

$$\bar{x}_j^{(l)} = \frac{1}{n_l} \sum_{i=1}^{n_l} x_{ij}^{(l)}, \quad (3.22)$$

2. odchylenie standardowe wartości j -tego symptomu w l -tej diagnozie

$$\sigma_j^{(l)} = \sqrt{\frac{1}{n_l - 1} \sum_{i=1}^{n_l} (x_{ij}^{(l)} - \bar{x}_j^{(l)})^2}, \quad (3.23)$$

3. Dolny kwartył $q_j^{(l)}$, jako maksymalna wartość $x_{ij}^{(l)} \in \mathbf{x}_j^{(l)}$ dla której

$$|X_j^*| \leq \frac{1}{4} n_l, \quad (3.24)$$

$$\text{gdzie } X_j^* = \{x_{ij}^{(l)} \leq q_j^{(l)}\}.$$

4. Górny kwartył $Q_j^{(l)}$, jako minimalna wartość $x_{ij}^{(l)} \in \mathbf{x}_j^{(l)}$ dla której

$$|X_j^{**}| \geq \frac{3}{4} n_l, \quad (3.25)$$

$$\text{gdzie } X_j^{**} = \{x_{ij}^{(l)} \geq Q_j^{(l)}\}.$$

Obliczenie pierwszych dwóch parametrów pozwala na przybliżenie wartości j -tego symptomu dla l -tej diagnozy rozkładem normalnym o funkcji gęstości

$$P_j^{(l)}(x_j) = \frac{1}{\sigma_j^{(l)} \sqrt{2\pi}} \cdot \exp\left(-\frac{(x_j - \bar{x}_j^{(l)})^2}{2(\sigma_j^{(l)})^2}\right). \quad (3.26)$$

Jeżeli obliczonych zostanie C funkcji gęstości opisujących wartości j -tego symptomu dla kolejnych C diagnoz, możliwe jest wyznaczenie punktu przecięcia się dwóch funkcji sąsiadujących ze sobą. Oznaczmy punkt przecięcia funkcji $P^{(l)}(x_j)$ i $P^{(l+1)}(x_j)$ jako $\kappa_j^{(l)}$. Ten punkt będzie także istotny w wyznaczaniu kształtów dla sąsiadujących ze sobą funkcji przynależności. Oznaczmy, dla przykładu, pierwszą funkcję gęstości jako $P^{(1)}(x_j)$ oraz drugą (sąsiadującą z P_1) jako $P^{(2)}(x_j)$ z założeniem $\bar{x}_j^{(1)} < \bar{x}_j^{(2)}$. Punkt $\kappa_j^{(1)}$ można obliczyć rozwiązując równanie

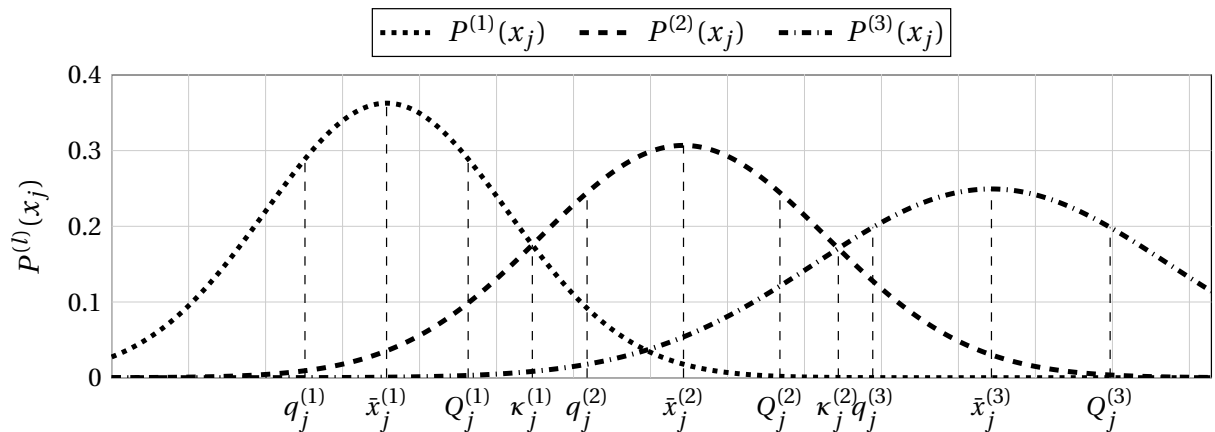
$$P^{(1)}(x_j) = P^{(2)}(x_j). \quad (3.27)$$

Na podstawie (3.26) i (3.27) łatwo zauważyć, że powyższe równanie ma charakter równania kwadratowego i jako punkt przecięcia $\kappa_j^{(l)}$ powinien zostać wybrany jeden z pierwiastków tego równania (x_{01} lub x_{02}). Możliwe jest jednak, że wartości średnie $\bar{x}_j^{(1)}$ i $\bar{x}_j^{(2)}$ nieznacznie się różnią między sobą. Wtedy oba punkty przecięcia mogą znaleźć się poza przedziałem

$[\bar{x}_j^{(1)}, \bar{x}_j^{(2)}]$. W takim wypadku, konieczne jest wybranie innego punktu aby utworzyć funkcje przynależności. Potrzeba wyznaczenia zastępczego punktu może pojawić się gdy dane uczące są nieliczne a stanowią jedyne źródło wiedzy. W wypadku takich problemów, punkt przecięcia dla tworzonych funkcji przynależności może być określony np. przez wartość położoną w połowie odległości między wartością $\bar{x}_j^{(1)}$ a $\bar{x}_j^{(2)}$ czyli $0.5 \cdot (\bar{x}_j^{(1)} + \bar{x}_j^{(2)})$. Mając na uwadze te oznaczenia, można przedstawić definicję punktu przecięcia się sąsiadujących funkcji przynależności jako

$$\kappa_j^{(l)} = \begin{cases} x_{01} \text{ jeżeli } x_{01} \in [\bar{x}_j^{(1)}, \bar{x}_j^{(2)}], \\ x_{02} \text{ jeżeli } x_{02} \in [\bar{x}_j^{(1)}, \bar{x}_j^{(2)}], \\ 0.5 \cdot (\bar{x}_j^{(1)} + \bar{x}_j^{(2)}) \text{ jeżeli } x_{01}, x_{02} \notin [\bar{x}_j^{(1)}, \bar{x}_j^{(2)}]. \end{cases} \quad (3.28)$$

Przykładową reprezentację trzech funkcji gęstości rozkładu normalnego opisujących trzy konkurujące diagnozy dla dziedziny j -tego symptomu wraz z zaznaczeniem wybranych parametrów przedstawia rysunek 3.4.



Rysunek 3.4: Wyznaczenie parametrów koniecznych do zdefiniowania funkcji przynależności dla trzech diagnoz

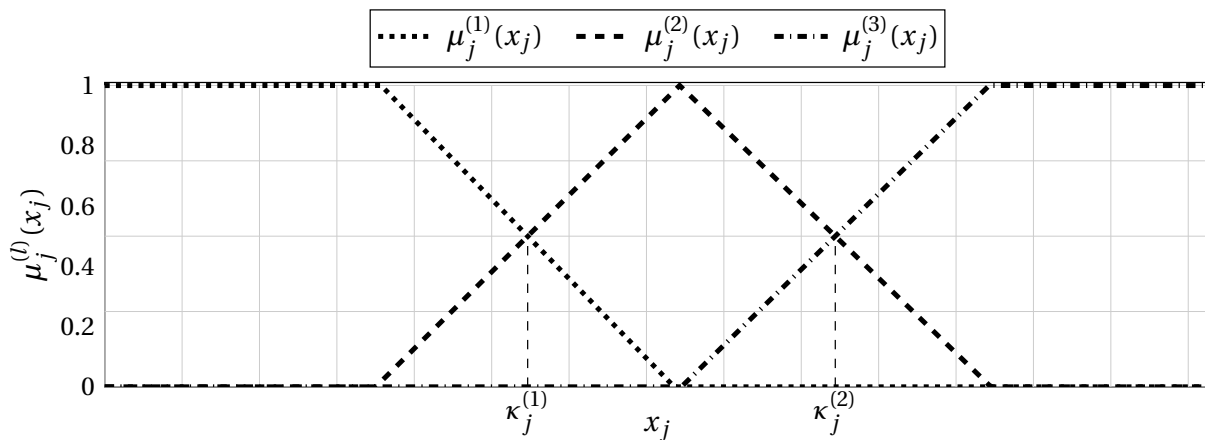
W poniższym opisie zostanie zaprezentowany sposób wyznaczania parametrów czterech funkcji przynależności: trójkątnej, trapezowej, uogólnionej dzwonowej i gaussowskiej łączonej. Użycie czterech funkcji przynależności pozwoli zbadać wpływ wyboru kształtu funkcji na wyodrębniane w eksperymentach reguły. Choć nic nie stoi na przeszkodzie aby użyć wszystkich kształtów opisanych w rozdziale 2, to użycie wybranych czterech kształtów podyktowane jest koniecznością ograniczenia objętości przedstawionych wyników oraz skupieniu się na funkcjach najłatwiej interpretowanych.

3.6.1 Trójkątna funkcja przynależności

Mając na uwadze definicję trójkątnej funkcji przynależności (2.8), parametry od (3.22) do (3.25) oraz wyznaczone punkty przecięcia sąsiadujących funkcji przynależności (3.28) parametry a, b, c mogą zostać określone jako:

- $a = \begin{cases} z \cdot \min_i(x_{ij}^{(l)}) & \text{dla } l = 1, \\ \bar{x}_j^{(l)} - 2(\bar{x}_j^{(l)} - \kappa_j^{(l-1)}) & \text{dla } 1 < l \leq C, \end{cases}$
- $b = \bar{x}_j^{(l)},$
- $c = \begin{cases} \bar{x}_j^{(l)} + 2(\kappa_j^{(l)} - \bar{x}_j^{(l)}) & \text{dla } 1 \leq l < C, \\ Z \cdot \max_i(x_{ij}^{(l)}) & \text{dla } l = C. \end{cases}$

Na rysunku 3.5 zaprezentowano trójkątne funkcje przynależności, które utworzono na podstawie funkcji gęstości z rysunku 3.4. Zarówno parametr a jak i c jest definiowany warunkowo-



Rysunek 3.5: Trójkątne funkcje przynależności dla symptomów opisanych funkcjami gęstości z rysunku 3.4

wo w zależności od numeru diagnozy, dla której tworzona jest funkcja. Na podstawie wcześniejszego założenia wartość l będzie się zmieniać od 1 do C , gdzie 1 oznacza diagnozę o najniższej a C o najwyższej wartości $\bar{x}_j^{(l)}$. Tym samym wartość a dla pierwszej i wartość c dla ostatniej funkcji przynależności zapewnia pełną przynależność przypadków o wartościach odpowiednio: minimalnych i maksymalnych dla danego symptomu. Wartość z odpowiednio mniejsza od $\min_i(x_{ij}^{(l)})$ i wartość Z odpowiednio większa od $\max_i(x_{ij}^{(l)})$ spowodują odpowiednio: lewostronne i prawostronne otwarcie odpowiedniej trójkątnej funkcji przynależności.

W przypadku parametrów dla $x_j^{(l)}$ o pośrednich wartościach $\bar{x}_j^{(l)}$ parametry a i c określone są tak aby zapewnić połowę przynależności wartościom zlokalizowanym w punkcie przecięcia $\kappa_j^{(l)}$, to znaczy

$$\mu_j^{(l)}(\kappa_j^{(l)}) = \mu_j^{(l+1)}(\kappa_j^{(l)}) = 0.5. \quad (3.29)$$

3.6.2 Trapezowa funkcja przynależności

Trapezowa funkcja przynależności jako uogólnienie trójkątnej funkcji przynależności pozwala na określenie pełnej przynależności nie tylko dla pojedynczej wartości rzeczywistej, określonej dla trójkątnej funkcji jako b , lecz dla przedziału. Do zdefiniowania tego przedziału wykorzystane mogą być wartości dolnego ($q_j^{(l)}$) i górnego ($Q_j^{(l)}$) kwartyla. Z uwagi na to,

że definicja trapezowych funkcji przynależności także uzależniona jest od ich wzajemnego położenia, wartości parametrów a , b , c , d określone są następująco:

- dla funkcji przynależności o najniższych wartościach nośnika ($l = 1$):

$$\begin{aligned} a &= z \cdot \min_i(x_{ij}^{(l)}), \\ b &= \min_i(x_{ij}^{(l)}), \\ c &= Q_j^{(l)} + \alpha \left| \kappa_j^{(l)} - Q_j^{(l)} \right|, \\ d &= c + 2(\kappa_j^{(l)} - c), \end{aligned}$$

- dla funkcji przynależności o pośrednich wartościach nośnika ($1 < l < C$):

$$\begin{aligned} b &= q_j^{(l)} - \alpha \left| q_j^{(l)} - \kappa_j^{(l-1)} \right|, \\ a &= b - 2(b - \kappa_j^{(l-1)}), \\ c &= Q_j^{(l)} + \alpha \left| \kappa_j^{(l)} - Q_j^{(l)} \right|, \\ d &= c + 2(\kappa_j^{(l)} - c), \end{aligned}$$

- dla funkcji przynależności o najwyższych wartościach nośnika ($l = C$):

$$\begin{aligned} b &= q_j^{(l)} - \alpha \left| q_j^{(l)} - \kappa_j^{(l-1)} \right|, \\ a &= b - 2(b - \kappa_j^{(l-1)}), \\ c &= \max_i(x_{ij}^{(l)}), \\ d &= Z \cdot \max_i(x_{ij}^{(l)}). \end{aligned}$$

Cel zastosowania wartości z i Z jest taki sam jak przy trójkątnych funkcjach przynależności (punkt 3.6.1). W powyższym zapisie występuje dodatkowo parametr α . Wpływa on na położenie wierzchołków trapezów b (dla funkcji tworzonych dla $l > 1$) oraz c (dla funkcji tworzonych dla $l < C$). Jeżeli α będzie równe zero, wierzchołki b i c będą określone przez wartość odpowiedniego kwartyla. Zwiększanie parametru α w przedziale $[0, 1)$ będzie przesunąć odpowiednie wierzchołki w stronę właściwych punktów $\kappa_j^{(l)}$. Tym samym parametr α umożliwia zmianę nachylenia zboczy trapezowej funkcji przynależności i określenie przedziału pełnej przynależności do zbioru rozmytego. W sytuacji, gdy $q_j^{(l)} \leq \kappa_j^{(l-1)}$ lub $Q_j^{(l)} \geq \kappa_j^{(l)}$ funkcja będzie określona niepoprawnie, tzn. niezgodnie z (3.29). W takim wypadku wystarczy dodanie warunku przy tworzeniu funkcji przynależności:

- jeżeli $q_j^{(l)} \leq \kappa_j^{(l-1)}$ ($1 < l \leq C$):

$$\begin{aligned} a &= 0.99 \cdot \kappa_j^{(l-1)}, \\ b &= 1.01 \cdot \kappa_j^{(l-1)} \end{aligned}$$

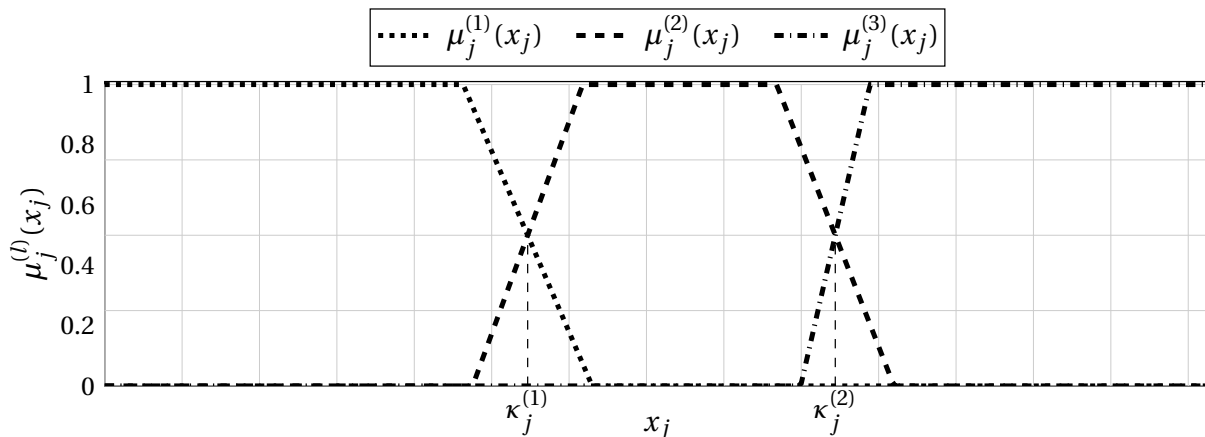
oraz

- jeżeli $Q_j^{(l)} \geq \kappa_j^{(l)}$ ($1 \leq l < C$):

$$c = 0.99 \cdot \kappa_j^{(l)},$$

$$d = 1.01 \cdot \kappa_j^{(l)}.$$

Dwa powyższe warunki, jeśli zaistnieje taka potrzeba, spowodują, że nachylenie funkcji przynależności będzie odpowiednio strome. Przykład utworzenia trapezowych funkcji przynależności przedstawiono na rysunku 3.6. Na rysunku 3.7 przedstawiono syntetyczny przykład wyznaczania trapezowej funkcji przynależności (dla $\alpha = 0$) bez konieczności korekcji parametrów (rysunek 3.7a) i z koniecznością korekcji parametrów c i d (rysunek 3.7b).



Rysunek 3.6: Trapezowe funkcje przynależności dla funkcji gęstości z rysunku 3.4

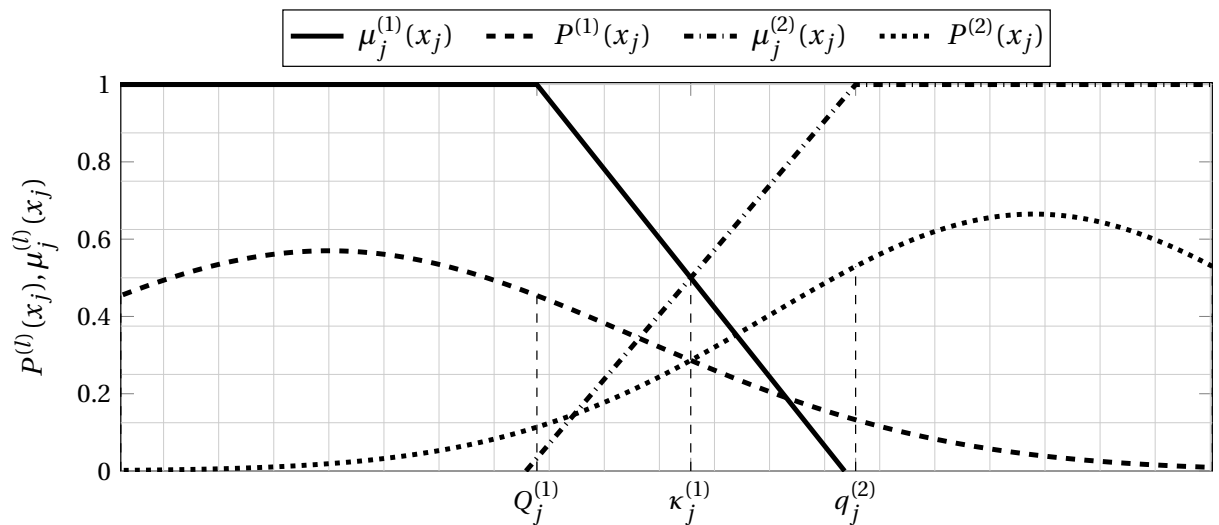
3.6.3 Uogólniona dzwonowa funkcja przynależności

Dla przedstawionej definicji (2.13) parametry a oraz c mogą być zdefiniowane następująco:

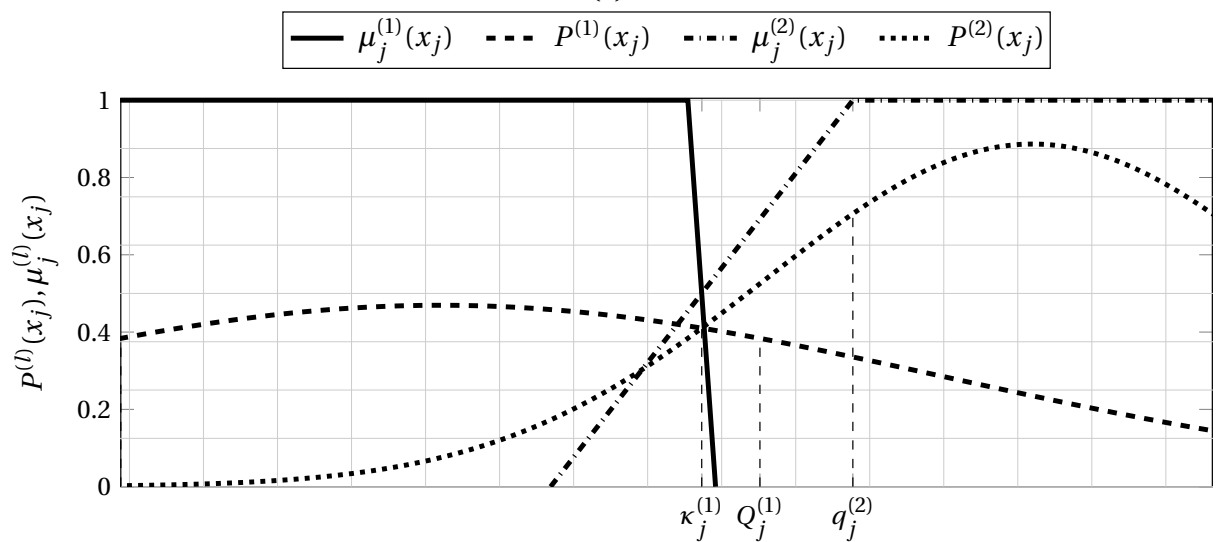
$$\bullet a = \begin{cases} \frac{1}{2} \cdot (\min_i(x_{ij}^{(l)}) + \kappa_j^{(l)}) & \text{dla } l = 1, \\ \frac{1}{2} \cdot (\kappa_j^{(l-1)} + \kappa_j^{(l)}) & \text{dla } 1 < l < C, \\ \frac{1}{2} \cdot (\kappa_j^{(l-1)} + \max_i(x_{ij}^{(l)})) & \text{dla } l = C, \end{cases}$$

$$\bullet c = \begin{cases} \frac{1}{2} \cdot (\kappa_j^{(l)} - \min_i(x_{ij}^{(l)})) & \text{dla } l = 1, \\ \frac{1}{2} \cdot (\kappa_j^{(l)} - \kappa_j^{(l-1)}) & \text{dla } 1 < l < C, \\ \frac{1}{2} \cdot (\max_i(x_{ij}^{(l)}) - \kappa_j^{(l-1)}) & \text{dla } l = C. \end{cases}$$

Dla parametru a określonego jak powyżej, wartość funkcji przynależności jest równa 0.5 w punktach przecięcia z sąsiednimi funkcjami. Parametr c decyduje o przedziale wartości dla których funkcja przyjmie pełną przynależność. Parametr b , może być strojony dla uzyskania pożądanego nachylenia zboczy. Przykład uogólnionych dzwonowych funkcji przynależności przedstawiono na rysunku 3.8.

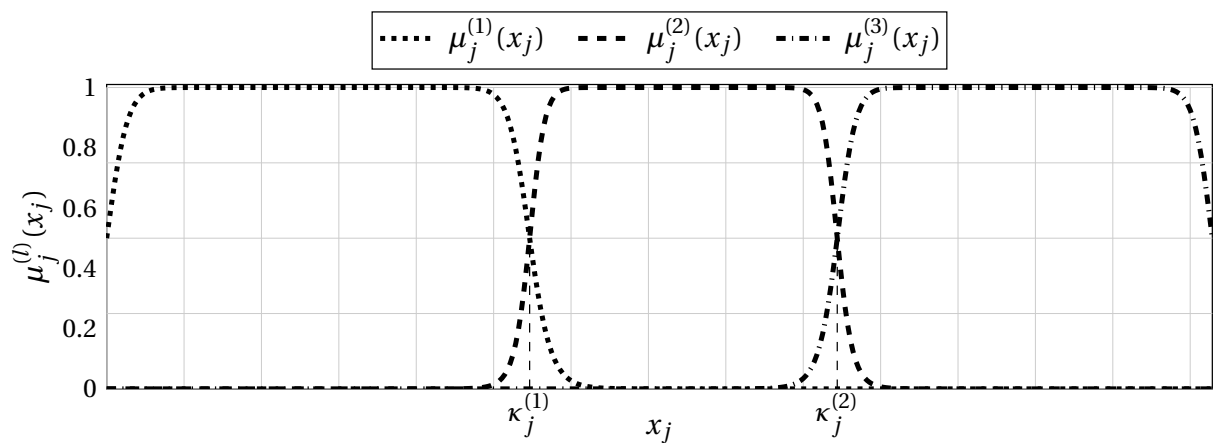


(a)



(b)

Rysunek 3.7: Trapezowe funkcje przynależności na podstawie danych opisanych przez $P^{(1)}$, $P^{(2)}$: (a) bez konieczności, (b) z koniecznością korekcji wartości jej parametrów



Rysunek 3.8: Uogólnione dzwonowe funkcje przynależności dla symptomów opisanych funkcjami gęstości z rysunku 3.4

3.6.4 Gaussowska łączona funkcja przynależności

Tak jak dla poprzednich, określenie gaussowskiej łączonej funkcji przynależności jest uzależnione od średnich danych dla różnych diagnoz. Funkcje te są jednostronnie otwarte dla rozkładów o ekstremalnych wartościach średnich. Konieczne jest więc uzależnienie definicji od uporządkowania diagnoz:

- dla funkcji przynależności o najniższych wartościach nośnika ($l = 1$):

$$\mu_j^{(1)} = \begin{cases} 1 & \text{dla } x_j \leq \bar{x}_j^{(l)}, \\ \exp \frac{-(x_j - \bar{x}_j^{(l)})^2}{2\sigma_2^2} & \text{dla } x_j > \bar{x}_j^{(l)}, \end{cases} \quad (3.30)$$

- dla funkcji przynależności o pośrednich wartościach nośnika ($1 < l < C$):

$$\mu_j^{(l)} = \begin{cases} \exp \frac{-(x_j - \bar{x}_j^{(l)})^2}{2\sigma_1^2} & \text{dla } x_j \leq \bar{x}_j^{(l)}, \\ \exp \frac{-(x_j - \bar{x}_j^{(l)})^2}{2\sigma_2^2} & \text{dla } x_j > \bar{x}_j^{(l)}, \end{cases} \quad (3.31)$$

- dla funkcji przynależności o najwyższych wartościach nośnika ($l = C$):

$$\mu_j^{(C)} = \begin{cases} \exp \frac{-(x_j - \bar{x}_j^{(l)})^2}{2\sigma_1^2} & \text{dla } x_j \leq \bar{x}_j^{(l)}, \\ 1 & \text{dla } x_j > \bar{x}_j^{(l)}. \end{cases} \quad (3.32)$$

Dla sąsiednich funkcji przynależności konieczne jest zapewnienie połowy przynależności w punktach przecięcia $\kappa_j^{(l)}$. Dlatego należy wyznaczyć wartości parametrów σ_1 oraz σ_2 (które występują w (3.30)-(3.32)) spełniających równania:

$$0.5 = \exp \frac{-(\kappa_j^{(l-1)} - \bar{x}_j^{(l)})^2}{2\sigma_1^2} \quad (3.33)$$

oraz

$$0.5 = \exp \frac{-(\kappa_j^{(l)} - \bar{x}_j^{(l)})^2}{2\sigma_2^2}, \quad (3.34)$$

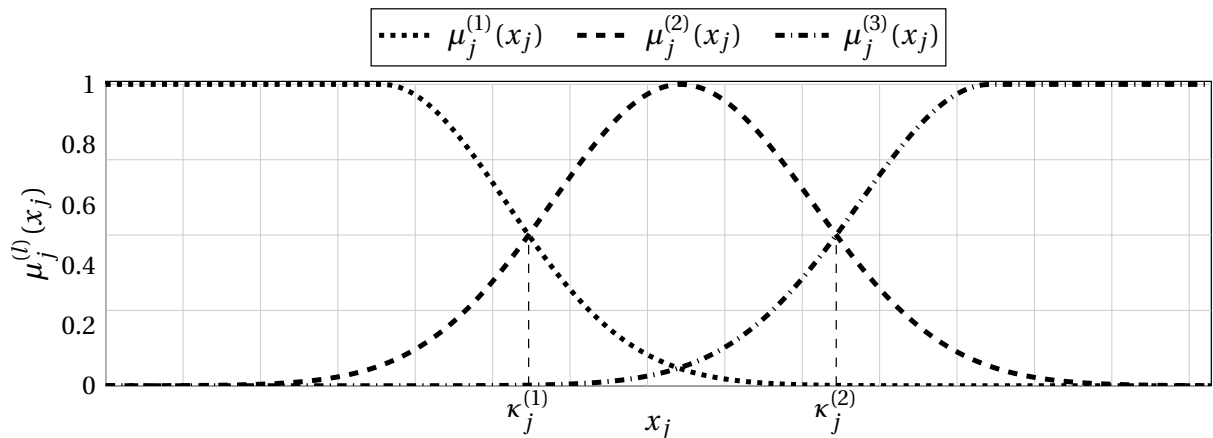
co po przekształceniu daje:

$$\sigma_1 = \sqrt{\frac{-(\kappa_j^{(l-1)} - \bar{x}_j^{(l)})^2}{2 \ln(0.5)}} \quad (3.35)$$

oraz

$$\sigma_2 = \sqrt{\frac{-(\kappa_j^{(l)} - \bar{x}_j^{(l)})^2}{2 \ln(0.5)}}. \quad (3.36)$$

Przykład gaussowskich łączonych funkcji przynależności przedstawiono na rysunku 3.9.



Rysunek 3.9: Gaussowskie łączone funkcje przynależności określone dla symptomów opisanych funkcjami gęstości z rysunku 3.4

3.7 Sposoby oceny reguł diagnostycznych

Ocena tworzonych reguł diagnostycznych ma kluczowe znaczenie dla jakości wspomaganie diagnozy, ponieważ wpływa bezpośrednio na wynik algorytmów wydobywania reguł. W początkowych podejściach do oceny pojedynczej reguły użyto wartości bazowego prawdopodobieństwa [94, 109], która określa wagę reguły dla wybranej diagnozy. W tym punkcie przedstawione zostanie jednak krytyczne spojrzenie na rozkład bazowego prawdopodobieństwa jako wstępną ocenę reguł. Jako rozwiązanie problemu, zaproponowane zostaną inne metody oceny rozmytych elementów ogniskowych, jako przesłanek reguł diagnostycznych.

3.7.1 Ocena reguł na podstawie wartości bazowego prawdopodobieństwa

Rozkład bazowego prawdopodobieństwa przedstawiony w punkcie 3.3 jest miarą związku elementów ogniskowych z odpowiednią diagnozą. Jeżeli przypadek $\mathbf{p}_i^{(l)}$ z bazy $\mathbf{X}^{(l)}$ jest dopasowany do r -tego elementu ogniskowego przynajmniej na poziomie zadanym przez próg η_e (3.11) to dla r -tej reguły przypadek interpretuje się jako należący do l -tej klasy. Taki wynik w statystyce określa się mianem prawdziwie dodatnim (ang. *True Positive*). Ponieważ do wyznaczenia bazowego prawdopodobieństwa są wybierane tylko przypadki ze zbioru l -tej diagnozy to wartość bazowego prawdopodobieństwa nie niesie żadnej informacji na temat niepożądanego dopasowania r -tego elementu ogniskowego z przypadkami z innych diagnoz.

Bazowe prawdopodobieństwo ma kluczowe znaczenie w wyznaczaniu miary przekonania i jego rola w tym procesie jest niepodważalna. Jednakże, jeśli chodzi o samą ocenę reguł, możliwe jest użycie i przetestowanie innych metod, które lepiej oddadą dopasowanie reguły do danego problemu medycznego.

3.7.2 Ocena na podstawie MCC

Sprawdzając dopasowanie elementu ogniskowego z wszystkimi przypadkami z bazy $X^{(l)}$, $l = 1, \dots, C$ jako wynik klasyfikacji na poziomie r -tej reguły można otrzymać cztery rezultaty określane jako:

- $TP^{(l)}$ (ang. *True Positive* – prawdziwie dodatni) – liczba przypadków l -tej diagnozy poprawnie rozpoznanych jako przypadki l -tej diagnozy,
- $FN^{(l)}$ (ang. *False Negative* – fałszywie ujemny) – liczba przypadków l -tej diagnozy niepoprawnie odrzuconych z l -tej diagnozy,
- $FP^{(l)}$ (ang. *False Positive* – fałszywie dodatni) – liczba przypadków l' -tej diagnozy ($l' \neq l$) niepoprawnie rozpoznanych jako przypadki l -tej diagnozy,
- $TN^{(l)}$ (ang. *True Negative* – prawdziwie ujemny) – liczba przypadków l' -tej diagnozy ($l' \neq l$) poprawnie odrzuconych z l -tej diagnozy.

Z uwagi na to, że w niniejszej rozprawie powyższe oznaczenia zostaną użyte także do oceny jakości wspomaganiania diagnozy, oznaczanie w tym punkcie zostaną zapisane małymi literami, aby jednoznacznie wskazywały na ocenę reguły. Tym samym miary poprawności klasyfikacji w ocenie reguł można obliczyć następująco:

$$tp_r^{(l)} = \sum_{i=1}^{n_l} I(\eta_r^{(l)}(\mathbf{p}_i^{(l)}) > \eta_e), \quad (3.37)$$

$$fn_r^{(l)} = n_l - \sum_{i=1}^{n_l} I(\eta_r^{(l)}(\mathbf{p}_i^{(l)}) > \eta_e) = n_l - tp_r^{(l)}, \quad (3.38)$$

$$fp_r^{(l)} = \sum_{\substack{1 \leq l' \leq C \\ l' \neq l}} \left(\sum_{i=1}^{n_{l'}} I(\eta_r^{(l)}(\mathbf{p}_i^{(l')}) > \eta_e) \right), \quad (3.39)$$

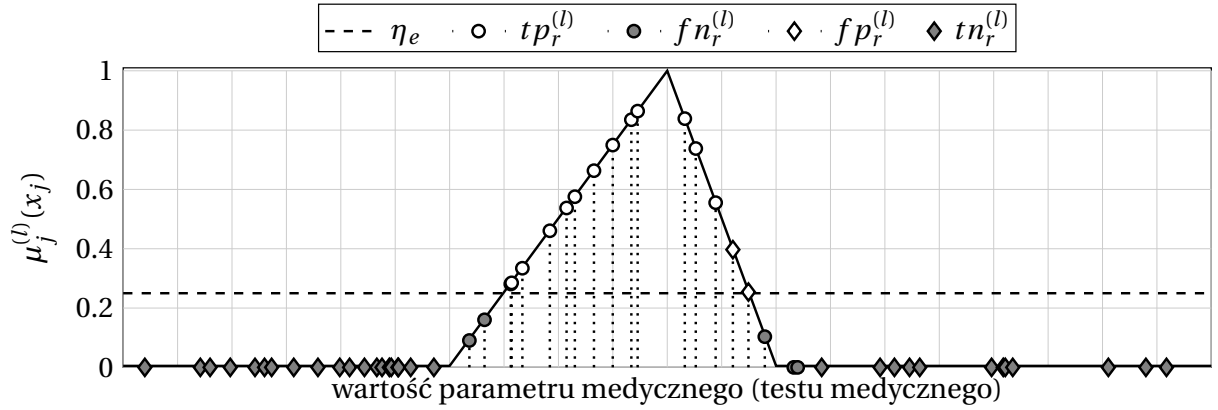
$$tn_r^{(l)} = \sum_{\substack{1 \leq l' \leq C \\ l' \neq l}} \left(n_{l'} - \sum_{i=1}^{n_{l'}} I(\eta_r^{(l)}(\mathbf{p}_i^{(l')}) > \eta_e) \right) = N - n_l - fp_r^{(l)}. \quad (3.40)$$

Na rysunku 3.10 przedstawiono cztery możliwe wyniki dopasowania przypadków do elementu ogniskowego.

Wskaźnik jakości klasyfikacji *MCC* (ang. *Matthews Correlation Coefficient*) [75] bierze pod uwagę wszystkie możliwe przypadki dopasowania i w celu oceny reguły może być opisany jako

$$MCC(s_r^{(l)}) = \frac{(tp_r^{(l)} \cdot tn_r^{(l)}) - (fp_r^{(l)} \cdot fn_r^{(l)})}{\sqrt{(tp_r^{(l)} + fp_r^{(l)}) \cdot (tp_r^{(l)} + fn_r^{(l)}) \cdot (tn_r^{(l)} + fp_r^{(l)}) \cdot (tn_r^{(l)} + fn_r^{(l)})}}. \quad (3.41)$$

Dla ułatwienia, podobnie jak przy wartości bazowego prawdopodobieństwa, wartość *MCC* obliczoną dla r -tego elementu ogniskowego l -tej diagnozy można oznaczyć jako $MCC_r^{(l)}$. Maksymalna wartość *MCC* może wynieść 1. Oznacza to, że do rozmytego elementu ogniskowego pasują wszystkie przypadki l -tej diagnozy, ale żaden przypadek z innych diagnoz



Rysunek 3.10: Przykład dopasowania danych do elementu ogniskowego. Okręgi reprezentują przypadki z l -tej diagnozy a romby przypadki z l' -tej diagnozy. Białe kształty wskazują na dopasowanie przypadku i elementu ogniskowego lepsze od przyjętego progu, a szare przeciwnie.

(występują tylko klasyfikacje prawdziwie dodatnie oraz prawdziwie ujemne). Wystąpienie klasyfikacji niepoprawnych ($fp_r^{(l)}$ oraz $fn_r^{(l)}$) pomniejsza MCC . Spadek MCC do zera oznacza klasyfikację nie lepszą od losowej. Otrzymanie wartości wskaźnika równej -1 wskazuje na całkowitą niezgodność wyniku klasyfikacji z danymi. Wskaźnik MCC uważa się za dobrą metodę wyrażenia macierzy pomyłek klasyfikatora za pomocą jednej wartości [98]. Jednak nie ilustruje on zmian w wartościach macierzy pomyłek. Z drugiej strony nietrudno przewidzieć, że MCC przyjmował będzie największe wartości gdy wartości $fp_r^{(l)}$ i $fn_r^{(l)}$ są najmniejsze i a $tp_r^{(l)}$ i $tn_r^{(l)}$ największe. Takie trafne dopasowanie może występować dla najbardziej złożonych elementów ogniskowych (z wieloma warunkami logicznymi). Tworzenie bardzo złożonych elementów ogniskowych nie jest jednak korzystne z punktu widzenia wspomagania diagnozy. Mając na uwadze nie tylko wysoką efektywność ale także czytelność i interpretowalność reguł zaproponowano wskaźnik oceny rozmytych elementów ogniskowych o nieco innych własnościach.

3.7.3 Ocena na podstawie współczynnika zgodności

Proponuje się wskaźnik, który będzie uwzględniał prawidłowe dopasowanie przypadków z l -tej diagnozy oraz niepożądane dopasowanie przypadków z innych diagnoz do elementu ogniskowego dla l -tej diagnozy. Niech wskaźnik ten będzie oznaczony przez CC i zdefiniowany następująco

$$CC(s_r^{(l)}) = CC_r^{(l)} = \max \left(0, \frac{\sum_{i=1}^{n_l} I(\eta_r^{(l)}(\mathbf{p}_i^{(l)}) > \eta_e) - \sum_{\substack{1 \leq l' \leq C \\ l' \neq l}} \left(\sum_{i=1}^{n_{l'}} I(\eta_r^{(l')}(\mathbf{p}_i^{(l')}) > \eta_e) \right)}{\sum_{r=1}^{n^{(l)}} \left(\sum_{i=1}^{n_l} I(\eta_r^{(l)}(\mathbf{p}_i^{(l)}) > \eta_e) \right)} \right) \quad (3.42)$$

$$= \max \left(0, \frac{tp_r^{(l)} - fp_r^{(l)}}{\sum_{i=1}^{n_l} tp_r^{(l)}} \right).$$

Pierwszy składnik w liczniku to suma poprawnie dopasowanych przypadków l -tej klasy do r -tego elementu ogniskowego czyli wartość $tp_r^{(l)}$. Od tej wartości odejmowany jest drugi element, który dotyczy sumy niepożądanych dopasowań przypadków z innych diagnoz (oznaczonych jako l' , gdzie $l' \neq l$) do r -tego elementu ogniskowego ($fp_r^{(l)}$). Mianownik stanowi suma wszystkich $tp_r^{(l)}$, $r = 1, \dots, n^{(l)}$ czyli dopasowań do wszystkich elementów ogniskowych w zbiorze $S^{(l)}$. Podzielenie licznika przez mianownik w (3.42) nie jest konieczne, ale umożliwia porównanie z wynikami uzyskanymi dla bazowego rozkładu prawdopodobieństwa. W ten sposób CC jest współczynnikiem, którego wartość zwiększa się wraz z liczbą poprawnych dopasowań ($tp_r^{(l)}$), a zmniejsza się przy niepożądanych dopasowaniach ($fp_r^{(l)}$).

3.8 Ocena reguł na przykładzie bazy *iris flowers*

Baza *iris flowers* nie jest bazą medyczną a zawiera 150 pomiarów związanych z kwiatami irysa z trzech równolicznych klas związanych z gatunkami *setosa*, *versicolor* oraz *virginica*. Niemniej jednak, baza ta zostanie wykorzystana jako przykładowa, ponieważ jest najczęściej używana do ilustracji różnych metod w szeroko pojętym rozpoznawaniu obrazów [34, 67, 94]. We wspomnianej bazie, pierwsze cztery parametry opisują dane osobnika, a piąty parametr jest przypisaniem do jednej z trzech wspomnianych klas (gatunków). Tym samym mamy do czynienia z czterema parametrami ($\rho = 4$) i trzema klasami ($C = 3$). Przywołując zależność (3.8), wiadomo, że dla każdej klasy można określić maksymalnie 15 elementów ogniskowych.

Na podstawie pełnego zbioru danych i metodyki przedstawionej w punkcie 3.6 zaprojektowano funkcje przynależności. Wszystkie projektowane funkcje przynależności dla bazy *iris flowers* znajdują się w Dodatku A na rysunkach A.1a-A.1d.

Początkowe zbiory zawierają następujące elementy ogniskowe:

$$\begin{aligned}
 s_1^{(l)} &: X_1 \text{ jest } A_1^{(l)}, \\
 s_2^{(l)} &: X_2 \text{ jest } A_2^{(l)}, \\
 s_3^{(l)} &: X_3 \text{ jest } A_3^{(l)}, \\
 s_4^{(l)} &: X_4 \text{ jest } A_4^{(l)}, \\
 &\vdots \\
 s_5^{(l)} &: X_1 \text{ jest } A_1^{(l)} \wedge X_2 \text{ jest } A_2^{(l)}, \\
 s_6^{(l)} &: X_1 \text{ jest } A_1^{(l)} \wedge X_3 \text{ jest } A_3^{(l)}, \\
 &\vdots \\
 s_{11}^{(l)} &: X_1 \text{ jest } A_1^{(l)} \wedge X_2 \text{ jest } A_2^{(l)} \wedge X_3 \text{ jest } A_3^{(l)}, \\
 s_{12}^{(l)} &: X_1 \text{ jest } A_1^{(l)} \wedge X_2 \text{ jest } A_2^{(l)} \wedge X_4 \text{ jest } A_4^{(l)}, \\
 &\vdots \\
 s_{15}^{(l)} &: X_1 \text{ jest } A_1^{(l)} \wedge X_2 \text{ jest } A_2^{(l)} \wedge X_3 \text{ jest } A_3^{(l)} \wedge X_4 \text{ jest } A_4^{(l)}.
 \end{aligned}$$

W tabeli 3.4 przedstawiono liczbę dopasowań elementów ogniskowych związanych z drugą klasą ($l = 2$) z przypadkami z każdej klasy w bazie *iris flowers*. Oznaczenie nagłówków kolumn prezentujących liczbę dopasowań elementów ogniskowych związanych z drugą klasą ($l = 2$) z przypadkami z klas pierwszej, drugiej i trzeciej uproszczono do postaci:

$$ld^{(2,1)} = \sum_{i=1}^{n_1} \left(I(\eta_r^{(2)}(\mathbf{p}_i^{(1)}) > \eta_e) \right), \quad (3.43)$$

$$ld^{(2,2)} = \sum_{i=1}^{n_2} \left(I(\eta_r^{(2)}(\mathbf{p}_i^{(2)}) > \eta_e) \right), \quad (3.44)$$

$$ld^{(2,3)} = \sum_{i=1}^{n_3} \left(I(\eta_r^{(2)}(\mathbf{p}_i^{(3)}) > \eta_e) \right) \quad (3.45)$$

i zgodnie z definicjami (3.37) i (3.39) w punkcie 3.7.2 warto zauważyć, że zachodzą równości:

$$tp_r^{(2)} = ld^{(2,2)} \quad (3.46)$$

oraz

$$fp_r^{(2)} = ld^{(2,1)} + ld^{(2,3)}. \quad (3.47)$$

Na podstawie analizy tabeli 3.4 można zauważyć, że maksymalną wartość bazowego prawdopodobieństwa otrzymują elementy ogniskowe $s_3^{(2)}$ i $s_4^{(2)}$, dla których liczba dopasowań z przypadkami z drugiej klasy wynosi 48. Jak jednak widać, jest kilka przypadków spoza klasy drugiej, które pasują do tych elementów ogniskowych. Natomiast ważny wydaje się element ogniskowy $s_{10}^{(2)}$, dla którego liczba dopasowań z przypadkami klasy drugiej wynosi 47, a do którego nie pasuje żaden przypadek z innych klas. Otrzymuje on jednak mniejszą wartość bazowego prawdopodobieństwa od $s_3^{(2)}$ czy $s_4^{(2)}$, choć w przeprowadzonej ocenie wydaje się od nich lepszy.

W tabeli 3.5 zawarto wartości oceny elementów ogniskowych ze zbioru $S^{(2)}$ dla bazy *iris flowers* za pomocą BPA i wskaźników zdefiniowanych w (3.41) i (3.42) w punkcie 3.7.2. Znormalizowane wartości ocen dla każdego elementu ogniskowego zostały także przedstawione na rysunku 3.11.

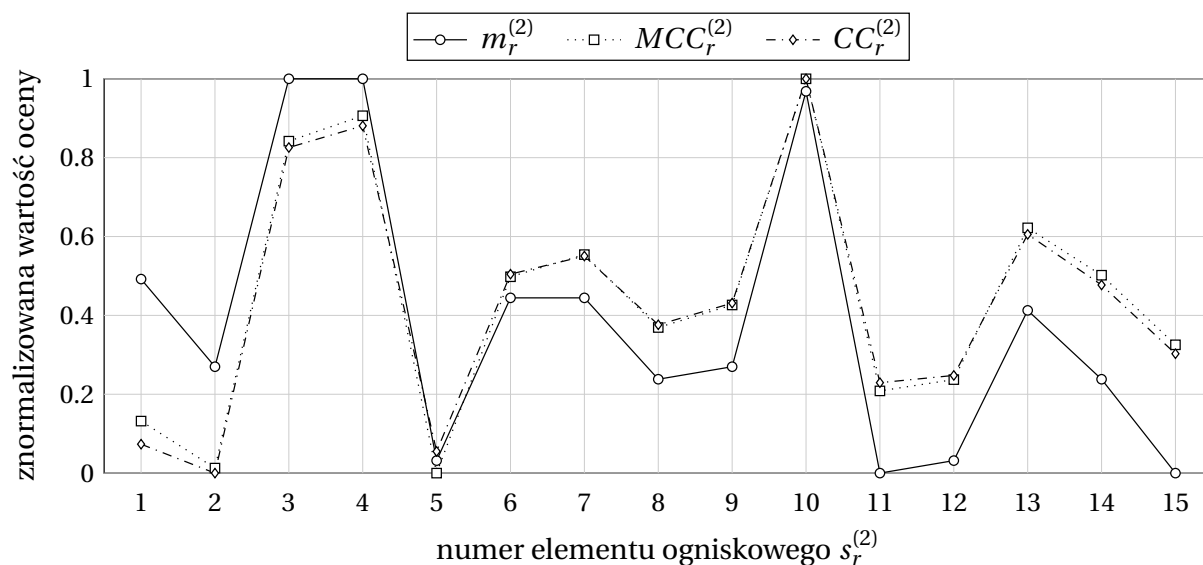
Na podstawie tabeli 3.5 oraz rysunku 3.11 można zauważyć, że wprowadzone zależności zmieniły ocenę elementów ogniskowych. Wskaźniki $MCC_r^{(2)}$ i $CC_r^{(2)}$ mają większe wartości dla $s_{10}^{(2)}$, niż dla $s_3^{(2)}$ i $s_4^{(2)}$. Tak więc $s_{10}^{(2)}$ został oceniony jako lepiej opisujący klasę *versicolor* niż dwa wspomniane elementy ogniskowe, inaczej niż to sugeruje BPA w tabeli 3.4. Na podstawie rysunku 3.11 dostrzega się tylko nieznaczne różnice między znormalizowanymi wartościami MCC i CC . Niemniej jednak, w przeprowadzanych badaniach dla baz medycznych

Tabela 3.4: Liczba przypadków z bazy *iris flowers* dopasowanych do elementów ogniskowych należących do zbioru $S^{(2)}$ i odpowiadające im wartości bazowego prawdopodobieństwa

$s_r^{(2)}$	$ld^{(2,1)}$	$ld^{(2,2)}$	$ld^{(2,3)}$	$m_r^{(2)}$
$s_1^{(2)}$	5	33	18	0.073
$s_2^{(2)}$	1	27	19	0.059
$s_3^{(2)}$	0	48	6	0.105
$s_4^{(2)}$	1	48	4	0.105
$s_5^{(2)}$	0	20	11	0.044
$s_6^{(2)}$	0	32	5	0.070
$s_7^{(2)}$	0	32	3	0.070
$s_8^{(2)}$	0	26	4	0.057
$s_9^{(2)}$	0	27	3	0.059
$s_{10}^{(2)}$	0	47	0	0.103
$s_{11}^{(2)}$	0	19	3	0.042
$s_{12}^{(2)}$	0	20	3	0.044
$s_{13}^{(2)}$	0	31	0	0.068
$s_{14}^{(2)}$	0	26	0	0.057
$s_{15}^{(2)}$	0	19	0	0.042

Tabela 3.5: Bazowe prawdopodobieństwo, *Matthews Correlation Coefficient* i współczynnik zgodności obliczone dla zbioru elementów ogniskowych $S^{(2)}$

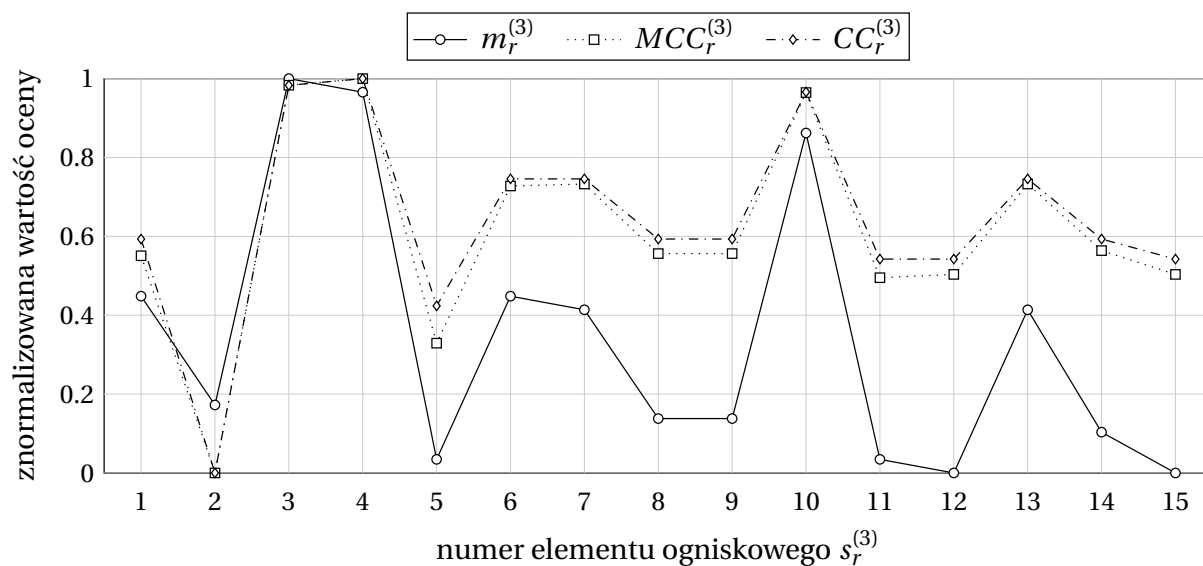
$s_r^{(2)}$	$tp_r^{(2)}$	$fp_r^{(2)}$	$tn_r^{(2)}$	$fn_r^{(2)}$	$m_r^{(2)}$	$MCC_r^{(2)}$	$CC_r^{(2)}$
$s_1^{(2)}$	33	23	77	17	0.073	0.419	0.022
$s_2^{(2)}$	27	20	80	23	0.059	0.346	0.015
$s_3^{(2)}$	48	6	94	2	0.105	0.884	0.092
$s_4^{(2)}$	48	5	95	2	0.105	0.897	0.095
$s_5^{(2)}$	20	11	89	30	0.044	0.338	0.020
$s_6^{(2)}$	32	5	95	18	0.070	0.645	0.059
$s_7^{(2)}$	32	3	97	18	0.070	0.680	0.064
$s_8^{(2)}$	26	4	96	24	0.057	0.566	0.048
$s_9^{(2)}$	27	3	97	23	0.059	0.601	0.053
$s_{10}^{(2)}$	47	0	100	3	0.103	0.955	0.103
$s_{11}^{(2)}$	19	3	97	31	0.042	0.466	0.035
$s_{12}^{(2)}$	20	3	97	30	0.044	0.484	0.037
$s_{13}^{(2)}$	31	0	100	19	0.068	0.722	0.068
$s_{14}^{(2)}$	26	0	100	24	0.057	0.648	0.057
$s_{15}^{(2)}$	19	0	100	31	0.042	0.539	0.042



Rysunek 3.11: Znormalizowane wartości ocen opartej na m , MCC oraz CC dla elementów ogniskowych utworzonych dla drugiej klasy bazy *iris flowers*

może być pożądanym użycie wszystkich opisanych metod oceny reguł ze względu na to, że dla większej liczby elementów ogniskowych różnice w ocenie mogą mieć kluczowe znaczenie.

Klasa *iris versicolor* została wybrana do przedstawienia sposobu oceny elementów ogniskowych ze względu na zauważone dopasowanie przypadków z innych klas do jej elementów ogniskowych. W klasie *iris setosa* analogiczne zjawiska nie występują gdyż jest to klasa łatwo rozdzielna od pozostałych. Natomiast w klasie *iris virginica* zauważa się podobny wpływ zaproponowanych metod oceny. Dlatego w tabeli 3.6 przedstawiono wartości macierzy pomyłek i wyznaczone oceny elementów ogniskowych dla trzeciej klasy a na rysunku 3.12 ich znormalizowane oceny. Można zauważyć różnice w ocenie elementów ogniskowych podobne do występujących dla *versicolor*.



Rysunek 3.12: Znormalizowane wartości ocen opartej na m , MCC oraz CC dla elementów ogniskowych utworzonych dla trzeciej klasy bazy *iris flowers*

Tabela 3.6: Bazowe prawdopodobieństwo, *Matthews Correlation Coefficient* i współczynnik zgodności obliczone dla zbioru elementów ogniskowych $S^{(3)}$

$s_r^{(l)}$	$tp_r^{(3)}$	$fp_r^{(3)}$	$tn_r^{(3)}$	$fn_r^{(3)}$	$m_r^{(3)}$	$MCC_r^{(3)}$	$CC_r^{(3)}$
$s_1^{(2)}$	31	11	89	19	0.074	0.535	0.048
$s_2^{(2)}$	23	38	62	27	0.055	0.077	0.000
$s_3^{(2)}$	47	4	96	3	0.112	0.896	0.102
$s_4^{(2)}$	46	2	98	4	0.110	0.910	0.105
$s_5^{(2)}$	19	9	91	31	0.045	0.351	0.024
$s_6^{(2)}$	31	2	98	19	0.074	0.683	0.069
$s_7^{(2)}$	30	1	99	20	0.071	0.687	0.069
$s_8^{(2)}$	22	2	98	28	0.052	0.540	0.048
$s_9^{(2)}$	22	2	98	28	0.052	0.540	0.048
$s_{10}^{(2)}$	43	1	99	7	0.102	0.880	0.100
$s_{11}^{(2)}$	19	2	98	31	0.045	0.489	0.040
$s_{12}^{(2)}$	18	1	99	32	0.043	0.496	0.040
$s_{13}^{(2)}$	30	1	99	20	0.071	0.687	0.069
$s_{14}^{(2)}$	21	1	99	29	0.050	0.546	0.048
$s_{15}^{(2)}$	18	1	99	32	0.043	0.496	0.040

3.9 Podsumowanie

W niniejszym rozdziale zaprezentowano sposób reprezentacji wiedzy medycznej za pomocą rozmytych reguł warunkowych i model wspomaganie diagnozy z użyciem miar przekonania. Do reprezentacji reguł diagnostycznych zaproponowano cztery kształty funkcji przynależności, które mogą być użyte w rozmytych elementach ogniskowych w przesłankach reguł. Badanie wpływu kształtu funkcji przynależności na działanie algorytmów wspomaganie diagnozy będzie przeprowadzone w rozdziale poświęconym eksperymentom. Dodatkowo, zaproponowano odmienny wskaźnik oceny reguł diagnostycznych, który także będzie przedmiotem badań w eksperymentach przeprowadzonych dla medycznych baz danych.

Rozdział 4

Algorytmy wydobywania reguł diagnostycznych

W niniejszym rozdziale przedstawione zostaną proponowane podejścia do wydobywania reguł diagnostycznych bezpośrednio z danych medycznych. Metody wykorzystujące teorię Dempstera-Shafera i teorię zbiorów rozmytych opisane w rozdziale 3 zapewniają uzyskanie łatwych w interpretacji rozmytych reguł warunkowych. Użyto technik wydobywania reguł diagnostycznych, których idea zbliżona jest do znanych metod selekcji cech zwanych metodami sekwencyjnymi [24]. Podejście to jest adekwatne, jeżeli tworzy się narzędzie, które w stosunkowo łatwy sposób ma tworzyć proste reguły. Dzięki iteracyjnemu algorytmowi wydobywania możliwe jest prześledzenie i interpretacja każdego z jego kroków. Podejście, które odchodzi od tradycyjnych metod eksploracji danych takich jak sieci neuronowe czy algorytmy genetyczne, pozwala na większą kontrolę procesu wydobywania reguł oraz wspomaganie diagnozy.

Należy podkreślić, że prezentowane algorytmy i przykłady ich działania będą w tym rozdziale dotyczyć zagadnienia wydobywania reguł traktując dane medyczne jako dane uczące. Wynik uczenia czyli wydobyty zbiór reguł diagnostycznych może już być poddany wstępnej ocenie choć o jego ostatecznej jakości będzie decydowała efektywność diagnozowania danych testowych czyli dotąd nieznanymi. Badaniu tej jakości czyli zdolności uogólniania poświęcone są najważniejsze punkty rozdziału 5.

4.1 Wyznaczanie początkowego zbioru rozmytych elementów ogniskowych

Proponowane algorytmy mają na celu znalezienie najlepszego podzbioru reguł diagnostycznych spośród pełnego zbioru reguł, które można utworzyć dla danego problemu medycznego. W tym punkcie opisano pierwszy krok wydobywania reguł diagnostycznych, który jest wspólny dla wszystkich proponowanych metod. Metoda tworzenia początkowych zbiorów zawierających maksymalną liczbę rozmytych elementów ogniskowych została omówio-

na w punkcie 3.2, a przykład utworzenia pełnego zbioru takich elementów ogniskowych został podany w punkcie 3.8, który poświęcono ocenie reguł dla bazy *iris flowers*.

Procedura 2 przedstawia kroki prowadzące do: (i) utworzenia początkowych zbiorów zawierających rozmyte elementy ogniskowe na podstawie danych odnoszących się do wszystkich rozpatrywanych diagnoz, (ii) wyznaczenia rozkładu bazowego prawdopodobieństwa oraz (iii) oceny reguł za pomocą jednej z trzech metod opisanych w punkcie 3.7.

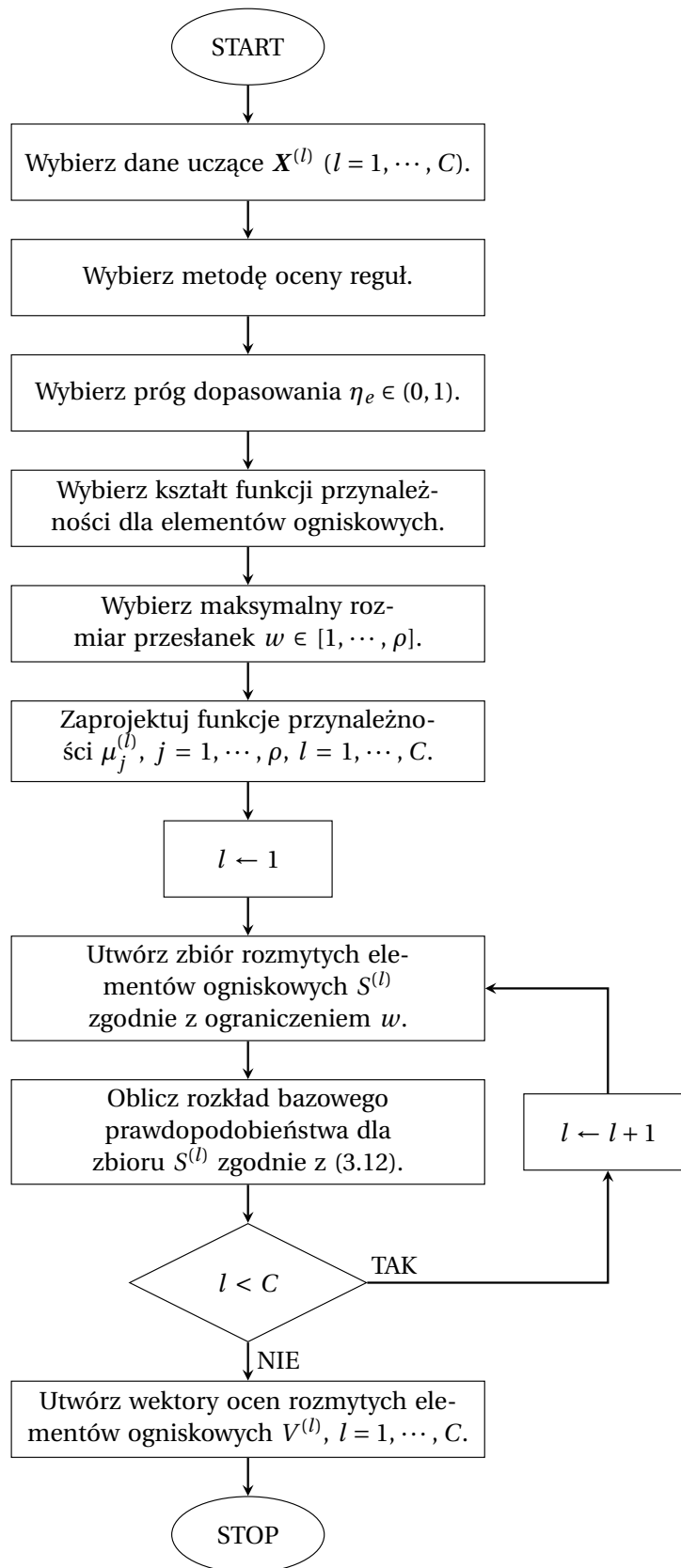
PROCEDURA 2

1. Wybierz:

- (a) dane uczące X , zawierające dane N pacjentów odpowiadających wszystkim rozpatrywanym diagnozom, tj. $X^{(l)}$, $l = 1, \dots, C$.
 - (b) metodę oceny reguł (na podstawie wartości m , MCC lub CC),
 - (c) próg dopasowania $\eta_e \in (0, 1)$ wymagany do oceny reguł,
 - (d) kształt funkcji przynależności dla rozmytych elementów ogniskowych (trójkątne, trapezowe, uogólnione dzwonowe lub gaussowskie łączone),
 - (e) maksymalną liczbę symptomów, które mogą być ujęte w złożonym rozmytym elemencie ogniskowym $w \in [1, \dots, \rho]$.
2. Dokonaj analizy danych uczących X i zaprojektuj funkcje przynależności wybranego kształtu dla wszystkich symptomów i diagnoz, tj. $\mu_j^{(l)}$, $j = 1, \dots, \rho$, $l = 1, \dots, C$.
 3. Utwórz zbiór rozmytych elementów ogniskowych $S^{(l)}$, zawierający maksymalną liczbę elementów ogniskowych na jaką pozwala ustalona w punkcie 1(e) wartość w .
 4. Oblicz rozkład bazowego prawdopodobieństwa dla zbioru $S^{(l)}$ na podstawie $X^{(l)}$ zgodnie z (3.13).
 5. Punkty 3 i 4 powtórz dla wszystkich rozpatrywanych diagnoz $l = 1, \dots, C$.
 6. Utwórz wektory $V^{(l)}$, $l = 1, \dots, C$, zawierające wartości $v_r^{(l)}$, $r = 1, \dots, n^{(l)}$, będące wartością oceny elementów ogniskowych na podstawie danych uczących X zgodnie z wybraną metodą:

$$\bigwedge_{\substack{l=1, \dots, C \\ r=1, \dots, n^{(l)}}} v_r^{(l)} = \begin{cases} m_r^{(l)} & \text{jeżeli wybrano bazowe prawdopodobieństwo,} \\ MCC_r^{(l)} & \text{jeżeli wybrano wskaźnik MCC,} \\ CC_r^{(l)} & \text{jeżeli wybrano wskaźnik CC.} \end{cases} \quad (4.1)$$

Procedurę 2 przedstawiono także za pomocą schematu blokowego na rysunku 4.1. Wynikiem jej działania dla danych uczących jest kompletny zbiór reguł diagnostycznych (ograniczony złożonością rozmytych elementów ogniskowych), które są ocenione za pomocą wartości zawartych w $V^{(l)}$. Na podstawie wartości ocen $V^{(l)}$ możliwe jest wskazanie lepiej lub gorzej dopasowanych reguł. Pozwala to na ujawnienie symptomów, które skutecznie odróżniają jedną diagnozę od innych, tzn. które są dla niej najbardziej charakterystyczne. Niemniej jednak, sama ocena nie pomaga w określeniu, które reguły powinny zostać wybrane aby utworzyć najlepszy możliwy zbiór reguł. Dalsze proponowane algorytmy są podejściami, które próbują w sposób iteracyjny poszukiwać najlepszych zbiorów reguł diagnostycznych dla danego problemu medycznego.



Rysunek 4.1: Schemat blokowy procedury tworzenia początkowych zbiorów elementów ogniskowych

Do wyznaczenia rozkładu bazowego prawdopodobieństwa oraz oceny reguł ustalono, że wymagany poziom dopasowania elementów ogniskowych i danych uczących zdeterninowany będzie przez próg η_e . W rozdziale 3 wprowadzono dwa sposoby obliczania miary przekonania: $Bel^{(l)}(\mathbf{p}^{(*)})$ (3.15) oraz $\widetilde{Bel}^{(l)}(\mathbf{p}^{(*)})$ (3.21). Wybór jednej z nich jest konieczny przed rozpoczęciem wydobywania reguł. Przy wyborze miary przekonania $Bel^{(l)}(\mathbf{p}^{(*)})$, do działania algorytmu wydobywania reguł diagnostycznych konieczne jest wybranie drugiego progu η_b . Jest on konieczny do sprawdzenia poziomu dopasowania danych pacjenta z elementami ogniskowymi podczas obliczenia wartości tej miary. Obliczenie drugiej miary przekonania $\widetilde{Bel}^{(l)}(\mathbf{p}^{(*)})$ nie wymaga zastosowania progu.

4.2 Algorytm eliminacji reguł

Pierwszym algorytmem, który jest modyfikacją metody zaproponowanej w [110], jest algorytm eliminacji reguł. Jak już wspomniano, początkowy zbiór reguł jest pełny (z założonym ograniczeniem liczby warunków w rozmytym elemencie ogniskowym) i utworzony za pomocą procedury 2 (rysunek 4.1). W każdej iteracji algorytmu eliminacji, reguły z najniższą wartością oceny są usuwane, a sam zbiór rozmytych elementów ogniskowych zmniejsza się. Załóżmy, że dla danych \mathbf{X} , w wyniku procedury 2 otrzymano początkowy zbiór reguł diagnostycznych $S^{(l)}$ oceniony za pomocą wybranej metody. Szczegółowy przebieg algorytmu eliminacji przedstawia procedura 3. Algorytm eliminacji reguł jest także przedstawiony za pomocą schematu blokowego na rysunku 4.2. W procedurach 3-5 oraz rysunkach 4.2-4.4 wybierana jest jedna z metod obliczania miary przekonania: $Bel^{(l)}(\mathbf{p}^{(*)})$ (3.15) lub $\widetilde{Bel}^{(l)}(\mathbf{p}^{(*)})$ (3.21). Aby poprawić czytelność procedur i nie wprowadzać nadmiarowego symbolu, zaniechano ich rozróżniania i w poszczególnych krokach zastosowano oznaczenie $BEL^{(l)}(\mathbf{p}^{(*)})$ z zastrzeżeniem, że zawsze dotyczy to tej metody obliczania miary przekonania, którą wybrano w kroku drugim procedur 3-5.

PROCEDURA 3

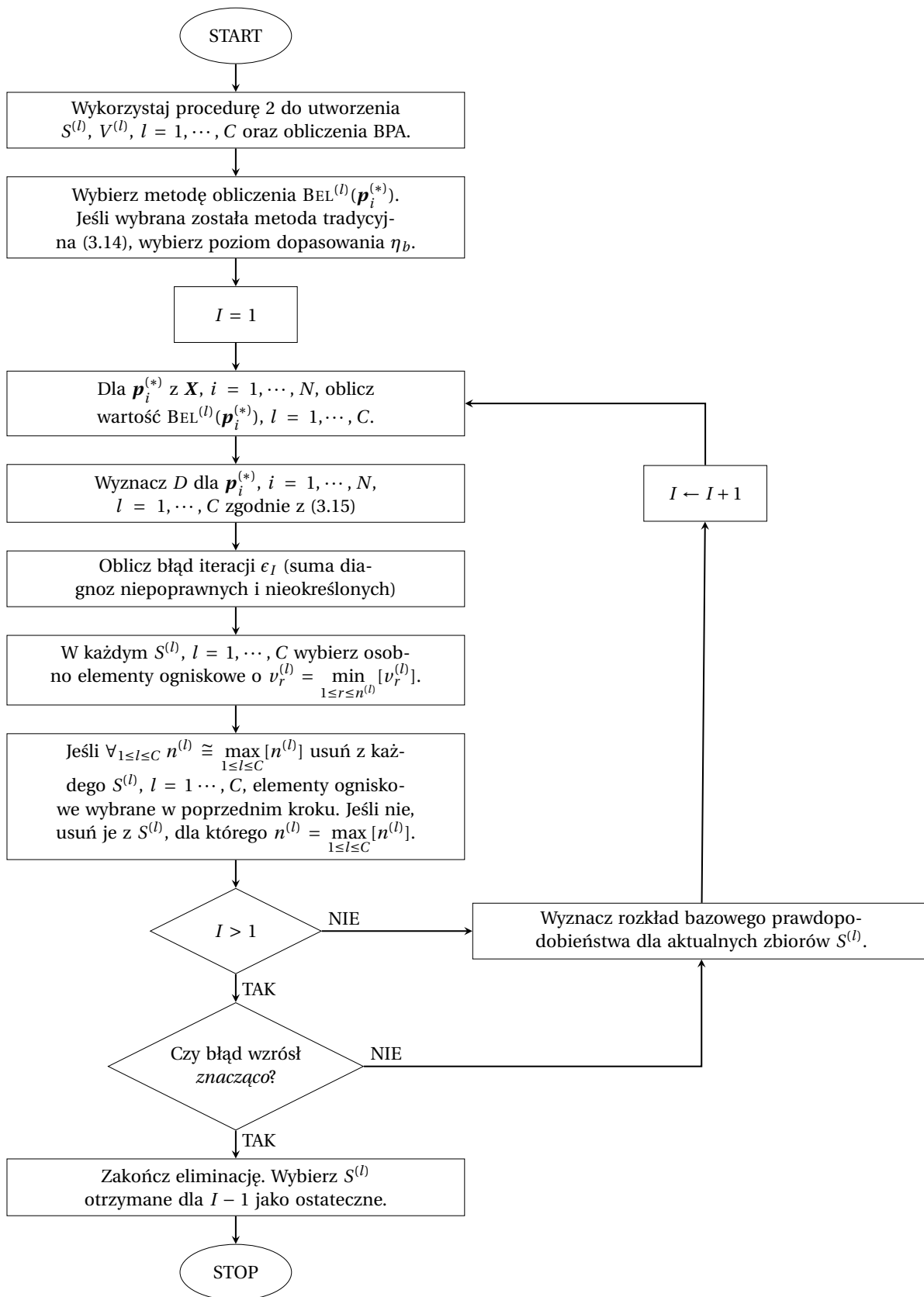
1. Wykonaj procedurę 2: utwórz początkowe zbiory elementów ogniskowych, oceń je i oblicz rozkład bazowego prawdopodobieństwa.
2. Wybierz metodę obliczenia miary przekonania (3.15) lub (3.21). Jeśli wybrana została miara przekonania obliczana tradycyjnie (3.15) to wybierz wymagany próg dopasowania η_b .
3. Dla każdego przypadku ze zbioru danych uczących \mathbf{X} oblicz wartości miar przekonania (sposobem wybranym w kroku drugim) dotyczące każdej rozpatrywanej diagnozy, tj. $BEL^{(l)}(\mathbf{p}_i^{(*)})$, $i = 1, \dots, N$, $l = 1, \dots, C$.
4. Wyznacz diagnozę dla każdego przypadku jako tą, dla której osiągnięto najwyższą wartość miary przekonania zgodnie z (3.16).
5. Oblicz błąd tej iteracji ϵ_l , który jest liczbą przypadków niepoprawnie zdiagnozowanych oraz przypadków, dla których wyznaczenie diagnozy wg (3.16) było niemożliwe.
6. W każdym zbiorze $S^{(l)}$ znajdź osobno elementy ogniskowe z minimalną wartością oceny $v_r^{(l)}$, $r = 1, \dots, C$, $l = 1, \dots, C$.

7. Sprawdź liczebność reguł w zbiorach $S^{(l)}$ dla kolejnych diagnoz $l = 1, \dots, C$. Jeśli jest *zblizona*, usuń z każdego zbioru $S^{(l)}$ elementy ogniskowe znalezione w poprzednim kroku. Jeśli nie, usuń je tylko ze zbioru $S^{(l)}$, w którym reguł jest najwięcej.
8. Jeśli to jest pierwsza iteracja, przejdź do punktu 10. Jeśli nie, przejdź do punktu 9.
9. Jeżeli błąd tej iteracji ϵ_l zwiększył się *znacznie* w stosunku do poprzedniej iteracji, przejdź do punktu 11. Jeśli nie, przejdź do punktu 10.
10. Wyznacz rozkład bazowego prawdopodobieństwa dla zbiorów $S^{(l)}$, w których dokonano zmian i rozpocznij następną iterację algorytmu od punktu 3.
11. Zakończ eliminację. Wybierz zbiory reguł otrzymane w poprzedniej iteracji jako ostateczne $\mathcal{S}^{(l)}$ (dla tej iteracji, dla której błąd diagnozy nie przekroczył warunku zakończenia algorytmu w punkcie 9.).

Dwa określenia użyte w algorytmie opisanym za pomocą procedury 3 są niejednoznaczne i w związku z tym wymagają wyjaśnienia. Pierwszym z nich jest użyte w siódmym kroku stwierdzenie, które dotyczy warunku eliminacji elementów ogniskowych w zbiorach $S^{(l)}$. Wymaganie, aby rozmiary zbiorów elementów ogniskowych $n^{(l)}$ dla wszystkich l były *zblizone*, podyktowane jest tym, że usuwanie elementów ogniskowych bez uwzględnienia proporcji liczebności zbiorów prowadzi szybko do sytuacji, w której liczba elementów ogniskowych w jednym ze zbiorów jest już niewielka, a w innych pozostaje duża. Może to prowadzić do lokalnie najlepszego rozwiązania. Dlatego też założono, że elementy ogniskowe powinny być usuwane ostrożniej, to znaczy tylko gdy ich liczba w zbiorach $S^{(l)}$ jest równa, tzn. $\forall_{1 \leq l \leq C} n^{(l)} = \max_{1 \leq l \leq C} (n^{(l)})$. Oczywiście możliwe jest złagodzenie tego warunku, określając maksymalną różnicę w proporcjach zbiorów zapisując warunek jako $\forall_{1 \leq l \leq C} n^{(l)} = \beta \max_{1 \leq l \leq C} (n^{(l)})$, gdzie β wynosi na przykład 0.8.

Drugim określeniem wymagającym wyjaśnienia jest warunek zakończenia algorytmu opisany w kroku dziewiątym. Zakończenie algorytmu nie jest podyktowane pierwszym wzrostem błędu diagnozy. Jak się okaże, w kolejnych iteracjach błąd może wzrosnąć, aby później dla jeszcze mniejszego zbioru elementów ogniskowych ponownie się obniżyć, co jest pożądaną sytuacją. Dlatego też należy w inny sposób zdecydować, kiedy algorytm powinien być zatrzymany, czyli np. po 50-procentowym wzroście liczby błędów w stosunku do liczby błędów w poprzedniej iteracji.

Powyższy algorytm, nie jest pozbawiony niedoskonałości. Pierwszą rzeczą, która mogłaby zostać poddana krytyce jest oczywiście warunek zakończenia algorytmu. Wybranie krytycznej różnicy błędów (wspomnianego procentowego wzrostu liczby błędów) warunkuje zakończenie algorytmu i będzie ono różne dla odmiennych danych medycznych. W zależności od założonego maksymalnego wzrostu błędu, zmieniać się będzie też wielkość uzyskiwanego zbioru reguł co zostanie przedstawione w rozdziale 5. Tym samym uzyskana zostanie inna wiedza, mniej lub bardziej ogólna. Drugim niuansiem związanym z prowadzeniem eliminacji reguł jest czas trwania tego algorytmu. Zależy on bezpośrednio od początkowej liczby reguł diagnostycznych, która może być bardzo wielka jeśli rozpatruje się rozmyte elementy ogniskowe o dużej złożoności i sama liczba symptomów w danym zagadnieniu medycznym jest znaczna. Tym samym przeprowadzanie iteracyjnej eliminacji reguł z bardzo



Rysunek 4.2: Schemat blokowy zmodyfikowanego algorytmu eliminacji reguł

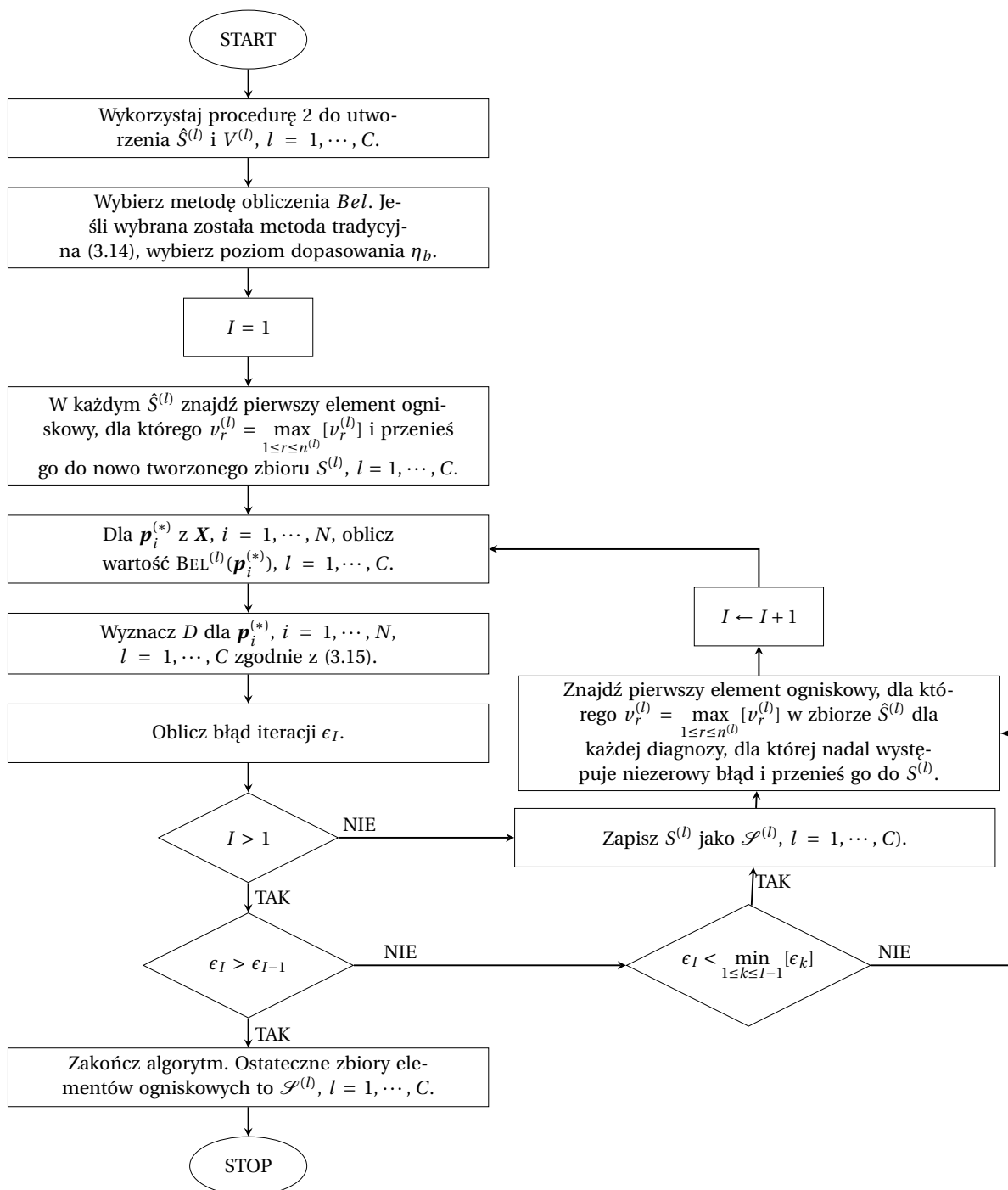
dużego zbioru staje się wyzwaniem pod względem złożoności obliczeniowej. Jakość algorytmu eliminacji reguł zostanie zbadana z użyciem danych medycznych w rozdziale 5. Można będzie wtedy stwierdzić, na ile te niedostatki algorytmu upośledzają jego działanie.

4.3 Algorytmy selekcji reguł

Odwrócenie kierunku algorytmu jest pierwszym krokiem, który wynika z krytycznej oceny przedstawionego w poprzednim punkcie algorytmu eliminacji reguł. Zaproponowany algorytm selekcji reguł, w odróżnieniu do algorytmu eliminacji, wybiera z początkowego zbioru reguł diagnostycznych (otrzymanego za pomocą procedury 2) pojedyncze najlepsze reguły i przenosi je do nowo tworzonego zbioru reguł. Procedura 4 przedstawia działanie pierwszego algorytmu selekcji reguł. Jego działanie jest także przedstawione na schemacie blokowym na rysunku 4.3.

PROCEDURA 4

1. Wykonaj procedurę 2: utwórz początkowe zbiory elementów ogniskowych, oznacz je jako $\hat{S}^{(l)}$ $l = 1, \dots, C$ i oceń je.
2. Wybierz metodę obliczenia miary przekonania (3.15) lub (3.21). Jeśli wybrana została miara przekonania obliczana tradycyjnie (3.15) to wybierz wymagany próg dopasowania η_b .
3. W każdym zbiorze $\hat{S}^{(l)}$ znajdź pierwszy element ogniskowy z maksymalną wartością oceny $v_r^{(l)}$ i przenieś go do nowo tworzonego zbioru $S^{(l)}$, $l = 1, \dots, C$.
4. Wyznacz rozkład bazowego prawdopodobieństwa zgodnie z (3.13) dla aktualnych zbiorów $S^{(l)}$, $l = 1, \dots, C$.
5. Dla każdego przypadku ze zbioru danych uczących X oblicz wartości miar przekonania (sposobem wybranym w kroku drugim) dotyczące każdej rozpatrywanej diagnozy, tj. $BEL^{(l)}(\mathbf{p}_i^{(l)})$, $i = 1, \dots, N$, $l = 1, \dots, C$.
6. Wyznacz diagnozę dla każdego przypadku jako tą, dla której osiągnięto najwyższą wartość miary przekonania zgodnie z (3.16).
7. Oblicz błąd tej iteracji ϵ_l , który jest liczbą przypadków niepoprawnie zdiagnozowanych oraz przypadków, dla których wyznaczenie diagnozy wg (3.16) było niemożliwe.
8. Jeżeli to jest pierwsza iteracja, przejdź do kroku 10. Jeżeli nie, przejdź do kroku 9.
9. Jeżeli:
 - (a) osiągnięto (jak dotąd) najniższy błąd, przejdź do kroku 10.
 - (b) błąd się nie zmienił, przejdź do kroku 11.
 - (c) błąd wzrósł, przejdź do kroku 12.
10. Zapisz aktualne zbiory $S^{(l)}$ jako najlepsze tzn. $\mathcal{S}^{(l)} = S^{(l)}$, $l = 1, \dots, C$.
11. Znajdź pierwszy element ogniskowy o maksymalnej wartości oceny w zbiorze $\hat{S}^{(l)}$ dla każdej diagnozy, dla której nadal występuje niezerowy błąd i przenieś go do tworzonego zbioru $S^{(l)}$. Rozpocznij kolejną iterację algorytmu od kroku 4.
12. Zakończ algorytm. Ostateczne zbiory elementów ogniskowych to $\mathcal{S}^{(l)}$, $l = 1, \dots, C$.



Rysunek 4.3: Schemat blokowy algorytmu selekcji reguł

Zauważmy, że zmiana kierunku algorytmu spowodowała jego uproszczenie, jednocześnie zyskał on na jednoznaczności. Mając na uwadze kompromis pomiędzy efektywnością a złożonością zbioru reguł, warunek zakończenia algorytmu podyktowany jest po prostu wzrostem liczby błędów. Tym samym nie jest sprawdzane czy wzrost liczby błędów jest znaczący i nie jest konieczne określanie tego kryterium z procedury 3. Dodatkowo reguły dla l -tej diagnozy są dodawane tylko do momentu kiedy w ramach l -tej diagnozy występuje niezerowy błąd.

Mimo, że przedstawiony w tym punkcie algorytm znacznie skróci proces wydobywania reguł diagnostycznych to uwagę zwraca punkt 9(b) prowadzący do punktu 11. Naturalnie, dodawanie kolejnej reguły dla zbioru dotyczącego l -tej diagnozy, jeżeli nie występuje błąd dla przypadków tej diagnozy, jest niepotrzebne. Niemniej jednak, rzadko kiedy osiąga się stuprocentową poprawność rozpoznawania przypadków diagnozy. Tym samym wydaje się, że warunek dodania reguły opisany w kroku 11. jest zbyt łagodny i rzadko może być spełniony, dlatego też zaproponowano jeszcze jeden algorytm selekcji reguł tym razem z warunkowym krokiem usuwania reguł. Pełne działanie algorytmu z warunkowym usuwaniem reguł przedstawia procedura 5.

PROCEDURA 5

1. Wykonaj procedurę 2: utwórz początkowe zbiory elementów ogniskowych, oznacz je jako $\hat{S}^{(l)}$ $l = 1, \dots, C$ i oceń je.
2. Wybierz metodę obliczenia miary przekonania (3.15) lub (3.21). Jeśli wybrana została miara przekonania obliczana tradycyjnie (3.15) to wybierz wymagany próg dopasowania η_b .
3. W każdym zbiorze $\hat{S}^{(l)}$ znajdź pierwszy element ogniskowy z maksymalną wartością oceny $\nu_r^{(l)}$ i przenieś go do nowo tworzonego zbioru $S^{(l)}$, $l = 1, \dots, C$.
4. Wyznacz rozkład bazowego prawdopodobieństwa zgodnie z (3.13) dla aktualnych zbiorów $S^{(l)}$, $l = 1, \dots, C$.
5. Dla każdego przypadku ze zbioru danych uczących X oblicz wartości miar przekonania (sposobem wybranym w kroku drugim) dotyczące każdej rozpatrywanej diagnozy, tj. $BEL^{(l)}(\mathbf{p}_i^{(l)})$, $i = 1, \dots, N$, $l = 1, \dots, C$.
6. Wyznacz diagnozę dla każdego przypadku jako tą, dla której osiągnięto najwyższą wartość wybranej miary przekonania zgodnie z (3.16).
7. Oblicz błąd tej iteracji ϵ_l , który jest liczbą przypadków niepoprawnie zdiagnozowanych oraz przypadków, dla których wyznaczenie diagnozy było niemożliwe (3.16).
8. Jeżeli to jest pierwsza iteracja, przejdź do kroku 12. Jeżeli nie, przejdź do kroku 9.
9. Usuń regułę dla l -tej diagnozy, która została dodana w tej iteracji, jeśli nie zmniejszyła się liczba błędnie zdiagnozowanych przypadków l -tej diagnozy, $l = 1, \dots, C$.
10. Jeżeli:
 - (a) liczba błędów się nie zmieniła, przejdź do kroku 11.
 - (b) osiągnięto (jak dotąd) najmniejszą liczbę błędów, przejdź do kroku 12.
 - (c) liczba błędów wzrosła, przejdź do kroku 14.
11. Jeżeli liczba reguł oraz błędów przestała się zmieniać przejdź do kroku 14. Jeśli nie, przejdź do kroku 13.

12. Zapisz aktualne zbiory $S^{(l)}$ jako najlepsze tzn. $\mathcal{S}^{(l)} = S^{(l)}$, $l = 1, \dots, C$.
13. Znajdź pierwszy element ogniskowy o maksymalnej wartości oceny w zbiorze $\hat{S}^{(l)}$ dla każdej diagnozy dla której, nadal występuje niezerowy błąd i przenieś go do tworzonego zbioru $S^{(l)}$. Przejdź do kroku 5.
14. Zakończ algorytm. Ostateczny zbiór elementów ogniskowych to $\mathcal{S}^{(l)}$.

Rysunek 4.4 przedstawia schemat proponowanego algorytmu. Główną różnicą między tym algorytmem a procedurą 4 jest wprowadzenie dodatkowego warunku jako punkt dziewiąty. Mianowicie, sprawdza się czy każda z dodanych w aktualnej iteracji reguł spowodowała zmniejszenie błędu diagnozy. Jeśli ten warunek nie jest spełniony, dodany element ogniskowy jest usuwany z $S^{(l)}$. Może się zdarzyć, że żadna z dodawanych reguł nie będzie poprawiała jakości rozpoznawania i wszystkie dodawane reguły będą natychmiast usuwane. Spowoduje to, że przestanie się zmieniać liczba elementów ogniskowych w zbiorach reguł oraz liczba błędnych diagnoz. Dlatego też ustalono, że taką sytuację należy potraktować jako kolejny warunek zatrzymania algorytmu (krok 11.). Zakończenie algorytmu może nastąpić np. jeśli przez dwie kolejne iteracje liczby błędów i reguł nie zmieniają się.

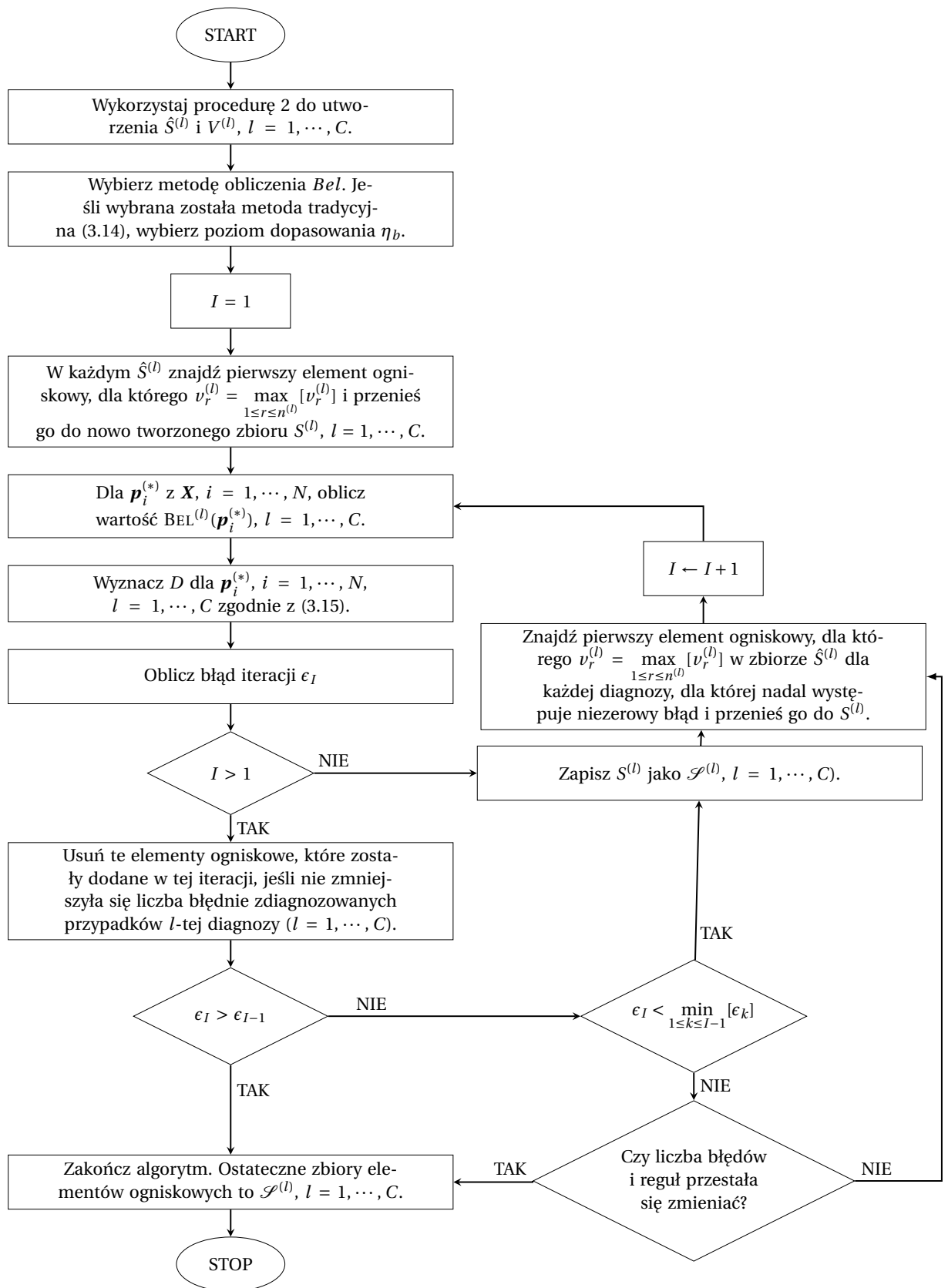
Aby prześledzić proces ekstrakcji reguł zaproponowanymi metodami, jako przykład zostanie użyta ponownie baza *iris flowers*. Porównanie działania algorytmów z użyciem tej bazy będzie przedstawione w punkcie 4.5. Przed tym jednak konieczne jest przedstawienie sposobu oceny wyniku działania algorytmów wydobywania reguł diagnostycznych. Od tej pory nazwa „Algorytm Eliminacji Reguł” będzie stosowana zamiennie ze skrótem „AER”. Skróty „ASR” oraz „AWR” będą z kolei odpowiadać nazwom: „Algorytm Selekcji Reguł” oraz „Algorytm Warunkowej selekcji Reguł”.

4.4 Metody oceny działania algorytmów wydobywania reguł diagnostycznych

Po spełnieniu warunku zakończenia algorytmów otrzymujemy wyodrębnione zbiory reguł diagnostycznych $\mathcal{S}^{(l)}$, $l = 1, \dots, C$, które charakteryzuje liczba wydobytych reguł oraz m.in. złożoność rozmytych elementów ogniskowych a przede wszystkim osiągnięta efektywność.

4.4.1 Efektywność wspomagania diagnozy

Efektywność jako stosunek liczby poprawnie zdiagnozowanych pacjentów do ich całkowitej liczby w zbiorze można obliczyć porównując diagnozę wyznaczoną na podstawie miary przekonania z diagnozą, do jakiej przypisany jest pacjent w zbiorze uczącym. Diagnozę przypisaną i -temu pacjentowi w bazie oznaczmy jako d_i . Przyjmuje ona wartości od 1 do C . Diagnozę wyznaczoną na podstawie miary przekonania oznaczmy jako δ_i . Przyjmuje ona



Rysunek 4.4: Schemat blokowy algorytmu warunkowej selekcji reguł

wartości od 1 do C oraz 0, gdzie 0 oznacza, że diagnoza nie jest podjęta. Stosując powyższe oznaczenia, efektywność diagnozy można obliczyć jako

$$E = \frac{1}{N} \sum_{i=1}^N I(d_i = \delta_i) \cdot 100\%, \quad (4.2)$$

gdzie $I(d_i = \delta_i)$ jest funkcją wskaźnikową przyjmującą wartość 1 dla poprawnej diagnozy oraz 0 dla niepoprawnej diagnozy. Dodatkowo, oznaczmy jako $E^{(l)}$ efektywność rozpoznawania przypadków l -tej diagnozy, obliczaną jako

$$E^{(l)} = \frac{1}{n_l} \sum_{i=1}^{n_l} I(d_i = \delta_i) \cdot 100\%. \quad (4.3)$$

Z uwagi na to, że wyznaczanie diagnozy można traktować jako zagadnienie klasyfikacji, można stosować podobne metody jej oceny. Szczególnie w obszarze medycznym sprawdza się tzw. macierz pomyłek klasyfikacji dwuklasowej, tzn. takiej, w której mamy do czynienia z dwoma diagnozami np. „pacjent zdrowy”, „pacjent chory”. Macierz pomyłek przedstawiono w tabeli 4.1. W przypadku kiedy mamy do czynienia z kilkoma diagnozami, oblicza się macierz

Tabela 4.1: Macierz pomyłek klasyfikacji dwuklasowej

	rozpoznany jako	
	chory	zdrowy
pacjent chory	TP	FN
pacjent zdrowy	FP	TN

pomyłek dla każdej diagnozy traktując pozostałe łącznie jako diagnoza l' ($l' \neq l$) [55, 97]. Do przedstawienia ogólnej macierzy pomyłek można skorzystać z oznaczeń występujących w funkcji wskaźnikowej w (4.2). Postać macierzy pomyłek z uwagi na l -tą diagnozę przedstawiona jest w tabeli 4.2. Interpretacja wartości w tabeli 4.2 staje się tożsama z wyjaśnieniem

Tabela 4.2: Macierz pomyłek dla l -tej diagnozy

		decyzja systemu	
		$\delta_i = l$	$\delta_i = l'$
prawdziwa diagnoza	$d_i = l$	$TP^{(l)}$	$FN^{(l)}$
	$d_i = l'$	$FP^{(l)}$	$TN^{(l)}$

przedstawionym w punkcie 3.7.2 a korzystając ze funkcji wskaźnikowej użytej w (4.2) oblicza się w następujący sposób:

$$TP^{(l)} = \sum_{i=1}^{n_l} I(d_i = \delta_i), \quad (4.4)$$

$$FN^{(l)} = n_l - TP^{(l)}, \quad (4.5)$$

$$TN^{(l)} = \sum_{i=1}^{n_l} I(d_i = \delta_i), \quad (4.6)$$

$$FP^{(l)} = n_l - TN^{(l)}. \quad (4.7)$$

Na podstawie (4.4)-(4.7) możliwe jest obliczenie tzw. miar pochodnych, które dają dodatkową informację o jakości wspomagania diagnozy. A są to czułość ($SE^{(l)}$ od ang. *Sensitivity*), specyficzność ($SP^{(l)}$ od ang. *Specificity*), pozytywna wartość prognostyczna ($PPV^{(l)}$ od ang. *Positive Prognostic Value*) oraz negatywna wartość prognostyczna ($NPV^{(l)}$ od ang. *Negative Prognostic Value*):

$$SE^{(l)} = \frac{TP^{(l)}}{TP^{(l)} + FN^{(l)}}, \quad (4.8)$$

$$SP^{(l)} = \frac{TN^{(l)}}{TN^{(l)} + FP^{(l)}}, \quad (4.9)$$

$$PPV^{(l)} = \frac{TP^{(l)}}{TP^{(l)} + FP^{(l)}}, \quad (4.10)$$

$$NPV^{(l)} = \frac{TN^{(l)}}{TN^{(l)} + FN^{(l)}}. \quad (4.11)$$

Można zauważyć, dodatkowo, że wartości $E^{(l)}$ oraz $SE^{(l)}$ odpowiadają sobie, bo:

$$E^{(l)} = SE^{(l)} \cdot 100\%. \quad (4.12)$$

Macierz pomyłek przedstawiona w tabeli 4.2 oraz miary pochodne (4.8-4.12) będą użyte przy badaniach związanych z rzeczywistą bazą medyczną. Dla danych benchmarkowych uwaga zostanie skupiona na badaniu efektywności uogólniania oraz liczebności zbiorów reguł diagnostycznych. Prezentacja większej liczby wskaźników jakości spowodowałoby nadmierny rozrost zamieszczonych wyników ze szkodą dla ich przejrzystej interpretacji.

4.4.2 Złożoność zbioru reguł diagnostycznych

Ocena zbioru reguł $\mathcal{S}^{(l)}$ pod względem efektywności jest jednoznaczna. Efektywność jest najważniejszym parametrem charakteryzującym algorytm wspomagania diagnozy. Jednakże, kolejnym parametrem jest złożoność zbioru reguł diagnostycznych. Można by sądzić, że im mniejszy zbiór reguł tym lepiej. Niestety wysoka efektywność wspomagania diagnozy i mała złożoność reguł są zwykle sprzecznymi wymaganiami. Znaczący to tyle, że aby uzyskać wiedzę bardziej szczegółową, która dokładniej opisze pacjentów ze zbioru uczącego, jest konieczne zwiększenie liczby reguł. W przeprowadzonych eksperymentach złożoność zbioru reguł ocenia się na podstawie liczby reguł w l -tej diagnozie, czyli $n^{(l)}$ oraz ogólnej liczby reguł

obliczanej jako

$$L = \sum_{l=1}^C n^{(l)}. \quad (4.13)$$

Dodatkowo, możliwa jest analiza liczebności rozmytych elementów ogniskowych złożonych z tej samej liczby warunków logicznych. Zdefiniujmy wektor Θ jako

$$\Theta = [\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(j)}, \dots, \theta^{(w)}], \quad (4.14)$$

gdzie w oznacza wybraną maksymalną liczbę warunków logicznych w elemencie ogniskowym a $\theta^{(j)}$ jest liczbą rozmytych elementów ogniskowych zawierających j warunków logicznych ($j = 1, \dots, w$) obliczaną jako

$$\theta^{(j)} = \sum_{l=1}^C \left(\sum_{r=1}^{n^{(l)}} I(\text{Card}(\mathbb{J}_r^{(l)}) = j) \right), \quad (4.15)$$

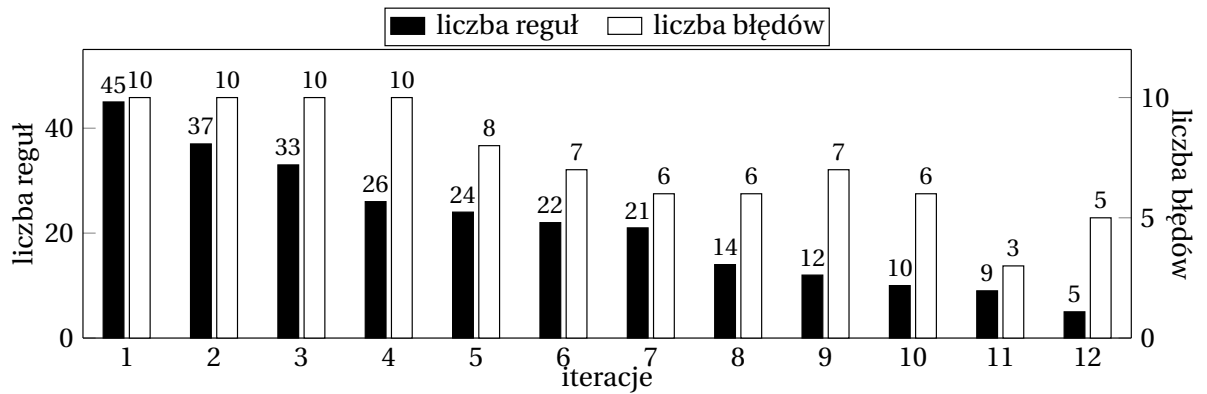
gdzie $\text{Card}()$ jest mocą zbioru, $I()$ jest funkcją wskaźnikową przyjmującą wartość 1 kiedy rozmyty element ogniskowy o określonej złożoności występuje w zbiorach reguł a 0 w przeciwnym przypadku. Przykładowo, jeżeli zostanie wydobyte 10 elementów ogniskowych (dla wszystkich diagnoz) i sześć z nich będzie prostymi rozmytymi elementami ogniskowymi, a cztery będą elementami ogniskowymi, z których każdy dotyczy dwóch symptomów, to $\theta^{(1)} = 6$, $\theta^{(2)} = 4$ i $\forall_{2 < j \leq w} \theta^{(j)} = 0$ a tym samym $\Theta = [6, 4, 0, \dots, 0]$. Taki zapis pozwoli sprawniej porównać wyniki otrzymane w przeprowadzonych eksperymentach.

4.5 Przykład wydobywania reguł dla bazy *iris flowers*

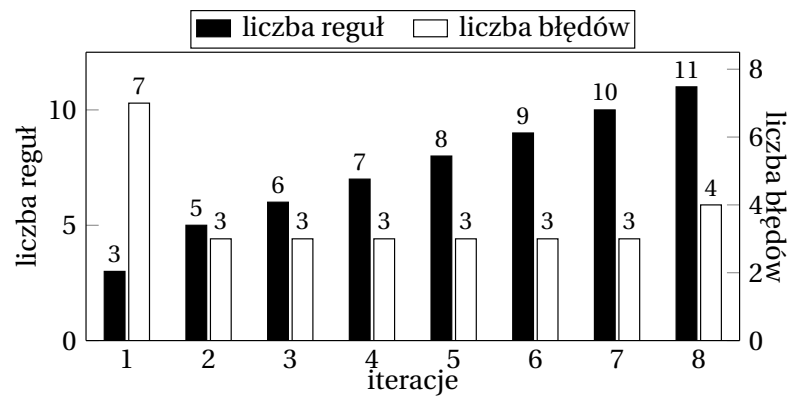
Obecnie zostanie przedstawiony przykład działania trzech algorytmów wydobywania reguł. Do ilustracji zostanie użyta baza *iris flowers*. Celem tego przykładu jest pokazanie różnic działania algorytmów pod względem liczby wydobywanych reguł i wyznaczenia błędnych diagnoz w kolejnych iteracjach. Dla uproszczenia założono, że we wszystkich algorytmach zostaną użyte: trójkątne funkcje przynależności w elementach ogniskowych, CC (3.42) jako metoda oceny reguł oraz tradycyjna miara przekonania (3.15). Wpływ tych wielkości na jakość wspomagania diagnozy będzie przedmiotem osobnych badań. Na początku obliczeń konieczne jest określenie progu η_e do wyznaczenia rozkładu bazowego prawdopodobieństwa oraz progu do obliczania miary przekonania (3.15) czyli η_b . Dla przedstawionego przykładu wybrano wartość 0.45 dla obu progów dla wszystkich trzech algorytmów. Przyczyna takiego wyboru progów zostanie wyjaśniona w dalszej części rozdziału.

Na rysunku 4.5-4.7 przedstawiono zaobserwowane zmiany liczby błędów oraz reguł dla kolejnych iteracji algorytmów przedstawionych na rysunkach 4.2-4.4.

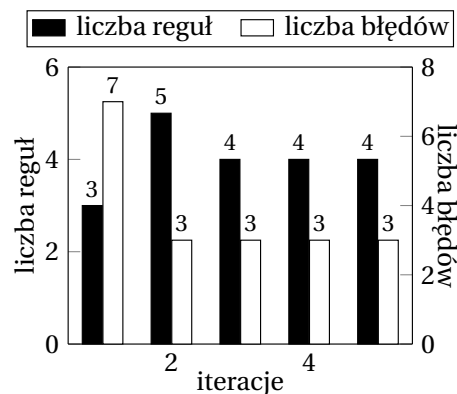
Rysunek 4.5 podsumowuje działanie algorytmu eliminacji (AER). Analizując ten rysunek można zauważyć, że zmniejszanie liczby reguł skutkuje pomniejszeniem liczby błędnych klasyfikacji. Do ósmej iteracji liczba reguł i liczba błędów maleją. W dziewiątej iteracji



Rysunek 4.5: Zmiana liczby reguł i błędów podczas działania algorytmu eliminacji reguł (AER) dla bazy *iris flowers*



Rysunek 4.6: Zmiana liczby reguł i błędów podczas działania algorytmu selekcji reguł (ASR) dla bazy *iris flowers*



Rysunek 4.7: Zmiana liczby reguł i błędów podczas działania algorytmu warunkowej selekcji reguł (AWR) dla bazy *iris flowers*

liczba błędnych klasyfikacji zwiększa się o jedną, a eliminacja reguł nie zostaje przerwana. Jak widać, ma to pożądaný skutek, ponieważ w następnej iteracji liczba błędów ponownie zostaje zmniejszona a dodatkowo udało się także pomniejszyć liczbę reguł. Najlepszy wynik pod względem liczby błędów osiągnięto w iteracji jedenastej. Tylko trzy błędy wskazano przy udziale dziewięciu elementów ogniskowych dla wszystkich rozpatrywanych klas. Naturalnie, eliminacja reguł nie będzie w nieskończoność zmniejszać liczby błędnych klasyfikacji. Kolejne pomniejszenie liczebności zbiorów elementów ogniskowych (iteracja dwunasta) powoduje, że usuwane zostają reguły, które pozostawały dotąd w zbiorze ze względu na ich najwyższe wartości oceny. Były to zatem reguły najlepiej dopasowane do danych uczących i niezbędne do osiągnięcia wysokiej efektywności klasyfikacji. Dlatego eliminacja przeprowadzona w iteracji dwunastej powoduje zwiększenie liczby błędów o ponad 50%. Taki wzrost błędów spełnia założony warunek zatrzymania algorytmu eliminacji reguł. Ostateczny zbiór wydobytych reguł $\mathcal{S}^{(l)}$, $l = 1, \dots, C$ stanowi zbiór otrzymany w iteracji jedenastej.

Pierwszą różnicą w algorytmie selekcji reguł (ASR – rysunek 4.6) w stosunku do algorytmu eliminacji (AER – rysunek 4.5) jest liczba jego iteracji. Tym razem zamiast usuwania elementów ogniskowych o najmniejszej wartości oceny (wybranej jako wartość bazowego prawdopodobieństwa), dodawane są elementy ogniskowe ocenione najlepiej. Po pierwszej iteracji ASR w każdym zbiorze reguł występuje tylko jeden rozmyty element ogniskowy. Jak widać, trzy najlepsze elementy ogniskowe powodują, że liczba popełnionych pomyłek przez algorytm jest już stosunkowo niewielka bo wynosi siedem błędnych klasyfikacji. W drugiej iteracji ASR zbiór reguł liczy pięć elementów ogniskowych. Dodane zostały już tylko dwa elementy ogniskowe a to oznacza, że jedna z klas bazy *iris flowers* jest już bezbłędnie rozpoznana. Tym samym warunek dotyczący dodawania elementów ogniskowych opisany w 11. kroku procedury 4 (patrz punkt 4.3) wydaje się użyteczny. Późniejsze iteracje związane są z dodawaniem tylko jednej reguły. To znaczy, że kolejna z klas bazy *iris flowers* jest bezbłędnie rozpoznana i nie ma potrzeby dodania do niej kolejnej reguły. Niemniej jednak, dodawanie kolejnych reguł dla pozostałej klasy nie zmniejsza błędów (od iteracji trzeciej do siódmej) dlatego nadal zbiór elementów ogniskowych osiągnięty w drugiej iteracji jest uznawany jako najlepszy. Kolejne dodawane reguły mają już co raz mniejszą wartość CC . W ósmej iteracji dodanie kolejnej reguły pogarsza jakość klasyfikacji co skutkuje zatrzymaniem algorytmu. Ostatecznym zbiorem reguł diagnostycznych $\mathcal{S}^{(l)}$, $l = 1, \dots, C$ jest zbiór uzyskany w iteracji drugiej, gdyż jest to najmniej liczny zbiór rozmytych elementów ogniskowych z najmniejszą liczbą błędów spośród wszystkich iteracji algorytmu.

Działanie algorytmu selekcji reguł z ich warunkowym usuwaniem (AWR) jest nieco bardziej złożone. Na rysunku 4.7 widać, że zmiana liczby elementów ogniskowych nie jest monotoniczna, jak to miało miejsce w algorytmie (ASR – rysunek 4.6). Jest to spowodowane spełnieniem warunku opisanego w dziewiątym kroku procedury 5 (punkt 4.3), który pozwala usunąć ostatnio dodany element ogniskowy ze zbioru reguł $S^{(l)}$, jeśli nie poprawił on klasyfikacji obiektów l -tej klasy. Pierwsze dwie iteracje AWR są tożsame z działaniem ASR przedstawionym na rysunku 4.6. W trzeciej iteracji nastąpiło usunięcie jednej z reguł, która w iteracji drugiej nie wpłynęła na jakość klasyfikacji odpowiedniej klasy. W wyniku takiego wyboru warunków działania algorytmu, od trzeciej iteracji elementy ogniskowe w zbiorach

reguł przestają się zmieniać. Jest to spowodowane tym, że kolejne dodawane warunkowo reguły są usuwane ze względu na to, że nie wpływają pozytywnie na poprawę jakości klasyfikacji. Taka sytuacja może jednak powodować niepotrzebne przedłużenie działania algorytmu selekcji. Dlatego, kiedy liczba reguł i liczba błędów nie zmieniają się, warunek zatrzymania algorytmu, opisany jako krok jedenasty procedury 5 (punkt 4.3), zostaje spełniony. Ostateczny wynik algorytmu selekcji z warunkowym usuwaniem reguł to cztery reguły i trzy błędne klasyfikacje.

W tabeli 4.3 zostały przedstawione rozmyte elementy ogniskowe uzyskane w przykładach opisanych powyżej wraz z wartościami bazowego prawdopodobieństwa dla każdego zbioru. Oznaczenia rozmytych elementów ogniskowych wprowadzono w punkcie 3.8 a ich numery odpowiadają oryginalnym numerom w początkowych zbiorach elementów ogniskowych poddanych ocenie.

Tabela 4.3: Zbiory elementów ogniskowych $\mathcal{S}^{(l)}$, $l = 1, \dots, 3$ uzyskane dla bazy *iris flowers* dzięki algorytmom: AER, ASR, AWR. Poniżej oznaczeń rozmytych elementów ogniskowych zapisano ich wartość bazowego prawdopodobieństwa.

algorytm	$\mathcal{S}^{(1)}$	$\mathcal{S}^{(2)}$	$\mathcal{S}^{(3)}$
AER	$\{s^{(3)}, s^{(4)}, s^{(10)}\}$ {0.34, 0.33, 0.33}	$\{s^{(3)}, s^{(4)}, s^{(10)}\}$ {0.34, 0.33, 0.33}	$\{s^{(3)}, s^{(4)}, s^{(10)}\}$ {0.35, 0.33, 0.32}
ASR	$\{s^{(3)}\}$ {1}	$\{s^{(10)}, s^{(4)}\}$ {0.495, 0.505}	$\{s^{(4)}, s^{(3)}\}$ {0.495, 0.505}
AWR	$\{s^{(3)}\}$ {1}	$\{s^{(10)}\}$ {1}	$\{s^{(4)}, s^{(3)}\}$ {0.495, 0.505}

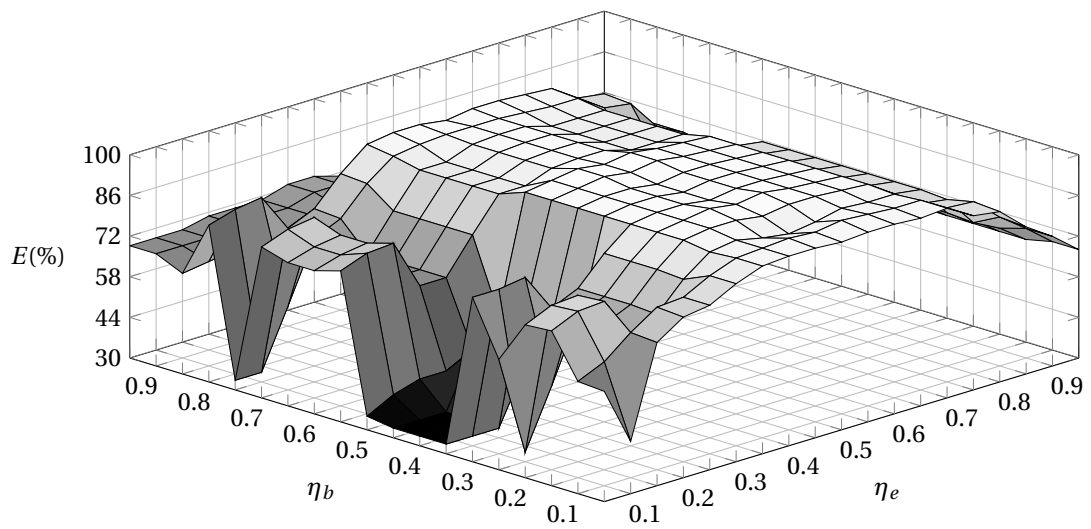
Nie powinno zaskakiwać, że dla trywialnego przykładu z bazą *iris flowers*, działanie algorytmów daje zbliżony wynik. Niemniej jednak, już dla tej bazy pewne zaproponowane udoskonalenia znajdują swoje odzwierciedlenie w wynikach. Jeżeli chodzi o wydobyte zbiory reguł przedstawione w tabeli 4.3, główną różnicą jest to, że dla algorytmu eliminacji (AER) niemożliwe było pozostawienie jednego elementu ogniskowego dla klasy pierwszej, choć, jak się okazuje, $s^{(3)}$ jest elementem ogniskowym wystarczającym do odróżnienia klasy *iris setosa* od pozostałych. Wybór tego elementu zapewniły natomiast algorytmy selekcji. Wszystkie algorytmy skupiły się na zmiennych trzeciej i czwartej, ale to algorytm warunkowej selekcji (AWR) osiąga tę samą efektywność co poprzednie przy wydobyciu najmniejszej liczby reguł.

4.6 Ocena zbioru reguł diagnostycznych

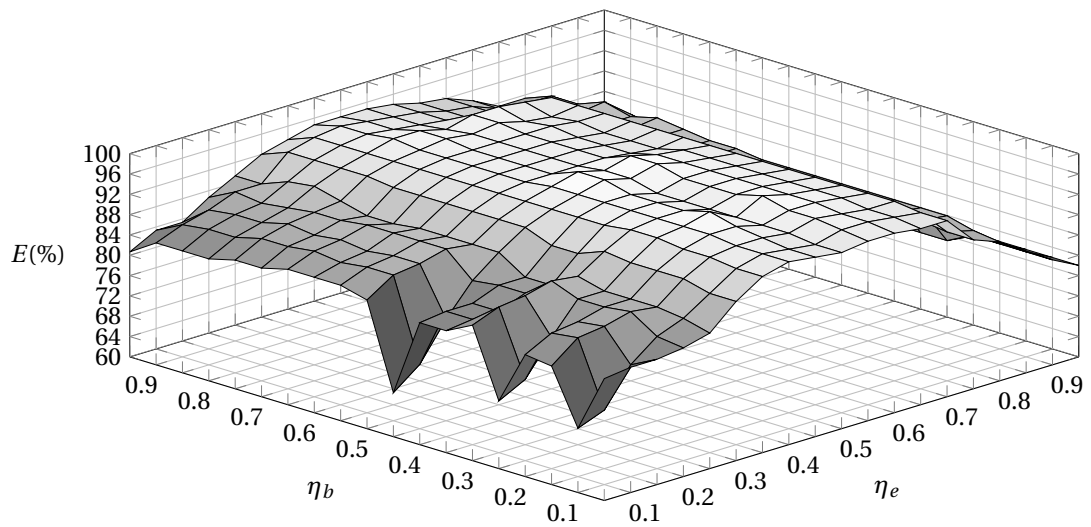
W poprzednim punkcie przedstawiono przykładowe wyniki otrzymane dla trzech zaproponowanych algorytmów wyodrębniania reguł. Do każdego algorytmu użyto tych samych danych, a także zastosowano tę samą metodę oceny i jednakowe kształty funkcji przynależności. Wybór wartości progów ($\eta_e = \eta_b = 0.45$) nie jest arbitralny. Na rysunku 4.8 przedstawiono jak zmienia się efektywność wyznaczania diagnoz (4.2) na podstawie reguł wydobytych przez trzy algorytmy w zależności od zmian progów η_e i η_b w zakresie od 0.05 do 0.95

z krokiem 0.05. Zmniejszenie wymaganego poziomu dopasowania do zera spowodowałoby, że wszystkie elementy ogniskowe byłyby uznane za dopasowane i byłyby włączone do obliczenia miary przekonania. Odpowiadałoby to nierozsądnemu założeniu, że wszystkie obserwowane u pacjenta objawy, nawet te, dla których dopasowanie do elementów ogniskowych jest zerowe, są uwzględnione podczas obliczenia $Bel^{(l)}$ (3.15). Z drugiej strony, przy wartości progów równej 1, dopasowanie wykazywałyby tylko symptomy o pełnej przynależności do elementu ogniskowego, co dla zbiorów opisanych funkcją trójkątną lub gaussowską łączoną dotyczy tylko jednej wartości. Stąd wynika, że strojenie progów η_e, η_b powinno być ograniczone do przedziału $(0,1)$.

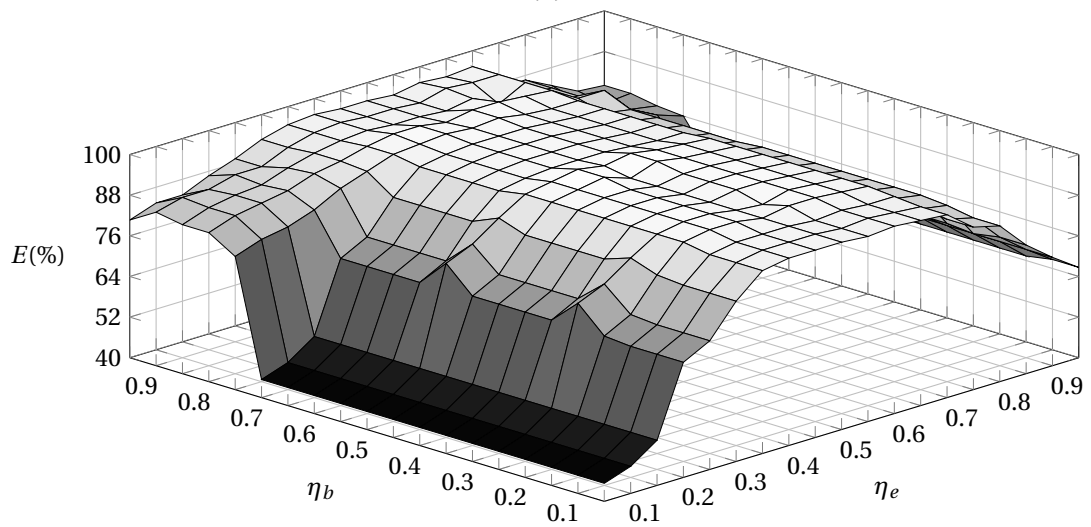
Na podstawie rysunków 4.8a-4.8c należy stwierdzić, że wybór progów ma duże znaczenie dla procesu wydobywania wiedzy z danych w postaci reguł. Spośród trzech algorytmów najmniejsze wahania wartości E możemy dostrzec dla ASR (rysunek 4.8b). Uwidacznia to skala trzeciej współrzędnej wykresu. Niemniej jednak, maksymalne wartości efektywności są osiągnięte dla średnich wartości wskaźników (kiedy są bliskie wartości 0.5) a dla małych wartości obu wskaźników (poniżej 0.3) efektywność znacznie maleje. Łatwiej dostrzec to na rysunkach 4.8a oraz 4.8c, gdzie wartości efektywności we wspomnianym zakresie znacznie spadają. Dodatkowo, czego nie zaprezentowano, każda z metod oceny reguł wpływa na osiąganą efektywność w zależności od przyjętych wartości progów. W związku z tym wartości progów wybrano tak, aby liczba błędów uzyskiwana trzema algorytmami była najniższa (dla każdej pary progów wzięto liczby błędów pochodzące od trzech algorytmów, uśredniono je i wybrano parę η_e i η_b związaną z najniższą średnią liczbą błędów). Pojawia się pytanie, którą parę progów wybrać dla danego problemu medycznego, tj. dla dostępnych danych uczących, aby jakość wspomaganie była najlepsza. To pytanie nasuwa konieczność wybrania kryterium oceny wyniku proponowanych algorytmów ekstrakcji reguł. Kryterium związane z najwyższą efektywnością, według którego ustalono parę progów η_e i η_b dla powyższych przykładów, może nie być najlepsze. Na rysunku 4.9 zaprezentowano wszystkie możliwe wyniki algorytmu przy zmianie progów η_e i η_b . Dla każdej pary progów obliczono dwie wartości: efektywność (4.2) i liczbę reguł (4.13). Taka para wartości opisujących rezultat jednego z algorytmów (ASR) w sposób jakościowy jest zaprezentowana na rysunku 4.9. jako punkt, którego odcięta stanowi osiągnięta efektywność, a rzędną – liczba reguł. Na podstawie tego wykresu można stwierdzić, że trzy wyniki osiągnęły maksymalną efektywność równą 98% (dla jedenastu, siedmiu i pięciu reguł). W dolnej części wykresu na rysunku 4.9 przedstawiono punkty odpowiadające wynikom związanym z najniższą liczbą reguł (trzy), dla których osiągnięto efektywności 93.33, 95.33% oraz 96%. Oczywiście wydaje się, że najbardziej pożądanym wynikiem byłoby jednoczesne osiągnięcie maksymalnej efektywności dla najmniejszej możliwej liczby reguł, np. 98% dla trzech reguł. Niestety wysoka efektywność wiąże się z uszczegółowieniem wiedzy a ono związane jest z powiększeniem zbioru reguł. Dlatego wspomniany wynik nie występuje na wykresie i może być nieosiągalny. Zwiększenie liczby reguł dla osiągnięcia pożądanej efektywności nie jest tak widoczne dla przykładu z bazą *iris flowers*, bo różnica między pięcioma regułami a minimalnie trzema jest nieduża. Niemniej jednak, dla przypadków medycznych z dużą liczbą symptomów w bazie, osiągnięcie najwyższej efektywności może wiązać się ze zbiorem reguł, który jest niemożliwy do interpretacji przez użytkowni-



(a)

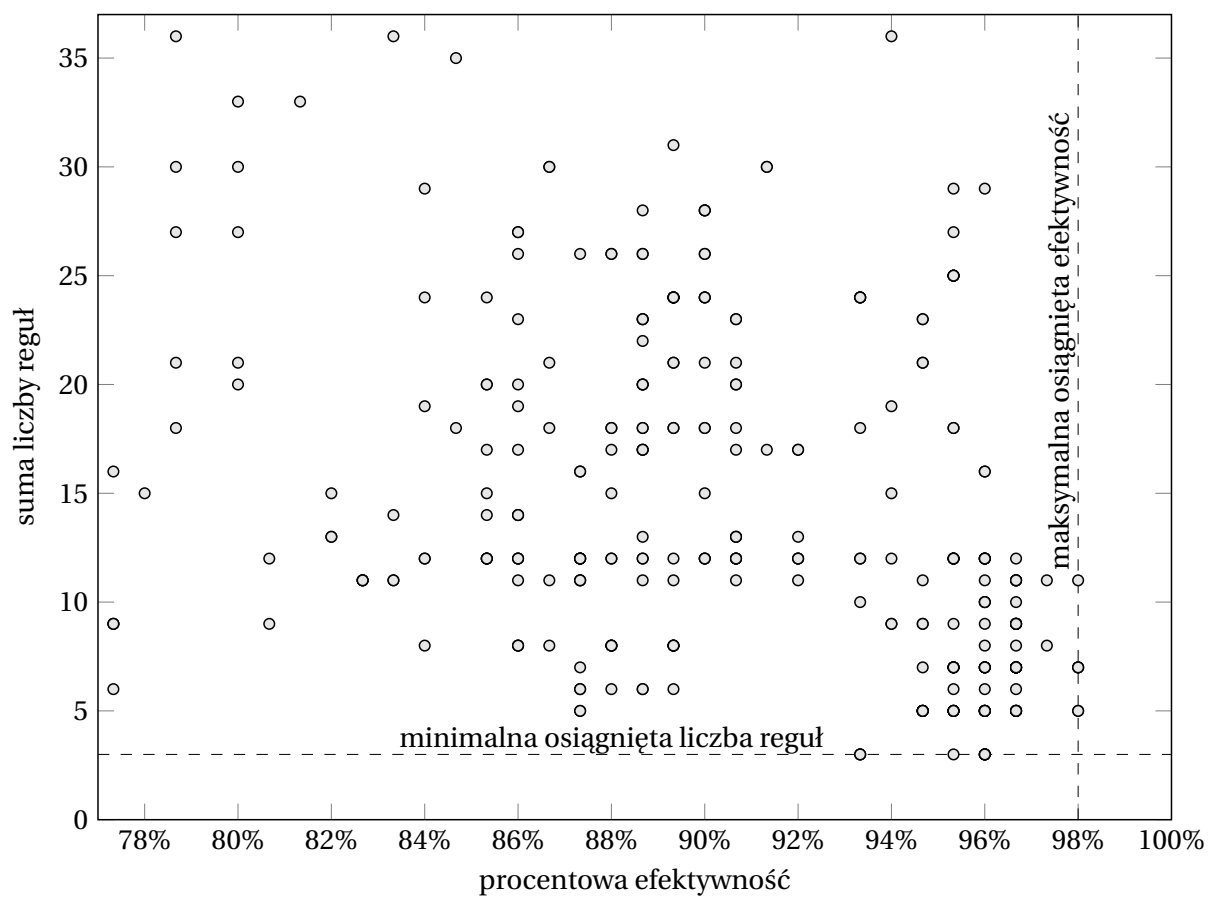


(b)



(c)

Rysunek 4.8: Wpływ zmiany progów η_e i η_b na efektywność klasyfikacji zbiorów reguł otrzymanych w wyniku działania algorytmów: (a) AER, (b) ASR, (c) AWR



Rysunek 4.9: Zależność całkowitej liczby reguł od efektywności klasyfikacji uzyskanych dla wszystkich kombinacji progów η_e i η_b w algorytmie wydobywania reguł ASR

ka ze względu na jego rozmiar. Dodatkowo, osiągnięcie wysokiej efektywności dla zbioru uczącego spowoduje szczegółowe dopasowanie się do niego. Z całą pewnością takie uszczegółowienie wiedzy będzie skutkowało słabą zdolnością uogólniania, którą sprawdza się po jakości rozpoznawania przypadków nieznanymi, określanymi także zbiorem danych testowych. Mimo, że wynik równy 98% efektywności dla bazy *iris flowers* przy pięciu regułach wydaje się wynikiem satysfakcjonującym, to dla złożonych problemów medycznych należy szukać rozwiązań, które będą stanowiły kompromis pomiędzy efektywnością zbioru reguł a ich złożonością.

W tym celu konieczne jest wprowadzenie wielkości umożliwiającej równoczesną ocenę tych parametrów. Zaproponowanym wskaźnikiem oceny reguł, który bierze pod uwagę jednocześnie liczbę reguł i poprawnych diagnoz jest przedstawiony poniżej parametr QI obliczany w następujący sposób

$$QI = \left(\frac{N_\epsilon}{N} \right) + \left(\frac{L}{C \cdot \sum_{k=1}^w \binom{\rho}{k}} \right). \quad (4.16)$$

Pierwszy składnik QI wyraża stosunek liczby błędów (N_ϵ) do liczby przypadków w bazie (N), a drugi składnik to stosunek liczby wydobytych reguł do liczby reguł utworzonych w początkowych zbiorach reguł (procedura 2 w punkcie 4.3). Taki sposób wydaje się lepszy od obliczania zwykłej sumy liczby błędów i reguł. Na przykład dla wyników: 5 błędów i 5 reguł, 7 błędów i 3 reguły czy 3 błędy i 7 reguł sumy błędów i reguł są jednakowe. Obliczanie QI dla tych trzech przypadków daje różne wyniki: 0.44, 0.17, 0.11, bo bada stopień zmniejszenia liczby błędów razem ze stopniem zmniejszenia liczby reguł. Zgodnie z zaproponowaną postacią QI , dla bazy *iris flowers* wybrano rozwiązanie, w którym uzyskano zbiór pięciu reguł diagnostycznych przedstawiony w tabeli 4.4. Określenia $A_3^{(1)}, \dots, A_4^{(3)}$ są zbiorami rozmytymi

Tabela 4.4: Zbiór reguł wydobyty na podstawie danych z bazy *iris flowers*

	$s_r^{(l)}$	$m_r^{(l)}$
$s_3^{(1)}$: X_3 jest $A_3^{(1)}$		1.00
$s_{10}^{(2)}$: X_3 jest $A_3^{(2)}$ i X_4 jest $A_4^{(2)}$		0.495
$s_4^{(2)}$: X_4 jest $A_4^{(2)}$		0.505
$s_4^{(3)}$: X_4 jest $A_4^{(3)}$		0.495
$s_3^{(3)}$: X_3 jest $A_3^{(3)}$		0.505

opisanymi funkcjami przynależności $\mu_3^{(1)}, \dots, \mu_4^{(3)}$ (patrz punkt 3.2). W tym przykładzie są to funkcje o kształtach trójkątnych. Są one zaprezentowane w dodatku A niniejszej rozprawy na rysunku A.1a. Pozostałe kształty funkcji przynależności można znaleźć na rysunkach A.1b-A.1d.

4.7 Podsumowanie

Zaproponowanie metod ekstrakcji reguł diagnostycznych na podstawie zbioru uczącego było głównym tematem niniejszego rozdziału. Każda z metod jest sekwencyjnym algorytmem eliminacji lub wyboru rozmytych elementów ogniskowych. Kolejność wyboru tych elementów uzależniona jest od oceny utworzonych zbiorów elementów ogniskowych, przeprowadzanej jako początkowy krok wszystkich zaproponowanych tu algorytmów. Aby zaprezentować przykład działania algorytmów wydobywania reguł użyto bazy *iris flowers*. Nie jest to baza medyczna, więc w rozdziale poświęconym wynikom przeprowadzonych badań nie będzie ona już omawiana. Mimo to, zaprezentowanie proponowanych algorytmów i metod oceny na jej przykładzie pozwoliło wstępnie dostrzec różnice w ich działaniu i powody kolejnych modyfikacji. W ostatnim punkcie tego rozdziału został zaproponowany wskaźnik *QI*. Jest on próbą oceny wyniku wydobywania reguł pod względem efektywności oraz złożoności zbioru reguł diagnostycznych. Dalsza część rozprawy będzie poświęcona bazom danych bezpośrednio związanym z zagadnieniami medycznymi. Warunki przeprowadzenia badań oraz ich wyniki zostaną przedstawione w rozdziale 5.

Rozdział 5

Wyniki przeprowadzonych badań

W niniejszym rozdziale przedstawione zostaną wyniki eksperymentów przeprowadzonych dla proponowanych metod wydobywania reguł diagnostycznych i wspomaganie diagnozy. Aby zachować czytelność prezentowanych treści, do eksperymentów wybrano pięć popularnych baz benchmarkowych związanych z zagadnieniami medycznymi. Bazy zawierają głównie dane numeryczne lub wyjątkowo symbole, ale interpretacja tych danych jest różna. Mamy więc do czynienia z danymi jakościowymi charakteryzującymi symptomy oraz wartościami liczbowymi, całkowitymi albo ułamkowymi będącymi wynikami przeprowadzonych testów. Osobną grupę danych jakościowych stanowią określenia symboliczne oznaczające płeć pacjenta lub wynik wywiadu lekarskiego. W tabeli 5.1 przedstawiono najistotniejsze parametry użytych baz: liczbę pacjentów N , liczbę symptomów ρ , nazwy poszczególnych diagnoz wraz z liczbą pacjentów przypisanych do każdej z nich. Możliwie dokładny opis symptomów przedstawiono w dodatku A niniejszej rozprawy.

Tabela 5.1: Opis użytych benchmarkowych baz medycznych

nazwa bazy	N	ρ	liczba przypadków w diagnozach
<i>thyroid disease</i> (zbiór <i>new-thyroid</i>) [99]	215	5	150 – eutyreoza 35 – niedoczynność tarczycy 30 – nadczynność tarczycy
<i>appendicitis</i> [120]	106	7	88 – ostre zapalenie wyrostka 18 – inny problem medyczny
<i>Pima Indians diabetes</i> [57]	768	8	268 – cukrzyca 500 – brak cukrzycy
<i>Wisconsin breast cancer</i> (zbiór <i>original</i>) [66]	699	9	458 – zmiany niezłośliwe 241 – zmiany złośliwe
<i>Cleveland heart disease</i> [66]	303	13	164 – brak choroby serca 139 – obecność chorób serca ¹

¹ W oryginalnej bazie wyróżnia się cztery diagnozy, natomiast m.in. z uwagi na jej mały rozmiar, podejścia opisywane w literaturze traktują ją jako zbiór dwóch diagnoz, np. [34].

Treść niniejszego rozdziału można podzielić na dwie zasadnicze części. Pierwsza z nich przedstawia działanie proponowanych metod koncentrując się głównie na wydobytych zbiorach reguł. Zbiór reguł jest bezpośrednim wynikiem każdej z metod, które na podstawie da-

nych pacjentów o znanych diagnozach (tzw. zbioru danych uczących) wyodrębniają reguły diagnostyczne najlepiej opisujące problem medyczny. W tej części przedstawiony zostanie wpływ kształtów funkcji przynależności na wynik wydobywania reguł, zbadany zostanie warunek zatrzymania algorytmu eliminacji reguł oraz złożoność reguł diagnostycznych. Dzięki tym wynikom możliwe będzie wskazanie istotnych różnic pomiędzy proponowanymi metodami oraz między sposobami oceny rozmytych elementów ogniskowych.

Druga część rozdziału jest poświęcona weryfikacji proponowanego systemu wspomaganie diagnozy. Opracowane metody wyodrębniania i oceny reguł będą testowane pod kątem zdolności uogólniania. Zostanie zbadane jak reguły diagnostyczne utworzone i wybrane na podstawie zbioru uczącego sprawdzają się we wspomaganie diagnozy pacjentów dotąd nieznanymi, czyli zawartymi w tzw. zbiorze danych testowych. Na podstawie wyników weryfikacji zostaną wybrane najlepsze z proponowanych w niniejszej rozprawie metod wspomaganie diagnozy.

Poza weryfikacją metod przeprowadzoną dla baz benchmarkowych, w badaniach wykorzystana zostanie także rzeczywista baza szpitalna dotycząca zagadnienia diagnozy stopnia zwłóknienia wątroby. Na zakończenie rozdziału, proponowane metody i algorytmy wyodrębniania zostaną porównane z kilkoma znanymi z literatury metodami klasyfikacji, które także są oparte na wyodrębnianiu reguł warunkowych.

5.1 Wpływ kształtu funkcji przynależności na rezultat wydobywania reguł

W tym punkcie, przy użyciu algorytmu eliminacji reguł (AER), przeanalizowano wpływ kształtu funkcji przynależności na wynik wydobywania reguł diagnostycznych. Wykorzystane zostaną funkcje przynależności trójkątne, trapezowe, uogólnione dzwonowe i gaussowskie łączone. Przetestowano, czy różne kształty funkcji powodują, że w wyniku działania algorytmu otrzymywane będą odmienne elementy ogniskowe i czy zmieni się efektywność wyznaczania diagnozy. W tabeli 5.2 zaprezentowano osiąganą efektywność (4.2) dla zbioru uczącego i liczbę reguł wydobytych dla każdego z proponowanych kształtów funkcji przynależności. Do tego eksperymentu wybrano ocenę reguł związaną z wartością bazowego prawdopodobieństwa oraz tradycyjną metodę obliczenia miary przekonania (3.15). Progi niepewności η_e i η_b mają bezpośredni wpływ na ocenę reguł jak i na obliczenie wspomnianej miary (3.15). Dlatego też w tabeli 5.2 przedstawiono wyniki z najlepszą osiągniętą efektywnością razem z wartościami wspomnianych progów. Dodatkowo, (co zostanie wyjaśnione w punkcie 5.3), ograniczono liczbę warunków logicznych w elemencie ogniskowym wybierając wartość parametru $w = 4$.

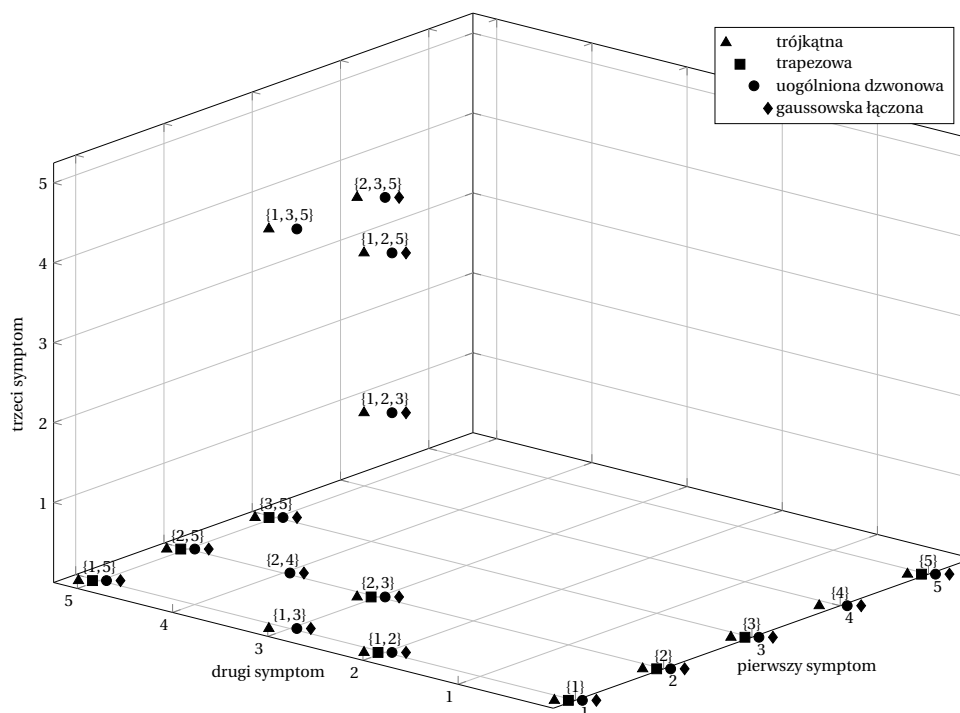
W tabeli 5.2 najwyższe procentowe efektywności E (4.2) diagnozy dla kolejnych baz zostały pogrubione. Można stwierdzić, że trójkątne funkcje przynależności pozwoliły na osiągnięcie najwyższej procentowej efektywności dla trzech z pięciu baz. Różnice procentowych wartości E (4.2) są większe dla baz liczniejszych (patrz tabela 5.1), np. *Pima Indians diabetes*. Dla tej bazy najwyższą efektywność osiągnięto dla gaussowskich łączonych funkcji przyna-

Tabela 5.2: Wpływ kształtu funkcji przynależności na rezultat algorytmu eliminacji reguł dla testowanych baz

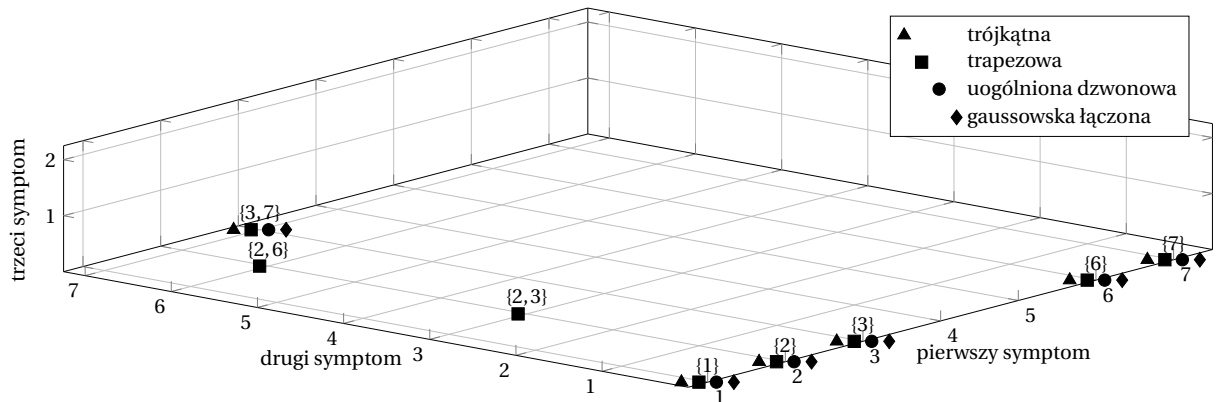
baza	funkcja przynależności	E [%]	L	η_e	η_b
<i>thyroid disease</i>	trójkątna	98.14	45	0.4	0.45
	trapezowa	99.07	26	0.6	0.25
	uogólniona dzwonowa	98.14	50	0.75	0.8
	gaussowska łączona	97.67	45	0.55	0.35
<i>appendicitis</i>	trójkątna	89.62	7	0.85	0.05
	trapezowa	82.08	18	0.3	0.05
	uogólniona dzwonowa	83.96	16	0.15	0.25
	gaussowska łączona	87.74	5	0.1	0.05
<i>Pima Indians diabetes</i>	trójkątna	76.69	12	0.15	0.15
	trapezowa	72.79	8	0.95	0.95
	uogólniona dzwonowa	74.74	22	0.05	0.65
	gaussowska łączona	76.82	8	0.2	0.15
<i>Wisconsin breast cancer</i>	trójkątna	97.42	13	0.5	0.25
	trapezowa	95.99	10	0.35	0.1
	uogólniona dzwonowa	96.42	18	0.2	0.05
	gaussowska łączona	97.28	26	0.2	0.25
<i>Cleveland heart disease</i>	trójkątna	85.48	18	0.25	0.45
	trapezowa	84.16	14	0.5	0.85
	uogólniona dzwonowa	83.83	12	0.2	0.2
	gaussowska łączona	84.82	15	0.2	0.75

leżności i jest to wynik wyższy o 4% od najlepszego wyniku znalezionej przy użyciu funkcji trapezowych. Z uwagi na to, że baza *Pima Indians diabetes* jest najbardziej liczna, to ta różnica efektywności równa się dodatkowym 30 błędnym diagnozom dla kształtów trapezowych. Z drugiej strony nietrudno zauważyć, że trapezowe funkcje przynależności pozwoliły osiągnąć najwyższą efektywność dla bazy *thyroid disease*. Warto odnotować, że liczba wydobytch reguł dla tego kształtu jest prawie o połowę niższa niż wynikowa liczebność zbiorów reguł dla innych kształtów.

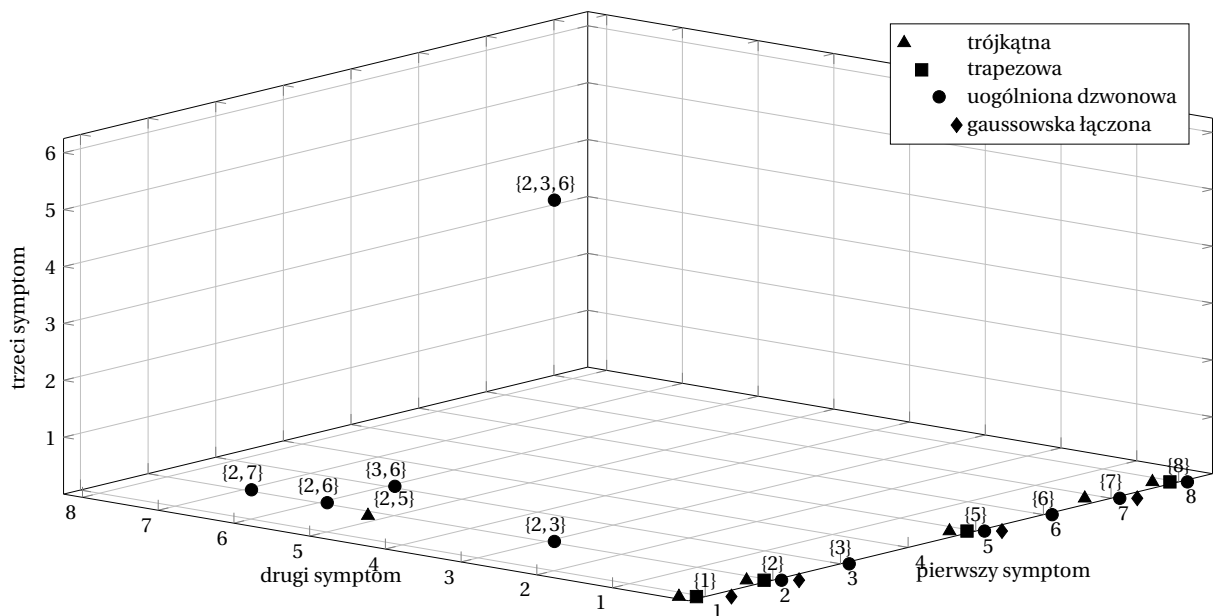
Z punktu widzenia wspomaganie diagnozy warto przeanalizować nie tylko efektywność i liczebność wyodrębnionych reguł, ale także to, czy kształt funkcji przynależności decyduje, które rozmyte elementy ogniskowe znajdują się w ostatecznym zbiorze reguł. Na rysunkach 5.1-5.5 przedstawiono w sposób graficzny wydobyte elementy ogniskowe w zależności od wybranych kształtów. Aby nie utrudnić interpretacji, zaprezentowano zbiory elementów ogniskowych tylko dla jednej diagnozy w każdej bazie. Każdy z elementów ogniskowych reprezentowany jest przez punkt we współrzędnych x, y, z gdzie wartość dla współrzędnej x jest numerem pierwszego symptomu w elemencie ogniskowym, y – numerem drugiego a z – trzeciego. Z uwagi na to, że na rysunkach 5.1-5.5 przedstawiono wykresy pseudo-trójwymiarowe, liczbę warunków ograniczono do trzech w pojedynczym elemencie ogniskowym ($w = 3$). Przykładowy element ogniskowy $s_r^{(l)}$: „ x_1 jest $A_1^{(l)}$ i x_3 jest $A_3^{(l)}$ i x_4 jest $A_4^{(l)}$ ” jest opisany przez indeksy symptomów $J_r^{(l)} = \{1, 3, 4\}$. Ten element jest prezentowany jako znacznik w punkcie o współrzędnych: $x = 1, y = 3$ i $z = 4$. Mniej złożone elementy ogniskowe mają zerowe wartości pozostałych współrzędnych. Przykładowo, prosty element ogniskowy $s_r^{(l)}$: „ x_2 jest $A_2^{(l)}$ ”, dla którego $J_r^{(l)} = \{2\}$, reprezentowany jest przez odpowiedni znacznik w punkcie o współrzędnych $x = 2, y = z = 0$.



Rysunek 5.1: Zbiór elementów ogniskowych jako wynik algorytmu eliminacji reguł dla czterech kształtów funkcji przynależności – baza *thyroid disease*

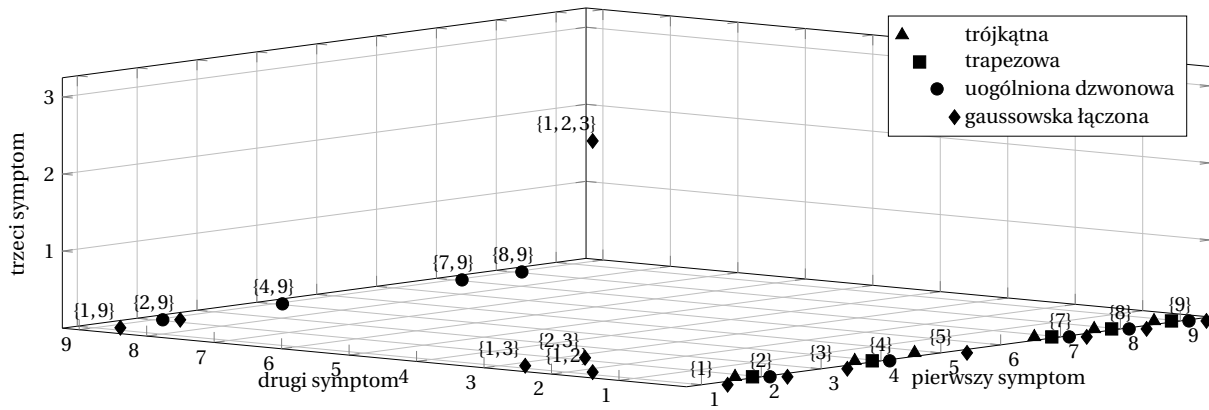


Rysunek 5.2: Zbiór elementów ogniskowych jako wynik algorytmu eliminacji reguł dla czterech kształtów funkcji przynależności – baza *appendicitis*

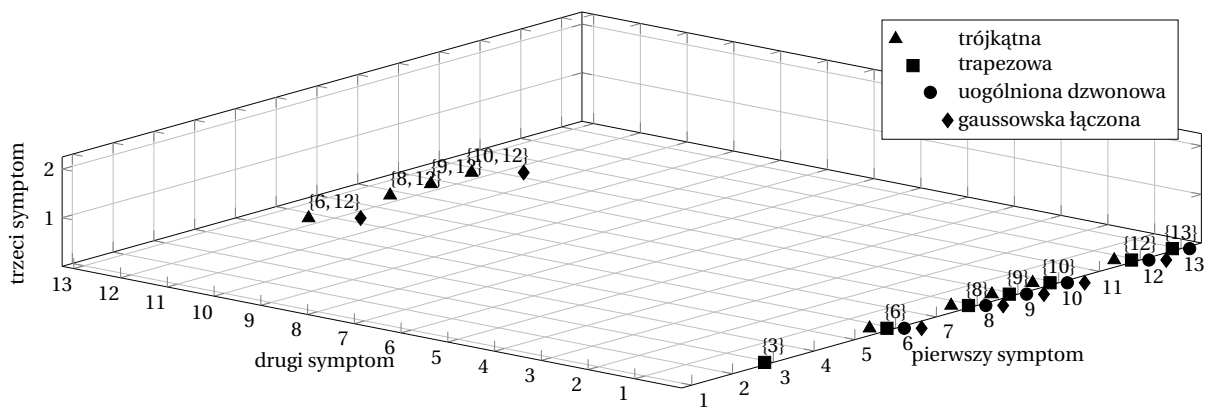


Rysunek 5.3: Zbiór elementów ogniskowych jako wynik algorytmu eliminacji reguł dla czterech kształtów funkcji przynależności – baza *Pima Indians diabetes*

Rysunki 5.1-5.5 pokazują, że mimo różnic w liczebności (patrz tabela 5.2), zmiana kształtu nie powoduje, że algorytm eliminacji wydobywa zupełnie inny zbiór elementów ogniskowych. Szczególnie dla bazy *thyroid disease* (rysunek 5.1) żaden z kształtów nie skutkuje wydobyciem indywidualnego elementu ogniskowego. Dla pozostałych baz występuje kilka elementów ogniskowych wydobytych tylko dla jednego z kształtów. Wiąże się to z tym, że dodanie kilku elementów ogniskowych pozwoliło nieznacznie zwiększyć efektywność (np. baza *Cleveland heart disease* – kształty trójkątne). Należy stwierdzić, że algorytm eliminacji reguł wydobywa zbliżone zbiory reguł niezależnie od kształtów funkcji przynależności. Okazuje się, że wartość wskaźnika oceny (w tym wypadku wartości bazowego prawdopodobieństwa) elementu ogniskowego niewiele się zmienia dla różnych kształtów tych funkcji. Dlatego w dalszej części rozprawy wyniki przeprowadzonych badań będą dotyczyły tylko jednego kształtu. Pozwoli to uniknąć nadmiaru prezentowanych wyników i zachować czytelność ich



Rysunek 5.4: Zbiór elementów ogniskowych jako wynik algorytmu eliminacji reguł dla czterech kształtów funkcji przynależności – baza *Wisconsin breast cancer*



Rysunek 5.5: Zbiór elementów ogniskowych jako wynik algorytmu eliminacji reguł dla czterech kształtów funkcji przynależności – baza *Cleveland heart disease*

analizy. Na podstawie przedstawionych wyników można wysunąć wniosek, że można wydobywać reguły oparte na jednym wybranym kształcie a później go dostrajać aby poprawić efektywność E . Celem wyboru najbardziej pożądanego kształtu funkcji przynależności, w tabeli 5.3 przedstawiono uszeregowanie kształtów dla poszczególnych baz pod względem osiągniętej efektywności. Dodatkowo w tabeli 5.4 przedstawiono taką samą ocenę, ale pod względem osiąganej liczebności reguł. W tabelach 5.3 i 5.4 wyniki oznaczano za pomocą cyfr rzymskich gdzie „I” oznacza wynik najlepszy a „IV” wynik najgorszy. Ostatni wiersz tabel 5.3 i 5.4 podsumowuje jaki najczęstszy wynik otrzymywano dla danego kształtu.

Tabela 5.3: Porównanie najlepszych kształtów funkcji przynależności wybranych dla każdej z baz na podstawie wartości efektywności przedstawionych w tabeli 5.2

baza	μ_T	μ_{TR}	μ_B	μ_Γ
<i>thyroid disease</i>	II	I	III	IV
<i>appendicitis</i>	I	IV	III	II
<i>Pima Indians diabetes</i>	II	IV	III	I
<i>Wisconsin breast cancer</i>	I	IV	III	II
<i>Cleveland heart disease</i>	I	III	IV	II
najczęstszy wynik	I	IV	III	II

μ_T – funkcja trójkątna, μ_{TR} – trapezowa, μ_B – uogólniona dzwonowa, μ_Γ – gaussowska łączona

Tabela 5.4: Porównanie najlepszych kształtów funkcji przynależności wybranych dla każdej z baz na podstawie liczebności zbiorów reguł diagnostycznych przedstawionych w tabeli 5.2

baza	μ_T	μ_{TR}	μ_B	μ_Γ
<i>thyroid disease</i>	II	I	III	II
<i>appendicitis</i>	II	IV	III	I
<i>Pima Indians diabetes</i>	II	I	III	I
<i>Wisconsin breast cancer</i>	II	I	III	IV
<i>Cleveland heart disease</i>	IV	II	I	III
najczęstszy wynik	II	I	III	I

μ_T – funkcja trójkątna, μ_{TR} – trapezowa, μ_B – uogólniona dzwonowa, μ_Γ – gaussowska łączona

Na podstawie tabeli 5.3 widać, że trójkątne funkcje przynależności osiągały najczęściej najwyższą notę, to znaczy najwyższą efektywność a funkcje gaussowskie łączone były pod tym względem na drugim miejscu. Z drugiej strony, funkcje trapezowe i gaussowskie łączone osiągały najczęściej najlepszy wynik jeśli brać pod uwagę liczebność zbiorów rozmytych elementów ogniskowych. W tym zestawieniu trójkątne kształty są na drugim miejscu. Na tej podstawie można sugerować użycie trójkątnych funkcji przynależności i gaussowskich łączonych jako tych, które będą łączyły możliwie najwyższą efektywność z najniższą liczbą reguł. Wykresy funkcji przynależności przedstawione w dodatku A pokazują, że przedział

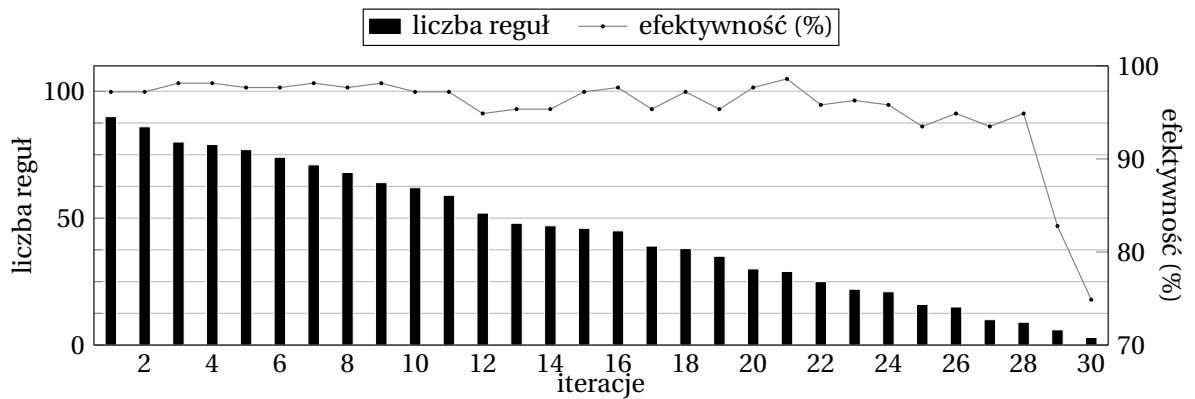
częściowej przynależności do zbiorów rozmytych opisanych funkcjami trójkątnymi i gaussowskimi łączonymi jest szerszy niż dla zaproponowanych kształtów trapezowych i uogólnionych dzwonowych. Dzięki temu dla trójkątnych funkcji przynależności i gaussowskich łączonych możliwe jest lepsze dostrojenie progów niepewności η_e oraz η_b . Trójkątne funkcje przynależności, które mają kształty przedziałowo-liniowe [93] wydają się mieć najbardziej intuicyjny charakter a ich ewentualne przekształcenie do kształtów gaussowskich łączonych jest możliwe [69]. Dlatego też użyto ich w kolejnych badaniach, których wyniki prezentuje dalsza część rozdziału.

5.2 Warunek zatrzymania algorytmu eliminacji reguł

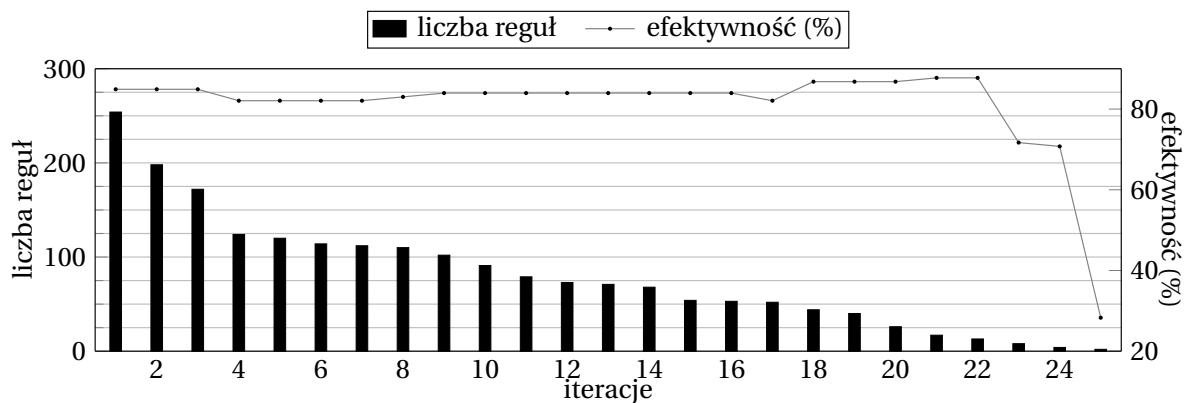
Opisując w punkcie 4.2 algorytm eliminacji reguł, zaznaczono, że jego zatrzymanie nie jest podyktowane pierwszym wzrostem liczby błędnie wyznaczonych diagnoz. Na rysunkach 5.6a-5.7b przedstawiono przykładowe działanie algorytmu eliminacji reguł (AER) dla pięciu baz danych. Na każdym z rysunków zaprezentowano jak zmienia się liczba reguł oraz podano efektywność rozpoznawania przypadków ze zbioru uczącego w kolejnych iteracjach. W celu analizy zagadnienia warunek zatrzymania dla tych przykładów nie został określony, a sama eliminacja zostaje przerwana dopiero kiedy w zbiorze reguł pozostaje minimalna liczba reguł (co daje trzy elementy ogniskowe dla bazy *thyroid disease* oraz dwa dla pozostałych baz). Na początku algorytmu ustalono, że eliminacja będzie przeprowadzona począwszy od maksymalnej możliwej do utworzenia liczby elementów ogniskowych w zbiorach reguł, czyli wartości parametru $w = \rho$, gdzie ρ jest liczbą symptomów występujących w poszczególnych bazach (tabela 5.1). Na podstawie obserwacji przedstawionych w punkcie 5.1, do utworzenia elementów ogniskowych wykorzystano kształty trójkątne. Wartość bazowego prawdopodobieństwa ponownie posłużyła za metodę oceny reguł.

Na podstawie rysunku 5.6a dotyczącego bazy *thyroid disease* można zauważyć wahania efektywności (oznaczonej poprzez punkty połączone liniami). Wskazują one, że wybór momentu przerwania algorytmu i uzyskania ostatecznego zbioru reguł nie jest zadaniem trywialnym. Można zauważyć, że 21. iteracja to moment, w którym zbiór 29 reguł uzyskuje najwyższą procentową efektywność równą 98.6%. Chociaż efektywność następnie spada, to w 28. iteracji uzyskany zostaje bardzo mały zbiór dziewięciu reguł, który pozwala uzyskiwać jej akceptowalną wartość równą 94.88%. Mniejsze zmiany efektywności mogą być zaobserwowane przy eliminacji dla bazy *appendicitis* (rysunek 5.6b). Tutaj w 4. i 17. iteracji eliminacja reguł spowodowała nieznaczne zmniejszenie efektywności (z 84.91 do 82.07%) ale do 21. iteracji efektywność jeszcze wzrasta i wynosi 87.33%. Utrzymuje się także po kolejnej eliminacji, kiedy liczba reguł zostaje zmniejszone do trzynastu.

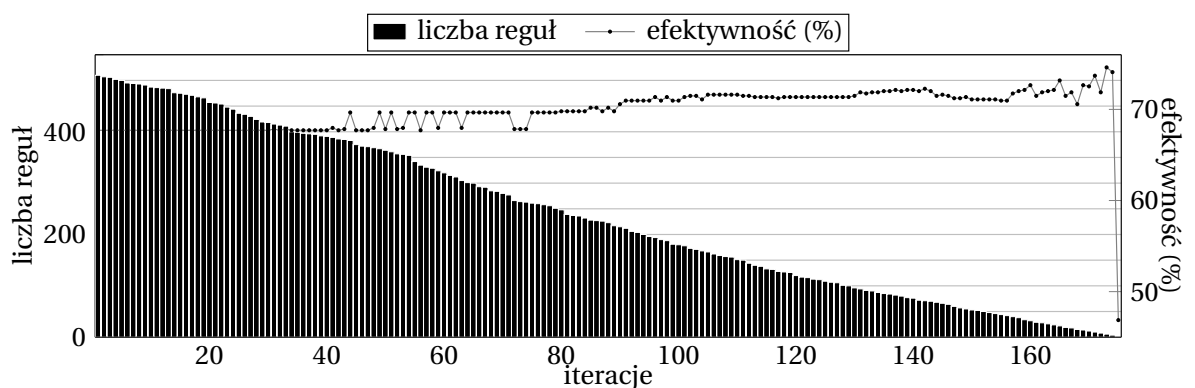
Analogiczna interpretacja zmian efektywności i liczby reguł na podstawie rysunków 5.6c-5.7b pozwala zauważyć, że działanie algorytmu eliminacji jest podobne jak dla baz *thyroid disease* oraz *appendicitis*. Nieco większe wahania efektywności obserwuje się dla bazy *Pima Indians diabetes* (rysunek 5.6c), a najłagodniejsze zmiany efektywności przedstawiają wyniki dla bazy *Wisconsin breast cancer* (rysunek 5.7a).



(a)

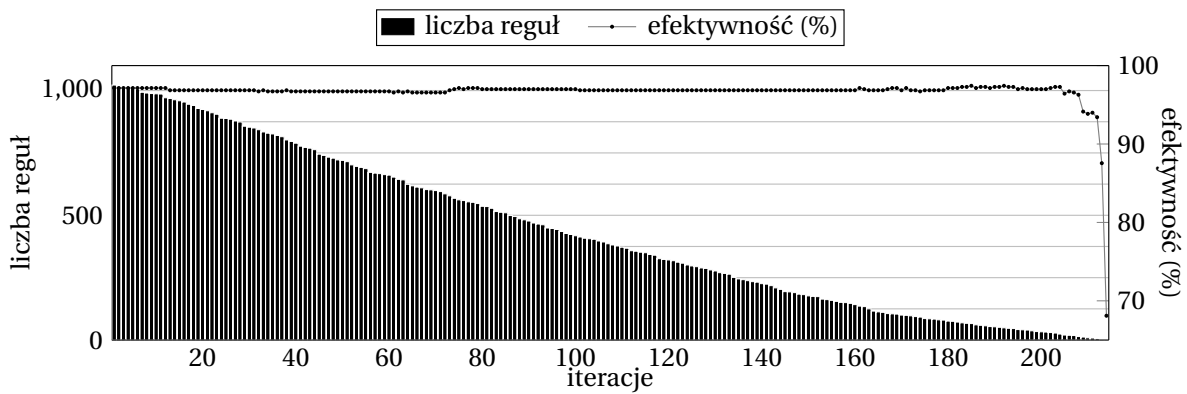


(b)

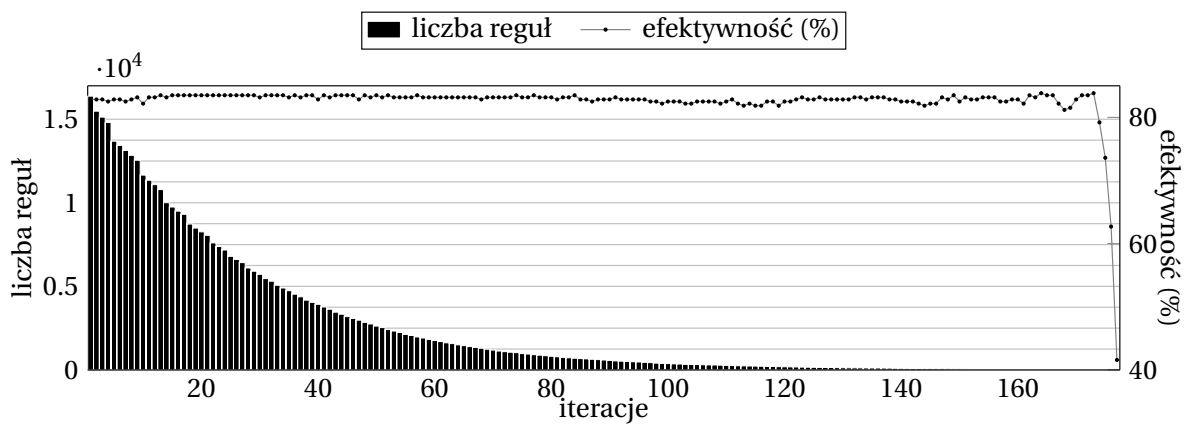


(c)

Rysunek 5.6: Zmiana liczby reguł oraz procentowej efektywności (4.2) w każdej iteracji algorytmu eliminacji reguł dla baz: (a) *thyroid disease*, (b) *appendicitis*, (c) *Pima Indians diabetes*

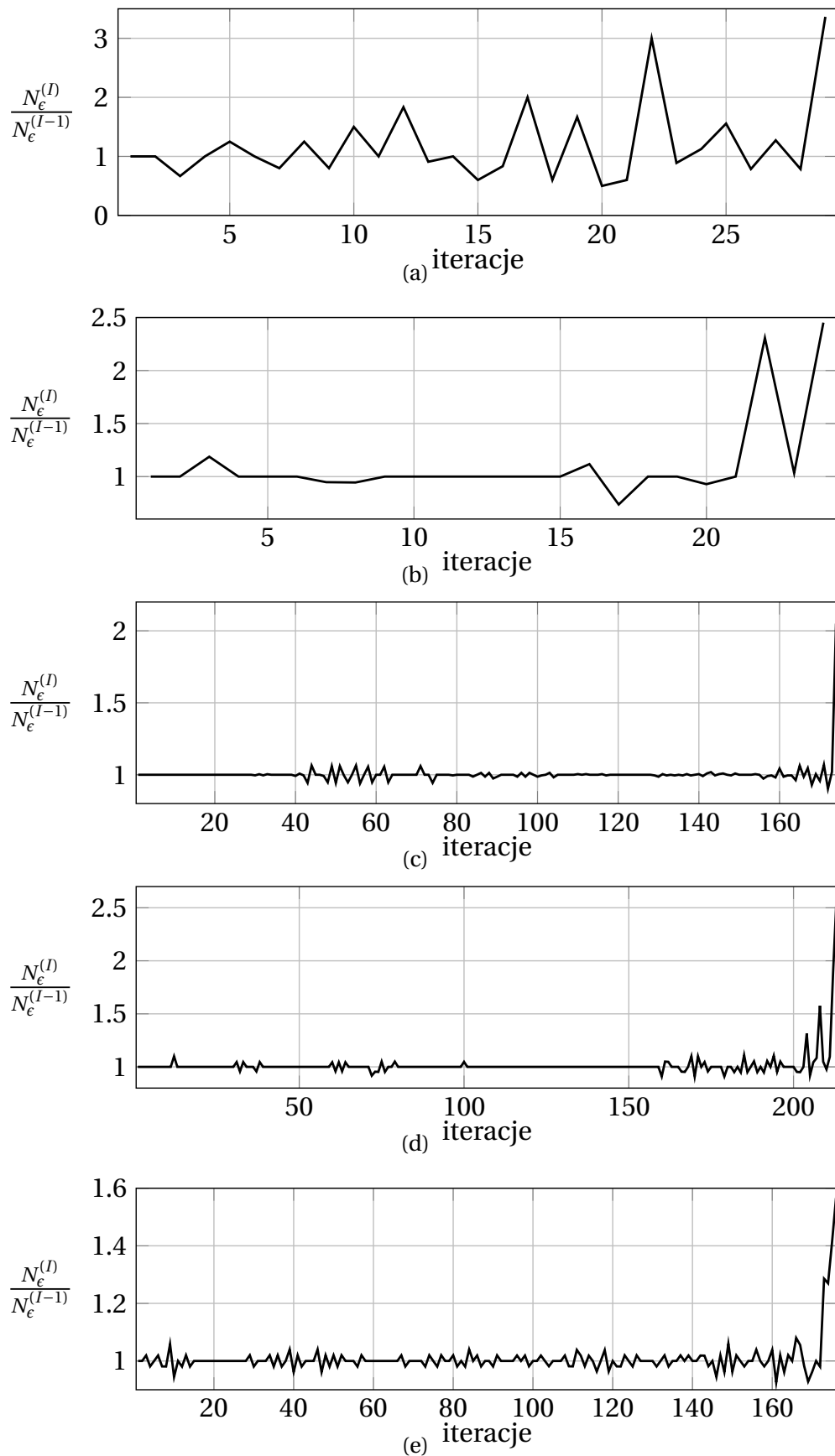


(a)



(b)

Rysunek 5.7: Zmiana liczby reguł oraz procentowej efektywności (4.2) w każdej iteracji algorytmu eliminacji reguł dla baz: (a) *Wisconsin breast cancer*, (b) *Cleveland heart disease*



Rysunek 5.8: Dynamika zmian liczby błędnych diagnoz w iteracjach algorytmu eliminacji reguł dla baz: (a) *thyroid disease*, (b) *appendicitis*, (c) *Pima Indians diabetes*, (d) *Wisconsin breast cancer*, (e) *Cleveland heart disease*

Na rysunku 5.8 przedstawiono dynamikę zmian liczby błędów. W proponowanym algorytmie ustalono, że warunek zatrzymania związany jest ze sprawdzaniem jak zwiększyła się liczba błędnych diagnoz w danej iteracji (oznaczymy ją jako $N_e^{(I)}$) w stosunku do iteracji poprzedniej, tj. $N_e^{(I-1)}$. Wartość $\frac{N_e^{(I)}}{N_e^{(I-1)}} < 1$ oznacza więc polepszenie efektywności E . Naturalnie $\frac{N_e^{(I)}}{N_e^{(I-1)}} = 1$ wskazuje na utrzymanie poprzedniej liczby błędów, a $\frac{N_e^{(I)}}{N_e^{(I-1)}} > 1$ oznacza, że liczba błędów wzrosła.

Na podstawie rysunku 5.8 można zauważyć, że wybranie akceptowalnego wzrostu błędu nie jest jednoznaczne i trudno wskazać uniwersalny warunek zatrzymania algorytmu dla wszystkich zbiorów danych. W przypadku bazy *thyroid disease*, celem ograniczenia liczby reguł, możliwe jest zaniechanie zatrzymania aż do trzykrotnego wzrostu liczby błędów w iteracji, co można było zauważyć na rysunku 5.8a. Pozostałe wykresy przedstawione na rysunku 5.8 pokazują, że warunek zatrzymania powinien wiązać się z mniejszym wzrostem liczby błędów, tj. gdy $\frac{N_e^{(I)}}{N_e^{(I-1)}} = 1.2, 1.5$ lub 2 . Dowodzi to, że wartość progowa $\frac{N_e^{(I)}}{N_e^{(I-1)}}$, po której należałoby zakończyć algorytm eliminacji reguł, staje się trudna w oszacowaniu. Należy więc stwierdzić, że sposób wydobywania reguł w postaci iteracyjnej eliminacji nie jest pozbawiony wad.

Poza wspomnianym warunkiem zatrzymania, algorytm eliminacji reguł (AER) jest wyzwaniem z uwagi na złożoność obliczeniową. Warto przypomnieć chociażby potrzebę obliczania rozkładu bazowego prawdopodobieństwa w każdej iteracji algorytmu eliminacji tzn. kiedy liczba elementów ogniskowych jest duża. Obie wspomniane cechy charakterystyczne tego podejścia stały się przyczyną poszukiwania efektywniejszego sposobu wydobywania reguł diagnostycznych opartych na teorii Dempstera-Shafera, co doprowadziło do zaproponowania algorytmu selekcji reguł (ASR) [96].

5.3 Złożoność wydobywanych reguł diagnostycznych

Rozpatrzmy znaczenie parametru w wydobywania reguł diagnostycznych, który wpływa bezpośrednio na początkowe zbiory reguł diagnostycznych, czyli ograniczenie złożoności elementów ogniskowych. Biorąc pod uwagę złożoność reguł diagnostycznych, można dążyć do tego, aby były one jak najprostsze, najlepiej z jednym warunkiem logicznym w elemencie ogniskowym. Nie znajduje to jednak potwierdzenia w diagnostyce klinicznej, w której prawie zawsze uwzględnia się także wspólne objawy [110]. Dlatego w proponowanej metodzie także złożone elementy ogniskowe także są tworzone i poddane ocenie. Niemniej jednak, nie wydaje się, aby koniecznym było wydobywanie elementów ogniskowych o złożoności przekraczającej możliwości skutecznej interpretacji. W [42] wspomina się o maksymalnych siedmiu warunkach logicznych w przesłance reguły jako liczbie pojedynczych pojęć, które są możliwe do jednoczesnej interpretacji przez człowieka powołując się na wyniki przedstawione w [79]. Nowsze źródło argumentuje, że taka maksymalna liczba pojęć może wynieść co najwyżej cztery [26].

Aby sprawdzić złożoność reguł diagnostycznych przeprowadzono ich wydobywanie dla wszystkich baz benchmarkowych z użyciem kolejnego algorytmu tj. ASR. Tym razem do oce-

ny reguł nie zostanie użyta tylko wartość bazowego prawdopodobieństwa, ale także zaczerpnięty z metod oceny klasyfikacji wskaźnik MCC (3.41) oraz nowy wprowadzony wskaźnik CC (3.42). Na przykładzie przedstawionym w punkcie 3.7 pokazano, że oba zaproponowane wskaźniki w różny sposób wartościują złożoność elementów ogniskowych. Dlatego słusznym wydaje się porównanie ich wpływu na wynik selekcji reguł (działanie ASR przedstawiono w punkcie 4.3). Reguły zostaną wyodrębnione dla omawianych baz benchmarkowych. Złożoność reguł będzie zbadana przy pomocy wektora Θ (4.14) zdefiniowanego w punkcie 4.4.2. Wartości w wektorze Θ ilustrują liczbę elementów ogniskowych z jednym warunkiem logicznych, z dwoma, z trzema, itd. Używając procedury 2 opisanej w punkcie 4.1, można utworzyć i ocenić wszystkie możliwe elementy ogniskowe (których liczba jest wykładniczo związana z liczbą symptomów).

W obecnych obliczeniach przedstawione zostaną wyniki wydobywania reguł, kiedy ich złożoność przy tworzeniu początkowych zbiorów została ograniczona do siedmiu warunków logicznych. Oczywiście, dla bazy *thyroid disease* to ograniczenie jest nadmiarowe i maksymalna złożoność elementów ogniskowych dla tej bazy jest równa liczbie symptomów czyli pięć. Do utworzenia rozmytych elementów ogniskowych użyto trójkątnych funkcji przynależności a tradycyjna metoda obliczania miary przekonania (3.15) użyta jest przy wyznaczeniu diagnozy.

Tabela 5.5: Złożoność uzyskanych zbiorów reguł diagnostycznych (4.14) dla każdej metody oceny w wyniku algorytmu selekcji reguł (ASR) dla bazy *thyroid disease*

metoda oceny	$\theta^{(1)}$	$\theta^{(2)}$	$\theta^{(3)}$	$\theta^{(4)}$	$\theta^{(5)}$	$E(\%)$
m	15	22	4	0	0	97.67
MCC	9	9	0	0	0	97.67
CC	2	7	6	0	0	98.14

Tabela 5.6: Złożoność uzyskanych zbiorów reguł diagnostycznych (4.14) dla każdej metody oceny w wyniku algorytmu selekcji reguł (ASR) dla bazy *appendicitis*

metoda oceny	$\theta^{(1)}$	$\theta^{(2)}$	$\theta^{(3)}$	$\theta^{(4)}$	$\theta^{(5)}$	$\theta^{(6)}$	$\theta^{(7)}$	$E(\%)$
m	9	3	0	0	0	0	0	87.74
MCC	2	5	4	1	0	0	0	87.74
CC	5	0	5	0	0	0	0	90.57

Tabela 5.7: Złożoność uzyskanych zbiorów reguł diagnostycznych (4.14) dla każdej metody oceny w wyniku algorytmu selekcji reguł (ASR) dla bazy *Pima Indians diabetes*

metoda oceny	$\theta^{(1)}$	$\theta^{(2)}$	$\theta^{(3)}$	$\theta^{(4)}$	$\theta^{(5)}$	$\theta^{(6)}$	$\theta^{(7)}$	$E(\%)$
m	9	1	0	0	0	0	0	77.73
MCC	4	4	0	0	0	0	0	76.95
CC	2	7	5	2	0	0	0	77.60

Tabela 5.8: Złożoność uzyskanych zbiorów reguł diagnostycznych (4.14) dla każdej metody oceny w wyniku algorytmu selekcji reguł (ASR) dla bazy *Wisconsin breast cancer*

metoda oceny	$\theta^{(1)}$	$\theta^{(2)}$	$\theta^{(3)}$	$\theta^{(4)}$	$\theta^{(5)}$	$\theta^{(6)}$	$\theta^{(7)}$	$E(\%)$
<i>m</i>	12	2	0	0	0	0	0	97.14
<i>MCC</i>	11	5	0	0	0	0	0	97.28
<i>CC</i>	11	7	0	0	0	0	0	97.14

Tabela 5.9: Złożoność uzyskanych zbiorów reguł diagnostycznych (4.14) dla każdej metody oceny w wyniku algorytmu selekcji reguł (ASR) dla bazy *Cleveland heart disease*

metoda oceny	$\theta^{(1)}$	$\theta^{(2)}$	$\theta^{(3)}$	$\theta^{(4)}$	$\theta^{(5)}$	$\theta^{(6)}$	$\theta^{(7)}$	$E(\%)$
<i>m</i>	11	7	0	0	0	0	0	85.48
<i>MCC</i>	8	10	0	0	0	0	0	87.13
<i>CC</i>	4	8	0	0	0	0	0	86.47

Analizując wyniki przedstawione w tabelach 5.5-5.9, można zauważyć, że bez względu na metodę oceny, bardzo rzadko wydobywane są reguły z elementami ogniskowymi o większej złożoności niż trzy warunki logiczne. Tylko dla dwóch spośród wszystkich wyników, wydobyto elementy ogniskowe o czterech warunkach logicznych. Metody oceny związane z *MCC* lub *CC* pozwalały osiągnąć wyniki o wyższych efektywnościach niż te przy użyciu bazowego prawdopodobieństwa, choć nie zawsze wiązało się to z najmniejszą liczbą reguł (np. wyniki dla *Wisconsin breast cancer* i *Cleveland heart disease*). Wyjątek stanowi baza *Pima Indians diabetes*, dla której ocena reguł w oparciu o bazowe prawdopodobieństwo okazała się najlepsza.

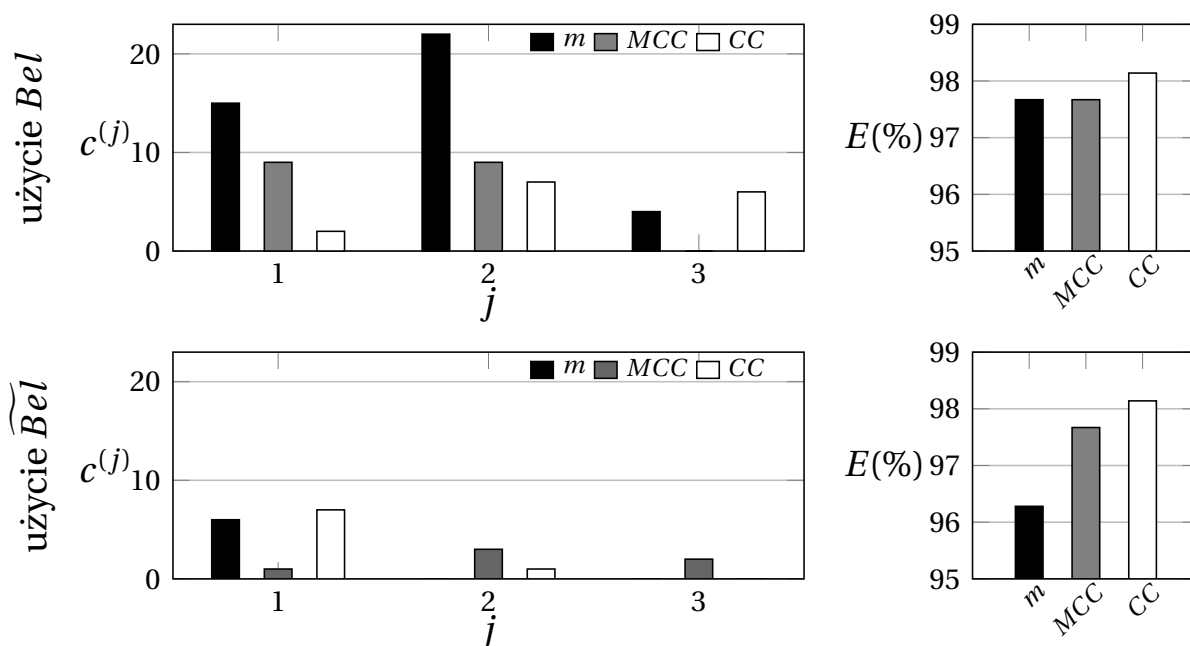
Najważniejszym wnioskiem z przeprowadzonego w tym punkcie badania jest spostrzeżenie, że bez względu na użytą metodę oceny (zaniedbując uzyskiwane efektywności) wyodrębniane zostają bardzo proste zbiory reguł. Podsumowując, nie wydaje się aby koniecznym było tworzenie początkowego zbioru reguł z tak złożonymi elementami ogniskowymi jak w tym punkcie, skoro i tak wybierane są spośród nich te najprostsze. To one osiągają najwyższe wartości ocen. Mając możliwość ograniczenia złożoności obliczeniowej związanej z utworzeniem tej nadmiarowej liczby reguł i wyznaczenia dla każdej z nich wartości oceny, nic nie stoi na przeszkodzie, aby tworzyć reguły z elementami ogniskowymi o liczbie warunków logicznych ograniczonych tylko do czterech.

5.4 Zmodyfikowana miara przekonania

W dotychczasowych rozważaniach przedstawiono działanie algorytmów eliminacji i selekcji reguł a każdy z przedstawionych wyników był związany z obliczaniem tradycyjnej miary przekonania Bel (3.15). Obecnie zostanie sprawdzone jaki wpływ na wydobywaną wiedzę, tzn. zbiór reguł diagnostycznych i efektywność diagnozy ma zastosowanie zmodyfikowanej miary przekonania \widetilde{Bel} (3.21) zamiast tradycyjnej. Aby jednoznacznie porównać wyniki otrzymane dwoma sposobami, obliczenia z punktu 5.3 dotyczącego badania złożoności

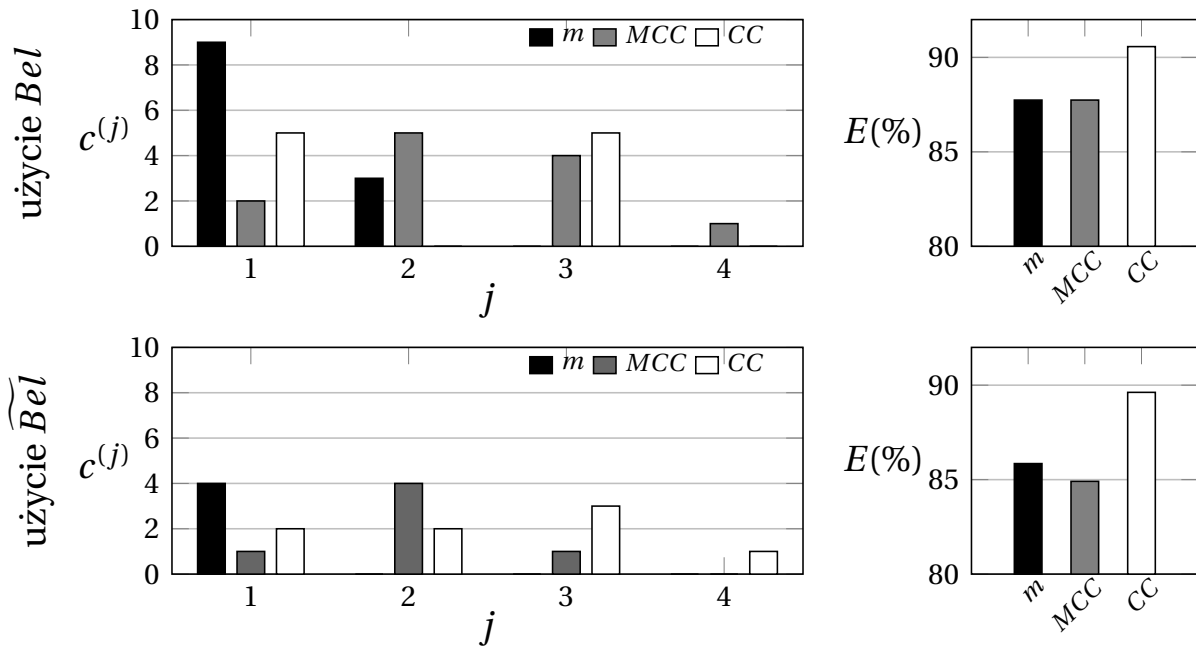
reguł zostaną powtórzone, przy czym zmieniona będzie tylko metoda obliczenia miary przekonania na \widetilde{Bel} (3.21). W celu czytelnej prezentacji, tym razem wyniki przedstawione w tabelach 5.5-5.9 oraz obliczone dla zmodyfikowanej miary przekonania, zostaną przedstawione na rysunkach.

Wyniki dla trzech metod oceny, uzyskane dla tradycyjnej miary Bel (wyniki z tabel 5.5-5.9), przedstawiono w górnej części rysunków 5.9-5.13 a dla zmodyfikowanej \widetilde{Bel} , w dolnej części. Wykresy słupkowe w lewej części tych rysunków przedstawiają liczebność elementów ogniskowych o danej złożoności (4.14) a wykresy słupkowe w prawej części – osiągnięte efektywności (4.2). Dla łatwiejszego porównania, zachowano tę samą skalę dla wykresów górnych i dolnych.

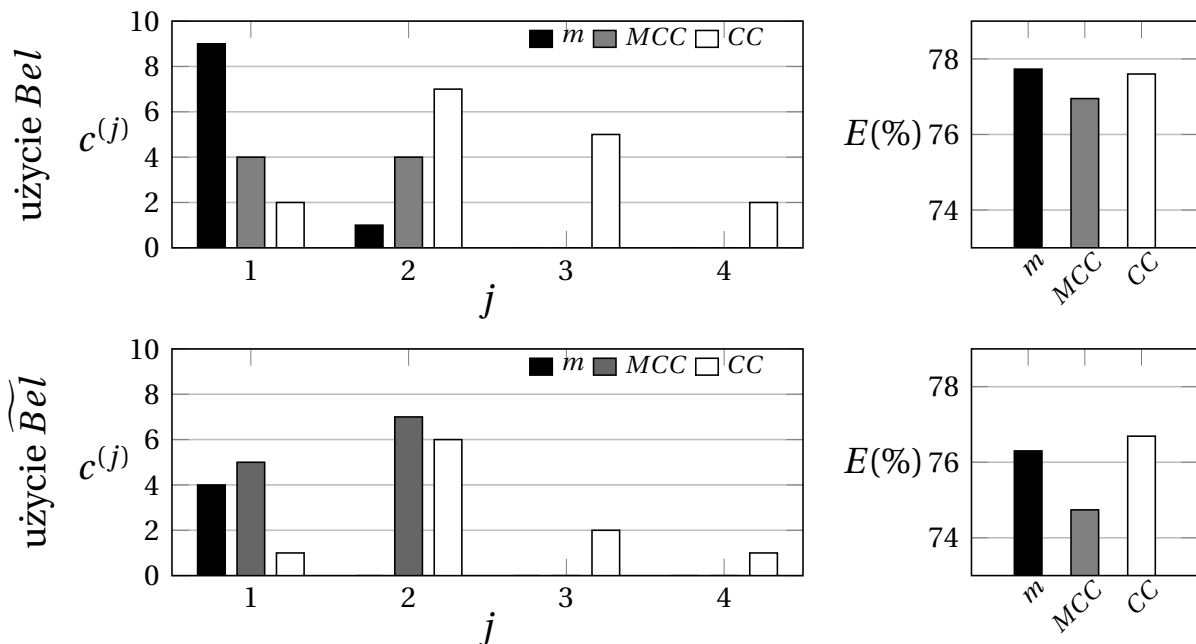


Rysunek 5.9: Porównanie wyników algorytmu selekcji reguł dla bazy *thyroid disease* otrzymanych za pomocą dwóch sposobów obliczania miary przekonania

Porównując górne i dolne wykresy słupkowe lewe (dotyczące liczebności rozmytych elementów ogniskowych $\theta^{(j)}$) można łatwo zauważyć, że użycie \widetilde{Bel} pozwala znacząco zmniejszyć złożoność reguł. Dla wszystkich metod oceny sumaryczna liczba reguł jest mniejsza. Oczywiście, zmniejszenie liczby reguł wpływa na uogólnienie wiedzy. Dlatego w większości przypadków efektywność E spada. Aby lepiej dostrzec różnice, w tabeli 5.10 przedstawiono jak przy użyciu \widetilde{Bel} zmienia się procentowo liczba reguł oraz procentowa efektywność uczenia dla trzech metod oceny. Można zauważyć, że modyfikacja miary przekonania diametralnie zmniejsza liczbę reguł przy ocenie związanej z wartością m , choć związane jest to ze spadkiem efektywności o 1-2%. Bardziej znaczące różnice w procentowej efektywności obserwuje się dla wskaźnika MCC (od zera do prawie 3%) a najmniejsze spadki dla wskaźnika CC (od zera do 1%). Dla tego wskaźnika zmiana miary przekonania, dwukrotnie nie spowodowała spadku efektywności, mimo że liczba reguł została zmniejszona prawie o połowę (dla *thyroid disease* – o 46.7% oraz *Wisconsin breast cancer* – o 44.4%). Tylko w jednym przypadku (dla bazy *Cleveland heart disease*) otrzymano więcej reguł przy spadku efektywności, choć nadal jest to lepszy wynik od otrzymanego przy udziale MCC (rysunek 5.13).



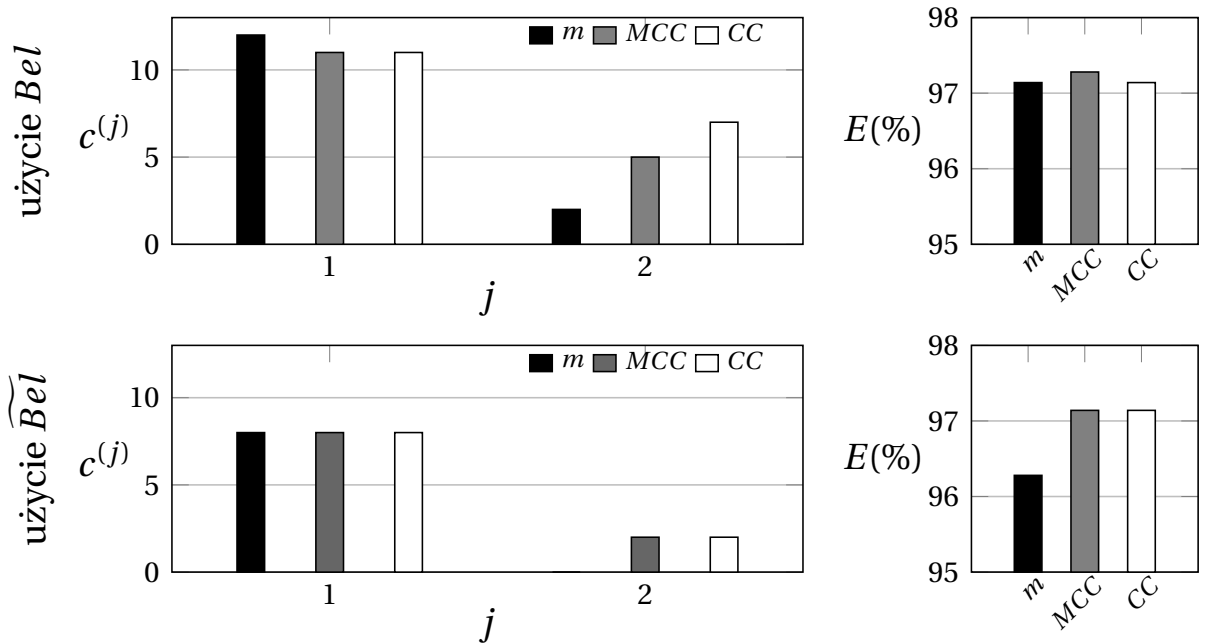
Rysunek 5.10: Porównanie wyników algorytmu selekcji reguł dla bazy *appendicitis* otrzymanych za pomocą dwóch sposobów obliczania miary przekonania



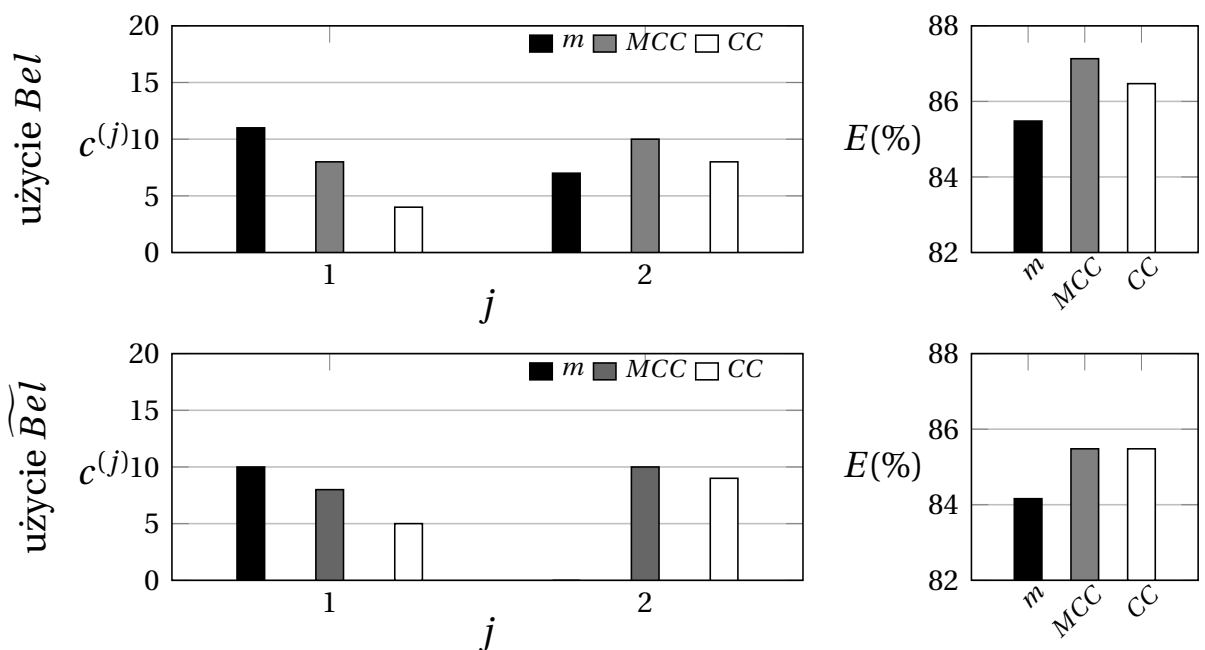
Rysunek 5.11: Porównanie wyników algorytmu selekcji reguł dla bazy *Pima Indians diabetes* otrzymanych za pomocą dwóch sposobów obliczania miary przekonania

Tabela 5.10: Procentowa zmiana efektywności uczenia (ΔE) i liczebności reguł diagnostycznych (ΔL) przy użyciu zmodyfikowanej miary przekonania (3.21) dla algorytmu selekcji reguł (ASR)

baza	<i>m</i>		<i>MCC</i>		<i>CC</i>	
	ΔE [%]	ΔL [%]	ΔE [%]	ΔL [%]	ΔE [%]	ΔL [%]
<i>thyroid disease</i>	-1.39	-85.4	0.00	-66.7	0.00	-46.70
<i>appendicitis</i>	-1.89	-66.7	-2.83	-50.0	-0.95	-20.00
<i>Pima Indians diabetes</i>	-1.40	-60.0	-2.21	+50.0	-0.91	-37.50
<i>Wisconsin breast cancer</i>	-0.86	-42.9	-0.14	-37.5	0.00	-44.40
<i>Cleveland heart disease</i>	-1.32	-44.4	-1.65	0.00	-0.99	+16.67



Rysunek 5.12: Porównanie wyników algorytmu selekcji reguł dla bazy *Wisconsin breast cancer* otrzymanych za pomocą dwóch sposobów obliczania miary przekonania



Rysunek 5.13: Porównanie wyników algorytmu selekcji reguł dla bazy *Cleveland heart disease* otrzymanych za pomocą dwóch sposobów obliczania miary przekonania

Należy pamiętać, że wyniki przedstawione na rysunkach 5.9-5.13 są osiągnięte dla progów niepewności dostrojonych tak aby w wyniku wydobywania uzyskać najwyższe wartości E . Dlatego warto zaznaczyć, że do osiągnięcia takiego wyniku dla miary Bel konieczne było strojenie progów η_e oraz η_b a dla miary \widetilde{Bel} tylko prog η_e dotyczącego ocen reguł (patrz punkt 3.5).

5.5 Wpływ warunkowej selekcji na wyniki wydobywania reguł

Ostatnim punktem tej części rozdziału będzie zbadanie, czy dodanie warunku usuwania reguł czyli zaproponowanie algorytmu warunkowej selekcji jest słuszne i czy ma ono znaczący wpływ na wynik wydobywania reguł. Aby ten punkt mógł stanowić podsumowanie dotychczasowych obserwacji dla każdej z baz benchmarkowych zostaną przedstawione wyniki dotyczące obu algorytmów selekcji (ASR i AWR) i obu metod wyznaczania miary przekonania (Bel i \widetilde{Bel}). Do oceny reguł zostaną kolejno użyte trzy używane dotąd wielkości (m , MCC oraz CC). Dla uproszczenia, wiedząc, że proponowane metody skutkują uzyskiwaniem prostych reguł (co zostało przedstawione w punkcie 5.3), w tabeli 5.11 wyszczególniono jedynie ich liczbę i uzyskaną efektywność.

Tabela 5.11: Procentowa efektywność diagnozy oraz liczebność reguł uzyskanych w wyniku algorytmów wydobywania reguł diagnostycznych ASR i AWR

baza	ocena	ASR, Bel	AWR, Bel	ASR, \widetilde{Bel}	AWR, \widetilde{Bel}
<i>thyroid disease</i>	m	97.67 (41)	97.21 (11)	96.28 (6)	96.28 (6)
	MCC	97.67 (18)	97.21 (12)	97.67 (6)	97.67 (6)
	CC	98.14 (15)	97.67 (12)	98.14 (8)	98.60 (11)
<i>appendicitis</i>	m	87.74 (12)	88.68 (8)	85.85 (4)	85.85 (4)
	MCC	87.74 (12)	87.74 (8)	84.91 (6)	86.79 (8)
	CC	90.57 (10)	90.57 (10)	89.62 (8)	89.62 (8)
<i>Pima Indians diabetes</i>	m	77.73 (10)	77.34 (8)	76.30 (4)	76.30 (4)
	MCC	76.95 (8)	76.95 (8)	74.74 (12)	74.22 (4)
	CC	77.60 (16)	77.08 (10)	76.69 (10)	76.04 (8)
<i>Wisconsin breast cancer</i>	m	97.14 (14)	96.85 (12)	96.28 (8)	96.28 (8)
	MCC	97.28 (16)	97.14 (12)	97.14 (10)	97.14 (10)
	CC	97.14 (18)	97.00 (8)	97.14 (10)	97.14 (10)
<i>Cleveland heart disease</i>	m	85.48 (18)	84.49 (10)	84.16 (10)	85.15 (10)
	MCC	87.13 (18)	85.48 (8)	85.48 (18)	84.82 (8)
	CC	86.47 (12)	85.81 (10)	85.48 (14)	84.82 (12)
wyniki średnie		89.50 (15.9)	89.15 (9.8)	88.39 (8.9)	88.45 (7.8)
mediana wyników		87.74 (15)	88.68 (10)	85.85 (8)	86.79 (8)

Analizując wyniki w tabeli 5.11 można zauważyć, że oprócz zmodyfikowanej metody obliczania \widetilde{Bel} , także algorytm warunkowej selekcji (AWR) pozwala ograniczyć liczbę reguł. Warunek sprawdzający wpływ dodanego elementu ogniskowego na poprawienie wyniku diagnozy okazuje się przydatny przy wydobywaniu reguł z benchmarkowych zbiorów uczących. Niemniej jednak, szczególną rolę w ograniczeniu liczby reguł pełni użyta modyfikacja miary przekonania \widetilde{Bel} . Choć średnie efektywności obliczone dla każdej kolumny tabeli 5.11 są zbliżone, to dzięki wspólnemu użyciu algorytmu warunkowej selekcji reguł i miary przekonania \widetilde{Bel} uzyskano najmniej złożone zbiory reguł diagnostycznych. Średnia efektywność uczenia jest o jeden procent mniejsza niż dla wyniku oznaczonego jako „ASR, Bel ” ale średnia liczba reguł jest o połowę mniejsza. Mediana wyników także potwierdza słuszność użycia algorytmu warunkowej selekcji reguł a także zmodyfikowanej miary przekonania.

5.6 Wybrane metody weryfikacji zdolności uogólniania

Uzyskane dotychczas wyniki pokazują, że dla omawianych baz benchmarkowych proponowane metody wydobywania reguł pozwalają uzyskiwać mało liczne zbiory reguł diagnostycznych. Wykorzystanie całej możliwej informacji czyli uwzględnienie wszystkich przypadków w bazie podczas wydobywania reguł, pozwala lepiej opisać dany problem medyczny. Możliwość wydobywania reguł, które są czytelne dla eksperta z dziedziny medycznej, należy już uznać za pewien sukces opracowanych metod. Jednakże, mając na uwadze wspomaganie diagnozy, koniecznym jest sprawdzenie jak zaproponowane metody radzą sobie w przypadku rozpoznawania danych nieznanymi.

Do zbadania zdolności uogólniania metod używanych w klasyfikacji danych, stosuje się m.in. procedury zwane sprawdzianami krzyżowymi (ang. *cross-validation*) [7]. Polegają one na podziale zbioru danych na zbiór uczący i testowy. Jako \mathbf{X} oznaczony zostanie pełny zbiór danych. Zawiera on oczywiście informację o wszystkich pacjentach danej bazy medycznej czyli zbiory $\mathbf{X}^{(l)}$ gdzie $l = 1, \dots, C$. Jako \mathbf{U} i \mathbf{T} oznaczone zostaną zbiory: uczący i testowy. W niniejszej rozprawie zostaną przeprowadzone dwie procedury określane wspólnym mianem metod niewyczerpujących: metoda rotacji z walidacją oraz metoda K -krotnej walidacji.

Podstawowa metoda rotacji polega na wykonaniu K losowych podziałów zbioru \mathbf{X} na $\mathbf{U}^{(k)}$ i $\mathbf{T}^{(k)}$, $k = 1, \dots, K$ z zachowaniem następujących warunków [55, 70]:

$$\mathbf{U}^{(k)} \cup \mathbf{T}^{(k)} = \mathbf{X}, \quad \mathbf{U}^{(k)} \cap \mathbf{T}^{(k)} = \emptyset \quad (5.1)$$

oraz

$$|\mathbf{U}^{(k)}| = \alpha|\mathbf{X}|, \quad |\mathbf{T}^{(k)}| = (1 - \alpha)|\mathbf{X}|, \quad (5.2)$$

gdzie $\alpha \in (0, 1)$. Zależności w (5.1) określają, że zbiór danych dzielony jest na wykluczające się zbiory uczący i testowy. Zależności w (5.2) pokazują, że podział zbioru danych na uczący i testowy definiuje parametr α . Zwykle, aby najlepiej badać zdolność uogólniania, należy przyjąć $\alpha \leq 0.5$, choć z punktu widzenia wydobywania wiedzy, ilość informacji z której wydo-

bywa się reguły, pomniejsza się. Dla każdego z K podziałów, wykonuje się procedurę uczenia (czyli w niniejszej rozprawie wydobywania reguł) na podstawie zbioru $\mathbf{U}^{(k)}$. Wynik uczenia charakteryzuje m.in. średnia efektywność uczenia obliczana jako

$$\bar{E}_U = \frac{1}{K} \sum_{k=1}^K E_U^{(k)}, \quad (5.3)$$

gdzie $E_U^{(k)}$ to efektywność (4.2) obliczona na podstawie wyniku uczenia w k -tej iteracji procedury czyli efektywność diagnozowania pacjentów ze zbioru $\mathbf{U}^{(k)}$. Po etapie uczenia testuje się zdolność uogólniania wydobytych reguł diagnozując pacjentów w zbiorze $\mathbf{T}^{(k)}$. W procedurze rotacji liczba podziałów (K) jest różna: 5, 10, 50 a nawet 100 [70]. Po jej wykonaniu o zdolności uogólniania świadczy średnia efektywność diagnozowania przypadków ze zbiorów testowych

$$\bar{E}_T = \frac{1}{K} \sum_{k=1}^K E_T^{(k)}, \quad (5.4)$$

gdzie $E_T^{(k)}$ to efektywność (4.2) obliczona na podstawie wyniku testu w k -tej iteracji procedury czyli efektywność diagnozowania pacjentów ze zbioru $\mathbf{T}^{(k)}$.

Z uwagi na to, że większość metod klasyfikacji oraz zaproponowana metoda wspomaganie diagnozy, zależą od doboru parametrów, przeprowadza się tzw. rotację z walidacją [55]. Różni się ona od standardowej rotacji tylko tym, że przeprowadza się najpierw np. $0.1 \cdot K$ iteracji i dla tych $0.1 \cdot K$ podziałów na podstawie wyniku uczenia i testu dostraja się wartości parametrów metody. W przypadku proponowanej metody są to progi niepewności η_e i η_b (druga wartość jest strojona tylko dla tradycyjnej miary przekonania (3.15)). Po wybraniu najlepszych parametrów (jako np. średnie z $0.1 \cdot K$ wybranych wartości) przeprowadza się metodę rotacji dla wszystkich K podziałów. Choć nic nie stoi na przeszkodzie aby dostrajać progi η_e i η_b w każdej procedurze uczenia, to chociażby dla ograniczenia obliczeń zastosowano wspomniano metodę rotacji z walidacją.

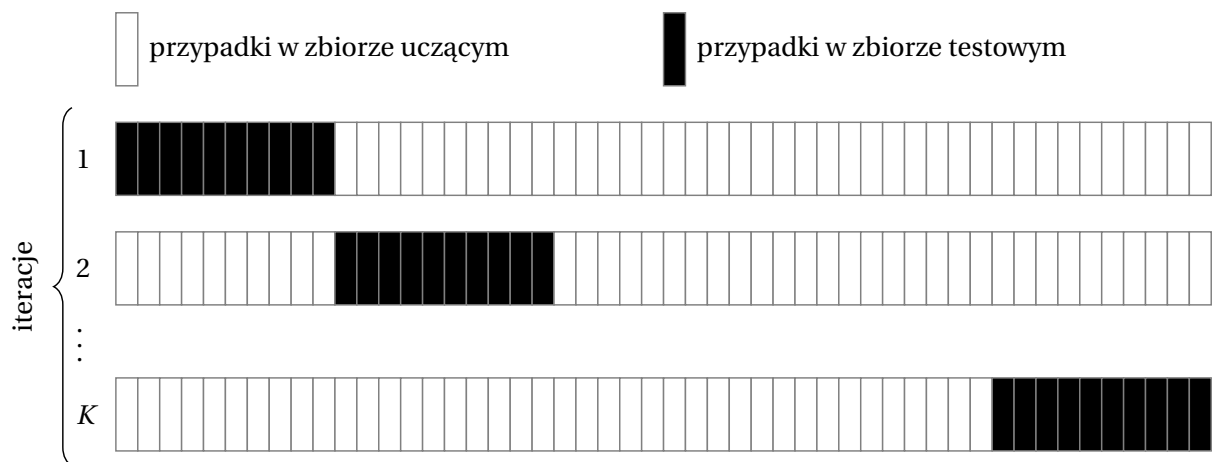


Rysunek 5.14: Schemat procedury rotacji ($\alpha = 0.5$)

Inna metoda polegająca na losowym podziale to metoda K -krotnej walidacji (ang. K -fold cross-validation). W tej metodzie zbiór X dzieli się na K możliwie równolicznych i wzajemnie wykluczających się zbiorów $F^{(k)}$:

$$|F^{(k)}| \cong \frac{1}{K} |X|, \quad \bigcup_{1 \leq k \leq K} F^{(k)} = X, \quad \forall_{\substack{1 \leq k \leq K \\ 1 \leq k' \leq K \\ k' \neq k}} F^{(k)} \cap F^{(k')} = \emptyset. \quad (5.5)$$

W każdej iteracji wybiera się jeden zbiór $F^{(k)}$, który jest użyty jako zbiór danych testowych, a pozostałe $K - 1$ zbiorów wybiera się jako zbiór uczący. Najczęściej $K = 5, 10$ [83]. Należy jednak zauważyć, że liczebność danych w zbiorze uczącym bezpośrednio zależy od przyjętej wartości K . Dla $K = 5$ wydobywanie reguł byłoby przeprowadzone za każdym razem na podstawie 80% danych, a dla $K = 10$ na podstawie 90% danych. Zostanie ona użyta w tym punkcie z uwagi na to, że jest to polecana [63] i w zasadzie najpopularniejsza metoda porównywania metod klasyfikacji i wspomaganie diagnozy medycznej [34, 56, 62, 67, 95, 128]. Jest ona także dostępna w środowiskach przeznaczonych do badań we wspomnianych obszarach [4, 40, 73]. Poglądowe schematy na rysunkach 5.14 i 5.15 przedstawiają realizację poszczególnych metod.



Rysunek 5.15: Schemat procedury K -krotnej walidacji ($K = 5$)

Prezentując użyte bazy benchmarkowe na początku rozdziału w tabeli 5.1 można było zauważyć, że mamy do czynienia z zbiorami w których procentowy udział diagnoz w bazie nie jest zachowany. Dlatego dla obu metod badania zdolności uogólniania zbiory uczące i testowe będą generowane losowo, ale będzie zachowana w nich taka sama proporcja diagnoz jak w pełnym zbiorze. Taką procedurę w literaturze określa się mianem *stratified cross-validation* [100]. Metodę rotacji z walidacją przeprowadzono w celu porównania ze sobą proponowanych metod a metoda k -krotnej walidacji do porównania proponowanych metod z wybranymi rozwiązaniami referencyjnymi.

5.7 Wpływ metody oceny rozmytych elementów ogniskowych na jakość wspomagania diagnozy oraz złożoność reguł diagnostycznych

W niniejszej rozprawie proponowane i badane są trzy algorytmy wydobywania reguł, trzy metody oceny oraz dwie metody obliczania miary przekonania. Na podstawie procedury rotacji z walidacją opisaną w punkcie 5.6, sprawdzone zostanie, która z metod oceny rozmytych elementów ogniskowych (m , MCC , CC) najlepiej łączy zdolność uogólniania ze złożonością reguł diagnostycznych. Do oceny kompromisu pomiędzy efektywnością diagnozy a złożonością reguł pomocny będzie wskaźnik QI (4.16). W tabeli 5.12 zaprezentowano wyniki otrzymane dla kolejnych baz medycznych w wyniku przeprowadzenia procedury rotacji z walidacją: dla każdego algorytmu wydobywania reguł i każdej metody wyznaczania miary przekonania. Główna uwaga jest skupiona na porównaniu średnich efektywności i liczebności rozmytych elementów ogniskowych uzyskanych na podstawie każdej z metod oceny. Tym samym w tabeli 5.12 można wyróżnić trzy części: lewa część tabeli dotyczy wyników otrzymanych przy użyciu wartości bazowego prawdopodobieństwa m w ocenie elementów ogniskowych, środkowa część – wartości MCC , prawa część - zaproponowanej wartości CC . Warto zauważyć, że jeden wiersz tabeli 5.12 dotyczy wyników tego samego algorytmu wydobywania reguł i tej samej bazy medycznej oraz wyników obliczeń przeprowadzonych z użyciem kolejnej metody oceny. Wartości \bar{E}_T są obliczane jak zdefiniowano (5.4), a wartości $\theta^{(j)}$ (4.15) są także zagregowane za pomocą wartości średniej ze 100 wyników z każdej iteracji. W przypadku procentowych średnich efektywności testu (5.4), wartości te zaokrąglono do setnych części procenta, a liczebności rozmytych elementów ogniskowych (4.15) zaokrąglono do liczb całkowitych.

Porównując osiągnięte wyniki dla trzech metod oceny, w każdym wierszu pogrubiono najlepszą średnią efektywność \bar{E}_T (5.4). Nietrudno zauważyć, że średnie efektywności występujące w jednym wierszu czasem różnią się one o małą wartość. Dlatego pogrubiono te wyniki, które dla danej metody oceny są najwyższe oraz różnią się statystycznie istotnie od pozostałych dwóch wyników (co sprawdzono metodą testowania różnic par średnich [115]). Można zauważyć, że jedynie najwyższy wynik osiągnięty dla oceny związanej z CC przedstawiony w wierszu pierwszym nie różnił się istotnie od pozostałych. Poza nim, wszystkie wyróżnione maksymalne wyniki przeszły pozytywnie test istotności różnic. Widać jednoznacznie, że najwyższa efektywność jest osiągana najczęściej dla oceny związanej z wprowadzonym wskaźnikiem CC (3.42). Wartość bazowego prawdopodobieństwa tylko pięciokrotnie osiągnęła najwyższą efektywność testu a wskaźnik MCC sześciokrotnie. Pozostałe 18 najlepszych wyników przypisuje się wskaźnikowi CC . W dole tabeli uwzględniono średnie i mediany z wszystkich wyników. Potwierdzają one powyższe obserwacje. Średnia efektywność osiągnięta dla wszystkich algorytmów jest prawie taka sama dla oceny związanej z wartościami m oraz MCC a dla CC jest największa. Mediana wyników także różnicuje wyniki na korzyść metody oceny związanej z wartością CC .

Tabela 5.12: Średnie efektywności uogólniania oraz złożoność reguł diagnostycznych uzyskiwanych w procedurach rotacji z walidacją ($K = 100$, $\alpha = 0.5$) dla wszystkich badanych baz benchmarkowych

metoda	baza	metoda oceny														
		m					MCC					CC				
		$\bar{E}_T[\%]$	$\theta^{(1)}$	$\theta^{(2)}$	$\theta^{(3)}$	$\theta^{(4)}$	$\bar{E}_T[\%]$	$\theta^{(1)}$	$\theta^{(2)}$	$\theta^{(3)}$	$\theta^{(4)}$	$\bar{E}_T[\%]$	$\theta^{(1)}$	$\theta^{(2)}$	$\theta^{(3)}$	$\theta^{(4)}$
AER, Bel	TH	91.05	9	7	3	1	91.54	8	9	4	1	92.51	5	9	4	1
	AP	79.13	9	13	12	6	78.44	3	4	4	2	81.71	3	2	2	2
	PI	72.18	9	4	3	2	72.6	4	5	2	1	73.72	2	5	3	1
	BC	93.67	9	4	1	0	93.59	3	4	2	0	93.21	3	2	1	0
	HD	82.26	17	38	44	35	83.41	20	119	321	412	82.36	7	33	47	29
AER, \widetilde{Bel}	TH	94.61	7	6	3	1	93.93	7	5	1	0	95.79	6	7	3	0
	AP	80.13	6	8	8	4	81.85	2	3	2	2	85.87	2	1	1	2
	PI	72.74	7	2	1	0	73.45	3	4	2	1	73.98	2	3	1	0
	BC	94.08	6	2	1	0	95.62	1	5	5	2	94.4	2	1	1	0
	HD	82.15	17	43	51	35	83.23	20	120	330	435	82.39	7	32	49	32
ASR, Bel	TH	91.92	8	1	0	0	88.07	5	6	1	0	88.85	7	3	0	0
	AP	80.12	8	3	0	0	82.15	4	6	3	1	84.90	4	6	6	3
	PI	72.61	8	0	0	0	73.94	3	5	1	0	74.46	3	5	3	0
	BC	96.06	10	3	1	1	95.33	6	6	2	0	94.81	5	7	4	2
	HD	81.23	11	1	0	0	80.88	6	5	0	0	80.44	4	10	2	0
ASR, \widetilde{Bel}	TH	91.19	5	0	0	0	93.74	1	3	1	0	95.17	6	2	0	0
	AP	81.73	5	3	2	1	82.46	3	4	4	3	85.29	3	3	4	2
	PI	72.38	5	0	0	0	73.37	3	3	0	0	74.2	2	3	1	0
	BC	94.94	8	2	0	0	95.29	4	3	1	0	95.53	3	4	1	0
	HD	81.03	9	1	0	0	81.07	5	4	0	0	78.75	2	7	2	0
AWR, Bel	TH	91.5	8	1	0	0	91.52	7	2	0	0	92.74	7	3	0	0
	AP	77.35	7	1	0	0	78.58	3	3	1	0	83.42	3	3	2	0
	PI	72.38	7	0	0	0	73.82	3	4	1	0	74.23	2	4	2	0
	BC	95.11	8	1	0	0	95.53	6	2	0	0	95.46	5	3	1	0
	HD	80.05	9	0	0	0	80.2	5	3	0	0	77.79	3	4	0	0
AWR, \widetilde{Bel}	TH	91.17	5	0	0	0	93.46	2	3	1	0	95.36	5	3	0	0
	AP	81.73	5	1	0	0	82.02	2	2	1	0	84.87	2	2	1	0
	PI	72.67	5	0	0	0	73.31	3	3	0	0	74.24	2	3	1	0
	BC	94.97	7	1	0	0	95.4	3	3	0	0	95.46	3	3	1	0
	HD	80.66	8	0	0	0	80.64	5	3	0	0	77.63	2	5	0	0
średnie	84.09	8	5	4	3	84.61	5	12	23	29	85.32	4	6	5	2	
mediany	81.73	8	1	0	0	82.31	4	4	1	0	84.89	3	3	1	0	

Teraz zostanie przeanalizowana złożoność wydobywanych rozmytych elementów ogniskowych (od $\theta^{(1)}$ do $\theta^{(4)}$). Poza algorytmami eliminacji reguł (wyniki: „AER, Bel” i „AER, \widetilde{Bel} ”), najczęściej wydobywano od kilku do kilkunastu reguł. Wśród nich rzadko otrzymywano rozmyte elementy ogniskowe o więcej niż trzech warunkach logicznych (wskazuje na to fakt, że znaczna część liczb w kolumnach opisanych nagłówkiem $\theta^{(4)}$ jest równa zero). Ocenę złożoności szczególnie dobrze ilustruje mediana przedstawiona w dole tabeli. Jest ona odporna na kilka odstających wartości (z bardzo dużą liczbą reguł) i trafniej prezentuje różnice pomiędzy użytymi metodami ocen.

Należy nadmienić, że poza możliwością wybrania najlepszej metody oceny, czasem występujące subtelne różnice w wynikach, wynikają przede wszystkim z faktu, że podczas procedury rotacji może się zdarzyć, że dla określonego zbioru uczącego zostaną wydobyte te same reguły osiągające podobne wyniki bez względu na wybraną metodę oceny. Niemniej jednak, wielokrotne osiąganie przez wskaźnik *CC* wyniku lepszego i istotnie różniącego się od pozostałych wypada uznać za dobrą przesłankę do wybrania go jako wskaźnik najlepszy.

Tabela 5.13 ponownie opisuje wyniki wykonywanych obliczeń, jednak tym razem celem podsumowania jest wskazanie średniej liczby reguł oraz zgodności z kompromisem liczba reguł-poprawność diagnozy. Ten ostatni ilustruje wartość wskaźnika *QI* (4.16). Średnie liczebności wydobytych reguł są obliczane jako

$$\bar{L} = \sum_{k=1}^K L^{(k)}, \quad (5.6)$$

gdzie $L^{(k)}$ to liczebność zbioru reguł (4.13) obliczona dla k -tej iteracji w procedurze rotacji. Na podstawie średniej efektywności i średniej liczebności reguł obliczona jest wartość wskaźnika *QI*. Użycie wartości bazowego prawdopodobieństwa tylko dwukrotnie pozwoliło osiągnąć najlepszy wynik, biorąc pod uwagę kompromis pomiędzy zdolnością uogólniania a złożonością zbiorów reguł diagnostycznych. Ocena związana z *MCC* uzyskała sześciokrotnie najlepszy wynik, ale to ocena związana z *CC* ponownie uzyskuje najczęściej najlepszy wynik, tj. najniższe wartości *QI*. Mediana liczebności reguł w dole tabeli 5.13 wskazuje, że są one do siebie bardzo zbliżone a to wpływa na mniej znaczne różnice w medianach wartości wskaźnika *QI*. Jednakże, uwzględniając osiągnięcie minimalnych wartości *QI* dla metody oceny związanej z *CC* i wcześniejsze wnioski dotyczące średnich efektywności, można twierdzić, że słusznym jest wprowadzenie wskaźnika *CC* w celu poprawy jakości wspomaganie diagnozy. Dla testowanych baz wykazano, że wskaźnik *QI* dobrze definiuje kompromis pomiędzy zdolnością uogólniania a złożonością reguł diagnostycznych.

5.8 Porównanie proponowanych algorytmów wydobywania reguł diagnostycznych

W tym punkcie zaprezentowane zostaną ponownie wyniki rotacji z walidacją ($K = 100$) ale będą one przedstawione w odmienny sposób. Przede wszystkim zostanie zbadany wpływ

Tabela 5.13: Średnie wyniki efektywności, liczebności reguł oraz wskaźnika QI w procedurach rotacji z walidacją ($K = 100$, $\alpha = 0.5$) dla wszystkich badanych baz benchmarkowych

metoda	baza	średnie wyniki								
		$\bar{E}_T[\%]$			\bar{L}			QI		
		m	MCC	CC	m	MCC	CC	m	MCC	CC
AER, Bel	TH	91.05	91.54	92.51	20	22	19	0.312	0.329	0.286
	AP	79.13	78.44	81.71	40	13	9	0.461	0.298	0.245
	PI	72.18	72.60	73.72	18	12	11	0.362	0.323	0.303
	BC	93.67	93.59	93.21	14	9	6	0.106	0.090	0.086
	HD	82.26	83.41	82.36	134	872	116	0.246	0.574	0.233
AER, \widetilde{Bel}	TH	94.61	93.93	95.79	17	13	16	0.232	0.205	0.209
	AP	80.13	81.85	85.87	26	9	6	0.358	0.233	0.188
	PI	72.74	73.45	73.98	10	10	6	0.325	0.306	0.282
	BC	94.08	95.62	94.40	9	13	4	0.087	0.071	0.068
	HD	82.15	83.23	82.39	146	905	120	0.254	0.591	0.235
ASR, Bel	TH	91.92	88.07	88.85	9	12	10	0.192	0.253	0.211
	AP	80.12	82.15	84.90	11	14	19	0.297	0.266	0.275
	PI	72.61	73.94	74.46	8	9	11	0.323	0.301	0.296
	BC	96.06	95.33	94.81	15	14	18	0.090	0.084	0.093
	HD	81.23	80.88	80.44	12	11	16	0.198	0.199	0.205
ASR, \widetilde{Bel}	TH	91.19	93.74	95.17	5	5	8	0.144	0.118	0.137
	AP	81.73	82.46	85.29	11	14	12	0.265	0.258	0.224
	PI	72.38	73.37	74.20	5	6	6	0.310	0.291	0.283
	BC	94.94	95.29	95.53	10	8	8	0.084	0.069	0.066
	HD	81.03	81.07	78.75	10	9	11	0.198	0.196	0.219
AWR, Bel	TH	91.50	91.52	92.74	9	9	10	0.185	0.185	0.184
	AP	77.35	78.58	83.42	8	7	8	0.299	0.271	0.222
	PI	72.38	73.82	74.23	7	8	8	0.322	0.296	0.289
	BC	95.11	95.53	95.46	9	8	9	0.084	0.072	0.071
	HD	80.05	80.20	77.79	9	8	7	0.208	0.204	0.227
AWR, \widetilde{Bel}	TH	91.17	93.46	95.36	5	6	8	0.155	0.132	0.135
	AP	81.73	82.02	84.87	6	5	6	0.234	0.211	0.193
	PI	72.67	73.31	74.24	5	6	6	0.307	0.292	0.279
	BC	94.97	95.40	95.46	8	6	7	0.078	0.066	0.063
	HD	80.66	80.64	77.63	8	8	7	0.201	0.199	0.228
średnie		84.093	84.615	85.318	20	68	17	0.231	0.233	0.201
mediany		81.730	82.305	84.885	10	9	9	0.233	0.222	0.221

trzech opracowanych algorytmów na zdolność uogólniania uzyskanych reguł i ich złożoność. Dodatkowo, analizie zostanie poddana miara przekonania \widetilde{Bel} (3.21). Wyniki zostaną tym razem przedstawione osobno dla poszczególnych baz benchmarkowych. W poprzednim punkcie metoda oceny reguł związana z wartością CC została wybrana jako najlepsza. Dlatego obecnie będą omówione tylko wyniki związane z tym wskaźnikiem.

W tabeli 5.14 przedstawiono wartości średnie dotyczące efektywności testu, liczby reguł oraz obliczony na ich podstawie wskaźnik QI . Najbardziej widoczną różnicą w tabeli jest liczba reguł uzyskiwana dla dwóch metod obliczania miary przekonania. Dla każdego algorytmu oraz każdej bazy użycie \widetilde{Bel} pozwala zmniejszyć liczbę reguł (nawet o ponad 50% jak dla bazy *Wisconsin breast cancer* przy algorytmie selekcji reguł). Większość wyników osiągniętych dla \widetilde{Bel} ma także większą efektywność. Sprawia to, że wartość wskaźnika QI także minimalizowana jest dla \widetilde{Bel} . Jedynie dla bazy *Cleveland heart disease*, kiedy liczba reguł zostaje pomniejszona dzięki użyciu zmodyfikowanej miary przekonania dla algorytmu selekcji (ASR), średnia efektywność diagnozy spada. Taki wynik sugeruje, że osiągnięcie rozwiązań kompromisowych (na których skupiają się proponowane metody) w przypadku tej bazy jest szczególnie trudne. Potwierdzą to także wyniki rozwiązań referencyjnych zaprezentowane w dalszej części rozdziału.

Skupiając uwagę na algorytmach wydobywania reguł, można zauważyć, że niemożliwe jest jednoznaczne wybranie najlepszego algorytmu jako rozwiązania osiągającego najwyższą wartość efektywności. Inaczej jest kiedy do oceny wyniku zostanie włączona liczba reguł i obliczona wartość wskaźnika QI . Jego wartości są najmniejsze najczęściej dla algorytmów warunkowej selekcji reguł (AWR), zarówno dla miary przekonania Bel , jak i \widetilde{Bel} . Niemniej jednak, użycie \widetilde{Bel} pozwala niejednokrotnie zmniejszyć wartość QI (jest to widoczne dla algorytmu selekcji reguł dla baz *thyroid disease*, *Pima Indians diabetes*, *Wisconsin breast cancer*).

Inne spojrzenie na wyniki procedury rotacji, przedstawione w tabeli 5.14, umożliwiają rozkłady efektywności (rysunki 5.16a-5.20a), liczebności reguł (rysunki 5.16b-5.20b) oraz wartości QI (rysunki 5.16c-5.20c) zilustrowane za pomocą wykresu typu „skrzynka z wąsami” (ang. *box and whiskers*). Pozioma linia zawarta w „skrzynkach” określa wartość mediany otrzymanych wyników. Dolna/górna granica „skrzynki” określa wartości dolnego/górnego kwartyla. Dolna/górna linia pozioma na końcach tzw. „wąsów” wyznaczona jest przez wynik minimalny/maksymalny nie będący wartością odstającą (tzn. wynik minimalny/maksymalny znajdujący się w odległości co najwyżej równej półtora rozstępu międzykwartylowego od dolnego/górnego kwartyla). Pozostałe punkty reprezentowane przez symbol * to właśnie wartości odstające.

Dla baz *thyroid disease*, *appendicitis* oraz *Wisconsin breast cancer* szczególnie widoczna jest poprawa rozkładu efektywności z jednoczesnym zmniejszeniem liczebności reguł dla algorytmów selekcji i selekcji warunkowej przy zastosowaniu zmodyfikowanej miary przekonania \widetilde{Bel} . Dla baz *Pima Indians diabetes* oraz *Cleveland heart disease* poprawa efektywności nie jest dostrzegalna, ale za to ujawnia się zmniejszenie i stabilność liczby wydobywanych reguł dla algorytmu warunkowej selekcji (zarówno dla miary Bel jak i \widetilde{Bel}). Z największą liczbą wyników odstających od reszty mamy szczególnie do czynienia jeśli chodzi o liczbę wydoby-

wanych reguł dla algorytmów eliminacji (dla obu metod obliczania miary przekonania). Dla bazy *Cleveland heart disease* maksymalna wartość na osi prezentującej liczbę reguł została ograniczona do 500 aby zachować czytelność wykresu, choć wartości odstające większe niż 500 wystąpiły dla obu wyników związanych z algorytmem eliminacji reguł.

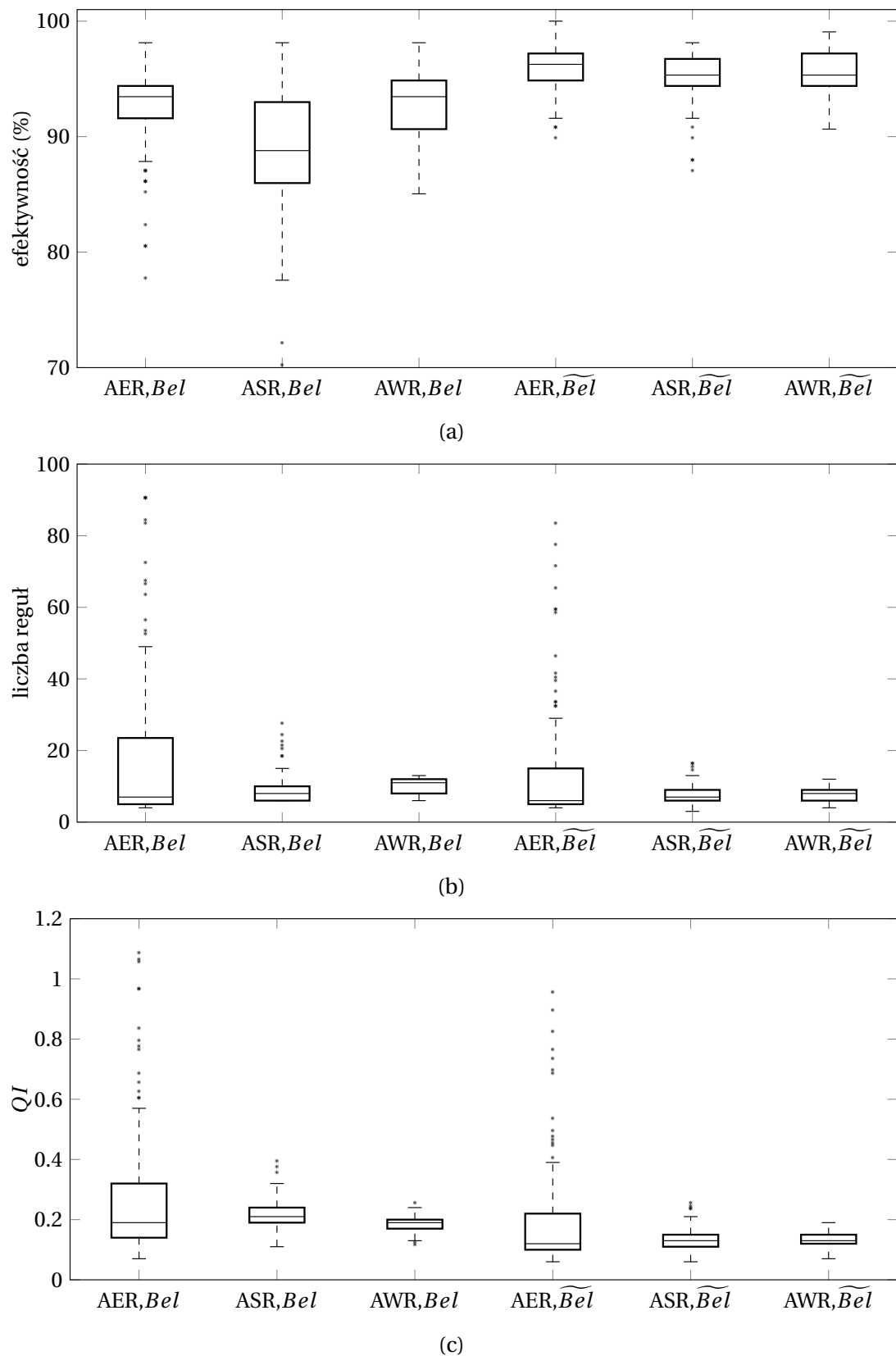
Tabela 5.14: Średnie efektywności, liczebności złożonych elementów ogniskowych oraz wartości wskaźnika QI uzyskiwane w procedurze rotacji z walidacją przedstawione osobno dla każdej z baz benchmarkowych

algorytm	miara przekonania					
	Bel			\widetilde{Bel}		
	$\overline{E_T}$	\overline{L}	QI	$\overline{E_T}$	\overline{L}	QI
baza <i>thyroid disease</i>						
AER	92.51	19	0.286	95.79	16	0.209
ASR	88.85	10	0.211	95.17	8	0.137
AWR	92.74	10	0.184	95.36	8	0.135
baza <i>appendicitis</i>						
AER	81.71	9	0.245	85.87	6	0.188
ASR	84.90	19	0.275	85.29	12	0.224
AWR	83.42	8	0.222	84.87	5	0.193
baza <i>Pima Indians diabetes</i>						
AER	73.72	11	0.303	73.98	6	0.282
ASR	74.46	11	0.296	74.20	6	0.283
AWR	74.23	8	0.289	74.24	6	0.279
baza <i>Wisconsin breast cancer</i>						
AER	93.21	6	0.086	94.40	4	0.068
ASR	94.81	18	0.093	95.53	8	0.066
AWR	95.46	9	0.071	95.46	7	0.063
baza <i>Cleveland heart disease</i>						
AER	82.36	116	0.233	82.39	120	0.235
ASR	80.44	16	0.205	78.75	11	0.219
AWR	77.79	7	0.227	77.63	7	0.228

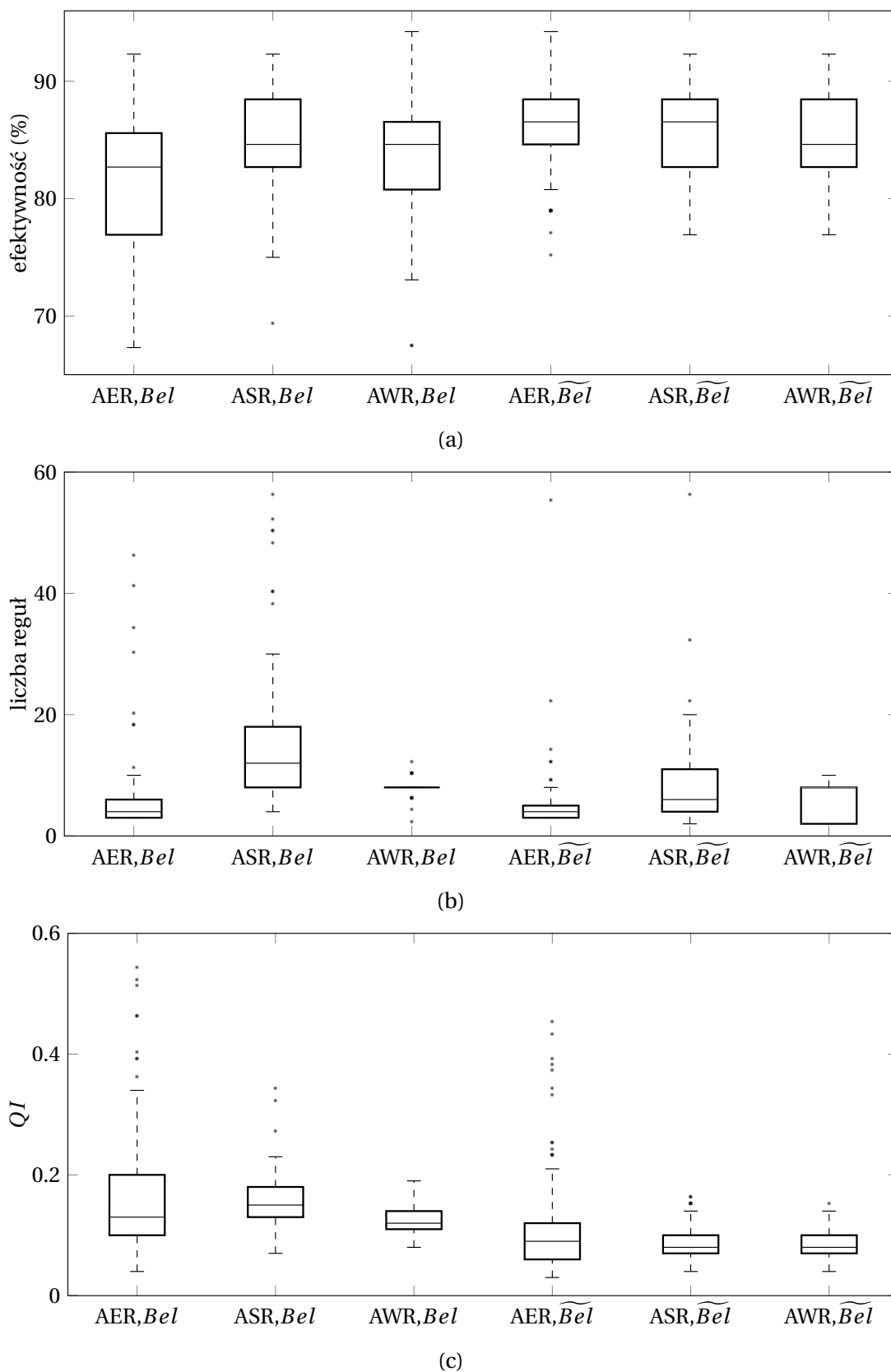
O pozytywnym rezultacie użycia algorytmów selekcji oraz zmodyfikowanej miary przekonania \widetilde{Bel} (3.21) świadczy także rozkład wartości QI (4.16) przedstawiony również na rysunkach 5.16c-5.20c. Najniżej położone rozkłady wartości QI wiążą się z najlepszymi rozwiązaniami. W przypadku bazy *thyroid disease* mamy do czynienia z wysoką efektywnością (czyli niewielką wartością ilorazu $\frac{N_c}{N}$), dlatego też rozkład QI jest bardzo podobny do rozkładu liczby reguł, bo związany jest głównie z ilorzem $\frac{L}{\sum_{k=1}^w \binom{w}{k}}$ występującym w (4.16). W przypadku reszty baz, QI lepiej łączy oba wspomniane składniki. Dla baz *appendicitis*, *Pima Indians diabetes* oraz *Wisconsin breast cancer* rozkłady otrzymane dla algorytmów selekcji i warunkowej selekcji (przy użyciu \widetilde{Bel}) są położone najniżej. Dla bazy *Cleveland heart disease* algorytm eliminacji przy użyciu \widetilde{Bel} osiągnął wynik konkurencyjny do obu algorytmów selekcji.

Niemniej jednak interpretacja rozkładów QI pozwala znaleźć algorytmy najbardziej stabilne i osiągające najlepsze rezultaty pod względem badanego kompromisu liczebności reguł i efektywności.

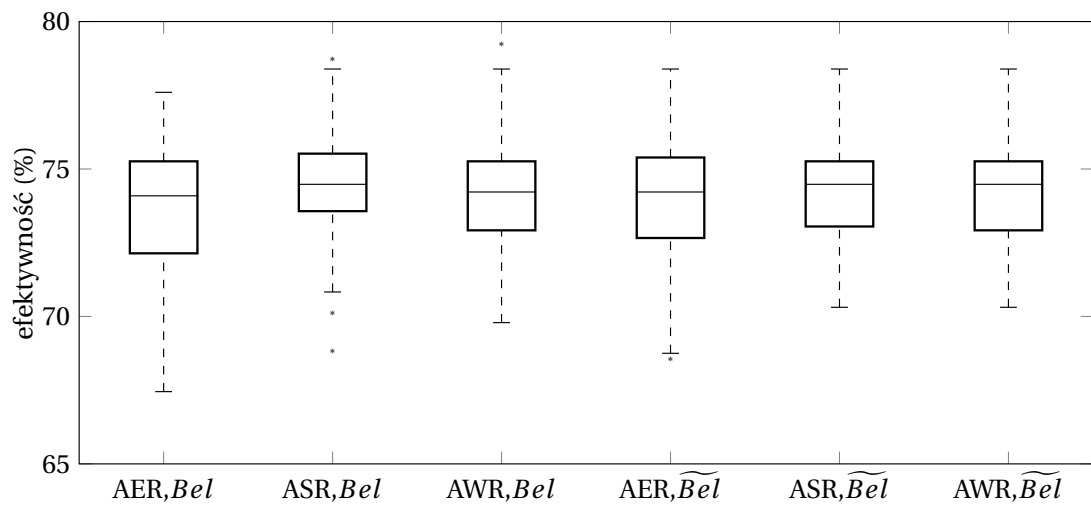
Na podstawie przedstawionych wyników wydaje się słusznym stwierdzenie, że jakość wydobywanych reguł diagnostycznych oceniona jako zdolność uogólniania i złożoność reguł diagnostycznych, jest lepsza dla algorytmu warunkowej selekcji (AWR) połączonej ze zmodyfikowaną miarą przekonania \widetilde{Bel} niż dla pozostałych algorytmów.



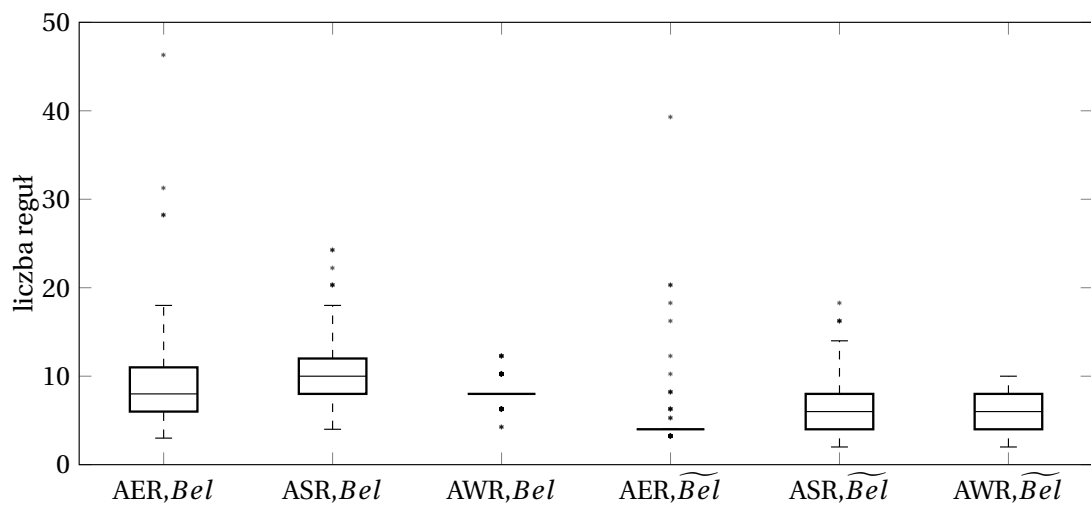
Rysunek 5.16: Wynik procedury rotacji dla bazy *thyroid disease* w postaci wykresu „skrzynka z wąsami” dla badanych algorytmów przy zastosowaniu wartości CC jako wskaźnika oceny reguł: (a) rozkład efektywności, (b) rozkład liczebności reguł, (c) rozkład wartości QI



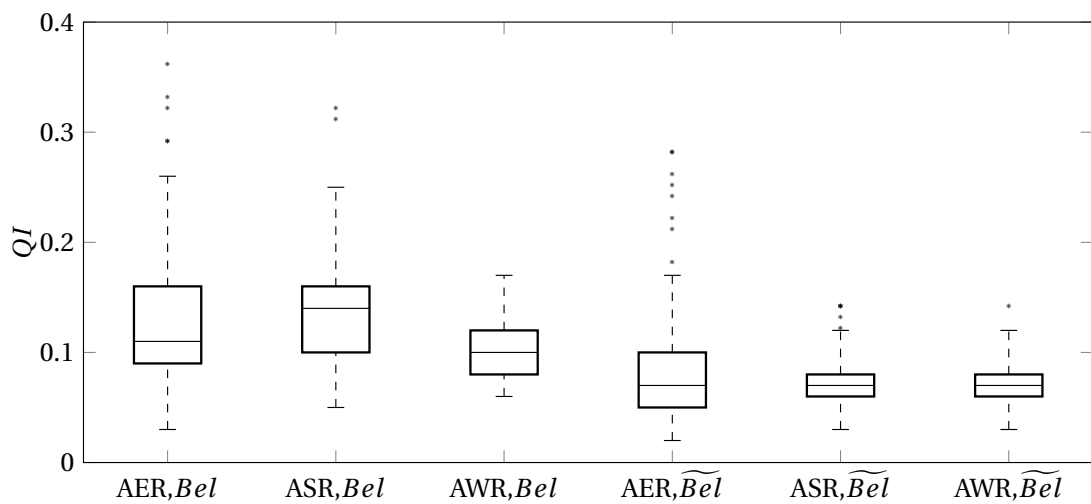
Rysunek 5.17: Wynik procedury rotacji dla bazy *appendicitis* w postaci wykresu „skrzynka z wąsami” dla badanych algorytmów przy zastosowaniu wartości CC jako wskaźnika oceny reguł: (a) rozkład efektywności, (b) rozkład liczebności reguł, (c) rozkład wartości QI



(a)

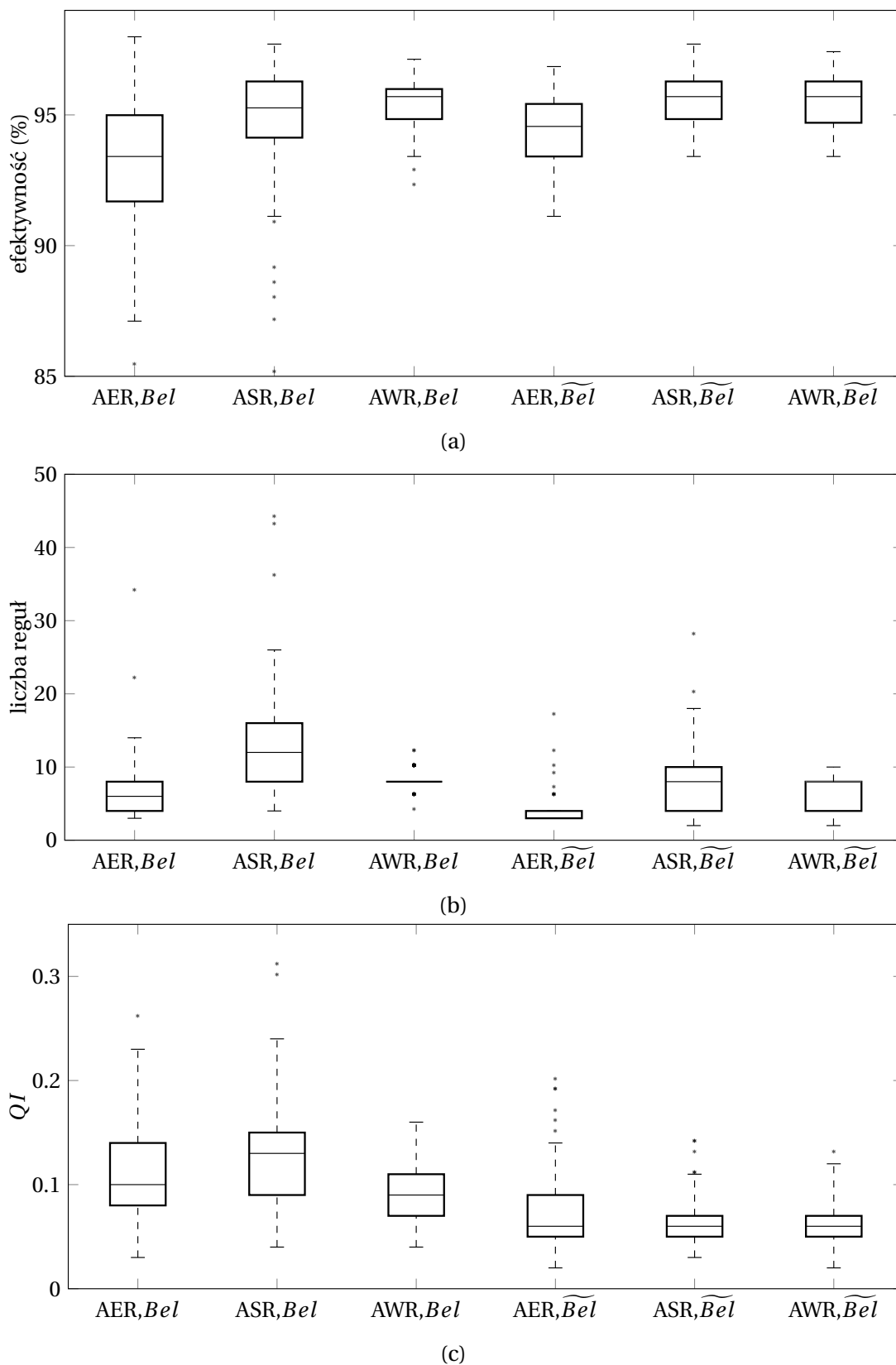


(b)

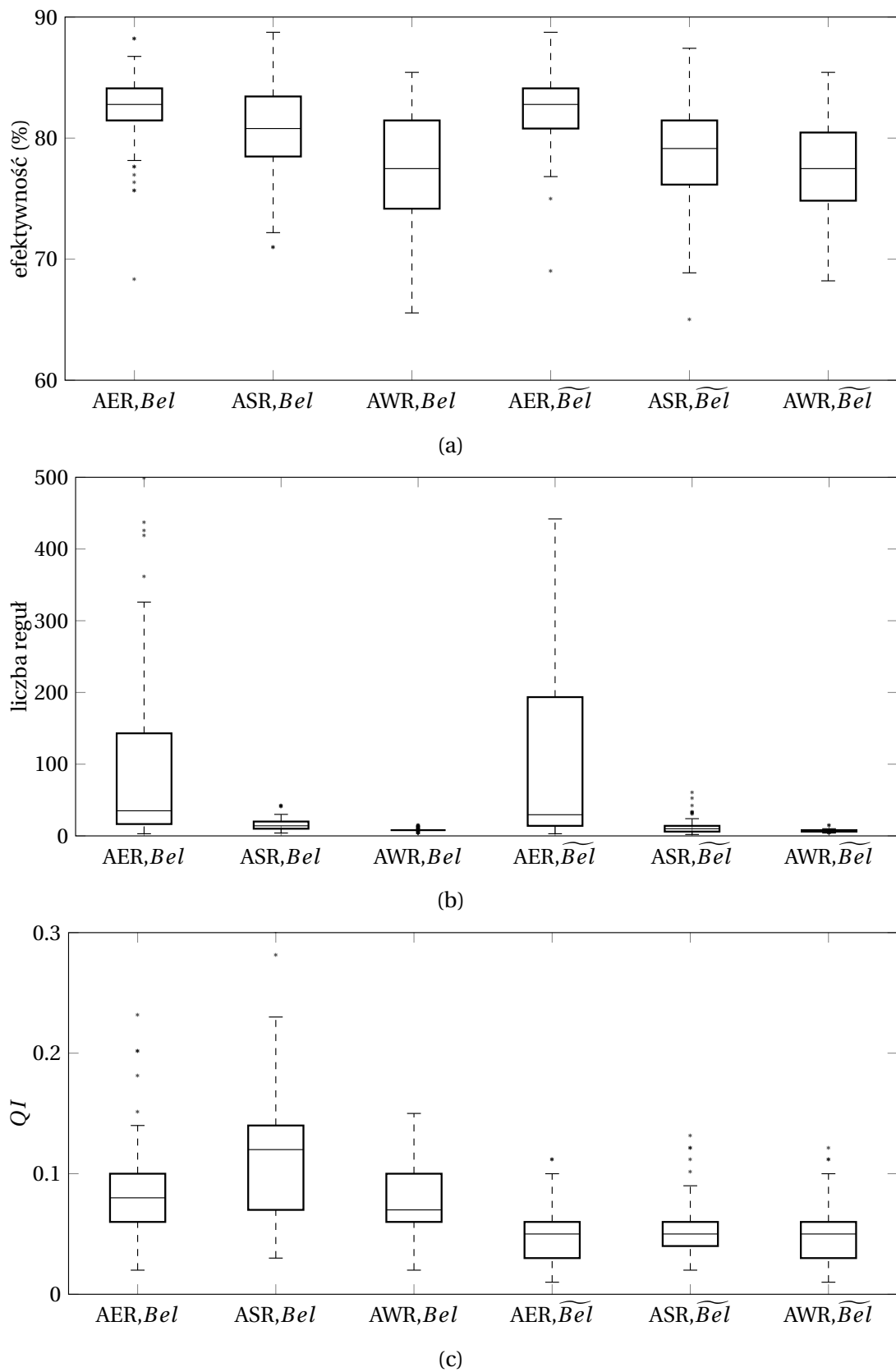


(c)

Rysunek 5.18: Wynik procedury rotacji dla bazy *Pima Indians diabetes* w postaci wykresu „skrzynka z wąsami” dla badanych algorytmów przy zastosowaniu wartości CC jako wskaźnika oceny reguł: (a) rozkład efektywności, (b) rozkład liczebności reguł, (c) rozkład wartości QI



Rysunek 5.19: Wynik procedury rotacji dla bazy *Wisconsin breast cancer* w postaci wykresu „skrzynka z wąsami” dla badanych algorytmów przy zastosowaniu wartości *CC* jako wskaźnika oceny reguł: (a) rozkład efektywności, (b) rozkład liczebności reguł, (c) rozkład wartości *QI*



Rysunek 5.20: Wynik procedury rotacji dla bazy *Cleveland heart disease* w postaci wykresu „skrzynka z wąsami” dla badanych algorytmów przy zastosowaniu wartości CC jako wskaźnika oceny reguł: (a) rozkład efektywności, (b) rozkład liczebności reguł, (c) rozkład wartości QI

5.9 Wydobywanie reguł diagnostycznych z rzeczywistej bazy medycznej

Dane użyte w poprzednich punktach zwykle służą jako typowe bazy benchmarkowe, ale nie są to tzw. dane surowe, czyli zebrane bezpośrednio przez lekarzy diagnostyków, ale bazy poddane wstępnemu opracowaniu [66]. Ważną częścią tego rozdziału jest więc omówienie proponowanych rozwiązań także dla rzeczywistej medycznej bazy danych. Wykorzystana zostanie do tego celu baza dotycząca diagnozy stopnia zwłóknienia wątroby pacjentów, u których zdiagnozowano wirusowe zapalenie wątroby typu C (w skrócie zwany wirusem HCV od ang. *hepatitis C virus*) [86, 89].

5.9.1 Diagnoza zwłóknienia wątroby

Zanim zostaną omówione problemy wydobywania wiedzy z omawianej bazy należy wspomnieć o trudności samej diagnozy stopnia zwłóknienia wątroby. Literatura medyczna wyróżnia kilka skal stopni zwłóknienia [89]. W omawianej bazie użyta jest skala METAVIR [41]. Każdemu z 290 pacjentów w bazie przypisano jeden z pięciu stopni. Ich nazwy [89] oraz procentowy udział w bazie to:

- F0: brak oznak zwłóknienia (6.90%),
- F1: włóknienie wrotne i okołowrotne bez tworzenia przegród (37.59%),
- F2: włóknienie wrotne i okołowrotne z pojedynczymi przegrodami (15.52%),
- F3: liczne przegrody, bez przebudowy marskiej (19.65%),
- F4: marskość (20.34%).

Przypisanie do jednego z pięciu stopni odbywa się na podstawie wzrokowej oceny fragmentu tkanki wątroby (hepatocytów) pobranego metodą biopsji. Niestety, jednoznaczne i bezbłędne zdiagnozowanie, szczególnie pośrednich stopni tj. F1-F3, jest zadaniem trudnym nawet dla doświadczonych lekarzy [97]. Nawet diagnoza wskazująca na brak zwłóknienia może być błędna, jeżeli w wyniku pobrania wycinków zostaną ominięte obszary zmienione chorobowo. Oba te problemy wskazują, że dla opisanych w zbiorze pacjentów przypisana diagnoza jest niepewna. Dodatkowo, biopsja jest badaniem inwazyjnym [59] więc poszukiwane są metody pośrednie [89]. Jedną z nich, zwaną *FibroTest* [85], na podstawie wyników kilku testów laboratoryjnych pozwala prognozować stopień zwłóknienia wątroby, a tym samym ograniczyć częstość wykonywanych biopsji. Lekarze uważają, że skuteczność metody *FibroTest* jest porównywalna z biopsją wycinka igłą o średnicy 15 mm (tzw. sztancą) a sama metoda łatwo powtarzalna i nieinwazyjna [47, 86].

Omawiana baza jest także trudna z punktu widzenia wydobywania wiedzy medycznej. Poza wspomnianą niepewnością wyniku biopsji, który uznaje się za diagnozę referencyjną, zbiór cechuje znacząca liczba danych brakujących. W bazie zawarta jest informacja na temat 26 wyników testów laboratoryjnych. Brakuje ok. 25% wyników, ponieważ baza była tworzona w ciągu kilkuletniej terapii i zmieniały się rodzaje wykonywanych testów. Co więcej, jak

pokazuje tabela 5.15, pacjenci w poważniejszym stanie są „lepiej” przebadani, tzn. mamy na ich temat więcej informacji, choć z drugiej strony stanowią oni mniejszość.

Tabela 5.15: Opis symptomów zebranych w bazie danych na temat pacjentów zarażonych wirusem HCV i procentowe liczby brakujących danych w całym zbiorze i w poszczególnych diagnozach

lp.	nazwa symptomu	liczba brakujących wartości w diagnozach			
		„L+M+H”	„L”	„M”	„H”
1	wiek	0.34%	0.00%	0.00%	1.69%
2	poziom hemoglobiny	58.28%	65.89%	66.67%	27.12%
3	liczba erytrocytów	58.28%	65.89%	66.67%	27.12%
4	liczba leukocytów	0.34%	0.00%	0.00%	1.69%
5	liczba trombocytów	0.34%	0.00%	0.00%	1.69%
6	Czas protrombinowy	26.55%	15.50%	30.39%	44.07%
7	PTP (małopłytkowość %)	3.10%	3.10%	1.96%	5.08%
8	APTT (czas trombolplastyny)	42.07%	37.98%	47.06%	42.37%
9	współczynnik INR	11.72%	5.43%	10.78%	27.12%
10	poziom enzymu ASPT	1.38%	1.55%	0.98%	1.69%
11	poziom enzymu ALAT	0.34%	0.00%	0.00%	1.69%
12	poziom fosfatazy alkalicznej (ALP)	5.17%	6.20%	2.94%	6.78%
13	poziom bilirubiny	6.21%	4.65%	6.86%	8.47%
14	poziom enzymu GGTP	2.76%	2.33%	2.94%	3.39%
15	poziom kreatyniny	59.66%	65.89%	67.65%	32.20%
16	poziom glukozy	61.72%	65.89%	68.63%	40.68%
17	stężenie sodu	63.45%	67.44%	71.57%	40.68%
18	stężenie potasu	63.10%	67.44%	70.59%	40.68%
19	poziom cholesterolu	20.34%	17.05%	12.75%	40.68%
20	poziom całkowity białka	15.86%	17.05%	12.75%	18.64%
21	poziom albumin	28.62%	25.58%	18.63%	52.54%
22	poziom albumin %	23.45%	31.01%	17.65%	16.95%
23	poziom α 1-globulin (%)	24.14%	31.01%	17.65%	20.34%
24	poziom α 2-globulin (%)	24.14%	31.01%	17.65%	20.34%
25	poziom β -globulin (%)	24.14%	31.01%	17.65%	20.34%
26	poziom γ -globulin (%)	23.45%	31.01%	17.65%	16.95%
liczba pacjentów		290	129	102	59
średni procent danych brakujących		24.96%	26.54%	24.92%	21.58%

Wspominano już, że baza obarczona jest niepewnością diagnoz. Co więcej, liczebność zbiorów odpowiadających poszczególnym diagnozom nie jest zbalansowana. Z tej przyczyny częstym podejściem badaczy zajmujących się analizą tej bazy jest zmniejszenie pięciostopniowej diagnozy do trzech kategorii [64, 88, 97]:

1. niski stopień zwłóknienia (F0 i F1),
2. średni stopień zwłóknienia (F2 i F3),
3. marskość (F4).

Utworzone diagnozy zostaną dla ułatwienia oznaczone kolejnymi symbolami „L”, „M” i „H” (od ang. *Low*, *Medium*, *High*). Tabela 5.15 zawiera nazwy wszystkich symptomów w bazie

wraz z informacją o tym, jaka jest procentowa liczba pustych rekordów w całym zbiorze oraz z podziałem na wyszczególnione wyżej diagnozy (kolejne kolumny tabeli).

Na podstawie tabeli 5.15 widać, że niektóre wyniki testów laboratoryjnych (np. zbadane poziomy sodu, potasu czy glukozy) opisują nie więcej niż 40% pacjentów a procent braków przekracza 60% w diagnozach oznaczonych przez „L” oraz „H”. W dole tabeli przedstawiono liczbę pacjentów przypisanych do poszczególnych diagnoz oraz średni procent liczby brakujących wartości. Na podstawie wartości średnich można stwierdzić, że liczba braków w diagnozach „L”, „M” oraz „H” jest zbliżona. Jednakże potwierdza się, że pacjenci przypisani do diagnoz wskazujących na większy stopień zwłóknienia są opisani w bardziej szczegółowy sposób.

Ogólna analiza omawianej bazy może utwierdzić w przekonaniu, że rzeczywiste bazy medyczne charakteryzują się wysokim poziomem niedoskonałości, czyli dużą liczbą brakujących wartości a przede wszystkim niepewnością diagnozy referencyjnej. Stąd wydobywanie efektywnej wiedzy z tych baz jest szczególnie trudne. Niemniej jednak, twórcy bazy używają jej w badaniach nad możliwością klasyfikowania pacjentów [88, 97]. Dlatego wydobywanie reguł umożliwiających wskazanie diagnozy i ich ocena mogą być bardzo cenne dla lekarzy. W kolejnym punkcie przedstawiono wynik procesu wydobywania reguł diagnostycznych na przykładzie tej rzeczywistej bazy medycznej.

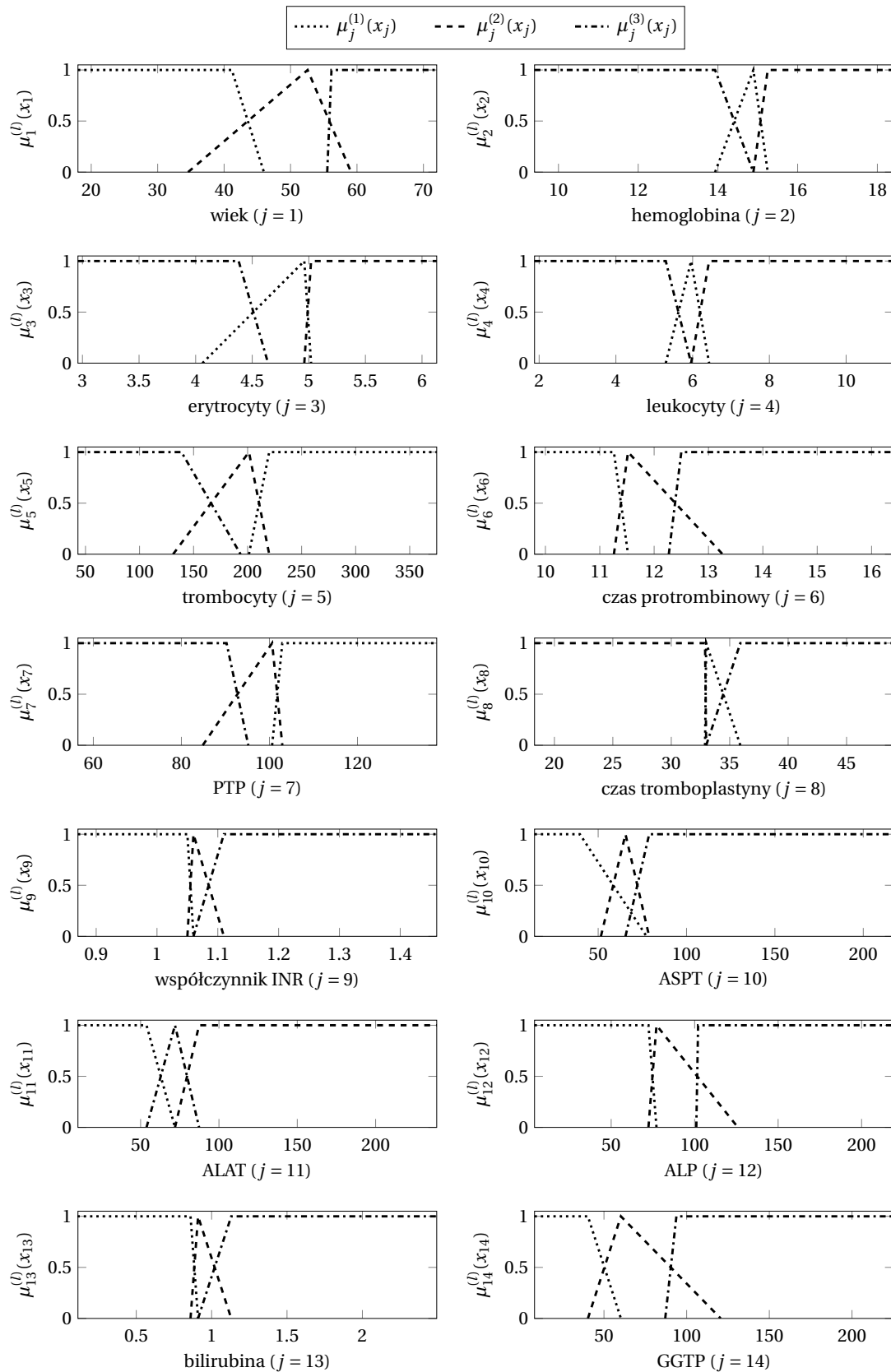
5.9.2 Funkcje przynależności i rozmyte elementy ogniskowe

Trójkątne funkcje przynależności zaprojektowano zgodnie z metodami przedstawionymi w punkcie 3.6. Utworzone funkcje modelujące wartości symptomów dla poszczególnych diagnoz przedstawiają rysunki 5.21-5.22.

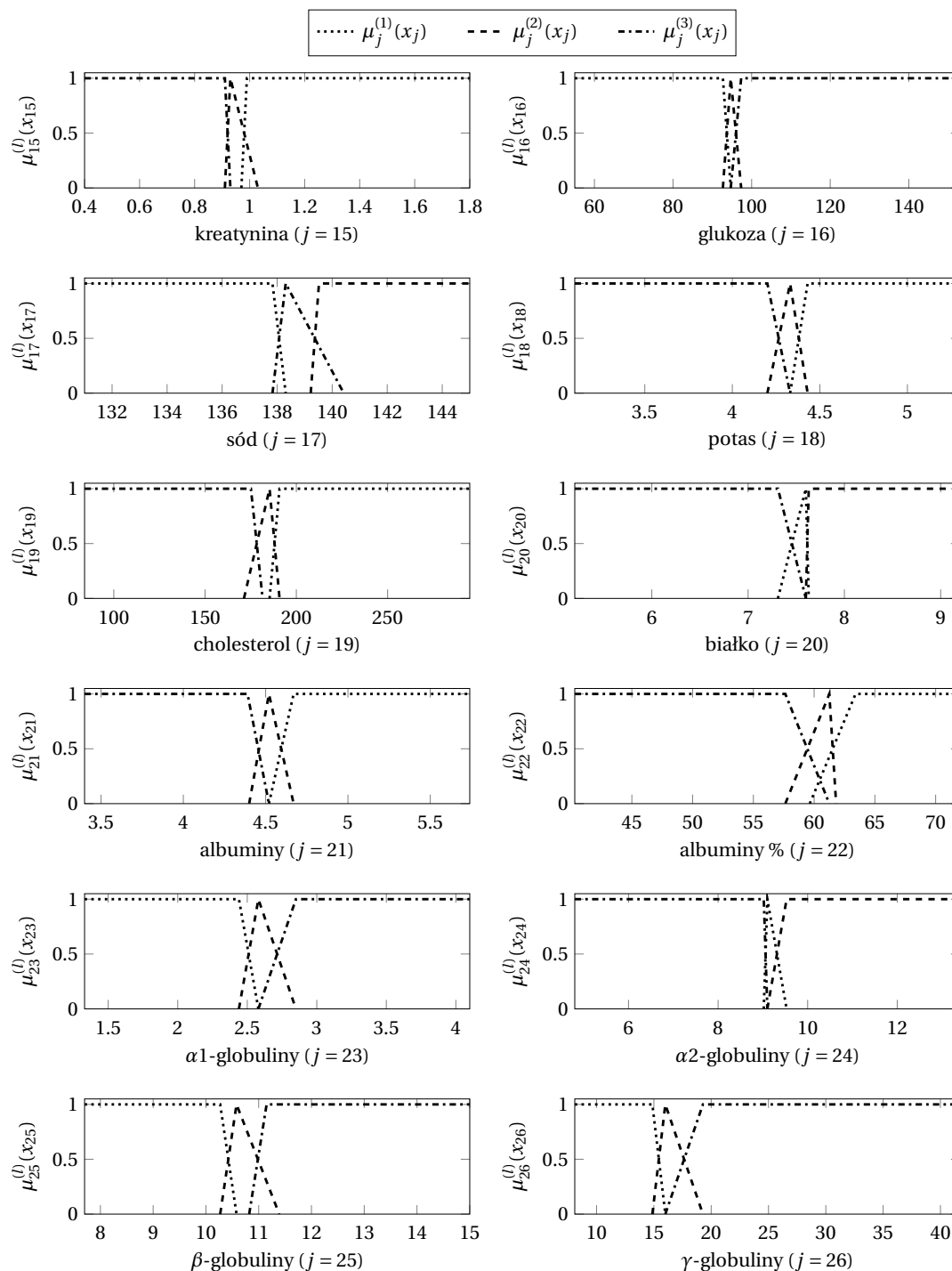
Warto zauważyć, że już same funkcje przynależności reprezentujące wartości lingwistyczne dają pewną wiedzę o omawianym problemie medycznym. Dzięki tym funkcjom można zauważyć m.in., że pogorszenie stanu wątroby związane jest z podwyższonymi poziomami enzymów ważnych dla jej prawidłowego funkcjonowania. Te enzymy są oznaczone skrótami ASPT, ALAT, ALP [11]. Z drugiej strony większa liczba erytrocytów i innych elementów morfotycznych (krwinek) wskazują na lepszy stan pacjenta. Mimo że są to funkcje przynależności utworzone tylko na podstawie dostępnych danych, to już na tym etapie mogą być poddane dyskusji z lekarzem diagnostykiem. Jego doświadczenie jak i krytyczne spojrzenie mogą wpłynąć na modyfikację tych funkcji.

5.9.3 Wyniki wydobywania reguł

W pierwszym podejściu użyto algorytmu warunkowej selekcji (AWR) ze zmodyfikowaną miarą przekonania (3.21). Początkowy zbiór reguł diagnostycznych został oceniony za pomocą zaproponowanego wskaźnika CC (3.42). W tym eksperymencie tworzone reguły o przesłankach zawierających maksymalnie dwa warunki logiczne. Wstępna ocena dopasowania reguł o większej złożoności ujawniła, że duża ich część otrzymuje niskie oceny, a co za tym idzie na pewno nie zostanie wybrana w trakcie algorytmu warunkowej selekcji. Taki zabieg daje także zysk obliczeniowy, gdyż liczba reguł, które trzeba ocenić i dla których



Rysunek 5.21: Trójkątne funkcje przynależności zaprojektowane dla bazy dotyczącej diagnozy stopnia zwłóknienia wątroby – symptomy od 1 do 14



Rysunek 5.22: Trójkątne funkcje przynależności zaprojektowane dla bazy dotyczącej diagnozy stopnia zwłóknienia wątroby – symptomy od 15 do 26

trzeba wyznaczyć rozkład bazowego prawdopodobieństwa spada do poniżej siedmiuset dla każdej z diagnoz. Z uwagi na niedoskonałość danych, na etapie oceny reguł, założono, że próg dopasowania η_e będzie strojony dokładniej, tj. w przedziale (0,1) z dokładnością 0.0005 (patrz punkt 4.6). W wyniku przeprowadzenia algorytmu warunkowej selekcji reguł ocenionych za pomocą CC przy wartości $\eta_e = 0.6195$ uzyskano zbiór 18 reguł diagnostycznych. Tabela 5.16 przedstawia zbiory przesłanek reguł diagnostycznych (czyli rozmyte elementy ogniskowe) wraz z odpowiadającymi im wartościami bazowego prawdopodobieństwa. Przesłanki te zapisano za pomocą stwierdzeń lingwistycznych, dzieląc wartości symptomów różnych diagnoz według funkcji przynależności na kategorie „niski”, „średni” lub „wysoki”. Aby uniknąć nieporozumienia, trzeba pamiętać, że określenia te, np. „wiek młodszy” lub „wiek starszy” odnoszą się do pacjentów rozpatrywanych w tym zagadnieniu medycznym a nie ogólnej populacji. Określenia występujące w przesłankach rozmytych reguł warunkowych zawsze związane są z konkretnym kształtem funkcji przynależności.

Tabela 5.16: Zbiór rozmytych elementów ogniskowych dla diagnozy stopnia zwłóknienia wątroby

$s_r^{(l)}$	$m_r^{(l)}$
$s_1^{(1)}$: wiek jest młodszy i poziom ASPT jest niski	0.1613
$s_2^{(1)}$: wiek jest młodszy	0.1883
$s_3^{(1)}$: wiek jest młodszy i poziom ALAT jest niski	0.1220
$s_4^{(1)}$: wiek jest młodszy i liczba trombocytów jest wysoka	0.1114
$s_5^{(1)}$: liczba trombocytów jest wysoka i poziom ASPT jest niski	0.1485
$s_6^{(1)}$: poziom ASPT jest niski	0.2679
$s_1^{(2)}$: poziom leukocytów jest wysoki i poziom ALAT jest wysoki	0.1558
$s_2^{(2)}$: poziom ALAT jest wysoki	0.2727
$s_3^{(2)}$: poziom ALAT jest wysoki i poziom $\alpha 2$ -globulin jest wysoki	0.1494
$s_4^{(2)}$: poziom leukocytów jest wysoki i poziom $\alpha 2$ -globulin jest wysoki	0.2143
$s_5^{(2)}$: poziom albumin jest średni	0.0974
$s_6^{(2)}$: wiek jest średni i poziom leukocytów jest wysoki	0.1104
$s_1^{(3)}$: poziom hemoglobiny jest niski i liczba trombocytów jest niska	0.1792
$s_2^{(3)}$: liczba erytrocytów jest niska i liczba trombocytów jest niska	0.1604
$s_3^{(3)}$: liczba trombocytów jest niska i poziom ALP jest wysoki	0.1698
$s_4^{(3)}$: wiek jest starszy i liczba erytrocytów jest niska	0.1415
$s_5^{(3)}$: poziom PTP jest niski i poziom ALP jest wysoki	0.1887
$s_6^{(3)}$: liczba trombocytów jest niska i poziom potasu jest niski	0.1604

Reguły przedstawione w tabeli 5.16 pozwalają osiągać 68.62% poprawnych diagnoz spośród wszystkich 290 pacjentów co jest wynikiem zbliżonym do podejść typowo klasyfikacyjnych przeprowadzonych do tej pory dla tej bazy [97]. Szczegółowe wyniki z uwzględnieniem każdej z diagnoz dostarcza macierz pomyłek oraz obliczone miary pochodne (4.8-4.9) przedstawione w tabeli 5.17.

Analizując macierz pomyłek w tabeli 5.17 nietrudno zauważyć, że wydobyty zbiór reguł diagnostycznych ma tendencję do diagnozowania pacjentów jako tych z lepszym stanem

Tabela 5.17: Macierz pomyłek i miary pochodne uzyskane dla wydobytych reguł diagnostycznych przedstawionych w tabeli 5.16

		diagnoza systemu			$SE^{(l)}$	$SP^{(l)}$
		L	M	H		
diagnoza	L	114	13	2	88.37%	63.35%
	M	43	53	4	51.96%	86.17%
	H	14	13	32	54.24%	96.54%
	$PPV^{(l)}$	66.67%	67.09%	84.21%	$E = 68.62\%$	
$NPV^{(l)}$	87.18%	77.51%	89.20%			

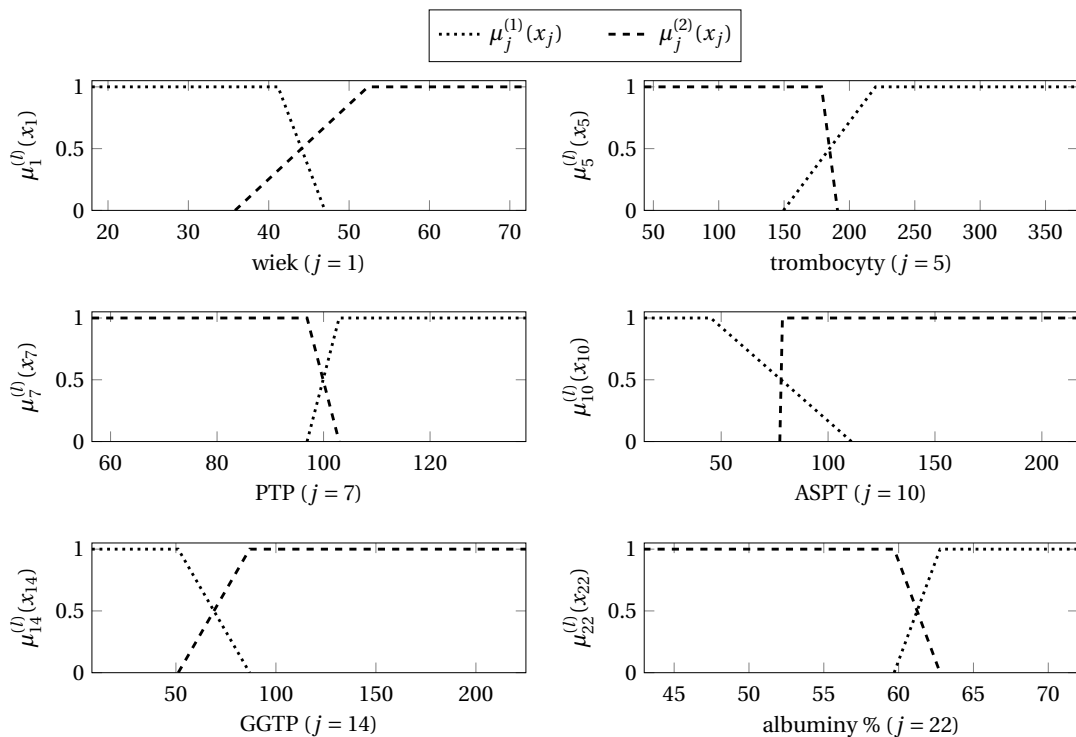
wątroby aniżeli wynikałoby to z diagnozy referencyjnej. Świadczą o tym niemałe liczby $FP^{(l)}$ (3.37) dla diagnozy „niski stan zwłóknienia” (kolumna z nagłówkiem „L”) równe 43 oraz 14. Poza tym, trzeci wiersz tabeli 5.17 ujawnia, że prawie połowa pacjentów przypisanych w bazie do diagnozy „marskość” („H”) zostaje rozpoznana jako pacjenci z niższym stopniem zwłóknienia wątroby. Powyższe obserwacje potwierdzają też czułości $SE^{(l)}$ obliczone dla diagnoz „M” i „H”, które wynoszą tylko nieco ponad 50%. Jest to jednak często osiągnięty wynik a próba zwiększenia $SE^{(2)}$ oraz $SE^{(3)}$ wiąże się z zmniejszeniem czułości $SE^{(1)}$ [97]. Skoro działanie proponowanego systemu jak i sprawdzonych do tej pory klasyfikatorów wykazują taką tendencję, przetestowano inne podejście. Reguły diagnostyczne zostały wydobyte przy innej interpretacji etykiet skali METAVIR: pacjentów sklasyfikowanych według niej do stopni F0 i F1 potraktowano jako pacjentów o „początkowym stopniu zwłóknienia wątroby” („Z1”) a pacjentów przypisanych do pozostałych stopni tj. F2, F3 oraz F4 do diagnozy „zaawansowany stopień zwłóknienia wątroby” („Z2”). Pozwala to zwiększyć liczebność pacjentów diagnozy „Z2” oraz zbliżyć się do wyrównania liczebności danych dla konkurujących diagnoz (129 pacjentów z diagnozą „Z1” oraz 161 pacjentów z diagnozą „Z2”). W takim ujęciu algorytm warunkowej selekcji dostarczy raczej ogólną wiedzę ale w obliczu niepewności diagnozy wydaje się to dozwolone, a nawet wskazane w celu zmniejszenia ryzyka zakwalifikowania pacjenta jako będącego w lepszym stanie niż w rzeczywistości.

Prezentacja wszystkich funkcji przynależności nie zostanie w tym miejscu powtórzona, tak jak dla rozpatrywanego problemu trzech diagnoz (rysunki 5.21 i 5.22). Przedstawione zostaną jedynie funkcje przynależności utworzone dla tych symptomów, dla których wydobyto rozmytą regułą warunkową. Dlatego najpierw przedstawiony zostanie zbiór reguł diagnostycznych wydobytych dzięki algorytmowi warunkowej selekcji reguł. Został on zaprezentowany w tabeli 5.18. Funkcje przynależności dla wydobytych elementów ogniskowych przedstawia rysunek 5.23.

Choć i tym razem w regułach diagnostycznych występują elementy ogniskowe takie jak „wiek młodszy”, „wiek starszy”, czy inne tożsame językowo określenia jak w przypadku prezentowanych wcześniej wyników, to należy zauważyć, że tym razem kształty funkcji przy-

Tabela 5.18: Zbiór rozmytych elementów ogniskowych dla diagnozy stopnia zwłóknienia wątroby – podejście z dwoma diagnozami

$s_r^{(l)}$	$m_r^{(l)}$
$s_1^{(1)}$: wiek jest młodszy i poziom ASPT jest niski	0.1971
$s_2^{(1)}$: wiek jest młodszy	0.2176
$s_3^{(1)}$: wiek jest młodszy i liczba trombocytów jest wysoka	0.1529
$s_4^{(1)}$: liczba trombocytów jest wysoka i poziom GGTP jest niski	0.2000
$s_5^{(1)}$: liczba trombocytów jest wysoka i poziom ASPT jest niski	0.2324
$s_1^{(2)}$: wiek jest starszy	0.2936
$s_2^{(2)}$: liczba trombocytów jest niska	0.2172
$s_3^{(2)}$: wiek jest starszy i liczba trombocytów jest niska	0.1671
$s_4^{(2)}$: poziom procentowy albumin jest niski	0.1742
$s_5^{(2)}$: liczba trombocytów jest niska i poziom PTP jest niski	0.1480



Rysunek 5.23: Trójkątne funkcje przynależności zaprojektowane dla bazy dotyczącej diagnozy stopnia zwłóknienia wątroby – symptomy występujące w regułach przedstawionych w tabeli 5.18

należności się zmieniły. W wyniku redukcji liczby diagnoz nie pojawiają się zbiory rozmyte o pośrednich wartościach nośnika a konkurujące ze sobą diagnozy są teraz opisane funkcjami przynależności o szerszych przedziałach częściowej przynależności.

Mając do czynienia tylko z dwoma diagnozami nie ma potrzeby prezentowania miar pochodnych w macierzy pomyłek, tak jak to miało miejsce w tabeli 5.17. W przypadku wykonywania procedury diagnostycznej, testu mającego charakter klasyfikacji dwuklasowej, aby ocenić jej jakość oblicza się jedną wartość SE oraz jedną SP . Wybiera się wtedy diagnozę, która uznawana jest za wynik pozytywny i poprawność wykrycia danej diagnozy uznaje się za przypadek *true positive*. Zwykle jako wynik pozytywny uznaje się wykrycie stanu chorobowego [55]. Dlatego, jako wynik pozytywny została wybrana diagnoza „Z2” czyli „zaawansowany stopień zwłóknienia”. Macierz pomyłek przedstawia tabela 5.19. Zmiana dokonana w przypisaniu diagnoz spowodowała powiększenie ogólnej efektywności wyznaczania diagnozy o 8% co daje o 22 więcej poprawnie zdiagnozowanych pacjentów. Nie jest to wynik jednoznacznie lepszy od poprzedniego z uwagi na to, że utracone zostało zróżnicowanie pomiędzy dwoma („M” i „H”), a nawet, biorąc pod uwagę oryginalną skalę METAVIR, trzema („F2”, „F3” i „F4”) kategoriami diagnostycznymi, interpretując je za tożsame („Z2”). Dodatkowo, zwiększenie ogólnej efektywności E zostało uzyskane kosztem pomniejszenia liczby poprawnie rozpoznanych pacjentów o niskim stanie zwłóknienia wątroby (spadek z 114 do 90 poprawnych diagnoz). Oba uzyskane wyniki przedstawione w tabelach 5.17 i 5.19 potwierdzają, że istotna jest nie tylko pewność diagnozy referencyjnej, która w omawianej bazie jest bardzo mała, ale także odpowiednie stopniowanie postępu choroby, czyli zwłóknienia wątroby. Przedstawione wyniki ilustrują wpływ zmiany tego stopniowania na wydobywanie wiedzy medycznej i wskazują, że być może, w istniejących warunkach warto jest poddać dyskusji obecne kategorie diagnostyczne. Tak jak dla poprzednich baz, uogólnienie wiedzy było sprawą kompromisu pomiędzy małą złożonością zbioru reguł a wysoką dokładnością, tak dla tej bazy również odpowiednią szczegółowością podziału diagnoz. Zauważmy, że zmniejszanie liczby diagnoz przeprowadza się także dla bazy *Cleveland heart disease* (patrz tabela 5.1).

Podsumowując, punkt ten poświęcono przedstawieniu możliwości wydobywania wiedzy z rzeczywistej bazy medycznej. W wyniku eksperymentów uzyskane zostały zbiory reguł diagnostycznych o bardzo małej złożoności. Spośród 26 symptomów w oryginalnym zbiorze danych algorytm AWR wybiera co najwyżej kilkanaście reguł związanych z nie więcej niż 12 wynikami testów laboratoryjnych. W pracy [97] istotne testy laboratoryjne, wybrane w oparciu o pewne metody selekcji, zostały skonfrontowane z opinią eksperta medycznego, który wskazał na 15 istotnych dla niego parametrów w zagadnieniu zwłóknienia wątroby. Reguły przedstawione w tabeli 5.16 związane są z testami laboratoryjnymi, które w dużej części pokrywają się z opinią tego eksperta, a także z prezentowanymi wynikami procesu selekcji cech. Proponowany algorytm wydobywania reguł ocenia je biorąc pod uwagę osobno każdą z diagnoz dlatego część symptomów (wyników testów) jest istotna tylko w odniesieniu do jednej diagnozy (patrz tabele 5.16-5.18). Takiego rozróżnienia nie zapewniają metody użyte w [97]. Wydaje się, że zaliczenie przypadków z diagnozy pośredniej do diagnozy związanej z zaawansowanym stopniem zwłóknienia, pozwoliło wydobyć lepsze w interpretacji funkcji

przynależności, skonstruowane na podstawie bardziej zbilansowanych zbiorów dwóch diagnoz.

Tabela 5.19: Macierz pomyłek i miary pochodne uzyskane dla wydobytych reguł diagnostycznych przedstawionych w tabeli 5.16

		diagnoza systemu		<i>SE</i>	<i>SP</i>
		Z1	Z2		
diagnoza	Z1	90	39	81.37%	69.77%
	Z2	30	131		
	<i>PPV</i>	77.06%		<i>E</i> = 76.21%	
	<i>NPV</i>	75.00%			

5.10 Porównanie wyników wydobywania reguł diagnostycznych z metodami referencyjnymi dla baz benchmarkowych

Na zakończenie tego rozdziału zaproponowane metody wydobywania reguł zostaną pogłębowo porównane z podobnymi rozwiązaniami. Do tego celu wykorzystano pakiet uczenia maszynowego KEEL [4]. Wśród wielu metod klasyfikacji gotowych do uruchomienia z poziomu narzędzia, wybrano kilka algorytmów opierających swoje działanie na tworzeniu reguł warunkowych oraz drzew decyzyjnych. Wybrane siedem metod to podejścia opisywane w publikacjach ukazanych w ciągu ostatnich kilku lat.

Tabela 5.20 przedstawia średnie efektywności uczenia (5.3) i testu (5.4) w procedurze zmodyfikowanej 10-krotnej walidacji (tzw. *distribution optimally balanced stratified cross-validation* – DOBSCV [83]). Ta metoda zapobiega zmianom rozkładów danych pomiędzy tworzonymi podziałami. Poszczególne podziały zapisano, co umożliwiło wykorzystanie ich w przeprowadzeniu tożsamej procedury także dla metod proponowanych w niniejszej rozprawie. Wybrano do tego dwa algorytmy selekcji tzn. ASR i AWR. Dodatkowo, w tabeli 5.20 wyszczególniono również liczbę wydobywanych reguł.

Wyniki uzyskanie dzięki procedurze 10-F DOBSCV są do siebie bardzo zbliżone (szczególnie wyniki średniej efektywności testu \bar{E}_T). Główną różnicą proponowanych metod jest najmniejszy spadek efektywności w teście \bar{E}_T względem wyniku uczenia \bar{E}_U . Dla większości metod referencyjnych (poza algorytmem sekwencyjnym FURIA [50] oraz algorytmem genetycznym GP-COACH [17]) wysoka średnia efektywność uczenia, związana często z niemałą liczbą reguł, skutkuje znacznie gorszym wynikiem testu. Algorytmy ASR i AWR (z użyciem CC (3.42) jako oceny reguł i zmodyfikowaną metodą obliczania miary przekonania (3.21)) wykazują największą stabilność efektywności. Średnia \bar{E}_T podaje także wprost procentową efektywność rozpoznawania wszystkich danych w zbiorze, co wynika z założonego podziału

Tabela 5.20: Średnie efektywności uczenia i testu oraz średnie liczby reguł otrzymanych w wyniku przeprowadzenia zmodyfikowanej procedury 10-F DOBSCV na testowych bazach danych za pomocą wybranych metod

metoda	TH			AP			PI			BC			HD		
	\bar{E}_U	\bar{E}_T	\bar{L}	\bar{E}_U	\bar{E}_T	\bar{L}	\bar{E}_U	\bar{E}_T	\bar{L}	\bar{E}_U	\bar{E}_T	\bar{L}	\bar{E}_U	\bar{E}_T	\bar{L}
AB.NC* [118]	85.9	82.7	6	92.9	85.0	4	91.9	76.8	28	99.2	97.1	9	93.1	78.8	11
C4.5-B** [43]	98.9	94.9	-	92.2	85.0	-	83.5	75.4	-	98.3	95.3	-	91.7	79.5	-
FURIA [50]	99.4	93.9	7	90.8	85.0	3	78.8	75.5	8	98.8	95.0	16	87.6	81.0	9
GP-C [17]	92.9	91.6	8	92.7	85.0	7	79.0	76.3	68	96.2	96.0	13	87.0	83.3	27
MPLCS [10]	99.1	92.5	5	96.0	83.1	4	84.7	75.8	8	98.9	95.9	5	94.9	75.1	7
SLAVE [44]	93.5	93.0	9	92.0	85.0	6	79.1	76.0	14	99.6	94.0	10	95.1	73.7	20
NSLV [44]	95.3	91.6	6	92.0	86.9	3	80.1	74.6	13	98.8	94.7	9	91.2	78.4	12
ASR	98.6	96.3	8	89.4	83.6	8	76.5	75.9	10	98.6	96.3	8	85.3	84.4	20
AWR	98.5	95.4	8	88.9	82.7	6	76.3	76.1	9	97.1	96.1	8	84.3	81.5	9

* - liczba reguł dla algorytmu AdaBoost.NC jest dodatkowo uśrednioną liczbą reguł w 10 klasyfikatorach tworzonych dla każdej iteracji procedury uczenia (tzn. że w faktyczna liczba reguł jest około 10-krotnie większa)

** - wynik działania C4.5-B w ramach pakietu KEEL nie prezentował wydobytych zbiorów reguł

w procedurach typu *k-fold cross-validation* [63]. Jak widać, algorytmy ASR i AWR wykazują większą efektywność diagnozy dla baz *thyroid disease*, *Wisconsin breast cancer* i *Cleveland heart disease*. Dla bazy *Pima Indians diabetes* proponowane metody osiągają porównywalne efektywności choć dla znacznie mniejszej liczby reguł. Żadna z metod referencyjnych wyszczególnionych w tabeli 5.20 nie miała możliwości zabezpieczenia przed tworzeniem przesłanek reguł o dużej złożoności. Dwie z wybranych metod [44] przedstawiały natomiast wielkość opisującą złożoność wydobywanych reguł, tzn. stosunek liczby warunków logicznych występujących we wszystkich regułach do całkowitej liczby reguł. Taką analizę zbiorów reguł przeprowadzono także dla reszty metod referencyjnych (poza algorytmem C4.5-B ze wspomnianych w tabeli 5.20 przyczyn) oraz algorytmów proponowanych w rozprawie. Wyniki tej analizy przedstawia tabela 5.21.

Tabela 5.21: Średnie stosunki liczb warunków logicznych do liczb reguł otrzymane dzięki analizie wyników procedury 10-F DOBSCV

metoda	TH	AP	PI	BC	HD	średnia
AB.NC	3.04	1.14	7.67	3.67	4.20	3.94
FURIA	1.55	1.05	1.15	1.43	1.13	1.26
GP-C	1.42	2.69	4.59	1.90	2.21	2.56
MPLCS	2.09	2.13	4.85	3.73	4.78	3.52
SLAVE	4.46	6.33	6.91	8.39	11.68	7.55
NSLV	2.27	2.17	3.43	2.44	3.39	2.74
ASR	1.61	2.43	2.25	1.36	1.89	1.91
AWR	1.73	2.22	2.15	1.80	1.60	1.90

Analizując tabelę 5.21 i zarazem mając na uwadze wyniki przedstawione w tabeli 5.20 można zauważyć, że wśród wszystkich metod algorytm FURIA [50] wydobywał reguły o najmniej złożonych przesłankach. Zwykle liczba reguł też była najmniejsza z jednym wyjątkiem (dla bazy *Wisconsin breast cancer*). Algorytm genetyczny GP-COACH [17] poza wynikiem dla bazy *Pima Indians diabetes* wydobywał także proste reguły, ale często ich sumaryczna liczba była bardzo duża. Proponowane w tej pracy metody dzięki ograniczeniu liczby warunków logicznych (punkt 4.1) pozwoliły osiągać bardzo proste zbiory reguł plasując się wspólnie na drugim miejscu (jak to pokazuje kolumna ze średnimi wynikami w tabeli 5.21) badanych metod. Należy jasno stwierdzić, że proponowane metody ASR i AWR zarówno pod względem osiągniętej efektywności uogólniania jak i wydobywanym zbiorom reguł okazują się konkurencyjne wobec metod omawianych w literaturze w ostatnim czasie.

Rozdział 6

Wnioski końcowe

Zjawiskiem typowym w obszarze medycznym, tak jak w innych dziedzinach, jest to, że wdrażanie technik cyfrowych (systemy monitorowania pacjentów, komputerowe systemy szpitalne, cyfrowe bazy danych) pozwala wygodnie gromadzić dane na temat pacjentów, a także powoduje rozwój samych technik badawczych (np. poprzez cyfrową akwizycję pomiarów) czy testów diagnostycznych. Z drugiej strony, przetwarzanie nieustannie gromadzonych zbiorów danych medycznych i uzyskiwanie na ich podstawie praktycznej wiedzy służącej poprawieniu jakości diagnozowania nie jest zadaniem łatwym z uwagi na niedoskonałość danych medycznych.

Niniejsza rozprawa przedstawia rozwiązanie wychodzące naprzeciw analizie niedoskonałych danych medycznych tj. nieprecyzyjności związanej z informacją na temat stanu pacjenta oraz sprzyjające uwzględnieniu niepewności wiedzy medycznej, która objawia się poprzez stopień pewności diagnoz stawianych przez lekarza. Opisana została technika wspomaganie decyzji w postaci reguł diagnostycznych wykorzystujących zalety dwóch teorii użytecznych w zagadnieniach przetwarzania danych z niepewnością: teorii Dempstera-Shafera i teorii zbiorów rozmytych. Użyty mechanizm wspomaganie diagnozy jest związany z regułami warunkowymi. Pojęcie zbioru rozmytego jest przydatne w reprezentacji przesłanki reguły. Zbiór rozmyty pozwala na określenie poziomu dopasowania objawu pacjenta z symptomem rozpatrywanej diagnozy, utożsamianym w teorii Dempstera-Shafera z elementem ogniskowym. Każdy element ogniskowy jako przesłanka reguły warunkowej związany jest z wartością przekonania o diagnozie, przy potwierdzeniu jego wystąpienia dla danego pacjenta. Ta wartość przekonania o diagnozie jest wartością bazowego prawdopodobieństwa, które stanowi wagę pojedynczej reguły i może pochodzić bezpośrednio od eksperta ludzkiego. Również element ogniskowy może być wprowadzony przez eksperta. W niniejszej pracy jednak zarówno określenie rozmytych elementów ogniskowych jak i rozkładu bazowego prawdopodobieństwa wykonuje się na podstawie danych uczących. Kluczowym aspektem rozprawy było zaproponowanie nowych sekwencyjnych algorytmów wydobywania reguł diagnostycznych z danych medycznych. W proponowanym rozwiązaniu nacisk postawiono na tworzenie efektywnych zbiorów reguł przy jednoczesnym ograniczeniu ich złożoności celem umożliwienia ich interpretacji i użycia przez lekarza podejmującego diagnozę. W proponowanych

algorytmach uwagę skupiono także na ocenie jakości reguł czego wynikiem było zaproponowanie wskaźnika oceny reguł, który skutecznie poprawił jakość wydobywanej wiedzy.

W niniejszej rozprawie z powodzeniem udowodniono przydatność proponowanych metod we wspomaganiu diagnozy. Przy użyciu ogólnodostępnych baz medycznych związanych ściśle z diagnozą medyczną przetestowano działanie proponowanych algorytmów wydobywania rozmytych reguł diagnostycznych. W wyniku działania proponowanych algorytmów uzyskiwane są zbiory reguł o małej liczebności i małej złożoności ich przesłanek (małej liczbie warunków logicznych). Metody proponowane na początku pracy, w trakcie badań były doskonalone. Odbywało się to głównie na etapie uczenia tj. wydobywania wiedzy z danych uczących. W wyniku tych działań, dopracowano rozwiązania z zastosowaniem pewnych technik (warunek zatrzymania algorytmu, warunki selekcji), które zabezpieczają przed nadmiernym dopasowaniem reguł do danych uczących. W rezultacie, pozwoliło to uzyskać kompromis pomiędzy złożonością reguł a pożądaną niezawodnością, gdyż wydobyte reguły cechują się satysfakcjonującą zdolnością uogólniania ocenianą po efektywności testu a jej wyniki identyfikują proponowane metody jako co najmniej konkurencyjne do znanych dotychczas metod.

Wydaje się tym samym, że zaplanowane zadania zostały zrealizowane, tj.:

1. Opracowano zmodyfikowany algorytm eliminacji reguł, w którym zbiór elementów ogniskowych powstaje z pewnego początkowego i nadmiarowego zbioru poprzez iteracyjne usuwanie elementów ogniskowych (przesłanek reguł diagnostycznych) z najniższą wartością bazowego prawdopodobieństwa.
2. Zaproponowano metodę wyznaczania kształtów funkcji przynależności dla elementów ogniskowych i zbadano ich wpływ na efektywność zaproponowanego algorytmu eliminacji reguł.
3. Zaproponowano kryterium (wskaźnik) oceny reguły diagnostycznej, które w wyniku działania proponowanych algorytmów, pozwoliło otrzymać mniejszą liczbę reguł z zachowaniem zbliżonej efektywności wyznaczania diagnozy, w porównaniu z dotychczasowym kryterium w postaci wartości bazowego prawdopodobieństwa.
4. Opracowano algorytm selekcji reguł, polegający na wyborze najlepiej dopasowanych elementów ogniskowych z ich zbioru początkowego.
5. Zmodyfikowano algorytm selekcji reguł poprzez dodanie kroku warunkowego usuwania reguł, które nie poprawiają efektywności diagnozy tj. uzyskano algorytm warunkowej selekcji reguł.
6. Wprowadzono wskaźnik oceny wyniku działania proponowanych algorytmów wydobywania reguł diagnostycznych. Wskaźnik ten uwzględnia kompromis pomiędzy efektywnością wyznaczania poprawnej diagnozy a złożonością zbioru reguł diagnostycznych.
7. Wykorzystano do wspomagania diagnozy zmodyfikowaną miarę przekonania, łączącą bezpośrednio wartość bazowego prawdopodobieństwa elementów ogniskowych z poziomami ich dopasowania z danymi diagnozowanego pacjenta.

Podsumowując, zrealizowane zadania i uzyskane rezultaty sugerują, że dla benchmarkowych i rzeczywistej bazy danych medycznych tezę rozprawy należy uznać za prawdziwą, to znaczy:

Wykorzystanie teorii Dempstera-Shafera i teorii zbiorów rozmytych poprzez zastosowanie iteracyjnych algorytmów wyodrębniania rozmytych reguł warunkowych na podstawie danych uczących poprawia jakość wspomaganie diagnozy medycznej. Jakość jest rozumiana jako efektywność uogólniania oraz możliwie mała złożoność uzyskanego zbioru rozmytych reguł warunkowych.

Prezentowana rozprawa stanowiła próbę rozwiązania problemu wspomaganie diagnozy medycznej w warunkach niedoskonałości danych. W opisywanych działaniach jako sprawę obligatoryjną uznano chęć uzyskania narzędzia, które przede wszystkim nadawać będzie się do interpretacji przez eksperta ludzkiego co narzuciło pewne ograniczenie złożoności metody wspomaganie diagnozy jak i algorytmów wydobywania reguł diagnostycznych. Niemniej jednak, z całą pewnością możliwe są dalsze kroki badawcze w tej tematyce z uwagi na to, że nie wyczerpano kierunków udoskonalania narzędzia wspomaganie diagnozy medycznej. Do takich działań mogą zaliczać się między innymi:

1. Uszczegółowienie procesu tworzenia funkcji przynależności dla elementów ogniskowych, to znaczy:
 - (a) użycie różnych kształtów funkcji przynależności indywidualnie dla różnych diagnoz i symptomów,
 - (b) określenie kryterium wyboru powyższych kształtów,
 - (c) wprowadzenie metody dostrajającej parametry funkcji przynależności w wydobytych regułach diagnostycznych w etapie uczenia.
2. Zaproponowanie nowej metody wyznaczania rozkładu bazowego prawdopodobieństwa, która będzie wykorzystywać miarę możliwości ale bez potrzeby wyboru prognozy dopasowania.
3. Przetwarzanie informacji wejściowej pacjentów w liczby rozmyte modelujące niedokładność pomiarów medycznych.
4. Przygotowanie systemu wspomaganie diagnozy do przetwarzania innych aspektów niedoskonałości danych medycznych np. danych sprzecznych czy odstających.

Dalsze działania w tej dziedzinie z pewnością mogą stanowić aktywny obszar badawczy. Najefektywniejszym wykorzystaniem przedstawionych wyników będzie implementacja proponowanych metod w systemie wspomaganie diagnozy.

Dodatek A

Testowane bazy

A.1 Opis baz i przykładowe zbiory reguł

A.1.1 *iris flowers*

Wśród 150 osobników w bazie każdy z nich opisany jest czterema parametrami, które dotyczą następujących wymiarów kwiatu irysa:

1. długość działki kielicha [cm],
2. szerokość działki kielicha [cm],
3. długość płątka korony [cm],
4. szerokość płątka korony [cm].

Na rysunkach A.1a-A.1d przedstawione zostały zbiory rozmyte opisane za pomocą czterech kształtów funkcji przynależności. Każda z nich opisuje zbiory rozmyte, które reprezentują wartości lingwistyczne: „mała”, „średnia” i „duża” w odniesieniu do poszczególnych osobników w bazie.

A.1.2 *thyroid disease*

Symptomy w tej bazie związane są z wynikami przeprowadzonych testów laboratoryjnych:

1. procentowy wynik testu T3RU (pomiar stężenia globuliny),
2. poziom tyroksyny (hormonu T4) w próbce,
3. poziom trójjodotyroniny (hormonu T3) w próbce,
4. poziom tyreotropiny (hormonu TSH) w próbce,
5. maksymalna zmiana poziomu TSH po podaniu 200 mg hormonu hamującego.

Podobnie jak baza *iris flowers* zawierała trzy klasy gatunków kwiatu irysa, tak baza *thyroid disease* jest bazą, w której pacjenci przypisani są do jednej z trzech diagnoz opisanych w rozdziale 5 w tabeli 5.1. Tym samym poszczególne funkcje przynależności (przedstawione na rysunkach A.2a-A.2b) opisywać będą jedną z wartości lingwistycznych dotyczących wyniku testu, np.: „niski”, „w normie” oraz „wysoki”. Przykładowy zbiór reguł diagnostycznych z wartościami bazowego prawdopodobieństwa przedstawiono w tabeli A.1.

Tabela A.1: Zbiór reguł diagnostycznych dotyczących stanu funkcjonowania tarczycy

$s_r^{(l)}$	$m_r^{(l)}$
$s_5^{(1)}$: zmiana poziomu tyreotropiny (TSH) jest w normie	0.50
$s_2^{(1)}$: poziom tyroksyny (T4) jest w normie	0.50
$s_2^{(2)}$: poziom tyroksyny (T4) jest wysoki	0.50
$s_3^{(2)}$: poziom trójjodotyroniny (T3) jest wysoki	0.50
$s_2^{(3)}$: poziom tyroksyny (T4) jest niski	0.50
$s_4^{(3)}$: poziom tyreotropiny (TSH) jest niski	0.50

Dla powyższego zbioru reguł wartości lingwistyczne występujące w elementach ogniskowych opisane są za pomocą trójkątnych funkcji przynależności przedstawionych na rysunku A.2a.

A.1.3 *appendicitis*

Opis symptomów dla powyższej bazy jest nieznanymi [120]. Znane są natomiast etykiety diagnoz. Dodatkowo, wartości wszystkich symptomów są liczbami rzeczywistymi i mieszczą się w one przedziale jednostkowym, co wskazuje na to, że przypisanie wartości dla danego symptomu wiązało się z wyrażeniem pewnej skali interpretowanej przez lekarza diagnostyka. Projektowane funkcje przynależności przedstawione zostały na rysunkach A.3a-A.4b. Określając symptom jako jego numer dla przykładowego rozwiązania możliwe jest zaprezentowanie bazy reguł diagnostycznych w następującej formie:

Tabela A.2: Zbiór reguł diagnostycznych dotyczących diagnozy stanu zapalnego wyrostka

$s_r^{(l)}$	$m_r^{(l)}$
$s_3^{(1)}$: wartość symptomu (3) jest wysoka	0.33
$s_6^{(1)}$: wartość symptomu (6) jest wysoka	0.34
$s_7^{(1)}$: wartość symptomu (7) jest wysoka	0.33
$s_{30}^{(2)}$: wartości symptomów: (1), (2) i (4) są niskie	0.31
$s_{31}^{(2)}$: wartości symptomów: (1), (2) i (5) są niskie	0.36
$s_{41}^{(2)}$: wartości symptomów: (1), (5) i (6) są niskie	0.33

Z uwagi na występowanie dwóch diagnoz w bazie *appendicitis* dla powyższej bazy reguł diagnostycznych stwierdzenia wartość niska bądź wartość wysoka są jednoznacznie opisane przez trójkątne funkcje przynależności z rysunku A.3a.

A.1.4 *Pima Indians diabetes*

Baza dotyczy diagnozy osób chorych na cukrzycę i zawiera informację medyczną na temat 768 kobiet wśród plemienia Pima wśród Indian Ameryki Północnej [57], z których każda pacjentka opisana jest następującymi informacjami:

1. liczba przeżytych ciąż,
2. stężenie glukozy,
3. ciśnienie diastoliczne,
4. grubość tkanki skórnej w okolicy mięśnia trójgłowego ramienia,
5. zawartość insuliny,
6. BMI,
7. obciążenie cukrzycą w rodzinie,
8. wiek.

Przykładowe rozwiązanie uzyskane dla trójkątnych funkcji przynależności (rysunek A.5a) związane jest za następującą bazą diagnostyczną:

Tabela A.3: Zbiór reguł diagnostycznych dotyczących diagnozy cukrzycy

$s_r^{(l)}$	$m_r^{(l)}$
$s_2^{(1)}$: stężenie glukozy jest niskie	0.24
$s_7^{(1)}$: obciążenie cukrzycą w rodzinie jest niskie	0.19
$s_5^{(1)}$: poziom insuliny jest niski	0.19
$s_8^{(1)}$: wiek jest młodszy	0.19
$s_1^{(1)}$: liczba ciąż jest mała	0.19
$s_6^{(2)}$: BMI jest wysoki	0.23
$s_2^{(2)}$: stężenie glukozy jest wysokie	0.21
$s_8^{(2)}$: wiek jest starszy	0.19
$s_3^{(2)}$: ciśnienie diastoliczne jest wysokie	0.19
$s_{19}^{(2)}$: stężenie glukozy jest wysokie i BMI jest wysoki	0.18

Wszystkie funkcje przynależności utworzone na podstawie tej bazy znajdują się na rysunkach A.5a-A.6b.

A.1.5 *Wisconsin breast cancer*

Baza ta dotyczy diagnozowania zmian złośliwych i niezłośliwych raka piersi pacjentów płci żeńskiej [66]. Wartości symptomów w bazie *Wisconsin breast cancer* związane są z wynikami testów laboratoryjnych pobranej krwi i mają następujący opis:

1. liczebność zgrupowanych komórek,
2. jednolitość rozmiaru komórek,
3. jednolitość kształtu komórek,
4. marginalna adhezja komórek,
5. rozmiar pojedynczej komórki nabłonkowej,
6. występowanie samego jądra komórkowego,
7. bezbarwność chromatyny,
8. prawidłowe jąderka w jądrach komórek,
9. występujące mitozy.

Podobnie jak dla bazy *appendicitis* wszystkie wartości symptomów są liczbami rzeczywistymi w jednostkowym zakresie. Każdy symptom jest jednak wielokrotnością wartości 0.1 i wyraża skalę subiektywnej zgody lekarza z wystąpieniem danego symptomu. Wartości bliskie 1.0 oznaczają dużą zgodność, a bliskie 0.0 na dużą niezgodność lekarza z sugerowanym objawem. W związku z tym, że baza dotyczy zagadnienia dwóch diagnoz (zmiany złośliwe i łagodne), w bazie reguł występują jedynie funkcje odnoszące do dwóch przeciwstawnych wartości lingwistycznych, np. mała i duża albo niska i wysoka. Przykładowa baza reguł diagnostycznych przedstawiona jest następująco:

Tabela A.4: Zbiór reguł diagnostycznych dotyczących diagnozy złośliwości nowotworu piersi

$s_r^{(l)}$	$m_r^{(l)}$
$s_{21}^{(1)}$: jednolitość rozmiaru komórek jest niska i częstość występowania samego jądra w komórkach jest mała	0.33
$s_{23}^{(1)}$: jednolitość rozmiaru komórek jest niska i liczba prawidłowych jąder jest mała	0.34
$s_{90}^{(1)}$: jednolitość rozmiaru komórek jest niska i częstość występowania samych jąder w komórkach jest mała i liczba prawidłowych jąder w komórkach jest mała	0.33
$s_2^{(2)}$: jednolitość rozmiaru komórek jest wysoka	0.33
$s_3^{(2)}$: jednolitość kształtu komórek jest wysoka	0.34
$s_6^{(2)}$: częstość występowania samych jąder jest duża	0.33

Wartości lingwistyczne występujące w regułach są opisane odpowiednich trójkątnych funkcji przynależności z rysunku A.5a. Kształty zbiorów rozmytych uzyskanych dla innych funkcji przynależności przedstawiono na rysunku A.5b–A.6b.

A.1.6 *Cleveland heart disease*

Ostatnia z testowanych baz dotyczy diagnozy chorób serca [66]. Trzydzieści symptomów w bazie *Cleveland heart disease* ma następujący opis (przy opisie symptomu zawarto informacje o typie wartości: N, gdy nominalna, R, gdy rzeczywista, L gdy logiczna):

- wiek (R),
- płeć (L): 0 – kobieta, 1 – mężczyzna,
- typ bólu w klatce piersiowej (N): 1 – typowa dusznica piersiowa, 2 – nietypowa dusznica piersiowa, 3 – inny ból, 4 – brak objawu,
- ciśnienie spoczynkowe (R),
- poziom cholesterolu (R),
- poziom cukru (L): 0 – w normie, 1 – podwyższony,
- wynik EKG spoczynkowego (N): 0 - normalny, 1 – anomalie odcinka ST-T, 2 – podejrzenie hipertrofii lewej komory,
- maksymalne zarejestrowane tętno (R),
- dusznica wywołana ćwiczeniami (L): 0 – nie, 1 – tak,
- obniżenie załamka ST wywołane ćwiczeniami (R),
- nachylenie załamka ST podczas ćwiczeń (N): 1 – narastające, 2 – płaskie, 3 – opadające,

12. liczba naczyń powiększonych (R),
13. rezultat znakowania izotopowego (N): 3 – normalny, 6 – stała wada, 7 - odwracalna wada.

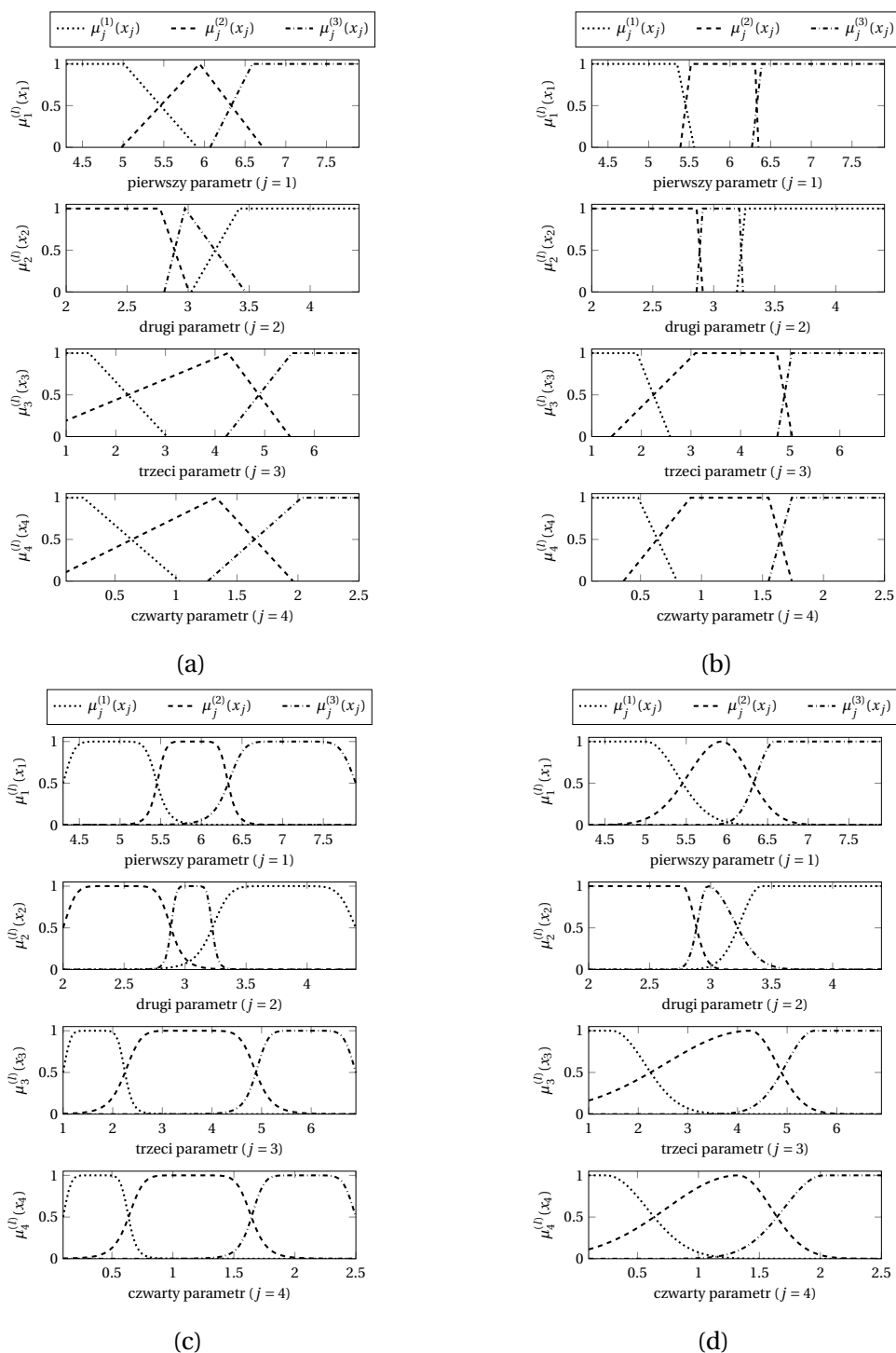
Przykładowa baza reguł diagnostycznych uzyskana dla bazy w części badawczej może być opisana w następujący sposób:

Tabela A.5: Zbiór reguł diagnostycznych dotyczących diagnozy choroby serca

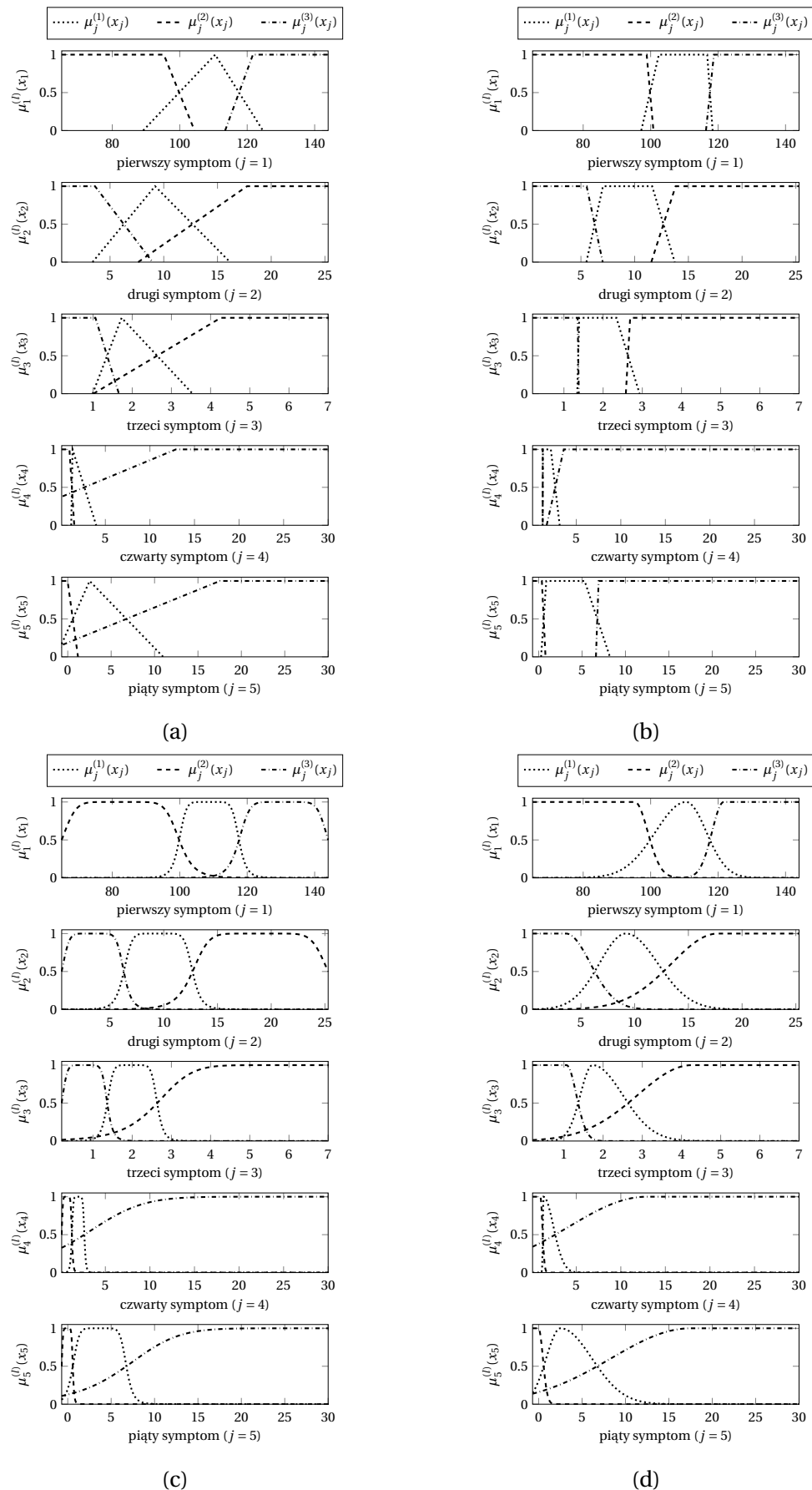
$s_r^{(l)}$	$m_r^{(l)}$
$s_{13}^{(1)}$: wynik badania talem jest niski	0.21
$s_{45}^{(1)}$: wartość nominalna bólu w klatce jest niska i liczba naczyń powiększonych jest niska	0.19
$s_{84}^{(1)}$: NIE wystąpiła dusznica i liczba naczyń powiększonych jest niska	0.21
$s_{85}^{(1)}$: NIE wystąpiła dusznica i wartość nominalna wyniku badania talem jest niska	0.19
$s_{91}^{(1)}$: liczba naczyń powiększonych jest niska i wartość nominalna wyniku badania talem jest niska	0.20
$s_3^{(2)}$: wartość nominalna typu bólu w klatce jest wysoka	0.29
$s_{13}^{(2)}$: wartość nominalna wyniku badania talem jest wysoka	0.28
$s_{45}^{(2)}$: wartość nominalna oznaczenia bólu w klatce jest wysoka i liczba naczyń powiększonych jest duża	0.20
$s_{46}^{(2)}$: wartość nominalna oznaczenia bólu w klatce jest wysoka i wartość nominalna wyniku badania talem jest wysoka	0.22

Kształty funkcji przynależności zaprezentowane są na rysunkach A.9–A.12.

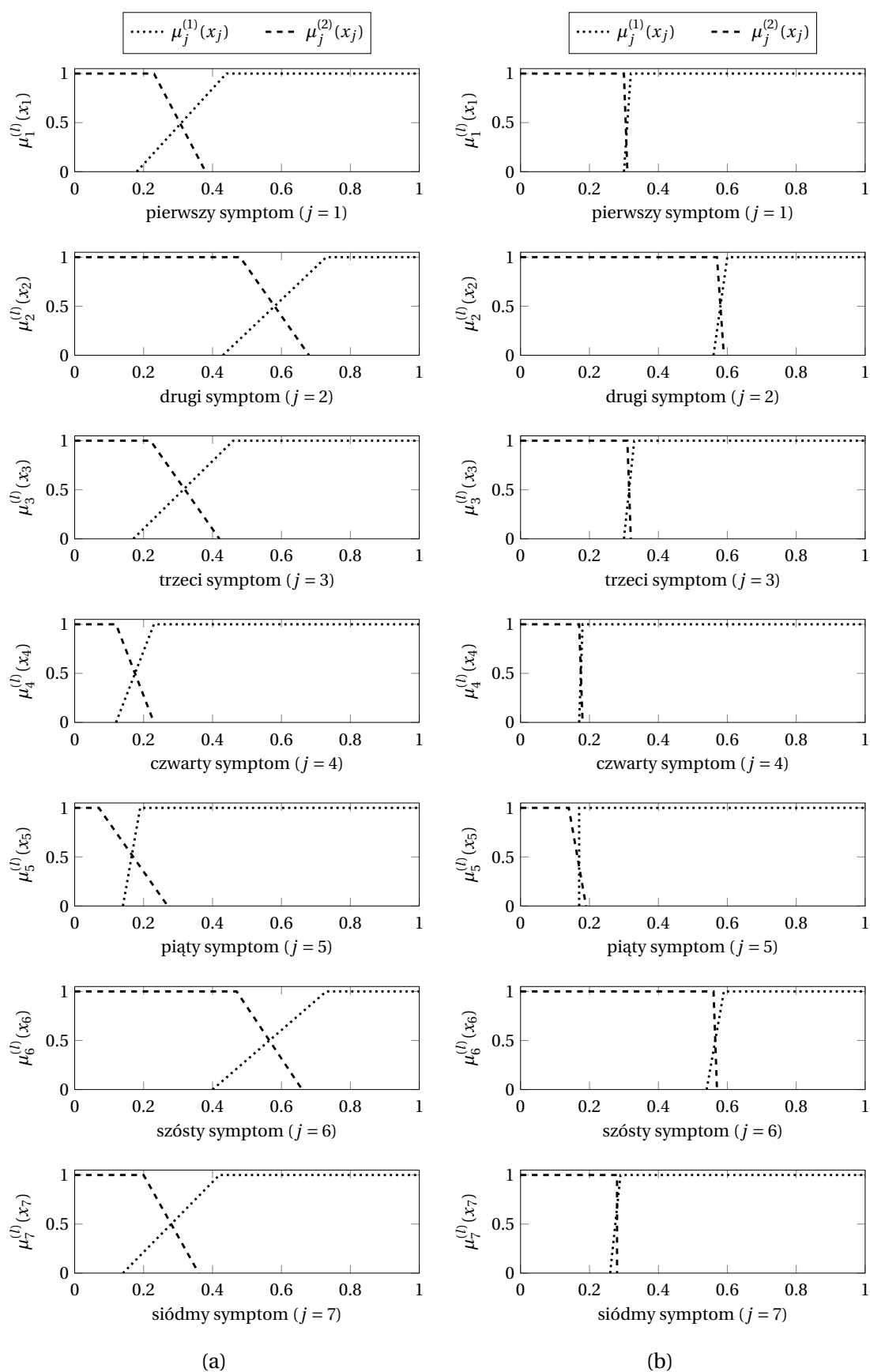
A.2 Funkcje przynależności

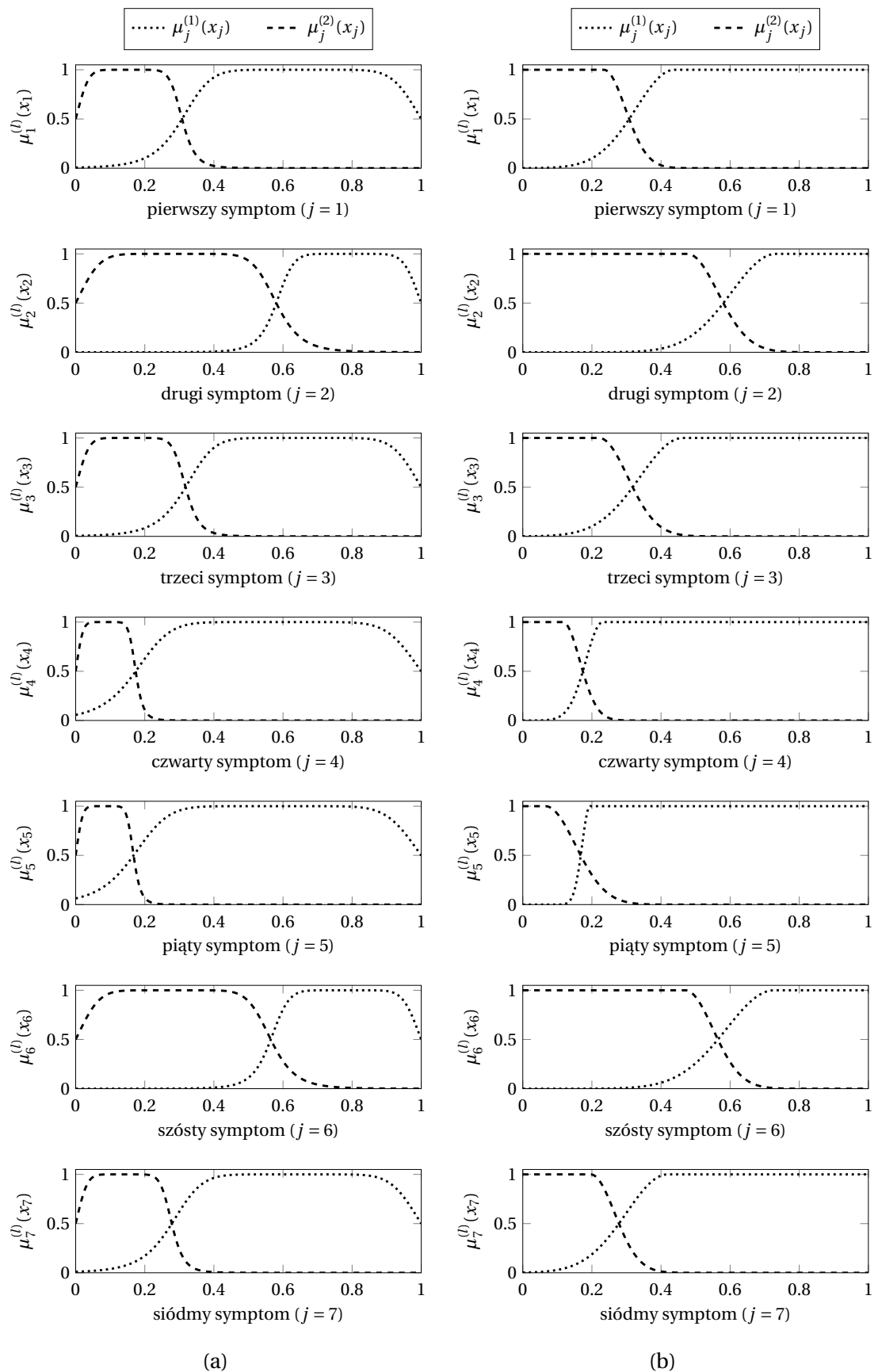


Rysunek A.1: Funkcje przynależności dla bazy *iris flowers*: (a) trójkątne, (b) trapezowe, (c) uogólnione dzwonowe, (d) gaussowskie łączone

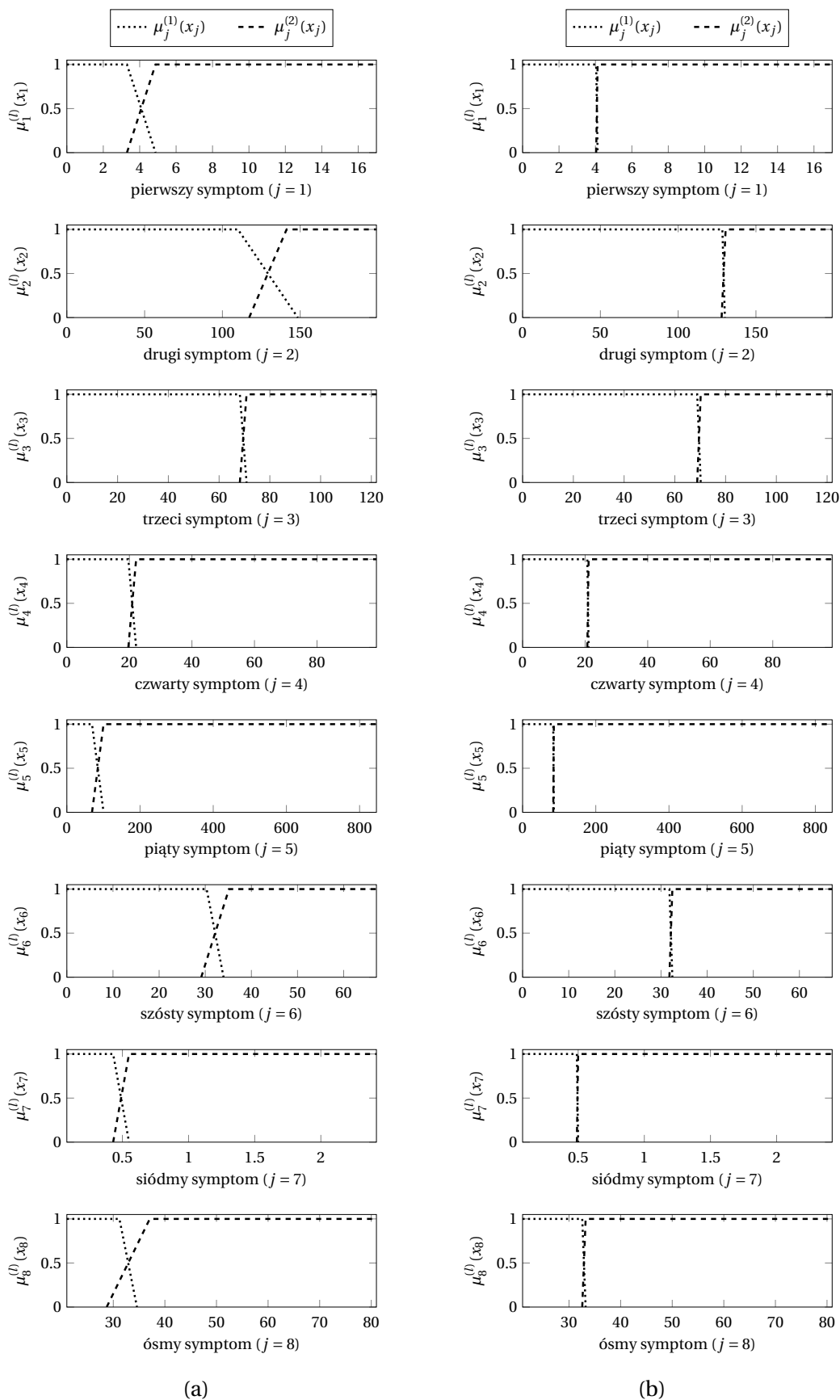


Rysunek A.2: Funkcje przynależności dla bazy *thyroid disease*: (a) trójkątne, (b) trapezowe, (c) uogólnione dzwonowe, (d) gaussowskie łączone

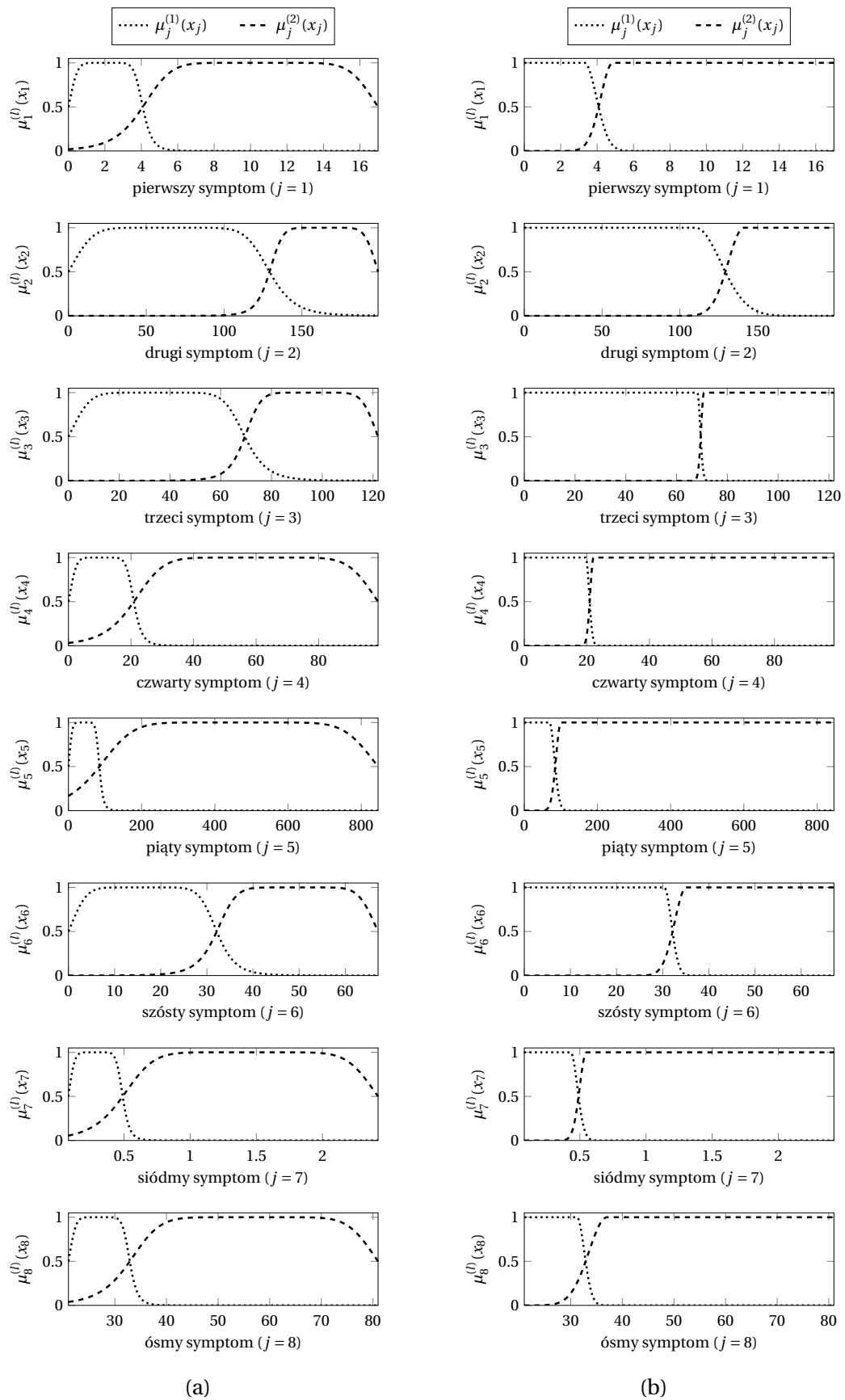
Rysunek A.3: Funkcje przynależności dla bazy *appendicitis*: (a) trójkątne, (b) trapezowe



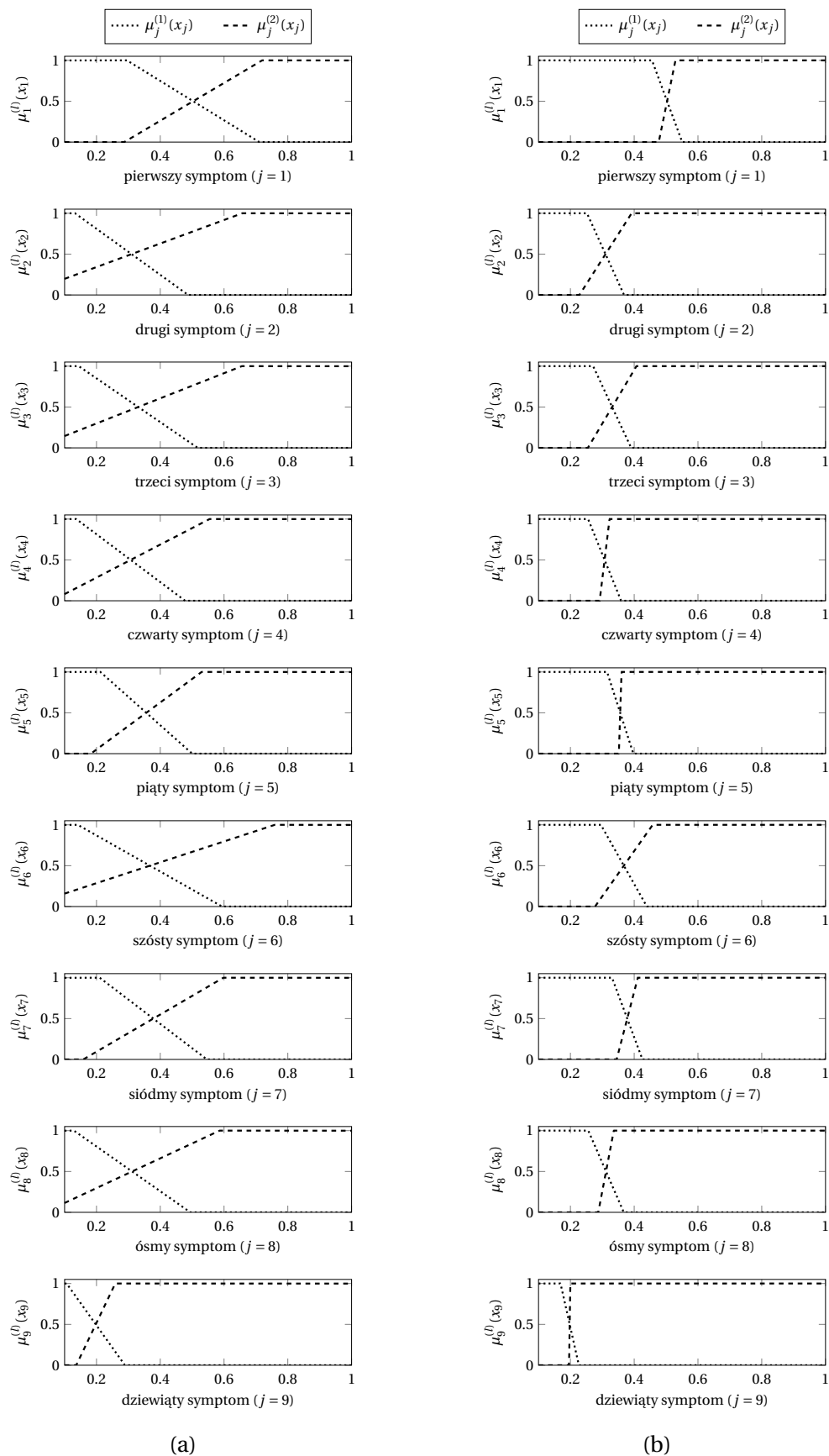
Rysunek A.4: Funkcje przynależności dla bazy *appendicitis*: (a) uogólnione dzwonowe, (b) gaussowskie łączone



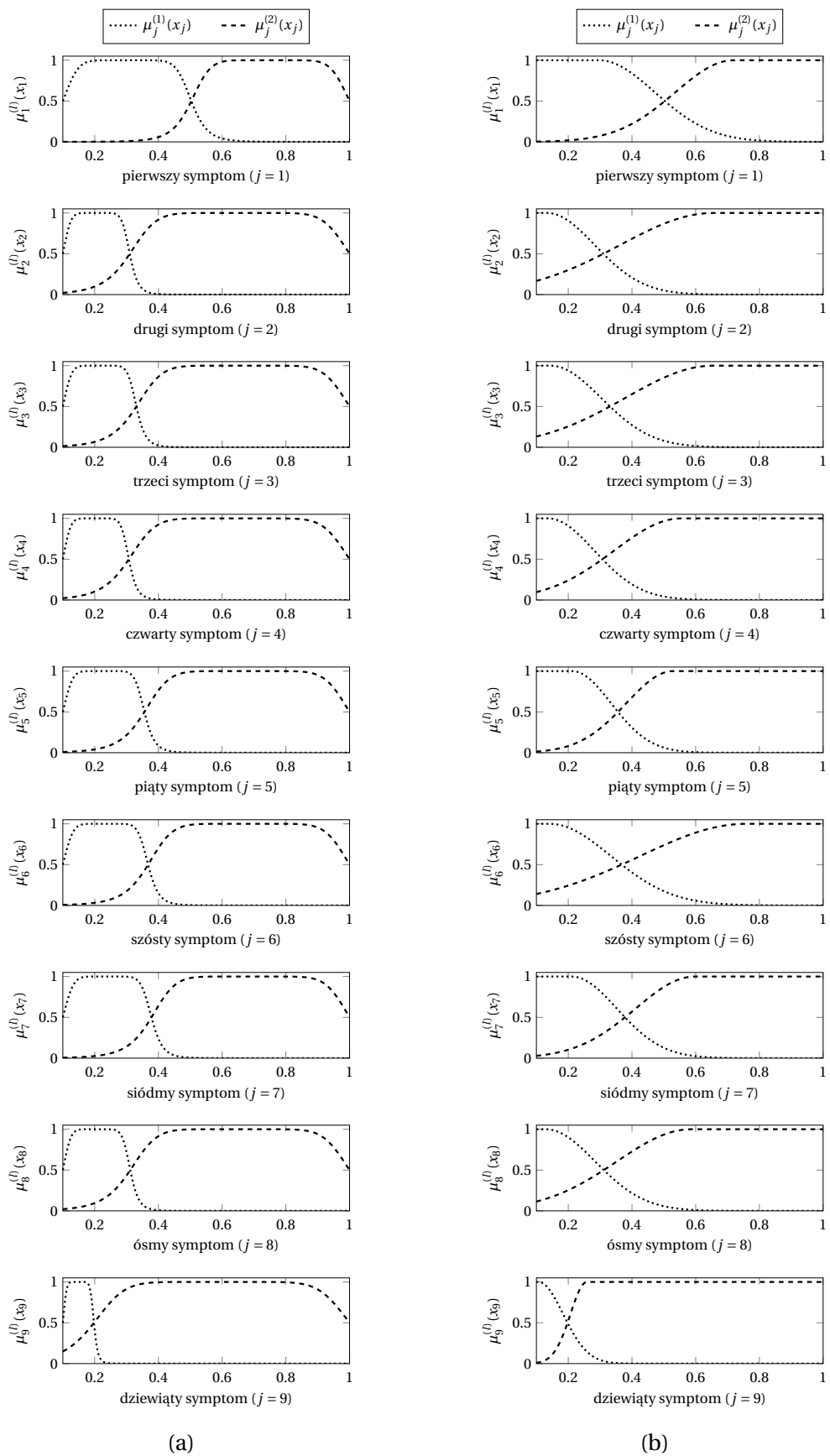
Rysunek A.5: Funkcje przynależności dla bazy *Pima Indians diabetes*: (a) trójkątne, (b) trapezowe



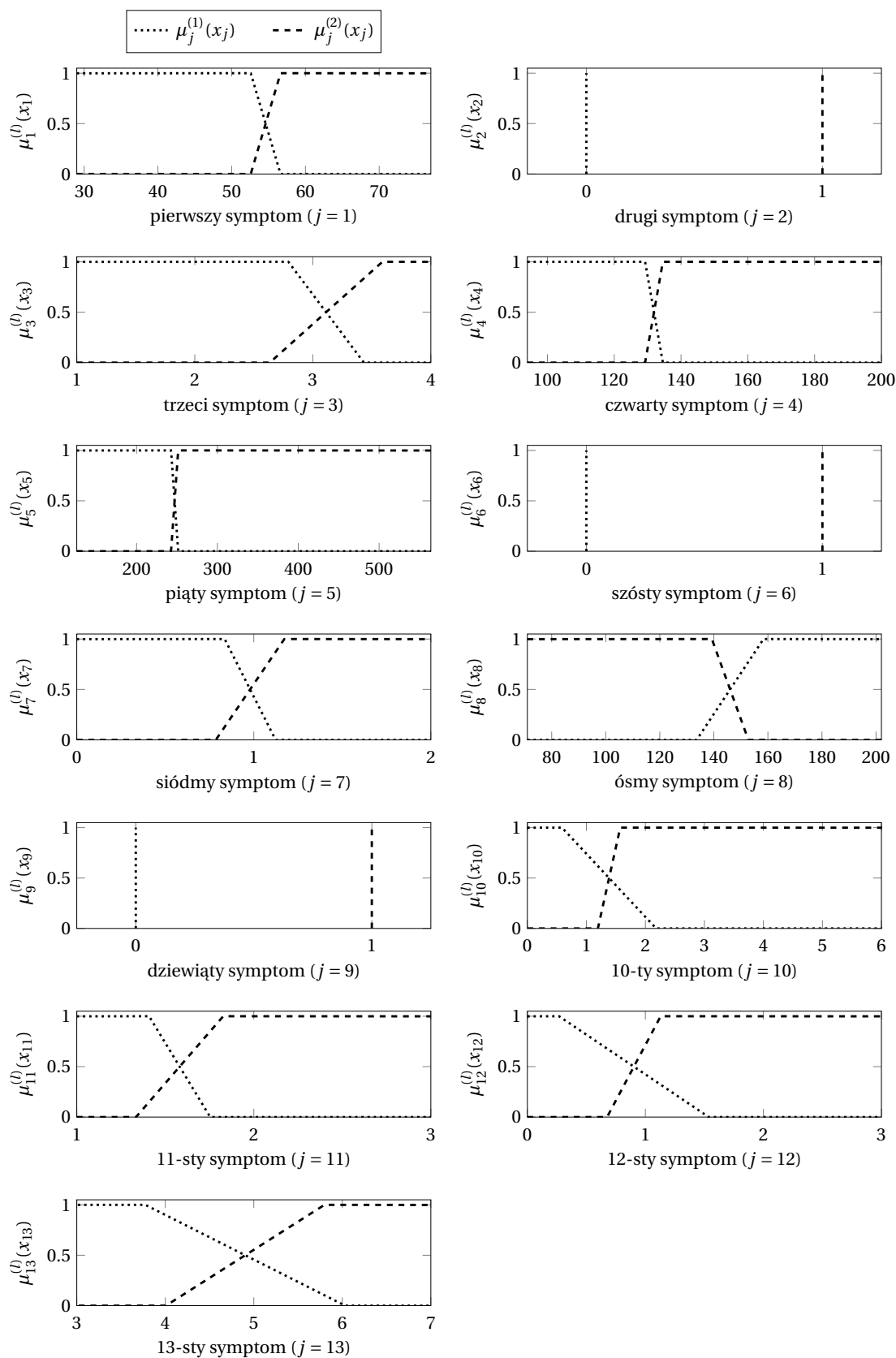
Rysunek A.6: Funkcje przynależności dla bazy *Pima Indians diabetes*: (a) uogólnione dzwonowe, (b) gaussowskie łączone

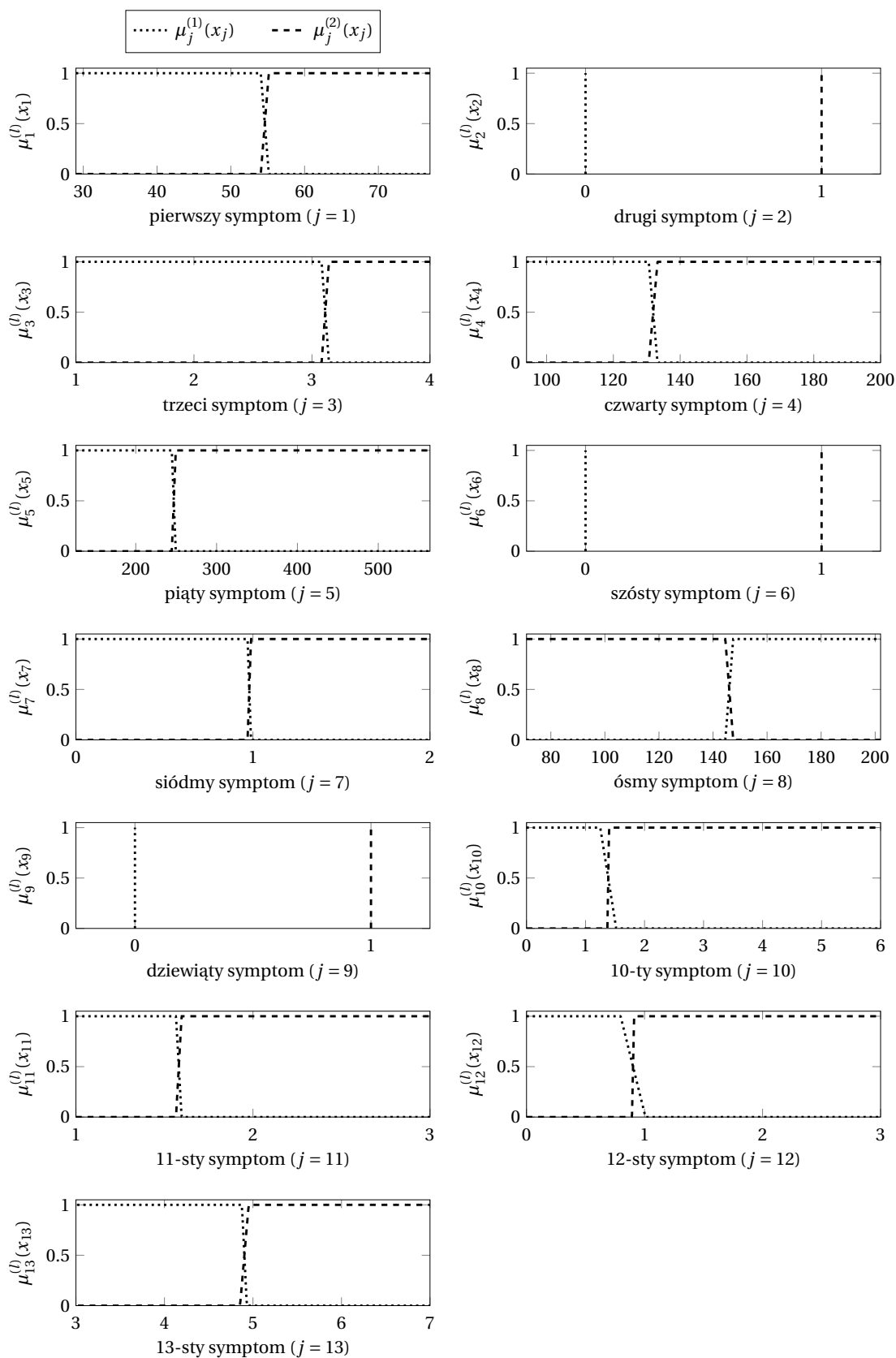


Rysunek A.7: Funkcje przynależności dla bazy *Wisconsin breast cancer*: (a) trójkątne, (b) trapezowe

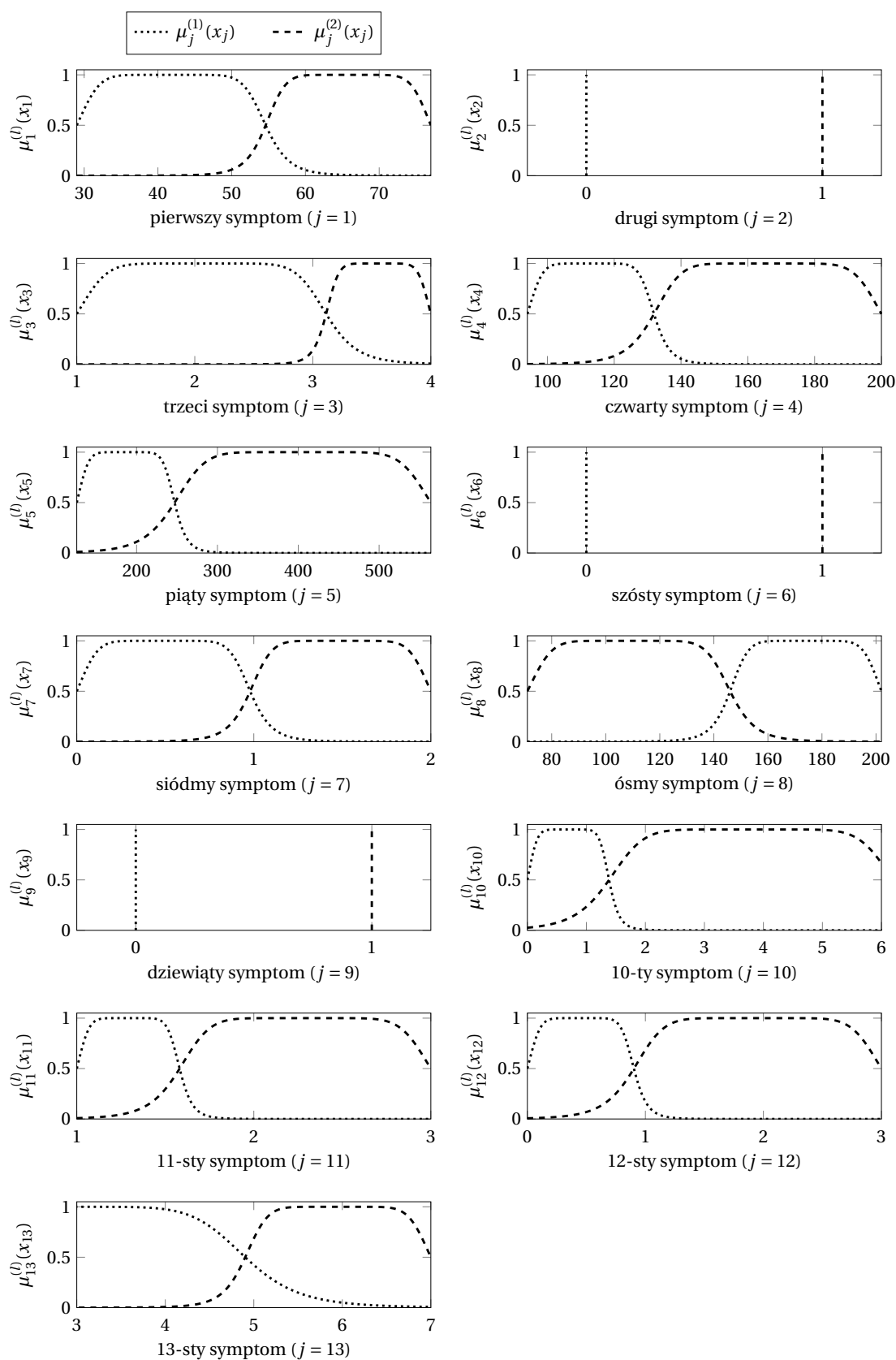


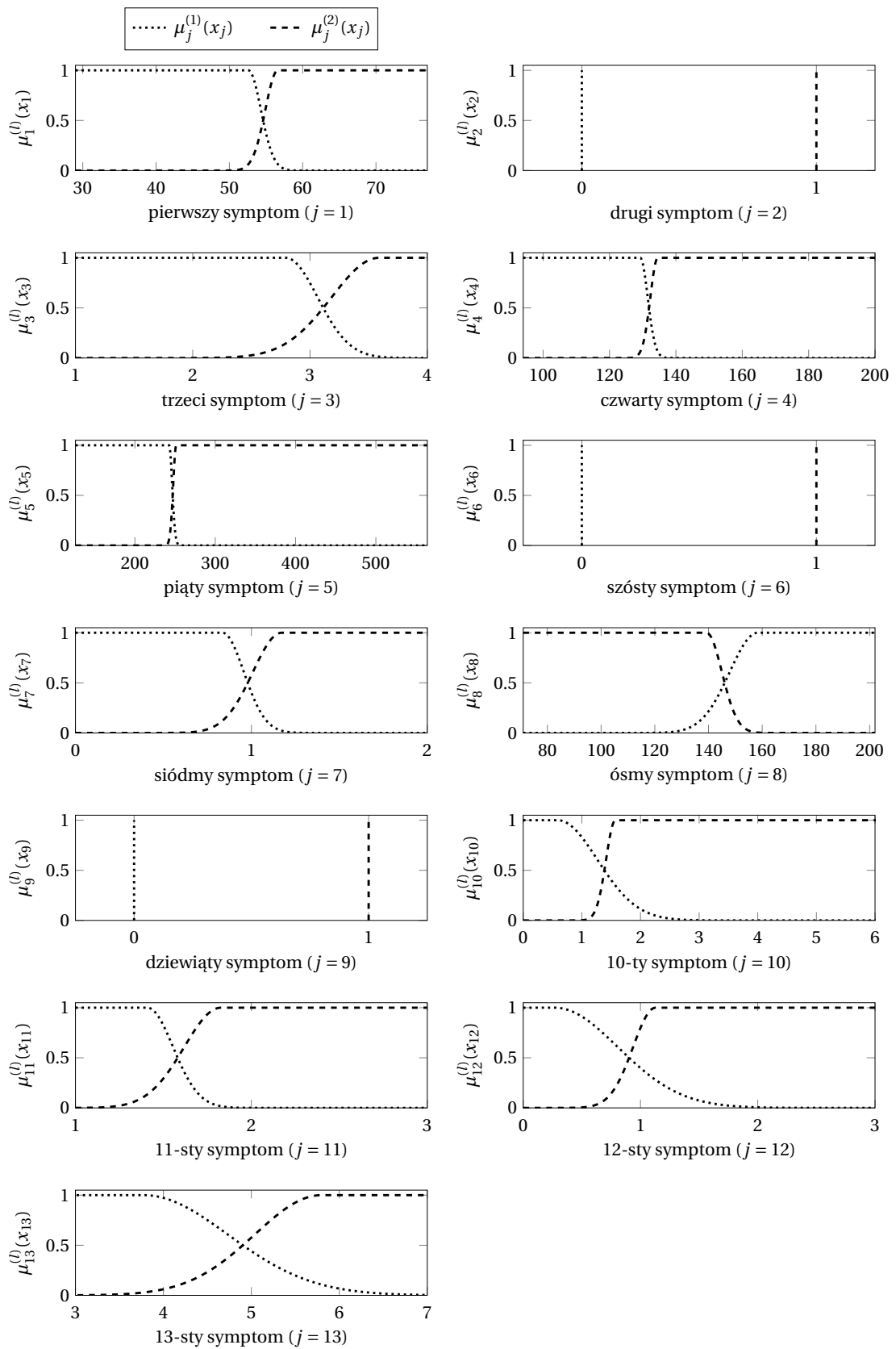
Rysunek A.8: Funkcje przynależności dla bazy *Wisconsin breast cancer*: (a) uogólnione dzwonowe, (b) gaussowskie łączone

Rysunek A.9: Trójkątne funkcje przynależności dla bazy *Cleveland heart disease*



Rysunek A.10: Trapezowe funkcje przynależności dla bazy *Cleveland heart disease*

Rysunek A.11: Uogólnione dzwonowe funkcje przynależności dla bazy *Cleveland heart disease*



Rysunek A.12: Gaussowskie łączone funkcje przynależności dla bazy *Cleveland heart disease*

Literatura

- [1] C. C. Aggarwal. *Outlier Analysis*, s. 237–263. Springer International Publishing, Cham, 2015.
- [2] R. Alcalá, Y. Nojima, F. Herrera, H. Ishibuchi. Multiobjective genetic fuzzy rule selection of single granularity-based fuzzy classification rules and its interaction with the lateral tuning of membership functions. *Soft Computing*, 15(12):2303–2318, 2011.
- [3] J. Alcalá-Fdez, R. Alcalá, F. Herrera. A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. *IEEE Transactions on Fuzzy Systems*, 19(5):857–872, 2011.
- [4] J. Alcalá-Fdez, L. Sánchez, S. García, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, J. C. Fernández, F. Herrera. KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 13(3):307–318, Feb 2009.
- [5] S. Alkhazaleh, A. R. Salleh. Fuzzy soft multiset theory. *Abstract and Applied Analysis*, 2012.
- [6] C. Andrieu, A. Doucet, E. Punskeya. *Sequential Monte Carlo methods for optimal filtering*, s. 79–95. Springer, New York, 2001.
- [7] S. Arlot, A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.
- [8] K. T. Atanassov. Intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, 20(1):87–96, 1986.
- [9] D. Ayres-de Campos, C. Y. Spong, E. Chandrachan. FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography. *International Journal of Gynecology & Obstetrics*, 131(1):13 – 24, 2015.
- [10] J. Bacardit, N. Krasnogor. Performance and efficiency of memetic Pittsburgh learning classifier systems. *Evolutionary Computation*, 17(3):307–342, 2009.
- [11] E. Bańkowski. *Biochemia: Podręcznik dla studentów uczelni medycznych – Wydanie trzecie*. Edra Urban & Partner, Wrocław, 2016.
- [12] N. Barakat, A. P. Bradley, M. N. H. Barakat. Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Transactions on Information Technology in Biomedicine*, 14(4):1114–1120, 2010.
- [13] J. A. Barnett. *Computational methods for a mathematical theory of evidence*, s. 197–216. Springer, Berlin, Heidelberg, 2008.
- [14] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994.

- [15] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:370–418, 1763.
- [16] B. A. Berg. *Markov chain Monte Carlo simulations and their statistic analysis*. World Scientific, Singapore, 2004.
- [17] F. J. Berlanga, A. J. Rivera, M. J. del Jesus, F. Herrera. GP-COACH: Genetic programming based learning of compact and accurate fuzzy rule based classification systems for high dimensional problems. *Information Sciences*, 180(8):1183–1200, 2010.
- [18] E. S. Berner. *Clinical Decision Support Systems: Theory and Practice*. Springer Publishing Company, Incorporated, 2nd edition, 2010.
- [19] E. Binaghi, P. Madella. Fuzzy Dempster-Shafer reasoning for rule-based classifiers. *Intelligent Systems*, 14:559–583, 1999.
- [20] E. Binaghi, P. Madella. A neural model for fuzzy Dempster-Shafer classifiers. *International Journal of Approximate Reasoning*, 25:89–121, 2000.
- [21] L. Bolc, W. Borodziejewicz, M. Wójcik. *Podstawy przetwarzania informacji niepewnej i niepełnej*. Współczesna Nauka i Technika: Informatyka. Państwowe Wydawnictwo Naukowe, 1991.
- [22] N. R. Cazarez-Castro, L. T. Aguilar, O. Castillo. Fuzzy logic control with genetic membership function parameters optimization for the output regulation of a servomechanism with nonlinear backlash. *Expert Systems with Applications*, 37(6):4368–4378, 2010.
- [23] B. Chandra, P. P. Varghese. Fuzzifying Gini Index based decision trees. *Expert Systems with Applications*, 36:8549–8559, 2009.
- [24] G. Chandrashekar, F. Sahin. A survey on feature selection methods. *Computers and Electrical Engineering*, 40:16–28, 2014.
- [25] C. Cornelis, G. Deschrijver, E. E. Kerre. Implication in intuitionistic fuzzy and interval-valued fuzzy set theory: construction, classification, application. *International Journal of Approximate Reasoning*, 35(1):55–95, 2004.
- [26] N. Cowan. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, 24(1):87–114, 2001.
- [27] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.
- [28] A. P. Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):205–247, 1968.
- [29] T. Denoeux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(5):804–813, 1995.
- [30] T. Denoeux. A neural network classifier based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(2):131–150, 2000.
- [31] D. Dubois. Possibility theory and statistical reasoning. *Computational Statistics & Data Analysis*, 51(1):47–69, 2006.
- [32] D. Dubois, H. Prade. Ranking fuzzy numbers in the setting of possibility theory. *Information Sciences*, 30(3):183–224, 1983.

- [33] D. Dubois, H. Prade. Random sets and fuzzy interval analysis. *Fuzzy Sets and Systems*, 42(1):87–101, 1991.
- [34] W. Duch, R. Adamczak, K. Grabczewski. A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. *IEEE Transactions on Neural Networks*, 12(2):277–306, 2001.
- [35] W. Duch, R. Setiono, J. M. Żurada. Computational intelligence methods for rule-based data understanding. *Proceeding of the IEEE*, 92(5):771–805, 2004.
- [36] L. Dymova, P. Sevastianov, P. Bartosiewicz. A new approach to the rule-base evidential reasoning: Stock trading expert system application. *Expert Systems with Applications*, 37(8):5564–5576, 2010.
- [37] ESH & ESC. European Society of Hypertension & European Society of Cardiology guidelines for the management of arterial hypertension. *Journal of Hypertension*, 21:1011–1053, 2003.
- [38] N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, V. K. Tabar. Knowledge discovery in medicine: current issue and future trend. *Expert Systems with Applications*, 41(9):4434–4463, 2014.
- [39] F. Feng. Generalized rough fuzzy sets based on soft sets. w: *International Workshop on Intelligent Systems and Applications*, s. 1–4, 2009.
- [40] E. Frank, M. Hall, I. H. Witten. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, Fourth Edition, 2016.
- [41] N. Frulio, H. Trillaud. Ultrasound elastography in liver. *Diagnostic and Interventional Imaging*, 94(5):515–534, 2013.
- [42] M. J. Gacto, R. Alcalá, F. Herrera. Interpretability of linguistic fuzzy rule-based systems: an overview of interpretability measures. *Information Sciences*, 181:4340–4360, 2011.
- [43] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, F. Herrera. Dynamic classifier selection for one-vs-one strategy: Avoiding non-competent classifiers. *Pattern Recognition*, 46(12):3412–3424, 2013.
- [44] D. Garcia, A. Gonzalez, R. Perez. Overview of the slave learning algorithm: A review of its evolution and prospects. *International Journal of Computational Intelligence Systems*, 7(6), 2014.
- [45] W. L. Gau, D. J. Buehrer. Vague sets. *Transactions on Systems, Man, and Cybernetics: Systems*, 23(2):610–614, 1993.
- [46] I. Guyon, A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [47] P. Halfon, M. Bourliere, R. Deydier, D. Botta-Fridlund, C. Renou, A. Tran, I. Portal, I. Allemand, J. J. Bertrand, A. Rosenthal-Allieri, M. Rotily, C. Sattonet, T. Benderitter, M. C. S. Paul, H. P. Bonnot, G. Penaranda, C. Degott, M. F. Masseyeff, D. Ouzan. Independent Prospective Multicenter Validation of Biochemical Markers (Fibrotest–Actitest) for the Prediction of Liver Fibrosis and Activity in Patients with Chronic Hepatitis C: The Fibropaca Study. *The American Journal Of Gastroenterology*, 101:547, 2006.

- [48] L. Han, S. Luo, J. Yu, L. Pan, S. Chen. Rule extraction from support vector machines using ensemble learning approach: An application for diagnosis of diabetes. *IEEE Journal of Biomedical and Health Informatics*, 19(2):728–734, 2015.
- [49] W. He, N. Williard, M. Osterman, M. Pecht. Prognostics of lithium-ion batteries based on Dempster-Shafer theory and the Bayesian Monte Carlo method. *Journal of Power Sources*, 196(23):10314–10321, 2011.
- [50] J. Hühn, E. Hüllermeier. FURIA: an algorithm for unordered fuzzy rule induction. *Data Mining and Knowledge Discovery*, 19(3):293–319, 2009.
- [51] X. Hong, C. Nugent, M. Mulvenna, S. McClean, B. Scotney, S. Delvin. Evidential fusion of sensor data for activity recognition in smart homes. *Pervasive and Mobile Computing*, 5(3):236–252, 2009.
- [52] The Franklin Institute. Lotfi Zadeh (materiał wideo), 2009. <https://www.fi.edu/laureates/lotfi-zadeh>, dostęp 05-02-2019.
- [53] H. Ishibuchi, Y. Nojima. Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning. *International Journal of Approximate Reasoning*, 44:4–31, 2007.
- [54] M. Ishizuka. Inference procedures under uncertainty for the problem-reduction method. *Information Sciences*, 28(3):179–206, 1982.
- [55] M. Jeżewski, R. Czabański, J. Łęski. The influence of cardiocogram signal feature selection method on fetal state assessment efficacy. *Journal of Medical Informatics & Technologies*, 23:51–58, 2014.
- [56] B. Jin, Y. C. Tang, Y.-Q. Zhang. Support vector machines with genetic fuzzy feature transformation for biomedical data classification. *Information Sciences*, 177(2):476–489, 2007.
- [57] R. S. Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Johns Hopkins APL Technical Digest*, 10:262–266, 1988.
- [58] B. Khaleghi, A. Khamis, F. O. Karray, S. N. Razavi. Multisensor data fusion: a review of the state-of-the-art. *Information Fusion*, 14(1):28–44, 2013.
- [59] W. Ray Kim, T. Berg, T. Asselah, R. Flisiak, S. Fung, S. C. Gordon, H. L. A. Janssen, P. Lampertico, D. Lau, J. D. Bornstein, R. E. Aguilar Schall, P. Dinh, L. J. Yee, E. B. Martins, S. G. Lim, R. Loomba, J. Petersen, M. Buti, P. Marcellin. Evaluation of apri and fib-4 scoring systems for non-invasive assessment of hepatic fibrosis in chronic hepatitis b patients. *Journal of Hepatology*, 64(4):773–780, 2016.
- [60] J. Klir, G. On fuzzy-set interpretation of possibility theory. *Fuzzy Sets and Systems*, 108:263–273, 1999.
- [61] M. A. Kłopotek. *Metody identyfikacji i interpretacje struktur rozkładów przekonań w teorii Dempstera-Shafera*. Instytut Podstaw Informatyki Polskiej Akademii Nauk, 1998.
- [62] R. Kohli, R. Krishnamurti, K. Jedidi. Subset-conjunctive rules for breast cancer diagnosis. *Discrete Applied Mathematics*, 154(7):1100–1112, 2006.
- [63] J. Koronacki, J. Ćwik. *Statystyczne systemy uczące się. Wydanie drugie*. Akademicka Oficyna Wydawnicza EXIT Andrzej Lang, 2015.

- [64] B. Krawczyk, M. Woźniak, T. Orczyk, P. Porwik. Cost sensitive hierarchical classifiers for non-invasive recognition of liver fibrosis stage. w: R. Burduk, K. Jackowski, M. Kurzyński, M. Woźniak, A. Żołnierek, (edytorzy), *Proceedings of the 8th International Conference of Computer Recognition Systems CORES 2013*, s. 639–647, Heidelberg, 2013. Springer International Publishing.
- [65] H. K. Lam. Polynomial fuzzy-model-based control systems: Stability analysis via piecewise-linear membership functions. *IEEE Transactions on Fuzzy Systems*, 19(3):588–593, June 2011.
- [66] M. Lichman. UCI machine learning repository <http://archive.ics.uci.edu/ml>, 2013. Dostęp: 16.02.2017.
- [67] X. D. Liu, X. Feng, W. Pedrycz. Extraction of fuzzy rules from fuzzy decision trees: An axiomatic fuzzy sets (afs) approach. *Data & Knowledge Engineering*, 84:1–25, 2013.
- [68] X. D. Liu, W. Pedrycz. The development of fuzzy decision trees in the framework of axiomatic fuzzy set logic. *Applied Soft Computing*, 7:325–342, 2007.
- [69] J. Łęski. *Systemy neuronowo-rozmyte*. Wydawnictwa Naukowo-Techniczne, Warszawa, 2008.
- [70] J. Łęski. Fuzzy $(c+p)$ -means clustering and its application to a fuzzy rule-based classifier: Toward good generalization and good interpretability. *IEEE Transactions of Fuzzy Systems*, 23(4):802–812, 2015.
- [71] R. P. S. Mahler. *Statistical Multisource-Multitarget Information Fusion*. Artech House, Inc., Norwood, MA, USA, 2007.
- [72] E. H. Mamdani, S. Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, 7(1):1–13, 1975.
- [73] W. L. Martinez, A. R. Martinez. *Computational Statistics Handbook with MATLAB, Third Edition*. Chapman & Hall/CRC Computer Science & Data Analysis. CRC Press, 2015.
- [74] Richard S Mathis. The Impacts of Innovation. *Science*, 328(5978):570, 2010.
- [75] B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2):442–451, 1975.
- [76] D. Meng, X. Zhang, K. Qin. Soft rough fuzzy sets and soft fuzzy rough sets. *Computers & Mathematics with Applications*, 62(12):4635–4645, 2011.
- [77] J. M. Merigo, M. Casanovas. Induced aggregation operators in decision making with the Dempster-Shafer belief structure. *International Journal of Intelligent Systems*, 24(8):934–954, 2009.
- [78] J. M. Merigo, M. Casanovas, L. Martinez. Linguistic aggregation operator for linguistic decision making based on the Dempster-Shafer theory of evidence. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 18(03):287–304, 2010.
- [79] G. Miller. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63:81–97, 1956.
- [80] S. Mitra, Y. Hayashi. Neuro-fuzzy rule generation: survey in soft computing framework. *IEEE Transactions on Neural Networks*, 11(3):748–768, 2000.

- [81] D. Molodtsov. Soft set theory – First results. *Computer & Mathematics with Applications*, 37(4-5):19–31, 1999.
- [82] R. Moore, R. Kearfott, M. Cloud. *Introduction to Interval Analysis*. Society for Industrial and Applied Mathematics, 2009.
- [83] J. G. Moreno-Torres, J. A. Saez, F. Herrera. Study on the impact of partition-induced dataset shift on k -fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1304–1312, 2012.
- [84] S. Nakariyakul, D. P. Casasent. An improvement on floating search algorithms for feature subset selection. *Pattern Recognition*, 42(9):1932–1940, 2009.
- [85] Y. Ngo, Y. Benhamou, V. Thibault, P. Ingiliz, M. Munteanu, P. Lebray, D. Thabut, R. Morra, D. Messous, F. Charlotte, F. Imbert-Bismut, D. Rousselot-Bonnefont, J. Moussalli, V. Ratziu, T. Poynard. An accurate definition of the status of inactive hepatitis B virus carrier by a combination of biomarkers (FibroTest-ActiTest) and viral load. *PLoS ONE*, 3(7), 2008.
- [86] Y. Ngo, M. Munteanu, D. Messous, F. Charlotte, F. Imbert-Bismut, D. Thabut, P. Lebray, V. Thibault, Y. Benhamou, J. Moussalli, V. Ratziu, T. Poynard. A prospective analysis of the prognostic value of biomarkers (FibroTest) in patients with chronic hepatitis C. *Clinical Chemistry*, 52(10):1887–1896, 2006.
- [87] H. Ogawa, K. S. Fu, J. T. P. Yao. An inexact inference for damage assessment of existing structures. *International Journal of Man-Machine Studies*, 22(3):295–306, 1985.
- [88] T. Orczyk, P. Porwik. Investigation of the impact of missing value imputation methods on the k -nn classification accuracy. w: M. Núñez, N. T. Nguyen, D. Camacho, B. Trawiński, (edytorzy), *Computational Collective Intelligence: 7th International Conference, ICCCI 2015, Madrid, Spain, September 21-23, 2015, Proceedings, Part II*, s. 557–565, Cham, 2015. Springer International Publishing.
- [89] A. Panasiuk. Metody oceny zapalenia, zwłóknienia i stłuszczenia wątroby. *Hepatologia*, 14:110–114, 2014.
- [90] Z. Pawlak. *Rough sets: theoretical aspects of reasoning about data*. Springer Netherlands, 1991.
- [91] Z. Pawlak. Rough set theory and its applications to data analysis. *Cybernetics and Systems*, 29(7):661–688, 1998.
- [92] W. Pedrycz. Shadowed sets: representing and processing fuzzy sets. *IEEE Transactions on Systems, Man and Cybernetics. Part B*, 28(1):103–109, 1998.
- [93] A. Piegat. *Modelowanie i sterowanie rozmyte*. EXIT, Warszawa, 1999.
- [94] **S. Porębski**, E. Straszecka. Membership functions for fuzzy focal elements. *Archives of Control Sciences*, 26(3):281–313, 2016.
- [95] **S. Porębski**, E. Straszecka. Extracting easily interpreted diagnostic rules. *Information Sciences*, 426:19–37, 2017.
- [96] **S. Porębski**, E. Straszecka. Diagnostic rule extraction using the Dempster-Shafer theory extended for fuzzy focal elements. w: M. Kurzyński, M. Woźniak, R. Burduk, (edytorzy), *Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017*, s. 63–72, Cham, 2018. Springer International Publishing.

- [97] P. Porwik, T. Orczyk, M. Lewandowski, M. Cholewa. Feature projection k-nn classifier model for imbalanced and incomplete medical data. *Biocybernetics and Biomedical Engineering*, 36(4):644–656, 2016.
- [98] D. M. W. Powers. Evaluation from precision recall and f-measure to ROC informedness markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.
- [99] J. R. Quinlan, P. J. Compton, K. A. Horn, L. Lazarus. Inductive knowledge acquisition: a case study. w: *Proceedings of the Second Australian Conference on Applications of Expert Systems*, s. 137–156, Boston, MA, USA, 1987. Addison-Wesley Longman Publishing Co., Inc.
- [100] P. Refaeilzadeh, Lei Tang, Huan Liu. *Cross-Validation*, s. 532–538. Springer US, Boston, MA, 2009.
- [101] P. Sevastianov. Numerical methods for interval and fuzzy number comparison based on the probabilistic approach and Dempster-Shafer theory. *Information Sciences*, 177(21):4645–4661, 2007.
- [102] P. Sevastianov, L. Dymova, P. Bartosiewicz. A framework for rule-base evidential reasoning in the interval setting applied to diagnosing type 2 diabetes. *Expert Systems with Applications*, 39(4):4190–4200, 2012.
- [103] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, New Jersey, 1976.
- [104] H. Singh, M. Gupta, T. Meitzler, Z.-G. Hou, K. Garg, A. Solo, L. A. Zadeh. Real-life applications of fuzzy logic. *Advances in Fuzzy Systems*, 2013(581879), 2013.
- [105] P. Smets. *Imperfect Information: Imprecision and Uncertainty*, s. 225–254. Springer US, Boston, MA, 1997.
- [106] P. Smets, R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.
- [107] E. Straszecka. An interpretation of focal elements as fuzzy sets. *International Journal of Intelligent Systems*, 18(7):821–835, 2003.
- [108] E. Straszecka. Combining uncertainty and imprecision in models of medical diagnosis. *Information Sciences*, 176(20):3026–3059, 2006.
- [109] E. Straszecka. *Matching Knowledge and Evidence in a Model of Medical Diagnosis*, s. 429–436. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
- [110] E. Straszecka. Combining knowledge from different sources. *Expert Systems*, 27(1):40–52, 2010.
- [111] E. Straszecka. *Measures of uncertainty and imprecision in medical diagnosis support*. Wydawnictwo Politechniki Śląskiej, Gliwice, 2010.
- [112] E. Straszecka, J. Straszecka. *Membership functions as combination of expert's knowledge with population information*, s. 322–327. Physica-Verlag HD, Heidelberg, 2003.
- [113] T. Sudkamp. The semantics of plausibility and possibility. *International Journal of General Systems*, 21(3):273–289, 1992.
- [114] M. Sugeno. *Theory of fuzzy integrals and its applications*. Ph.D. thesis. Tokyo Institute of Technology, Tokyo, Japan, 1974.

- [115] R. Tadeusiewicz, A. Izworski, J. Majewski. *Biometria*. Wydawnictwa AGH, Kraków, 1993.
- [116] M. G. Tsipouras, T. P. Exarchos, D. I. Fotiadis, A. P. Kotsia, K. V. Vakalis, K. K. Naka, L. K. Michalis. Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. *IEEE Transactions on Information Technology in Biomedicine*, 12(4):447–458, 2008.
- [117] M. Vatankhah, V. Asadpour, R. Fazel-Rezai. Perceptual pain classification using AN-FIS adapted RBF kernel support vector machine for therapeutic usage. *Applied Soft Computing*, 13(5):2537–2546, 2013.
- [118] S. Wang, X. Yao. Multiclass imbalance problems: Analysis and potential solutions. *IEEE Transactions on Systems and Man and Cybernetics Part B*, 42(4):1119–1130, 2012.
- [119] Z. Wang, G. J. Klir. *Fuzzy measure theory*. Springer US, 2013.
- [120] S. M. Weiss, C. A. Kulikowski. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann Publishers Inc., San Francisco, USA, 1991.
- [121] S.T. Wierzchocń. *Metody reprezentacji i przetwarzania informacji niepewnej w ramach teorii Dempstera-Shafera*. Instytut Podstaw Informatyki Polskiej Akademii Nauk, 1996.
- [122] R. R. Yager. Generalized probabilities of fuzzy events from fuzzy belief structures. *Information Sciences*, 28(1):45–62, 1982.
- [123] R. R. Yager. On the theory of bags. *International Journal of General Systems*, 13(1):23–37, 1986.
- [124] R. R. Yager, L. Liu. *Classic Works of the Dempster-Shafer Theory of Belief Functions*, volume 219 of *Studies in Fuzziness and Soft Computing*. Springer, Berlin, 2008.
- [125] R. R. Yager, Filev D. P. Including probabilistic uncertainty in fuzzy logic controller modeling using Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(8):1221–1230, 1995.
- [126] J. B. Yang, J. Liu, J. Wang, H. S. Sii, H. W. Wang. Belief rule-base inference methodology using the evidential reasoning approach – RIMER. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 36(2):266–285, 2006.
- [127] J. B. Yang, Y. M. Wang, D. L. Xu, K. S. Chin. The evidential reasoning approach for MADA under both probabilistic and fuzzy uncertainties. *European Journal of Operational Research*, 171(1):309–343, 2006.
- [128] D.-Y. Yeh, C.-H. Cheng, Y.-W. Chen. A predictive model for cerebrovascular disease using data mining. *Expert Systems with Applications*, 38(7):8970–8977, 2011.
- [129] J. Yen. Generalizing the Dempster-Schafer theory to fuzzy sets. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(3):559–570, 1990.
- [130] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.
- [131] L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning—I. *Informations Sciences*, 8(3):199–249, 1975.
- [132] L. A. Zadeh. Possibility theory and soft data analysis. w: L. Cobb, R. M. Thrall, (edytorzy), *Mathematical Frontiers of the Social and Policy Sciences*, s. 69–129. Westview Press, 1981.

-
- [133] L. A. Zadeh. A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. *The AI Magazine*, 7(2):85–90, 1986.
- [134] L. A. Zadeh. *Fuzzy Logic = Computing with Words*, s. 3–23. Physica-Verlag HD, Heidelberg, 1999.
- [135] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 100, Supplement 1:9–34, 1999.
- [136] C. Zhang, Y. Hu, F. T. S. Chan, R. Sadiq, Y. Deng. A new method to determine basic probability assignment using core samples. *Knowledge-Based Systems*, 69:140–149, 2014.
- [137] A.-X. Zhu, L. Yang, B. Li, C. Qin, T. Pei, B. Liu. Construction of membership functions for predictive soil mapping under fuzzy logic. *Geoderma*, 155(3):164 – 174, 2010.
- [138] L. M. Zouhal, T. Denoeux. An evidence-theoretic k-nn rule with parameter optimization. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 28(2):263–271, 1998.

